MATH 3339 - 03 15951
Statistics for the Sciences
Inferences on Proportions and Regression

Wendy Wang
wwang60@central.uh.edu

Lecture 34 - 3339
Outline

1. Confidence Interval for the Difference of Two Proportions

2. Inference for the Regression Parameters
Popper Set Up

- Fill in all of the proper bubbles.
- Make sure your ID number is correct.
- Make sure the filled in circles are very dark.
- This is popper number 30.
Comparing Two Proportions

What is the difference between the proportion of m&ms that are blue in the plain m&ms compared to the peanut m&ms?

From a random sample of plain m&ms and peanut m&ms we get the following results.

<table>
<thead>
<tr>
<th>Candy type</th>
<th>n</th>
<th>Number of Blue</th>
<th>Sample proportion ((\hat{p}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>plain</td>
<td>81</td>
<td>28</td>
<td>(\hat{p}_{\text{plain}} = \frac{28}{81} = 0.3458)</td>
</tr>
<tr>
<td>peanut</td>
<td>100</td>
<td>20</td>
<td>(\hat{p}_{\text{peanut}} = \frac{20}{100} = 0.2)</td>
</tr>
</tbody>
</table>

We want to know what is the difference of the proportion of m&ms that are blue for all of plain and peanut m&ms. That is, estimate:

\[
p_{\text{plain}} - p_{\text{peanut}}
\]
Two-sample problems assumptions

The goal of inference is to compare the responses in two groups.

1. Each group is considered to be a \textbf{simple random sample} from two \textbf{distinct} populations.
2. The population sizes are both \textbf{at least ten times} the sizes of the samples.
3. The number of successes and failures in \textbf{both} samples must all be $\geq 10$. 
Confidence intervals for comparing two proportions

Choose an SRS of $n_1$ from a large population having proportion $p_1$ of successes and an independent SRS of size $n_2$ from another population having proportion $p_2$ of successes.

1. Point estimate: $D = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$

2. Confidence level: $C$ a percent predetermined in the problem if not use 95%.

3. Critical value: $z^*$ is the value for the standard Normal density curve with area $C$ between $-z^*$ and $z^*$.

4. Confidence interval:

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

5. Interpret
Determine a 95% confidence interval for the difference of the proportion of m&ms that are blue for all of plain and peanut m&ms.

From a random sample of plain m&ms and peanut m&ms we get the following results.

<table>
<thead>
<tr>
<th>Candy type</th>
<th>n</th>
<th>Number of Blue</th>
<th>Sample proportion ($\hat{p}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>plain</td>
<td>81</td>
<td>$X_1 = 28$</td>
<td>$\hat{p}_{\text{plain}} = \frac{28}{81} = 0.3458$</td>
</tr>
<tr>
<td>peanut</td>
<td>100</td>
<td>$X_2 = 20$</td>
<td>$\hat{p}_{\text{peanut}} = \frac{20}{100} = 0.2$</td>
</tr>
</tbody>
</table>

\[ C = 95\% = 0.95 \]

\[
(0.3458 - 0.2) \pm \text{qnorm}(\frac{1.95}{2}) \sqrt{\frac{\hat{p}_{\text{plain}}(1-\hat{p}_{\text{plain}})}{n_1} + \frac{\hat{p}_{\text{peanut}}(1-\hat{p}_{\text{peanut}})}{n_2}}
\]

\[
(0.01578, 0.2756)
\]

Our 95% confidence interval for $p_1 - p_2$ is $(0.01578, 0.2756)$. 
R code

```r
prop.test(x=c(28, 20), n=c(81, 100), conf.level = 0.95, correct=FALSE)
```

2-sample test for equality of proportions without continuity correction

data:  c(28, 20) out of c(81, 100)
X-squared = 4.8738, df = 1, p-value = 0.02727
alternative hypothesis: two.sided
95 percent confidence interval: 
0.01578192 0.27557610
sample estimates:
prop 1  prop 2
0.345679  0.200000
Assumptions for Two-Sample Proportion Test

1. Both samples must be independent SRSs from the populations of interest.

2. The population sizes are both at least ten times the sizes of the samples.

3. The number of successes and failures in both sample must all be at least 10.

Test statistic:

\[ z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \]
Left-handedness

Is the proportion of left-handed students higher in honors classes than in academic classes? Two hundred academic and one hundred honors students from grades 6 - 12 were selected throughout a school district and their left handedness was recorded. The sample information is:

<table>
<thead>
<tr>
<th>Sample</th>
<th>Honors</th>
<th>Academic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Number of left-handed students</td>
<td>18</td>
<td>32</td>
</tr>
</tbody>
</table>

Is there sufficient evidence at the 1% significance level to conclude that the proportion of left-handed students is greater in honor classes?

\[ H_0: p_1 = p_2 \quad Ha: p_1 > p_2 \Rightarrow \text{Right-tailed test} \]

\[ \alpha = 1\% = 0.01 \]

\[ Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{\frac{18}{100} - \frac{32}{200}}{\sqrt{\frac{0.18(0.82)}{100} + \frac{0.16(0.84)}}{200}} = 0.433 \]
\[ P\text{-value} = P( z > 0.4331) \]
\[ = 1 - \text{pnorm}(0.4331) \]
\[ = 0.3333 > \alpha = 0.01 \]

