ICS Part-II · Punjab Board · English Medium

Statistics 2nd Year

Complete Notes

freebooks.pk

Table of Contents

Prepared by freebooks.pk — original student notes covering the Punjab Curriculum and Textbook Board (PTB/PCTB) syllabus. Each chapter includes key concepts, definitions, formulas, diagrams, solved examples, short and long questions, MCQs, revision and exam tips.

Chapter 10: Normal Distribution

The normal distribution is the single most important probability distribution in all of statistics. It shows up everywhere -- heights and weights of individuals, IQ scores, measurement errors -- because it is also the limiting form that many other probability distributions approach, and it underlies the central limit theorem that makes so much of inferential statistics possible. This chapter opens the second half of the ICS Statistics course: everything from Chapter 11 onward (sampling distributions, estimation, hypothesis testing) leans on the normal distribution introduced here.

The chapter builds the normal probability density function and its key properties, then introduces the standard normal random variable Z -- a single reference distribution (mean 0, variance 1) that lets us answer probability questions for ANY normal distribution using one shared set of tables, through the simple transformation Z = (X - mu)/sigma.

Learning Objectives

- State the normal probability density function and identify its two parameters, mu and sigma
- List the key properties of the normal distribution: symmetry, mode, special areas, moments
- Standardize any normal random variable X into the standard normal variable Z
- Use the standard normal cumulative distribution function Phi(z) to calculate probabilities
- Find quantiles and percentiles of the standard normal distribution and de-standardize them back to X
- Solve applied problems involving normally distributed variables: scores, heights, dimensions, times
- Work backward to find mu or sigma (or both) given probability information

Key Concepts

10.1 The Normal Probability Density Function

A continuous random variable X is normally distributed if its probability density function is f(x) = [1/(sigma.sqrt(2.pi))] . e^[-(1/2)((x-mu)/sigma)^2], for -infinity < x < infinity, where mu can be any real number and sigma must be positive. A normal distribution is fully characterized by just these two parameters -- mu (its mean) and sigma (its standard deviation) -- and is commonly written X ~ N(mu, sigma-squared).

The graph of this function is called a normal curve: unimodal, symmetrical, and bell-shaped. The parameter sigma controls how flat or peaked the curve is -- decreasing sigma (with mu fixed) makes the curve more sharply peaked (higher probability of X being close to mu); increasing sigma flattens it. Changing mu (with sigma fixed) simply slides the same-shaped curve left or right along the x-axis.

10.1 Properties of the Normal Distribution

The normal distribution has several defining properties worth memorizing. It is a continuous distribution ranging from -infinity to +infinity, with total area under the curve equal to 1 (P(-infinity < X < infinity) = 1). It is unimodal with its single peak (mode) at x = mu, where the maximum ordinate is f(mu) = 1/(sigma.sqrt(2.pi)). It is perfectly symmetrical, so mean = median = mode = mu, the lower and upper quartiles are equidistant from mu, and all odd-order moments about the mean are zero.

Three 'special areas' recur constantly in problems: P(mu-sigma < X < mu+sigma) = 0.6827 (about 68%), P(mu-2sigma < X < mu+2sigma) = 0.9545 (about 95%), and P(mu-3sigma < X < mu+3sigma) = 0.9973 (about 99.7%) -- together known as the empirical rule. The quartile deviation is Q.D.(X) = 0.6745.sigma, the mean deviation is M.D.(X) = 0.7979.sigma, and the variance/SD are simply sigma-squared and sigma. The curve is asymptotic to the x-axis (never touches it) and has points of inflexion at x = mu - sigma and x = mu + sigma. A useful reproductive property: if X1 ~ N(mu1, sigma1^2) and X2 ~

N(mu2, sigma2^2) are independent, then X1+X2 ~ N(mu1+mu2, sigma1^2+sigma2^2).

10.2 The Standard Normal Random Variable

Because the normal cumulative distribution function cannot be integrated in closed form, and because there are infinitely many possible (mu, sigma) combinations, statisticians instead standardize: any normal random variable X can be transformed into the standard normal random variable Z = (X-mu)/sigma, which always has mean 0 and variance 1, regardless of X's original mu and sigma. This lets a single set of tables answer probability questions for every normal distribution.

The standard normal density function is written phi(z) = [1/sqrt(2.pi)].e^(-z^2/2), and its cumulative distribution function is written Phi(z) = P(Z <= z) -- the area under the standard normal curve up to z. Because the curve is symmetric, phi(-z) = phi(z). Four handy results follow directly from the definition: P(Z<=a) = Phi(a); P(Z>=a) = 1-Phi(a); P(a<=Z<=b) = Phi(b)-Phi(a); and using symmetry, Phi(-a) = 1-Phi(a), so P(|Z|<a) = 2.Phi(a)-1 and P(|Z|>a) = 2.Phi(-a). Phi(z) and its inverse Phi-inverse(p) are both read from standard tables.

10.2 Quantiles, Percentiles, and De-standardizing

The p-th quantile (or 100p-th percentile) of Z is the value z_p such that Phi(z_p) = p, i.e., z_p = Phi-inverse(p) -- found by looking up p in the cumulative table and reading off the corresponding z. Once a quantile z_p of the standard normal distribution is known, it can be de-standardized back into the original scale of any normal variable X using x_p = mu + sigma.z_p, obtained by rearranging Z = (X-mu)/sigma into X = mu + sigma.Z.

For applied problems on any normal variable X ~ N(mu, sigma-squared), the same standardization logic applies directly to probability statements: P(X<=a) = Phi[(a-mu)/sigma]; P(X>=a) = 1-Phi[(a-mu)/sigma]; P(a<=X<=b) = Phi[(b-mu)/sigma] - Phi[(a-mu)/sigma]. This single technique -- standardize, look up in the Z table, and de-standardize when needed -- solves essentially every problem type in this chapter, from finding probabilities given X-values to finding X-values (or even mu or sigma) given probabilities.

Important Definitions

What is a normal distribution?: A continuous probability distribution described by f(x) = [1/(sigma.sqrt(2.pi))].e^[-(1/2)((x-mu)/sigma)^2], fully determined by its mean mu and standard deviation sigma, producing a symmetric, unimodal, bell-shaped curve.

What is the standard normal distribution?: The special case of the normal distribution with mean 0 and variance 1, obtained from any normal variable X via the transformation Z = (X-mu)/sigma.

What is Phi(z)?: The standard normal cumulative distribution function: Phi(z) = P(Z<=z), the area under the standard normal curve up to the point z.

What is phi(z)?: The standard normal probability density function (ordinate): phi(z) = [1/sqrt(2.pi)].e^(-z^2/2), the height of the standard normal curve at the point z.

What is a quantile (percentile) of the standard normal distribution?: The value z_p such that P(Z<=z_p) = p; found as z_p = Phi-inverse(p) using the inverse standard normal table.

What does 'de-standardizing' mean?: Converting a standard normal value z back to the original scale of X using X = mu + sigma.Z.

What is the empirical rule for the normal distribution?: Approximately 68.27% of values lie within 1 standard deviation of the mean, 95.45% within 2 standard deviations, and 99.73% within 3 standard deviations.

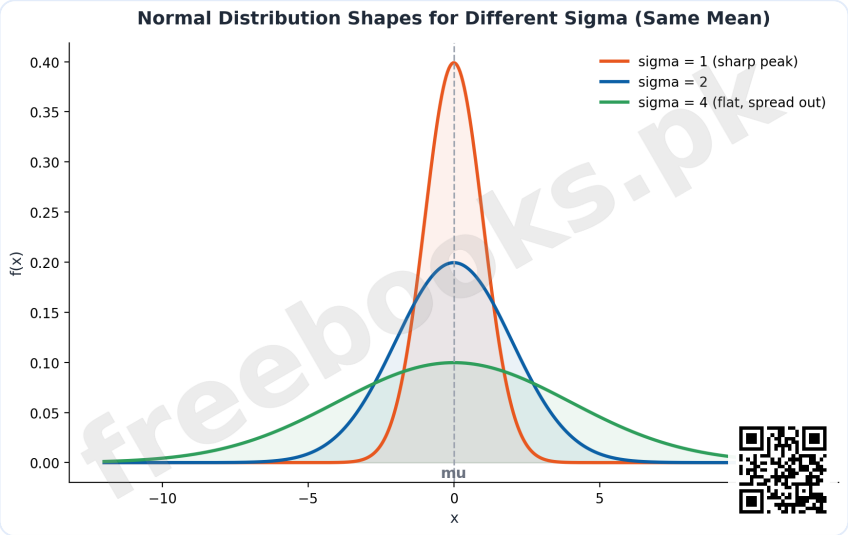
Key Facts and Relations

Topic	Key Fact / Relation
Normal PDF	f(x) = [1/(sigma.sqrt(2.pi))] e^[-(1/2)((x-mu)/sigma)^2]
Standardization	Z = (X - mu) / sigma
De-standardizing	X = mu + sigma.Z
Standard normal PDF	phi(z) = [1/sqrt(2.pi)] e^(-z^2/2)

Topic	Key Fact / Relation
Standard normal CDF	Phi(z) = P(Z <= z)
Symmetry results	Phi(-a) = 1-Phi(a); P(|Z|<a) = 2Phi(a)-1; P(|Z|>a) = 2Phi(-a)
Quartile deviation	Q.D.(X) = 0.6745 sigma
Mean deviation	M.D.(X) = 0.7979 sigma
Special areas (empirical rule)	mu+/-sigma: 68.27% | mu+/-2sigma: 95.45% | mu+/-3sigma: 99.73%
Quantile of X	x_p = mu + sigma . z_p, where z_p = Phi-inverse(p)
Moments about mean	mu1=0, mu2=sigma^2, mu3=0, mu4=3sigma^4; Beta1=0, Beta2=3

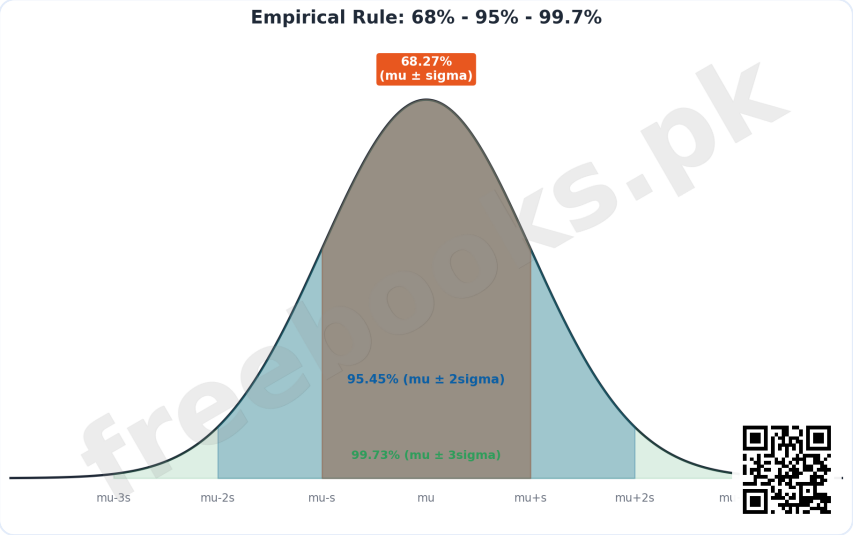
Diagrams

Normal Distribution Shapes for Different Sigma: Three normal curves with the same mean but different standard deviations, showing how a smaller sigma produces a sharper, more peaked curve while a larger sigma produces a flatter, more spread-out curve



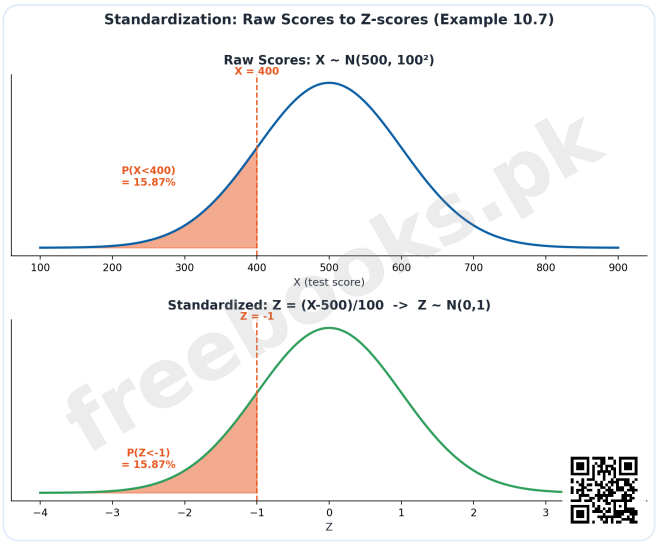
Normal Distribution Shapes for Different Sigma

Empirical Rule: 68-95-99.7: The standard normal curve with the areas within 1, 2, and 3 standard deviations of the mean shaded and labelled 68.27%, 95.45%, 99.73%, illustrating the special areas under the normal curve



Empirical Rule: 68-95-99.7

Standardization: Raw Scores to Z-scores: Two aligned normal curves -- the raw test-score distribution (mean 500, SD 100) on top and the standard normal Z distribution below -- with the region X<400 (Z<-1) shaded on both, illustrating the standardization technique from Example 10.7



Standardization: Raw Scores to Z-scores

Short Questions & Answers

What are the two parameters of a normal distribution and what do they control?

Mu (the mean) controls the location/center of the curve; sigma (the standard deviation) controls its spread -- smaller sigma gives a sharper peak, larger sigma gives a flatter curve.

Why do we standardize a normal random variable before using probability tables?

Because there are infinitely many possible normal distributions (one for every combination of mu and sigma), so probabilities are tabulated only once for the single standard normal distribution (mean 0, variance 1), and any normal variable is converted to this common scale via Z=(X-mu)/sigma.

State the empirical rule (68-95-99.7 rule) for a normal distribution.

About 68.27% of values fall within 1 standard deviation of the mean, about 95.45% within 2 standard deviations, and about 99.73% within 3 standard deviations.

Why is the mode of a normal distribution equal to its mean?

Because the normal distribution is perfectly symmetric and unimodal, its single peak occurs exactly at x=mu, and by symmetry the mean, median, and mode all coincide at that same point.

What is the maximum ordinate of a standard normal curve, and where does it occur?

The maximum ordinate is phi(0) = 1/sqrt(2.pi) ≈ 0.3989, occurring at z=0.

How do you find the value of X corresponding to a given percentile of a normal distribution?

First find the corresponding z-value using the standard normal table (z_p = Phi-inverse(p)), then de-standardize using x_p = mu + sigma.z_p.

Long Questions & Answers

Explain the normal probability density function and its key properties, and explain why the normal distribution is considered the most important distribution in statistics.

The normal distribution occupies a uniquely central position in statistical theory and practice, and understanding why requires looking both at its mathematical structure and at the practical reasons it appears so often in real data. Formally, a continuous random variable X is said to be normally distributed if its probability density function takes the specific form f(x) equals one over sigma times the square root of 2 pi, multiplied by e raised to the power of negative one-half times the quantity (x minus mu) divided by sigma, all squared, for x ranging across the entire real number line from negative infinity to positive infinity. This formula might look intimidating at first glance, but its behavior is governed entirely by just two parameters: mu, which can be any real number and represents the mean of the distribution, and sigma, which must always be strictly positive and represents the standard deviation. Because exactly these two numbers completely determine the shape and location of the entire curve, statisticians commonly abbreviate the statement 'X follows a normal distribution with mean mu and variance sigma-squared' using the compact notation X tilde N(mu, sigma-squared). The graph traced out by this density function is called the normal curve, and it possesses several visually and mathematically striking characteristics: it is unimodal, meaning it has exactly one peak; it is perfectly symmetrical around that peak; and it forms the instantly recognizable bell shape that gives the distribution its popular nickname. The specific role played by each of the two parameters becomes clear once we consider what happens as each is varied independently of the other. Holding mu fixed while decreasing sigma causes the curve to become more sharply peaked and narrow, concentrating probability more tightly around the mean and thereby making values close to mu considerably more likely relative to values far away from it. Conversely, holding mu fixed while increasing sigma causes the curve to flatten out and spread wider, distributing probability more evenly across a broader range and correspondingly making values close to mu comparatively less likely. Meanwhile, if sigma is instead held constant while mu itself is varied, the overall shape of the curve remains completely unchanged, and only its horizontal position along the

number line shifts, sliding to center on whatever new value mu has taken. Beyond this basic shape-and-location behavior, the normal distribution obeys several formal properties that prove immensely useful when solving problems. Because the entire distribution is a valid continuous probability distribution, the total area beneath the curve across its entire infinite range must equal exactly 1, and because probability under any continuous distribution corresponds to area rather than height at a single point, the probability of X falling in any particular interval is found by calculating the corresponding area under the curve above that interval. Because of the distribution's perfect symmetry, several familiar summary measures all collapse onto the exact same single value: the mean, the median, and the mode of any normal distribution are always identical, all located precisely at mu, and this same symmetry also guarantees that the lower and upper quartiles sit at exactly equal distances on either side of mu, and that every odd-numbered moment calculated about the mean equals exactly zero. Perhaps the most practically useful properties of all are the so-called special areas that hold true for absolutely every normal distribution regardless of its specific mu and sigma values: no matter which particular normal distribution is being considered, exactly 68.27 percent of its total probability always falls within one standard deviation of the mean, 95.45 percent always falls within two standard deviations, and 99.73 percent always falls within three standard deviations -- a remarkably consistent pattern collectively known as the empirical rule, which turns out to be enormously useful for quickly estimating probabilities without needing to consult any table at all. As for why this particular distribution earns its status as the single most important distribution across the whole of statistics, three distinct but mutually reinforcing reasons stand out clearly. First, purely as an empirical matter of observation, an enormous number of naturally occurring variables genuinely do follow something extremely close to a normal distribution in real life, including the heights and weights of individuals within a population, standardized intelligence test scores, and the small random errors that inevitably accompany essentially any careful physical measurement. Second, on a more theoretical level, the normal distribution frequently emerges mathematically as the limiting form that numerous entirely different probability distributions converge toward under suitable conditions, which means it can often serve as a remarkably accurate practical approximation to those other distributions even in situations where the underlying process generating the data isn't literally normal

to begin with. Third, and perhaps most profoundly important of all for the rest of statistical theory, the normal distribution is also precisely the limiting distribution described by the celebrated central limit theorem, a foundational and far-reaching result explored in depth in the following chapter, which explains why sample means calculated from repeated random sampling tend to follow an approximately normal distribution even when the underlying population itself is not normally distributed at all -- and it is this specific theoretical property above all others that ultimately underlies the vast majority of statistical inference techniques used throughout the entire discipline, from constructing confidence intervals to performing hypothesis tests.

Explain the concept of the standard normal random variable, describe the process of standardization and de-standardization, and explain how the standard normal table is used to solve problems involving probabilities, quantiles, and unknown parameters.

Since a genuinely unlimited number of different normal distributions exist mathematically, one for literally every possible combination of a mean mu and a standard deviation sigma, it would be entirely impractical, in fact essentially impossible, to construct and maintain a separate complete probability table for every single one of these countless distributions individually. Statisticians resolve this fundamental practical difficulty through an elegant technique called standardization, which converts any and every normal random variable into one single common, universally shared reference distribution before any table lookup is ever performed. Specifically, given any normal random variable X that follows a distribution with mean mu and standard deviation sigma, the standardized variable Z is defined directly by the simple transformation formula Z equals the quantity (X minus mu), all divided by sigma. What makes this particular transformation so remarkably powerful and useful in practice is a clean mathematical guarantee: no matter what specific numerical values the original mu and sigma of X happened to be, the resulting standardized variable Z that comes out the other end of this transformation will always, without any exception whatsoever, have a mean of exactly 0 and a variance of exactly 1. This special, universally shared distribution is formally called the standard normal distribution, conventionally written as Z tilde N(0,1), and because literally every normal distribution can be converted into this exact same standard form through the

identical transformation process, one single unified probability table can then be constructed just once, published, and subsequently reused again and again to solve probability problems for every conceivable normal distribution, regardless of that particular distribution's own specific mu and sigma values. The standard normal distribution comes equipped with its own dedicated and specially reserved mathematical notation precisely because it gets referenced so extremely frequently throughout the rest of statistical work. Its probability density function, representing the height or ordinate of the curve at any given point z, is written using the lowercase Greek letter phi, as phi(z) equals one over the square root of 2 pi, multiplied by e raised to the power of negative z-squared divided by 2. Its cumulative distribution function, representing instead the total accumulated area under that same curve running from negative infinity all the way up to some specified point z, is written using the corresponding uppercase Greek letter, as Phi(z) equals P(Z is less than or equal to z). This particular uppercase Phi(z) function is unquestionably the single most frequently used and referenced tool throughout this entire chapter, since it directly supplies the answer to the single most fundamental and recurring question posed again and again in problems of this type: what proportion, or equivalently what probability, of a standard normal distribution's total area lies at or below some particular specified point along the z-axis. Because the underlying standard normal curve possesses perfect mirror symmetry around the central value of zero, several extremely useful and time-saving relationships follow immediately and directly as natural consequences of that inherent symmetry, all readily available without requiring any additional or separate calculation whatsoever: the probability that Z exceeds some point a can always be found quickly as simply 1 minus Phi(a); the probability that Z falls somewhere between two specified points a and b can always be found as the straightforward difference Phi(b) minus Phi(a); and by that same underlying symmetry, Phi of negative a always equals exactly 1 minus Phi(a), a relationship which itself then leads directly to two further convenient shortcut formulas frequently needed for absolute-value probability statements, specifically that P(the absolute value of Z is less than a) equals 2 times Phi(a) minus 1, and that P(the absolute value of Z is greater than a) equals 2 times Phi(negative a). Beyond simply calculating probabilities in the forward direction starting from some given z-value, problems in this chapter frequently also demand working the entire process in reverse: given some specific target

probability or percentage value already in hand, the task instead becomes finding whichever particular z-value corresponds to producing exactly that stated probability. This particular kind of reverse lookup defines what is formally called a quantile, or equivalently and interchangeably a percentile, of the standard normal distribution: the p-th quantile, conventionally denoted using the symbol z sub p, is specifically defined as that unique value satisfying the defining equation Phi(z sub p) equals p, which is written more compactly and directly using standard inverse-function notation as z sub p equals Phi inverse of p, and this particular inverse value is found in practice simply by locating the target probability p somewhere within the body of the standard cumulative probability table and then reading directly across to whichever corresponding z-value produced that exact same tabulated probability in the first place. Once any particular quantile of the standard normal distribution Z has successfully been determined through this table-lookup process, that resulting standardized value can then, whenever it becomes genuinely necessary or useful for the problem at hand, be converted straight back once again into the original measurement scale belonging to whatever specific normal variable X was actually being studied in the first place, through a process referred to quite naturally and appropriately as de-standardizing. This reverse conversion is accomplished simply and directly by algebraically rearranging the original standardization formula Z equals (X minus mu) divided by sigma into its equivalent inverted form X equals mu plus sigma times Z, meaning that the p-th quantile expressed specifically in the original X-scale, conventionally written as x sub p, can always be calculated directly and immediately as x sub p equals mu plus sigma multiplied by z sub p. Taken together as a single unified, integrated procedure, this complete three-stage process, comprising standardizing the specific problem at hand into the common Z-scale, then consulting the single shared standard normal table to either find a probability or alternatively to find a percentile as the specific problem demands, and finally de-standardizing that resulting z-value back into the original problem's own particular scale whenever that final conversion step happens to be required, together forms the single foundational technique that runs consistently through virtually every single problem type encountered throughout this entire chapter -- whether the specific task at hand is calculating some probability directly from a given X-value, or instead working entirely in reverse to find some unknown X-value starting from a given target probability, or even

solving the comparatively more advanced style of problem that requires determining an entirely unknown mean mu or unknown standard deviation sigma (or occasionally even solving simultaneously for both unknowns at once) by making skillful and deliberate use of whatever specific probability information happens to already be given as part of that particular problem's stated conditions.

Multiple Choice Questions (MCQs)

The normal distribution is fully characterized by which two parameters? (A) Median and mode (B) Mean (mu) and standard deviation (sigma) (C) Range and IQR (D) Skewness and kurtosis

Correct answer: (B) Mean (mu) and standard deviation (sigma). A normal distribution X ~ N(mu, sigma^2) is completely determined by its mean mu and standard deviation sigma.

The total area under any normal curve equals: (A) 0 (B) 0.5 (C) 1 (D) Infinity

Correct answer: (C) 1. Like all valid probability density functions, the total area under the normal curve is exactly 1.

In a normal distribution, mean, median, and mode: (A) Are always different (B) Are always equal (C) Sum to zero (D) Equal sigma

Correct answer: (B) Are always equal. Due to the perfect symmetry of the normal distribution, mean = median = mode = mu.

The standard normal distribution has mean and variance: (A) Mean 1, variance 0 (B) Mean 0, variance 1 (C) Mean mu, variance sigma^2 (D) Mean 100, variance 15

Correct answer: (B) Mean 0, variance 1. The standard normal distribution Z ~ N(0,1) always has mean 0 and variance 1.

The standardization formula is: (A) Z = X + mu/sigma (B) Z = (X - mu)/sigma (C) Z = X.sigma - mu (D) Z = mu/X

Correct answer: (B) Z = (X - mu)/sigma. Standardization converts any normal variable X into Z using Z = (X-mu)/sigma.

Approximately what percent of a normal distribution lies within 2 standard deviations of the mean? (A) 68.27% (B) 95.45% (C) 99.73% (D) 50%

Correct answer: (B) 95.45%. By the empirical rule, about 95.45% of values lie within mu +/- 2sigma.

Phi(z) represents: (A) The ordinate of the standard normal curve at z (B) P(Z <= z) (C) The variance of Z (D) The mean of X

Correct answer: (B) P(Z <= z). Phi(z) is the standard normal cumulative distribution function, giving P(Z<=z).

If P(Z<a) = 1 - P(Z<-a), this reflects which property? (A) Skewness (B) Kurtosis (C) Symmetry of the standard normal curve (D) The empirical rule

Correct answer: (C) Symmetry of the standard normal curve. This relationship, Phi(-a)=1-Phi(a), follows directly from the symmetry of the standard normal curve about zero.

The formula to de-standardize a z-value back to the original scale is: (A) X = mu - sigma.Z (B) X = mu + sigma.Z (C) X = sigma/mu (D) X = mu.Z

Correct answer: (B) X = mu + sigma.Z. De-standardizing uses X = mu + sigma.Z, the inverse of the standardization formula.

The maximum ordinate of the standard normal curve occurs at: (A) z = 1 (B) z = -1 (C) z = 0 (D) z = mu

Correct answer: (C) z = 0. The standard normal curve peaks at z=0, where phi(0) = 1/sqrt(2.pi) ≈ 0.3989.

Quick Revision Summary

- Normal PDF: f(x) = [1/(sigma.sqrt(2pi))] e^[-(1/2)((x-mu)/sigma)^2], parameters mu and sigma
- Symmetric, unimodal, bell-shaped; mean=median=mode=mu; total area = 1
- Empirical rule: mu+/-sigma = 68.27% | mu+/-2sigma = 95.45% | mu+/-3sigma = 99.73%
- Standardization: Z = (X-mu)/sigma; standard normal Z~N(0,1)
- Phi(z) = P(Z<=z); P(Z>a)=1-Phi(a); P(a<Z<b)=Phi(b)-Phi(a)
- Symmetry: Phi(-a)=1-Phi(a); P(|Z|<a)=2Phi(a)-1; P(|Z|>a)=2Phi(-a)

- Quantile: z_p = Phi-inverse(p); de-standardize: x_p = mu + sigma.z_p
- Q.D.(X) = 0.6745 sigma; M.D.(X) = 0.7979 sigma; Var(X)=sigma^2

Exam Tips

- Always sketch a quick bell curve and shade the region asked for — it prevents sign errors with Phi(a) vs 1-Phi(a)
- For 'P(X between a and b)' problems, standardize BOTH endpoints before subtracting Phi values
- For 'find X given a probability' problems, always look up the z-value FIRST, then de-standardize with x = mu + sigma.z
- Remember Phi(-z) = 1 - Phi(z) — most tables only list positive z, so use this to handle negative z-values
- When solving for unknown mu or sigma, set up the standardization equation and solve algebraically — don't guess
- Watch for continuity correction language ('less than 60mm' on discrete/measured data may mean 'less than 59.5' in a problem's intended interpretation)

Chapter 11: Sampling Techniques and Sampling Distributions

Almost nothing in real-world statistics is based on studying an entire population -- it is far too slow, expensive, or simply impossible. Instead, statisticians study a carefully chosen sample and use it to draw conclusions about the whole population. This chapter builds the foundation for that entire process: how populations and samples are defined, how a representative sample is actually selected (probability vs non-probability methods, simple random sampling, stratified sampling), and what kinds of errors can creep in along the way.

The second and more theoretical half of the chapter introduces sampling distributions -- the probability distribution of a statistic (like the sample mean or sample proportion) across all possible samples of a given size. This concept, and especially the Central Limit Theorem introduced here, is the single most important theoretical bridge in the entire course: it is what allows the normal distribution from Chapter 10 to be used for inference about sample means and proportions in every chapter that follows (Estimation, Hypothesis Testing).

Learning Objectives

- Define population, sample, sampling, parameter, and statistic, and distinguish clearly between each pair
- Distinguish probability sampling from non-probability sampling and describe common methods of each
- Explain and apply simple random sampling with and without replacement, including counting the number of possible samples
- Describe stratified sampling and explain when it is preferred over simple random sampling
- Distinguish sampling error from non-sampling error, and define accuracy, precision, and bias
- Explain what a sampling distribution of a statistic is and define its standard error

- State and apply the properties of the sampling distribution of the sample mean, with and without the finite population correction
- State the Central Limit Theorem and explain its importance in statistical inference
- Find the mean and standard error of the sampling distributions of the difference of two means, sample proportion, difference of two proportions, and sample variance

Key Concepts

11.1-11.2 Population and Sample

A population is the totality of all observations (or all sampling units) possessing some common, specified characteristic that a researcher wants to study -- it may be finite (a limited, countable number of units, such as all registered voters in a city) or infinite (an unlimited number of units, such as all possible outcomes of tossing a coin indefinitely, or all items that could ever be produced by an ongoing production process).

A sample is a part of the population selected in such a way that it is expected to represent the characteristics of the whole population reasonably well. Sampling is the procedure used to select such a representative sample. A sample survey studies only a part of the population, whereas a census (complete enumeration) studies every single unit. Samples are generally preferred over a census because they save time, cost far less, are often more feasible for large or inaccessible populations, and -- perhaps counter-intuitively -- can sometimes give MORE accurate results than a census, since a smaller, well-managed data collection effort allows more careful measurement and closer supervision, reducing non-sampling errors. The limitations of sampling are that it can never give perfectly exact results (only estimates) and requires careful, expert planning to avoid a biased or unrepresentative sample.

11.3 Sampling Design

A sampling design is the complete plan for selecting a sample from a given population. Its building blocks are the sampling unit (the individual element, or group of elements, that will actually be selected, e.g., a household, a student, a

factory) and the sampling frame -- a complete list of all sampling units in the population, from which the sample is actually drawn (e.g., a voter list, a list of registered students). A good sampling frame should be accurate, complete, and up to date; a poor frame is one of the most common sources of non-sampling error.

Sampling methods fall into two broad families: non-probability sampling, in which the chance of any particular element being included in the sample cannot be determined in advance (examples include convenience sampling, judgment/ purposive sampling, and quota sampling) -- these are quick and cheap but their results cannot be generalized to the population with a known level of confidence; and probability sampling (also called random sampling), in which every element of the population has a known, non-zero probability of being selected. Probability sampling is the standard for scientific inference because it allows the sampling error to be objectively measured and controlled. Selection may be done with replacement (a selected unit is returned to the population and can be chosen again) or without replacement (a selected unit is removed and cannot be chosen again).

11.6 Simple Random Sampling

Simple random sampling (SRS) is the most basic probability sampling method: a sample of size n is drawn from a population of size N such that every possible sample of that size has an equal chance of being selected. It can be carried out physically using the lottery method (writing all N units on identical slips, mixing thoroughly, and drawing n of them) or, more commonly and more rigorously, using a table of random numbers (or a random-number generator), where units of the population are numbered and n numbers are read off the table according to a fixed, unbiased rule.

Theorem 11.1 gives the counting rule that underlies every sampling-distribution calculation in this chapter: if sampling is done WITH replacement, the number of possible ordered samples of size n from a population of size N is N raised to the power n (N^n) -- because each of the n draws independently has N possible outcomes. If sampling is done WITHOUT replacement, the number of possible ordered samples of size n is N(N-1)(N-2)...(N-n+1), written compactly as NPn (N permute n) -- because the first draw has N possible outcomes, the second has

only N-1 remaining, the third only N-2, and so on, since a selected unit cannot be chosen again. These two counting formulas (N^n and NPn) are the starting point of virtually every worked example on sampling distributions in this chapter -- they tell you exactly how many rows the sampling distribution table will have before you even start building it.

11.7 Stratified Sampling

Stratified sampling is used when the population is heterogeneous but can be divided into non-overlapping, internally homogeneous subgroups called strata (for example, dividing a student population into strata by grade level, or a national population into strata by province). A simple random sample is then drawn independently from each stratum, and the stratum samples are combined into the final overall sample. Stratified sampling is preferred over plain SRS whenever the strata differ meaningfully from each other but units within the same stratum are fairly similar, because it guarantees representation from every subgroup and typically produces estimates with a smaller standard error than an SRS of the same total size drawn from the unstratified population.

11.8 Errors in Sampling

No sample-based estimate is ever perfectly exact; the discrepancy between an estimate and the true population value can arise from two very different sources. Sampling error is the difference E = T - theta between a sample statistic T (such as the sample mean) and the true population parameter theta (such as the population mean) that arises purely because only a sample, and not the entire population, was observed -- it exists even if every single measurement in the sample was made perfectly, and it can generally be reduced by increasing the sample size. Non-sampling errors, by contrast, can occur even in a complete census, and arise from sources such as faulty questionnaire design, non-response, data-entry mistakes, or interviewer bias -- these do NOT necessarily shrink as the sample size grows, and in fact tend to grow more troublesome in very large surveys where quality control becomes harder to maintain.

