1 Research Project - Data

2 Synthesis
   - Overview
   - One numerical variable
   - Numerical vs. categorical
   - One categorical variable
   - Two categorical variables
Public Bundle

**research-quality data sets**

This is a bundle that gathers public data sets that might be interesting to researchers in a variety of fields in one place.

**Lending Club Loan Data**

[www.lendingclub.com/info/download-data.action](http://www.lendingclub.com/info/download-data.action)

Added by [hmason](http://bitly.com/bundles/hmason/1) 2 years ago

*This page contains downloadable statistics data.*

**Curators**

[hmason](http://bitly.com/bundles/hmason/1)

**What's a bundle?**

[hmason](http://bitly.com/bundles/hmason/1) used bitly to curate this fascinating collection of links.

Like what you see? Create your own
The City of Chicago is on Github: Street center lines in Chicago, Building footprints in Chicago, Bike routes in Chicago, Pedway routes in Chicago, Bike rack locations in Chicago (thecangelog.com)

USDA Food Nutrient Database as MySQL export. Ready-made Shelf Data (github.com)

Looking for good ideas for a MATLAB programming assignment using a large data set. (self.datasets)

Free "low-end" ZIP database and reference material on how its compiled and used (greatdata.com)

Does anyone have access to "CountryData" or the PRS Group datasets?? Would be eternally grateful (details inside) (prsgroup.com)

Anonymising data (r-bloggers.com)

FDI losses (caused by Arab Spring / other Political Violence & Conflict) (self.datasets)

[Request] Princeton UNCTAD Bilateral FDI dataset (self.datasets)
The Data and Story Library

DASL (pronounced "dazzle") is an online library of datafiles and stories that illustrate the use of basic statistics methods. We hope to provide data from a wide variety of topics so that statistics teachers can find real-world examples that will be interesting to their students. Use DASL’s powerful search engine to locate the story or datafile of interest.
Data Sets: A List in Flux

Posted on November 7, 2012 by andy

One person likes this. Be the first of your friends.

After my Pinterest post, I got a little bit hooked, mostly because I realized that it was a visual way for me to see my bookmarks. This makes it easier for me to find the information I am looking for quickly. One problem is that it requires an image, so I quickly realized that the links for data sets wouldn’t work so well on Pinterest.
Getting Data in R

In lab you have been given nicely formated data that can be directly loaded into R using either load or source. This will rarely happen outside of this class, and for your project you will need to convert your data into a format R can read.

- Ideally use a plaintext format: csv, tab delimited, etc. (read.csv, read.table, read.delim)
- Avoid proprietary formats (usually doable but require extra work)
- Programs like Excel are useful to convert and clean up data
- Find your data early, if you run into trouble ask for help
Research Project - Data

Synthesis

Overview
One numerical variable
Numerical vs. categorical
One categorical variable
Two categorical variables
Inference

Design of studies

Bayesian inference

Frequentist inference (CLT & simulation)

Modeling (numerical response)

Exploratory data analysis

Probability

numerical

- one mean & median
- two means & medians
- many means

categorical

- one proportion
- two proportions
- many proportions

1 explanatory

many explanatory
# Exploratory data analysis

<table>
<thead>
<tr>
<th>Variable(s)</th>
<th>Visualization</th>
<th>Summary statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>Bar plot</td>
<td>Frequency table, relative frequency table, proportion</td>
</tr>
<tr>
<td>Numerical</td>
<td>Dot plot, histogram, box plot</td>
<td>Mean, median, SD, z-score, range, IQR, five number summary (min, Q1, median, Q3, max)</td>
</tr>
<tr>
<td>Categorical vs. categorical</td>
<td>Segmented bar plot, mosaicplot</td>
<td>Contingency table, difference in proportions</td>
</tr>
<tr>
<td>Numerical vs. categorical</td>
<td>Side-by-side box plots</td>
<td>Statistics by group, difference in means</td>
</tr>
<tr>
<td>Numerical vs. numerical</td>
<td>Scatterplot</td>
<td>Correlation</td>
</tr>
</tbody>
</table>
Inference - past and future

- one categorical variable
  - proportion: $Z$ if $np \geq 10$ and $n(1 - p) \geq 10$
Inference - past and future

- one categorical variable
  - proportion: $Z$ if $np \geq 10$ and $n(1 - p) \geq 10$
- one numerical variable:
  - mean: $Z$ if $n \geq 30$, $T$ if $n < 30$
Inference - past and future

- one categorical variable
  - proportion: $Z$ if $np \geq 10$ and $n(1 - p) \geq 10$

- one numerical variable:
  - mean: $Z$ if $n \geq 30$, $T$ if $n < 30$

- categorical vs. categorical variable
  - categorical variables have 2 levels - difference of proportions
    - $Z$ if $n_1p \geq 10$, $n_1(1 - p) \geq 10$, $n_2p \geq 10$, $n_2(1 - p) \geq 10$
  - categorical variables more than 2 levels - $\chi^2$ test
Inference - past and future

- one categorical variable
  - proportion: $Z$ if $np \geq 10$ and $n(1 - p) \geq 10$

