Review

Collin Nolte

May 3, 2022

Course Review

Big picture stuff

Statistics? Data?

Distributions

Statistics we have known and loved

Data Reduction

Reducing entirety of dataset to meaningful summaries

Measures of Centrality:

- Mean
- Median
- Skew

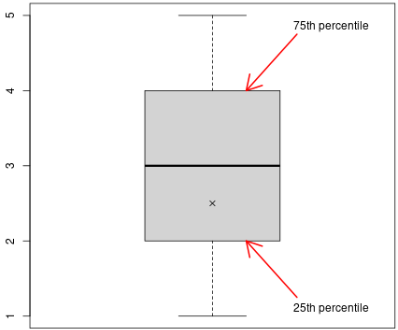
Measures of Dispersion:

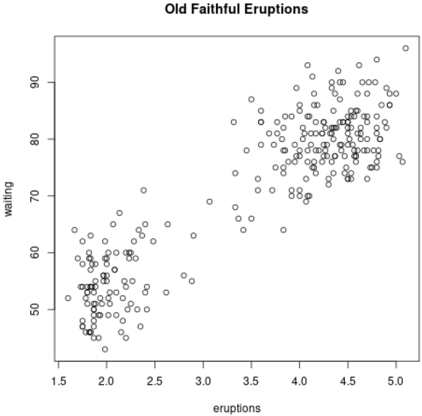
- Variance/standard deviation
- Quantiles
- IQR

Measures of Association:

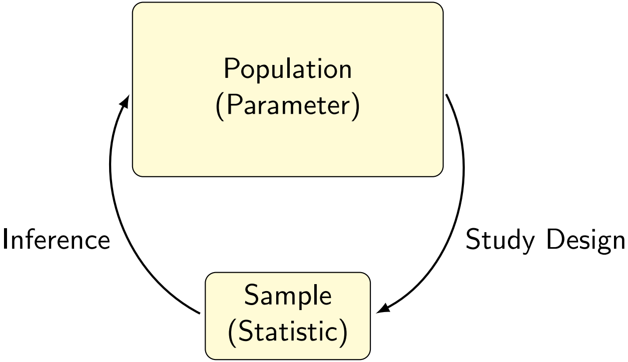
- Correlation (Spearman, Pearson)
- Scatterplots

Data Reduction





Statistical framework



Study Design

Types of Studies:

- Observational
  + Case-control
  + Longitudinal
  + Retrospective
- Clinical

Types of Bias:

- Sample Bias
- Confounding
- Extrapolation

Bias vs Variance

Statistic

A *statistic* is any value that can be computed from a sample

This includes (but is not limited to): mean, median, variance, max/min, ratios, and differences

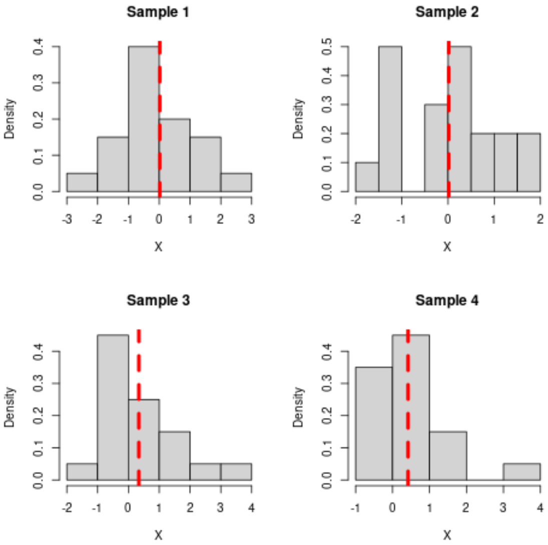
Randomness

A statistic is computed form a sample, which is randomly selected from a population

As one sample may not be identical to another sample, we may assume that the derived statistics are not identical either

If we are only able to collect a single sample, what are we able to say about the statistic derived? Can we find a range of likely values?