Two further concepts describe the quality of an estimate. Accuracy refers to how close an estimate is to the true population value -- a highly accurate estimate has very little sampling error. Precision refers to how closely repeated estimates

(from repeated samples of the same size and design) cluster around each other, regardless of whether they are close to the true value -- it is measured by the standard error of the estimator; a small standard error means high precision. Bias is defined as the difference between the expected value of an estimator T and the true parameter, Bias = E(T) - theta; an unbiased estimator has Bias = 0, meaning it is correct 'on average' across all possible samples, even though any single sample estimate may still differ from the true value due to ordinary sampling error.

11.9-11.10 Parameters, Statistics, and the Idea of a Sampling Distribution

A parameter is any descriptive numerical measure calculated from the ENTIRE population -- since the population is fixed, a parameter is a constant with one single true (though often unknown) value. Key parameters include the population total tau = sum of all x-j, the population mean mu = tau/N, the population variance sigma-squared = [sum of (x-j - mu)^2]/N, and the population proportion pi = k/N (where k is the number of elements possessing some attribute).

A statistic is any descriptive numerical measure calculated from a SAMPLE -- because the particular sample drawn will differ from one instance to the next, a statistic is a random variable, taking a different value for each possible sample; it therefore has its own probability distribution. Key statistics include the sample total (sum of x-i), the sample mean x-bar = (sum of x-i)/n, the sample proportion p = x/n, and two versions of sample variance: the divide-by-n version s-squared = [sum of (x-i - x-bar)^2]/n and the divide-by-(n-1) version s-hat-squared = [sum of (x-i - x-bar)^2]/(n-1) -- the two are related by n.s^2 = (n-1).s-hat^2.

The sampling distribution of a statistic is the probability distribution formed by the values of that statistic computed from EVERY possible sample of a given size that could be drawn from the population. The standard error of a statistic is defined as the standard deviation of its sampling distribution -- it measures how much the statistic is expected to vary from one random sample to the next, and it is the single most important number needed for statistical inference (confidence intervals and hypothesis tests, covered in the next two chapters) because it quantifies the precision of an estimate.

11.11-11.12 Sampling Distribution of the Sample Mean and the Central Limit Theorem

The sampling distribution of the sample mean X-bar is the probability distribution of the means of all possible simple random samples of size n drawn from a population with mean mu and variance sigma-squared. Its properties are given by a sequence of theorems. Theorem 11.2: the mean of the sampling distribution of X-bar always equals the population mean, mu-Xbar = E(X-bar) = mu, regardless of sample size or replacement method -- so X-bar is an unbiased estimator of mu. Theorem 11.3: when sampling is done with replacement from an infinite population (or a finite one with replacement), the variance of X-bar is the population variance divided by the sample size, sigma-Xbar-squared = sigma-squared/n, so the standard error is sigma-Xbar = sigma/sqrt(n) -- notice the standard error shrinks as the sample size n grows, which is exactly why larger samples give more precise estimates.

Theorem 11.4 extends this to sampling WITHOUT replacement from a finite population of size N: the variance picks up an extra multiplier called the finite population correction (fpc), sigma-Xbar-squared = (sigma-squared/n).[(N-n)/(N-1)]. The fpc is always less than or equal to 1, so sampling without replacement always gives an equal or SMALLER standard error than sampling with replacement -- and when the sample size n is small relative to the population size N (a small 'sampling fraction' n/N), the fpc is so close to 1 that it can often be safely ignored in practice.

Theorem 11.5 states that if the ORIGINAL population itself is normally distributed, then X-bar is exactly normally distributed with mean mu and variance sigma-squared/n, for ANY sample size (even n=1). But most real populations are not normal -- which is where Theorem 11.6, the Central Limit Theorem (CLT), becomes indispensable: for a LARGE sample size, the sampling distribution of X-bar is approximately normal with mean mu and variance sigma-squared/n, REGARDLESS of the shape of the original population's distribution. The larger the sample size, the better this normal approximation becomes. This is the single most powerful and far-reaching result in this chapter: it means the standardized variable Z = (X-bar - mu)/(sigma/sqrt(n)) can be treated as approximately standard normal even when nothing at all is known about the shape of the

underlying population -- and it is this fact alone that makes the confidence intervals and hypothesis tests of Chapters 12 and 13 possible for virtually any kind of data.

11.13 Sampling Distribution of the Difference Between Two Sample Means

When comparing two populations (for example, comparing the average lifetime of light bulbs from two different manufacturers), we work with the difference X-bar-1 minus X-bar-2 between two INDEPENDENT sample means. Theorem 11.7: the mean of this difference always equals the difference of the population means, mu = mu-1 - mu-2. Theorem 11.8: for independent samples with replacement (or from infinite populations), the variance of the difference is the SUM of the two individual sampling variances, sigma-squared = sigma-1-squared/n-1 + sigma-2-squared/n-2 -- variances of independent random variables add even though we are taking a DIFFERENCE, because Var(A-B) = Var(A) + Var(B) when A and B are independent. Theorem 11.9 adds the finite population correction to each term for sampling without replacement from finite populations, and Theorem 11.10 confirms that if both original populations are normal, the difference X-bar-1 - X-bar-2 is exactly normally distributed (and by the Central Limit Theorem, approximately normal for large samples even if the populations are not normal).

11.14-11.15 Sampling Distribution of Sample Proportion and Its Difference

When the characteristic of interest is qualitative with two outcomes (success/failure, yes/no), the population proportion is pi = k/N and the sample proportion is P = X/n, where X counts the number of 'successes' in the sample. Theorem 11.11: for sampling with replacement (or from an infinite/Bernoulli population), P has mean mu-P = pi and variance sigma-P-squared = pi(1-pi)/n -- in fact the exact sampling distribution of P in this case is directly linked to the binomial distribution from Chapter 9. The shape of the sampling distribution of P is right-skewed when pi < 0.5, left-skewed when pi > 0.5, and symmetric when pi = 0.5, but as n grows large it approaches a normal shape by the Central Limit Theorem. Theorem 11.12 adds the finite population correction for sampling without

replacement, linking the sampling distribution of P in that case to the hypergeometric distribution, also from Chapter 9.

Just as with means, we often compare two proportions from two independent Bernoulli populations. Theorem 11.13: the mean of the difference P1 - P2 equals pi-1 - pi-2. Theorem 11.14: for independent samples with replacement, the variance of the difference is again the sum of the two individual variances, sigma-squared = pi-1(1-pi-1)/n-1 + pi-2(1-pi-2)/n-2. Theorem 11.15 adds the finite population correction to each term for sampling without replacement from finite populations.

11.16-11.17 Other Sampling Distributions: Sample Variance

Besides the mean and the proportion, every other sample statistic -- the sample median, the sample variance, the sample standard deviation -- has its own distinct sampling distribution, and a full specification of any sampling distribution must always state three things together: the population being sampled, the particular statistic being computed, and the sample size being used, since changing any one of these three produces a genuinely different sampling distribution. For the sample variance S-squared = [sum of (X-i - X-bar)^2]/n (computed with the divide-by-n convention), the sampling distribution has the property mu-S-squared = E(S-squared) = [(n-1)/n].sigma-squared -- notice this is slightly LESS than the true population variance sigma-squared, which is exactly why the alternative divide-by-(n-1) version of sample variance is preferred as an unbiased estimator in later chapters on estimation.

Important Definitions

What is a population?: The totality of all observations (sampling units) possessing some common, specified characteristic that a study wants to examine; may be finite or infinite.

What is a sample?: A part of the population selected in such a way that it is expected to represent the characteristics of the whole population.

What is the difference between a parameter and a statistic?: A parameter is a descriptive measure computed from the entire population and is therefore a

fixed constant; a statistic is a descriptive measure computed from a sample and is therefore a random variable that varies from sample to sample.

What is a sampling frame?: A complete list of all the sampling units in a population, from which a sample is actually drawn.

What is the difference between probability and non-probability sampling?: In probability sampling every element of the population has a known, non-zero chance of selection, so results can be generalized with a known confidence level; in non-probability sampling the selection chances cannot be determined, so results cannot be reliably generalized to the population.

What is sampling error?: The difference E = T - theta between a sample statistic and the true population parameter, arising purely because only a sample rather than the whole population was observed; it generally decreases as sample size increases.

What is the difference between accuracy and precision?: Accuracy is how close an estimate is to the true population value; precision is how closely repeated estimates cluster around each other (measured by the standard error), regardless of whether they are close to the true value.

What is a sampling distribution?: The probability distribution formed by the values of a statistic computed from every possible sample of a given size that could be drawn from a population.

What is the standard error of a statistic?: The standard deviation of the sampling distribution of that statistic; it measures how much the statistic is expected to vary from sample to sample.

What is the finite population correction (fpc)?: The factor (N-n)/(N-1) applied to the variance of a sampling distribution when sampling is done without replacement from a finite population; it is always <=1 and can often be ignored when the sample is a small fraction of the population.

State the Central Limit Theorem.: For a large sample size, the sampling distribution of the sample mean X-bar is approximately normal with mean mu and variance sigma-squared/n, regardless of the shape of the original population's distribution; the approximation improves as n increases.

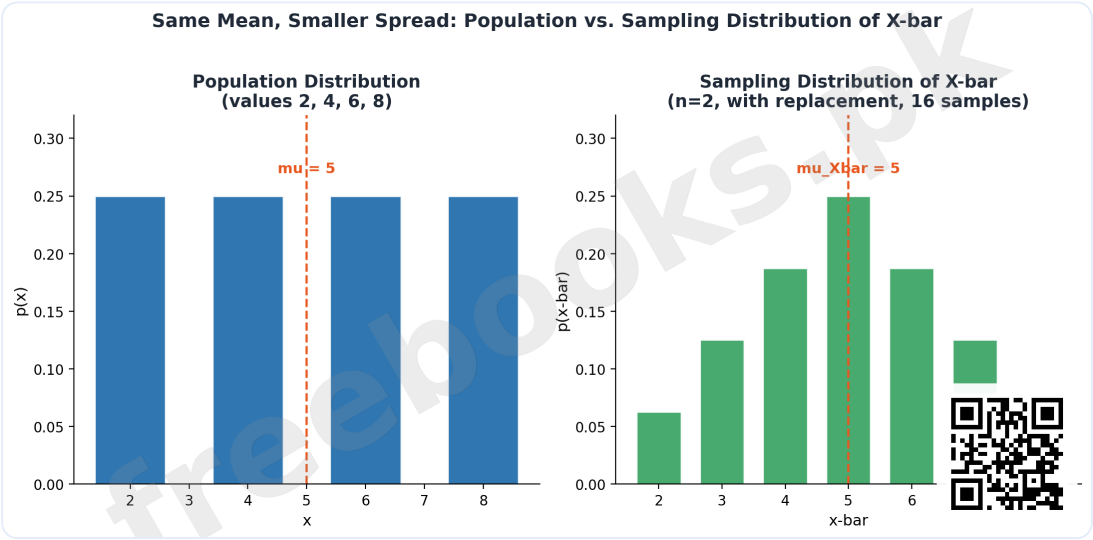
Key Facts and Relations

Topic	Key Fact / Relation
Population parameters	mu = (sum x_j)/N; sigma^2 = (sum x_j^2)/N - mu^2; pi = k/N
Sample statistics	x-bar = (sum x_i)/n; s^2 = [sum(x_i - x-bar)^2]/n; p = x/n
Number of samples (Theorem 11.1)	With replacement: N^n | Without replacement: NPn = N(N-1)...(N-n+1)
Mean of X-bar (Thm 11.2)	mu_Xbar = E(X-bar) = mu
Variance of X-bar, with replacement (Thm 11.3)	sigma_Xbar^2 = sigma^2/n; SE = sigma/sqrt(n)
Variance of X-bar, without replacement (Thm 11.4)	sigma_Xbar^2 = (sigma^2/n).[(N-n)/(N-1)]
Central Limit Theorem (Thm 11.6)	For large n: X-bar ~approx N(mu, sigma^2/n), regardless of population shape
Z for sample mean	Z = (X-bar - mu)/(sigma/sqrt(n))
Difference of means: mean & variance (Thm 11.7-11.8)	mu = mu1-mu2; sigma^2 = sigma1^2/n1 + sigma2^2/n2
Sample proportion: mean & variance (Thm 11.11-11.12)	mu_P = pi; sigma_P^2 = pi(1-pi)/n [x fpc if without replacement]
Difference of proportions: mean & variance (Thm 11.13-11.14)	mu = pi1-pi2; sigma^2 = pi1(1-pi1)/n1 + pi2(1-pi2)/n2
Mean of sample variance (S^2, divide by n)	mu_(S^2) = E(S^2) = [(n-1)/n].sigma^2

Diagrams

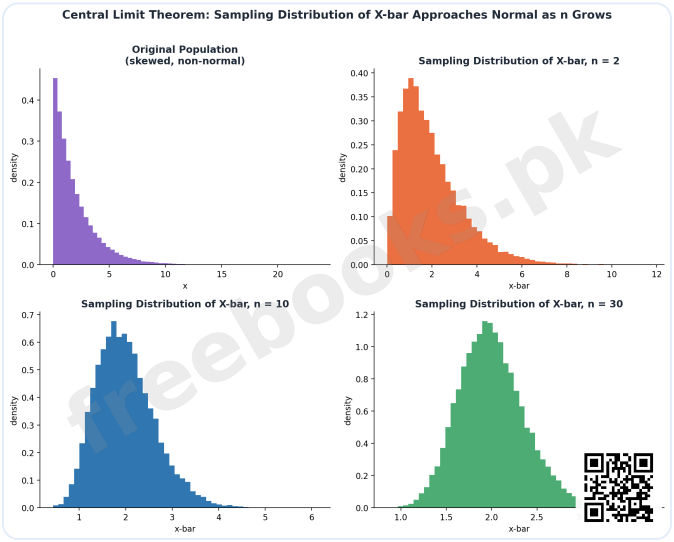
Population Distribution vs. Sampling Distribution of the Mean: A side-by-side comparison of a flat (uniform) population distribution for the values 2,4,6,8 against the mound-shaped sampling distribution of the sample mean for all 16

possible samples of size 2 drawn with replacement -- both centred at the same mean of 5, but the sampling distribution of X-bar is visibly narrower, illustrating Theorem 11.2 (equal means) and Theorem 11.3 (reduced variance)



Population Distribution vs. Sampling Distribution of the Mean

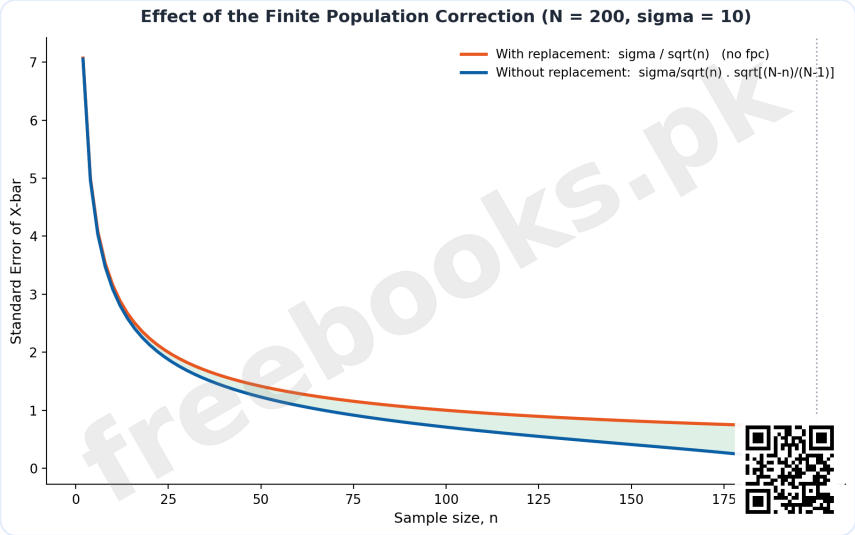
The Central Limit Theorem in Action: Three panels showing a skewed (non-normal) population distribution alongside simulated sampling distributions of the sample mean for increasing sample sizes n=2, n=10, and n=30, illustrating how the sampling distribution of X-bar becomes progressively more bell-shaped and narrower as n grows, regardless of the population's original skewed shape



The Central Limit Theorem in Action

Effect of the Finite Population Correction Factor: A line chart showing the standard error of the sample mean against sample size n for a fixed population

size N, comparing the with-replacement formula sigma/sqrt(n) against the without-replacement formula that includes the finite population correction, showing how the fpc pulls the standard error down further as the sample size approaches the population size



Effect of the Finite Population Correction Factor

Short Questions & Answers

Distinguish between a parameter and a statistic with one example each.

A parameter is a fixed numerical measure from the whole population, e.g., the population mean mu; a statistic is a numerical measure from a sample, and varies from sample to sample, e.g., the sample mean x-bar.

Why is sampling generally preferred over a complete census?

Sampling saves time and cost, is more feasible for large or infinite populations, and -- because a smaller data collection effort allows closer supervision -- can sometimes even be more accurate than a census by reducing non-sampling errors.

What is the key difference between sampling error and non-sampling error?

Sampling error arises only because a sample rather than the whole population was studied and generally shrinks as sample size grows; non-sampling error can occur even in a full census (e.g., from bad questionnaires or data-entry mistakes) and does not necessarily shrink with sample size.

Using Theorem 11.1, how many distinct samples of size n=3 can be drawn without replacement from a population of N=6 units?

NPn = N(N-1)(N-2) = 6(5)(4) = 120 possible samples.

Why does the finite population correction factor make the standard error smaller, not larger?

Because the fpc, (N-n)/(N-1), is always less than or equal to 1 when n < N, so multiplying the with-replacement variance by this factor can only reduce (or leave unchanged) the resulting standard error -- sampling without replacement gives more information per unit sampled.

Why is the Central Limit Theorem so important for statistical inference?

It guarantees that the sampling distribution of the sample mean is approximately normal for large samples regardless of the shape of the original population, which allows normal-distribution-based methods (confidence intervals, hypothesis tests) to be applied to sample means from virtually any kind of population.

Long Questions & Answers

Explain the concept of a sampling design, distinguishing probability from non-probability sampling, and describe simple random sampling and stratified sampling as two concrete probability sampling methods.

Before any statistical analysis of sample data can be trusted, careful thought has to be given to precisely how that sample was actually chosen in the first place, since the method of selection determines whether the resulting sample can be reasonably expected to represent the population it was drawn from. A sampling design is the complete, deliberate plan governing exactly how a sample will be selected from a given population, and it rests on two foundational building blocks. The first is the sampling unit, which is simply the individual element or defined group of elements that will actually be selected during the sampling process -- this might be a single household in a survey of family income, an individual student in a survey of academic performance, or an entire factory in a survey of industrial output, depending entirely on what the study is designed to measure. The second building block is the sampling frame, which is a complete,

itemized list of every single sampling unit that exists within the population under study, and it is from this list that the actual sample will eventually be drawn -- common real-world examples include an electoral register listing every registered voter, or an official enrollment list containing every currently registered student at an institution. The quality of this sampling frame matters enormously in practice, because a frame that is inaccurate, incomplete, or badly out of date is one of the single most common and consequential sources of non-sampling error encountered in real survey work, since units that are missing from the frame have no chance whatsoever of ever being selected into the sample, regardless of how careful the subsequent random selection procedure itself might be. Once a sampling frame has been properly established, the actual sampling methods available for selecting units from it fall broadly into two fundamentally different families. The first family is non-probability sampling, in which the precise probability of any particular individual element ending up included in the final sample simply cannot be calculated or determined in advance -- common examples include convenience sampling, where units are selected purely because they happen to be easy to reach, judgment or purposive sampling, where a researcher's own expert opinion decides which units to include, and quota sampling, where predetermined numbers from various subgroups are filled without genuine randomization within each subgroup. Such non-probability methods are certainly quick to carry out and inexpensive, but because the underlying probability of selection is unknown, the results they produce cannot be legitimately or objectively generalized back to the wider population with any statistically defensible, quantifiable level of confidence. The second, generally far more scientifically rigorous family is probability sampling, also frequently referred to as random sampling, in which every single element within the population is guaranteed to have some known, non-zero probability of ultimately being selected into the sample. This particular property is precisely what allows the resulting sampling error to be objectively measured, quantified, and rigorously controlled through the mathematical machinery of sampling distributions covered later in this very chapter, which is exactly why probability sampling methods form the accepted foundation of essentially all serious scientific and statistical inference. Simple random sampling, commonly abbreviated as SRS, represents the most basic and conceptually fundamental probability sampling method of all: a sample of some particular size n is drawn

from a population of size N in such a careful manner that every conceivable possible sample of that exact size has been given an precisely equal chance of ultimately being the one selected. In practical terms, SRS can be physically carried out through the lottery method, which involves writing every single population unit onto identical, indistinguishable slips of paper, thoroughly mixing the entire collection of slips together, and then drawing out exactly n of them at random; alternatively, and considerably more commonly in modern practice, SRS is instead carried out using a table of random numbers or an appropriate computerized random-number generator, whereby every unit in the population first gets assigned its own unique identifying number, after which n of these numbers are systematically read off according to some fixed, entirely unbiased selection rule. A crucial theoretical result underlying essentially every single sampling-distribution calculation that appears throughout the remainder of this chapter is the precise counting rule for determining exactly how many distinct possible samples actually exist: when sampling is carried out with replacement, the total number of distinct ordered samples of size n is N raised to the power n, conventionally written N^n, since each and every one of the n individual draws independently offers exactly N possible outcomes; but when sampling is instead carried out without replacement, the total count of distinct ordered samples of size n instead becomes N multiplied by (N-1) multiplied by (N-2), continuing on down through (N-n+1), a quantity conventionally denoted as NPn, precisely because each successive draw after the very first one necessarily has one fewer remaining possible outcome available, since previously selected units simply cannot be chosen again a second time. Stratified sampling represents an important and frequently more efficient refinement upon plain, unstructured simple random sampling, and it becomes especially valuable specifically in situations where the overall population being studied is genuinely heterogeneous as a whole, yet can nevertheless be meaningfully divided into a number of distinct, non-overlapping subgroups, formally termed strata, within each of which the individual member units are comparatively far more similar and homogeneous to one another -- a good illustrative example would be dividing an entire national population up into separate strata corresponding to its various different provinces, or alternatively dividing an entire student population up into separate strata according to each student's particular grade level. Under this particular sampling design, an entirely independent simple random sample is

then drawn separately from within each individual stratum, after which all of these individual stratum-level samples are subsequently combined together to form the single final overall sample used for the study. Stratified sampling proves genuinely preferable to an unstratified simple random sample of the exact same total overall size specifically whenever the various strata differ substantially and meaningfully from each other in their underlying characteristics, precisely because this particular sampling design guarantees that every single relevant subgroup of the population will be adequately represented within the final sample, and because it also typically manages to produce final estimates carrying a meaningfully smaller standard error compared to what an equivalently-sized unstratified simple random sample would have produced from that same overall heterogeneous population.

State the Central Limit Theorem in full, and explain its relationship to Theorems 11.2 through 11.5 on the sampling distribution of the sample mean, including why it is considered the single most important theoretical result of this chapter.

The theoretical heart of this entire chapter builds up in a careful, deliberate sequence of theorems, each one adding a further layer of generality on top of the last, and understanding precisely how each theorem in this sequence relates to the ones immediately before and after it is absolutely essential for genuinely grasping why the Central Limit Theorem occupies the uniquely central position that it does in modern statistical theory. Theorem 11.2 establishes the very first and most fundamental building block of the entire sequence: no matter which particular sample size n happens to be used, and entirely regardless of whether the actual sampling itself is conducted with or without replacement, the mean of the sampling distribution of the sample mean X-bar always turns out to equal precisely the true population mean mu itself, formally expressed as mu-Xbar equals E(X-bar) equals mu. This particular property carries an enormously important practical implication: it guarantees that the sample mean X-bar constitutes what statisticians formally term an unbiased estimator of the unknown population mean mu, meaning that even though any single individual sample's calculated mean will typically differ somewhat from the true population mean purely due to ordinary sampling error, the AVERAGE value of X-bar computed across every single conceivable sample that could possibly be drawn

will nevertheless land exactly precisely on the true population mean, with absolutely no systematic bias whatsoever pulling estimates consistently either too high or too low. Theorem 11.3 then proceeds to add crucial additional information specifically regarding the VARIABILITY, rather than merely the central tendency, of this same sampling distribution: specifically for the particular case of sampling with replacement, the variance of the sampling distribution of X-bar works out to equal the original population variance sigma-squared divided by the sample size n, yielding the well known standard error formula sigma-Xbar equals sigma divided by the square root of n. This particular formula carries a critically important practical takeaway of its own: as the sample size n is progressively increased, this standard error correspondingly and continuously shrinks smaller and smaller, which is exactly the precise underlying mathematical reason why larger samples reliably yield noticeably more precise, tightly clustered estimates of the unknown population mean than smaller samples do. Theorem 11.4 then extends this same basic variance result specifically to the alternative case of sampling WITHOUT replacement from a finite population, introducing in the process the finite population correction factor (N-n) divided by (N-1), a quantity that is always mathematically guaranteed to be less than or equal to 1 whenever the sample size n happens to be smaller than the population size N, and this multiplicative correction factor consequently always shrinks the resulting standard error down somewhat further still whenever sampling without replacement is actually employed, compared against what sampling with replacement of that identical sample size would otherwise have produced. Theorem 11.5, meanwhile, takes an important but comparatively more restrictive and narrowly limited step further: specifically IF the original underlying population itself happens to already be normally distributed, then the resulting sampling distribution of X-bar is not merely approximately normal but is instead EXACTLY normally distributed with mean mu and variance sigma-squared over n, and this particular exact result additionally holds true regardless of whatever specific sample size n happens to be used, even including the extreme case of a sample size as small as just a single solitary observation, n equals 1. The single crucial practical limitation of Theorem 11.5, however, is that its validity depends entirely and completely upon the rather strong underlying assumption that the original population being sampled from is already normally distributed in the first place -- and this particular assumption is one that, in the overwhelming majority

of genuinely real-world practical situations encountered by working statisticians, either cannot be verified with any real confidence at all, or else can be confidently verified as being straightforwardly false. This is precisely the exact point at which Theorem 11.6, the Central Limit Theorem itself, becomes so extraordinarily indispensable and powerful within the whole of statistical theory. The Central Limit Theorem states, in its complete and proper full form, that for a sufficiently large sample size, the sampling distribution of the sample mean X-bar is approximately normally distributed with mean mu and variance sigma-squared divided by n, and this remarkable approximate normality result holds true entirely regardless of whatever particular shape or form the original underlying population's own probability distribution actually happens to take -- the original population could be heavily and noticeably skewed, could be flat and uniform in shape, could be distinctly bimodal with two separate peaks, or could take on virtually any other conceivable non-normal shape whatsoever, and yet the resulting sampling distribution of its sample mean will nevertheless still end up looking approximately bell-shaped and normal, provided only that the sample size being used is sufficiently large. Moreover, the overall quality and accuracy of this particular normal approximation itself continues to improve steadily and progressively for larger and larger sample sizes, meaning that even population distributions that are quite dramatically and severely non-normal in their original underlying shape will nevertheless still yield a sampling distribution of X-bar that already looks remarkably close to normal once the sample size employed becomes reasonably large, generally taken in most practical contexts to be somewhere in the vicinity of thirty or more individual observations. The overwhelming practical importance of this single specific result, considered across the whole of statistics as a discipline, is genuinely difficult to overstate. Because the Central Limit Theorem removes entirely the otherwise crucially restrictive assumption of Theorem 11.5 that the original population itself must already be normally distributed, it consequently allows the standardized variable Z, defined as the quantity (X-bar minus mu) divided by (sigma divided by the square root of n), to be legitimately treated as approximately following a standard normal distribution for large samples drawn from virtually ANY kind of population distribution whatsoever, however irregular or unusual its original shape might happen to be. It is precisely this single specific fact, considered in isolation, that ultimately makes it genuinely possible to construct valid

confidence intervals for an unknown population mean, and equally to conduct valid hypothesis tests concerning that same unknown population mean, using the exact same standard normal-distribution-based methodology explored in full technical detail back in Chapter 10 -- and this crucial theoretical connection is exactly the specific reason why the Central Limit Theorem is so very widely and consistently regarded, by working statisticians and instructors alike, as constituting the single most important and far-reaching theoretical bridge to be found anywhere within the whole of this particular statistics course, directly connecting the probability theory developed earlier in the book to the practical methods of statistical inference on estimation and hypothesis testing that immediately follow in the two chapters yet to come.

Multiple Choice Questions (MCQs)

A parameter differs from a statistic in that a parameter is: (A) A random variable (B) A constant computed from the entire population (C) Always unknown (D) Computed only from a sample

Correct answer: (B) A constant computed from the entire population. A parameter is a fixed, constant value computed from the whole population; a statistic is a random variable computed from a sample.

Using Theorem 11.1, the number of possible samples of size n=2 drawn WITH replacement from a population of N=5 is: (A) 10 (B) 20 (C) 25 (D) 5

Correct answer: (C) 25. With replacement: N^n = 5^2 = 25.

Using Theorem 11.1, the number of possible samples of size n=2 drawn WITHOUT replacement from a population of N=5 is: (A) 25 (B) 20 (C) 10 (D) 5

Correct answer: (B) 20. Without replacement: NPn = N(N-1) = 5(4) = 20.

The finite population correction factor (N-n)/(N-1) is: (A) Always greater than 1 (B) Always equal to 1 (C) Always less than or equal to 1 (when n<N) (D) Always negative

Correct answer: (C) Always less than or equal to 1 (when n<N). Since n < N in practical sampling, (N-n)/(N-1) is always <= 1, which is why sampling without replacement reduces the standard error.

Sampling error, unlike non-sampling error, arises: (A) Only in a complete census (B) Purely because a sample rather than the whole population was studied (C) From data-entry mistakes (D) From poorly worded questions

Correct answer: (B) Purely because a sample rather than the whole population was studied. Sampling error exists purely because only a sample (not the full population) is observed, and generally decreases as sample size increases.

Precision of an estimator is measured by its: (A) Bias (B) Mean (C) Standard error (D) Population size

Correct answer: (C) Standard error. Precision refers to how tightly repeated estimates cluster together, measured by the standard error of the estimator.

By Theorem 11.2, the mean of the sampling distribution of X-bar equals: (A) sigma^2/n (B) mu (C) 0 (D) sigma

Correct answer: (B) mu. Theorem 11.2 states mu_Xbar = E(X-bar) = mu, regardless of sample size or replacement method.

The Central Limit Theorem states that for a large sample size, the sampling distribution of X-bar is approximately normal: (A) Only if the population is already normal (B) Regardless of the shape of the population distribution (C) Only for infinite populations (D) Only when sampling without replacement

Correct answer: (B) Regardless of the shape of the population distribution. The power of the CLT is precisely that it holds for large n regardless of the population's original shape.

For sampling with replacement, the variance of the sampling distribution of the difference between two independent sample means X1-bar - X2-bar is: (A) sigma1^2/n1 - sigma2^2/n2 (B) sigma1^2/n1 + sigma2^2/n2 (C) (sigma1^2+sigma2^2)/(n1+n2) (D) sigma1.sigma2/(n1.n2)

Correct answer: (B) sigma1^2/n1 + sigma2^2/n2. Variances of independent random variables ADD even when taking a difference: Var(X1-bar - X2-bar) = sigma1^2/n1 + sigma2^2/n2 (Theorem 11.8).

The mean of the sampling distribution of the sample proportion P equals: (A) n (B) pi(1-pi) (C) pi (D) N/n

Correct answer: (C) pi. Theorem 11.11 states mu_P = E(P) = pi, the true population proportion.

Quick Revision Summary

- Population = whole group (parameter, constant); Sample = part studied (statistic, random variable)
- Sampling frame = list of all sampling units; probability sampling gives every unit a known chance of selection
- Theorem 11.1: with replacement N^n samples; without replacement NPn = N(N-1)...(N-n+1) samples
- Sampling error: T-theta, shrinks with larger n | Non-sampling error: from bad frames/questions/entry, doesn't shrink with n
- Accuracy = closeness to true value | Precision = closeness of repeated estimates (measured by SE) | Bias = E(T)-theta
- Thm 11.2-11.3: mu_Xbar=mu; sigma_Xbar^2=sigma^2/n (with replacement)
- Thm 11.4: without replacement, multiply variance by fpc = (N-n)/(N-1), always <=1
- Thm 11.6 (CLT): for large n, X-bar ~approx N(mu, sigma^2/n) regardless of population shape
- Difference of means: mu=mu1-mu2, sigma^2=sigma1^2/n1+sigma2^2/n2 (variances ADD for independent samples)
- Sample proportion: mu_P=pi, sigma_P^2=pi(1-pi)/n [x fpc if without replacement]

Exam Tips

- Always identify FIRST whether sampling is with or without replacement -- it changes both the counting formula (N^n vs NPn) and the variance formula (fpc or not)
- When a question says 'infinite population' or doesn't mention N, treat it as sampling with replacement -- no fpc needed
- For difference-of-means or difference-of-proportions problems, remember variances ADD even though you're computing a difference

- Build the sampling distribution table systematically: list all samples, compute the statistic for each, then tally into a frequency/probability table before computing mean and variance
- The CLT applies to means and proportions for LARGE n -- don't assume normality for small samples unless the population itself is stated to be normal
- Double-check verification questions (e.g., 'verify mu_Xbar = mu') by computing both sides independently and confirming they match numerically

Chapter 12: Estimation

Chapter 11 built the theoretical machinery of sampling distributions -- how a statistic like the sample mean or sample proportion varies from sample to sample, and how the Central Limit Theorem lets us treat that variation as approximately normal for large samples. Chapter 12 puts that machinery to work for the first time: this is where statistics actually starts answering real questions about an unknown population using nothing but sample data. Statistical inference has two main branches -- estimation (this chapter) and hypothesis testing (the next chapter) -- and both rest entirely on the sampling-distribution theory just developed.

This chapter covers both flavours of estimation: point estimation, which produces a single best-guess number for an unknown parameter, and interval estimation, which produces a range of plausible values together with a stated level of confidence. Interval estimation -- confidence intervals -- is the centrepiece of the chapter, covering the population mean, the population proportion, and the difference between two means or two proportions, across a range of practical situations (known/unknown variance, large/small samples, normal/non-normal populations).

