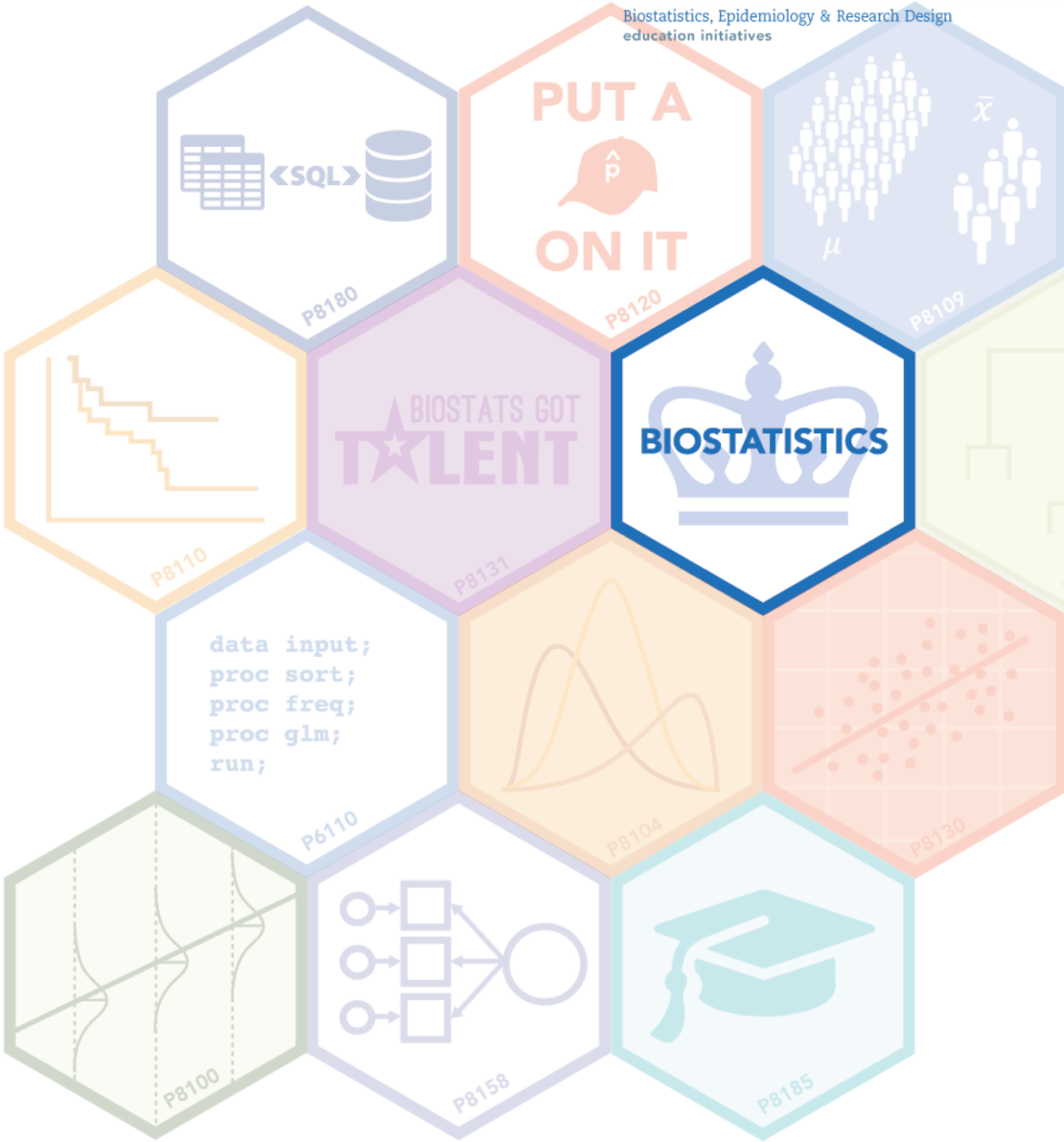
GETTING STARTED WITH R (PART 2)

Christine Mauro, PhD

November 26, 2018



Last Time

- Overview of R and R Studio
- Importing Data
  - Read CSV files using readr package
- Examining Data Attributes
  - Data structure, type and dimensionality
- Manipulating Data (Data Wrangling)
  - Select, Filter, Mutate, Arrange
  - Stacking and Merging

Today's Outline

- Descriptive Statistics
  - Continuous/Categorical data
- Data Visualization
  - Histogram
  - Box-plot
  - Scatterplot
- Basic Hypothesis Testing
  - T-tests and ANOVA
  - Chi-squared and Fisher's Exact test

Application

- Risk Factors Associated with Low Birthweight: lowbwt_ALL.csv

- The data on 189 births were collected at Baystate Medical Center, Springfield, Mass. during 1986. The dataset contains an indicator of low infant birth weight as a response and several risk factors associated with low birth weight. The actual birth weight is also included in the dataset.