- one numerical variable:
  - mean: $Z$ if $n \geq 30$, $T$ if $n < 30$

- categorical vs. categorical variable
  - categorical variables have 2 levels - difference of proportions
    - $Z$ if $n_1 p \geq 10$, $n_1 (1 - p) \geq 10$, $n_2 p \geq 10$, $n_2 (1 - p) \geq 10$
  - categorical variables more than 2 levels - $\chi^2$ test

- numerical vs. categorical variable:
  - categorical variable has 2 levels - difference of means
    - $Z$ if $n \geq 30$, $T$ if $n < 30$
  - categorical variable has more than 2 levels - ANOVA
Inference - past and future

- one categorical variable
  - proportion: $Z$ if $np \geq 10$ and $n(1 - p) \geq 10$
- one numerical variable:
  - mean: $Z$ if $n \geq 30$, $T$ if $n < 30$
- categorical vs. categorical variable
  - categorical variables have 2 levels - difference of proportions
    - $Z$ if $n_1p \geq 10$, $n_1(1 - p) \geq 10$, $n_2p \geq 10$, $n_2(1 - p) \geq 10$
  - categorical variables more than 2 levels - $\chi^2$ test
- numerical vs. categorical variable:
  - categorical variable has 2 levels - difference of means
    - $Z$ if $n \geq 30$, $T$ if $n < 30$
  - categorical variable has more than 2 levels - ANOVA
- numerical vs. numerical variable - Regression
Inference - past and future

- one categorical variable
  - proportion: $Z$ if $np \geq 10$ and $n(1 - p) \geq 10$
- one numerical variable:
  - mean: $Z$ if $n \geq 30$, $T$ if $n < 30$
- categorical vs. categorical variable
  - categorical variables have 2 levels - difference of proportions
    - $Z$ if $n_1 p \geq 10$, $n_1 (1 - p) \geq 10$, $n_2 p \geq 10$, $n_2 (1 - p) \geq 10$
  - categorical variables more than 2 levels - $\chi^2$ test
- numerical vs. categorical variable:
  - categorical variable has 2 levels - difference of means
    - $Z$ if $n \geq 30$, $T$ if $n < 30$
    - categorical variable has more than 2 levels - ANOVA
- numerical vs. numerical variable - Regression
- numerical vs. many numerical and categorical variables - Multiple Regression
Example - Salmon contamination

Researchers tested 150 randomly selected farm raised salmon for organic contaminants. They found the mean concentration of carcinogenic insecticide mirex to be 0.0913 ppm, with standard deviation 0.0495 ppm. As a safety recommendation to recreational fishers, the EPA’s recommended screening value for mirex is 0.08 ppm. Are farmed salmon contaminated beyond the level permitted by the EPA?
Salmon - set-up

\[ n = 150, \ \bar{x} = 0.0913, \ s = 0.0495 \]

- cases:
Salmon - set-up

\[ n = 150, \, \bar{x} = 0.0913, \, s = 0.0495 \]

- cases: 150 salmon
- variable(s):
Salmon - set-up

\[ n = 150, \bar{x} = 0.0913, \ s = 0.0495 \]

- cases: 150 salmon
- variable(s): only 1 - concentration of contaminant (numerical)
- parameter of interest:
Salmon - set-up

\[ n = 150, \ \bar{x} = 0.0913, \ s = 0.0495 \]

- cases: 150 salmon
- variable(s): only 1 - concentration of contaminant (numerical)
- parameter of interest: \( \mu \), true mean concentration of contaminant in all farmed salmon
- point estimate:
Salmon - set-up

$n = 150, \bar{x} = 0.0913, s = 0.0495$

- cases: 150 salmon
- variable(s): only 1 - concentration of contaminant (numerical)
- parameter of interest: $\mu$, true mean concentration of contaminant in all farmed salmon
- point estimate: $\bar{x}$, mean concentration of contaminant in sampled farmed salmon
- test:
Salmon - set-up

\[ n = 150, \bar{x} = 0.0913, s = 0.0495 \]

- cases: 150 salmon
- variable(s): only 1 - concentration of contaminant (numerical)
- parameter of interest: \( \mu \), true mean concentration of contaminant in all farmed salmon
- point estimate: \( \bar{x} \), mean concentration of contaminant in sampled farmed salmon
- test: test of population mean against a null value, \( Z \) test since \( n > 30 \), (\( T \) test will yield similar results)
  - independence: random sample, 150 < 10% of all farmed salmon → concentration of contaminant in one salmon in the sample is independent of another
  - sample size/skew: the distribution isn’t extremely skewed, and \( n > 30 \)
- hypotheses:
Salmon - set-up

\[ n = 150, \ \bar{x} = 0.0913, \ s = 0.0495 \]

- cases: 150 salmon
- variable(s): only 1 - concentration of contaminant (numerical)
- parameter of interest: \( \mu \), true mean concentration of contaminant in all farmed salmon
- point estimate: \( \bar{x} \), mean concentration of contaminant in sampled farmed salmon
- test: test of population mean against a null value, \( Z \) test since \( n > 30 \), (\( T \) test will yield similar results)
  - independence: random sample, 150 < 10\% of all farmed salmon \( \rightarrow \) concentration of contaminant in one salmon in the sample is independent of another
  - sample size/skew: the distribution isn’t extremely skewed, and \( n > 30 \)
- hypotheses: \( H_0 : \mu = 0.08; H_A : \mu > 0.08 \)
Salmon - p-value