Randomness



Sample 1
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
X
Sample 2
Density
0.0
0.1
0.2
0.3
0.4
0.5
-2
-1
0
1
2
X
Sample 3
Density
0.0
0.1
0.2
0.3
0.4
-2
-1
0
1
2
3
4
X
Sample 4
Density
0.0
0.1
0.2
0.3
0.4
-1
0
1
2
3
4
X

Central Limit Theorem

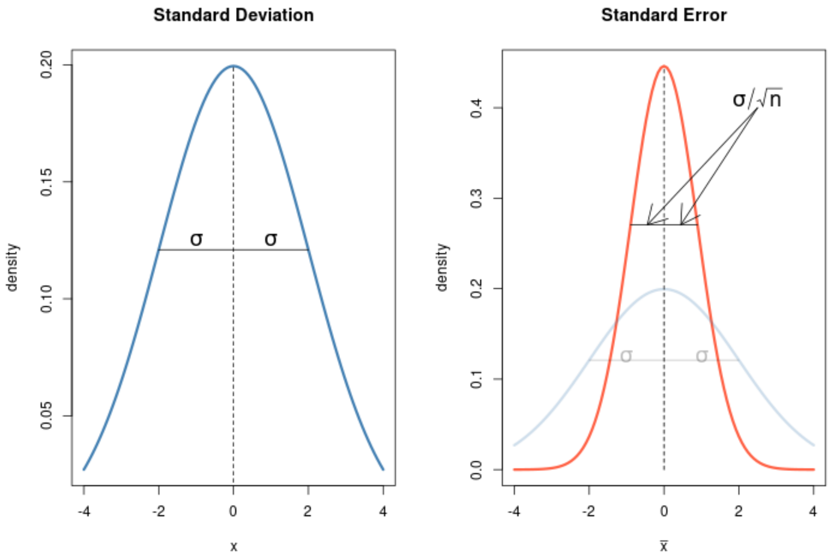
The Central Limit Theorem has been backbone of most of what we have worked with this semester

It states that for a population X with mean μ and variance σ^2, for any sample $\{X_1, \ldots, X_n\}$ of size n, the sample means follows an approximately normal distribution:

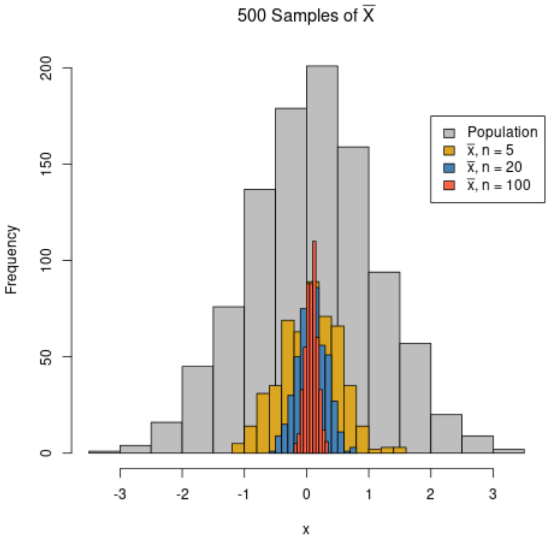
$$\overline{X} \sim N(\mu, \sigma^2/n)$$

1. Does not require that the population be normal (though it helps)
2. Works for statistics beyond the sample mean
3. Larger sample $==$ more normal

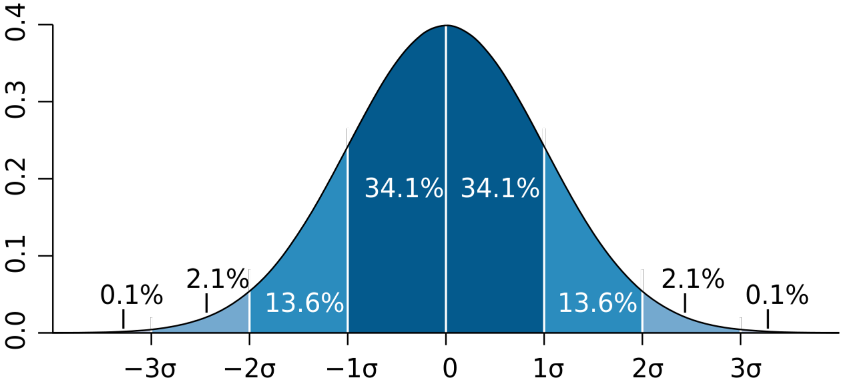
Standard Deviation with n



Sample Mean Distribution



Distribution of Statistic



Basic Hypothesis Testing

The formal process of scientific investigation

1. Define the *null hypothesis* as a declarative, unambiguous statement
2. Collect observational or experimental data
3. Compare the results to what would have been expected based on the null hypothesis (statistical inference)
4. Either *reject* or *fail to reject* the null hypothesis based on the *strength of the evidence*

Basic Hypothesis Testing, cont.