Learning Objectives

- Distinguish statistical inference, estimation, and hypothesis testing, and explain the role each plays
- Distinguish a point estimator from a point estimate, and construct point estimators for the mean, variance, and proportion
- Explain unbiasedness and determine whether a given estimator is biased or unbiased
- Identify the best (minimum variance unbiased) estimator for common population parameters
- Construct pooled estimators for the mean, variance, and proportion from two samples

- Explain what a confidence interval is and interpret the meaning of the confidence coefficient
- Construct confidence intervals for a population mean under known and unknown variance, and for large and small samples (Z and t)
- Construct confidence intervals for a population proportion
- Construct confidence intervals for the difference between two population means and two population proportions

Key Concepts

12.1-12.2 Statistical Inference and Estimation

Statistical inference is the field concerned with drawing conclusions about a population's distribution by using observed sample data governed by that distribution. It has two main branches: estimation of parameters (determining likely values for an unknown population parameter) and testing of hypotheses (deciding between competing claims about a parameter, covered in the next chapter). Statistical estimation is the procedure of making a judgment about the unknown value of a population parameter using sample observations -- since examining the entire population is usually impracticable, we instead estimate its parameters from a sample, with the accuracy of the estimate controllable in part by increasing the sample size.

Estimation is further divided into two types. Point estimation produces a single number, intended as the best guess for the unknown parameter. Interval estimation instead produces a range of values, together with a stated degree of confidence that the range actually contains the true parameter -- and it is interval estimation that gives us information about the PRECISION of an estimate, something a single point estimate can never convey on its own.

12.3-12.4 Point Estimator, Point Estimate, and Unbiasedness

A point estimator is a sample statistic used to estimate the unknown true value of a population parameter -- it is always a function of the sample observations and, being computed from a random sample, is itself a random variable with its own probability distribution; it is denoted by a capital letter such as T. A point estimate is the specific numerical value of the estimator computed from one

particular observed sample -- a fixed number, denoted by a small letter such as t. This distinction (estimator = random variable / formula; estimate = the actual number you get) is fundamental and recurs throughout the chapter.

Unbiasedness is the most important desirable property of an estimator: the distribution of a good estimator should be centred at the true value of the parameter it estimates. An estimator T is unbiased for a parameter theta if E(T) = theta -- meaning that, averaged across every possible sample, T gives the correct value with no systematic tendency to over- or under-estimate. If E(T) is not equal to theta, T is biased, and the amount of the bias is Bias = E(T) - theta. The sample mean X-bar is always an unbiased estimator of the population mean mu. Interestingly, the sample variance S-squared = [sum(X_i - X-bar)^2]/n (divide by n) is a BIASED estimator of sigma-squared -- its expected value is [(n-1)/n].sigma-squared, slightly less than the truth -- which is exactly why the alternative version S-hat-squared = [sum(X_i - X-bar)^2]/(n-1) (divide by n-1) is preferred: it is unbiased, E(S-hat-squared) = sigma-squared.

12.5-12.6 Best Estimator and Pooled Estimators

Among all unbiased estimators of a parameter, the best (or minimum variance) estimator is the one whose sampling distribution has the smallest variance -- i.e., no other unbiased estimator is more precise. For a population with unknown mean mu and unknown variance sigma-squared, the best estimators are the sample mean X-bar (for mu) and the unbiased sample variance S-hat-squared = [sum(X_i - X-bar)^2]/(n-1) (for sigma-squared). For a population with unknown proportion pi, the best estimator is the sample proportion P = X/n.

When two independent samples are available from the SAME population, their information can be combined into a single, more precise pooled estimator. The pooled estimator of the mean is X-bar-p = (n1.X-bar-1 + n2.X-bar-2)/(n1+n2) -- essentially a weighted average of the two sample means, weighted by sample size. The pooled estimator of the variance is S-p-squared = [(n1-1).S-hat-1-squared + (n2-1).S-hat-2-squared]/(n1+n2-2) -- a weighted average of the two sample variances. The pooled estimator of a proportion is pi-hat = (X1+X2)/(n1+n2) = (n1.P1 + n2.P2)/(n1+n2).

12.7 The Concept of Interval Estimation

A point estimate alone tells you nothing about how far it might be from the truth -- however good the estimator, a single number can never be expected to exactly equal the parameter, and looking at just that one value gives no sense of its precision. Interval estimation solves this by extending the idea of an error bound to produce a whole INTERVAL of plausible values, likely (with a stated, known probability) to contain the true parameter. This is the concept of the confidence interval: an interval (L, U), computed from sample data, such that BEFORE sampling, P(L < theta < U) = 1 - alpha for a specified high probability 1 - alpha.

The interval (L, U) is then called a 100(1-alpha)% confidence interval for theta, and 1-alpha is called the confidence coefficient (or confidence level). L and U -- the lower and upper confidence limits -- are themselves random variables, since they are computed from sample data; theta itself is the fixed, constant, unknown quantity. The width U-L measures the precision of the estimate: a narrower interval is more precise. Precision can be increased either by decreasing the standard error of the estimate (i.e., increasing the sample size) or by decreasing the confidence coefficient -- but these two goals pull in opposite directions: for any fixed sample size, demanding higher confidence necessarily widens the interval. It is crucial to interpret a confidence interval correctly: 1-alpha is the long-run proportion of intervals (across many repeated samples) that would contain theta -- it is NOT the probability that theta lies in this one particular already-computed interval, since theta is a fixed constant, not a random variable.

12.8 Confidence Interval for the Population Mean, mu

The exact formula for a confidence interval for mu depends on three things: whether the population is normal, whether sigma-squared is known, and whether the sample size is large or small. When the population is normal with KNOWN variance sigma-squared, the sampling distribution of X-bar is exactly normal, so the two-sided 100(1-alpha)% confidence interval is x-bar +/- z_(1-alpha/2). (sigma/sqrt(n)), where z_(1-alpha/2) is the standard normal value cutting off alpha/2 in the upper tail.

When the population is NOT normal (or its shape is unknown), but the sample is large (conventionally n>30), the Central Limit Theorem justifies treating X-bar as approximately normal regardless of the population's shape, giving the same-looking approximate interval x-bar +/- z_(1-alpha/2).(sigma/sqrt(n)) -- and when sigma is also unknown (the usual real-world case), it is simply replaced by the sample estimate s-hat, since for large n this substitution barely affects the result: x-bar +/- z_(1-alpha/2).(s-hat/sqrt(n)). When sampling is done WITHOUT replacement from a finite population and the sampling fraction n/N exceeds 5%, the finite population correction is added: x-bar +/- z_(1-alpha/2).(sigma/sqrt(n)).sqrt[(N-n)/(N-1)] -- this fpc can be safely ignored when n is less than 5% of N.

When the population IS normal but sigma-squared is UNKNOWN and the sample is SMALL (n<=30), the statistic T = (X-bar - mu)/(S-hat/sqrt(n)) follows Student's t-distribution with v = n-1 degrees of freedom rather than the standard normal distribution -- reflecting the extra uncertainty introduced by estimating sigma from a small sample. The resulting confidence interval is x-bar +/- t_(v,1-alpha/2).(s-hat/sqrt(n)), using the t-table instead of the Z-table. For large degrees of freedom, the t-distribution converges to the standard normal, so the two approaches agree as n grows.

12.9 Confidence Interval for the Population Proportion, pi

When estimating a population proportion pi (e.g., the fraction of defective items, or the fraction of voters favouring a candidate), the sampling distribution of the sample proportion P=X/n is approximately normal for large n (with pi not too close to 0 or 1), by the same Central Limit Theorem logic. The resulting two-sided 100(1-alpha)% confidence interval is p +/- z_(1-alpha/2).sqrt[p(1-p)/n], where p is the observed sample proportion. As with the mean, if sampling is done without replacement from a finite population and n exceeds 5% of N, the finite population correction sqrt[(N-n)/(N-1)] is multiplied in; otherwise it may be dropped.

12.10-12.11 Comparative Studies and CI for the Difference Between Two Means

Many practical questions involve comparing two populations rather than describing just one -- is a new fertilizer more effective than the old one, does a

new machine pack more precisely than an old one, do two departments differ in output. Two samples used for such comparisons are independent if the selection from one population has no bearing on the selection from the other; they are dependent (or paired/matched) if each observation in one sample is deliberately paired with a corresponding observation in the other.

For two INDEPENDENT samples from normal populations with KNOWN variances (any sample size), X-bar-1 - X-bar-2 is exactly normal with mean mu1-mu2 and variance sigma1-squared/n1 + sigma2-squared/n2, giving the confidence interval (x-bar-1 - x-bar-2) +/- z_(1-alpha/2).sqrt[sigma1^2/n1 + sigma2^2/n2]. When both samples are large (n1,n2 > 30), the same formula works approximately for ANY population shape (by the CLT), with the sample variances s-hat-1-squared and s-hat-2-squared substituted for the unknown sigma1-squared and sigma2-squared.

When both samples are SMALL, populations are normal, and the two (unknown) population variances are assumed EQUAL, a pooled estimate of the common variance is first computed, S-p-squared = [(n1-1).S-hat-1^2 + (n2-1).S-hat-2^2]/(n1+n2-2), and then the confidence interval uses the t-distribution with v=n1+n2-2 degrees of freedom: (x-bar-1 - x-bar-2) +/- t_(v,1-alpha/2).s-p.sqrt[1/n1 + 1/n2].

12.12 Confidence Interval for the Difference Between Two Proportions

To compare the incidence rate of some characteristic across two populations (e.g., defect rates from two production lines, approval rates in two regions), independent samples of sizes n1 and n2 give sample proportions P1=X1/n1 and P2=X2/n2. For large samples, the CLT ensures P1-P2 is approximately normal with mean pi1-pi2 and (estimated) standard error sqrt[P1(1-P1)/n1 + P2(1-P2)/n2], giving the two-sided confidence interval (p1-p2) +/- z_(1-alpha/2).sqrt[p1(1-p1)/n1 + p2(1-p2)/n2].

Important Definitions

What is the difference between a point estimator and a point estimate?: A point estimator is a sample statistic (a formula/random variable, e.g., X-bar)

used to estimate a parameter; a point estimate is the specific numerical value that estimator takes for one particular observed sample.

What does it mean for an estimator to be unbiased?: An estimator T is unbiased for a parameter theta if E(T) = theta -- i.e., averaged across all possible samples, it gives the correct value with no systematic over- or under-estimation.

What is bias?: The amount by which a biased estimator's expected value differs from the true parameter: Bias = E(T) - theta.

What is the best (minimum variance) estimator?: Among all unbiased estimators of a parameter, the one whose sampling distribution has the smallest variance -- i.e., the most precise unbiased estimator available.

What is a pooled estimator?: An estimator that combines information from two independent samples of the same population into a single, more precise estimate, e.g., the pooled sample mean or pooled sample variance.

What is a confidence interval?: An interval (L,U), computed from sample data, such that prior to sampling, P(L<theta<U) = 1-alpha for a specified high probability 1-alpha, called the confidence coefficient.

What is the confidence coefficient?: The probability 1-alpha that the confidence interval procedure produces an interval containing the true parameter; commonly 90%, 95%, 98%, or 99%.

How should a confidence interval be correctly interpreted?: As a long-run statement: in repeated sampling, 100(1-alpha)% of all such intervals constructed would contain the true parameter -- it is NOT the probability that the parameter lies in this one specific already-computed interval.

What is the t-distribution used for in this chapter?: For constructing confidence intervals for a mean when the population is normal, the variance is unknown, and the sample size is small (n<=30); it accounts for the extra uncertainty of estimating sigma from a small sample.

What is the finite population correction and when is it used?: The factor sqrt[(N-n)/(N-1)], applied to the standard error when sampling without

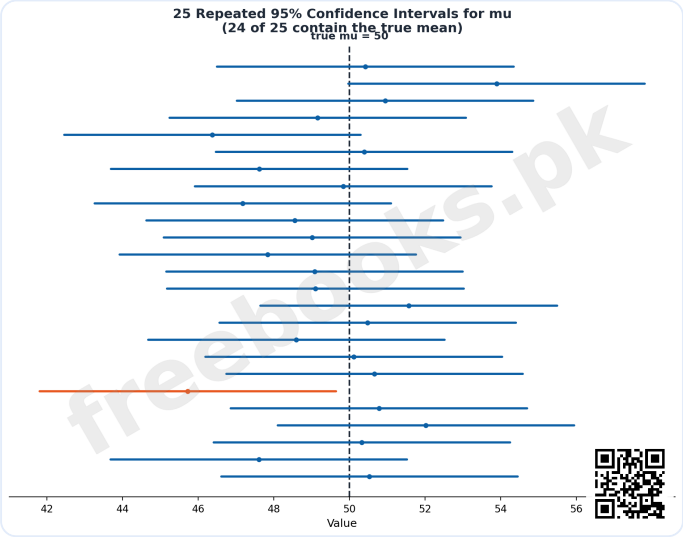
replacement from a finite population and the sample is more than 5% of the population (n>0.05N); it may be ignored otherwise.

Key Facts and Relations

Topic	Key Fact / Relation
Point estimators (mean, variance, proportion)	X-bar = (sum Xi)/n; S-hat^2 = [sum(Xi-Xbar)^2]/(n-1); P = X/n
Unbiasedness / Bias	Unbiased: E(T) = theta. Bias = E(T) - theta
Pooled mean / variance / proportion	Xbar_p = (n1.Xbar1+n2.Xbar2)/(n1+n2); Sp^2 = [(n1-1)S1^2+(n2-1)S2^2]/(n1+n2-2); pi-hat = (X1+X2)/(n1+n2)
CI for mu, sigma known	x-bar +/- z_(1-a/2) . sigma/sqrt(n)
CI for mu, large n, sigma unknown	x-bar +/- z_(1-a/2) . s-hat/sqrt(n)
CI for mu, without replacement (fpc)	x-bar +/- z_(1-a/2) . (sigma or s-hat)/sqrt(n) . sqrt[(N-n)/(N-1)]
CI for mu, small n, normal, sigma unknown	x-bar +/- t_(v,1-a/2) . s-hat/sqrt(n), v = n-1
CI for proportion, pi	p +/- z_(1-a/2) . sqrt[p(1-p)/n]
CI for mu1-mu2, known variances / large samples	(xbar1-xbar2) +/- z_(1-a/2) . sqrt[sigma1^2/n1 + sigma2^2/n2]
CI for mu1-mu2, small samples, equal unknown variance	(xbar1-xbar2) +/- t_(v,1-a/2) . sp.sqrt[1/n1+1/n2], v=n1+n2-2
CI for pi1-pi2	(p1-p2) +/- z_(1-a/2) . sqrt[p1(1-p1)/n1 + p2(1-p2)/n2]
Confidence coefficient breakdown	1-alpha = confidence level; alpha/2 in each tail; common z-values: 90%->1.645, 95%->1.960, 98%->2.326, 99%->2.576

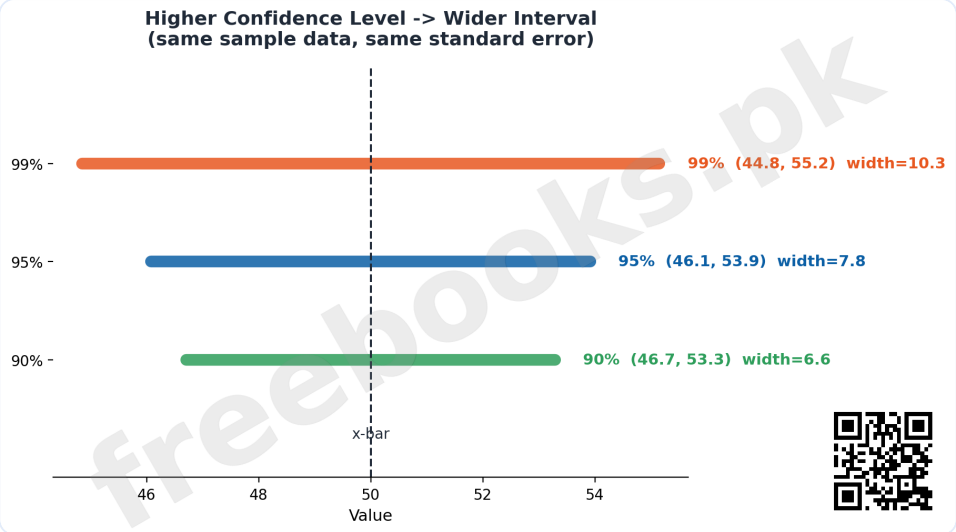
Diagrams

Repeated Confidence Intervals for mu: A simulation of 25 different 95% confidence intervals, each computed from a different random sample of the same population, plotted as horizontal line segments against a fixed vertical line at the true population mean -- illustrating that about 95% of such intervals capture the true mean while a few, purely by chance, miss it



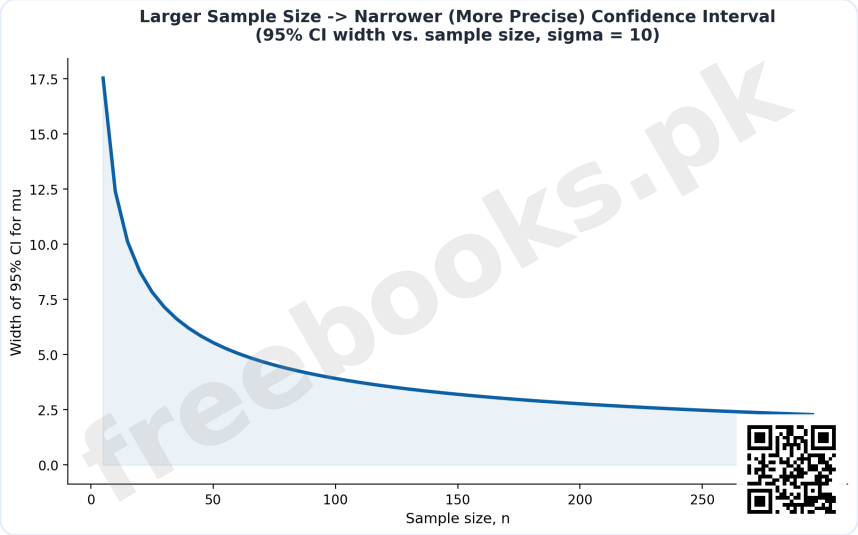
Repeated Confidence Intervals for mu

Effect of Confidence Level on Interval Width: Three nested confidence intervals (90%, 95%, 99%) computed from the same sample data and centred at the same sample mean, showing how demanding a higher confidence level necessarily widens the interval



Effect of Confidence Level on Interval Width

Effect of Sample Size on Interval Width: A line chart showing how the width of a 95% confidence interval for the mean shrinks as the sample size n increases, illustrating that larger samples give more precise (narrower) interval estimates



Effect of Sample Size on Interval Width

Short Questions & Answers

Distinguish between a point estimate and an interval estimate.

A point estimate is a single number intended as the best guess for a parameter and gives no information about its precision; an interval estimate is a range of values, together with a stated confidence level, that conveys both an estimate and its precision.

Why is the sample variance S^2 = sum(Xi-Xbar)^2/n considered a biased estimator, and what is used instead?

Because E(S^2) = [(n-1)/n].sigma^2, which is systematically less than the true sigma^2; the unbiased alternative S-hat^2 = sum(Xi-Xbar)^2/(n-1), which satisfies E(S-hat^2)=sigma^2, is used instead.

What does a 95% confidence interval actually mean?

That if the same sampling and interval-construction procedure were repeated many times, approximately 95% of the resulting intervals would contain the true population parameter -- not that there is a 95% chance the parameter is in this one specific interval.

How can the precision (width) of a confidence interval be improved?

By increasing the sample size (which reduces the standard error) or by decreasing the confidence coefficient (accepting a lower level of confidence) -- these two levers work in opposite directions.

When should the t-distribution be used instead of the standard normal (Z) distribution for a confidence interval on the mean?

When the population is normal, the population variance is unknown, and the sample size is small (typically n<=30); the t-distribution accounts for the added uncertainty of estimating sigma from limited data.

When is the finite population correction factor included in a confidence interval formula, and when can it be dropped?

It is included when sampling without replacement and the sample is more than 5% of the population (n>0.05N); it can safely be dropped when the sample is a small fraction of the population (n<0.05N).

Long Questions & Answers

Explain the concepts of point estimation and unbiasedness in detail, including the distinction between an estimator and an estimate, and explain with an example why the sample variance divided by n is a biased estimator while the version divided by n-1 is unbiased.

Point estimation forms the conceptual starting point of the entire theory of statistical estimation, and grasping it properly requires carefully distinguishing between two closely related but genuinely different ideas that are extremely easy to conflate in casual usage: the estimator itself, and any particular estimate produced by that estimator. A point estimator is formally defined as a sample statistic used to estimate the unknown true value of some population parameter, and crucially, an estimator is always expressed as a specific mathematical function of the sample observations themselves, written in the general form T equals g of X-one, X-two, and so on through X-n. Because this function is applied to a RANDOM sample, whose particular observed values will necessarily differ from one sampling occasion to the next, the estimator T is itself properly understood as a random variable, possessing its very own complete probability distribution -- namely, the sampling distribution explored in exhaustive detail throughout the whole of the immediately preceding chapter. By long-standing

statistical convention, an estimator is denoted using a capital letter, such as T or U, precisely to emphasize this random-variable status. A point estimate, by clear contrast, is the single specific numerical value that this same estimator happens to take on once one particular sample has actually been drawn and observed in practice -- it is calculated using the exact identical formula as the estimator, but is instead applied to the concrete observed numbers x-one, x-two, and so forth rather than to the abstract random variables X-one, X-two, and so on, yielding a single fixed constant conventionally denoted using a lowercase letter such as t. To make this crucial estimator-versus-estimate distinction fully concrete: the sample mean formula X-bar equals the sum of the X-i's divided by n is the ESTIMATOR of the unknown population mean mu, while the specific number 28 obtained after actually computing that formula on one particular observed sample is the ESTIMATE. Once this foundational distinction between estimator and estimate has been firmly established, the single most important theoretical property that any good point estimator should ideally possess is what statisticians call unbiasedness. The underlying intuition is that the probability distribution of a genuinely good estimator ought to be properly centred, in some precise and well-defined statistical sense, directly around the true value of whatever parameter that estimator is attempting to estimate, and because the expected value of a random variable is precisely the standard mathematical way of describing the centre of its distribution, this intuitive requirement translates directly into the formal mathematical condition that E of T should equal theta, where theta represents the true, generally unknown parameter value being estimated. An estimator T satisfying this particular condition is formally termed unbiased, meaning that, if one could somehow average the resulting point estimate across every single conceivable sample that could possibly have been drawn from the population, that average would land exactly precisely on the true parameter value, entirely without any systematic tendency whatsoever to consistently overestimate or consistently underestimate that true value. Should an estimator instead fail to satisfy this defining condition, so that E of T is not equal to theta, that particular estimator is instead termed biased, and the specific numerical amount of this systematic discrepancy is formally called its bias, calculated directly as Bias equals E of T minus theta. A particularly instructive and genuinely important illustration of unbiasedness in action, one that recurs constantly throughout the whole of this chapter and indeed

throughout the whole of subsequent statistical theory, concerns the two competing formulas commonly used for calculating a sample's variance. The first of these two formulas, conventionally written S-squared and calculated as the sum of the squared deviations (X-i minus X-bar) all divided by the sample size n, might initially seem like the entirely natural and obvious choice, since it exactly parallels the corresponding formula used for calculating the TRUE population variance sigma-squared. However, it can rigorously be proven, and is furthermore demonstrated very concretely and directly through the specific worked textbook example using the small five-value population 2, 4, 6, 8, and 10, that the expected value of this particular divide-by-n version of sample variance works out to E of S-squared equals the quantity (n-1) divided by n, all multiplied by the true population variance sigma-squared -- a resulting value that is always, without a single exception, strictly LESS than sigma-squared itself for any finite sample size n, meaning this particular version of sample variance is provably and systematically biased, in fact biased in a very specific and entirely predictable downward direction. The clear underlying reason behind this systematic downward bias is that the sample mean X-bar, purely by its own defining mathematical construction as the specific value that itself minimizes the sum of squared deviations from any single sample's own data points, will always fit that PARTICULAR sample's own data at least marginally more closely than the true, but of course still fundamentally unknown, population mean mu ever possibly could, and this subtle but entirely real overfitting effect systematically and predictably shrinks the resulting calculated sum of squared deviations, and therefore correspondingly the resulting calculated variance as well, to a value that runs slightly too low, on average, across repeated sampling. The specific alternative formula, conventionally written S-hat-squared and calculated instead as that exact same sum of squared deviations but this time divided by the quantity (n-1) rather than simply by n, was deliberately introduced and adopted for one precise purpose: to correct for and thereby entirely eliminate this systematic downward bias just described. It can indeed be rigorously proven, and is furthermore likewise demonstrated very concretely through that exact same worked textbook example, that the expected value of this particular divide-by-(n-1) version instead works out to E of S-hat-squared equals sigma-squared exactly, precisely and completely with no remaining systematic bias whatsoever. It is exactly this specific proven property of complete unbiasedness that explains

why S-hat-squared, rather than the seemingly more natural-looking plain divide-by-n version S-squared, is the particular formula that is consistently and specifically singled out and recommended for use as the preferred BEST estimator of an unknown population variance throughout the whole of the remainder of this chapter and, more broadly, throughout essentially the whole of the wider discipline of statistical estimation theory.

Explain what a confidence interval is, how the confidence coefficient should be correctly interpreted, and describe how the choice of formula for a confidence interval on the population mean changes depending on whether the population variance is known, whether the sample is large or small, and whether the population itself is normal.

A single point estimate, no matter how carefully and rigorously it has been calculated, suffers from one deeply fundamental and entirely unavoidable practical limitation: it provides absolutely no built-in information whatsoever regarding its own precision, meaning it gives no direct sense at all of just how large or how small the inevitable margin of sampling error attached to that estimate is actually likely to be. Confidence interval estimation was specifically developed by statisticians as the direct, principled solution to this exact particular limitation. Formally, a confidence interval for some unknown population parameter theta is defined as an interval, conventionally written using the two endpoints L and U, that is calculated directly from the observed random sample data in such a specific way that, importantly, BEFORE that sample was ever actually drawn, the probability that this particular calculated interval will end up containing the true unknown value of theta is equal to some specified, deliberately chosen high probability, conventionally written as 1 minus alpha. This particular quantity 1 minus alpha carries its own specific, dedicated statistical name: it is called the confidence coefficient, or equivalently and interchangeably the confidence level, and in ordinary everyday practical usage, this same confidence coefficient is very frequently instead expressed simply as an easily interpretable percentage figure, such as the specific values 90 percent, 95 percent, 98 percent, or 99 percent, all four of which happen to be particularly common and frequently encountered choices across real-world applications. Correctly and precisely interpreting exactly what this confidence coefficient

genuinely does and does not actually mean turns out to be one of the single most consistently misunderstood, yet also one of the single most fundamentally and critically important, conceptual points found absolutely anywhere within the whole of introductory statistics. The single, absolutely crucial technical point to properly grasp here is that theta itself, the true underlying population parameter under active investigation, is always and forever a single FIXED, constant, unchanging quantity -- theta itself is simply not literally free to randomly vary or randomly move around from one sampling occasion to another. It is instead specifically the two calculated endpoints L and U themselves, rather than theta, that are the genuinely random quantities under discussion here, precisely because both L and U are calculated directly as specific functions of whatever particular random sample data happened to actually be observed and collected on that given occasion. Consequently, the technically correct and fully precise interpretation of, say, a stated 95 percent confidence interval must always necessarily be framed and expressed in strictly LONG-RUN, repeated-sampling terms: if the identical underlying sampling and interval-construction procedure being used here were to be repeated over and over again a very large number of times, then approximately 95 percent of all of the many resulting intervals produced across all of those many repetitions would successfully turn out to contain the true, fixed value of theta, while the small remaining 5 percent, purely as an unavoidable and inescapable consequence of the ordinary operation of pure random chance, unfortunately would not manage to capture it. It would consequently be a clear and definite technical error to instead claim or assert that there specifically exists a 95 percent probability that theta itself happens to lie somewhere within this one single, particular, already fully computed interval, precisely because, once that specific interval has actually already been computed from already-observed sample data, it has then already either successfully captured the fixed value of theta, or else it has definitively failed to do so -- there simply remains no further meaningful randomness left over at that particular later point for any further probability statement to sensibly attach itself to. The precision of any given confidence interval, meaning specifically how narrow or how wide that resulting interval actually turns out to be, is very directly and immediately governed by the interval's own overall total width, calculated simply as U minus L, and this particular width can only ever be reduced, thereby correspondingly improving the resulting estimate's overall precision, through the

deliberate application of one or the other of exactly two possible available levers: either by actively increasing the sample size actually used, which correspondingly reduces the underlying standard error, or alternatively by deliberately choosing to accept a lower, less demanding stated confidence coefficient, which of course simultaneously and unavoidably lowers the overall associated reliability of that resulting interval at exactly the very same time. When it specifically comes to constructing a confidence interval targeting the population mean mu in particular, the single specific formula that should properly be applied in any given situation depends critically and directly upon three quite separate, entirely distinct underlying considerations, considered carefully and specifically together as a full and complete package. The first and most straightforward specific case to properly handle arises whenever the underlying population itself is already known with confidence to be normally distributed, AND its own true variance sigma-squared happens to already be precisely and exactly known in advance -- under this particular specific combination of conditions, the sampling distribution of the sample mean X-bar can then be shown to be EXACTLY normally distributed for literally any sample size whatsoever, even including the extreme case of a sample size as small as just one single solitary observation, which correspondingly yields the fully exact confidence interval formula x-bar plus or minus z-sub-(1 minus alpha over 2) multiplied by sigma divided by the square root of n. The second broad specific case that commonly and frequently arises in genuine real-world practical situations instead occurs whenever the population's own true variance sigma-squared is NOT already precisely known in advance, which is of course overwhelmingly the far more typical and realistic situation actually encountered when working with genuine real-world data -- but provided that the sample size being used is nevertheless still sufficiently large, conventionally understood in practice to specifically mean larger than 30 individual observations, the Central Limit Theorem developed extensively back in the immediately preceding chapter then steps directly in to fully justify treating the sample mean X-bar as being at least approximately normally distributed, and this important justification furthermore continues to hold true regardless of whatever particular shape the underlying original population distribution itself actually happens to take, meaning it is not even strictly necessary for the original population to be normal at all in this particular large-sample case; the unknown sigma is then simply and

directly replaced within that same formula by its own best available unbiased sample estimate, s-hat, yielding the resulting large-sample APPROXIMATE confidence interval x-bar plus or minus z-sub-(1 minus alpha over 2) multiplied by s-hat divided by the square root of n. The third and generally most theoretically demanding specific case of all to properly handle arises specifically whenever the sample size being used is instead small, conventionally taken in practice to specifically mean n less than or equal to 30, AND the population's own true variance sigma-squared remains genuinely unknown at the very same time -- under this particular specific combination of conditions, the Central Limit Theorem's own large-sample justification described immediately above can simply no longer be safely or legitimately invoked or relied upon, and so an additional, further, quite crucial assumption must instead now be explicitly added directly into the mix: that the underlying population itself is indeed genuinely normally distributed. Under this specific full combination of conditions taken together, it can then formally and rigorously be shown that the particular standardized statistic T, defined and calculated as the quantity (X-bar minus mu), all divided by (S-hat divided by the square root of n), follows what is specifically termed Student's t-distribution, carrying v equals n minus 1 associated degrees of freedom, rather than instead simply following the ordinary standard normal distribution that was used throughout both of the two preceding cases already discussed above -- and this particular resulting t-distribution is itself somewhat more spread out and noticeably flatter than the standard normal distribution is, a specific mathematical feature that directly and appropriately reflects the additional layer of statistical uncertainty that is inevitably introduced into any such calculation whenever the true, unknown value of sigma itself must first be separately estimated from what is, by the very definition and terms of this particular case, necessarily only quite limited underlying sample data to begin with, which correspondingly yields the properly matching small-sample confidence interval formula x-bar plus or minus t-sub-(v, 1 minus alpha over 2) multiplied by s-hat divided by the square root of n, calculated using the appropriate t-table value looked up specifically according to the correct number of degrees of freedom rather than instead using the ordinary standard normal Z-table that had already been used throughout each of the two earlier cases discussed above.

Multiple Choice Questions (MCQs)

A point estimator is best described as: (A) A fixed constant (B) A sample statistic that is a random variable (C) The true population parameter (D) A confidence interval

Correct answer: (B) A sample statistic that is a random variable. A point estimator is a formula/statistic computed from sample data; since the sample varies, the estimator is itself a random variable.

An estimator T is unbiased for parameter theta if: (A) E(T) = 0 (B) E(T) = theta (C) Var(T) = 0 (D) T = theta always

Correct answer: (B) E(T) = theta. Unbiasedness means the expected value of the estimator equals the true parameter, E(T)=theta.

The sample variance S^2 = sum(Xi-Xbar)^2/n is: (A) An unbiased estimator of sigma^2 (B) A biased estimator of sigma^2 (C) Equal to sigma^2 always (D) Undefined

Correct answer: (B) A biased estimator of sigma^2. E(S^2) = [(n-1)/n]sigma^2, which is less than sigma^2, so S^2 (divide by n) is biased; the divide-by-(n-1) version is unbiased.

The 'best' estimator among all unbiased estimators is the one with: (A) The largest variance (B) The smallest variance (C) A bias of zero only (D) The largest sample size

Correct answer: (B) The smallest variance. The best (minimum variance) unbiased estimator is the most precise one -- the unbiased estimator with the smallest possible variance.

A 95% confidence interval means: (A) There is a 95% chance the parameter lies in this specific interval (B) 95% of the population lies in the interval (C) In repeated sampling, 95% of such intervals would contain the true parameter (D) The sample mean equals the parameter 95% of the time

Correct answer: (C) In repeated sampling, 95% of such intervals would contain the true parameter. The correct long-run interpretation: 95% of intervals constructed this way, across repeated sampling, would contain the true (fixed) parameter.

Increasing the sample size, holding confidence level fixed, causes a confidence interval's width to: (A) Increase (B) Decrease (C) Stay the same (D) Become undefined

Correct answer: (B) Decrease. A larger sample size reduces the standard error, which narrows (decreases) the confidence interval.

The t-distribution is used instead of the Z-distribution for a confidence interval on the mean when: (A) The sample size is large and sigma is known (B) The population is normal, sigma is unknown, and n is small (C) The population is not normal and n is large (D) Never -- Z is always used

Correct answer: (B) The population is normal, sigma is unknown, and n is small. The t-distribution applies specifically when the population is normal, sigma^2 is unknown, and the sample size is small (n<=30).

For a confidence interval on a population proportion, the standard error is estimated using: (A) sigma/sqrt(n) (B) sqrt[p(1-p)/n] (C) s-hat/sqrt(n) (D) sqrt[sigma1^2/n1+sigma2^2/n2]

Correct answer: (B) sqrt[p(1-p)/n]. The estimated standard error of the sample proportion P is sqrt[p(1-p)/n].

