Hypothesis Testing

UC Berkeley, MIDS w203

Statistics for Data Science

September 28, 2022

Introducing the Two-Sample t-Test

Motivating the Two-Sample t-Test

An Important Data Science Question

Is group A different from group B?

Examples:

- Are customers who get a birthday gift less likely to leave than those who don't?
- Do patients who take Vitamin W get over the flu faster than patients who don't?
- Do democracies or autocracies start more wars?

Example Scenario

From the Journal of Empirical Fashion

	New Yorkers	San Franciscans
black outfits	12.1	13.3
sample size	50	50

Is this evidence that San Franciscans have more black outfits than New Yorkers *in expectation*?

Two-Sample Model Framework

Basic Model Setup

Suppose $(X_1, .., X_{n_1})$ are i.i.d. with mean μ_X.

Suppose $(Y_1, .., Y_{n_2})$ are i.i.d. with mean μ_Y.

Null Hypothesis

$H_0 : \mu_X = \mu_Y$ (The two population means are equal)

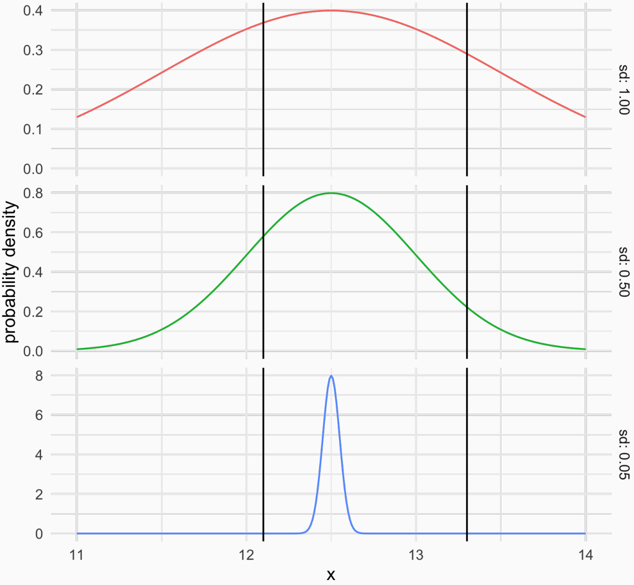
Alternative hypotheses

$H_1 : \mu_X \neq \mu_Y$ (best choice in most cases)

$H_2 : \mu_X > \mu_Y$

$H_3 : \mu_X < \mu_Y$

Comparing Population Means



Comparing Population Means

Two-Sample t-Test

$$t = \frac{\overline{X} - \overline{Y}}{\text{Estimate of Standard Deviation}}$$

Assess whether two populations have the same expectation **while accounting for variability**

The Two-Sample z-Test

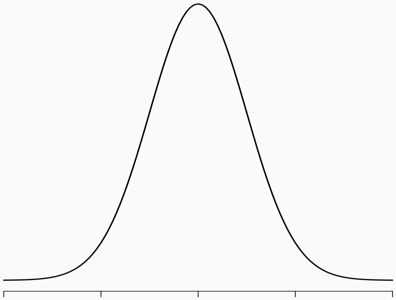
Two-Sample z-Test

Assumptions

Suppose $(X_1, .., X_{n_X})$ are i.i.d. with mean μ_X.

Suppose $(Y_1, .., Y_{n_Y})$ are i.i.d. with mean μ_Y.

Assume $X \sim N(\mu_X, \sigma_X)$. $Y \sim N(\mu_Y, \sigma_Y)$. We know σ_X, σ_Y.



From the Journal of Empirical Fashion

	New Yorkers	San Franciscans
black outfits	12.1	13.3
sample size	50	50

Let $X_1, \ldots, X_{50}$ rep. New Yorkers. Assume iid, mean μ_X.
Let $Y_1, \ldots, Y_{50}$ rep. San Franciscans. Assume iid, mean μ_Y.
Assume $\sigma_X = \sigma_Y = 2$.

General Procedure for Testing

Three steps:

1. Specify model, null hypothesis, rejection criterion
2. Calculate statistic
3. Plot statistic on the null distribution to get the p value.

