Hypothesis Testing

Cohen Chapter 5

EDUC/PSY 6600

"I'm afraid that I rather
give myself away when I explain,"
said he.
"Results without causes
are much more impressive."

-- Sherlock Holmes

The Stock-Broker's Cat

Two Types of Research Questions

Do groups
significantly differ
on 1 or more characteristics?

Comparing group means, counts, or proportions

- t-tests
- ANOVA
- χ^2 tests

Two Types of Research Questions

Do groups *significantly* differ on 1 or more characteristics?

Comparing group means, counts, or proportions

- t-tests
- ANOVA
- χ^2 tests

Is there a *significant* relationship among a set of variables?

Testing the association or dependence

- Correlation
- Regression

Inferential Statistics

Descriptive statistics are limited

- Rely only on **raw** data distribution
- Generally describe **one** variable only
- Do not address **accuracy** of estimators or hypothesis testing
- How **precise** is sample mean or does it differ from a given value?
- Are there between or within **group differences** or **associations**?

Inferential Statistics

Descriptive statistics are limited

- Rely only on **raw** data distribution
- Generally describe **one** variable only
- Do not address **accuracy** of estimators or hypothesis testing
- How **precise** is sample mean or does it differ from a given value?
- Are there between or within **group differences** or **associations**?

Goals of inferential statistics

- Hypothesis testing
  - p-values
- Parameter estimation
  - confidence intervals

Repeated sampling

- Estimators will vary from sample to sample
- Sampling or random error is variability due to chance

Causality and Statistics

Causality depends on **evidence** from outside statistics:

- Phenomenological (educational, behavioral, biological) credibility
- Strength of association, ruling out occurrence by chance alone
- Consistency with past research findings
- Temporality
- Dose-response relationship
- Specificity
- Prevention

Causality and Statistics

Causality depends on **evidence** from outside statistics:

- Phenomenological (educational, behavioral, biological) credibility
- Strength of association, ruling out occurrence by chance alone
- Consistency with past research findings
- Temporality
- Dose-response relationship
- Specificity
- Prevention

Causality is often a **judgmental** evaluation of **combined** results from **several studies**

z-Scores and Statistical Inference

Probabilities of z-scores used to determine how **unlikely** or **unusual** a single case is relative to other cases in a sample

Small probabilities
(p-values)
reflect **unlikely** or **unusual scores**

Not frequently interested in whether **individual scores** are unusual relative to others, but whether scores from **groups of cases** are unusual.

Sample mean, $\bar{x}$ or M, summarizes **central tendency** of a group or sample of subjects

Steps of a Hypothesis test

1. State the Hypotheses
   - Null & Alternative

2. Select the Statistical Test & Significance Level
   - α level
   - One vs. Two tails

3. Select random sample and collect data

4. Find the Region of Rejection
   - Based on α & # of tails

5. Calculate the Test Statistic
   - Examples include: z, t, F, χ^2

6. Write the Conclusion
   - Statistical decision must by in context!

Steps of a Hypothesis test

1. State the Hypotheses
   - Null & Alternative
2. Select the Statistical Test & Significance Level
   - α level
   - One vs. Two tails
3. Select random sample and collect data
4. Find the Region of Rejection
   - Based on α & # of tails
5. Calculate the Test Statistic
   - Examples include: z, t, F, χ^2
6. Write the Conclusion
   - Statistical decision must by in context!

Definition of a p-value:

The probability of observing
a test statistic
as extreme or more extreme
IF
the NULL hypothesis is true.

Stating Hypotheses

Hypotheses are always specified in terms of **population**

- Use μ for the population mean, not $\bar{x}$ which is for a sample

If you are comparing TWO population MEANS:

Null Hypothesis

$$H_0 : \mu_1 = \mu_2$$

Research or Alternative Hypothesis options...

