Lecture 10: Classical Tests & the LRT Framework

One Principle · Many Tests · One Decision Flowchart

Previously, on Lecture 9...

Logic: Assume H_0 (nothing happening). Ask: how surprising is my data under H_0?

p-value: $P(\text{this extreme or more} \mid H_0)$. **NOT** $P(H_0 \mid \text{data})$.

Errors: Type I (α) = false alarm. Type II (β) = missed detection. Power $= 1 - \beta$.

Tables: z-table when σ known. t-table when σ unknown (df $= n - 1$).

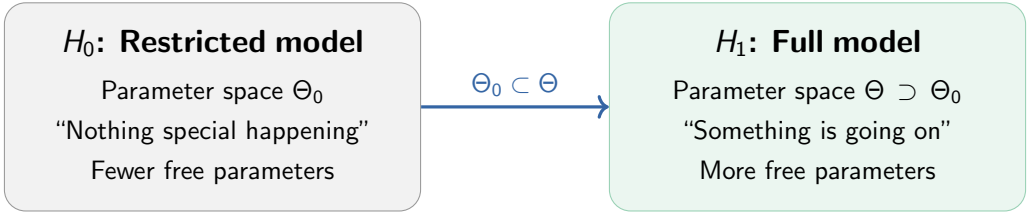
Tools: Permutation tests (no assumptions). Multiple testing: Holm/BF (FWER) or BH (FDR).

Today: The framework is clear. But which **specific test** do I actually use?

Part I: The Unifying Principle

Why so many tests? Because they all come from one idea.

Nested Models: Simple vs Complex



Every test asks: does the full model fit *significantly* better than the restricted model?

$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ (1 free param vs 0 $\Rightarrow k = 1$)

$H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ (1 mean vs 2 $\Rightarrow k = 1$)

H_0: rows & columns independent vs H_1: associated $(k = (r-1)(c-1))$

The Likelihood Ratio Test (LRT)

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

Numerator: best fit *with* constraint (H_0). **Denominator:** best fit *without* constraint.

The Likelihood Ratio Test (LRT)

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

Numerator: best fit *with* constraint (H_0). **Denominator:** best fit *without* constraint.

Λ close to 1

H_0 fits almost as well as H_1.
No evidence against H_0.
$\Rightarrow$ fail to reject.

Λ close to 0

H_0 fits much worse than H_1.
Strong evidence against H_0.
$\Rightarrow$ reject H_0.

In practice, work with $-2 \log \Lambda$ (large = evidence against H_0).

Wilks' Theorem: $-2\log\Lambda \to \chi^2$

$$-2\log\Lambda \xrightarrow{d} \chi^2_k \quad \text{as } n \to \infty$$

where $k = \dim(\Theta) - \dim(\Theta_0)$ = number of constraints imposed by H_0.

Test one mean
$k = 1$
$-2\log\Lambda \sim \chi^2_1$

Two means equal
$k = 1$
$-2\log\Lambda \sim \chi^2_1$

$r \times c$ independence
$k = (r-1)(c-1)$
$-2\log\Lambda \sim \chi^2_{(r-1)(c-1)}$

One theorem produces all the classical tests.
The z-test, t-test, and χ^2-test are all special cases of the LRT.

Caveat: Wilks' theorem is **asymptotic** ($n \to \infty$). Poor approximation for small n or boundary hypotheses.

LRT Worked Example: Testing a Normal Mean

Setup: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ (known σ^2). $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

LRT Worked Example: Testing a Normal Mean

Setup: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ (known σ^2). $\quad H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

Numerator (best fit under H_0): $\quad L(\mu_0) = \prod \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i - \mu_0)^2/2\sigma^2}$

Denominator (best fit overall): $\quad L(\hat{\mu}) = L(\bar{X})$ $\quad$ (MLE is $\bar{X}$)

LRT Worked Example: Testing a Normal Mean

Setup: $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ (known σ^2). $\quad H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

Numerator (best fit under H_0): $\quad L(\mu_0) = \prod \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i-\mu_0)^2/2\sigma^2}$

Denominator (best fit overall): $\quad L(\hat{\mu}) = L(\bar{X})$ $\quad$ (MLE is $\bar{X}$)

Log-likelihood ratio:

$$-2\log\Lambda = -2\big[\ell(\mu_0) - \ell(\bar{X})\big] = \frac{n(\bar{X}-\mu_0)^2}{\sigma^2} = \left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 = Z^2$$

LRT Worked Example: Testing a Normal Mean

Setup: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ (known σ^2). $\quad H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

Numerator (best fit under H_0): $\quad L(\mu_0) = \prod \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i-\mu_0)^2/2\sigma^2}$

