Hypothesis Testing

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

Assignment 1, Programming Problem “C) 9.”

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

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

Hypothesis Testing

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

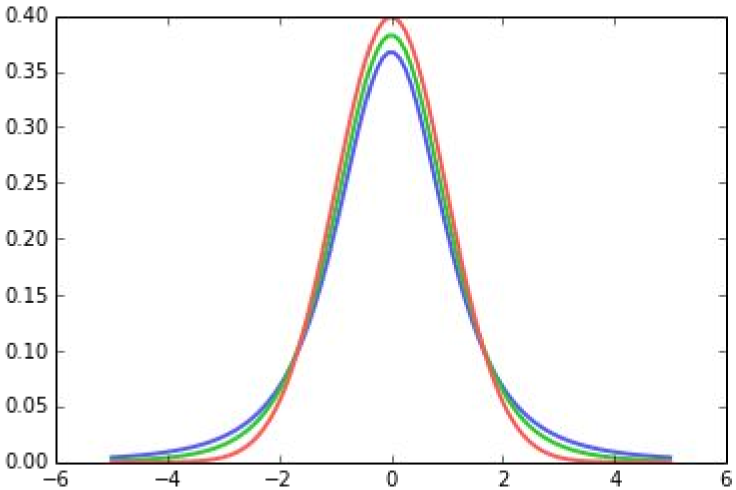
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1 X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

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

Hypothesis Testing

t-test: comparing means of distributions

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



$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)

Hypothesis Testing

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)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent, same variance)

Hypothesis Testing

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)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{X_1X_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$$

(assuming independent,
same variance)

Hypothesis Testing

Important logical question:

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

Hypothesis Testing

Important logical question:

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

no.

Hypothesis Testing

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?

Hypothesis Testing

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

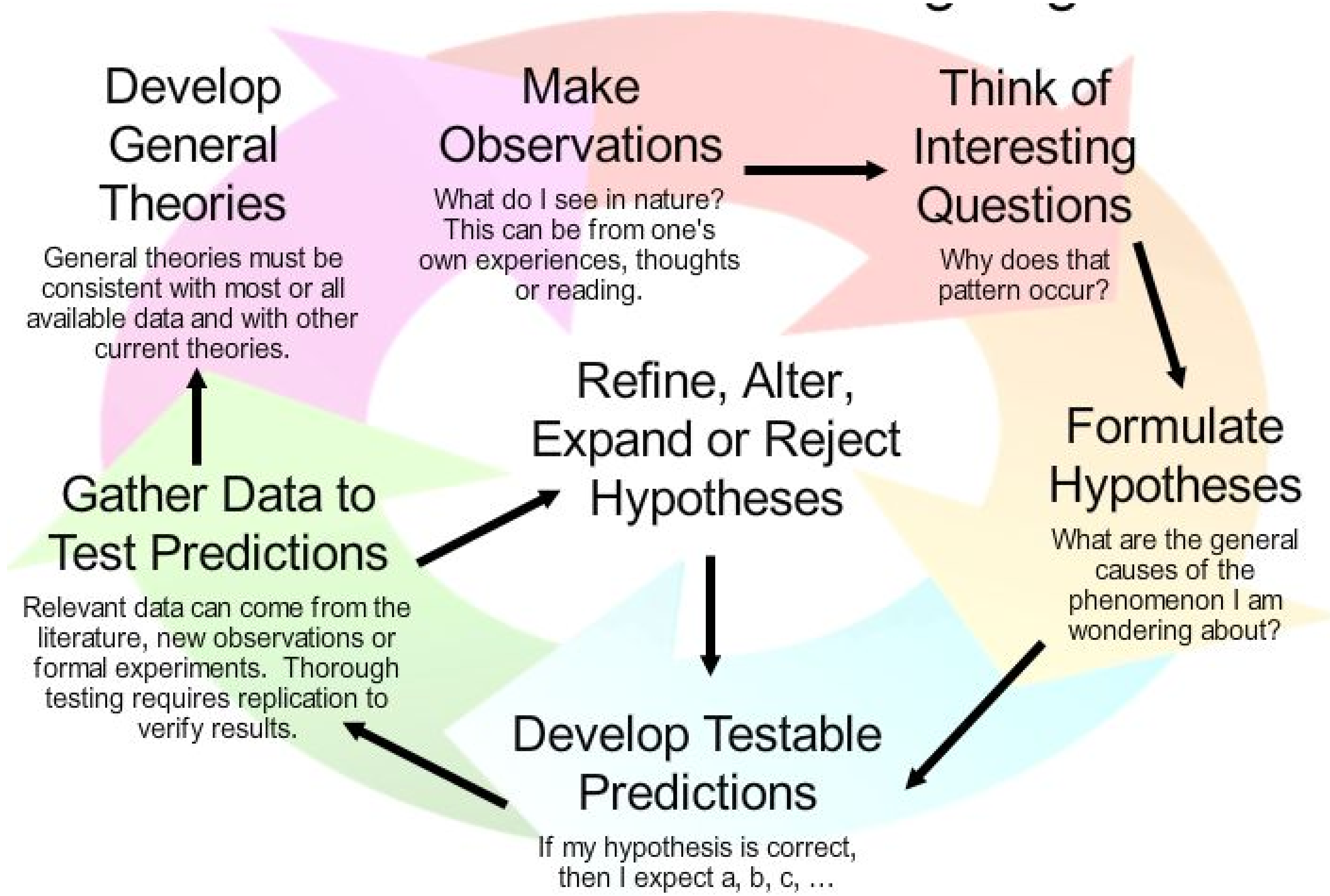
		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Multi-test Correction