For the difference between two means with small samples and equal but unknown variances, the confidence interval uses: (A) The standard normal distribution only (B) The pooled variance and the t-distribution with v=n1+n2-2 (C) The F-distribution (D) The chi-square distribution

Correct answer: (B) The pooled variance and the t-distribution with v=n1+n2-2. This case uses the pooled variance estimate Sp^2 and the t-distribution with n1+n2-2 degrees of freedom.

The finite population correction factor is applied when: (A) Sampling with replacement from an infinite population (B) Sampling without replacement and the sample exceeds 5% of the population (C) The sample size is always small (D) The population variance is unknown

Correct answer: (B) Sampling without replacement and the sample exceeds 5% of the population. The fpc is applied for sampling without replacement from a finite population when n exceeds 5% of N; it can be dropped otherwise.

Quick Revision Summary

- Estimator (random variable, capital letter) vs Estimate (fixed number, lowercase letter)
- Unbiased: E(T)=theta | Bias = E(T)-theta | Best estimator = unbiased with minimum variance
- Best estimators: mu -> X-bar; sigma^2 -> S-hat^2 (divide by n-1); pi -> P=X/n
- Pooled mean: weighted average by n1,n2 | Pooled variance: Sp^2=[(n1-1)S1^2+(n2-1)S2^2]/(n1+n2-2)
- CI = point estimate +/- (table value) x (standard error); width measures precision
- 1-alpha = confidence coefficient; long-run interpretation, NOT probability for one specific interval
- CI for mu: sigma known -> Z; large n, sigma unknown -> Z with s-hat; small n, normal, sigma unknown -> t (v=n-1)
- CI for pi: p +/- z.sqrt[p(1-p)/n]
- CI for mu1-mu2: known var/large n -> Z; small n equal var -> pooled t (v=n1+n2-2)
- CI for pi1-pi2: (p1-p2) +/- z.sqrt[p1(1-p1)/n1+p2(1-p2)/n2]
- fpc = sqrt[(N-n)/(N-1)], used when sampling without replacement and n>0.05N

Exam Tips

- Always identify FIRST which of the three mean-CI cases applies: sigma known (Z, exact); large n sigma unknown (Z, approx); small n normal sigma unknown (t)
- Never confuse the confidence coefficient (1-alpha) with a probability statement about one already-computed interval -- it's a long-run repeated-sampling statement
- For difference-of-means problems, check whether variances are assumed equal (pooled t) or samples are simply large (Z with separate variances)
- Remember z-values by heart: 90%->1.645, 95%->1.960, 98%->2.326, 99%->2.576 -- these appear constantly

- For small-sample t-interval problems, always compute degrees of freedom v=n-1 (or n1+n2-2 for two samples) BEFORE looking up the table value
- Check the 5% rule (n vs 0.05N) before deciding whether to include the finite population correction factor

Chapter 13: Hypothesis Testing

Chapter 12 answered the question 'what is a plausible range of values for an unknown parameter?' through confidence intervals. This chapter answers a different but closely related question: 'is a specific claimed value of that parameter believable, given what the sample shows?' This is hypothesis testing -- the second and last major branch of statistical inference, and by far the most heavily used tool in applied statistics, business decision-making, and scientific research. It is also the largest chapter in the book (70 pages), because it methodically works through the same battery of situations covered in Chapter 12 -- mean, proportion, difference of two means (independent and paired/dependent samples), difference of two proportions -- but now framed as formal decision procedures rather than interval estimates.

The chapter builds a single, disciplined framework -- state hypotheses, choose a significance level, pick a test statistic, define a rejection region, compute the observed value, and draw a conclusion -- and then applies that exact same framework, unchanged, to every situation that follows. Master the framework once in section 13.1, and the rest of the chapter is just plugging different test statistics (Z or T) into the same seven-step procedure.

Learning Objectives

- Define statistical hypothesis, null hypothesis, and alternative hypothesis, and correctly formulate them from a verbal claim
- Distinguish simple from composite hypotheses, and one-tailed from two-tailed tests
- Define Type-I and Type-II errors, and the level of significance and level of confidence
- List and apply the seven standard steps of a hypothesis test
- Test hypotheses about a population mean using the Z and T test statistics, choosing correctly between them
- Test hypotheses about a population proportion, with and without the continuity correction

- Test hypotheses about the difference between two population means for independent samples (known variance, pooled t, large-sample Z)
- Test hypotheses about the difference between two population means for dependent (paired/matched) samples
- Test hypotheses about the difference between two population proportions

Key Concepts

13.1 The Elements of a Test of Hypothesis

A statistical hypothesis is an assertion or conjecture about the distribution of one or more random variables -- typically phrased quantitatively in terms of a population parameter (e.g., mu=25, pi=0.4). Hypothesis testing is the procedure used to determine whether such an assumption is supported by an observed random sample. The null hypothesis, H0, is the hypothesis tested for possible rejection under the assumption that it is true -- it is always a statement of 'no effect' or 'the status quo' (a coin is unbiased, a drug is ineffective, there is no difference between two methods). The alternative hypothesis, H1, is the hypothesis we are willing to accept if H0 is rejected -- it is often called the research hypothesis, because it expresses what the researcher actually believes and is trying to support with evidence.

Formulating H0 and H1 correctly follows a fixed procedure: first state the alternative hypothesis as whatever the researcher is trying to find support for (an inequality: 'less than', 'greater than', or 'not equal to'), then state the null hypothesis as the complementary claim with an equality sign built in. The three matched pairs are: H0: theta>=theta0 vs H1: theta<theta0 (left-tailed); H0: theta<=theta0 vs H1: theta>theta0 (right-tailed); H0: theta=theta0 vs H1: theta!=theta0 (two-tailed). A simple hypothesis completely specifies the population distribution (both its functional form and every parameter value); a composite hypothesis does not (e.g., specifying only the mean while leaving the variance unspecified or as an inequality).

The test statistic is the sample statistic used to decide whether to reject H0 -- commonly Z, T, chi-square, or F. The rejection (critical) region is the set of test-statistic values for which H0 is rejected; the nonrejection region is the complementary set; the critical values are the boundary points separating them.

A two-tailed test places the rejection region equally in both tails of the sampling distribution (used when H1 is theta!=theta0); a one-tailed test places it entirely in one tail -- a left-tailed test when H1 is theta<theta0 (looking for a definite decrease), and a right-tailed test when H1 is theta>theta0 (looking for a definite increase).

Because a decision is being made from incomplete (sample) information, two kinds of error are possible. A Type-I error is rejecting H0 when H0 is actually true. A Type-II error is accepting (not rejecting) H0 when H1 is actually true. The probability of a Type-I error is denoted alpha and is called the level of significance (also the size of the critical region/test) -- a small, pre-assigned value such as 0.05 or 0.01, fixed BEFORE the sample is drawn so the results cannot bias the choice. The probability of a Type-II error is denoted beta. The level of confidence, 1-alpha, is the probability of correctly accepting a true H0. A hypothesis test concludes either by rejecting H0 in favour of H1 (if the test statistic falls in the rejection region), or by failing to reject H0 (there is insufficient evidence to reject it -- this is NOT the same as proving H0 true).

13.1.25 The Seven Steps of a Hypothesis Test

Every hypothesis test in this chapter follows the identical disciplined sequence. Before looking at the sample: (1) identify the population and state the conditions required for the test's validity; (2) formulate and state H0 and H1; (3) decide and specify the level of significance, alpha; (4) select the appropriate test statistic and its sampling distribution assuming H0 is true; (5) find the critical value(s) of the test statistic for the chosen alpha; (6) establish the rejection region; (7) state the decision rule (reject H0 if the test statistic falls in the rejection region, otherwise do not reject). After the sample is collected: (8) calculate the observed value of the test statistic from the sample; (9) draw the conclusion -- reject or do not reject H0 -- and state it in plain, practical terms.

13.2 Test of Hypothesis About a Population Mean, mu

The three possible hypothesis pairs about mu (against a hypothesised value mu0) are H0: mu>=mu0 vs H1: mu<mu0; H0: mu<=mu0 vs H1: mu>mu0; and H0: mu=mu0 vs H1: mu!=mu0. The choice of test statistic mirrors Chapter 12's confidence-interval logic exactly: when the population is normal with KNOWN

variance sigma^2 (any sample size), use Z = (X-bar - mu0)/(sigma/sqrt(n)), which is exactly standard normal under H0. When the population is normal with UNKNOWN variance and the sample is SMALL (n<=30), use T = (X-bar - mu0)/(S-hat/sqrt(n)), which follows a t-distribution with v=n-1 degrees of freedom under H0. When the sample is LARGE (n>30), the Central Limit Theorem justifies using Z = (X-bar - mu0)/(S-hat/sqrt(n)) as an approximation regardless of the population's shape or whether sigma is known.

In every case the mechanics are the same: compute the observed value of the test statistic from the sample, compare it against the critical value(s) that mark off the rejection region at the chosen alpha, and reject H0 if the observed value falls in that region.

13.3 Test of Hypothesis About a Population Proportion, pi

For a hypothesised proportion pi0, the three hypothesis pairs are analogous to those for the mean. Since the sampling distribution of the sample proportion P=X/n is (for large n) approximately normal with mean pi0 and variance pi0(1-pi0)/n, the test statistic is Z = (P - pi0)/sqrt[pi0(1-pi0)/n], or equivalently in terms of the count X: Z = (X - n.pi0)/sqrt[n.pi0(1-pi0)]. For extra precision, a continuity correction can be applied (since X is discrete but is being approximated by a continuous normal distribution): Z = [(X +/- 0.5) - n.pi0]/sqrt[n.pi0(1-pi0)] -- using a plus sign when X<n.pi0 (or P<pi0) and a minus sign when X>n.pi0 (or P>pi0).

13.4 Difference Between Two Means -- Independent Samples

For comparing mu1 and mu2 via independent samples, against a hypothesised difference delta0, three cases parallel Chapter 12 exactly. Known variances (any sample size, normal populations): Z = [(X-bar-1 - X-bar-2) - delta0] / sqrt[sigma1^2/n1 + sigma2^2/n2], exactly standard normal under H0. Small samples (n1,n2<=30), normal populations, equal but unknown variance: first pool the two sample variances into Sp^2 = [(n1-1).S-hat-1^2 + (n2-1).S-hat-2^2]/(n1+n2-2), then use T = [(X-bar-1 - X-bar-2) - delta0] / [Sp.sqrt(1/n1 + 1/n2)], which follows a t-distribution with v=n1+n2-2 degrees of freedom. Large samples (n1,n2>30), any population shape: the CLT justifies Z = [(X-bar-1 - X-bar-2) - delta0] / sqrt[S-hat-1^2/n1 + S-hat-2^2/n2] as an approximation, with no assumption of equal variances required.

13.5 Difference Between Two Means -- Dependent (Paired) Samples

When observations naturally come in matched pairs (before/after measurements on the same subject, or two products tested on the same set of items), the two samples are dependent rather than independent -- pairing lets each subject act as its own control, filtering out subject-to-subject variability that would otherwise inflate the noise in an independent-samples comparison. For each pair, define the difference D = Y - X (after minus before, or Candidate-A minus Candidate-B, etc.). The population of these differences has mean mu_D = mu2 - mu1 and its own variance sigma_D^2 -- and testing mu_D reduces the two-sample problem to an ordinary ONE-sample test on the differences, using exactly the machinery from section 13.2.

Given n paired differences d1, d2, ..., dn with mean d-bar = (sum d_i)/n and standard deviation s-hat_D = sqrt{[sum(d_i - d-bar)^2]/(n-1)}, the test statistic for H0: mu_D = delta0 (delta0 is usually 0, i.e. 'no difference') is T = (D-bar - delta0)/(s-hat_D/sqrt(n)), following a t-distribution with v=n-1 degrees of freedom. Two assumptions are required: the n differences must be a random sample from the population of differences, and that population of differences must be (approximately) normally distributed -- though this normality assumption can be relaxed for large n (>30) by the Central Limit Theorem, exactly as with a single-sample mean test.

13.6 Difference Between Two Proportions

For comparing pi1 and pi2 from two independent samples (sizes n1, n2, sample proportions P1=X1/n1, P2=X2/n2), two distinct situations arise. To test the GENERAL hypothesis H0: pi1-pi2=delta0 for some non-zero delta0, use Z = [(P1-P2) - delta0] / sqrt[P1(1-P1)/n1 + P2(1-P2)/n2], using the two sample proportions separately in the standard error. To test specifically whether the two populations have the SAME proportion of successes, H0: pi1=pi2 (i.e. delta0=0), it is more efficient to POOL the two samples into a single combined estimate of the common proportion, pi-hat = (X1+X2)/(n1+n2) = (n1.P1+n2.P2)/(n1+n2), and use Z = (P1-P2) / sqrt{pi-hat(1-pi-hat).(1/n1 + 1/n2)} -- exactly mirroring the pooled-variance logic used for two means with equal variance in section 13.4.

Important Definitions

What is a null hypothesis (H0)?: A statement of 'no effect' or 'the status quo' about a population parameter, tested for possible rejection under the assumption that it is true.

What is an alternative hypothesis (H1)?: The hypothesis accepted when H0 is rejected; it expresses the research claim the investigator is trying to support with sample evidence.

What is a Type-I error?: Rejecting the null hypothesis H0 when H0 is actually true; its probability is denoted alpha, the level of significance.

What is a Type-II error?: Failing to reject (accepting) the null hypothesis H0 when the alternative hypothesis H1 is actually true; its probability is denoted beta.

What is the level of significance?: The maximum probability of a Type-I error the researcher is willing to risk, denoted alpha; a small, pre-assigned value (e.g., 0.05 or 0.01) fixed before sampling.

What is the rejection (critical) region?: The set of values of the test statistic for which H0 is rejected in favour of H1.

What is the difference between a one-tailed and a two-tailed test?: A one-tailed test places the entire rejection region in one tail (used when H1 specifies a direction, < or >); a two-tailed test splits the rejection region equally between both tails (used when H1 is !=).

What is a simple hypothesis versus a composite hypothesis?: A simple hypothesis completely specifies the population distribution (functional form and all parameter values); a composite hypothesis leaves at least one parameter unspecified or given as an inequality.

Why are dependent (paired) samples used instead of independent samples?: Pairing lets each subject act as its own control, removing subject-to-subject variability from the comparison and increasing the power to detect a real difference, especially with small samples.

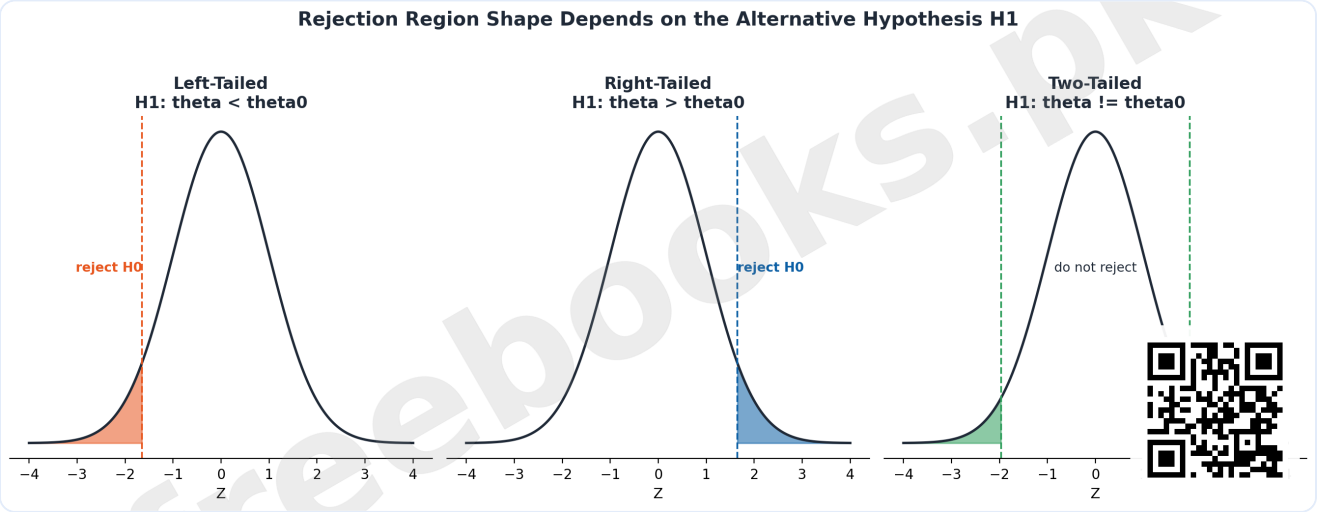
What is the pooled estimate of a common proportion, pi-hat, used for?: To test H0: pi1=pi2 (that two independent populations have the same proportion of successes), by combining both samples' successes into one estimate, pi-hat=(X1+X2)/(n1+n2), used in the standard error of P1-P2.

Key Facts and Relations

Topic	Key Fact / Relation
Test statistic for mu, sigma known / large n	Z = (X-bar - mu0) / (sigma/sqrt(n)) [or S-hat/sqrt(n) if sigma unknown, large n]
Test statistic for mu, small n, normal, sigma unknown	T = (X-bar - mu0) / (S-hat/sqrt(n)), v = n-1
Test statistic for proportion pi	Z = (P - pi0) / sqrt[pi0(1-pi0)/n]
Continuity-corrected Z for proportion	Z = [(X +/- 0.5) - n.pi0] / sqrt[n.pi0(1-pi0)]
Difference of means, known variances	Z = [(Xbar1-Xbar2) - delta0] / sqrt[sigma1^2/n1 + sigma2^2/n2]
Difference of means, small n, pooled variance	T = [(Xbar1-Xbar2) - delta0] / [Sp.sqrt(1/n1+1/n2)], v=n1+n2-2
Pooled variance, Sp^2	Sp^2 = [(n1-1)S1^2 + (n2-1)S2^2] / (n1+n2-2)
Difference of means, large samples	Z = [(Xbar1-Xbar2) - delta0] / sqrt[S1^2/n1 + S2^2/n2]
Paired-sample test (mu_D)	T = (D-bar - delta0) / (S_D-hat/sqrt(n)), v = n-1; D-bar = (sum d_i)/n
Difference of proportions, general delta0	Z = [(P1-P2) - delta0] / sqrt[P1(1-P1)/n1 + P2(1-P2)/n2]
Difference of proportions, testing pi1=pi2 (pooled)	Z = (P1-P2) / sqrt{pi-hat(1-pi-hat)(1/n1+1/n2)}, pi-hat=(X1+X2)/(n1+n2)
Error probabilities	alpha = P(reject H0 | H0 true) = P(Type-I); beta = P(accept H0 | H1 true) = P(Type-II)

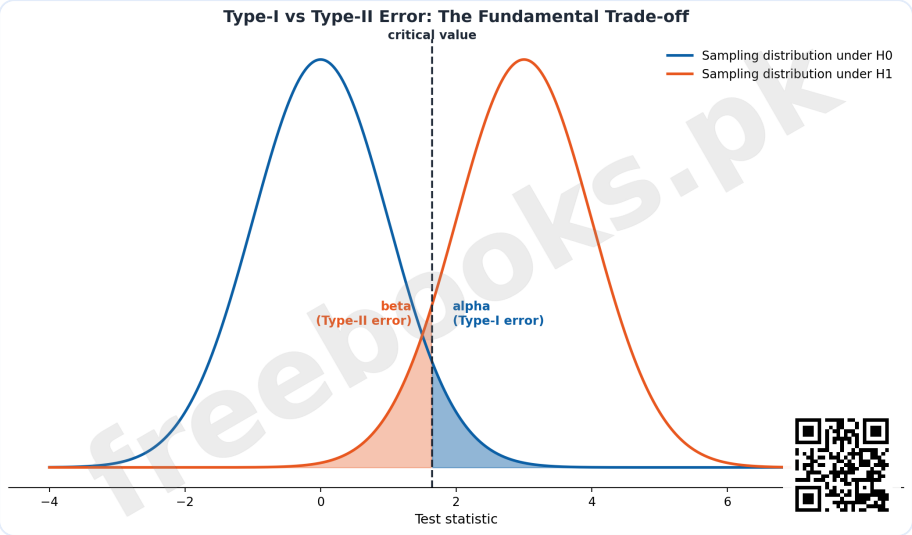
Diagrams

One-Tailed vs Two-Tailed Rejection Regions: Three standard normal curves side by side showing the shaded rejection region for a left-tailed test (Z<-z_alpha), a right-tailed test (Z>z_alpha), and a two-tailed test (split equally in both tails, Z<-z_alpha/2 or Z>z_alpha/2), illustrating how the shape of H1 determines where the critical region is placed



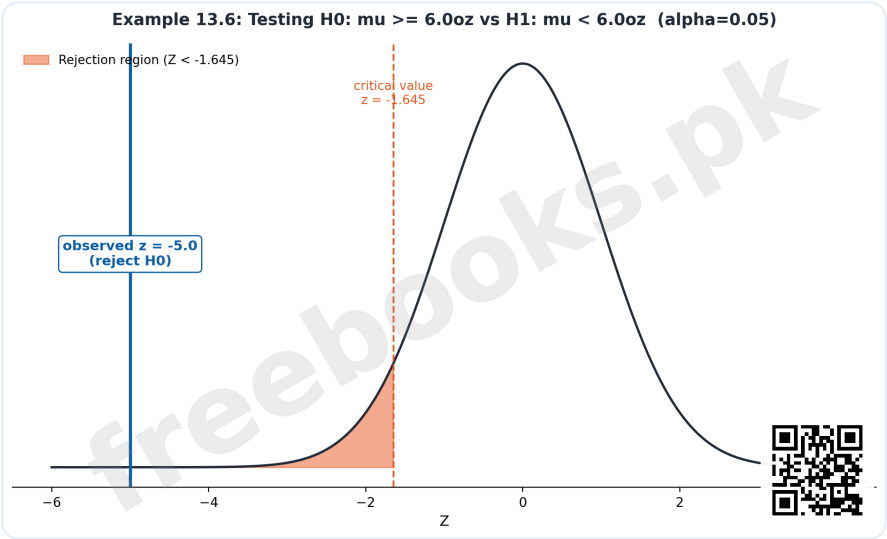
One-Tailed vs Two-Tailed Rejection Regions

Type-I and Type-II Error Trade-off: A two-panel diagram showing the sampling distribution of the test statistic under H0 and under H1 side by side (overlapping), with the alpha region (Type-I error, area under the H0 curve past the critical value) and the beta region (Type-II error, area under the H1 curve on the non-rejection side of the critical value) shaded and labelled, illustrating the fundamental trade-off between the two error types



Type-I and Type-II Error Trade-off

Worked Z-Test Example: Coffee Jar Weight: A standard normal curve for Example 13.6 (H0: mu>=6.0oz vs H1: mu<6.0oz, alpha=0.05) with the left-tail rejection region (Z<-1.645) shaded and the observed test statistic z=-5.0 plotted well inside the rejection region, visually confirming the decision to reject H0



Worked Z-Test Example: Coffee Jar Weight

Short Questions & Answers

Distinguish between the null hypothesis and the alternative hypothesis.

The null hypothesis (H0) is a statement of 'no effect' or the status quo, tested for possible rejection; the alternative hypothesis (H1) is the research claim accepted only if H0 is rejected, and expresses what the researcher is trying to demonstrate.

What is the difference between a Type-I error and a Type-II error?

A Type-I error rejects a true H0 (probability alpha); a Type-II error fails to reject a false H0, i.e., accepts H0 when H1 is actually true (probability beta).

Why is the level of significance, alpha, fixed BEFORE the sample is collected?

So that the chosen risk level cannot be influenced or biased by the actual sample results -- fixing alpha in advance keeps the test objective.

When should a one-tailed test be used instead of a two-tailed test?

A one-tailed test is used when the alternative hypothesis specifies a direction (a definite increase or decrease, i.e. H1 uses < or >); a two-tailed test is used when H1 simply states 'not equal to', with no specified direction.

Why does testing paired (dependent) samples reduce to a one-sample t-test on the differences?

Because each pair's difference D=Y-X collapses two measurements into a single number per subject; testing mu_D=mu2-mu1 then uses exactly the same one-sample machinery as testing an ordinary population mean, just applied to the sample of differences.

Why is a pooled estimate of variance (or proportion) needed for small-sample two-group tests?

Because two separate sample estimates (S1^2 and S2^2, or P1 and P2) of the same underlying population quantity will differ due to sampling error alone; pooling combines information from both samples into a single, more reliable estimate assumed common to both populations.

Long Questions & Answers

Explain the complete logical structure of a hypothesis test -- the null and alternative hypotheses, Type-I and Type-II errors, the level of significance, and the seven-step testing procedure -- using a concrete example to illustrate each concept.

Hypothesis testing provides a formal, disciplined framework for using sample evidence to decide between two competing claims about an unknown population parameter, and understanding this framework properly requires working carefully

through each of its interlocking pieces in turn, since they all depend on one another. Every hypothesis test begins with two competing statements about the population parameter under investigation. The null hypothesis, always denoted H0, is a statement of 'no effect' or 'the status quo' -- it is the specific claim that is put on trial and tested for possible rejection, under the working assumption that it is true unless the sample evidence strongly contradicts it. Typical null hypotheses include claims like 'this coin is unbiased,' 'this new drug is no more effective than the existing one,' or 'there is no difference between two production methods.' The alternative hypothesis, always denoted H1, is the competing claim that the researcher actually suspects or hopes to demonstrate -- it is often called the research hypothesis precisely because it usually expresses what the investigator set out to prove in the first place. Crucially, these two hypotheses must be formulated as an exhaustive, mutually exclusive pair covering every possible true value of the parameter, and the standard convention is to build the equality sign into H0 (using >=, <=, or =) while H1 takes the strict, direction-expressing complement (<, >, or !=). Because any decision reached from a hypothesis test is based on incomplete information, drawn from only a sample rather than the entire population, two distinct kinds of mistake are always possible, and distinguishing them carefully is essential. A Type-I error occurs when the null hypothesis is rejected even though it was actually true -- in effect, a false alarm, concluding an effect exists when it doesn't. A Type-II error occurs when the null hypothesis is NOT rejected even though the alternative hypothesis was actually true -- a missed detection, failing to find an effect that genuinely exists. These two kinds of error are asymmetric in an important sense: a hypothesis test as constructed in this chapter always fixes a small, pre-chosen risk of Type-I error, denoted alpha and called the level of significance, chosen deliberately BEFORE any sample data is even collected, specifically so this choice cannot be influenced or biased by whatever the data eventually turn out to show. The complementary probability, 1-alpha, is called the level of confidence, representing the probability of correctly failing to reject a true null hypothesis. Once H0, H1, and alpha have been settled, an appropriate test statistic must be selected -- a sample-based quantity, such as Z or T, whose exact sampling distribution under the assumption that H0 is true is known and can be looked up. Depending on the specific shape of H1, the rejection region -- the range of test-statistic values that would lead to rejecting H0 -- is placed either entirely in one

tail of that sampling distribution (a one-tailed test, used when H1 specifies a direction such as 'less than' or 'greater than') or split equally between both tails (a two-tailed test, used when H1 simply asserts 'not equal to', with no particular direction specified in advance). To make this entire framework fully concrete, consider testing whether a coffee-packing machine, claimed to fill jars with a mean of 6.0 ounces, might actually be underfilling them. Since the concern is specifically about UNDERfilling, the research (alternative) hypothesis is H1: mu<6.0, making this a one-sided, left-tailed test, with the corresponding null hypothesis H0: mu>=6.0. Suppose the significance level is set at alpha=0.05 before any sample is drawn, and a random sample of 100 jars is then weighed. Since the population standard deviation is assumed known here, the appropriate test statistic is Z=(X-bar-mu0)/(sigma/sqrt(n)), which follows a standard normal distribution if H0 is true. Looking up the critical value for a left-tailed test at alpha=0.05 gives z_alpha=-1.645, so the rejection region is Z<-1.645: any observed z-value more negative than -1.645 leads to rejecting H0. If the observed sample mean turns out to be 5.9 ounces, the observed test statistic computes to z=(5.9-6.0)/(0.2/sqrt(100))=-5.0, and because -5.0 is far more negative than the critical value of -1.645, it falls squarely inside the rejection region, so H0 is rejected: the evidence supports concluding the machine really is underfilling the jars, at the 5% significance level. If a Type-I error were made here, it would mean concluding the machine underfills when it actually doesn't; if a Type-II error had instead been made (by failing to reject H0 when the machine truly was underfilling), the company would continue believing everything is fine when customers are, in fact, being shortchanged.

Compare and contrast the test of the difference between two means for independent samples versus dependent (paired) samples, explaining when each design is appropriate, how each test statistic is constructed, and why pairing can be advantageous.

Comparing two populations' means is one of the most common tasks in applied statistics -- deciding whether a new fertilizer outperforms an old one, whether two machines fill containers to the same average weight, or whether a training programme actually improves worker productivity -- and the correct testing procedure depends critically on HOW the two samples being compared were actually collected, specifically on whether they are independent of one another

or deliberately paired (dependent). Two samples are independent if the selection of observations from one population has no bearing whatsoever on the selection of observations from the other population -- for instance, drawing 25 items from one machine's output and, entirely separately, drawing 36 items from a different machine's output. Two samples are instead dependent, or paired/matched, if each observation in one sample is deliberately linked to a specific, corresponding observation in the other sample, forming genuine pairs -- the classic example being the same group of individuals measured twice, once before and once after some treatment or intervention, such as recording each patient's weight before and after a diet programme. For independent samples, testing H0 concerning mu1-mu2 proceeds through one of three parallel cases depending on what is known or assumed about the two population variances and how large the two samples are. When both populations are normal and their variances sigma1-squared and sigma2-squared are already known, the test statistic Z=[(X-bar-1-X-bar-2)-delta0]/sqrt(sigma1-squared/n1+sigma2-squared/n2) is exactly standard normal under H0, for any sample size whatsoever. When both samples are instead small (conventionally n1 and n2 both no larger than 30), the populations are normal, and their two variances, though unknown, are assumed EQUAL to some common value sigma-squared, the two separate sample variance estimates S1-hat-squared and S2-hat-squared must first be combined into a single pooled estimate, Sp-squared=[(n1-1)S1-hat-squared+(n2-1)S2-hat-squared]/(n1+n2-2), essentially a weighted average of the two sample variances weighted by their respective degrees of freedom; the resulting test statistic T=[(X-bar-1-X-bar-2)-delta0]/[Sp.sqrt(1/n1+1/n2)] then follows a t-distribution with v=n1+n2-2 degrees of freedom. When both samples are instead large (greater than 30), the Central Limit Theorem removes the need for either a normality assumption or an equal-variance assumption altogether, and the test statistic simply becomes Z=[(X-bar-1-X-bar-2)-delta0]/sqrt(S1-hat-squared/n1+S2-hat-squared/n2), using each sample's own separate variance estimate directly, with no pooling required. Dependent (paired) samples call for an entirely different, and in an important sense much SIMPLER, approach. Rather than working with the two original sets of measurements directly, the key manoeuvre is to collapse each matched pair down into a single number: the difference D=Y-X for that pair (for instance, weight-after minus weight-before, for each individual person in the study). Once this collapsing step has been carried out, the entire

two-sample comparison problem has been transformed into an ordinary ONE-sample problem, concerning only the population mean of these differences, denoted mu_D=mu2-mu1, and every formula and technique from the single-sample mean test developed earlier in this very chapter can then be applied directly and without modification, just substituted onto the sample of differences rather than onto the original raw measurements: specifically, the test statistic T=(D-bar-delta0)/(S_D-hat/sqrt(n)), calculated from the n paired differences' own sample mean D-bar and own sample standard deviation S_D-hat, follows a t-distribution with v=n-1 degrees of freedom, exactly mirroring the single-sample small-sample t-test structure. The genuine practical advantage of deliberately using a paired design, whenever it is actually feasible to do so, is that pairing allows each individual subject to effectively serve as their own personal control or baseline, which very efficiently strips out subject-to-subject variability that would otherwise remain baked into, and would inflate, the noise present in a comparison based on two entirely separate, unrelated independent samples. Consider, for example, comparing the effectiveness of two different weight-loss diets: if diet A were tested on one entirely separate group of people while diet B were tested on a completely different second group of people (an independent-samples design), any observed difference between the two groups' outcomes would be muddied and confounded together with all sorts of pre-existing individual differences between the particular people who happened to be randomly assigned into each of the two separate groups -- differences in starting weight, metabolism, activity level, genetics, and so on. If, instead, the very same group of individuals could somehow be tested on both diets sequentially (or, more practically, closely matched pairs of highly similar individuals could each be split one-to-a-diet), then each person's own individual baseline characteristics would automatically cancel out entirely when the paired difference D is calculated for that person, leaving only the genuine treatment effect of interest still remaining in the resulting differences -- and this systematic removal of extraneous subject-to-subject variability from the comparison is exactly why paired designs generally achieve substantially greater statistical power to correctly detect a real, genuinely existing difference, particularly whenever the available sample sizes are comparatively small, compared to what an equivalently-sized independent-samples design testing the very same underlying research question would typically be able to achieve.

Multiple Choice Questions (MCQs)

The null hypothesis H0 is best described as: (A) The claim the researcher hopes to prove (B) A statement of 'no effect' or the status quo, tested for possible rejection (C) Always a one-tailed statement (D) Only used for testing proportions

Correct answer: (B) A statement of 'no effect' or the status quo, tested for possible rejection. H0 is always a 'no effect'/status-quo statement, assumed true and tested for possible rejection.

A Type-I error occurs when: (A) H0 is accepted when it is false (B) H0 is rejected when it is actually true (C) H1 is rejected when it is true (D) The sample size is too small

Correct answer: (B) H0 is rejected when it is actually true. Type-I error = rejecting a true H0; its probability is alpha, the level of significance.

A Type-II error occurs when: (A) H0 is rejected when it is true (B) H0 is not rejected when H1 is actually true (C) The test statistic is miscalculated (D) Alpha is set too low

Correct answer: (B) H0 is not rejected when H1 is actually true. Type-II error = failing to reject H0 (accepting it) when H1 is actually true; its probability is beta.

If H1 is stated as theta < theta0, the appropriate test is: (A) Two-tailed (B) Right-tailed (C) Left-tailed (D) No test is possible

Correct answer: (C) Left-tailed. H1 with '<' indicates a definite decrease is being sought, requiring a one-sided LEFT-tailed test.

