Parametric and Non-Parametric Tests: Comprehensive Tutorial

Introduction

Statistical tests are divided into two major categories based on whether they assume specific underlying probability distributions. Understanding when to use each type is crucial for valid statistical inference.

Parametric Tests

What Are Parametric Tests?

Parametric tests make assumptions about the population parameters and the underlying distribution of the data. They typically assume that the data comes from a normal (bell-shaped) distribution and use the actual values of the data points in calculations.

Key Characteristics of Parametric Tests

- **Distribution assumption**: Data are normally distributed
- **Data type**: Work with interval or ratio data
- **Information used**: Use actual data values
- **Statistical power**: Generally have greater statistical power (ability to detect true effects)
- **Parameter focus**: Make inferences about population parameters (means, variances)
- **Variance assumption**: Often assume equal variances across groups (homogeneity of variance)

Assumptions for Parametric Tests

Before using parametric tests, verify these assumptions:

1. Normality: The data should be approximately normally distributed. Check using Q-Q plots, histograms, or formal tests like Shapiro-Wilk test (H_0: data is normal; if $p > 0.05$, data is approximately normal).

2. Homogeneity of Variance: Variance should be similar across groups. Test using Levene's test or Bartlett's test (H_0: variances are equal; if $p > 0.05$, assumption is met).

3. Independence: Observations should be independent of each other. This is ensured through proper study design rather than statistical testing.

4. Scale of Measurement: Data should be interval or ratio level (not categorical or ordinal).

Assumption Testing: Detailed Guide

Before conducting parametric tests, you must formally test assumptions. Violating assumptions compromises the validity of your results. This section covers the most important assumption tests.

Test for Normality

Shapiro-Wilk Test

Purpose: Tests whether a sample comes from a normally distributed population.

Hypotheses:

- H_0: Data are normally distributed
- H_1: Data are not normally distributed

Interpretation:

- If $p > 0.05$: Fail to reject H_0 → Data are approximately normal ✓
- If $p \leq 0.05$: Reject H_0 → Data significantly deviate from normality ✗

Important Notes:

- With large samples ($n > 50$), even small deviations from normality become statistically significant
- With small samples ($n < 30$), test has low power to detect non-normality
- Consider both statistical test AND visual inspection

When to use: Best for small to moderate sample sizes ($n < 5000$).

Kolmogorov-Smirnov Test

Purpose: Tests if data follow a specified distribution (usually normal).

Hypotheses:

- H_0: Data follow the specified distribution (usually normal)
- H_1: Data do not follow the specified distribution

Interpretation: Same as Shapiro-Wilk ($p > 0.05$ = normal).

When to use: Good for moderate sample sizes; less powerful than Shapiro-Wilk for normality testing.

Anderson-Darling Test

Purpose: Tests goodness of fit to a distribution, particularly sensitive at the tails.

Hypotheses:

- H_0: Data follow the specified distribution
- H_1: Data do not follow the specified distribution

Advantages over Shapiro-Wilk: More sensitive to deviations in the tails of the distribution.

When to use: When extreme values are important (financial data, safety data).

Visual Methods for Assessing Normality

Histogram: Plot frequency distribution of data.

- Normal data shows roughly bell-shaped, symmetric curve
- Skewed data shows asymmetry (tail on one side)
- Multimodal data shows multiple peaks

Q-Q Plot (Quantile-Quantile Plot): Plots data quantiles against theoretical normal quantiles.

- Normal data: Points roughly follow diagonal line
- Deviation from line indicates non-normality
- S-shaped pattern suggests heavy tails
- Curved pattern suggests skewness

Box Plot: Shows distribution shape and outliers.

- Symmetric box = normally distributed
- Unequal whiskers = skewed data
- Isolated points beyond whiskers = potential outliers

Detailed Example: Testing Normality of Test Scores

Scenario: You have test scores from 50 students: 65, 72, 78, 82, 85, 88, 72, 95, 68, 75, ... (50 values total)

Step 1: Visual Inspection

Create histogram:

- Shows somewhat bell-shaped distribution

- Slight left skew (tail extends toward lower scores)
- One outlier at 45 (student who didn't study)

Create Q-Q plot:

- Points mostly follow diagonal
- Slight deviation at lower tail (where outlier is)
- Overall pattern acceptable

Step 2: Conduct Shapiro-Wilk Test

Using statistical software with test scores:

- Test statistic: $W = 0.968$
- p-value = 0.183

Interpretation:

- $p = 0.183 > 0.05$
- Fail to reject H_0
- Conclude: Test scores are approximately normally distributed ✓

Step 3: Decision

Result: Data are normal enough for parametric tests like t-test or ANOVA.

However, the slight left skew visible in histogram suggests considering:

- Robust t-test (Welch's) instead of standard t-test
- Trimmed means approach
- Or non-parametric alternative if you want extra caution

Test for Homogeneity of Variance

Parametric tests comparing groups (t-test, ANOVA) assume equal variances across groups. Test this assumption before proceeding.

Levene's Test

Purpose: Tests whether variances are equal across groups.

Hypotheses:

- H_0: Variances are equal across groups: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$
- H_1: At least one group has different variance

Formula: Levene's test uses absolute deviations from group medians (or means), making it robust.

Interpretation:

- If $p > 0.05$: Fail to reject H_0 → Equal variances ✓
- If $p \leq 0.05$: Reject H_0 → Unequal variances ✗

Advantages:

- Robust to non-normality (doesn't require normality)
- Recommended over Bartlett's test in most situations

When to use: For testing equality of variances in any situation.

Bartlett's Test

Purpose: Alternative test for homogeneity of variance.

Hypotheses: Same as Levene's test.

Key difference: Assumes data are normally distributed (sensitive to departure from normality).

Interpretation: Same as Levene's ($p > 0.05$ = equal variances).

When to use: Only when you've confirmed normality first; less robust than Levene's.

Visual Method: Plotting Standard Deviations

Create bar plot of standard deviations for each group:

- Similar bar heights = equal variances ✓
- Very different heights = unequal variances ✗
- Rule of thumb: Largest SD ÷ Smallest SD < 2 is usually acceptable

---

Detailed Example: Testing Equal Variances in Teaching Methods

Scenario: Remember the teaching methods example?

- Method A: n=25, mean=78, SD=8
- Method B: n=28, mean=82, SD=9

Question: Are the variances significantly different?

Step 1: Visual Inspection

Plot standard deviations:

- Method A: SD = 8
- Method B: SD = 9
- Ratio: 9/8 = 1.125 (ratio < 2, suggests equal variances)

Step 2: Conduct Levene's Test

Using statistical software:

- Test statistic: F = 0.34
- p-value = 0.561

Interpretation:

- p = 0.561 > 0.05
- Fail to reject H_0
- Conclude: Variances are equal across methods ✓

Step 3: Decision

Result: Use standard Student's t-test (which assumes equal variances).

If p had been < 0.05, you would use Welch's t-test instead (doesn't assume equal variances).

Test for Independence

Purpose: Verify that observations are independent (not influenced by each other).

Why it matters: Violating independence inflates Type I error rates and makes p-values unreliable.

Testing method: Primarily through research design, not statistical tests.

Durbin-Watson Test

Purpose: Tests for autocorrelation in time series or sequential data.

Hypotheses:

- H_0: No autocorrelation between observations
- H_1: Autocorrelation exists

Interpretation:

- DW ≈ 2: No autocorrelation ✓
- DW < 2: Positive autocorrelation (consecutive values positively related)
- DW > 2: Negative autocorrelation (consecutive values negatively related)
- Range: DW from 0 to 4

When to use: When data are collected sequentially over time (stock prices, temperature measurements, etc.).

Design-Based Verification

Most independence issues are caught through study design:

Random assignment: Each subject/observation independently assigned to conditions **Randomized order**: Data collected in random order (not systematically) **No repeated measures on same unit**: Each subject measured once (unless paired design) **Blinded data collection**: Collector unaware of hypothesis/condition

---

Detailed Example: Testing Independence in Sales Data

Scenario: Store collects daily sales for 60 consecutive days to compare two promotional strategies:

- Days 1-30: Promotion A
- Days 31-60: Promotion B

Question: Are daily sales independent, or does each day's sales depend on previous day?