- $H_A : \mu > 0.08 \rightarrow \text{one-tailed}$
Salmon - p-value

- $H_A : \mu > 0.08 \rightarrow$ one-tailed
- p-value: $P($obtaining a random sample of 150 farmed salmon with an average concentration of contaminant $\geq 0.0913$ ppm, if the true average concentration is $0.08$ ppm$)$
Salmon - p-value

- $H_A : \mu > 0.08 \rightarrow$ one-tailed
- p-value: $P$(obtaining a random sample of 150 farmed salmon with an average concentration of contaminant $\geq 0.0913$ ppm, if the true average concentration is 0.08 ppm)
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score, which requires
  - an observed value (sample mean = 0.0913),
  - a true mean (null value that we assume to be true = 0.08), and
  - a measure of variability (SE = 0.004).
Salmon - p-value

- $H_A : \mu > 0.08 \rightarrow$ one-tailed
- p-value: $P(\text{obtaining a random sample of 150 farmed salmon with an average concentration of contaminant} \geq 0.0913 \text{ ppm, if the true average concentration is 0.08 ppm})$
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score, which requires:
  - an observed value (sample mean = 0.0913),
  - a true mean (null value that we assume to be true = 0.08), and
  - a measure of variability (SE = 0.004).
Salmon - p-value

- $H_A : \mu > 0.08 \rightarrow$ one-tailed
- p-value: $P($obtaining a random sample of 150 farmed salmon with an average concentration of contaminant $\geq 0.0913$ ppm, if the true average concentration is 0.08 ppm$)$
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score, which requires
  - an observed value (sample mean = 0.0913),
  - a true mean (null value that we assume to be true = 0.08), and
  - a measure of variability ($SE = 0.004$).

\[
Z = \frac{0.0913 - 0.08}{0.004} = 2.83
\]
Salmon - p-value

- $H_A : \mu > 0.08 \rightarrow$ one-tailed
- p-value: $P(\text{obtaining a random sample of 150 farmed salmon with an average concentration of contaminant} \geq 0.0913 \text{ ppm, if the true average concentration is 0.08 ppm})$
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score, which requires
  - an observed value (sample mean = 0.0913),
  - a true mean (null value that we assume to be true = 0.08), and
  - a measure of variability (SE = 0.004).

\[
Z = \frac{0.0913 - 0.08}{0.004} = 2.83
\]

\[
p - \text{value} = P(Z > 2.83) = 0.0023
\]
Salmon - p-value (using R)

```r
inference(salmon[, "Mirex"], est = "mean", type = "ht",
    method = "theoretical", null = 0.08,
    alternative = "greater")
```

```
## Single mean
## Summary statistics: mean = 0.0913 ; sd = 0.0495 ; n = 150
## H0: mu = 0.08
## HA: mu > 0.08
## Standard error = 0.004
## Test statistic: Z = 2.804
## p-value: 0.0025
```

*Results slightly different due to rounding.*
Salmon - confidence interval

- Confidence level: 95%
Salmon - confidence interval

- Confidence level: 95%
- Theoretical: Using a critical value based on the normal distribution ($z^*$):

$$\text{point estimate} \pm ME \quad \Rightarrow \text{point estimate} \pm z^* \times SE$$
Salmon - confidence interval

- Confidence level: 95%
- Theoretical: Using a critical value based on the normal distribution ($z^*$):

\[
\text{point estimate } \pm ME \\
\equiv \text{point estimate } \pm z^* \times SE
\]

\[
0.0913 \pm 1.96 \times 0.004 \approx 0.0913 \pm 0.008
\]
Salmon - confidence interval

- Confidence level: 95%
- Theoretical: Using a critical value based on the normal distribution ($z^*$):

\[
\text{point estimate } \pm ME \\
= \text{point estimate } \pm z^* \times SE
\]

\[
0.0913 \pm 1.96 \times 0.004 \approx 0.0913 \pm 0.008 \\
= (0.0833, 0.0993)
\]
Salmon - confidence interval (using R)

```
inference(salmon[, "Mirex"], est = "mean", type = "ci",
         method = "theoretical")
```

```r
## Single mean
## Summary statistics: mean = 0.0913 ; sd = 0.0495 ; n = 150
## Standard error = 0.004
## 95 % Confidence interval = ( 0.0834 , 0.0993 )
```

*Results are slightly different due to rounding.*
Example - Reading

An educator believes that new reading activities for elementary school children will improve reading comprehension scores. She randomly assigns third graders to an 8-week program in which some will use these activities (18 students) and others will experience traditional teaching methods (20 students). At the end of the experiment, both groups can take a reading comprehension exam. The distributions of their scores are shown below. Is there a difference between the average reading scores of students subjected to new and traditional methods?
Reading - set-up

- cases:
Reading - set-up

- cases: 38 students
- variable(s):
Reading - set-up

- cases: 38 students
- variable(s): (1) score - numerical, (2) treatment - categorical
- parameter of interest:
Reading - set-up

- cases: 38 students
- variable(s): (1) score - numerical, (2) treatment - categorical
- parameter of interest: $\mu_{new} - \mu_{trad}$, difference between average scores of all students subjected to new and traditional methods
- point estimate:
Reading - set-up

