VU LEARNING HUB

vulearninghub.xyz

HANDOUT

STA644

Non-Parametric Statistics

STA644 by VU Learning HUB

Prepared for VU students • Free study resources

https://vulearninghub.xyz

STA644

Non-Parametric Statistics

Study Handout

Complete lecture-wise notes for Virtual University students

About This Handout

This handout covers the complete STA644 course: distribution-free statistical methods used when data do not satisfy the normality or other assumptions required by classical parametric tests. Each lesson presents the concept, the test statistic and its sampling distribution, a worked example, and exam-focused pointers. Learn the decision rules and the conditions for each test — VU papers frequently ask you to choose the correct test for a given situation and to perform the ranking procedure step by step.

Table of Contents

Lesson 1	Introduction to Non-Parametric Methods
Lesson 2	Measurement Scales and Their Role
Lesson 3	The Sign Test — Concept and Procedure
Lesson 4	Wilcoxon Signed-Rank Test (One Sample / Paired)
Lesson 5	Mann-Whitney U Test (Two Independent Samples)
Lesson 6	Kruskal-Wallis Test (Three or More Independent Samples)
Lesson 7	Friedman Test (Randomized Block / Repeated Measures)
Lesson 8	Spearman Rank Correlation
Lesson 9	Kendall's Tau and Other Rank Correlations
Lesson 10	Chi-Square Goodness-of-Fit Test
Lesson 11	Chi-Square Test of Independence
Lesson 12	Runs Test for Randomness
Lesson 13	Kolmogorov-Smirnov Tests
Lesson 14	Median Test and Other Quick Tests
Lesson 15	Non-Parametric Confidence Intervals
Lesson 16	Power, Efficiency, and Choosing the Right Test
Lesson 17	Handling Ties and Small Samples
Lesson 18	Non-Parametric Regression and Smoothing Concepts
Lesson 19	Revision: Complete Test-Selection Drill

Lesson 1: Introduction to Non-Parametric Methods

Learning Objectives

- Explain what non-parametric (distribution-free) statistics means
- Compare parametric vs non-parametric approaches
- Identify situations where non-parametric tests are required
- State advantages and disadvantages of non-parametric methods

Parametric statistical methods, such as the t-test, z-test, ANOVA and Pearson correlation, make strong assumptions about the population: most commonly that the data come from a normal distribution, that variances are equal, and that measurements are on an interval or ratio scale. Non-parametric (distribution-free) methods make no such distributional assumptions. They work with ranks, signs, or orderings of the data rather than the raw numerical values, so they remain valid even when the population distribution is unknown, skewed, or heavy-tailed.

The price of this flexibility is statistical power. When the parametric assumptions truly hold, parametric tests are more powerful — they are more likely to detect a real effect. But when assumptions fail (small samples from skewed populations, ordinal data such as Likert-scale responses, or data with extreme outliers), non-parametric tests are the safer and often the only valid choice. A good statistician always checks assumptions first and reaches for the non-parametric alternative when they are violated.

Non-parametric tests are used in four main situations: (1) the data are ordinal or nominal (ranks, preferences, categories); (2) the sample is too small to verify normality; (3) the population is clearly non-normal (e.g., income data, waiting times); (4) outliers make means unreliable, so medians and ranks are more meaningful. Most non-parametric tests are tests about the median (or about stochastic dominance) rather than the mean.

Key Terms & Definitions

- Parametric test — A test that assumes a specific population distribution (usually normal) and makes inferences about parameters like the mean.
- Non-parametric test — A distribution-free test that makes no assumption about the population distribution; typically based on ranks or signs.
- Power of a test — Probability of correctly rejecting a false null hypothesis; parametric tests have higher power when their assumptions hold.
- Robustness — The ability of a procedure to remain valid when its assumptions are mildly violated; non-parametric tests are more robust.

Lesson Summary

- Non-parametric tests are distribution-free and work on ranks/signs rather than raw values.
- Use them for ordinal data, small samples, non-normal populations, or outlier-prone data.
- They are less powerful than parametric tests when parametric assumptions hold, but safer otherwise.

Exam Pointers — Very Important

★ Define distribution-free statistics and give two situations where non-parametric tests are preferred.

★ Distinguish parametric vs non-parametric tests with one example each — a very common 3-mark question.

Lesson 2: Measurement Scales and Their Role

Learning Objectives

- Name and describe the four measurement scales
- Explain why the scale of measurement decides which test is valid
- Give examples of nominal, ordinal, interval and ratio data

Every statistical analysis begins with the measurement scale. Nominal data are labels with no order — gender, blood type, religion, yes/no answers. Only counts and proportions are meaningful; the mode is the only measure of center. Ordinal data have a meaningful order but unknown spacing — Likert scales (strongly agree … strongly disagree), class ranks, pain scores. Medians and percentiles are meaningful, but means are technically not.

Interval data have equal spacing but no true zero — temperature in Celsius, IQ scores, calendar years. Ratio data have equal spacing and a true zero — height, weight, income, time. Parametric tests require interval or ratio data; non-parametric tests can handle ordinal data and, in many cases, nominal data (e.g., the chi-square test).

This is why the choice of test always starts with two questions: what is the measurement scale, and is the population approximately normal? A t-test on ordinal Likert data is invalid; a Mann-Whitney or Wilcoxon test is the correct replacement. Examiners love to give a scenario (e.g., 'customer satisfaction rated 1–5, n=12 per group') and ask which test applies — ordinal scale plus small sample points straight to a non-parametric test.

Key Terms & Definitions

- Nominal scale — Categories without order; only equality/inequality is meaningful.
- Ordinal scale — Ordered categories with unknown intervals; medians and ranks are meaningful.
- Interval scale — Equal intervals, no true zero; addition/subtraction meaningful.
- Ratio scale — Equal intervals with a true zero; all arithmetic operations meaningful.

Lesson Summary

- Nominal: labels only. Ordinal: ordered ranks. Interval: equal steps, no true zero. Ratio: equal steps with true zero.
- Parametric tests need interval/ratio data; non-parametric tests accept ordinal (and some accept nominal).
- Always identify the scale before choosing a test.

