Freebooks.pk

Free Notes • Past Papers • Guides for Students

ICS Part-II Statistics — Punjab Board

CHAPTER 15

Association

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

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.

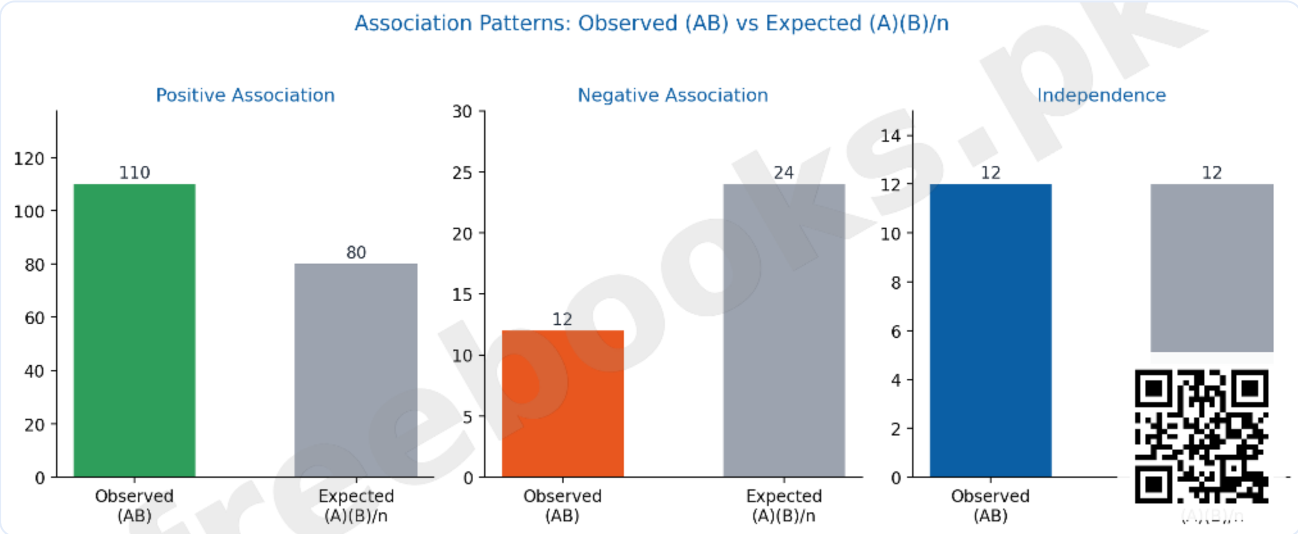
Key Facts and Relations

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

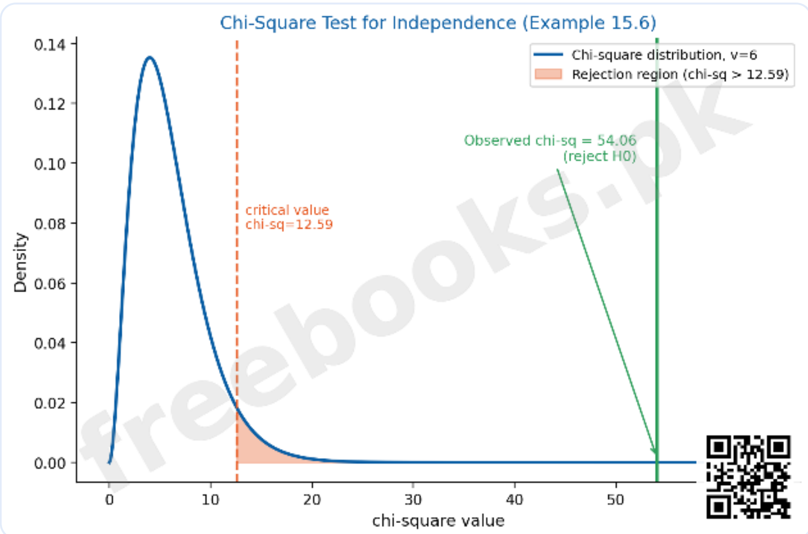
Association Patterns: Positive, Negative, Independent: A 3-panel grouped bar chart comparing the observed frequency (AB) against the expected-under-independence frequency (A)(B)/n for three worked cases: positive association

((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



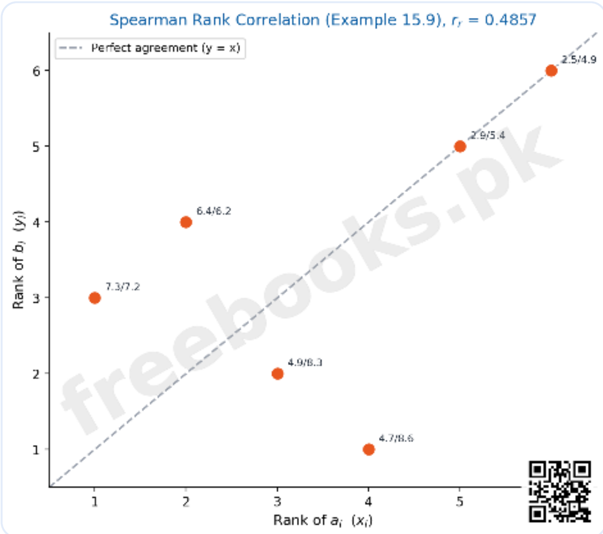
Association Patterns: Positive, Negative, Independent

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 H_0 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