STAT 110: Week 9

University of Otago

Outline

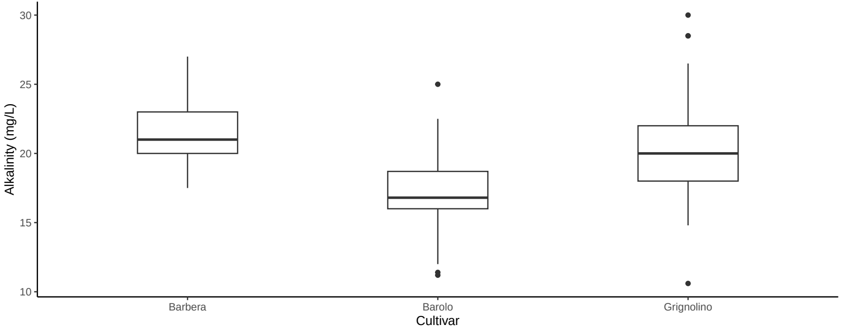
- Fitting ANOVA model
- Understanding ANOVA table
  - Comparing the variance within a group, to the variance between groups
- Look at multiple comparisons
  - Pairwise differences

Recall: chemical composition of Italian wines

- We are looking at alkalinity of the wine (measured in mg/L)
    - Three cultivars: barbera, barolo, grignolino
- Import the data

```
wine = read.csv('wine.csv')
```

- Look at the data



Recall: ANOVA

- One-way ANOVA model with K groups
  - Outcome variable in group j is normally distributed with mean μ_j and variance σ^2
- We want to know how the mean outcome differs among groups
  - Potential problems with multiple comparisons
- Are there any differences in mean outcome among the groups?
- This takes the form of a hypothesis test
  - $\text{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_K$
  - H_A : at least one mean is different

In R

- As with categorical variables with 2 levels
  - Special case of linear regression
  - Categorical variables can be included in R as factors

```
wine$cultivar = as.factor(wine$cultivar)
```

- We can then fit a linear regression model

```
m_wine = lm(alkalinity ~ cultivar, data = wine)
```

- This fits the ANOVA model
- Problem: output from `m_wine` is not in a convenient form
  - Output is in terms of particular pairwise comparisons

In R

- We use the aov function to get the results in more convenient form

```
a_wine_lm = aov(m_wine)
```

- We can also use aov directly

```
a_wine = aov(alkalinity ~ cultivar, data = wine)
```

- The output we will consider is an ANOVA table

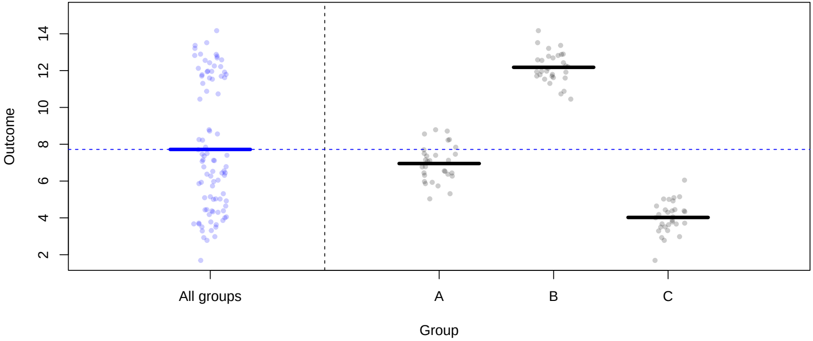
```
summary(a_wine)

##               Df Sum Sq Mean Sq F value  Pr(>F)
## cultivar       2    573     286    35.8 9.4e-14 ***
## Residuals    175   1401       8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

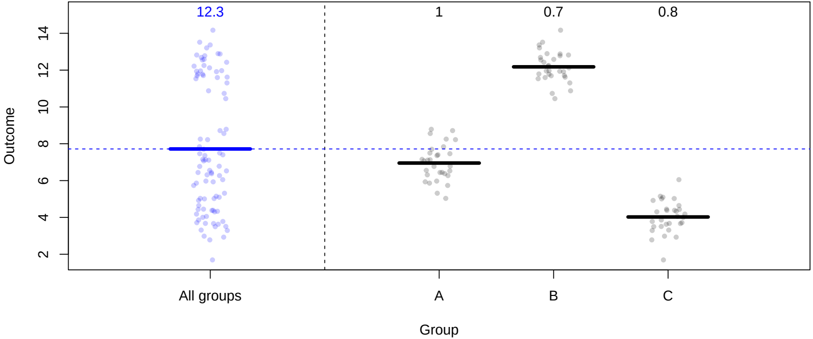
- Take a graphical look at the ANOVA model to help explain what this tells us

Understanding ANOVA (analysis of variance)

- Left plot (blue): plot of all outcome variables (irrespective of group)
- Right three plots (black): plot of outcome variables by group
- Solid horizontal lines: means
  - Dashed blue line is the overall mean

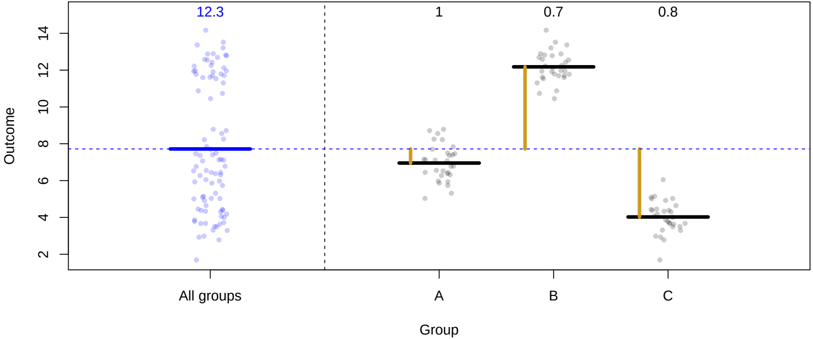


Comparing variance



- The sample variance for each group is given on the plot above
  - Combined data (blue): outcomes are highly variable
  - Data from each group (black; A, B, C): outcomes have much lower variability
- The group variable has explained a lot of the variability in the data

Comparing variance



- Overall variability partitioned into:
  - Variability in group means (indicated by gold lines)
  - Variability within the groups (points around their mean)
- This is the information summarized in the ANOVA table

ANOVA table

- The ANOVA table for the wine data is