- The dataset consists of the following 10 variables:
  - low: indicator of birth weight less than 2.5kg
  - age: mother's age in years
  - lwt: mother's weight in pounds at last menstrual period
  - race: mothers race ("white", "black", "other")
  - smoke: smoking status during pregnancy (yes/no)
  - ht: history of hypertension (yes/no)
  - ui: presence of uterine irritability (yes/no)
  - ftv: physician visit during the first trimester (yes/no)
  - ptl: previous premature labor (yes/no)
  - bwt: birth weight in grams

Let's Get Started

- Step 1: Open your R project from last time (double click file to open)
- Step 2: Create a new script and save.
- Step 3: Load in lowbwt_ALL.csv and packages.

```
#################################################################
# November 26, 2018
# Christine Mauro
#
# Getting Started with R - Part 2
#################################################################

library(dplyr)
library(readr)

############################## Loading Data ##########################


lowbirth = read_csv(file = "./lowbwt_All.csv")
names(lowbirth)
lowbirth
```

DESCRIPTIVE STATISTICS

Descriptive Stats

- Describe the basic features of the data
- Continuous Variables
  - Measures of central tendency (e.g., mean, median)
  - Measures of variability/spread (e.g., standard deviation, interquartile range, range)
- Categorical Variables
  - Counts and Percentages

skimr package

library(skimr)

```
> skim(lowbirth)
Skim summary statistics
 n obs: 189
 n variables: 10

── Variable type:character ──────────────────────────────────────────────────────────────────────
 variable missing complete   n min max empty n_unique
     race       0      189 189   5   5     0        3

── Variable type:integer ────────────────────────────────────────────────────────────────────────
 variable missing complete   n     mean     sd  p0  p25  p50  p75 p100     hist
      age       0      189 189    23.24    5.3  14   19   23   26   45 ▂▅▅▅▂▁▁▁
      bwt       0      189 189 2944.66  729.02 709 2414 2977 3475 4990 ▁▂▅▅▅▅▅▂▁
      ftv       0      189 189     0.47    0.5   0    0    0    1    1 ▇▁▁▁▁▁▁▇
       ht       0      189 189     0.063   0.24  0    0    0    0    1 ▇▁▁▁▁▁▁▁
      low       0      189 189     0.31    0.46  0    0    0    1    1 ▇▁▁▁▁▁▁▂
      lwt       0      189 189   129.69   30.65 80  110  121  140  250 ▂▅▅▂▁▁▁▁
      ptl       0      189 189     0.16    0.37  0    0    0    0    1 ▇▁▁▁▁▁▁▂
    smoke       0      189 189     0.39    0.49  0    0    0    1    1 ▇▁▁▁▁▁▁▇
       ui       0      189 189     0.15    0.36  0    0    0    0    1 ▇▁▁▁▁▁▁▂
```

skimr updated

```
> lowbirth2 <- mutate_at(lowbirth, vars(race, ftv, ht, low, ptl, smoke, ui), as.factor)
> skim(lowbirth2)
Skim summary statistics
 n obs: 189
 n variables: 10

── Variable type:factor ────────────────────────────────────────────────────────────────────────────────────────────────
 variable missing complete   n n_unique                        top_counts ordered
      ftv       0      189 189        2             0: 100, 1: 89, NA: 0   FALSE
       ht       0      189 189        2             0: 177, 1: 12, NA: 0   FALSE
      low       0      189 189        2             0: 130, 1: 59, NA: 0   FALSE
      ptl       0      189 189        2             0: 159, 1: 30, NA: 0   FALSE
     race       0      189 189        3 whi: 96, oth: 67, bla: 26, NA: 0   FALSE
    smoke       0      189 189        2             0: 115, 1: 74, NA: 0   FALSE
       ui       0      189 189        2             0: 161, 1: 28, NA: 0   FALSE

── Variable type:integer ───────────────────────────────────────────────────────────────────────────────────────────────
 variable missing complete   n    mean     sd  p0  p25  p50  p75 p100     hist
      age       0      189 189   23.24   5.3  14   19   23   26   45 ▂▅▇▅▂▁▁▁
      bwt       0      189 189 2944.66 729.02 709 2414 2977 3475 4990 ▁▂▅▇▇▅▂▁
      lwt       0      189 189  129.69  30.65  80  110  121  140  250 ▅▇▅▂▁▁▁▁
>
```

Summary Function (base R)

```
> summary(lowbirth2)
 low          age             lwt           race     smoke    ht       ui       ftv      ptl          bwt
 0:130   Min.   :14.00   Min.   : 80.0   black:26   0:115    0:177    0:161    0:100    0:159    Min.   : 709
 1: 59   1st Qu.:19.00   1st Qu.:110.0   other:67   1: 74    1: 12    1: 28    1: 89    1: 30    1st Qu.:2414
         Median :23.00   Median :121.0   white:96                                                Median :2977
         Mean   :23.24   Mean   :129.7                                                           Mean   :2945
         3rd Qu.:26.00   3rd Qu.:140.0                                                           3rd Qu.:3475
         Max.   :45.00   Max.   :250.0                                                           Max.   :4990
```

Descriptive Stats: Continuous Variables

- Base R functions:

mean(mydata)	Mean of all numeric variables
mean(mydata$myvar)	Mean of a selected numeric variable from the dataset
median(mydata$myvar)	Median: the 50th percentile
var(mydata$myvar)	Variance
sd(mydata$myvar)	Standard Deviation
min(mydata$myvar)	Minimum value
max(mydata$myvar)	Maximum value
range(mydata$myvar)	Range: Min-Max
quantile(mydata$myvar)	Quartiles; Interquartile Range: 25th – 75th percentiles

```
> mean(lowbirth2$bwt)
[1] 2944.656
> sd(lowbirth$bwt)
[1] 729.0224
> quantile(lowbirth2$bwt, c(.25, .75))
 25%  75%
2414 3475
>
```

Summarize function (dplyr)

