

Ch1 & Ch2

1. Statistical Methods: Descriptive vs Inferential
2. Basic Terminology
3. Research Studies: Observational Study vs Experimental study

1. Statistical Methods

1. **Descriptive Statistical Methods:** collect data and describe them.
2. **Inferential Statistical Methods:** collect data, analyze, interpret, and make conclusions based on the data.

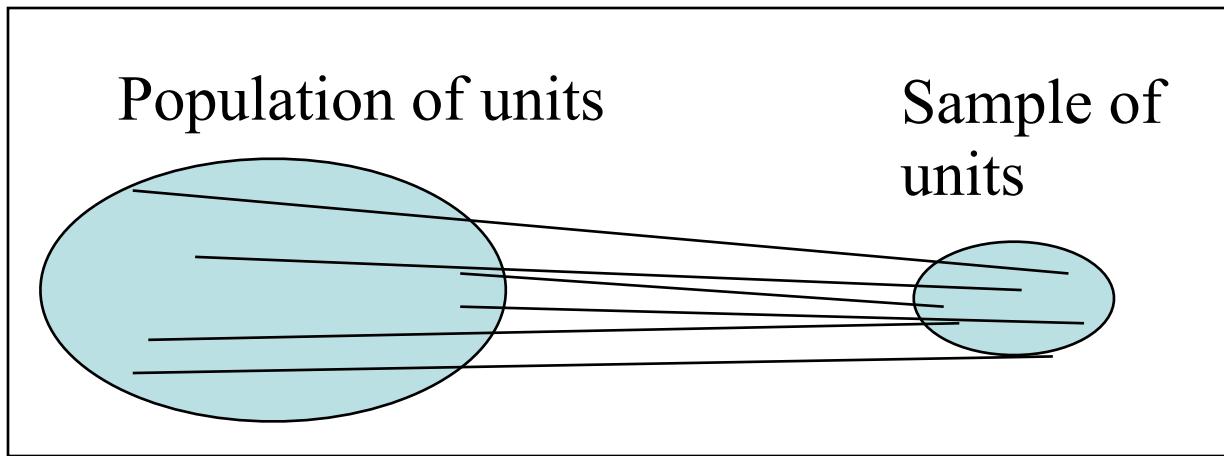
In this course, we will concentrate on inferential methods for the research study but the distinction may not always be clear.

2. Basic Terminology

- **Observation Unit:** The unit upon which data are collected.
Ex: US Adult
- **Population:** Complete set of units of interest.
Ex: All US Adults
- **Sample:** A subset of the population that is actually measured.
Ex: 100 individuals selected based on random sample of SSNs.
- **Census:** when the sample equals the population
Ex: US Census

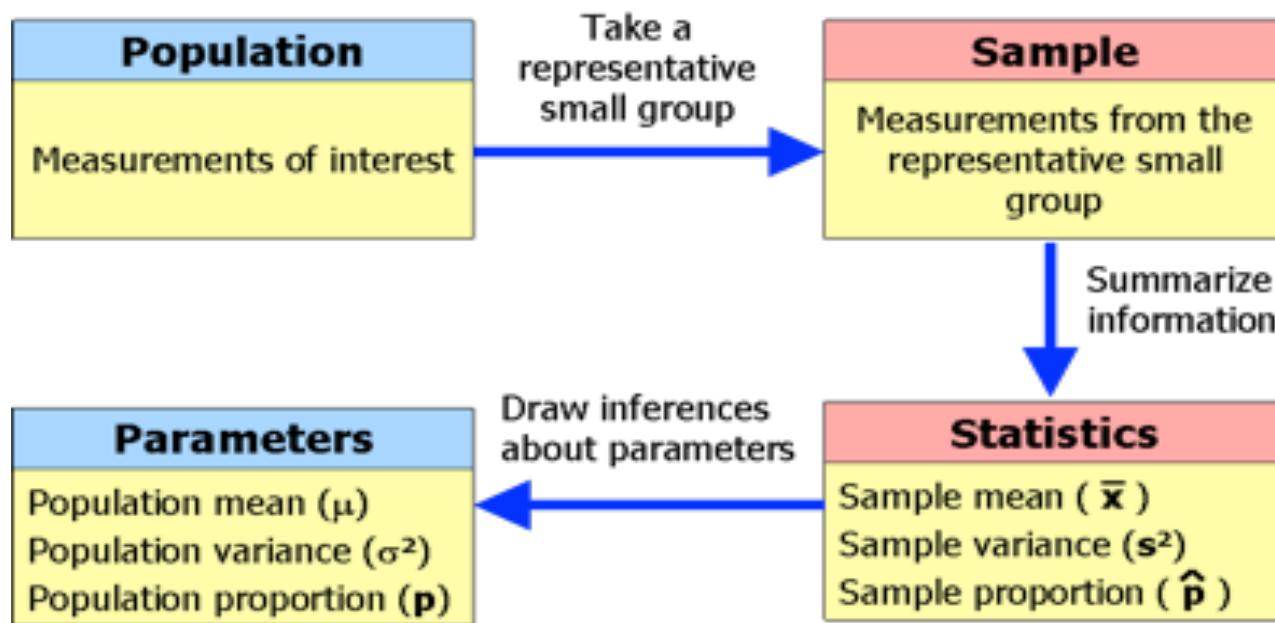
- **Variable:** information of interest about each individual item in a population.
Ex: Height, Weight, Age, Gender
- **Statistic:** numerical descriptive measure for a sample.
Ex: Average height of 100 individuals in sample
Note: Sample average is denoted \bar{y} (y bar).
- **Parameter:** numerical descriptive measure for a population.
Ex: Average height of all US adults
Note: Population average is denoted μ (mu).

Return to Statistical Methods



- **Descriptive** statistical methods involve describing the sample.
Ex: Describe the height values of the 100 U.S. adults.
- **Inferential** statistical methods involve making statements about the population based on the sample.
Ex: Make statements about the heights of all U.S. adults based on the sample.

The Process for Statistical Inference



First principle of statistical inference: You make inference to the population from which you sample.

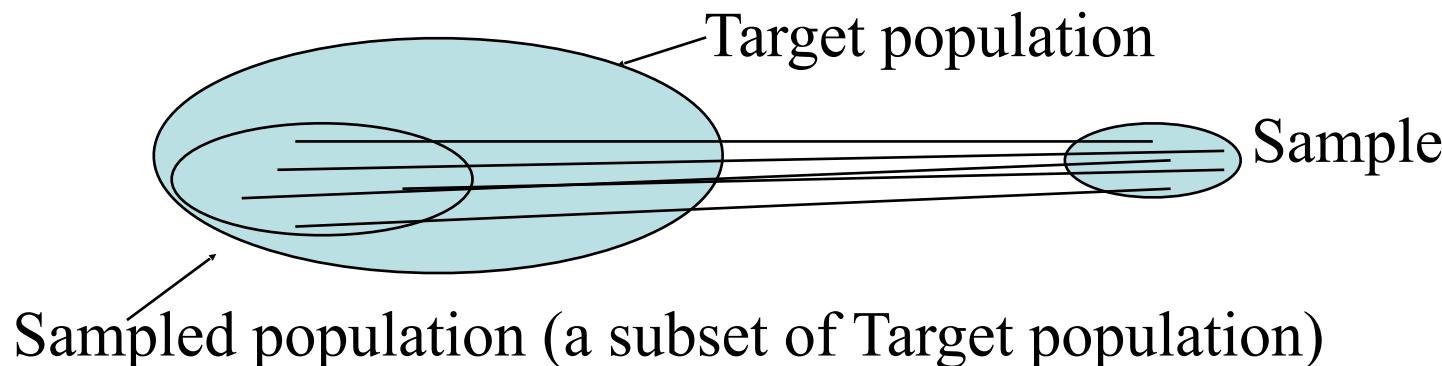
Ex1: A company is interested in implementing a new manufacturing process. Researchers observes several lab runs of the new process. The objective is to find out how the process will work in large scale production

Target vs Sampled Populations

- **Target Population:** the population you would like to sample.
- **Sampled Population:** the population you actually do sample.

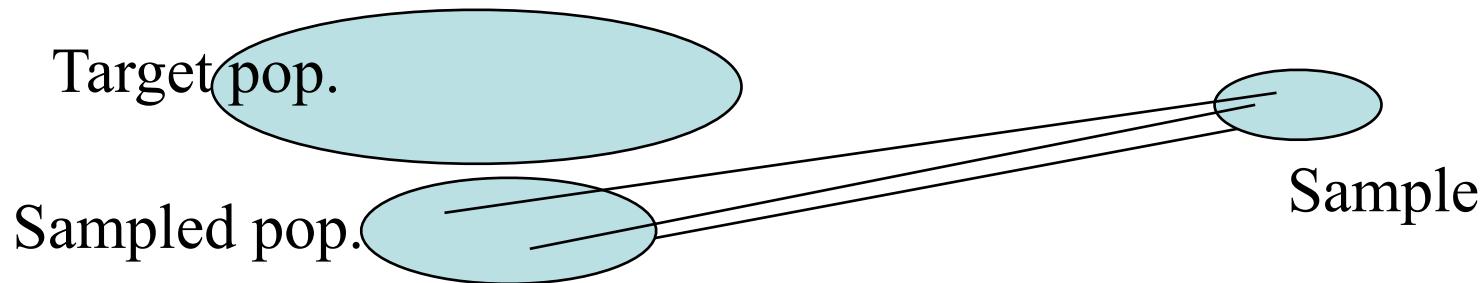
Often, the sampled population is a subset of the target population.

Ex: Researcher interested in comparing wheat damage for three application levels of a particular pesticide. They perform a field trial where they inoculate several fields with one type of pest.



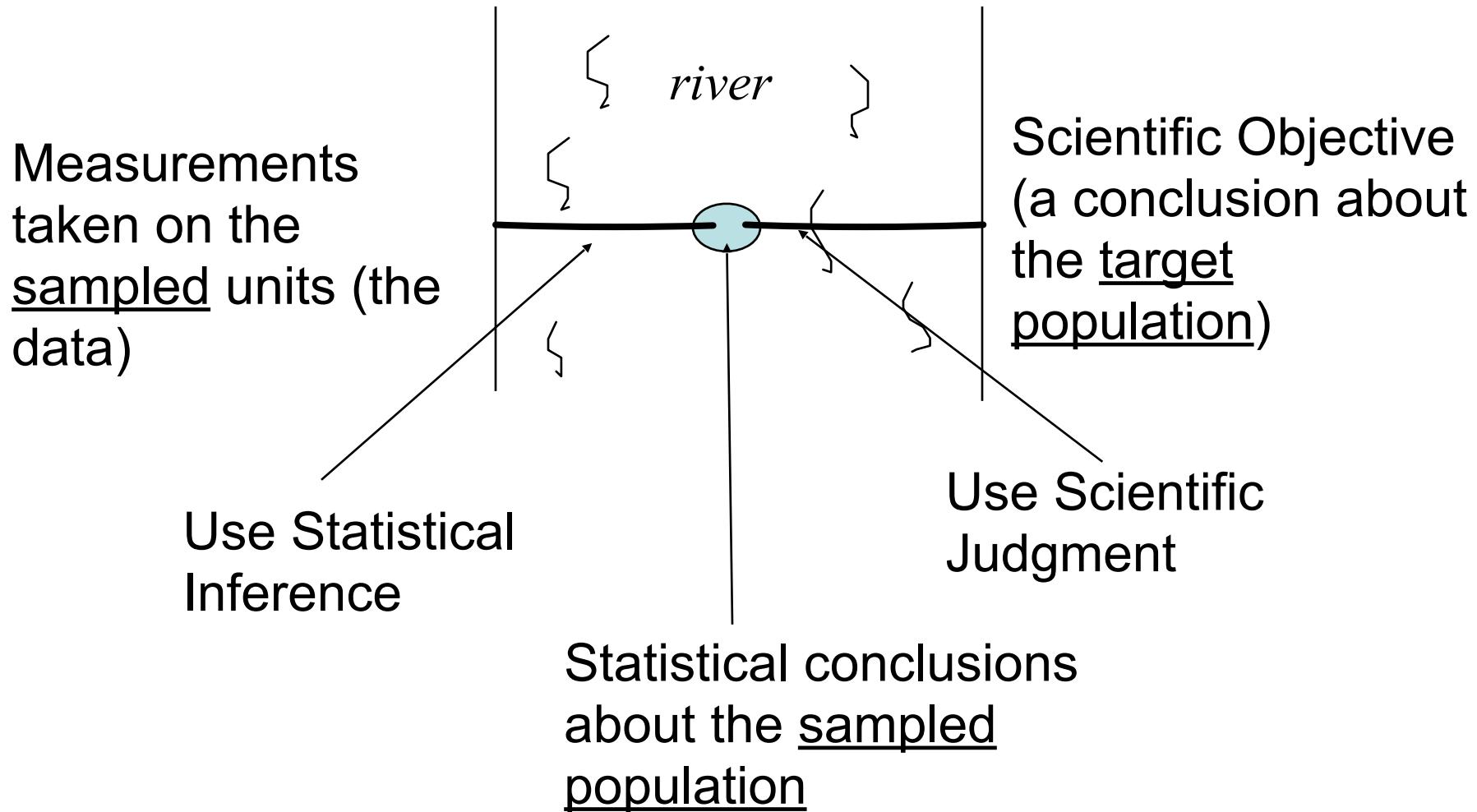
Other times, the Sampled population is not a subset of the Target population.

Ex: From Wikipedia (08/11/16): “A **model organism** is a (non-human) species that is extensively studied to understand particular biological phenomena, with the expectation that discoveries made in the organism model will provide insight into the workings of other organisms.”



You make statistical inference to the sampled population. You make scientific judgments about whether that is close enough to the target population to satisfy your objectives.

“River diagram” (Cornfield and Tukey)



Recall that in this course, we will concentrate on **inferential methods for the research study**.

Now we are going to look at different types of research studies.

3. research studies

- The type of **conclusion** we draw depends on the **study method** used.
- Two types of conclusions from the research study:
 - Correlation (association)
 - Causation (cause and effect relationship)
- Two types of study methods:
 - Observational study
 - Experimental study

Observational Studies

This study method observes individuals and measures variables of interest, but does **not** attempt to manipulate or influence the variables of interest.

Types of observational studies:

1. A **sample survey** is a study that provides information about a population at a particular point in time.
2. A **prospective study** is a study that observes a population in the present *using sample survey* and proceeds to follow the subjects in the sample forward in time in order to record the occurrence of specific outcomes.
3. A **retrospective study** is a study that observes a population in the present *using sample survey* and also collects information about the subjects in the sample regarding the occurrence of specific outcomes that have already taken place.

Sampling Methods for Surveys:

Sampling Frame: list of sampling units from which sampling is done.

1. **Simple Random Sampling:** select units randomly from the sampling frame.
2. **Systematic Sampling:** take units from the sampling frame at a regular interval: e.g. by selecting every 15th person on a list of the population.
3. **Stratified Random Sampling:** Organize the frame into groups (or strata) of like units. Sample independently within each stratum. The objective is to gain efficiency by sampling less intensively in strata that have low variability. e.g. randomly select 5 female and 5 male.

4. Cluster Sampling: researchers divide a population into smaller groups known as clusters. They then randomly select among these clusters to form a sample.

Note: In stratified random sampling, we take a simple random sample within each group. In cluster sampling, we take a simple random sample of groups and then observe all items within the selected groups.

Observational studies are valuable for discovering trends and possible **association**. However, it is **NOT possible** for observational studies to demonstrate a **causal relationship**.

Example: Suppose that we observe that a kid is violent (A) and happens to watch a lot of violent TV shows (B):
Possible scenarios for the cause and effect relationship among the events :

- He could be violent because he is learning the behaviour (B causes A)
- He could be watching violent TV because he likes violence (A causes B)
- He could be experiencing a mental health issue (Both A and B cause C or C cause A and B)

Experimental Studies

- To really understand **cause-and-effect relationships**, **experimental studies** have to be used.
- In an experimental study, we must **deliberately** change the input variables and observe changes in the output.
- Generally, any experimental study has two aspects:
 1. The design of experiment (collecting data)
 2. Statistical analysis of the data: the method of analysis depends directly on the design employed

Basic concepts:

- An **experiment** can be defined as a series of runs in which purposeful changes are made to the input variables so that we may observe and identify the reasons for changes that may be observed in the output response.
 - Each **experimental run** is a **test**.
- A **factor** is an input variable studied in the experiment.
- In order to study the effect of a factor on the response, two or more values of the factor are used. These values are referred to as **levels**.
- A **treatment** is a combination of factor levels.

Experimental Study

Example

We want to study the effects of water and sunlight on the growth of tomato plants, where water is measured by the amount of water a tomato plant needs per week, and sunlight is measured by the number of hours of sunlight a tomato plant needs per day.

- factors of interest: water and sunlight
- levels of each factors
 - water:
eg: 1 inch, 3 inch (labeled by “ α ” and “ β ”, respectively)
 - sunlight:
eg: 3 hours, 6 hours (labeled by “A” and “B”, respectively)
- 4 treatments
 $(\alpha, A) (\alpha, B) (\beta, A) (\beta, B)$

“Statistical” design of experiment: The process of planning the experiment so that the appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions.

Three basic principles:

1. Randomization
2. Replication
3. Blocking

CH 3: Data Description

1. Types of data: Categorical, Numerical
2. Summary Statistics for a Numerical variable
3. Summary Graphs for a Numerical variable
4. Mean vs Median

1. Types of Data

- **Categorical/Qualitative/Factor Variables**: can be placed into categories.

Examples: Transport Type

Note: Can be coded as numbers (eg:
Car = 0, Other = 1)

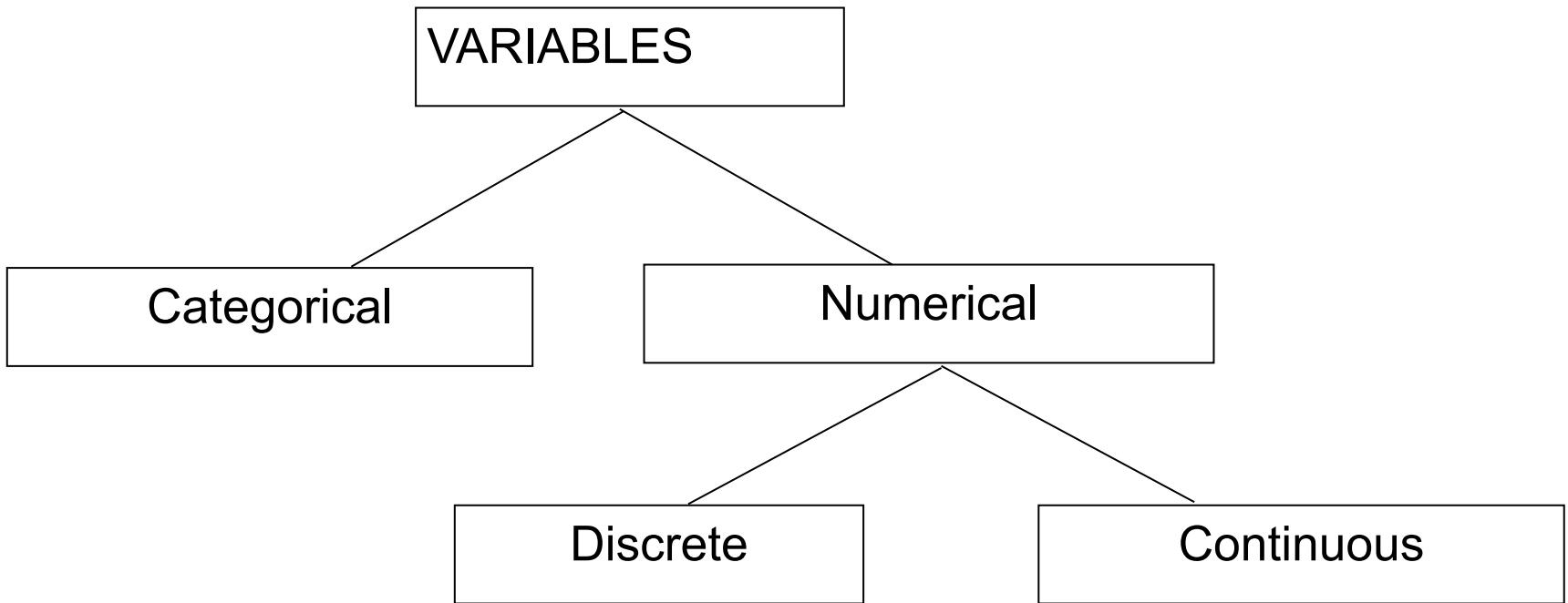
- **Numerical/Quantitative Variables**: have numerical values with natural ordering.

- **Discrete Variables** can only take some values; often obtained by counting.

Examples: Number of Children, Credit Hours

- **Continuous Variables** can take any value within a given interval.

Examples: Height, Weight



2. Summary Statistics for a Numerical variable

Measures of Central Tendency

- The **mode** is the value that occurs most often (with the highest frequency).
- The **median** is the middle value in the ordered data set. (Half above, half below)
- The **mean** (denoted \bar{y}) is the sum of the values divided by the number of observations.

n = sample size

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \left(\sum_{i=1}^n y_i \right) / n$$

Percentiles (and Quartiles)

- The **pth percentile** of a set of n measurements arranged in order is the value that has p% of the measurements below it.
- Hence the median is the 50th percentile.
- Q1 is the 25th percentile and Q3 is the 75th percentile.
- The “**five number summary**” includes **min, Q1, median, Q3 and max** values for a data set.
- We will see that a **boxplot** is the graphical display of the five number summary.

Measures of Variability (or Spread)

- The **range** is the difference between the largest and smallest values. Range = max – min.
- The **interquartile range (IQR)** is the difference between Q3 (the 75th percentile) and Q1 (the 25th percentile). IQR = Q3 – Q1
- The **sample standard deviation (s)** and **sample variance (s²)** also measure variability. Cannot be negative!

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

An Important Reminder

- Recall that the population of measurements is a complete set of observations. A sample is a subset of observations selected from the population of interest.
- The population mean is denoted μ (mu); the sample mean is denoted \bar{y} .
- The population standard deviation is denoted σ (sigma); the sample standard deviation is denoted s .

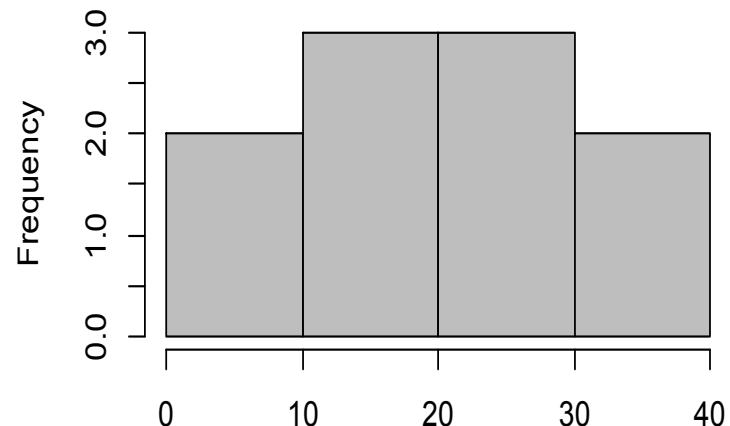
3. Summary Graphs for a Numerical Variable

Common graphics for a single numerical variable are histograms and boxplots.

Histograms

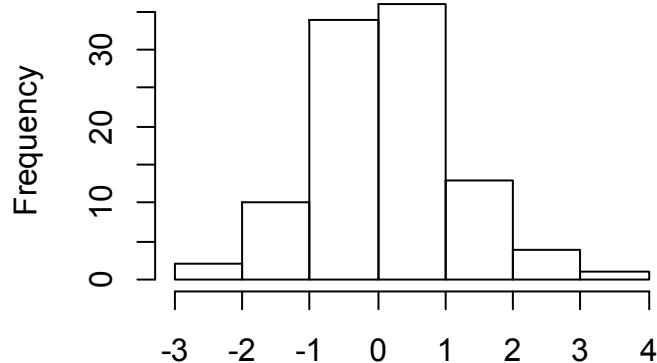
- Start with some equally spaced intervals.
- Count the # of observations (or frequency) that fall into each interval.
- Relative frequency is the frequency divided by the total # of observations (n).
- Histogram is a graph of the frequencies or relative frequencies.

Interval	Freq	Rel Freq
0 - 9	2	$2/10 = 0.2$
10 – 19	3	$3/10 = 0.3$
20 – 29	3	$3/10 = 0.3$
30 - 40	2	$2/10 = 0.2$

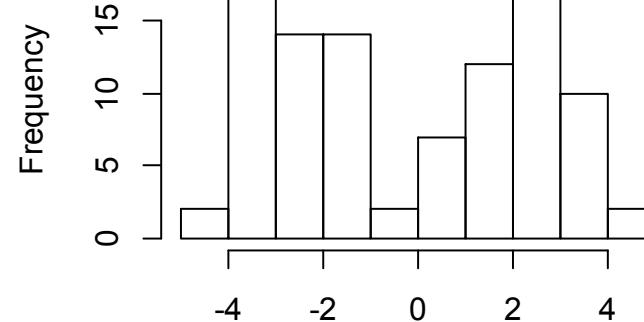


Example Histograms

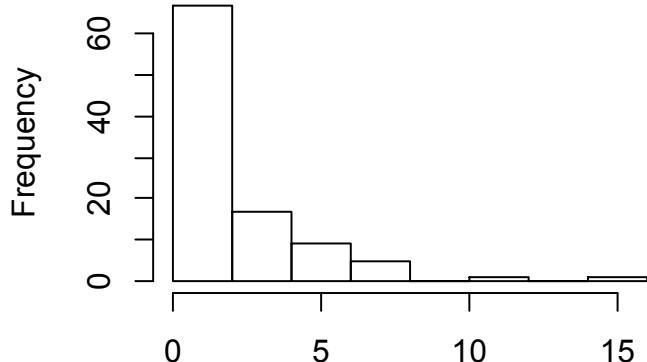
Symmetric, Unimodal



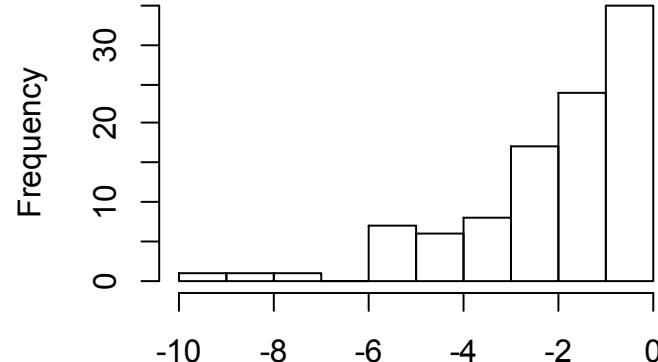
Bimodal



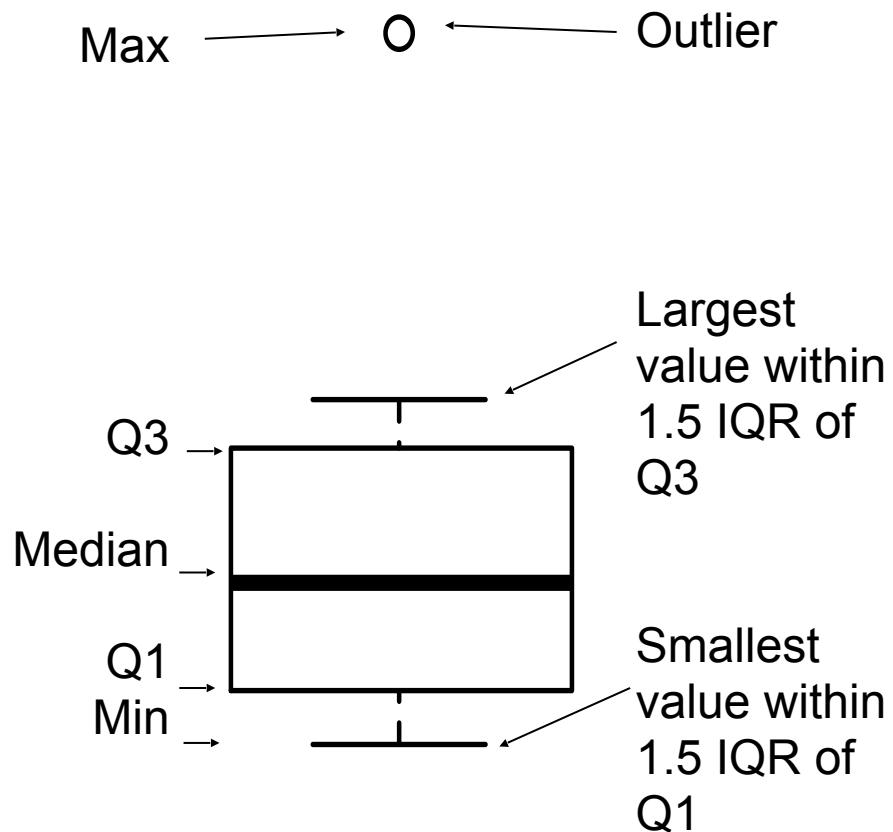
Skewed Right



Skewed Left



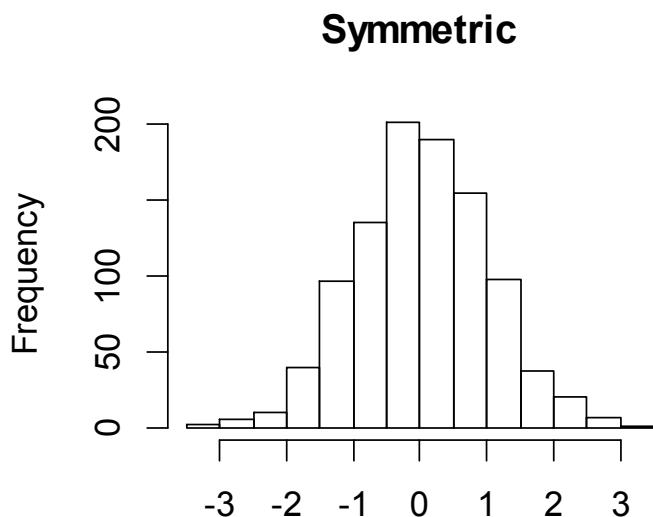
Boxplots



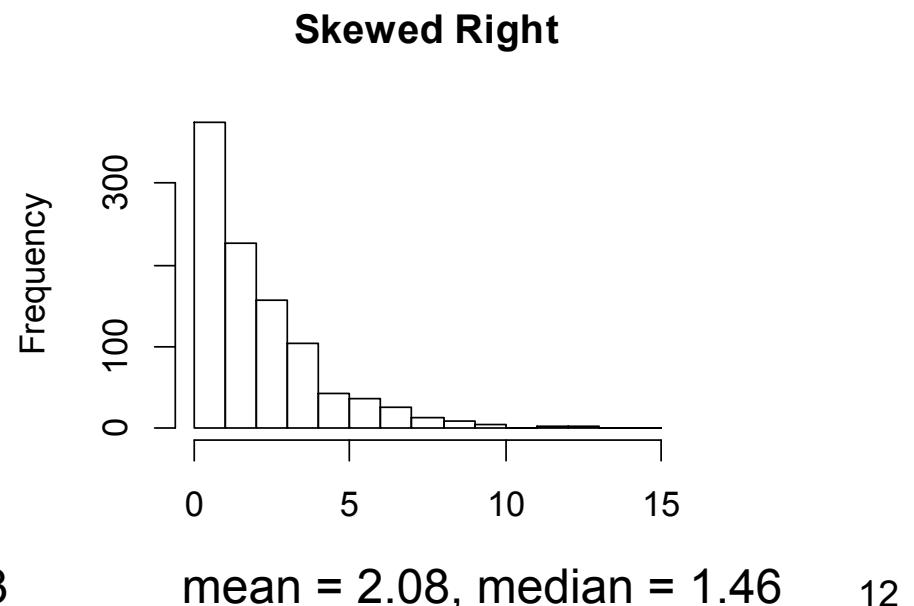
- The boxplot is a graph of the 5 number summary (min, Q1, median, Q3, max) with outliers marked.
- One definition of an **outlier** is a value that lies more than 1.5 IQR from Q1 or Q3. Recall that $IQR = Q3 - Q1$.

4. Mean vs Median

- For symmetric distributions, the sample mean and median will be close.
- For skewed distributions, the sample mean and median can be very different.



mean = 0.03, median = 0.03



mean = 2.08, median = 1.46

Should I report Mean or Median?

- Most often, people will report mean (and standard deviation or SE).
- But if the distribution is skewed, they may choose to report median (and range or IQR).
- Ideally, choice of mean or median should be driven by the research question not the shape of the distribution.
- If average or cumulative value is of interest, use the mean.
- If “typical” value is of interest, use the median.

CH 4 Probability Distributions

1. Probability, random variables and probability distributions
2. The Normal (Gaussian) distribution
3. The “Empirical Rule” and Chebyshev’s Rule
4. Sampling distribution of the sample mean

1. Random Variables

- **Probability** is a numerical quantity that expresses the likelihood of an event.
Probabilities take values between 0 and 1.
- The probability of an event can be interpreted as the relative frequency (proportion of times) the event occurs in an indefinitely long series of repetitions of the chance operation.
- **Example:** Single flip of a fair coin.
 $P(\text{Heads}) = 0.5$
- In a long series of tosses of a fair coin, we expect to get Heads about 50% of the time.

- A **random variable** (RV) is a variable whose value is the outcome of a random event.
- A **probability distribution** for a RV is a description of the probabilities for all possible outcomes. Total probability equals 1.
 - For **discrete** RVs, the distribution can be summarized as a table, graph or formula. Sum of probabilities must equal 1.
 - For **continuous** RVs, the distribution is summarized as a formula to describe a curve. The area under the curve must equal 1.

Example of a Discrete RV

Let Y be a random variable that gives the outcome of a single roll of a fair die.

Table:

Formula:

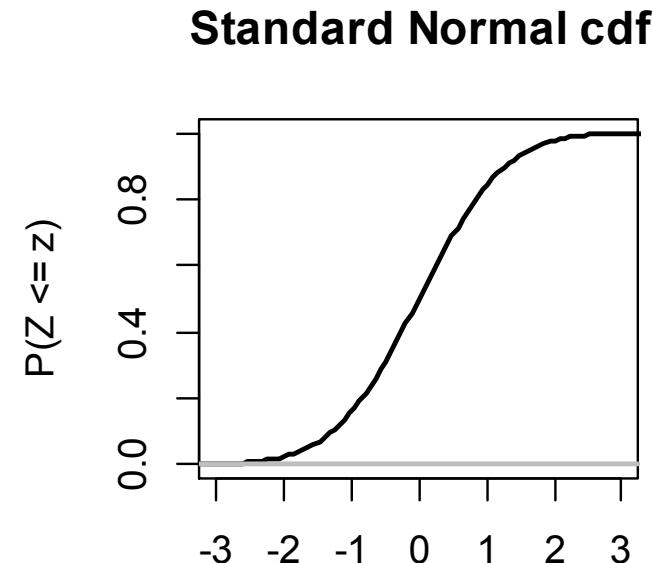
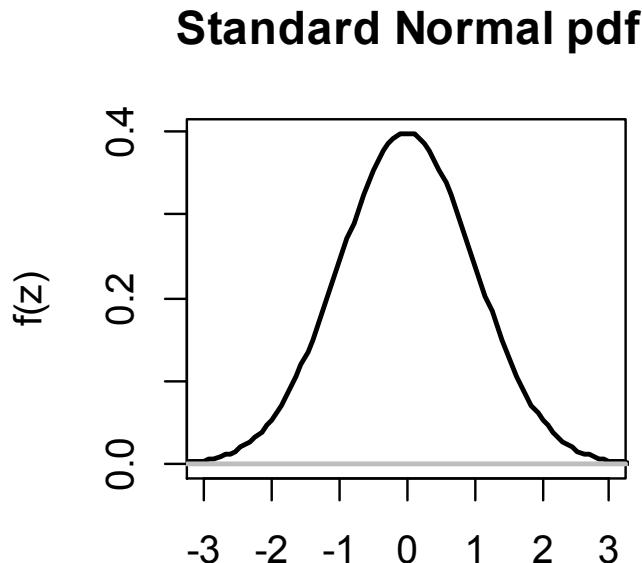
Graph:

2. Continuous Example: the Normal (Gaussian) family of distributions

- Many populations can be described by a normal distribution.
- Each normal distribution is defined by it's mean (μ) and standard deviation (σ).
- If a variable Y follows a normal distribution with mean μ and standard deviation σ , then we write $Y \sim N(\mu, \sigma)$.

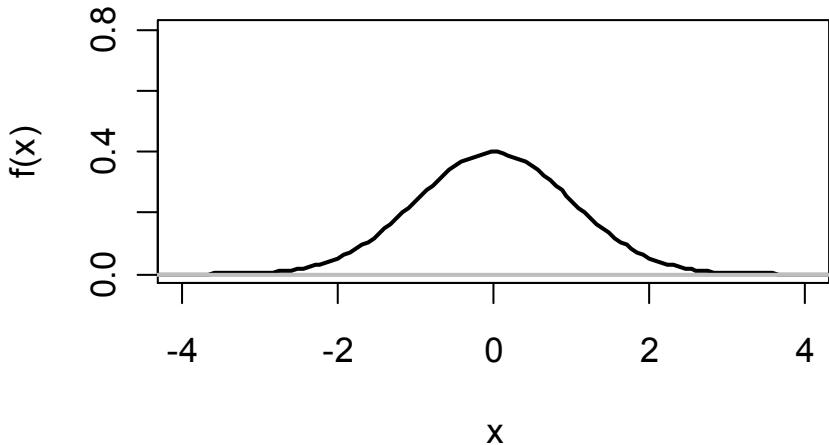
Normal pdf vs cdf

- The normal probability density function (**pdf**) is like a smooth relative histogram.
 - General form is $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$
 - The total area under the curve must equal 1.
- The normal cumulative distribution function (**cdf**) gives the $P(Y \leq y)$.

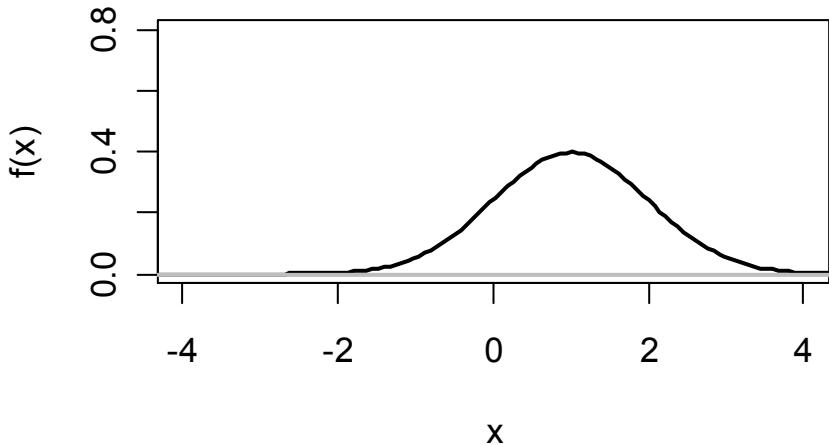


Normal Distribution Examples

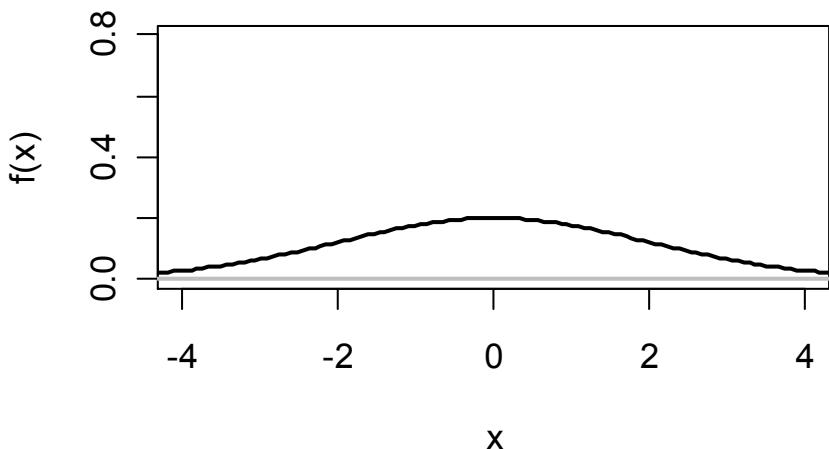
$\mu = 0, \sigma = 1$



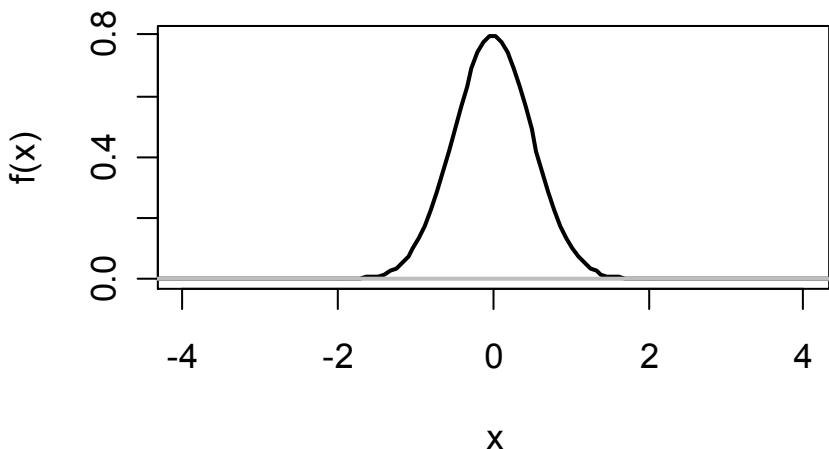
$\mu = 1, \sigma = 1$



$\mu = 0, \sigma = 2$



$\mu = 0, \sigma = 0.5$



Example: Assume Z is normal with $\mu=0$, $\sigma=1$

(“Standard Normal”, $Z \sim N(0, 1)$).

- Ex1: Find $P(Z \leq 1.31)$. (Use Table 1 in O&L)

R: `pnorm(1.31)`

- Ex2: Find $P(Z > 1.72)$

R: $1 - pnorm(1.72)$

- Ex3: Find z such that $P(Z > z) = 0.95$.
(Use Table 1 from the inside out.)

R: $qnorm(0.05)$

Ex4. $P(-1 < Z < +1)$

Ex5. Find z , such that $P(-z < Z < z) = 99\% = .99$

Practice finding these using R on own

Standardizing Variables

- If Y has a normal distribution with mean μ and standard deviation σ ($Y \sim N(\mu, \sigma)$),
- Then $Z = (Y - \mu)/\sigma$ has a standard normal distribution ($Z \sim N(0, 1)$).
- Strategy for solving problems for non-standard normal distributions:
 - Standardize both sides (subtract mean and divide by standard deviation)
 - Calculate probabilities based on standard normal distribution using Table 1 or R function `pnorm`.

Example: Suppose $Y \sim N(\mu=5, \sigma=2)$.

- Ex6: Find $P(Y \leq 8)$

R: `pnorm((8-5)/2)` or `pnorm(8,mean=5,sd=2)`

- Ex7: Find y such that $P(Y \leq y)=0.975$.

R: `2*qnorm(0.975)+5` or `qnorm(0.975,mean=5,sd=2)`

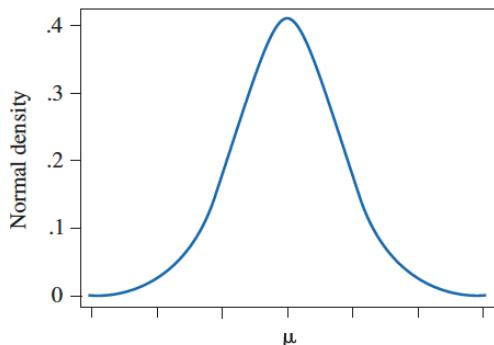
3. The Empirical Rule

For normal distributions (with mean μ and standard deviation σ):

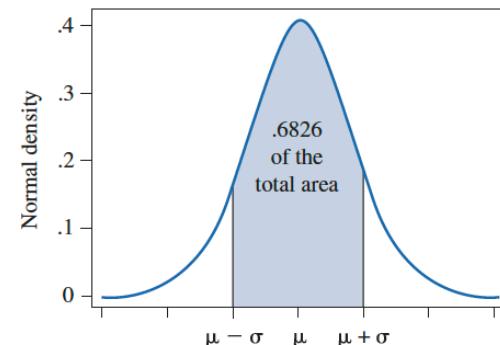
Approx. 68% of the data lie within $\mu \pm 1\sigma$

Approx. 95% of the data lie within $\mu \pm 2\sigma$

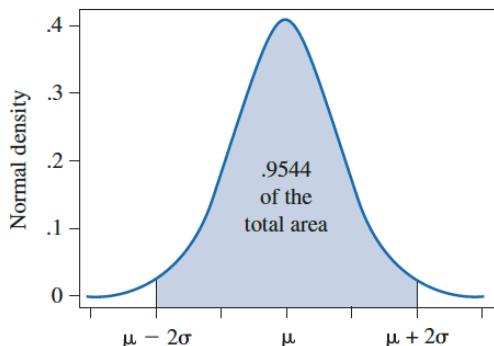
Approx. 99.7% of the data lie within $\mu \pm 3\sigma$



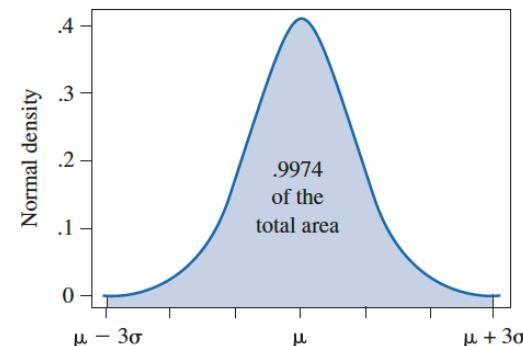
(a) Normal curve



(b) Area under normal curve within 1 standard deviation of mean



(c) Area under normal curve within 2 standard deviations of mean



(d) Area under normal curve within 3 standard deviations of mean

Chebyshev's Rule

For any distribution:

At least 75% of the data lie within $\mu \pm 2\sigma$

At least 88.8% of the data lie within $\mu \pm 3\sigma$

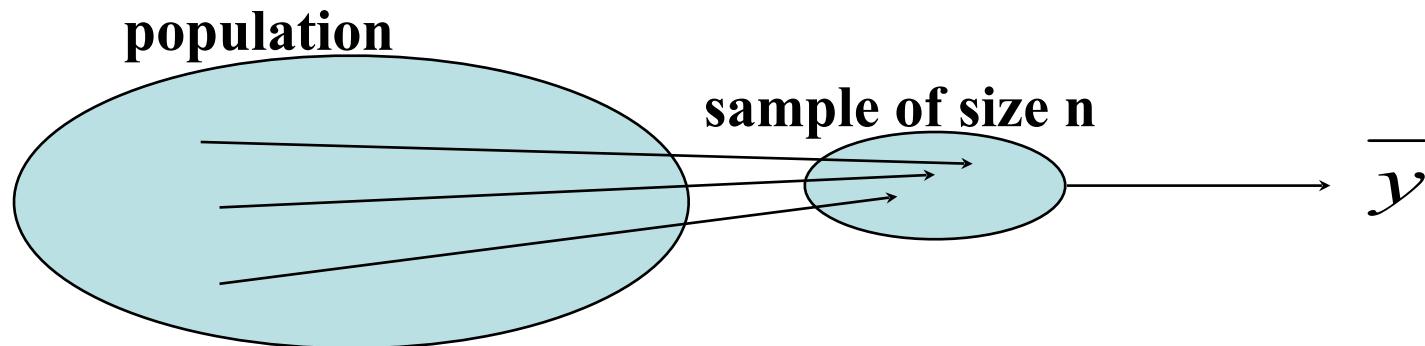
At least 93.75% of the data lie within $\mu \pm 4\sigma$

NOTES:

1. This is weaker than the Empirical rule ($75\% < 95\%$)
2. The general version of Chebyshev's rule is:

At least $(1 - 1/k^2) \times 100\%$ of the data lie within $\mu \pm k\sigma$

4. Sampling distribution of the sample mean



- We can imagine repeating the procedure (taking another sample of size n and finding another sample mean). Suppose we repeated this 1000 times. What distribution would these means have?
- In practice, we don't usually take repeat samples; we are imagining what would happen if we did, in order to better interpret the one sample we do take.

Let \bar{y} be **the mean of a sample** of size n taken from a population with mean μ and standard deviation σ .
→ \bar{y} is a statistic.

The sample mean \bar{y} is also a random variable, as it varies from sample to sample in a way that cannot be predicted with certainty.

When the sample mean is thought of as a random variable, we will write \bar{Y} , and then we write \bar{y} for the values that it takes.

If \bar{y} is the mean of a sample of size n taken from a population (Y) with mean μ and standard deviation σ , then \bar{Y} itself is a random variable.

Hence, there are two kinds of RV's being discussed here: (1) individual Y and (2) \bar{Y} .
Neither of these is assumed to be normal (so far).

- If the distribution of population is **normal**, the sampling distribution of \bar{Y} will be **exactly normal** (for **any** n).
- If the distribution of population is **non-normal**, the sampling distribution of \bar{Y} will be **approximately normal** as n gets large. (***The Central Limit Theorem***).
 - The closer the distribution of population is to normal, the smaller the n required for the distribution of \bar{y} to be approximately normal.
- **Mean** of \bar{Y} is μ .
- **Standard deviation** of \bar{Y} is σ/\sqrt{n} .

Ch5

- The Ch 5 notes includes:
 - Estimation and Standard Error (SE)
 - Confidence intervals (CI)
 - Hypothesis testing
 - Power calculations
 - And more!
- The Ott & Longnecker textbook presents CI and testing (**Z-based**) approaches for the case where σ is **known** or sample size is very **large**. We will NOT use those approaches. Instead we will focus on **t-based approaches** that work for a wider range of scenarios.

Ch 5.1: Confidence Interval for a Single Mean

1. Point estimation of a mean
 - Standard error of the estimated mean
2. Confidence interval for a mean

1. Estimate of population mean (μ)

Cattle Example:

We are interested in the level of a particular hormone in the meat for a large herd of cattle. We are specifically interested in estimating the population mean (μ).

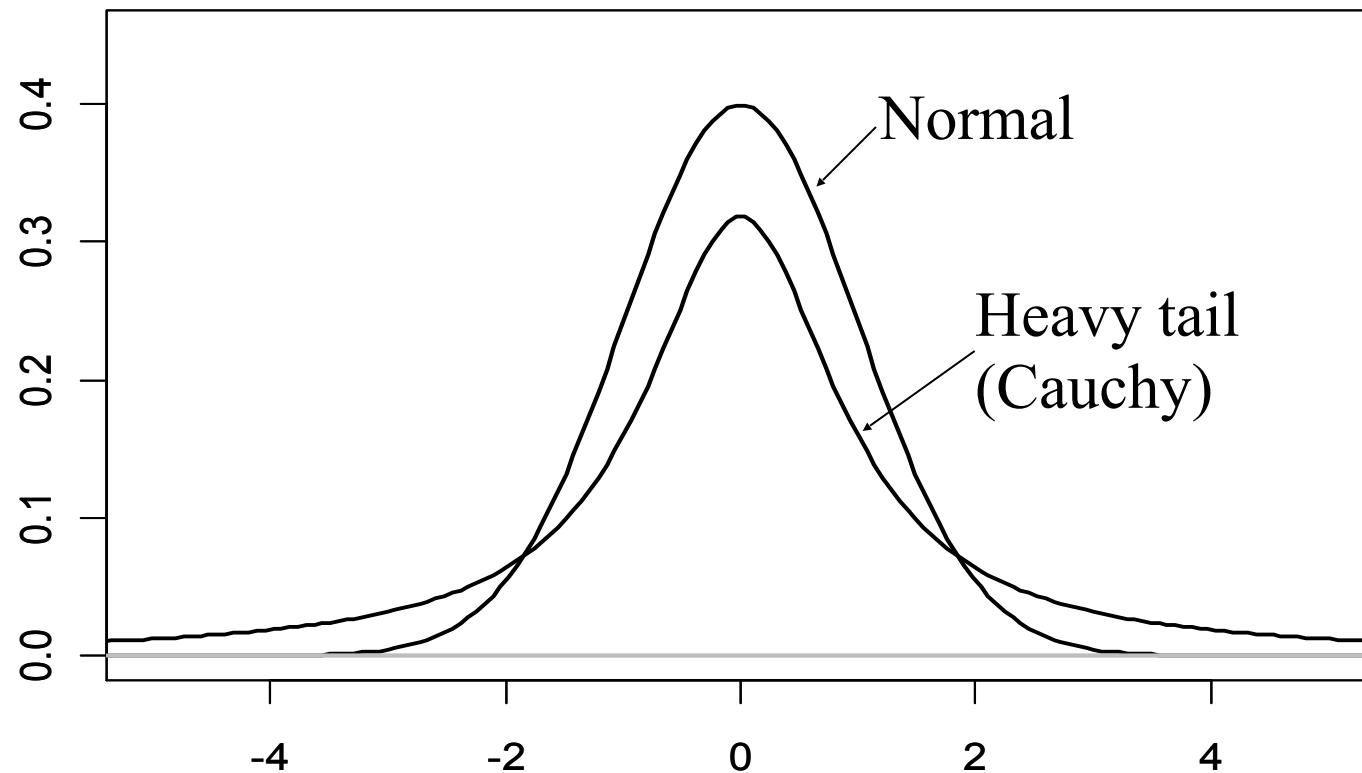
Let Y be the measured hormone level in the meat of a randomly selected animal. (Y is a random variable.)

Sample $n = 20$ values: y_1, y_2, \dots, y_n

Let: $\bar{y} = \text{sample mean} = 14.62 \text{ units}$
 $s = \text{sample std. dev.} = 2.73 \text{ units}$

Sample mean $\bar{y} = \hat{\mu}$ is almost always used as estimate of μ .

When is it not used?



Use medians (or trimmed means) to estimate the mean of heavy-tailed symmetric distributions.

Standard Error (SE) of the sample mean \bar{Y}

Standard Error (SE) is an indication of **the precision** of a (point) estimate. Often, an estimate (eg. a sample mean) is presented along with a corresponding SE. Later, we will see that the SE is used in the calculation of test statistics and confidence intervals.

The standard error of **a statistic** is (1) the standard deviation of its **sampling distribution** or (2) an estimate of that standard deviation.

Recall that $\bar{Y} \sim N(\mu, \sigma/\sqrt{n})$. The **actual/true** standard error of \bar{Y} is $\frac{\sigma}{\sqrt{n}}$.

The **standard error** of the sample mean \bar{Y} is: $SE = \frac{s}{\sqrt{n}}$, when σ is unknown.

Cattle Example:

In the cattle herd example a sample of $n = 20$ is taken. Recall:

$$\bar{y} = 14.62 \text{ and } s = 2.73$$

$$\text{Then } SE = \frac{s}{\sqrt{n}} = \frac{2.73}{\sqrt{20}} = 0.61$$

sample standard deviation (s) v.s. standard error (s/\sqrt{n})

- Use standard deviation (s) when you want to describe the sample (those individuals you collected).
- Use standard error ($SE = s/\sqrt{n}$) when you want to describe the accuracy of **sample mean** as an estimator of μ . SE is more common. SE used in formal inference (CI or test).
- As long as the sample size is given, you can calculate s from SE or vice versa. Specifically, $SE = s/\sqrt{n}$.

2. Confidence Interval for Population Mean (μ)

We will often want to make a conclusion or inference about a population parameter based on a single sample.

One of the most common types of inference is to construct what is called a **confidence interval**, which states how confident we are that this interval indeed captures the true value of the parameter.

Later, we will be looking at confidence intervals for different population parameters.

The general form:

$$\text{Estimate} \pm \text{Margin of Error}$$

where **margin of error** tells how accurate the estimate is and is based on the variability of the estimate.

Deriving a 95% Confidence Interval for Population Mean (μ)

Assume (just temporarily) that σ is known.

Apply the Empirical Rule to the sampling distribution of \bar{Y} :

After rearranging the terms using algebra we get

This gives,

a random interval that will contain the true μ with probability approximately 0.95.