Exam Pointers — Very Important

★ Match a scenario to the correct scale — examiners test this to see if you can pick the right test.

Lesson 3: The Sign Test — Concept and Procedure

Learning Objectives

- State the null hypothesis of the sign test
- Perform the sign test for a single median
- Use the binomial distribution to find critical values
- Apply the sign test to paired data

The sign test is the simplest non-parametric test. It tests a hypothesis about the population median. The null hypothesis is H0: median = M0 (some claimed value). The procedure: subtract M0 from each observation, record only the sign (+ or −), discard zeros (they carry no information), and count the number of plus signs. Under H0 each observation is equally likely to be above or below M0, so the number of plus signs follows a binomial distribution with n = number of non-zero signs and p = 0.5.

For small n, exact binomial probabilities give the p-value. For large n (n > 25), the normal approximation applies: $z = (x - n/2) / \sqrt{(n/4)}$, where x is the smaller of the plus/minus counts (for a two-sided test). Compare |z| with the critical value (e.g., 1.96 at $\alpha = 0.05$). The sign test is easy but wasteful — it ignores the magnitudes of the differences, so it has the lowest power among non-parametric tests. It is mainly useful as a quick check or when only the direction of difference is recorded.

The sign test also applies to paired samples (before/after studies): compute each pair's difference, take signs, and proceed exactly as above. Example: ten patients' pain scores before and after treatment give 8 minuses (improvement) and 2 pluses. With $n = 10$ and $p = 0.5$, $P(X \leq 2)$ is small, so we reject H0 and conclude the treatment reduces pain.

Key Terms & Definitions

- Sign test — Tests H0: median = M0 using only the signs of $(x_i - M_0)$; zeros are discarded.
- Binomial model — Under H0, the number of + signs ~ Binomial(n, 0.5).
- Normal approximation — For $n > 25$ use $z = (x - n/2)/\sqrt{(n/4)}$.
- Paired sign test — Sign test applied to the differences of matched pairs.

Example / Illustration

- Claim: the median delivery time is 30 minutes. Sample of 12 times: 8 are above 30, 4 below, 0 equal. $n = 12$, $x = 8$ plus signs. Two-sided test at $\alpha = 0.05$: $P(X \geq 8) + P(X \leq 4)$ under Binomial(12, 0.5) $\approx 0.388 > 0.05$, so do not reject H0 — the data do not contradict the claimed median.

Lesson Summary

- Subtract M0, keep signs, discard zeros, count plus signs.
- Test statistic ~ Binomial(n, 0.5); normal approximation for $n > 25$.
- Works for one median or for paired differences; ignores magnitudes, hence low power.

Exam Pointers — Very Important

★ Steps of the sign test and why zeros are dropped — frequently asked as a short question.

★ Compute a sign test for $n \leq 15$ using binomial tables.

Lesson 4: Wilcoxon Signed-Rank Test (One Sample / Paired)

Learning Objectives

- Explain how the signed-rank test improves on the sign test
- Rank absolute differences and attach signs
- Compute the test statistic W and apply the decision rule
- Use the normal approximation for large samples

The Wilcoxon signed-rank test is the non-parametric counterpart of the one-sample (or paired) t-test, and it is more powerful than the sign test because it uses the magnitudes of the differences through their ranks. Procedure for H0: median = M0: compute $d_i = x_i - M_0$, discard zeros, rank the absolute values $|d_i|$ from smallest (rank 1) to largest,

average the ranks of ties, then re-attach the sign of each di to its rank. The test statistic W is the sum of the positive ranks (equivalently, the smaller of the positive-rank sum and negative-rank sum for a two-sided test).

Under H0 the expected sum of positive ranks is n(n+1)/4, with variance n(n+1)(2n+1)/24. For n ≤ 30, compare W with critical values from the Wilcoxon signed-rank table; for n > 30 use the normal approximation z = (W − μW)/σW. Reject H0 when the rank sum is too small (or too large) to be plausible under H0. The test assumes the distribution of differences is symmetric about the median — a weaker assumption than normality, but still an assumption worth remembering for exams.

For paired data (before/after, matched pairs), compute di = after − before for each pair and apply the same procedure — this is the non-parametric paired t-test. Example: 8 students' scores before/after a workshop; rank the absolute gains, sum positive ranks; a large positive sum with small p-value indicates the workshop raised scores.

Key Terms & Definitions

- Signed-rank test — Non-parametric analogue of the one-sample/paired t-test; uses ranks of absolute differences with signs re-attached.
- Tied ranks — When |di| values tie, assign the average of the ranks they would occupy.
- W statistic — Sum of the positive ranks (or the smaller rank sum for two-sided tests).
- Symmetry assumption — The signed-rank test assumes the difference distribution is symmetric about the median.

Example / Illustration

- Differences: +3, −1, +5, +2, −4, +6 (n=6, no zeros). |d| ranked: 1→1, 2→2, 3→3, 4→4, 5→5, 6→6. Positive ranks: 3+5+2+6 = 16; negative: 1+4 = 5. W = min(16,5) = 5. For n=6, two-sided α=0.05 critical value is 0; since 5 > 0, do not reject H0.

Lesson Summary

- Rank |di|, attach signs, sum the positive ranks → W.
- E(W) = n(n+1)/4 under H0; use tables for n ≤ 30, normal approximation beyond.
- Paired version: apply to pair differences; assumes symmetric differences.

Exam Pointers — Very Important

★ Ranking with ties (average ranks) is the most-tested mechanical step — practice it.

★ State the symmetry assumption; examiners use it to distinguish from the sign test.

Lesson 5: Mann-Whitney U Test (Two Independent Samples)

Learning Objectives

- State when the Mann-Whitney test is used
- Combine, rank, and compute the U statistic
- Apply the normal approximation for large samples
- Interpret results as stochastic dominance / median shift

The Mann-Whitney U test (also called the Wilcoxon rank-sum test) is the non-parametric analogue of the independent two-sample t-test. It tests whether two independent samples come from populations with the same distribution — in practice, whether one population's values tend to be larger than the other's (a shift in medians). Procedure: pool both samples, rank all observations from 1 (smallest) to N (largest), average tied ranks, then sum the ranks for sample 1 (R1). The statistic is $U1 = n1 \cdot n2 + n1(n1+1)/2 - R1$, and $U2 = n1 \cdot n2 - U1$; the test statistic U is the smaller of the two.