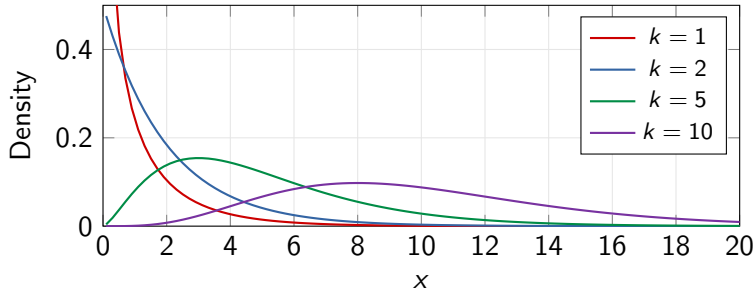
Denominator (best fit overall): $\quad L(\hat{\mu}) = L(\bar{X})$ $\quad$ (MLE is $\bar{X}$)

Log-likelihood ratio:

$$-2\log\Lambda = -2\big[\ell(\mu_0) - \ell(\bar{X})\big] = \frac{n(\bar{X}-\mu_0)^2}{\sigma^2} = \left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 = Z^2$$

The LRT statistic is just the z-**statistic squared**: $-2\log\Lambda = Z^2 \sim \chi^2_1$.
By Wilks' theorem, $k = 1$ (one constraint: $\mu = \mu_0$). Everything checks out!
Replace σ with $S \Rightarrow$ you get the t-**test**. Same idea, different distribution.

The χ^2 Distribution



Key facts: Mean $= k$, Variance $= 2k$. Sum of k independent $N(0,1)^2$ values. Skewed right; becomes symmetric as k grows.
Decision: Reject H_0 when $-2\log\Lambda > \chi^2_{k,\alpha}$ (the upper-α critical value from the χ^2 table).

Part II: Tests for Means

Each test below is a special case of the LRT.

One-sample, paired, and two-sample

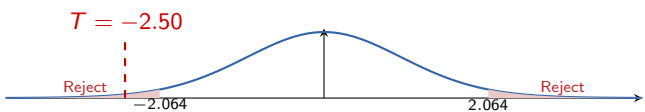
One-Sample t-Test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \text{under } H_0$$

Example: Coffee shop claims cups contain $\mu_0 = 350$ mL. We measure $n = 25$: $\bar{X} = 345$, $S = 10$.

$$T = \frac{345 - 350}{10/\sqrt{25}} = \frac{-5}{2} = -2.50$$



$t_{24,0.025} = 2.064$. Since $|T| = 2.50 > 2.064$: **reject** H_0. The cups are underfilled.

Assumptions: (1) Approx. Normal (or $n \geq 30$), (2) independent. Non-Normal + small n: Wilcoxon (Part IV).

One-sided: $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ — one tail only.

Paired t-Test: Before vs After

If you have paired observations (X_i, Y_i), compute differences $D_i = X_i - Y_i$ and apply a one-sample t-test on D_i.

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D \neq 0$$

$$T = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t_{n-1}$$

Example: Blood pressure before/after drug, $n = 8$ patients.

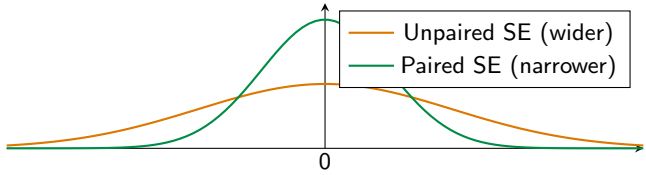
Before	148	142	136	134	138	140	132	144
After	140	138	132	130	135	137	130	140
D_i	8	4	4	4	3	3	2	4

$\bar{D} = 4.0$, $S_D = 1.77$, $T = \dfrac{4.0}{1.77/\sqrt{8}} = \dfrac{4.0}{0.626} = 6.39$.

$t_{7,0.025} = 2.365$. Since $|T| = 6.39 \gg 2.365$: **reject** H_0. The drug works. Cohen's $d = \bar{D}/S_D = 4.0/1.77 = 2.26$ (very large).

Why Pairing Matters

Each subject is their own control — removes between-subject variability



Unpaired (two-sample)

Compares group means.
Variability includes *between-subject* differences.
More noise ⇒ lower power.

Paired

Compares *within-subject* change.
Removes between-subject variability.
Less noise ⇒ higher power.

Always pair when you can!