The Two-Sample t-Test

Types of Two-Sample t-Tests

1. Student's t-Test
2. Welch's t-Test

The Two-Sample z-Test

$$z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

- Problem : we don't know σ_X or σ_Y

Student's t-Test

- Estimate a single "pooled" standard deviation, s.
- Substitute s for both s_X and s_Y.

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s^2}{n_X} + \frac{s^2}{n_Y}}}$$

Theorem

if $\sigma_X = \sigma_Y$, t has a T distribution with $n_X + n_Y - 2$ degrees of freedom.

degrees of freedom (*df*)

Number of independent pieces of information that vary given estimated parameters

- **One sample t-test**
    - Model has one parameter (the mean)
    - Given the sample mean, and $n-1$ observations, can compute the last one.
    - $df = n - 1$
- **Student's two-sample t-test**
    - Model has two parameters (μ_X and μ_Y)
    - Given the sample means, $n_X - 1$ observations for X and $n_Y - 1$ observations for Y, can compute the rest
    - $df = n_X + n_Y - 2$

Student's t-Test Summary

- Tests if mean of X equals mean of Y.
- Uses a pooled estimate for standard deviation.
- **Major disadvantage:** Only valid if $\sigma_X = \sigma_Y$.

Welch's t-Test

- Compute two sample standard deviations: s_X and s_Y.
- Substitute s_X for σ_X and s_Y for σ_Y.

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

Theorem

t has *approximately* a T distribution. The degrees of freedom are given by

$$v = \frac{\left(s_X^2/n_X + s_Y^2/n_Y\right)^2}{\frac{\left(s_X^2/n_X\right)^2}{n_X - 1} + \frac{\left(s_Y^2/n_Y\right)^2}{n_Y - 1}}$$

Choosing a Two-Sample t-Test

Some authors recommend a two step process:

1. Use Levine's test for equal variances ($H_0 : \sigma_X = \sigma_Y$)
2. If non-significant, proceed with Student's t-Test

Our advice: always use Welch's t-Test

- Power is almost as high as for Student's test
- We never know for sure if variances are equal
- This is the default in most statistical software

Welch's Two-Sample t-Test Assumptions

- **Metric Scale:** $X_1, X_2, ..X_{n_X}$ and $Y_1, Y_2, ..., Y_{n_Y}$ are random variables measured on a metric scale.
- **Independence:** X's are iid, Y's are iid, and X's and Y's are mutually independent.
- **Normality:** The distribution of the X's is normal and the distribution of the Y's is normal
  - The CLT guarantees normality for large samples
  - Main concern is strong skewness with a small sample

Practical Significance of the T-Test

Practical Significance for the T-Test

After using a t-test to assess statistical significance, it is important to assess practical significance.

> *Your main goal is to explain to your audience why they should or should not care about the effect.*

Three common effect size measures:

1. Difference in means
2. Cohen's d
3. Correlation r

Difference in Means

Difference in means

$$\overline{X}_A - \overline{X}_B$$

- Answers the question "*How different are these groups?*"
- Often makes great headlines and is a good choice if units are familiar
- But lacks context in its calculation
- People who eat chocolate live 1.5 years longer than those who do not each chocolate

Cohen's d

Cohen's d

Cohen's d is a measure of difference of means standardized by the variance in the data.

$$\frac{\overline{X}_A - \overline{X}_B}{s}$$

Where s is a pooled standard deviation: $\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2}}$

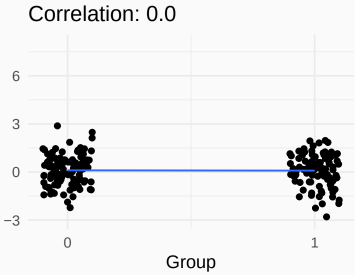
- Answers the question "*How many standard deviations apart are the groups?*"
- The difference in sarcasm score between frequentists and Bayesians is $d = 0.54$ standard deviations.

Correlation

Biserial Correlation

Correlation answers the question “How strong is the relationship between group identity and the outcome?”

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

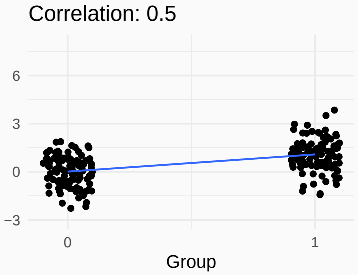


Correlation

Biserial Correlation

Correlation answers the question “How strong is the relationship between group identity and the outcome?”

