PSY 501: Inferential Statistics

Week 12

Outline

Outline

Inferential Statistics

Purpose: to make claims about populations based on data collected from samples

Example experiment:

- Participants:
  - Group A – gets treatment to improve memory
  - Group B – gets no treatment (control)
- After treatment period, test both groups for memory
- Results:
  - Group A: memory score of 80%
  - Group B: memory score of 76%
- Is the 4% difference a "real" difference (statistically significant), or is this just a *sampling error*?

How to test hypotheses

Steps:

1. State your hypotheses
2. Set your decision criteria
3. Collect data from your sample(s)
4. Compute some test statistics
5. Make a decision about your null hypotheses
   - Reject H_0
   - Fail to reject H_0

Testing hypotheses

Step 1: State your hypotheses

- Null hypothesis (H_0)
  - "There are no differences between groups"
  - This is the hypothesis that you are testing!
- Alternative hypothesis
  - "Not all groups are equal"
  - Regardless of what it seems like you are doing, you are NOT trying to prove the alternative hypothesis
  - If you reject the null hypothesis, then you're left with support for the alternative(s) (NOT proof!)

Testing hypotheses

Step 1: State your hypotheses

In our example memory experiment:

- Null (H_0): Mean of group A $=$ Mean of group B
- Alternative (H_A): Mean of group A $\neq$ Mean of group B

Based on previous work, we predict that the treatment should improve memory; that is, Group A $>$ Group B. This is the alternative hypothesis. This is NOT the one we'll test with inferential stats. Rather, we'll test H_0.

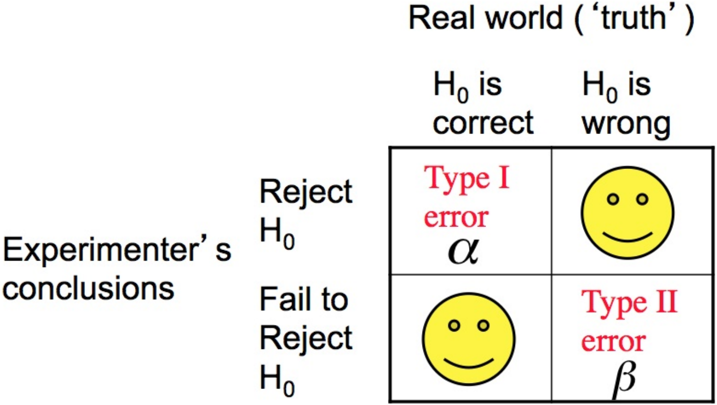
Testing Hypotheses

Step 1: State your hypotheses

Step 2: Set your decision criteria

- Your alpha level will be your guide for when to
  - "reject the null hypothesis"
  - "fail to reject the null hypothesis"
- This could be the correct conclusion or the incorrect conclusion. There are TWO different ways to go wrong here:
  - *Type I Error*
  - *Type II Error*

Error types


Real world ('truth')
H_0 is correct
H_0 is wrong
Experimenter's conclusions
Reject H_0
Fail to Reject H_0
Type I error
α
Type II error
β

Error types

Type I Error (α): Concluding that there is an effect when there really isn't

- The chance of doing this is called the "significance level"
- Goal: minimize this possibility
- Pick low level of α (0.05 and 0.01 most common in psychology)

Type II Error (β): Concluding that there isn't an effect when there really is

- Minimize this by maximizing statistical power $(1 - \beta)$
- Statistical power $=$ how likely you are able to detect an effect if it is really there

Testing Hypotheses

Step 1: State your hypotheses

Step 2: Set your decision criteria

Step 3: Collect data

Step 4: Compute your test statistics

- Descriptive stats (means, SD, etc.)
- Inferential stats (t-tests, ANOVAs, etc.)