$$H_1 : \mu_1 \neq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

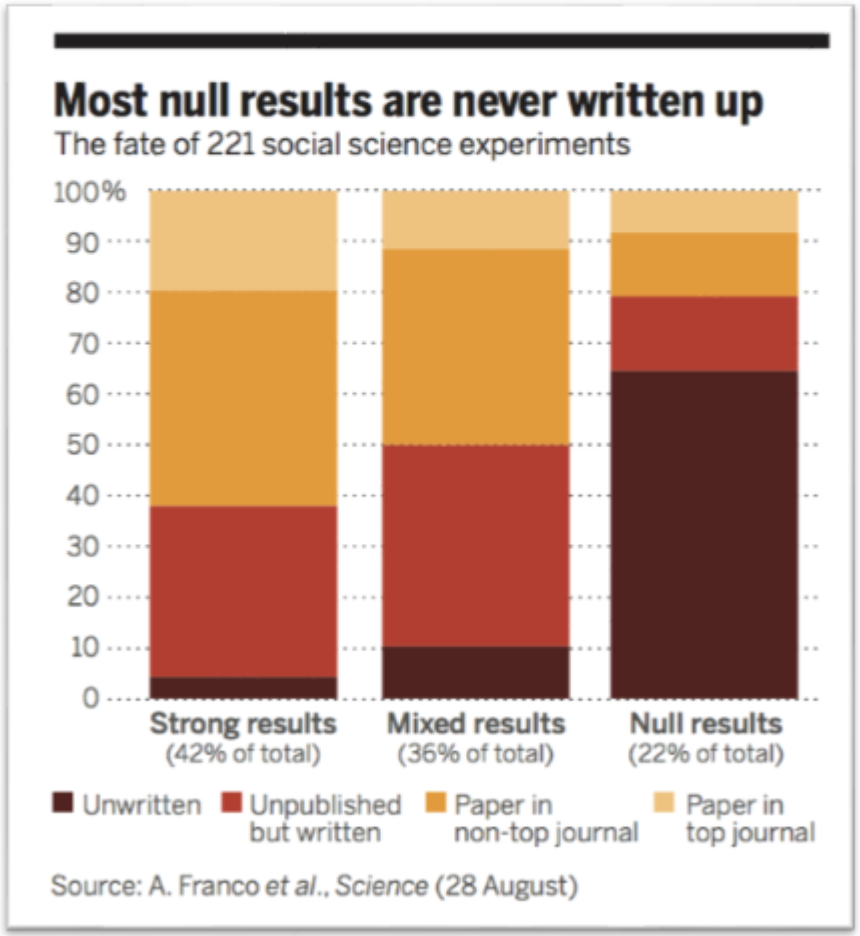
$$H_1 : \mu_1 > \mu_2$$

Innocent Until Proven Guilty

IF there is Not enough statistical evidence to reject

Judgment suspended until further evidence evaluated:

- "Inconclusive"
- Larger sample?
- Insufficient data?



Rejecting the Null Hypothesis

Assumption:

The NULL hypothesis is TRUE in the POPULATION

IF:

The p-value is very **SMALL**

- How small? `(p-value \lt \alpha)`

THEN:

We have evidence AGAINST the NULL hypothesis

- It is **UNLIKELY** we would have observed a sample that extreme **JUST DUE TO RANDOM CHANCE**...

Rejecting the Null Hypothesis

Assumption:

The NULL hypothesis is TRUE in the POPULATION

IF:

The p-value is very **SMALL**

- How small? `(p-value \lt \alpha)`

THEN:

We have evidence AGAINST the NULL hypothesis

- It is **UNLIKELY** we would have observed a sample that extreme **JUST DUE TO RANDOM CHANCE**...

Criteria:

May judge by either...

- the p-value $< \alpha$
  -OR-
- test statistic $<$ Critical Value

Conclusion:

We either **REJECT** or **FAIL TO REJECT** the Null hypothesis

We NEVER **ACCEPT**
the **ALTERNATIVE** hypothesis!!!

ONE tail or TWO?

2-tailed test

$H_1 : \mu_1 \neq \mu_2$

1-tailed test

Suggests a directionality in results!

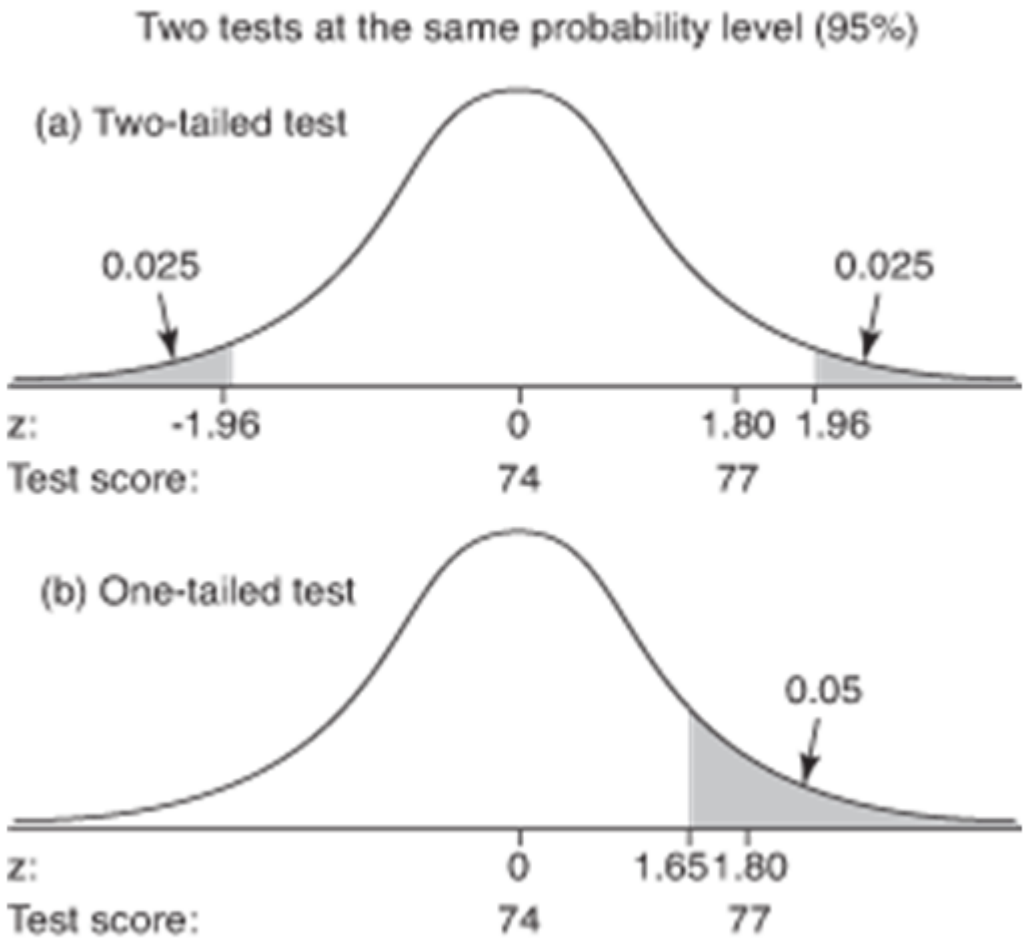
$H_1 : \mu_1 < \mu_2$ -OR- $H_1 : \mu_1 > \mu_2$