Step 1: Create Time Series Plot

Plot sales over 60 days:

- Days show random up-and-down pattern
- No obvious trend over time
- No clear clustering (suggesting independence)

Step 2: Conduct Durbin-Watson Test

Using statistical software on sales data:

- Durbin-Watson = 1.94
- Range for no autocorrelation: approximately 1.5 to 2.5

Interpretation:

- DW = 1.94 ≈ 2
- No significant autocorrelation detected ✓
- Sales on consecutive days are independent

Step 3: Decision

Result: Assumption of independence is met. Can proceed with parametric test (t-test comparing promotions).

If DW had been 1.2 (positive autocorrelation):

- Good sales days followed by good sales days
- Would suggest using time series methods or accounting for dependence
- Could be due to: day-of-week effects, promotions building momentum, etc.

Practical Decision Tree for Assumption Testing

Before conducting parametric test:

1. **Check Normality**
    - Create histogram and Q-Q plot
    - Conduct Shapiro-Wilk test
    - If $p > 0.05$ AND visuals look acceptable → Continue with parametric test
    - If $p \leq 0.05$ OR visuals show severe non-normality → Consider non-parametric test or transformation
2. **Check Homogeneity of Variance** (if comparing groups)
    - Compare standard deviations across groups
    - Conduct Levene's test
    - If $p > 0.05$ AND ratio of largest SD to smallest SD < 2 → Use standard parametric test
    - If $p \leq 0.05$ OR ratio > 2 → Use Welch's version or non-parametric test
3. **Check Independence**
    - Verify random assignment and randomized data collection

- For time series data, conduct Durbin-Watson test
- If concerns → Use methods accounting for dependence (mixed models, GEE)

4. **Check Scale of Measurement**
    - Verify data are interval/ratio (not ordinal/nominal)
    - If truly categorical → Use chi-square or other categorical test
    - If ordinal → Consider non-parametric test

When Assumptions Are Violated

Option 1: Data Transformation

Transform data to better meet assumptions:

- Log transformation: Use for right-skewed data (positive values only)
- Square root transformation: Use for count data
- Reciprocal transformation: Use for right-skewed data with zeros
- Box-Cox transformation: Automatic optimal transformation

Example: If test scores range 0-100 but cluster at high end (left-skewed), try log(101 - score) to reverse skew.

Option 2: Use Non-Parametric Alternative

- Non-normal, two independent groups → Mann-Whitney U instead of t-test
- Non-normal, paired groups → Wilcoxon Signed-Rank instead of paired t-test
- Non-normal, 3+ groups → Kruskal-Wallis instead of ANOVA
- Unequal variances → Welch's t-test or Welch's ANOVA

Option 3: Robust Methods

- Trimmed means: Calculate mean after removing extreme values (e.g., remove top/bottom 5%)
- Bootstrapping: Resample data with replacement to create sampling distribution
- Permutation tests: Calculate p-value by permuting group assignments

Option 4: Larger Sample Size

- Parametric tests become more robust with larger samples
- Central Limit Theorem: Sample means become normally distributed even if individual data aren't
- Rule of thumb: n > 30 per group helps even with non-normal data

Summary Table: Assumption Tests

Assumption	Test	H_0	Interpretation	Action if $p < 0.05$
Normality	Shapiro-Wilk	Data normal	$p > 0.05$ = normal	Use non-parametric or transform
Homogeneity	Levene's	Variances equal	$p > 0.05$ = equal	Use Welch's version
Independence	Durbin-Watson	No autocorrelation	$DW \approx 2$	Account for dependence
Normality	Kolmogorov-Smirnov	Data normal	$p > 0.05$ = normal	Use non-parametric or transform
Variance	Bartlett's	Variances equal	$p > 0.05$ = equal (if normal)	Use Welch's version

Best Practices for Assumption Testing

1. **Always visualize first**: Histograms and Q-Q plots often reveal issues better than formal tests
2. **Use multiple approaches**: Don't rely solely on p-values; combine formal tests with visual inspection
3. **Consider context**: With very large samples, minor violations may be statistically significant but practically unimportant
4. **Pre-register your plan**: Decide before analysis how you'll handle assumption violations
5. **Report what you tested**: Transparent reporting helps readers understand your decisions
6. **Be conservative with p-values**: Use $\alpha = 0.01$ or 0.10 for assumption tests (rather than 0.05) to reduce Type I error
7. **Small p-values in assumption tests don't mean you must abandon parametric tests**: With large n, minor deviations are detected. Use judgment about practical significance.

Common Parametric Tests

One-Sample t-test

Purpose: Tests whether a sample mean differs significantly from a known or hypothesized population mean.

Assumptions: Normal distribution, independence of observations.

Formula: $t = (\bar{x} - \mu_0) / (s / \sqrt{n})$

Where $\bar{x}$ is sample mean, μ_0 is hypothesized population mean, s is sample standard deviation, n is sample size.

Detailed Example: Product Lifespan Testing

A company claims their product lasts 1000 hours. You test 30 units and find mean = 950 hours. Does it differ from 1000?

Step 1: Set Up the Hypotheses

- H_0: $\mu = 1000$ (The true population mean lifespan is 1000 hours; the company's claim is correct)
- H_1: $\mu \neq 1000$ (The true population mean lifespan differs from 1000 hours; the company's claim is wrong)

This is a two-tailed test because we're checking for difference in either direction.

Step 2: Check Assumptions

- Verify that lifespan data is approximately normally distributed (check with histogram or Shapiro-Wilk test)
- Confirm observations are independent (each unit tested separately, no overlap)

Step 3: Collect Sample Data

- Sample size (n) = 30 units tested
- Sample mean ($\bar{x}$) = 950 hours (average lifespan of your 30 units)
- Sample standard deviation (s) = Let's say 80 hours (measure of variability in your sample)
- Hypothesized population mean (μ_0) = 1000 hours (company's claim)

Step 4: Calculate the Test Statistic

$t = (\bar{x} - \mu_0) / (s / \sqrt{n})$

$t = (950 - 1000) / (80 / \sqrt{30})$

$t = -50 / (80 / 5.477)$

$t = -50 / 14.61$

$t = -3.42$

Step 5: Interpret the Test Statistic

The t-statistic of -3.42 means your sample mean is 3.42 standard errors below the hypothesized population mean. The negative sign indicates the sample mean (950) is below the claimed mean (1000).

Step 6: Find the p-value

With t = -3.42 and degrees of freedom (df) = n - 1 = 29:

- Using a t-table or statistical software, p-value ≈ 0.002

This p-value means: If the true population mean really is 1000 hours, there's only a 0.2% chance of observing a sample mean as extreme as 950 hours (or more extreme in either direction).

Step 7: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value (0.002) < α (0.05), we **reject H_0**

Conclusion: There is statistically significant evidence that the true mean lifespan differs from 1000 hours. Based on your sample, the product appears to last significantly less than claimed.

Step 8: Calculate Effect Size

Cohen's $d = (\bar{x} - \mu_0) / s = (950 - 1000) / 80 = -50/80 = -0.625$

This is a medium effect size (0.5 to 0.8 range), indicating a practically meaningful difference, not just a statistically significant one.

Real-World Interpretation

- Statistical significance: The difference is unlikely due to random chance (p = 0.002)
- Practical significance: Products last about 50 hours less than claimed, which is a meaningful difference
- Business implication: You might investigate manufacturing quality, demand a refund from the supplier, or adjust marketing claims
- Confidence interval: A 95% CI around the sample mean would show the likely range of true population mean (approximately 920-980 hours)

Two-Sample t-test (Independent Samples)

Purpose: Compares means between two independent groups.

Assumptions: Both groups normally distributed, independent observations, equal variances.

Formula: $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s^2_{pooled} \times (1/n_1 + 1/n_2))}$

Variations:

- **Equal variances** (Student's t): Assumes both groups have equal variance
- **Unequal variances** (Welch's t): Doesn't assume equal variance; recommended when group sizes or variances differ substantially

Detailed Example: Comparing Teaching Methods

A school wants to know if two different teaching methods produce different student outcomes. They randomly assign students to either Method A or Method B and measure test scores.