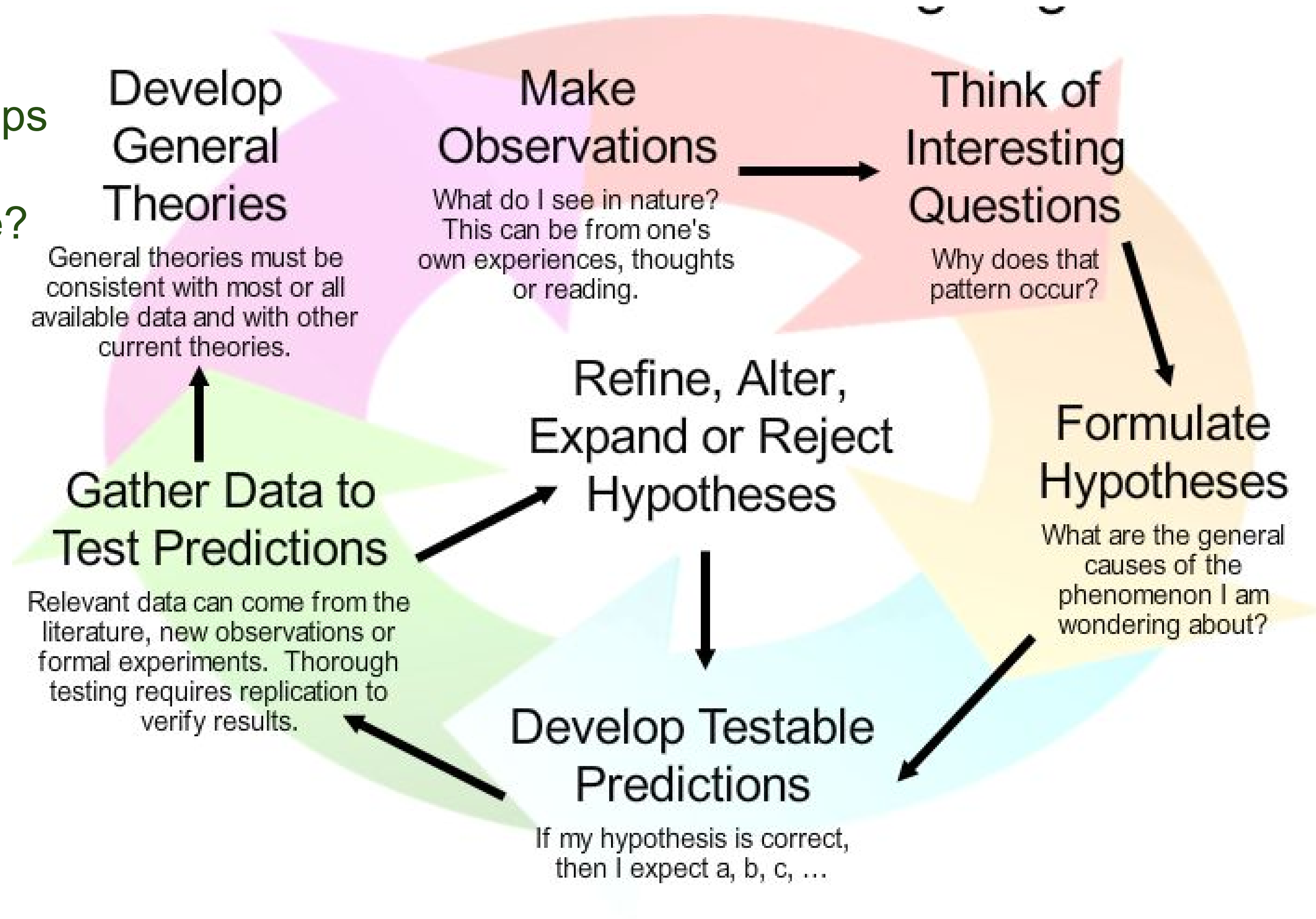
If alpha = .05, and I run 40 variables through significance tests, just by chance, how many are likely to be significant?