```
##             Df Sum Sq Mean Sq F value  Pr(>F)    
## cultivar     2    573     286    35.8 9.4e-14 ***
## Residuals  175   1401       8                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- To explain what this represents we will use the table:

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	$\text{GMS} = \frac{\text{GSS}}{\text{DF}}$	$\text{F} = \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	$\text{RMS} = \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

ANOVA table: rows

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	GMS $= \frac{\text{GSS}}{\text{DF}}$	F $= \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	RMS $= \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

- Group row: describes the variation between group means
  - Variation represented by gold bar in plot above
- Residuals row: describes the variation within each group
- Total row: describes the variation when we combine across groups
  - Data represented in blue in plot above
  - This row is not in R output

ANOVA table: columns

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	GMS $= \frac{\text{GSS}}{\text{DF}}$	F $= \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	RMS $= \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

- Mean Sq[uare]
  - Group (GMS): related to the between-group variance
  - Residual (RMS): estimate of within-group variance
- F value: ratio of group mean square and residual mean square
- Df: degrees of freedom
- Sum Sq: sum of squares
  - Convenient when calculating by hand

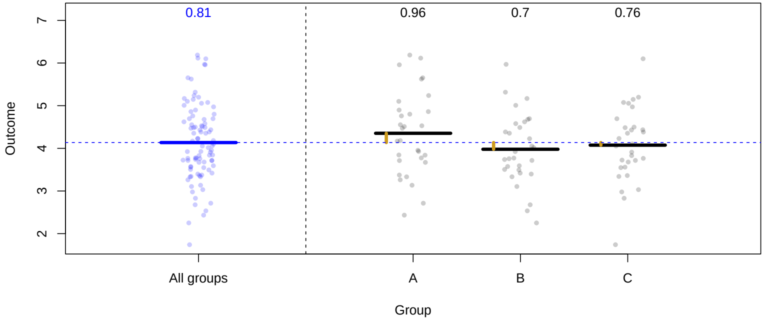
ANOVA table

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	$\text{GMS} = \frac{\text{GSS}}{\text{DF}}$	$\text{F} = \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	$\text{RMS} = \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

- If the groups explain a lot of variability (like our plots above)
  - The group mean square will be large relative to residual mean square
  - F-value will be relatively large
    - ANOVA table below is for data from plots above

```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## group         2   1024     512     635 <2e-16 ***
## Residuals    87     70       1                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example II: group does not explain much variation



- The group mean square will not be large relative to residual mean square
- The F-value is not large

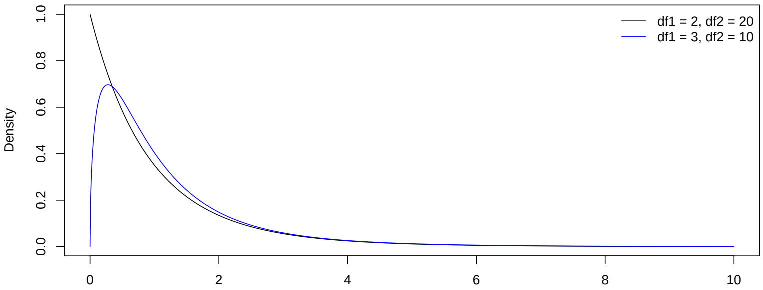
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2    2.3   1.135    1.41   0.25
## Residuals   87   70.1   0.805
```

ANOVA table: F column

- The F-value is comparing the variance among groups (the variability in the group means) to the variance within the groups
  - It is a measure of how much variation in the data is explained by the groups compared to unexplained variation
- If the null hypothesis is true
  - Data come from the ANOVA model with all means equal ($\mu_1 = \mu_2 = \ldots = \mu_k$)
    - The data are normally distributed with the same mean and variance
  - F-statistic will have an F-distribution with Df (group), Df (residual) degrees of freedom
- We can use this to find a p-value
  - Quantify the incompatibility between the data and null hypothesis
  - Are the data unusual given that the null hypothesis is true (group means are the same)
- If null hypothesis is true, we expect an F-value of around 1