Step 1: Organize Your Data

- Group 1 (Method A): $n_1 = 25$ students, mean = 78, SD = 8
- Group 2 (Method B): $n_2 = 28$ students, mean = 82, SD = 9

The difference in sample means is 78 - 82 = -4 points (Method B students scored 4 points higher on average).

Step 2: Set Up the Hypotheses

- H_0: $\mu_1 = \mu_2$ (The true population means are equal; there is no real difference in teaching effectiveness)
- H_1: $\mu_1 \neq \mu_2$ (The true population means differ; one teaching method is more effective)

This is a two-tailed test because we're checking whether either method is better.

Step 3: Check Assumptions

Normality: Examine histograms or Q-Q plots for both groups to verify approximately normal distributions.

Independence: Confirm students were randomly assigned to teaching methods (no overlap).

Equal Variances: Check if SD_1 (8) and SD_2 (9) are reasonably similar.

- Use Levene's test: H_0: $\sigma_1^2 = \sigma_2^2$
- Since the SDs are close (8 vs 9), equal variance assumption is probably reasonable
- If this assumption is violated, use Welch's t-test instead

Step 4: Calculate Pooled Variance

For Student's t-test with equal variances, combine variance information from both groups:

$s^2_{pooled} = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] / (n_1 + n_2 - 2)$

$s^2_{pooled} = [(25 - 1)(8^2) + (28 - 1)(9^2)] / (25 + 28 - 2)$

$s^2_{pooled} = [(24 \times 64) + (27 \times 81)] / 51$

$s^2_{pooled} = [1,536 + 2,187] / 51$

$s^2_{pooled} = 3,723 / 51$

$s^2_{pooled} = 73.0$

Standard error of the difference:
$SE = \sqrt{(s^2_{pooled} \times (1/n_1 + 1/n_2))}$

$SE = \sqrt{(73.0 \times (1/25 + 1/28))}$

$SE = \sqrt{(73.0 \times (0.04 + 0.0357))}$

$SE = \sqrt{(73.0 \times 0.0757)}$

$SE = \sqrt{5.53}$

$SE = 2.35$

Step 5: Calculate the Test Statistic

$t = (\bar{x}_1 - \bar{x}_2) / SE$

$t = (78 - 82) / 2.35$

$t = -4 / 2.35$

$t = -1.70$

Step 6: Determine Degrees of Freedom

$df = n_1 + n_2 - 2 = 25 + 28 - 2 = 51$

Step 7: Find the p-value

With $t = -1.70$ and $df = 51$:

- Using a t-table or statistical software, p-value ≈ 0.095

This p-value means: If the two teaching methods truly produced equal test scores in the population, there's a 9.5% chance of observing a difference as large as 4 points (or larger) between sample means.

Step 8: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value (0.095) > α (0.05), we **fail to reject H_0**

Conclusion: There is insufficient evidence to conclude that the two teaching methods produce significantly different test scores. The observed 4-point difference could reasonably be due to random sampling variation.

Step 9: Report Effect Size

Cohen's d = $(\bar{x}_1 - \bar{x}_2)$ / spooled

where spooled = $\sqrt{73.0} = 8.54$

d = (78 - 82) / 8.54 = -4 / 8.54 = -0.47

This is a small-to-medium effect size (0.4 to 0.6 range).

Step 10: Calculate Confidence Interval

95% CI for difference in means = $(\bar{x}_1 - \bar{x}_2) \pm$ t_critical × SE

t_critical for df=51, α=0.05 (two-tailed) ≈ 2.008

95% CI = -4 ± (2.008 × 2.35) = -4 ± 4.72 = (-8.72, 0.72)

Real-World Interpretation

Statistical Result: With p = 0.095, the difference is not statistically significant at the 0.05 level. It's marginally close to significance (this is sometimes called "approaching significance").

Practical Meaning: Method B students scored 4 points higher on average, but we can't confidently say this difference is real rather than due to chance. The 95% confidence interval (-8.72 to 0.72) means we're 95% confident the true difference between methods is somewhere in this range, which includes zero.

Important Caveat: The effect size (d = -0.47) suggests a meaningful practical difference despite lack of statistical significance. This could be due to insufficient sample size. With larger samples, we might reach statistical significance.

Possible Next Steps:

1. Increase sample size and retest
2. Investigate whether there are other factors affecting performance (student motivation, class size, teacher experience)
3. Consider whether a 4-point difference is practically important for educational decisions (if it is, Method B might be worth adopting despite non-significance)
4. Look at the confidence interval rather than just the p-value for decision-making

When to Use Welch's t-test Instead

If Levene's test showed unequal variances ($p < 0.05$), or if you noticed substantially different SDs (like 8 vs 15), use Welch's t-test instead:

$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$

This formula doesn't assume equal variances and is increasingly recommended as the default choice.

Paired t-test

Purpose: Compares two related/paired samples (e.g., before-after measurements on same subjects).

Assumptions: Differences are normally distributed, paired observations are independent.

Formula: $t = \bar{D} / (sD / \sqrt{n})$

Where $\bar{D}$ is mean of differences, sD is standard deviation of differences.

Detailed Example: Blood Pressure Medication Study

A pharmaceutical company tests whether their new blood pressure medication is effective. They measure systolic blood pressure (mmHg) for 20 patients before taking the medication and again 4 weeks after starting treatment.

Step 1: Organize the Data

Here's sample data from 10 of the 20 patients (imagine similar data for remaining 10):

Patient	Before	After	Difference (D)
1	145	138	-7
2	152	148	-4
3	148	145	-3
4	160	150	-10
5	155	152	-3
6	158	154	-4
7	150	149	-1
8	162	155	-7
9	148	147	-1
10	156	151	-5
...	...	...	...
All 20 patients	**Average**	**Average**	**$\bar{D}$ = -4.8**

Key point: The difference is calculated as After - Before. Negative values mean blood pressure decreased (medication worked).

Step 2: Set Up the Hypotheses

- H_0: $\mu D = 0$ (The true mean difference is zero; the medication has no effect)
- H_1: $\mu D \neq 0$ (The true mean difference is not zero; the medication has an effect)

This is a two-tailed test because we're checking whether the medication changes blood pressure in either direction.

Step 3: Check Assumptions

Normality of differences: Check whether the differences (not the original before/after values) are approximately normally distributed using:

- Histogram of the 20 difference values
- Q-Q plot
- Shapiro-Wilk test (H_0: differences are normal; if $p > 0.05$, assumption met)

Independence: Each patient's measurement pair is independent (no overlap between patients).

Step 4: Calculate Summary Statistics for Differences

Based on all 20 patients' difference values (D):

- Sample size: $n = 20$
- Mean difference: $\bar{D} = -4.8$ mmHg (on average, blood pressure dropped 4.8 points)
- Standard deviation of differences: $sD = 5.2$ mmHg (measure of variability in how much each person's BP changed)

Why standard deviation matters: Some patients may have dropped 10 mmHg while others only 1 mmHg. The sD captures this variability.

Step 5: Calculate Standard Error

The standard error tells you how much variation to expect in sample mean differences:

$SE = sD / \sqrt{n} = 5.2 / \sqrt{20} = 5.2 / 4.47 = 1.16$ mmHg

This means if you repeated the study many times, the sample mean difference would vary with a standard deviation of about 1.16 mmHg.

Step 6: Calculate the Test Statistic

$t = \bar{D} / SE = -4.8 / 1.16 = -4.14$

What this means: The sample mean difference is 4.14 standard errors away from zero. This is a large value, suggesting the observed difference is substantial relative to the natural variation.

Step 7: Determine Degrees of Freedom

$df = n - 1 = 20 - 1 = 19$

Step 8: Find the p-value

With $t = -4.14$ and $df = 19$:

- Using a t-table or statistical software, p-value ≈ 0.0007

This p-value means: If the medication truly had no effect ($\mu D = 0$), there's only a 0.07% chance of observing a sample mean difference as extreme as -4.8 mmHg (or larger in magnitude in either direction).