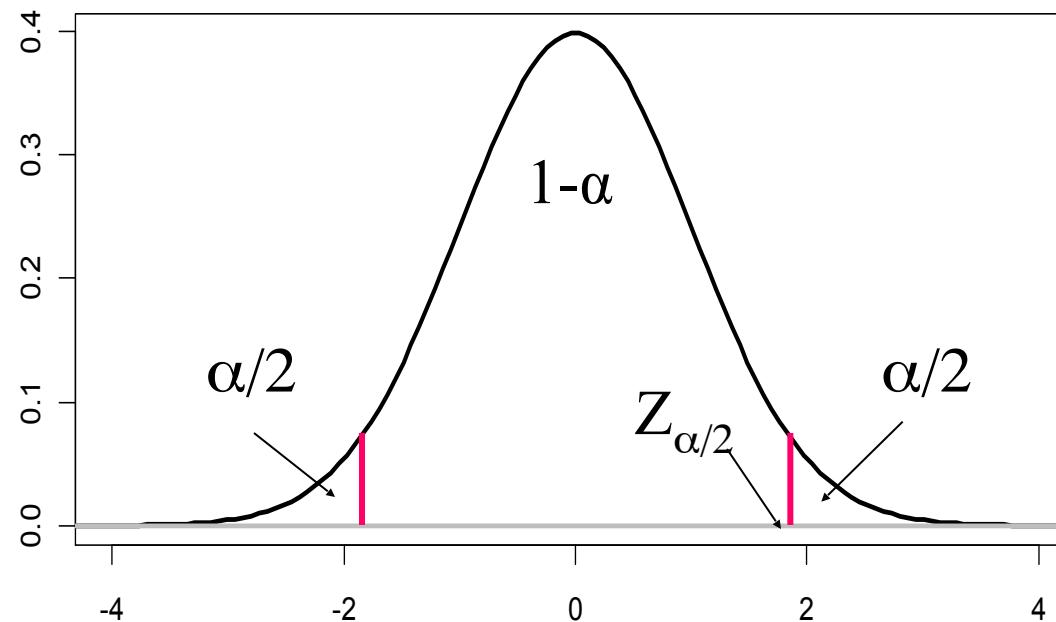
Such an interval is called a **95% confidence interval**.

Note #1: The multiplier 2 in the formula (taken from the Empirical Rule) is only approximate.

Let Z_α be the value such that the probability of being greater than Z_α is α . Hence $P(Z > Z_\alpha) = \alpha$, where α is the **significance level** and you will see this again when discussing the hypothesis testing.

For Confidence Intervals, we want the total area on both tails to be α , so we use $Z_{\alpha/2}$ as the multiplier.

Hence $P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$.



CI%	α	$\alpha/2$	$Z_{\alpha/2}$	R code
99%	0.01	0.005	2.58	<code>qnorm(0.995)</code>
95%	0.05	0.025	1.96	<code>qnorm(0.975)</code>
90%	0.10	0.05	1.645	<code>qnorm(0.950)</code>

- By default, we will use 95% confidence interval, corresponding to $\alpha = 0.05$.
- Common $Z_{\alpha/2}$ values shown here. For other values:
 - R: `qnorm(1-alpha/2)`
 - R: `qnorm(alpha/2, lower.tail = FALSE)`
- $Z_{\alpha/2}$ indicates an *upper* tail area of $\alpha/2$. This corresponds to a “greater than” probability.
- But by default, R returns “less than” probabilities. To change this, use `lower.tail = FALSE` option as shown above.

Note #2: The previous interval cannot be used for data analysis because **it requires the value of true σ** which cannot be determined from the data.

When we replace σ with its estimate (s), we use **Student's t distribution** (instead of the normal distribution) to construct the confidence interval.

The exact shape of a Student's t distribution depends on a quantity called **degrees of freedom** (df) which is related to sample size. Specifically for CI for single mean (or one-sample t-test), $df = n - 1$.

Using quantity $t_{\alpha/2, n-1}$ value, the interval between $-t_{\alpha/2, n-1}$ and $+t_{\alpha/2, n-1}$ contains $100(1-\alpha)\%$ of the area under the curve.

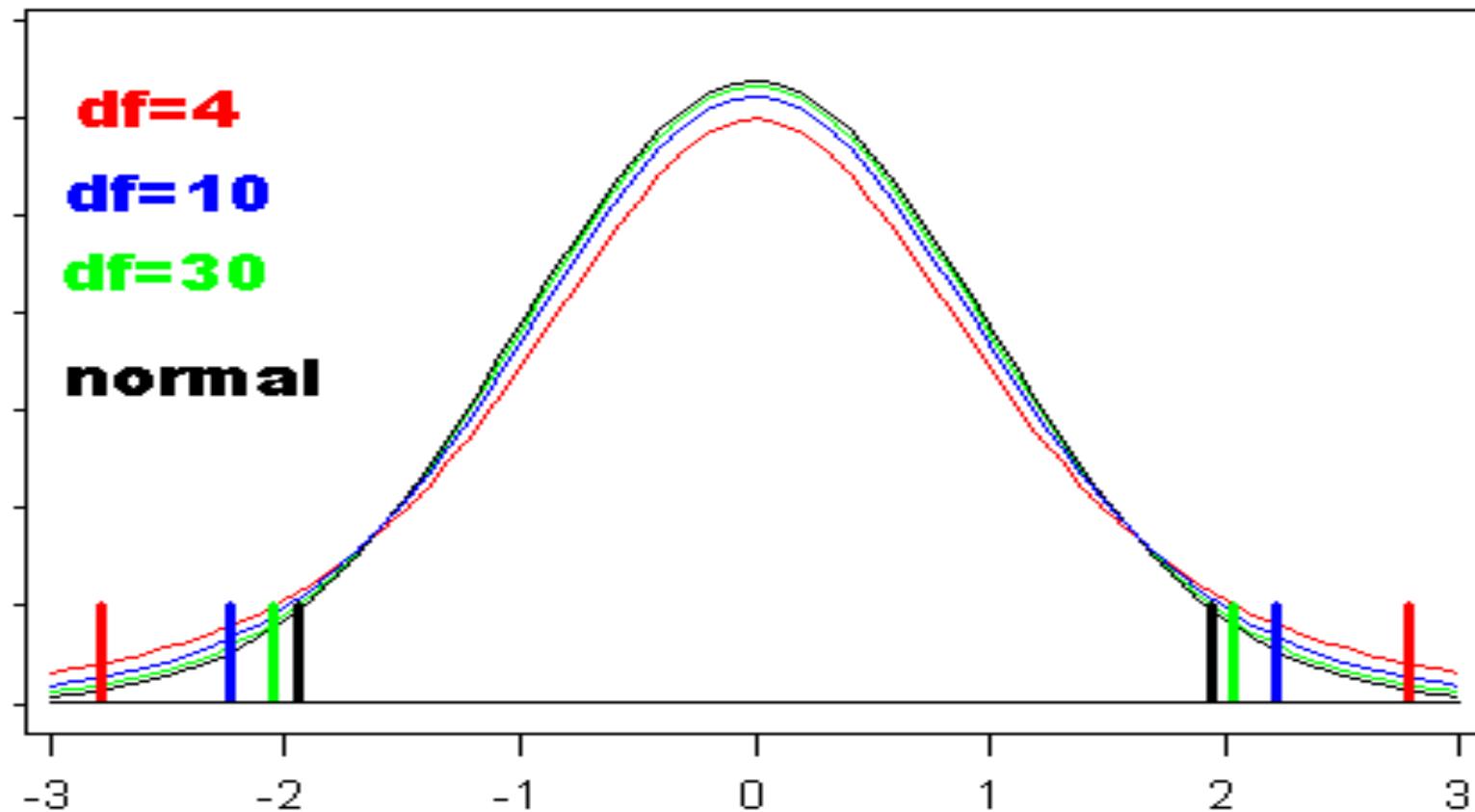
Hence $P(-t_{\alpha/2, n-1} < t < t_{\alpha/2, n-1}) = 1-\alpha$.

For $t_{\alpha/2}$ “table values”:

- R: `qt (1-alpha/2, df)`
- R: `qt (alpha/2, df, lower.tail = FALSE)`

The t distribution is symmetric and bell shaped like the normal curve, but has a larger standard deviation.

As the df increase, the t -curves approach the normal curve; thus the normal curve can be regarded as a t curve with infinite df ($df=\infty$).



The lines are mark the edge of a 95% interval for each distribution.

The $(1-\alpha)100\%$ confidence interval for μ is:

$$\bar{y} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

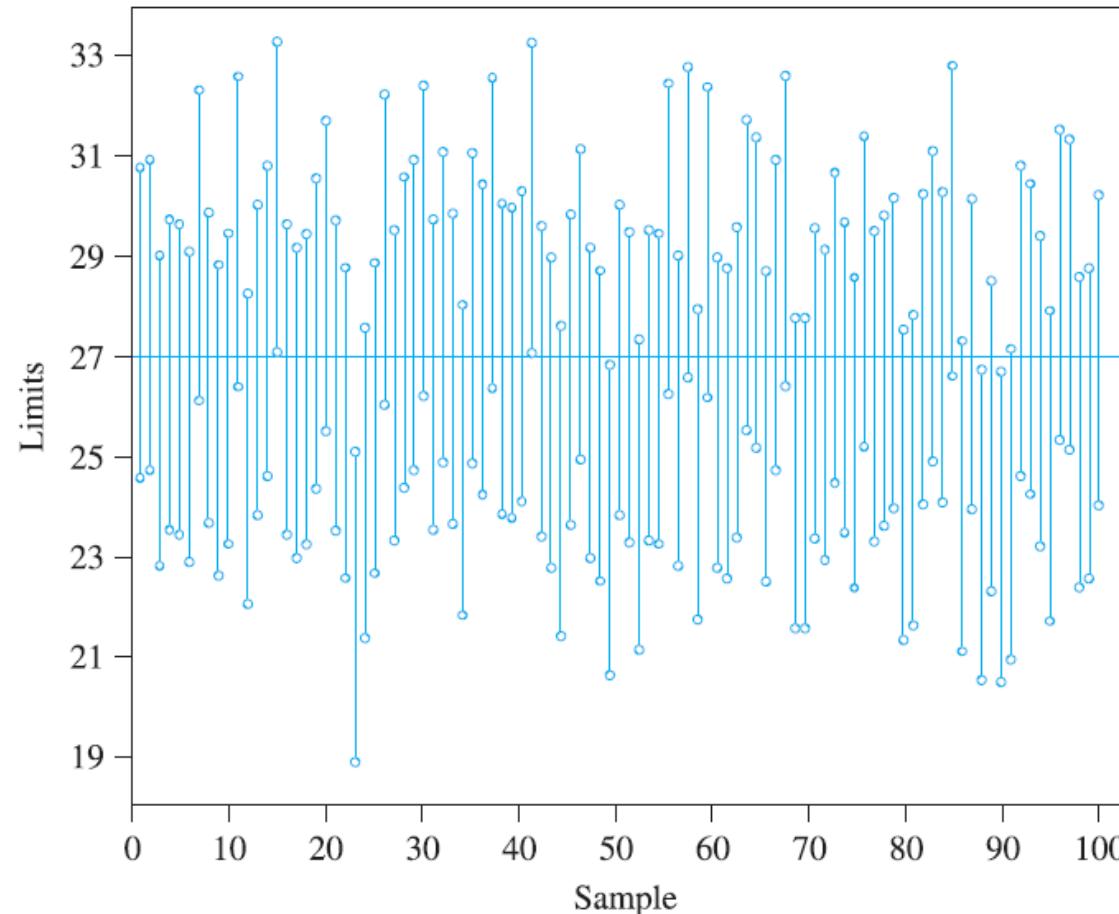
where the “table value” $t_{\alpha/2, n-1}$ is determined from the Student’s t-distribution with $df = n-1$.

- Margin of Error = $ME = t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$
- **Interpretation:** We can be 95% confident that μ is contained in this interval.
- **Assumptions:** Random sample, independent observations, normally distributed data and/or large sample size.

Confidence Interval Interpretation

- We can be 95% confident that μ is contained in a specific interval, which means that if we were to repeat our estimation method on many random samples, 95% of the intervals would contain the true population mean.
- Ott & Longnecker also discuss a simulation showing that approximately 95% of intervals capture the true mean (μ).

FIGURE 5.3
Fifty interval estimates of the population mean (27)



Cattle Example (“by hand” for illustration):

$$\bar{y} = 14.62, s = 2.73, n = 20$$

Interpretation:

We are 95% confident that the true population mean hormone concentration is between 13.34 and 15.90.

CI for a single mean using R

```
> t.test(CattleData$Hormone)
```

One Sample t-test

data: CattleData\$Hormone

t = 23.951, **df = 19**, p-value = 1.174e-15

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

13.34239 15.89761

sample estimates:

mean of x

14.62

By default we get 95% confidence interval, but also a test (p-value). We will discuss the one-sample t-test later.

Note #3: All of the formulas assume that the distribution of individual observations is normal. If that is not true, it causes a problem because the distribution of the sample mean is not quite normal, and using Student's t distribution is not quite right.

However, the confidence intervals, even under non-normality, are generally satisfactory if the distribution is not too skewed and/or the sample size is large. (because of CLT)

There are statistical procedures (histogram, qqplots, tests of normality) that are used to assess the validity of the normality assumption. These will be discussed later.

Note #4: The CI formula given here assumes that the sample size is much smaller than the population size (so the population size is considered to be effectively infinite). A “finite population correction” is available, but not discussed in this class.

Note #5: Confidence vs Prediction

The **confidence interval** gives a range that is likely to contain the unknown **population mean**. It does not tell us anything about the distribution of individual values!

A **prediction interval** is a range that is likely to contain the response value of **a single new observation**. The prediction interval is always wider than the corresponding confidence interval because of the added uncertainty involved in predicting a single response versus mean response.

$$(1 - \alpha) 100\% \text{ Prediction Interval: } \bar{y} \pm t_{\alpha/2} s \sqrt{1 + (1/n)}$$

Ch5

- The Ch 5 notes includes:
 - Estimation and Standard Error (SE)
 - Confidence intervals (CI)
 - **Hypothesis testing**
 - Power calculations
 - And more!
- The Ott & Longnecker textbook presents CI and testing (**Z-based**) approaches for the case where σ is **known** or sample size is very **large**. We will NOT use those approaches. Instead we will focus on **t-based approaches** that work for a wider range of scenarios.

Chapter 5.2: Hypothesis Test for a Single Mean

1. Intro to Hypothesis Testing
2. Hypothesis Test for Single Mean (two-sided)
 - Using rejection region
 - Using P-value
3. Hypothesis Test v.s. Confidence Interval (two-sided)
4. Type I and Type II errors and power of a test

1. Intro to Hypothesis Testing

A hypothesis test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess.

Return to the problem of the hormone levels in cattle.

Say the owner of the herd claims that the average value for the herd is 12. That is, he is claiming that $\mu=12$. Do we have evidence against this claim?

The approach taken is to assume the claim is true (called the “**null hypothesis**”), and see if the data are consistent with that assumption, or consistent with some other value for μ (the “**alternative hypothesis**”)

Null Hypothesis: $H_0: \mu=12 (\mu_0)$

Alternative Hypothesis: $H_a: \mu \neq 12$

Notes about Hypotheses

- Hypotheses are statements about **population parameters** (ex: μ = population mean).
- H_0 , H_a , μ_0 are motivated by a specific research question. These can (should!) be specified before looking at the data.
- The **null hypothesis (H_0)** is the claim that is initially assumed to be true. Typically corresponds to the current status or understanding.
- The **alternative hypothesis (H_a)** is the assertion that is contradictory to H_0 . This hypothesis typically corresponds to new discovery or understanding.
- H_0 , H_a cover all possible outcomes.

Conclusions from a test

The null hypothesis will be rejected in favor of the alternative only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0 , we will continue to believe in the truth of the null hypothesis.

The two possible conclusions from a hypothesis-testing analysis are **Reject H_0** or **Fail to reject H_0** .

2. Overview of Formal Hypothesis Testing

1. State the null and alternative hypotheses and choose a significance level α . This can (should) be done before any data is collected. Hypotheses are based on specific research questions.
2. Collect data, check assumptions, calculate summary statistics and test statistic.
3. **A. Define the rejection region (RR) based on a table value**
OR
B. Calculate the p-value (based on test statistic)
4. Make a decision (Reject H_0 or Fail to Reject H_0) by
 - A. Comparing test statistic to the rejection region
OR
 - B. Comparing p-value to α .
5. Draw conclusions.

NOTE: In STAT511 by “default”, we will use $\alpha = 0.05$.

A formal hypothesis test for μ (“two-sided” alternative)

$$H_0: \mu = \mu_0 \quad H_a: \mu \neq \mu_0$$

Test Statistic:

$$t = \frac{\bar{y} - \mu_0}{(s / \sqrt{n})} \sim T_{n-1}, \text{ under } H_0$$

Decision:

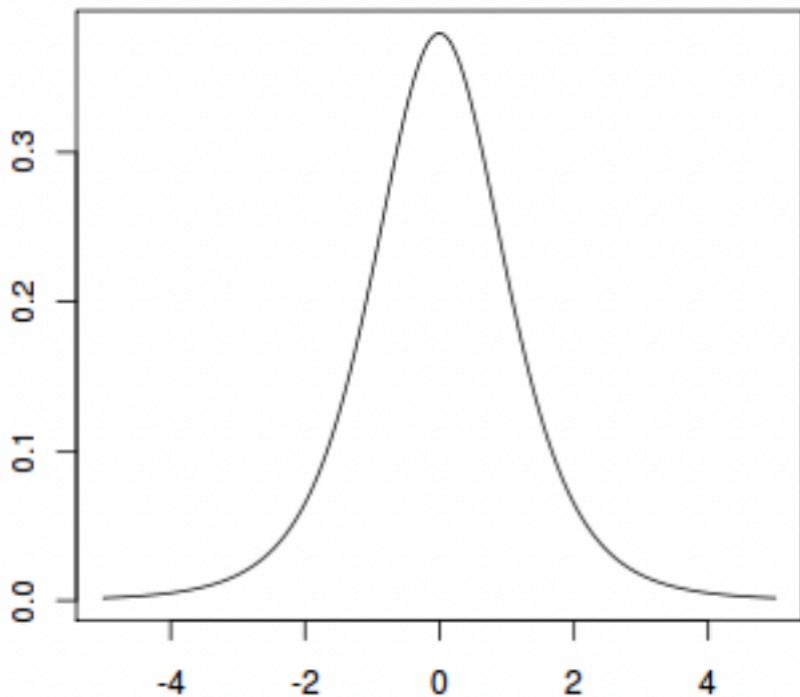
A: RR: Reject H_0 if $|t| > t_{\alpha/2, n-1}$

B: P-value: Reject H_0 if P-value = $P(|T| > |t|) < \alpha$

A: Using RR

$$H_0: \mu = \mu_0 \quad H_a: \mu \neq \mu_0$$
$$t = \frac{\bar{y} - \mu_0}{(s / \sqrt{n})} \sim T_{n-1}, \text{ under } H_0$$

RR : Reject H_0 if $|t| > t_{\alpha/2, n-1}$



Cattle example: $\bar{y} = 14.62$ & $s = 2.73$

1. State Hypotheses:

3. Test Statistic (TS):

Interpretation: \bar{y} is SE's away from the hypothesized value.

3. Define Rejection Region (RR):

4. Conclusion:

B: Using P-value

Review of P-value:

p-value is probability of observing a value of the test statistic **as or more supportive of H_a** than the actual observed value, **given H_0 is true.**

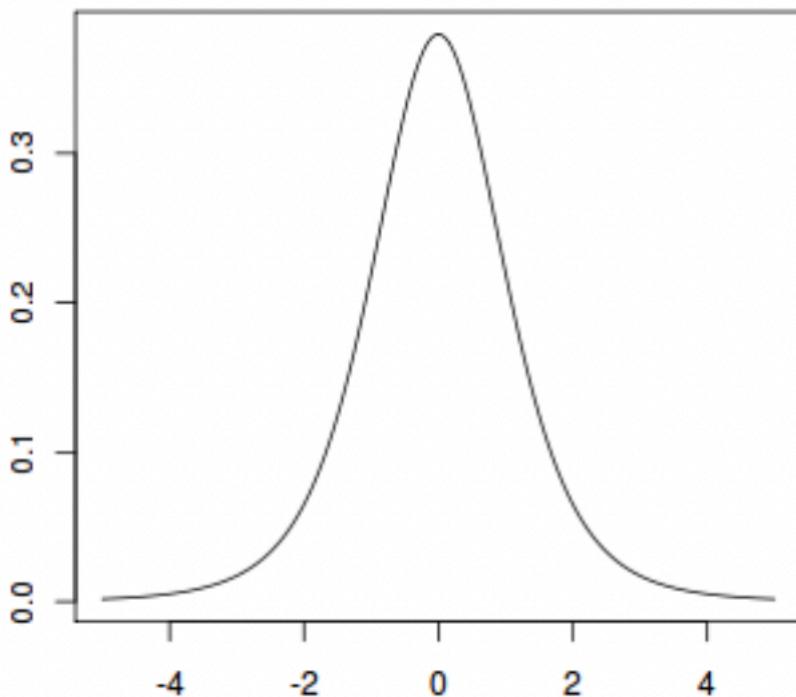
If p-value is small enough, we conclude that it would have been very unusual to observe such an extreme value of the test statistic if the null hypothesis was true.

Hence, small p-values support the alternative hypothesis.

$$H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$$

$$t = \frac{\bar{y} - \mu_0}{(s / \sqrt{n})} \sim T_{n-1}, \text{ under } H_0$$

P-value: Reject H_0 if $P(|T| > |t|) < \alpha$



Return to the **Cattle example** ($\alpha = 0.05$, $n = 20$, $df = 19$):

1. **Hypotheses:** $H_0: \mu = 12$ vs $H_a: \mu \neq 12$
2. **Test Statistic:** $t = +4.29$
3. **P-value:** area under curve outside the interval $(-t, t)$ (for 2-sided tests)
4. **Conclusion:**

One-sample t-test using R

```
> t.test(CattleData$Hormone, mu = 12)
```

One Sample t-test

```
data: CattleData$Hormone
t = 4.2911, df = 19, p-value = 0.0003943
alternative hypothesis: true mean is not
equal to 12
95 percent confidence interval:
13.34208 15.89792
sample estimates:
mean of x
14.62
```

The “mu = 12” option specifies $H_0: \mu=12$ vs $H_a: \mu\neq12$.
The confidence interval is not effected by the choice of mu.

3. Two-sided Test v.s. Confidence Interval

If H_0 is true at level α , then we would expect the null hypothesized mean (μ_0) to be within the $(1-\alpha)100\%$ confidence interval of μ (with some high level of confidence).

If μ_0 falls outside the confidence interval, we Reject H_0 .

If μ_0 falls within the confidence interval, we Fail to Reject H_0 .

Cattle Example: 95% CI = (13.34, 15.90)

$H_0: \mu=12$ ($=\mu_0$) vs $H_a: \mu \neq 12$

Since $\mu_0=12$ falls outside the confidence interval, we Reject H_0 .

We have evidence that $\mu \neq 12$. In other words, we have evidence against the owner's claim.

Note: Since our interval will not contain μ about 5% of the time, we must acknowledge that there is a 5% chance, an $\alpha=0.05$ probability, that our interval will cause us reject by mistake.

Comments

1. Usually hypothesis tests are set up so that the thing we want to prove is the alternative hypothesis. But “Failure to reject H_0 ” does not mean we really believe it is true.
2. NEVER report just a p-value! Estimate, SE and sample size are required for the reader to make sense of the results. We will discuss ASA guidance about p-values in the CH6 notes.
3. In STAT511 we will often ask you to make a conclusion (Reject H_0 or Fail to Reject H_0). But this type of conclusion is almost never included in an article.
4. Most research problems do not really require a decision. They require an estimate and an indication of its accuracy. **A Confidence Interval may be a better answer than a hypothesis test.**
 - Note about the Cattle example: The storyline/motivation for the formal test is weak here. In practice, I think the confidence interval is a better fit for this scenario.

4. Type I and Type II Errors and Power

Decision	Truth	
	H_0 True	H_0 False
Reject H_0	$P(\text{Type I Error}) = \alpha$	$P(\text{Correct}) = 1 - \beta = \text{Power}$
Fail to Reject H_0	$P(\text{Correct}) = 1 - \alpha$	$P(\text{Type II Error}) = \beta$

- **Type I error** is the error of rejecting H_0 , when H_0 is true. We denote $P(\text{Type I error}) = \alpha$. False positive.
- **Type II error** is the error of not rejecting H_0 , when H_0 is false. We denote $P(\text{Type II error}) = \beta$. False negative.
- **Power** is the probability of rejecting H_0 , when H_0 is in fact false. Power = $1 - \beta$.
- Common Procedure: Control α at a very small value (0.05 is the typical, but 0.01 and 0.10 are also used), and let β fall where it may.

Significance Level (α) v.s. The Power

- significance level (α)

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

- Power

$$1 - \beta = 1 - P(\text{type II error}) = P(\text{reject } H_0 \mid H_0 \text{ is false})$$

For the **Cattle example**:

$$H_0: \mu = 12 \text{ vs } H_a: \mu \neq 12$$

Reject H_0 if $|t| > 2.093$.

1. $P(\text{Type I Error}) = \alpha$ is the probability rejecting H_0 , when H_0 is true.
2. Power ($= 1 - \beta$) is the probability of rejecting H_0 , when H_0 is in false (hence H_a is true). In order to calculate power we need to conjecture a specific alternative. For example: $\mu = 15$. More on power later.

Ch5

- The Ch 5 notes includes:
 - Estimation and Standard Error (SE)
 - Confidence intervals (CI)
 - **Hypothesis testing**
 - Power calculations
 - And more!
- The Ott & Longnecker textbook presents CI and testing (**Z-based**) approaches for the case where σ is **known** or sample size is very **large**. We will NOT use those approaches. Instead we will focus on **t-based approaches** that work for a wider range of scenarios.

Chapter 5.3: Inference for a Single Mean

1. One-sided tests
2. One sample t-test summary
3. Checking the assumption of normality
4. Simulation for two-sided test and CI

1. Intro to One-sided tests

A **two-sided test** is of the form:

$$H_0: \mu = \mu_0 \text{ vs } H_a: \mu \neq \mu_0$$

A **one-sided tests** are of one of these forms:

$$H_0: \mu \leq \mu_0 \text{ vs } H_a: \mu > \mu_0$$

$$H_0: \mu \geq \mu_0 \text{ vs } H_a: \mu < \mu_0$$

The form of the **test statistic is the same**, regardless of whether we are interested in a one- or two-sided alternative.

But the **rejection region and p-value are different**.

The Rejection Region is only on the side supporting H_a .

The p-value takes area only in the direction that supports H_a .

Notes: Two-sided tests are the “default”, one-sided tests should not be used unless there is some compelling reason. The choice of hypotheses should be driven by the research question and determined before looking at the data!

Return to **Cattle example**:

Say our research hypothesis is that $\mu > 12$.

Hence $H_0: \mu \leq 12$ vs $H_a: \mu > 12$.

Test Statistic $t = +4.29$ (same as two-sided case!).

But for this “**greater than**” alterative, (large) **positive** test statistics support H_a .

Rejection Region:

Reject H_0 if $t > t_{\alpha, n-1}$.

R: `qt(1-alpha, df = n-1)`

Result: $t = +4.29 > 1.73 = t_{\alpha, n-1}$

->Reject H_0

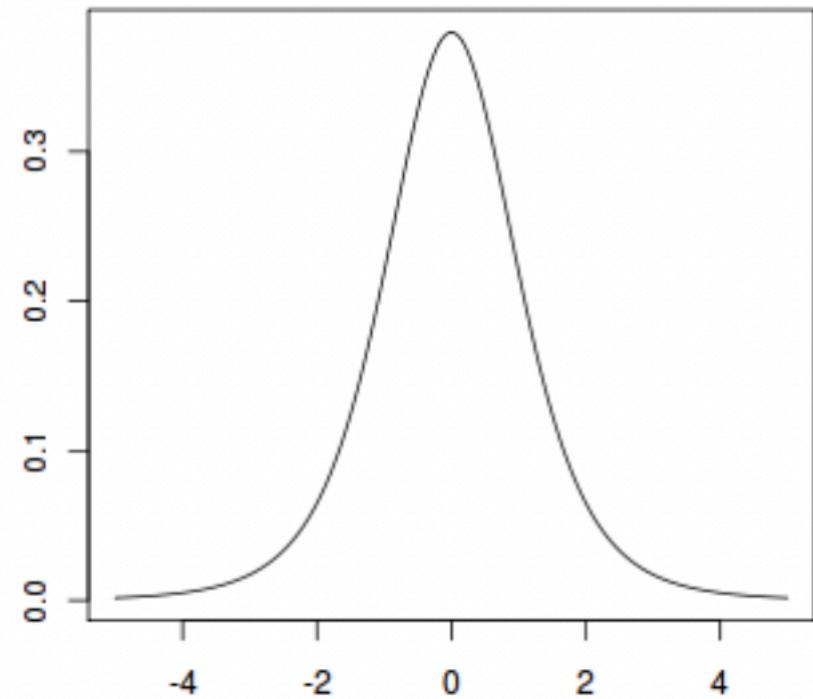
p-value:

$P(T \geq t) = P(T \geq 4.29)$

R: `1-pt(t, df=n-1)`

Result: $p = 0.002 < 0.05 = \alpha$

->Reject H_0



One-sided, One-sample t-test using R

```
> t.test(CattleData$Hormone, mu = 12,  
alternative = "greater")
```

One Sample t-test

```
data: CattleData$Hormone  
t = 4.2911, df = 19, p-value = 0.0001971  
alternative hypothesis: true mean is greater than 12  
95 percent confidence interval:  
 13.56426      Inf  
sample estimates:  
mean of x  
 14.62
```

The “mu” and “alternative” options specify $H_0: \mu \leq 12$ vs $H_a: \mu > 12$. Note that a “one-sided” confidence interval is given. We will not use these in this class.

Notes about One-sided tests

1. Two-sided tests are the “default”, one-sided tests should not be used unless there is some compelling reason.
2. The choice of hypotheses should be driven by the research question and determined before looking at the data!
3. By default, R will return two-sided p-values, but will give you the option to calculate a one-sided p-value.
4. Given the test statistic and two-sided p-value, you can also calculate the appropriate one-sided p-value.
 - Can always calculate p-value given test statistic (using `pt()`)
 - If the test statistic (or just estimated value) supports H_a , then one-sided p-value = (two-sided p-value)/2
 - If the test statistic (or just the estimated value) is on the opposite side relative to the side that supports H_a , then one-sided p-value = 1 – (two-sided p-value)/2

EPA Example: If there is evidence that the mean level of a contaminant is **above** the allowed maximum of 10, then remediation will be initiated (need to take action). Hence $H_0: \mu \leq 10$ vs $H_a: \mu > 10$.

The inspector takes $n = 30$ random samples from the area of interest. From a **two-sided** test we find: $t = -2.20$, $p = 0.036$.

Note: Since our test statistic t is negative this implies $\bar{y} < 10$. Based on this information alone, we clearly *cannot* conclude that $\mu > 10$!

How do we calculate the one-sided p-value ($P(T \geq -2.2)$)?

Option 1:

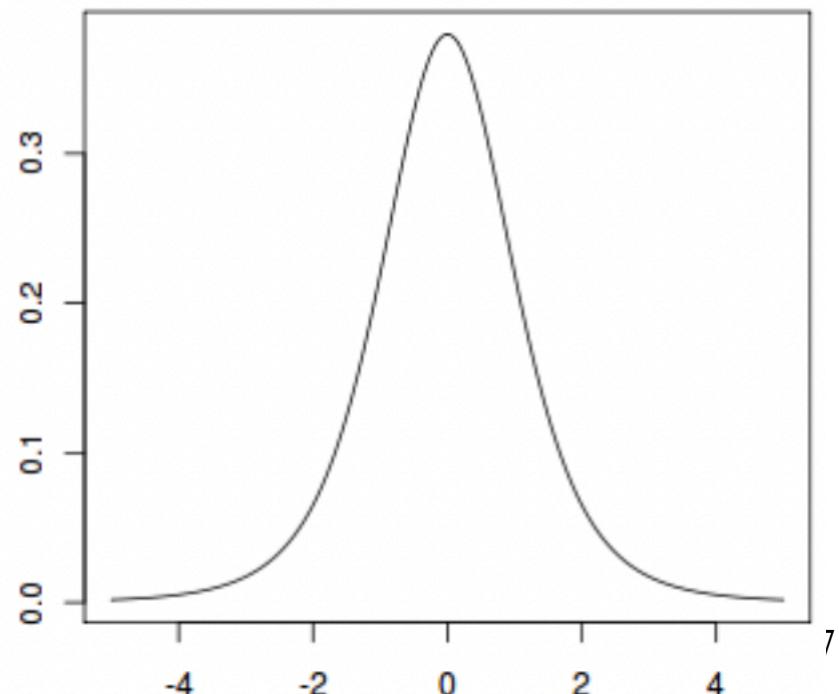
$$1 - pt(-2.20, \text{ df } = 29)$$

$$p = 0.982$$

Option 2:

$$1 - (0.036/2)$$

$$p = 0.982$$



Note about one-sided Confidence “Intervals”

- In R (and some other software packages), if you ask for a one-sided test you will get a one-sided CI.
 - Interpretation of one-sided lower limit for Cattle Example:
We can be 95% confident that the population mean is at least 13.56.
- For this class, when we ask for a confidence interval we mean a **“standard” two-sided** CI (unless specifically stated otherwise).
- Want both a one-sided test and a “standard” two-sided CI? Just run the `t.test()` function twice!

2. One-Sample t-test for Single Population Mean (μ)

Assumptions: Random sample, independent observations, normally distributed data and/or large sample size.

$$H_0 : \mu = \mu_0$$

Test Statistic: $t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}} \sim T_{n-1}$, under H_0

H_a Form:

- (1) $H_a: \mu > \mu_0$
- (2) $H_a: \mu < \mu_0$
- (3) $H_a: \mu \neq \mu_0$

Reject H_0 if:

- $t \geq t_{\alpha, n-1}$
- $t \leq -t_{\alpha, n-1}$
- $|t| \geq t_{\alpha/2, n-1}$

R code:

```
qt(1-alpha, df=n-1)  
qt(alpha, df=n-1)  
qt(1-alpha/2, df=n-1)
```

H_a Form:

- (1) $H_a: \mu > \mu_0$
- (2) $H_a: \mu < \mu_0$
- (3) $H_a: \mu \neq \mu_0$

P-value:

- $P(T \geq t)$
- $P(T \leq t)$
- $2 * P(T \geq |t|)$

R code:

```
1-pt(t, df=n-1)  
pt(t, df=n-1)  
2 * (1-pt(abs(t),  
df=n-1))
```

p-values for one and two-sided tests

For this example, the test statistic $t = +2$.

Alternative

Hypothesis:

Greater Than
 $(H_a : \mu > \mu_0)$

Less Than
 $(H_a : \mu < \mu_0)$

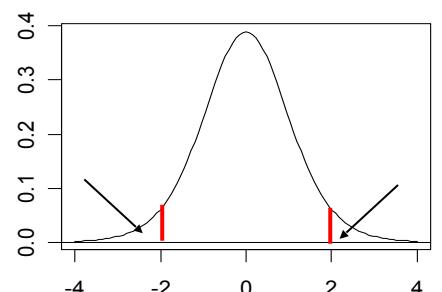
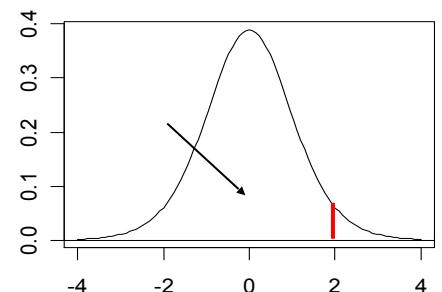
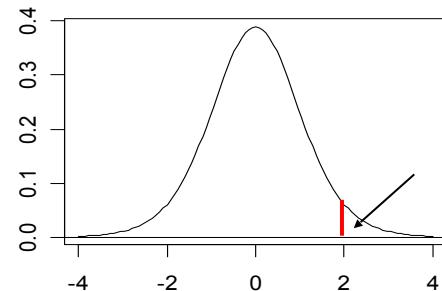
Not Equal
 $(H_a : \mu \neq \mu_0)$

P-value:

Area to the right of t

Area to the left of t

Area beyond $-t$ and $+t$



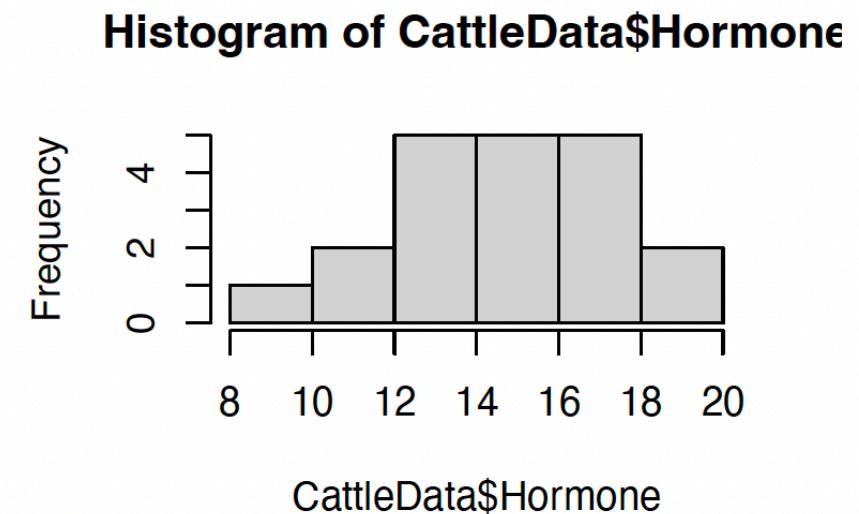
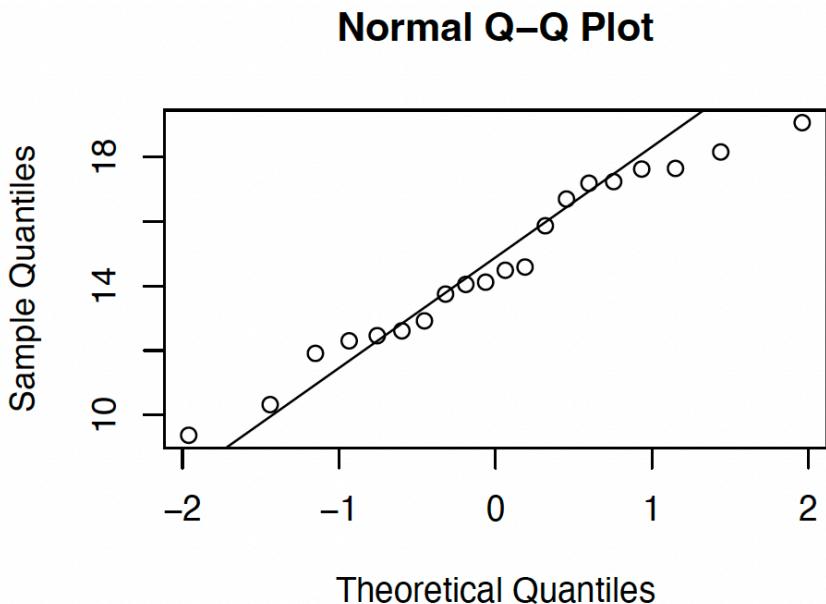
Some clarification.....

- In CH5, One-SAMPLE t-test refers to inference on a single mean (μ).
- In CH6, we will discuss the two-SAMPLE t-test for comparing two means (μ_1, μ_2) from independent samples.
- One or Two SIDED test refers to the form of H_a .
- We can have any combination of a 1 or 2 SAMPLE and 1 or 2 SIDED tests!

3. Checking Normality

A **Q-Q plot** or **Quantile-Quantile plot** is a common graphical way to check data for non-normality. Quantile is another term for percentile. A Q-Q plot is a plot of the quantiles of a data set versus the quantiles of a reference theoretical distribution.

If the plot is a straight line, it supports the idea that the data comes from the normal distribution.



Testing for Normality in R

QQplots can be generated in R using qqnorm() function.

Two options are available in R for testing for normality:

1. Shapiro-Wilk Test - shapiro.test()
2. Kolmogorov-Smirnov Test - ks.test()

See “**One Sample t-test**” example.

Notes about tests of normality:

1. **Both tests (SW, KS) use the H_0 : data are normally distributed.**
So, small p-values indicate that we should reject H_0 and conclude that the data are not normally distributed.
2. Histograms and QQ plots are usually more informative and useful than the tests, because small sample sizes generally “pass” the test (high p-value, no evidence against normality), and large sample sizes generally “fail” (small p-value, evidence against normality).

4. Simulation for two-sided test and CI

This is not a basic data analysis example! However, use of summarise (from the dplyr package) can be useful for practical data analysis.

In practice, we don't know the value of the true population mean (μ). So, we don't know whether (1) our CI has captured the true value or (2) whether we have made the correct decision in the hypothesis test.

Simulation gives us the ability to generate data where the truth is known and see how a method(s) performs.

For this simulation, we use rnorm() to generate random observations from the standard normal distribution (with mean = 0 and standard deviation = 1). Specifically, we work with 1000 samples (SampleID goes from 1 to 1000), each of size $n = 25$ where we know that $\mu = 0$.

Then for each SampleID (with $n=25$ observations), we use t.test to (1) calculate the 95% CI and (2) test $H_0: \mu = 0$. Using $\alpha = 0.05$, any p-value < 0.05 represents a false rejection (type I error).

Simulation Results

1. $952/1000 = 0.952 \approx 95\%$ of the confidence intervals include $\mu = 0$ (the true population mean). This is expected for 95% confidence intervals!
2. We find that $48/1000 = 0.048 \approx 5\%$ of the p-values are less than 0.05. This is expected when using $\alpha = 0.05$. (Still a 5% chance of a type I error or false positive.)
3. The Sample IDs that yield a CI not including $\mu = 0$ are the same as the samples that have a p-value < 0.05 . This is expected because the CI and hypothesis test will give the same conclusion for the two-sided test.
4. We find that the observed t test statistics follow a t-distribution with $df = 24$. This is the expected distribution “under the null hypothesis”.

Ch5

- The Ch 5 notes includes:
 - Estimation and Standard Error (SE)
 - Confidence intervals (CI)
 - Hypothesis testing
 - **Power calculations**
 - And more!
- The Ott & Longnecker textbook presents CI and testing (**Z-based**) approaches for the case where σ is **known** or sample size is very **large**. We will NOT use those approaches. Instead we will focus on **t-based approaches** that work for a wider range of scenarios.

Chapter 5.4: Sample Size & Power Calculations

1. Idea of Sample Size and Power Calculations
2. Sample Size through the CI
3. Power Calculation (using R)
 - One-sided Power Calculation
 - Two-sided Power Calculation
 - Determine sample size using the power

1. Idea of Sample Size and Power Calculations

- These calculations are done either before experiments (before any data is collected) or after experiments (after statistical analysis has been done).
- Before experiments: We want to determine a reasonable sample size for the study and have the desired power for a test so that we can achieve our research goals.
 - Sample size justification should match your planned analysis (and hence your research goals).
 - Calculations are based on conjectures. Coming up with reasonable conjectures is often the hardest part!
- After the data analysis, we also want to make sure that the test has a large enough power and know how large the sample we need is if we want to achieve the desired power.
- Power corresponds to a hypothesis test. Recall that power is the probability of rejecting H_0 , given H_a is true.

2. Sample Size Calculation for Desired CI

Find the n required so that **the expected width** of a $(1-\alpha)100\%$ Confidence Interval is approximately **2 margin of error** (ME):

$$ME \approx \frac{\text{the expected width}}{2}$$

A $(1-\alpha)100\%$ CI of the mean: $\bar{y} \pm ME$

where ME tells us how big/small the uncertainty involved in estimating the population parameter is.

- When σ is known, $ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. The formula of ME depends on n and σ .
 - If a conjectured value of σ is used, we can simply obtain the sample size by $n = \frac{(z_{\alpha/2})^2 \sigma^2}{ME^2}$. (It is used to quickly determine a rough sample size.)
eg. You want a 95% confidence interval of width less than 6 mm (and hence $ME < 3$ mm), and you conjecture that $\sigma = 4$ mm.

$$n > \frac{(1.96)^2 4^2}{3^2} = 6.8295 \rightarrow n = 7$$

- When σ is unknown, $ME = t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$. It depends on n and s.
- Note that $t_{\alpha/2,n-1}$ will vary if using different n.
→ cannot get the explicit formula of n from ME.
- Instead, we will consider two approaches to find n from $ME = t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$.

Using a conjectured value for s, then

Two Approaches:

- A. Use R to calculate ME for a range of n values.
- B. Iteratively solve for n: Since $t_{\alpha/2,n-1}$ depends on n, we need to (1) use starting value for $t_{\alpha/2,n-1}$ and plug into the calculation then (2) update/repeat the calculation with the updated value of $t_{\alpha/2,n-1}$ to find n.

Example for Approach A.

You want a 95% confidence interval of width less than 6 mm (and hence $ME < 3$ mm), and you conjecture that $\sigma = 4\text{mm}$.

Use R to try values of n between 5 and 15 with $s=4$.

It is also the conjectured value of s

Based on these results, a value of $n=10$ will result in a 95% $ME < 3$ (or a total CI width < 6)

n	ME
5	4.967
6	4.198
7	3.699
8	3.344
9	3.075
10	2.861
11	2.687
12	2.541
13	2.417
14	2.31
15	2.215

Example for Approach B.

$$ME = t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \rightarrow n = \frac{(t_{\alpha/2,n-1})^2 s^2}{ME^2}$$

You want a 95% confidence interval of width less than 6 mm ($E = 3$ mm), and you conjecture that $\sigma = 4$ mm.

1. Set initially $t_{\alpha/2,n-1} = 2$. This gives,

It is also the conjectured value of s

$$n = \frac{(2)^2 4^2}{3^2} = 8 \text{ (rounded up)}$$

2. With a ballpark estimate of n , we can now update $t_{\alpha/2,n-1}$.

$$df = n-1 = 8-1 = 7. \quad t_{\alpha/2,n-1} = 2.365.$$

$$n = \frac{(2.365)^2 4^2}{3^2} = 9.9 \cong 10$$

3. Check the resulting value based on $n=10$. In this case, when $n=10$ $ME = 2.861 < 3 \rightarrow OK!$

Comments:

- Before data collected, no matter which way we will consider, coming up with reasonable conjecture for σ is often the hardest part!
 - Sometimes, we may consider a very small sample of data before collecting the full data. We can use this small dataset to obtain s which then can be used as the conjecture of σ in the process of getting appreciate size for the large dataset.
- Sample size justification is often required for grant proposals, Animal Care protocols or Human Subjects committees.
- Always round sample size up (to next integer value)!
- You will also see how to use power to determine sample size soon.

3. Power Calculations

We will consider how to use R to

- Compute power for **one-sided** t-tests
- Compute power for **two-sided** t-tests
- Compute sample size from the desired power

A. Use R to compute power for one-sided t-tests.

Recall that power is the probability of rejecting H_0 , given H_a is true.

We need to make conjectures about true μ (under H_a) and σ .

(Coming up with reasonable conjectures is often the hardest part!)

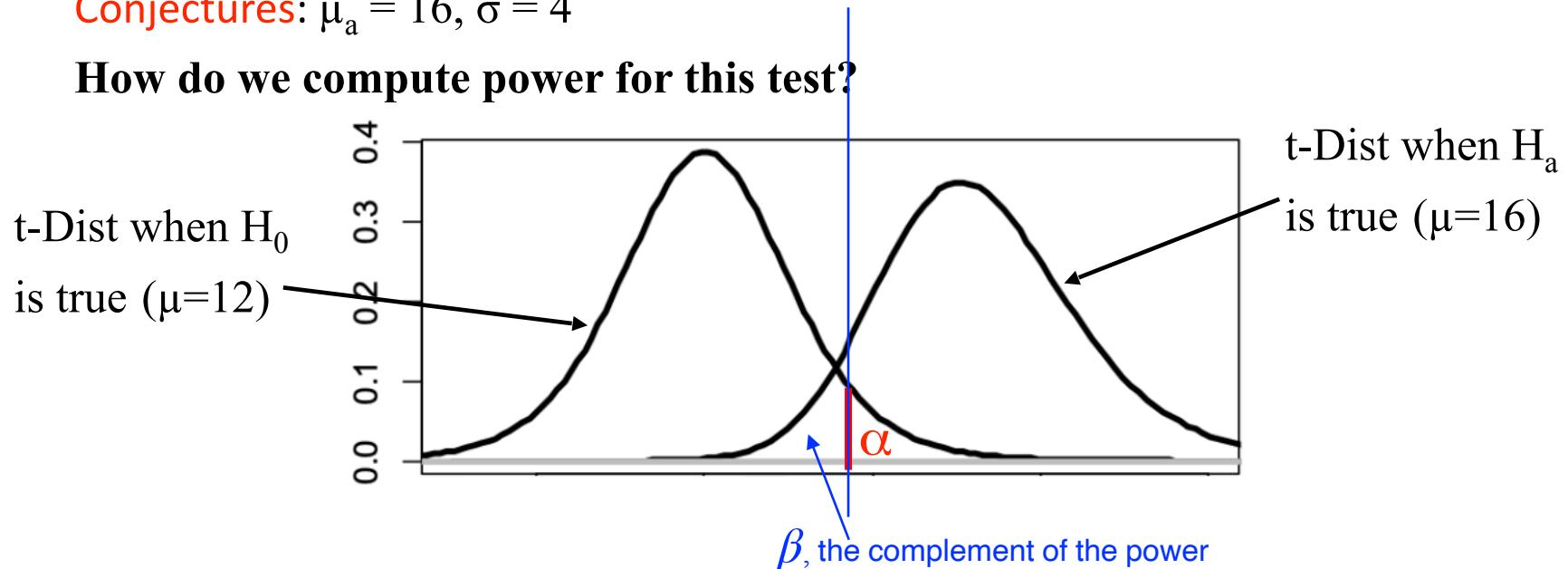
Example: $\alpha=0.05$, $n=10$, $df = 9$

$H_0: \mu \leq 12$ ($=\mu_0$) vs $H_a: \mu > 12$

Rejection Region: We will reject H_0 if $t > t_{\alpha, n-1} = 1.8333$

Conjectures: $\mu_a = 16$, $\sigma = 4$

How do we compute power for this test?



If H_a is true, then the distribution of t is **not centered at zero**; it is “non-central” with “noncentrality” parameter given by

$$\lambda = \frac{\mu_a - \mu_0}{\sigma / \sqrt{n}}$$

(**Note:** Some authors define this with an additional 2 in the denominator of λ . We follow R notation which does not have a 2 in the denominator)

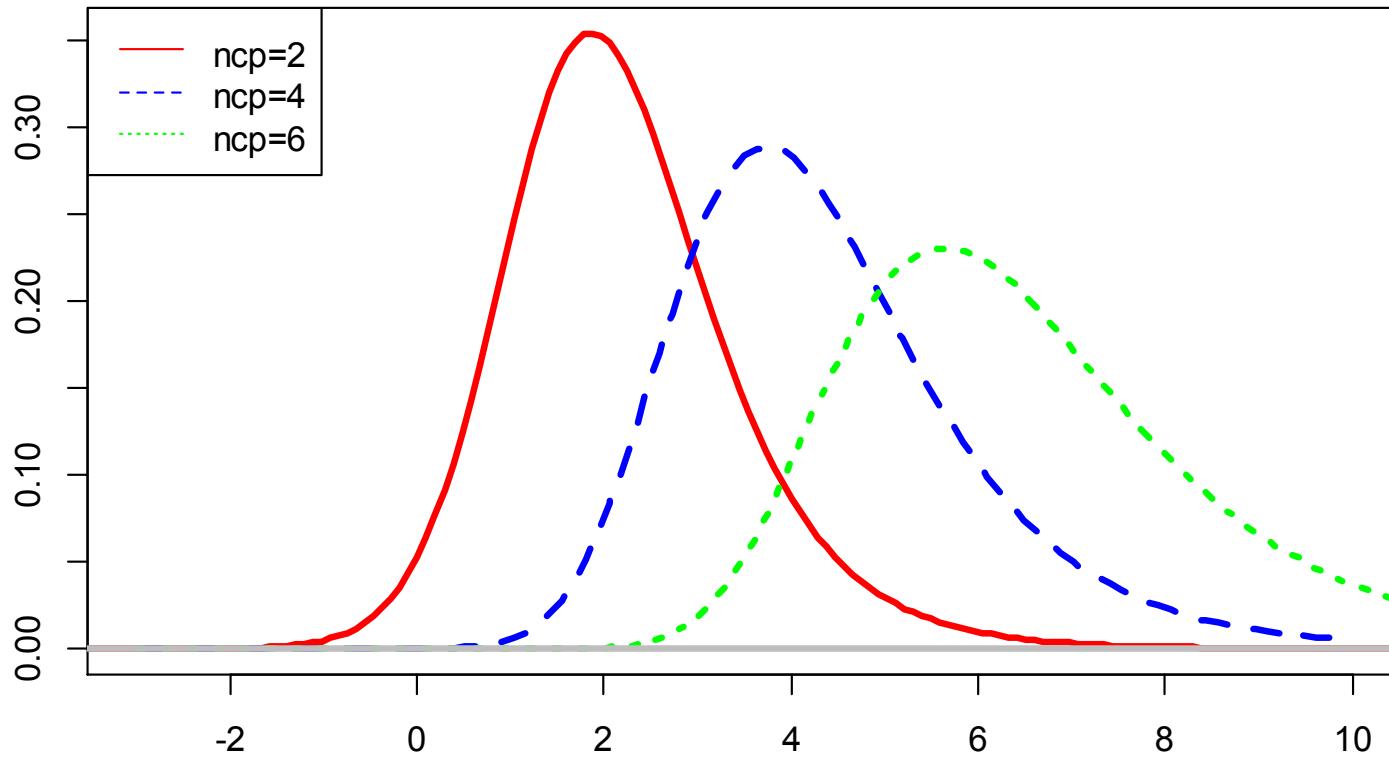
For the example, the value of λ is $\lambda = \frac{16 - 12}{4.0 / \sqrt{10}} = 3.16$

Power can then be computed in R using the function **pt**

```
power= 1-pt(1.833, df = 9, ncp = 3.16)
      = 0.898
```

In practice, we can use **power.t.test()** to compute power.

Example non-central t-distributions



Power for One-Sample t-test in R

```
>power.t.test(n = 10, delta = 4, sd = 4,  
sig.level = 0.05, type = "one.sample",  
alternative = "one.sided")
```

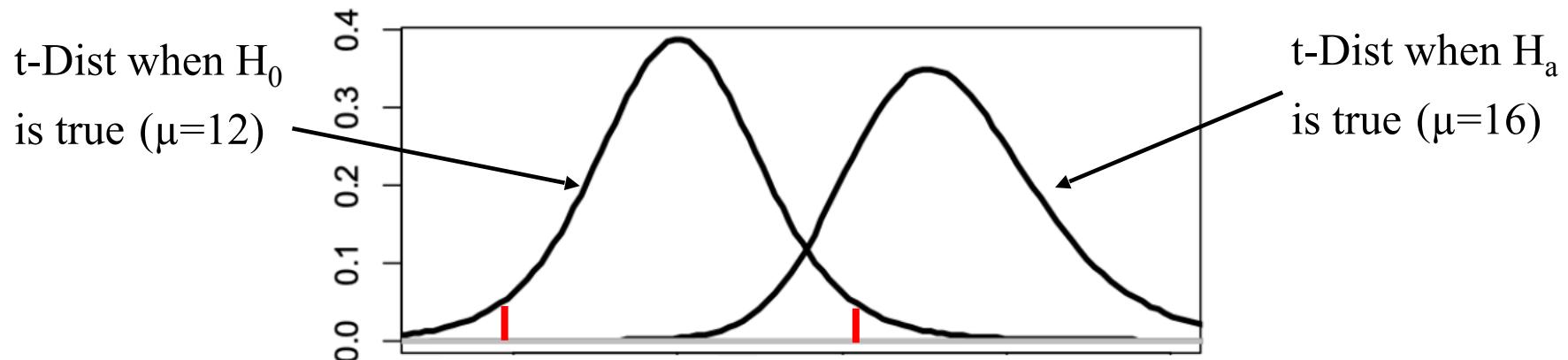
One-sample t test power calculation

```
    n = 10  
    delta = 4  
    sd = 4  
    sig.level = 0.05  
    power = 0.897517  
    alternative = one.sided
```

B. Use R to compute power for two-sided t-tests.

Power for a two-sided t-test can be computed with a minor modification of the one-sided case. $H_0: \mu = 12$ vs $H_a: \mu \neq 12$. $n=10$ (so $df=9$).

Reject if $|t| > t_{\alpha/2} = 2.262$. Sum area under the non-central curve from both tails.



Power can then be computed in R using the function `pt`

```
power= 1-pt(2.262, 9, 3.16)+pt(-2.262, 9, 3.16)  
= 0.803
```

We also can use `power.t.test()` to do this.

Using power to determine sample size

```
>power.t.test(n = 5:15, delta = 4, sig.level =  
0.05, type = "one.sample", alternative = "one.sided")
```

One-sample t test power calculation

n = 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
delta = 4 → Conjectured diff $|\mu_a - \mu_0|$
sd = 4 → Conjectured sd (σ)
sig.level = 0.05 → Significance level (α)
power = 0.5797374, 0.6769450, 0.7543959,
0.8150194, 0.8618137, 0.8975170, 0.9244891, 0.9446895,
0.9597032, 0.9707857, 0.9789162
alternative = one.sided

Sample size

The power

In practice, we can use **power.t.test()** to compute power (for fixed n) or compute n (to achieve a certain level of power).