Under H0, $E(U) = n1 \cdot n2/2$ with variance $n1 \cdot n2(n1+n2+1)/12$ (adjusted for ties if needed). For small samples use the Mann-Whitney table; for $n1, n2 > 10$ use $z = (U - \mu U)/\sigma U$ with a continuity correction of 0.5. Reject H0 when U is too small — a small U means one sample's ranks cluster at one end, i.e., the populations differ. Strictly, the test detects any distributional difference, but it is most powerful (and usually interpreted) for a location shift when the two distributions have similar shape.

Example: test scores of two teaching methods ($n1 = n2 = 8$). If method B's ranks are systematically higher, U will be small and significant, supporting B's superiority — without ever assuming normality of scores.

Key Terms & Definitions

- Mann-Whitney U test — Non-parametric two-independent-sample test based on pooled ranks; analogue of the independent t-test.
- Rank sum R1 — Sum of the pooled ranks belonging to sample 1.
- U statistic — $U1 = n1 \cdot n2 + n1(n1+1)/2 - R1$; test uses $U = \min(U1, U2)$.
- Stochastic dominance — Interpretation: $P(X > Y) \neq 0.5$, i.e., one population tends to give larger values.

Example / Illustration

- Sample A: 12, 15, 9 ($n1=3$); Sample B: 20, 18, 25, 22 ($n2=4$). Pooled ranks: 9→1, 12→2, 15→3, 18→4, 20→5, 22→6, 25→7. $R1 = 1+2+3 = 6$. $U1 = 3 \cdot 4 + 3 \cdot 4/2 - 6 = 12+6-6 = 12$; $U2 = 12-12 = 0$. $U = 0$ → significant at $\alpha=0.05$ (critical value 0): B scores are stochastically larger.

Lesson Summary

- Pool both samples, rank, sum ranks of sample 1 → compute U1, U2; test U = min(U1, U2).
- Small U ⇒ samples differ; use normal approximation when n1, n2 > 10.
- Non-parametric replacement for the independent two-sample t-test.

Exam Pointers — Very Important

★ Full Mann-Whitney computation for small samples is a classic 5–10 mark question — master the ranking and the U1 formula.

Lesson 6: Kruskal-Wallis Test (Three or More Independent Samples)

Learning Objectives

- Recognize Kruskal-Wallis as the non-parametric ANOVA
- Compute the H statistic from group rank sums
- Use the chi-square approximation with k−1 degrees of freedom
- Know when post-hoc pairwise tests are needed

The Kruskal-Wallis test extends the Mann-Whitney idea to three or more independent groups — it is the non-parametric one-way ANOVA. H0: all k populations have the same distribution (equal medians, in the location-shift interpretation). Procedure: pool all N observations, rank them 1 to N (average ties), compute each group's rank sum Ri, then $H = [12/(N(N+1))] \cdot \Sigma(Ri^2/ni) - 3(N+1)$. Under H0, H is approximately chi-square with (k − 1) degrees of freedom, provided each ni ≥ 5.

Reject H0 when H exceeds the χ^2 critical value — a large H means the group rank sums differ more than chance allows. Like ANOVA, Kruskal-Wallis is an omnibus test: significance only tells you that at least one group differs, not which. Follow up with pairwise Mann-Whitney tests (with a Bonferroni correction) to locate the differences. The tie correction divides H by $1 - \Sigma(t^3-t)/(N^3-N)$, where t is each tie-group size; it matters only with many ties.

Example: crop yields under 3 fertilizers (n = 6 each). Rank all 18 yields; if one fertilizer's rank sum is much higher, H becomes large and significant — evidence that fertilizer affects yield, with no normality assumption on yields.

Key Terms & Definitions

- Kruskal-Wallis test — Non-parametric one-way ANOVA for k ≥ 3 independent samples, based on pooled ranks.
- H statistic — $H = [12/(N(N+1))]\Sigma(Ri^2/ni) - 3(N+1)$; approx. χ^2 with k−1 df.
- Omnibus test — A test that detects any difference among groups without identifying which groups differ.
- Tie correction — Divides H by $1 - \Sigma(t^3-t)/(N^3-N)$; used when many tied observations exist.

Example / Illustration

- Three groups, n=4 each. Rank sums: R1=10, R2=24, R3=44, N=12. $H = [12/(12\cdot13)]\cdot(100/4+576/4+1936/4) - 39 = (12/156)\cdot653 - 39 \approx 50.23 - 39 = 11.23$. $\chi^2(2)$ critical at 0.05 is 5.99; 11.23 > 5.99 → reject H0: the groups differ.

Lesson Summary

- Rank all N observations, get group rank sums Ri, compute H.
- Compare H with $\chi^2(k-1)$; needs ni ≥ 5 per group for the approximation.
- Significant H ⇒ follow up with pairwise tests to find which groups differ.

Exam Pointers — Very Important

★ The H formula and its $\chi^2(k-1)$ distribution are must-know; practice one full numerical.

Lesson 7: Friedman Test (Randomized Block / Repeated Measures)

Learning Objectives

- Identify the Friedman test as the non-parametric repeated-measures ANOVA
- Rank within blocks and compute the Fr statistic
- Apply the chi-square approximation
- Contrast Friedman with Kruskal-Wallis

The Friedman test is the non-parametric analogue of the repeated-measures (randomized block) ANOVA: k treatments are compared across n blocks (subjects), where each block receives every treatment. Because observations within a block are dependent, we rank within each block separately (ranks 1 to k per block), then sum ranks across blocks for each treatment (Rj). H0: the treatments have identical effects. The statistic is $Fr = [12/(n\cdot k(k+1))] \cdot \Sigma Rj^2 - 3n(k+1)$, approximately chi-square with (k−1) df.

Reject H0 for large Fr. The key distinction from Kruskal-Wallis: Friedman ranks within blocks (handling the dependence of repeated measures on the same subject), while

Kruskal-Wallis pools independent samples. Example: 10 tasters rate 4 biscuit brands — each taster is a block; rank the 4 brands within each taster, sum brand ranks across tasters, compute Fr. A significant result means at least one brand is rated differently.