Step 5: Make a decision about your null hypothesis

- Reject H_0: "statistically significant differences"
- Fail to reject H_0: "no statistically significant differences"

Outline

Testing Hypotheses

Statistical significance

Constructing a statistical test

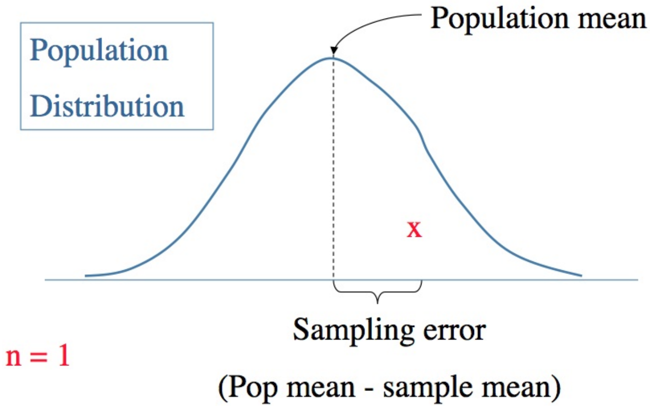
Tests you should know

So what is "statistical significance"?

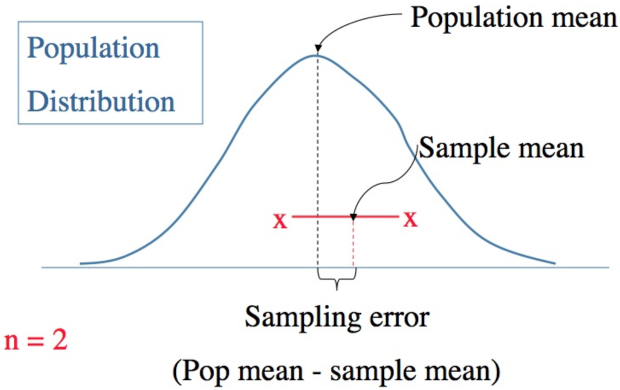
When you "reject your null hypothesis",

- This means that the observed difference is above what you'd expect by <u>chance alone</u>
- <u>Chance</u> is determined by estimating how much *sampling error* there is
- Factors affecting "chance":
  - Sample size
  - Population variability

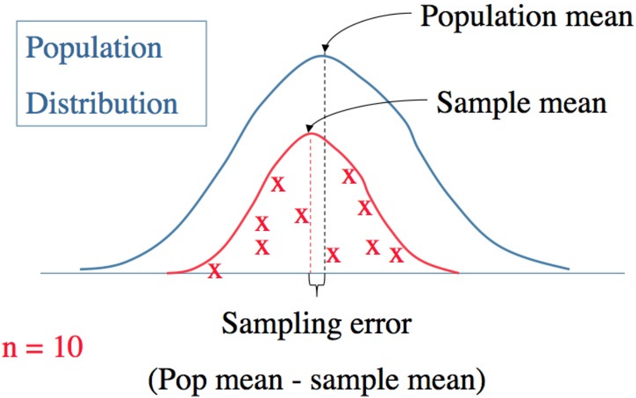
Sampling error



Sampling error

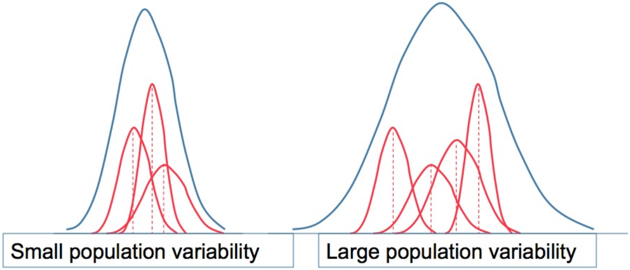


Sampling error



Typically, the larger the sample size, the smaller the sampling error

Sampling error

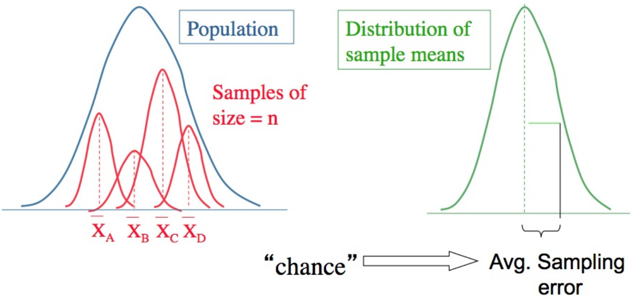


Typically, the narrower the population distribution, the narrower the range of possible samples, and hence the smaller the sampling error

Sampling error

These two factors combine to impact the *distribution of sample means*.