“Factors” that affect Power

- **n (sample size):** As n increases, so does power.
- **Magnitude of difference ($|\mu_a - \mu_0|$):** As magnitude of difference increases, so does power.
- **σ (standard deviation):** As σ increases, power decreases.
- Also α and whether test is one or two-sided.

How to see this?

1. λ (non-centrality parameter): As λ increases, so does power.

$$\lambda = \frac{\mu_a - \mu_0}{\sigma / \sqrt{n}}$$

2. Graphs
3. Intuition

Comments about Power

1. **Typically, people strive to achieve 80% or 90% power.**
2. Always round sample size up (to next integer value)!
3. The power calculation is based on conjectured values! If possible, use pilot data or published articles to come up with conjectures for μ_a and σ .
4. Another way to think about the conjecture for μ_a is to think about what would be a “meaningful difference” (between $\mu_a - \mu_0$). We want our study to be powerful enough to detect a meaningful difference.
5. Here is another approach for coming up with a conjecture for σ .
Sometimes people have an expected range of values (min, max). The empirical rule (based on the normal distribution) tells us that almost all (>99%) of the data should fall within 3 standard deviations of the mean.
Then $\sigma = (\text{max} - \text{min})/6$. O&L use a more conservative denominator of 4.
6. We can do “what if?” calculations. Programs can be altered so that n is fixed and σ varies. Graphs can be added.
7. Formulas in O&L using z_α are only for larger n. We use the R which is based on the t distribution.

8. Sometimes power calculations are done to show lack of effect. In this case failure to reject H_0 is used to argue that H_0 is true. Power calculations are used to argue that if there had been an effect of a specified size, we probably would have rejected the null hypothesis. This argument is faulty. It is better to use a confidence interval to argue lack of an important effect.
9. Using `power.t.test()`, we can find the power corresponding to a certain sample size by specifying `n`. We can find the sample size needed to achieve a certain level of power by specifying `power`. Ex: `power.t.test(n=10,...)` vs `power.t.test(power=0.90,...)`
10. The R package `pwr` contains some additional power calculations.
11. Use Lenth's online power calculator to compute power.
<http://homepage.stat.uiowa.edu/~rlenth/Power/>
12. For more discussion about power and sample size justification see the article "Some Practical Guidelines for Effective Sample Size Determination" (2001) by Lenth.

Ch6 Inference comparing two population central values

- The Ch 6 notes includes:
 - Two independent samples
 - **Two-sample t-tests and CIs (notes06.1)**
 - Sample size and power calculations (notes06.2)
 - Paired samples (notes06.4)
 - Practical considerations (notes06.5)

Ch 6.1: Two independent samples - t-tests and CIs

1. Rat Lead Example

Considering **equal** variances for two populations

2. Pooled two-sample t-test
3. Pooled two-sample CI

Considering **unequal** variances for two populations

4. Welch-Satterthwaite t-test
5. Welch-Satterthwaite CI
6. Pooled v.s. Welch-Satterthwaite t-tests

1. Rat Lead Example

Rat Lead Example: Twenty rats were randomly assigned to two groups. 10 rats in the Control group received a standard diet. 10 rats in the Deficient group received a calcium deficient diet. For both groups, a 0.15% lead-acetate solution was available to drink. The amount of solution (Y) consumed by each rat was measured.

The goal of the study is to compare mean lead consumption for the two treatments. It seems obvious, but notice that use of a Control treatment is critical! This provides a benchmark to which the Deficient treatment is compared.

Summary Statistics:	
Control	Deficient
$\bar{y}_C = 5.06$	$\bar{y}_D = 8.56$
$s_C = 1.189$	$s_D = 1.471$
$n_C = 10$	$n_D = 10$

Let μ_1 = population mean for group 1
 μ_2 = population mean for group 2
 σ_1 = population standard deviation for group 1
 σ_2 = population standard deviation for group 2

We want to make inference about the **difference** between population means using sample means.

Hypothesis test: Are the means the same?

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2 & (\text{or } \mu_1 - \mu_2 = 0) \\ H_A : \mu_1 \neq \mu_2 & (\text{or } \mu_1 - \mu_2 \neq 0) \end{array}$$

Strategy:

Estimate the difference; standardize it; then evaluate whether the result is far enough from zero to reject H_0 .

Let μ_1 = population mean for group 1
 μ_2 = population mean for group 2
 σ_1 = population standard deviation for group 1
 σ_2 = population standard deviation for group 2

We want to make inference about the **difference** between population means using sample means.

Hypothesis test: Are the means the same?

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2 & (\text{or } \mu_1 - \mu_2 = 0) \\ H_A : \mu_1 \neq \mu_2 & (\text{or } \mu_1 - \mu_2 \neq 0) \end{array}$$

Strategy:

Estimate the difference; standardize it; then evaluate whether the result is far enough from zero to reject H_0 .

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\text{std. error of } (\bar{y}_1 - \bar{y}_2)}$$

General form of t :

$$t = \left[\frac{\text{est} - (\text{hyp.val.})}{\text{std. error of (est)}} \right]$$

Standard Error and Assumptions

In many sampling situations, we will select **independent** random samples from two populations. Therefore, \bar{y}_1 and \bar{y}_2 are independent. If two populations are **normally distributed** and/or we have large sample sizes, then \bar{y}_1 and \bar{y}_2 are normally distributed.

Since \bar{y}_1 and \bar{y}_2 are **independent** then their sum or difference is also normally distributed:

$$Var(\bar{y}_1 - \bar{y}_2) = Var(\bar{y}_1) + Var(\bar{y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If we assume that $\sigma_1 = \sigma_2 = \sigma$ (**equal variance**) then

$$Var(\bar{y}_1 - \bar{y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Hence the std. error of $(\bar{y}_1 - \bar{y}_2)$ = $s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

We estimate σ by a “pooled estimate”

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(Think of s_p^2 as a weighted average of s_1^2 and s_2^2)

$$df = n_1 + n_2 - 2$$

Standard Error and Assumptions

In many sampling situations, we will select **independent** random samples from two populations. Therefore, \bar{y}_1 and \bar{y}_2 are independent. If two populations are **normally distributed** and/or we have large sample sizes, then \bar{y}_1 and \bar{y}_2 are normally distributed.

Since \bar{y}_1 and \bar{y}_2 are **independent** then their sum or difference is also normally distributed:

If we assume **equal variances** of two populations then

Hence the std. error of $\bar{y}_1 - \bar{y}_2$

We estimate σ by a “pooled estimate”

2. Pooled Two-Sample t-test (Equal Variances)

Assumptions: Independent random samples, equal variances, normally distributed data and/or large sample sizes.

Test Statistic: $t = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{df}$ under H_0

where $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ and $df = n_1 + n_2 - 2$.

Reject H_0

Reject region:

$$t \geq t_{\alpha, df}$$

$$t \leq -t_{\alpha, df}$$

$$|t| \geq t_{\alpha/2, df}$$

P-value:

$$P(T \geq t) < \alpha$$

$$P(T \leq t) < \alpha$$

$$2 * P(|T| \geq |t|) < \alpha$$

R code:

`qt(1-alpha, df)`

`qt(alpha, df)`

`qt(1-alpha/2, df)`

R code:

`1-pt(t, df)`

`pt(t, df)`

`2 * (1-pt(abs(t), df))`

Rat Lead Example: Two-sided alternative

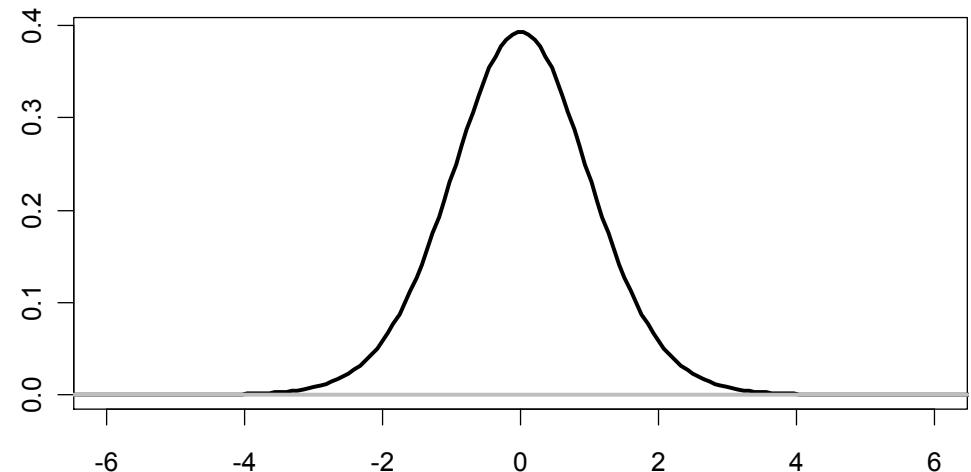
1. Hypotheses:

2. Test Statistic:

3. Rejection Region:

4. Conclusion:

Summary Statistics:	
Control	Deficient
$\bar{y}_C = 5.06$	$\bar{y}_D = 8.56$
$s_C = 1.189$	$s_D = 1.471$
$n_C = 10$	$n_D = 10$



Rat Lead Example: Two-sided alternative

1. Hypotheses:

$$H_0: \mu_C - \mu_D = 0 \text{ vs } H_a: \mu_C - \mu_D \neq 0$$

2. Test Statistic:

$$s_p = \sqrt{\frac{9(1.189)^2 + 9(1.471)^2}{18}} = 1.337$$

$$t = \frac{(5.06 - 8.56) - 0}{1.337 \sqrt{\frac{1}{10} + \frac{1}{10}}} = -5.85$$

3. Rejection Region:

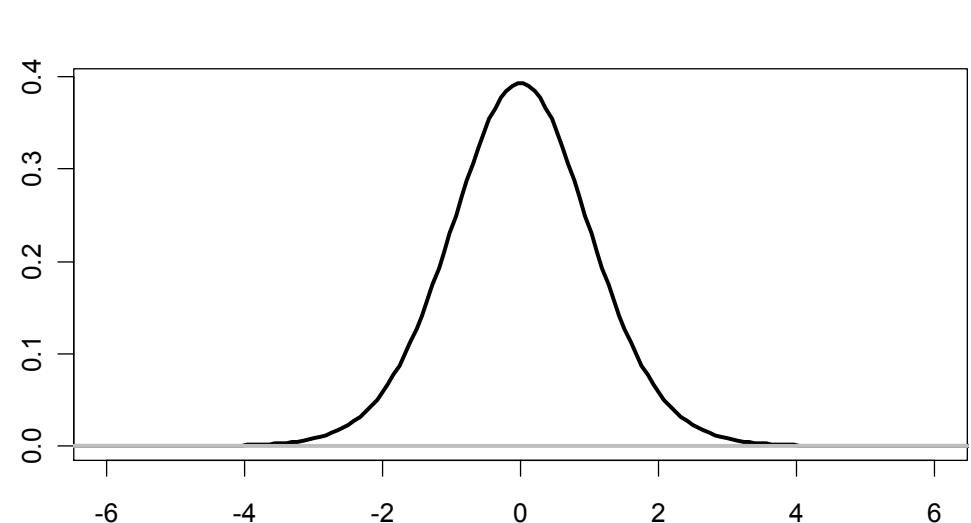
$$\alpha = 0.05, df = 10 + 10 - 2 = 18$$

Reject H_0 if $|t| > t_{\alpha/2, df} = 2.101$

4. Conclusion:

$$|t| = 5.85 > t_{\alpha/2, df} = 2.101$$

Reject H_0 . We have evidence of a difference between the true population means. Rats on calcium deficient diet consume more lead solution.



p-value (in R):

$$\begin{aligned} p &= 2 * (1 - pt(5.85, df=18)) \\ &= 0.000015 \end{aligned}$$

Rat Lead Example: One-sided alternative

Suppose we want only to consider the research alternative that Deficient group consumes **more** of the lead acetate solution than the Control group. (In practice, this should be decided in advance!)

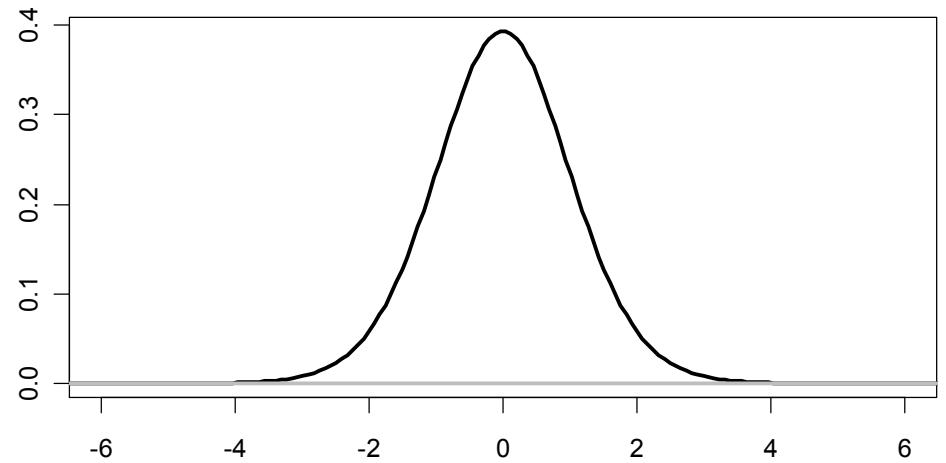
1. Hypotheses:

2. Test Statistic (same as before):

3. Rejection Region:

4. Conclusion:

Summary Statistics:	
Control	Deficient
$\bar{y}_C = 5.06$	$\bar{y}_D = 8.56$
$s_C = 1.189$	$s_D = 1.471$
$n_C = 10$	$n_D = 10$



Rat Lead Example: One-sided alternative

Suppose we want only to consider the research alternative that Deficient group consumes **more** of the lead acetate solution than the Control group. (In practice, this should be decided in advance!)

1. Hypotheses:

$$H_0: \mu_C - \mu_D \geq 0 \text{ vs } H_a: \mu_C - \mu_D < 0$$

2. Test Statistic (same as before):

$$t = \frac{5.06 - 8.56}{1.337 \sqrt{\frac{1}{10} + \frac{1}{10}}} = -5.85$$

3. Rejection Region:

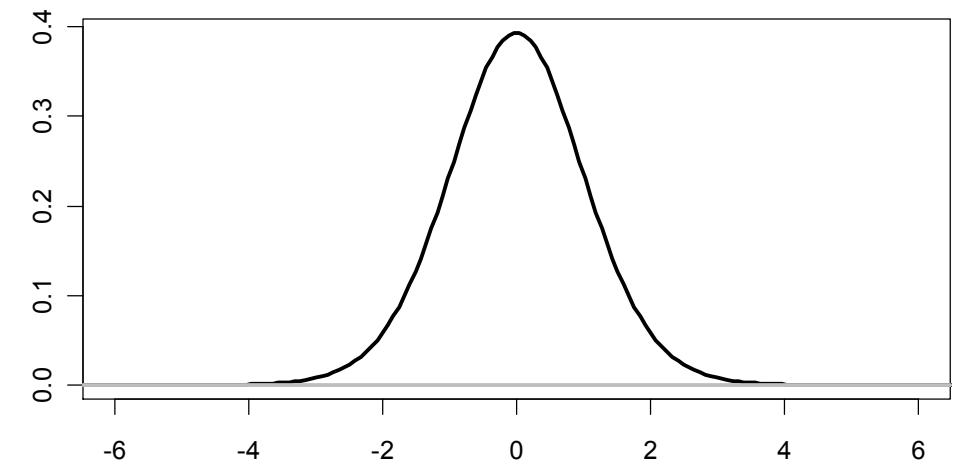
$$\alpha=0.05, df=10+10-2=18$$

Reject H_0 if $t < -t_{\alpha, df} = -1.734$.

4. Conclusion:

$$t = -5.85 < -t_{\alpha, df} = -1.734$$

Reject H_0 . We have evidence that $\mu_C < \mu_D$. Rats on calcium deficient diet consume more lead solution.



p-value (in R):

$$\begin{aligned} p &= pt(-5.85, df=18) \\ &= 0.000007 \end{aligned}$$

3. Pooled Two-Sample CI (Equal Variances)

The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2, df} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the table value $t_{\alpha/2, df}$ is determined from the Student's t-distribution with $df = n_1 + n_2 - 2$.

Assumptions: Independent random samples, equal variances, normally distributed data and/or large sample sizes.

Rat Lead Example (95% CI):

Summary Statistics:

Control	Deficient
$\bar{y}_C = 5.06$	$\bar{y}_D = 8.56$
$s_C = 1.189$	$s_D = 1.471$
$n_C = 10$	$n_D = 10$

3. Pooled Two-Sample CI (Equal Variances)

The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2, df} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the table value $t_{\alpha/2, df}$ is determined from the Student's t-distribution with $df = n_1 + n_2 - 2$.

Assumptions: Independent random samples, equal variances, normally distributed data and/or large sample sizes.

Rat Lead Example (95% CI, df=18):

$$(5.06 - 8.56) \pm (2.101)(1.337) \sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$-3.50 \pm 1.26 \rightarrow (-4.76, -2.24)$$

Since the CI does not include 0, we have evidence that there is a difference between the population means. The CI will give same conclusion as two-sided test.

Pooled Two-Sample t-test and CI in R

```
> t.test(y ~ trt, var.equal=TRUE, data=ratlead)
```

Two Sample t-test

```
data: y by trt
t = -5.8507, df = 18, p-value = 1.532e-05
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
-4.756813 -2.243187
sample estimates:
mean in group control mean in group deficient
5.06                      8.56
```

4. Welch-Satterthwaite t-test (Unequal variances)

Assumptions: Independent random samples, unequal variances, normally distributed data and/or large sample sizes.

Test Statistic: $t = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ under H_0

where $df' = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (1 - c)^2(n_1 - 1)}$ where $c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}$

Reject H_0

H_a Form:

- (1) $H_a: \mu_1 - \mu_2 > \Delta_0 \quad t \geq t_{\alpha, df'}$
- (2) $H_a: \mu_1 - \mu_2 < \Delta_0 \quad t \leq -t_{\alpha, df'}$
- (3) $H_a: \mu_1 - \mu_2 \neq \Delta_0 \quad |t| \geq t_{\alpha/2, df'}$

H_a Form:

- (1) $H_a: \mu_1 - \mu_2 > \Delta_0 \quad P(T \geq t) < \alpha$
- (2) $H_a: \mu_1 - \mu_2 < \Delta_0 \quad P(T \leq t) < \alpha$
- (3) $H_a: \mu_1 - \mu_2 \neq \Delta_0 \quad 2 * P(|T| \geq |t|) < \alpha$

R code:

`qt(1-alpha, df')`
`qt(alpha, df')`
`qt(1-alpha/2, df')`

R code:

`1-pt(t, df')`
`pt(t, df')`
`2 * (1-pt(abs(t), df'))`

5. Welch-Satterthwaite t CI (Unequal variances)

The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2, df'} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the table value $t_{\alpha/2, df'}$ is determined from the Student's t-distribution with $df' =$ Welch-Satterthwaite df.

Assumptions: Independent random samples, normally distributed data and/or large sample sizes.

Welch-Satterthwaite t-test and CI in R

```
> t.test(y ~ trt, var.equal=FALSE, data=ratlead)
```

Welch Two Sample t-test

```
data: y by trt
t = -5.8507, df = 17.241, p-value = 1.822e-05
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
-4.760793 -2.239207
sample estimates:
mean in group control mean in group deficient
5.06                      8.56
```

Comments on the Welch-Satterthwaite t-test

1. In R the `t.test()` function has `var.equal = FALSE` as the default.
2. The use of the t-distribution is approximate for the Satterthwaite t test, but this approach is good and very common!
3. When $n_1=n_2$, then the t-test statistics will be the same whether or not we assume equal variances. The df will still be different.

5. How do we decide between Pooled or Satterthwaite tests?

Recall that the Pooled variance two-sample t-test assumes equal variances, while the Welch-Satterthwaite t-test allows unequal variances.

Pooled variance t-test is generally used unless the variances are believed to be unequal. See simulation results from O&L and later in these notes.

In Chapter 7, we will discuss a formal test of $H0: \sigma_1^2 = \sigma_2^2$

For now, you can use a rule of thumb:

If $\frac{s_{max}}{s_{min}} < 2$ then assume “equal” variances (use Pooled t-test).

If $\frac{s_{max}}{s_{min}} \geq 2$ then do not assume equal variances (use Welch-Satterthwaite t-test)

Rat Lead Example: Allowing Unequal Variances

Using the rule of thumb from the previous slide:

$$s_D/s_C = 1.471/1.189 = 1.237 < 2$$

So, in practice we could use the pooled variance t-test. But for illustration, we will rerun the analysis allowing unequal variances. Results are very similar to original analysis!

Note: Satterthwaite df' = 17.24

Hypothesis Test:

Because the sample sizes are equal, the test statistic is unchanged.

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5.06 - 8.56}{\sqrt{\frac{(1.189)^2}{10} + \frac{(1.471)^2}{10}}} = -5.85$$

We can calculate the two-sided p-value using R:

`p=2*(1-pt(5.85, df = 17.24))`

Result: p=0.000018

How Sensitive is t-test to Unequal Variances?

In these simulations, 1000 samples are from normal populations with equal **means** ($H_0: \mu_1 - \mu_2 = 0$ is true) but different **variances**. Pooled variance and Welch-Satterthwaite t-tests were run using stated $\alpha = 0.05$. Empirically estimate the Type I error rate (proportion of times that the H_0 is falsely rejected).

n_1	n_2	σ_1	σ_2	Pooled t-test	WS t-test
15	15	1	2	0.060	0.055
10	20	1	2	0.017	0.044
20	10	1	2	0.114	0.059

Conclusions from the simulation:

1. If sample sizes are equal, or nearly equal, then the effect of unequal variances on the t-test is minimal. It doesn't matter "much" whether Pooled or Welch-Satterthwaite test is used; Pooled test is preferred unless s_1^2 and s_2^2 are very different (in which case, use Welch-Satterthwaite).
2. If sample sizes are unequal, then the effect of unequal variances on the t-test is more serious:
 - A. When the group with the smaller sample size has the larger variance, there are too many false rejections. This is considered serious because there are too many false claims of significance.
 - B. When the group with the larger sample size has the larger variance, there are too few false rejections. This is considered not as serious, because we would have too few false claims of significance (conservative), but it wastes power.

Ch6 Inference comparing two population central values

- The Ch 6 notes includes:
 - Two independent samples
 - Two-sample t-tests and CIs (notes06.1)
 - **Sample size and power calculations (notes06.2)**
 - Paired samples (notes06.4)
 - Practical considerations (notes06.5)

Ch 6.2: Sample Size and Power for two-sample t-test

1. Sample size corresponding to CI Width (or ME)
2. Two-sample One-sided Power Calculation
3. Two-sample Two-sided Power Calculation
4. Writing up a sample size justification

Sample Size and Power for Two Sample t-tests

- These calculations are done either before experiments (before any data is collected) or after experiments (after statistical analysis has been done).
- Before experiments: We want to determine a reasonable sample size for the study and have the desired power for a test so that we can achieve our research goals.
 - Sample size justification should match your planned analysis (and hence your research goals).
 - Calculations are based on conjectures. Coming up with reasonable conjectures is often the hardest part!
- After the data analysis, we also want to make sure that the test has a large enough power and know how large the sample we need is if we want to achieve the desired power.
- Power corresponds to a hypothesis test. Recall that power is the probability of rejecting H_0 , given H_a is true.

We will consider several cases:

1. Find the **n (per group)** required so that the expected width of a $100(1-\alpha)\%$ CI is approximately 2 margin of error (ME).
2. Compute power of tests
 - Use R to compute power for two-sample one-sided t-test.
 - Use R to compute power for two-sample two-sided t-test.

NOTE: When planning (before collecting data), people typically use:

- $n_1=n_2$ (equal sample sizes for the two groups)
- $\sigma_1=\sigma_2$ (equal standard deviations for the two groups)

1. Sample size corresponding to CI Width/ME

Find the n required so that the expected width of a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is approximately 2ME: $ME \approx \frac{\text{the expected width}}{2}$

A $100(1-\alpha)\%$ CI of $\mu_1 - \mu_2$: $\bar{y}_1 - \bar{y}_2 \pm ME$

where ME tells us how big/small the uncertainty involved in estimating the population parameter is.

- When σ is known, $ME = z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. When $n_1 = n_2 = n$, we have $ME = z_{\alpha/2}\sigma\sqrt{\frac{2}{n}}$. The formula of ME depends on n and σ .

→ If a conjectured value of σ is used, we can simply obtain the sample size by $n = \frac{2(z_{\alpha/2})^2\sigma^2}{ME^2}$. (It is used to quickly determine a rough sample size.)

eg. You want a 95% confidence interval for the difference between μ_1 and μ_2 to have total width about 10mg (or ME=5mg), and you conjecture that $\sigma=4\text{mg}$.

$$\frac{2(1.96)^2 4^2}{5^2} = 4.917248 \rightarrow n = 5$$

This is n per group.

- When σ is unknown, $ME = t_{\alpha/2, n_1+n_2-2} s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.
- If $n_1 = n_2 = n$, we have $ME = t_{\alpha/2, 2n-2} s_P \sqrt{\frac{2}{n}}$, which is the case we consider.
- Note that $t_{\alpha/2, 2n-2}$ will vary if using different n .
 → cannot get the explicit formula of n from ME.
- Instead, we will consider two approaches to find n from

$$ME = t_{\alpha/2, 2n-2} s_P \sqrt{\frac{2}{n}}$$

Using a conjectured value for s , then

Two Approaches:

- Calculate ME for a range of n values.
- Iteratively solve for n .

Example for Approach A.

You want a 95% confidence interval for the difference between μ_1 and μ_2 to have total width about 10mg (or ME=5mg), and you conjecture that $\sigma=4\text{mg}$.

It is also the conjectured value of s_p .

Use R to try values of n between 5 and 15 with s=4.

Based on these results, a value of $n_1=n_2=7$ will result in $ME < 5$ (or a total CI width < 10)

n1	n2	ME
5	5	5.83
6	6	5.15
7	7	4.66
8	8	4.29
9	9	4.00
10	10	3.76

Example for Approach B.

You want a 95% confidence interval for the difference between μ_1 and μ_2 to have total width about 10mg (ME=5mg), and you conjecture that $\sigma=4\text{mg}$.

It is also the conjectured
value of s_p .

$$ME = t_{\alpha/2, 2n-2} s_p \sqrt{\frac{2}{n}} \rightarrow n = 2 \frac{(t_{\alpha/2, 2n-2})^2 s_p^2}{ME^2}$$

1. Take **initially** $t_{\alpha/2, 2n-2} = 2$. $n = \frac{2(2)^2 4^2}{5^2} = 5.1 \cong 5$
2. With a ballpark estimate of n , we can now update $t_{\alpha/2, 2n-2}$.

$$df=2n-2=8 \rightarrow t_{\alpha/2, 2n-2} = 2.306$$

$$n = \frac{2(2.306)^2 4^2}{5^2} = 6.8 \cong 7$$

Note: This is n per group.

3. Check the resulting value based on $n1=n2=7$. $ME = 4.66 < 5 \rightarrow \text{OK!}$

2. Two-sample One-sided Power Calculation

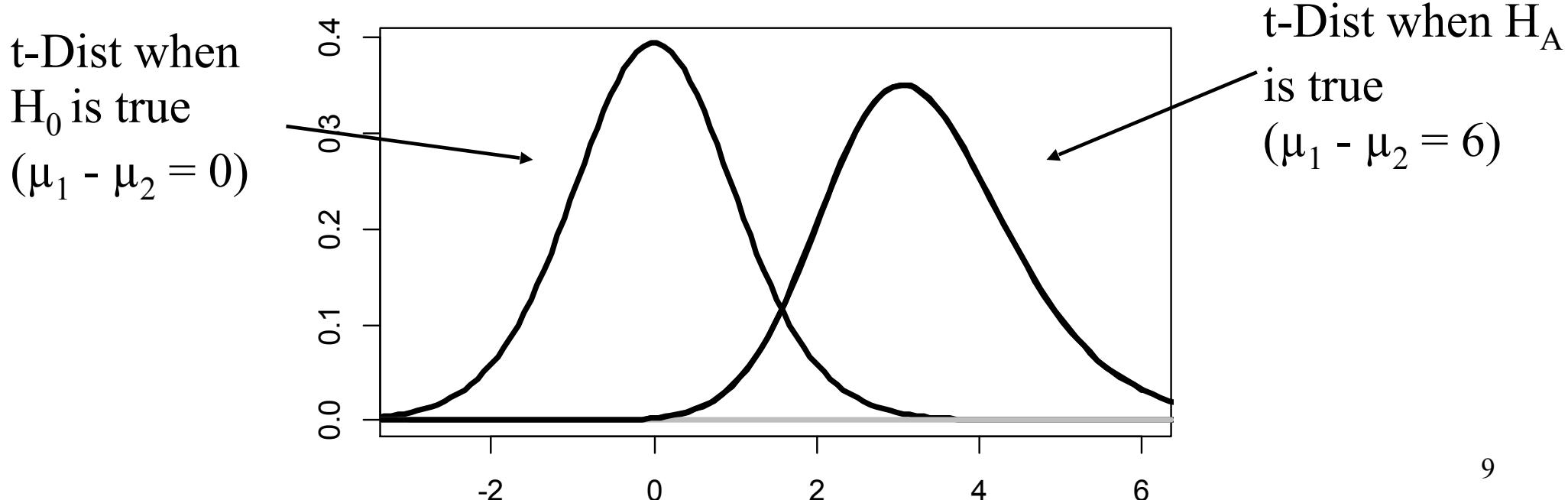
Recall that power is the probability of rejecting H_0 , given H_a is true. We need to make conjectures about true **difference $\mu_1 - \mu_2$** (under H_a) and σ . (Coming up with reasonable conjectures is often the hardest part!)

Example: $\alpha=0.05$, $n_1=n_2=n=9$, $df=9+9-2=16$

Step 1: Set up the hypothesis test:

$$H_0: \mu_1 - \mu_2 \leq 0 \quad H_a: \mu_1 - \mu_2 > 0 \quad (\mu_1 > \mu_2)$$

We will reject H_0 if $t > t_{\alpha, df} = 1.746$ (the Rejection Region).



Step 2: Identify your conjectures about σ and about the true means.

We conjecture: $\sigma = 4$, $\mu_1 = 18$, $\mu_2 = 12$. So, $\mu_1 - \mu_2 = 6$.

When the alternative is true, the distribution of t is not centered at zero; it is “non-central”, centered at its “noncentrality” parameter:

$$\lambda = \frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{2}{n}}} = \frac{18 - 12}{4.0 \sqrt{\frac{2}{9}}} = 3.18$$

The noncentrality parameter describes the difference between the true means in terms of std. deviation of $(\bar{y}_1 - \bar{y}_2)$.

Step 3: Power can then be computed in R:

power = **1-pt(1.746, df = 16, ncp = 3.18)** = 0.919

In practice, we can use **power.t.test()** to compute power (for fixed n) or n (to achieve a certain level of power).

Power for Two-Sample t-test in R

```
>power.t.test(n=9, delta=6, sd=4,  
sig.level=0.05, type="two.sample",  
alternative="one.sided")
```

Two-sample t test power calculation

n = 9 → Sample size (per group!)
delta = 6 → Conjectured diff $|\mu_1 - \mu_2|$
sd = 4 → Conjectured std deviation (σ)
sig.level = 0.05 → Significance level (α)
power = 0.9189915
alternative = one.sided

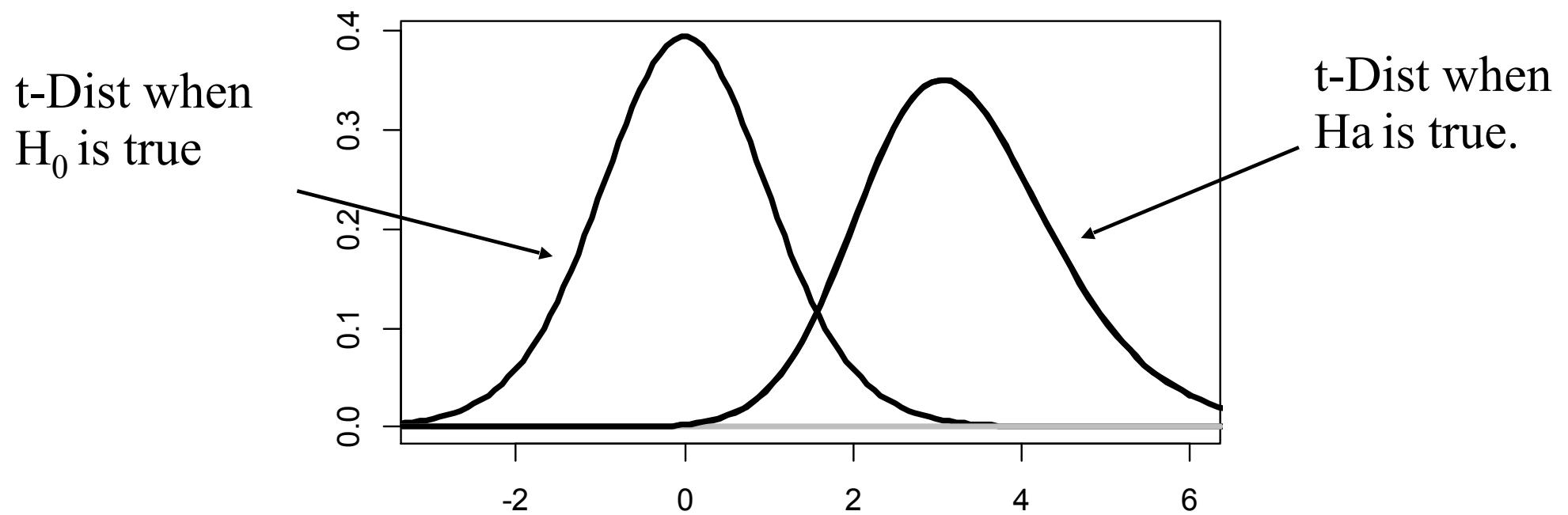
NOTE: n is number in *each* group

3. Two-sample Two-sided Power Calculation

$$H_0: \mu_1 - \mu_2 = 0 \quad H_a: \mu_1 - \mu_2 \neq 0$$

A modification/continuation of previous example.

Reject H_0 if $|t| > t_{\alpha/2, df} = 2.120$ ($df = 2*9-2=16$) and sum power from both tails.



For the example with $n = 9$ per group, the critical value with $df=16$ is $t_{\alpha/2, df} = 2.120$. Then power can be computed in R using:

```
power = pt(-2.120, df=16, ncp=3.18)
+ (1-pt(2.120, df=16, ncp=3.18))
```

We use `power.t.test()` to compute the power or sample size.

Power for $n=9$ (per group) is 0.847.

Notes:

- Power for the two-sided test is **lower** than power for the one-sided test. (This is because the one-sided test was able to “concentrate” its power in only the one direction.)
- The R package `pwr` contains some additional power calculations, including the option to allow for unequal sample sizes.
- Use Lenth’s online power calculator to compute power:

<http://homepage.stat.uiowa.edu/~rlenth/Power/>

- When using other power calculators (besides R or Lenth), watch out for whether the n per group (n_1, n_2) or total n (n_1+n_2) is given!!!!!!

4. Writing up a sample size justification

We want to give the reader enough information to recreate our results. At the same time, we want to keep things brief.

With sample size justifications, I try to keep things simple, but realistic.

In practice, I often try several “what if” calculations. But when I write things up, I tend to report the power for a single set of conditions (ex: sample size and conjectures).

If your “conjectured” values come from a published article, consider providing the reference. If your “conjectured” values come from pilot data, say so.

If you have multiple response variables, you can focus on the most important or run multiple power calculations.

Example1: We calculate power for a two-sample t-test with $\alpha = 0.05$. Based on a difference between means of 6 and standard deviation of 4 and $n = 9$ subjects per group, the power was found to be 0.85.

Example2: We calculate power for a two-sample t-test with $\alpha = 0.05$. Based on a difference between means of 6 and standard deviation of 4, to achieve 90% power, a sample size of $n=11$ subjects per group is required.

Ch6 Inference comparing two population central values

- The Ch 6 notes includes:
 - Two independent samples
 - Two-sample t-tests and CIs (notes06.1)
 - Sample size and power calculations (notes06.2)
 - **Paired samples (notes06.4)**
 - Practical considerations (notes06.5)

Ch 6.4: Comparing means using Paired Samples

1. Dog Benzedrine Example
2. Paired t-test

1. Dog Benzedrine Example

Dog Benzedrine Example (Problem 6.36 from textbook):

A study was conducted to investigate the effect of benzedrine on the heart rate of dogs. The response variable is heart rate (pbm).

A total of $n = 14$ dogs each got both placebo (P) and benzedrine (B) in a randomized order. A “wash-out” period was allowed between treatments. This is **paired data** because each dog experiences both treatments.

P	B
250	258
271	285
243	245
252	250
266	268
272	278
293	280
296	305
301	319
298	308
310	320
286	293
306	305
309	313

Summary Statistics:	
Placebo	Benzedrene
$\bar{y}_1 = 282.36$	$\bar{y}_2 = 287.64$
$s_1 = 23.14$	$s_2 = 25.39$
$n_1 = 14$	$n_2 = 14$

Comments on Dogs Example:

- This is an experiment because the researcher decides how to assign treatments (*randomly assigned*) - deliberately change the values of the input variables).
- Note that dogs were likely recruited rather than *randomly selected*.
- This is called a **paired comparison design**, which is a special case of randomized block design. - Dogs are blocks.
- The paired comparison design allows us to account for dog-to-dog variability because each dog serves as their own control. This is especially helpful with dogs (and humans!) because we expect lots of dog to dog variability (ex: small dogs vs large dogs).
- Use of placebo is important.
- Washout period is also important for this example to allow time between treatments. But not always clear how to choose.
- **Not appropriate to use two-sample t-test** here because we do not have independent observations! Instead we use **paired t-test**.

Analysis of Paired Data

When we have paired data, we consider the differences, d_1, \dots, d_{n_d} with $d_i = y_{Pi} - y_{Bi}$, for $i = 1, \dots, n$, where n_d is the same as the number of observations for each treatment.

We denote the mean of the d_i 's as \bar{d}

$$\bar{d} = \frac{\sum_{i=1}^{n_d} d_i}{n_d} = \frac{\sum_{i=1}^{n_d} (y_{Pi} - y_{Bi})}{n_d} = \bar{y}_1 - \bar{y}_2$$

- The mean of the difference is equal to the difference of the means.*

Caution: Not a two-sample t-test, because the samples are not independent.

- using paired t-test
(Indeed, it is one-sample t-test of differences d_i 's.)

P	B	Diff
250	258	-8
271	285	-14
243	245	-2
252	250	2
266	268	-2
272	278	-6
293	280	13
296	305	-9
301	319	-18
298	308	-10
310	320	-10
286	293	-7
306	305	1
309	313	-4

Summary Statistics:		
Placebo	Benzedrene	Differences
$\bar{y}_1 = 282.3$	$\bar{y}_2 = 287.64$	$\bar{d} = -5.29$
$s_1 = 23.14$	$s_2 = 25.39$	$s_d = 7.63$
$n_1 = 14$	$n_2 = 14$	$n_d = 14$

2. Paired t-test and CI

Assumptions: random sample, paired data, normally distributed differences and/or large sample size.

Confidence Interval: $\bar{d} \pm t_{\alpha/2, df} \frac{s_d}{\sqrt{n_d}}$

Test Statistic: $t = \frac{\bar{d} - \Delta_0}{s_d / \sqrt{n_d}}$

Note: n_d is the same as the number of observations for each group

$$df = n_d - 1 = \# \text{ subjects} - 1$$

H_a Form:

(1) $H_a: \mu_D > \Delta_0$

P-value

$$P(T \geq t)$$

R code:

$$1 - pt(t, df)$$

(2) $H_a: \mu_D < \Delta_0$

$$P(T \leq t)$$

$$pt(t, df)$$

(3) $H_a: \mu_D \neq \Delta_0$ $2 * P(T \geq |t|)$

$$2 * (1 - pt(\text{abs}(t), df))$$

Note: Rejection region approach can also be used.

Dog Benzedrine Example:

Test with $\alpha=0.05$:

1. $H_0: \mu_D = 0$ vs $H_a: \mu_D \neq 0$

or $H_0: \mu_P - \mu_B = 0$ vs $H_a: \mu_P - \mu_B \neq 0$

2. $t = -5.286 / (7.630 / \sqrt{14}) = -2.59$

3. P-value = 0.0224 (using R)

$$2 * (1 - pt(2.59, df = 13)) \text{ or } 2 * pt(-2.59, df = 13)$$

Since $p < \alpha=0.05$, Reject H_0 . We find evidence that there is a difference between the means.

95% Confidence Interval

$$-5.286 \pm 2.160 \frac{7.630}{\sqrt{14}}$$
$$(-9.691, -0.881)$$

Since the CI does not include 0 we conclude that there is a difference between the means.

Paired t-test and CI in R

```
> t.test(DogData$P, DogData$B, paired = T)
```

Paired t-test

```
data: DogData$P and DogData$B
t = -2.592, df = 13, p-value = 0.02234
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
-9.6912541 -0.8801745
sample estimates:
mean of the differences
-5.285714
```

Notes about Paired t-test

1. The most obvious form of pairing occurs when we observe both treatments on each subject. However, other types of pairing can exist. Classic example: identical twins.
2. The primary advantage of pairing is that differences often have smaller variability because subject to subject variability is accounted for. In other words, by accounting for subject to subject variability, we are better able to detect the treatment difference.
3. Paired t-test requires the differences to be normally distributed, not the individual observations!

Ch6 Inference comparing two population central values

- The Ch 6 notes includes:
 - Two independent samples
 - Two-sample t-tests and CIs (notes06.1)
 - Sample size and power calculations (notes06.2)
 - Paired samples (notes06.4)
 - **Practical considerations (notes06.5)**

Chapter 6.5: Practical Considerations

1. Fundamental principles of experimental design
2. Statistical vs practical significance
3. Guidance on p-values
4. Writing up results
5. Independence

1. Fundamental principles of experimental design

“Statistical” design of experiment: The process of planning the experiment so that the appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions.

Three basic principles:

1. Randomization
2. Replication
3. Blocking

Randomization

- The allocation of experimental units to treatments, is randomly determined, which prevents subjective assignment.
 - An experimental unit is a generic term that refers to a basic unit, to which a treatment is applied.
- The order in which the individual runs of the experiment are to be performed is randomly determined.
- Most statistical methods require that the observations are independently distributed random variables. Randomization usually makes this assumption valid.
- It is the best strategy when facing possible systematic variation in the experimental units.

an inaccuracy in observations which is the result of factors that are not under statistical control.

eg. testing water samples for harmful bacteria — having no control over the vitamin content that might be present and have an effect of the bacteria.

Replication

- Replication: an **independent** repeat run of each input value combination.
- Two properties :
 1. This allows the experimenter to obtain an **estimate of experimental/random error** (the fluctuation that occurs from one repetition to another), which is a basic unit of measurement to determine whether observed differences in the data are really statistically significant.
 2. If the sample mean (\bar{y}) is used to estimate the true mean response for one of the factor levels, then this allows for a more precise estimate of the parameter.
- Difference from repeated measurements:
 - **replication**: the treatment is applied to **different (multiple)** experimental units.
 - **repeated measurements**: the treatment is applied to **the same** experimental units in multiple times.

Blocking

- Blocking: a design technique which deals with nuisance factors.
- Nuisance factors are factors that may influence the experimental response but we are **not** interested in them. Blocking is used to reduce or eliminate the variability transmitted from nuisance factors.
 - eg. we want to know if caffeine really cause higher memory retention. We suspect people of similar ages might see similar effects from caffeine.
- A block is a group of homogeneous (or like) units.
 - eg. we can block these people by putting those with similar ages in the same group, i.e. young adults, middle-aged adults, senior citizens.
- Randomization is performed within each block.
- For blocking to be effective, the units should be arranged so that the within-block variation is much smaller than the between-block variation.
- In general, the strategy is **block what you know and randomize what you don't.**

2. Statistical vs practical significance

“Statistical significance” of a study is generally determined by a statistical test (or CI).

We have already seen that statistical significance depends on the magnitude of the difference ($\bar{y}_1 - \bar{y}_2$) but also the sample size (n).

Practical significance generally has to do with the magnitude of the difference.

Practical significance is somewhat harder to define (at least for me) because every reader must judge for themselves what is practically significant.

Especially when the sample size is large, it is possible to obtain a small p-value even if the estimated difference is very small. In other words, it is possible to have statistical significance without any practical significance!

3. Guidance on p-values

ASA Statement on Statistical Significance and p-values (2016)

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.
 - The p-value is **NOT** a statement about the truth of a null hypothesis. We **CANNOT** conclude that there is “no difference” based on a large p-value.
 - **Absence of evidence is not evidence of absence.**
3. Scientific conclusions and business or policy decisions should **not** be based only on whether a p-value passes a specific threshold.
 - Cannot justify “bright line” rule at $p = 0.05$ (or any other value).
 - Even rare cases that require yes/no decision, want to consider many contextual factors not just p-value!

4. Proper inference requires full reporting and transparency.
 - Conducting multiple analyses of the data and reporting only those with certain p -values (typically those passing a significance threshold) renders the reported p -values essentially uninterpretable.
5. A p -value, or statistical significance, does not measure the size of an effect or the importance of a result.
 - Statistical significance is not equivalent to scientific, human, or economic significance.
6. By itself, a p -value does not provide a good measure of evidence regarding a model or hypothesis.
 - Researchers should recognize that a p -value without context or other evidence provides limited information.

Moving to a World Beyond “ $p < 0.05$ ” (The American Statistician, 2019)

Don’t say “Statistically Significant”

- We are NOT recommending that the calculation and use of continuous p-values be discontinued.
- When p-values are used, they should be reported as continuous quantities (e.g. $p = 0.08$).
- For the integrity of scientific publishing and research dissemination, therefore, whether a p-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight.

An Idea: Consider writing most of the text for the Abstract BEFORE running the statistical analysis. This forces the author to identify the most important research questions and variables. After analysis, simply add in the actual results.

Acronym: ATOM

Accept uncertainty.

- Accompany every point estimate with a measure of uncertainty such as a standard error or interval estimate.

Be thoughtful, open and modest.

- Thoughtful researchers begin above all else with clearly expressed objectives.
- They invest in producing solid data.
- They consider not one but a multitude of data analysis techniques.
- Thoughtful research includes careful consideration of the definition of meaningful difference. As a researcher you should communicate this up front, before data are collected and analyzed.

4. Writing Up Results

We want to give the reader enough information to recreate our results. At the same time, we want to keep things brief.

We should never present just a p-value without giving the reader more information! It is critical to provide information about the estimated values, variability and sample size. Do not “round off” the p-value to $<$ or $>$ 0.05.

CI's can also be used as an alternative to hypothesis testing.

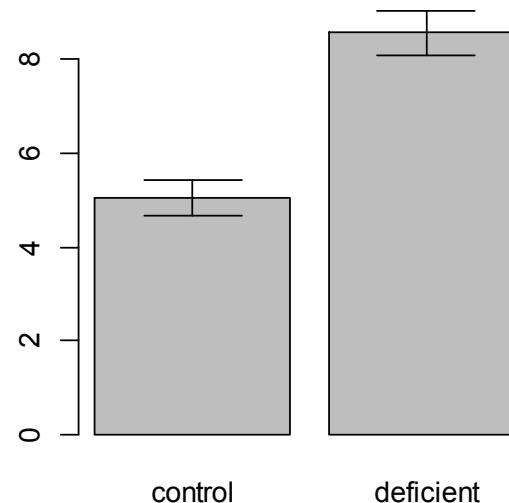
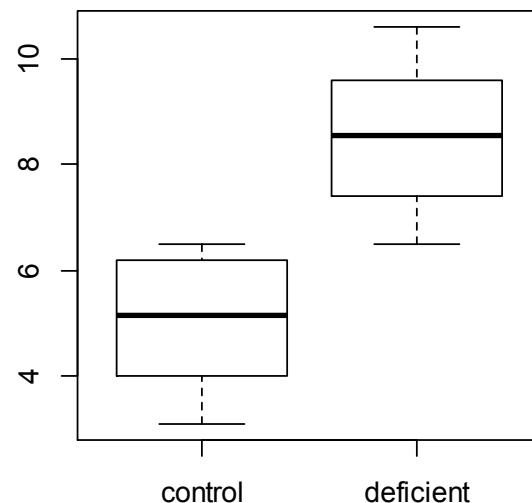
I find that the summary statistics (means, SE (or sd) and sample size) for each group are most helpful.

Only present a table or graph if there is something to say about it!

Rat Lead Example:

R software was used for statistical analysis. The two-sample t-test assuming equal variance was used to compare mean lead consumption for the two groups. The table below gives the mean (and SE) for each group and the p-value from the two-sample t-test.

Response	Control (n=10)	Deficient (n=10)	p-value
Lead Consumption	5.06 (0.38)	8.56 (0.47)	<0.001



5. Independence

What does it mean to have “independent” observations?
How can this assumption be violated?

Suppose an investigator is investigating an anti-fungal treatment on plants. They have two treatments under consideration: Mock and Active. Treatments are applied to the leaves of plants.

We consider several possible designs.

Design1: A total of 12 plants are grown to a fixed age. 6 plants are randomly assigned to receive Mock trt and 6 are randomly assigned to receive Active trt. The treatment is applied to a single leaf from each plant. We record a total of 12 measurements.

5. Independence

What does it mean to have “independent” observations?
How can this assumption be violated?

Suppose an investigator is investigating an anti-fungal treatment on plants. They have two treatments under consideration: Mock and Active. Treatments are applied to the leaves of plants.

We consider several possible designs.

Design1: A total of 12 plants are grown to a fixed age. 6 plants are randomly assigned to receive Mock trt and 6 are randomly assigned to receive Active trt. The treatment is applied to a single leaf from each plant. We record a total of 12 measurements.

Different plants -> Independent Obs -> 2 sample t ($df = 10$)

Design2: A total of 6 plants are grown to a fixed age. Each plant has one leaf treated with Active and another leaf treated with Mock. We record a total of 12 measurements.

Paired observations -> Paired t-test or CI.

Design3: 2 plants are grown to a fixed age. One plant has 6 leaves treated with Active trt. The other plant has 6 leaves treated with Mock trt. We record a total of 12 measurements.

Do not use this design! Unreplicated or “pseudo” replicated. NOT independent observations.

Design2: A total of 6 plants are grown to a fixed age. Each plant has one leaf treated with Active and another leaf treated with Mock. We record a total of 12 measurements.

Design3: 2 plants are grown to a fixed age. One plant has 6 leaves treated with Active trt. The other plant has 6 leaves treated with Mock trt. We record a total of 12 measurements.

Do not use this design! Unreplicated or “pseudo” replicated. NOT independent observations.

Design4: A total of 6 plants are grown to a fixed age. 3 plants have Active trt applied to 2 leaves. 3 plants have Mock trt applied to 2 leaves. We record a total of 12 measurements.

Two levels of replication:

“Bio” reps = plants

“Tech” reps = leaves within plants

Not independent observations, do NOT use 2 sample t-test with $df = 10$.

Option1: Average over 2 leaves per plant and use total of 6 observations for analysis. Use 2 sample t-test with $df = 4$.

Option2: Mixed model (nested) analysis discussed in STAT512.

Pseudo replication was defined by Hurlbert in 1984 as “the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated, or replicates are not statistically independent.”

From Manly:

“When dependent data are analyzed as if they are independent, the sample size used is larger than the effective number of independent observations....To avoid this, a good rule to follow is that statistical inferences should be based on only one value from each independently sampled unit, unless the dependence in the data is properly handled in the analysis.”

Chapter 7: Inference for Variance (σ^2) or Standard Deviation (σ)

1. Inference for a Single Variance or Standard Deviation
2. Comparing Two Variances or Standard Deviations

Some Perspective

CH5: Inference about a single mean (μ)

CH6: Inference about difference between two means ($\mu_1 - \mu_2$)

CH7: Inference about one or two standard deviations (σ) or variances (σ^2)

Considering parametric methods:

CH5,6: Normal and t distributions

- Both symmetric
- Positive or Negative values

CH 7: χ^2 and F distributions

- Skewed, not symmetric
- Strictly positive

Why study standard deviations (or variances)?

1. Sometimes σ_1 and/or σ_2 are the parameters of interest.

Confidence intervals or tests about a single σ .

Compare σ_1, σ_2 using confidence intervals or tests for the ratio.

Example:

Two machines - both making a part that is supposed to be x units wide.

machine 1 produces parts with mean μ_1 and std. dev. σ_1 .

machine 2 produces parts with mean μ_2 and std. dev. σ_2 .

Say, we can adjust the settings for either machine until

$\mu_1 = x$ and $\mu_2 = x$. Then the machine with the smaller std.dev. will produce the better (more consistent) product.

2. When comparing population means, some textbooks recommend formally evaluate whether the pooled (equal variance) t-test or the Welch-Satterthwaite (unequal variance) t-test should be used. Using a confidence interval for this is fine. I recommend against just using a hypothesis test to test against $H_0: \sigma_1 = \sigma_2$

A few more comments....

The Chapter 7 methods are much less commonly used compared to other material in this course. Variances and standard deviations measure how spread-out data are, and it is not so common to see real-world research projects in which assessing variance is key.

Inference about variance(s) or standard deviation(s) is more common in manufacturing settings. This includes pharmaceutical manufacturing.

Chapter 7.1: Inference for a Single Variance (σ^2) or Standard Deviation (σ)

1. Lab example
2. Chi-square test for σ^2 (or σ)
3. Confidence interval for σ^2 (or σ)
4. Simulation Study

In this course, we will see the first of 3 different “chi-square” (χ^2) tests of Chapter 7 in the textbook. The chi-square test for contingency tables (Ch10) is the one you might remember from an undergrad stats class.

In general, tests are often named after the distribution of the test statistic used. χ^2 distribution is used as a reference distribution, much like z or t .

1. Lab Example

To study the **precision** of a lab instrument we do $n = 10$ serum cholesterol determinations on the same (well mixed) blood sample. A perfect instrument would yield the same result each time, but our imperfect instrument yields: $\bar{y}=210$, $s = 10.2$.

We do not know the true mean, so we cannot evaluate whether the 210 is too high or low. **That is an “accuracy” issue.** We are interested in “precision”, i.e. variance.

The manufacturer claims that the true standard deviation of results from this instrument is 5.0 mg/dl or less. Are the data consistent with this claim, or do they contradict this claim?

This suggests a one-sided alternative (using $\sigma_0 = 5$):

$H_0: \sigma \leq 5$ vs $H_a: \sigma > 5$ or equivalently

$H_0: \sigma^2 \leq 25$ vs $H_a: \sigma^2 > 25$

Rejection of H_0 in favor of H_a would be evidence against the claim.

2. Chi-square Test for σ^2 (or σ)

Assumptions: Random sample, independent observations, normally distributed data.

H₀ Form: $\sigma^2 \leq \sigma_0^2$

Test Statistic: $\chi_0^2 = (n-1)s^2/\sigma_0^2 \sim \chi_{df}^2$, under H_0

Notes: df = n - 1, test statistic takes **only positive values**. In STAR 511, we will only consider **right-tailed** χ^2 tests.