NO computational differences

ONLY the $p - value$ differs:

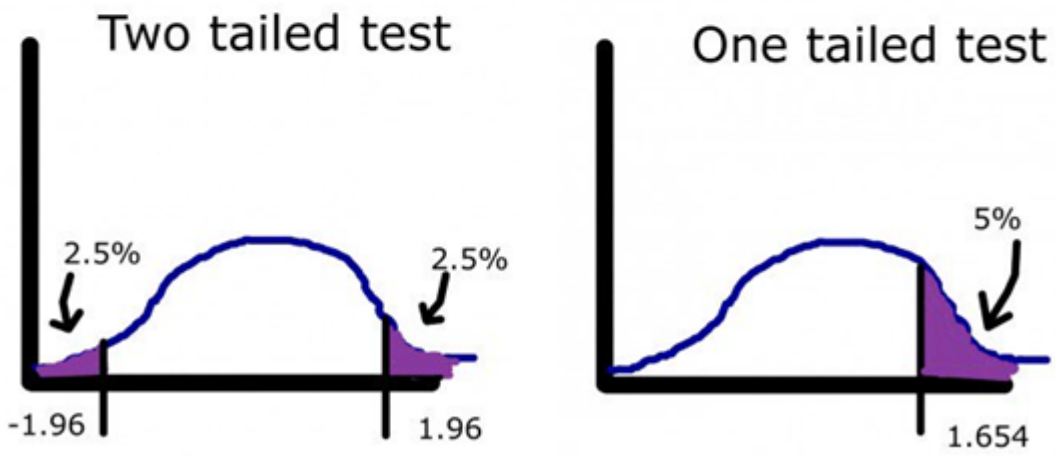
$2\ tail\ p - value = \mathbf{2} \times 1\ tail\ p - value$

- IF: 1-sided: $p = .03$
- THEN: 2-sided: $p = .06$



ONE tail or TWO?

Some circumstances may warrant a 1-tailed test, BUT...
We generally **prefer** and default to a 2-tailed test!!!



More conservative = 2 tails

Rejection region is distributed in both tails

- e.g.: $\alpha = .05$ distributed across both tails
  - (2.5% in each tail)
- If we know outcome, why do study?
  - Looks suspicious to reviewer's?
  - "significant results at all costs!"