Assumptions are light: blocks are independent, and within each block the k observations are ranked (ordinal data suffice). No normality, no sphericity assumption as in parametric repeated-measures ANOVA.

Key Terms & Definitions

- Friedman test — Non-parametric repeated-measures ANOVA; ranks treatments within each block.
- Block — A subject or matched set receiving all k treatments; controls for subject-to-subject variation.
- Fr statistic — $[12/(nk(k+1))]\Sigma Rj^2 - 3n(k+1)$, approx. $\chi^2(k-1)$.
- Within-block ranking — Each block's k observations get ranks 1..k, removing between-block differences.

Example / Illustration

- 4 judges (blocks) rank 3 designs. Rank sums: R1=5, R2=9, R3=10 (n=4, k=3). $Fr = [12/(4\cdot3\cdot4)]\cdot(25+81+100) - 3\cdot4\cdot4 = (12/48)\cdot206 - 48 = 51.5 - 48 = 3.5$. $\chi^2(2)$ critical = 5.99 → do not reject: no evidence the designs differ.

Lesson Summary

- Rank treatments 1..k inside each block; sum ranks per treatment → Rj.
- Fr ≈ $\chi^2(k-1)$; large Fr ⇒ treatments differ.
- Use for repeated measures / matched blocks; the dependent-sample cousin of Kruskal-Wallis.

Exam Pointers — Very Important

★ Friedman vs Kruskal-Wallis: independent samples → Kruskal-Wallis; same subjects/blocks → Friedman. This distinction is heavily tested.

Lesson 8: Spearman Rank Correlation

Learning Objectives

- Compute Spearman's rho from ranked data
- Test the significance of a rank correlation
- Explain when Spearman is preferred over Pearson
- Interpret the strength and direction of rho

Spearman's rank correlation coefficient (rho, ρs) measures the strength of a monotonic relationship between two variables. Rank each variable separately (1 to n, averaging ties), compute the differences di between the paired ranks, then $\rho s = 1 - 6\Sigma di^2$ /

$(n(n^2-1))$. It ranges from -1 (perfect decreasing monotonic) to $+1$ (perfect increasing monotonic); 0 means no monotonic association. Unlike Pearson's r, Spearman needs no linearity and no normality — only that the relationship be monotonic, and it works on ordinal data.

Spearman is the right choice when: the relationship looks curved but monotonic (e.g., diminishing returns), the data are ordinal (ranks, ratings), outliers would distort Pearson, or normality fails. Significance is tested via $t = \rho_s\sqrt{((n-2)/(1-\rho_s^2))}$ with $n-2$ df (for $n > 10$), or exact tables for small n. H0: $\rho_s = 0$ (no monotonic association).

Example: 8 students ranked by study hours and by exam position. If higher study ranks consistently pair with better exam ranks, Σd_i^2 is small and ρ_s is near $+1$ — strong evidence that more study associates with better performance, without assuming any linear form.

Key Terms & Definitions

- Spearman's rho — $\rho_s = 1 - 6\Sigma d_i^2/(n(n^2-1))$; measures monotonic association between ranked variables.
- Monotonic relationship — As one variable increases, the other consistently increases (or decreases), not necessarily linearly.
- Tied ranks — Average the ranks for equal values before computing di.
- Significance test — $t = \rho_s\sqrt{((n-2)/(1-\rho_s^2))} \sim t(n-2)$ for $n > 10$.

Example / Illustration

- $n=6$, $\Sigma d_i^2=4$. $\rho_s = 1 - 6\cdot4/(6\cdot35) = 1 - 24/210 = 0.886$. $t = 0.886\cdot\sqrt{(4/(1-0.785))} = 0.886\cdot\sqrt{(18.6)} \approx 3.82$; t(4) critical (0.05, two-sided) = 2.776 → significant positive monotonic association.

Lesson Summary

- Rank both variables, take rank differences di, apply the rho formula.
- Captures any monotonic (not just linear) relationship; robust to outliers.
- Test with $t(n-2)$; the non-parametric alternative to Pearson's r.

Exam Pointers — Very Important

★ Compute rho for a small dataset — a standard numerical; remember to average tied ranks.

Lesson 9: Kendall's Tau and Other Rank Correlations

Learning Objectives

- Define concordant and discordant pairs
- Compute Kendall's tau
- Compare tau with Spearman's rho
- Know when tau is preferred

Kendall's tau (τ) is another rank correlation, based on a beautifully simple idea: look at every pair of observations and ask whether they are concordant (both variables rank in the same order) or discordant (opposite order). With C concordant and D discordant pairs out of n(n−1)/2 total, τ = (C − D) / (n(n−1)/2). Like Spearman's rho it lies in [−1, 1] and tests H0: no association, but tau has a direct probabilistic interpretation: it is the difference between the probability of concordance and discordance.

Tau is generally preferred over rho when: the sample is small, there are many tied ranks (tau-b adjusts for ties), or you want an interpretable effect size — τ = 0.6 means concordant pairs exceed discordant by 60% of all pairs. Spearman's rho is more widely reported and slightly easier to compute by hand, and the two usually agree in direction and rough magnitude (|τ| is typically smaller than |ρs| for the same data). For significance, both use normal approximations for larger n.

Example: ranking 5 job candidates by interview score and by test score — count concordant vs discordant pairs among the 10 pairs; τ = (C−D)/10. This pair-counting logic also underlies the Mann-Whitney U test, which is mathematically equivalent to a tau-type comparison between two samples.

Key Terms & Definitions

- Concordant pair — A pair where the two variables order the same way (both higher or both lower).
- Discordant pair — A pair where the variables order oppositely.
- Kendall's tau — τ = (C−D)/(n(n−1)/2); probability difference of concordance vs discordance.
- Tau-b — Tie-adjusted version of tau for data with many tied ranks.

Example / Illustration

- 5 pairs ranked: (1,2),(2,1),(3,3),(4,5),(5,4). Pairs: total 10. Concordant: (1,2)-(3,3)? order (1<3, 2<3) yes... counting carefully gives C=7, D=3. τ = (7−3)/10 = 0.4 — moderate positive association.