Step 9: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value (0.0007) < α (0.05), we **reject H_0**

Conclusion: There is statistically significant evidence that the medication affects blood pressure. Based on the negative mean difference, the medication significantly reduces blood pressure.

Step 10: Calculate Effect Size

Cohen's d for paired samples = $\bar{D} / sD = -4.8 / 5.2 = -0.92$

This is a large effect size (>0.8), indicating the medication has a practically meaningful impact.

Step 11: Calculate Confidence Interval

95% CI for mean difference = $\bar{D} \pm t\_critical \times SE$

$t\_critical$ for df=19, α=0.05 (two-tailed) ≈ 2.093

95% CI = $-4.8 \pm (2.093 \times 1.16) = -4.8 \pm 2.43 = (-7.23, -2.37)$

Real-World Interpretation

Statistical Result: With p = 0.0007, the result is highly statistically significant. The medication clearly affects blood pressure.

Practical Meaning: On average, patients experienced a 4.8 mmHg reduction in systolic blood pressure. We're 95% confident the true population mean reduction is between 2.37 and 7.23 mmHg.

Clinical Significance: A 4.8 mmHg reduction is clinically meaningful. For context:

- Normal blood pressure: <120 mmHg
- Elevated: 120-129 and <80
- Stage 1 Hypertension: 130-139 or 80-89
- Stage 2 Hypertension: ≥140 or ≥90

For a hypertensive patient, nearly a 5 mmHg drop can move them into a lower risk category.

Advantages of Paired Design

The paired t-test is more powerful than independent samples t-test because:

1. **Controls for individual differences**: Each person serves as their own control. If some patients naturally have higher BP, this doesn't affect the analysis
2. **Reduces noise**: We only care about changes within each person, not absolute values
3. **Smaller sample size needed**: You need fewer subjects because the design is more efficient

Example of why this matters: Two patients might have very different baseline BPs (one 140, one 155), but both might drop the same amount (5 mmHg). The paired test focuses on this common change, whereas comparing groups might get confused by the different starting points.

What if Results Were Different?

Scenario 1: If p-value = 0.08 (not significant)

- Conclusion: Insufficient evidence that medication affects BP
- Action: Might need larger sample, longer trial period, or redesigned study

Scenario 2: If $\bar{D}$ = -1.5 mmHg (smaller effect)

- Statistical significance depends on variability
- Even small changes can be significant if very consistent across patients
- Clinical significance would be questionable for such small change

Checking Assumptions in Practice

Before conducting the paired t-test:

1. **Look at differences histogram**: Should be roughly bell-shaped, not heavily skewed or bimodal
2. **Q-Q plot**: Points should roughly follow the diagonal line

3. **Check for outliers**: One patient with -25 mmHg change while others have -2 to -7 might indicate a data entry error or unusual response

Non-Parametric Alternative: If differences are severely non-normal, use the **Wilcoxon Signed-Rank Test** instead, which doesn't assume normality but tests whether the median difference differs from zero.

One-Way ANOVA (Analysis of Variance)

Purpose: Tests whether means differ significantly across three or more independent groups.

Assumptions: Normal distribution in each group, homogeneity of variance, independent observations.

Key Concept: Partitions total variance into between-group variance (signal) and within-group variance (noise).

F-statistic: F = (Between-group variance) / (Within-group variance)

Larger F values indicate greater differences between group means.

Detailed Example: Comparing Sales Across Four Regional Stores

A retail company wants to know if average daily sales differ significantly across their four regional stores. They collect 30 days of sales data from each store.

Step 1: Organize the Data

Region	Sample Size	Mean Daily Sales ($)	Std Dev ($)
North	30	4,250	520
South	30	4,580	610
East	30	3,920	480
West	30	4,420	550
Overall	**120**	**4,293**	**580**

The overall mean sales across all regions is $4,293. But each region varies from this average:

- South is about $287 above average (highest)
- East is about $373 below average (lowest)

Step 2: Set Up the Hypotheses

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (All four regional stores have equal mean daily sales; regional location doesn't affect sales)

- H_1: At least one region has a different mean (Regional location affects sales; at least one store differs from the others)

This is not a specific alternative—we're just testing whether any difference exists among the four groups.

Step 3: Check Assumptions

Normality: For each of the four regions, verify daily sales are approximately normally distributed:

- Create histograms for each region
- Run Shapiro-Wilk test for each group
- Look for severe skewness or outliers

Homogeneity of Variance: Check if the variability (SD) is similar across regions:

- Levene's test: H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
- SDs range from 480 to 610 (roughly similar), so assumption is probably met
- If Levene's test $p < 0.05$, assumption is violated; consider transformation or use Welch's ANOVA

Independence: Each day's sales are independent; sales on different days don't affect each other.

Step 4: Calculate the Grand Mean and Deviations

Grand mean (overall average) = \$4,293

Deviations from grand mean for each region:

- North: 4,250 - 4,293 = -\$43
- South: 4,580 - 4,293 = +\$287
- East: 3,920 - 4,293 = -\$373
- West: 4,420 - 4,293 = +\$127

Step 5: Calculate Between-Group Variance

Between-group variance measures how far each group mean is from the overall mean:

$SS\_between = \Sigma\ n\_i(\bar{x}\_i - \bar{x}\_grand)^2$

$SS\_between = 30(-43)^2 + 30(287)^2 + 30(-373)^2 + 30(127)^2$

$SS\_between = 30(1,849) + 30(82,369) + 30(139,129) + 30(16,129)$

$SS\_between = 55,470 + 2,471,070 + 4,173,870 + 483,870$

$SS\_between = 7,184,280$

Degrees of freedom: df_between = k - 1 = 4 - 1 = 3

Mean square between: MS_between = SS_between / df_between = 7,184,280 / 3 = 2,394,760

Step 6: Calculate Within-Group Variance

Within-group variance measures average variability around each group's mean (how "noisy" the data is):

$SS_{within} = \Sigma(n_i - 1)s_i^2$

$SS_{within} = (30-1)(520)^2 + (30-1)(610)^2 + (30-1)(480)^2 + (30-1)(550)^2$

$SS_{within} = 29(270,400) + 29(372,100) + 29(230,400) + 29(302,500)$

$SS_{within} = 7,841,600 + 10,790,900 + 6,681,600 + 8,772,500$

$SS_{within} = 34,086,600$

Degrees of freedom: df_within = N - k = 120 - 4 = 116

Mean square within: MS_within = SS_within / df_within = 34,086,600 / 116 = 293,851

Step 7: Calculate the F-Statistic

F = MS_between / MS_within = 2,394,760 / 293,851 = 8.15

What this means: The between-group variance is 8.15 times larger than the within-group variance. This suggests meaningful differences between regions.

Step 8: Determine Degrees of Freedom

- df_numerator = 3 (between groups)
- df_denominator = 116 (within groups)

Step 9: Find the p-value

With F(3, 116) = 8.15:

- Using F-table or statistical software, p-value ≈ 0.00008

This p-value means: If all four regions truly had equal mean sales, there's only a 0.008% chance of observing an F-statistic as large as 8.15.

Step 10: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value (0.00008) < α (0.05), we **reject H_0**

Conclusion: There is statistically significant evidence that mean daily sales differ across the four regional stores. At least one region has significantly different sales than the others.

Step 11: Calculate Effect Size

Eta-squared (η^2) = SS_between / SS_total

SS_total = SS_between + SS_within = 7,184,280 + 34,086,600 = 41,270,880

η^2 = 7,184,280 / 41,270,880 = 0.174

This means regional location explains about 17.4% of the variance in daily sales. This is a large effect size.