For a normal population with unknown variance and a small sample (n<=30), the correct test statistic for the mean is: (A) Z, using sigma (B) T, using S-hat, with v=n-1 degrees of freedom (C) Chi-square (D) F

Correct answer: (B) T, using S-hat, with v=n-1 degrees of freedom. Small sample, normal population, unknown variance: use the T statistic with n-1 degrees of freedom.

The continuity correction is applied when testing a hypothesis about: (A) The difference of two means (B) A population proportion, approximating the discrete binomial with the normal distribution (C) Paired samples (D) The pooled variance

Correct answer: (B) A population proportion, approximating the discrete binomial with the normal distribution. The continuity correction adjusts for approximating the discrete binomial (X or P) using the continuous normal distribution.

For comparing two independent means with small samples and equal but unknown variances, the correct test statistic is: (A) Z with pooled variance (B) T with pooled variance Sp^2, v=n1+n2-2 (C) Z with separate variances (D) T with v=n1+n2-1

Correct answer: (B) T with pooled variance Sp^2, v=n1+n2-2. This case requires the pooled variance estimate Sp^2 and a t-distribution with v=n1+n2-2 degrees of freedom.

Testing the difference between two means using paired (dependent) samples reduces to: (A) A two-sample Z-test (B) A one-sample t-test on the differences D=Y-X (C) A chi-square test (D) A test of two proportions

Correct answer: (B) A one-sample t-test on the differences D=Y-X. Pairing collapses the two-sample problem into an ordinary one-sample t-test performed on the differences.

To test H0: pi1=pi2 (equal proportions in two populations), the standard error is estimated using: (A) Separate P1 and P2 only (B) A pooled proportion pi-hat=(X1+X2)/(n1+n2) (C) The population variance (D) The t-distribution

Correct answer: (B) A pooled proportion pi-hat=(X1+X2)/(n1+n2). Testing pi1=pi2 specifically uses a pooled estimate of the single common proportion, pi-hat, combining both samples.

The level of significance, alpha, must be: (A) Calculated after the sample is collected (B) Fixed before the sample is drawn (C) Always equal to 0.5 (D) The same as the level of confidence

Correct answer: (B) Fixed before the sample is drawn. Alpha is chosen and fixed BEFORE sampling, so the decision rule cannot be influenced by the observed data.

Quick Revision Summary

- H0 = status quo/no-effect (tested for rejection) | H1 = research claim (accepted only if H0 rejected)

- Type-I error = reject true H0, probability alpha | Type-II error = accept false H0, probability beta
- Left-tailed: H1 uses '<' | Right-tailed: H1 uses '>' | Two-tailed: H1 uses '!='
- 7 steps: state hypotheses -> set alpha -> pick test statistic -> find critical value(s) -> rejection region -> decision rule -> compute & conclude
- Mean test: sigma known -> Z; small n normal sigma unknown -> T (v=n-1); large n -> Z (CLT)
- Proportion test: Z=(P-pi0)/sqrt[pi0(1-pi0)/n]; continuity correction adds/subtracts 0.5
- Diff of means, independent: known var -> Z; small n equal var -> pooled T (v=n1+n2-2); large n -> Z
- Diff of means, paired/dependent: reduces to 1-sample T-test on differences D=Y-X, v=n-1
- Diff of proportions: general delta0 -> separate P1,P2 in SE; testing pi1=pi2 -> pooled pi-hat

Exam Tips

- Always write H0 with the equality sign (>=, <=, or =) and H1 as the strict, direction-matching complement
- Match the tail of the test to H1's inequality: '<' = left tail, '>' = right tail, '!=' = two tails (split alpha/2 each side)
- For mean tests, check THREE things before picking Z or T: is sigma known, is the population normal, is n large or small
- For paired-sample problems, compute the differences FIRST (consistently: always after-minus-before, or always A-minus-B) before doing anything else
- When testing pi1=pi2 specifically, don't forget to pool: use pi-hat in the standard error, not separate p1 and p2
- State your final conclusion in plain managerial language, not just 'reject H0' -- explain what that means for the actual question being asked

Chapter 14: Simple Linear Regression and Correlation

Chapters 12 and 13 dealt with a single variable at a time -- estimating or testing claims about one population mean or proportion. This chapter opens a new area of statistics devoted to relationships BETWEEN two variables. Two very different but closely related questions are asked about any pair of variables: is there a mathematical rule that lets us predict one variable from the other (regression), and how strong or close is the linear association between them (correlation)? Regression analysis fits the 'best' straight line through a scatter of paired observations and uses it to estimate or predict the value of one variable (the dependent, or regressand) from a given value of the other (the independent, or regressor). Correlation analysis, by contrast, treats both variables symmetrically and measures the degree to which they move together, without singling out either one as dependent.

The chapter builds the whole machinery from first principles: the distinction between a functional and a statistical relation, the population regression model and its sample counterpart, the method of least squares used to fit the best line, the sample correlation coefficient r and its properties, and finally the tight algebraic relationship that ties regression coefficients and the correlation coefficient together. A recurring and important theme is that regression and correlation are mathematically linked but conceptually distinct -- and that neither one, however strong, proves that one variable causes changes in the other.

Learning Objectives

- Distinguish a functional relation from a statistical relation between two variables
- Define regressor, regressand, regression function, and regression curve
- State and apply the least squares principle to fit a straight line to paired data
- Derive and use the least squares estimates a and b of the population regression line

- Use coding and scaling to simplify regression computations for equally spaced x-values
- State the limitations of linear regression
- Define positive, negative, and no correlation between two variables
- Compute and interpret the sample correlation coefficient r and list its key properties
- Relate the correlation coefficient r to the two regression coefficients b_yx and b_xy
- Explain why a high correlation does not, by itself, establish causation

Key Concepts

14.1 Relations Between Variables

A functional relation between two variables x and y is a perfect relation: the value of the dependent variable y is uniquely and exactly determined by the value of the independent variable x through a mathematical formula, y=f(x). Observations plotted on a graph all fall exactly on the line or curve of the relationship -- this exactness is the defining feature of a functional relation. A statistical relation, by contrast, is not exact: the value of y is not uniquely determined merely from knowledge of x, and plotted observations scatter around, rather than fall exactly on, a line or curve. Most real relationships in business, economics, and the social sciences -- income and expenditure, fertilizer and crop yield, advertising cost and sales, a father's height and his son's height -- are statistical, not functional, relations.

A separate question from the mathematical form of a relationship is whether it is causal: does a genuine cause-and-effect mechanism link the two variables, or do they merely tend to move together? Regression analysis and correlation analysis, on their own, make no assertions whatsoever about causality -- establishing cause requires evidence and reasoning that goes well beyond fitting a line or computing a correlation coefficient.

14.2 Regression Analysis

Regression analysis provides a method for estimating an average (often linear) relationship between two or more variables, allowing the investigator to explain

and predict the value of one variable from known values of the others. In the simplest case, only two variables are involved: the regressor (also called the predictor, independent, controlled, or explanatory variable, usually denoted x) is the variable that forms the basis for estimation, its values fixed in advance by the experimenter and free of random error; the regressand (also called the response, predictand, dependent, or explained variable, usually denoted Y) is the variable whose value is being estimated or predicted, and which is subject to random variation even for a fixed value of x.

If mu_Y|x = E(Y|x) denotes the expected value of the distribution of Y for a given, fixed value of x, then the simple regression is defined as mu_Y|x = f(x), where f(x) may be linear, quadratic, exponential, or any other functional form describing how the regressor relates to the response. Because the points only are given and the function f(x) must be 'worked backwards' or regressed from them, this function is called the regression function. The regression curve is the locus of these expected values mu_Y|x traced out as x varies -- the curve joining the expected value of Y's distribution at each possible value of x.

14.3 Curve Fitting

Curve fitting is the process of estimating, from an observed sample, the parameters of the population regression function relating a response variable to a regressor variable. The least squares principle, due to the French mathematician Adrien Legendre, states that among all possible curves that could be fitted to a set of observed data, the sum of squares of the residuals (the observed values minus their corresponding fitted/estimated values) should be made as small as possible; the curve achieving this minimum is called the least squares fit. Using this principle avoids the personal bias that would otherwise creep in if a curve were fitted to data purely 'by eye'.

Before selecting the form of curve to fit, a scatter diagram is drawn: a set of points in a rectangular coordinate system, with x plotted horizontally and y plotted vertically, where each point represents one observed pair (x_i, y_i). A visual inspection of the scatter diagram reveals the apparent nature of the relationship -- whether the points tend to run from lower-left to upper-right (a direct relationship, Y tends to increase as x increases), from upper-left to lower-right (an inverse relationship, Y tends to decrease as x increases), whether the

pattern looks straight (linear) or bent (curvilinear), and how tightly or loosely the points cluster around the apparent pattern.

14.4 Simple Linear Regression

When the simple regression describes the expected value of the dependent variable Y as a LINEAR function of the independent variable x, the regression is called simple linear regression, written mu_Y|x = alpha + beta.x. Here alpha is the intercept of the line along the y-axis (the value of mu_Y|x when x=0), and beta -- the simple linear regression coefficient -- is the change in the mean of Y's probability distribution per unit increase in x, i.e. the slope of the line, which remains constant at every value of x. Geometrically, beta is measured by tan(theta), where theta is the angle the line makes with the positive x-axis: if beta is positive the line slopes upward, and if beta is negative the line slopes downward.

14.5 The Simple Linear Regression Model

For a fixed value x_i of the regressor, the associated response is a random variable Y_i with mean mu_Y|xi = alpha + beta.x_i and constant variance sigma^2 (assumed the same for every value of x -- the variance does not depend on x, even though the mean does). The deviation of Y_i from its mean is the random error, epsilon_i = Y_i - (alpha + beta.x_i), and the resulting population regression model is Y_i = alpha + beta.x_i + epsilon_i, where alpha and beta are the (unknown) population intercept and slope. This model is 'simple' because there is only one regressor, 'linear in parameters' because no parameter appears as an exponent or is multiplied by another parameter, and 'linear in the independent variable' because x appears only to the first power.

Since alpha and beta are unknown population parameters, they must be estimated from sample data (x1,y1), (x2,y2), ..., (xn,yn). Let a be the best estimate of alpha and b be the best estimate of beta; then the sample simple linear regression line is y-hat = a + b.x, and for any specific x_i, y-hat_i = a + b.x_i is the best point estimate of mu_Y|xi. The residual (or deviation, or prediction error) for observation i is e_i = y_i - y-hat_i = y_i - a - b.x_i, giving the sample regression model y_i = a + b.x_i + e_i.

The covariance of two variables X and Y, denoted s_xy, measures their linear mutual variability: s_xy = [sum(x_i - xbar)(y_i - ybar)]/n = [sum(x_i.y_i)]/n - xbar.ybar. Its sign shows the direction of mutual variability -- positive when the variables tend to move together, negative when they move in opposite directions. If z_i = x_i +/- y_i, then s_z^2 = s_x^2 + s_y^2 +/- 2.s_xy in general, reducing to s_z^2 = s_x^2 + s_y^2 when X and Y are independent (s_xy=0).

Theorem 14.1 (Least Squares Point Estimation): given n observed pairs (x_i, y_i), the least squares line y-hat = a + b.x has slope b = [n.sum(x_i.y_i) - (sum x_i)(sum y_i)] / [n.sum(x_i^2) - (sum x_i)^2] = [sum(x_i-xbar)(y_i-ybar)] / [sum(x_i-xbar)^2] = s_xy/s_x^2, and intercept a = [sum y_i - b.sum x_i]/n = ybar - b.xbar. The least squares line always passes through the point of means (xbar, ybar), and the least squares estimate of mu_Y|xi is y-hat_i = ybar + b(x_i - xbar).

Coding and Scaling (Theorem 14.2): the sample regression coefficient b is independent of a change of origin but NOT independent of a change of scale; if u_i=(x_i-p)/h and v_i=(y_i-q)/k, then b_yx = (k/h).b_vu. When the x-values are equally spaced at interval h, a special coding takes the origin at the middle value (odd n) or the average of the two middle values (even n), and rescales by h (or h/2), making sum(u_i)=0 and greatly simplifying the arithmetic; the resulting line of Y on u, y-hat=a+bu with a=ybar and b=sum(u_i.y_i)/sum(u_i^2), is then transformed back to the original x-scale.

Properties of the least squares line: the sum of residuals is zero (sum e_i = 0); the sum of the observed y-values equals the sum of the fitted values (sum y_i = sum y-hat_i), so ybar equals the mean of the fitted values too; the sum of squared residuals, sum(e_i^2) = sum(y_i^2) - a.sum(y_i) - b.sum(x_i.y_i), is the smallest possible value achievable by ANY straight line through the data; and the line always passes through the point of means (xbar, ybar) -- the 'centre of gravity' of the data.

Limitations of Linear Regression: (i) it applies only to relationships that can genuinely be described by a straight line -- a scatter diagram should be examined first to check this; (ii) the least-squares procedure will always produce a 'best-fit' line, even when no real linear relationship exists at all, so a formal test of significance is needed to confirm the slope b is 'real' and not just sampling noise; (iii) the regression equation is asymmetrical -- the equation predicting Y

from x cannot simply be algebraically rearranged to predict x from Y, a separate regression of x on Y must be fitted; (iv) the regression equation is reliable only over the range of x actually observed in the sample, and should not be extrapolated far beyond that range.

14.6 Simple Linear Correlation

Simple linear correlation measures the strength or closeness of a linear relationship between two variables, addressing two related questions: are the two variables related at all, and how closely does one variable's regression line fit the scatter of observed points? Positive (direct) correlation occurs when the two variables tend to move in the same direction (both increase or both decrease together), corresponding to regression lines with positive slopes; it is called perfect positive correlation if every point lies exactly on a straight line with positive slope. Negative (inverse) correlation occurs when the variables move in opposite directions, corresponding to negative-sloped regression lines; perfect negative correlation means every point lies exactly on a line with negative slope. No correlation exists when one least squares regression line is horizontal and the other is vertical -- equivalently, if X and Y are independent, Cov(X,Y)=0, which implies the correlation coefficient rho=0.

14.7 Correlation Analysis

In pure correlation problems, both variables X and Y are random, and the relationship is considered simultaneously and symmetrically -- unlike regression, where one variable is singled out as dependent. Typical correlation examples include heights and weights of persons, ages of husbands and wives at marriage, and marks in two different subjects. Given a random sample of n pairs (x_i,y_i) from a bivariate population, the sample correlation coefficient r (or r_xy) is defined as r = s_xy/(s_x.s_y), which expands into several algebraically equivalent computing formulas, most commonly r = [n.sum(x_i.y_i) - (sum x_i)(sum y_i)] / sqrt{[n.sum(x_i^2)-(sum x_i)^2][n.sum(y_i^2)-(sum y_i)^2]}.

Properties of r: (1) r is symmetrical in X and Y, i.e. r_xy = r_yx; (2) r equals the covariance of the two variables measured in standard units, r=Cov(z_x,z_y); (3) the MAGNITUDE |r| is unaffected by a change of origin or scale in either variable, though r changes sign if the variables are multiplied/divided by constants of

opposite sign; (4) r always lies between -1 and +1 inclusive; (5) |r| is the geometric mean of the two regression coefficients, r = +/-sqrt(b_yx . b_xy), with the sign matching the common sign of b_yx and b_xy; (6) r is zero whenever one of the two variables is constant. Theorem 14.3 formalises property (3): the correlation coefficient is independent of the origin and the scale of measurement of the variables (up to a possible sign flip if the scale change reverses direction).

Goodness of Fit and the Regression-Correlation Link: for a bivariate sample there are two distinct least squares lines -- Y on X, y-hat=a_yx+b_yx.x, and X on Y, x-hat=a_xy+b_xy.y -- with b_yx=s_xy/s_x^2=r.s_y/s_x and b_xy=s_xy/s_y^2=r.s_x/s_y. Because s_x and s_y are always positive, s_xy, b_yx, b_xy, and r always share the same sign, and if any one of these four quantities is zero, all the others must be zero too. Theorem 14.4 states the two regression lines always intersect at the point of means (xbar, ybar). Theorem 14.5 states that r is the slope of the regression lines when both variables are expressed in standard (z-score) units. Theorem 14.6 states that the two regression lines coincide (become identical) exactly when every sample point lies exactly on a single straight line, i.e. when |r|=1.

Correlation and Causation: a high correlation coefficient shows only that the observed data are consistent with a linear association -- it does NOT, by itself, establish that one variable causes changes in the other. A high correlation between two variables may arise because X causes Y, because Y causes X, because some third factor Z affects both X and Y simultaneously (a common cause, sometimes called a spurious or nonsense correlation), or purely by chance. Only further, careful investigation beyond the correlation coefficient itself -- controlled experiments, theoretical reasoning, checking for confounding variables -- can establish whether a causal relationship genuinely exists.

Important Definitions

What is a functional relation between two variables?: A perfect, exact relation y=f(x) in which the value of the dependent variable is uniquely determined by the independent variable; all plotted observations fall exactly on the line or curve.

What is a statistical relation between two variables?: A relation in which the value of the dependent variable is not uniquely determined by the independent variable; plotted observations scatter around, rather than fall exactly on, the line or curve of the relationship.

What are the regressor and the regressand?: The regressor (x) is the independent/predictor variable forming the basis of estimation, fixed in advance and free of random error; the regressand (Y) is the dependent/response variable being predicted, subject to random variation for any fixed x.

What is the least squares principle?: The principle, due to Adrien Legendre, that the best-fitting curve to a set of data is the one for which the sum of squares of the residuals (observed minus estimated values) is as small as possible.

What is a scatter diagram?: A plot of points in a rectangular coordinate system, one point per observed pair (x_i, y_i), used to visually assess the nature (linear/curvilinear, direct/inverse) and strength of the relationship between two variables before fitting a curve.

What is the simple linear regression coefficient, beta?: The relative change in the expected value of the dependent random variable Y per unit increase in the independent variable x; it is the (constant) slope of the population regression line mu_Y|x = alpha + beta.x.

What is the covariance of two variables, s_xy?: A measure of the linear mutual variability of two variables, s_xy = [sum(x_i-xbar)(y_i-ybar)]/n; positive when the variables move together, negative when they move oppositely.

What is the sample correlation coefficient, r?: A number between -1 and +1, r = s_xy/(s_x.s_y), measuring the strength and direction of the linear relationship between two random variables X and Y.

What does 'no correlation' mean between two variables?: One least squares regression line is horizontal and the other is vertical; equivalently, if X and Y are independent then Cov(X,Y)=0 and the correlation coefficient rho=0.

Why doesn't a high correlation prove causation?: A high correlation only shows the data are consistent with a linear association; it may instead arise

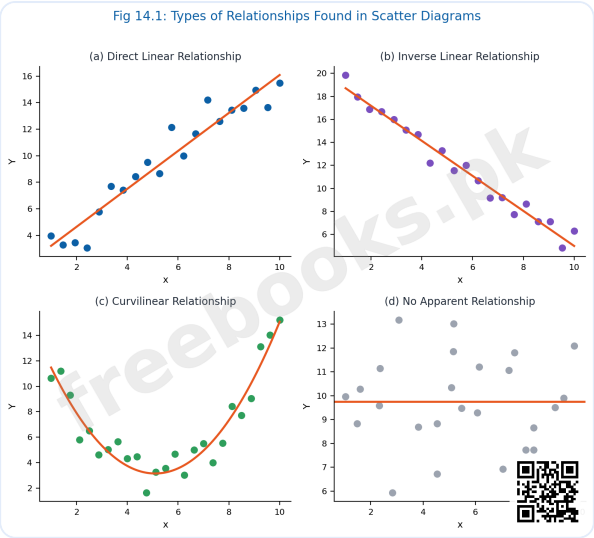
because Y causes X, because a third factor affects both variables, or purely by chance -- further investigation is needed to establish a genuine causal link.

Key Facts and Relations

Topic	Key Fact / Relation
Population simple linear regression	mu_Y|x = alpha + beta.x; Y_i = alpha + beta.x_i + epsilon_i
Sample simple linear regression	y-hat = a + bx; y_i = a + b.x_i + e_i, e_i = y_i - y-hat_i
Least squares slope, b	b = [n.sum(x_i y_i) - (sum x_i)(sum y_i)] / [n.sum(x_i^2) - (sum x_i)^2] = s_xy/s_x^2
Least squares intercept, a	a = [sum y_i - b.sum x_i]/n = ybar - b.xbar
Covariance, s_xy	s_xy = [sum(x_i-xbar)(y_i-ybar)]/n = [sum x_i y_i]/n - xbar.ybar
Sample correlation coefficient, r	r = s_xy/(s_x.s_y) = [n.sum(x_i y_i)-(sum x_i)(sum y_i)] / sqrt{[n.sum x_i^2-(sum x_i)^2][n.sum y_i^2-(sum y_i)^2]}
Regression coefficients via r	b_yx = s_xy/s_x^2 = r.s_y/s_x; b_xy = s_xy/s_y^2 = r.s_x/s_y
r as geometric mean of regression coefficients	r = (+/-) sqrt(b_yx . b_xy), sign = common sign of b_yx and b_xy
Regression equation of Y on X (via r)	y-hat = ybar + (r.s_y/s_x)(x - xbar)
Regression equation of X on Y (via r)	x-hat = xbar + (r.s_x/s_y)(y - ybar)
Coding and scaling relation	u_i=(x_i-p)/h, v_i=(y_i-q)/k => b_yx = (k/h).b_vu
Sum of squared residuals identity	sum(e_i^2) = sum(y_i^2) - a.sum(y_i) - b.sum(x_i y_i) [minimum value over all straight lines]

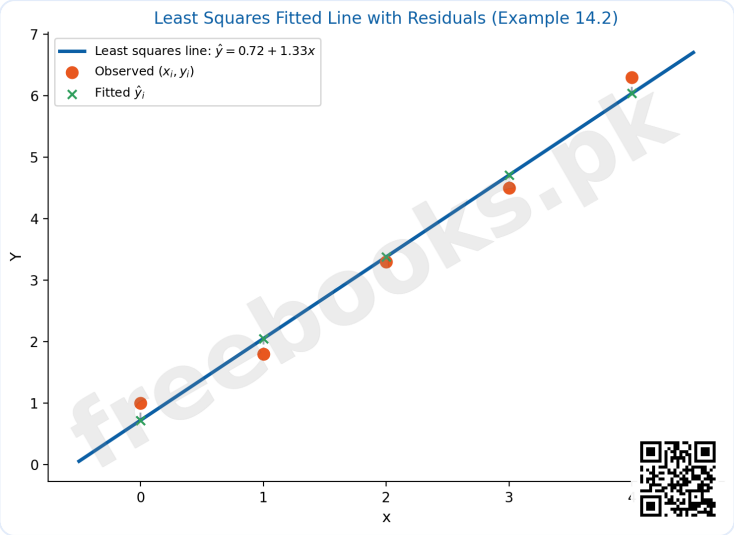
Diagrams

Scatter Diagram Patterns: A 2x2 panel of scatter diagrams illustrating the main relationship patterns covered in the chapter: direct linear relationship (positive slope, points rising left to right), inverse linear relationship (negative slope, points falling left to right), a curvilinear relationship (points following a bending curve rather than a line), and no apparent relationship (points scattered with no discernible pattern), matching Fig 14.1 of the textbook



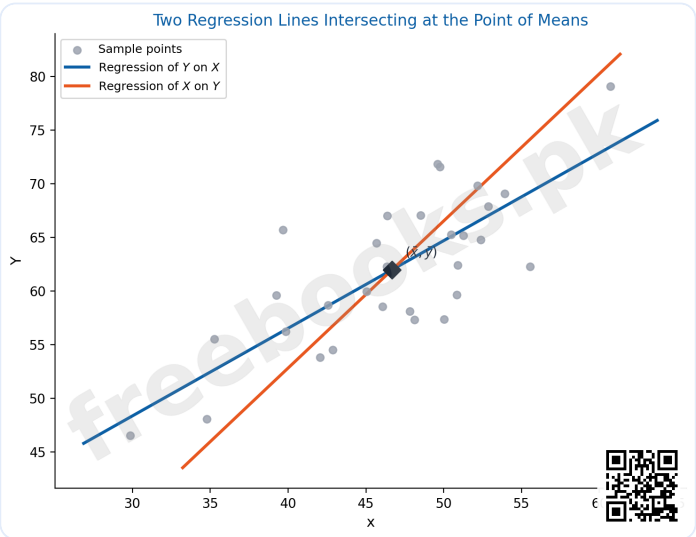
Scatter Diagram Patterns

Least Squares Fitted Line with Residuals: A scatter diagram of Example 14.2's data (x=0,1,2,3,4; y=1.0,1.8,3.3,4.5,6.3) with the least squares regression line y-hat=0.72+1.33x drawn through the points, and a vertical dashed segment from each observed point down (or up) to the fitted line representing its residual e_i, illustrating the least squares principle of minimizing the sum of squared residuals



Least Squares Fitted Line with Residuals

Two Regression Lines Intersecting at the Point of Means: A scatter diagram with both least squares lines drawn -- the regression of Y on X and the regression of X on Y -- shown crossing exactly at the point (xbar, ybar), illustrating Theorem 14.4 and how the angle between the two lines narrows as the correlation |r| approaches 1 (lines nearly coincide) and widens as |r| approaches 0 (lines nearly perpendicular)



Two Regression Lines Intersecting at the Point of Means

Short Questions & Answers

Distinguish between a functional relation and a statistical relation.

A functional relation is exact -- y=f(x) uniquely determines y from x, and all points fall exactly on the curve; a statistical relation is not exact -- y is not

uniquely determined by x, and observed points scatter around, rather than lie exactly on, the relationship.

Define the regressor and the regressand.

The regressor (x) is the independent, predictor variable, fixed in advance and free of random error; the regressand (Y) is the dependent, response variable whose value is predicted from x and which is subject to random variation.

State the least squares principle.

Among all possible curves that could be fitted to a set of observed data, the least squares fit is the one for which the sum of squares of the residuals (observed values minus their fitted/estimated values) is the smallest possible.

What does it mean for the least squares line to 'pass through the point of means'?

The fitted regression line y-hat=a+bx always passes exactly through the point (xbar, ybar) -- when x_i=xbar, the fitted value y-hat_i equals ybar exactly; this is one of the guaranteed properties of the least squares solution.

Why is r said to be the geometric mean of the two regression coefficients?

Algebraically, r = +/-sqrt(b_yx . b_xy), because b_yx=s_xy/s_x^2 and b_xy=s_xy/s_y^2, so their product equals s_xy^2/(s_x^2.s_y^2) = r^2; the sign of r matches the common sign of b_yx and b_xy.

Why can a strong correlation not, by itself, prove causation?

A high correlation only shows the data are consistent with a linear association; the same pattern could arise because X causes Y, because Y causes X, because a third factor affects both, or purely by chance -- further investigation beyond the correlation coefficient is required to establish causation.

Long Questions & Answers

Trace the full development of the simple linear regression model from the population regression line through to the sample least squares line, explaining the role of each symbol and using Example 14.1 to illustrate the computations.

Regression analysis begins by recognising that in most real-world situations the relationship between two variables is statistical rather than functional -- for a given value of the independent, non-random variable x, the dependent variable Y does not take on one single fixed value but instead behaves as a random variable, possessing its own probability distribution with a mean mu_Y|x=E(Y|x) and a variance sigma_Y|x^2=Var(Y|x). It is assumed that this mean, mu_Y|x, changes in some constant, systematic way as x changes, while the variance sigma_Y|x^2 is assumed constant across all values of x, equal to a single common value sigma^2 -- an assumption sometimes called homoscedasticity. When it is reasonable to assume that all these conditional means mu_Y|x lie along a single straight line, the population regression line is written mu_Y|x = alpha + beta.x, where alpha is the population y-intercept (the value of mu_Y|x when x=0) and beta is the population slope, also called the population regression coefficient, representing the change in the mean of Y per unit increase in x. For a specific value x_i drawn from x's domain, the associated random variable Y_i has mean mu_Y|xi=alpha+beta.x_i and constant variance sigma^2; the deviation of Y_i from this mean is called the random error or disturbance, epsilon_i=Y_i-(alpha+beta.x_i), leading to the population regression model Y_i=alpha+beta.x_i+epsilon_i, which is simple (one regressor), linear in its parameters, and linear in x. In practice, alpha, beta, mu_Y|x, and sigma^2 are all unknown population parameters that must be ESTIMATED from a sample of n observed pairs (x_1,y_1), (x_2,y_2), ..., (x_n,y_n). It is customary to let a denote the best (least squares) estimate of alpha, b denote the best estimate of beta, and y-hat denote the resulting estimate of mu_Y|x, giving the sample simple linear regression line y-hat=a+bx, with y-hat_i=a+b.x_i being the best point estimate of mu_Y|xi for any specific x_i. The difference between an actually observed y_i and its fitted value y-hat_i is the residual, e_i=y_i-y-hat_i=y_i-a-b.x_i, an estimate of the unobservable population error epsilon_i, giving the sample regression model y_i=a+b.x_i+e_i. The values a and b that make this

fitted line the 'best' one are found using the method of least squares, formalised in Theorem 14.1: b=[n.sum(x_iy_i)-(sum x_i)(sum y_i)]/[n.sum(x_i^2)-(sum x_i)^2], which can also be written as s_xy/s_x^2, and a=[sum y_i - b.sum x_i]/n=ybar-b.xbar. Example 14.1 of the textbook illustrates the full computation concretely: eight students' matriculation scores x_i (480,490,510,510,530,550,610,640) are paired with their college grade point averages y_i (2.7,2.9,3.3,2.9,3.1,3.0,3.2,3.7). Summing gives sum x_i=4320, sum y_i=24.8, sum(x_i^2)=2355800, and sum(x_iy_i)=13492, with n=8. Substituting into the least squares formula gives b=[8(13492)-(4320)(24.8)]/[8(2355800)-(4320)^2]=0.00435, and then a=[24.8-0.00435(4320)]/8=0.751, so the best fitted line is y-hat=0.751+0.00435x. This estimated line can now be used for prediction: for a student scoring x=600 on matriculation, the estimated mean GPA is y-hat=0.751+0.00435(600)=3.361. Several important properties are guaranteed to hold for any least squares line, and can be checked against this example: the sum of the residuals sum(e_i) is exactly zero (any apparent nonzero result is due only to rounding); the sum of the observed y-values equals the sum of the fitted y-hat values, so the mean of the fitted values equals ybar; the sum of squared residuals sum(e_i^2) achieves the smallest possible value of any straight line through this data (that is the entire point of the least squares principle); and the fitted line always passes exactly through the point of means (xbar, ybar) -- here xbar=4320/8=540 and ybar=24.8/8=3.1, and indeed y-hat=0.751+0.00435(540)=3.10 confirms this.

Define the sample correlation coefficient r, list its key properties, explain how it connects algebraically to the two regression coefficients b_yx and b_xy, and explain why correlation does not establish causation.

Correlation analysis, unlike regression analysis, treats both variables X and Y symmetrically as jointly random -- there is no distinguished 'independent' or 'dependent' variable, only a bivariate population from which pairs (x_i,y_i) are randomly sampled, such as the heights and weights of a group of people or the marks obtained by students in two different subjects. Given a random sample of n such pairs, the sample correlation coefficient, denoted r or r_xy, is defined as r=s_xy/(s_x.s_y), the covariance of X and Y divided by the product of their individual standard deviations; this expands algebraically into several

computationally convenient equivalent forms, the most commonly used being r=[n.sum(x_iy_i)-(sum x_i)(sum y_i)]/sqrt{[n.sum(x_i^2)-(sum x_i)^2][n.sum(y_i^2)-(sum y_i)^2]}. This coefficient r has a set of well-established properties that make it such a useful summary measure. First, r is symmetrical with respect to X and Y, meaning r_xy=r_yx -- unlike the two regression coefficients, which generally differ depending on which variable is treated as dependent. Second, r can be interpreted as the covariance between the two variables after each has been converted into standard (z-score) units, r=Cov(z_x,z_y). Third, and very usefully in practice, the MAGNITUDE of r, |r|, is completely unaffected by a change of origin (adding or subtracting a constant from either variable) or a change of scale (multiplying or dividing by a constant of the SAME sign in both variables) -- this is precisely why, in Example 14.5 of the textbook, computing r directly from the raw height and weight data gives exactly the same value, 0.956, as computing it from deviations taken from arbitrarily assumed means of 70 and 120 respectively; only if the variables are rescaled by constants of OPPOSITE sign does r flip in sign. Fourth, r always lies within the closed interval [-1,+1], with the extreme values +1 and -1 occurring only when every single sample point lies exactly on a straight line with positive or negative slope respectively (perfect correlation), and a value near zero indicating little or no linear association (though a strong NONLINEAR association could still be present even when r is near zero, since r only measures LINEAR association). Fifth, and most directly connecting correlation back to regression, |r| is the geometric mean of the two regression coefficients b_yx (the slope of Y regressed on X) and b_xy (the slope of X regressed on Y): r=+/-sqrt(b_yx . b_xy), with the sign of r matching the common sign shared by b_yx and b_xy (since s_x and s_y are always positive, the sign of s_xy determines the sign of both regression coefficients and of r together, and if any one of s_xy, b_yx, b_xy, or r is zero then all four must be zero). This tight relationship can also be written directly in terms of r: b_yx=r.s_y/s_x and b_xy=r.s_x/s_y, which lets the two regression equations be rewritten using r explicitly as y-hat=ybar+(r.s_y/s_x)(x-xbar) and x-hat=xbar+(r.s_x/s_y)(y-ybar). A further set of theorems ties the geometry together neatly: the two regression lines always intersect at the point of means (xbar,ybar) regardless of the value of r; r itself equals the common slope of both regression lines when the data are expressed in standardised (z-score) units; and the two regression lines become IDENTICAL (coincide exactly) precisely when every

sample point lies on a single straight line, that is, when |r|=1 -- conversely, when r=0, the two lines are at right angles to one another (one horizontal, one vertical), reflecting 'no correlation'. Despite this rich and precise mathematical structure, it is essential to recognise that even a correlation coefficient very close to +1 or -1 does NOT, by itself, prove that one variable causes changes in the other. A high correlation is consistent with (but does not distinguish between) several very different underlying explanations: X may genuinely cause Y; Y may instead cause X; some third, unmeasured factor Z may independently influence both X and Y, generating a 'spurious' correlation between them even though neither causes the other directly (the textbook's own example is a high correlation between city temperature and birth rate, which plainly reflects no direct causal mechanism at all); or the observed correlation may simply be a product of chance, especially in small samples. Establishing genuine causation requires evidence well beyond the correlation coefficient itself -- controlled experiments, a plausible causal mechanism, ruling out confounding third factors, and, ideally, replication of the finding in independent samples.