Detour: F-distribution

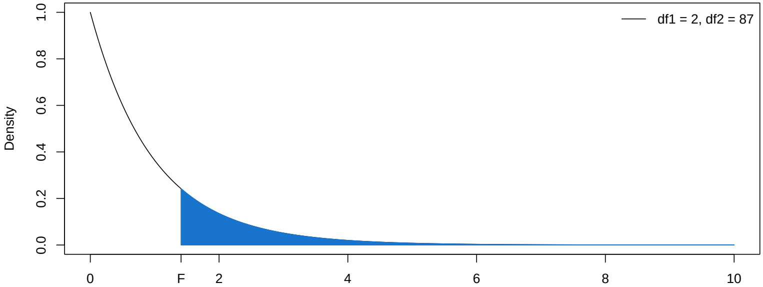
- The F-distribution is a distribution for positive random variables



  - It is asymmetric (positively skewed)
  - It has two parameters:
    - Degrees of freedom for the numerator (df1)
    - Degrees of freedom for the denominator (df2)

Finding a p-value

- An extreme F-value is as large, or larger, than that observed
  - Indicative of groups explaining as much, or more, variation in the data



- The blue area is given by `1-pf(F, df1, df2)`
  - `pf(F, df1, df2)` gives probability of a value less than F

Example II

- The ANOVA table for example II is

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## group         2    2.3   1.135    1.41   0.25
## Residuals    87   70.1   0.805
```

- The observed F-statistic is 1.41
  - df1 is degrees of freedom for group: 2
  - df2 is degrees of freedom for residuals: 87
- The p-value is

```
1-pf(1.41, 2, 87)

## [1] 0.25
```

- In practice: refer to the `Pr(>F)` column in the output

In R: wine data

- The ANOVA table for the wine data is

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## cultivar      2    573     286    35.8 9.4e-14 ***
## Residuals   175   1401       8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The F-value is large, p-value is small
  - p-value $< \alpha$: evidence of incompatibility between data and null hypothesis
  - Data are (highly) unusual if all the means were truly the same
  - Providing evidence that at least one of the means differ
- Which groups have means that appear to differ?

Pairwise comparisons of group means

- To compare each group, there are (potentially) many comparisons
  - If we have $K = 3$ groups: 3 comparisons
  - If we have $K = 5$ groups: 10 comparisons
  - If we have $K = 10$ groups: 45 comparisons
- E.g. for $K = 3$: conduct hypothesis tests or find confidence intervals:
  - CI for $\mu_1 - \mu_2$; hypothesis test with $\text{H}_0 : \mu_1 - \mu_2 = 0$
  - CI for $\mu_1 - \mu_3$; hypothesis test with $\text{H}_0 : \mu_1 - \mu_3 = 0$
  - CI for $\mu_2 - \mu_3$; hypothesis test with $\text{H}_0 : \mu_2 - \mu_3 = 0$

Multiple comparisons

- The problem with multiple tests (or multiple confidence intervals) is that properties no longer hold. For hypothesis testing:
  - α gives the type I error rate for a single test
    - Probability of α of a 'false positive' given that the null hypothesis is true
  - In each test, there is a chance of a false positive (type I error)
  - With multiple tests, the overall chance of a type I error increases
  - Overall type I error rate: referred to as the family-wise error rate
    - Probability of making at least one type I error when performing multiple tests
  - Multiple comparisons increase the family wise error rate
    - e.g. if we perform 10 independent tests with $\alpha = 0.05$, then the probability of at least one type I error is $1 - 0.95^{10} = 0.4$, if the null hypothesis is true in each instance
    - Probability found using complements

Tukey HSD

- Tukey's honest significant difference (HSD) is a multiple comparison approach designed for ANOVA models
- If the sample sizes are the same in each group
  - Family-wise error rate is exactly α
- If the sample sizes are different among groups
  - It is conservative (family-wise error rate is less than α)
- The Tukey approach finds corrected confidence intervals and p-values
- It is easily implemented in R: `TukeyHSD`

In R: wine data