Step 12: Post-Hoc Tests (Pairwise Comparisons)

Since ANOVA is significant, we need to identify *which* regions differ from each other. Use Tukey HSD (Honestly Significant Difference) test:

Tukey HSD critical value ≈ 1,100 (for 4 groups, df=116, α=0.05)

Compare all pairs:

Comparison	Difference	Significant?
South vs East	4,580 - 3,920 = 660	No (660 < 1,100)
South vs North	4,580 - 4,250 = 330	No
South vs West	4,580 - 4,420 = 160	No
North vs East	4,250 - 3,920 = 330	No
West vs East	4,420 - 3,920 = 500	No
North vs West	4,250 - 4,420 = -170	No

Interesting finding: The overall ANOVA is significant ($p < 0.05$), but pairwise comparisons show no single pair differs significantly. This can happen when:

- All groups are moderately different from each other (none drastically so)
- The cumulative effect of these small differences creates overall significance
- Sample size is large enough to detect these moderate differences

Real-World Interpretation

Statistical Finding: Sales differ significantly across regions ($p = 0.00008$, $\eta^2 = 0.174$).

Practical Meaning:

- South store averages $330 more per day than North, and $660 more than East
- East store is the lowest performer
- Regional location explains 17.4% of sales variation (other factors like foot traffic, staffing, local competition explain the remaining 82.6%)

Business Implications:

1. Investigate East store: Why is it underperforming? (Staff training, location issues, local competition)
2. Study South store: What's working well there? (Marketing, staff quality, customer service)
3. Set differential sales targets based on region if appropriate
4. Consider whether differences are due to legitimate factors (population density, demographics) or controllable factors (management, staffing)

Why Not Just Do Multiple t-tests?

A common mistake is conducting separate t-tests:

- North vs South: t-test
- North vs East: t-test
- North vs West: t-test
- South vs East: t-test
- South vs West: t-test
- East vs West: t-test

This is 6 tests! With $\alpha = 0.05$ per test, the overall Type I error rate balloons to approximately 26% (instead of 5%). You'd likely find false positives by chance alone.

ANOVA avoids this by testing all groups simultaneously, then using post-hoc tests with adjusted significance levels.

Assumptions Violated?

If Levene's test showed unequal variances ($p < 0.05$), use **Welch's ANOVA** instead, which doesn't assume equal variances.

If normality is violated, use the non-parametric **Kruskal-Wallis test** instead, which tests whether groups differ without assuming normal distributions.

Pearson Correlation

Purpose: Measures the linear relationship between two continuous variables.

Assumes: Both variables approximately normally distributed, linear relationship, homoscedasticity.

Formula: $r = \Sigma[(xi - \bar{x})(yi - \bar{y})] / \sqrt{(\Sigma(xi - \bar{x})^2 \times \Sigma(yi - \bar{y})^2)}$

Ranges from -1 to +1.

Example: Testing correlation between study hours and exam scores.

Linear Regression

Purpose: Predicts one variable (Y) from another (X) and tests the significance of the relationship.

Assumptions: Linear relationship, homogeneity of variance of residuals, normality of residuals, independence of observations.

Model: $\hat{Y} = a + bX$

Tests whether slope b differs significantly from zero.

Non-Parametric Tests

What Are Non-Parametric Tests?

Non-parametric tests make few or no assumptions about the underlying distribution of the population. They're often called "distribution-free" tests and work with ranks or categories rather than actual values. They're more flexible but generally have less statistical power than parametric tests.

Key Characteristics of Non-Parametric Tests

- **Distribution assumption**: No assumption about underlying distribution
- **Data type**: Work with ordinal, nominal, or non-normal interval/ratio data
- **Information used**: Use ranks, signs, or categories; ignore actual values
- **Statistical power**: Generally lower power than parametric counterparts
- **Robustness**: More robust to outliers and violations of normality
- **Flexibility**: Can handle skewed data, ordinal data, and small samples

Common Non-Parametric Tests

Mann-Whitney U Test (Wilcoxon Rank-Sum Test)

Purpose: Non-parametric alternative to independent samples t-test. Compares central tendencies of two independent groups.

Data requirement: Ordinal or non-normal interval/ratio data.

Procedure:

1. Combine all data from both groups
2. Rank all values from smallest to largest
3. Sum ranks for each group separately
4. Calculate U statistic comparing the rank sums

Example: Comparing pain levels (ordinal scale: mild, moderate, severe) between two treatment groups.

Advantages over t-test: Doesn't assume normality, handles ordinal data, less affected by outliers.

Wilcoxon Signed-Rank Test

Purpose: Non-parametric alternative to paired t-test. Tests whether paired observations differ.

Data requirement: Ordinal or non-normal interval/ratio data for paired samples.

Procedure:

1. Calculate differences for each pair
2. Rank absolute differences
3. Assign signs based on whether difference is positive or negative
4. Compare sum of positive ranks to sum of negative ranks

Example: Testing whether ratings of a product differ before and after a marketing campaign (same 50 consumers rate before and after).

Advantages: Doesn't assume normality of differences, handles ordinal data, robust to outliers.

Kruskal-Wallis Test

Purpose: Non-parametric alternative to one-way ANOVA. Tests whether three or more independent groups differ.

Data requirement: Ordinal or non-normal interval/ratio data.

Procedure:

1. Rank all observations across all groups
2. Calculate H statistic based on rank sums for each group
3. Compare to chi-square distribution

Example: Comparing customer satisfaction (ordinal scale) across five different store locations.

Advantages over ANOVA: Doesn't assume normality, robust to outliers, appropriate for ordinal data.

Post-hoc tests: If significant, use Mann-Whitney U tests with Bonferroni correction for pairwise comparisons.

Spearman's Rank Correlation (ρ)

Purpose: Non-parametric alternative to Pearson correlation. Measures monotonic relationship between two variables.

Data requirement: Ordinal data or interval/ratio data that violates normality assumptions.

Procedure:

1. Rank values for each variable separately
2. Calculate Pearson correlation on the ranks

Formula: $\rho = 1 - (6\Sigma d^2) / (n(n^2 - 1))$

Where d is difference in ranks for each pair, n is number of pairs.

Example: Correlation between movie critic rankings and audience rankings (both ordinal).

Advantages: Captures non-linear monotonic relationships, handles ordinal data, robust to outliers.

Kendall's Tau

Purpose: Another non-parametric correlation measure, similar to Spearman's.

Interpretation: Based on concordant vs. discordant pairs.

Advantage: Often more accurate for smaller samples than Spearman's.

Friedman Test

Purpose: Non-parametric alternative to repeated measures ANOVA. Tests whether three or more paired observations differ.

Data requirement: Ordinal or non-normal interval/ratio data with repeated measurements.

Example: Rating satisfaction with three different coffee brands (same 30 people taste all three).

Advantages: No normality assumption, appropriate for ordinal data, handles repeated measures.

Chi-Square Test of Independence

Purpose: Tests relationship between two categorical variables (whether they're associated or independent).

Data requirement: Categorical (nominal) data in contingency table format.

Formula: $\chi^2 = \Sigma[(O - E)^2 / E]$

Where O = observed frequency, E = expected frequency under independence.

Detailed Example: Product Preference by Gender

A marketing company wants to know if product preference is related to customer gender. They survey 300 customers about which of three products they prefer: Product A, Product B, or Product C.

Step 1: Create a Contingency Table

Here's the observed data (actual counts from the survey):

	Product A	**Product B**	**Product C**	**Row Total**
Male	45	38	27	110
Female	42	85	63	190
Column Total	**87**	**123**	**90**	**N = 300**

Observations:

- 110 males and 190 females surveyed (unequal split)
- Males seem to prefer A relatively more (45/110 = 41%)
- Females seem to prefer B relatively more (85/190 = 45%)
- Females represent larger share of B and C purchasers

But are these differences statistically significant, or just due to random sampling variation?

Step 2: Set Up the Hypotheses

- H_0: Gender and product preference are independent (no association; preference doesn't depend on gender)
- H_1: Gender and product preference are associated (preference differs by gender)

Step 3: Calculate Expected Frequencies

Under the null hypothesis (independence), the expected frequency for each cell is:

E = (Row Total × Column Total) / N

For each cell:

Male × Product A: E = (110 × 87) / 300 = 32.1

Male × Product B: E = (110 × 123) / 300 = 45.1

Male × Product C: E = (110 × 90) / 300 = 33.0

Female × Product A: E = (190 × 87) / 300 = 54.9

Female × Product B: E = (190 × 123) / 300 = 77.9

Female × Product C: E = (190 × 90) / 300 = 57.0

Expected Frequencies Table:

	Product A	**Product B**	**Product C**
Male	32.1	45.1	33.0
Female	54.9	77.9	57.0

Interpretation: If gender didn't matter, you'd expect males to be distributed across products proportionally: 29% to A, 41% to B, 30% to C.