Lesson Summary

- τ = (C − D)/total pairs; interpretable as P(concordance) − P(discordance).
- Better than rho for small n and tied data (tau-b); |τ| usually < |ρs|.
- Shares its pair logic with the Mann-Whitney test.

Exam Pointers — Very Important

★ Define concordant/discordant pairs and compute tau for n=5 — quick 5-mark question.

Lesson 10: Chi-Square Goodness-of-Fit Test

Learning Objectives

- State the purpose of the goodness-of-fit test
- Compute expected frequencies under H0
- Calculate the χ^2 statistic and degrees of freedom
- Check the test's validity conditions

The chi-square goodness-of-fit test checks whether observed categorical frequencies match a hypothesized distribution. H0 specifies the expected proportions (e.g., a die is fair: each face 1/6; or Mendel's 9:3:3:1 ratio). Compute expected counts Ei = n·pi, then $\chi^2 = \Sigma(Oi - Ei)^2/Ei$, which under H0 follows a chi-square distribution with (k − 1) df, where k is the number of categories. Reject H0 when χ^2 is too large — large discrepancies between observed and expected are implausible under H0.

Validity conditions: observations must be independent counts, categories mutually exclusive and exhaustive, and expected frequencies should be at least 5 in most cells (combine sparse categories if needed). The test is always right-tailed because only large χ^2 values contradict H0. Example: 60 die rolls give faces 8,12,9,11,7,13; Ei = 10 each; $\chi^2 = (4+4+1+1+9+9)/10 = 2.8$; $\chi^2(5)$ critical = 11.07 → do not reject: no evidence the die is biased.

The same machinery tests any fully specified distribution — uniform, binomial with given p, Poisson with given λ — making it the most versatile non-parametric test for categorical data.

Key Terms & Definitions

- Goodness-of-fit — Tests whether observed frequencies match hypothesized probabilities.
- Expected frequency — Ei = n·pi; must be ≥ 5 in most cells for the χ^2 approximation.
- χ^2 statistic — $\Sigma(Oi-Ei)^2/Ei \sim \chi^2(k-1)$ under H0.
- Right-tailed test — Only large χ^2 values reject H0; there is no 'too good' rejection region in the basic test.

Example / Illustration

- Mendel's peas: observed 315 round-yellow, 108 round-green, 101 wrinkled-yellow, 32 wrinkled-green (n=556). Expected 9:3:3:1 → 312.75, 104.25, 104.25, 34.75. $\chi^2 \approx 0.47$; $\chi^2(3)$ critical = 7.81 → do not reject: data consistent with 9:3:3:1.

Lesson Summary

- Ei = n·pi; $\chi^2 = \Sigma(Oi-Ei)^2/Ei$; df = k−1; right-tailed.
- Needs independent counts and Ei ≥ 5 (combine categories otherwise).
- Tests any hypothesized categorical distribution.

Exam Pointers — Very Important

★ A goodness-of-fit numerical with combining of low-Ei categories is a favourite long question.

Lesson 11: Chi-Square Test of Independence

Learning Objectives

- Set up a contingency table
- Compute expected frequencies under independence
- Perform the χ^2 test with (r−1)(c−1) df
- Interpret association between two categorical variables

The chi-square test of independence asks whether two categorical variables are associated. Data form an r × c contingency table of observed counts Oij. Under H0 (independence), expected counts are Eij = (row total × column total)/grand total. The statistic $\chi^2 = \Sigma\Sigma(Oij - Eij)^2/Eij$ follows χ^2 with (r−1)(c−1) df. A large χ^2 rejects independence — the variables are associated.

Example: 200 voters classified by gender (rows) and party preference (columns). If men and women prefer parties differently, observed counts deviate from Eij and χ^2 is significant. Important: significance means association, not causation, and not strength — for strength use Cramér's $V = \sqrt{(\chi^2/(n\cdot\min(r-1,c-1)))}$. Validity needs the same conditions as goodness-of-fit; for 2×2 tables with small counts, Fisher's exact test is the exact alternative, and Yates' continuity correction is sometimes applied.

The test of homogeneity is mathematically identical but samples from separate populations (e.g., defect rates across 3 factories) — H0: the distributions are the same across populations. Examiners often ask you to distinguish independence (one sample, two variables) from homogeneity (multiple samples, one variable).

Key Terms & Definitions

- Contingency table — r × c table of observed counts for two categorical variables.
- Expected under independence — Eij = (row i total × column j total)/N.
- Degrees of freedom — (r−1)(c−1).
- Cramér's V — Effect-size measure of association strength: $\sqrt{(\chi^2/(n \cdot \min(r-1,c-1)))}$.

Example / Illustration

- 2×2 table: 40/60 and 30/70. Row totals 100,100; col totals 70,130; N=200. E11 = 100·70/200 = 35, E12=65, E21=35, E22=65. χ^2 = 25/35+25/65+25/35+25/65 ≈ 2.20; $\chi^2(1)$ critical = 3.84 → do not reject: no significant association.

Lesson Summary

- Eij = row·col/N; $\chi^2 = \Sigma\Sigma(Oij-Eij)^2/Eij$; df = (r−1)(c−1).
- Significant χ^2 ⇒ variables are associated (not independent).
- Independence = one sample/two variables; homogeneity = several samples/one variable.

Exam Pointers — Very Important

★ Contingency-table numericals are guaranteed exam material — practice the Eij computation until automatic.

Lesson 12: Runs Test for Randomness

Learning Objectives

- Explain what a 'run' is
- Test whether a sequence is random
- Apply the normal approximation for the number of runs
- Use the runs test above/below the median

The runs test (Wald-Wolfowitz) checks whether a sequence of observations is random. A run is a maximal consecutive subsequence of identical symbols — e.g., in HHHTTHT, the runs are HHH, TT, H, T: 4 runs. Too few runs suggests clustering/trend; too many suggests oscillation/alternation. With n1 symbols of type 1 and n2 of type 2, the number of runs R under randomness has mean $\mu R = 2n1n2/(n1+n2) + 1$ and variance $\sigma^2 R = 2n1n2(2n1n2-n1-n2)/((n1+n2)^2(n1+n2-1))$.

