Example: Communities with higher population have different amounts of violent crimes (per capita) than those with lower population.

Assignment 1, Programming Problem "C) 9."



$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent, same variance) t statistic for 2 samples

Degrees of Freedom: the number of values that are free to vary

The number of observations available to measure a parameter in a distribution. In other words, what is the minimum *i*, such that given *i* observations one could determine the parameter?

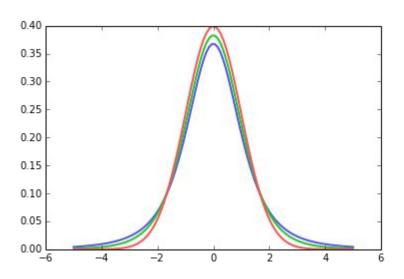
Examples: mean, variance

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent, same variance) t statistic for 2 samples

t-test: comparing means of distributions

Remember, t identifies an x in a distribution (Student's t distribution) P(T < t; df)





$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent, same variance)

t-test: comparing means of distributions

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{(s_1^2/n_1)^2}{df_1} + \frac{(s_1^2/n_1)^2}{df_2}}$$
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(assuming independent, different variance)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent, same variance)

t-test: comparing means of distributions

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{(s_1^2/n_1)^2}{df_1} + \frac{(s_1^2/n_1)^2}{df_2}}$$
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(assuming independent, different variance)

$$t = \frac{\bar{X}_1 - \mu_0}{\frac{s_1}{\sqrt{n_1}}}$$

(compared to theoretical mean)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent, same variance)

Important logical question:

Does failure to reject the null mean the null is true?



Important logical question:

Does failure to reject the null mean the null is true?

no.

Important logical question:

Does failure to reject the null mean the null is true?

no.

Traditionally, one of the most common reasons to fail to reject the null: n is too small (not enough data)

Thought experiment: If we have infinite data, can the null ever be true?

Important logical question:

Does failure to reject the null mean the null is true?

no.

Traditionally, one of the most common reasons to fail to reject the null: n is too small (not enough data)

Thought experiment: If we have infinite data, can the null ever be true?

Big Data problem: "everything" is significant. Thus, consider "effect size"

Type I, Type II

		True state of nature	
		H_0	H_A
Our	Reject H_0	Type I error	correct decision
decision	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Multi-test Correction

If alpha = .05, and I run 40 variables through significance tests, just by chance, how many are likely to be significant?



The Scientific Method

Develop General Theories

General theories must be consistent with most or all available data and with other current theories.

Gather Data to Test Predictions

Relevant data can come from the literature, new observations or formal experiments. Thorough testing requires replication to verify results.

Make Observations

What do I see in nature?
This can be from one's own experiences, thoughts or reading.

Refine, Alter, Expand or Reject Hypotheses

Develop Testable
Predictions

If my hypothesis is correct, then I expect a, b, c, ...

Think of Interesting Questions

Why does that pattern occur?

Formulate Hypotheses

What are the general causes of the phenomenon I am wondering about?

The Scientific Method

Which steps are most subjective?

Develop General Theories

General theories must be consistent with most or all available data and with other current theories.

Gather Data to Test Predictions

Relevant data can come from the literature, new observations or formal experiments. Thorough testing requires replication to verify results.

Make Observations

What do I see in nature?
This can be from one's own experiences, thoughts or reading.

Refine, Alter, Expand or Reject Hypotheses

Develop Testable
Predictions

If my hypothesis is correct, then I expect a, b, c, ...

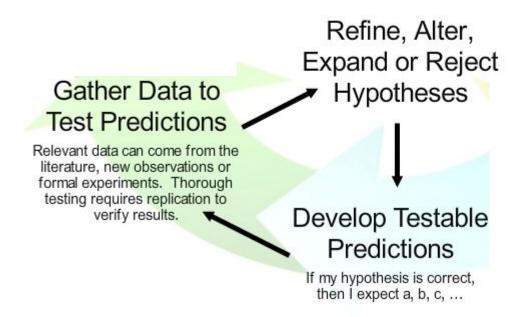
Think of Interesting Questions

Why does that pattern occur?

Formulate Hypotheses

What are the general causes of the phenomenon I am wondering about?

The Scientific Method Potential Effects of Big Data



Resampling Techniques

"nonparametric" tests

The permutation test:

- t_{obs} = Compute observed score
- passes = 0
- for 1 to *B*:
 - o randomly permute the data, keeping the same sizes per class
 - \circ t_{B} = compute score on permuted data
 - \circ if $t_B > (or <) t_{obs}$: passes+=1
- p_value = passes/B

Application: comparing two distributions, especially when they are unknown.