- Similar to summary function, but stores results as a tibble (data set).
  - Useful for later calculations
  - Can use with "group_by" function

```
> summarize(lowbirth2, mean_bwt = mean(bwt), median_bwt = median(bwt), sd_bwt = sd(bwt))
# A tibble: 1 x 3
  mean_bwt median_bwt sd_bwt
     <dbl>      <int>  <dbl>
1    2945.       2977   729.
```

```
> as.data.frame(summarize(lowbirth2, mean_bwt = mean(bwt), median_bwt = median(bwt), sd_bwt = sd(bwt)))
  mean_bwt median_bwt   sd_bwt
1 2944.656       2977 729.0224
```

Descriptive Stats: Continuous Variables

- Summary statistics for each level of another categorical variable --> use group_by function (dplyr)

Example: summary stats of birthweight 'bwt' by 'race'

```
group_by(lowbirth2, race) %>%
summarize(mean_bwt = mean(bwt), median_bwt = median(bwt), sd_bwt = sd(bwt))
```

```
> group_by(lowbirth2, race) %>%
+ summarize(mean_bwt = mean(bwt), median_bwt = median(bwt), sd_bwt = sd(bwt))
# A tibble: 3 x 4
  race  mean_bwt median_bwt sd_bwt
  <fct>    <dbl>      <dbl>  <dbl>
1 black    2720.       2849   639.
2 other    2804.       2835   721.
3 white    3104.       3076   728.
>
```

Descriptive Stats: Categorical Variables

- Row, column, and total frequencies
- Two- and three-way tabulations
- R functions:

`tbl <- table(mydata$var1, mydata$var2)`	Two-way table
`prop.table(tbl, 1)`	Row proportions
`prop.table(tbl, 2)`	Column proportions
`prop.table(tbl)`	Total proportions
`xtabs(~var1+var2+var3, data=mydata)`	3-way cross-tabulation

Let's try it

```
> tbl  <- table(lowbirth2$race, lowbirth2$smoke)
> tbl

        0  1
  black 16 10
  other 55 12
  white 44 52
> prop.table(tbl, 1)

               0         1
  black 0.6153846 0.3846154
  other 0.8208955 0.1791045
  white 0.4583333 0.5416667
> prop.table(tbl, 2)

               0         1
  black 0.1391304 0.1351351
  other 0.4782609 0.1621622
  white 0.3826087 0.7027027
>
```

Descriptive Stats: Categorical Variables

- 3-way tabulation

  R function: xtabs(~var1+var2+var3, data=mydata)

Example: two-way tables of 'race' x 'smoke' stratified by the levels of history of hypertension 'ht'

```
> xtabs(~race+smoke+ht, data=lowbirth2)
, , ht = 0

       smoke
race     0  1
  black 14  9
  other 51 12
  white 43 48

, , ht = 1

       smoke
race     0  1
  black  2  1
  other  4  0
  white  1  4
```

DATA VISUALIZATION

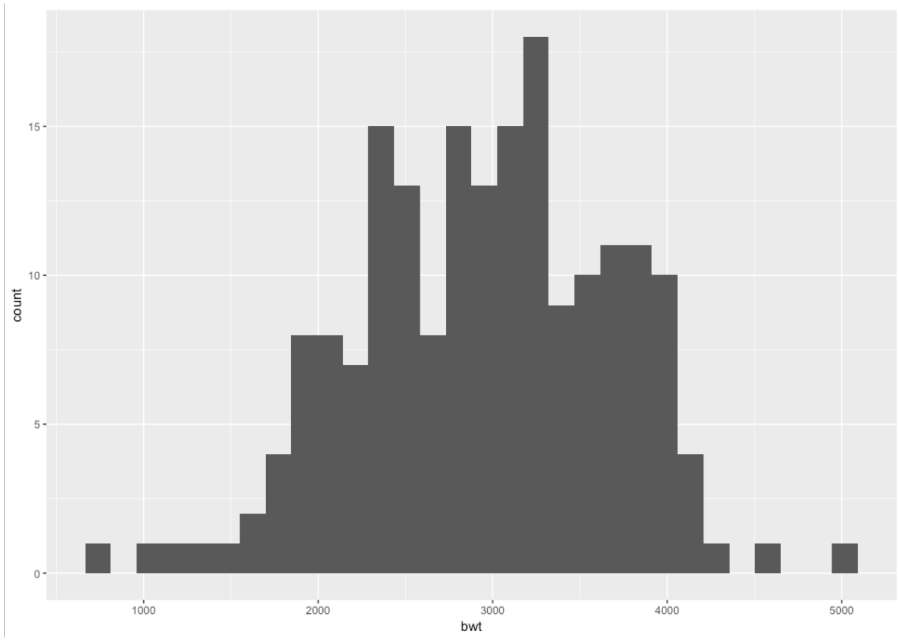
Data Visualization

- Use ggplot2 package!
- Histogram
  - Shows the underlying frequency distribution of continuous data
- Box-plot
  - Shows the underlying distribution of continuous data based on the five number summary: min, 1st quartile, median, 3rd quartile, max.
- Scatter plot
  - Shows the relationships between two continuous (numeric) variables, each plotted of one of the axes
- Barplot
  - Useful for summarizing categorical data

Data Visualization: Histograms (ggplot2)