```
TukeyHSD(a_wine)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = alkalinity ~ cultivar, data = wine)
##
## $cultivar
##                      diff   lwr     upr p adj
## Barolo-Barbera     -4.38 -5.68 -3.0792 0.000
## Grignolino-Barbera -1.18 -2.43  0.0712 0.069
## Grignolino-Barolo   3.20  2.02  4.3791 0.000
```

Interpretation: wine data

- Intepret the adjusted confidence intervals, e.g.
  - We are 95% confident that the difference in mean alkalinity between the Grignolino and Barolo cultivars is between 2.02 and 4.38
- Interpret the adjusted p-values, e.g.
  - The p-value for the difference between Grignolino and Barbera cultivars is 0.069.
  - As p-value $> \alpha$ there is no evidence that the observed difference is unusual given the null hypothesis that the two means are the same
  - Note: the uncorrected p-value is 0.027

ANOVA: big picture

- We have looked at fitting one-way a ANOVA model
  - One-way refers to one categorical predictors: cultivar (for wine example)
  - Two-way ANOVA: have two categorical predictors
- There might be many other potential predictors (categorical or continuous)
  - e.g. vineyard, climate (temperature, rainfall), fertilizer used, etc
- Recall: ANOVA is a special case of linear regression
  - We can use multiple linear regression to include these other variables
- There are lots of possible extensions
- There are also lots of ways to get ourselves into trouble
- These more complex models are explored in STAT 210

Summary

- Looked at the ANOVA summary table
  - Group: the variation between group means
  - Residuals: the variation within a group
  - F-value: comparing the variance within a group, to the variance between groups
- F-distribution to find p-value
- Look at multiple comparisons for pairwise differences
  - Tukey's honest significant difference
  - See multiple comparisons in general context later in the course

Outline

- Previous
  - Exploring (normal) models for continuous data
    - Single mean
    - Two independent groups
    - Paired data
    - Multiple independent groups
    - Linear regression
- Today
  - Consider data that are not continuous
  - Explore models for binary data

How well can you putt?

- What is the probability a pro golfer will sink a 6 ft putt?
- Data on professional golfers from 6 feet:
  - 272 attempts, 149 successes

Problem

- We have been working with models for continuous outcome variables
- This is not continuous data
- It is binary data
  - Each observation is yes/no, success/failure, 1/0
  - Each putt will either go in (success), or not (failure)
- Such data arises all the time
  - Will you support candidate X in the next election?
  - Did the chick successfully fledge?
  - Did the participant select option A (or B)?
  - Did the home team win the football match?
- We need a model for binary data
  - Probability distribution for binary data

Bernoulli distribution

- Recall: discrete probability distributions
- Random variable Y with two possible outcomes: success/failure
  - Represent success with 1
  - Represent failure with 0
- These two outcomes have associated probabilities
  - Earlier in semester: we assigned them actual numbers, e.g. 0.6 and 0.4
  - Now: represent the probability of success with an (unknown) parameter: p
- That gives the probability distribution

i	1	2	Total
y_i	0	1	
$\Pr(Y = y_i)$	$1 - p$	p	1

Bernoulli distribution: properties

- Recall: we found means and variances of discrete probability distributions

$$E[Y] = \sum_{i=1}^{k} y_i \Pr(Y = y_i)$$

$$\mathsf{Var}(Y) = \sum_{i=1}^{k} (y_i - E[Y])^2 \Pr(Y = y_i)$$

- Using these we can find the mean and variance of a Bernoulli distribution

$$E[Y] = p$$

$$\mathsf{Var}(Y) = p(1-p)$$

- Extension: Confirm these using the expectation and variance formulae above

Binary to binomial

- We may be interested in cases where there are many binary trials
  - Flip a coin 15 times
  - Record the success/failure of 272 putts
- The number of successes from multiple trials has a binomial distribution, if:

1. The trials are binary
   - The outcome can be represented as success / failure
2. The number of trials n, is fixed
   - e.g. the number of trials does not depend on the number of successes (or failures) you see
3. The trials are independent
   - The outcome of one trial does not affect the outcome of another
4. The probability of success, p, is the same for each trial
   - The probability of success does not change from one trial to another

Binary to binomial

- Let's think about the simplest case
  - Y_1 and Y_2 are two (independent) random variables
  - Each of them has a Bernoulli distribution with probability of success p
- Our interest is in the random variable $X = Y_1 + Y_2$
  - Number of successes from two trials
- If we had two professionals putting from 6 foot
  - X is a random variable that represents how many putts go in

Binomial distribution: $n = 2$

- The probability distribution of $X = Y_1 + Y_2$ is

i	1	2	3	Total
x_i	0	1	2	
$\Pr(X = x_i)$	$(1-p)^2$	$2p(1-p)$	p^2	1

$$\begin{aligned} Pr(X = 0) &= \Pr(Y_1 = 0 \text{ and } Y_2 = 0) \\ &= \Pr(Y_1 = 0)\,\Pr(Y_2 = 0) \qquad \text{multiplication rule: independence} \\ &= (1-p) \times (1-p) \end{aligned}$$

Binomial distribution: $n = 2$

- The probability distribution of $X = Y_1 + Y_2$ is

i	1	2	3	Total
x_i	0	1	2	
$\Pr(X = x_i)$	$(1-p)^2$	$2p(1-p)$	p^2	1

$$\begin{aligned}
\Pr(X = 1) &= \Pr(Y_1 = 1 \text{ and } Y_2 = 0) + \Pr(Y_1 = 0 \text{ and } Y_2 = 1) \\
&= \Pr(Y_1 = 1)\Pr(Y_2 = 0) + \Pr(Y_1 = 0)\Pr(Y_2 = 1) \qquad \text{independence} \\
&= p(1-p) + (1-p)p
\end{aligned}$$

Binomial distribution: general

- In general, the number of successes from n independent Bernoulli trials is:
  - $X = Y_1 + Y_2 + \ldots + Y_n$
- For moderate or large values of n
  - Possible, but extremely tedious, to write out full probability distribution
- We have a shortcut: we can find the probability of x successes from n independent Bernoulli trials

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Binomial distribution: general

- The probability of x successes from n independent Bernoulli trials is

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ is the number of ways to obtain x successes from n trials[1]

- For each of these, the probability of observing those x successes is $p^x(1-p)^{n-x}$
  - E.g. there are two ways to see $x = 1$ success from $n = 2$ trials (see above)
    - Each of those has probability $p(1-p)$
  - E.g. there are 3003 ways to see $x = 5$ successes from $n = 15$ trials
    - Each of these has probability $p^5(1-p)^{10}$

[1] $x! = x \times (x-1) \times \ldots \times 3 \times 2 \times 1$, e.g. $3! = 3 \times 2 \times 1 = 6$. $x!$ is read as x factorial.

Binomial distribution: general

- The probability of x successes from n independent Bernoulli trials is

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- We can use this to find the expectation and variance
  - The mean of a binomial distribution is $E[X] = np$
  - The variance of a binomial distribution $\text{Var}(X) = np(1-p)$
- If there are $n = 100$ putts with probability of success $p = 0.2$, then
  - $E[X] = np = 100 \times 0.2 = 20$
  - $\text{Var}(X) = np(1-p) = 100 \times 0.2 \times 0.8 = 16$
  - $\text{sd}(X) = \sqrt{\text{Var}(X)} = 4$

Binomial probabilities in R

- We don't have to calculate the long form of that equation
  - We can use the R function `dbinom`
- Example: what is $\Pr(X = 1)$ when $p = 0.2$ and $n = 2$