- That is, the distribution of all possible sample means of a particular sample size that can be drawn from the population



Significance

A "statistically significant difference" means:

- the researcher is concluding that there is a difference above and beyond chance
- there is a probability of 5% that the researcher has made a type I error

Note: "statistical significance" is not the same thing as theoretical significance

- statistical difference does not imply an important difference!

Non-significance

Failing to reject the null hypothesis

- Remember, you cannot accept a null hypothesis (unless you're Bayesian, but that's another course)
- Usually, you should check to see if you made a Type II error (failed to detect a difference that really is there)
  - Check the statistical power of your test
    - Sample size is too small
    - Effects that you're looking for are really small
  - Check your controls; maybe too much variability

Outline

Memory example

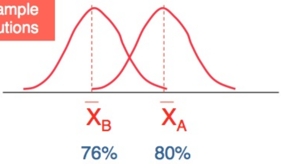
- Example Experiment:
  - Group A - gets treatment to improve memory
  - Group B - gets no treatment (control)
- After treatment period test both groups for memory
- **Results:**
  - Group A's average memory score is 80%
  - Group B's is 76%
- Is the 4% difference a "real" difference (*statistically significant*) or is it just *sampling error*?

About populations

$H_0: \mu_A = \mu_B$

H_0: there is no difference between Grp A and Grp B

Two sample distributions

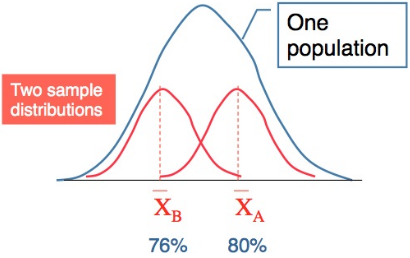


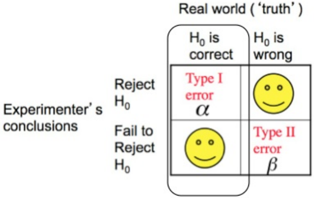
		Real world ('truth')	
		H_0 is correct	H_0 is wrong
Experimenter's conclusions	Reject H_0	Type I error α	
	Fail to Reject H_0		Type II error β

Memory example

Two possibilities in the "real world":

- H_0 is true (no treatment effect)

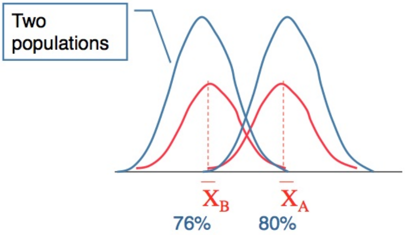


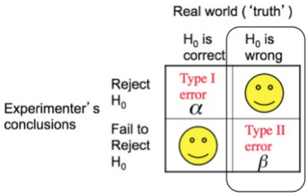


Memory example

Two possibilities in the "real world":

- H_0 is false (there is a treatment effect)



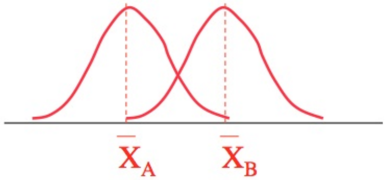


Note: people who get the treatment form a new population (the "treatment population")

A “generic” statistical test

Why might the samples be different?

- Need to analyze the different sources of variability between groups

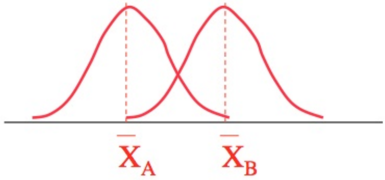


- ER: Random sampling error
- ID: Individual differences
- TR: The effect of a treatment

A “generic” statistical test

A generic test statistic (ratio of sources of variability):

$$\text{Computed test statistic} = \frac{\text{Observed difference}}{\text{Difference from chance}} = \frac{\color{blue}{\text{TR}} + \text{ID} + \text{ER}}{\text{ID} + \text{ER}}$$