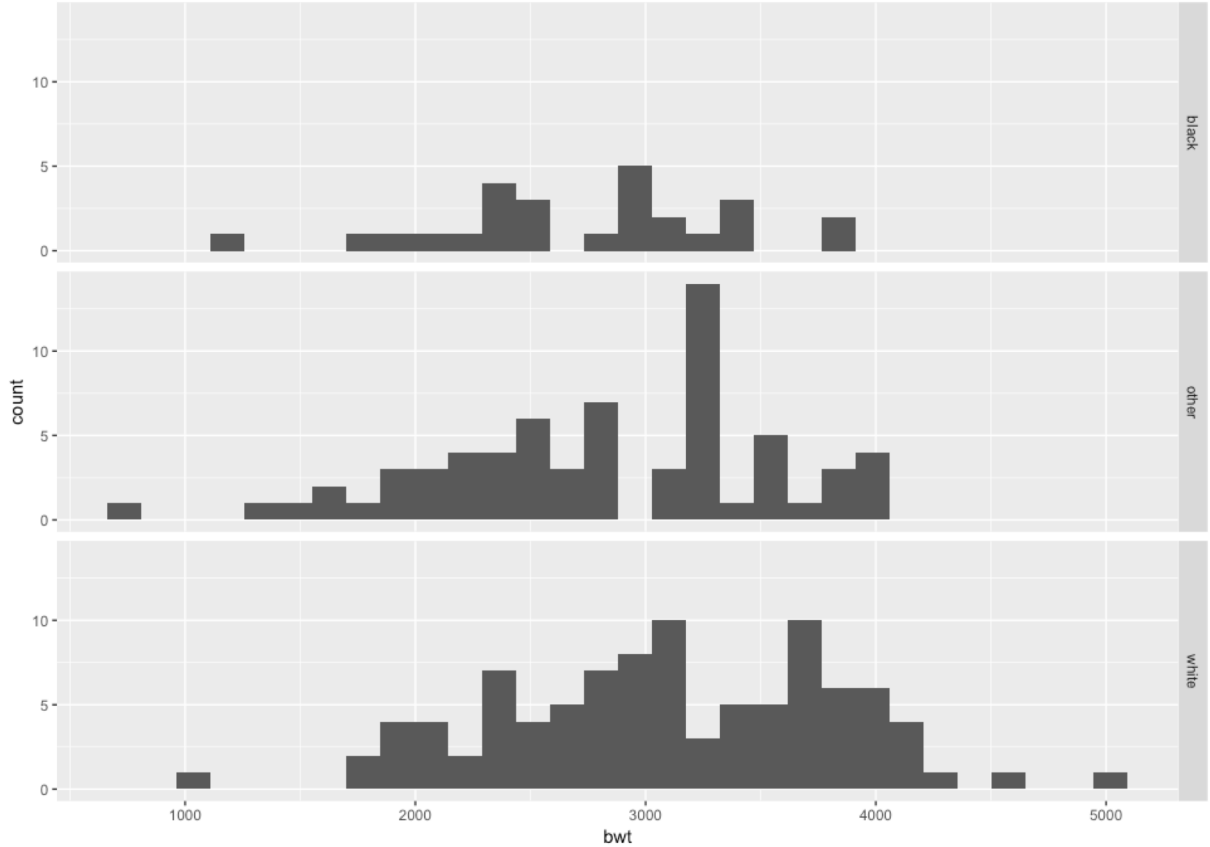
- Histogram plot of one continuous variable.

```
##### Data Visualization

library(ggplot2)

#histrogram of birthweight
ggplot(lowbirth2, aes(x = bwt)) +
  geom_histogram()

```



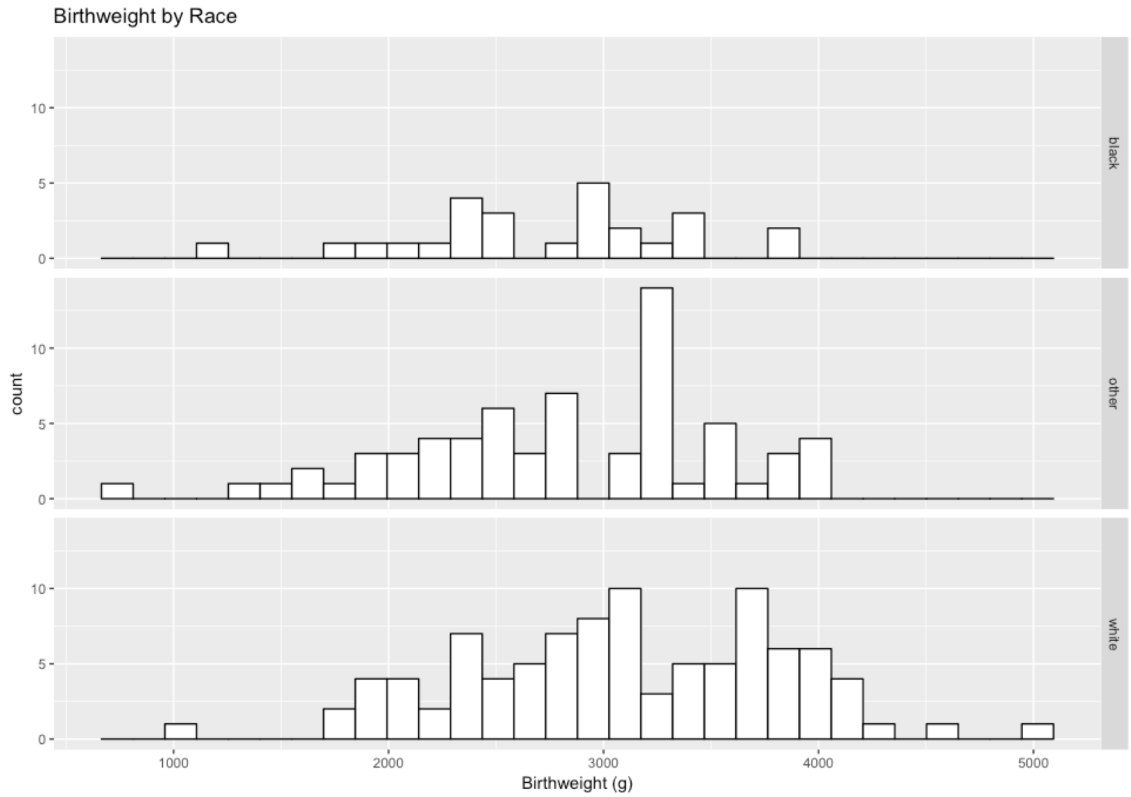
Histograms by Groups

```
#histogram of birthweight by race
ggplot(lowbirth2, aes(x = bwt)) +
  geom_histogram() +
  facet_grid(race ~ .)
```