H_a Form:

$H_a: \sigma^2 > \sigma_0^2$

Reject H₀ if:

Rejection Region:

$$\chi_0^2 > \chi_{\alpha, n-1}^2$$

R code:

`qchisq(1-alpha, df=n-1)`

P-value:

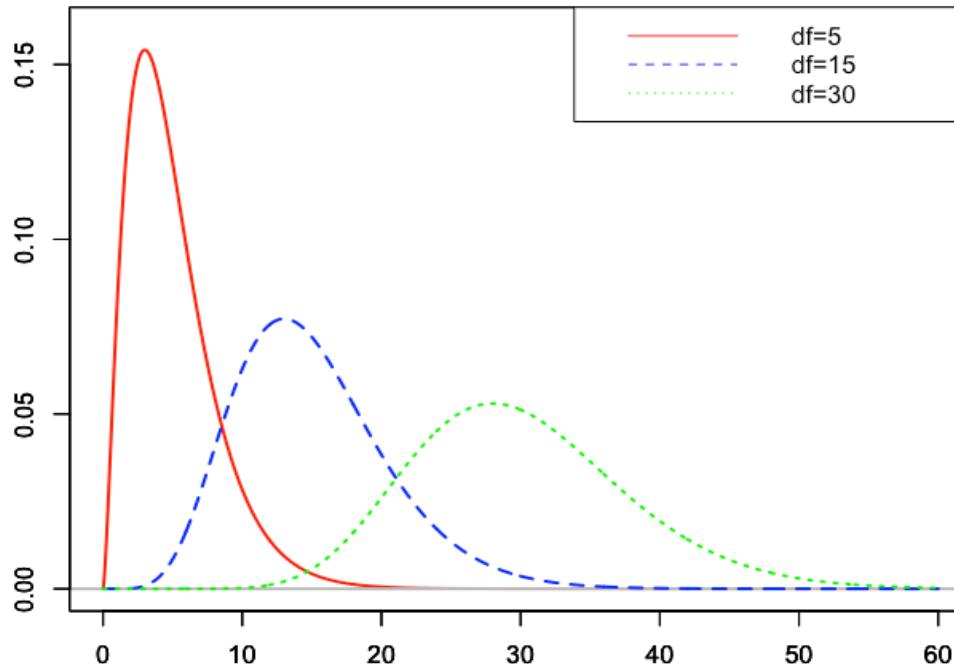
$$P(\chi^2 \geq \chi_0^2)$$

`1-pchisq(chi0^2, df=n-1)`

Comments:

1. Note that $H_a: \sigma^2 > \sigma_0^2$ is equivalent to $H_a: \sigma > \sigma_0$. But either way we use, the test statistic is calculated using s^2 and σ_0^2 .
2. If H_0 is true, then the test statistic $\chi^2 = (n-1)s^2/\sigma_0^2$ has a chi-square distribution with degrees of freedom (df) = n-1.
3. The mean and variance of the chi-square distribution are given by $\mu = df$ and $\sigma^2 = 2df$.

Example Chi-Square Distributions

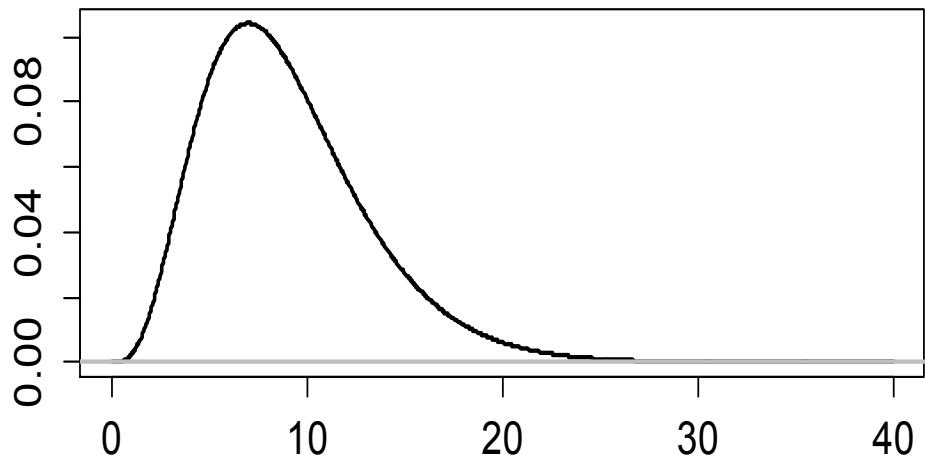


Lab Example

1. $H_0: \sigma \leq 5.0$ vs. $H_a: \sigma > 5.0$

2. Test Statistic:

$$\chi_0^2 = \frac{(n - 1)s^2}{\sigma_0^2} = \frac{(10 - 1)(10.2)^2}{5^2} = 37.45$$



3. Make a decision

(1) Rejection Region:

$$df = n - 1 = 10 - 1 = 9$$

Using R: $\chi_{\alpha, n-1}^2 = \text{qchisq}(0.95, df = 9) = 16.92$

Since $\chi_0^2 = 37.45 > \chi_{\alpha, n-1}^2 = 16.92 \rightarrow \text{Reject } H_0$.

(2) p-value:

Using R: $P\text{-value} = 1 - \text{pchisq}(37.45, df = 9) = 0.0000219$

P-value = 0.0000219 which is very small, close to 0. $\rightarrow \text{Reject } H_0$.

3. Derivation of Confidence Interval for σ^2

A Confidence interval is derived by starting with a probability statement about the test statistic, then manipulating the inequalities so that the parameter is in the middle.

The $\chi^2_{0.025,df}$ is the value corresponding to the upper area probability 0.025 of χ^2 distribution with degree freedom df.

$$P \left(\chi^2_{0.975,df} < \frac{(n - 1)s^2}{\sigma^2} < \chi^2_{0.025,df} \right) = 0.95$$

$$P \left(\frac{1}{\chi^2_{0.025,df}} < \frac{\sigma^2}{(n - 1)s^2} < \frac{1}{\chi^2_{0.975,df}} \right) = 0.95$$

$$P \left(\frac{(n - 1)s^2}{\chi^2_{0.025,df}} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_{0.975,df}} \right) = 0.95$$

The $\chi^2_{0.975,df}$ is the value corresponding to the upper area probability 0.975 of χ^2 distribution with degree freedom df.

Confidence interval for σ^2 (or σ)

Assumptions: Random sample, independent observations, normally distributed data.

Note: $df = n - 1$

The $100(1-\alpha)\%$ confidence interval for σ^2 is given by:

The $\chi^2_{\alpha/2, df}$ is the value

corresponding to the upper area probability $\alpha/2$ of χ^2 distribution with degree freedom df .

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2, df}}, \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2), df}} \right)$$

The $\chi^2_{(1-\alpha/2), df}$ is the value corresponding to the upper area probability $(1 - \alpha/2)$ of χ^2 distribution with degree freedom df .

The $100(1-\alpha)\%$ confidence interval for σ is given by:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2, df}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2), df}}} \right)$$

For Lab example, the 95% CI for σ is:

$$\sqrt{\frac{(10-1)(10.2)^2}{19.02}} < \sigma < \sqrt{\frac{(10-1)(10.2)^2}{2.700}} \quad \text{i.e., } 7.02 < \sigma < 18.62$$

$qchisq(0.975, df = 9)$

$qchisq(0.025, df = 9)$

Comments about Confidence Interval for σ

- CI **not symmetric** about the point estimate. It is skewed to the right.
- You will reject (in a two-sided test, $\alpha=0.05$) any σ not in the 95% confidence interval.
- Textbook writes $\chi^2_{\alpha/2}$ as χ^2_U and $\chi^2_{1-\alpha/2}$ as χ^2_L , referring to “upper” and “lower” table values.
- Tests and Confidence Intervals assume that individual observations are from the normal distribution.
- This hypothesis test and confidence interval are **very affected** by outlier-prone distributions (heavy-tailed distributions) and skewed distributions.

4. Simulation to study effect of Outliers and Skewness

Data simulated under H_0 . Test using $\alpha=0.05$. Type 1 error rates recorded. For details see simulation in CH7 of textbook.

$$H_0: \sigma^2 \leq 100 \text{ vs } H_a: \sigma^2 > 100$$

	Normal	Uniform	t	Gamma(1)	Gamma(.1)
n=10	0.047	0.004	0.083	0.134	0.139
n=20	0.052	0.006	0.103	0.139	0.175
n=50	0.049	0.004	0.122	0.156	0.226

Conclusion: Far too many rejections with heavy tails (t) or skewness (gamma). Far too few rejections with short tails (unif).

Chapter 7.2: Comparing Two Variances or Standard Deviations

1. F test to compare two variances (σ_1^2/σ_2^2)
2. Confidence interval for σ_1^2/σ_2^2
3. Sensitivity to assumptions
4. Levene's test to compare two variances

Lab Example 2:

We now compare two instruments that measure serum cholesterol. Machine 2 is a newer machine that is supposed to give more consistent results.

We do $n_1=10$ determinations using machine 1, and $n_2=15$ determinations using machine 2, all on the same (well mixed) sample.

We observe: $s_1=15.4$, $s_2=12.3$

Suppose we want to see if there is strong evidence that machine 2 gives more consistent results.

This suggests:

$H_0: \sigma_1^2/\sigma_2^2 \leq 1$ vs $H_a: \sigma_1^2/\sigma_2^2 > 1$ or equivalently

$H_0: \sigma_1^2 \leq \sigma_2^2$ vs $H_a: \sigma_1^2 > \sigma_2^2$ or equivalently

$H_0: \sigma_1 / \sigma_2 \leq 1$ vs $H_a: \sigma_1 / \sigma_2 > 1$ or equivalently

$H_0: \sigma_1 \leq \sigma_2$ vs $H_a: \sigma_1 > \sigma_2$

1. F test to compare two variances (σ_1^2/σ_2^2)

Assumptions: Independent, random samples, normally distributed data.

H₀ Form: $H_0: \sigma_1^2/\sigma_2^2 \leq 1$ or $\sigma_1^2 \leq \sigma_2^2$ or $\sigma_1/\sigma_2 \leq 1$ or $\sigma_1 \leq \sigma_2$

Test Statistic: $F_0 = s_1^2/s_2^2 \sim F_{df_1, df_2}$, under H_0

Note: df_1 = numerator df= $n_1 - 1$ and df_2 =denominator df= $n_2 - 1$, F_0 takes only positive values. As with χ^2 tests, we will only consider **right sided F-tests**.

H_a Form:

$$H_a: \sigma_1^2/\sigma_2^2 > 1$$

Reject H₀ if:

Rejection Region:

$$F_0 > F_{\alpha, df_1, df_2}$$

R code:

```
qf(1-alpha, df1, df2)
```

P-value:

$$P(F \geq F_0)$$

```
1-pf(F0, df1, df2)
```

Comments:

1. The F distribution looks a lot like a chi-square distribution.
2. Tests that involve ratios of variances will typically use an F-test statistic.
3. Note that $H_a: \sigma_1^2/\sigma_2^2 > 1$ is equivalent to $H_a: \sigma_1^2 > \sigma_2^2$. But either way we use, the test statistic is calculated using s_1^2 and s_2^2 .
4. If H_0 is true, then the test statistic $F_0 = s_1^2/s_2^2$ has a F distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

Lab Example 2 : Right One-sided Alternative

1. $H_0: \sigma_1^2/\sigma_2^2 \leq 1$ vs $H_a: \sigma_1^2/\sigma_2^2 > 1$

2. Test Statistic:

$$F_0 = s_1^2/s_2^2 = 15.42/12.32 = 1.57$$

3. Make a decision

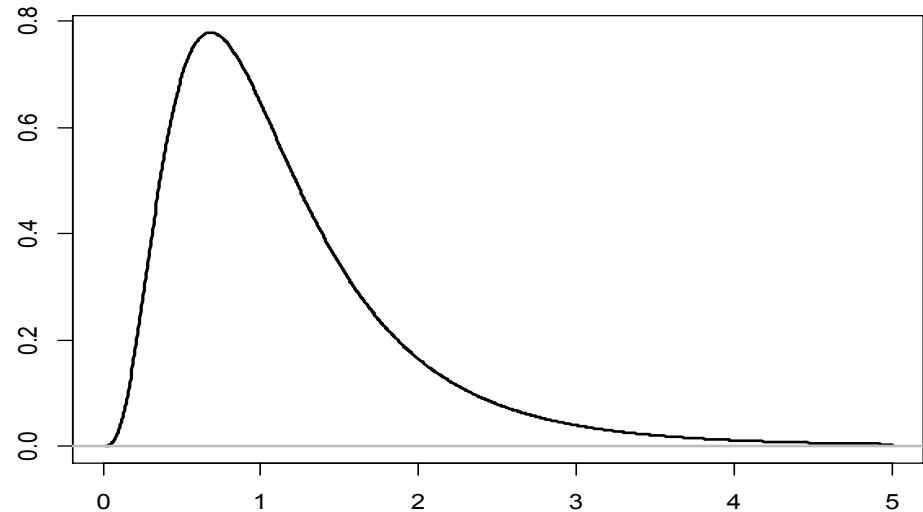
(1) Rejection Region:

Using R: $F_{0.05, df1, df2} = qf(0.95, df1=9, df2=14) = 2.65$

Since $F_0 = 1.57 < 2.65$, we fail to Reject H_0 .

(2) P-value:

Using R: $P\text{-value} = 1 - pf(1.57, df1 = 9, df2 = 14) = 0.22$,
which is too large \rightarrow fail to Reject H_0



2. Confidence Interval for σ_1/σ_2 (We do not need to make these by hand)

Assumptions: Independent, random samples, normally distributed data.

$$\sqrt{\frac{s_1^2}{s_2^2} F_L} < \frac{\sigma_1}{\sigma_2} < \sqrt{\frac{s_1^2}{s_2^2} F_U}$$

In the notation of text

$$F_L = \frac{1}{F_{\alpha/2, dfn, dfd}} \text{ and } F_U = \frac{1}{F_{1-\alpha/2, dfn, dfd}} = F_{\alpha/2, dfd, dfn}$$

CI for 2 Variances in R

```
> test=var.test(y ~ trt, data = RatLead)
> test
  F test to compare two variances
data: y by trt
F = 0.653, num df = 9, denom df = 9, p-value =
0.5356
alternative hypothesis: true ratio of variances
is not equal to 1
95 percent confidence interval:
 0.162208 2.629170
sample estimates:
ratio of variances 0.6530487
```

This is CI for the ratio of variances (σ_1^2/σ_2^2)

The CI for the ratio of standard deviations (σ_1/σ_2) is

```
> sqrt(test$conf.int)
[1] 0.4027505 1.6214716
attr(,"conf.level")
[1] 0.95
```

3. Sensitivity to Assumptions

- F-test for comparing two variances, and the CI based on the F-tests, assume normality of the individual observations.
- Failure of this assumption causes far too many rejections in the tests and far too low coverage in the CIs.
- This is a similar result to the single variance case.

4. Levene's test to compare two variances

- Levene's test is a commonly used robust test for equality of variances. Can think of testing $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_a: \sigma_1^2 \neq \sigma_2^2$
- Can be used to test equality of two (or more) variances **without the assumption of normality**.
- Idea: Compute the absolute values of the deviations from the sample means ($|y_{ij} - \bar{y}_i|$) to construct the test statistic, then do a t-test to compare groups.
- Alternative: When the deviations are calculated using the median (instead of the mean), the test is called the Brown-Forsythe test. This is the default in R.
- In practice, people use Levene's test more often than the F-test to test $H_0: \sigma_1^2 = \sigma_2^2$. However, F-tests that serve a slightly different purpose, which will be turning up a lot when we get to ANOVA and linear regression.

Levene's Test in R

```
> library(car)
> leveneTest(Aphids ~ Trt, data = Aphids)
```

Levene's Test for Homogeneity of Variance
(center = median)

	Df	F value	Pr(>F)
group	1	0.7575	0.3935
	22		

Ch8 Inference about more than two population central values

The Ch 8 notes includes:

- **The analysis of variance (ANOVA) for comparing several mean values (notes08.1)**
- More with one-way ANOVA and Kruskal-Wallis test (notes08.2)

Ch 8.1: The analysis of variance (ANOVA) for comparing several population means

1. ANOVA model
2. The idea of ANOVA
3. ANOVA table and F-test

Rice Example:

Goal is to compare effects of 4 acids on the growth of rice seedlings.

$t = 4$ acids: control, acetic, propionic, butyric (4 trts)

$n_T = 20$ dishes (with shoots) randomly assigned to trts ($n = 5$ dishes/acid).

Let $y_{ij} = \text{dry weight after seven days for the } j^{\text{th}} \text{ dish of the } i^{\text{th}} \text{ treatment}$ (acid), where $i=1, \dots, t$, $j=1, \dots, n_i$. In this example, $t = 4$ and $n_i = n = 5$.

Let $\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4$ be the sample means and $\mu_1, \mu_2, \mu_3, \mu_4$ be the population means (unknown).

We could use two-sample t-tests to compare every 2 groups, but there are some problems with this idea:

1. In order to test all possible pairs, we would need to run 6 tests for this example. Inconvenient!
2. We get many different estimated pooled variances.
3. Multiple testing problem. We will discuss this in Chapter 9.

1. ANOVA Model

Two equivalent model statements for one-way ANOVA

1. **Means Model (No Intercept)** : $y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \dots, t$, and $j = 1, \dots, n_i$,

where

- y_{ij} is the j^{th} observation for the i^{th} treatment
- μ_i is **the mean of the i^{th} treatment**
- ϵ_{ij} is a random error with the assumption that $\epsilon'_{ij} s \stackrel{iid}{\sim} N(0, \sigma^2)$

2. **Effects Model (Default)** : $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, \dots, t$, and $j = 1, \dots, n_i$,

Where

- μ is the overall mean
- α_i is the i^{th} group/treatment effects (NOT Type I error!) with $\sum_{i=1}^t \alpha_i = 0$,
which represents the deviation from μ when treatment i is applied.
- ϵ_{ij} is a random error with the assumption that $\epsilon'_{ij} s \stackrel{iid}{\sim} N(0, \sigma^2)$

Statistical Hypothesis

- Recall that the model for the data can be expressed either in

$$y_{ij} = \mu_i + \epsilon_{ij}, i = 1, \dots, t, \text{ and } j = 1, \dots, n_i,$$

or

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, \dots, t, \text{ and } j = 1, \dots, n_i.$$

- Since we are interested in testing the equality of the t treatment means, the hypotheses are

the means model: $H_0 : \mu_1 = \dots = \mu_t$

$$H_a : \mu_i \neq \mu_j \text{ for at least one pair } (i, j)$$

or, **the effects model:** $H_0 : \alpha_i = 0 \text{ for } i = 1, \dots, t$

$$H_a : \alpha_i \neq 0 \text{ for at least one } i$$

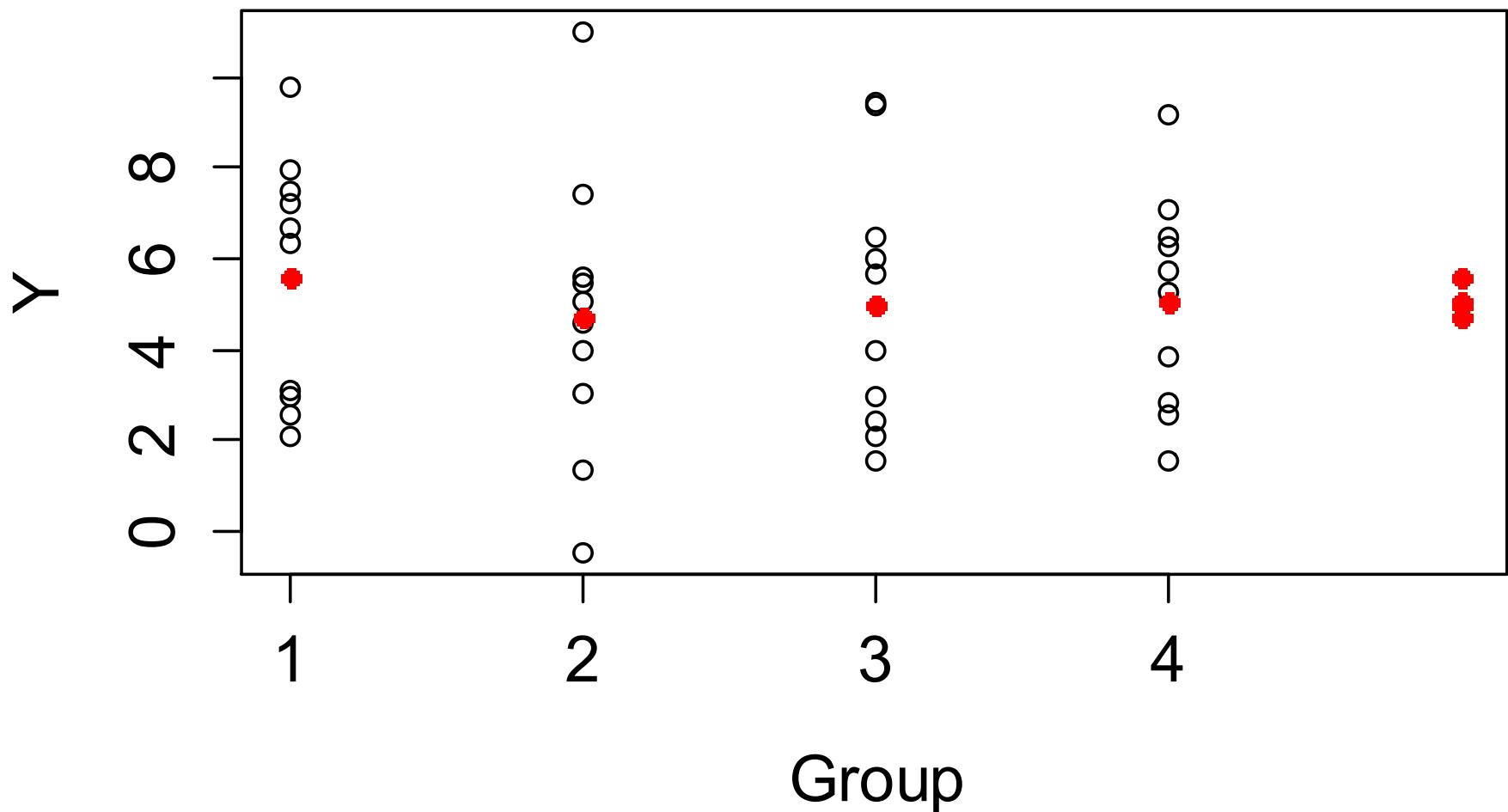
If H_0 is true (using either notation), then the model may be written:

$$y_{ij} = \mu + \epsilon_{ij}, \text{ where } \mu \text{ is the overall mean.}$$

2. The idea of ANOVA

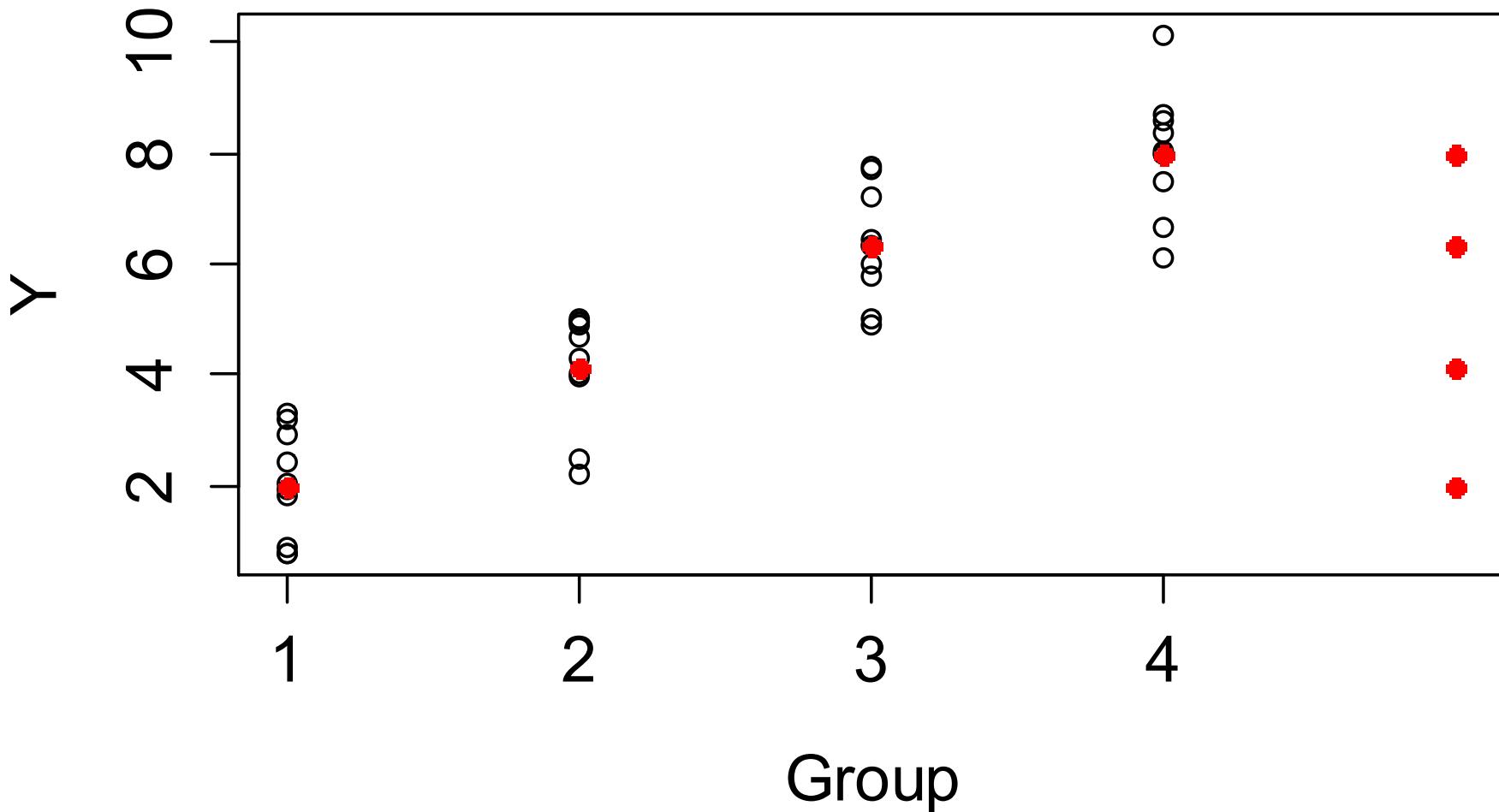
Example of True $H_0 (\mu_1 = \mu_2 = \mu_3 = \mu_4)$

Variation between group means is not large compared to variation within groups.



Example of False H_0

Variation between group means is large compared to variation within groups.



3. The ANOVA table and F-test

- The name “Analysis of Variance” stems from a **partitioning** of the total variability in the response variable into components that are consistent with a model for the experiment.
- Notation :
 - $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$: sum of observations under the i^{th} treatment
 - $\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$: average of the observations under the i^{th} treatment
 - $y_{\cdot\cdot} = \sum_{i=1}^t \sum_{j=1}^{n_i} y_{ij}$: grand sum for all $n_T = n_1 + \dots + n_t$ observations
 - $\bar{y}_{\cdot\cdot} = \frac{1}{n_T} \sum_{i=1}^t \sum_{j=1}^{n_i} y_{ij}$: grand average for all $n_T = n_1 + \dots + n_t$ observations
- Variability will be measured by the “sum of squares”.

Sum of Squares

- Total corrected sum of squares : deviations from the **grand average**

$$SSTotal = \sum_{i=1}^t \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

- Between-treatment sum of squares: how much the **treatment averages** vary about the **grand average**.

$$SSTrt = \sum_{i=1}^t \sum_{j=1}^{n_i} (\bar{y}_{i\bullet} - \bar{y}_{..})^2 = \sum_{i=1}^t n_i (\bar{y}_{i\bullet} - \bar{y}_{..})^2$$

- Within-treatment (residual) sum of squares : deviations from **treatment averages**

$$SSResid = \sum_{i=1}^t \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2$$

- Equation:

$$SSTotal = SSTrt + SSResid$$

- A pitfall in the SS : The more terms we add, the greater the variability becomes.
 - If we add treatments, we add SST_{Trt} .
 - If we add subjects, we add SS_{Resid} .
- We need to consider the number of treatments or subjects to compare these two variabilities. Thus, we take the average of SS by dividing SS by the '**degrees of freedom**'.

- **Degree of freedom**
 - The number of “free” components, or the number of independent values that a statistical analysis can estimate.
 - Typically, it is equal to (# of observations - # of parameters estimated)
- In the data structure,
 - There are n_T total observations; thus, $SSTotal$ has $n_T - 1$ degrees of freedom.
 - There are t treatment means, so $SSTrt$ has $t-1$ degrees of freedom.
 - There are n_i replicates within the i-th treatment providing $n_i - 1$ degrees of freedom with which to estimate the experimental error.
Because there are t treatments, we have $(n_1 - 1) + \dots + (n_t - 1) = n_T - t$ degrees of freedom for $SSResid$.
- **Mean squares**
 - $MSTotal = SSTotal/(n_T - 1)$
 - $MStrt = SSTrt/(t - 1)$
 - $MSResid = SSResid/(n_T - t)$

We test H_0 by forming an F-ratio based on two estimates of variance σ^2 :

$$F = \frac{s_B^2}{s_W^2} = \frac{MSTrt}{MSResid}, \quad \text{where}$$

s_W^2 is an estimate of σ^2 formed by pooling sample variances.

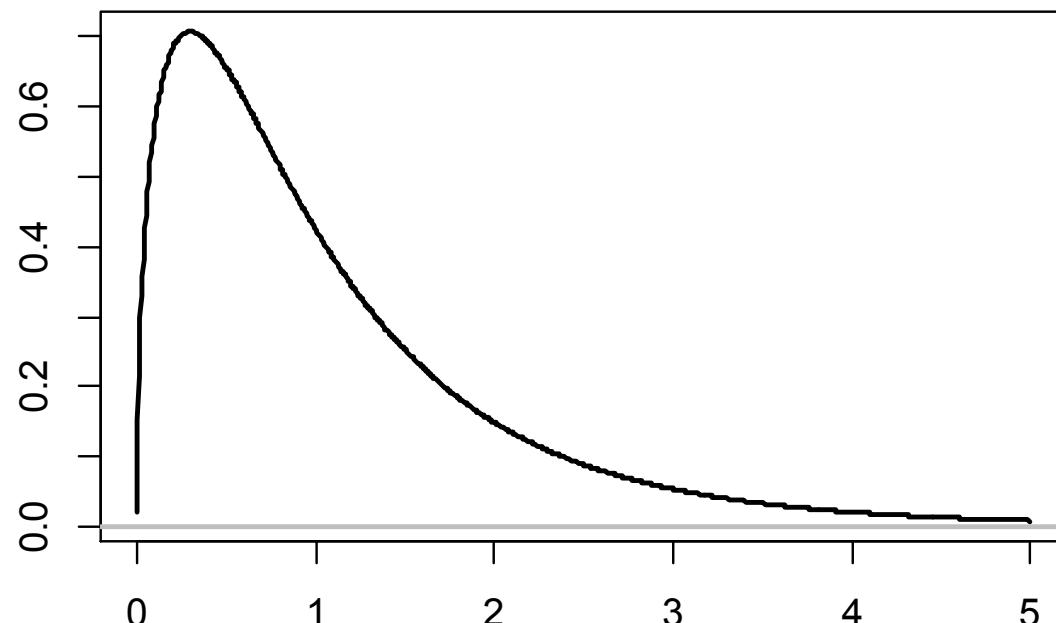
s_B^2 is an estimate of σ^2 formed using only sample means.

$$s_W^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_t - 1)s_t^2}{n_T - t} = \frac{SSResid}{n_T - t} = MSResid$$

$$s_B^2 = \frac{\sum_{i=1}^t n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2}{t - 1} = \frac{SSTrt}{t - 1} = MSTrt$$

s_1^2, \dots, s_t^2 are sample variances of treatment groups $1, \dots, t$, respectively.

- s_W^2 is a pooled estimate of the variance σ^2 that is valid whether H_0 is true or not, because it is based on differences within treatment groups.
- s_B^2 is an estimate of $\sigma^2 + [\sum n_i(\mu_i - \mu)^2]/(t-1)$ where $\mu = (\sum n_i \mu_i)/n_T$. Hence, when H_0 is true, s_B^2 is a valid estimate of σ^2 . If H_0 is false, then s_B^2 will tend to be too big (because it is “contaminated” by mean differences).
- If H_0 is true, the F-statistic will be distributed with the F-distribution with $df_1=t-1$ and $df_2=n_T-t$, where t is # of trts, n_T is the total sample size.



ANOVA Table

The test procedure is summarized in table below. This is called an analysis of variance (ANOVA) table.

Source	SS	df	MS = SS/df	F-test
Trt	$SSTrt$	$t-1$	$s_B^2 = SSTrt/(t-1)$	$F=MStrt/MSResid$
Resid	$SSResid$	n_T-t	$s_W^2 = SSResid/(n_T-t)$	
Total	$SSTotal$	n_T-1		

ANOVA F-test for the Equality of Means

Assumptions: Random sample, independent observations, normally distributed residuals, equality of variances.

$H_0: \mu_1 = \mu_2 = \dots = \mu_t$ vs. $H_a: \mu_i \neq \mu_j$ for at least one pair (i, j)

Or $H_0: \alpha_i = 0$ for $i = 1, \dots, t$ vs. $H_a: \alpha_i \neq 0$ for at least one i

Test Statistic: $F_0 = \frac{MSTrt}{MSResid}$

P-value: $P(F \geq F_0)$

R code: `1-pf(F0, df1, df2)`

Notes:

- This test has a NON-directional alternative.
- $df1 = ndf = t-1 = dfTrt$ and $df2 = ddf = n_T - t = df Resid$
- $MSResid = MSError = MSWithin = s_W^2 = \hat{\sigma}^2$
- $MSTrt = MSBetween$

Rice Example: One Way ANOVA

- In R, use lm() or aov().
- **NOTE:** Be sure “trt” variable is a factor! Check using str().

```
> OneWayFit <- lm(weight ~ trt, data = rice)
> anova(OneWayFit)
          Df  Sum Sq Mean Sq F value    Pr(>F)
trt          3 1.2199  0.4066   103.5 1.08e-10 ***
Residuals   16 0.0628  0.0039
```

Reject H_0 because p-value < 0.0001

Based on the P-value for the F-test we conclude that there is extremely strong evidence indicating differences among the treatment means.

But which means are different?

— We will look at this in Chapter 9.

Comments on lm() Function:

The linear model **lm()** function can be used to fit a broad class of linear models. This includes one-way ANOVA (this chapter) and simple linear regression (CH11) and many others. Again, be sure `trt` is defined as a factor! This can be checked using `str()`.

The analysis of variance **aov()** function is a “wrapper” for the `lm()` function.

From `?aov`:

- The main difference between `lm()` and `aov()` is the way `print`, `summary`, etc handle the fit. `aov()` is expressed in the traditional language of ANOVA rather than that of linear models.
- `aov()` is designed for balanced designs, and the results can be hard to interpret without balance. Beware that missing values in the response will likely lose the balance.

Note: In STAT512, we will use `lm()` and `Anova()` function from the `car` package!

A note about ANOVA vs t-test when there are t=2 groups:

An ANOVA with t=2 groups is equivalent to running a two-sample t-test assuming equal variances.

The p-values testing $H_0: \mu_1 = \mu_2$ will be identical.

The test statistics are related as follows $F = t^2$.

Note that F-test is non-directional whereas the t-test can easily accommodate a one-sided alternative.

Ch8 Inference about more than two population central values

The Ch 8 notes includes:

- The analysis of variance (ANOVA) for comparing several mean values (notes08.1)
- **More with one-way ANOVA and Kruskal-Wallis test (notes08.2)**

Ch 8.2: More with one-way ANOVA and Kruskal-Wallis test

1. Sample size and power for the ANOVA F-test
2. Checking ANOVA Assumptions
3. Remedies of the failure of ANOVA Assumptions
 - A. Transformations
 - B. Kruskal-Wallis Test

1. Sample size and power for the ANOVA F-test

As with all of the other sample size calculations, we need:

- 1) A conjecture about the within-group standard deviation σ .
- 2) Identification of the true alternative that we want to detect:
conjectures for $\mu_1, \mu_2, \dots, \mu_t$.

Given the alternative, power is a function of the “noncentrality” parameter for the F-distribution:

$$\lambda = \frac{n \sum_{i=1}^t (\mu_i - \bar{\mu})^2}{\sigma^2}$$

$\bar{\mu}$ is the average of $\mu_1, \mu_2, \dots, \mu_t$.

Power increases as sample size and the differences among the true means increase, and decreases as the error standard deviation increases.

Power can be computed in R using the **pf** function:

```
fcrit=qf(0.95, dfn, dfd)  
power = 1 - pf(fcrit, dfn, dfd, lambda)
```

Power can also be computed using `power.anova.test()` (See CH8p2_R.pdf)

Or using Lenth to get the power: <http://homepage.stat.uiowa.edu/~rlenth/Power/>

2. Checking the ANOVA assumptions

1. Random sample, independent observations
2. Residuals are normally distributed: $\varepsilon_{ij} \sim N(0, \sigma^2)$

QQ plot of residuals
(Test of normality for residuals)
(eg. Shapiro-Wilks test)

If the observations come from a normal distribution, we would expect to get approximately a straight line.
3. Equality (Homogeneity) of variances: $\text{Var}(\varepsilon_{ij}) = \sigma^2$

Plot of residuals vs fitted (predicted) values
(Levene's test)

If the residuals roughly form a "horizontal band" around the 0 line, it suggests that the variances of the error terms are equal.

We will be using the **residuals** to check assumptions 2 and 3.

Fitted/predicted values $\hat{y}_{ij} = \bar{y}_{i\bullet}$ are the sample means.

Residuals are calculated as $e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\bullet}$

Note: Both diagnostic plots (and more) can be generated in R by applying the `plot()` function to a `lm` or `aov` object.

Checking for Normality of Residuals:

1. A Q-Q plot of the residuals is a useful graphical tool for checking normality. The Q-Q plot easily constructed in R. A histogram of the residuals would also work for this purpose.
2. Tests of normality can be used (eg: Shapiro-Wilks test). Plots are usually more helpful than the formal tests.

Important Note: When you check the data for normality, check the residuals, not the observations themselves. The errors (i.e., deviations from group means) are assumed to follow a common normal distribution but the observations themselves may come from *different* normal distributions. So, when the means are very different, the combined data will often look non-normal, even when the residuals are close to normal.

Checking for Equality of Variance (Homogeneity of Variance):

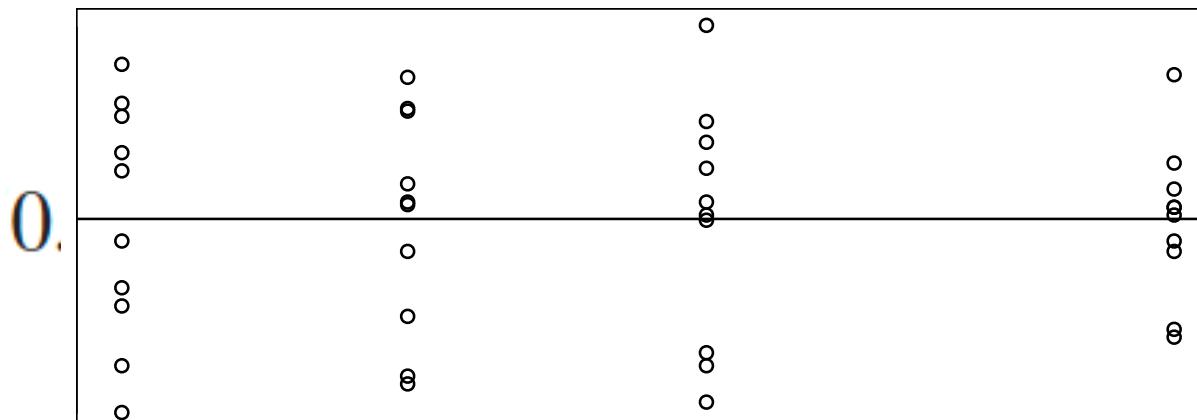
Check plot of residuals versus predicted values (or fitted values). Plots are usually more helpful than the formal tests.

Fitted/predicted values $\hat{y}_{ij} = \bar{y}_{i\bullet}$ are the sample means.

Residuals are calculated as $e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\bullet}$.

Primarily interested in checking for equal “scatter” around the horizontal 0 line (representative of equal variance), but sometimes we can also detect “skew” and/or outliers in the residual diagnostic plot.

Megaphone shape is common when assumption of equal variances is NOT met.



3. Remedies of the failure of ANOVA Assumptions

A. Transformations

- Common transformations
- Box-Cox transformation
- Interpretation

B. Kruskal-Wallis Test

3A. Transformations

If ANOVA assumptions are not satisfied on the original scale (Y), then you can consider using a transformed response variable.

Transforming the response makes the model harder to interpret, so we don't want to do it unless it's really necessary.

The literature may have examples of transformations that are common in a particular field. **But the way to decide what transformation is appropriate is by fitting the model with the transformed response and checking the diagnostic plots.**

Common transformations:

- **Square root:** $YT = \text{sqrt}(Y)$ or $YT = (Y)^{0.5}$ (in R)

Example: Count data (calls to switchboard, insects in trap, etc)

- **Power:** $YT = 1/Y$ or $YT = Y^2$ (in R)

- **Log:** $YT = \log(Y)$ or $YT = \log_{10}(Y)$ (in R)

Examples: Chemical concentrations or hormone levels

Important Note: Watch out for $y=0$ values which will be undefined after log transformation! If this is not properly accounted for, then these values will be treated as “missing”. A simple, common solution is to add a small positive constant before log transformation. You might use $y_T = \log(y+1)$ when the y 's are 0 to 20, and use $y_T = \log(y+0.01)$ when the y 's are 0 to 0.25. These are just suggestions!

- The use of transformations like $YT = Y^{(0.75)}$ is rare. Transformations like $YT = Y^{(0.63)}$ are seldom used.

Box-Cox transformation:

A systematic method for choosing a transformation is the Box-Cox transformation. This approach should only be used for $y > 0$!

The general form of the Box-Cox transformation is:

$$g(y_i) = (y_i^\lambda - 1)/\lambda$$

where λ is a constant to be determined from the data.

If $\lambda = 1$, then no transformation is needed.

If $\lambda = 2$, then model Y^2 (instead of Y).

If $\lambda = -1$, then model $Y^{-1} = 1/Y$ (instead of Y).

If $\lambda = 0.5$, then model $Y^{0.5} = \sqrt{Y}$ (instead of Y).

If $\lambda = 0$, then model $\log(Y)$ (instead of Y)

We will use the `boxcox()` function from the MASS package to create a Box-Cox plot to choose λ .

Interpretation after transformation

The means of the transformed variables are not the same as the means of the original variables. However, if the means are significantly different based on the analysis in the transformed scale, it is reasonable to conclude that the means in the original scale are also significantly different.

What should a researcher report when it was necessary to do the ANOVA using a transformed scale?

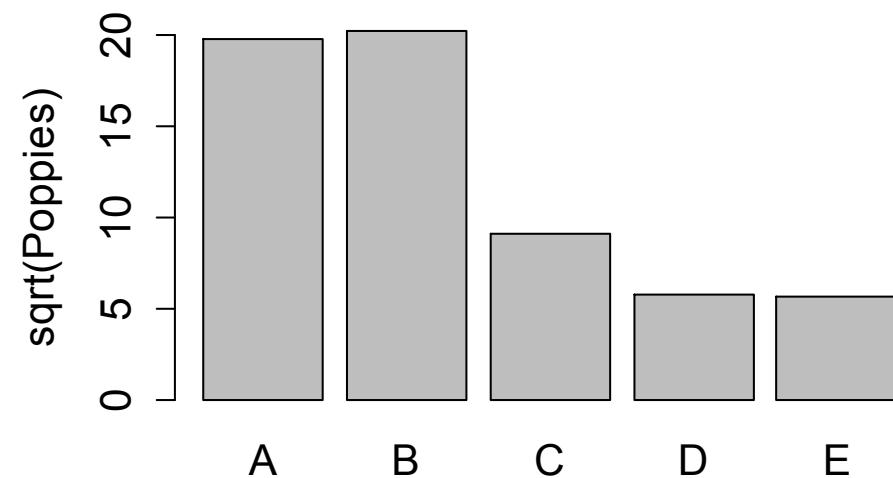
Three options:

1. Present the means on the transformed scale.
2. Present the means on the original scale, but note that analysis was done on transformed data.
3. Back transform to the original scale.

Option 1: Present the means (or graphs) and the comparison of means in the transformed scale. This is a reasonable option when the scale is common in that field of study (e.g. log in chemistry, sqrt in radiological sciences.)

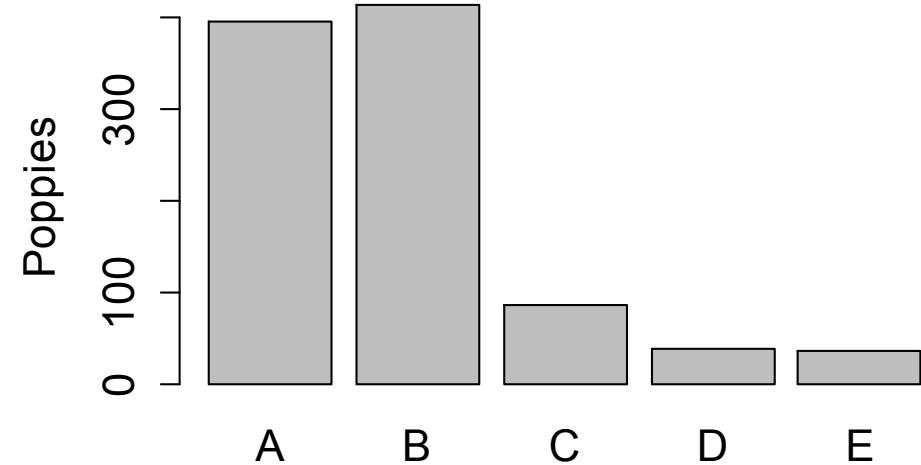
Example: Poppy plants in oats (see R example).

Trt	Orig Scale	Sqrt Scale	Back Transformed
A	394.75	19.83	393.09
B	413.00	20.23	409.07
C	86.75	9.08	82.47
D	37.75	5.75	33.11
E	35.25	5.64	31.89



Option 2: Present the means (or graphs) in the original scale, but with the comparison of means based on the transformed scale. Describe what you have done in your methods section and in a footnote to the table or graph of means.

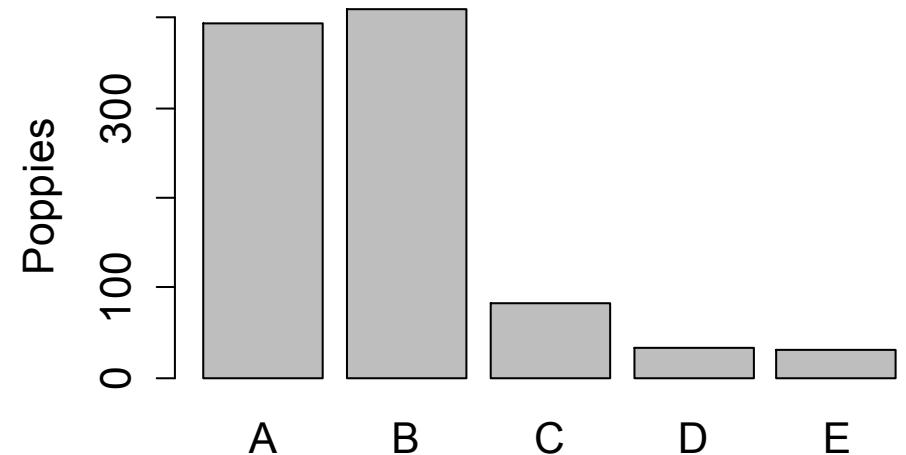
Trt	Orig Scale	Sqrt Scale	Back Transformed
A	394.75	19.83	393.09
B	413.00	20.23	409.07
C	86.75	9.08	82.47
D	37.75	5.75	33.11
E	35.25	5.64	31.89



Option 3: Compute the means using the transformed variable, but “back-transform” the means before presentation. To back-transform a square root transformed variable, square the mean.

Backtransformed means can be substantially lower than the means in the original scale, particularly when the transformation is log.

Trt	Orig Scale	Sqrt Scale	Back Transformed
A	394.75	19.83	393.09
B	413.00	20.23	409.07
C	86.75	9.08	82.47
D	37.75	5.75	33.11
E	35.25	5.64	31.89



Comments about interpretation after **log** transformation:

There can be interpretational advantages to do an analysis in the **log** scale. Because $\log(x/y) = \log(x) - \log(y)$, differences between means in the log analysis can be interpreted as ratios in the original scale.

In bioinformatics, it is common to use a log₂ transformation to satisfy ANOVA assumptions, but also because the differences can be interpreted as log₂ fold change (FC) values.

$$\log_2^{(2/1)} = +1, \log_2^{(1/2)} = -1$$

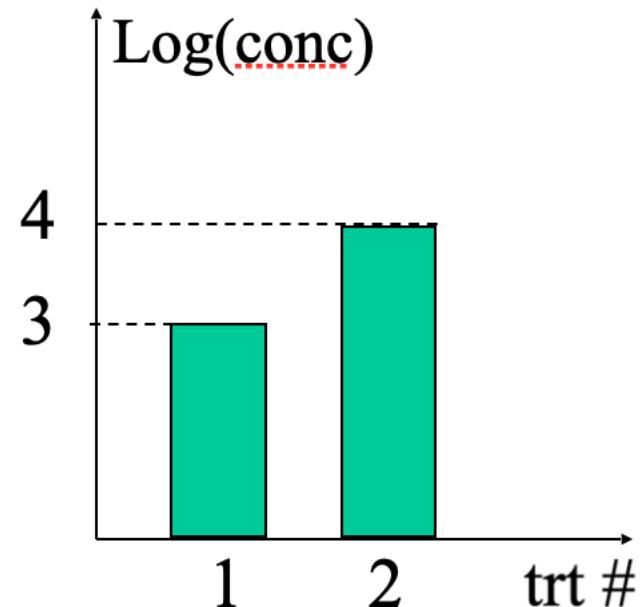
When using a log transformation, it is common to see people back transform the difference to a “ratio” scale after analysis.

A Lognormal Example:

Let \bar{x}_1 and \bar{x}_2 be the sample means of log-scaled observations for trt1 and trt2, respectively.

Let \bar{x}_{1O} and \bar{x}_{2O} be the sample means of original-scaled observations for trt1 and trt2, respectively.

$$\rightarrow \bar{x}_{1O} = e^{\bar{x}_1} \text{ and } \bar{x}_{2O} = e^{\bar{x}_2}$$



$$\text{Est. trt diff (log scale)} = \bar{x}_1 - \bar{x}_2 = 3 - 4 = -1$$

$$\text{Est. ratio of trt means (orig. scale)} = \bar{x}_{1O}/\bar{x}_{2O} = e^{\bar{x}_1}/e^{\bar{x}_2} = e^{\bar{x}_1 - \bar{x}_2}$$

$$e^{\bar{x}_1 - \bar{x}_2} = e^{3-4} = e^{-1} = 1/2.718 = 0.368 \approx 37\%$$

Trt1 concentration is 37% Trt2 concentration!

A C.I. for this % is given by: (e^{LCL}, e^{UCL})

(where (LCL,UCL) is a C.I. for the difference in the means of the log data)

3B. The Kruskal-Wallis Test

- The Kruskal-Wallis test is a non-parametric alternative to the one-way ANOVA F-test. This test does NOT require the assumption of normality.
- The Kruskal-Wallis test is an extension of the Wilcoxon Rank Sum test for more than two groups. Both are rank-based methods and hence robust to outliers.
- This test is typically thought of as a test of **medians**. But it is more correct to think of it as a test of shift in distribution.
- In R, use `kruskal.test()`.

Ch8 Inference about more than two population central values

The Ch 8 notes include:

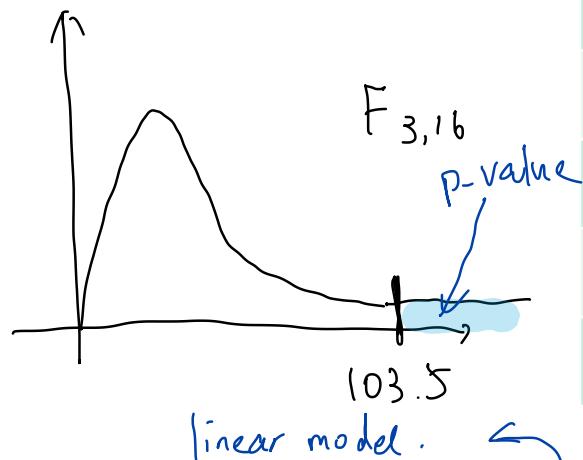
- The analysis of variance (ANOVA) for comparing several mean values
- **More on one-way ANOVA**

Ch 8.2: More on one-way ANOVA

1. ANOVA “by hand”
2. Checking ANOVA assumptions
3. Sample size and power for the ANOVA F-test
4. Remedies of the failure of ANOVA assumptions (reading only)
 - A. Transformations
 - B. Kruskal-Wallis Test

1. ANOVA “by hand”

Rice data: Do we have evidence of any differences in the population mean weight of the rice seedlings across the treatment groups?



	control	acetic	propion	butyric
	4.23	3.75	3.75	3.68
	4.38	3.68	3.65	3.69
	4.25	3.81	3.82	3.64
	4.3	3.84	3.69	3.56
	4.25	3.76	3.73	3.73

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

vs.

$$H_a: \text{not all } \mu_i \text{'s are the same}$$

```
> OneWayFit <- lm(weight ~ trt, data = rice)
> anova(OneWayFit)
```

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
trt	4-1	3 1.2199	0.4066	103.5	1.08e-10 ***
Residuals	16	0.0628	0.0039		

20-4

Within group.

$$\frac{MS_B}{MS_W}$$

↓ indicating large Between-group variability

$$\bar{y}_{1\cdot} = 4.282$$

$$\bar{y}_{\cdot\cdot} = 3.8595$$

$$\bar{y}_{2\cdot} = 3.768$$

$$\bar{y}_{3\cdot} = 3.728$$

$$\bar{y}_{4\cdot} = 3.660$$

control	acetic	propion	butyric
4.23	3.75	3.75	3.68
4.38	3.68	3.65	3.69
4.25	3.81	3.82	3.64
4.3	3.84	3.69	3.56
4.25	3.76	3.73	3.73

$$BSS = n_1 (\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})^2 + n_2 (\bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot})^2 + n_3 (\bar{y}_{3\cdot} - \bar{y}_{\cdot\cdot})^2 + n_4 (\bar{y}_{4\cdot} - \bar{y}_{\cdot\cdot})^2$$

$\uparrow \quad \uparrow \quad \uparrow$
5 4.282 3.8595

$$= 1.219855$$

$$WSS = (y_{11} - \bar{y}_{1\cdot})^2 + \dots + (y_{15} - \bar{y}_{1\cdot})^2 + \dots + (y_{41} - \bar{y}_{4\cdot})^2 + \dots + (y_{45} - \bar{y}_{4\cdot})^2$$

$\uparrow \quad \uparrow \quad \dots \quad \uparrow \quad \uparrow$
4.23 4.282 . . . 4.25 4.282 . . . 3.68 3.660 . . . 3.73 3.660

$$= 0.06284$$

$$4 \quad MS_B = \frac{BSS}{g-1} = \frac{1.219855}{4-1} = 0.41$$

$$MS_w = \frac{WSS}{n_T - g} = \frac{0.06284}{20 - 4} = 0.0039$$

$$F = \frac{MS_B}{MS_w} = 103.531$$

① F critical

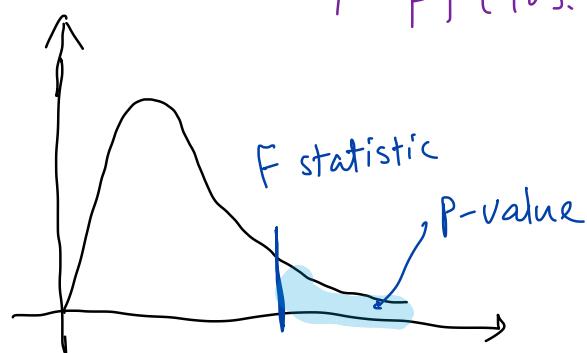
$$f_{\text{critical}} = q_f(1 - \alpha, df_1 = g - 1, df_2 = n_T - g)$$

$$= q_f(0.95, 3, 16) = 3.239$$

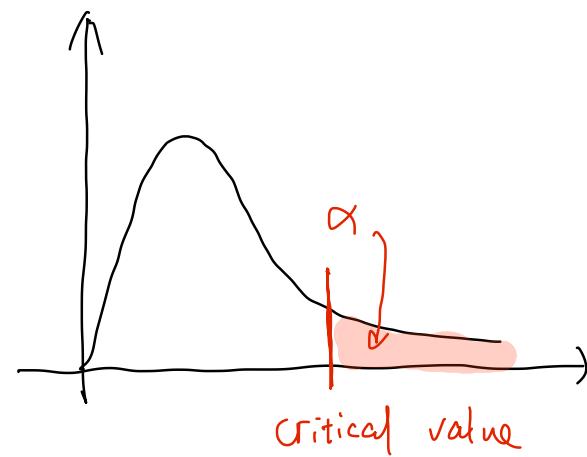
② p-value = 1 - pf(F, df1, df2)

$$= 1 - pf(103.5311, 3, 16) =$$

$$1.08 \times 10^{-10}$$



control	acetic	propion	butyric
4.23	3.75	3.75	3.68
4.38	3.68	3.65	3.69
4.25	3.81	3.82	3.64
4.3	3.84	3.69	3.56
4.25	3.76	3.73	3.73



pf(..., lower.tail = FALSE)

EXAMPLE 8.2

A clinical psychologist wished to compare three methods for reducing hostility levels in university students and used a certain test (HLT) to measure the degree of hostility. A high score on the test indicated great hostility. The psychologist used 24 students who obtained high and nearly equal scores in the experiment. Eight were selected at random from among the 24 problem cases and were treated with method 1. Seven of the remaining 16 students were selected at random and treated with method 2. The remaining nine students were treated with method 3. All treatments were continued for a one-semester period. Each student was given the HLT test at the end of the semester, with the results shown in Table 8.9. Use these data to perform an analysis of variance to determine whether there are differences among mean scores for the three methods. Use $\alpha = .05$.

of observations doesn't need to be the same across groups.