- cases: 38 students
- variable(s): (1) score - numerical, (2) treatment - categorical
- parameter of interest: $\mu_{new} - \mu_{trad}$, difference between average scores of all students subjected to new and traditional methods
- point estimate: $\bar{x}_{new} - \bar{x}_{trad}$, difference between average scores of students in this study subjected to new and traditional methods
- test:
Reading - set-up

- cases: 38 students
- variable(s): (1) score - numerical, (2) treatment - categorical
- parameter of interest: $\mu_{new} - \mu_{trad}$, difference between average scores of all students subjected to new and traditional methods
- point estimate: $\bar{x}_{new} - \bar{x}_{trad}$, difference between average scores of students in this study subjected to new and traditional methods
- test: compare two population means of independent groups (unpaired)
- hypotheses:
cases: 38 students

variable(s): (1) score - numerical, (2) treatment - categorical

parameter of interest: \( \mu_{new} - \mu_{trad} \), difference between average scores of all students subjected to new and traditional methods

point estimate: \( \bar{x}_{new} - \bar{x}_{trad} \), difference between average scores of students in this study subjected to new and traditional methods

test: compare two population means of independent groups (unpaired)

hypotheses: (two-tailed)
\[
H_0 : \mu_{new} = \mu_{trad} \\
H_A : \mu_{new} \neq \mu_{trad}
\]

test statistic:
Reading - set-up

- cases: 38 students
- variable(s): (1) score - numerical, (2) treatment - categorical
- parameter of interest: $\mu_{new} - \mu_{trad}$, difference between average scores of all students subjected to new and traditional methods
- point estimate: $\bar{x}_{new} - \bar{x}_{trad}$, difference between average scores of students in this study subjected to new and traditional methods
- test: compare two population means of independent groups (unpaired)
- hypotheses: (two-tailed)
  \[ H_0 : \mu_{new} = \mu_{trad} \]
  \[ H_A : \mu_{new} \neq \mu_{trad} \]
- test statistic: $T$ since $n_1 < 30$ and $n_2 < 30$
Test statistic: $T = 2.254$

Degrees of freedom: 17

<table>
<thead>
<tr>
<th>one tail</th>
<th>0.100</th>
<th>0.050</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>two tails</td>
<td>0.200</td>
<td>0.100</td>
<td><strong>0.050</strong></td>
<td><strong>0.020</strong></td>
<td>0.010</td>
</tr>
<tr>
<td>df 16</td>
<td>1.34</td>
<td>1.75</td>
<td>2.12</td>
<td>2.58</td>
<td>2.92</td>
</tr>
<tr>
<td>17</td>
<td>1.33</td>
<td>1.74</td>
<td><strong>2.11</strong></td>
<td><strong>2.57</strong></td>
<td>2.90</td>
</tr>
<tr>
<td>18</td>
<td>1.33</td>
<td>1.73</td>
<td>2.10</td>
<td>2.55</td>
<td>2.88</td>
</tr>
<tr>
<td>19</td>
<td>1.33</td>
<td>1.73</td>
<td>2.09</td>
<td>2.54</td>
<td>2.86</td>
</tr>
<tr>
<td>20</td>
<td>1.33</td>
<td>1.72</td>
<td>2.09</td>
<td>2.53</td>
<td>2.85</td>
</tr>
</tbody>
</table>

p-value: between 0.01 and 0.05
Reading - p-value (using R)

```
inference(reading[,"score"], group = reading[,"treatment"],
est = "mean", type = "ht", method = "theoretical",
null = 0, alternative = "twosided")
```

## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_new = 18, mean_new = 16.7778, sd_new = 6.6999
## n_trad = 20, mean_trad = 11.55, sd_trad = 7.5983
## Observed difference between means (new-trad) = 5.2278
##
## H0: mu_new - mu_trad = 0
## HA: mu_new - mu_trad != 0
## Standard error = 2.32
## Test statistic: T = 2.254
## Degrees of freedom: 2.24
## p-value = 0.0378
Confidence level: 95%
Confidence level: 95%

Theoretical: Using a critical value based on the t distribution \( (t^*) \):

\[
\text{point estimate} \pm ME = \text{point estimate} \pm t^* \times SE
\]
Confidence level: 95%

Theoretical: Using a critical value based on the t distribution ($t^*$):

$$\text{point estimate} \pm ME = \text{point estimate} \pm t^* \times SE$$

$$5.23 \pm 2.11 \times 2.32 \approx 5.23 \pm 4.90$$
Confidence level: 95%

Theoretical: Using a critical value based on the t distribution ($t^*$):

$$\text{point estimate} \pm ME = \text{point estimate} \pm t^* \times SE$$

$$5.23 \pm 2.11 \times 2.32 \approx 5.23 \pm 4.90 = (0.33, 10.13)$$
inference(reading[,"score"], group = reading[,"treatment"],
est = "mean", type = "ci", method = "theoretical")

## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_new = 18, mean_new = 16.7778, sd_new = 6.6999
## n_trad = 20, mean_trad = 11.55, sd_trad = 7.5983
## Observed difference between means (new-trad) = 5.2278
##
## Standard error = 2.3196
## 95 % Confidence interval = ( 0.3339 , 10.1217 )
Example - Breast Cancer