```
dbinom(x = 1, size = 2, prob = 0.2)

## [1] 0.32
```

- The arguments are:
  - `x = 1`: the number of successes x
  - `size = 2`: the number of trials n
  - `prob = 0.2`: the probability of success p
- Check that it gives the correct answer: we know it should be $2p(1 - p)$

```
2*0.2*(1-0.2)

## [1] 0.32
```

More examples

- If we take 15 putts where there is a probability of 0.7 of making the putt
- What is the probability that we make 10 putts?
- We have $x = 10$, $n = 15$, $p = 0.7$

```
dbinom(x = 10, size = 15, prob = 0.7)

## [1] 0.206
```

- What is the probability of making 70 putts out of 100 putts with probability 0.6

```
dbinom(x = 70, size = 100, prob = 0.6)

## [1] 0.01001
```

Back to the data

- We want to estimate the probability of a professional golfer making a 6 foot putt
- What is our statistical model?
  - Each putt is the outcome of an independent Bernoulli trial with probability p
  - Equivalently, the total number of successful putts is binomially distributed
- We want to estimate a parameter (population) with a statistic (sample)
  - (Reasonably) obvious statistic: sample proportion x/n
- For golf data:

$$\hat{p} = \frac{x}{n} = \frac{149}{272} = 0.548$$

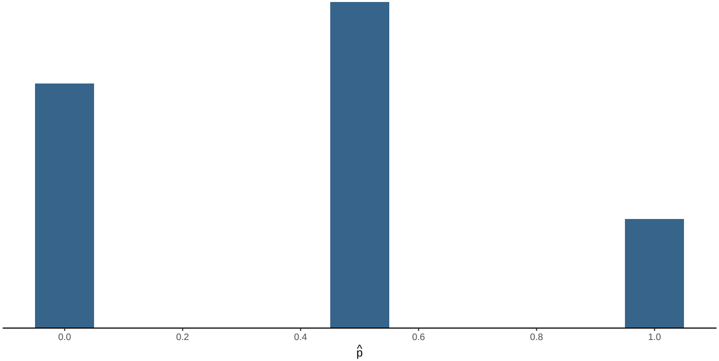
- Recall: $\hat{p}$ is the estimate of parameter p

Confidence interval

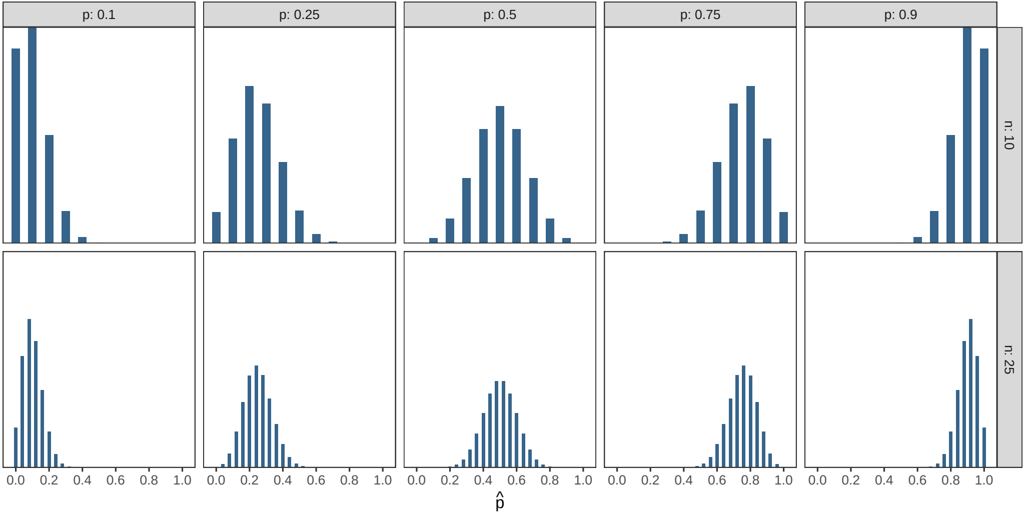
- How do we find a confidence interval?
- Recall: normal model
  - Found the sampling distribution
  - Obtained a confidence interval from the sampling distribution
- Can we do the same thing here?
  - The sampling distribution is the distribution of $\hat{p}$ if we take repeated samples
- Look at it graphically

Sampling distribution for $\hat{p}$: Start small with $n = 2$ and $p = 0.4$

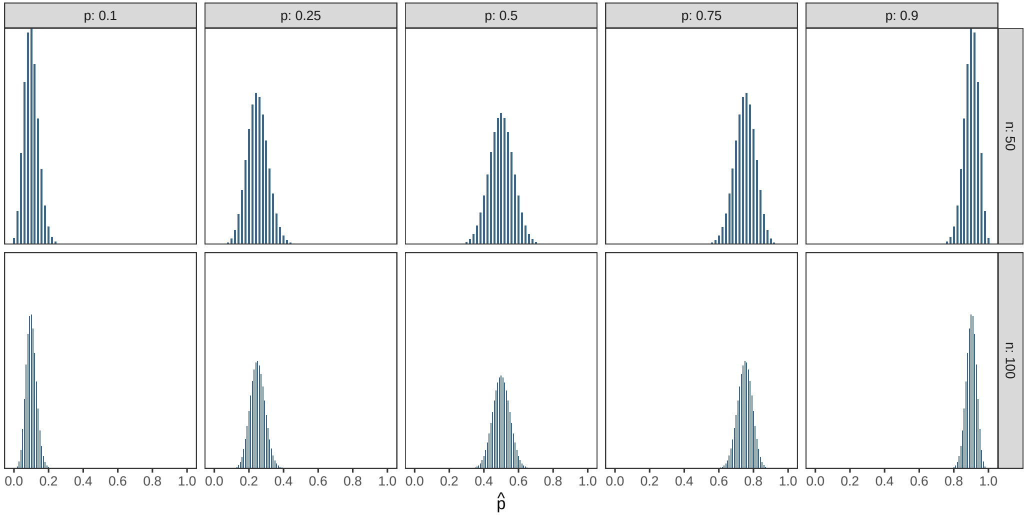
- There are three possibilities:
  - Observe $x = 0$ with probability 0.36: estimate $\hat{p} = 0$
  - Observe $x = 1$ with probability 0.48: estimate $\hat{p} = 0.5$
  - Observe $x = 2$ with probability 0.16: estimate $\hat{p} = 1$