Method	Test Scores									Mean	Standard Deviation	Sample Size
	1	96	79	91	85	83	91	82	87			
2	77	76	74	73	78	71	80			75.571	3.101	7
3	66	73	69	66	77	73	71	70	74	71.000	3.674	9

Solution The null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_a : At least one of the population means differs from the rest.

For $n_1 = 8$, $n_2 = 7$, and $n_3 = 9$, we have a total sample size of $n_T = 24$. Using the sample means given in the table, we compute the overall mean of the 24 data values:

$$\bar{y}_{..} = \frac{\sum_{i=1}^3 n_i \bar{y}_i / n_T}{n_T} = \frac{(8(86.750) + 7(75.571) + 9(71.000))}{24} = 1,861.997/24$$

weighted average of group-specific means

$$= 77.5832$$

Using this value along with the means and standard deviations in Table 8.9, we can compute the three sums of squares as follows:

$$\begin{aligned} SSB &= \sum_{i=1}^3 n_i (\bar{y}_i - \bar{y}_{..})^2 = 8(86.750 - 77.5832)^2 + 7(75.571 - 77.5832)^2 \\ &\quad + 9(71 - 77.5832)^2 = 1,090.6311 \end{aligned}$$

and

$$\begin{aligned} SSW &= \sum_{i=1}^3 (n_i - 1)s_i^2 = (8 - 1)(5.625)^2 + (7 - 1)(3.101)^2 + (9 - 1)(3.674)^2 \\ &= 387.1678 \end{aligned}$$

Finally, $TSS = SSB + SSW = 1,090.6311 + 387.1678 = 1,477.7989$. The AOV table for these data is given in Table 8.10.

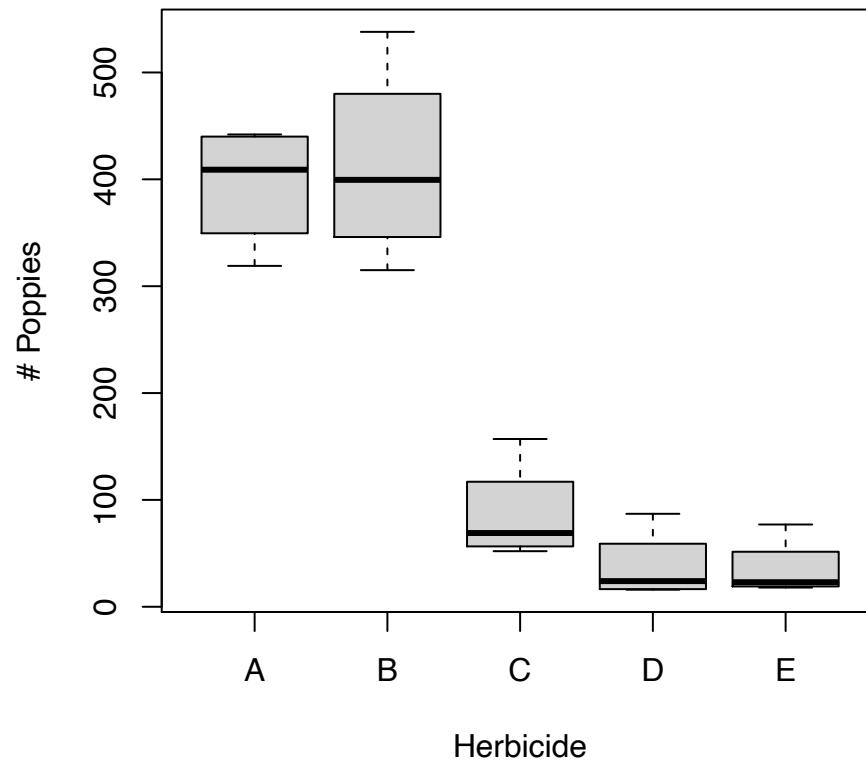
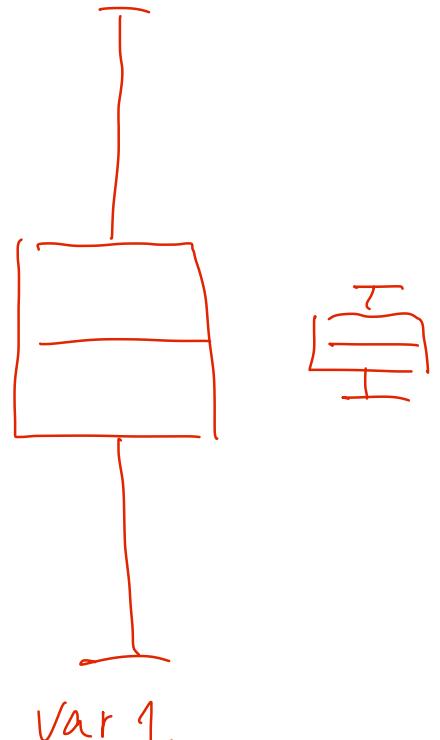
Source	SS	df	MS	F	p-value
Between samples	1,090.6311	2	545.316	545.316/18.4366 = 29.58	<.001
Within samples	387.1678	21	18.4366		
Total	1,477.7989	23			

2. Checking ANOVA assumptions

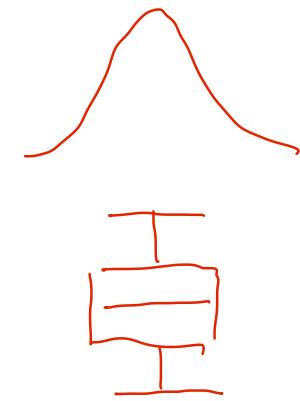
1. Random sample, independent observations
(Guarantees F has an F distribution)
2. Data are **normally distributed** (within each group)
QQ plot, or formal test of normality (e.g., Shapiro-Wilks test)
3. **Equality** (homogeneity) **of variances** (across groups)
Plot of residuals vs fitted (predicted) values, or formal test (e.g., Levene's test)

We will be using the **residuals** to check assumptions 2 and 3. Recall the ANOVA model $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, \dots, g$, and $j = 1, \dots, n_i$.

- Assumption 2 is $\epsilon_{ij} \sim \text{Normal}$
- Assumption 3 is $\text{Var}(\epsilon_{ij}) \equiv \sigma^2$
- In summary, $\epsilon_{ij} \sim N(\text{mean} = 0, \text{sd} = \sigma)$
- Fitted/predicted values $\hat{y}_{ij} = \bar{y}_i$. are the sample means by group
- Residuals are calculated as $e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i$.



Normality
&
equal variance



```
Fit1 <- lm(Plants ~ Trt, data = poppies)
```

R example using the poppies dataset

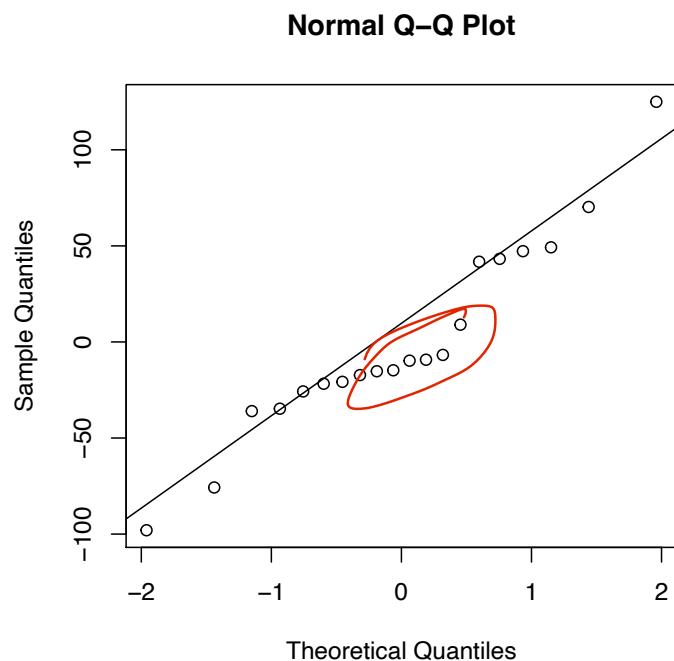
- residuals(Fit1) returns the residuals
- fitted(Fit1) returns the fitted values



Checking for Normality of Residuals:

1. A **Q-Q plot of the residuals** is a useful graphical tool for checking normality. The Q-Q plot easily constructed in R. A histogram of the residuals would also work for this purpose.
2. Tests of normality can be used (eg: Shapiro-Wilks test). Plots are usually more helpful than the formal tests.

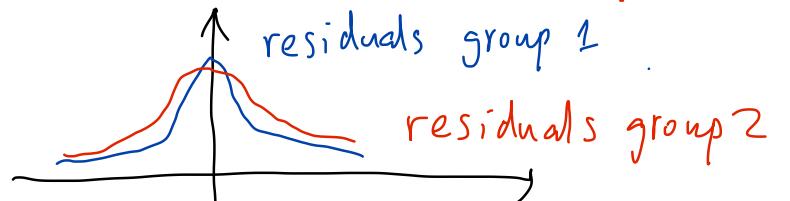
```
qqnorm(residuals(Fit1))  
qqline(residuals(Fit1))  
shapiro.test(residuals(Fit1))
```



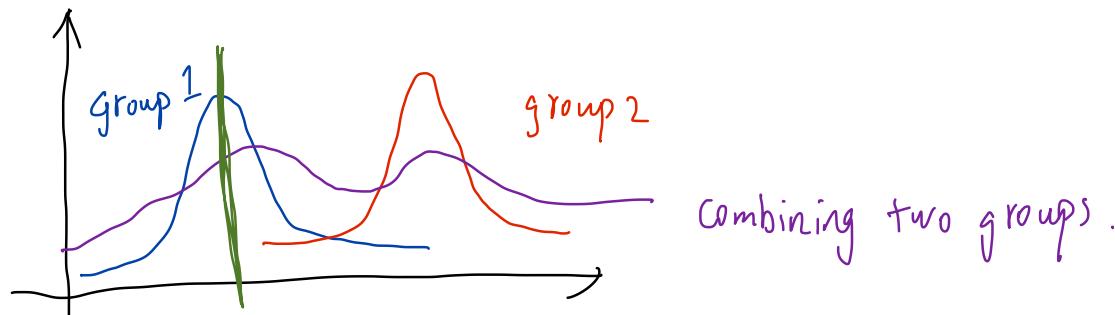
If the residuals come from a normal distribution, we would expect to get approximately a straight line.

1. convenience
2. combines information from different groups .

Note: Can also do the normality check for raw data separately for each group. However, no need to check normality for combined data across groups!



- The observations within each group are assumed to follow a normal distribution with group-specific mean and a common variance.
- The errors (i.e., deviations from group-specific means) follow a common normal distribution with zero mean.
- The combined observations across groups come from a mixture of *different* normal distributions (different means). Therefore, the combined observations do not necessarily follow a normal!



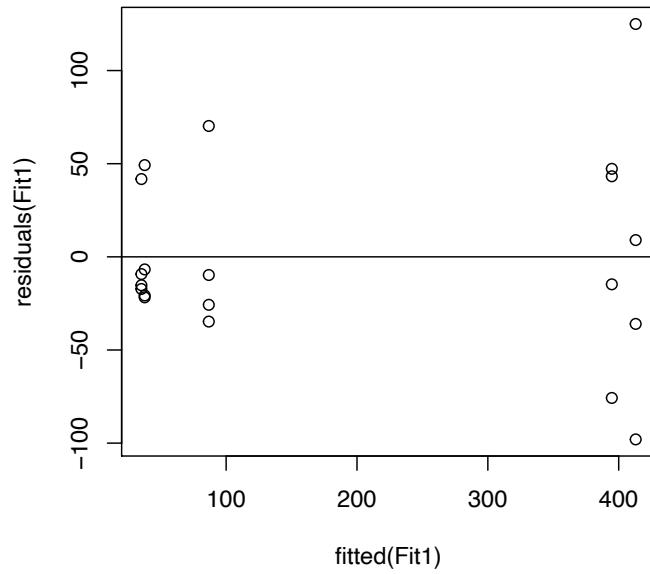
Checking for Equality of Variance:

1. A plot of residuals versus predicted values (or fitted values) is a useful graphical tool for checking equal variance. Primarily interested in checking for equal “scatter” around the horizontal 0 line (representative of equal variance), but sometimes we can also detect “skew” and/or outliers in the residual diagnostic plot. Funnel shape is common when assumption of equal variances is NOT met.
2. Tests of equal variance can be used (eg: Levene’s test). Plots are usually more helpful than the formal tests.

```

plot(fitted(Fit1), residuals(Fit1))
abline(h = 0)
leveneTest(Plants ~ factor(Trt), data = poppies)

```



If the residuals roughly form a "horizontal band" around the 0 line, it suggests that the variances of the error terms are equal.

Some pattern / funnel shape in residual vs fitted plot, but may not be strong enough to refute the equal variance assumption.

Note: Can also draw a boxplot of residuals by group, which provides a straightforward visualization of variability within each group.

Note: Both diagnostic plots (and more) can be generated in R by applying the plot() function to a lm or aov object.

```
plot(Fit1)
```

↓
generates 4 plots. first 2 are qqnorm of residuals
and residuals vs fitted.

3. Sample size and power for the ANOVA F-test

As with all of the other sample size calculations, we need:

- 1) A conjecture about the within-group standard deviation σ .
- 2) Identification of the true alternative that we want to detect:

conjectures for $\mu_1, \mu_2, \dots, \mu_g$.

Conjectured population mean for group i

Typically, equal sample size n is assumed for all groups.

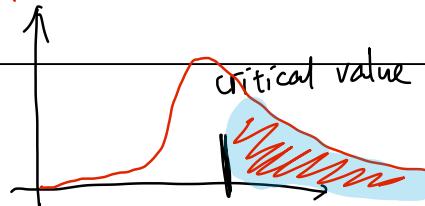
$$\bar{\mu} = \frac{\mu_1 + \dots + \mu_g}{g}$$

Given the alternative, power is a function of the “noncentrality” parameter for the F-distribution:

$$\lambda = \frac{n \sum_{i=1}^g (\mu_i - \bar{\mu})^2}{\sigma^2}$$

where $\bar{\mu}$ is the average of $\mu_1, \mu_2, \dots, \mu_g$. Conjectured within-group sd.

```
f_critical = qf(1-alpha, df1=g-1, df2=n*g-g)
power = 1-pf(f_critical, df1=g-1, df2=n*g-g,
               ncp=lambda)
```



Power can also be computed using `power.anova.test()`

```
power.anova.test(groups = g, n = n,  
                  between.var = var(mu_all),  
                  within.var = sigmasq,  
                  sig.level = alpha)
```

- `between.var` is $\frac{1}{g-1} \sum_{i=1}^g (\mu_i - \bar{\mu})^2$.
- `within.var` is σ^2 .

To get sample size for a specific power, remove `n =` and add `power = .`

4. Remedies of the failure of ANOVA assumptions (reading only)

A. Transformations

- Common transformations
- Box-Cox transformation
- Interpretation

B. Kruskal-Wallis Test

3A. Transformations

If ANOVA assumptions are not satisfied on the original scale (Y), then you can consider using a transformed response variable.

Transforming the response makes the model harder to interpret, so we don't want to do it unless it's really necessary.

The literature may have examples of transformations that are common in a particular field. **But the way to decide what transformation is appropriate is by fitting the model with the transformed response and checking the diagnostic plots.**

Common transformations:

- **Square root:** $YT = \sqrt{Y}$ or $YT = (Y)^{0.5}$ (in R)

Example: Count data (calls to switchboard, insects in trap, etc)

- **Power:** $YT = 1/Y$ or $YT = Y^2$ (in R)

- **Log:** $YT = \log(Y)$ or $YT = \log_{10}(Y)$ (in R)

Examples: Chemical concentrations or hormone levels

Important Note: Watch out for $y=0$ values which will be undefined after log transformation! If this is not properly accounted for, then these values will be treated as “missing”. A simple, common solution is to add a small positive constant before log transformation. You might use $y_T = \log(y+1)$ when the y ’s are 0 to 20, and use $y_T = \log(y+0.01)$ when the y ’s are 0 to 0.25. These are just suggestions!

- The use of transformations like $YT = Y^{0.75}$ is rare. Transformations like $YT = Y^{0.63}$ are seldom used.

Box-Cox transformation:

A systematic method for choosing a transformation is the Box-Cox transformation. This approach should only be used for $y > 0$!

The general form of the Box-Cox transformation is:

$$g(y_i) = (y_i^\lambda - 1)/\lambda$$

where λ is a constant to be determined from the data.

If $\lambda = 1$, then no transformation is needed.

If $\lambda = 2$, then model Y^2 (instead of Y).

If $\lambda = -1$, then model $Y^{-1} = 1/Y$ (instead of Y).

If $\lambda = 0.5$, then model $Y^{0.5} = \sqrt{Y}$ (instead of Y).

If $\lambda = 0$, then model $\log(Y)$ (instead of Y)

We will use the `boxcox()` function from the MASS package to create a Box-Cox plot to choose λ .

Interpretation after transformation

The means of the transformed variables are not the same as the means of the original variables. However, if the means are significantly different based on the analysis in the transformed scale, it is reasonable to conclude that the means in the original scale are also significantly different.

What should a researcher report when it was necessary to do the ANOVA using a transformed scale?

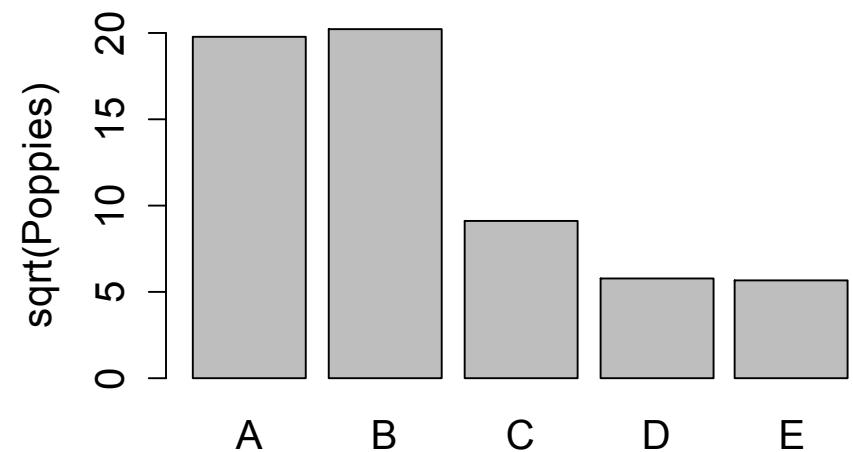
Three options:

1. Present the means on the transformed scale.
2. Present the means on the original scale, but note that analysis was done on transformed data.
3. Back transform to the original scale.

Option 1: Present the means (or graphs) and the comparison of means in the transformed scale. This is a reasonable option when the scale is common in that field of study (e.g. log in chemistry, sqrt in radiological sciences.)

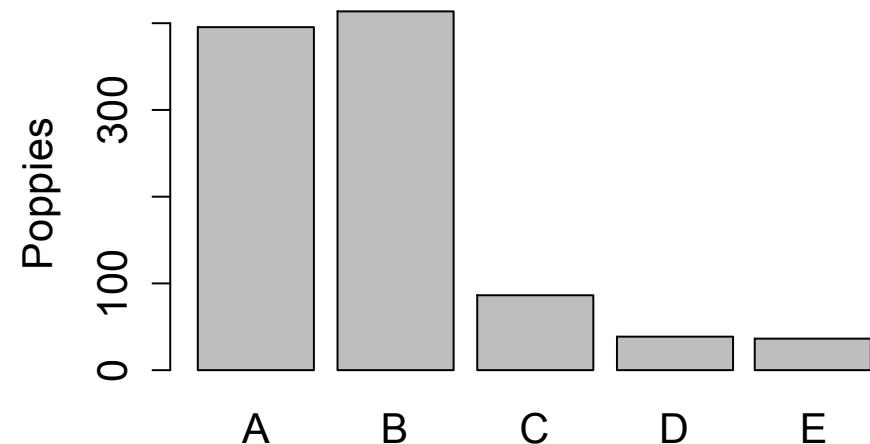
Example: Poppy plants in oats (see R example).

Trt	Orig Scale	Sqrt Scale	Back Transformed
A	394.75	19.83	393.09
B	413.00	20.23	409.07
C	86.75	9.08	82.47
D	37.75	5.75	33.11
E	35.25	5.64	31.89



Option 2: Present the means (or graphs) in the original scale, but with the comparison of means based on the transformed scale. Describe what you have done in your methods section and in a footnote to the table or graph of means.

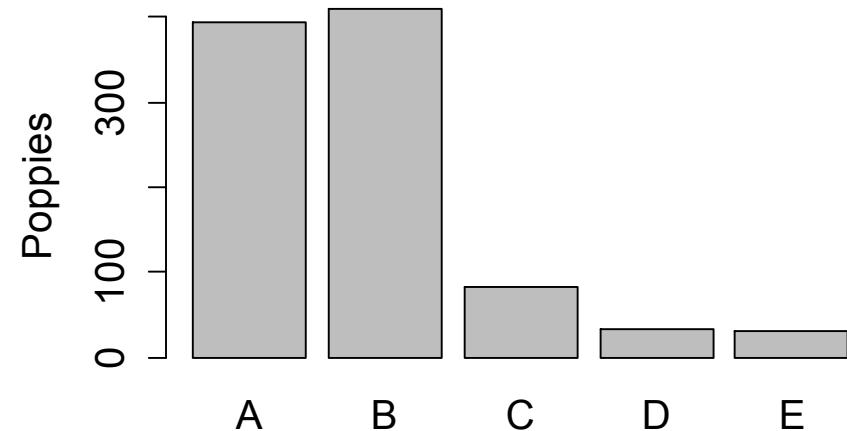
Trt	Orig Scale	Sqrt Scale	Back Transformed
A	394.75	19.83	393.09
B	413.00	20.23	409.07
C	86.75	9.08	82.47
D	37.75	5.75	33.11
E	35.25	5.64	31.89



Option 3: Compute the means using the transformed variable, but “back-transform” the means before presentation. To back-transform a square root transformed variable, square the mean.

Backtransformed means can be substantially lower than the means in the original scale, particularly when the transformation is log.

Trt	Orig Scale	Sqrt Scale	Back Transformed
A	394.75	19.83	393.09
B	413.00	20.23	409.07
C	86.75	9.08	82.47
D	37.75	5.75	33.11
E	35.25	5.64	31.89



Comments about interpretation after **log** transformation:

There can be interpretational advantages to do an analysis in the **log** scale. Because $\log(x/y) = \log(x) - \log(y)$, differences between means in the log analysis can be interpreted as ratios in the original scale.

In bioinformatics, it is common to use a log₂ transformation to satisfy ANOVA assumptions, but also because the differences can be interpreted as log₂ fold change (FC) values.

$$\log_2^{(2/1)} = +1, \log_2^{(1/2)} = -1$$

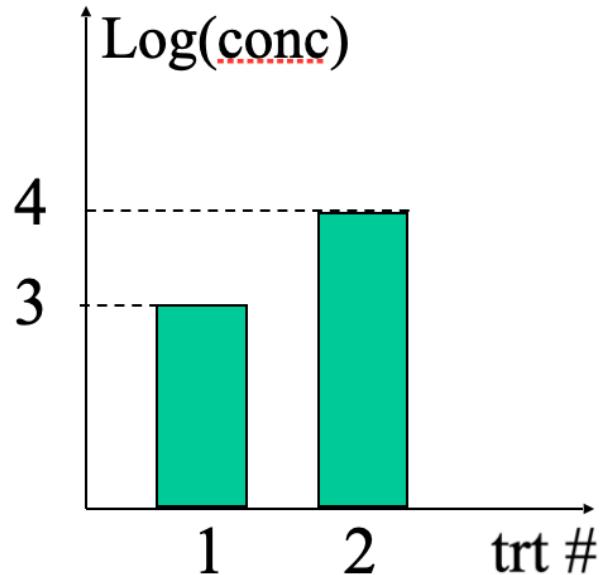
When using a log transformation, it is common to see people back transform the difference to a “ratio” scale after analysis.

A Lognormal Example:

Let \bar{x}_1 and \bar{x}_2 be the sample means of log-scaled observations for trt1 and trt2, respectively.

Let \bar{x}_{1O} and \bar{x}_{2O} be the sample means of original-scaled observations for trt1 and trt2, respectively.

$$\rightarrow \bar{x}_{1O} = e^{\bar{x}_1} \text{ and } \bar{x}_{2O} = e^{\bar{x}_2}$$



$$\text{Est. trt diff (log scale)} = \bar{x}_1 - \bar{x}_2 = 3 - 4 = -1$$

$$\text{Est. ratio of trt means (orig. scale)} = \bar{x}_{1O}/\bar{x}_{2O} = e^{\bar{x}_1}/e^{\bar{x}_2} = e^{\bar{x}_1 - \bar{x}_2}$$

$$e^{\bar{x}_1 - \bar{x}_2} = e^{3-4} = e^{-1} = 1/2.718 = 0.368 \approx 37\%$$

Trt1 concentration is 37% Trt2 concentration!

A C.I. for this % is given by: (e^{LCL}, e^{UCL})

(where (LCL,UCL) is a C.I. for the difference in the means of the log data)

3B. The Kruskal-Wallis Test

- The Kruskal-Wallis test is a non-parametric alternative to the one-way ANOVA F-test. This test does NOT require the assumption of normality.
- The Kruskal-Wallis test is an extension of the Wilcoxon Rank Sum test for more than two groups. Both are rank-based methods and hence robust to outliers.
- This test is typically thought of as a test of **medians**. But it is more correct to think of it as a test of shift in distribution.
- In R, use `kruskal.test()`.

Ch9 Multiple Comparison

After using an ANOVA test to analyze overall differences among treatment means, many researchers use **multiple comparison procedures** to determine which treatment means differ from each other.

The hard part is to understand why a multiple testing adjustment is needed and what effect it has on the analysis.

- We will use a simulation study to see why the adjustment is needed.
- To see the effect, we will show calculations for various margin of errors.

The Ch 9 notes includes:

- **Pairwise comparison (notes09.1)**
- Comparison of Treatments to a Control and comparison for contrasts (notes09.2)

Ch 9.1: Pairwise comparison

1. Fisher's Least Significant Difference Method
2. Comparisonwise vs Experimentwise Error Rate
3. Bonferroni's Method
4. Tukey's Method
5. Comparison of Methods
6. Implementation of Methods in R

Clover Example: (Steele and Torrie)

- Red Clover inoculated with $t = 6$ bacteria strains (Strain). The response variable is nitrogen content (N). The goal of the study is to compare means for the six treatments (bacteria strains).
- Five pots are randomly assigned to each of six treatments in a greenhouse experiment ($t = 6$ trts, $n = 5$ pots/trt, $n_T = 30$).
- The one-way ANOVA F-test (Ch8) is the overall test to see if any difference among treatment means.

```
> OneWayFit <- lm(N ~ Strain, data = Clover)
> anova(OneWayFit)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Strain      5 847.05 169.409   14.37 1.485e-06
Residuals  24 282.93  11.789
```

- Based on results from the one-way ANOVA F-test ($F = 14.37$, p-value < 0.0001), we reject H_0 and conclude there is some difference among treatment means.

- But which means are different?
 - further consider **multiple comparison procedures**
 - Notes09.1 focuses on **pairwise comparison** and we will look at
 - Fisher's Least Significant Difference method
 - Bonferroni's Method
 - Tukey's Method

1. Fisher's Least Significant Difference (LSD) method

Assumptions: Random sample, independent observations, normally distributed residuals, equal variances.

The Fisher's LSD test:

$$H_0: \mu_i = \mu_j \text{ vs. } H_1: \mu_i \neq \mu_j \text{ for any pair } (i, j) \text{ with } 1 \leq i, j \leq t$$

Test Statistic: $t_0 = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{df}$, under H_0 ,

where $\hat{\sigma} = \sqrt{s_W^2} = \sqrt{MSResid}$ and $df = dfResid = n_T - t$.

(Two-sided) P-value: $2 * P(t \geq |t_0|)$

R code: $2 * (1 - pt(\text{abs}(t_0), df))$

The $(1 - \alpha)100\%$ CI for the difference between two means:

$$\bar{y}_{i.} - \bar{y}_{j.} \pm t_{\frac{\alpha}{2}, df} \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Margin of error (ME)

where $\hat{\sigma} = \sqrt{s_W^2} = \sqrt{MSResid}$ and $df = dfResid = n_T - t$

- If $n_i = n_j = n$, the ME reduces to: $ME = t_{\alpha/2, df} \hat{\sigma} \sqrt{\frac{2}{n}}$.

NOTE: The Fisher's LSD test and its corresponding CI look VERY similar to methods from CH6 (comparing two means assuming equal variances). The only difference:

$$\hat{\sigma}^2 = S_p^2 \text{ for the t-test} — \hat{\sigma}^2 = MSResid \text{ for the Fisher's LSD test}$$

Comments about Fisher's LSD method

- Note that we also reject $H_0 (\mu_i - \mu_j = 0)$ using the CI - it does not contain 0.
Equivalently, reject H_0 (Two means are “significantly” different) if $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > ME$.
- Using this approach is equivalent to using the p-value or rejection region to make a decision about a test.

Return to the Clover example:

t = 6 treatments,

n = 5 observations per treatment

$$\hat{\sigma}^2 = s_W^2 = MSResid = 11.79$$

$$dfResid = 30 - 6 = 24$$

$$qt(0.975, df = dfResid) = 2.064$$

$$ME = t_{\alpha/2, df} \hat{\sigma} \sqrt{\frac{2}{n}} = 2.064 \sqrt{11.79 \times \frac{2}{5}} = 4.5$$

Thus, two means are “significantly” different if $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > 4.5$.

compos - 3D0k1	-10.12
compos - 3D0k13	5.44
compos - 3D0k4	4.06
compos - 3D0k5	-5.28
compos - 3D0k7	-1.22
3D0k1 - 3D0k13	15.56
3D0k1 - 3D0k4	14.18
3D0k1 - 3D0k5	4.84
3D0k1 - 3D0k7	8.90
3D0k13 - 3D0k4	-1.38
3D0k13 - 3D0k5	-10.72
3D0k13 - 3D0k7	-6.66
3D0k4 - 3D0k5	-9.34
3D0k4 - 3D0k7	-5.28
3D0k5 - 3D0k7	4.06

More comments about Fisher's LSD method

- It is a **one-at-a-time** method - one **particular** pair of two treatments is tested at the significance level α (the error rate).
 - Error rate α is the probability of type I error.
- α in **Fisher's method** is the error rate for doing a single pairwise test.
- However, in many problems, the experimenter may wish to conduct several or all pairwise tests **simultaneously**. Then what happens to the error rate for doing these tests simultaneously?
 - Fisher's method is the **unadjusted comparison**, meaning not adjusted for multiple testing.

2. Comparisonwise vs Experimentwise Error Rate

- Recall that **false rejection (or type I error)** means that we reject H_0 when H_0 is really true.
- **Comparisonwise error rate (CER)** is the probability of a false rejection on a single test. This is what we have focused on so far. The textbook also calls this the “individual comparisons” error rate.
eg: Using Fisher’s test to compare a single pairwise treatment means.
- **Experimentwise error rate (EER)** is the probability of having at least one false rejection in the group of tests from a single experiment or study.
eg: For the Clover data with $t=6$ treatments, we have 15 pairwise comparisons.
- The Fisher’s method for comparing all pairs of treatment means controls the **comparisonwise error rate (CER)** for each individual pairwise comparison but **does not control the experimentwise error rate (EER)** for a set of pairwise comparisons.

To look at this vividly, we consider simulations below.

- Consider two scenarios, $t = 5$ and 10 treatment groups with $n = 10$ observations per group
- For each scenario,
 1. Generate data with true $H_0: \mu_1 = \mu_2 = \dots = \mu_t$
 2. Using Fisher's method for pairwise comparisons (unadjusted) with $\alpha = 0.05$
 3. Repeat Steps 1-2 1000 times
- Results:
 - With $t = 5$ treatment groups, the observed EER is 29%.
 - With $t = 10$ treatment groups, the observed EER is 61%.
- In other words, 29% and 61% of all tests with $t = 5$ and 10 , respectively, in which treatments are really the same, will find evidence of at least one difference using unadjusted comparisons.
- The scientific community is generally unwilling to accept such high error rates. How do we get control of the EER?
 - Bonferroni's Method and Tukey's Method

3. Bonferroni's Method

- Bonferroni's method controls the experimentwise error rate for any set of m tests (not restricted to pairwise comparisons).
- Let α_E represent the experimentwise error rate (EER) and α_C represent the comparisonwise error rate (CER). Bonferroni's inequality states that $\alpha_E \leq m\alpha_C$.
- Hence, if we want to control the experimentwise error rate, α_E , at a fixed level α , we need to use $\alpha_C = \alpha/m$ for each of the m tests.
- To incorporate the Bonferroni adjustment with m tests:
 - Calculate Bonferroni adjusted p-values by multiplying the (unadjusted) p-values by m . If p-values come out to be greater than 1, just report a value of 1.
 - Calculate the Bonferroni adjusted ME using α/m (instead of α)

Return to the Clover example:

$t = 6$ treatments $\rightarrow m = 15$ tests of pairwise comparisons

$n = 5$ observations per treatment

$$\hat{\sigma}^2 = s_W^2 = MSResid = 11.79$$

$$dfResid = 30 - 6 = 24$$

$$qt(1 - (0.05/m)/2, df = 24) = 3.258$$

$$BonME = t_{(\alpha/m)/2, df} \hat{\sigma} \sqrt{\frac{2}{n}} = 3.258 \sqrt{11.79 \times \frac{2}{5}} = 7.1$$

$$\text{Recall that } FisherME = UnadjME = t_{\alpha/2, df} \hat{\sigma} \sqrt{\frac{2}{n}} = 2.064 \sqrt{11.79 \times \frac{2}{5}} = 4.5$$

BonME > FisherME, hence we will find evidence of fewer differences using Bonferroni's method.

Comments about Bonferroni's Method

1. **Bonferroni's method is NOT commonly used for pairwise comparisons.** (will see why via simulations soon.)

2. However, it is still a handy test to be aware of because
 - (1) it can be used for any set of m tests (not just pairwise comparisons after ANOVA)
 - (2) it is very easy to implement “by hand”.

4. Tukey's Method

- John Tukey proposed a method that can be used to run all pairwise comparisons while controlling EER.
- This method is based on a “Honestly Significant Difference”, which gives

$$TukeyME = \frac{q_\alpha(t, df)}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

- If $n_i = n_j = n$, the TukeyME reduces to:

$$TukeyME = q_\alpha(t, df) \hat{\sigma} \sqrt{\frac{1}{n}}$$

- $q_\alpha(t, df)$: the upper α percentage points of *studentized range statistic*, q , where t is the number of treatments, and df is the number of degrees of freedom associated with the *MSResid* ($df = dfResid = n_T - t$).
- $\hat{\sigma} = \sqrt{s_W^2} = \sqrt{MSResid}$

- Find the Tukey q using R:

`qtukey((1-\alpha), t, df)`

where t = number of treatments, $df = dfResid$.

Return to the Clover example:

$t = 6$ treatments $\rightarrow m = 15$ tests of pairwise comparisons

$n = 5$ observations per treatment

$$\hat{\sigma}^2 = s_W^2 = MSResid = 11.79$$

$$dfResid = 30 - 6 = 24$$

$$qtukey(0.95, 6, df = 24) = 4.37$$

$$TukeyME = q_\alpha(t, df) \hat{\sigma} \sqrt{\frac{1}{n}} = 4.37 \sqrt{11.79 \times \frac{1}{5}} = 6.7$$

$$\text{Recall that } FisherME = UnadjME = t_{\alpha/2, df} \hat{\sigma} \sqrt{\frac{2}{n}} = 2.064 \sqrt{11.79 \times \frac{2}{5}} = 4.5$$

TukeyME > FisherME, hence we will find evidence of fewer differences using Tukey's method.

Comments about Tukey's Method

1. Tukey's method controls maximum EER, without the need for "F-protection". So you can consider Tukey adjusted pairwise comparisons regardless of F-test results.
2. In other words, it is possible to have F test with $p\text{-value} < \alpha$, but have no evidence of differences from pairwise comparisons after Tukey adjustment.
3. The TukeyME is an (adjusted) ME for pairwise comparisons. It can also be used to construct CIs for pairwise comparisons.

5. Comparison of Methods

Asking which method is the best is really asking which is the “correct” α .
The answer depends on the seriousness of a type I error.
Below are the ME values for the Clover data.

Fisher/Unadjusted

4.5

Tukey

6.7

Bonferroni

7.1

lower type I error rate, lower power

fewer differences (“conservative”)

higher type I error rate, higher power

more differences (“liberal”)

Comments about Multiple Testing Adjustments

1. Using a multiple testing adjustment yields higher p-values (or wider confidence intervals) than unadjusted.
2. Using a multiple testing adjustment, running more tests yields higher p-values (or wider confidence intervals) than unadjusted.
3. Both Tukey and Bonferroni methods control the EER. However, Bonferroni is more conservative and hence will yield evidence of fewer differences. Thus, **Bonferroni's method is NOT commonly used for pairwise comparisons.**
4. Tukey adjustment is very common. Tukey controls EER and is simple. Many people are familiar with it. It is slightly conservative.
5. Though Bonferroni is too conservative for running all pairwise comparisons, it can be a good choice for other situations.
6. It is very important to report what method was used.
7. We can also use the ME to calculate “simultaneous confidence intervals”. For example, the probability that all these intervals simultaneously contain the true differences is 0.95.

Simulation for comparing three methods

- Consider two scenarios, $t = 5$ and 10 treatment groups with $n = 10$ observations per group
- For each scenario,
 1. Generate data with true $H_0: \mu_1 = \mu_2 = \dots = \mu_t$
 2. Using unadjusted pairwise comparisons (Fisher's LSD method) with $\alpha = 0.05$
 3. Repeat Steps 1-2 1000 times

Recall:

CER = comparison-wise error rate

EER = experiment-wise error rate.

Conclusions from Simulation

Fisher/Unadjusted: $CER \approx 0.05$; $EER \gg 0.05$

With $t = 5$ treatment groups, the observed EER is 29%.

With $t = 10$ treatment groups, the observed EER is 61%.

Unadjusted method controls CER but not EER.

Tukey: $CER \ll 0.05$; $EER \approx 0.05$

With $t = 5$ treatment groups, the observed EER is 5.5%.

With $t = 10$ treatment groups, the observed EER is 4.0%.

Tukey's method controls EER by reducing the CER.

Bonferroni: $CER \ll 0.05$; $EER < 0.05$

With $t = 5$ treatment groups, the observed EER is 3.9%.

With $t = 10$ treatment groups, the observed EER is 2.8%

Bonferroni's method over-controls the EER. Hence we say this method is “conservative”.

6. Implementation of Methods in R

- In R, we will use the `emmeans()` function from the `emmeans` package to get p-values for all pairwise comparisons.
- Emmeans stands for “expected marginal means”. These are model based estimates.
- Important Note: With different methods, we have to specify the option “adjust” in `emmeans()`:
 - `adjust = “none”` for Fisher’s (unadjusted) method
 - `adjust = “bonferroni”` for Bonferroni’s method
 - `adjust` with the default setting for Tukey’s method

Clover Example: Pairwise Comparisons using emmeans()

```
> library(emmeans)
> OneWayFit <- lm(N ~ Strain, data = Clover)
> emout <- emmeans(OneWayFit, ~ Strain)
```

```
> pairs(emout, adjust = "none")
contrast      estimate    SE  df t.ratio p.value
compos - 3DOK1   -10.12  2.17 24  -4.660  0.0001
compos - 3DOK13    5.44  2.17 24   2.505  0.0194
compos - 3DOK4    4.06  2.17 24   1.870  0.0738
compos - 3DOK5   -5.28  2.17 24  -2.431  0.0229
compos - 3DOK7   -1.22  2.17 24  -0.562  0.5794
3DOK1 - 3DOK13  15.56  2.17 24   7.166 <.0001
3DOK1 - 3DOK4   14.18  2.17 24   6.530 <.0001
3DOK1 - 3DOK5    4.84  2.17 24   2.229  0.0354
3DOK1 - 3DOK7    8.90  2.17 24   4.099  0.0004
3DOK13 - 3DOK4  -1.38  2.17 24  -0.636  0.5311
3DOK13 - 3DOK5  -10.72  2.17 24  -4.937 <.0001
3DOK13 - 3DOK7  -6.66  2.17 24  -3.067  0.0053
3DOK4 - 3DOK5   -9.34  2.17 24  -4.301  0.0002
3DOK4 - 3DOK7   -5.28  2.17 24  -2.431  0.0229
3DOK5 - 3DOK7    4.06  2.17 24   1.870  0.0738
```

```

> pairs(emout, adjust = "bonferroni")

```

contrast	estimate	SE	df	t.ratio	p.value
compos - 3DOK1	-10.12	2.17	24	-4.660	0.0015
compos - 3DOK13	5.44	2.17	24	2.505	0.2914
compos - 3DOK4	4.06	2.17	24	1.870	1.0000
compos - 3DOK5	-5.28	2.17	24	-2.431	0.3431
compos - 3DOK7	-1.22	2.17	24	-0.562	1.0000
3DOK1 - 3DOK13	15.56	2.17	24	7.166	<.0001
3DOK1 - 3DOK4	14.18	2.17	24	6.530	<.0001
3DOK1 - 3DOK5	4.84	2.17	24	2.229	0.5317
3DOK1 - 3DOK7	8.90	2.17	24	4.099	0.0062
3DOK13 - 3DOK4	-1.38	2.17	24	-0.636	1.0000
3DOK13 - 3DOK5	-10.72	2.17	24	-4.937	0.0007
3DOK13 - 3DOK7	-6.66	2.17	24	-3.067	0.0793
3DOK4 - 3DOK5	-9.34	2.17	24	-4.301	0.0037
3DOK4 - 3DOK7	-5.28	2.17	24	-2.431	0.3431
3DOK5 - 3DOK7	4.06	2.17	24	1.870	1.0000

P value adjustment: bonferroni method for 15 tests

```

> pairs(emout)
contrast      estimate    SE  df t.ratio p.value
compos - 3DOK1     -10.12 2.17 24  -4.660 0.0012
compos - 3DOK13      5.44 2.17 24   2.505 0.1622
compos - 3DOK4      4.06 2.17 24   1.870 0.4435
compos - 3DOK5     -5.28 2.17 24  -2.431 0.1852
compos - 3DOK7     -1.22 2.17 24  -0.562 0.9926
3DOK1 - 3DOK13     15.56 2.17 24   7.166 <.0001
3DOK1 - 3DOK4     14.18 2.17 24   6.530 <.0001
3DOK1 - 3DOK5      4.84 2.17 24   2.229 0.2617
3DOK1 - 3DOK7      8.90 2.17 24   4.099 0.0049
3DOK13 - 3DOK4    -1.38 2.17 24  -0.636 0.9871
3DOK13 - 3DOK5    -10.72 2.17 24  -4.937 0.0006
3DOK13 - 3DOK7    -6.66 2.17 24  -3.067 0.0528
3DOK4 - 3DOK5     -9.34 2.17 24  -4.301 0.0030
3DOK4 - 3DOK7     -5.28 2.17 24  -2.431 0.1852
3DOK5 - 3DOK7      4.06 2.17 24   1.870 0.4435

```

P value adjustment: tukey method for comparing a family of 6 estimates

“Simple summary” statistics vs emmeans()

- Recall that emmeans stands for “expected marginal means”. These are model based estimates.
- For one-way ANOVA, the simple means and emmeans will be the same, even if the sample sizes are not balanced across groups.
- However, there will be a difference between the “simple” SE and the model based SE (from emmeans):

$$\text{“Simple” SE for group } i = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_i}{\sqrt{n}} \quad \text{v.s. Model based SE for group } i = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_W}{\sqrt{n}}$$

- The difference is that the simple SE allows standard deviation to be estimated **separately** for each group, while the model based SE uses a pooled estimate of standard deviation, $s_W = \sqrt{MSResid}$ (this makes sense because the ANOVA model assumes equal variance.).
- Note that if sample sizes are equal, the model based SE will be the same for all groups. This is true for the clover example.

Clover Example

```
> SumStats
  Strain n   mean      sd      se
1  compos 5  18.7  1.60  0.716
2  3DOK1  5  28.8  5.80  2.59
3  3DOK13 5  13.3  1.43  0.638
4  3DOK4  5  14.6  4.12  1.84
5  3DOK5  5  24.0  3.78  1.69
6  3DOK7  5  19.9  1.13  0.505

> library(emmeans)
> OneWayFit <- lm(N ~ Strain, data = Clover)
> emout <- emmeans(OneWayFit, ~ Strain)
> emout
Strain emmean    SE df lower.CL upper.CL
compos     18.7 1.54 24     15.5     21.9
3DOK1      28.8 1.54 24     25.7     32.0
3DOK13     13.3 1.54 24     10.1     16.4
3DOK4      14.6 1.54 24     11.5     17.8
3DOK5      24.0 1.54 24     20.8     27.1
3DOK7      19.9 1.54 24     16.8     23.1
```

Generating a “CLD” display using emmeans()

- We can also generate a “CLD” display using emmeans(). It will generate a warning, but the results are fine. Any two means that are not “significantly” different will be indicated with a shared number.
- Recall that we reject H_0 (Two means are “significantly” different) if $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > ME$. Using this approach is equivalent to using p-value or rejection region.
- Either way, the results are sometimes summarized in a simple CLD/lines display:
 - Order the means first.
 - Then identify any two means that are not “significantly” different with a shared line or number or letter.
- WARNING: If there are unequal sample sizes for groups, then “statistical significance” is based on magnitude of difference and sample size. For this reason, there is no fixed LSD (=ME) value when sample sizes are unequal. Resulting p-values, CIs, etc are fine. But “lines”/CLD display should be interpreted with caution.

Clover Example: CLD/Lines Display “by hand”

$$FisherME = UnadjME = t_{\alpha/2, \text{df}} \hat{\sigma} \sqrt{\frac{2}{n}} = 2.064 \sqrt{11.79 \times \frac{2}{5}} = 4.5$$

Compare $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > 4.5$

Two means are “significantly” different if $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > 4.5$.

```
> CLD(emout, adjust = "none")
```

Strain	emmmean	SE	df	lower.CL	upper.CL	.group
3D0k13	13.3	1.54	24	10.1	16.4	1
3D0k4	14.6	1.54	24	11.5	17.8	12
compos	18.7	1.54	24	15.5	21.9	23
3D0k7	19.9	1.54	24	16.8	23.1	34
3D0k5	24.0	1.54	24	20.8	27.1	4
3D0k1	28.8	1.54	24	25.7	32.0	5

Recall: Two means that are not “significantly” different are indicated with a shared line or number (or letter).

Ch9 Multiple Comparison

After using an ANOVA test to analyze overall differences among treatment means, many researchers use **multiple comparison procedures** to determine which treatment means differ from each other.

The Ch 9 notes includes:

- Pairwise comparison (notes09.1)
- **Comparison of Treatments to a Control and comparison for contrasts (notes09.2)**

Ch 9.2: Comparison of Treatments to a Control and comparison for contrasts

1. Dunnett's Method for Comparisons against a Control
2. (Linear) Contrasts
3. Multiple Comparisons for Contrasts
4. False Discovery Rate (Optional)
5. More on Multiple Comparison (Optional)

1. Dunnett's Method for Comparisons against a Control

- In some cases, the primary objective of an experiment is to compare a single “control” treatment to all other (“active”) treatments.
- Dunnett’s method is designed to control **the experimentwise error rate** in the family of comparisons of individual means vs control.
- Note that this is a subset of all pairwise comparisons, since we only consider $t-1$ tests.
- We focus on the core of equal sample sizes ($n_1 = \dots = n_t = n$) and have

$$DunnettME = t_{Dunnett} \sqrt{\frac{2\hat{\sigma}^2}{n}} = t_{Dunnett} \sqrt{\frac{2\hat{s}_W^2}{n}} = t_{Dunnett} \sqrt{\frac{2MSResid}{n}}$$

- We can use Table 11 in the textbook to find the Dunnett table value with $k = t = \# \text{ treatments}$ and $v = df = dfResid$
- Unfortunately there is no convenient way to calculate the Dunnett table value in R. (For this reason, we are unlikely to ask for a calculation of the Dunnett table value.)

Return to the Clover example:

$t = 6$ treatments,

$n = 5$ observations per treatment

$$\hat{\sigma}^2 = s_W^2 = MSResid = 11.79$$

$$dfResid = 30 - 6 = 24$$

From Table 11: $t_{Dunnett} = 2.76$

$$s_W^2 = MS Resid = \hat{\sigma}^2 = 11.79$$

$$DunnettME = 2.76 \sqrt{\frac{2(11.79)}{5}} = 5.99$$

For comparison: DunnettME = 5.99 is smaller than TukeyME = 6.71, but larger than FisherME = UnadjME= 4.5.

Comments about Dunnett's Method:

1. In R, we will calculate Dunnett adjusted p-values using the emmeans package. Specifically:
`emmeans(OneWayFit, dunnett ~ trt)`
2. Using R emmeans, the first group is used as “control”. May need to reorder factor levels so that the control group of interest is first. See Clover Example in CH9p1_R.pdf or CH9p2_R.pdf for reordering factor levels.
3. Dunnett's method increases power, compared to Tukey, by reducing the number of tests.
4. Tukey's method can also be used, when we are interested in comparing “active” treatments verses “control”. However, since Tukey controls EER when comparing all pairs of treatments, it will have unnecessarily low power. We are “paying a price” for testing comparisons that aren't really of interest.

2. (Linear) Contrasts in the One-way ANOVA

- In many cases, ANOVA and pairwise comparisons of means will address all research questions.
- But occasionally, **contrasts** are required to address additional comparisons of interest.
- A linear contrast (usually called just a **contrast**) provides one measure of the difference in treatment means.

- A linear parameter $l = a_1\mu_1 + a_2\mu_2 + \cdots + a_t\mu_t = \sum_{i=1}^t a_i\mu_i$ is called a **contrast** if $\sum_{i=1}^t a_i = 0$.
 - The contrast is described by a list of coefficients (a_1, a_2, \dots, a_t).
 - eg. $l_1 = \mu_1 - \mu_2, l_2 = 2\mu_3 - \mu_2 - \mu_1, l_3 = 2\mu_1 - 3\mu_2 + \mu_3$
- Obviously, substituting the μ_i 's by sample means yields an (unbiased) estimator of l , $\hat{l} = a_1\bar{y}_1 + a_2\bar{y}_2 + \cdots + a_t\bar{y}_t = \sum_{i=1}^t a_i\bar{y}_i$.

A few important notes:

1. In order to estimate and test a contrast in R, we supply the coefficients (a_i 's).
2. The number (and order) of coefficients needs to match the number (and order) of the treatments/groups.
3. Some a_i 's may be zero.

Wheat Contrasts Example: An experiment is performed to compare yield for $t =$ four varieties (A, B, C, D) of wheat. Varieties A and B are similar in that they are classified as “resistant” to a particular disease. Varieties C and D are classified as “susceptible”.

Sample means: A = 5.54 , B = 5.16, C = 4.06 , D = 7.42

Contrast Example #1: Compare variety A vs B

$$H_0: \mu_A - \mu_B = 0 \quad \text{or} \quad 1 \times \mu_A - 1 \times \mu_B + 0 \times \mu_c + 0 \times \mu_D = 0$$

Coefficients (a_i 's): (1, -1, 0, 0)

(Check Sum to Zero: $+1 - 1 + 0 + 0 = 0$)

$$\text{Estimate} = \bar{y}_A - \bar{y}_B = 5.54 - 5.16 = 0.38$$

*This is just a “pairwise comparison”. So, we can get an estimate and test of this comparison using `emmeans` and `pairs`.

Contrast Example #2: Compare average of resistant varieties (A, B) versus average of susceptible varieties (C, D).

$$H_0: \frac{\mu_A + \mu_B}{2} - \frac{\mu_C + \mu_D}{2} = 0$$

$$l = \frac{1}{2}\mu_A + \frac{1}{2}\mu_B - \frac{1}{2}\mu_C - \frac{1}{2}\mu_D$$

Coefficients (a_i 's): (0.5, 0.5, -0.5, -0.5)

(Check Sum to Zero: $+0.5 + 0.5 - 0.5 - 0.5 = 0$)

$$\begin{aligned} \text{Estimate} &= 0.5\bar{y}_A + 0.5\bar{y}_B - 0.5\bar{y}_C - 0.5\bar{y}_D \\ &= 0.5 * 5.54 + 0.5 * 5.16 - 0.5 * 4.06 - 0.5 * 7.42 = -0.39 \end{aligned}$$

*This comparison is not included in the pairwise comparisons!

Contrast Example #3: Compare average of variety D versus average of other varieties

$$H_0 : -1\left(\frac{\mu_A + \mu_B + \mu_C}{3}\right) + 1\mu_D = 0$$

$$l = -\frac{1}{3}\mu_A - \frac{1}{3}\mu_B - \frac{1}{3}\mu_C + 1\mu_D$$

Coefficients (a_i 's): (-1/3, -1/3, -1/3, 1)

Contrast Example #4: Compare average of variety C versus average of varieties A and B.

$$H_0 : -1\left(\frac{\mu_A + \mu_B}{2}\right) + 1\mu_C + 0\mu_D = 0$$

$$l = -\frac{1}{2}\mu_A - \frac{1}{2}\mu_B + 1\mu_C + 0\mu_D$$

Coefficients (a_i 's): (-0.5, -0.5, 1, 0)

*These comparison are not included in the pairwise comparisons! How to make these comparisons?

SE for Contrasts

We have seen how to set up contrasts (by specifying coefficients) and estimate them (using sample means).

However, any tests or confidence intervals require first calculating the standard error of \hat{l} which is given by

$$SE(\hat{l}) = \sqrt{\hat{\sigma}^2 \sum_{i=1}^t \frac{a_i^2}{n_i}} = \sqrt{s_W^2 \sum_{i=1}^t \frac{a_i^2}{n_i}} = \sqrt{MSResid \sum_{i=1}^t \frac{a_i^2}{n_i}}.$$

With equal sample sizes ($n_1 = \dots = n_t = n$), we have that

$$SE(\hat{l}) = \sqrt{\frac{MSResid}{n} \sum_{i=1}^t a_i^2}$$

Since a contrast is a linear combination of independent normal means, it is also **normally** distributed and thus hypotheses about a contrast can be tested using a t-test, or equivalently, an F-test. (remember $t^2 = F$)

CI and Test for Contrasts

Let $l = a_1\mu_1 + a_2\mu_2 + \dots + a_t\mu_t$ Then $\hat{l} = a_1\bar{y}_{1.} + a_2\bar{y}_{2.} + \dots + a_t\bar{y}_{t.}$

A $(1-\alpha) \times 100\%$ Confidence Interval for a contrast l :

$$\hat{l} \pm t_{\alpha/2, df} \sqrt{\hat{\sigma}^2 \sum_{i=1}^t \frac{a_i^2}{n_i}}$$

Hypothesis Test for a contrast l :

$$H_0 : l = l_0 \text{ vs } H_a : l \neq l_0$$

$$t_0 = \frac{\hat{l} - l_0}{\sqrt{\hat{\sigma}^2 \sum_{i=1}^t \frac{a_i^2}{n_i}}} \sim T_{df} \text{ under } H_0$$

(Two-sided) P-value: $2 * P(T \geq |t_0|)$

R code: `2 * (1-pt(abs(t0), df))`

Notes on Contrasts

1. In R, we will estimate and test contrasts using the `contrast()` function within the `emmeans` package.
2. The number (and order) of coefficients (a_i 's) needs to match the number (and order) of the treatments/groups. Some coefficients (a_i 's) may be zero.
3. In many cases, pairwise comparisons of means answers all of the relevant research questions. Occasionally, a specific contrast is needed to answer the research question.
4. Pairwise (unadjusted) comparisons of 2 means is a special case of a contrast.

3. Multiple Comparisons for contrasts

- NOTE: For this course, we will generally not adjust contrasts for multiple testing. We will use **unadjusted** p-values.
- A contrast is ***a priori*** if it was selected by the experimenter as being important for testing at the outset of the experiment (**before** looking at the data).
- It is usually considered acceptable to test a **small** number of *a priori* contrasts **without making any multiple comparison adjustments**.
- For **larger groups** of *a priori* contrasts, use **Bonferroni's** method. You will pay a penalty for testing a lot of contrasts, as Bonferroni adjusted p-values are calculated by multiplying the unadjusted p-values by the number of tests.