Choosing Alpha

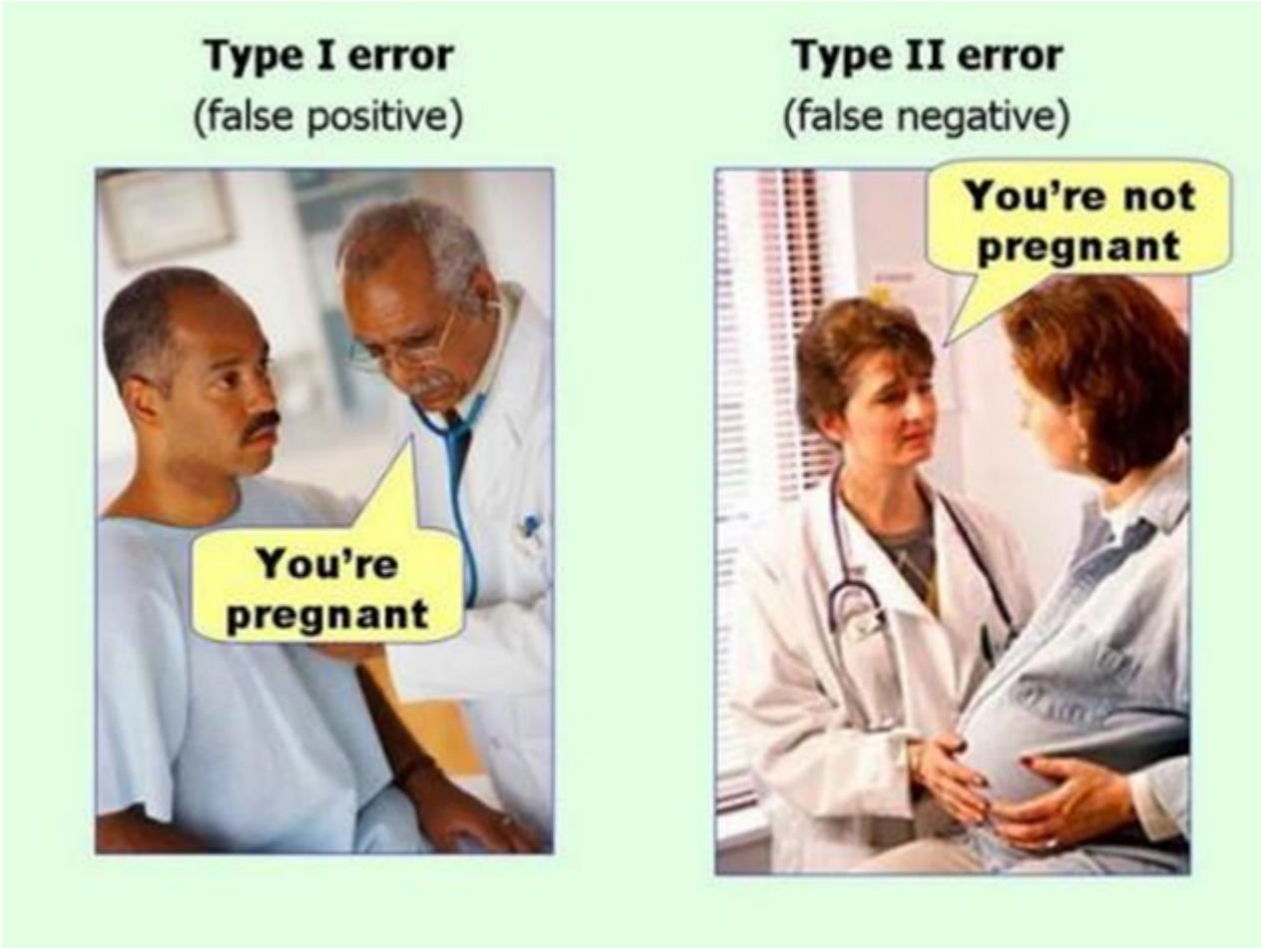
Alpha = probability of making a **type I error**

type I error

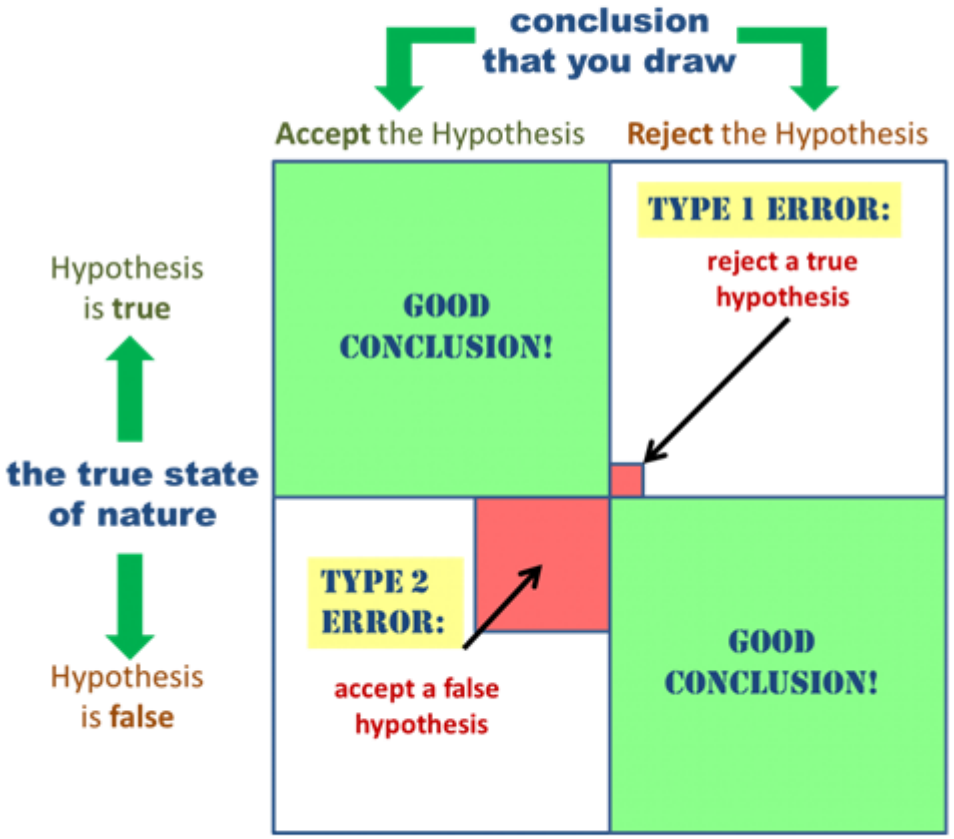
- We reject the NULL when we should not
- The risk of "false positive" results

type II error

- We FAIL to reject the NULL when we should
- The risk of "false negative" results



