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

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

Testing the association or dependence

- Correlation
- Regression

Understanding Statistical Inference - statistics help

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

Smoking and Lung Cancer: From Association to Causation

Causality and Statistics:

Hill's View Points and "Diversity of Evidence"

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

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

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, $\bar{x}$ *(for formulas)* or M *(for APA)*, summarizes **central tendency** of a group or sample of subjects

Hypothesis testing: step-by-step, p-value, t-test for difference of two means - Statistics Help

Hypothesis Testing - Introduction

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.

How P-Values Help Us Test Hypotheses: Crash Course Statistics #21

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...

$$H_1 : \mu_1 \neq \mu_2 \quad \text{or} \quad \mu_1 < \mu_2 \quad \text{or} \quad \mu_1 > \mu_2$$

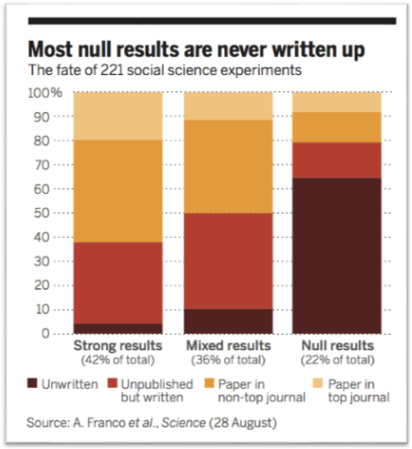
Statistical Significance and p-Values Explained Intuitively

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? $p - value < \alpha$

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

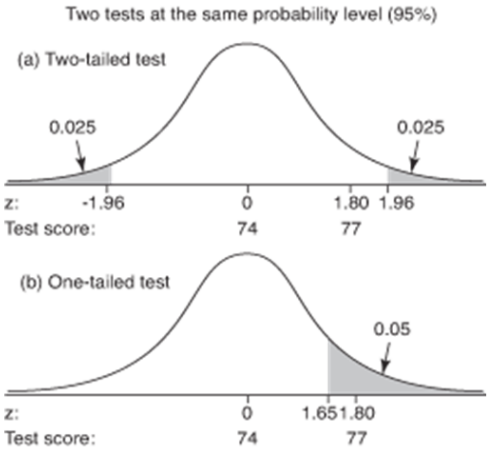
Suggests a directionality in results!

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

NO computational differences

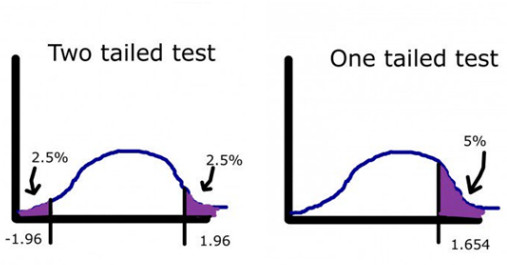
$2\ tail\ p-value = \mathbf{2}\times 1\ tail\ p-value$

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



ONE tail or TWO?

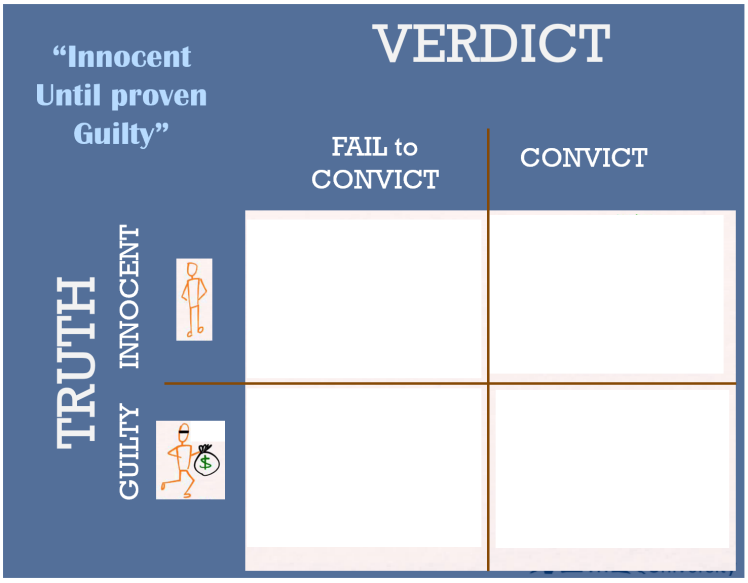
Some circumstances may warrant a 1-tailed test, BUT...
We generally **prefer** and default to a 2-tailed test!!!

