ICS Part-II Statistics — Punjab Board

CHAPTER 12

Estimation

Statistics • 2nd Year • ICS • Punjab Board • Free PDF Notes

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

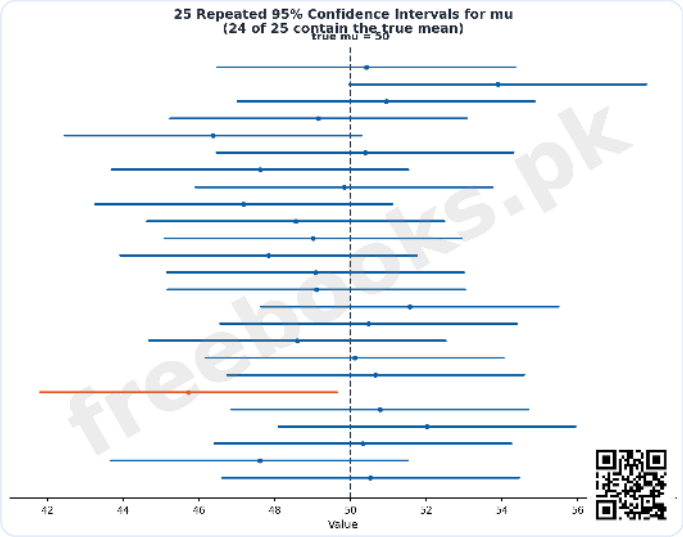
replacement from a finite population and the sample is more than 5% of the population (n>0.05N); it may be ignored otherwise.

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

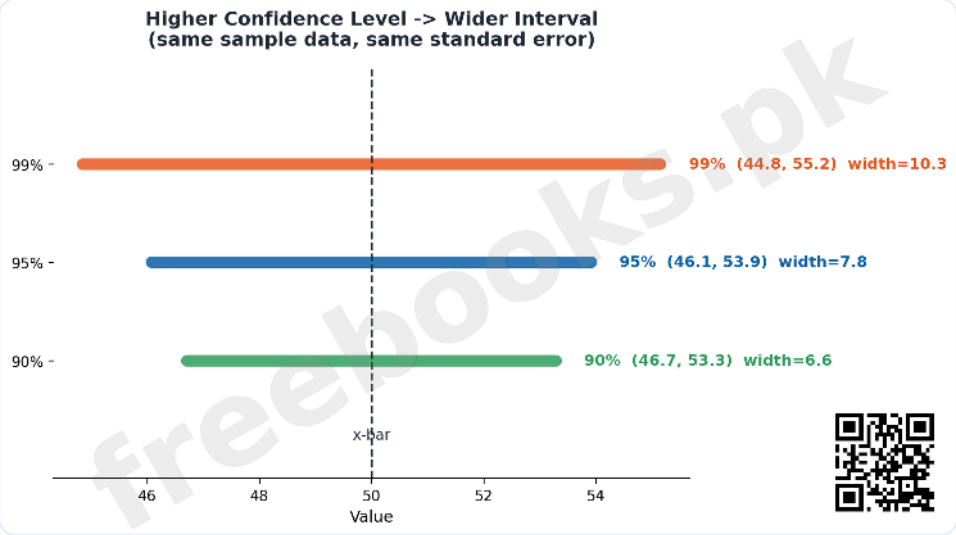
Diagrams

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



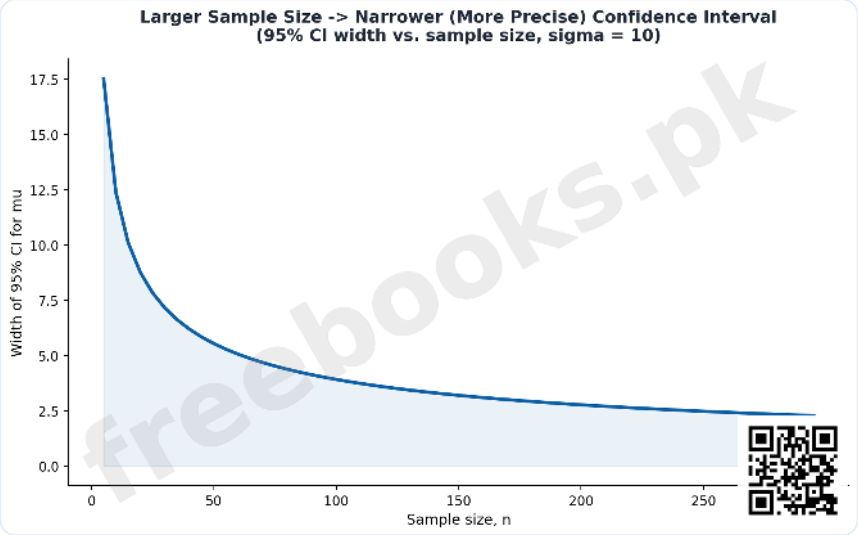
Repeated Confidence Intervals for mu

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