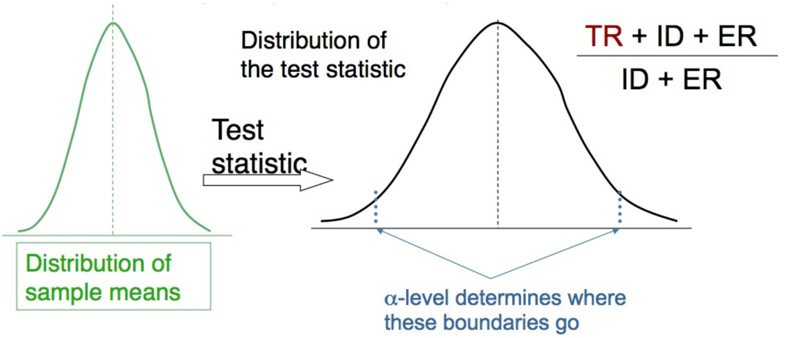


- ER: Random sampling error
- ID: Individual differences
- TR: The effect of a treatment

A "generic" statistical test

To reject H_0, you want a computed test statistic that is "large enough"

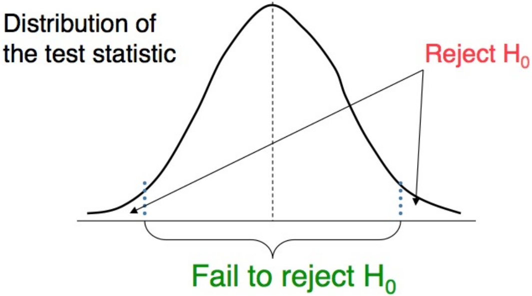
- reflects a large Treatment Effect (TR)



A "generic" statistical test

To reject H_0, you want a computed test statistic that is "large enough"

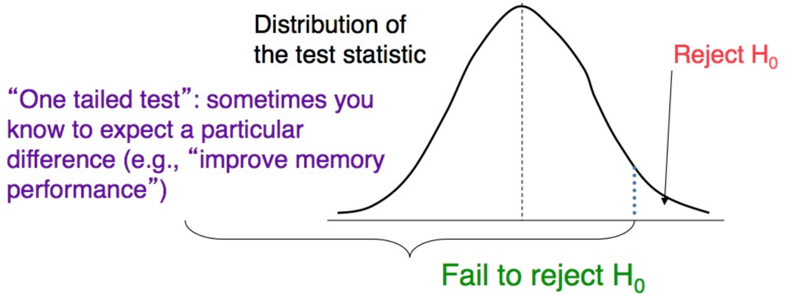
- reflects a large Treatment Effect (TR)



A "generic" statistical test

To reject H_0, you want a computed test statistic that is "large enough"

- reflects a large Treatment Effect (TR)



A “generic” statistical test

Things that affect the computed test statistic

- Size of the treatment effect
  - The bigger the effect, the bigger the test statistic
- Difference expected by chance (sampling error)
  - Sample size
  - Variability in the population

Outline

Testing Hypotheses

Statistical significance

Constructing a statistical test

Tests you should know

Some inferential statistical tests

1 factor with two groups

- T-tests
  - Between-subjects design: Independent-samples t-test
  - Within-subjects design: Paired-samples t-test

1 factor with more than two groups

- Analysis of Variance (ANOVA) (either between-subjects or repeated measures)

Multiple factors

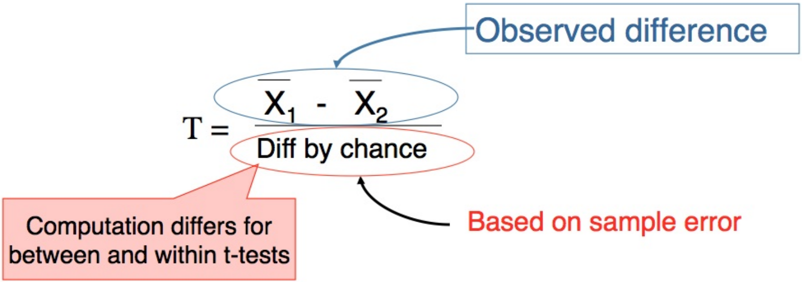
- Factorial ANOVA

T-tests

Design

- 2 separate experimental conditions
- Depends on "degrees of freedom"
  - Based on the size of the sample and the kind of t-test