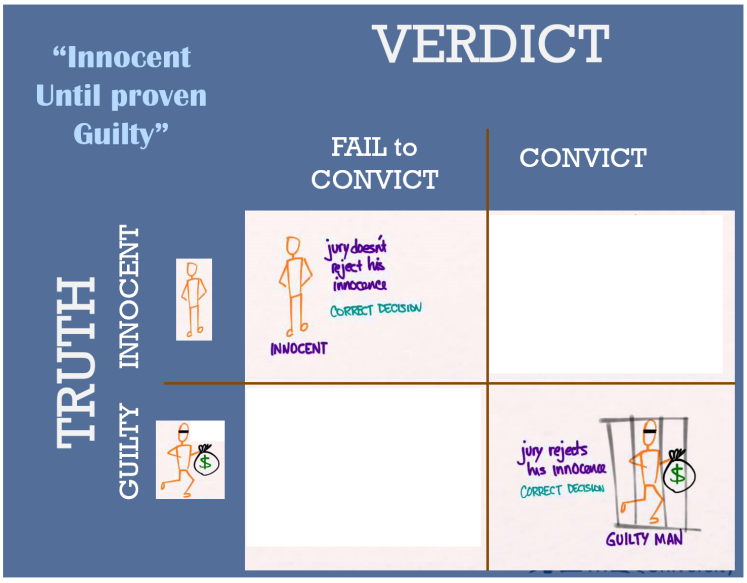


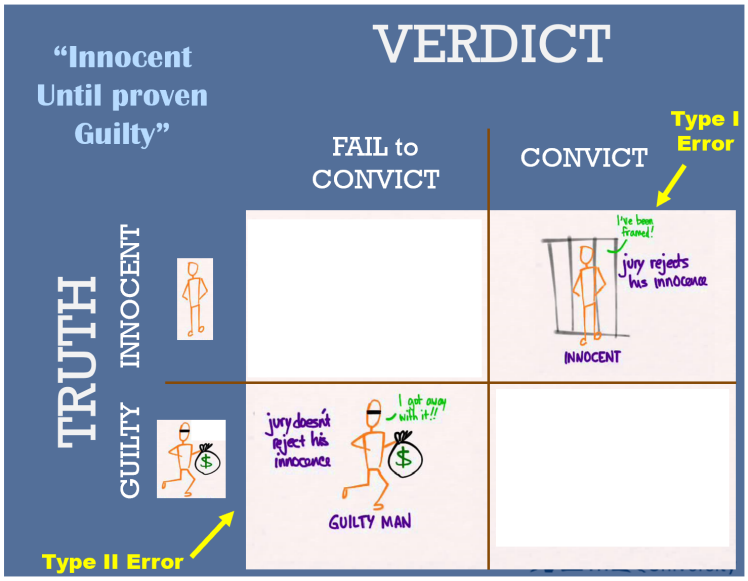
More conservative = 2 tails

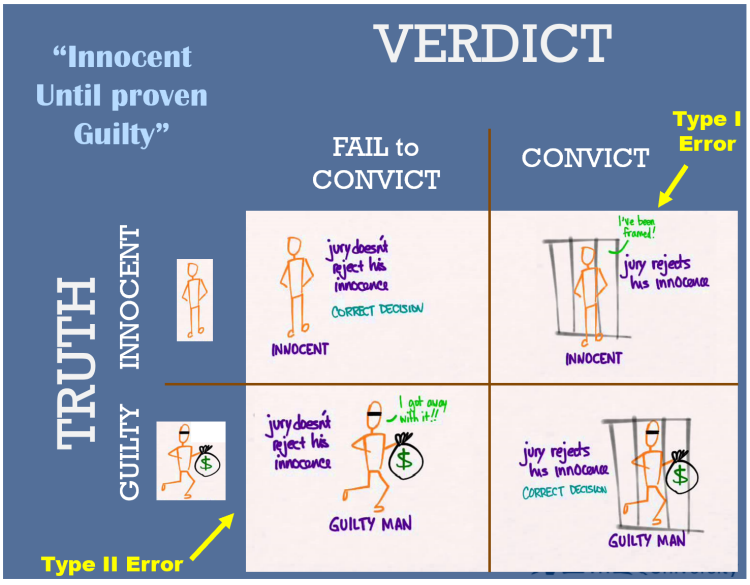
Rejection region is distributed in both tails

- 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!"