- A linear contrast is ***a posteriori*** if it was selected by the experimenter after looking at the data.
- **Scheffe's method** can be used for controlling **maximum EER** when testing a large number of *a priori* contrasts or testing any number of *a posteriori* contrasts. **Scheffe's method** is even more conservative than Bonferroni's method.
 - We will not cover the details of Scheffe's method in these notes!
 - Steel and Torrie call the process of searching the means for contrasts that show evidence of differences “Data Dredging”. Scheffe's method is designed for such contrasts.

4. False Discovery Rate (Optional)

- So far in these notes, we have focused on “classical” multiple testing scenarios. Pairwise comparisons of means from ANOVA is an example.
- In large scale multiple testing scenarios (hundreds or thousands of tests), different methods are used which tend to focus on the false discovery rate (FDR).
- $\text{FDR} = (\# \text{ False Rejections}) / (\text{Total } \# \text{ of Rejections})$
- For example, suppose a genomics experiment was conducted and 100 genes were identified as “differentially expressed” with a FDR of 5%. Then we would expect 5/100 of our differentially expressed genes to be “false positives”.

- Notice that the idea of FDR (proportion of rejections that are false) is very different from the idea of EER (proportion of experiments with at least one false rejection).
- FDR methods will be considerably more lenient (meaning more false rejections) but have higher power to detect a difference for a particular comparison. Again, this is a trade off.
- Most common FDR method is Benjamini-Hochberg but there are MANY other methods for controlling the FDR.

5. More on Multiple Comparison (Optional)

Other Multiple Testing Methods from R

From glht() multcomp package:

- Bonferroni
- Dunnett
- Holm
- Shaffer
- Tukey
- Westfall

From p.adjust():

- holm
- hochberg
- hommel
- bonferroni
- BH = fdr (Benjamini & Hochberg)
- BY (Benjamini & Yekutieli)

Power for Pairwise Comparisons with Lenth

Using Lenth's online power calculator:

<http://homepage.stat.uiowa.edu/~rlenth/Power/>

Ch10 Categorical Data

In CH5-9, we focused on inference for means and standard deviations. **Mean and standard deviation** are parameters for center and spread for a **numerical variables** (ex: hormone level, lead consumption).

In CH10, we focus on **categorical variables** (ex: infested or not infested, cold or no cold). In this case, the parameter of interest is **proportion**.

CH5-9: Normal, t, χ^2 , F distributions

- All continuous distributions

CH10: **Binomial and Poisson distributions**

- Both discrete distributions
- BUT sometimes these discrete distributions are approximated by continuous distributions!

The Ch 10 notes includes:

Binomial Distribution

- **Binomial Distribution and Its Approximation (notes10.1)**
- Inference for a Single Proportion π (notes10.2)
- Comparing ≥ 2 Proportions (notes10.3)
- Contingency Tables: Tests for Independence and Homogeneity (notes10.4)
- Odds Ratios (notes10.5)

Poisson Distribution (notes10.6)

Chapter 10.1: Binomial Distribution and Its Approximation

1. Binomial Distribution
2. Normal Approximation to the Binomial

1. Binomial Distribution

Consider an experiment that has only two outcomes (a binary response).
Examples: 0 or 1, Heads or Tails, Event or No Event, Success or Failure

Let $\pi = P(\text{observe a } 1)$; then $1-\pi = P(\text{observe a } 0)$.

Note: π represents a probability between 0 and 1. ($\pi \neq 3.14159$ here.)

Repeat the experiment **n** times **independently**

Let $Y = \text{number of } 1\text{'s observed out of the } n \text{ trials}$
= sum of the outcomes of the n trials

An experiment involving n repeats is called a **binomial experiment**, and Y is called a **binomial random variable**.

For $Y \sim \text{bin}(n, \pi)$, the probability mass function (pmf) is

$$P(Y = y) = \frac{n!}{y!(n-y)!} (\pi)^y (1 - \pi)^{(n-y)}, \text{ for } y = 0, 1, \dots, n$$

- $\text{mean}(Y) = \mu_Y = n\pi$ and $\text{var}(Y) = \sigma_Y^2 = n\pi(1-\pi)$
- Note $n! = n \cdot (n-1) \cdot (n-2) \dots 1$, and $0!$ defined as 1.

Example :

$n = 4$ (4 trials)

$\pi = 0.6$ (probability of event)

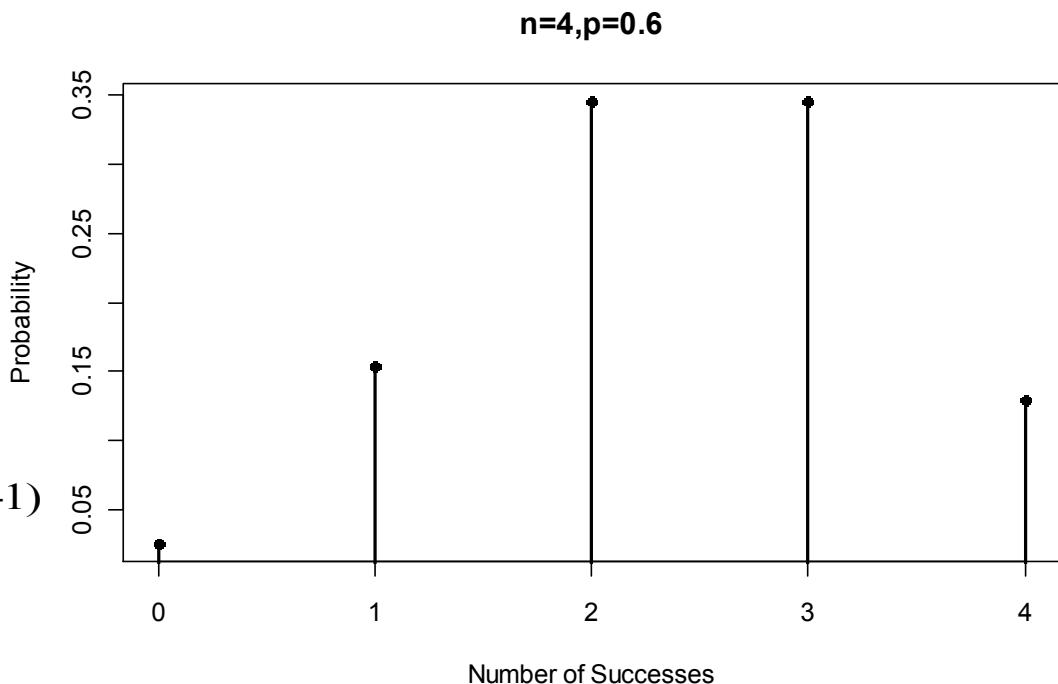
$y = 1$ (total # of events)

$$P(Y = 1)$$

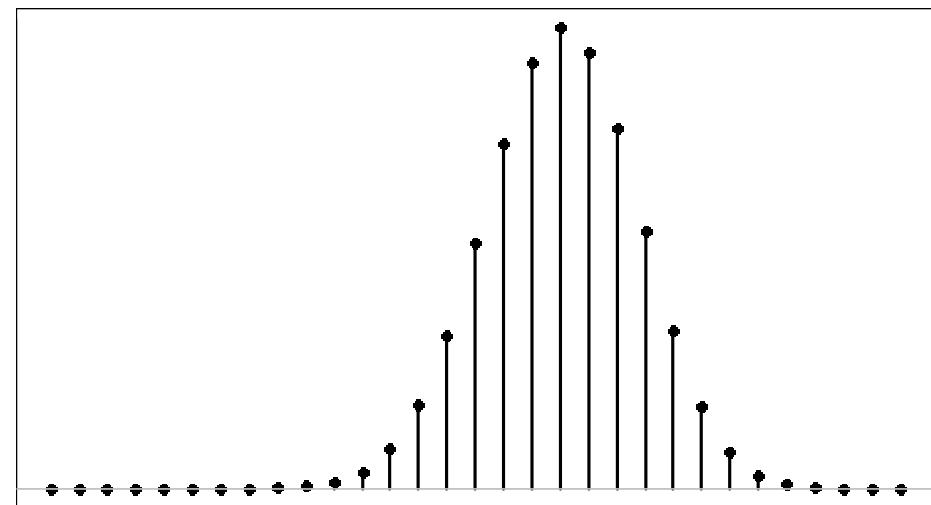
$$= \frac{4!}{1!(4-1)!} (0.6)^1 (1 - 0.6)^{(4-1)}$$

$$= 0.1536$$

Note : $0! = 1$ and $y^0 = 1$



With $n=30$ and $\pi=0.6$ the distribution looks close to **normal**.



Cumulative distribution functions (cdfs) gives the probability of being less than or equal to a value.

Example: For the Binomial distribution

$$P(Y \leq y) = \sum_{k=0}^y \frac{n!}{k!(n-k)!} (\pi)^k (1-\pi)^{(n-k)}$$

NOTE: Be careful about inequalities!

$$P(Y < y) \neq P(Y \leq y) !$$

$$P(Y \leq 2) = P(Y=0) + P(Y=1) + P(Y=2)$$

$$\begin{aligned} P(Y < 2) &= P(Y=0) + P(Y=1) \\ &= P(Y \leq 1) \end{aligned}$$

Binomial Probabilities in R

- In R, `dbinom()` gives $P(Y=y)$ or `pbinom()` gives $P(Y \leq y)$.

Example: Suppose $n=15$ and $\pi=0.5$.

```
> dbinom(5, size = 15, prob = 0.5)
[1] 0.09164429
> pbinom(5, size = 15, prob = 0.5)
[1] 0.1508789
> pbinom(5, 15, 0.5) - pbinom(4, 15, 0.5)
[1] 0.09164429
```

Mean and standard deviation of a binomial random variable

The mean and standard deviation of a population of outcomes from imaginary replications of the experiment.

If $Y \sim \text{bin}(n, \pi)$, then

$$\text{mean}(Y) = \mu_Y = n\pi \text{ and } \text{sd}(Y) = \sigma_Y = \sqrt{n\pi(1 - \pi)}$$

Example: $n=4$ $\pi=0.6$

$$\mu_Y = n\pi = (4)(0.6) = 2.4$$

$$\sigma_Y = \sqrt{n\pi(1 - \pi)} = \sqrt{(4)(0.6)(1 - .6)} = 0.98$$

2. Normal Approximation to the Binomial

Let X be the result of a single trial with probability of success π .

Let $X = 1$ in case of a success and 0 in case of failure.

Observe n independent trials X_1, X_2, \dots, X_n .

By the central limit theorem, the sample mean \bar{X} has a distribution that is approximately normal for large n .

$$\begin{aligned} Y &= \sum X_i = \# \text{successes} \\ &= n \left(\sum \frac{X_i}{n} \right) = n \bar{X} \end{aligned}$$

Y is a binomial random variable (RV), but Y is just a multiple of \bar{X} and \bar{X} is approximately normal, so Y must also be approximately normal.

Example: See binomial distribution with $n=30$ and $\pi=0.6$ (slide 5).

Normal Approximation to the Binomial (*Example*)

Example: Give a dose of insecticide to 30 flies. Assuming that the probability that an individual fly dies is 0.7, what is the probability that 25 or more die? $P(Y \geq 25) = 1 - P(Y < 25) = 1 - P(Y \leq 24)$

Let Y represent the number of dead flies. Then $Y \sim \text{bin}(n=30, \pi=0.7)$. (Here an “event” is a dead fly.)

EXACT binomial probability using R:

$$1 - \text{pbinom}(24, \text{size} = 30, \text{prob} = 0.7) = 0.077$$

Normal Approximation (without Continuity Correction):

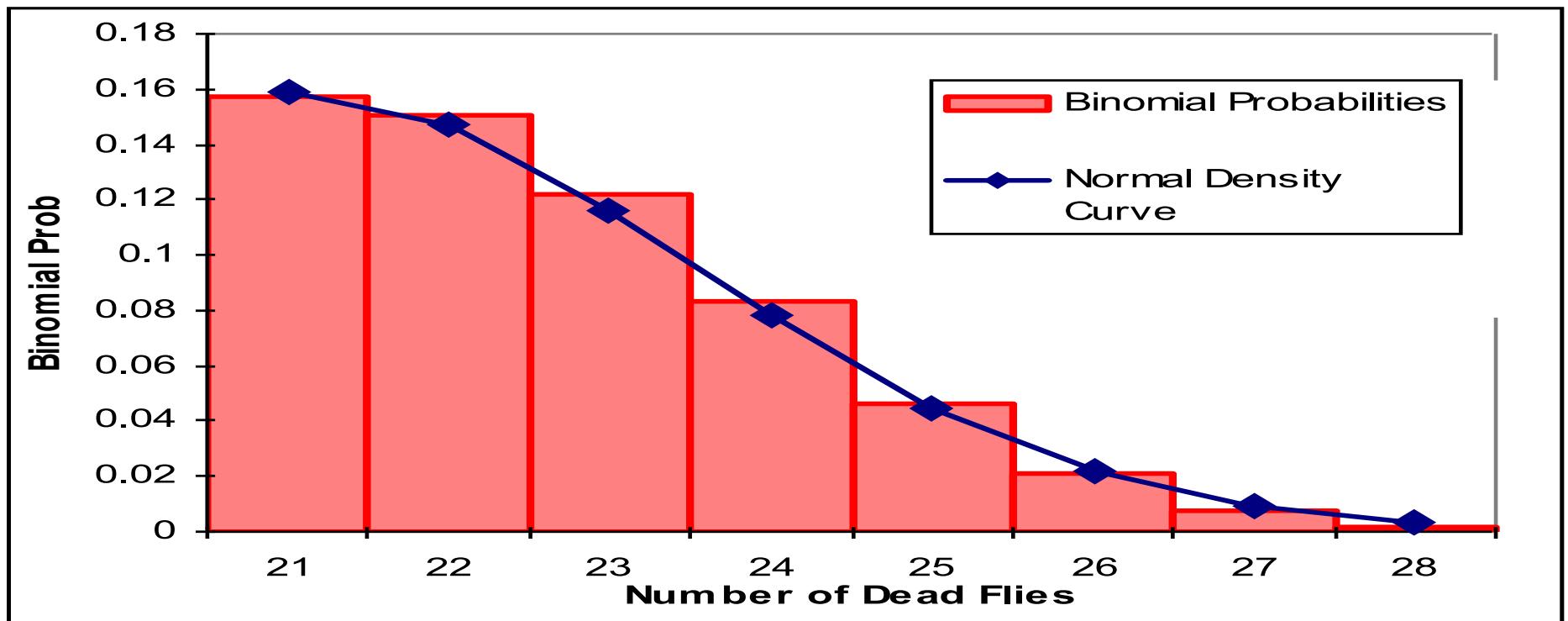
$$\begin{aligned}\mu_Y &= n\pi = 30(0.7) = 21 \\ \sigma_Y &= \sqrt{n\pi(1-\pi)} = \sqrt{30(0.7)(1-0.7)} = 2.51\end{aligned}$$

Consider $Y \sim N(\mu = 21, \sigma = 2.51)$

$$\begin{aligned}P(Y \geq 25) &= 1 - P(Y \leq 24) \\ &= 1 - \text{pnorm}(24, \text{mean} = 21, \text{sd} = 2.51) \\ &= 0.116\end{aligned}$$

Normal Approximation to the Binomial (*continuity correction*)

Continuity Correction: Use 24.5 in place of 24 to include whole 24 box. The idea extends to other situations.



Consider $Y \sim N(\mu = 21, \sigma = 2.51)$

$$\begin{aligned} P(Y \geq 25) &= 1 - P(Y \leq 24.5) \\ &= 1 - \text{pnorm}(24.5, \text{mean} = 21, \text{sd} = 2.51) \\ &= 0.082 \end{aligned}$$

If we can calculate exact probabilities, **why are we talking about the normal approximation and continuity correction?**

Because we will see that many standard approaches for inference for proportions are based on a normal approximation.

Examples include large sample CI and Z-test for a single proportion and chi-square test for contingency tables.

In R, we will see that these methods will usually include a “continuity correction” by default.

We will not be too concerned about the exact details of the continuity correction because it is most important when sample sizes are small. And in those cases, we prefer (exact) small sample methods. For example, Fisher’s Exact Test.

Ch10 Categorical Data

In CH5-9, we focused on inference for means and standard deviations. **Mean and standard deviation** are parameters for center and spread for a **numerical variables** (ex: hormone level, lead consumption).

In CH10, we focus on **categorical variables** (ex: infested or not infested, cold or no cold). In this case, the parameter of interest is **proportion**.

CH5-9: Normal, t, χ^2 , F distributions

- All continuous distributions

CH10: **Binomial and Poisson distributions**

- Both discrete distributions
- BUT sometimes these discrete distributions are approximated by continuous distributions!

The Ch 10 notes includes:

Binomial Distribution

- Binomial Distribution and Its Approximation (notes10.1)
- **Inference for a Single Proportion π (notes10.2)**
- Comparing ≥ 2 Proportions (notes10.3)
- Contingency Tables: Tests for Independence and Homogeneity (notes10.4)
- Odds Ratios (notes10.5)

Poisson Distribution (notes10.6)

Chapter 10.2: Inference for a Single Proportion π

Large Sample Inference (Normal Approximation)

1. Confidence interval and test
2. Sample size
 - A. Sample size through CI
 - B. Sample size through the power of test

Small Sample Inference (Exact Binomial)

3. Test and confidence interval
4. Power of test about π

Nonparametric method

5. Back to Sign Test for median

Large Sample Inference

(Using Normal Approximation)

Aphid Example #1: A researcher is interested in estimating the true proportion of fields in Larimer County that are infested with Russian Wheat Aphid (π).

Note that we can interpret π as the true proportion that are infested or the probability that a randomly selected field will have aphids.

Based on a random sample of $n = 100$ fields, $y = 53$ are found to be infested. Hence we estimate $\hat{\pi} = 53/100 = 0.53$.

$$\hat{\pi} = \frac{y}{n} = \frac{\# \text{ events}}{\# \text{ trials}}$$

- $\hat{\pi}$ is a statistic and we will consider its sampling distribution.
- It turns out that, in repeated samples, the **sampling distribution** of $\hat{\pi}$ is **approximately normal** distribution with mean π and variance $\frac{\pi(1 - \pi)}{n}$.
- The above sampling distribution of $\hat{\pi}$ is the foundation for the large sample inference of π including CI and test.

1. Confidence interval and test for single proportion

An approximate $(1-\alpha)100\%$ confidence interval for π is:

$$\hat{\pi} \pm Z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

where the table value $Z_{\alpha/2}$ is obtained from the Normal distribution using `qnorm(1-alpha/2)`.

Assumptions: Random sample, independent observations, large sample size.

Rule of thumb for sample size: $n\hat{\pi} \geq 5$ and $n(1 - \hat{\pi}) \geq 5$.

95% CI for the Aphid Example:

$$\begin{aligned}\hat{\pi} &= \frac{53}{100} = 0.53 & 0.53 \pm 1.96 \sqrt{\frac{(0.53)(1-0.53)}{100}} \\ && 0.53 \pm 0.098, \quad \text{i.e.,} \quad (0.432, 0.628)\end{aligned}$$

CI for Single Proportion in R

- In R, use `prop.test()`

For the Aphids Example #1 ($n=100$, $y=53$):

```
> prop.test(53, 100, correct = FALSE)

  1-sample proportions test without
continuity correction

data: 53 out of 100, null probability 0.5
X-squared = 0.36, df = 1, p-value = 0.5485
alternative hypothesis: true p is not equal to
0.5

95 percent confidence interval:
0.4328886 0.6248918

sample estimates:
p
0.53
```

Checking for Large Sample Size

Rule of thumb for large sample size: $n\hat{\pi} \geq 5$ and $n(1 - \hat{\pi}) \geq 5$.

Note that although we say “large sample” normal approximation, the rule of thumb is based on both n and $\hat{\pi}$.

The big picture idea is that the **normal approximation to the binomial distribution “fails” if π gets too close to 0 or 1**. The reason is that at these extremes, the binomial distribution is skewed (not “bell shaped”).

The exact binomial CI and test will be covered in the next section.

Aphid Example #1:

Intuitively with $n = 100$ and $\hat{\pi} = 0.53$, the large sample approximation is reasonable.

$$100 * 0.53 = 53 \geq 5 \text{ AND } 100 * (0.47) = 47 \geq 5$$

Both conditions are satisfied, so the large sample normal approximation is adequate.

Formulas used by textbook and R

For **Large sample** confidence interval for π , R and the book use (different) modifications of what is given in these notes. The differences will be most noticeable when sample size is small (but then we will want to use the exact binomial CI and test).

WAC CI (O&L): $\tilde{y} = y + 0.5z_{\alpha/2}^2 \quad \tilde{n} = n + z_{\alpha/2}^2 \quad \tilde{\pi} = \frac{\tilde{y}}{\tilde{n}}$

R prop.test(): $\left(\hat{\pi} + z^* \right) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^*}{2n}} / (1 + 2z^*)$

where $z^* = (z_{\frac{\alpha}{2}}^2)/2n$

Question: How can we tell what formula R is using?

Answer: Help is not helpful, so we need to use the source!

1. Confidence interval and test for single proportion

Assumptions: Random sample, independent observations, large sample size. Rule of thumb for sample size: $n\hat{\pi} \geq 5$ and $n(1 - \hat{\pi}) \geq 5$.

$$H_0: \pi = \pi_0$$

Test Statistic: $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$ under H_0

Reject H_0 if

H_a Form:

- (1) $H_a: \pi > \pi_0$
- (2) $H_a: \pi < \pi_0$
- (3) $H_a: \pi \neq \pi_0$

Reject region

- $z \geq z_\alpha$
- $z \leq -z_\alpha = z_{1-\alpha}$
- $|z| \geq z_{\alpha/2}$

R code:

`qnorm(1-alpha)`
`qnorm(alpha)`
`qnorm(1-alpha/2)`

H_a Form:

- (1) $H_a: \pi > \pi_0$
- (2) $H_a: \pi < \pi_0$
- (3) $H_a: \pi \neq \pi_0$

P-value:

- $P(Z \geq z)$
- $P(Z \leq z)$
- $2 * P(Z \geq |z|)$

R code:

`1-pnorm(z)`
`pnorm(z)`
`2 * (1-pnorm(abs(z)))`

For the Aphid Example:

$H_0: \pi \leq 0.5$ versus $H_a: \pi > 0.5$

Test Statistic:

$$z = \frac{0.53 - 0.50}{\sqrt{\frac{0.50(1-0.50)}{100}}} = \frac{0.03}{0.05} = 0.6$$

Rejection Rule:

$$\text{qnorm}(0.95) = 1.645$$

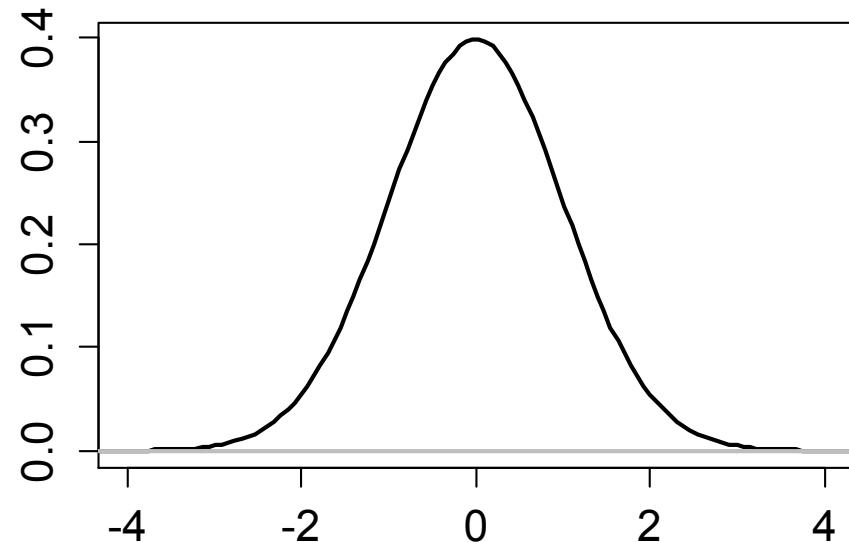
Reject H_0 if $z > z_{0.05} = 1.645$

(One-sided) p-value:

$$1 - \text{pnorm}(0.6) = 0.274$$

Decision: We fail to Reject H_0 .

We cannot conclude that the population proportion of infested fields is greater than 0.5.



Test for Single Proportion in R

- In R, use `prop.test()`

For the Aphids Example #1 ($n=100$, $y=53$):

```
> prop.test(53, 100, p=0.5, alternative =  
  "greater", correct = FALSE)  
  
 1-sample proportions test without  
 continuity correction  
  
data: 53 out of 100, null probability 0.5  
x-squared = 0.36, df = 1, p-value = 0.2743  
alternative hypothesis: true p is greater than  
 0.5  
  
95 percent confidence interval:  
0.4481999 1.0000000  
sample estimates:  
p  
0.53
```

This is the one-sided CI. We are NOT going to discuss the one-sided CI in this course. We focus on the two-sided CI.

Notes on the Large Sample Normal Approximation

1. When forming the large sample CI use $\hat{\pi}$ to compute the standard error, but when testing use π_0 to compute the estimated standard error.
2. Due to previous point, the set of π that would not be rejected by the large sample hypothesis test is slightly different from the set of π that are in the CI.
3. The X-squared statistic is equal to Z^2 where the formula for Z is given on slide 16. To get the sign of Z, we look at $\hat{\pi} - \pi_0$ (numerator of Z test statistic).
4. The formulas presented assume a large population. A finite population correction (FPC) is possible, but not covered here.

More detail about the Continuity Correction

1. Yates continuity correction is possible (default in R).
2. The idea of the continuity correction is to better approximate binomial distribution (discrete) with normal (continuous). However the effect of the continuity correction is most noticeable when the sample size is small (in which case we will use the exact binomial approach for testing and CI).
3. I am fine with the continuity correction, but I will not ask for (or show) a continuity correction when doing hand calculations. Hence, I use `correct = FALSE` here to more closely match the hand calculations.

ME Reported for Polls in the News

Polls reported in the news should give the sampling dates, number of participants and margin of error.

Note that the formula for ME depends on the estimated proportion:

$$ME = Z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

However, often a single ME will be reported even if several different questions (with different estimates) were asked.

$$95\%ME = 1.96 \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq 2 \sqrt{\frac{0.5(1-0.5)}{n}} = \frac{1}{\sqrt{n}}$$

Hence, the news will often report the ME as $1/\sqrt{n}$!

EXTRA: Denver Post Ghost Article (10/26/17)

34 percent of people say they believe in ghosts, according to a pre-Halloween poll by the Associated Press and Ipsos. That's the same proportion who believe in UFOs, exceeding the 19 percent of people who accept the existence of spells or witchcraft.

Forty-eight percent believe in extrasensory perception, or ESP. But nearly half of you knew we were about to tell you that, right?

A smaller but still substantial 23 percent say they have actually seen a ghost or believe they have been in one's presence, with the most likely candidates for such visits including single people, Catholics and those who never attend religious services.

Spells and witchcraft are more readily believed by urban dwellers, minorities and lower-earning people. Those who find credibility in ESP are more likely to be better educated and white – 51 percent of college graduates compared to 37 percent with a high school diploma or less.

The poll, conducted Oct 16-18, involved telephone interviews with **1013 adults** and had a **margin of sampling error of plus or minus 3.1 percentage points**.

$$95\% \text{ ME} = 1.96 \sqrt{\hat{\pi}(1 - \hat{\pi})/n} \leq 2\sqrt{0.5(1 - 0.5)/n} = 1/\sqrt{n}$$
$$1/\sqrt{1013} = 0.0314 = 3.1\%$$

2A. Sample Size based on CI width

Recall that the (large sample) confidence interval for π has the form

$$\hat{\pi} \pm Z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Let E represent the desired margin of error (ME) of the confidence interval is:

$$E = Z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Solving this for n we get

$$n = \frac{Z_{\alpha/2}^2 \hat{\pi}(1-\hat{\pi})}{E^2}$$

We need a **conjecture for $\hat{\pi}$** or we can substitute $\hat{\pi} = 0.5$ corresponding to a “worst case” scenario. The choice $\hat{\pi} = 0.5$ will give the largest possible sample size that may be needed, a conservative answer for the required sample size.

Example: Suppose that we want to find the sample size required to achieve a 95% ME for π to be ≤ 0.10 . We do not have a specific conjecture for π , so we use the “worst case” conjecture $\pi = 0.50$.

95%ME -> $Z_{\alpha/2} = 1.96$

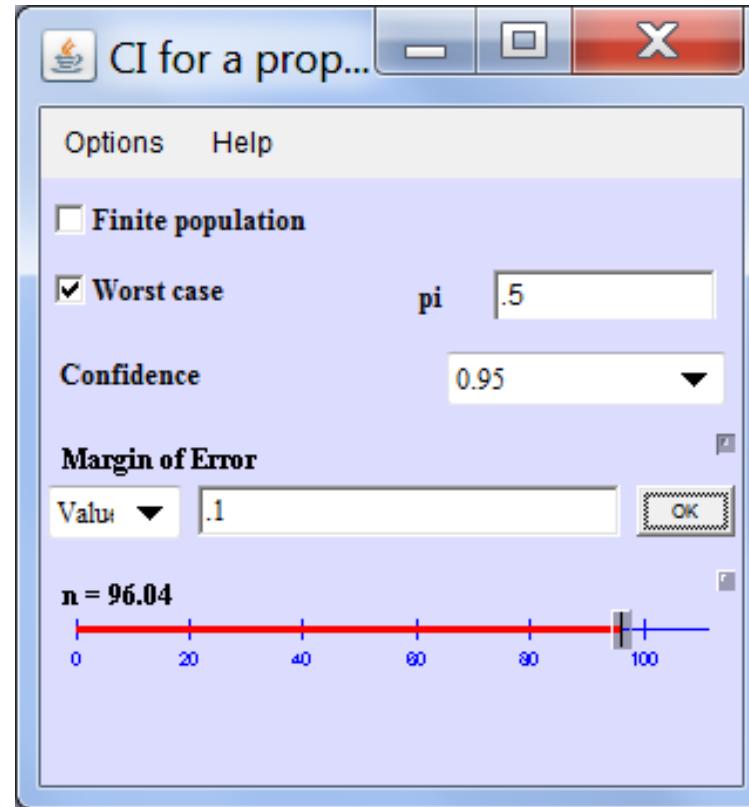
$$n = \frac{(1.96)^2(0.5)(1 - 0.5)}{(0.1)^2} = 96.04$$

Round up to n=97!

Sample Size based on CI width using Lenth

<http://homepage.stat.uiowa.edu/~rlenth/Power/>

- Choose CI for one Proportion.
- Here we calculate the required sample size to achieve a 95% ME of 0.1
- We use the “worst case scenario” conjecturing that $\pi=0.5$



2B. Sample Size from power of a hypothesis test

$H_0: \pi \leq \pi_0$ versus $H_a: \pi > \pi_0$ (one-sided)

Test Statistic: $z = \frac{\hat{\pi} - \pi_0}{\sigma_{\pi}}$ where $\sigma_{\pi} = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$

Reject H_0 if $z > z_{\alpha}$

Calculating power:

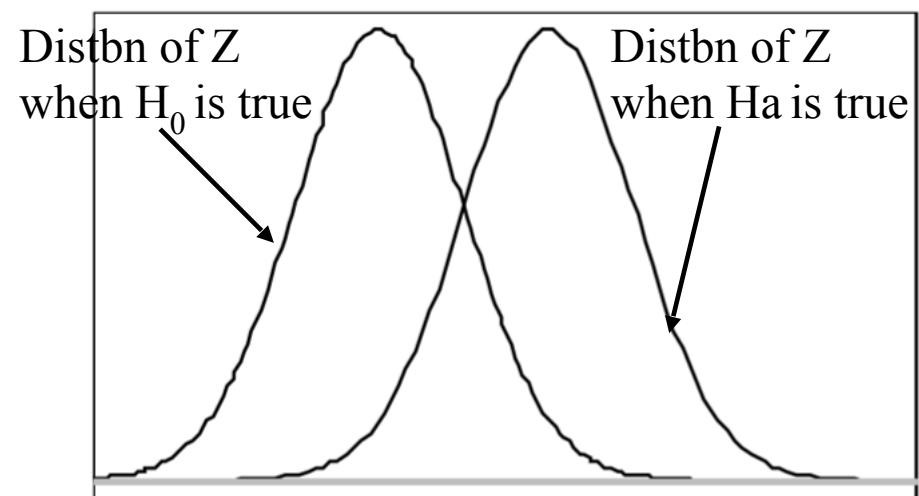
For a specific conjectured value of π , the distribution of Z is approximately normal with variance approximately 1.0 and mean:

$$\lambda = \frac{\pi - \pi_0}{\sqrt{(\pi_0(1 - \pi_0))/n}}$$

We can use this information to calculate power.

Example: Suppose we want to calculate power for testing $H_a: \pi > 0.5$ using $n = 100$ and $\alpha = 0.05$. (1) Hence we Reject H_0 if $Z > Z_\alpha = 1.645$. (2) We conjecture that $\pi=0.6$.

$$\lambda = \frac{\pi - \pi_0}{\sqrt{(\pi_0(1 - \pi_0))/n}} = \frac{0.6 - 0.5}{\sqrt{(0.5(1 - 0.5))/100}} = 2$$



Calculate power “by hand” using R:
 $1 - pnorm(1.645, \text{mean} = 2, \text{sd} = 1) = 0.6388$

$$d = \frac{\pi - \pi_0}{\sqrt{\pi_0(1 - \pi_0)}}$$

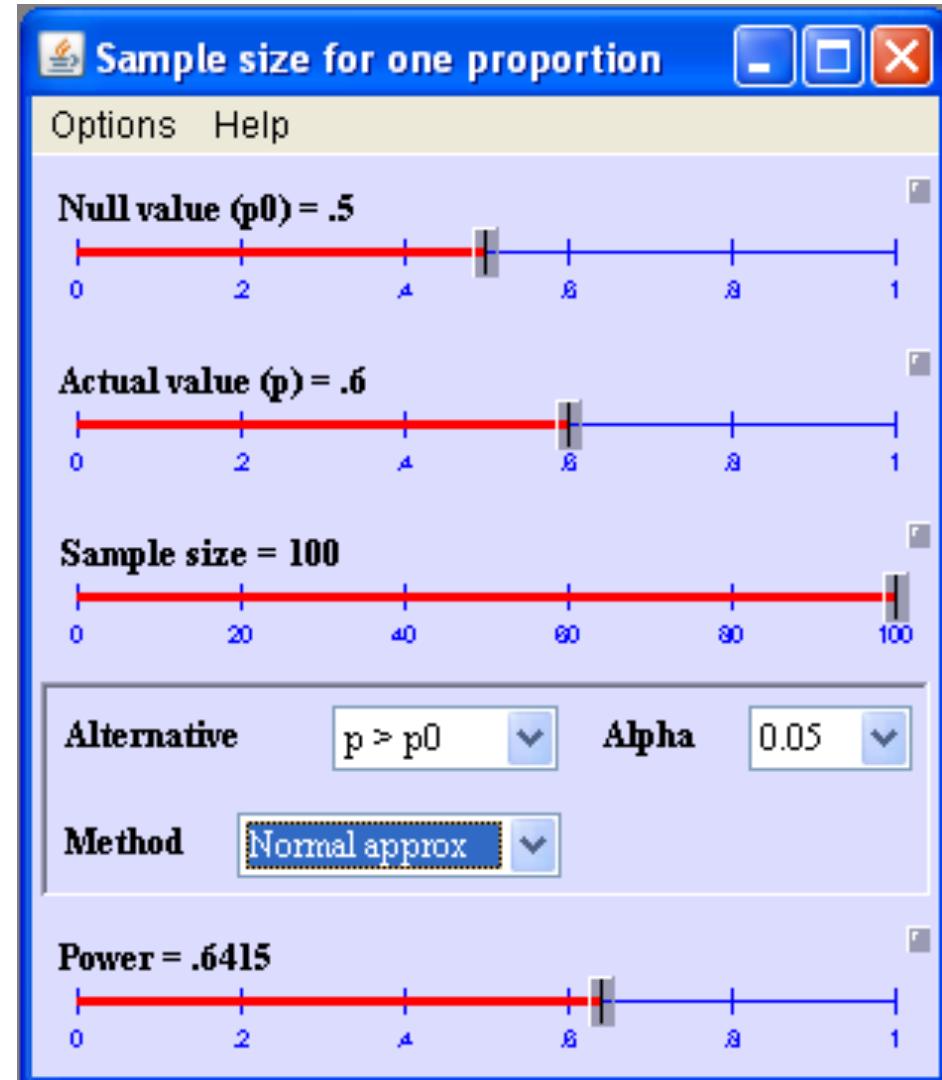
Or using `pwr.norm.test(d = 0.1/0.5, n=100, sig.level = 0.05, alternative = "greater")` in package `pwr`.

- This function can also be used to find sample size n for the desired power.
- Eg. `pwr.norm.test(d = 0.1/0.5, power=0.6, sig.level = 0.05, alternative = "greater")`

Power for Large Sample Test about π using Lenth

<http://homepage.stat.uiowa.edu/~rlenth/Power/>

- Choose Test of one Proportion.
- Here we calculate power for HA: $\pi > 0.5$ using $n=100$ based on a conjectured value of $\pi=0.6$.
- Since $n=100$ the large sample normal approximation is appropriate.



Small Sample Inference

(Using Exact Binomial)

3. Test and confidence interval

Aphids Example #2: Suppose we want to test

$$H_0: \pi \leq 0.25 \text{ versus } H_a: \pi > 0.25$$

Say that $n=8$ plots are sampled and of these $y = 4$ are found to be infested.

Test statistic is Y , which has a binomial distribution.

$$\text{Under } H_0, Y \sim \text{bin}(n=8, \pi=0.25)$$

P-value = $P(Y \geq y_{\text{obs}})$ In our example, $y_{\text{obs}}=4$.

$$\begin{aligned} \text{p-value} &= P(Y \geq 4) = 1 - P(Y \leq 3) \\ &= 1 - \text{pbinom}(3, \text{size} = 8, \text{prob} = 0.25) \\ &= 0.114 \end{aligned}$$

Since p-value $> \alpha = 0.05$, we Fail to Reject H_0 .

Exact binomial test can be run using `binom.test()`.

One and two-sided alternatives can be done.

An exact confidence interval for π can also be obtained by using `binom.test()`.

The exact confidence interval is constructed as the set of values of π that are not rejected in the exact hypothesis test on the previous page.

This cannot be done easily by hand, because you have to test all possible π values. Use R.

If $n = 8$ and the observed response $y = 4$ then, a 95% CI is $(0.157, 0.843)$.

Test & CI for Single Proportion in R

For the Aphids Example #2 ($n=8$, $y=4$):

- One-sided Test

```
> binom.test(4, 8, p=0.25, alternative="greater")
      Exact binomial test
data: 4 and 8
number of successes = 4, number of trials = 8,
p-value = 0.1138
alternative hypothesis: true probability of
success is greater than 0.25
95 percent confidence interval:
0.1929029 1.0000000
sample estimates:
probability of success
                           0.5
```

For the Aphids Example #2 ($n=8$, $y=4$):

- Two-sided confidence interval

```
> binom.test(4, 8, p=0.25)

Exact binomial test

data: 4 and 8
number of successes = 4, number of trials = 8,
p-value = 0.1138
alternative hypothesis: true probability of
success is not equal to 0.25
95 percent confidence interval:
0.1570128 0.8429872
sample estimates:
probability of success
0.5
```

4. Power for binomial test about π

Example: Compute power for an exact binomial test with $n=20$:

$H_0: \pi \leq 0.25$ versus $H_a: \pi > 0.25$ (one-sided) ($\alpha = 0.05$)

Step 1: Define the Reject Region.

Reject H_0 when $Y \geq Y_{\text{crit}}$

Find Y_{crit} such that under H_0 ($\pi_0=0.25$),
 $P(Y \geq Y_{\text{crit}}) \leq \alpha = 0.05$.

Using `pbinom()` with $n=20$ and $\pi=0.25$:

$$P(Y \geq 9) = P(Y > 8) = 0.0409$$

Therefore, we Reject H_0 if $Y \geq 9$.

The test is slightly conservative,
meaning $\alpha < 0.05$.

k	$P(Y \leq k)$	$P(Y > k)$
0	0.0032	0.9968
1	0.0243	0.9757
2	0.0913	0.9087
3	0.2252	0.7748
4	0.4148	0.5852
5	0.6172	0.3828
6	0.7858	0.2142
7	0.8982	0.1018
8	0.9591	0.0409
9	0.9861	0.0139
10	0.9961	0.0039
11	0.9991	0.0009
....		

Step 2: Use a specific conjectured value to calculate power.

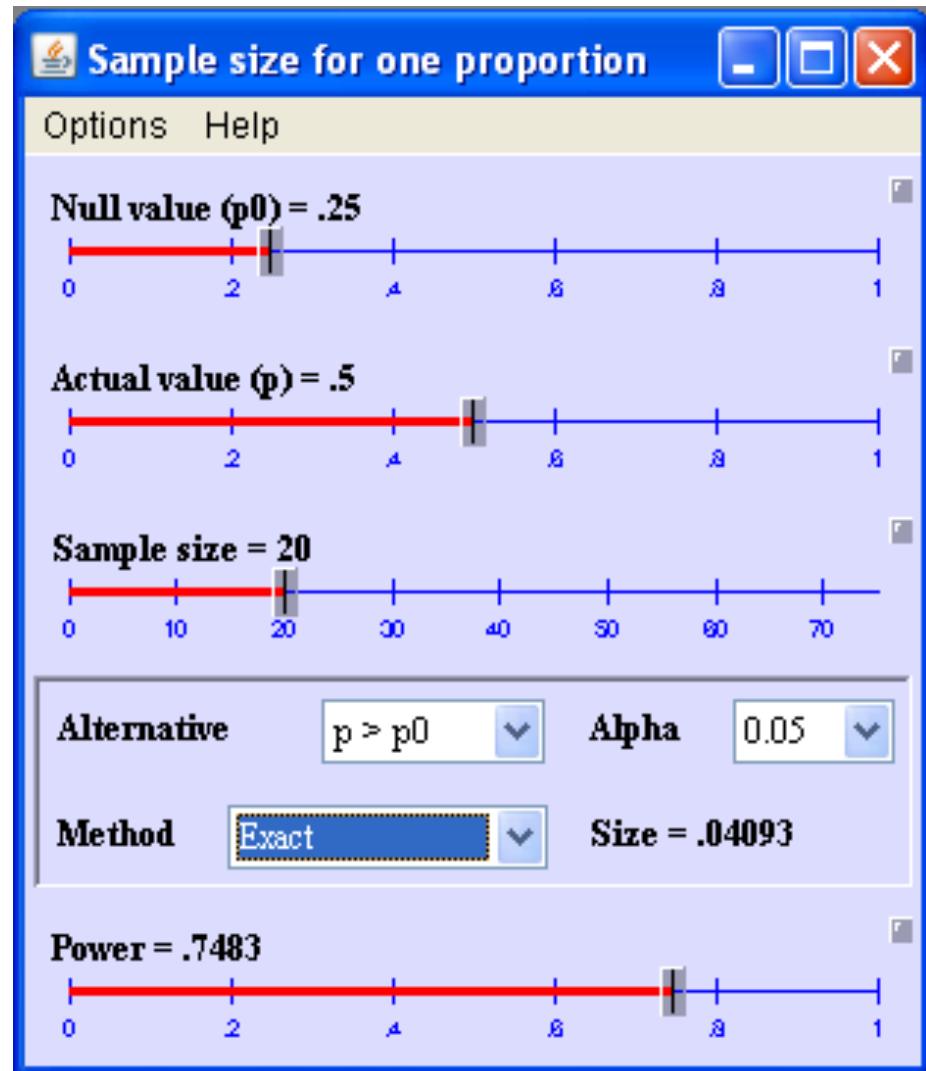
For a conjectured value of $\pi=0.5$, we calculate the power as:

$$\begin{aligned}\text{Power} &= P(Y \geq Y_{\text{crit}}) = P(Y \geq 9) \\ &= 1 - P(Y \leq 8) \\ &= 1 - \text{pbinom}(8, 20, 0.5) \\ &= 0.7483\end{aligned}$$

Power for Small Sample Test about π using Lenth

<http://homepage.stat.uiowa.edu/~rlenth/Power/>

- Choose Test of one Proportion
- Here we calculate power for $H_a: \pi > 0.25$ using $n=20$ based on a conjectured value of $\pi=0.5$.
- Since $n=20$ the small sample exact test is appropriate.



Nonparametric method

5. Return to the Sign Test for the Median

In chapter 5, we used the sign test to make inference about a single median. The sign test is based on the binomial test of a single proportion.

Suppose we are testing $H_0: M \leq M_0$ vs $H_a: M > M_0$.

If the null hypothesis was true (M_0 was the population median) then we would expect 50% of the differences $(y_i - M_0)$ to be positive.

To calculate the test statistic, we look at the sign of each difference $(y_i - M_0)$. The test statistic (s) is the number of positive differences.

To calculate the p-value we compare the test statistic (s) to the binomial distribution with $n = \# \text{ observations}$, $\pi = 0.5$.

Example: Suppose we are testing $H_0: M \leq 2$ vs $H_a: M > 2$. Based on a sample of size $n = 8$, we find that $s = 5$ observations are greater than 2. Then the p-value is $P(Y \geq 5) = 1 - P(Y \leq 4)$.

```
> InData <- c(0, 1, 1, 3, 5, 7, 9, 10)
> 1 - pbinom(4, size = 8, prob = 0.5)
[1] 0.3632813
> library(BSDA)
> SIGN.test(InData, md=2, alternative =
"greater")
```

```
One-sample Sign-Test
data: InData
s = 5, p-value = 0.3633
alternative hypothesis: true median is greater
than 2
```

Ch10 Categorical Data

In CH5-9, we focused on inference for means and standard deviations. **Mean and standard deviation** are parameters for center and spread for a **numerical variables** (ex: hormone level, lead consumption).

In CH10, we focus on **categorical variables** (ex: infested or not infested, cold or no cold). In this case, the parameter of interest is **proportion**.

CH5-9: Normal, t, χ^2 , F distributions

- All continuous distributions

CH10: **Binomial and Poisson distributions**

- Both discrete distributions
- BUT sometimes these discrete distributions are approximated by continuous distributions!

The Ch 10 notes includes:

Binomial Distribution

- Binomial Distribution and Its Approximation (notes10.1)
- Inference for a Single Proportion π (notes10.2)
- **Comparing ≥ 2 Proportions (notes10.3)**
- Contingency Tables: Tests for Independence and Homogeneity (notes10.4)
- Odds Ratios (notes10.5)

Poisson Distribution (notes10.6)

Ch 10.3: Comparing ≥ 2 Proportions

Comparing **Two** Proportions

Independent samples (**Large Samples**)

1. Inference: CI and Test
2. Power and Sample Size of Test

Paired samples (**Nonparametric** method)

3. McNemar's test

Comparing **More Two** Proportions

4. Chisquare Goodness of Fit (GOF) Test

More

5. Extra: Mendel

Skiers Example: Researchers were interested in examining the effect of Vitamin C for prevention of colds.

279 French skiers randomly divided into two groups.

139 (n_1) are given Vitamin C, of these 17 (y_1) developed colds.

140 (n_2) are given Placebo, of these 31 (y_2) developed colds

	Cold	No Cold	Total
Vitamin C	17	122	139
Placebo	31	109	140

$$\hat{\pi}_{VitC} = 17/139 = 0.122$$

$$\hat{\pi}_P = 31/140 = 0.221$$

Of those taking Vitamin C, we estimate that 12% will get colds.

Of those taking Placebo, we estimate that 21% will get colds.

Note: We will analyze this data 3 ways! Z-test for comparing proportions, chi-square test for contingency tables, odds ratios.

1. Large Sample Inference for two independent proportions

We are interested in the inference for the difference between two populations, $\pi_1 - \pi_2$.

We will introduce those methods developed based on **Large sample**, meaning **Normal approximation** considered.

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}$$

$$SE(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{(\hat{\pi}_1)(1-\hat{\pi}_1)}{n_1} + \frac{(\hat{\pi}_2)(1-\hat{\pi}_2)}{n_2}}$$

Confidence Interval for $\pi_1 - \pi_2$

An approximate $(1-\alpha)100\%$ confidence interval for $\pi_1 - \pi_2$ is:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2} \sqrt{\frac{(\hat{\pi}_1)(1-\hat{\pi}_1)}{n_1} + \frac{(\hat{\pi}_2)(1-\hat{\pi}_2)}{n_2}}$$

where the table value $z_{\alpha/2}$ is determined from the Normal distribution using `qnorm(1-alpha/2)`.

Assumptions: Independent random samples, large sample sizes.

Hypothesis Test for $\pi_1 - \pi_2$

Assumptions: Independent random samples with large sample sizes.

$$H_0: \pi_1 - \pi_2 = 0$$

Test Statistic:
$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \text{ under } H_0$$

where

$$\hat{\pi} = \frac{y_1 + y_2}{n_1 + n_2}$$

H_a Form:

(1) $H_a: \pi_1 - \pi_2 > 0$

P-value:

$$P(Z \geq z)$$

(2) $H_a: \pi_1 - \pi_2 < 0$

$$P(Z \leq z)$$

(3) $H_a: \pi_1 - \pi_2 \neq 0$

$$2 * P(Z \geq |z|)$$

R code:

`1-pnorm(z)`

`pnorm(z)`

`2 * (1-pnorm(abs(z)))`

Rejection rule approach also possible but not shown.

For the Skiers example:

$$H_0: \pi_1 - \pi_2 = 0 \text{ vs } H_a: \pi_1 - \pi_2 \neq 0$$

Recall: $\hat{\pi}_1 = 0.122$, $\hat{\pi}_2 = 0.221$

$$\hat{\pi} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{17 + 31}{139 + 140} = 0.172$$

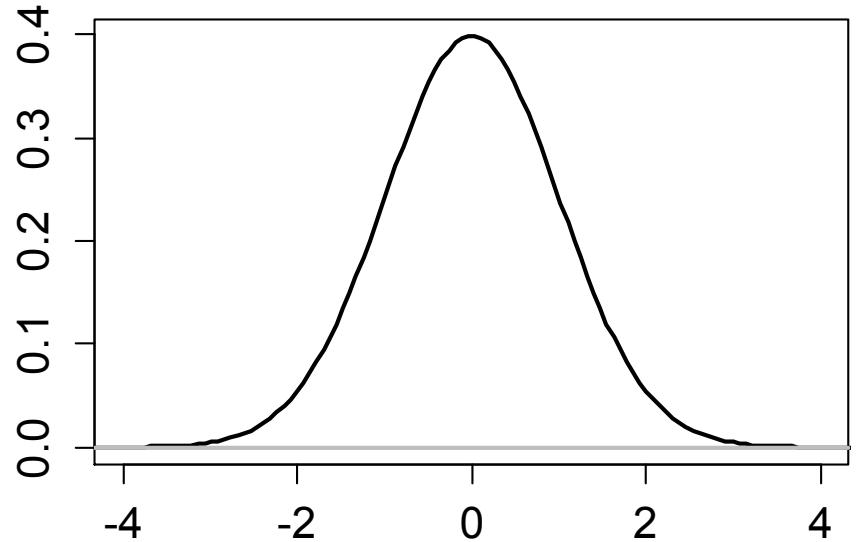
$$SE(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{0.172(1 - 0.172) \left(\frac{1}{139} + \frac{1}{140} \right)} = 0.0452$$

Test Statistic:

$$Z = (0.122 - 0.221)/0.0452 = -2.19$$

(Two-sided) p-value:

$$2 * (1 - \text{pnorm}(2.19)) = 0.028$$



Conclusion: Reject H0. We have evidence that Vitamin C reduces incidence of colds (in the population from which the sample was taken).

Large Sample Test & CI for $\pi_1 - \pi_2$ in R

- In R, use prop.test()

For the Skiers Example:

```
> prop.test(c(17, 31), c(139, 140), correct =  
FALSE)  
2-sample test for equality of proportions  
without continuity correction  
data: c(17, 31) out of c(139, 140)  
X-squared = 4.8114, df = 1, p-value = 0.02827  
alternative hypothesis: two.sided  
95 percent confidence interval:  
-0.18685917 -0.01139366  
sample estimates:  
prop 1     prop 2  
0.1223022 0.2214286
```

Notes on the Large Sample test and CI for $\pi_1 - \pi_2$

1. When forming the CI, we used a standard error expression that involved separate estimates of π_1 and π_2 , but, when testing used, an estimate involving a common estimate of π_1 and π_2 , which we called $\hat{\pi}$.
2. Due to previous point, rejecting $H_0: \pi_1 = \pi_2$ is not exactly equivalent to a CI for $\pi_1 - \pi_2$ **not** containing zero.
3. **The (two-sided) Z-test is equivalent to the commonly used chi-squared (χ^2) test for 2x2 tables which we will discuss later in these notes.**
The χ^2 statistic is equal to Z^2 where the formula for Z is given on slide 7.
4. Above test and CI require large enough samples, so that the normal approximation to the Binomial distribution is adequate.
5. A continuity correction (called the Yates correction) is possible. This is used by default in R. I am fine with the continuity correction.
6. **When sample sizes are small, tests of $H_0: \pi_1 = \pi_2$ can be done using “Fisher’s Exact Test” to be discussed later in the notes.**

2. Power and Sample Size of Hypothesis Test

Example: We want to design an experiment to test

$$H_0: \pi_1 \geq \pi_2 \text{ vs } H_a: \pi_1 < \pi_2.$$

Use conjectured proportions: $\pi_1 = 0.15$ $\pi_2 = 0.35$.

1. Calculate power when $n_1 = n_2 = 50$
2. Calculate sample size when power=0.75

Comment: When making inference about proportions, larger sample sizes are often required (as compared to making inference about means).

1. In R, use power.prop.test() to calculate the power

```
> power.prop.test(n = 50, p1 = 0.15, p2 = 0.35, sig.level =  
  0.05, alternative = "one.sided")  
Two-sample comparison of proportions power calc  
  n = 50  
  p1 = 0.15  
  p2 = 0.35  
  sig.level = 0.05  
  power = 0.7526999  
  alternative = one.sided
```

NOTE: **n is number in *each* group**

2. In R, use power.prop.test() to calculate the power

```
> power.prop.test(power = 0.75, p1 = 0.15, p2 = 0.35, sig.level =  
  0.05, alternative = "one.sided")  
Two-sample comparison of proportions power calc  
  n = 49.64163  
  p1 = 0.15  
  p2 = 0.35  
  sig.level = 0.05  
  power = 0.75  
  alternative = one.sided
```

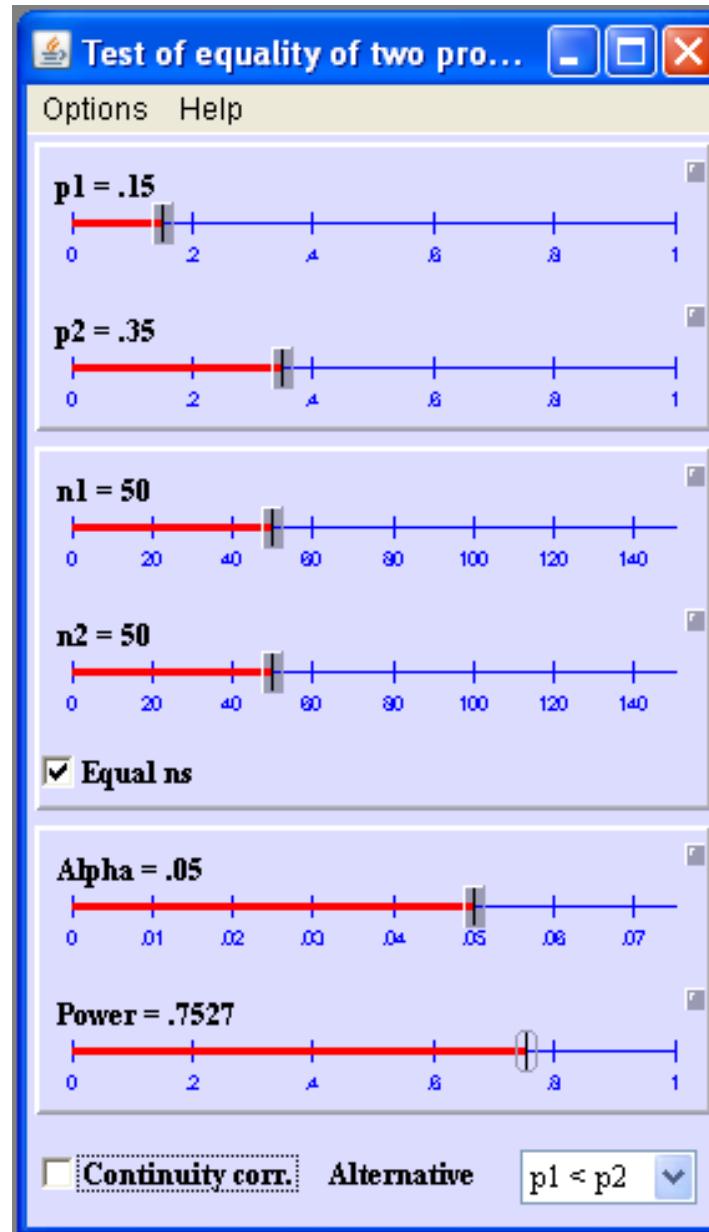
NOTE: **n is number in *each* group**

Power calculation using Lenth (Optional)

[http://
homepage.stat.uiowa.edu/
~rlenth/Power/](http://homepage.stat.uiowa.edu/~rlenth/Power/)

Note: This calculator will produce an error if p_1 and/or p_2 are too close to zero or one.