Step 4: Calculate the Chi-Square Statistic

For each cell, calculate $(O - E)^2 / E$:

Cell	**O**	**E**	**O - E**	**$(O - E)^2$**	**$(O - E)^2/E$**
Male × A	45	32.1	12.9	166.41	5.18
Male × B	38	45.1	-7.1	50.41	1.12
Male × C	27	33.0	-6.0	36.00	1.09
Female × A	42	54.9	-12.9	166.41	3.03
Female × B	85	77.9	7.1	50.41	0.65
Female × C	63	57.0	6.0	36.00	0.63

Sum: $\chi^2 = 5.18 + 1.12 + 1.09 + 3.03 + 0.65 + 0.63 =$ **11.70**

Step 5: Determine Degrees of Freedom

df = (rows - 1) × (columns - 1) = (2 - 1) × (3 - 1) = 1 × 2 = 2

Step 6: Find the p-value

With $\chi^2 = 11.70$ and df = 2:

- Using a chi-square table or statistical software, p-value ≈ 0.0028

This p-value means: If gender and product preference were truly independent, there's only a 0.28% chance of observing a chi-square value as extreme as 11.70.

Step 7: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value (0.0028) < α (0.05), we **reject H_0**

Conclusion: There is statistically significant evidence that product preference is associated with gender. Gender and product preference are not independent.

Step 8: Examine the Pattern

To understand the association, look at the differences between observed and expected:

Males compared to expected:

- Product A: 45 observed vs 32.1 expected (+12.9 more) — males prefer A more
- Product B: 38 observed vs 45.1 expected (-7.1 fewer) — males prefer B less
- Product C: 27 observed vs 33.0 expected (-6.0 fewer) — males prefer C less

Females compared to expected:

- Product A: 42 observed vs 54.9 expected (-12.9 fewer) — females prefer A less
- Product B: 85 observed vs 77.9 expected (+7.1 more) — females prefer B more
- Product C: 63 observed vs 57.0 expected (+6.0 more) — females prefer C slightly more

Pattern: Males prefer Product A, females prefer Products B and C.

Step 9: Calculate Effect Size

Cramér's V measures strength of association for chi-square tests:

$V = \sqrt{(\chi^2 / (N \times (\min(r,c) - 1)))}$

$V = \sqrt{(11.70 / (300 \times (2-1)))} = \sqrt{(11.70 / 300)} = \sqrt{0.039} = 0.197$

Interpretation: 0.197 is a small-to-medium effect size (0.1 = small, 0.3 = medium, 0.5 = large).

Step 10: Calculate Percentages for Clarity

It's often clearer to examine conditional percentages:

	Product A	Product B	Product C	Total
Male	40.9%	34.5%	24.5%	100%
Female	22.1%	44.7%	33.2%	100%

Clear pattern emerges:

- 40.9% of males prefer A, but only 22.1% of females
- 44.7% of females prefer B, but only 34.5% of males

Real-World Interpretation

Statistical Finding: Product preference and gender are significantly associated ($p = 0.0028$, Cramér's $V = 0.197$).

Practical Meaning:

- Males strongly favor Product A (41% vs expected 29%)
- Females favor Product B (45% vs expected 41%)
- The relationship is statistically significant but effect size is modest

Business Implications:

1. **Marketing strategy**: Target Product A ads toward males, Products B/C toward females
2. **Product development**: If considering male-focused features, integrate them into Product A
3. **Store placement**: Place Products B/C in areas frequented by female customers
4. **Inventory management**: Stock Product A more heavily for male-heavy locations

Assumptions and Conditions

For chi-square test validity:

1. **Expected frequency ≥ 5**: All expected cells must have frequency ≥ 5
    - Our table shows all E values ≥ 32.1, so assumption is satisfied
    - If violated, combine categories or use Fisher's exact test
2. **Independence of observations**: Each surveyed person counted only once

3. **Categorical data**: Data must be counts in categories, not continuous measurements

What if Expected Frequencies Were Low?

If some expected frequencies were < 5:

- Combine related categories (e.g., combine Products B and C if low counts)
- Use **Fisher's Exact Test** (for 2×2 tables)
- Use **Monte Carlo simulation** for larger tables
- Collect larger sample size

Comparing to Other Tests

Why not use t-test? Gender is categorical (binary), product preference is categorical (3 options). We're analyzing counts in categories, not continuous measurements. T-tests are inappropriate here.

Why not just use percentages? Percentages describe the data but don't tell us if differences are statistically significant. Chi-square provides statistical inference and a p-value.

Non-Parametric Note: Chi-square IS already a non-parametric test—it makes no assumptions about underlying distributions. It's appropriate for categorical data where parametric tests (assuming normality) don't apply.

Mann-Whitney U Test (Wilcoxon Rank-Sum Test)

Purpose: Non-parametric alternative to independent samples t-test. Compares central tendencies of two independent groups.

Data requirement: Ordinal or non-normal interval/ratio data.

Formula: $U = n_1n_2 + (n_1(n_1+1))/2 - R_1$

Where n_1 and n_2 are sample sizes, R_1 is sum of ranks for group 1.

Detailed Example: Comparing Patient Pain Levels After Two Treatments

A hospital compares two pain management approaches. Group 1 receives standard medication, Group 2 receives a new herbal supplement. Patients rate pain on a 1-10 ordinal scale (1 = no pain, 10 = severe pain) after treatment.

Step 1: Organize the Data

Group 1 (Standard Med)	Group 2 (Herbal Supplement)
4, 5, 3, 6, 2, 5, 4, 7	3, 2, 1, 4, 2, 3, 2, 1, 3, 2
n_1 = 8 patients	n_2 = 10 patients

Step 2: Set Up the Hypotheses

- H_0: The two treatments produce equal pain relief (distributions are the same)
- H_1: The two treatments differ in pain relief (distributions are different)

Step 3: Check Assumptions

Normality check: Create histograms for both groups

- Group 1 histogram shows slight skewness (more patients with higher pain ratings)
- Group 2 histogram also skewed (clustered at lower values)
- Shapiro-Wilk test $p < 0.05$ confirms non-normality

Since data violate normality, Mann-Whitney U is appropriate (parametric t-test would be risky).

Step 4: Rank All Data Combined

Combine all values and rank from smallest to largest, keeping track of which group each came from:

Value	Group	Rank
1	2	1
1	2	2
2	1	3
2	2	4
2	2	5
2	2	6
3	1	7
3	2	8
3	2	9
3	2	10
4	1	11
4	1	12
4	2	13
5	1	14
5	1	15
6	1	16
7	1	17

Note on ties: When values tie (like three 3's), give each the average rank. For the three 2's in positions 4-6, each gets rank (4+5+6)/3 = 5.

Sum of ranks for each group:

- Group 1: 3 + 7 + 11 + 12 + 14 + 15 + 16 + 17 = $R_1 = 95$
- Group 2: 1 + 2 + 4 + 5 + 6 + 8 + 9 + 10 + 13 = $R_2 = 58$

Step 5: Calculate U Statistics

$U_1 = n_1n_2 + (n_1(n_1+1))/2 - R_1$

$U_1 = (8)(10) + (8 \times 9)/2 - 95 = 80 + 36 - 95 =$ **21**

$U_2 = n_1n_2 + (n_2(n_2+1))/2 - R_2$

$U_2 = (8)(10) + (10 \times 11)/2 - 58 = 80 + 55 - 58 =$ **77**

Verification: $U_1 + U_2$ should equal $n_1n_2 = 80$. Indeed, 21 + 77 = 98. (Note: Due to rounding in ranks, may be off by 1)

Convention: Use the smaller U value: U = 21

Step 6: Find the p-value

With U = 21, $n_1 = 8$, $n_2 = 10$:

- Using Mann-Whitney U table or statistical software, p-value ≈ 0.047

This p-value means: If the two treatments produced identical pain relief, there's only a 4.7% chance of observing data this extreme (one group having much lower ranks than the other).