Same principle, but increase the number of trials

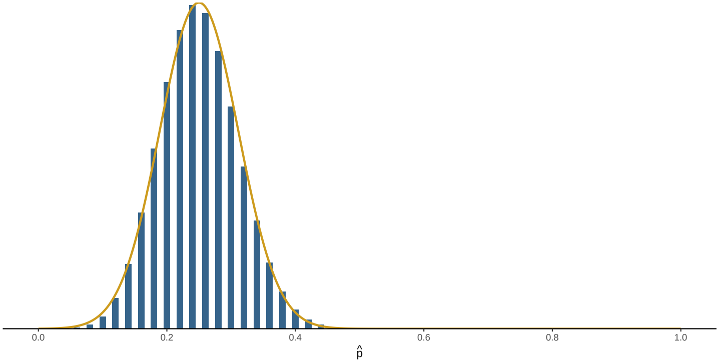


Increase the number of trials some more



Sampling distribution

- As the sample size gets larger, the sampling distribution looks increasingly normal
  - Normal pdf given in gold
- Example: $n = 50$, $p = 0.25$



Sampling distribution

- We can approximate the sampling distribution by a normal distribution
  - Provided n is large enough
- There are various rules of thumb used to determine if the normal approximation is appropriate
- One of these is
  - $np > 10$ and $n(1-p) > 10$
- As we saw on the plots above, this reflects that
  - The sampling distribution is increasingly normal as n increases
  - When p is close to 0 or 1 it takes a larger n for it to approach normality
- In practice we use $n\hat{p}$ and $n(1-\hat{p})$ to check if a normal approximation is reasonable

Sampling distribution

- We can approximate the sampling distribution by a normal distribution
  - Provided n is large enough
- The mean and variance are

$$E[\hat{p}] = p$$

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

- So the standard error: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- Extension: Derive $E[\hat{p}]$ and $\text{Var}(\hat{p})$
  - We have $\hat{P} = \frac{X}{n}$ where $E[X] = np$ and $\text{Var}(X) = np(1-p)$

Confidence interval in R

- We use the normal approximation to find a confidence interval: `prop.test`

```
n = 272; x = 149
prop.test(x, n)

##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 2.3, df = 1, p-value = 0.13
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.48656 0.60766
## sample estimates:
##       p
## 0.54779
```

- We are 95% confident that the probability of a professional golfer making a putt from 6 feet is between 0.487 and 0.608

Hypothesis test

- We can also test the hypothesis
  - $\text{H}_0 : p = p_0$
  - $\text{H}_A : p \neq p_0$
- `prop.test` defaults to $p_0 = 0.5$
  - It can be changed with option `p`, e.g. `p = 0.4`