A survey was conducted in the US of 10,000 women with ages between 50-54 whose mothers had been diagnosed with breast cancer at some point in their life. 400 of the survey respondents reported themselves as also having had breast cancer. If the prevalence rate of breast cancer for US women in this age group is about 2% does this survey show evidence for the heritability of breast cancer?
Breast Cancer - set-up

\[ n = 10000, \hat{p} = 0.04 \]

cases:
Breast Cancer - set-up

\[ n = 10000, \hat{p} = 0.04 \]

- cases: 10000 women whose mothers had breast cancer
- variable(s):
Breast Cancer - set-up

\[ n = 10000, \hat{p} = 0.04 \]

- cases: 10000 women whose mothers had breast cancer
- variable(s): only 1 - respondent’s breast cancer status (categorical)
- parameter of interest:
Breast Cancer - set-up

\[ n = 10000, \hat{p} = 0.04 \]

- cases: 10000 women whose mothers had breast cancer
- variable(s): only 1 - respondent’s breast cancer status (categorical)
- parameter of interest: \( p \), true proportion of women with breast cancer (whose mothers also had breast cancer)
- point estimate:
Breast Cancer - set-up

\[ n = 10000, \hat{p} = 0.04 \]

- cases: 10000 women whose mothers had breast cancer
- variable(s): only 1 - respondent’s breast cancer status (categorical)
- parameter of interest: \( p \), true proportion of women with breast cancer (whose mothers also had breast cancer)
- point estimate: \( \hat{p} \), proportion of women with breast cancer in survey sample
- hypotheses:
Breast Cancer - set-up

\[ n = 10000, \hat{p} = 0.04 \]

- cases: 10000 women whose mothers had breast cancer
- variable(s): only 1 - respondent’s breast cancer status (categorical)
- parameter of interest: \( p \), true proportion of women with breast cancer (whose mothers also had breast cancer)
- point estimate: \( \hat{p} \), proportion of women with breast cancer in survey sample
- hypotheses: \( H_0 : p = 0.02; H_A : p > 0.02 \)
Breast Cancer - p-value

- $H_A : p > 0.02 \rightarrow$ one-tailed
Breast Cancer - p-value

- $H_A : p > 0.02 \rightarrow$ one-tailed
- p-value: $P(\text{obtaining a sample of 10000 women where 400 or more have had breast cancer, if the true prevalence breast cancer is 0.02})$
Breast Cancer - p-value

- $H_A : p > 0.02 \rightarrow$ one-tailed
- p-value: $P(\text{obtaining a sample of 10000 women where 400 or more have had breast cancer, if the true prevalence breast cancer is 0.02})$
- Conditions: expected number of success (200) and failures (9800) $\geq 10$
Breast Cancer - p-value

- $H_A : p > 0.02 \rightarrow$ one-tailed
- p-value: $P(\text{obtaining a sample of 10000 women where 400 or more have had breast cancer, if the true prevalence breast cancer is 0.02})$
- Conditions: **expected** number of success (200) and failures (9800) $\geq 10$
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score
Breast Cancer - p-value

- $H_A : p > 0.02 \rightarrow \text{one-tailed}$
- p-value: $P(\text{obtaining a sample of 10000 women where 400 or more have had breast cancer, if the true prevalence breast cancer is 0.02})$
- Conditions: **expected** number of success (200) and failures (9800) $\geq 10$
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score
- Because we are conducting a hypothesis test on a proportion,
  
  $$SE = \sqrt{p_0(1 - p_0)/n} = \sqrt{0.02(1 - 0.02)/10000} = 0.0014$$
Breast Cancer - p-value

- \( H_A : p > 0.02 \rightarrow \text{one-tailed} \)
- p-value: \( P(\text{obtaining a sample of 10000 women where 400 or more have had breast cancer, if the true prevalence breast cancer is 0.02}) \)
- Conditions: \textbf{expected} number of success (200) and failures (9800) \( \geq 10 \)
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a \( Z \) score
- Because we are conducting a hypothesis test on a proportion,

\[
SE = \sqrt{p_0(1 - p_0)/n} = \sqrt{0.02(1 - 0.02)/10000} = 0.0014
\]

\[
Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.04 - 0.02}{0.0014} = 14.29
\]
Breast Cancer - p-value

- $H_A: p > 0.02 \rightarrow$ one-tailed
- p-value: $P(\text{obtaining a sample of 10000 women where 400 or more have had breast cancer, if the true prevalence breast cancer is 0.02})$
- Conditions: **expected** number of success (200) and failures (9800) $\geq 10$
- Since the sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score
- Because we are conducting a hypothesis test on a proportion,

$$SE = \sqrt{p_0(1 - p_0)/n} = \sqrt{0.02(1 - 0.02)/10000} = 0.0014$$

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.04 - 0.02}{0.0014} = 14.29$$

$$p-value = P(Z > 14.29) \approx 0$$
Breast Cancer - p-value (using R)

```r
inference(bc, est = "proportion", type = "ht",
         method = "theoretical", null = 0.02,
         alternative = "greater", success="yes")
```

```r
## Single proportion -- success: yes
## Summary statistics: p_hat = 0.04 ; n = 10000
## H0: p = 0.02
## HA: p > 0.02
## Check conditions: number of expected successes = 200 ; number of expected failures = 9800
## Standard error = 0.0014
## Test statistic: Z = 14.286
## p-value = 0
```
Breast Cancer - confidence interval