Step 7: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value (0.047) < α (0.05), we **reject H_0**

Conclusion: There is statistically significant evidence that the two treatments differ in pain relief. The herbal supplement appears more effective (lower pain ratings).

Step 8: Calculate Effect Size

Effect size for Mann-Whitney U (rank-biserial correlation):

$r = 1 - (2U) / (n_1 \times n_2) = 1 - (2 \times 21) / (8 \times 10) = 1 - 0.525 = 0.475$

This is a medium-to-large effect size.

Real-World Interpretation

Statistical Finding: Pain levels differ significantly between treatments (p = 0.047, r = 0.475).

Practical Meaning: The herbal supplement group had lower pain ratings overall. Median pain for Group 1 (standard) appears around 4-5, while Group 2 (herbal) median appears around 2-3.

Clinical Interpretation:

- Herbal supplement is about 2 points more effective on 1-10 scale

- This is clinically meaningful for pain management
- Effect size (r = 0.475) indicates a substantial difference

Advantages over t-test:

- Doesn't assume normal distribution (data were skewed)
- Doesn't use actual values (robust to outliers)
- Appropriate for ordinal pain ratings
- Uses ranks, making it less sensitive to extreme values

Why Mann-Whitney Instead of t-test?: If we'd used a parametric t-test on this non-normal data:

- Could give misleading results
- Assumptions violated, so p-value less trustworthy
- Mann-Whitney U is more appropriate and robust

Wilcoxon Signed-Rank Test

Purpose: Non-parametric alternative to paired t-test. Tests whether paired observations differ.

Data requirement: Ordinal or non-normal interval/ratio data for paired samples.

Detailed Example: Customer Satisfaction Before and After Website Redesign

An e-commerce company redesigned their website and wants to know if customer satisfaction improved. They surveyed 15 customers on a 1-10 scale before and after the redesign.

Step 1: Organize the Data

Customer	Before	After	Difference	|D|	Rank of |D|	Signed Rank
1	6	8	+2	2	6.5	+6.5
2	5	7	+2	2	6.5	+6.5
3	7	7	0	-	-	-
4	4	8	+4	4	11.5	+11.5
5	6	9	+3	3	8.5	+8.5
6	5	6	+1	1	1.5	+1.5
7	8	8	0	-	-	-
8	5	9	+4	4	11.5	+11.5
9	7	6	-1	1	1.5	-1.5
10	4	6	+2	2	6.5	+6.5

11	6	7	+1	1	1.5	+1.5
12	5	8	+3	3	8.5	+8.5
13	7	9	+2	2	6.5	+6.5
14	6	5	-1	1	1.5	-1.5
15	5	7	+2	2	6.5	+6.5

Step 2: Set Up the Hypotheses

- H_0: $\mu D = 0$ (No change in satisfaction; redesign had no effect)
- H_1: $\mu D \neq 0$ (Satisfaction changed; redesign had an effect)

Step 3: Check Assumptions

Check normality of differences:

- Histogram shows roughly symmetric distribution (roughly normal)
- Shapiro-Wilk test p = 0.08 (doesn't reject normality)
- But with only n=15, even parametric paired t-test is reasonable

However, Wilcoxon test still appropriate and more robust to any non-normality.

Step 4: Calculate Differences

D = After - Before

Most differences are positive (satisfaction increased), with two negative (decreased).

Step 5: Rank the Absolute Differences

Ignore the sign temporarily and rank by magnitude:

- |D| = 0: Two zeros (customers 3, 7) are excluded from analysis. n goes from 15 to 13.
- |D| = 1: Four values (1, 1, 1, 1) → Average ranks = (1+2+3+4)/4 = 2.5... Actually, they get ranks 1.5 each (positions 1-2), 1.5 (positions 3-4), etc. More precisely: ranks 1-4 average to 2.5

Let me recalculate precisely:

- |D| = 1: Positions 1,2,3,4 → ranks 1.5, 1.5, 1.5, 1.5
- |D| = 2: Positions 5-10 → ranks 7.5 (average of 5-10)
- |D| = 3: Positions 11,12 → ranks 11.5
- |D| = 4: Positions 13,14 → ranks 13.5

Step 6: Assign Signs to Ranks

Add back the original sign direction:

- Positive differences get positive signed ranks: +6.5, +6.5, +11.5, +8.5, +1.5, +11.5, +6.5, +1.5, +8.5, +6.5, +6.5
- Negative differences get negative signed ranks: -1.5, -1.5

Step 7: Calculate Test Statistic

Sum of positive signed ranks: $T_+ = 6.5 + 6.5 + 11.5 + 8.5 + 1.5 + 11.5 + 6.5 + 1.5 + 8.5 + 6.5 + 6.5 =$ **83**

Sum of negative signed ranks: $T_- = -1.5 + -1.5 =$ **-3**

Test statistic: $T = \min(|T_+|, |T_-|) = \min(83, 3) =$ **3**

(The smaller of the two absolute values)

Step 8: Find the p-value

With T = 3 and n = 13 (excluding zero differences):

- Using Wilcoxon table or statistical software, p-value ≈ 0.0005

This p-value means: If the redesign had no effect, there's only a 0.05% chance of observing data this extreme (such a large imbalance between positive and negative differences).

Step 9: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value (0.0005) < α (0.05), we **reject H_0**

Conclusion: There is highly significant evidence that customer satisfaction changed after the website redesign. Based on the 11 positive vs. 2 negative differences, satisfaction increased.

Step 10: Calculate Effect Size

Effect size (rank-biserial correlation):

$r = 1 - (2T) / (n(n+1)) = 1 - (2 \times 3) / (13 \times 14) = 1 - 0.033 =$ **0.967**

This is a very large effect size (approaching 1.0).

Real-World Interpretation

Statistical Finding: Satisfaction significantly increased after redesign (p = 0.0005, r = 0.967).

Practical Meaning:

- 11 of 13 customers (85%) experienced increased satisfaction
- Average increase approximately 2 points on 1-10 scale
- Very consistent effect (nearly everyone improved)

Business Implication: Website redesign was highly successful. The vast majority of customers are more satisfied.

Advantages over Paired t-test:

- Doesn't assume normality of differences
- Uses ranks rather than actual values (robust to outliers)
- More appropriate for ordinal satisfaction ratings
- Clear direction of effect (most customers improved)

Kruskal-Wallis Test

Purpose: Non-parametric alternative to one-way ANOVA. Tests whether three or more independent groups differ.

Data requirement: Ordinal or non-normal interval/ratio data.

Detailed Example: Customer Satisfaction Across Three Service Channels

A company offers customer support through three channels: Phone, Email, and Chat. They want to know if satisfaction differs by channel. They survey 45 customers (15 per channel) on satisfaction using a 1-10 ordinal scale.

Step 1: Organize the Data

Phone	**Email**	**Chat**
9, 8, 7, 9, 8	6, 5, 7, 6, 5	8, 7, 9, 8, 7
8, 9, 7, 8, 9	6, 7, 6, 5, 6	7, 8, 7, 9, 8
8, 7, 9, 8, 7	5, 6, 7, 6, 5	8, 9, 7, 8, 9
$n_1 = 15$	**$n_2 = 15$**	**$n_3 = 15$**
Mean = 8.1	**Mean = 6.0**	**Mean = 8.0**

Step 2: Set Up the Hypotheses

- H_0: All three channels produce equal satisfaction (distributions are the same)
- H_1: At least one channel differs in satisfaction

Step 3: Check Assumptions

Normality: Histograms show:

- Phone: Clustered at high values (8-9)
- Email: Clustered at low values (5-7)
- Chat: Spread across 7-9

Shapiro-Wilk tests show non-normal distributions in each group. This violates ANOVA assumptions, making Kruskal-Wallis appropriate.

Independence: Each customer surveyed once through one channel.