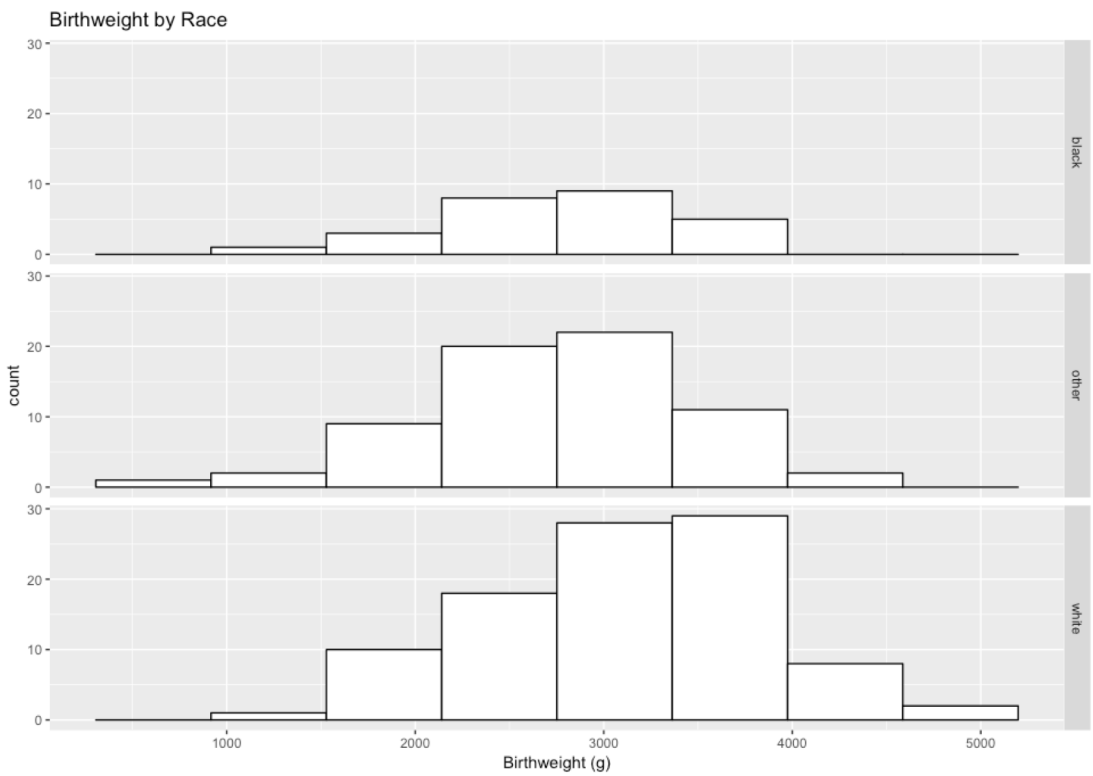
Histograms - Formatting

```
## add title and axis labels; nicer formats
ggplot(lowbirth2, aes(x = bwt)) +
  geom_histogram(colour="black", fill="white") +
  facet_grid(race ~ .) +
  ggtitle("Birthweight by Race") +
  labs(x = "Birthweight (g)")
```



Histograms – Fix Bins!

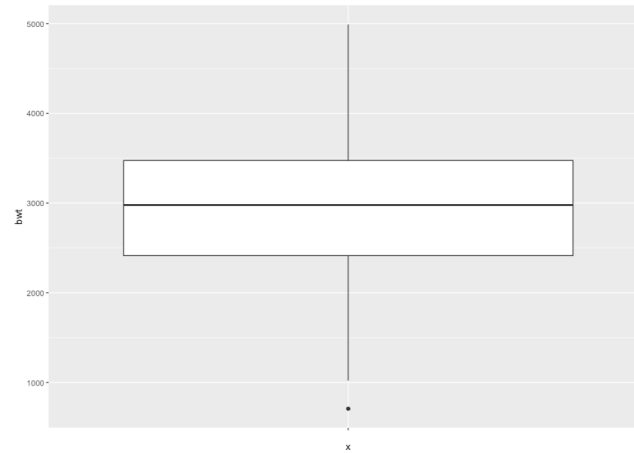
```
## add title and axis labels; nicer formats - FIX BINS
ggplot(lowbirth2, aes(x = bwt)) +
  geom_histogram(colour="black", fill="white", bins = 8) +
  facet_grid(race ~ .) +
  ggtitle("Birthweight by Race") +
  labs(x = "Birthweight (g)")
```