```r
prop.test(x, n, p = 0.4)
```

- For the putting data with $p_0 = 0.5$ we have a p-value of 0.13
  - This quantifies the incompatibility between the data and null hypothesis
  - Since p-value $> \alpha = 0.05$ there is no evidence that the data are unusual given the null hypothesis is true
    - The data we have observed would not be unusual if professionals truly sank 50% of their putts from 6 feet

Summary

- Introduced binary data
- Bernoulli distribution for binary observations
- The number of successes from multiple binary trials have binomial distribution
  - Several conditions need to be satisfied
- Use a binomial model to find:
  - Confidence interval for p
  - Hypothesis test
    - We will look more into these in the next lecture

Outline

- A closer look at confidence intervals and hypothesis tests for p
- Extending the model
  - Compare probabilities between two (independent) groups
- Difference in proportions: $p_1 - p_2$
  - Confidence interval
  - Hypothesis test

Recall: Golf putting

- What is the probability a professional golfer makes a putt from 6 feet?
  - $n = 272$ putts with $x = 149$ made

```
n = 272; x = 149
prop.test(x, n)

##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 2.3, df = 1, p-value = 0.13
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.48656 0.60766
## sample estimates:
##       p
## 0.54779
```

Finding confidence interval for p

- We found the confidence interval in R
  - We haven't yet described where it comes from (like we normally do)
- It turns out there are many possible confidence intervals for p
  - The `binomCI` package in R gives the choice of 15 (!) different intervals
- Why are there so many many intervals?
  - There are many reasons
  - Most obvious: because the 'standard' confidence interval doesn't work well

Confidence intervals for p

- The 'standard' confidence interval can be written as

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

  - Estimate: $\hat{p}$
  - Multiplier: sampling distribution is approximate normal
    - Multiplier is $z_{1-\alpha/2}$
  - Standard error: $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$
    - Estimate this: $s_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
- Commonly called a Wald interval
- Similar to what we had for μ

Problems with the Wald interval

- The Wald interval is not very reliable, particularly when n not large, and p close to 0 or 1
  - Despite this it is still commonly used and seen in textbooks
- Recall: what is a confidence interval?
  - If we collect multiple datasets with n binary observations from the population of interest and calculate a confidence interval for each:
    - Then 95% of the intervals, on average, should contain the true p ($\alpha = 0.05$)
- The Wald interval does a poor job of this
  - The interval tends to contain the true value (p) less often than it is supposed to
    - e.g. when $n = 50$ and $p = 0.06$ fewer than 81% of intervals will contain the true p
    - Particularly poor when np or $n(1 - p)$ is small

What about the interval that R gives?

- `prop.test` finds the Wilson (score) interval
- Comparing the Wilson interval to the Wald interval:
  - Both are based on a normal approximation to the binomial
  - The Wilson interval is asymmetric
    - It is not found using: estimate ± multiplier × standard error
  - It has improved performance when p is close to 0 or 1
    - It is reasonable to use even if $np < 10$ or $n(1-p) < 10$
  - We will not delve into the detail
    - It is more complicated
    - Extension: more information is provided at this link for those who may be interested
- In practice: use Wilson interval found using `prop.test`

Continuity correction

- By default `prop.test` adopts a continuity correction
  - For confidence intervals and hypothesis tests
- A continuity correction is adjustment that reflects that we are approximating a discrete distribution (binomial) with a continuous distribution (normal)
  - We make an adjustment of ± 0.5
- If X is a random variable with a binomial distribution, and Z is a random variable with a normal distribution that approximates X, a continuity correction is
  - $\Pr(X \leq 10) \approx \Pr(Z < 10.5)$
  - $\Pr(X \geq 5) \approx \Pr(Z > 4.5)$
- It is conservative: makes confidence intervals wider (increases p-value)
- It can be turned off using option `correct = FALSE`
  - We will use the default settings in `prop.test`

What about the hypothesis test?

- We may wish to test the hypotheses:
  - $\text{H}_0 : p = p_0$
  - $\text{H}_A : p \neq p_0$
- A test statistic can be found using:

$$z = \frac{\text{estimate} - \text{null}}{\text{standard error}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Two things to note:
  - Find standard error assuming null hypothesis is true: $\sigma_{\hat{p}} = \sqrt{\dfrac{p_0(1-p_0)}{n}}$
  - Find p-value from a (standard) normal distribution
    - That's why the test statistic is z, not t

Hypothesis test: golf

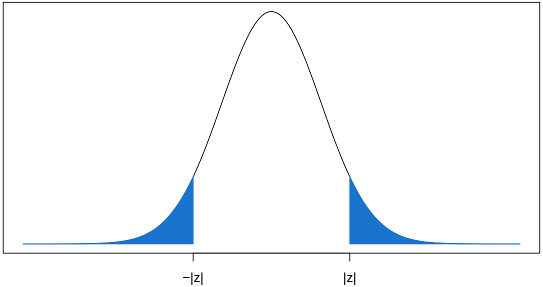
- To test if putting probability is different from $50/50$: $p_0 = 0.5$

```
# estimate of p
phat = x/n
p0 = 0.5
# Find standard error under H0
se = sqrt(p0*(1-p0)/n)
# Find test statistic
z = (phat - p0)/se
# Find pvalue
pval = 2*pnorm(-abs(z))
pval