- Confidence level: 99%
Breast Cancer - confidence interval

- Confidence level: 99%
- Conditions: **observed** number of success (400) and failures (9600) \(\geq 10\)
Breast Cancer - confidence interval

- Confidence level: 99%
- Conditions: **observed** number of success (400) and failures (9600) \( \geq 10 \)
- Theoretical: Using a critical value based on the Z distribution (\( z^* \)):

\[
\text{point estimate} \pm ME = \text{point estimate} \pm z^* \times SE
\]
Breast Cancer - confidence interval

- Confidence level: 99%
- Conditions: observed number of success (400) and failures (9600) \( \geq 10 \)
- Theoretical: Using a critical value based on the Z distribution \((z^*)\):

\[
\text{point estimate} \pm ME = \text{point estimate} \pm z^* \times SE
\]

For a confidence interval,

\[
SE = \sqrt{\hat{p}(1 - \hat{p})/n} = 0.00196
\]
Breast Cancer - confidence interval

- Confidence level: 99%
- Conditions: observed number of success (400) and failures (9600) \( \geq 10 \)
- Theoretical: Using a critical value based on the Z distribution (\( z^* \)):

\[
\text{point estimate} \pm ME = \text{point estimate} \pm z^* \times SE
\]

For a confidence interval,

\[
SE = \sqrt{\hat{p}(1 - \hat{p})/n} = 0.00196
\]

\[
0.04 \pm 2.58 \times 0.00196 \approx 0.04 \pm 0.0050
\]
Breast Cancer - confidence interval

- Confidence level: 99%
- Conditions: observed number of success (400) and failures (9600) \( \geq 10 \)
- Theoretical: Using a critical value based on the Z distribution \((z^*)\):

\[
\text{point estimate} \pm ME \\
= \text{point estimate} \pm z^* \times SE
\]

For a confidence interval,

\[
SE = \sqrt{\hat{p}(1 - \hat{p})/n} = 0.00196
\]

\[
0.04 \pm 2.58 \times 0.00196 \approx 0.04 \pm 0.0050 \\
= (0.035, 0.045)
\]
inference(bc, est = "proportion", type = "ci",
    method = "theoretical", success="yes",
    conflevel = 0.99)

Single proportion -- success: yes
Summary statistics: \( p_{\hat{}} = 0.04 \); \( n = 10000 \)
Check conditions: number of successes = 400 ; number of failures = 9600
Standard error = 0.002
99 % Confidence interval = ( 0.035 , 0.045 )
Example - Breast Cancer & Age

It is theorized that an important risk factor for breast cancer is age at first birth. An international study was set up to test this hypothesis. Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did not have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: (1) women whose age at first birth was less than or equal to 29 years and (2) women whose age at first birth was 30 years. The following results were found among women with at least one birth: 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth greater than 30. How can we assess whether this difference is significant or simply due to chance?
Breast Cancer & Age - set-up

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer (case)</th>
<th>No Breast Cancer (Controls)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 29$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\geq 30$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Breast Cancer & Age - set-up

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth greater than 30.

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer (case)</th>
<th>No Breast Cancer (Controls)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth greater than 30.

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer (case)</th>
<th>No Breast Cancer (Controls)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 30</td>
<td></td>
<td>683</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3220</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Breast Cancer & Age - set-up

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth greater than 30.

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer (case)</th>
<th>No Breast Cancer (Controls)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 29</td>
<td>2537</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 30</td>
<td>683</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3220</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Breast Cancer & Age - set-up

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth greater than 30.

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer (case)</th>
<th>No Breast Cancer (Controls)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 29</td>
<td>2537</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 30</td>
<td>683</td>
<td>1498</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3220</td>
<td>10245</td>
<td></td>
</tr>
</tbody>
</table>
Breast Cancer & Age - set-up

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth greater than 30.

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer (case)</th>
<th>No Breast Cancer (Controls)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 29</td>
<td>2537</td>
<td>8747</td>
<td></td>
</tr>
<tr>
<td>≥ 30</td>
<td>683</td>
<td>1498</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3220</td>
<td>10245</td>
<td></td>
</tr>
</tbody>
</table>
Breast Cancer & Age - set-up

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth greater than 30.

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer (case)</th>
<th>No Breast Cancer (Controls)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 29</td>
<td>2537</td>
<td>8747</td>
<td>11284</td>
</tr>
<tr>
<td>≥ 30</td>
<td>683</td>
<td>1498</td>
<td>2181</td>
</tr>
<tr>
<td>Total</td>
<td>3220</td>
<td>10245</td>
<td>13465</td>
</tr>
</tbody>
</table>
Breast Cancer & Age - set-up

\[ n_{\text{case}} = 3220, \ n_{\text{ctrl}} = 10245, \ \hat{p}_{\text{case}} = 0.212, \ \hat{p}_{\text{ctrl}} = 0.146 \]

- cases:
Breast Cancer & Age - set-up

\[ n_{\text{case}} = 3220, \ n_{\text{ctrl}} = 10245, \ \hat{p}_{\text{case}} = 0.212, \ \hat{p}_{\text{ctrl}} = 0.146 \]