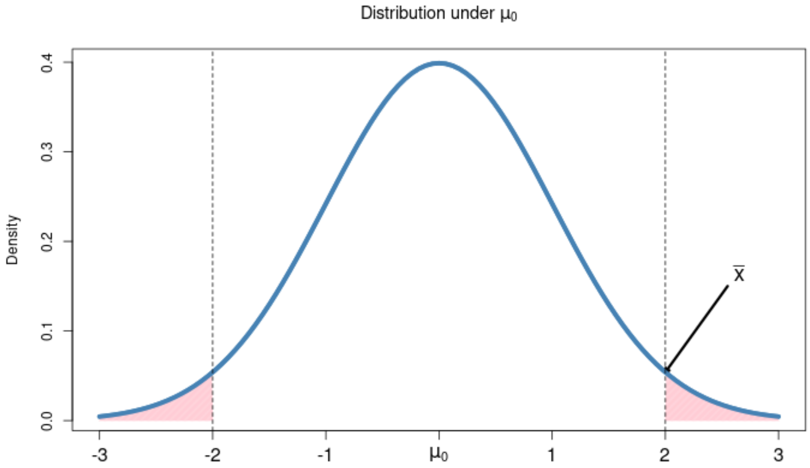
$$H_0 : \mu = \mu_0$$

Given a hypothesis, μ_0, and an observed sample statistic, say, $\overline{x}$, we ask ourselves, "Is this difference due to chance, or is the null hypothesis incorrect?"

Frequently, we reduce this down to a single metric, the *p-value*:

$$p = P(\text{observed data} \mid H_0)$$

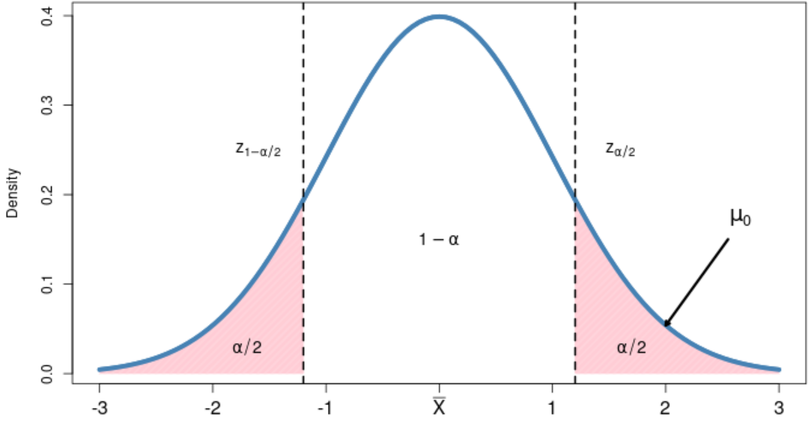
p-values



Distribution under μ_0
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
μ_0
1
2
3
$\bar{X}$

Confidence Intervals

$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$



Testing Errors

In actuality, a null hypothesis is either true or false, and based on the data, we may reject or fail to reject this null. As a consequence, there are two ways in which we might make a mistake.

Test Result	True State of Nature: H_0 True	True State of Nature: H_0 False
Fail to reject H_0	Correct $(1-\alpha)$	Incorrect Type II Error (β)
Reject H_0	Incorrect Type I Error (α)	Correct $(1-\beta)$

- Type I error $= P(\text{Reject } H_0 | H_0 \text{ true}) =$ false alarm
- Type II error $= P(\text{Fail to reject } H_0 | H_A \text{ true}) =$ missed opportunity

Type I Error

If drug A doesn't work, I want to know

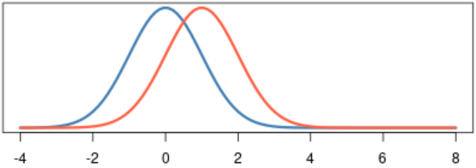
- "If H_0 is true, I want to be reasonably sure"
- "I can be more confident by collecting more evidence"
- "Evidence, in this case, would mean that my observed $\overline{X}$ is extreme, given the null distribution based on H_0"
- "I can set my threshold for how much evidence I would need by my choice of α, the Type I error rate"
- "Smaller values of α indicate that I need stronger evidence. This requires that I have a smaller p-value, with $p < \alpha$"

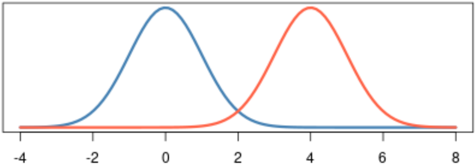
Type II Error

If drug A does work, I want to know

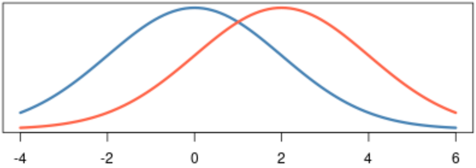
- "If H_0 is false, I want to be sure to reject it"
- "This means I want to be more confident about my estimate of μ"
- "This is difficult to do if there is a lot of variability. I can reduce the amount of variability by increasing my sample size"
- "This can be expensive, though, so I should know how many I need in order to be reasonably sure I have enough. This is called estimating my *power*, $(1-\beta)$"
- "This will also depend on my *effect size*. A larger effect size requires less evidence, while a smaller effect size requires more"