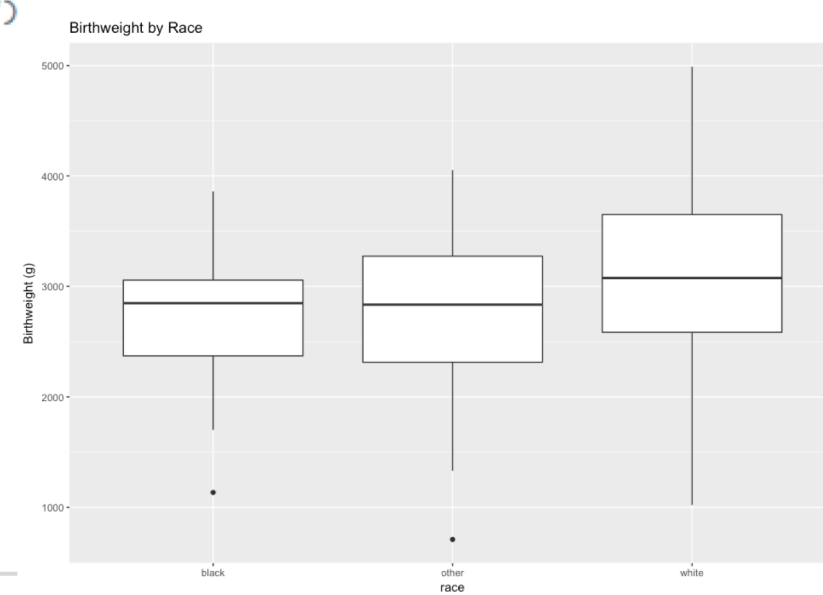


Data Visualization: Boxplots

Boxplot of one continuous variable; or continuous variable by categorical variable.

```
#boxplot of birthweight
ggplot(lowbirth2, aes(x= "", y = bwt)) +
  geom_boxplot()


#boxplot of birthweight by race
ggplot(lowbirth2, aes(x= race, y = bwt)) +
  geom_boxplot() +
  ggtitle("Birthweight by Race") +
  labs(y = "Birthweight (g)")
```

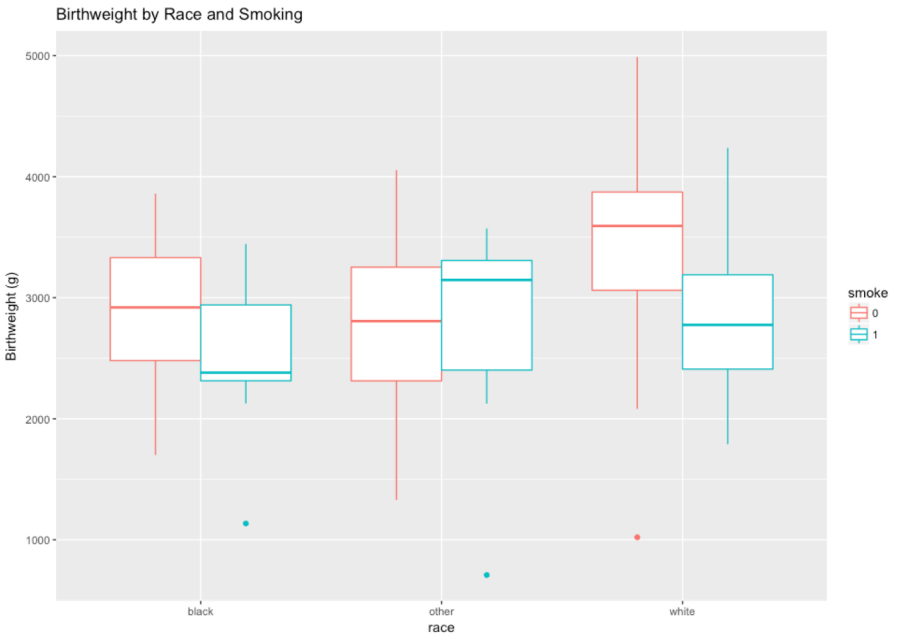




Data Visualization: Grouped Boxplots

A grouped boxplot is a boxplot where each category is subdivided in several groups.

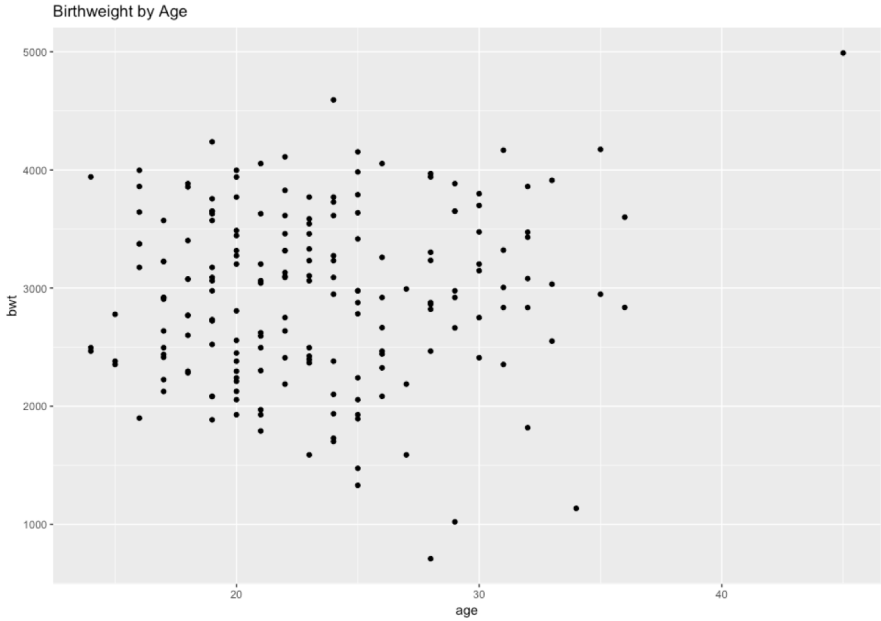
```
#boxplot of birthweight by race and smoking status
ggplot(lowbirth2, aes(x= race, y = bwt, color=smoke)) +
  geom_boxplot() +
  ggtitle("Birthweight by Race and Smoking") +
  labs(y = "Birthweight (g)")
```