This is because the large sample normal approximation “fails” in this situation.



3. Inference for the paired proportions

Opinions Example (from Agresti):

- This data is from the General Social Survey (GSS).
- Subjects ($n = 1144$) were asked whether, to help the environment, they would be willing to (1) pay higher taxes (TaxInc) and (2) accept a cut in living standards (LSDec).
- Because each subject was asked both questions, we have paired data for which the responses not independent.
- Note that there are 905 (79%) concordant pairs and 239 discordant pairs.
 $678 + 227$ $107 + 132$

	LSDec No	LSDec Yes	Total
TaxInc No	678	107	785
TaxInc Yes	132	227	359
Total	810	334	1144

$$\hat{\pi}_{\text{LSDec}} = 334/1144 = 0.292 \quad \hat{\pi}_{\text{TaxInc}} = 359/1144 = 0.314$$

McNemar's test

- McNemar's test is a nonparametric test used on paired nominal data.
- It's used when you are interested in finding a change in proportion for the paired data.
- Let π_1 and π_2 be the proportions of pairs responding Yes for response 1 (TaxInc) and response 2 (LSDec), respectively.
- The null hypothesis for McNemar's test can be phrased as:
 $H_0: \pi_1 = \pi_2$
or
 $H_0: \text{Among discordant pairs, } \Pr(\text{yes/no}) = \Pr(\text{no/yes}) = 0.5$
- There is an exact version of McNemar's test based on the binomial distribution.
- We will use `mcnemar.test()` from R which uses a large sample normal approximation.

McNemar's test of paired proportions in R

- In R, use `mcnemar.test()`

```
> Opinions <- matrix(c(678, 107, 132, 227),  
+ byrow = TRUE, nrow = 2)  
> Opinions  
     [,1] [,2]  
[1,]   678   107  
[2,]   132   227  
> mcnemar.test(Opinion)  
McNemar's Chi-squared test with continuity  
correction  
data:  Opinions  
McNemar's chi-squared = 2.41, df = 1,  
p-value = 0.1206
```

So for this example, we fail to Reject H_0 . We cannot conclude that there is a difference in proportions (for those willing to accept LSDec vs TaxInc).

4. Comparing More Than Two Proportions

Maize Example (Snedecor and Cochran): Two types of maize were crossed. A sample of $n=1301$ plants is taken and 4 types of maize are observed. Hence $k = 4$ categories.

	Green (1)	Gold (2)	Green Striped (3)	Green/Gold Striped (4)	Total
Observed	$n_1=773$	$n_2=231$	$n_3=238$	$n_4=59$	1301
H_0	$\pi_1=9/16$	$\pi_2=3/16$	$\pi_3=3/16$	$\pi_4=1/16$	1

Question: Are these data consistent with the Mendelian laws of inheritance? These laws would imply that in the long run the four types would occur with the following proportions:
 $\pi_1=9/16$, $\pi_2=3/16$, $\pi_3=3/16$, $\pi_4=1/16$.

- (Pearson's) Chisquare Goodness of Fit (GOF) Test

(Pearson's) GOF Test

Assumptions: Independent observations, large sample size.

Rule of thumb for sample size: No E_i can be less than 1, and no more than 20% of E_i 's can be less than 5.

H_0 : $\pi_i = \pi_{i0}$ for categories $i=1,\dots,k$. (π_{i0} are specified probabilities or proportions.)

H_a : At least one of the cell probabilities differs from the hypothesized value.

Test statistic: $\chi^2_0 = \sum_{i=1}^k \frac{(n_i - E_i)^2}{E_i} \sim \chi^2_{df}$ under H_0

where $df = k - 1$ with $k = \#$ categories, and $E_i = n\pi_{i0}$ (the expected count under H_0)

P-value: $P(\chi^2 \geq \chi^2_0)$ **R code:** `1-pchisq(chisq, df = k - 1)`

Notes:

- This test has a NON-directional alternative.

Maize Example

$$H_0 : \pi_1 = 9/16, \pi_2 = 3/16, \pi_3 = 3/16, \pi_4 = 1/16$$

$$H_a : \text{not } H_0$$

	Green (1)	Gold (2)	Green Striped (3)	Green/Gold Striped (4)
Observed	n1=773	n2=231	n3=238	n4=59
	1301(9/16)	1301(3/16)	1301(3/16)	1301(1/16)
Expected	E1=731.8	E2=243.9	E3=243.9	E4=81.3

$$\begin{aligned} \chi^2 &= \frac{(773 - 731.8)^2}{731.8} + \frac{(231 - 243.9)^2}{243.9} + \frac{(238 - 243.9)^2}{243.9} \\ &\quad + \frac{(59 - 81.3)^2}{81.3} = 9.27 \end{aligned}$$

P-value: `1-pchisq(9.27, df = 3)` = 0.0259

Conclusion: Reject H0. We find evidence against Mendel's law.

Chisquare GOF Test in R

- In R, use `chisq.test()`

```
> chisq.test(c(773, 231, 238, 59),  
             p = c(9/16, 3/16, 3/16, 1/16),  
             correct = FALSE)
```

Chi-squared test for given probabilities

```
data: c(773, 231, 238, 59)  
X-squared = 9.2714, df = 3, p-value = 0.02589
```

Notes on the Chisquare GOF Test

1. The more common test is the chisquare test for contingency tables. This will be covered in the next group of notes.
2. The null hypothesized probabilities should be motivated by the research question! These null hypothesized probabilities must sum to one.
3. Rule of thumb for sample size: No E_i can be less than 1, and no more than 20% of E_i 's can be less than 5. Note that these rules are based on expected cell counts.
4. In some cases, it may be reasonable to combine small categories.
5. The expected cell counts can be calculated using R.
6. Another use of the expected cell counts is to identify which category or categories are “causing” the rejection. We do this by comparing observed vs expected counts. For the Maize example, the Green and Green/Gold Striped categories show noticeable differences (comparing observed vs expected).

5. EXTRA: Gregor Mendel

- The famous statistician Fisher did some re-analysis of Mendel’s famous data. Fisher combined several experiments into one test.
- H_0 : Mendel’s law holds.
 H_a : Data were “edited” so that they are closer to the theoretically expected E_i ’s than they would be if “editing” was not done.
- In this (very rare!) scenario, chi-square GOF p-value was calculated as $P(\chi^2 \geq \chi_0^2)$.
- $\chi_0^2 = 42$, $df = 84$, $p\text{-value} = 0.00004$
- Conclusion: Reject H_0 . The observed data was closer to expected than random data would be. This implies that the data may have been edited.
- Fisher “I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made.”

EXTRA: Gregor Mendel (Wikipedia 10/13/17)



Mendel (1822-1884) gained posthumous recognition as the **founder of the modern science of genetics.**

He became a friar in part because it enabled him to obtain an education without having to pay for it himself.

Mendel worked as a substitute high school teacher. In 1850, he failed the oral part, the last of three parts, of his exams to become a certified high school teacher. In 1851, he was sent to the University of Vienna to so that he could get more formal education. Mendel returned to his abbey in 1853 as a teacher, principally of physics. In 1856, he took the exam to become a certified teacher and **again failed the oral part.** In 1867, he became the abbot of his monastery.

He studied astronomy and meteorology, founding the 'Austrian Meteorological Society' in 1865. The majority of his published works was related to meteorology.

Mendel is best known for his studies involving pea plants. Between 1856 and 1863 Mendel cultivated and tested some 28,000 plants, the majority of which were pea plants. His experiments led him to make two generalizations, **the Law of Segregation and the Law of Independent Assortment, which later came to be known as Mendel's Laws of Inheritance.**

Mendel presented his paper, "Versuche über Pflanzenhybriden" ("Experiments on Plant Hybridization") in 1865. It generated a few favorable reports in local newspapers, but **was ignored by the scientific community**.

When Mendel's paper was published in 1866, it was seen as essentially **about hybridization rather than inheritance**, had little impact, and was **only cited about three times over the next thirty-five years**. Notably, Charles Darwin was unaware of Mendel's paper, and it is envisaged that if he had, genetics as we know it now might have taken hold much earlier.

After Mendel's death three researchers, each from a different country, published their rediscovery of Mendel's work within a two-month span in the Spring of 1900.

In 1936, R.A. Fisher, a prominent statistician and population geneticist, reconstructed Mendel's experiments, analyzed results from the F2 (second filial) generation and found the ratio of dominant to recessive phenotypes (e.g. green versus yellow peas; round versus wrinkled peas) to be **implausibly and consistently too close to the expected ratio of 3 to 1**. Fisher asserted that "the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations".

Ch10 Categorical Data

In CH5-9, we focused on inference for means and standard deviations. **Mean and standard deviation** are parameters for center and spread for a **numerical variables** (ex: hormone level, lead consumption).

In CH10, we focus on **categorical variables** (ex: infested or not infested, cold or no cold). In this case, the parameter of interest is **proportion**.

CH5-9: Normal, t, χ^2 , F distributions

- All continuous distributions

CH10: **Binomial and Poisson distributions**

- Both discrete distributions
- BUT sometimes these discrete distributions are approximated by continuous distributions!

The Ch 10 notes includes:

Binomial Distribution

- Binomial Distribution and Its Approximation (notes10.1)
- Inference for a Single Proportion π (notes10.2)
- Comparing ≥ 2 Proportions (notes10.3)
- **Contingency Tables: Tests for Independence and Homogeneity (notes10.4)**
- Odds Ratios (notes10.5)

Poisson Distribution (notes10.6)

Ch10.4: Contingency Tables: Tests for Independence and Homogeneity

Contingency Tables

1. Chisquare test for independence
2. Homogeneity
 - A. Chisquare test
 - B. Fisher's Exact Test (**small sample size**)
3. Extra: Lady Tasting Tea

1. Chi-square Test of Independence for Contingency Tables

Skiers Example (2x2 Contingency Table):

	Cold	No Cold	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

H_0 : No association between treatment and cold status (Independence)

- Chi-square test is used.

Chi-square Test

Assumptions: Independent observations, large sample size.

Rule of thumb for sample size: No E_i can be less than 1, and no more than 20% of E_i 's can be less than 5.

H_0 : The row and column variables are independent (no association).

H_a : The row and column variables are dependent (associated).

Test statistic: $\chi^2 = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{df}$ under H_0

where $df = (\#rows - 1)(\#cols - 1)$ and $E_{ij} = \frac{(\text{i}^{\text{th}} \text{ row total})(\text{j}^{\text{th}} \text{ column total})}{\text{grand total}}$

P-value: $P(\chi^2 \geq \chi^2_0)$

R code: `1-pchisq(χ²₀, df)`

Notes:

- This test has a NON-directional alternative.

Skiers Example (2x2 Contingency Table):

	Cold	No Cold	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

Expected Counts:

	Cold	No Cold
Vitamin C	$139 * 48 / 279 = 23.91$	$139 * 231 / 279 = 115.09$
Placebo	$140 * 48 / 279 = 24.09$	$140 * 231 / 279 = 115.91$

2x2 Contingency Table

	Cold	No Cold	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

Table of Expected Counts

	Cold	No Cold
Vitamin C	23.91	115.09
Placebo	24.09	115.91

H_0 : The treatment and cold status are independent (no association).

H_a : The treatment and cold status are dependent (associated).

Test Statistic: $\chi^2_0 = 4.811$ (calculation not shown)

$$df = (2-1)*(2-1) = 1$$

P-value: $1 - \text{pchisq}(4.811, df = 1) = 0.0282$

Conclusion: Reject H_0 .

We have evidence of an association between treatment and cold status.

Chi Square Test in R

- In R, use `chisq.test()`

For the Skiers Example:

```
> Skiers<-  
  matrix(c(17,122,31,109),byrow=TRUE,nrow=2)  
> Skiers  
      [,1] [,2]  
[1,]    17   122  
[2,]    31   109  
> chisq.test(Skiers,correct=FALSE)  
    Pearson's Chi-squared test  
data: Skiers  
X-squared = 4.8114, df = 1, p-value = 0.02827
```

Rat Tumors Example (3x2Table): Three groups of 100 rats were given different doses of a drug scheduled for testing in humans.

	Yes Tumors	No Tumors	Total
Control	10	90	100
Low Dose	14	86	100
High Dose	19	81	100
Total	43	257	300

H_0 : Tumors and dose are independent (no association).

H_a : Tumors depends on dose (associated).

Test statistic: $\chi^2 = 3.31$, $df = (3-1)(2-1) = 2$

p-value = 0.191

Conclusion: Fail to reject H_0 . We do not find evidence that probability of tumors depends on dose.

Note: Logistic regression is a reasonable alternative analysis that would allow for (1) dose to be treated as continuous or (2) pairwise comparisons between doses.

Opinions Example (2x2 Table):

1397 people were surveyed and asked two questions:

1. Do you favor hand gun registration (GR)?
2. Do you favor the death penalty (DP)?

	DP Yes	DP No	Total
GR Yes	784	236	1020
GR No	311	66	377
Total	1095	302	1397

H_0 : No association between opinion about GR and DP.

H_A : Some association between opinion about GR and DP.

Test statistic: $\chi^2 = 5.15$, $df = (2-1)(2-1) = 1$

p-value = 0.023

Conclusion: Reject H_0 . We have evidence of an association between an association between opinion about GR and DP.

Note: This is an observational study. The skiers example was an experiment. But the analysis is identical.

2A. Chisquare Test of Homogeneity for Contingency Tables

- An implicit assumption of our discussion surrounding the Chisquare test of independence is that the data result from a single random sample from the whole population.
- Often, separate random samples are taken from the subpopulations defined by the row (or column) variable.
- For example, 68 patients having the skin disease resulted from separate samples (of respective sizes 47 and 21) from the two age categories rather than from a single random sample of 68 patients.

Age Category	Severity		Total
	Moderate	Servere	
I	15	32	47
II	1	20	21
Total	16	52	68

- The research hypothesis is that there is a difference in the distribution of subpopulation units (patients in 2 age categories) into 2 levels of Severity.
- $H_0 : \pi_{AgeI,Moderate} = \pi_{AgeII,Moderate}$ (equality of proportions)
- The test is called a test of homogeneity of distributions.
- **The mechanics of the test of homogeneity and the test of independence are identical.** However, note that the sampling scheme and conclusions are different.

Skiers Example (2x2 Contingency Table):

	Cold	No Cold	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

$$\hat{\pi}_{\text{VitC}} = 17/139$$

$$\hat{\pi}_P = 31/140$$

$H_0: \pi_{\text{VitC}} = \pi_P$ (Equality of proportions) vs $H_a: \pi_{\text{VitC}} \neq \pi_P$

Test Statistic: $\chi^2_0 = 4.811$ (calculation not shown)

$$df = (2-1)*(2-1) = 1$$

P-value: $1 - \text{pchisq}(4.811, df = 1) = 0.0282$

Conclusion: Reject H_0 .

We have evidence that Vitamin C reduces incidence of colds (in the population from which the sample was taken).

Rat Tumors Example (3x2Table): Three groups of 100 rats were given different doses of a drug scheduled for testing in humans.

	Yes Tumors	No Tumors	Total
Control	10	90	100
Low Dose	14	86	100
High Dose	19	81	100
Total	43	257	300

$$H_0: \pi_{Ctrl} = \pi_{Low} = \pi_{High}$$

H_a : Not all the proportions are the same

Test statistic: $\chi^2 = 3.31$, $df = (3-1)(2-1) = 2$

p-value = 0.191

Conclusion: Fail to reject H_0 . We do not find evidence that these probabilities of tumors are different.

Note: Logistic regression is a reasonable alternative analysis that would allow for (1) dose to be treated as continuous or (2) pairwise comparisons between doses.

$$\hat{\pi}_{Ctrl} = 10/100 = 0.10$$

$$\hat{\pi}_{Low} = 14/100 = 0.14$$

$$\hat{\pi}_{High} = 19/100 = 0.19$$

Notes about the Chi-Squared Test for Contingency Tables

1. The (two-sided) two-sample Z-test comparing two proportions is equivalent to the chi-squared (χ^2) test for 2x2 tables. Specifically $Z^2 = \chi^2$ with identical p-values.
2. Reordering rows or columns (or transposing rows and columns) will give the same result.
3. The chi-squared test can be used for tables with any number of rows and columns. However, the expected values for each cell must be reasonably large. Sometimes it may be reasonable to combine categories to achieve this.
4. Rule of thumb for sample size: No E_i can be less than 1, and no more than 20% of E_i 's can be less than 5. Note that these rules are based on expected cell counts.
5. When sample size is small, use Fisher's Exact Test for homogeneity for contingency table.

2B. Fisher's Exact Test (FET) of Homogeneity for Contingency Tables

- Fisher's Exact Test (FET) is an alternative to the chi-square test for homogeneity for contingency tables that can be used with **small sample sizes**.
- This test was designed for sampling in which both row and column marginal totals are fixed (set by the experimenter). But in practice, it is commonly used even if this condition is not satisfied.
- It is called an “exact test” or a “randomization test” because for fixed row and column marginal totals, we can list out all possible combinations of cell counts.
- For this test it is not traditional to report a test statistic. The “test statistic” is just a particular cell count!
- Because it is an exact test, it can take awhile to run for larger tables.
- For 2x2 tables, an equivalent approach is based on the hypergeometric distribution. (Not covered here.)

Birds Example (2x2 Table):

Blue-winged and Golden-winged warblers of Southeastern Michigan were tested using tape recordings. If they responded to recorded songs from only their own species they were termed “discriminators”. If they responded to songs from both species, they were “non-discriminators”. We are interested in whether the proportion of discriminators differ by species.

	Disc Yes	Disc No	Total
Blue	4	6	10
Gold	3	9	12
Total	7	15	22

$$\hat{\pi}_B = 4/10 = 0.40$$

$$\hat{\pi}_G = 3/12 = 0.25$$

$$H_0: \pi_B = \pi_G \text{ vs } H_a: \pi_B \neq \pi_G$$

$$\text{P-value (FET)} = 0.6517$$

Conclusion: Fail to Reject H_0 . We do not have evidence of a difference in proportion that discriminate comparing the two species.

FET for Contingency Tables in R

- In R, use `fisher.test()`
- For the Birds Example:

```
> Birds<-matrix(c(4, 6, 3, 9), nrow=2, byrow=TRUE)
> Birds
      [,1] [,2]
[1,]    4    6
[2,]    3    9
> fisher.test(Birds)

Fisher's Exact Test for Count Data

p-value = 0.6517


alternative hypothesis: true odds ratio is not equal
to 1
95 percent confidence interval:
0.2308102 18.5028997
sample estimates:
odds ratio
1.936549
```

3. EXTRA: Lady Tasting Tea (Wikipedia 10/13/17)

In the design of experiments in statistics, the **lady tasting tea** is a randomized experiment devised by Ronald Fisher and reported in his book *The Design of Experiments* (1935). The experiment is the original exposition of Fisher's notion of a null hypothesis, which is "never proved or established, but is possibly disproved, in the course of experimentation".

Fisher's description is less than 10 pages in length and is notable for its simplicity and completeness regarding terminology, calculations and design of the experiment. The example is loosely based on an event in Fisher's life. **The test used was Fisher's exact test.**

The lady in question (Dr. Muriel Bristol) claimed to be able to tell whether the tea or the milk was added first to a cup. Fisher proposed to give her eight cups, four of each variety, in random order. One could then ask what the probability was for her getting the specific number of cups she identified correct, but just by chance.

The null hypothesis was that the lady had no ability to distinguish the teas.

David Salsburg reports that a colleague of Fisher, H. Fairfield Smith, revealed that in the test, **the woman got all eight cups correct**. The chance of someone who just guesses getting all correct, assuming she guesses that four had the tea put in first and four the milk, would be only 1 in 70 (the combinations of 8 taken 4 at a time).

Ch10 Categorical Data

In CH5-9, we focused on inference for means and standard deviations. **Mean and standard deviation** are parameters for center and spread for a **numerical variables** (ex: hormone level, lead consumption).

In CH10, we focus on **categorical variables** (ex: infested or not infested, cold or no cold). In this case, the parameter of interest is **proportion**.

CH5-9: Normal, t, χ^2 , F distributions

- All continuous distributions

CH10: **Binomial and Poisson distributions**

- Both discrete distributions
- BUT sometimes these discrete distributions are approximated by continuous distributions!

The Ch 10 notes includes:

Binomial Distribution

- Binomial Distribution and Its Approximation (notes10.1)
- Inference for a Single Proportion π (notes10.2)
- Comparing ≥ 2 Proportions (notes10.3)
- Contingency Tables: Tests for Independence and Homogeneity (notes10.4)
- **Odds Ratios (notes10.5)**

Poisson Distribution (notes10.6)

Ch10.5: Odds Ratios

1. Odds
2. Odds Ratio
3. Case Control Studies
4. Simpson's Paradox
5. Breslow-Day (BD) Test
6. Cochran-Mantel-Haenszel (CMH) Test
7. Designed Experiments versus Observational Studies

1. Odds

- So far in Chapter 10 we have been focusing on **probabilities** and **proportions**.
- Another way to summarize count data on categorical variables is to use **odds** and **odds ratios**.
- Odds is the ratio of the probability that an event happens to the probability that it does not happen.
- **Odds = $P(\text{yes})/(1-P(\text{yes})) = P(\text{yes})/P(\text{no})$.**
- Example: Suppose $P(\text{yes}) = 0.8$.
Then odds = $0.8/0.2 = 4$.
Often expressed as odds of 4:1.
- While probabilities and proportions may be easier to interpret, odds and odds ratios have some nice mathematical properties.
- Odds ratios are important for interpreting case-control studies and logistic regression results.

2. Odds Ratios

Odds ratio (λ) is **the ratio** of the odds (of an event) for **two groups**.

Skiers Example:

	No Cold	Cold	Total
Placebo	109	31	140
Vitamin C	122	17	139

$$\text{Odds ratio} = \lambda = \frac{\text{odds of cold in VitC group}}{\text{odds of cold in Placebo group}} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

$$\text{Estimate by } \hat{\lambda} \cong \frac{\left(\frac{17}{139}\right) / \left(\frac{122}{139}\right)}{\left(\frac{31}{140}\right) / \left(\frac{109}{140}\right)} = \frac{(17/122)}{(31/109)} = \frac{(17 \times 109)}{(31 \times 122)} = 0.49$$

Interpretation: Odds of getting a cold when taking Vitamin C are estimated to be about half the odds of getting a cold when taking Placebo.

Odds Ratios and their CIs in R

```
> library(epitools)
> oddsratio(Skiers, method = "wald")
$data
      NoCold YesCold Total
Placebo     109       31    140
VitC        122       17    139
Total       231       48    279

$measure
                               NA
odds ratio with 95% C.I. estimate      lower      upper
                           Placebo 1.0000000      NA      NA
                           VitC   0.4899524 0.2569419 0.9342709

$p.value
                  NA
two-sided midp.exact fisher.exact chi.square
  Placebo          NA          NA          NA
  VitC   0.02951602 0.03849249 0.02827186
```

Comments on odds ratios in R:

- We can calculate odds ratio in R using `oddratio()` from the `epitools` package.
- When calculating odds ratios in R, it helps to have
 - reference/control group in the first row and
 - “event” of interest in the last column.
- The results include the estimated odds ratio and its corresponding confidence interval.
- The results also include p-values of Fisher’s exact test and chi-square test.

General comments on odds ratios:

1. $\lambda = 1$ indicates equal odds for the two groups. This can be interpreted as no difference.
2. The odds ratio is not affected if an entire row or column is multiplied by a constant. This is important for analyzing case-control studies.
3. If the rows (or columns) are interchanged, then $\lambda_{\text{new table}} = 1/\lambda_{\text{old table}}$.
4. Odds ratio can be used for larger tables by combining rows and/or columns down to 2 by 2, or by looking at a series of 2 by 2 subtables.
5. The textbook gives the formula for calculating CI for odds ratio. The CI will include the estimate, but is **not symmetric** (due to the way the interval is calculated).
6. For odds ratio, it is somewhat more common to present CI. But chi-square or Fisher's Exact tests still apply.
7. If a CI for λ does not include 1, we can conclude there is a difference between the odds for the two groups. If a CI for λ includes 1, then we cannot conclude a difference.

3. Case Control Studies

- A case control study is a special type of retrospective, observational study.
- This type of study is often used when the event of interest (typically a disease or particular cause of death) is relatively rare.
- First, subjects with and without the event of interest are identified. (In prospective study, the event of interest would typically correspond to the “response”.) For example, identify subjects with and without a particular disease. Often, subjects are “matched” based on other factors (ex: age or location).
- Next, information is gathered about the subjects from their past concerning risk factors that are potentially associated with the disease.
- Important Note: We **cannot** estimate the probability of the event from a case control study. We certainly do **not** have a random sample, since we start by identifying subjects with and without the event.
- The odds ratio is an appropriate analysis for case-control studies.

Birth Control (Case-Control) Example:

- The goal of this 1975 study was to look for an association between birth control use and myocardial infarction (MI = “heart attack”) in women under age 45. (Note current birth control formulations have changed considerably since 1975!)
- 58 married women under 45 being treated for MI (“cases”) were identified. Then each was matched with approximately three “controls” (women of similar age, weight, etc) without MI.
- All subjects then classified on whether they had used oral contraceptives.
- Important note: The investigators decide the number of cases and controls. (In this case, approximately 25% of subjects are cases). Hence, we cannot estimate the probability of the event from a case control study but we can estimate the odds ratio.
- The simple odds ratio (next slide) does not account for the matching of cases with controls. Alternate analysis approaches include conditional logistic regression and propensity score matching.

Birth Control Example:

Contraceptive
practice
 $(p=0.004)$

Myocardial infarction

	No	Yes	Total
Never Used	132	35	167
Used	34	23	57
Total	166	58	224

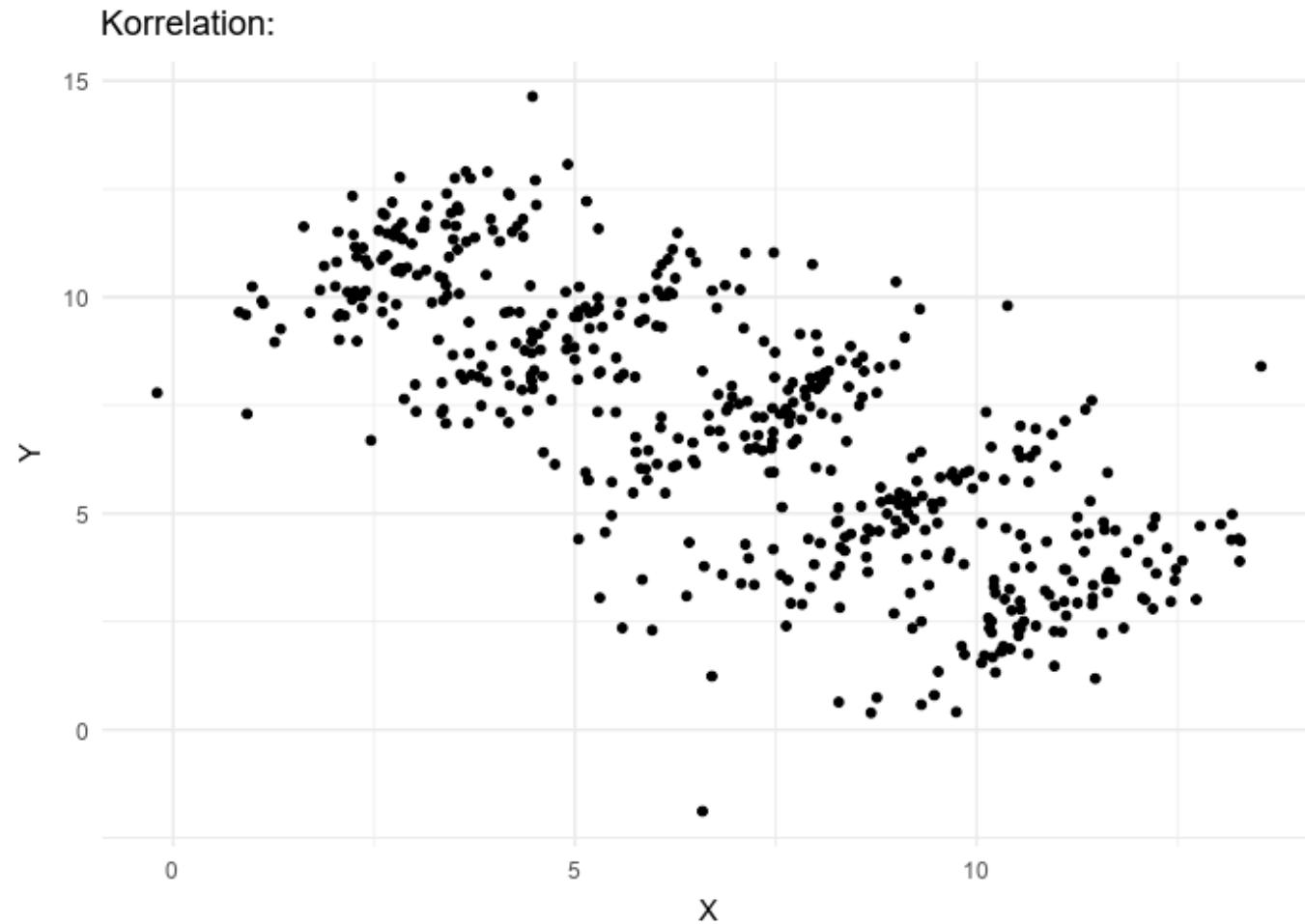
$$\hat{\lambda} = \frac{(23/34)}{(35/132)} = 2.55 \quad 95\% \text{ C.I. } (1.34, 4.87)$$

Conclusion: The odds of myocardial infarction are estimated to be 2.55 times higher among those that have used oral contraceptives. Based on the CI (completely above 1) and the chi-square p-value (0.004), we have evidence of an association between contraceptive use and odds of heart attack.

4. Simpson's Paradox

- It is important to note that the vast majority of STAR511 is focused on inference with one response variable and one predictor variable.
- But in this section, we will have a brief discussion of the impact of a third variable.
- Simpson's paradox refers to the scenario in which a trend appears in several different groups but disappears or reverses when these groups are combined.
- In some situations, important information may be lost when groups are combined.
- We will discuss formal ways to check for this, but a simple starting point is to consider the groups. Do the easy things first!

Simpson's paradox can occur with categorical or numerical responses.
<https://commons.wikimedia.org/w/index.php?curid=62007681>



Berkley Example:

- This is classic example using 1975 graduate admissions data from six departments.
- We start by looking at “aggregate” or “combined” data (“ignoring” department).

	Rejected	Admitted	Total
Male	1493	1198	2691
Female	1278	557	1835

- Overall, 44.5% of males were admitted vs 30.4% of females.
- The estimated odds of admission for males vs females is found to be $\hat{\lambda} = 1.84$ (95% CI = (1.62, 2.09), p-value < 0.001). This supports that men have higher odds of admission as compared to women.

- Logical next step is to look at individual departments.
- Most (4/6) departments have higher odds of admission for women!

	A	B	C	D	E	F
Odds Ratio (M vs F)	0.349	0.802	1.133	0.921	1.222	0.827

- Since the odds ratio noticeably varies by department, the combined results are not representative. Report by department!
- But it is also important to note that other things vary by department including admission rate and proportion of male applicants.
- Note that the 3 departments with highest proportion female applicants also have the lowest admission rates!

	A	B	C	D	E	F
Prop Admitted	0.644	0.632	0.351	0.339	0.252	0.064
Prop Male	0.884	0.957	0.354	0.527	0.327	0.522

5. The Breslow-Day (BD) Test

- Breslow-Day is a test for equality of odds ratios across **some number of “groups”**.
- $H_0: \lambda_1 = \lambda_2 = \dots = \lambda_g$ vs H_a : Not all the λ 's are the same.
- The Breslow-Day test can be done in R using the `BreslowDayTest()` function from the `DescTools` package.
- An alternative approach is to use logistic regression and test for interaction (for example, between department and gender).

Berkley Example:

- The BD p-value = 0.0021, so we Reject H_0 and conclude that the odds ratios are not the same for all departments.
- Hence, we should not combine information across departments!

6. The Cochran-Mantel-Haenszel (CMH) Test

- CMH is a tests for “average” λ (averaging across several groups).
- $H_0: \lambda_{CMH} = 1$ vs $H_a: \lambda_{CMH} \neq 1$.
- Note: We should check BD test first. If we have evidence of differences between groups (small BD p-value), we should NOT use CMH.
- The CMH test can be done in R using the `cmh.test()` function from the `lawstat` package.
- An alternative approach is to use logistic regression with additive effects.

Drug Clinic Example:

In this designed experiment, one drug is tested at three clinics.

Clinic	Drug group	Improved	Not Improved	Total
1	Drug	40 (80%)	10	50
	Placebo	15 (30%)	35	50
	<i>Total</i>	55	45	100
2	Drug	35 (70%)	15	50
	Placebo	20 (40%)	30	50
	<i>Total</i>	55	45	100
3	Drug	43 (86%)	7	50
	Placebo	31 (62%)	19	50
	<i>Total</i>	74	26	100

Drug Clinic Example:

1. First consider the odds ratio for each clinic:

$$\lambda_1 = 9.33, \lambda_2 = 3.50, \lambda_3 = 3.77$$

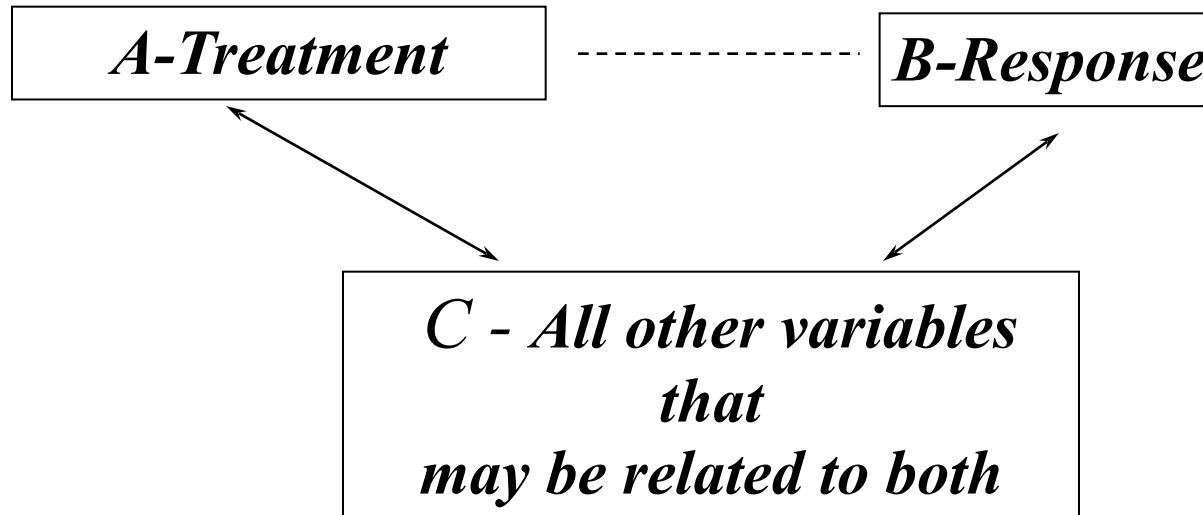
2. Based on the BD test ($p = 0.245$), we fail to reject H_0 . We do not have evidence that odds ratios are not the same for all clinics.
3. Hence, it makes sense to combine information across the clinics, (controlling for clinic). Using CMH we find:

$$\lambda_{CMH} = 4.90, p < 0.001$$

Based on the CMH test (p -value < 0.001) we conclude that the odds ratio is different from 1. The odds of improvement are higher for the active treatment as compared to placebo.

7. Designed Experiments versus Observational Studies

Question: Why can you conclude more in a designed study (with treatments randomly assigned) than in an observational study?



Randomization of subjects (experimental units) to treatment groups helps achieve (approximate) balance with respect to the “confounding variables”. It “breaks the circuit” between A and C.

Chapter 11: Linear Regression and Correlation

In STAT511 we focus on analyses with a single response (or dependent) variable (Y) and a single predictor (or independent) variable (X). In R: `lm(Y ~ X)`, `plot(Y ~ X)`

- **Continuous response with a categorical predictor** -> two-sample t-test (CH6) or one-way ANOVA (CH8) to compare means
Example Rat lead: Y = amount of solution consumed (#),
X = treatment group (Control or Deficient)
- **Categorical response with categorical predictor** -> chi-squared test, FET or Z-test (CH10) to compare proportions
Example Skiers: Y = cold status (Yes or No),
X = treatment group (Vitamin C or Placebo).
- **Continuous response with continuous predictor** -> consider regression to estimate slope or correlation. The most important instance of regression methodology is **linear regression (CH11)**.
Example Corn Yield: Y = corn yield (#), X = fertilizer (#)
- **Categorical (binary) response with continuous predictor** -> logistic regression (CH12-notes12.2)
Example Beetle Kill: Y = status (dead or alive), X = pesticide dose (#)

The Ch 11 notes focus on simple linear regression (only one predictor):

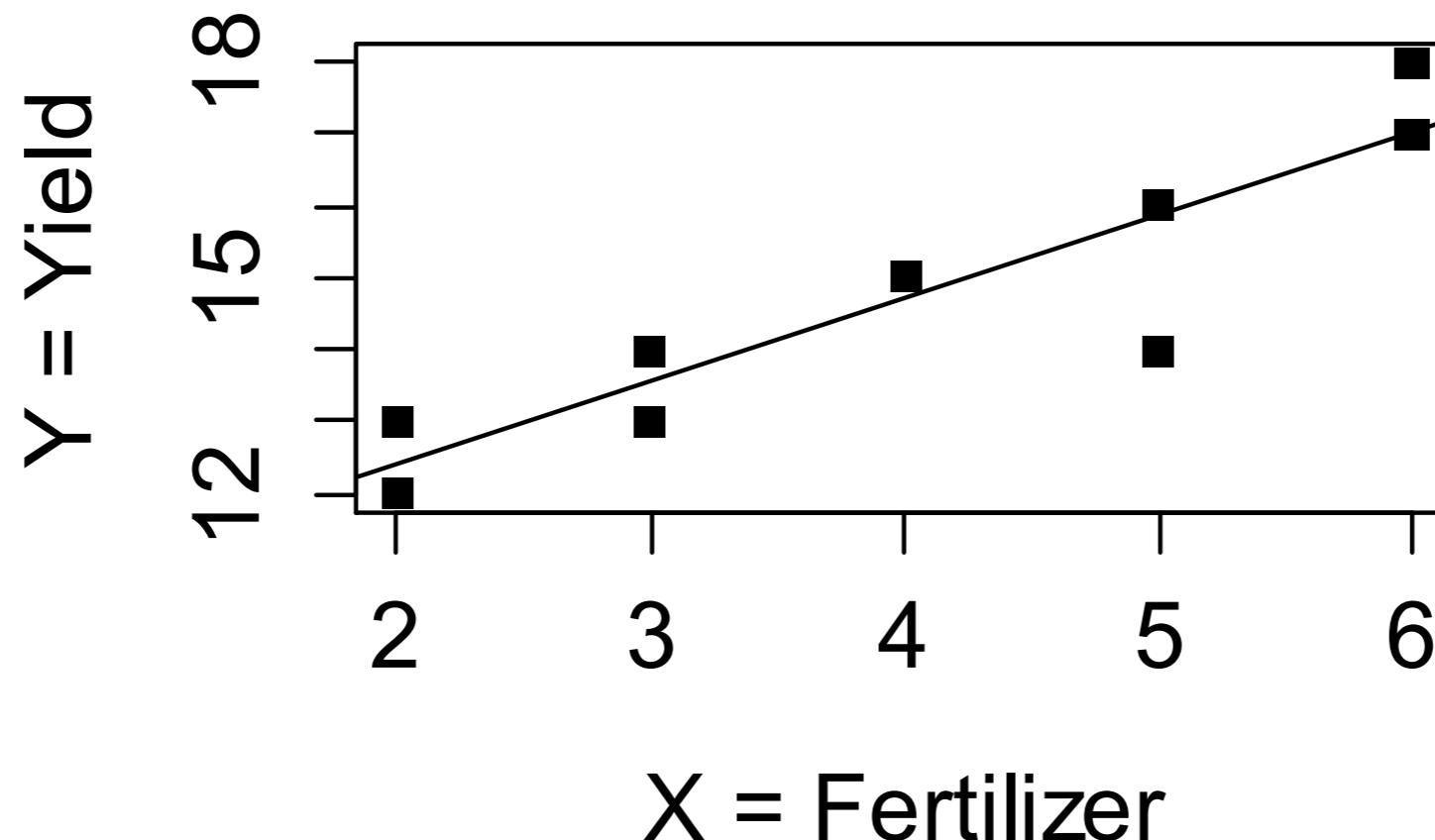
- **Simple linear regression I (notes11.1)**
- Simple linear regression II (notes11.2)
- Linear regression special topics (notes11.3)
- Correlation (notes11.4)

Corn Example:

A study was done where different quantities of fertilizer (lbs/plot) were applied to fields. The response is corn yield (bu/plot)

Considering the data, the yield of corn (Y) is approximately linearly related to fertilizer applied (X).

Hence, simple linear regression is reasonable here.



Chapter 11.1: Simple Linear Regression I

1. Simple Linear Regression Model
2. Least Square Estimation for Simple Linear Regression
3. Interpretation of Slope (β_1) and Intercept (β_0) in the model
4. Properties of $\hat{\beta}_1$ and $\hat{\beta}_0$
5. EXTRA: Sir Francis Galton

1. Simple linear regression Model

Simple linear regression

It is typically used to model the linear relationship between two variables Y and X .

- Y : response variable or dependent variable
 - The variable whose values we want to explain
 - Its values depend on something else
- X : explanatory variable or predictor variable
 - The variable that explains the other one

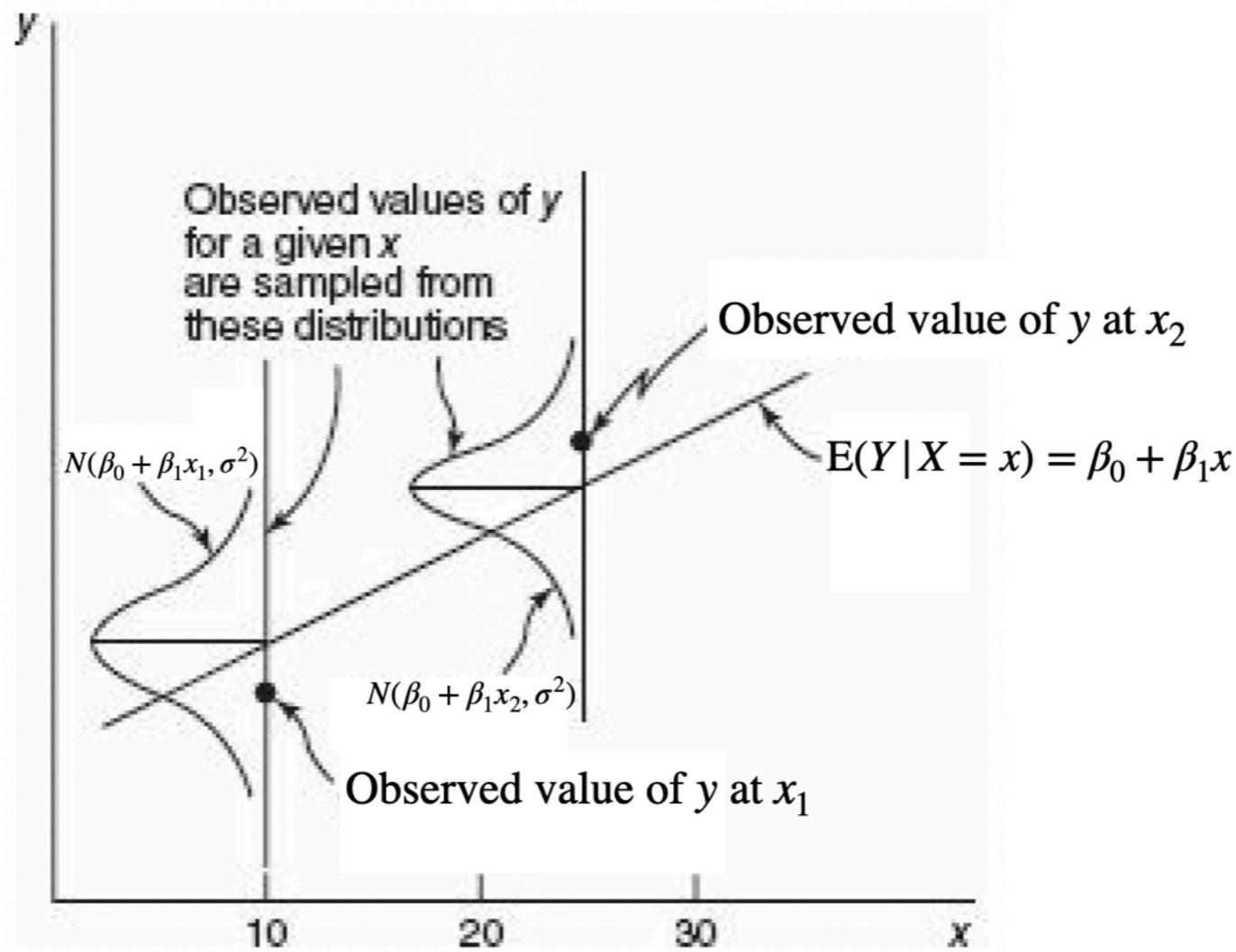
Our interest centers on how the distribution of Y changes as X is varied.

- For the distribution of $Y|X = x$, three important components of this distribution are
 - Center: *the mean function*, which we define by $E(Y|X = x)$,
 - Spread: *the variance function*, defined by $\text{Var}(Y|X = x)$, and
 - Shape: it can be seen from the graph

- In simple linear regression, $Y|X = x$ is *normally* distributed, where
 - $E(Y|X = x)$: It is a linear equation (the true trend in the scatterplot), i.e.,

$$E(Y|X = x) = \beta_0 + \beta_1 x,$$
 where the unknown parameters β_0 and β_1 determine the intercept and the slope of a specific straight line, respectively.
 - ◆ The slope β_1 plays a special role:
 - $\beta_1 = 0$ implies no dependence
 - $\beta_1 > 0$ implies positive dependence
 - $\beta_1 < 0$ implies negative dependence
 - $\text{Var}(Y|X = x)$: A frequent assumption in fitting linear regression models is that the variance function is the same for every value of x . This is usually written as

$$\text{Var}(Y|X = x) = \sigma^2,$$
 where σ^2 is a generally unknown positive constant.



- That is, Given X at x , $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$, where β_0, β_1 and σ^2 are unknown parameter and need to be determined.
- We see that only $E(Y|X=x)$ depends on x ; other aspects of the distribution of $Y|X=x$ do not.

- Because the variance $\sigma^2 > 0$, the observed value of the response y at $X = x$ will typically not equal its expected value $E(Y|X = x)$.
- A **statistical error**, or ϵ , is the difference between the observed data y and the expected value $E(Y|X = x)$, i.e.,

$$\epsilon = y - E[Y|X = x] = y - \beta_o - \beta_1 x.$$
 - ϵ is random variable, and $\epsilon \sim N(0, \sigma^2)$.
- A more familiar form of the linear model is

$$y = \beta_o + \beta_1 x + \epsilon,$$
 where $\epsilon \sim N(0, \sigma^2)$

- We collect data in pairs. The standard notation used to designate this is:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),$$

where x_i denotes the i th value of variable X , and y_i denotes the i th value of variable Y .

- Then we obtain the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ for } i = 1, \dots, n,$$

where:

- $E(\epsilon_i) = 0$
- Constant variance: $\text{Var}(\epsilon_i) = \sigma^2$
- The errors ϵ_i 's are all independent, meaning that the value of the error for one case gives no information about the value of the error for another case.
- Errors ϵ_i 's are often assumed to be normally distributed.
 $\iff \epsilon_1, \dots, \epsilon_n \stackrel{iid}{\sim} N(0, \sigma^2)$

- The distributions of y_i 's depend on β_0, β_1 and σ^2 , which have to be estimated. - How to do this?

2. Least Square Estimation

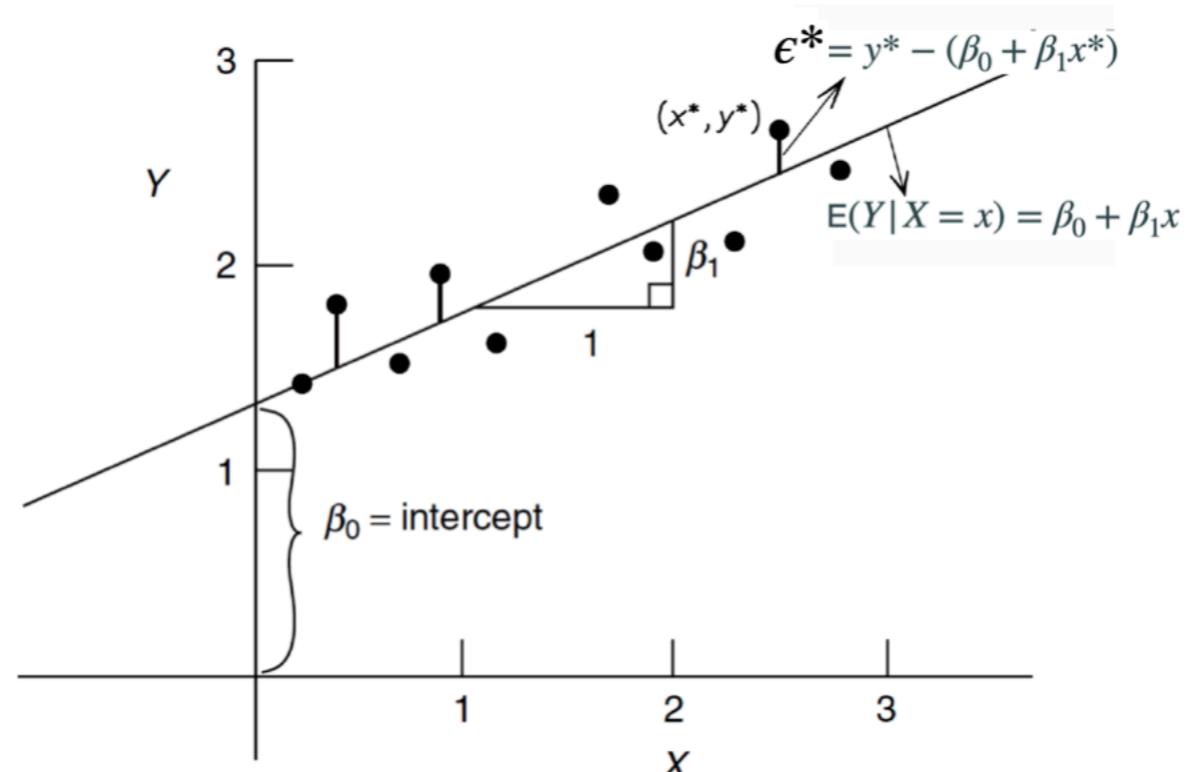
Recall that

- The model for simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ for } i = 1, \dots, n,$$

where $\epsilon_1, \dots, \epsilon_n \stackrel{iid}{\sim} N(0, \sigma^2)$

We first look at the estimation of β_0 and β_1 .



- The idea is to make the fitted line as close as possible to the data.

- Formally, we estimate β_0 and β_1 by minimizing $\sum_{i=1}^n \left(y_i - E(Y|X=x_i) \right)^2 = \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2$.
- For $L(\beta_0, \beta_1) = \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2$, we set

$$\frac{d}{d\beta_0} L(\beta_0, \beta_1) = 0 \text{ and } \frac{d}{d\beta_1} L(\beta_0, \beta_1) = 0.$$
(Solving these partial derivatives for β_0 and β_1)
 - Web Demo: <https://www.desmos.com/calculator/zvrc4lg3cr>
- After much algebra, we have that

$$\hat{\beta}_1 = S_{xy}/S_{xx} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where

- sum of square S_{xx} : $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$
- sum of cross-product S_{xy} : $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

In practice, we estimate the slope and intercept using lm().

- Fit=lm(y~x, data=mydata) # computes the model and stores in ‘Fit’
- summary(Fit) # displays the estimated coefficients

The **fitted value** for the i th case, denoted \hat{y}_i , is given by $\hat{E}(Y|X = x_i)$, i.e.,

$$\hat{y}_i = \hat{E}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ for } i = 1, \dots, n.$$

In R:

- `plot(y~x, data=mydata)` # creates a scatterplot
- `abline(my.model)` # superimposes the best fitted line on the scatterplot

The **residual** for the i th case, denoted e_i , is given by

$$e_i = y_i - \hat{E}(Y|X = x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \text{ for } i = 1, \dots, n,$$

which should be compared with the equation for the statistical error for the i th case, ϵ_i ,

$$\epsilon_i = y_i - E(Y|X = x_i) = y_i - (\beta_0 + \beta_1 x_i) \text{ for } i = 1, \dots, n.$$

3. Interpreting Slope and Intercept

- Slope is often more useful than intercept
- Pay attention to the context of the data
- The linear model includes the mean function:

$$E(Y|X = x_i) = \beta_0 + \beta_1 x_i \text{ for } i = 1, \dots, n.$$

- **Interpreting the intercept**

$$E(Y|X = 0) = \beta_0 + \beta_1 0 = \beta_0$$

- Therefore, the **intercept** represents the **mean** value of Y when $x_i = 0$.
- This is not always useful or even meaningful information.
 - ◆ eg. Suppose that a valid model relating price of homes in Westwood to size of the homes in square-feet is given by $\hat{E}(\text{price}) = 173000 + 344 \times \text{sqft}$
Interpret the intercept:
While it seems strange to think of a home with 0 square-feet, this might be telling us that the mean value of empty lots (lots with homes of 0 square-feet) is \$173,000.
- Warning: Often $x = 0$ is far beyond the range of the data collected, and so the linearity of the association may no longer hold.

- **Interpreting the slope**

$$\Delta E(Y) = E(Y|X = x_i + 1) - E(Y|X = x_i) = (\beta_0 + \beta_1(x_i + 1)) - (\beta_0 + \beta_1x_i) = \beta_1$$

- Slope: A 1 unit difference in x is **associated with** a β_1 unit increase in the **mean** of Y.
- But in statistics, we also need to be aware of the context, and report an interpretation that makes sense and is true in the context of the data.
 - ♦ eg.

$$\hat{E}(\text{price}) = 173000 + 344 \times \text{sqft}$$

Interpret the slope:

Each additional square foot of the home is associated with a mean of 344 more dollars

- **Note:** intercept and slope are unit dependent.

Return to the Corn example:

Here we calculate the estimated slope and intercept “by hand”, but the values are also given in the computer output.

$$\bar{x} = 4.0$$

$$\bar{y} = 14.7$$

$$S_{xy} = 23$$

$$S_{xx} = 20$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{23}{20} = 1.15$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 14.7 - 1.15 \times 4 = 10.10$$

Interpretation of the intercept and slope:

Intercept: Predicted yield (y) with no fertilizer applied (x=0) is estimated to average 10.10 bu/plot. (This prediction should be taken with caution, because 0 is beyond the range of the data).

Slope: A one lb/plot increase in fertilizer (1 unit increase in x) is associated with a 1.15 bu/plot predicted increase in yield (y).