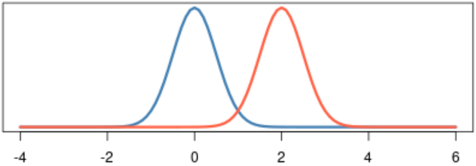
Magnitude/Effect Size



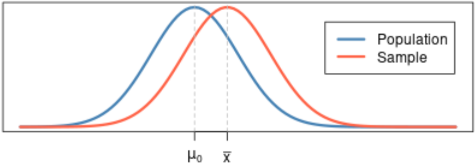


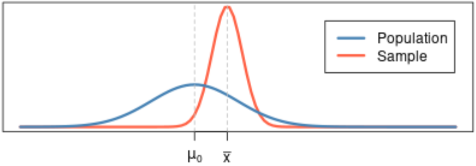
Variability





Sample Size





t-tests

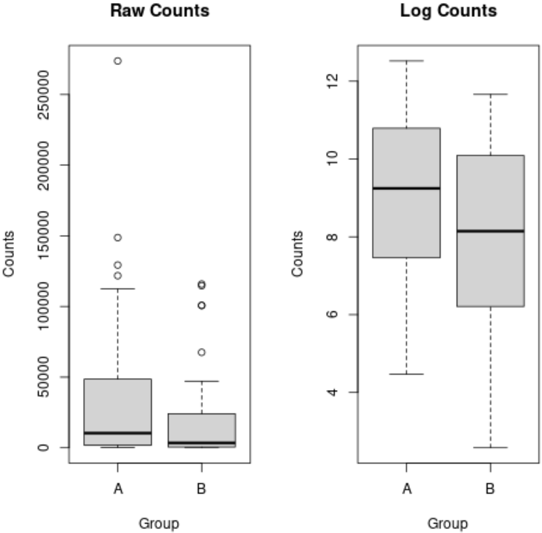
Useful in testing for effect or differences between groups

$H_0 : \mu = \mu_0$ or $H_0 : \mu_A - \mu_B = 0$

May be paired or unpaired

Approximately normal as sample size increases

t-test


Raw Counts
Counts
250000
200000
150000
100000
50000
0
A
B
Group
Log Counts
Counts
12
10
8
6
4
A
B
Group

χ^2 Test

Population	Variable +	Variable −	Total
A	a	b	$a+b$
B	c	d	$c+d$
Total	$a+c$	$b+d$	N

- Independent or homogenous
- *p*-value does *NOT* indicate magnitude of relationship
- Alternatives: Fisher's exact test, binomial test
- Transmission Disequilibrium Test (TDT)

Regression

Describes a linear relationship

$$Y = X_1\beta_1 + X_2\beta_2 + \cdots + X_n\beta_n + \epsilon$$

Where:

- $\epsilon \sim N(0, \sigma^2)$
- β_i describes change in Y given change in X_i, *everything else equal*
- Collinearity
- Less is more

Principal Components

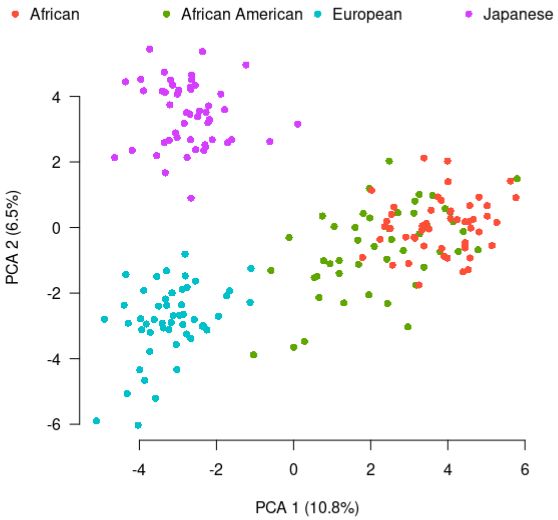
Method of creating p new covariates out of p old covariates (linear combination)

Ordered by amount of variability in the data

New covariates are linearly independent

Can serve as easy form of data reduction (i.e., keeping first two or three)

Principal Components



THE END