The Scientific Method



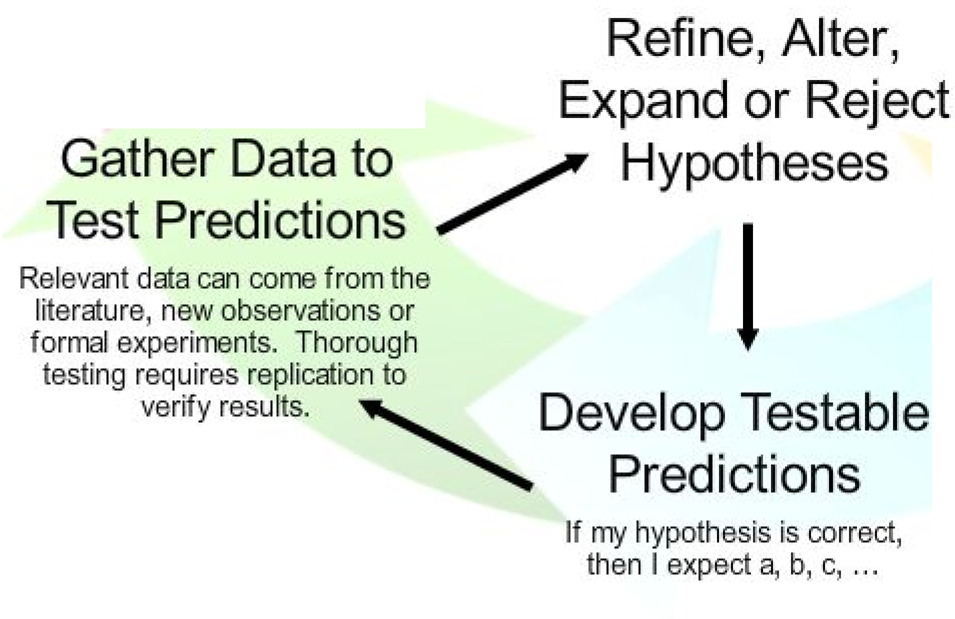
The Scientific Method

Which steps are most subjective?



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:*
  - randomly permute the data, keeping the same sizes per class
  - t_B = compute score on permuted data
  - if t_B > (or <) t_{obs}: passes+=1
- p_value = passes/*B*

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