Multiple Choice Questions (MCQs)

A relation y=f(x) in which the value of y is uniquely and exactly determined by x is called: (A) A statistical relation (B) A functional relation (C) A correlation (D) A regression curve

Correct answer: (B) A functional relation. An exact relation, where every point falls precisely on the curve, is called a functional relation.

The independent, predictor variable in a regression problem is called the: (A) Regressand (B) Response variable (C) Regressor (D) Residual

Correct answer: (C) Regressor. The regressor (x) forms the basis of estimation/ prediction; the regressand (Y) is the variable being predicted.

The least squares principle chooses the fitted curve that minimizes: (A) The sum of the residuals (B) The sum of squares of the residuals (C) The correlation coefficient (D) The sample size

Correct answer: (B) The sum of squares of the residuals. Least squares minimizes sum(e_i^2), the sum of squared residuals, not the (unsquared) sum which is always zero for the optimal line.

In the population regression model Y_i = alpha + beta.x_i + epsilon_i, beta represents: (A) The y-intercept (B) The change in mean of Y per unit increase in x (C) The variance of Y (D) The correlation coefficient

Correct answer: (B) The change in mean of Y per unit increase in x. Beta is the population regression coefficient / slope -- the change in mu_Y|x per unit increase in x.

The least squares regression line always passes through the point: (A) (0,0) (B) (xbar, ybar) (C) (max x, max y) (D) (median x, median y)

Correct answer: (B) (xbar, ybar). One of the guaranteed properties of the least squares line is that it passes exactly through the point of means (xbar, ybar).

If the values of x are multiplied by a positive constant h and y by a positive constant k, the regression coefficient b_yx is: (A) Unaffected (B) Multiplied by k/h (C) Multiplied by h/k (D) Set to zero

Correct answer: (B) Multiplied by k/h. By the coding/scaling theorem, b_yx = (k/h).b_vu -- the regression coefficient DOES change with a change of scale (unlike r's magnitude).

The sample correlation coefficient r always lies in the range: (A) 0 to 1 (B) -1 to 0 (C) -1 to +1 (D) -infinity to +infinity

Correct answer: (C) -1 to +1. By its mathematical construction, r always satisfies -1<=r<=1.

If b_yx = 0.8 and b_xy = 0.45 (both positive), the correlation coefficient r equals: (A) 1.25 (B) 0.6 (C) 0.36 (D) -0.6

Correct answer: (B) 0.6. r = +sqrt(b_yx . b_xy) = sqrt(0.8*0.45) = sqrt(0.36) = 0.6 (positive since both regression coefficients are positive).

The two regression lines (Y on X, and X on Y) coincide exactly when: (A) r = 0 (B) r = 0.5 (C) |r| = 1 (D) The sample size is large

Correct answer: (C) |r| = 1. The two regression lines become identical only when every point lies exactly on one line, i.e. when |r|=1 (perfect correlation).

A high correlation coefficient between two variables: (A) Always proves X causes Y (B) Never has any practical meaning (C) Does not, by itself, establish a causal relationship (D) Means the two regression lines are at right angles

Correct answer: (C) Does not, by itself, establish a causal relationship. Correlation and regression make no assertion about causality; a high r may reflect direct causation, reverse causation, a common third cause, or chance.

Quick Revision Summary

- Functional relation = exact, y=f(x) | Statistical relation = not exact, points scatter around the curve
- Regressor (x) = independent/predictor, fixed in advance | Regressand (Y) = dependent/response, random
- Least squares principle: minimize sum(e_i^2), the sum of SQUARED residuals
- Population model: mu_Y|x = alpha + beta.x, Y_i = alpha + beta.x_i + epsilon_i
- Sample model: y-hat = a + bx; b = s_xy/s_x^2; a = ybar - b.xbar
- Least squares line ALWAYS passes through (xbar, ybar); sum(e_i)=0; sum(y_i)=sum(y-hat_i)
- Covariance s_xy: positive = variables move together, negative = move oppositely
- Correlation r = s_xy/(s_x s_y); |r| unaffected by change of origin/scale; -1<=r<=1
- r = +/-sqrt(b_yx . b_xy); b_yx = r.s_y/s_x; b_xy = r.s_x/s_y
- Two regression lines meet at (xbar, ybar); coincide when |r|=1; right angles when r=0
- Regression/correlation never proves causation -- could be X causes Y, Y causes X, a third factor, or chance

Exam Tips

- Always draw or imagine the scatter diagram FIRST -- it tells you whether a straight-line fit even makes sense before you compute anything
- Remember b (slope) uses s_xy/s_x^2, NOT s_xy/(s_x.s_y) -- that second formula is for r, a very common mix-up
- When x-values are equally spaced, use the coding trick (origin at the middle, unit = the common interval) to avoid large, error-prone numbers
- Check your answer: does the fitted line pass through (xbar, ybar)? If not, recheck your arithmetic

- |r| is unaffected by adding/subtracting or multiplying/dividing by a POSITIVE constant -- use this to simplify awkward raw data before computing r
- Never claim causation from r alone in an exam answer -- always note that regression/correlation analysis makes no assertion about cause and effect

Chapter 15: Association

Chapters 12-14 dealt with quantitative (numerically measurable) variables -- means, proportions, and the strength of linear relationships between two measured variables. Many real observations, however, are only qualitative: people classified as satisfied/neutral/dissatisfied with their jobs, manufactured items graded excellent/good/poor/scrap, households owning no car/one car/two-or-more cars. This chapter develops the parallel machinery for such categorical data, called attributes. Just as correlation measures the strength of a linear relationship between two quantitative variables, association measures the strength of relationship between two qualitative attributes.

The chapter builds from the simplest case -- two attributes, each either present or absent, forming a 2x2 table -- through Yule's coefficient of association Q, up to the general r x c contingency table and the chi-square test of statistical independence, the workhorse test for categorical data used throughout applied statistics. It closes with Spearman's rank correlation coefficient, which measures the degree of agreement between two rankings of the same set of objects -- a natural bridge back to the correlation ideas of Chapter 14, but applied to ranks rather than raw measurements.

Learning Objectives

- Distinguish an attribute (qualitative variable) from a quantitative variable
- Define class, class frequency, dichotomy, positive/negative attributes, and ultimate class frequency
- Test whether the data on two attributes are consistent
- State the condition for independence of two attributes and compute expected frequencies under independence
- Define positive association, negative association, and complete association/disassociation
- Compute and interpret Yule's coefficient of association, Q
- Construct and interpret an r x c contingency table

- State and apply the chi-square test of statistical independence, including Yates' correction for 2x2 tables
- Compute Pearson's coefficient of mean square contingency, C, and its maximum possible value
- Compute and interpret Spearman's coefficient of rank correlation

Key Concepts

15.1 Multinomial Populations

When each element of a population is assigned to one, and only one, of more than two attribute categories, the population is called a multinomial population. This generalises the binomial idea (success/failure, two categories) to any number of mutually exclusive categories -- for example, classifying manufactured items as excellent, good, poor, or scrap.

15.2 Attribute (Qualitative Variable)

An attribute is a characteristic that varies only in QUALITY, not quantity, from one individual to another -- marital status, education level, blindness, smoking, beauty. An attribute cannot be measured numerically; data on it is obtained simply by noting whether each object possesses the attribute, and counting how many do and do not. A class is a set of objects sharing a given characteristic, and a class frequency is the number of observations falling into a class. Dichotomy is the process of dividing objects into two mutually exclusive, complementary classes according to whether or not they possess a particular attribute; classification can continue indefinitely by further dichotomising on additional attributes.

Notation: capital letters A, B, ... denote presence of attributes; Greek letters alpha, beta, ... denote their absence ('not A', 'not B'). Class frequencies are written in brackets: (A) = number possessing A; (alphaB) = number possessing B but not A. Classes represented purely by positive attributes (A, B, AB) are positive classes; purely negative attributes (alpha, beta, alphabeta) are negative classes; mixed classes (Abeta, alphaB) are contrary classes. The order of a class equals the number of attributes specifying it (order 1: (A), (alpha); order 2: (AB), (Abeta), (alphaB), (alphabeta)); the class of order zero is the total n. For k

attributes, the total number of class frequencies (including n) is 3^k, and the ultimate class frequencies -- those of the HIGHEST order -- number 2^k.

Every class frequency can be expressed as the sum of its two subclasses from further dichotomy, e.g. n=(A)+(alpha), (A)=(AB)+(Abeta). Data are consistent if no class frequency computed this way is negative; the necessary and sufficient condition for consistency is that no ULTIMATE class frequency is negative. Inconsistency may arise from wrong counting, arithmetic errors, or misprints -- but consistency itself is no guarantee that the counting was accurate.

15.3 Independence of Attributes

Two attributes A and B are independent if the proportion possessing both equals the product of the proportions possessing each separately: (AB)/n = [(A)/n] x [(B)/n], equivalently (AB) = (A)(B)/n. This is the categorical-data analogue of statistical independence of events, P(A and B)=P(A)P(B). Under independence, every cell of the 2x2 table can be filled in purely from the row and column totals, using this same multiplication rule for each of the four cells.

15.4 Association of Attributes (Correlation of Qualitative Variables)

Two attributes are associated if they are NOT independent, i.e. (AB) != (A)(B)/n. Association is positive (the attributes are 'simply associated') if (AB) > (A)(B)/n -- the attributes occur together more often than chance/independence would predict. Association is negative (the attributes are 'disassociated') if (AB) < (A)(B)/n -- they occur together less often than chance would predict. Disassociation does NOT imply independence; it is a distinct, opposite-direction relationship.

Complete (perfect positive) association exists if neither attribute can occur without the other except possibly in one direction: if (A)=(B), all A's are B's and vice versa; if (A)<(B), all A's are B's; if (B)<(A), all B's are A's. Complete disassociation (perfect negative association) exists if none of the A's are B's and none of the alpha's are beta's -- the two attributes are mutually exclusive.

Yule's coefficient of association, Q, measures the strength of association: Q = [(AB)(alphabeta) - (Abeta)(alphaB)] / [(AB)(alphabeta) + (Abeta)(alphaB)]. Q always lies between -1 and +1: Q=0 means independence, Q=+1 means

complete association, Q=-1 means complete disassociation. Q is the categorical-data counterpart of the correlation coefficient r from Chapter 14.

15.5 Two Dimensional Count Data: Contingency Table

When attribute A has r categories and attribute B has c categories, a simple random sample produces an r x c contingency table (a term coined by Karl Pearson) of observed cell frequencies o_ij, i=1,...,r and j=1,...,c. The cell frequency is the count falling in a particular (row, column) combination. The i-th row total is o_i. = sum over j of o_ij; the j-th column total is o_.j = sum over i of o_ij; and the grand total is n = sum of all o_ij = sum of all row totals = sum of all column totals.

15.6 Test for Statistical Independence

The chi-square test of independence tests H0: the row attribute and column attribute are independent (P(Ai and Bj)=P(Ai)P(Bj) for all i,j) against H1: they are not independent for at least one cell. Under H0, the expected frequency for cell (i,j) is e_ij = (o_i. x o_.j)/n -- the row total times the column total, divided by the grand total, exactly mirroring the independence condition of section 15.3. The test statistic is chi-square = sum over all cells of [(o_ij - e_ij)^2 / e_ij], which follows an approximate chi-square distribution with v=(r-1)(c-1) degrees of freedom for large n. Large discrepancies between observed and expected frequencies inflate chi-square; if observed and expected agree exactly, chi-square=0. The critical region for level of significance alpha is chi-square > chi-square_{v; 1-alpha} (upper tail only -- large chi-square, not small, indicates departure from independence).

A commonly used rule of thumb requires each expected frequency e_ij to be at least 5 for the chi-square approximation to be adequate; if some cells have smaller expected frequencies, neighbouring rows or columns can be combined (pooled), which reduces the degrees of freedom by one for each such combination performed.

Yates' Correction for Continuity applies specifically to 2x2 tables (v=1 degree of freedom), improving the approximation of the discrete data to the continuous chi-square distribution: Adjusted chi-square = sum of [(|o-e|-0.5)^2 / e]. The

correction reduces the value of chi-square (making rejection of H0 less likely), and should only be omitted when |o-e| is already less than 0.5.

Pearson's Coefficient of Mean Square Contingency, C = sqrt[chi-square / (n + chi-square)], measures the STRENGTH of an association once independence has been rejected, on a scale that (unlike raw chi-square) does not depend purely on sample size. Its maximum possible value depends on the table's dimensions: max C = sqrt[(q-1)/q], where q is the smaller of the number of rows or columns -- so C should always be interpreted relative to this ceiling, not against a fixed universal scale like the correlation coefficient r.

15.7 Rank Correlation

Rank correlation measures the correlation between the RANKS (not the raw measurements) assigned to individuals on two variables -- for instance, two judges each ranking the same n contestants. Spearman's coefficient of rank correlation, r_r, is the ordinary Pearson correlation coefficient computed on the two sets of ranks, and simplifies to r_r = 1 - [6.sum(d_i^2)] / [n(n^2-1)], where d_i = x_i - y_i is the difference between an individual's two ranks. Like r, r_r always lies between -1 and +1: r_r near +1 indicates strong agreement between the two rankings, r_r near -1 indicates strong disagreement (reversed order), and r_r near 0 indicates little or no relationship between the rankings. Spearman's method is especially useful when the underlying data are not available as precise numerical measurements but can still be meaningfully ranked, or when the assumptions needed for Pearson's r (e.g. approximate linearity, interval-scale data) are in doubt.

Important Definitions

What is an attribute?: A characteristic that varies only in quality, not quantity, from one individual to another (e.g. marital status, smoking, beauty); it cannot be measured numerically, only noted as present or absent.

What is dichotomy?: The process of dividing objects into two mutually exclusive, complementary classes according to whether or not they possess a particular attribute.

What is an ultimate class frequency?: The frequency of a class of the HIGHEST order for the attributes under study; for k attributes there are 2^k ultimate class frequencies.

When are data on class frequencies said to be consistent?: When no class frequency (computed from the given data) is negative; the necessary and sufficient condition is that no ULTIMATE class frequency is negative.

When are two attributes A and B independent?: When (AB) = (A)(B)/n, i.e. the proportion possessing both attributes equals the product of the proportions possessing each attribute separately.

What is positive association and negative association?: Positive association (simple association): (AB) > (A)(B)/n, the attributes occur together more than chance predicts. Negative association (disassociation): (AB) < (A)(B)/n, they occur together less than chance predicts.

What is Yule's coefficient of association, Q?: A measure of the strength of association between two attributes, Q=[(AB)(alphabeta)-(Abeta)(alphaB)]/[(AB)(alphabeta)+(Abeta)(alphaB)], lying between -1 (complete disassociation) and +1 (complete association), with Q=0 meaning independence.

What is a contingency table?: An r x c two-way frequency table (a term coined by Karl Pearson) of observed cell counts, cross-classifying a sample by r categories of one attribute and c categories of another.

What is Pearson's coefficient of mean square contingency, C?: A measure of the strength of association derived from the chi-square statistic, C=sqrt[chi-square/(n+chi-square)], whose maximum possible value sqrt[(q-1)/q] depends on the table's dimensions (q = smaller of rows/columns).

What is Spearman's coefficient of rank correlation, r_r?: The ordinary correlation coefficient computed between two sets of RANKS (rather than raw measurements) assigned to the same n individuals, r_r=1-[6.sum(d_i^2)]/[n(n^2-1)], lying between -1 and +1.

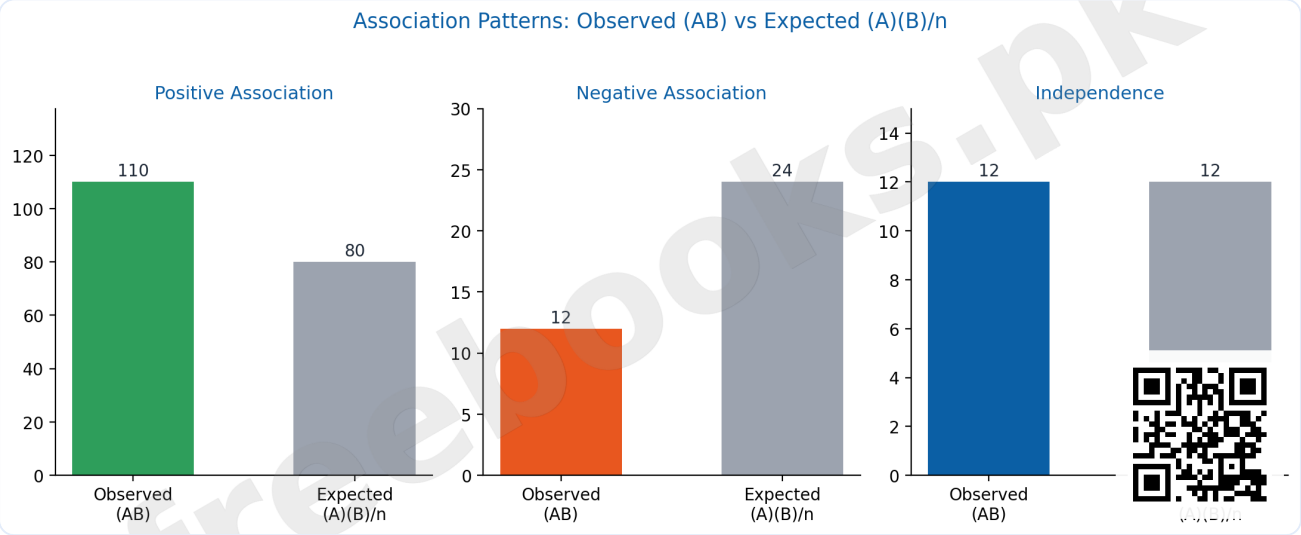
Key Facts and Relations

Topic	Key Fact / Relation
Total / ultimate class frequencies (k attributes)	Total class frequencies = 3^k; Ultimate class frequencies = 2^k
Independence condition	(AB) = (A)(B)/n
Positive / negative association	Positive: (AB) > (A)(B)/n; Negative: (AB) < (A)(B)/n
Yule's coefficient of association, Q	Q = [(AB)(alphabeta) - (Abeta)(alphaB)] / [(AB)(alphabeta) + (Abeta)(alphaB)]
Expected frequency under independence (contingency table)	e_ij = (o_i. x o_.j) / n
Chi-square test statistic	chi-sq = sum_i sum_j [(o_ij - e_ij)^2 / e_ij], v = (r-1)(c-1)
Yates' continuity-corrected chi-square (2x2 tables)	Adjusted chi-sq = sum [(|o-e| - 0.5)^2 / e]
Pearson's coefficient of mean square contingency	C = sqrt[chi-sq / (n + chi-sq)]
Maximum possible value of C	max C = sqrt[(q-1)/q], q = smaller of (rows, columns)
Mean square contingency	phi^2 = chi-sq / n
Spearman's rank correlation coefficient	r_r = 1 - [6.sum(d_i^2)] / [n(n^2-1)], d_i = x_i - y_i (difference of ranks)
Critical region for independence test	Reject H0 if chi-sq > chi-sq_{v; 1-alpha} (upper tail only)

Diagrams

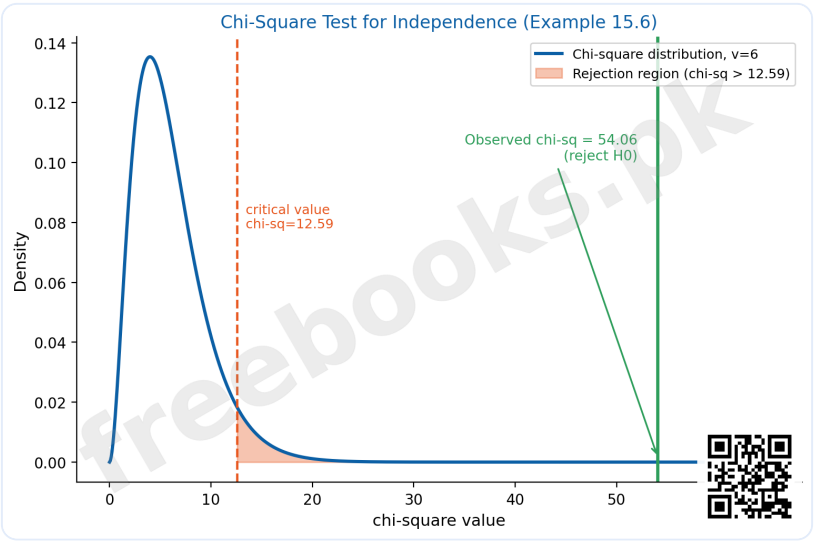
Association Patterns: Positive, Negative, Independent: A 3-panel grouped bar chart comparing the observed frequency (AB) against the expected-under-independence frequency (A)(B)/n for three worked cases: positive association

((AB) noticeably taller than expected), negative association ((AB) noticeably shorter than expected), and independence ((AB) equal to expected), illustrating how the sign of (AB)-(A)(B)/n determines the direction of association



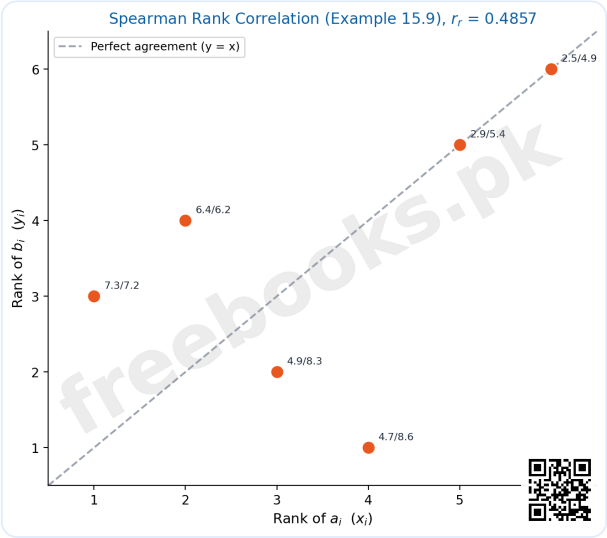
Association Patterns: Positive, Negative, Independent

Chi-Square Test for Independence (Example 15.6): A chi-square distribution curve with v=6 degrees of freedom, with the upper-tail rejection region beyond the critical value chi-sq=12.59 (at alpha=0.05) shaded, and the observed test statistic chi-sq=54.06 from the degree/hobby contingency table plotted well inside the rejection region, visually confirming the decision to reject the null hypothesis of independence



Chi-Square Test for Independence (Example 15.6)

Spearman Rank Correlation (Example 15.9): A scatter plot of the six paired ranks (x_i, y_i) from Example 15.9 with a reference line y=x (perfect agreement) overlaid, showing how closely the two judges'/measurements' rankings track each other, illustrating the moderate positive rank correlation r_r=0.4857 obtained for this data



Spearman Rank Correlation (Example 15.9)

Short Questions & Answers

Distinguish between an attribute and a (quantitative) variable.

An attribute is a qualitative characteristic that cannot be measured numerically and is only noted as present or absent (e.g. marital status); a variable is a quantitative characteristic that can be measured numerically (e.g. height, income).

What is the necessary and sufficient condition for the consistency of class-frequency data?

No ULTIMATE class frequency (the frequency of the highest-order class) should be negative; if any ultimate class frequency computed from the data is negative, the data are inconsistent.

Distinguish between association and disassociation of two attributes.

Association (positive) occurs when (AB) > (A)(B)/n -- the attributes occur together more than chance predicts; disassociation (negative association) occurs when (AB) < (A)(B)/n -- they occur together less than chance predicts. Disassociation does not imply independence.

What does the degrees of freedom v=(r-1)(c-1) represent in a chi-square test of independence for an r x c contingency table?

It is the number of cell frequencies that can be freely chosen once the row and column totals are fixed; the remaining cells are then determined by the requirement that sub-totals and totals agree with the observed data.

Why is Yates' correction applied to 2x2 contingency tables?

To improve the approximation of the discrete chi-square statistic (based on whole-number counts) to the continuous chi-square distribution, by reducing the absolute value of each deviation |o-e| by 0.5 before squaring.

Why can raw chi-square not be used directly to compare the strength of association across contingency tables of different sizes?

Because the magnitude of chi-square depends on the number of degrees of freedom (the table's dimensions), not purely on the strength of association; Pearson's coefficient C standardises this onto a common scale, though even C's maximum value still depends on the table's dimensions.

Long Questions & Answers

Explain the full logic of the chi-square test for statistical independence in a contingency table -- the null and alternative hypotheses, the expected frequencies, the test statistic, and the role of degrees of freedom -- illustrating each step using Example 15.6 (degree vs hobby).

The chi-square test of independence addresses a very natural question about categorical (qualitative) data: when a sample of individuals is cross-classified by two different attributes at once, does knowing an individual's category on one attribute tell you anything about their likely category on the other, or are the two classifications statistically unrelated? This question is formalised through a contingency table -- a two-way table of observed frequencies o_ij, with r rows representing the categories of one attribute and c columns representing the categories of the other, a name and format due to Karl Pearson. The null hypothesis H0 states that the two attributes are statistically independent, meaning that for every cell (i,j), the probability of falling into that cell equals the product of the row category's marginal probability and the column category's

marginal probability, exactly mirroring the multiplication rule for independent events introduced much earlier in probability theory. The alternative hypothesis H1 simply states that this independence fails for at least one cell, i.e. some form of association exists between the two classifications. To test H0, we cannot compare the observed frequencies directly to some fixed benchmark; instead we must first compute what the cell frequencies WOULD look like if H0 were exactly true, given the row and column totals we actually observed. This expected frequency for cell (i,j), denoted e_ij, is computed as e_ij = (o_i. times o_.j) divided by n -- the i-th row total times the j-th column total, divided by the overall sample size -- which is precisely the independence condition (AB)=(A)(B)/n from earlier in this chapter, generalised from a simple 2x2 table to a full r x c table. The test statistic then measures the overall discrepancy between what was actually observed and what independence would predict: chi-square = the sum, over every cell in the table, of (observed minus expected) squared, divided by that cell's expected frequency. Squaring the deviations ensures that positive and negative discrepancies don't cancel out, and dividing by e_ij weights each squared discrepancy relative to the size of frequency expected in that cell, so that a given absolute discrepancy counts for more in a cell where only a few observations were expected than in a cell where hundreds were expected. If H0 is exactly true and every observed frequency matched its expected frequency perfectly, chi-square would equal exactly zero; the further the observed frequencies drift from the expected ones, the larger chi-square becomes, so only unusually LARGE values of chi-square provide evidence against H0 -- the rejection region is entirely in the upper (right-hand) tail of the chi-square distribution, at chi-square greater than the critical value chi-square_{v;1-alpha}. The appropriate degrees of freedom, v=(r-1)(c-1), reflect the fact that once the row totals and column totals of the table are fixed (which happens automatically, since these are simply the observed marginal totals), only (r-1) times (c-1) of the interior cells can be freely chosen -- the remaining cells are then completely determined by the requirement that every row and column must sum correctly to its observed total. Example 15.6 makes all of this concrete: 492 candidates are cross-classified by university degree (mathematics, chemistry, physics -- 3 columns) and hobby (music, craftwork, reading, drama -- 4 rows), giving a 4x3 table with v=(4-1)(3-1)=6 degrees of freedom. At alpha=0.05, the critical value is chi-square_{6;0.95}=12.59, so the decision rule is: reject H0 if the computed

chi-square exceeds 12.59. Working out each cell's expected frequency (row total times column total, divided by 492) and comparing against the corresponding observed frequency, cell by cell, produces a total computed chi-square of 54.06 -- far beyond the critical value of 12.59 -- so H0 is decisively rejected: degree and hobby are NOT independent; some genuine association exists between a candidate's field of study and their choice of hobby. Because H0 was rejected, it becomes meaningful to ask how STRONG this association is, using Pearson's coefficient of mean square contingency, C=sqrt[chi-square/(n+chi-square)]=sqrt[54.06/(492+54.06)]=0.315, interpreted against its maximum possible value for a table this shape, sqrt[(q-1)/q]=sqrt[2/3]=0.8165 (since q, the smaller table dimension, is 3 -- the number of degree categories).

Compare Yule's coefficient of association (Q) for a simple 2x2 table with Spearman's coefficient of rank correlation (r_r), explaining what each measures, how each is computed, and when each would be the appropriate tool to use.

Although Yule's coefficient of association Q and Spearman's coefficient of rank correlation r_r both produce a single number between -1 and +1 that summarises the strength and direction of a relationship between two things, they are built for genuinely different kinds of data and answer genuinely different questions, and understanding the distinction is essential to choosing the right tool. Yule's Q applies specifically to the very simplest possible categorical-data situation: TWO attributes, each of which is simply either present or absent for every individual in the sample -- forming exactly a 2x2 table with four cell frequencies, conventionally labelled (AB), (Abeta), (alphaB), and (alphabeta), representing individuals who possess both attributes, only the first, only the second, and neither, respectively. Q is defined algebraically as [(AB)(alphabeta) minus (Abeta)(alphaB)], all divided by [(AB)(alphabeta) plus (Abeta)(alphaB)] -- essentially comparing the product of the two 'agreement' cells (both present, both absent) against the product of the two 'disagreement' cells (present/absent combinations), on a normalised scale. When Q=0, the two attributes are exactly independent, meaning (AB) equals exactly (A)(B)/n, the frequency you would expect purely from the two attributes' separate marginal rates occurring together by chance. When Q is positive, the two attributes are positively associated (occurring together MORE often than chance alone would predict), reaching the

extreme Q=+1 only in the case of complete association, where possessing one attribute effectively guarantees possessing the other (in at least one direction). When Q is negative, the attributes are negatively associated or disassociated (occurring together LESS often than chance would predict), reaching Q=-1 only under complete disassociation, where the two attributes are essentially mutually exclusive. Spearman's r_r, by contrast, is built for an entirely different setting: two variables that CAN in principle be ranked, or perhaps only observed AS ranks in the first place (such as two different judges each ranking the very same set of n contestants from best to worst, or a beauty contest where absolute numerical scores were never even recorded, only relative orderings). Rather than working with cell counts, r_r is computed by first converting each variable's n observations into an ordinal ranking from 1 to n, then computing, for each individual, the simple difference d_i between their rank on the first variable and their rank on the second variable, and finally combining all these rank-differences into the single formula r_r = 1 minus [6 times the sum of the squared differences d_i-squared], all divided by [n times (n-squared minus 1)]. When the two rankings agree perfectly (every individual ranks in exactly the same relative position on both variables), every d_i equals zero and r_r reaches its maximum value of exactly +1; when the two rankings are perfectly and completely REVERSED (the individual ranked first on one variable is ranked last on the other, and so on symmetrically), r_r reaches its minimum possible value of exactly -1; and when there is no discernible relationship whatsoever between the two rankings, r_r settles near zero. The essential practical distinction, then, comes down to the nature of the underlying data being related: Q is the appropriate tool whenever both variables of interest are genuinely categorical, dichotomous (yes/no, present/absent) attributes rather than measurements that could sensibly be ranked or ordered along a continuous scale -- for example, whether a patient was vaccinated (yes/no) against whether they contracted a disease (yes/no). Spearman's r_r, by contrast, is the appropriate tool whenever the underlying data consists of, or can meaningfully be reduced to, an ORDERING of individuals along some scale -- for example, comparing how closely two different examiners' rankings of the same set of essays agree with one another, even if the examiners never assigned precise numerical scores at all, only relative orderings from best to worst.

Multiple Choice Questions (MCQs)

A characteristic that varies only in quality, not quantity, is called: (A) A variable (B) An attribute (C) A contingency table (D) A rank

Correct answer: (B) An attribute. An attribute is a qualitative characteristic (e.g. marital status); it cannot be measured numerically, only noted as present or absent.

For k attributes, the number of ULTIMATE class frequencies is: (A) 3^k (B) 2^k (C) k^2 (D) k!

Correct answer: (B) 2^k. The ultimate class frequencies are those of the highest order; for k attributes there are 2^k of them.

Two attributes A and B are independent if: (A) (AB) > (A)(B)/n (B) (AB) < (A)(B)/n (C) (AB) = (A)(B)/n (D) (AB) = 0

Correct answer: (C) (AB) = (A)(B)/n. Independence means the proportion possessing both attributes equals the product of their separate proportions: (AB)=(A)(B)/n.

If (AB) > (A)(B)/n, the two attributes A and B are: (A) Independent (B) Negatively associated (C) Positively associated (D) Inconsistent

Correct answer: (C) Positively associated. When the observed joint frequency exceeds the expected-under-independence value, the attributes are positively associated.

Yule's coefficient of association Q always lies in the range: (A) 0 to 1 (B) -1 to 0 (C) -1 to +1 (D) -infinity to +infinity

Correct answer: (C) -1 to +1. Like the correlation coefficient r, Yule's Q always lies between -1 (complete disassociation) and +1 (complete association).

In an r x c contingency table, the expected frequency for cell (i,j) under independence is: (A) o_ij / n (B) (o_i.)(o_.j)/n (C) o_i. + o_.j (D) chi-square / n

Correct answer: (B) (o_i.)(o_.j)/n. e_ij = (row total)(column total)/n -- the same multiplication logic as the independence condition (AB)=(A)(B)/n.

The degrees of freedom for a chi-square test of independence in an r x c table is: (A) r + c (B) rc (C) (r-1)(c-1) (D) r x c - 1

Correct answer: (C) (r-1)(c-1). v=(r-1)(c-1), since only that many cells can be freely chosen once the row/column totals are fixed.

Yates' continuity correction is applied specifically to: (A) Any contingency table (B) 2x2 tables (v=1) (C) Tables with more than 10 categories (D) Spearman's rank correlation

Correct answer: (B) 2x2 tables (v=1). Yates' correction improves the discrete-to-continuous approximation specifically for 2x2 tables, where v=1 degree of freedom.

The maximum possible value of Pearson's coefficient of contingency C depends on: (A) The sample size n only (B) The value of alpha (C) The dimensions of the table (q = smaller of rows/columns) (D) Nothing -- it is always 1

Correct answer: (C) The dimensions of the table (q = smaller of rows/columns). max C = sqrt[(q-1)/q], where q is the smaller of the number of rows or columns -- so C's ceiling varies by table shape.