Choosing Alpha



We want α to be SMALL, but we can't just make too tiny, since the trade off is increasing the type II error rate

DEFAULT is $\alpha = .05$ (5% = 1 in 20 & seems *rare* to humans) **BUT** there is nothing magical about it

Let it be LARGER value, $\alpha = .10$, **IF** we'd rather not miss any potential relationship and are okay with some false positives

- Ex) screening genes, early drug investigation, pilot study

Set it SMALLER, $\alpha = .01$, **IF** false positives are costly and we want to be more stringent

- Ex) changing a national policy, mortgaging the farm

Assumptions of a 1-sample z-test

Sample was drawn at random (at least as representative as possible)

- Nothing can be done to fix NON-representative samples!
- Can not statistically test

Assumptions of a 1-sample z-test

Sample was drawn at random (at least as representative as possible)

- Nothing can be done to fix NON-representative samples!
- Can not statistically test

SD of the sampled population = SD of the comparison population

- Very hard to check
- Can not statistically test

Assumptions of a 1-sample z-test

Sample was drawn at random (at least as representative as possible)

- Nothing can be done to fix NON-representative samples!
- Can not statistically test

SD of the sampled population = SD of the comparison population

- Very hard to check
- Can not statistically test

Variables have a normal distribution

- Not as important if the sample is large (Central Limit Theorem)
- IF the sample is far from normal &/or small n, might want to transform variables
  - Look at plots: histogram, boxplot, & QQ plot (straight 45 degree line)
  - Skewness & Kurtosis: Divided value by its SE & $> \pm 2$ indicates issues
  - Shapiro-Wilks test (small N): p < .05 ??? not normal
  - Kolmogorov-Smirnov test (large N)

APA: results of a 1-sample z-test

- State the alpha & number of tails prior to any results
- Report exact p-values (usually 2 decimal places), except for `p < .001`

APA: results of a 1-sample z-test

- State the alpha & number of tails prior to any results
- Report exact p-values (usually 2 decimal places), except for `p < .001`

Example Sentence:

A one sample z test showed that the difference in the quiz scores between the current sample (N = 9, M = 7.00, SD = 1.23) and the hypothesized value (6.000) were statistically significant, z = 2.45, p = .040.

EXAMPLE: 1-sample z-test

After an earthquake hits their town, a random sample of townspeople yields the following anxiety score:

72, 59, 54, 56, 48, 52, 57, 51, 64, 67

Assume the general population has an anxiety scale that is expressed as a T score, so that $\mu = 50$ and $\sigma = 10$.

EXAMPLE: 1-sample z-test

After an earthquake hits their town, a random sample of townspeople yields the following anxiety score:

72, 59, 54, 56, 48, 52, 57, 51, 64, 67

Assume the general population has an anxiety scale that is expressed as a T score, so that $\mu = 50$ and $\sigma = 10$.

1. Null/Alt Hypotheses

$H_0: \mu = 50$

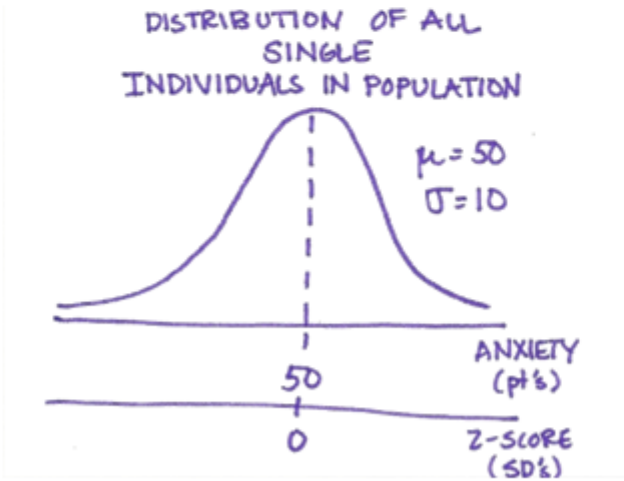
$H_1: \mu \neq 50$

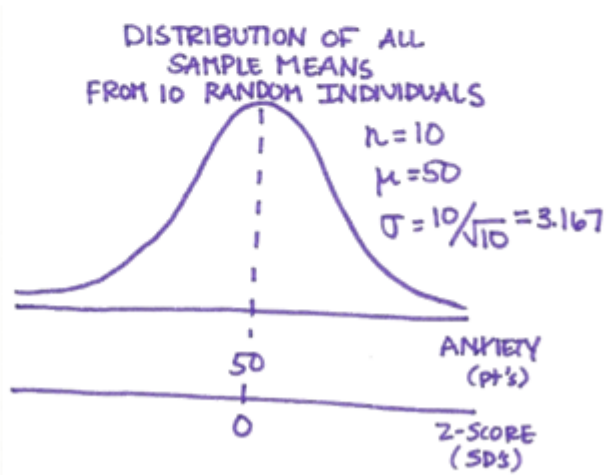
2. Choose Test Stat, α, & # tails

CLT: mean of repeated SRS → normally dist.