Fail to reject \( H_0 \)

There is no evidence that the proportion of left-handed students in the honors class is higher in the honors class.
9.3-9.4: Inference for the Regression Parameters
Least-Squares Regression

- The **least-squares regression line (LSRL)** of $Y$ on $X$ is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

- The linear regression model is: $Y = \beta_0 + \beta_1 x + \epsilon$
  - $Y$ is dependent variable (response).
  - $x$ is the independent variable (explanatory).
  - $\beta_0$ is the population intercept of the line.
  - $\beta_1$ is the population slope of the line.
  - $\epsilon$ is the error term which is assumed to have mean value 0. This is a random variable that incorporates all variation in the dependent variable due to factors other than $x$.
  - The variability: $\sigma$ of the response $y$ about this line. More precisely, $\sigma$ is the standard deviation of the deviations of the errors, $\epsilon_i$ in the regression model.

- We will gather information from a sample so we will have the least squares estimates model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. 
Least-Squares Regression

Formulas:

\[
\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x
\]

\[
\hat{\beta}_1 = \text{cor}(x, y) \cdot \frac{S_y}{S_x}
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]
$R^2$ is the percent (fraction) of variability in the response variable ($Y$) that is explained by the least-squares regression with the explanatory variable.

- This is a measure of how successful the regression equation was in predicting the response variable.

- The closer $R^2$ is to one (100%) the better our equation is at predicting the response variable.

- We will look later at how this is calculated.

- In the R output it is the **Multiple R-squared** value.
Is this good at predicting the response?

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line.

\[
\text{residual} = \text{observed } y - \text{predicted } y
\]

- We can determine residuals for each observation.
- The closer the residuals are to zero, the better we are at predicting the response variable.
- We can plot the residuals for each observation, these are called the residual plots.
Residual Plot
Examining a residual plot

- A **curved pattern** shows that the relationship is not linear.

- **Increasing spread** about the zero line as $x$ increases indicates the prediction of $y$ will be less accurate for larger $x$. **Decreasing spread** about the zero line as $x$ increases indicates the prediction of $y$ to be more accurate for larger $x$.

- Individual points with larger residuals are considered outliers in the vertical ($y$) direction.

- Individual points that are extreme in the $x$ direction are considered outliers for the $x$-variable.
Example 2

The following data on $x =$ frying time (sec) of tortilla chips and $y =$ moisture content (%) of tortilla chips.

<table>
<thead>
<tr>
<th>$x$</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>45</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>16.3</td>
<td>9.7</td>
<td>8.1</td>
<td>4.2</td>
<td>3.4</td>
<td>2.9</td>
<td>1.9</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Show the residual plot.

```r
> fry = c(5,10,15,20,25,30,45,60)
> moisture = c(16.3,9.7,8.1,4.2,3.4,2.9,1.9,1.3)
>
> chips.lm = lm(moisture ~ fry)
> plot(resid(chips.lm), xlab = "fry", main = "Residual Plot of Chips")
> abline(h=0)
```
Residual Plot of Chips

```
> library(ggplot2)
> data(chips)
> chips$lfit <- predict(lm(weight ~ size, data = chips))
> chips$resid <- residuals(lm(weight ~ size, data = chips))
> ggplot(chips, aes(size, resid)) + geom_line() + geom_point()
```

- `resid(chips.lm)` represents the residuals from the linear model of weight vs. size.
- `fry` is the variable on the x-axis.
- The y-axis represents the residuals.
Estimating the Regression Parameters

In the simple linear regression setting, we use the slope \( b_1 \) and intercept \( b_0 \) of the least-squares regression line to estimate the slope \( \beta_1 \) and intercept \( \beta_0 \) of the population regression line.

The standard deviation, \( \sigma \), in the model is estimated by the regression standard error

\[
s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\sum \text{all residuals}^2}{n - 2}}
\]

Recall that \( y_i \) is the observed value from the data set and \( \hat{y}_i \) is the predicted value from the equation.

In R, \( s \) is the called the **Residual Standard Error** in the last paragraph of the summary.
Determining if the Model is Good

- For the sample we can use $R^2$ and the residuals to determine if the equation is a good way of predicting the response variable.

- Another way to determine if this equation is a good way of predicting the response variable is to determine if the explanatory variable is needed (significant) in the equation.