"Innocent Until proven Guilty"
VERDICT
FAIL to CONVICT
CONVICT
TRUTH
INNOCENT
GUILTY


"Innocent Until proven Guilty"
VERDICT
FAIL to CONVICT
CONVICT
TRUTH
INNOCENT
GUILTY
jury doesn't reject his innocence
CORRECT DECISION
INNOCENT
jury rejects his innocence
CORRECT DECISION
GUILTY MAN


"Innocent Until proven Guilty"
VERDICT
FAIL to CONVICT
CONVICT
Type I Error
TRUTH
INNOCENT
GUILTY
I've been framed!
jury rejects his innocence
INNOCENT
jury doesn't reject his innocence
I got away with it!!
GUILTY MAN
Type II Error


"Innocent Until proven Guilty"
VERDICT
FAIL to CONVICT
CONVICT
Type I Error
TRUTH
INNOCENT
GUILTY
jury doesn't reject his innocence
CORRECT DECISION
INNOCENT
I've been framed!
jury rejects his innocence
INNOCENT
jury doesn't reject his innocence
I got away with it!!
GUILTY MAN
jury rejects his innocence
CORRECT DECISION
GUILTY MAN
Type II Error

Choosing Alpha

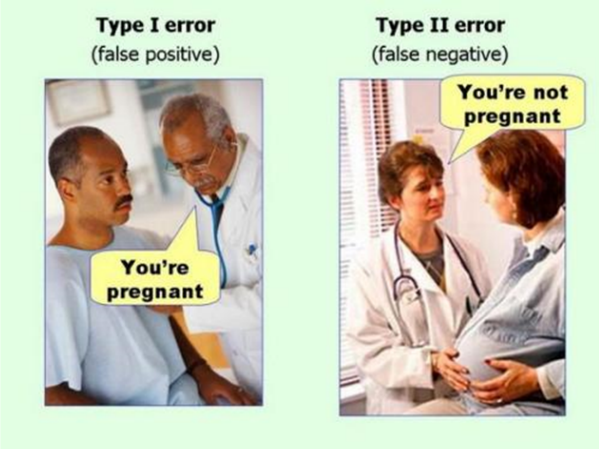
Alpha = probability of making a **type I error**

type I error

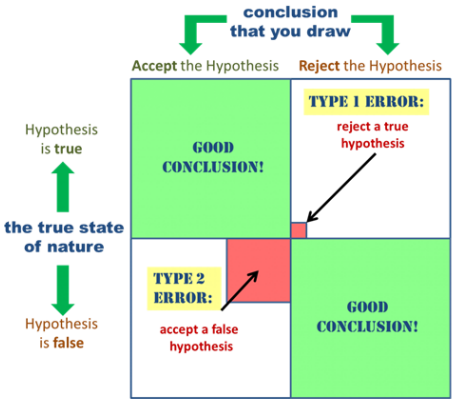
- We reject the NULL when we should not
- The risk of "false positive" results

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 trade off (type II error rate)

DEFAULT is $\alpha = .05$ **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

1. Sample was drawn at RANDOM *(at least as representative as possible)*

- Nothing can be done to fix a NON-representative samples!
- Can **NOT** statistically test

2. SD of the sampled population = SD of the comparison population

- Nearly impossible to check, can **NOT** statistically test

3. Variable has a NORMAL distribution in the population

- **NOT** as important if the sample is large, due to the **Central Limit Theorem**
- **CAN** statistically test:
  - Visual inspection of a histogram, boxplot, and/or QQ plot *(straight 45 degree line)*
  - Calculate the Skewness & Kurtosis... less clear guidelines
  - Conduct Shapiro-Wilks test *($p < .05$??? not normal)*

> For more information see this blogpost Is This Normal? Shapiro-Wilk Test in R To The Rescue and this article Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, as well as the R help page with `?shapiro.test`.