→ So use the z-stat

$\alpha = .05$ & 2 tails (default)





3. SRS data → Sample Mean

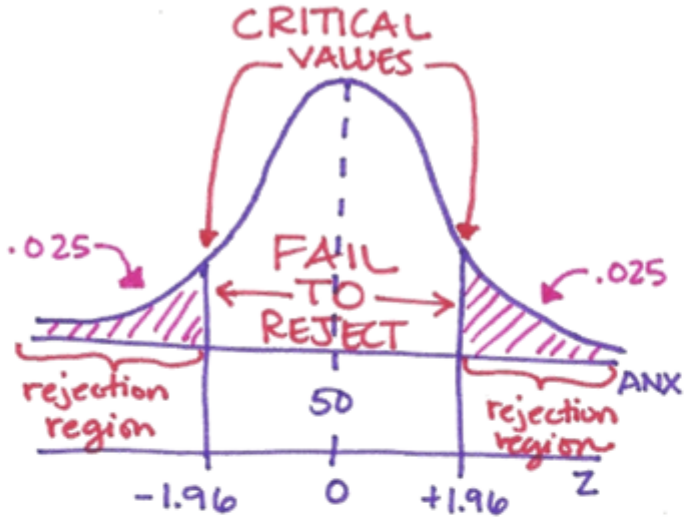
$$\bar{X} = \frac{\sum X}{n} = \frac{580}{10} = \mathbf{58}$$

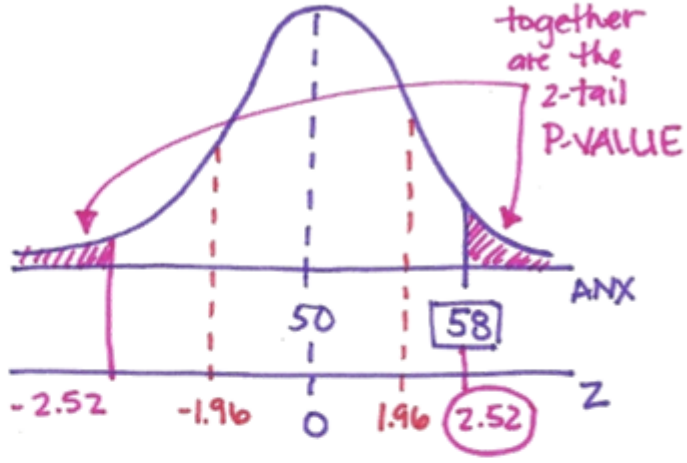
4. Rejection Region?

.05 in **BOTH** tails, so .025 in **EACH** tail …

→ Critical z = +/- 1.96 …

→ Reject if Z-score is > 1.96 or < -1.96





5. Calculate the Test Stat

Distribution of all sample means:

$$Mean_{mean} = \mu_{\bar{X}} = \mu_{pop} = 50$$

$$SE_{mean} = \sigma_{\bar{X}} = \frac{\sigma_{pop}}{\sqrt{n}} = \frac{10}{\sqrt{10}} = 3.167$$

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{58-50}{3.167} = \mathbf{2.52}$$

6. Conclusion

Z-stat falls in the rejection region

evidence the population's mean is not 50

"reject the Null"

"After the earthquake, townspeople's anxiety levels are higher than 50, on average."

Cautions About Significance Tests

Statistical significance

- only says whether the effect observed is likely to be due to chance alone, because of random sampling
- may not be practically important

That's because *statistical* significance doesn't tell you about the **magnitude of the effect**, only that there **is** one.

An *effect* could be too small to be **relevant**.

And with a large enough sample size, significance can be reached even for the tiniest effect.

- EX) A drug to lower temperature is found to reproducibly lower patient temperature by 0.4 degrees Celsius, $p < 0.01$. But clinical benefits of temperature reduction only appear for a 1 decrease or larger.

STATISTICAL significance does NOT mean PRACTICAL significance!!!

Cautions About Significance Tests

Don't ignore lack of significance

"Absence of evidence is not evidence of absence."

Having no proof of who committed a murder
does not imply that the murder was not committed.

Indeed, failing to find statistical significance in results is *not* rejecting the null hypothesis. This is very different from actually accepting it. The sample size, for instance, could be too small to overcome large variability in the population.

When comparing two populations, lack of significance does NOT imply that the two samples come from the same population. They could represent two very distinct populations with similar mathematical properties.