- cases: 13465 women (hospital patients) with at least one child
- variable(s):
Breast Cancer & Age - set-up

\[ n_{\text{case}} = 3220, \quad n_{\text{ctrl}} = 10245, \quad \hat{p}_{\text{case}} = 0.212, \quad \hat{p}_{\text{ctrl}} = 0.146 \]

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest:
Breast Cancer & Age - set-up

\[ n_{\text{case}} = 3220, \quad n_{\text{ctrl}} = 10245, \quad \hat{p}_{\text{case}} = 0.212, \quad \hat{p}_{\text{ctrl}} = 0.146 \]

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: \( p_{\text{case}} - p_{\text{ctrl}} \)
- point estimate:
Breast Cancer & Age - set-up

\[ n_{\text{case}} = 3220, \quad n_{\text{ctrl}} = 10245, \quad \hat{p}_{\text{case}} = 0.212, \quad \hat{p}_{\text{ctrl}} = 0.146 \]

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: \( p_{\text{case}} - p_{\text{ctrl}} \)
- point estimate: \( \hat{p}_{\text{case}} - \hat{p}_{\text{ctrl}} \)
- test:
Breast Cancer & Age - set-up

- \( n_{\text{case}} = 3220, \ n_{\text{ctrl}} = 10245, \ \hat{p}_{\text{case}} = 0.212, \ \hat{p}_{\text{ctrl}} = 0.146 \)

- cases: 13465 women (hospital patients) with at least one child

- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical

- parameter of interest: \( p_{\text{case}} - p_{\text{ctrl}} \)

- point estimate: \( \hat{p}_{\text{case}} - \hat{p}_{\text{ctrl}} \)

- test: compare two population proportion of independent groups

- hypotheses:
Breast Cancer & Age - set-up

\[ n_{case} = 3220, \ n_{ctrl} = 10245, \ \hat{p}_{case} = 0.212, \ \hat{p}_{ctrl} = 0.146 \]

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: \( p_{case} - p_{ctrl} \)
- point estimate: \( \hat{p}_{case} - \hat{p}_{ctrl} \)
- test: compare two population proportion of independent groups
- hypotheses: (two-tailed)
  \[ H_0 : p_{case} = p_{ctrl} \]
  \[ H_A : p_{case} \neq p_{ctrl} \]
- Note: \( p_{case} = P(\text{age} \geq 30 | \text{case}) \) and \( p_{ctrl} = P(\text{age} \geq 30 | \text{ctrl}) \)
Breast Cancer & Age - p-value

- $H_A : p \neq 0 \rightarrow$ two-tailed
Breast Cancer & Age - p-value

- $H_A : p \neq 0 \rightarrow$ two-tailed
- We need to calculated the pooled proportion:

$$\hat{p}_{pooled} = \frac{2181}{13465} = 0.162$$
Breast Cancer & Age - p-value

- $H_A : p \neq 0 \rightarrow$ two-tailed
- We need to calculate the pooled proportion:
  \[ \hat{p}_{\text{pooled}} = \frac{2181}{13465} = 0.162 \]
- Conditions:
  - case: expected number of successes (522) and failures (2698) are $\geq 10$
  - control: expected number of successes (1659) and failures (8586) are $\geq 10$
Breast Cancer & Age - p-value

- $H_A : p \neq 0 \rightarrow$ two-tailed
- We need to calculated the pooled proportion:

$$\hat{p}_{pooled} = \frac{2181}{13465} = 0.162$$

- Conditions:
  - case: expected number of successes (522) and failures (2698) are $\geq 10$
  - control: expected number of successes (1659) and failures (8586) are $\geq 10$
- The sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score
Breast Cancer & Age - p-value

- $H_A : p \neq 0 \rightarrow$ two-tailed
- We need to calculated the pooled proportion:
  
  $$\hat{p}_{pooled} = \frac{2181}{13465} = 0.162$$

- Conditions:
  - case: expected number of successes (522) and failures (2698) are $\geq 10$
  - control: expected number of successes (1659) and failures (8586) are $\geq 10$

- The sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score

- Because we are conducting a hypothesis test on the difference of two proportions,

  $$SE = \sqrt{\hat{p}(1-\hat{p})(1/n_{case} + 1/n_{ctrl})} = 0.0074$$
Breast Cancer & Age - p-value

- $H_A : p \neq 0 \rightarrow$ two-tailed
- We need to calculate the pooled proportion:

\[ \hat{p}_{\text{pooled}} = \frac{2181}{13465} = 0.162 \]

- Conditions:
  - case: **expected** number of successes (522) and failures (2698) are $\geq 10$
  - control: **expected** number of successes (1659) and failures (8586) are $\geq 10$

- The sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score

- Because we are conducting a hypothesis test on the difference of two proportions,

\[ SE = \sqrt{\hat{p}(1-\hat{p})(1/n_{\text{case}} + 1/n_{\text{ctrl}})} = 0.0074 \]

\[ Z = \frac{\hat{p}_{\text{case}} - \hat{p}_{\text{ctrl}} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92 \]
Breast Cancer & Age - p-value