For n1, n2 > 10, $z = (R - \mu R)/\sigma R$ (with continuity correction) tests H0: the sequence is random, two-sided. Example: 20 coin tosses with only 6 runs — far below μR ≈ 11 — indicates non-randomness (clustering). The test also applies to numerical data: label

each observation + if above the median and − if below (drop equals), then test the resulting +/− sequence for randomness — a check for trends or cycles over time, e.g., in quality-control measurements.

The two-sample Wald-Wolfowitz runs test extends the idea: merge two samples, label by sample, and count runs — few runs indicate the samples come from different distributions.

Key Terms & Definitions

- Run — A maximal block of identical consecutive symbols in a sequence.
- μR, $\sigma^2 R$ — Mean and variance of the number of runs under randomness.
- Randomness test — Two-sided: too few runs ⇒ clustering/trend; too many ⇒ systematic alternation.
- Median runs test — Convert data to +/− about the median, then test the symbol sequence.

Example / Illustration

- Sequence: +++−−−+++−− (n1=6 plus, n2=4 minus). Runs: +++, −−−, +++, −− → R=4. $\mu R = 2 \cdot 6 \cdot 4/10 + 1 = 5.8$. R=4 is close to 5.8 → no evidence against randomness.

Lesson Summary

- Count runs R; compare with $\mu R = 2n1n2/(n1+n2)+1$ via z for large samples.
- Too few runs = clustering; too many = alternation.
- Numerical data: dichotomize about the median first.

Exam Pointers — Very Important

★ Compute R and μR for a short sequence — quick marks; state both one-sided interpretations.

Lesson 13: Kolmogorov-Smirnov Tests

Learning Objectives

- Explain the K-S one-sample test using the empirical CDF
- Compute the D statistic
- Describe the two-sample K-S test
- Compare K-S with chi-square goodness-of-fit

The Kolmogorov-Smirnov one-sample test compares a sample with a fully specified continuous distribution. Build the empirical CDF $Fn(x)$ = (number of observations ≤ x)/n — a step function — and measure its maximum vertical distance from the hypothesized CDF $F0(x)$: $D = \max|Fn(x) - F0(x)|$. Under H0 (data follow F0), D is small; large D rejects H0. Critical values come from the K-S table (depending on n and α). Unlike chi-square,

K-S needs no binning, works for small samples, and is more powerful for continuous data.

Example: test whether 10 service times follow an exponential distribution with $\lambda = 0.2$. Compute F0 at each observation, Fn as i/n, find the largest gap D; if D exceeds the critical value (0.409 for n=10, α=0.05), reject. The two-sample K-S test generalizes this: $D = \max|F1(x) - F2(x)|$ over the two empirical CDFs, testing whether two samples come from the same continuous distribution — a non-parametric alternative to the t-test that detects any distributional difference (location, spread, or shape).

K-S is sensitive near the center of the distribution; the Anderson-Darling test weights the tails more heavily. For exams, remember: K-S = maximum CDF gap, no grouping needed, exact for continuous distributions.

Key Terms & Definitions

- Empirical CDF — Fn(x) = #{xi ≤ x}/n; the sample's step-function CDF.
- D statistic — max|Fn(x) − F0(x)|; the largest vertical gap between empirical and hypothesized CDFs.
- Two-sample K-S — max|F1(x) − F2(x)|; tests equality of two continuous distributions.
- No binning — Advantage over χ^2: uses every observation's exact position.

Example / Illustration

- n=5 data: 1.2, 2.1, 2.8, 3.5, 4.4 vs Uniform(0,5): F0 = x/5. At x=2.1: Fn=0.4, F0=0.42, gap 0.02; largest gap ≈ 0.1 < critical 0.565 → do not reject uniformity.

Lesson Summary

- D = maximum |empirical CDF − hypothesized CDF|; large D rejects H0.
- One-sample: test against a specified distribution; two-sample: compare two samples.
- More powerful than χ^2 for continuous data; no arbitrary binning.

Exam Pointers — Very Important

★ State the D formula and explain why K-S beats chi-square for continuous data — theory favourite.

Lesson 14: Median Test and Other Quick Tests

Learning Objectives

- Perform the median test for k independent samples
- Describe Mood's median test procedure
- Know the control median and quantile tests
- Choose among quick non-parametric tests

The median test (Mood's median test) is the simplest k-sample test: find the grand median of all pooled observations, classify each observation as above or below it, and form a k×2 contingency table of counts. Under H0 (all populations have the same median), apply the chi-square test with (k−1) df. It is quick and needs only the grand median, but it is less powerful than Kruskal-Wallis because it discards all ranking information beyond the median split.

Related quick tests include the control median test (compare each treatment against a control using the control's median as the split point) and quantile tests (split at quartiles or other quantiles instead of the median, detecting differences in spread as well as location). These tests shine when data collection only recorded whether each observation exceeded a threshold — common in field studies and quality control.

Choose: two independent samples → Mann-Whitney (powerful) or median test (quick); k independent samples → Kruskal-Wallis (powerful) or median test (quick); paired data → Wilcoxon signed-rank; repeated measures → Friedman. The 'quick' tests trade power for simplicity and minimal assumptions.

Key Terms & Definitions

- Median test — k×2 table of above/below grand-median counts; $\chi^2(k-1)$.
- Control median test — Uses the control group's median as the classification threshold.
- Quantile test — Splits at quartiles/deciles to detect spread differences too.
- Power trade-off — Quick tests are simpler but less powerful than rank-based tests.

Example / Illustration

- Two machines' outputs; grand median = 50. Machine A: 8 above, 2 below; B: 3 above, 7 below. $\chi^2(1) = \Sigma(O-E)^2/E$ with E=5.5,4.5: $\approx 5.05 > 3.84$ → medians differ.

Lesson Summary

- Grand median → above/below counts → $\chi^2(k-1)$.
- Fast but low-power; Kruskal-Wallis/Mann-Whitney preferred when full data exist.
- Quantile variants can detect differences in spread.

Exam Pointers — Very Important

★ When would you use the median test instead of Kruskal-Wallis? (Answer: only threshold data available / need a quick robust check.)