Let's Apply This to the Cancer Dataset

Read in the Data

```
library(tidyverse)     # Loads several very helpful 'tidy' packages
library(rio)           # Read in SPSS datasets
library(psych)         # Lots of nice tid-bits
library(car)           # Companion to "Applied Regression"
```

```
cancer_raw <- rio::import("cancer.sav")
```

Read in the Data

```
library(tidyverse)     # Loads several very helpful 'tidy' packages
library(rio)           # Read in SPSS datasets
library(psych)         # Lots of nice tid-bits
library(car)           # Companion to "Applied Regression"
```

```
cancer_raw <- rio::import("cancer.sav")
```

And Clean It

```
cancer_clean <- cancer_raw %>%
  dplyr::rename_all(tolower) %>%
  dplyr::mutate(id = factor(id)) %>%
  dplyr::mutate(trt = factor(trt,
                             labels = c("Placebo",
                                        "Aloe Juice"))) %>%
  dplyr::mutate(stage = factor(stage))
```

Descriptive Statistics

Skewness & Kurtosis

```
cancer_clean %>%
  dplyr::select(age, totalcw4) %>%
  psych::describe()
```

```
         vars  n  mean    sd median trimmed   mad min max range  skew
age         1 25 59.64 12.93     60   59.95 11.86  27  86    59 -0.31
totalcw4    2 25 10.36  3.47     10   10.19  2.97   6  17    11  0.49
         kurtosis   se
age         -0.01 2.59
totalcw4    -1.00 0.69
```

Tests for Normaility - Shapiro-Wilks

```
cancer_clean %>%
  dplyr::pull(age) %>%
  shapiro.test()
```

```
    Shapiro-Wilk normality test

data:  .
W = 0.98317, p-value = 0.9399
```

```
cancer_clean %>%
  dplyr::pull(totalcw4) %>%
  shapiro.test()
```

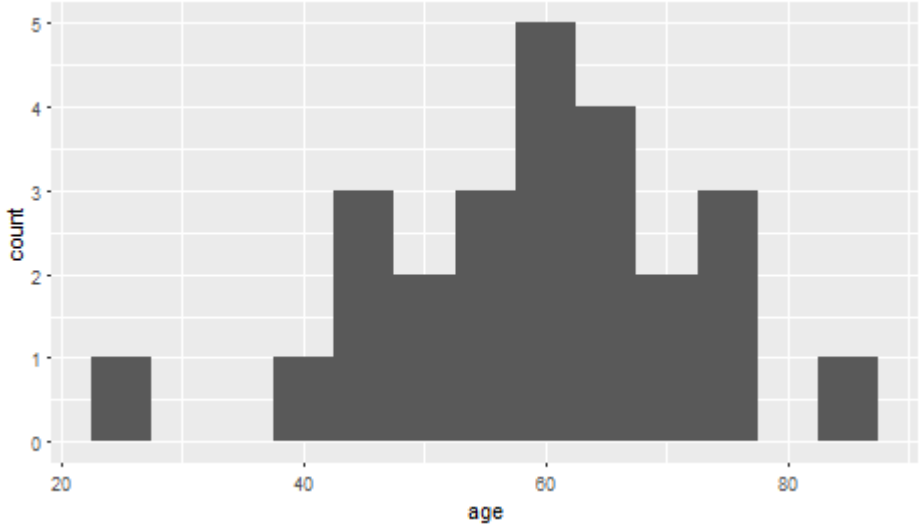
```
    Shapiro-Wilk normality test

data:  .
W = 0.9131, p-value = 0.03575
```

Plots to Check for Normaility - Age

Histogram

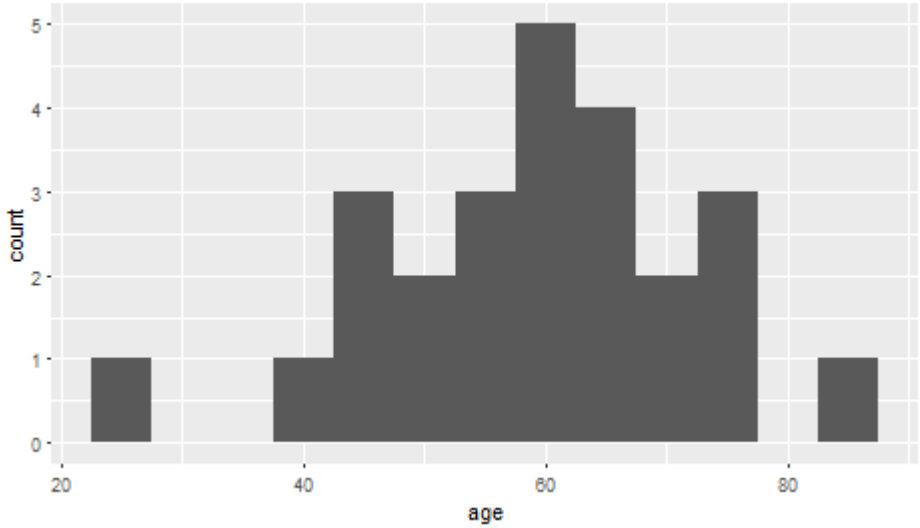
```
cancer_clean %>%
  ggplot(aes(age)) +
  geom_histogram(binwidth = 5)
```