## [1] 0.115
```

```
z

## [1] 1.58
```



Hypothesis testing in R

- `prop.test` conducts the hypothesis test in a slightly different way
  - By default it uses a continuity correction
  - Uses χ^2 test statistic[2] rather than z
    - Performing the same test, but in a different way
    - Details are outside the scope of the course (see STAT 270)
  - If the correction was turned off (`correct = FALSE`)
    - Obtain an identical p-value to our procedure above
  - Alternatively, we could include a continuity correction in our p-value calculation
    - We would find an identical p-value to that from `prop.test`
    - Details outside the scope of the course

[2] χ is the greek letter chi, pronounced kai (rhymes with sky).

Data: Smallpox in Boston

- Data are 6224 observations from individuals in Boston in 1721 who were exposed to smallpox[3]
  - Inoculated: yes or no
  - Result: lived or died
- We are interested in comparing the probability of death for those who were inoculated to those who were not

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

[3] This is the same data that we saw in week 2.

Models for binomial data

- We don't have the tools to answer the question
  - We only know how to estimate p, not compare p across two groups
- We can look at model extensions for binomial data that parallel those we explored for normal models, e.g.
  - Comparing two or more independent groups
  - Regression-type models: probability of success depends on predictor variables
    - Called logistic regression
  - Defer many of these extensions to later courses (i.e. STAT 210)
- For smallpox data: two independent binomials
  - Inoculated: modelled as binomial with probability p_1
    - $x_1 = 238$, $n_1 = 244$
  - Not inoculated: modelled as binomial with probability p_2
    - $x_2 = 5136$, $n_2 = 5980$

Big picture

- We want to compare the survival between inoculated and uninoculated
- There are multiple ways we could do this, e.g.
  - Difference in probabilities: $p_1 - p_2$
  - Ratio of probabilities (also called relative risk): p_1/p_2
- We will focus on $p_1 - p_2$
- It is straightfoward to estimate this difference
  - $\hat{p}_1 - \hat{p}_2$
- We also know those estimates are uncertain
  - Found from data (a sample from the population)
  - Find a confidence interval

Confidence interval for $p_1 - p_2$

- Find a confidence interval using

$$\text{estimate} \pm \text{multiplier} \times \text{standard error}$$

- Estimate: $\hat{p}_1 - \hat{p}_2$
- Multiplier: we again approximate the sampling distribution with normal
  - Multiplier is $z_{1-\alpha/2}$
- Standard error: $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$
  - Estimate this with: $s_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Wald confidence interval for $p_1 - p_2$

- Putting this together we have the $100(1-\alpha)\%$ Wald confidence interval:

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- This is the interval returned by `prop.test` when we have two groups
- As with the Wald interval for p
  - The interval is not that reliable if either n_1 or n_2 is small and either p_1 or p_2 is close to 0 or 1
  - Improved confidence intervals do exist
    - e.g. the Newcombe interval is based on Wilson interval
    - Such intervals can be found in other R packages
- We will use the Wald interval in `prop.test`

In R

```
x = c(238, 5136); n = c(244, 5980) # smallpox data
prop.test(x, n)

## 
##  2-sample test for equality of proportions with continuity correction
## 
## data:  x out of n
## X-squared = 26, df = 1, p-value = 3e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0931 0.1400
## sample estimates:
## prop 1 prop 2 
##  0.975  0.859
```

- We are 95% confident that the probability of survival was between 0.093 and 0.14 higher for those who were inoculated compared to those who were not

Hypothesis test

- Both p_1 and p_2 are conditional probabilities
  - p_1 is the survival probability given inoculated
  - p_2 is the survival probability given not inoculated
- If $p_1 = p_2$ then survival does not depend on inoculation
  - Survival and inoculation are independent
- We can test the hypotheses:
  - $\text{H}_0 : p_1 - p_2 = 0$ (this is equivalent to $p_1 = p_2$)
  - $\text{H}_A : p_1 - p_2 \neq 0$ (this is equivalent to $p_1 \neq p_2$)

Hypothesis test

- A test statistic can be found using:

$$z = \frac{\text{estimate} - \text{null}}{\text{standard error}}$$

- Estimate is $\hat{p}_1 - \hat{p}_2$
- Null value is 0
- We need the standard error assuming null hypothesis is true
  - The two groups have the same probability: $p_1 = p_2$
  - The null hypothesis doesn't specify what this value is
    - Let's call it p^*

Hypothesis test

- The standard error is: $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}}$
  - This is the standard error above evaluated at $p_1 = p_2 = p^*$
- We don't know p^*
  - Estimate it: $\hat{p}^* = \frac{\text{total success}}{\text{total trials}} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$
  - $\hat{p}^*$ is sometimes call the pooled proportion
- Use this to estimate the standard error: $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n_1} + \frac{\hat{p}^*(1-\hat{p}^*)}{n_2}}$
- This hypothesis test is found using `prop.test`. As with the test for p:
  - It uses a different test statistic (χ^2 vs z)
  - Includes a continuity correct by default

Hypothesis test: in R

- Using prop.test to find the p-value

```
prop.test(x,n)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 26, df = 1, p-value = 3e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0931 0.1400
## sample estimates:
## prop 1 prop 2
##  0.975  0.859
```

Interpretation

- The p-value quantifies the incompatibility between the null hypothesis and the data
    - The p-value $< \alpha = 0.05$, which suggests the data are unusual if the two groups (inoculated and uninoculated) truly had the same probability of survival

Summary

- Look at estimating p
  - Confidence intervals:
    - Wald interval can be unreliable
    - `prop.test` using more reliable alternative
  - Hypothesis tests
- Explored comparison between two groups: $p_1 - p_2$
  - Confidence intervals
  - Hypothesis test