Lesson 15: Non-Parametric Confidence Intervals

Learning Objectives

- Construct a CI for the median using order statistics
- Explain the binomial basis of the interval
- Construct a CI for the difference of two medians
- Interpret distribution-free intervals

Confidence intervals need not be parametric. A distribution-free CI for the population median uses order statistics: sort the sample $x(1) \le \dots \le x(n)$; then $(x(L), x(U))$ is a $(1-\alpha)$ CI for the median, where L and U are chosen so that $P(\text{Binomial}(n,0.5) \le L-1) \approx \alpha/2$ and symmetrically for U. For large n, $L \approx (n - z\cdot\sqrt{n})/2$ and $U \approx (n + z\cdot\sqrt{n})/2 + 1$. The interval is exact in coverage (conservatively) and valid for any continuous distribution.

Example: $n = 15$, $\alpha = 0.05$. Binomial(15, 0.5): $P(X \le 3) = 0.018$, so take $L = 4$, $U = 12$: the 4th and 12th order statistics bracket the median with ≥95% confidence. The method generalizes: a CI for the difference of two medians uses the Walsh averages or the Hodges-Lehmann estimator (median of all pairwise differences), with critical values from the Wilcoxon/Mann-Whitney tables.

These intervals inherit the robustness of their tests — they remain valid under skewness and outliers where the t-interval fails. The Hodges-Lehmann estimator (median of pairwise averages/differences) is the natural point estimator paired with signed-rank/Mann-Whitney intervals.

Key Terms & Definitions

- Order statistics — Sorted sample $x(1)\dots x(n)$; the CI endpoints are selected order statistics.
- Binomial selection — L, U chosen from Binomial(n, 0.5) tail probabilities.
- Hodges-Lehmann estimator — Median of pairwise differences (or averages); the location estimator behind rank CIs.
- Distribution-free CI — Valid coverage without distributional assumptions.

Example / Illustration

- n=15 sorted data; L=4, U=12 → 95% CI for median = (x(4), x(12)) = (38, 61). No normality assumed.

Lesson Summary

- Median CI = $(x(L), x(U))$ with L, U from Binomial(n, 0.5) tails.
- Large-n shortcut: $L \approx (n - z\sqrt{n})/2$.
- Hodges-Lehmann estimator pairs naturally with these intervals.

Exam Pointers — Very Important

★ Explain the order-statistic CI for the median — a strong theory question showing real understanding.

Lesson 16: Power, Efficiency, and Choosing the Right Test

Learning Objectives

- Define asymptotic relative efficiency (ARE)
- Quote the ARE of common non-parametric tests
- Build a decision flowchart for test selection
- Explain why non-parametric tests are 'safe' choices

Asymptotic relative efficiency (ARE) compares tests by the sample-size ratio needed for equal power as $n \to \infty$. The celebrated results: the Wilcoxon signed-rank test has ARE 0.955 vs the t-test under normality — it loses almost nothing when normality holds — and its ARE can be arbitrarily large under heavy-tailed distributions. The sign test's ARE is only 0.637. The Mann-Whitney test similarly has ARE 0.955 vs the two-sample t-test under normality. In other words, rank tests sacrifice at most ~5% efficiency in the best case for parametric tests, while gaining enormous robustness in the worst case.

This is the practical takeaway: when in doubt, the non-parametric test is the safe choice — you lose little if assumptions hold and avoid catastrophe if they don't. The decision flowchart: (1) one sample? → sign/signed-rank (median) or K-S/χ^2 (distribution); (2) two independent samples? → Mann-Whitney; (3) two paired samples? → signed-rank; (4) k independent? → Kruskal-Wallis; (5) k repeated/blocked? → Friedman; (6) association of two variables? → Spearman/Kendall (ordinal/continuous) or χ^2 (categorical); (7) randomness? → runs test.

Also remember the scale rule: nominal → χ^2 tests; ordinal → rank tests; interval/ratio + normal → parametric; interval/ratio + non-normal → rank tests. Walking through this logic explicitly in an exam answer earns method marks even if the arithmetic slips.

Key Terms & Definitions

- Asymptotic relative efficiency — ARE; sample-size ratio for equal power as $n \to \infty$.
- 0.955 rule — Wilcoxon/Mann-Whitney lose only ~5% efficiency vs t-tests under normality.
- Decision flowchart — Scale → design → sample size → test; show your reasoning in exams.
- Safe choice principle — Rank tests cost little under normality and protect against assumption failure.

Lesson Summary

- Wilcoxon/Mann-Whitney ARE ≈ 0.955 vs t-tests under normality; far better under heavy tails.
- Follow the flowchart: design + scale determine the test.
- When unsure, choose the non-parametric test — the efficiency cost is tiny.

Exam Pointers — Very Important

★ 'ARE of the signed-rank test is 0.955' — interpret this statement. A guaranteed theory question.

Lesson 17: Handling Ties and Small Samples

Learning Objectives

- Apply the average-rank rule for ties
- State tie corrections for H, Fr, and variance formulas
- Use exact tables for small samples
- Apply continuity corrections in normal approximations

Ties (equal observations) need care in every rank test. The universal rule: assign each tied observation the average of the ranks they would have occupied. Example: values 12, 15, 15, 18 get ranks 1, 2.5, 2.5, 4. This preserves the total rank sum. With few ties, corrections are negligible; with many ties, adjust: Kruskal-Wallis H is divided by $1 - \Sigma(t^3-t)/(N^3-N)$; the Mann-Whitney variance loses $\Sigma(t^3-t)/12$ terms; Spearman's formula needs the tie-corrected version (or compute Pearson's r on the averaged ranks — the safest general method).

For small samples, never use the normal approximation — use exact tables: Wilcoxon signed-rank tables ($n \leq 30$), Mann-Whitney U tables, Spearman critical values, K-S tables, binomial tables for the sign test. The normal approximation generally needs $n > 25$–30 (or $n1, n2 > 10$ for Mann-Whitney). When using the approximation, apply the continuity correction: subtract 0.5 from |observed − expected| before dividing by σ — this noticeably improves accuracy near the boundary.