Plots to Check for Normaility - Age

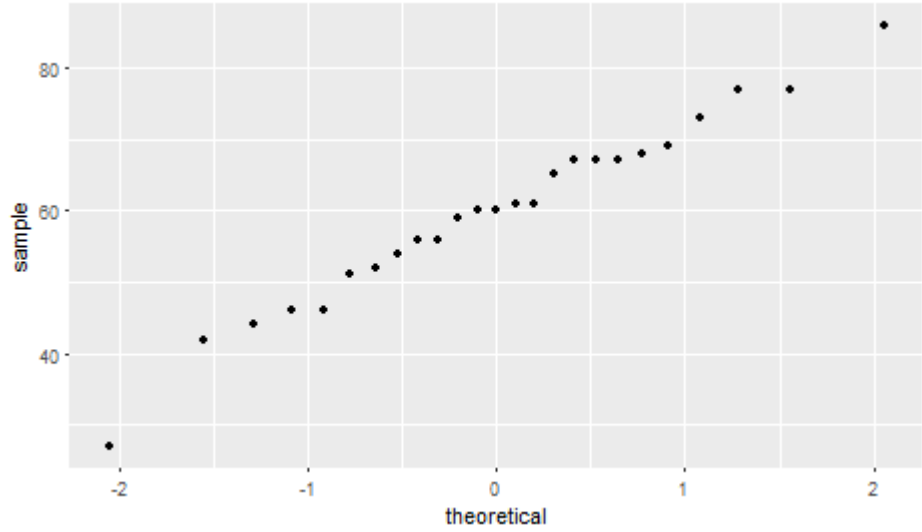
Histogram

```
cancer_clean %>%
  ggplot(aes(age)) +
  geom_histogram(binwidth = 5)
```



Q-Q Plot

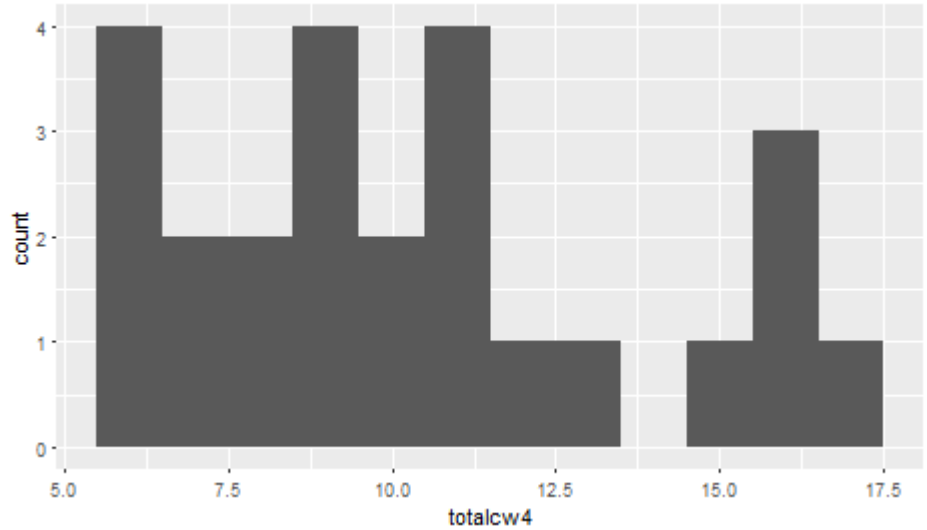
```
cancer_clean %>%
  ggplot(aes(sample = age)) +
  geom_qq()
```



Plots to Check for Normaility - Week 4

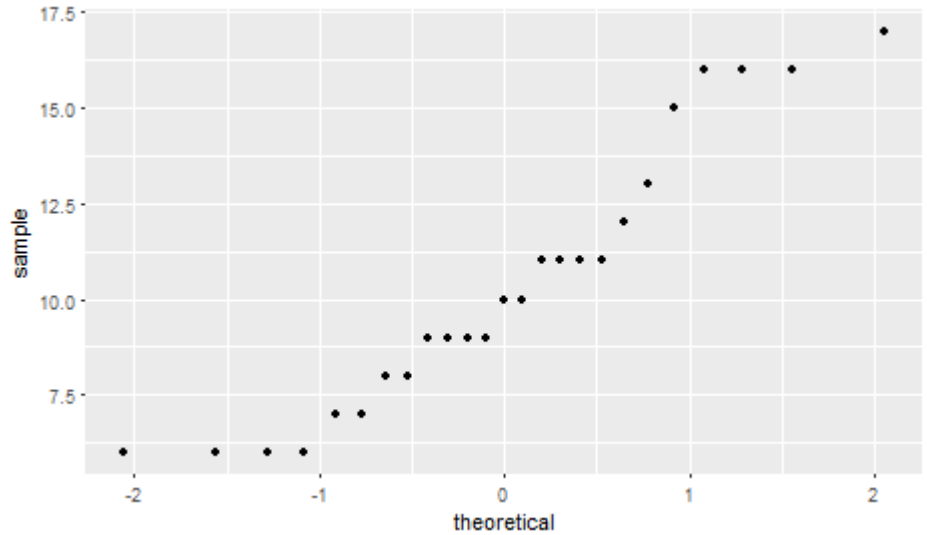
Histogram

```
cancer_clean %>%
  ggplot(aes(totalcw4)) +
  geom_histogram(binwidth = 1)
```



Q-Q Plot

```
cancer_clean %>%
  ggplot(aes(sample = totalcw4)) +
  geom_qq()
```



Questions?

Next Topic

Confidence Interval Estimation & The t-Distribution