Two-Sample t-Test: Comparing Two Groups

Two independent groups: n_1 observations with $(\bar{X}_1, S_1)$ and n_2 with $(\bar{X}_2, S_2)$

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

Equal variances assumed: $$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

Pooled SD: $$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}$$

Two-Sample t-Test: Comparing Two Groups

Two independent groups: n_1 observations with $(\bar{X}_1, S_1)$ and n_2 with $(\bar{X}_2, S_2)$

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

Equal variances assumed: $$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

Pooled SD: $$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}$$

Welch's t-test (no equal-variance assumption — default choice):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \qquad \text{df} = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$$

Always default to Welch's. Correct whether variances are equal or not. The pooled version only gains a tiny bit of power when $\sigma_1 = \sigma_2$.

R and Python use Welch by default. Don't "test for equal variances first, then decide" — this two-step procedure distorts your Type I error.

Two-Sample Example: Teaching Methods

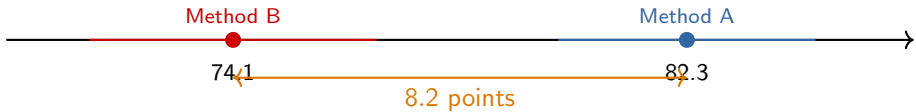
Method A ($n_1 = 10$): $\bar{X}_1 = 82.3$, $S_1 = 8.5$. Method B ($n_2 = 12$): $\bar{X}_2 = 74.1$, $S_2 = 10.2$.

Welch's t-test:

$$T = \frac{82.3 - 74.1}{\sqrt{8.5^2/10 + 10.2^2/12}} = \frac{8.2}{\sqrt{7.225 + 8.67}} = \frac{8.2}{3.99} = 2.06$$

df ≈ 19.8, $t_{19.8,\,0.025} \approx 2.09$.

$|T| = 2.06 < 2.09$: **fail to reject** at $\alpha = 0.05$ (just barely!).



Cohen's $d \approx 0.87$ (large effect!). Underpowered study — not the effect's fault.

Test for a Proportion

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$$

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

(Use p_0 in SE, not $\hat{p}$!)

Example: Is a coin fair? $n = 200$ flips, $k = 115$ heads, $\hat{p} = 0.575$.

$$Z = \frac{0.575 - 0.5}{\sqrt{0.5 \times 0.5/200}} = \frac{0.075}{0.0354} = 2.12$$

p-value $= 2(1 - \Phi(2.12)) = 2 \times 0.017 = 0.034 < 0.05$: **reject** H_0. Evidence the coin is biased.

Rule of thumb: use this test when $np_0 \geq 10$ and $n(1-p_0) \geq 10$.
For two proportions: $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ where $\hat{p} = \frac{k_1+k_2}{n_1+n_2}$
(pooled under H_0: $p_1 = p_2$).

Part III: Chi-Squared Tests

When the data is counts, not measurements

Goodness-of-Fit: Does the Data Match a Distribution?

Observed counts $O_1, \ldots, O_k$ across k categories, total n.
Expected under H_0: $E_i = n \cdot p_i^0$.

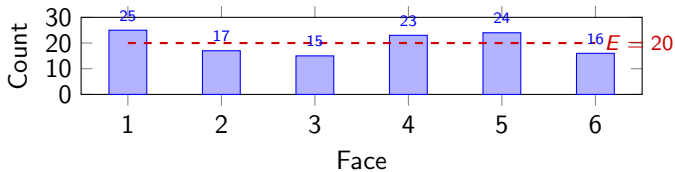
$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \;\dot{\sim}\; \chi^2_{k-1}$$

Goodness-of-Fit: Does the Data Match a Distribution?

Observed counts $O_1, \ldots, O_k$ across k categories, total n.
Expected under H_0: $E_i = n \cdot p_i^0$.

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \;\dot{\sim}\; \chi^2_{k-1}$$

Example: Is a die fair? Roll $n = 120$ times. $H_0 : p_i = 1/6 \Rightarrow E_i = 20$.



$X^2 = \frac{25}{20} + \frac{9}{20} + \frac{25}{20} + \frac{9}{20} + \frac{16}{20} + \frac{16}{20} = 5.0$. $\quad \text{df} = 5$. $\quad \chi^2_{5,\,0.05} = 11.07$.
$5.0 < 11.07$: **fail to reject**. No evidence the die is unfair.

Assumption: All $E_i \geq 5$. If some expected counts < 5, merge categories or use an exact test.

Test of Independence: Two Categorical Variables

Setup: Contingency table with r rows and c columns. H_0: variables are independent.

$$\text{Under } H_0: \quad E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{col } j \text{ total})}{n}$$

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{\cdot}{\sim} \chi^2_{(r-1)(c-1)}$$

Test of Independence: Two Categorical Variables

Setup: Contingency table with r rows and c columns. H_0: variables are independent.

$$\text{Under } H_0: \quad E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{col } j \text{ total})}{n}$$

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \;\dot{\sim}\; \chi^2_{(r-1)(c-1)}$$

Example: Smoking and lung cancer ($n = 200$).