APA: results of a 1-sample z-test

- State the alpha & number of tails in the methods section, prior to the results section
- When used in a sentence, spell out mean and standard deviation
- When included in a table, figure, or within parentheses, use abbreviates: *n*, *M*, *SD*
- Report most values to TWO decimal places *usually*
- Report exact p-values to THREE decimal places *usually*, 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 (6.00) 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

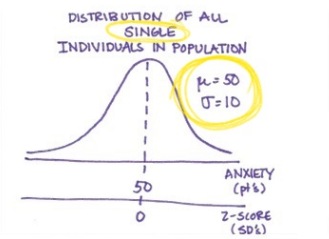
Assume the general population has an anxiety scale that is expressed as a T score, so that $\mu = 50$ and $\sigma = 10$.

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 $\sigma = 10$.



EXAMPLE: 1-sample z-test

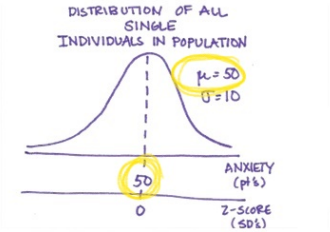
After an earthquake hits their town, a random sample of townspeople yields the following anxiety score:

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

$H_0: \mu = 50$
$H_1: \mu \neq 50$



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

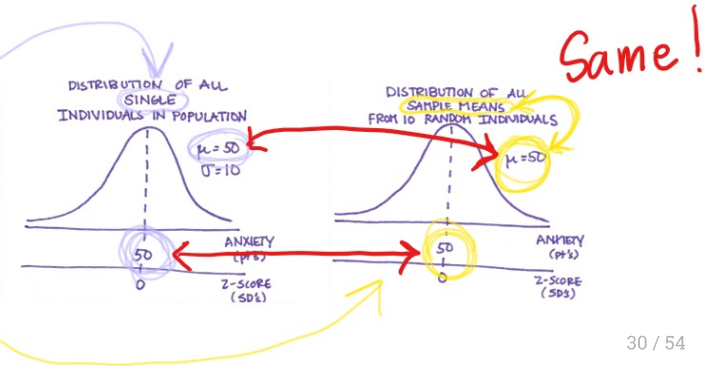
Assume the general population has an anxiety scale that is expressed as a T score, so that $\mu = 50$ and $\sigma = 10$.

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
α = .05 & 2 tails (default)



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

$H_0: \mu = 50$

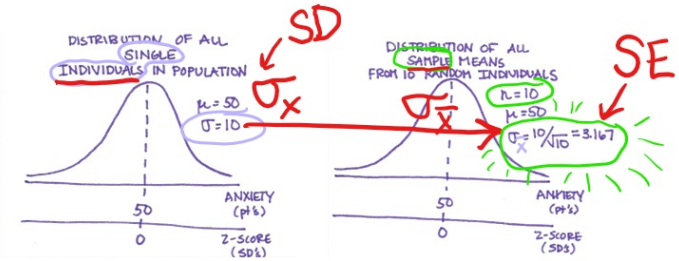
$H_1: \mu \neq 50$

2. Choose Test Stat, α, & # tails

CLT: mean of repeated SRS → normally dist.

→ So use the z-stat

α = .05 & 2 tails (default)



3. SRS data → Sample Mean

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

4. Rejection Region?

5. Calculate the Test Stat

6. Conclusion

3. SRS data → Sample Mean

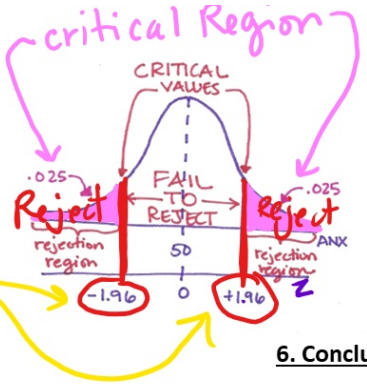
$$\bar{X} = \frac{\sum X}{n} = \frac{580}{10} = \mathbf{58}$$

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

6. Conclusion

3. SRS data → Sample Mean

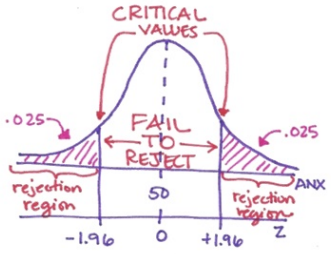
$$\bar{X} = \frac{\sum X}{n} = \frac{580}{10} = \mathbf{58}$$

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:

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

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

6. Conclusion

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

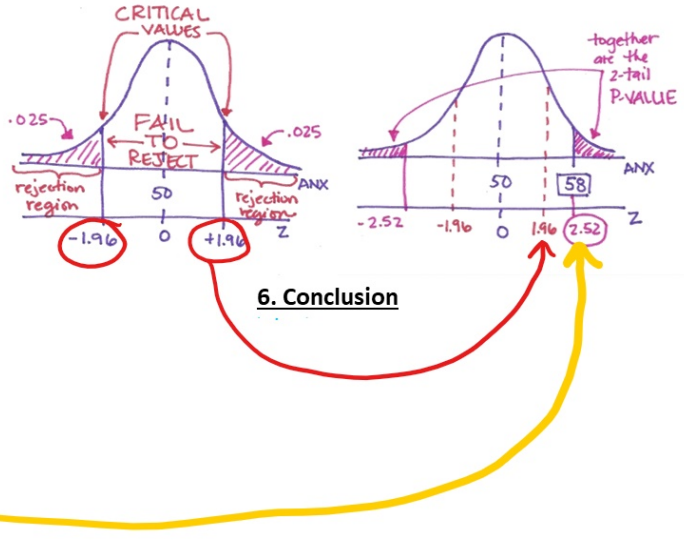
5. Calculate the Test Stat

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

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

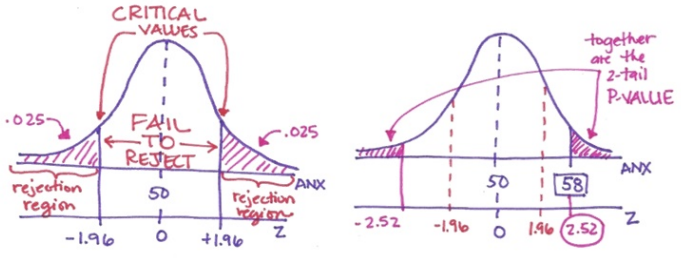
5. Calculate the Test Stat

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"

3. SRS data → Sample Mean

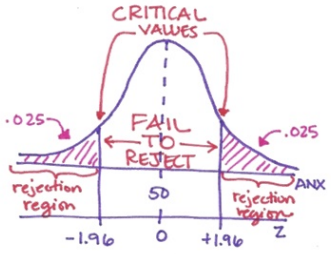
$$\bar{X} = \frac{\sum X}{n} = \frac{580}{10} = \mathbf{58}$$

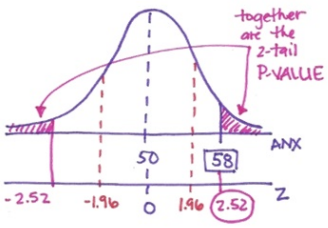
4. Rejection Region?

.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."

P-Value Problems: Crash Course Statistics #22

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 likely **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.

Good statistical practice is an essential component of good scientific practice, the statement observes and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

"The p-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post p<0.05 era.'"

"Over time it appears the p-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the **'file-drawer effect'** in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as **'p-hacking'** and **'data dredging'** that emphasize the search for small p-values over other statistical and scientific reasoning."

In light of misuses of and misconceptions concerning p-values, the statement notes that statisticians often supplement or even replace p-values with other approaches. These include methods "that emphasize estimation over testing, such as confidence, credibility or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates."

"The contents of the ASA statement and the reasoning behind it **are not new** — statisticians and other scientists have been writing on the topic for decades," Utts said.

APA Task Force on p-values

The six principles, which are elaborated in the statement, are:

1. P-values **CAN** indicate how incompatible the data are with a specified *statistical model*.
2. P-values **DO NOT** measure the *probability* that the studied hypothesis is true, or the probability that the data were produced by *random* chance alone.
3. Scientific conclusions and business or policy decisions **SHOULD NOT** be based only on whether a p-value passes a specific threshold.
4. Proper inference requires **FULL REPORTING** and **TRANSPARENCY**.
5. A p-value or statistical significance **DOES NOT** measure the **SIZE** of an effect or the **IMPORTANCE** of a result.
6. By itself, a p-value **DOES NOT** provide a good measure of evidence regarding a model or hypothesis.