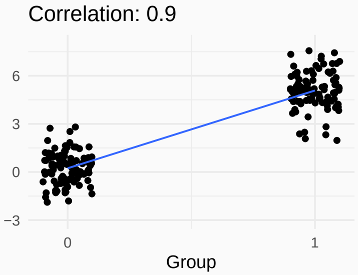
$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$



Biserial Correlation

Correlation answers the question "How strong is the relationship between group identity and the outcome?"

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$



Practical Significance is about Context

- How strong is the same relationship between *different* groups?
- How strong is a *different* relationship between the same group?
- What is the underlying dispersion in the data?
- What is a meaningful anchor or reference point that you can use for context?

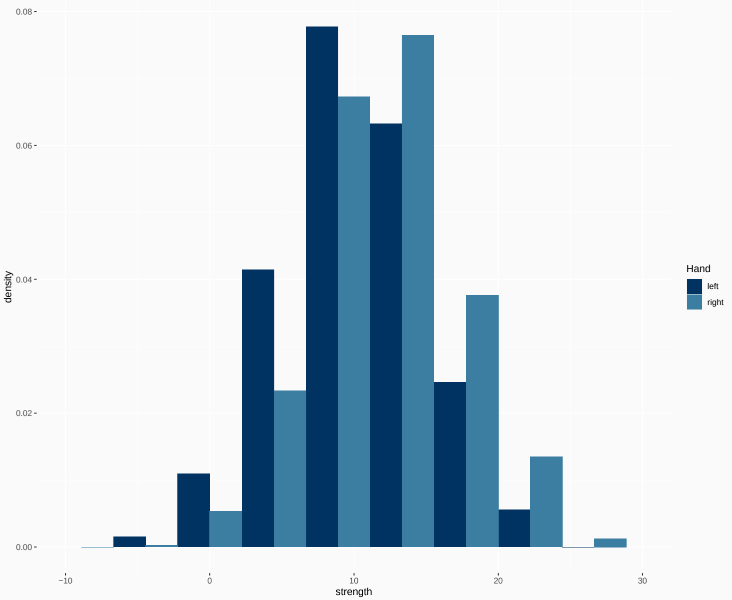
The Paired t-Test

Climbing grip

Suppose you randomly sample 30 Berkeley students. For each student i, you measure right-hand strength (R_i) and left-hand strength (L_i).

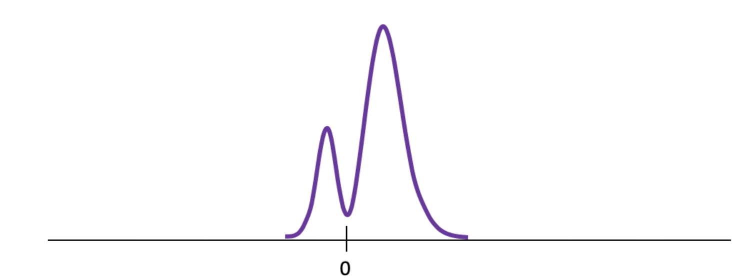
- You conduct a t-test with $H_0 : \mathrm{E}[R] = \mathrm{E}[L]$
- **Problem**: Grip strength varies a lot person-to-person, $\Rightarrow$ t-test has low power.