- These tests of significance and confidence intervals in regression analysis are based on assumptions about the error term $\epsilon$. 
Assumptions about the error term $\epsilon$

1. The error term $\epsilon$ is a random variable with a mean or expected value of zero, that is $E(\epsilon) = 0$, an estimate for $\epsilon$ is the residuals for each value of the X-variable.

   $$\text{residual} = \text{observed } y - \text{predicted } y$$

2. The variance of $\epsilon$, denoted by $\sigma^2$, is the same for all values of $x$. The estimate for $\sigma^2$ is

   $$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}.$$  

3. The values of $\epsilon$ are independent.

4. The error term $\epsilon$ is a normally distributed random variable.

5. The residual plots help us assess the fit of a regression line and determine if the assumptions are met.
Definitions of Regression Output

1. The **error sum of squares**, denoted by $SSE$ is
   \[ SSE = \sum (y_i - \hat{y}_i)^2 \]

2. A quantitative measure of the total amount of variation in observed values is given by the **total sum of squares**, denoted by $SST$.
   \[ SST = \sum (y_i - \bar{y})^2 \]

3. The **regression sum of squares**, denoted $SSR$ is the amount of total variation that *is* explained by the model
   \[ SSR = \sum (\hat{y}_i - \bar{y})^2 \]

4. The **coefficient of determination**, $r^2$ is given by
   \[ r^2 = \frac{SSR}{SST} \]
> anova(shelf.lm)
Analysis of Variance Table

Response: sold

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>space</td>
<td>1</td>
<td>20535</td>
<td>20535</td>
<td>21.639</td>
<td>0.0009057 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>10</td>
<td>9490</td>
<td>949</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ 1

Wendy Wang  wwang60@central.uh.edu
Conditions for regression inference

- The sample is an SRS from the population.
- There is a linear relationship in the population.
- The standard deviation of the responses about the population line is the same for all values of the explanatory variable.
- The response varies Normally about the population regression line.
t Test for Significance of $\beta_1$

- **Hypothesis**
  
  \[ H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0 \]

- **Test statistic**
  
  \[ t = \frac{\text{observed} - \text{hypothesized}}{\text{standard deviation of observed}} \]
  
  observed = $b_1$
  
  hypothesized = 0

  \[ \text{standard error} = SE_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \]

  With degrees of freedom $df = n - 2$.

- **P-value**: based on a $t$ distribution with $n - 2$ degrees of freedom.

- **Decision**: Reject $H_0$ if $p$-value $\leq \alpha$.

- **Conclusion**: If $H_0$ is rejected we conclude that the explanatory variable $x$ can be used to predict the response variable $y$. 

Wendy Wang  wwang60@central.uh.edu
Testing $\beta_1$

1. We want to test: $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ for the coffee sales.

2. Test statistic: $t = \frac{(7.4-0)}{1.591} = 4.652$

3. $P$-value: $2 \times P(T > 4.652) = 0.000906$

4. Decision: Reject the Null hypothesis

5. Conclusion: $\beta_1$ is significantly not zero, thus shelf space can be used to predict the number of units sold.
> shelf.lm=lm(sold~space)
> summary(shelf.lm)

Call:
  lm(formula = sold ~ space)

Residuals:
   Min     1Q   Median     3Q    Max
-42.00  -26.75    5.50  21.75   41.00

Coefficients:
             Estimate Std. Error t_value Pr(>|t|)
(Intercept)  145.000     21.783   6.657  5.66e-05 ***
space        7.400      1.591   4.652 0.000906 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.81 on 10 degrees of freedom
Multiple R-squared:  0.6839, Adjusted R-squared:  0.6523
F-statistic: 21.64 on 1 and 10 DF,  p-value: 0.000906
Because elderly people may have difficulty standing to have their heights measured, a study looked at predicting overall height from height to the knee. Here are data (in centimeters, cm) for five elderly men:

<table>
<thead>
<tr>
<th>Knee Height (cm)</th>
<th>57.7</th>
<th>47.4</th>
<th>43.5</th>
<th>44.8</th>
<th>55.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Height(cm)</td>
<td>192.1</td>
<td>153.3</td>
<td>146.4</td>
<td>162.7</td>
<td>169.1</td>
</tr>
</tbody>
</table>

1. Test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

2. Give a conclusion for the relationship between using knee length to predict overall height.
Confidence Intervals for $\beta_1$

If we want to know a range of possible values for the slope we can use a confidence interval.

- Remember confidence intervals are  

  \[
  \text{estimate} \pm t^* \times \text{standard error of the estimate}
  \]

- Confidence interval for $\beta_1$ is  

  \[
  b_1 \pm t_{\alpha/2, n-2} \times SE_{b_1}
  \]

- Where $t^*$ is from table D with degrees of freedom $n - 2$ where $n =$ number of observations.

- In R we can get this by `confint(name.lm, level = 0.95)`.

```r
> confint(shelf.lm)
2.5 %   97.5 %
(Intercept)  96.464405 193.53560
space           3.855461 10.94454
```
Inferences Concerning $\hat{\mu}_Y$

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ where $x^*$ is some fixed value of $x$. Then,

1. The mean value of $\hat{Y}$ is

$$E(\hat{Y}) = \beta_0 + \beta_1 x^*$$

2. The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$$

3. $\hat{Y}$ has a normal distribution.

4. The $100(1 - \alpha)\%$ confidence interval for $\mu_Y$ that is the expected value of $Y$ for a specific value of $x^*$, is

$$\hat{\mu}_Y(x^*) \pm t_{\alpha/2, n-2} \sqrt{\sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)}$$
```r
> predict(shelf.lm, newdata=data.frame(space=12), interval="c", level = 0.9)
fit  lwr    upr
1 233.8 217.6177 249.9823
```