Spearman's rank correlation coefficient r_r = 1 - [6.sum(d_i^2)]/[n(n^2-1)] equals +1 when: (A) The two rankings are completely reversed (B) The two rankings agree perfectly (every d_i = 0) (C) n is very large (D) Chi-square equals zero

Correct answer: (B) The two rankings agree perfectly (every d_i = 0). When every individual has the identical rank on both variables, every d_i=0, and the formula gives r_r=1 (perfect agreement).

Quick Revision Summary

- Attribute = qualitative characteristic (present/absent) | Variable = quantitative, numerically measurable
- Dichotomy = splitting objects into two mutually exclusive classes on one attribute
- Order of class = number of attributes specifying it; ultimate class frequencies = highest order, count = 2^k
- Consistency: data are consistent iff no ULTIMATE class frequency is negative
- Independence: (AB) = (A)(B)/n | Positive association: (AB) > (A)(B)/n | Negative: (AB) < (A)(B)/n

- Yule's Q = [(AB)(ab)-(Ab)(aB)]/[(AB)(ab)+(Ab)(aB)]; Q=0 independent, Q=+1 complete association, Q=-1 complete disassociation
- Contingency table: r rows x c columns of observed o_ij; e_ij = (row total)(col total)/n
- Chi-square = sum[(o-e)^2/e], v=(r-1)(c-1); reject H0 (independence) if chi-square > critical value (upper tail only)
- Yates' correction (2x2 only): subtract 0.5 from |o-e| before squaring, to improve the continuous approximation
- Pearson's C = sqrt[chi-square/(n+chi-square)]; max C = sqrt[(q-1)/q], q = smaller of rows/columns
- Spearman's r_r = 1 - 6.sum(d_i^2)/[n(n^2-1)], d_i = rank difference; range -1 to +1, like ordinary r

Exam Tips

- Never confuse 'disassociation' with 'independence' -- disassociation is a specific NEGATIVE relationship, not the absence of one
- Always check e_ij >= 5 in every cell before trusting the chi-square approximation; pool rows/columns if needed and reduce df accordingly
- Apply Yates' correction ONLY to 2x2 tables (v=1); it is not used for larger tables
- The chi-square rejection region is always in the UPPER tail only -- there is no 'too good a fit' rejection region in this test
- When reporting C, always state it alongside its maximum possible value sqrt[(q-1)/q] for that table shape -- a raw C value alone is not comparable across differently-shaped tables
- For rank correlation, first double-check that ranks (not raw scores) have been assigned consistently in the same direction (both high-to-low or both low-to-high) before computing d_i

Chapter 16: Analysis of Time Series

Almost every business, economic, or scientific measurement recorded repeatedly over time -- annual wheat production, monthly sales, hourly temperature -- forms a time series. Unlike the cross-sectional data considered in earlier chapters, a time series has an inherent ORDER: the position of an observation in time is itself meaningful information, not just another label. This chapter develops the standard framework for describing and decomposing such data: the four classical components (secular trend, seasonal variations, cyclical fluctuations, and irregular movements), the two standard models (multiplicative and additive) that combine them, and four practical methods for isolating and measuring the most important of these components, the secular trend -- free hand curve, semi-averages, moving averages, and least squares.

The mathematics of secular trend estimation by least squares directly reuses the machinery built in Chapter 14 for fitting a straight line, applied here with time itself as the independent variable, plus an extension to fitting a quadratic (parabolic) curve when a straight line does not adequately describe the pattern. A recurring practical theme is the CODING of the time variable -- shifting and rescaling the origin so that sum(x)=0, which dramatically simplifies the arithmetic of every method in the chapter.

Learning Objectives

- Define a time series and distinguish it from cross-sectional data
- Construct a historigram and identify the signal and noise components of a series
- Identify and describe the four classical components of a time series: secular trend, seasonal variations, cyclical fluctuations, and irregular movements
- Distinguish the multiplicative model from the additive model for combining time series components
- Code the time variable to simplify trend calculations
- Estimate secular trend using the method of free hand curve
- Estimate secular trend using the method of semi-averages

- Compute k-period (and centred) moving averages, and use them to smooth a time series
- Fit a linear and a quadratic secular trend using the method of least squares
- Shift the origin of an already-fitted trend equation to a new reference time period

Key Concepts

16.1 Time Series

A time series is the sequence y1, y2, ..., yn of n observations of a variable Y, recorded according to their time of occurrence t1, t2, ..., tn. Symbolically Y is expressed as a function of time, y=f(t)+e, where f(t) is the systematic, determined component of variation (the signal) and e is the random, irregular component (the noise). Observations in a time series are usually recorded at equidistant points in time -- hourly temperature, annual wheat yield, monthly sales, daily attendance.

A historigram is a graph of a time series: time t is plotted on the x-axis, the observed values of Y on the y-axis, and successive points are joined by line segments. Constructing a historigram is always the first step in examining a time series, before any formal analysis is attempted.

16.2 Components of a Time Series

A time series may be composed of up to four distinct components, each the result of a well-defined but different underlying cause; a given series need not contain all four. The secular (long-term) trend, T, is a smooth, steady, gradual movement of the series in the same direction (up or down) over a fairly long period, reflecting forces such as population growth, technological change, or shifting long-run demand -- for example the decline in death rate due to advances in science, or steadily rising demand for wheat as population grows.

Seasonal variations, S, are short-term movements that recur regularly, with a fixed periodicity of a year or less (daily, weekly, monthly, yearly) -- caused by the calendar seasons, religious festivals, or social customs, such as increased electricity use in summer or an after-Eid sale spike. Cyclical fluctuations, C, are

longer-term oscillations about the trend that recur over a period longer than a year but WITHOUT a fixed periodicity (unlike seasonal variations) -- caused by the business cycle, which has four phases: trough (depression, the lowest point), expansion (recovery, the upswing), peak (boom/prosperity, the highest point), and recession (contraction, the downswing). Cyclical fluctuations are less predictable than seasonal variations or trend, making them more consequential for business planning.

Irregular movements, I, are unpredictable, unsystematic, non-recurring changes caused by random events such as wars, floods, earthquakes, strikes, or fires -- also called erratic, accidental, or random variations.

16.3 Analysis of Time Series

Analysing a time series means decomposing it into its separate components for individual study, mainly to support estimation and forecasting. Two standard models combine the four components: the multiplicative model, Y=TxSxCxI, treats trend T in the series' original units while S, C, and I are unitless percentage index numbers; the additive model, Y=T+S+C+I, treats all four components in the series' original units. The multiplicative model is conventionally treated as the standard model for time series analysis.

Coded Time Variable: to simplify trend calculations, the time variable t is coded as x=(t-tbar)/h (or a rescaled version), where tbar=(first period + last period)/2 and h is the constant interval between periods. Since sum(t-tbar)=0 by construction, the coding guarantees sum(x)=0=sum(x^3)=sum(x^5)=..., which is what dramatically simplifies the least squares normal equations used later in the chapter. For an odd number of periods, x=0 is assigned to the exact middle period; for an even number of periods, x=0 falls halfway between the two middle periods (giving half-integer codes unless the unit of measurement is redefined as half a period, which restores whole-number codes).

16.4 Estimation of Secular Trend

Four methods are used to estimate secular trend, in increasing order of objectivity and mathematical rigour. The Method of Free Hand Curve plots the historigram and then draws, by eye, a smooth line or curve that best represents the general movement -- often anchored through the point (mean time, mean y)

for extra reliability. It is simple and quick, and smooths out seasonal variations, but is a rough, personally-biased method: different analysts will draw different lines from the same data.

The Method of Semi-Averages divides the observed series into two equal halves (omitting or duplicating the middle value if n is odd), computes the average of each half, and draws a straight line through the two resulting points, each plotted at its half's midpoint in time. Given the two semi-averages y1' (at x1) and y2' (at x2), the trend line's slope is b=(y2'-y1')/(x2-x1) and intercept a=y1'-b.x1; equivalently, for the whole series, b=4(S2-S1)/n^2 where S1, S2 are the sums of the first and second halves and n is the total number of periods. This method is objective and quick, and gives a better approximation than free hand curve (since it rests on a definite mathematical rule), but is heavily affected by extreme values (because it uses arithmetic means) and is only appropriate when the underlying trend is genuinely linear.

The Method of Moving Averages replaces each observation with the average of itself and a fixed number k of neighbouring periods, sliding this window one period at a time across the whole series: a1=(1/k)sum of the first k values, a2=(1/k)sum of values 2 through k+1, and so on; equivalently, using successive k-period totals S1, S2, S3,..., each new total is obtained from the previous one by dropping the oldest value and adding the next one, S2=S1+y_{k+1}-y1, and each average a_i=S_i/k. Each moving average is plotted against the MIDDLE of its k-period window. When k is odd, this middle falls exactly on an observed time period; when k is even, it falls between two periods, requiring a further 2-period 'centring' average of the moving averages themselves to realign them with the actual observed periods (yielding CENTRED moving averages). Moving averages smooth out periodic (seasonal/cyclical) fluctuations most effectively when the averaging period is chosen to match the periodicity of those fluctuations, but the method sacrifices trend values at both ends of the series, is still sensitive to extreme observations, and provides no actual mathematical equation for extrapolating the trend into the future.

The Method of Least Squares fits a definite mathematical curve -- typically a straight line y-hat=a+bx or a parabola y-hat=a+bx+cx^2, with x the coded time value -- by minimizing the sum of squared deviations of observed values from

the fitted curve, exactly the same least squares principle developed in Chapter 14 for ordinary regression. For a straight line, the normal equations sum(y)=na+b.sum(x) and sum(xy)=a.sum(x)+b.sum(x^2) give b=[n.sum(xy)-sum(x)sum(y)]/[n.sum(x^2)-(sum x)^2] and a=ybar-b.xbar; when the time variable has been coded so sum(x)=0, these collapse to the much simpler a=ybar=sum(y)/n and b=sum(xy)/sum(x^2). For a parabola, with x coded so sum(x)=0=sum(x^3), the normal equations similarly simplify to give c=[n.sum(x^2 y)-sum(x^2)sum(y)]/[n.sum(x^4)-(sum x^2)^2], a=[sum(y)-c.sum(x^2)]/n, and b=sum(xy)/sum(x^2). The fitted line/curve always passes through the point (xbar, ybar); the residuals sum to zero (sum y = sum y-hat); and the smaller the sum of squared residuals, the better the fit -- which is exactly how a quadratic trend can be objectively confirmed to fit better than a linear one for a given dataset, by comparing their two sum(e^2) values directly.

Shifting the Origin: given an already-fitted trend line y-hat=a+bx (linear) with the origin at some reference time period, moving the origin k periods FORWARD means substituting (x+k) for x, and k periods BACKWARD means substituting (x-k) for x; only the intercept changes (to a+/-bk), never the slope b. For a quadratic trend, the same substitution is made into y-hat=a+bx+cx^2 and the result expanded and regrouped in powers of the new x, changing all three constants a, b, and c.

Important Definitions

What is a time series?: The sequence y1, y2, ..., yn of n observations of a variable Y, recorded according to their time of occurrence t1, t2, ..., tn, usually at equidistant points in time.

What are the signal and the noise in a time series?: The signal, f(t), is the systematic, determined component of variation; the noise, e, is the random, irregular component; symbolically y=f(t)+e.

What is a historigram?: A graph of a time series, with time t on the x-axis and the observed values of Y on the y-axis, successive points joined by line segments -- the first step in examining any time series.

What is secular trend?: A smooth, steady, gradual movement of a time series in the same direction (up or down) over a fairly long period of time, reflecting long-run forces such as population growth or technological change.

What are seasonal variations?: Short-term movements that recur regularly with a fixed periodicity of a year or less, caused by calendar seasons, festivals, or social customs.

What are cyclical fluctuations?: Longer-term oscillations about the trend recurring over more than a year, WITHOUT a fixed periodicity, caused by the business cycle (trough, expansion, peak, recession).

What are irregular movements?: Unpredictable, non-recurring, unsystematic changes caused by random events such as wars, strikes, floods, or fires; also called erratic or accidental variations.

What is the multiplicative model of a time series?: The model Y=TxSxCxI, in which trend T is in the series' original units while S, C, and I are unitless percentage index numbers; conventionally the standard model for time series analysis.

What is the coded time variable, and why is it used?: A rescaled/shifted time variable x=(t-tbar)/h chosen so that sum(x)=0 (and sum(x^3)=0), which greatly simplifies the least squares normal equations used to estimate secular trend.

Why must k-period moving averages be 'centred' when k is even?: Because an even-period moving average falls exactly halfway between two observed time periods; a further 2-period moving average of these averages realigns ('centres') them with the actual observed periods.

Key Facts and Relations

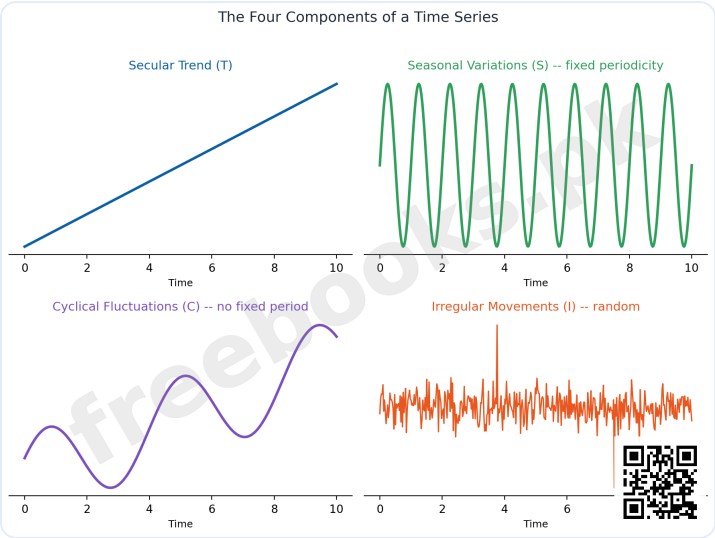
Topic	Key Fact / Relation
Time series model	y = f(t) + e (f(t) = signal, e = noise)
Multiplicative and additive models	Multiplicative: Y = T x S x C x I; Additive: Y = T + S + C + I

Topic	Key Fact / Relation
Coded time variable (general)	x = (t - tbar)/h, tbar = (first period + last period)/2
Semi-averages trend slope and intercept	b = (y2' - y1')/(x2 - x1); a = y1' - b.x1; [b = 4(S2-S1)/n^2 for whole series]
k-period moving average (successive)	a1 = S1/k; a2 = a1 + (y_{k+1} - y1)/k; S2 = S1 + y_{k+1} - y1
Least squares linear trend (general)	b = [n.sum(xy) - sum(x)sum(y)] / [n.sum(x^2) - (sum x)^2]; a = ybar - b.xbar
Least squares linear trend (coded, sum x=0)	a = ybar = sum(y)/n; b = sum(xy)/sum(x^2)
Quadratic (parabolic) trend model	y-hat = a + bx + cx^2
Quadratic trend estimates (coded, sum x=0=sum x^3)	c = [n.sum(x^2 y) - sum(x^2)sum(y)] / [n.sum(x^4) - (sum x^2)^2]; a = [sum(y) - c.sum(x^2)]/n; b = sum(xy)/sum(x^2)
Shifting the origin (linear trend)	y-hat = a + b(x +/- k) = (a +/- bk) + bx
Sum of squared residuals (goodness of fit)	sum(e^2) = sum(y^2) - a.sum(y) - b.sum(xy) [linear]; minus c.sum(x^2 y) also, for quadratic
Properties of the least squares trend line	Passes through (xbar, ybar); sum(y - y-hat) = 0, so sum(y) = sum(y-hat)

Diagrams

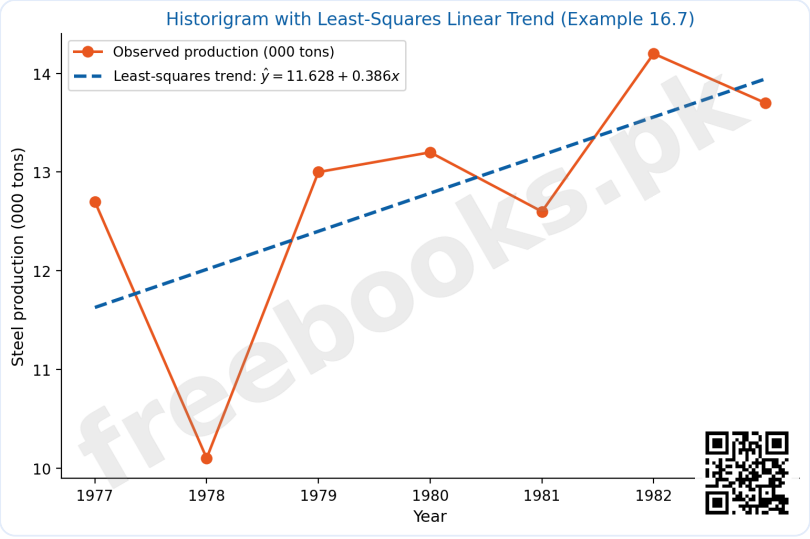
The Four Components of a Time Series: A 4-panel illustration showing idealised patterns for each classical component: secular trend (a smooth rising straight line), seasonal variations (a regular repeating up-down wave pattern within each year), cyclical fluctuations (a longer, irregular-period wave oscillating around a rising trend), and irregular movements (a jagged, unpredictable random

pattern with occasional spikes), illustrating how a real time series is the combination of some or all of these



The Four Components of a Time Series

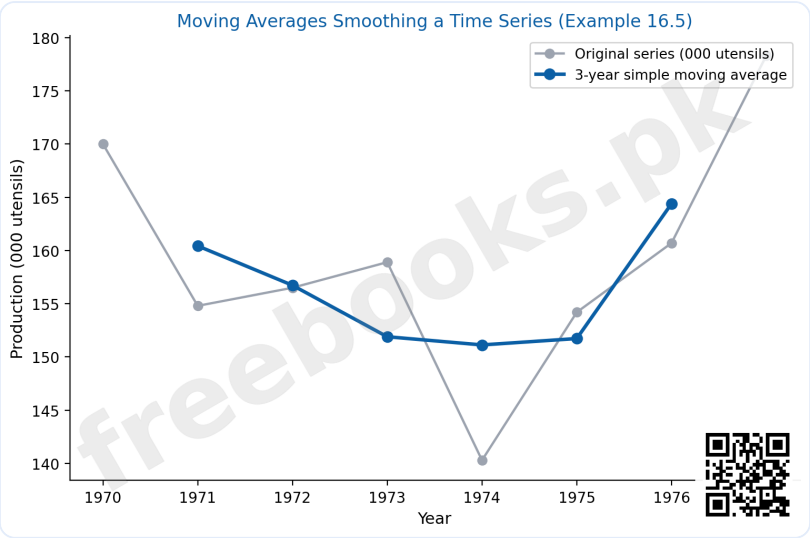
Historigram with Least-Squares Linear Trend (Example 16.7): A historigram (line plot) of steel production 1977-1983, with the least-squares fitted straight trend line y-hat=11.628+0.386x overlaid, showing how the fitted line captures the general upward movement while smoothing over the year-to-year fluctuations



Historigram with Least-Squares Linear Trend (Example 16.7)

Moving Averages Smoothing a Time Series (Example 16.5): The original (jagged) yearly production series for silver utensils 1970-1977 plotted alongside its 3-year simple moving average, showing visually how the moving average

smooths out short-term fluctuations to reveal the underlying movement, while losing values at both ends of the series



Moving Averages Smoothing a Time Series (Example 16.5)

Short Questions & Answers

Define a time series and give two examples.

A time series is the sequence of observations of a variable recorded according to their time of occurrence, usually at equidistant points; examples include the monthly sales of a store and the annual wheat yield of a country.

Distinguish between seasonal variations and cyclical fluctuations.

Seasonal variations recur with a FIXED periodicity of a year or less (daily/weekly/monthly/yearly) caused by calendar seasons or customs; cyclical fluctuations recur over MORE than a year WITHOUT a fixed periodicity, caused by the business cycle, and are less predictable.

Why is the time variable coded before estimating secular trend?

Coding shifts and rescales the origin so that sum(x)=0 (and sum(x^3)=0 for quadratic trend), which greatly simplifies the least squares normal equations, reducing them to simple formulas like a=ybar and b=sum(xy)/sum(x^2).

State one merit and one demerit of the method of moving averages.

Merit: an appropriately chosen period eliminates seasonal/cyclical fluctuations effectively. Demerit: it does not provide trend values at the beginning and end of the series, and gives no mathematical equation for forecasting.

How is a k-period moving average 'centred' when k is even?

By computing a further 2-period moving average of the original k-period moving averages, which realigns them to correspond with the actual observed time periods rather than falling between two of them.

What effect does shifting the origin of a linear trend equation have on its slope and intercept?

The slope b never changes; only the intercept changes, becoming a+bk (shifting k periods forward) or a-bk (shifting k periods backward).

Long Questions & Answers

Describe the four classical components of a time series in detail, and explain how the multiplicative and additive models combine them, illustrating with realistic examples for each component.

A time series records the same variable repeatedly over successive points in time, and a central insight of this chapter is that the resulting sequence of values is rarely a single, uniform kind of movement -- instead, it is generally the combined outcome of up to four distinct, separately-caused kinds of variation acting together, and separating these out is the whole purpose of time series analysis. The secular (or long-term) trend, denoted T, captures the smooth, gradual, persistent movement of the series in one consistent direction -- either steadily rising or steadily falling -- over a fairly long span of time, reflecting slow-moving underlying forces that don't reverse direction from one period to the next. Classic examples include the long-run decline in death rates as medical science and public health advance, the steadily increasing demand for wheat as a country's population grows year after year, or the gradual, sustained shift in automobile demand toward smaller, more fuel-efficient models as fuel costs and environmental awareness rise over decades. Seasonal variations, denoted S, are an entirely different kind of movement: short-term fluctuations that repeat with a FIXED, predictable periodicity of one year or less -- daily, weekly, monthly, or annually -- driven by causes tied directly to the calendar itself, such as the physical seasons, recurring religious festivals, or established social customs. A department store's weekly sales pattern (busier on weekends), a sharp jump in electricity consumption every summer as air conditioning use rises, or the

predictable post-Eid clearance sale spike are all textbook seasonal variations: the KEY diagnostic feature is that both the timing and the approximate size of the fluctuation repeat reliably, period after period, in a way that can be anticipated in advance. Cyclical fluctuations, denoted C, superficially resemble seasonal variations in that they too are recurring, wave-like oscillations around the underlying trend line, but they differ in one crucial respect: they do NOT have a fixed periodicity. A business cycle -- the classic source of cyclical fluctuation in economic time series -- might last two years in one instance and seven years in the next, moving irregularly through four recognised phases: a trough (or depression), the lowest point of the cycle; an expansion (or recovery), the upswing phase as the trough phase reverses and activity begins climbing; a peak (or boom/prosperity), the highest point the cycle reaches; and finally a recession (or contraction), where the upswing turns and begins declining again toward the next trough. Because cyclical fluctuations lack the seasonal component's fixed, predictable rhythm, they are inherently harder to forecast and, correspondingly, more dangerous for business and economic planning than seasonal variations, even though both are technically 'recurring' in character. Finally, irregular (or erratic, accidental, random) movements, denoted I, capture whatever variation is left over once trend, seasonal, and cyclical effects have all been accounted for -- unpredictable, non-recurring, one-off disturbances caused by genuinely random external events such as wars, natural disasters like floods or earthquakes, labour strikes, factory fires, or sudden political upheavals; a steel strike that delays production for a single week, or a factory fire that halts output for three weeks, are exactly this kind of component, precisely because neither event follows any discernible pattern that could ever be anticipated or built into a forecasting model. Once all four of these conceptually distinct components have been identified, the analyst must decide how they combine together to produce the ACTUAL observed value Y at any given moment, and two standard mathematical models are used for this purpose. The multiplicative model expresses this combination as a straightforward product, Y=TxSxCxI, under which the trend component T is expressed in the series' own original units (e.g. actual tons of steel, or actual rupees of sales), while the remaining three components S, C, and I are all expressed as unitless PERCENTAGE index numbers that scale the trend value up or down (a seasonal index of 120% for December sales would indicate December sales run 20% above what the trend alone would predict, for

instance). By long-standing convention, this multiplicative model is treated as the STANDARD model for time series analysis throughout applied statistics. The alternative additive model instead expresses the combination as a simple sum, Y=T+S+C+I, under which ALL FOUR components -- not just the trend -- must be expressed consistently in the series' own original units (e.g. all measured in actual rupees, or actual units sold, rather than mixing raw units for trend with percentage indices for the other three); this model implicitly assumes that the absolute SIZE of the seasonal, cyclical, and irregular fluctuations stays roughly constant over time, regardless of how large or small the underlying trend level happens to be at that particular moment, which is often a less realistic assumption for economic and business series (where fluctuations more commonly scale up proportionally as the trend level itself grows larger) -- helping to explain why the multiplicative model has become the conventional default choice in practice.

Compare the four methods of estimating secular trend -- free hand curve, semi-averages, moving averages, and least squares -- explaining the mechanics, and the relative merits and demerits, of each, and explain why least squares is generally considered the most rigorous of the four.

Estimating the secular trend of a time series -- isolating the smooth, long-run directional movement from the noisier seasonal, cyclical, and irregular fluctuations superimposed on top of it -- can be approached through four progressively more rigorous and objective methods, and understanding the trade-offs between them is essential to choosing the right one for a given practical situation. The method of free hand curve is the simplest and most intuitive: after plotting the historigram of the observed series, the analyst simply draws, purely by eye, a smooth straight line or curve through the scatter of points that seems to best capture the general upward or downward movement, ideally anchored through the point representing the overall mean of the time period and the overall mean of the observed values, since this is a reasonable condition the true trend line should satisfy. This method is fast, requires no calculation whatsoever, and effectively smooths out the noisier seasonal fluctuations by eye -- but its fundamental weakness is that it is entirely subjective: given the identical dataset, two different analysts, or even the same

analyst on two different days, could easily draw two visibly different trend lines, and there is no principled way to adjudicate which one is objectively 'more correct,' making the method unreliable for any application requiring reproducibility or defensible precision. The method of semi-averages introduces a first, modest dose of mathematical objectivity: the observed series is split into exactly two equal halves (with the single middle observation either dropped entirely, or duplicated into both halves, if the total number of periods happens to be odd), the arithmetic mean of each half is computed and plotted at that half's own midpoint in time, and a single straight line is then drawn connecting these two computed points, with its slope and intercept determined by ordinary two-point coordinate geometry rather than by eye. This produces a genuinely objective result, in the specific sense that any two analysts working from the identical data will always arrive at EXACTLY the same trend line -- a clear improvement over the free hand method -- but it carries its own serious limitations: because it relies on simple arithmetic means, it remains highly sensitive to unusually extreme values occurring within either half, and more fundamentally, the method is only appropriate at all when the underlying trend is genuinely linear (or very close to it); it offers no mechanism whatsoever for handling a visibly curved, non-linear trend pattern. The method of moving averages takes a rather different approach: instead of attempting to fit any single overall curve to the whole series at once, it works locally, replacing each individual observation with the average of itself and a symmetric window of k neighbouring periods, sliding this averaging window one period forward at a time across the entire length of the series (dropping the oldest observation from the window and including the next new one each time it slides). Because averaging over several adjacent periods tends to cancel out short-term seasonal and irregular noise while a genuine underlying trend movement tends to persist and reinforce itself across the averaging window, this method can be remarkably effective at revealing the trend -- PROVIDED the chosen averaging period k is deliberately matched to the actual periodicity of whatever seasonal or cyclical fluctuation needs to be eliminated (for instance, using a 12-month moving average to eliminate a fluctuation with an annual, 12-month period). Its notable drawbacks, however, are that no trend values can be computed at all near the very beginning and very end of the series (since a full symmetric window of k periods simply isn't available there), the resulting moving-average values remain

sensitive to any individual extreme observations that happen to fall within a given averaging window, and critically, the method never produces any actual mathematical EQUATION for the trend -- meaning it cannot, by itself, be used to forecast or extrapolate values into future time periods beyond the observed data, unlike the other three methods. The method of least squares is generally regarded as the most rigorous and, correspondingly, the most widely used of the four available methods, precisely because it fits a definite, explicit mathematical equation -- most commonly a straight line y-hat=a+bx, though a quadratic (parabolic) curve y-hat=a+bx+cx^2 is used instead whenever a straight line visibly fails to capture genuine curvature present in the data -- chosen specifically so as to make the sum of the SQUARED deviations between every observed value and its corresponding fitted trend value as small as mathematically possible, following exactly the identical least squares principle already developed for ordinary two-variable regression back in Chapter 14, merely with coded time itself now standing in as the independent variable x. Because this method yields a definite, fully-specified equation rather than merely a hand-drawn line or a table of local averages, it can be used both to compute a precise trend value at any observed historical time period AND to extrapolate forward, generating genuine numerical forecasts for future time periods purely by substituting new, larger values of x into the fitted equation. Additionally, because different candidate models (for instance, a straight line versus a quadratic curve, fitted to that exact same underlying dataset) can be directly and objectively compared against one another simply by comparing their two respective sums of squared residuals side by side -- the model producing the smaller sum of squared residuals is, by the very definition of the least squares criterion itself, unambiguously the better-fitting model for that particular dataset -- the least squares method offers a level of built-in, principled rigour that none of the three preceding methods can genuinely match, which is precisely why it has become the standard, default method of choice whenever a reasonably precise, forecastable trend equation is genuinely required in serious practical or academic work.

Multiple Choice Questions (MCQs)

A time series is best described as: (A) Any collection of numerical data (B) A sequence of observations recorded according to their time of occurrence (C) Only annual data (D) Data with no order at all

Correct answer: (B) A sequence of observations recorded according to their time of occurrence. A time series is specifically the sequence y1,...,yn recorded according to time of occurrence t1,...,tn -- order in time is essential.

A graph of a time series, with time on the x-axis, is called a: (A) Histogram (B) Scatter diagram (C) Historigram (D) Ogive

Correct answer: (C) Historigram. A historigram plots the observed values of Y against time t, joined by line segments -- distinct from a histogram (a bar chart of frequency distributions).

Which time series component recurs with a FIXED periodicity of one year or less? (A) Secular trend (B) Seasonal variations (C) Cyclical fluctuations (D) Irregular movements

Correct answer: (B) Seasonal variations. Seasonal variations have a fixed periodicity (daily/weekly/monthly/yearly); cyclical fluctuations recur but WITHOUT a fixed periodicity.

The four phases of a business cycle, in order, are: (A) Peak, trough, recession, expansion (B) Trough, expansion, peak, recession (C) Expansion, peak, trough, recession (D) Recession, trough, peak, expansion

Correct answer: (B) Trough, expansion, peak, recession. The standard order is: trough (depression) -> expansion (recovery) -> peak (boom) -> recession (contraction) -> back to trough.

In the multiplicative model Y = T x S x C x I, the components S, C, and I are expressed as: (A) Original units of Y, like T (B) Unitless percentage index numbers (C) Always equal to 1 (D) Negative numbers

Correct answer: (B) Unitless percentage index numbers. In the multiplicative model, only T is in Y's original units; S, C, and I are unitless percentage index numbers.

Coding the time variable so that sum(x) = 0 is done primarily to: (A) Make the data more accurate (B) Simplify the least squares normal equations (C) Eliminate seasonal variation (D) Remove irregular movements

Correct answer: (B) Simplify the least squares normal equations. sum(x)=0 (and sum(x^3)=0) collapses the least squares normal equations to simple formulas like a=ybar, b=sum(xy)/sum(x^2).

The main weakness of the free hand curve method for estimating trend is that it is: (A) Too mathematically complex (B) Subject to personal bias (C) Unable to handle any data at all (D) Only usable for seasonal data

Correct answer: (B) Subject to personal bias. Free hand curve fitting is drawn by eye, so different analysts can draw different trend lines from the same data -- it is highly subjective.

A k-period moving average needs to be 'centred' when: (A) k is odd (B) k is even (C) The trend is quadratic (D) n is less than 10

Correct answer: (B) k is even. When k is even, the moving average falls between two observed periods; a further 2-period average 'centres' it back onto an observed period.

If a linear trend equation is y-hat = a + bx with origin at year 1980, shifting the origin FORWARD by k years gives a new intercept of: (A) a - bk (B) a + bk (C) a x bk (D) a / bk

Correct answer: (B) a + bk. Shifting the origin forward by k substitutes (x+k) for x, giving y-hat=a+b(x+k)=(a+bk)+bx -- the new intercept is a+bk.

When comparing a linear trend and a quadratic trend fitted to the same data, the better-fitting model is the one with: (A) The larger slope b (B) The smaller sum of squared residuals, sum(e^2) (C) The larger intercept a (D) More data points

Correct answer: (B) The smaller sum of squared residuals, sum(e^2). By the least squares criterion, the model with the smaller sum(e^2) fits the observed data more closely and is preferred.

Quick Revision Summary

- Time series: y=f(t)+e, signal f(t) + noise e; historigram = time-series line graph
- 4 components: Secular trend (T, long-term smooth) | Seasonal (S, fixed periodicity <=1yr) | Cyclical (C, no fixed periodicity, business cycle) | Irregular (I, random/unpredictable)
- Business cycle phases: trough -> expansion -> peak -> recession
- Multiplicative: Y=TxSxCxI (T in original units, S/C/I unitless %) | Additive: Y=T+S+C+I (all original units)
- Coded time variable: x=(t-tbar)/h chosen so sum(x)=0=sum(x^3), simplifying least squares
- Free hand curve: quick but subjective/biased | Semi-averages: objective but only for linear trend, sensitive to extremes
- Moving averages: smooth out fixed-period fluctuations; no trend values at ends; no forecasting equation
- Least squares linear: b=sum(xy)/sum(x^2), a=ybar (when sum x=0) | Quadratic: adds c=[n.sum(x^2y)-sum(x^2)sum(y)]/[n.sum(x^4)-(sum x^2)^2]
- Best-fit model = smaller sum(e^2) = sum(y^2)-a.sum(y)-b.sum(xy)[-c.sum(x^2y)]
- Shifting origin (linear): only intercept changes, a+/-bk; slope b unchanged

Exam Tips

- Always draw the historigram first -- visual inspection tells you whether to expect a linear or curved trend before you compute anything
- Remember: seasonal = FIXED periodicity, cyclical = NO fixed periodicity -- this is the single most-tested distinction in this chapter
- When n is even and using semi-averages, the two midpoint x-values (x1, x2) will be at the centre of each half, not at integer years -- don't forget this when computing b=(y2'-y1')/(x2-x1)
- For moving averages, always double-check whether k is odd (no centring needed) or even (must centre with a further 2-period average)

- When coding time for least squares, decide up front whether the unit is 1 year or 1/2 year for EVEN numbers of periods -- this changes b (but not the trend values themselves) by a factor of 2
- To compare linear vs quadratic fits, compute BOTH sum(e^2) values and pick the smaller one -- don't just eyeball which curve 'looks' better

Chapter 17: Orientation of Computers

The final chapter of the Statistics Part-II syllabus steps away from probability and inference to introduce the machine that now performs almost all real-world statistical computation: the computer. It covers the computer's basic capabilities and history, the physical (hardware) and logical (software) building blocks of a computer system, how data moves in and out through input and output devices, the layers of software from the operating system up to application programs and programming languages, and finally the number systems -- decimal, binary, octal, and hexadecimal -- that explain how a machine built from simple on/off electrical switches can represent and process any kind of data.