Data Visualization: Scatterplot

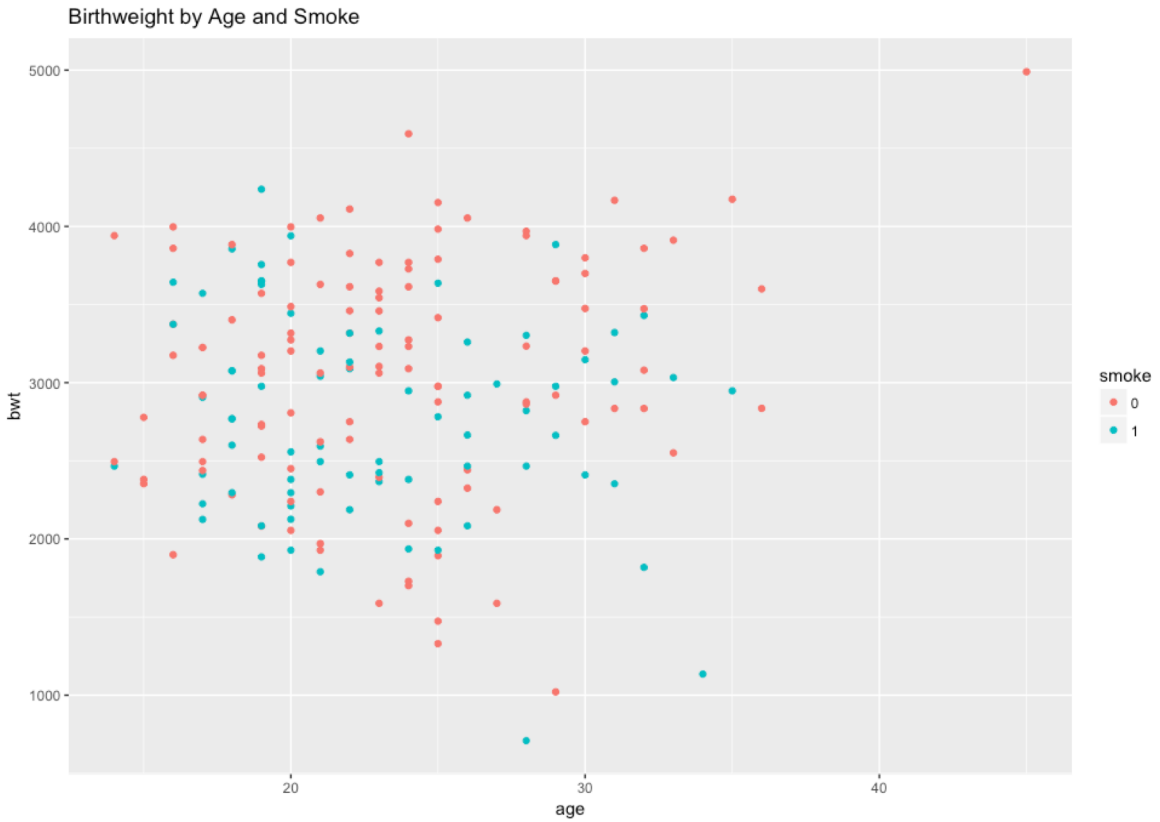
- Shows the relationship / trend between two <u>continuous</u> variables

```
#scatterplot of birthweight by age
ggplot(lowbirth2, aes(x= age, y = bwt)) +
  geom_point() +
  ggtitle("Birthweight by Age")
```



Data Visualization: Multiple Scatterplots

```
#scatterplot of birthweight by age and smoking status
ggplot(lowbirth2, aes(x= age, y = bwt, color=smoke)) +
  geom_point() +
  ggtitle("Birthweight by Age and Smoke")
```



BASIC HYPOTHESIS TESTING

Hypothesis Testing

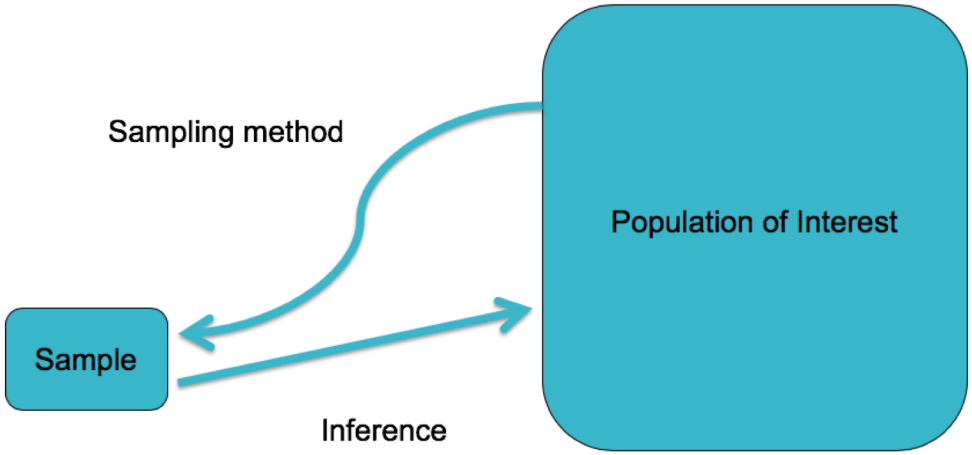
- Hypothesis testing provides a framework for making decisions about the population based on data from a sample.