	Cancer	No cancer	Total
Smoker	**40** (E=25)	**60** (E=75)	100
Non-smoker	**10** (E=25)	**90** (E=75)	100
Total	50	150	200

$X^2 = \frac{(40-25)^2}{25} + \frac{(60-75)^2}{75} + \frac{(10-25)^2}{25} + \frac{(90-75)^2}{75} = 9 + 3 + 9 + 3 = \mathbf{24.0}$

$\text{df} = (2-1)(2-1) = 1$. $\chi^2_{1,\,0.05} = 3.84$. $24.0 \gg 3.84$: **reject**. Strong association.

Cramér's $V = \sqrt{24/200} = 0.35$ (medium). $E_{ij} < 5$? Fisher's exact test. χ^2 tests association, not causation.

Part IV: When Assumptions Fail

Nonparametric alternatives: no Normal required

Mann–Whitney U Test

Comparing two independent groups — nonparametric alternative to two-sample t

Idea: Replace values with their *ranks*, then compare rank sums.

A's ranks: $2 + 3 + 5 + 7 + 9 = 26$. **B's ranks:** $1 + 4 + 6 + 8 + 10 = 29$.

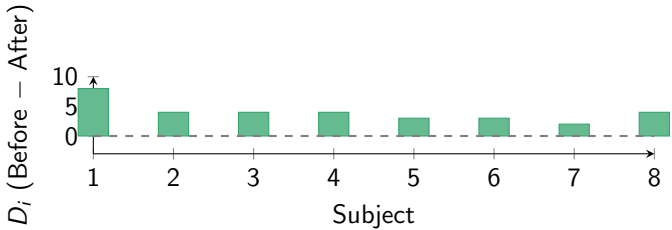
If groups are identical, rank sums should be similar.

When to use: skewed data, outliers, ordinal data, small n with non-Normal distributions.

Wilcoxon Signed-Rank Test

Paired nonparametric alternative to the paired t-test

Compute differences D_i, rank the $|D_i|$, then compare the sum of positive vs negative ranks.



All $D_i > 0$: all positive ranks.
$W^+ = 1 + 2 + \cdots + 8 = 36, \quad W^- = 0$
$W = \min(W^+, W^-) = 0$
Very small $\Rightarrow$ reject H_0.

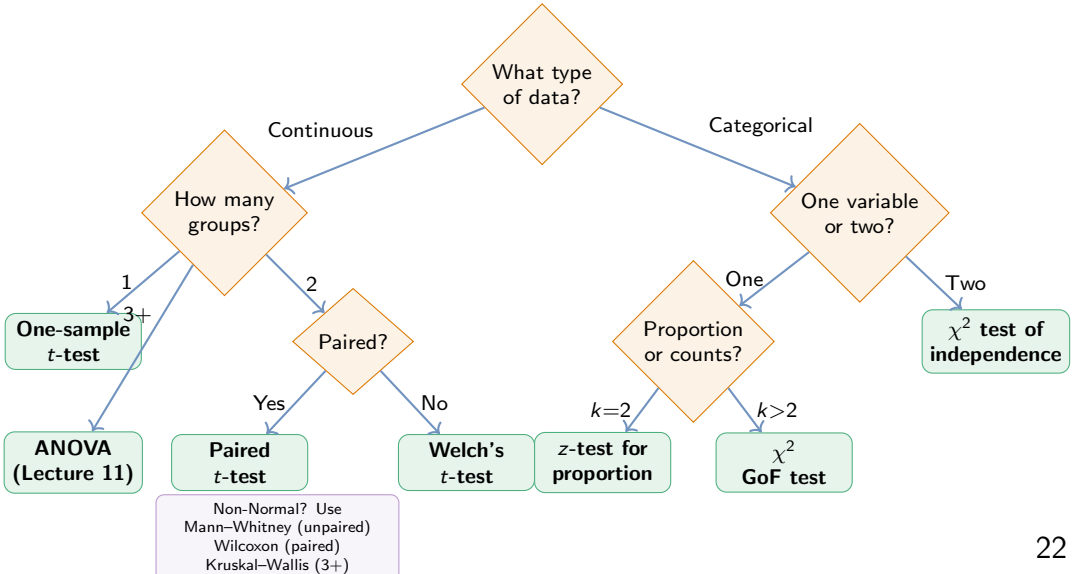
When to use:

Paired data, but not Normal.
Ordinal measurements.
Small samples with outliers.