This chapter is conceptual and definitional rather than computational: there are no statistical formulas or worked numerical examples as in earlier chapters, and success here depends on precisely knowing terms, classifications, and the roles different hardware and software components play within a complete computer system.

Learning Objectives

- Define a computer and describe its key capabilities: speed, data storage, data processing, accuracy, and diligence
- Outline the historical milestones in the development of the computer, from Pascal's Pascaline to the modern stored-program computer
- Distinguish digital, analog, and hybrid computers
- Classify computers by size and processing power into micro, mini, mainframe, and super computers
- Distinguish hardware from software and identify the four major hardware components of a personal computer
- Describe the functions of the Control Unit and the Arithmetic Logic Unit (ALU) within the CPU
- Distinguish main memory (RAM) from secondary storage, and identify common input and output devices

- Distinguish system software from application software, and list the functions of an operating system
- Distinguish low-level from high-level programming languages, and describe assemblers, interpreters, and compilers
- Convert numbers between the decimal, binary, octal, and hexadecimal number systems, and explain why computers use binary internally

Key Concepts

17.1 Introduction to Computer

A computer is an electronic device used to store and process data to solve problems according to a set of instructions given to it; the word comes from 'compute', meaning to calculate. A modern computer's key capabilities are: Speed -- the number of instructions processed per second, measured in megahertz (MHz) or gigahertz (GHz), where hertz is the number of electronic pulses generated per second by the processor's clock. Data Storage -- the ability to store large amounts of data in memory and retrieve it at very high speed; data itself is a combination of characters, numbers, and symbols collected for a specified purpose. Data Processing -- a series of operations (arithmetic, logical, classification, arrangement, transmission) performed on data to produce results, called output or information. Accuracy -- a computer performs millions of error-free operations per second, PROVIDED the input data and program instructions given to it are themselves correct. Diligence -- a computer can work for long hours without tiring, and its accuracy is unaffected by how long it has been running.

17.2 History of Computer

Blaise Pascal (France) built the first mechanical adding machine, the 'Pascaline', in 1642, capable of addition and subtraction; Gottfried Wilhelm von Leibnitz modified it in 1671, adding a 'multiplier wheel' enabling all four basic arithmetic operations. Charles Babbage (UK) designed the 'Analytical Engine' in 1833 -- the first programmable computer, consisting of a storage unit, a mill (for arithmetic), and a control unit, programmed using punched cards. The 'Harvard Mark-I' was developed at Harvard University between 1937 and 1943. In 1943, J.W. Mauchly and J.P. Eckert developed ENIAC (Electronic Numerical Integrator and Calculator)

at the Moore School of Engineering, USA, completed in 1946; its programs were stored externally on tape and executed sequentially. In 1944, John von Neumann proposed storing the computer's program ELECTRONICALLY INSIDE the computer itself -- the final conceptual breakthrough underlying every modern computer's design.

17.3 Types of Computers

Computers are of three types based on how they process signals. A Digital Computer works only with two signals, 0 and 1, counting numbers/digits and giving output in digital form; data and instructions are stored as coded 0's and 1's; examples include digital watches and digital thermometers. An Analog Computer does not operate directly with digital signals -- it measures physical quantities and gives output on a continuous scale (a graph or a dial reading); examples include a dial clock, a thermometer, and a weighing machine; its results are less accurate than a digital computer's. A Hybrid Computer combines features of both, accepting input and giving output in either analog or digital form; a modem is a common example.

17.4 Classification of Computers

By speed and memory size, computers are classified into four groups. Micro Computers (personal computers) are designed for one user at a time, commonly used in offices, homes, and schools, with processing speed measured in millions of instructions per second (MIPS); laptops and notebooks are micro computers. Mainframe Computers are very large, very high-speed computers used by large organisations (banks, insurance companies, research institutes, weather bureaus) -- the largest IBM S/390 mainframe can support 50,000 users while executing over 1.6 billion instructions per second. Mini Computers, introduced in the 1960s, are smaller than mainframes but can still handle far more data than personal computers, used for large-organisation record-keeping, experimental data analysis, or factory process control; they cost from about $18,000 to $500,000. Super Computers are the most powerful computers built, capable of over 1 trillion calculations per second, used by nuclear scientists to model fission/fusion and to map the human genome; they can cost tens of millions of dollars and consume enormous amounts of electricity.

17.5-17.6 Hardware, Software, and Hardware Components

Hardware is the physical parts of a computer (CPU, monitor, mouse, keyboard, etc.); Software is the set of instructions given to a computer to solve a problem or control its operation, prepared in a programming language (e.g. Microsoft Word, Excel). A personal computer's critical hardware falls into four categories: the Central Processing Unit (CPU), Main Memory, Input/Output Devices, and Secondary Storage.

The CPU is the 'brain' of the computer, where data is manipulated, and has two basic parts. The Control Unit manages all the computer's resources -- like a traffic cop directing the flow of data through the CPU and to other devices -- interpreting each coded instruction (add, subtract, store) and moving data between input devices, memory, the ALU, and output devices accordingly; it is the logical hub/nervous system of the computer. The Arithmetic Logic Unit (ALU) performs all arithmetic operations (+, -, x, /, raised to a power) and logical/ comparison operations (=, !=, >, <, >=, <=) on data, using high-speed registers built directly into the CPU to hold data currently being processed.

Main Memory, also called RAM (Random Access Memory) or primary storage, temporarily stores data and program instructions WHILE they are being processed; each storage location has an address, like a post-box number; reading data does not destroy it, but writing new data to a location erases whatever was there before. Memory is measured in bytes (the storage needed for one character): 1 Kilobyte (KB) = 1,024 bytes; 1 Megabyte (MB) = 1,048,576 bytes; 1 Gigabyte (GB) = 1,073,741,824 bytes; 1 Terabyte (TB) = 1,099,511,627,776 bytes. Cache memory is a small amount of very high-speed memory placed between the CPU and main memory to store the most frequently used instructions/data, speeding up processing whenever the needed item is already present in cache (a 'cache hit').

Secondary (auxiliary) storage holds programs and data when they are NOT being processed (unlike RAM, which is only temporary). Storage media are the physical materials data is stored on; storage devices read/write to that media. Two main storage technologies are used: Magnetic storage (diskettes, hard disks, high-capacity floppy disks, disk cartridges, magnetic tape) and Optical storage (CD-

ROM, DVD-ROM, CD-R, CD-RW, Photo CD); read/write heads (similar to a tape recorder's) read from or write to a spinning magnetic disk.

17.7 Input Devices and Output Devices

Input devices translate people-readable data into computer-readable ('0'/'1', off/on) form, and fall into three categories: keyboards, pointing devices, and source data entry devices. A Keyboard has alphabet keys, numeric keys, and special keys (F1-F12, Alt, Ctrl, Shift, Tab, Caps Lock, Enter, etc.) -- the standard keyboard has 101 keys, though 104/106/110-key versions also exist. Pointing devices control the on-screen cursor: a Mouse is rolled on a desktop to move the pointer, with functions point, click, drag, drop, and right-click; a Trackball is a stationary device rotated by the fingers/palm, well suited to portable computers; a Joystick has a vertical handle mounted on a base; a Touch pad is a small flat surface the user slides a finger across; a Light Pen is a light-sensitive stylus used directly on the display screen.

Output devices receive data from the CPU in binary code and convert it to a readable form. Output is either sent to secondary storage (for later reuse as input) or delivered to people, split into Softcopy Output (temporary, erased when the computer switches off, e.g. a screen display -- delivered via monitors, PC projectors, sound systems, collectively called Visual Display Units or VDUs) and Hardcopy Output (permanent, printed on paper, e.g. via a printer or a plotter).

17.8-17.13 Software, Operating Systems, Programming Languages, and Program Development

System software consists of all programs, including the operating system, that control the computer's own equipment -- starting it up, loading/executing/storing application programs, storing/retrieving files, formatting disks, sorting data, and translating instructions into machine language; it is classified into operating systems, utilities, and language translators. An Operating System (OS) is an integrated set of programs managing a computer's hardware resources, loaded into main memory every time the computer starts. Its functions include: processor management, memory management, input/output management, file management, job-priority scheduling, automatic job-to-job transition, interpreting commands, coordinating compilers/assemblers/utilities, ensuring data security

and integrity, producing error/debugging aids, and maintaining a system usage log. Common operating systems include DOS, WINDOWS, OS/2, UNIX, and LINUX. DOS (Disk Operating System), the most widely used OS on early personal computers, is text-driven -- the user types command lines that the computer executes; MS-DOS and PC-DOS, both originally from Microsoft (1981), are the best-known versions, with later versions adding a tree-structured directory scheme and a menu-driven 'DOS shell'.

Application software consists of programs that tell a computer how to produce information for a specific real-world task -- word processing, desktop publishing, spreadsheets, databases, presentation graphics, communications, electronic mail, personal information management, and project management are the most widely used categories on personal computers.

A programming language is a means of communication between a user/ programmer and the computer, used to write programs (code) that solve problems; each language's own rules for writing valid programs are called its syntax. Programming languages are of two types. Low-level languages are close to machine code: Machine Language uses binary strings of 0's and 1's directly executed by the computer (the most fundamental language); Assembly Language uses symbolic codes (mnemonics) instead of raw binary, translated into machine code by an assembler. High-level languages are close to human (English-like) language, each with its own syntax: examples include ALGOL (ALGOrithmic Language), BASIC (Beginners All-purpose Symbolic Instruction Code), COBOL (COmmon Business Oriented Language), PASCAL (named after the mathematician Pascal), FORTRAN (FORmula TRANslation), and C (a general-purpose language widely used in science and elsewhere).

A language processor (translator) converts a source program into machine code (strings of 0's and 1's). There are three types: an Assembler translates assembly language into machine code; an Interpreter translates and executes a source program ONE instruction at a time; a Compiler translates an entire program written in a high-level language (typically all at once, before execution begins).

A computer program is a detailed set of instructions directing a computer to process data into information; it is also called software. Program development follows five standard steps: (1) Review specification -- the programmer reviews

the system analyst's specification; (2) Design -- the programmer determines and documents the specific actions the computer will take; (3) Code -- the programmer writes the actual program instructions; (4) Test -- the written program is tested to confirm it performs as intended; (5) Finalize documentation -- explanatory information about steps 1-4 is organised and completed.

17.14-17.16 Number Systems and Data Representation

A number system is a set of digits, symbols, and rules used to express quantities, and is named after its base -- the total number of distinct digits it uses. The Decimal Number System has base 10 (digits 0-9); each digit's value depends on its position (its 'weight'), e.g. 3046 = 3x10^3 + 0x10^2 + 4x10^1 + 6x10^0. The Binary Number System has base 2 (digits 0, 1 only), e.g. 1011 (binary) = 1x2^3 + 0x2^2 + 1x2^1 + 1x2^0 = 11 (decimal). The Octal Number System has base 8 (digits 0-7). The Hexadecimal Number System has base 16 (digits 0-9 plus letters A-F representing 10-15).

To a computer, everything -- letters, punctuation, sounds, pictures, even the computer's own instructions -- is ultimately represented as numbers, specifically as binary strings of 0's and 1's (e.g. the letter 'H' is stored as 0100 1000). Computers use the binary system because their physical electronic/electrical components (transistors, magnetic materials, wires) can naturally represent only TWO distinguishable states -- on/off, conducting/non-conducting, magnetised/non-magnetised, pulse-present/pulse-absent -- for which the two-digit binary system is the most natural fit; because only two digits need to be handled (versus ten for decimal), computer circuit design is simplified, yielding cheaper and more reliable circuits; and everything that can be computed in base 10 can equally be computed in binary. Octal and hexadecimal are used alongside binary because they represent binary values in a much more COMPACT form while still converting to/from binary very efficiently -- an 8-digit binary number, for instance, can be represented by just a 2-digit hexadecimal number.

Important Definitions

What is a computer?: An electronic device used to store and process data to solve different problems according to a set of instructions given to it; the word comes from 'compute', meaning to calculate.

What is hardware?: The physical parts of a computer -- all the physical devices or units that make up a computer system, such as the CPU, monitor, mouse, and keyboard.

What is software?: The set of instructions given to a computer to solve a problem or to control its operation, prepared in a computer programming language.

What is the Central Processing Unit (CPU)?: The 'brain' of the computer, where data is manipulated; it has two basic parts, the Control Unit and the Arithmetic Logic Unit (ALU).

What is main memory (RAM)?: Also called primary storage, the memory contained in the processor unit that TEMPORARILY stores data and program instructions while they are being processed.

What is secondary (auxiliary) storage?: Storage that holds programs and data when they are NOT being processed, unlike main memory which only holds them temporarily during processing.

What is an operating system (OS)?: An integrated set of programs that manages a computer system's hardware resources, loaded into main memory every time the computer is started, to improve performance, efficiency, and ease of use.

What is the difference between a low-level and a high-level programming language?: A low-level language (machine or assembly language) is close to machine code; a high-level language (e.g. BASIC, COBOL, FORTRAN, PASCAL, C) is close to human (English-like) language and uses its own syntax.

What is a compiler?: A translator (language processor) that converts an entire program written in a high-level language into machine code.

What is the binary number system, and why is it used in computers?: A number system with base 2, using only the digits 0 and 1; it is used in computers because their electronic components can naturally represent only two states (on/off), simplifying circuit design and making binary the most efficient representation for digital hardware.

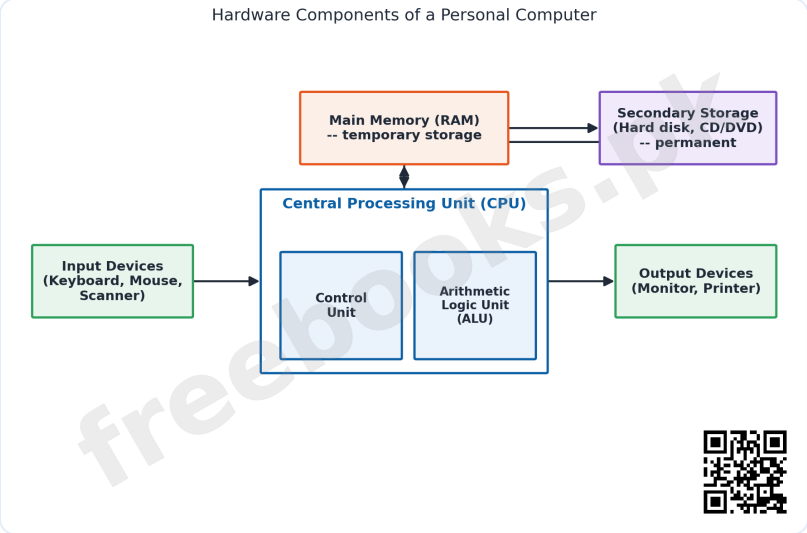
Key Facts and Relations

Topic	Key Fact / Relation
Computer speed units	Hertz (Hz) = pulses/vibrations per second; Megahertz (MHz) = 10^6 Hz; Gigahertz (GHz) = 1000 MHz
Memory/storage units (bytes)	1 KB = 1,024 B; 1 MB = 1,048,576 B; 1 GB = 1,073,741,824 B; 1 TB = 1,099,511,627,776 B
Number system bases	Decimal = base 10 (0-9); Binary = base 2 (0-1); Octal = base 8 (0-7); Hexadecimal = base 16 (0-9, A-F)
Positional (weighted) value of a digit	N = d_n.b^n + ... + d_1.b^1 + d_0.b^0 (b = base of the number system)
Decimal example	3046 (base 10) = 3x10^3 + 0x10^2 + 4x10^1 + 6x10^0
Binary example	1011 (base 2) = 1x2^3 + 0x2^2 + 1x2^1 + 1x2^0 = 11 (decimal)
CPU = Control Unit + ALU	Control Unit manages/directs data flow; Arithmetic Logic Unit (ALU) performs arithmetic and logical operations
RAM vs Secondary Storage	RAM (primary storage) = temporary, holds data WHILE processing; Secondary storage (auxiliary) = holds data when NOT being processed
Software categories	System software (OS, utilities, language translators) + Application software (word processing, spreadsheet, database, etc.)
Low-level vs high-level languages	Low-level: Machine language (binary), Assembly language (mnemonics) | High-level: BASIC, COBOL, FORTRAN, PASCAL, C, ALGOL
Language processors	Assembler: assembly -> machine code | Interpreter: translates & executes one instruction at a time | Compiler: translates entire high-level program at once
	Review specification -> Design -> Code -> Test -> Finalize documentation

Topic	Key Fact / Relation
Five steps of program development	

Diagrams

Hardware Components of a Personal Computer: A block diagram showing the four critical hardware categories -- Input Devices, Central Processing Unit (with its Control Unit and Arithmetic Logic Unit shown as sub-blocks), Main Memory, and Output Devices -- with Secondary Storage connected alongside, and arrows showing the flow of data between them, illustrating how data moves from input through processing and memory to output

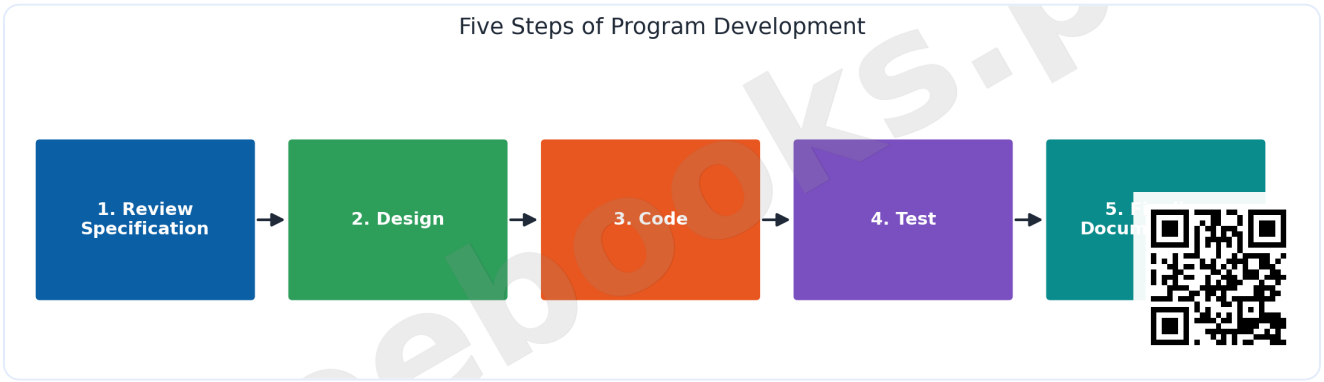


Hardware Components of a Personal Computer

Number System Conversion Table (0-16): A table/chart showing decimal numbers 0 through 16 alongside their binary, octal, and hexadecimal representations, visually illustrating how the same quantity is expressed differently depending on the base of the number system used

Number System Conversion Table (0-16)

Five Steps of Program Development: A horizontal flow diagram of the five standard steps of computer program development, in order: Review Specification, Design, Code, Test, and Finalize Documentation, connected by arrows to show the sequential nature of the process



Five Steps of Program Development

Short Questions & Answers

Define a computer and state two of its key capabilities.

A computer is an electronic device that stores and processes data to solve problems according to given instructions. Two key capabilities: Speed (millions of instructions processed per second) and Accuracy (error-free calculation, provided input data and instructions are correct).

Distinguish between hardware and software.

Hardware is the physical parts of a computer (CPU, monitor, keyboard, etc.); software is the set of instructions given to the computer to solve a problem or control its operation, prepared in a programming language.

What are the two basic parts of the CPU, and what does each do?

The Control Unit manages and directs the flow of data through the CPU and to other devices; the Arithmetic Logic Unit (ALU) performs all arithmetic and logical (comparison) operations on data.

Distinguish between main memory (RAM) and secondary storage.

Main memory (RAM/primary storage) temporarily holds data and instructions WHILE they are being processed; secondary storage (auxiliary storage) holds programs and data when they are NOT being processed, providing permanent storage.

What is the difference between an interpreter and a compiler?

An interpreter translates and executes a source program one instruction at a time; a compiler translates an entire high-level-language program into machine code, typically all at once before execution.

Why do computers use the binary number system rather than decimal?

Because their electronic components (transistors, magnetic materials, wires) can naturally represent only two distinguishable states (on/off), for which the two-digit binary system is the most natural and efficient fit, also simplifying circuit design.

Long Questions & Answers

Describe the four major hardware components of a personal computer -- the CPU, main memory, input/output devices, and secondary storage -- explaining the role each plays and how they work together to process data into information.

A complete personal computer, despite the enormous variety of tasks it can perform, is built from just four critical categories of hardware components working together, and understanding how information actually flows between these four categories is essential to understanding how a computer functions as a whole system. The Central Processing Unit, universally described as the 'brain'

of the computer, is the place where data is actually manipulated and processed; in an average micro computer the entire CPU is fabricated onto a single chip called a microprocessor, while larger systems such as mainframes and supercomputers may distribute processing tasks across many separate processing chips working in parallel. The CPU itself is built from two distinct basic parts that must cooperate closely: the Control Unit, which manages all of the computer's resources and directs the flow of data through the CPU and out to other devices -- functioning much like a traffic policeman directing traffic, or like the nervous system of a living body -- interpreting each basic coded instruction (such as 'add', 'subtract', or 'store') and orchestrating exactly where data needs to move next (from an input device into memory, from memory into the Arithmetic Logic Unit for processing, from the Arithmetic Logic Unit back into memory, and finally from memory out to an output device); and the Arithmetic Logic Unit itself, which is the part of the processor where the actual arithmetic operations (addition, subtraction, multiplication, division, raising to a power) and logical/comparison operations (testing whether one value is equal to, greater than, less than, or not equal to another) are physically carried out on the data, using a small number of extremely high-speed storage locations called registers, built directly into the CPU chip itself, to hold whichever specific pieces of data are currently being worked on at that instant. Working alongside the CPU is Main Memory, more commonly called RAM (Random Access Memory) or primary storage, which is physically housed within the processor unit and whose entire purpose is to hold data and program instructions TEMPORARILY, for exactly as long as they are actively being processed by the CPU -- each individual storage location within RAM has its own unique address, functioning conceptually much like a post-office box number, which the computer uses (entirely automatically, without any need for the user to think about it) both to retrieve previously stored data (a read operation, which never destroys or alters the data being read) and to store brand-new data (a write operation, which DOES permanently overwrite and erase whatever data happened to occupy that exact memory address beforehand). Input and Output Devices form the third essential category, and exist specifically because a computer would otherwise be entirely unable to interact meaningfully with the people using it: input devices such as a keyboard, a mouse, a scanner, or a digital camera translate people-readable information into the '0' and '1' electrical-signal form that the computer's internal circuitry can

actually process, while output devices such as a monitor, a printer, or a plotter perform the exact reverse operation, receiving the CPU's raw binary-coded results and converting them back into a form that is genuinely readable and useful to a human being, whether that form is a temporary softcopy display on a screen or a permanent hardcopy printout on paper. Finally, Secondary Storage (also called auxiliary storage) exists to solve a problem that main memory, by its very nature, cannot solve on its own: because RAM only holds data and programs temporarily WHILE they are actively being processed, and because RAM's contents are lost entirely whenever the computer is switched off, a genuinely useful computer system additionally requires a place to keep program files and their associated data safely and PERMANENTLY, for whenever they are not currently in active use -- this permanent-storage role is filled by secondary storage devices and their associated storage media, most commonly employing either magnetic storage technology (hard disks, diskettes, magnetic tape) or optical storage technology (CD-ROM, DVD-ROM, and related formats), or in some modern devices a genuine combination of both underlying technologies working together. Taken together, these four hardware categories form a single, continuously repeating processing cycle: input devices first bring raw data INTO the system; that raw data, along with the relevant program instructions needed to process it, is then temporarily held in main memory for as long as active processing requires; the CPU's Control Unit and Arithmetic Logic Unit then actually carry out whatever specific processing operations the loaded program instructions call for; and finally, the processed results are delivered back OUT to the user through output devices, while secondary storage remains available throughout this entire cycle to permanently preserve whatever data or completed results genuinely need to persist beyond that one individual processing session.

Explain the classification of programming languages into low-level and high-level languages, describe the role of language processors (assemblers, interpreters, and compilers) in bridging them to machine code, and explain why high-level languages are generally preferred for most modern programming despite machine language being what a computer actually executes directly.

Every single instruction a computer ultimately carries out must, at the deepest hardware level, exist as a string of binary digits -- 0's and 1's -- because this is the only form of instruction that the computer's own electronic circuitry can directly execute without any further translation whatsoever; this most fundamental level is called machine language, and a program written directly in machine language, however tedious and error-prone this would be for a human programmer to actually write by hand, requires no additional translation step at all before the computer can run it. Because writing entire, genuinely useful programs directly as raw strings of binary digits is extraordinarily impractical, time-consuming, and error-prone for human programmers to do reliably, programming languages have been organised, historically, into two broad conceptual categories reflecting how CLOSE each language sits to genuine machine code versus how close it sits to ordinary, familiar human language. Low-level languages sit at the machine-code end of this spectrum: besides raw machine language itself, the other principal low-level language is assembly language, which replaces cryptic, entirely numeric binary instruction codes with somewhat more human-memorable symbolic codes, technically called mnemonics (short alphabetic abbreviations standing in for particular machine operations, such as ADD or SUB, rather than their raw numeric binary equivalents) -- assembly language remains extremely closely tied to the specific underlying hardware's own instruction set and architecture, offering the programmer very fine-grained, precise control over exactly what the machine does at every single step, but at the real cost of being tedious to write, difficult to genuinely understand at a glance, and entirely non-portable from one type of computer hardware to a different type. High-level languages, by clear contrast, sit at the opposite, human-language end of the very same spectrum: languages such as BASIC, COBOL, FORTRAN, PASCAL, ALGOL, and C all allow a programmer to express computational logic and intent using instructions, keywords, and

overall syntax that read far more like structured, disciplined English than like any kind of raw machine code, each such high-level language governed by its own particular, well-defined set of syntax rules that any valid program written in that specific language must always correctly follow. Because the physical computer hardware itself, however, can genuinely only ever directly execute raw machine code -- and absolutely nothing else -- some kind of translation step is always unavoidably necessary to bridge the substantial gap between whatever a human programmer has actually written and what the computer's own circuitry can directly run, and this essential bridging role is performed by a category of specialised software collectively known as language processors, or equivalently, translators. An assembler is the specific translator used for assembly-language programs, converting the programmer's symbolic mnemonic codes directly into their exactly corresponding machine-code binary equivalents, essentially performing a fairly direct, largely one-to-one substitution between assembly instructions and machine instructions. For high-level languages, two rather different translation strategies exist side by side, each carrying its own distinctive practical trade-offs. An interpreter translates and immediately executes a source program strictly ONE individual instruction at a time, in sequential order -- this approach makes interactive testing, debugging, and rapid experimentation considerably easier and faster for the programmer (since any given instruction's effects can be observed by the programmer essentially immediately, without needing to wait for the entire surrounding program to finish translating first), but interpreted programs generally run measurably more slowly overall than compiled ones do, precisely because every single instruction must be freshly re-translated again, completely from scratch, every single time that same program is actually run. A compiler, by contrast, instead translates an ENTIRE high-level-language program, all at once and in its complete entirety, into a fully self-contained machine-code version, BEFORE the program is ever actually executed for the very first time -- this approach requires a distinct, separate compilation step to be performed up front before the program can be run at all, but compiled programs subsequently run substantially faster afterwards (since no further translation overhead of any kind remains necessary at actual run-time), which is precisely why compilers are strongly favoured for large-scale, performance-critical, or genuinely commercial-grade software of essentially every kind. High-level languages have become overwhelmingly dominant in

modern programming practice for several closely interconnected, mutually reinforcing reasons: they are dramatically easier for human programmers to read, correctly write, and reliably maintain over time, since their syntax deliberately and closely resembles familiar structured English rather than opaque binary or cryptic assembly mnemonics; a single high-level-language program can very often be compiled or interpreted to run correctly on many quite different types of underlying computer hardware, provided only that a suitable compiler or interpreter for that particular high-level language has been made available for each specific target machine -- a portability advantage that low-level languages, tied intimately to one specific machine's own particular architecture, simply cannot offer in anything like the same way; and high-level languages allow a programmer to focus their own attention productively on solving the actual real-world problem itself at hand, rather than being forced to spend the majority of their available time and effort managing extremely low-level, tedious hardware-specific implementation details that a good compiler or interpreter can, and reliably does, handle correctly and automatically on the programmer's own behalf.

Multiple Choice Questions (MCQs)

A computer is best defined as: (A) A device that only performs arithmetic (B) An electronic device that stores and processes data according to given instructions (C) A type of software (D) A type of programming language

Correct answer: (B) An electronic device that stores and processes data according to given instructions. A computer is an electronic device used to store and process data to solve problems according to a set of given instructions.

Which scientist designed the 'Analytical Engine', the first programmable computer? (A) Blaise Pascal (B) Gottfried Leibnitz (C) Charles Babbage (D) John von Neumann

Correct answer: (C) Charles Babbage. Charles Babbage, a UK mathematician, designed the Analytical Engine in 1833, the first programmable computer.

A computer that operates using only two signals, 0 and 1, is called a: (A) Analog computer (B) Digital computer (C) Hybrid computer (D) Mini computer

Correct answer: (B) Digital computer. A digital computer works with digits, operating with only two signals (0 and 1), storing data and instructions as coded binary values.

The physical parts of a computer, such as the CPU, monitor, and keyboard, are collectively called: (A) Software (B) Firmware (C) Hardware (D) Application programs

Correct answer: (C) Hardware. Hardware refers to all the physical devices or units that make up a computer system.

Within the CPU, which unit performs arithmetic and logical operations on data? (A) Control Unit (B) Arithmetic Logic Unit (ALU) (C) Main Memory (D) Secondary Storage

Correct answer: (B) Arithmetic Logic Unit (ALU). The ALU is the part of the processor where all arithmetic (add, subtract, etc.) and logical (comparison) operations are actually performed.

RAM (main memory) is best described as: (A) Permanent storage that never loses data (B) Temporary storage used while data is being processed (C) A type of input device (D) A type of output device

Correct answer: (B) Temporary storage used while data is being processed. RAM, also called primary storage, temporarily stores data and program instructions ONLY while they are being processed.

1 Megabyte (MB) is equal to: (A) 1,000 bytes (B) 1,024 bytes (C) 1,048,576 bytes (D) 1,000,000,000 bytes

Correct answer: (C) 1,048,576 bytes. The actual value of 1 MB is 1,048,576 bytes (1,024 x 1,024), though it is often approximated as 1,000,000 bytes.

Which of the following is an example of a high-level programming language? (A) Machine language (B) Assembly language (C) FORTRAN (D) Binary code

Correct answer: (C) FORTRAN. FORTRAN (FORmula TRANslation) is a high-level language, close to human/English-like syntax; machine and assembly languages are low-level.

A translator that converts an entire high-level-language program into machine code before execution is called a: (A) Assembler (B) Interpreter (C) Compiler (D) Operating system

Correct answer: (C) Compiler. A compiler translates the entire program written in a high-level language into machine code, typically all at once before running it.

Computers use the binary number system primarily because: (A) Binary numbers are easier for humans to read (B) Electronic components can naturally represent only two distinguishable states (on/off) (C) Binary uses fewer digits to write large numbers (D) Decimal numbers cannot be stored electronically

Correct answer: (B) Electronic components can naturally represent only two distinguishable states (on/off). Electronic/electrical components (transistors, magnetic materials, wires) can only indicate two states (on/off), making binary the most natural and efficient fit.

Quick Revision Summary

- Computer capabilities: Speed (MHz/GHz) | Storage | Data Processing | Accuracy | Diligence
- History: Pascal's Pascaline (1642) -> Leibnitz's Multiplier Wheel (1671) -> Babbage's Analytical Engine (1833) -> Harvard Mark-I -> ENIAC (1946) -> von Neumann's stored program (1944)
- 3 types by signal: Digital (0/1) | Analog (continuous scale) | Hybrid (both)
- 4 classes by size/speed: Micro (personal) | Mini | Mainframe | Super
- Hardware = physical parts | Software = instructions (System software + Application software)
- CPU = Control Unit (manages data flow) + ALU (arithmetic/logical operations)
- RAM (primary, temporary) vs Secondary Storage (auxiliary, permanent); magnetic vs optical storage
- Input: keyboard, mouse, trackball, joystick, touch pad, light pen | Output: softcopy (screen) vs hardcopy (printer/plotter)
- OS functions: processor/memory/I-O/file management, job priority, security, error-detection aids
- Low-level: machine language, assembly language | High-level: BASIC, COBOL, FORTRAN, PASCAL, C, ALGOL
- Language processors: Assembler (assembly->machine) | Interpreter (line-by-line) | Compiler (whole program at once)

- Number systems: Decimal(10), Binary(2), Octal(8), Hexadecimal(16) -- binary used because hardware has only 2 states

Exam Tips

- Don't confuse the Control Unit (directs/manages data flow) with the ALU (actually performs arithmetic/logical operations) -- this distinction is frequently tested
- Remember RAM = temporary/while-processing, Secondary storage = permanent/not-currently-processing -- the single biggest source of confusion in this chapter
- Memorize the byte units precisely: 1 KB=1,024 B, 1 MB=1,048,576 B, 1 GB=1,073,741,824 B -- exam MCQs often test the EXACT actual value, not the rounded approximation
- For number system conversions, always write out the positional (weighted) expansion first (d_n.b^n+...+d_0.b^0) before calculating -- this avoids simple arithmetic slips
- Keep language-processor roles straight: Assembler=assembly language only; Interpreter=line-by-line; Compiler=whole program at once, both for high-level languages
- Learn the five program-development steps IN ORDER: Review specification -> Design -> Code -> Test -> Finalize documentation -- a common short-answer question