- $H_A : p \neq 0 \rightarrow$ two-tailed
- We need to calculated the pooled proportion:

$$\hat{p}_{pooled} = \frac{2181}{13465} = 0.162$$

- Conditions:
  - case: **expected** number of successes (522) and failures (2698) are $\geq 10$
  - control: **expected** number of successes (1659) and failures (8586) are $\geq 10$
- The sampling distribution is nearly normal (according to the CLT), we can find this probability using a $Z$ score
- Because we are conducting a hypothesis test on the difference of two proportions,

$$SE = \sqrt{\hat{p}(1 - \hat{p})(1/n_{case} + 1/n_{ctrl})} = 0.0074$$

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

p-value $= P(Z > 8.92) + P(Z < -8.92) \approx 0$
Breast Cancer & Age - p-value (using R)

```r
inference(z[,"age"], group=z[,"bc"],
    est = "proportion", type = "ht",
    method = "theoretical", null = 0.02,
    alternative = "greater", success=">30")
```

---

```r
## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: >30
## Summary statistics:
##
##    group
## data case ctrl Sum
## <29 2537  8747 11284
## >30  683  1498  2181
## Sum 3220 10245 13465
##
## Observed difference between proportions (case-ctrl) = 0.0659
```

...
Breast Cancer & Age - p-value (using R)

```
inference(z[,"age"], group=z[,"bc"],
    est = "proportion", type = "ht",
    method = "theoretical", null = 0,
    alternative = "greater", success=">30")

...```

```r
## H0: p_case - p_ctrl = 0
## HA: p_case - p_ctrl > 0
## Pooled proportion = 0.162
## Check conditions:
##   case : number of expected successes = 522 ; number of expected failures = 2698
##   ctrl : number of expected successes = 1659 ; number of expected failures = 8586
## Standard error = 0.007
## Test statistic: Z = 8.853
## p-value = 0
```
Breast Cancer & Age - confidence interval

- Confidence level: 98%
Breast Cancer & Age - confidence interval

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distribution ($z^*$):

$$\text{point estimate} \pm ME = \text{point estimate} \pm z^* \times SE$$
Breast Cancer & Age - confidence interval

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distribution ($z^*$):

  \[
  \text{point estimate} \pm ME \\
  = \text{point estimate} \pm z^* \times SE
  \]

For a confidence interval,

\[
SE = \sqrt{\hat{p}_{\text{case}}(1 - \hat{p}_{\text{case}})/n_{\text{case}} + \hat{p}_{\text{ctrl}}(1 - \hat{p}_{\text{ctrl}})/n_{\text{ctrl}}}
= \sqrt{0.212(1 - 0.212)/3220 + 0.146(1 - 0.146)/10245}
= 0.008
\]
Breast Cancer & Age - confidence interval

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distribution ($z^*$):

  \[
  \text{point estimate} \pm ME = \text{point estimate} \pm z^* \times SE
  \]

For a confidence interval,

\[
SE = \sqrt{\hat{p}_{\text{case}}(1 - \hat{p}_{\text{case}})/n_{\text{case}} + \hat{p}_{\text{ctrl}}(1 - \hat{p}_{\text{ctrl}})/n_{\text{ctrl}}}
\]

\[
= \sqrt{0.212(1 - 0.212)/3220 + 0.146(1 - 0.146)/10245}
\]

\[
= 0.008
\]

\[
(0.212 - 0.146) \pm 2.33 \times 0.008 \approx 0.066 \pm 0.0186
\]
Breast Cancer & Age - confidence interval

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distribution ($z^*$):

$$\text{point estimate} \pm ME = \text{point estimate} \pm z^* \times SE$$

For a confidence interval,

$$SE = \sqrt{\hat{p}_{case}(1 - \hat{p}_{case})/n_{case} + \hat{p}_{ctrl}(1 - \hat{p}_{ctrl})/n_{ctrl}}$$

$$= \sqrt{0.212(1 - 0.212)/3220 + 0.146(1 - 0.146)/10245}$$

$$= 0.008$$

$$(0.212 - 0.146) \pm 2.33 \times 0.008 \approx 0.066 \pm 0.0186$$

$$= (0.0474, 0.0846)$$
Breast Cancer & Age - confidence interval (using R)

```r
inference(z[,"age"], group = z[,"bc"],
    est = "proportion", type = "ci",
    method = "theoretical", success=">30",
    conflevel = 0.98)
```

## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: >30
## Summary statistics:
##
<table>
<thead>
<tr>
<th>group</th>
<th>data</th>
<th>case</th>
<th>ctrl</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;29</td>
<td>2537</td>
<td></td>
<td>8747</td>
<td>11284</td>
</tr>
<tr>
<td>&gt;30</td>
<td>683</td>
<td>1498</td>
<td></td>
<td>2181</td>
</tr>
<tr>
<td>Sum</td>
<td>3220</td>
<td>10245</td>
<td>13465</td>
<td></td>
</tr>
</tbody>
</table>
## Observed difference between proportions (case-ctrl) = 0.0659

## Check conditions:
## case : number of successes = 683 ; number of failures = 2537
## ctrl : number of successes = 1498 ; number of failures = 8747
## Standard error = 0.008
## 98 % Confidence interval = ( 0.0473 , 0.0845 )