Step 4: Rank All Data Combined

Combine all 45 satisfaction ratings and rank from 1 to 45:

Sample ranking (simplified):

- Value 5 appears 8 times → ranks 1-8 → average rank = 4.5 each
- Value 6 appears 10 times → ranks 9-18 → average rank = 13.5 each
- Value 7 appears 12 times → ranks 19-30 → average rank = 24.5 each
- Value 8 appears 11 times → ranks 31-41 → average rank = 36 each
- Value 9 appears 4 times → ranks 42-45 → average rank = 43.5 each

Sum of ranks for each group:

- Phone (mostly 8-9): $R_1 = 35 \times 8.1 \approx$ **545** (approximate)
- Email (mostly 5-7): $R_2 = 35 \times 6.0 \approx$ **200** (approximate)
- Chat (mostly 7-9): $R_3 = 35 \times 8.0 \approx$ **535** (approximate)

Step 5: Calculate Kruskal-Wallis H Statistic

$H = [12 / (N(N+1))] \times \Sigma(R^2_i/n_i) - 3(N+1)$

Where N = total sample size (45), R_i = sum of ranks for group i, n_i = group size

$H = [12 / (45 \times 46)] \times [(545^2/15) + (200^2/15) + (535^2/15)] - 3(46)$

$H = [12 / 2070] \times [19,792 + 2,667 + 19,092] - 138$

$H = 0.0058 \times 41,551 - 138$

$H = 240.8 - 138$

$H = \mathbf{102.8}$

Step 6: Determine Degrees of Freedom

$df = k - 1 = 3 - 1 = 2$

(where k is number of groups)

Step 7: Find the p-value

With $H = 102.8$ and $df = 2$:

- Using chi-square table (H approximately follows chi-square distribution), p-value < 0.0001

This p-value means: If all three channels produced equal satisfaction, there's less than 0.01% chance of observing an H value as extreme as 102.8.

Step 8: Make a Decision

Using significance level $\alpha = 0.05$:

- Since p-value < 0.0001 < α (0.05), we **reject H_0**

Conclusion: There is highly significant evidence that customer satisfaction differs across the three support channels. Email provides significantly lower satisfaction than Phone or Chat.

Step 9: Calculate Effect Size

Epsilon-squared (ε^2):

$\varepsilon^2 = (H - k + 1) / (N - k) = (102.8 - 3 + 1) / (45 - 3) = 100.8 / 42 = \mathbf{2.4}$

Note: This calculation can exceed 1; an alternative is:

$\eta^2 = (H - k + 1) / (N - 1) = 100.8 / 44 = \mathbf{2.3}$

These large values (>0.14) indicate strong effect.

Step 10: Post-Hoc Comparisons

Since H is significant, perform pairwise Mann-Whitney U tests to identify which channels differ:

- Phone vs Email: U-statistic p < 0.0001 (highly different)

- Phone vs Chat: U-statistic p = 0.92 (not different)
- Email vs Chat: U-statistic p < 0.0001 (highly different)

Apply Bonferroni correction: With 3 comparisons, use $\alpha = 0.05/3 = 0.0167$ instead of 0.05.

Real-World Interpretation

Statistical Finding: Satisfaction differs significantly across channels ($H = 102.8$, $p < 0.0001$).

Practical Meaning:

- **Phone**: Mean satisfaction 8.1 (highest; customers most satisfied)
- **Email**: Mean satisfaction 6.0 (lowest; customers least satisfied)
- **Chat**: Mean satisfaction 8.0 (comparable to phone)
- Phone and Chat perform equally well; Email significantly underperforms

Business Implications:

1. **Investigate Email**: Why are email support customers less satisfied? (Slow response time, impersonal tone, unclear answers)
2. **Study Phone/Chat**: What makes these channels successful? (Personal interaction, quick resolution, clear communication)
3. **Improve Email**: Train staff, improve templates, reduce response time
4. **Resource allocation**: Consider shifting resources from email to phone/chat
5. **Customer choice**: Let customers prefer phone or chat if available

Advantages over ANOVA:

- Doesn't assume normal distributions
- Robust to ordinal satisfaction ratings
- Handles non-linear relationships
- More trustworthy with non-normal data (email clustered at 5-7, phone at 8-9)

When Assumptions Violated: If Kruskal-Wallis is also inappropriate (very discrete data, extreme outliers), consider:

- Fisher's exact test (for 2×2 categorical data)
- Median test (tests if medians equal)
- Permutation tests

Comparison Table

Aspect	Parametric	Non-Parametric
Distribution assumption	Assumes normal distribution	No distribution assumption
Data type	Interval/Ratio	Ordinal, Nominal, or non-normal Interval/Ratio
Statistical power	Higher	Lower
Effect of outliers	Sensitive	Robust
Sample size	Generally needs moderate to large	Works with small samples
Computational complexity	Moderate	Generally simpler
Parameter focus	Tests means, variances	Tests medians, ranks, distributions

---

Choosing Between Parametric and Non-Parametric Tests

Use Parametric Tests When:

- Data are approximately normally distributed
- Data are interval or ratio scale
- Sample size is reasonably large (n > 30 helps even if not perfectly normal)
- Homogeneity of variance assumption is met
- You want maximum statistical power

Use Non-Parametric Tests When:

- Data violate normality (confirmed by Shapiro-Wilk or Q-Q plot)
- Data are ordinal or nominal
- Sample size is very small
- Data contain severe outliers that can't be removed
- Homogeneity of variance is severely violated
- When working with ranked or categorical data inherently

Decision Tree

1. **What type of data do you have?**
    - Categorical/Nominal → Use Chi-Square or similar categorical tests
    - Ordinal → Use non-parametric tests
    - Interval/Ratio → Go to step 2
2. **Is data approximately normal?**
    - No → Use non-parametric test
    - Yes → Go to step 3
3. **Are variances roughly equal?**
    - No → Use Welch's t-test (parametric) or Mann-Whitney U (non-parametric)
    - Yes → Use standard parametric test
4. **How many groups/samples?**
    - 1 sample: One-sample t-test vs. Sign test
    - 2 independent samples: Independent t-test vs. Mann-Whitney U
    - 2 paired samples: Paired t-test vs. Wilcoxon Signed-Rank
    - 3+ independent: ANOVA vs. Kruskal-Wallis
    - 3+ paired: Repeated measures ANOVA vs. Friedman test
    - Correlation: Pearson vs. Spearman/Kendall

Practical Example: Choosing the Right Test

Scenario: You're testing whether a new training program improves employee productivity.

Data collected: Productivity scores (1-10 scale) from 25 employees, measured before and after training.

Decision process:

1. Data type: Ordinal (1-10 scale) or Interval?
    - If treated as ordinal → Use Wilcoxon Signed-Rank Test
    - If treated as interval → Check normality
2. Check normality of differences:
    - Plot histogram of before-after differences
    - Run Shapiro-Wilk test

- If $p < 0.05$ → Data not normal → Use Wilcoxon Signed-Rank Test
- If $p > 0.05$ → Data normal → Use Paired t-test

3. Conclude and report:
    - State which test was used and why
    - Report test statistic and p-value
    - Describe practical significance with effect size

Common Mistakes to Avoid

Using parametric tests with clearly non-normal data: Always check assumptions first.

Ignoring ordinal nature of data: A 1-10 rating scale is often better analyzed with non-parametric methods.

Assuming larger sample size eliminates non-normality concerns: While large samples help, severe non-normality should still lead to non-parametric tests.

Choosing test for power alone: Validity is more important than power. Invalid results are worthless.

Not checking for outliers: Before choosing a test, examine your data for extreme values that might need handling.

Applying parametric procedures to ranks already converted to ordinal: Use non-parametric tests on ordinal data; don't convert to intervals artificially.

Conclusion

Both parametric and non-parametric tests are valuable tools. Parametric tests offer greater power when their assumptions are met, while non-parametric tests provide flexibility and robustness when assumptions are violated. The key is understanding your data, checking assumptions, and selecting the appropriate test. When in doubt, consult your data's characteristics rather than defaulting to familiar tests.