Exam tip: if a question gives tied data and asks for Spearman's rho, the expected method is Pearson's r computed on the (averaged) ranks — state this explicitly to show you know the shortcut formula fails with ties.

Key Terms & Definitions

- Average-rank rule — Tied observations share the mean of their would-be ranks.
- Tie correction — Adjusts H, variances, and rho when many ties exist.
- Exact tables — Required for small n; normal approximations need n ≳ 25–30.
- Continuity correction — Subtract 0.5 from |O−E| in normal approximations of discrete statistics.

Example / Illustration

- Data 5, 7, 7, 7, 9 → ranks: 1, 3, 3, 3, 5 (the three 7s average ranks 2,3,4). Check: 1+3+3+3+5 = 15 = 5·6/2 ✓.

Lesson Summary

- Average tied ranks; verify the rank sum still equals n(n+1)/2.
- Use exact tables for small n; continuity-correct the normal approximation.
- With ties, compute Spearman as Pearson's r on averaged ranks.

Exam Pointers — Very Important

★ Tied-rank assignment with the sum check is a common 2–3 mark part-question.

Lesson 18: Non-Parametric Regression and Smoothing Concepts

Learning Objectives

- Describe the idea of non-parametric regression
- Explain the Theil-Sen median slope estimator
- Contrast with ordinary least squares
- Know where rank methods appear in modern statistics

Non-parametric ideas extend to regression. The Theil-Sen estimator fits a line by computing the slope between every pair of points and taking the median of those slopes — a robust alternative to least squares that resists up to ~29% outliers without distortion. The intercept is then the median of (yi − slope·xi). It is the regression analogue of the rank tests: no normality, no equal variance, just robustness. Confidence intervals for the slope come from Kendall's tau distribution, neatly tying the course together.

More broadly, non-parametric regression (kernel smoothing, LOESS, splines) fits curves without assuming a functional form — the data determine the shape. While STA644 focuses on classical rank tests, knowing that 'non-parametric' also means 'no assumed model form' completes the picture: distribution-free inference and model-free fitting are

two faces of the same philosophy.

In modern practice, rank-based methods remain essential: A/B testing with skewed metrics (use Mann-Whitney), clinical trials with ordinal endpoints (use Wilcoxon), genomics and finance with heavy tails (use rank correlations). The course's tests are not museum pieces — they are daily tools wherever assumptions are doubtful.

Key Terms & Definitions

- Theil-Sen estimator — Median of all pairwise slopes; robust line fit (≈29% breakdown point).
- Breakdown point — Fraction of bad data an estimator tolerates; 29% for Theil-Sen vs 0% for least squares.
- Non-parametric regression — Curve fitting without an assumed functional form (kernels, LOESS, splines).
- Robustness in practice — Rank methods dominate in A/B testing, clinical ordinal endpoints, heavy-tailed data.

Example / Illustration

- Points (1,2),(2,3),(3,9),(4,5): pairwise slopes include 1, 3.5, 1, ...; median slope ≈ 1.2 vs least-squares slope pulled up by (3,9). Theil-Sen resists the outlier.

Lesson Summary

- Theil-Sen: median pairwise slope — robust regression without assumptions.
- Non-parametric = distribution-free inference + model-free fitting.
- Rank tests are everyday tools in tech, medicine, and finance.

Exam Pointers — Very Important

★ Describe the Theil-Sen estimator and its breakdown advantage — a strong final-lesson question.

Lesson 19: Revision: Complete Test-Selection Drill

Learning Objectives

- Match 15 exam-style scenarios to the correct test
- State each test's H0 and statistic from memory
- Avoid the classic selection traps
- Plan time for numerical questions

Final drill — for each scenario name the test: (1) median battery life vs claimed 500h, n=14 → sign test or signed-rank; (2) before/after training scores, skewed → Wilcoxon signed-rank; (3) two teaching methods, ordinal ratings → Mann-Whitney; (4) 4 fertilizers, independent plots → Kruskal-Wallis; (5) 12 patients × 3 drugs (each patient tries all) → Friedman; (6) study hours vs GPA, curved relationship → Spearman; (7) gender vs voting

preference → χ^2 independence; (8) die fairness, 120 rolls → χ^2 goodness-of-fit; (9) coin-toss sequence random? → runs test; (10) service times exponential? → K-S; (11) 3 machines vs grand median → median test; (12) CI for median, n=20 → order-statistic interval; (13) interview vs test ranks, n=6 → Kendall's tau; (14) two samples same distribution? → two-sample K-S or Mann-Whitney; (15) trend in monthly defects → median runs test.

Classic traps: paired vs independent (read the design!); Friedman vs Kruskal-Wallis (blocks vs independent groups); goodness-of-fit vs independence (one variable vs two); Spearman vs Pearson (monotonic vs linear); sign vs signed-rank (direction only vs magnitudes). Time plan for numericals: ranking 2 min, statistic 3 min, critical value + conclusion 2 min — and always state H0, the test, and the conclusion in words.

Memorize this spine: sign→Binomial; signed-rank→W, $E=n(n+1)/4$; Mann-Whitney→$U1=n1n2+n1(n1+1)/2-R1$; Kruskal-Wallis→$H\approx\chi^2(k-1)$; Friedman→$Fr\approx\chi^2(k-1)$; Spearman→$1-6\Sigma d^2/(n(n^2-1))$; $\chi^2\rightarrow\Sigma(O-E)^2/E$; runs→$\mu R=2n1n2/(n1+n2)+1$; K-S→$D=\max|Fn-F0|$.

Key Terms & Definitions

- Scenario matching — The #1 exam skill: design + scale → test.
- Trap pairs — Paired/independent, Friedman/Kruskal-Wallis, fit/independence, Spearman/Pearson, sign/signed-rank.
- Formula spine — The nine formulas above — memorize cold.
- Conclusion in words — Always translate 'reject H0' into a plain-English finding.

Lesson Summary

- Drill the 15 scenarios until test selection is instant.
- Dodge the five classic trap pairs.
- Memorize the nine-formula spine; budget ~7 minutes per numerical.

Exam Pointers — Very Important

★ A full scenario-based paper is exactly this drill — if you can do these 15, you can do the exam.