Formula:



T-tests

Reporting your results (minimal):

1. The observed difference between conditions
2. Kind of t-test
3. Computed T-statistic
4. Degrees of freedom for the test
5. The "p-value" for the test

Examples:

- "The mean of the treatment group was 12 points higher than the control group. An independent samples t-test yielded a significant difference, $t(24) = 5.67$, $p < 0.05$"

T-tests

Reporting your results (minimal):

1. The observed difference between conditions
2. Kind of t-test
3. Computed T-statistic
4. Degrees of freedom for the test
5. The "p-value" for the test

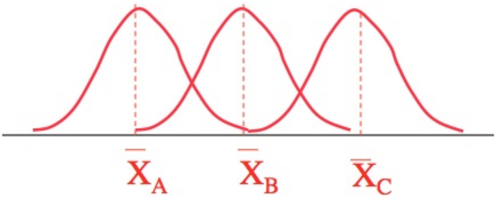
Examples:

- "The mean score of the post test was 12 points higher than the pretest. A paired-samples t-test demonstrated that this difference was statistically significant, $t(24) = 5.67$, $p < 0.05$"

Analysis of Variance

More than two groups

- 1 factor ANOVA
- Factorial ANOVA
- Within- and Between-subjects factors are possible
- Sometimes called GLM (General Linear Model)
- Test statistic is an **F-ratio**
- Degrees of freedom
  - Several to keep track of
  - Number depends on the design

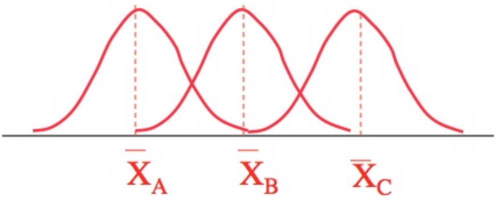


Analysis of Variance

More than two groups

- Now we can't just compute a simple difference score since there is more than one difference
- Instead, we analyze the "variance" among and between conditions

$$\text{F-ratio} = \frac{\text{Observed variance}}{\text{Variance from chance}}$$



Analysis of Variance

Null hypothesis:

H_0: all the groups are equal

$$\overline{X}_A = \overline{X}_B = \overline{X}_C$$

The ANOVA tests this one!!

Alternative hypotheses

H_A: not all the groups are equal

Do further tests to pick between these

$$\overline{X}_A \neq \overline{X}_B \neq \overline{X}_C \qquad \overline{X}_A \neq \overline{X}_B = \overline{X}_C$$

$$\overline{X}_A = \overline{X}_B \neq \overline{X}_C \qquad \overline{X}_A = \overline{X}_C \neq \overline{X}_B$$

Analysis of Variance

Reporting your results (minimal):

1. The observed differences
2. Kind of test
3. Computed F-ratio
4. Degrees of freedom for the test
5. The "p-value" for the test
6. Any post-hoc or planned comparison results

Examples:

- "The mean score Group A was 12, Group B was 25, and Group C was 27. A one-way ANOVA was conducted and the results yielded a significant difference, $F(2, 25) = 5.67$, $p < 0.05$." Post-hoc tests revealed that the differences between groups A and B and A and C were significant (respectively, $t(1) = 5.67$, $p < 0.05$ and $t(1) = 6.02$, $p < 0.05$.) Groups B and C did not differ significantly from each other.

Factorial ANOVAs

Reporting your results:

- The observed differences
  - Because there may be a lot of these, typically presented in a table instead of directly in text.
- Kind of design
  - e.g., "2 x 2 between-subjects design"
- Computed F-ratios
  - May see separate paragraphs for each factor and for interactions
- Degrees of freedom
  - Each F-ratio will have its own set of df's
- The "p-value" of the test
  - More typically, we just say "all tests were conducted with an alpha level of 0.05"
- Post-hoc or planned comparisons
  - Typically only those that are theoretically interesting are presented