> See Statistical Methods in Psychology Journals: Guidelines and Explanations, *(1999) by Leland Wilkinson and the Task Force on Statistical Inference APA Board of Scientific Affairs*

Some ways to improve the 'Crisis of Lack of Replicability'

- Be completely transparent in reporting, including data cleaning/wrangling and statistical analysis
- Make all data *(deidentified)* open source and freely available repositories
- Reduce focus on *NEW* findings and incentives *REPLICATION* studies
- Increase statistical power, often requiring larger sample sizes, more complex study design, or more sophisicated sattistical analyses
- Reduce publication bias
    - interpret p-values correctly *(not overstate)*
    - lower the reliance on p-values and the strict .05 cut-off
    - employ pre-registrations processes

Let's Apply This to the Cancer Dataset

Testing normality in the population, based on a sample

Read in the Data

```
library(tidyverse)    # Loads several very helpful 'tidy' packages
library(haven)        # Read in SPSS datasets
library(furniture)    # Nice tables (by our own Tyson Barrett)
library(psych)        # Lots of nice tid-bits
```

```
cancer_raw <- haven::read_spss("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))
```

The Cancer Dataset

```
<div id="htmlwidget-03746a5f6d563ca6a4bc" style="width:100%;height:auto;" class="datatables html-
<script type="application/json" data-for="htmlwidget-03746a5f6d563ca6a4bc">{"x":{"filter":"none",
```


Descriptive Statistics

Skewness & Kurtosis - Age & Week 4

```
cancer_clean %>%                          # start with the dataset name
  dplyr::select(age, totalcw4) %>%   # select your variables
  psych::describe()                       # calculate descriptive statistics
```

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

Tests for Normaility - Age

> In our population, does age follow the normal distribution?

The Shapiro-Wilks test:

- H_0: In the population, age DOES follow the normal distribution
- H_1: In the population, age does NOT follow the normal distribution

```
cancer_clean %>%                # start with the dataset name
  dplyr::pull(age) %>%        # pull out the variable in question
  shapiro.test()                # run the Shapiro Wilks test of normality (in base R)
```

```
    Shapiro-Wilk normality test

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

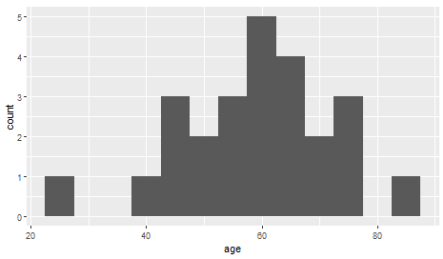
Big p-value...fail to reject the Null...

> **Conclusion**: The Shapiro-Wilks test on this sample provides **no evidence** that distribution of age in the population is **not** normally distributed, $W = 0.98$, p = .940.

Plots to Check for Normaility - Age

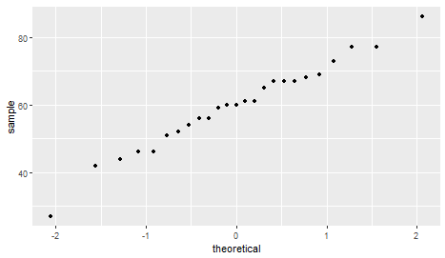
Histogram

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



Q-Q Plot

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



Tests for Normaility - Week 4

> In our population, does total oral condition at 4 weeks follow the normal distribution?

The Shapiro-Wilks test:

- H_0: In the population, totalcw4 DOES follow the normal distribution
- H_1: In the population, totalcw4 does NOT follow the normal distribution

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

```
    Shapiro-Wilk normality test

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

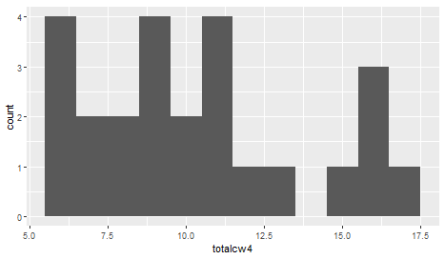
Tiny p-value...reject the Null...

> **Conclusion:** The Shapiro-Wilks test on this sample **does** provide **evidence** that distribution of total oral condition at 4 weeks in the population is **NOT** normally distributed, $W = 0.91$, p = .036.

Plots to Check for Normaility - Week 4

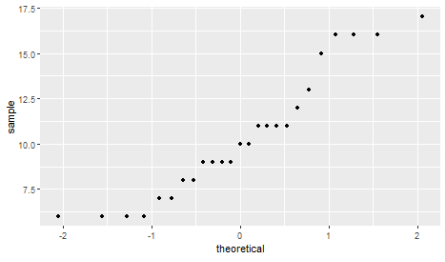
Histogram

```
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