Trade-off: less power than t-test when data actually is Normal.

3+ groups: Kruskal–Wallis (L11).

Which Test Do I Use?



Classical Tests in Python

```python
from scipy import stats

# One-sample t-test: H0: mu = 350
stats.ttest_1samp(data, popmean=350)

# Paired t-test
stats.ttest_rel(before, after)

# Two-sample Welch's t-test (default)
stats.ttest_ind(group_a, group_b, equal_var=False)

# Chi-squared test of independence
stats.chi2_contingency([[40,60],[10,90]])

# Nonparametric alternatives
stats.mannwhitneyu(group_a, group_b)  # unpaired
stats.wilcoxon(before - after)        # paired
```

Each function returns `(statistic, p_value)`. Always check which alternative hypothesis is default.

Summary: Classical Tests & LRT

LRT: Compare best fit under H_0 vs overall. $-2\log\Lambda \sim \chi^2_k$ (Wilks' theorem).

One-sample t: $T = (\bar{X} - \mu_0)/(S/\sqrt{n}) \sim t_{n-1}$. The workhorse of statistics.

Paired t: Compute differences first, then one-sample t. More powerful than unpaired.

Welch's t: Two independent means. No equal-variance assumption. Default choice.

Proportions: Z-test with SE $= \sqrt{p_0(1-p_0)/n}$. Use when $np \geq 10$.

χ^2 **GoF:** Do observed counts match expected? $\sum(O-E)^2/E \sim \chi^2_{k-1}$.

χ^2 **independence:** Are two categorical variables associated? df $= (r-1)(c-1)$.

Nonparametric: Mann–Whitney (independent), Wilcoxon (paired) when normality fails.

Practical: Classical Tests

1. **One-sample t-test:**
   - Coffee shop data (n=25, $\bar{X}$=345, S=10, μ_0=350)
   - Compute T, find p from t-table, verify with `scipy.stats.ttest_1samp`
2. **Paired t-test:**
   - Blood pressure before/after. Compute by hand and with `ttest_rel`
3. **Welch's t-test:**
   - Teaching method data. Use `ttest_ind(equal_var=False)`
   - Compare p with the permutation test from L9
4. **χ^2 independence test:**
   - Smoking/cancer 2×2 table. Compute by hand
   - Verify with `scipy.stats.chi2_contingency`

Homework

1. A manufacturer claims mean weight is 500 g. Sample $n = 16$: $\bar{X} = 495$, $S = 8$. Test at $\alpha = 0.05$. Compute the 95% CI. Does it contain 500?

2. 12 runners' times (s) before and after training:
   Before: 58.2, 57.1, 60.3, 55.8, 59.6, 62.1, 56.4, 61.2, 58.7, 57.5, 60.8, 59.3.
   After: 56.1, 55.8, 58.7, 54.2, 57.3, 59.8, 55.1, 59.0, 56.4, 56.0, 58.2, 57.1.
   (a) Paired t-test. (b) Unpaired t-test on the same data. Why does the p-value change?

3. Survey of 300 people on preferred transport: bus 90, metro 120, car 60, bicycle 30. Test whether preferences are equally distributed (χ^2 goodness-of-fit).

4. Study method (group/solo) vs exam result (pass/fail): Group+pass 70, Group+fail 30, Solo+pass 55, Solo+fail 45. Test independence at $\alpha = 0.05$.

5. Startup salaries (thousands): Team A: 120, 45, 38, 200, 52. Team B: 55, 62, 48, 70, 58. (a) Why is a t-test problematic here? (b) Use Mann–Whitney U (`scipy.stats.mannwhitneyu`).

Recommended Visualizations & Resources

Interactive: Seeing Theory — Frequentist Inference (Brown)

`seeing-theory.brown.edu/frequentist-inference` — hypothesis tests, χ^2, t-tests with interactive sliders.

Video: StatQuest — t-Tests Explained

Clear walkthrough of one-sample, paired, and two-sample t-tests. Also: "Chi-Square Tests."

Interactive: R Psychologist — Independent Samples t-Test

`rpsychologist.com/d3/tdist/` — visualize t-distribution, rejection regions, and effect sizes.

Reading: Wasserman, "All of Statistics" — Ch. 10–11

Hypothesis testing, LRT framework, t-tests, χ^2 tests. Concise and rigorous.

Python: `scipy.stats`

`ttest_1samp`, `ttest_ind`, `ttest_rel`, `chi2_contingency`, `mannwhitneyu`, `wilcoxon`.

Questions?

Next: Lecture 11 — ANOVA & A/B testing