Linear Regression in R

- In R, use `lm(Y ~ X)`
- For the Corn example:

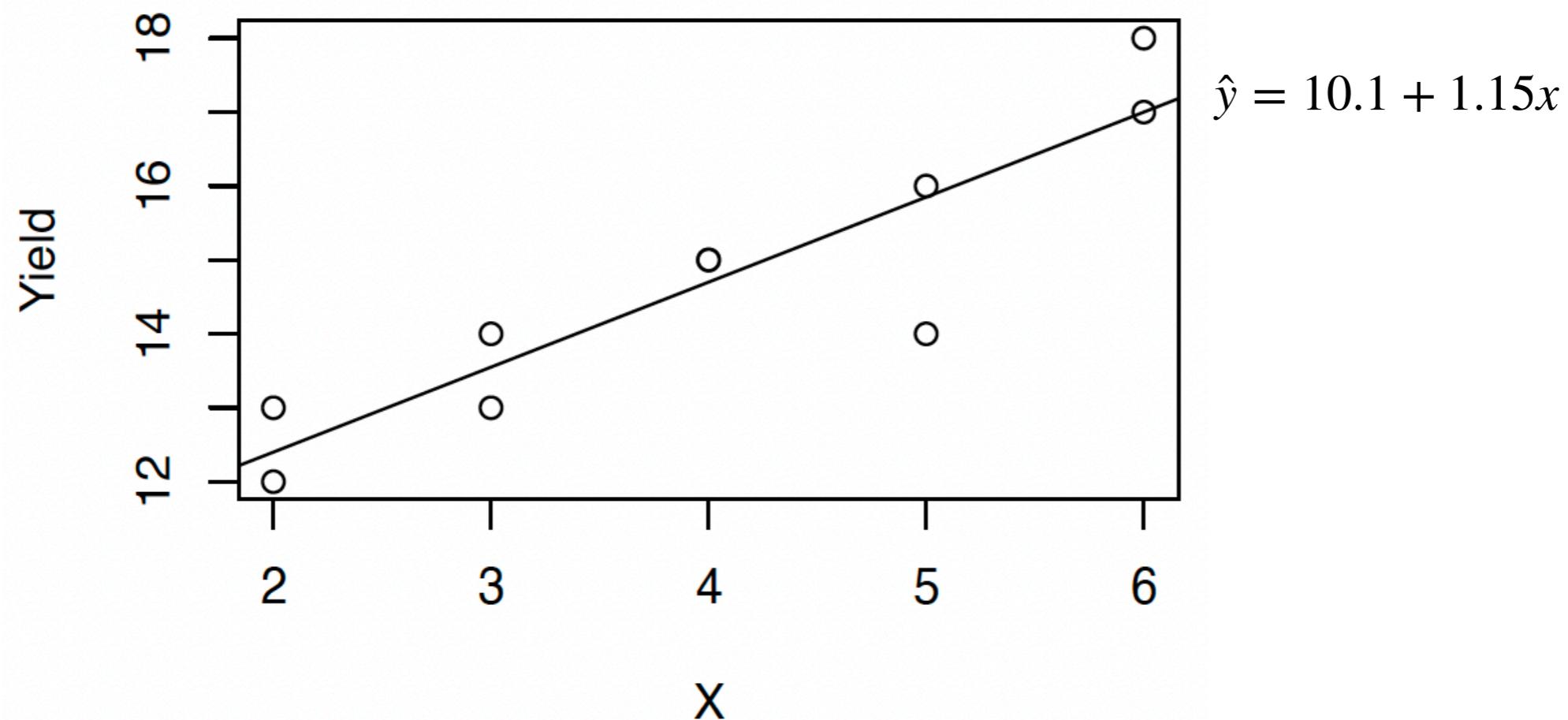
```
> Fit<-lm(Yield ~ X, data=Corn)
> summary(Fit)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.1000   0.7973 12.67 1.42e-06
X            1.1500   0.1879  6.12 0.000283

Residual standard error: 0.8404 on 8 degrees of freedom
```

Recall that we also used the `lm()` function to fit the one-way ANOVA model.

```
> plot(Yield ~ X, data = Corn)  
> abline(Fit)
```



4. Properties of $\hat{\beta}_1$ and $\hat{\beta}_0$

- Recall that $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.
 - They can be rewritten as $\hat{\beta}_1 = \sum_{i=1}^n c_i y_i$ and $\hat{\beta}_0 = \sum_{i=1}^n d_i y_i$, where $c_i = \frac{x_i - \bar{x}}{S_{xx}}$ and $d_i = 1/n - c_i x_i$.
 - Both $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear combinations of y_i 's, and hence linear combinations of errors ϵ_i 's.
 - The following properties can be easily shown under the assumptions of ϵ_i 's.
- Property 1:** Both $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased, i.e.,
- $$E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0.$$
- **Remark:** By Gauss-Markov Theorem, the LS estimators are **the best linear unbiased estimators** (BLUE).

Property 2: We have that $\text{Var}(\hat{\beta}_1) = \frac{1}{S_{xx}} \sigma^2$,

$$\text{Var}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \sigma^2, \text{ and}$$

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_0) = -\frac{\bar{x}}{S_{xx}} \sigma^2$$

Property 3: $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$ and $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \sigma^2\right)$

- Note that
 1. **Property 1** and **Property 2** do not require a distributional assumption concerning the errors.
 2. The assumption of normality of errors is needed for **Property 3** when the sample size is small.
 - If the sample size is large enough, the central limit theorem shows that LS estimators will be approximately normally distributed.

5. Estimator of σ^2 and Residuals

- σ^2 can be estimated by

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2,$$

where $e_i = y_i - \hat{y}_i$ is the *residual* and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the *fitted value* of y_i for x_i .

- Note that
 - $E(\hat{\sigma}^2) = \sigma^2$
 - $n - 2$ is the residual degree of freedom (*df*)
(residual *df* = total number of observations - number of parameters in the mean function)
- In other words, σ^2 is estimated by the variance of the residuals.

Two ways to obtain the estimated σ^2

- (a)

```
anova(Fit)
## Analysis of Variance Table
##
## Response: Yield
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1 26.45  26.4500 37.451 0.0002832 ***
## Residuals  8  5.65  0.7062
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- (b)

```
Fit <- lm(Yield ~ X, data = Corn)
summary(Fit)

##
## Call:
## lm(formula = Yield ~ X, data = Corn)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -1.8500 -0.3000  0.2250  0.4125  1.0000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.1000    0.7973 12.67 1.42e-06 ***
## X            1.1500    0.1879  6.12 0.000283 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error 0.8404 on 8 degrees of freedom
## Multiple R-squared:  0.824, Adjusted R-squared:  0.802
## F-statistic: 37.45 on 1 and 8 DF, p-value: 0.0002832
```

$\hat{\sigma}^2$ is the squared residual standard error, i.e., $\hat{\sigma}^2 = s^2 = (0.8404)^2 = 0.7062722$

- standard deviation of ϵ_i 's (σ) is estimated by residual standard error ($\hat{\sigma}$ or s).

Standardized (or Studentized) Residuals

"Raw" residual: $e_i = y_i - \hat{y}_i$

$$\text{SE}(e_i) = s \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}$$

Standardized residual: $\frac{e_i}{\text{SE}(e_i)}$

Standardized (or studentized) residuals are residuals that have been “standardized” by dividing each residual by its SE.

- They have approximately a t-distribution, which is approximately normal for moderate sample sizes, so we expect that about 95% of the standardized residuals will be between -2 and +2. Values greater in absolute value than 3.5 are usually considered **outliers**.
- In R, standardized residuals can be found using `rstandard(ModelObject)`.

Chapter 11: Linear Regression and Correlation

In STAT511 we focus on analyses with a single response (or dependent) variable (Y) and a single predictor (or independent) variable (X). In R:
`lm(Y ~ X), plot(Y ~ X)`

- **Continuous response with a categorical predictor** -> two-sample t-test (CH6) or one-way ANOVA (CH8) to compare means
Example Rat lead: Y = amount of solution consumed (#),
X = treatment group (Control or Deficient)
- **Categorical response with categorical predictor** -> chi-squared test, FET or Z-test (CH10) to compare proportions
Example Skiers: Y = cold status (Yes or No),
X = treatment group (Vitamin C or Placebo).
- **Continuous response with continuous predictor** -> consider regression to estimate slope or correlation. The most important instance of regression methodology is **linear regression (CH11)**.
Example Corn Yield: Y = corn yield (#), X = fertilizer (#)
- **Categorical (binary) response with continuous predictor** -> logistic regression (CH12-notes12.2)
Example Beetle Kill: Y = status (dead or alive), X = pesticide dose (#)

The Ch 11 notes focus on simple linear regression (only one predictor):

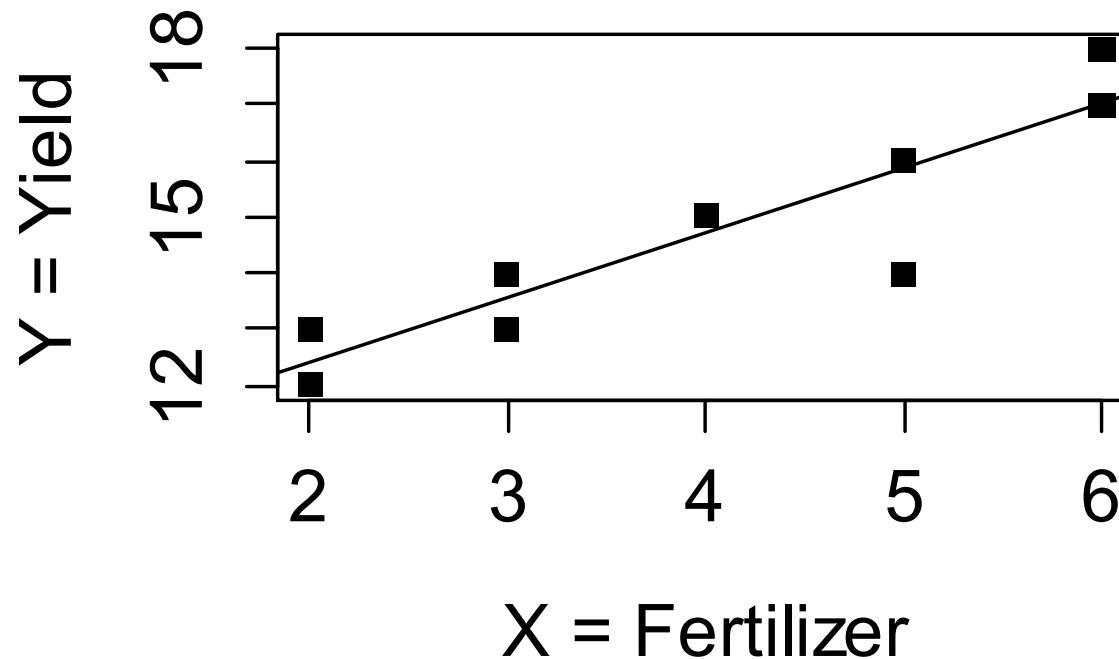
- Simple linear regression I (notes11.1)
- **Simple linear regression II (notes11.2)**
- Linear regression special topics (notes11.3)
- Correlation (notes11.4)

Corn Example:

A study was done where different quantities of fertilizer (lbs/plot) were applied to fields. The response is corn yield (bu/plot)

Considering the data, the yield of corn (Y) is approximately linearly related to fertilizer applied (X).

Hence, simple linear regression is reasonable here.



Chapter 11.2: Simple Linear Regression II

1. Inference for the Intercept and Slope
2. Inference for Mean response and Future Response
3. Regression ANOVA Table
4. Remedies for the failure of model assumptions
 - Transformations
 - Adding quadratic terms
5. F-test for Lack of Fit

1. Inference for the Intercept and Slope

Review:

Sampling distributions of the intercept and slope estimators are as follows:

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right) \text{ and } \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1 , respectively. Their variances can be calculated by replacing σ^2 by s^2 in the above formulas:

$$Var(\hat{\beta}_0) = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \quad Var(\hat{\beta}_1) = \frac{s^2}{S_{xx}}, \text{ where}$$

- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ and
- $s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$ (see P22 of Notes11.1 for obtaining s^2 or s from R functions)

Their standard errors are

$$SE(\hat{\beta}_0) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad SE(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}$$

Test and Confidence Interval for β_0 (Intercept)

Assumptions : Regression model Assumptions.

Hypotheses:

$$H_0: \beta_0 = b_0 \text{ vs } H_a: \beta_0 \neq b_0$$

Test Statistic: $t = \frac{\hat{\beta}_0 - b_0}{SE(\hat{\beta}_0)} \sim T_{df} \text{ under } H_0$

where $df = n - 2$

(Two-sided) P-value: $2 * P(T \geq |t|)$. **R code:** $2 * (1 - pt(abs(t), df = n - 2))$

Confidence Interval: $\hat{\beta}_0 \pm t_{\alpha/2, df} SE(\hat{\beta}_0)$

NOTES:

- Inference for the intercept is rarely of research interest.
- Test of $H_0: \beta_0 = 0$ is most common.

Test and Confidence Interval for β_1 (Slope)

Assumptions : Regression Model Assumptions.

Hypotheses:

$$H_0: \beta_1 = b_1 \text{ vs } H_a: \beta_1 \neq b_1$$

Test Statistic: $t = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)} \sim T_{df} \text{ under } H_0$

where $\text{df} = n - 2$

(Two-sided) P-value: $2 * P(T \geq |t|)$ **R code:** $2 * (1 - pt(\text{abs}(t), df = n - 2))$

Confidence Interval: $\hat{\beta}_1 \pm t_{\alpha/2, df} SE(\hat{\beta}_1)$

NOTES:

- Inference for the slope is more commonly of research interest.
- Test of $H_0: \beta_1 = 0$ is most common. This is considered a test of linear association between Y and X.

Tests and Confidence Intervals in R

- In R, use `lm(Y~X)` . Use `confint()` for CIs.

For the Corn example:

```
> LMFit<-lm(Yield~X,data=Corn)
```

```
> summary(LMFit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.1000	0.7973	12.67	1.42e-06
X	1.1500	0.1879	6.12	0.000283

```
> confint(LMFit)
```

	2.5 %	97.5 %
(Intercept)	8.2615130	11.938487
X	0.7166645	1.583336

Tests of $H_0: \beta_0=0$ and $H_0: \beta_1=0$

95% Confidence Intervals for β_0 and β_1

2. Inference for Mean response and Future Response

Mean Response vs. Future Response

- Both have the same value, but they are used to answer different questions:
 - ◆ Given $X = x^*$, what's mean value of $Y|X = x^*$?
 - mean response: $\mu^* = E[Y|X = x^*] = \beta_0 + \beta_1 x^*$, which is unknown because β_0, β_1 are unknown.
 - The natural estimator for μ^* is $\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ (**Estimation**)
 - ◆ Given $X = x^*$, what's the value of individual y ?
 - future response: Let y^* be a future response at $X = x^*$.
 - y^* is predicted by \hat{y}^* (**Prediction**), where $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$, which has the same value as $\hat{\mu}^*$.
- Therefore, given $X = x^*$, there are **two** interpretations of $\hat{\beta}_0 + \hat{\beta}_1 x^*$:
 1. The estimated expected *average* value of Y over *all units* with $X = x^*$.
 2. The predicted value of Y for a *single* (new) unit with $X = x^*$.

Though the estimate (\hat{y}) is the same for either of these scenarios, the standard error of the estimates are different (with a larger SE for the prediction case).

NOTE: Caution should be used when making predictions beyond the range of the data because the linear relationship might only hold in a certain range. (**Extrapolation**)

Confidence Interval for mean response:

The $100(1-\alpha)\%$ confidence interval for $\mu^* = E[Y|X=x^*]$

$$\hat{\mu}^* \pm t_{\alpha/2, df} SE(\hat{\mu}^*) = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2, df} \sqrt{\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right) s^2}$$

Prediction Interval for future response:

Let y^* be a future response at $X = x^*$. We use $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ to predict y^* .

The $100(1-\alpha)\%$ prediction interval for y^* at $X = x^*$:

$$\hat{y}^* \pm t_{\alpha/2, df} SE(\hat{y}^*) = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2, df} \sqrt{\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right) s^2}$$

Notes:

- $df = n-2$ and the critical value: $qt(1-\alpha/2, df = n - 2)$
- s^2 is the estimate of σ^2
- The prediction interval is wider than the corresponding CI, as the prediction case has a larger SE.

- In R, use `predict()`.

Name of the set of new data (a set of predictor values we want to estimate)

For the Corn example:

```
> LMFit<-lm(Yield~X,data=Corn)
> newdata <- data.frame(X = 5.5)
> predict(LMFit, newdata, interval = "confidence")
      fit      lwr      upr
1 16.425 15.53166 17.31834
> predict(LMFit, newdata, interval = "predict")
      fit      lwr      upr
1 16.425 14.29107 18.55893
```

Notes:

The “new data” needs to have the same column/variable name as was used to fit the model.

95% intervals are returned by default. This can be changed using the `level =` option.

3. Regression ANOVA Table

- Fundamental to understanding the ANOVA table are the concepts of "nested models".
- One model is nested inside another if, by adding variables, you get the other model.

For example:

$$y_i = \beta_0 + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

If we add a term for x to the first model, we get the second

$y_i = \beta_0 + \epsilon_i$ is nested inside the more complex model.

- In simple linear regression,
 - The null model: $y_i = \beta_0 + \epsilon_i$
 - The full model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- For these two models, which one we use?

- The null hypothesis will favor the simpler model (with fewer variables) because simplicity is prized in science.

$$y_i = \beta_0 + \epsilon_i$$

H_0 : null model is true

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

H_1 : full model is true.

In this case, this is the same thing as saying that our null hypothesis is that $\beta_1 = 0$.

Previously, to test this hypothesis, we first fit the linear model using `lm()` and then used `summary()`

For the Corn example:

```
> LMFit<-lm(Yield~X, data=Corn)
> summary(LMFit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.1000	0.7973	12.67	1.42e-06
X	1.1500	0.1879	6.12	0.000283

Testing $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$

$t=6.12 \rightarrow P\text{-value}=0.000283 \rightarrow$ the slope is significant.

Another Approach

- Using Analysis of Variance (ANOVA) method
- Using function `anova(LMFit)`
- ANOVA approach allows us to compare “**nested**” models

For the Corn example:

```
> LMFit<-lm(Yield~X,data=Corn)  
> anova(LMFit)
```

Analysis of Variance Table

$$\text{Mean Sq} = \frac{\text{Sum Sq}}{\text{Df}}$$

$$F \text{ value} = \frac{SS_{reg}/Df_{reg}}{RSS/Df_{residual}}$$

Response: Yield

	Df	Sum Sq	Mean Sq	F value	Pr (>F)	
X	1	26.45	26.4500	37.451	0.0002832	***
Residuals	8	5.65	0.7062			

- "Residuals" Sum Sq = SSResid: $SS_{Resid} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- "X" Sum Sq is called "Sum of Squares due to Regression", denoted by SSReg: $SS_{Reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

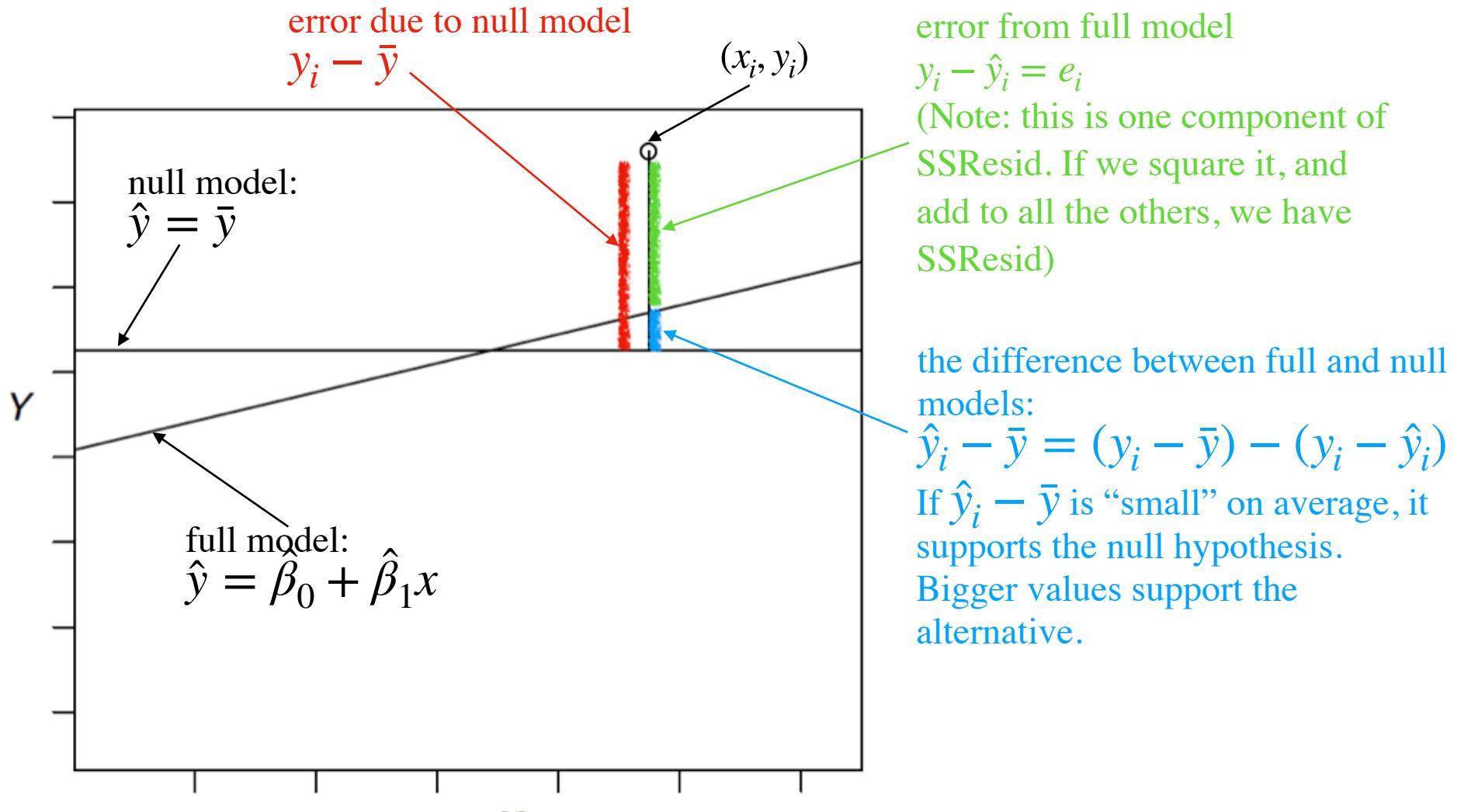
Two components, $SSResid$ and $SSReg$, in ANOVA table are obtained by decomposing “total variation”, i.e., $SSTotal = \sum_{i=1}^n (y_i - \bar{y})^2$.

- We then demonstrate the decomposition of the total variation, i.e.,

$$SSTotal = SSReg(\text{Variation “explained by the regression line”})$$

+

$$SSResid(\text{Variation unexplained})$$



$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$SSTotal = SSReg + SSResid$

To sum up,

<u>Source</u>	<u>SS</u>	<u>df</u>	<u>MS = SS/df</u>	<u>F-test</u>
Regression	SSReg	1	MSReg	$F = MSReg/MSResid$
Residual	SSResid	n-2	MSResid	
Total	SSTotal	n-1		

Recall that our reason for looking at the ANOVA table is to compare two models:

Null model: $y_i = \beta_0 + \epsilon_i$

Full model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \iff$ to see if the slope is 0.

Which is best?

$SSTotal = \sum_{i=1}^n (y_i - \bar{y})^2 \iff$ Residual Sum Square of null model, as \bar{y} is the fitted value of y_i 's.

$SSResid = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \iff$ Residual Sum Square of full model

$\rightarrow SSReg = SSTotal - SSResid$: measures the improvement in adding the slope to the model.

Therefore, we are able to compare null model with full mode via considering SSReg.

\rightarrow If the predictor (X) is useful, SSResid will be small. \rightarrow SSReg will be big.

How do we know if $SSReg$ is “big”?

- coefficient of determination (R^2)
 - We compare $SSReg$ to the total variability, $SSTotal$.
- F-test
 - We compare $SSReg$ to $SSResid$

Notes:

- Both (R^2 and F-test) come from the regression ANOVA table.

3A. Coefficient of determination (R^2)

R^2 helps us choose between the null model and the full model by comparing the $SSReg$ to the total variation ($SSTotal$).

$$R^2 = \frac{SSReg}{SSTotal}$$

R output of `anova()` does not show $SSTotal$ in ANOVA table (only showing $SSReg$ and $SSResid$), but it can be calculated as: $SSTotal = SSReg + SSResid$.

Interpretation of R^2 : **Proportion of variation in Y that is explained by the linear regression on X.**

Connection between R^2 and $SSResid$:

$$\text{We have that } R^2 = \frac{SSReg}{SSTotal} = \frac{SSTotal - SSResid}{SSTotal}$$

- If $R^2 = 1$, then $SSResid = 0$ and we fit the data perfectly.
- If R^2 is small, then lots of $SSResid$ and our fit is not good.

Therefore, large R^2 means adding the slope to the mode is a good thing.

Note that R^2 is the square of the correlation coefficient between the observed and fitted response values.
You will see this again in Notes 10.4.

For the Corn example:

```
> LMFit<-lm(Yield~X,data=Corn)
```

```
> anova(LMFit)
```

Analysis of Variance Table

Response: Yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	26.45	26.4500	37.451	0.0002832 ***
Residuals	8	5.65	0.7062		

$$R^2 = (26.45)/(26.45 + 5.65) = 0.824$$

```
> summary(LMFit)
```

Call:

```
lm(formula = Yield ~ X, data = Corn)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.8500	-0.3000	0.2250	0.4125	1.0000

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.1000	0.7973	12.67	1.42e-06 ***
X	1.1500	0.1879	6.12	0.000283 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8404 on 8 degrees of freedom

Multiple R-squared: 0.824, Adjusted R-squared: 0.802

F-statistic: 37.45 on 1 and 8 DF, p-value: 0.0002832

Coefficient of determination

82.4% of the variability in yield (y) is explained by the linear regression on fertilizer (x).

3B. F-test

We choose a model using R^2 , which compares the $SSReg$ to the $SSTotal$.

But we can also compare $SSReg$ to $SSResid$, since if $SSReg$ is big, $SSResid$ must be small.

But, instead of using $SSResid$, we use $MSResid = SSResid/(n-2)$ for statistical purpose, like a mean.

Do same for the $SSReg$: we use $MSReg = SSReg/1 = SSReg$

- Degree of freedom for $SSReg$ is 1 for just simple linear regression (only 1 X).

$$\text{We then create } F = \frac{SSReg/1}{SSResid/(n - 2)} = \frac{MSReg}{MSResid}$$

If the full model is good, $SSReg$ is big and $SSResid$ is small. $\rightarrow F$ is big.

In other words, large F means adding the slope to the model is a good thing.

Formally, we test the hypotheses

$$H_0 : E(Y|x) = \beta_0$$

$$H_a : E(Y|x) = \beta_0 + \beta_1 x$$

In the simple linear regression, these are equivalent to $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$

Test statistic: $F_0 = \frac{SSReg/1}{SSResid/(n-2)} = \frac{MSReg}{MSResid} \sim F(1, n-2)$ under H_0

P-value in R:

$$1 - pf(F_0, \text{df1} = 1, \text{df2} = n-2)$$

For the Corn example:

```
> LMFit<-lm(Yield~X,data=Corn)
> anova(LMFit)
Analysis of Variance Table

Response: Yield
          Df  Sum Sq Mean Sq F value    Pr(>F)
X           1  26.45  26.4500 37.451 0.0002832 ***
Residuals  8   5.65  0.7062

```

Recall that the corresponding t-test gives the following results.

```
> LMFit<-lm(Yield~X, data=Corn)
```

```
> summary(LMFit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.1000	0.7973	12.67	1.42e-06
X	1.1500	0.1879	6.12	0.000283

$t^2 = (6.12)^2 = 37.451 = F \text{ test statistic value.}$

t and F

- Take any variable that follows a t distribution and square it. The new variable follows an F distribution.
- In practice, the t-statistic that tests **the slope** is the square root of the F statistic that tests **the same slope**. (Hold for simple linear regression - only one X)

4. Regression Model Assumptions

The simple linear regression model carries with it some assumptions:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

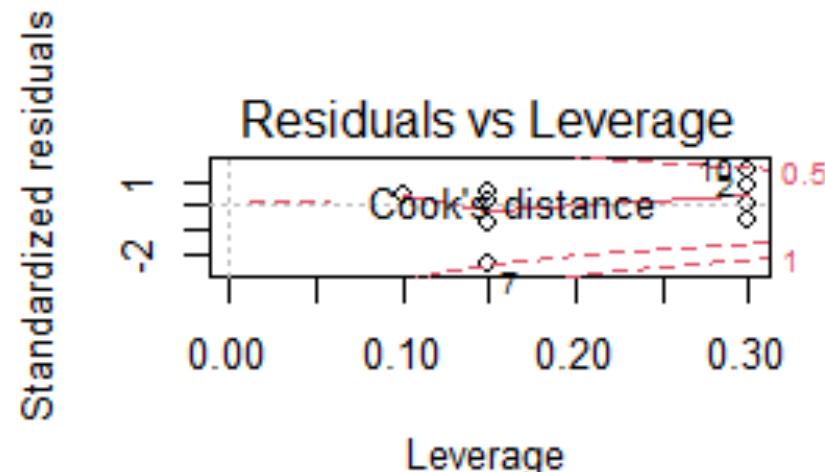
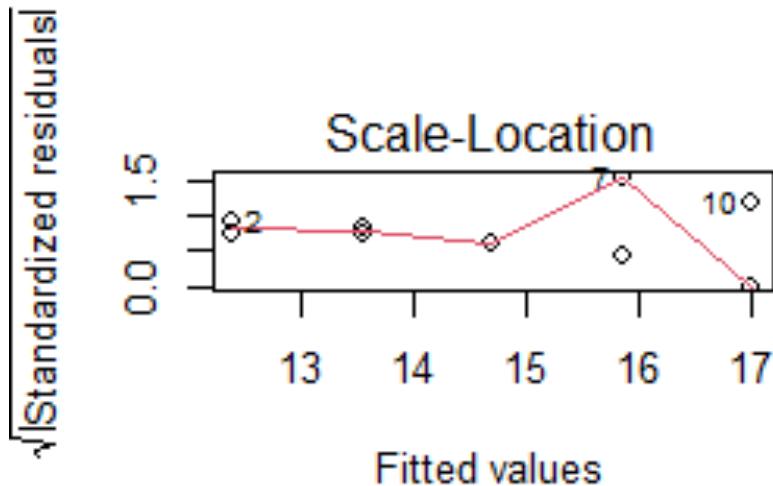
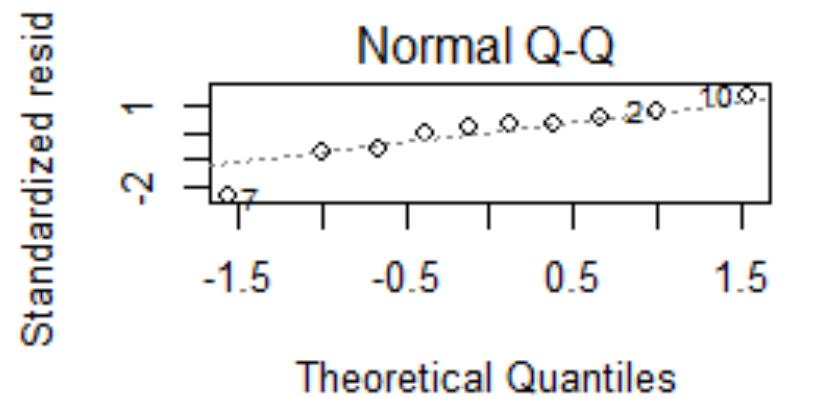
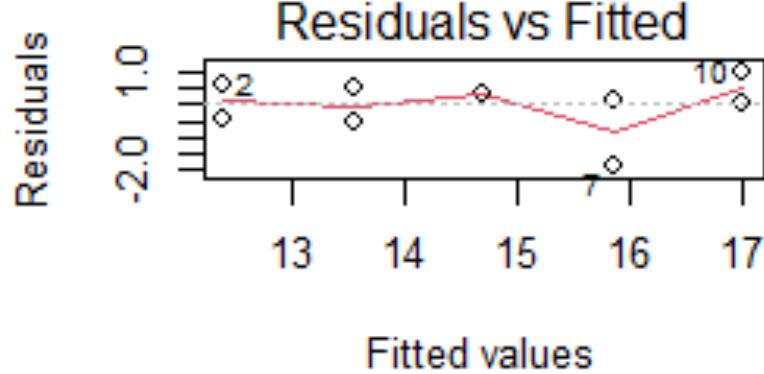
- 1. Independence:** Observations (and ε_i 's) are independent.
- 2. Linear Response:** $E(\varepsilon_i) = 0$ for each x .
 - Scatter plot of Y vs X : should show linear trend.
 - Plot of residuals vs fitted values: should not show a trend.
- 3. Equal Variance:** $\text{Var}(\varepsilon_i) = \sigma^2$ for each x .
 - Plot of residuals vs fitted values: should show equal scatter.
- 4. Normality:** ε_i 's are normally distributed.
 - QQ plot of residuals: should be linear.

You should always look at a scatterplot of the raw data, but diagnostic plots (based on residuals) can be used to check assumptions.

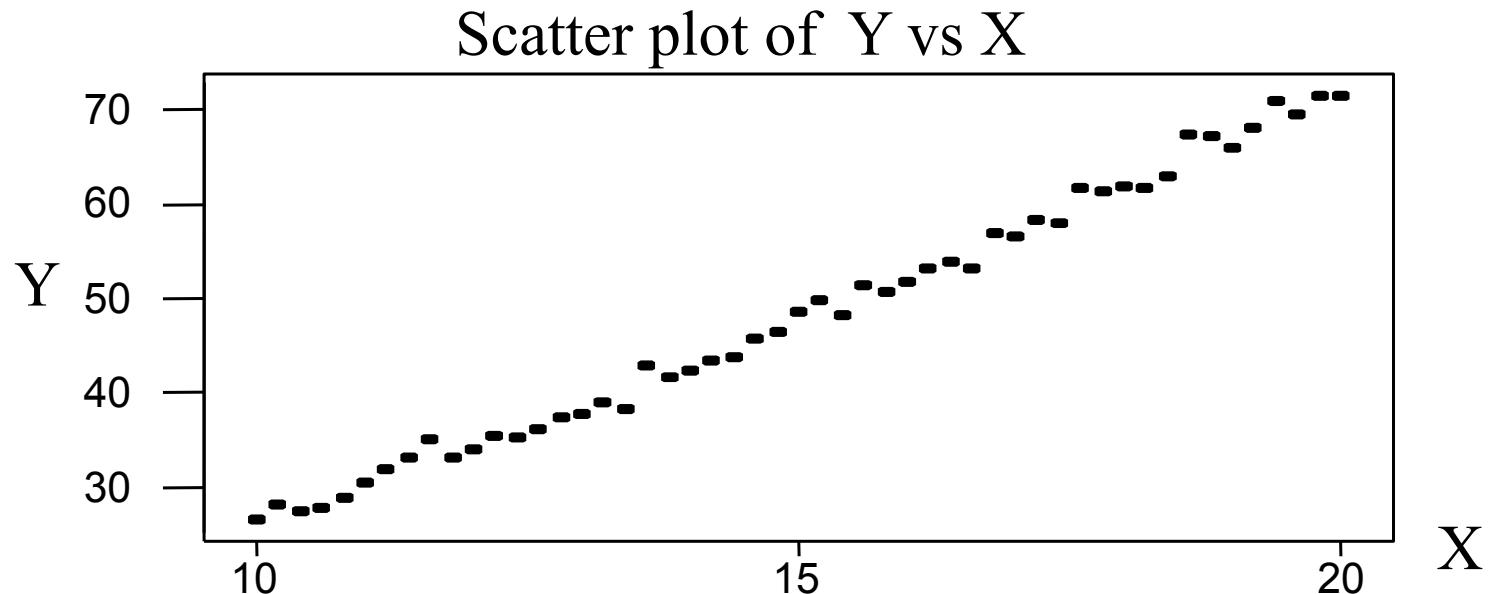
Use `plot(LMFit)` to generate diagnostic plots.

Diagnostic Plots in R

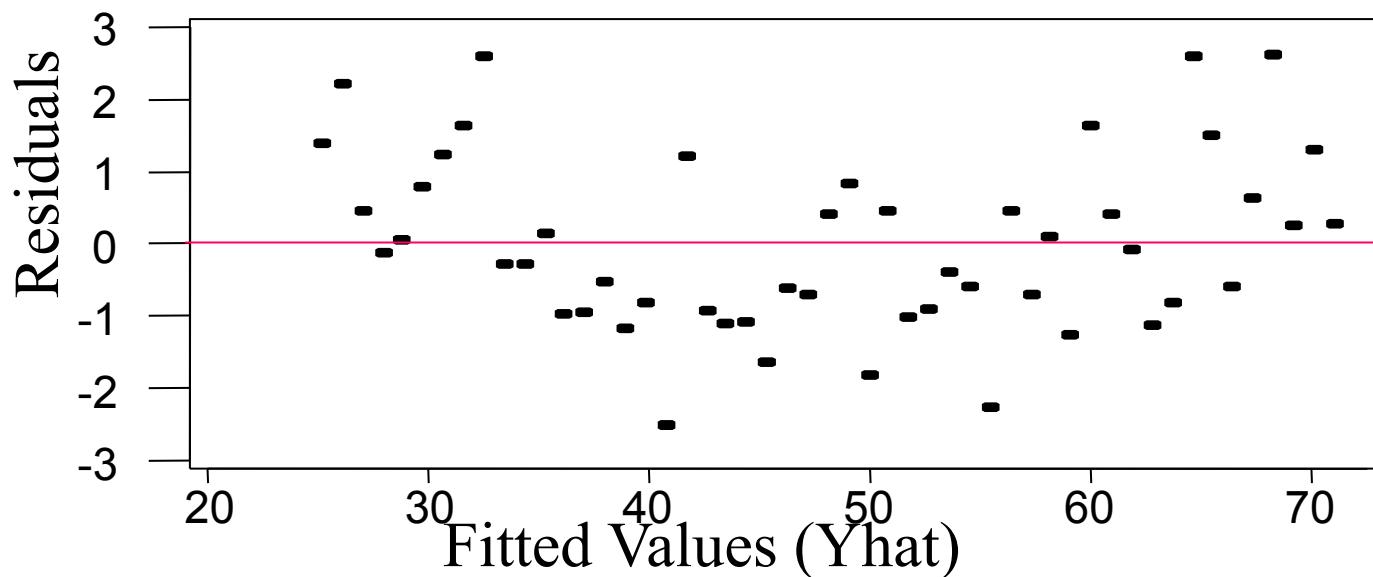
- In R, use `plot(LMFit)`
- For the Corn example:



Checking Regression Assumptions: *Example #1*

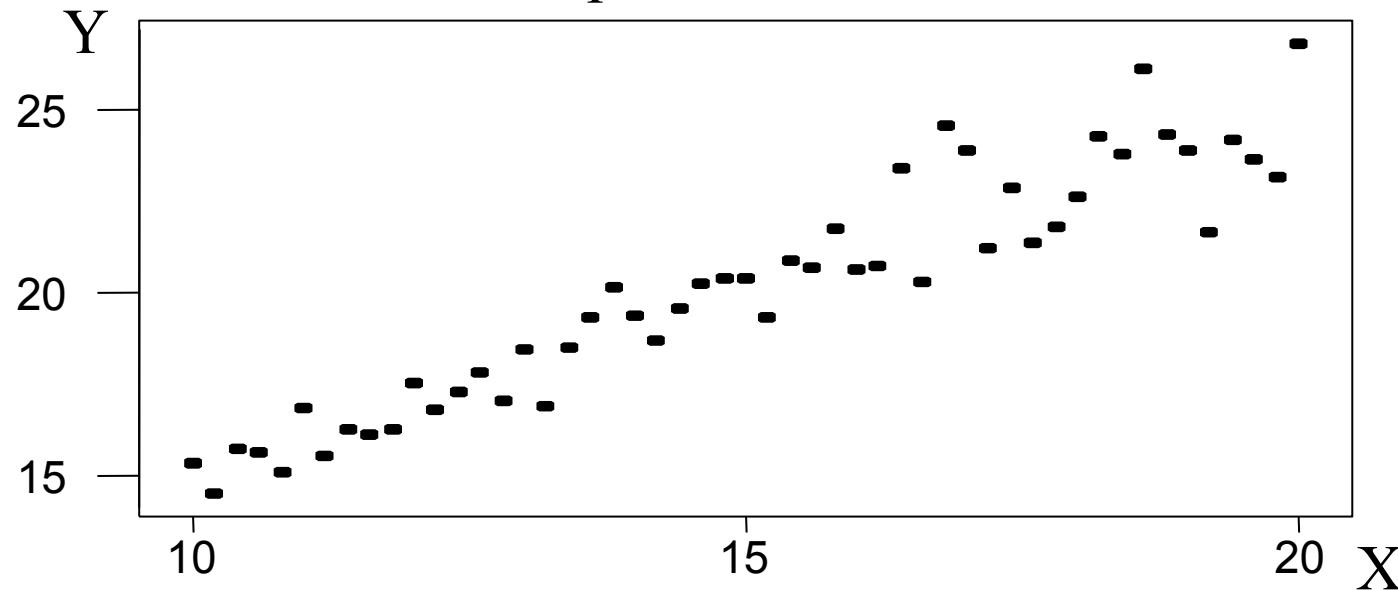


Plot of Residuals vs Fitted Values (Same Data!)

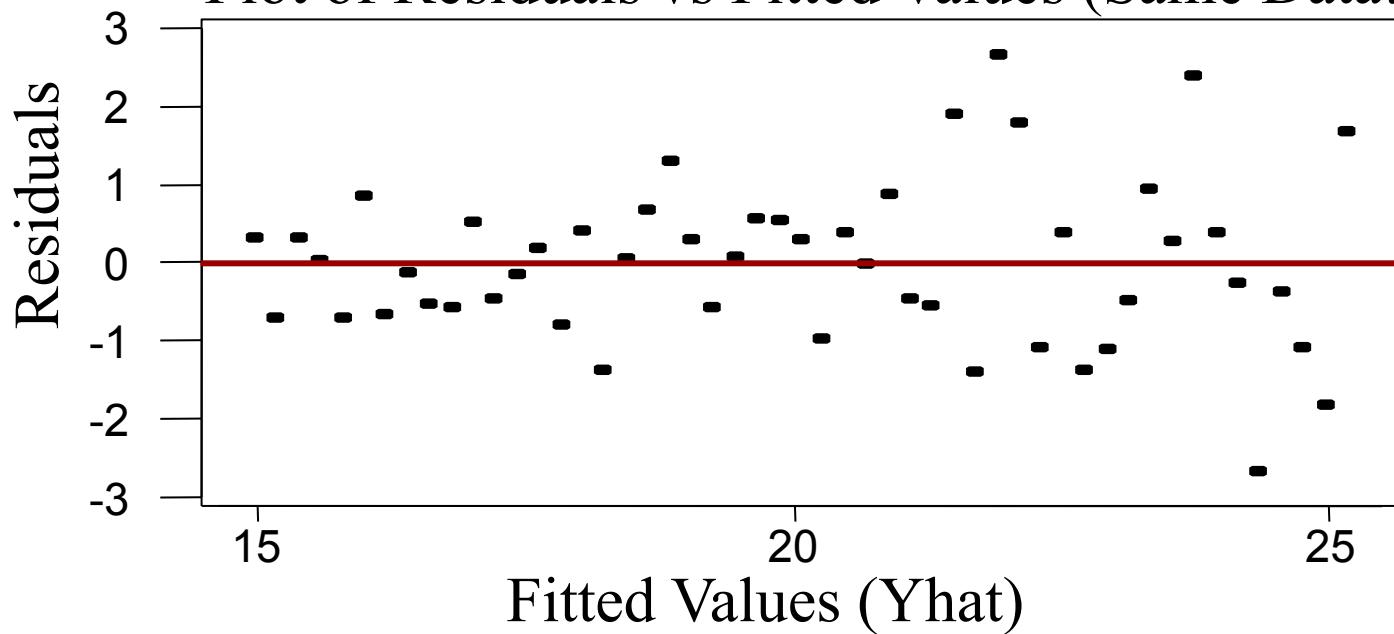


Checking Regression Assumptions: *Example #2*

Scatter plot of Y vs X

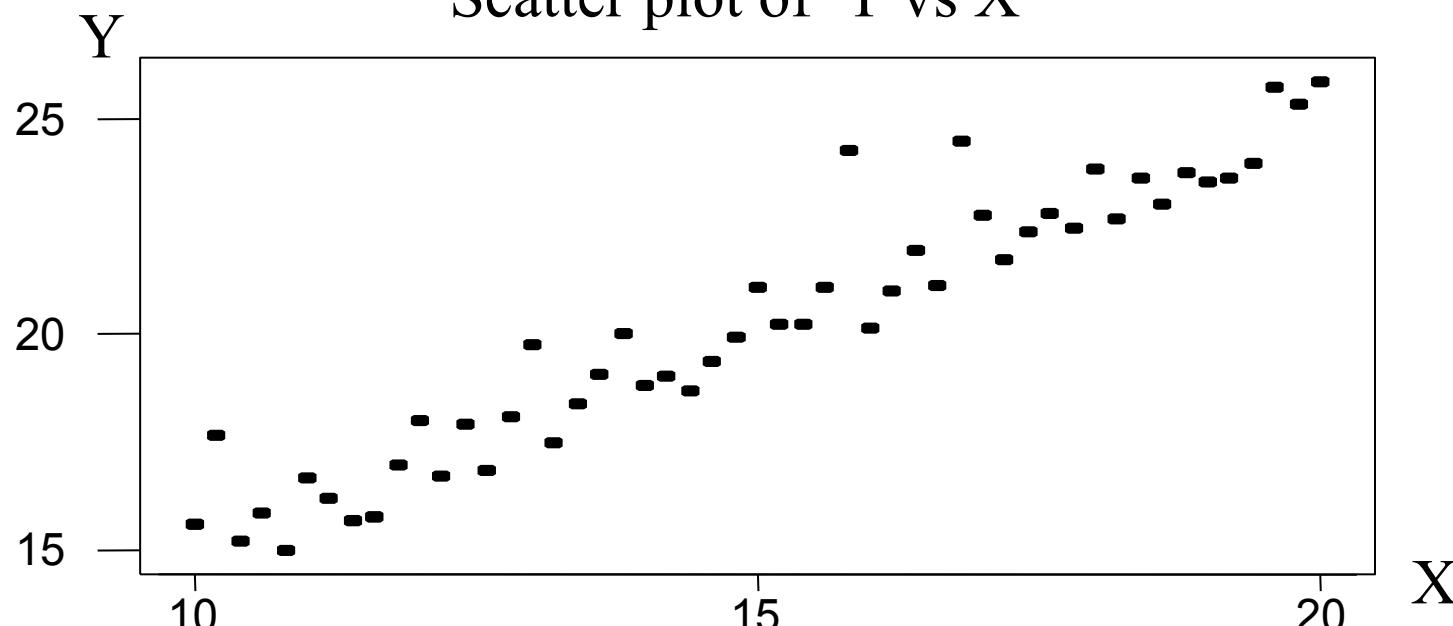


Plot of Residuals vs Fitted Values (Same Data!)

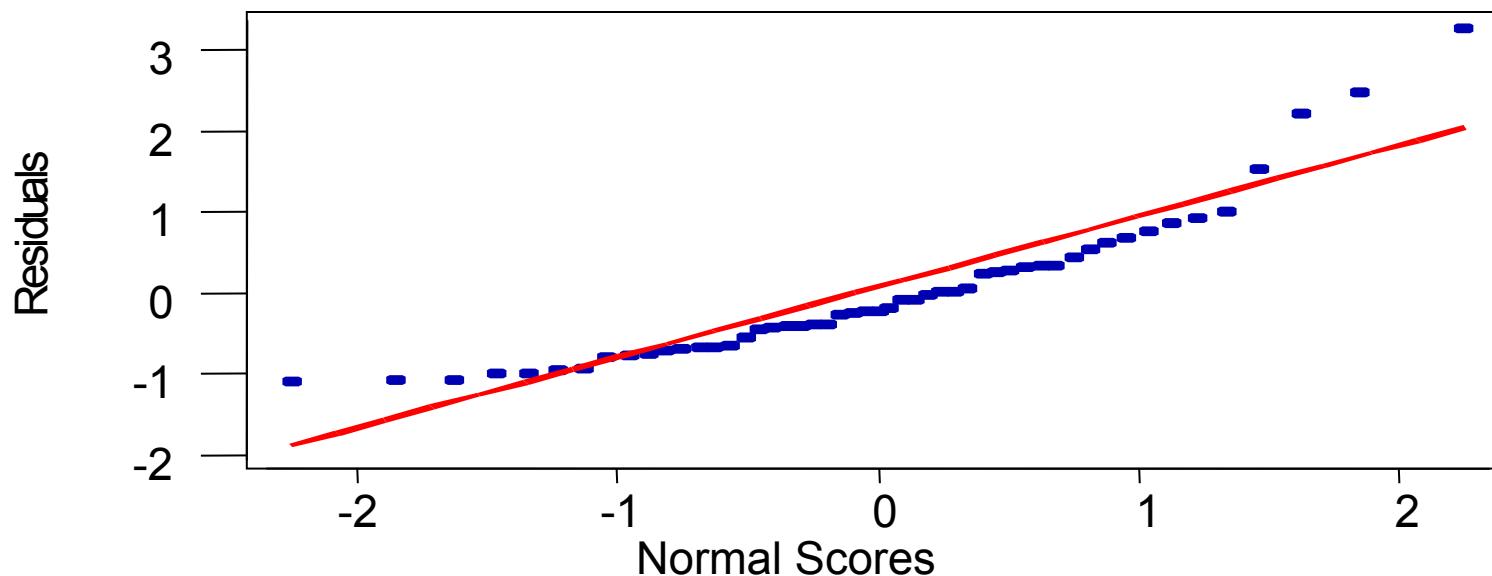


Checking Regression Assumptions: Example #3

Scatter plot of Y vs X



QQPlot of Residuals



5. Remedies for the failure of model assumptions

A. adding quadratic terms

This topic will be discussed further in STAR512.

We mention it here only because it is an approach sometimes used to deal with the failure of model assumptions. When a simple linear regression model does not fit due to curvature(eg. see a quadratic pattern in (1) scatter plot of Y versus X or (2) residuals versus fitted values), one could consider a quadratic regression.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

This model can be fit in R using:

```
lm(y ~ poly(x, 2, raw=TRUE))  
lm(y ~ x + I(x^2))
```

Or just create a new variable:

```
x2 <- x*x  
lm(y ~ x + x2)
```

B. transformations

Transformation is a common approach to help satisfy model assumptions.

The most common reasons to transform are:

1. To achieve linearity
2. To achieve normality and equality of variances.

We can transform the predictor (X), response (Y) or both.

Considerations for transformation include:

1. Whether assumptions are satisfied!
2. Researcher preference. Often, the experimenter does not want to interpret the results in some transformed scale, e.g log(dollars).

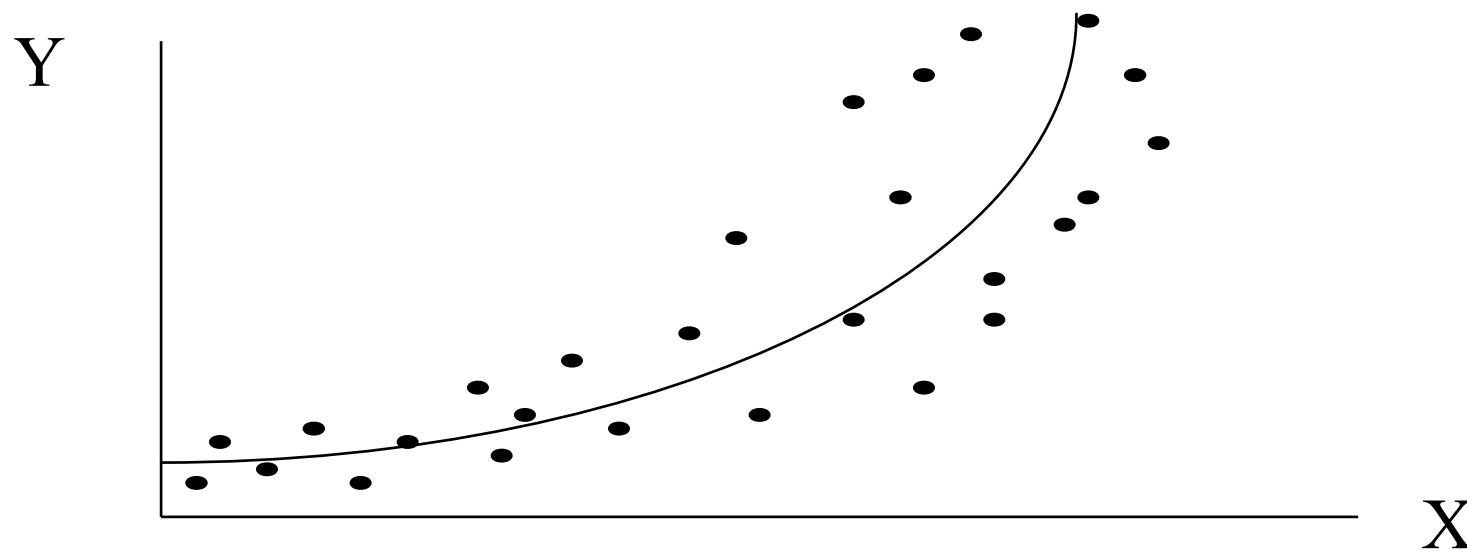
Example: In the plot below, Y looks like an exponential function of X. But we also see evidence of unequal variance.

We have two possible approaches:

1. Transform Y by regressing $\log(Y)$ on X
2. Transform X by regressing Y on e^X

To decide, consider whether one of the approaches above can achieve **both** linearity and equality of variance.

We can also consider non-linear regression (covered in Notes 12.2)



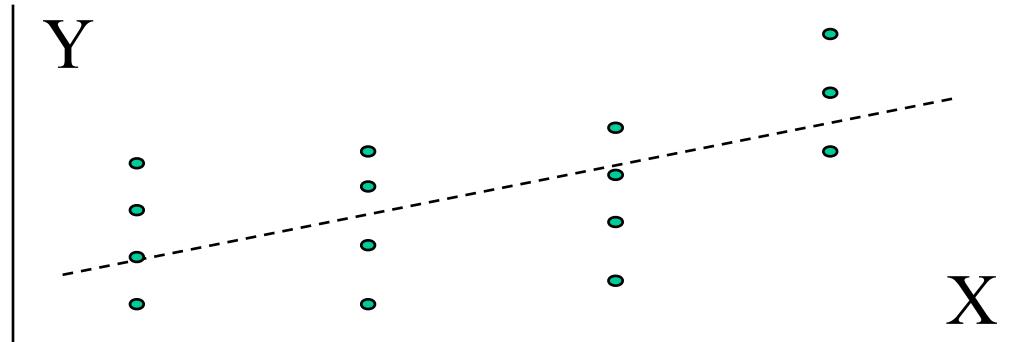
6. An F-test for “lack of fit”

The easiest way to examine possible lack of fit is by plotting (1) scatter plot of Y versus X and (2) residuals versus fitted values. Curvature or other trend in either of these plots might suggest lack of fit.

In this section, we will (briefly) discuss a formal test for lack of fit.

- We consider the case where there are repeated Y observations at the same X value.
- In this case, we can compare two different models for the data:
 - linear regression (2 parameters) V.S one-way ANOVA (t parameters).
 - We compare the estimated σ^2 values from the two model approaches.

The lack of fit test determines whether the group means are farther from the line than they should be if linear regression was appropriate.



An F-test for “lack of fit”

H_0 : The linear regression model is appropriate.

H_a : The linear regression model is not appropriate.

Test Statistic:

$$F = \frac{(\text{SSResid}_{\text{Reg}} - \text{SSResid}_{\text{ANOVA}}) / (\text{dfResid}_{\text{Reg}} - \text{dfResid}_{\text{ANOVA}})}{\text{MSResid}_{\text{ANOVA}}}$$

Reject H_0 if $F > F_{\alpha, df1, df2}$

where $df1 = t - 2 = \text{dfResid}_{\text{Reg}} - \text{dfResid}_{\text{ANOVA}}$, $df2 = \text{dfResid}_{\text{ANOVA}}$

NOTES:

1. Values of X that are very close can be altered to be the same to get groups with repeats.
2. To do the test in R, run the analysis as a regression and a one-way ANOVA. Then compare the models using the `anova()` function. See “**Lack of Fit**” Example.

Corn Example: First fit the regression and one-way ANOVA analyses in R. See “**Lack of Fit**” Example.

$$F = \frac{(\text{SSResid}_{\text{Reg}} - \text{SSResid}_{\text{ANOVA}})(\text{dfResid}_{\text{Reg}} - \text{dfResid}_{\text{ANOVA}})}{\text{MSResid}_{\text{ANOVA}}}$$
$$= \frac{(5.65 - 3.5)/(8 - 5)}{0.70} = 1.0238$$

$$\text{pvalue} = 1 - \text{pf}(1.0238, \text{df1}=3, \text{df2}=5) = 0.4564$$

Conclusion: Fail to Reject H0. There is no evidence of lack of fit for the corn regression.