Paired t-test

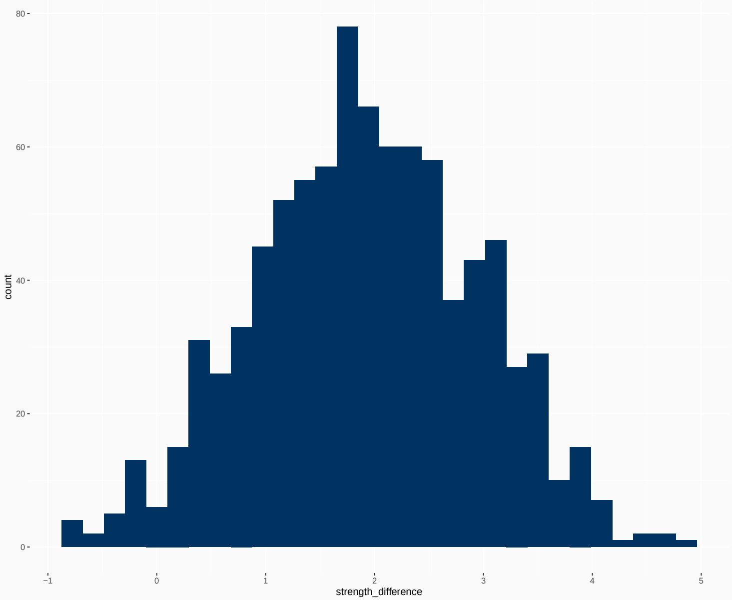


Paired t-Test

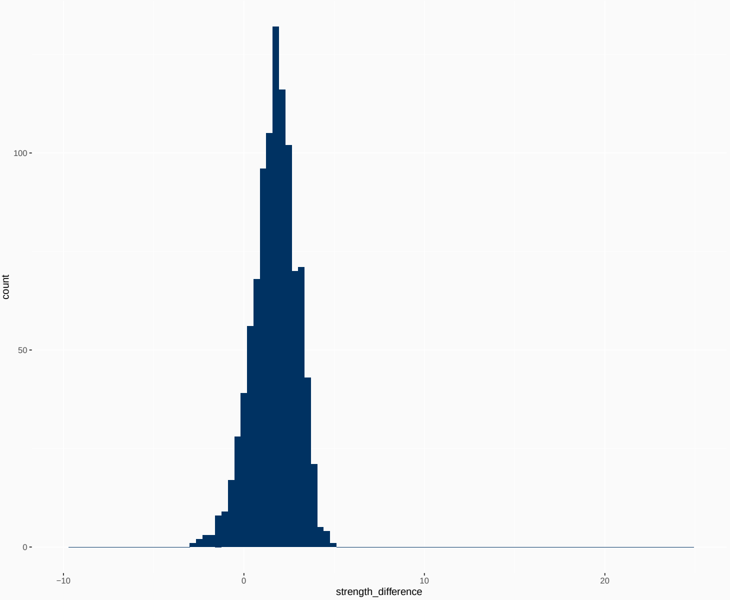
- **Idea:** For any *particular* subject i, the difference between right-hand strength and left-hand stregth, $R_i - L_i$, will usually be small.
- Within-person variation is small.



Paired t-Test



Paired t-Test



Paired t-Test

Paired t-test

A *paired t-test*, sometimes called a *dependent t-test*, builds an explicit dependency between data. Instead, perform a one-sample t-test with $H_0 : \mathrm{E}[R_i - L_i] = 0$.

- This dependency must actually exist
- Cannot simply change the test

Unpaired vs. Paired t-Test

Unpaired

- $t = \frac{\overline{A} - \overline{B}}{\sigma_{A\&B}}$

Paired

- $t = \frac{\overline{A} - \overline{B}}{\sigma_{(A-B)}}$

Paired t-Test Assumptions

- A and B have a metric scale with the same units.
- There is a natural pairing between observations for A and for B.
    - pre-test and post-test for same individual
    - response to two types of stimulus for same mouse
    - responses for a pair of spouses
- Each pair (A_i, B_i) is drawn i.i.d.
- The distribution of $A - B$ is sufficiently normal given the sample size.

Introduction to Non-parametric Tests

Non-parametric Tests

- t-test is parametric, like all the tests we've seen so far
    - Assumes the population comes from a parametric family of distributions
    - Typically the normal curves
- It is not always possible to meet this assumption

Non-parametric Tests (cont.)

Large sample

- No Problem
- central limit theorem tells us that the sampling distribution of the mean will be approximately normal, so t-tests are valid
- Parametric tests are generally valid for large samples

Non-parametric Tests (cont.)

Small sample

- *t*-test is fairly robust to deviations from normality, but you should look at your distribution and see how non-normal it is
- Suppose you have a small sample and you suspect you have a major deviation from normality
- You might be able to transform the variable to make it more normal, but that can alter the meaning and make results harder to interpret

An alternative is to use a *non-parametric* test

Non-parametric Test Details

- Non-parametric tests can be also called **distribution- free tests**
    - Still involve assumptions, but they are less restrictive than those of parametric tests
- Many tests work on principle of ranking data
    - List the scores from lowest to highest – each score gets a rank, so higher scores have higher ranks
    - Only consider ranks instead of looking at the metric value of the variable
    - Use the order of variables to construct statistics that we can use to test hypotheses

Non-parametric Test Details (cont.)

Advantages

- Population distribution doesn't have to be normal
- Easier to justify a rank-based test

Disadvantages

- We throw out metric information
- Rule of thumb: if you throw away information, you lose statistical pwoer

Rank-Based Tests for Ordinal Variables

- Rank-based tests are especially useful when we have an ordinal variable
    - eg. a Likert variable such as "how do you feel about a presidential campaign?"
    - Neutral, support, strongly support, etc.
- It is hard to argue that the difference between neutral and support is the same as the difference between support and strongly support

Love Tester Example

Do you trust that the difference between harmless and mild is the same as the difference between burning and passionate?

Rank-Based Tests for Ordinal Variables (cont.)

If you run a *t*-test in these cases, you impose a linear structure on your variable, treating it as metric

- This method may or may not be reasonable
- If you use a rank-based test that is okay–you are asking whether one group tends to rank below or above another
- The ranks are still meaningful

Conclusion

- There are some situations in which you should consider non-parametric tests
- Coye is going to tell you more about the specifics

Wilcoxon Rank-sum Test for Independent Groups

Parametric and Non-parametric Tests for Comparing Only Two Groups

Type of Design	**Parametric Tests**	**Non-parametric Tests**
Two independent samples	Independent samples t test	Wilcoxon rank-sum test (Mann-Whitney test)
Two dependent Samples	Dependent samples t test	Wilcoxon signed-rank test

Comparing Two Independent Conditions: Wilcoxon Rank-Sum Test

- Data are ranked from lowest to highest across groups
- This provides **potential rank** scores
- If the same score occurs more than once then all scores of the same value receive the average of the potential ranks for those scores

ID	Group	Score	Potential Rank	Final Rank
1	A	10	1	1
2	A	11	2	2.5
3	B	11	3	2.5
4	B	12	4	4
5	A	20	5	6
6	B	20	6	6
7	B	20	7	6
8	A	33	8	8

- This gives us the **final rank** scores

Comparing Two Independent Conditions: Wilcoxon Rank-Sum Test

ID	Group	Score	Potential Rank	Final Rank
1	A	10	1	1
2	A	11	2	2.5
3	B	11	3	2.5
4	B	12	4	4
5	A	20	5	6
6	B	20	6	6
7	B	20	7	6
8	A	33	8	8

Calculating the Wilcoxon Rand-Sum Test

- After assigning final ranks, add up all the final ranks for each of the two groups
- Subtract the mean rank for a group of the same size as our groups
    - Otherwise, larger groups would always have larger values
    - For example, the mean group for a group of four = 1 + 2 + 3 + 4 = 10
- Our final calculation in therefore:
    - W = sum of ranks - mean rank

Calculating the Wilcoxon Rank-Sum Test (cont.)

ID	Group	Score	Potential Rank	Final Rank
1	A	10	1	1
2	A	11	2	2.5
3	B	11	3	2.5
4	B	12	4	4
5	A	20	5	6
6	B	20	6	6
7	B	20	7	6
8	A	33	8	8

- Group A: W = sum of ranks (17.5) - mean rank (10) = 7.5

Interpretation of the Wilcoxon Rank-Sum Test

Default is a two-sided test, like a t test

Null hypothesis: There is no difference in ranks
Alternative hypothesis: There is a difference in ranks

- You can also do a one-directional test if you hypothesize that one particular group will have higher ranks than the other
- Always two values for W (one for each group)
- Lowest score for W is typically used as the test statistic

Interpretation of the Wilcoxon Rank-Sum Test (cont.)

- For small sample sizes ($N < 40$), R calculates the *p* value with the Monte Carlo methods
    - ie. simulated data are used to estimate the statistic
- For larger samples, R calculates the *p* value with a normal approximation method
    - Assumes that the sampling distribution of the *W* statistic is normal, not the data
    - Normal approximation method helpful because it calculates a z statistic in the process of calculating the *p* value

Effect Size for the Wilcoxon Rank-Sum Test

Effect Size Correlation

$$r = \frac{Z}{\sqrt{N}}$$

Divide the z statistic by the square root of the total sample size

r	Effect Size
0.10	Small
0.30	Medium
0.50	Large