The null hypothesis (H_0)
vs.
The alternative hypothesis (H_1)

- Null hypothesis is often of "no difference" in population
- Note that our decisions will always be with respect to the null hypothesis:
  - Reject the null, Fail to reject the null!

Hypothesis Testing

- Can the differences in my sample be explained by chance (i.e. sampling variability)?
  - Reject the null -> observed differences are likely not due to chance!
  - We infer from our sample back to the population:



Student's T-test

- A T-test is an analysis of one- or two-population means
  - Used for continuous (numeric) data (outcomes)
  - One-sample T-test (compare one population mean)
  - Two-sample T-test (compare two populations means)
    - Independent or paired test
- Always remember to check model assumptions before inferences
  - Normality (for small samples)
    - Histograms or QQplots/Normality tests (not covered here)
  - Independent observations within the group(s) (not repeated)

Two-Sample T-test

- Hypothesis to be tested (two-sided):

$$H_0\colon\ \mu_1 = \mu_2 \text{ vs } H_1\colon \mu_1 \neq \mu_2$$

- In the two-sample case, you FIRST need to test for the equality of variances
  - Testing the equality of variances implies testing the hypotheses:

$$H_0\colon \sigma_1^2 = \sigma_2^2 \text{ vs } H_1\colon \sigma_1^2 \neq \sigma_2^2$$

Two-Sample T-test

Example: is the birthweight of babies born to smokers significantly different than the birthweight to babies born to non-smokers?

R function to test equality of variances:

var.test(cont ~ binary, data=mydata)

R function for two-sample independent t-test:

t.test(cont ~ binary, data=mydata, var.equal=FALSE, paired=FALSE)

Options:

- Default var.equal = FALSE
  - Can be changed to TRUE if variances are unequal
- Default paired = FALSE
  - Can be changed to TRUE if data is paired, e.g., pre/post tests from same subject

Two-Sample T-test

Outcome: Continuous
Predictor: Binary

Example: is the birthweight of babies born to smokers significantly different than the birthweight to babies born to non-smokers?

```
> var.test(bwt ~ smoke, data = lowbirth2)

        F test to compare two variances

data:  bwt by smoke
F = 1.2993, num df = 114, denom df = 73, p-value = 0.229
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8469514 1.9550579
sample estimates:
ratio of variances
          1.299335

> t.test(bwt ~ smoke, data=lowbirth2, var.equal= TRUE)

        Two Sample t-test

data:  bwt by smoke
t = 2.6336, df = 187, p-value = 0.009156
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  70.69274 492.73382
sample estimates:
mean in group 0 mean in group 1
       3054.957        2773.243
```

Not enough evidence to declare inequality of variances.

Interpretation: At 0.05 significance level, we reject the null hypothesis (p-value=0.009) and conclude that the true birthweight means for smokers and non-smokers are significantly different.

One-Sample T-test

- Hypothesis to be tested (two-sided):

$$H_0\colon \mu = \mu_0 \text{ vs } H_1\colon \mu \neq \mu_0$$

Example: is the average age of mother's in this population different from 26?

R function:

```
t.test(mydata$myvar, mu = mu_null)
```

Options:

- Default alternative = 'two-sided'
  - Can be changed to alternative = 'less' or alternative = 'greater'
- Default alpha = 0.05

One-Sample T-test

Outcome:
Continuous

Example: is the average age of mother's in this population different from 26?

```
> t.test(lowbirth2$age, mu=26)

        One Sample t-test

data:  lowbirth2$age
t = -7.1659, df = 188, p-value = 1.707e-11
alternative hypothesis: true mean is not equal to 26
95 percent confidence interval:
 22.47779 23.99840
sample estimates:
mean of x
  23.2381
```

95% CI: (22.48, 23.99).
We are 95% confident that the true mean mother's age is b/w approximately 23 and 24 yrs.

Interpretation: At 0.05 significance level, we reject the null hypothesis (p-value <0.0001) and conclude that the true mean mother's age is not equal to 26.

Analysis of Variance (ANOVA)

- Use to compare the (continuous) outcomes across 3 or more groups
- Model assumptions:
  - Independent samples
  - Responses within the groups are independent and identically distributed (i.i.d)
  - Residuals are normally distributed
  - Equality of variances across groups

Analysis of Variance (ANOVA)

Example: is the birthweight of babies significantly different by race?

R function:

lm(cont_outcome~cat_predictor, data = mydata)

Notes:

- R also has an aov() function, but lm() is broader including linear regression models
- Better to declare the categorical variable/predictor as a *factor*, o/w it will be considered a continuous measurement.

Analysis of Variance (ANOVA)

Example: is the birthweight of babies significantly different by race?

```
> example1 = lm(bwt~race, data=lowbirth2)
> anova(example1)
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value   Pr(>F)
race        2  5070608 2535304  4.9719 0.007879 **
Residuals 186 94846445  509927
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Outcome: Continuous

Predictor: Categorical (3 or more levels)

Interpretation: At 0.05 significance level, we reject the null hypothesis (p-value=0.008) and conclude that there is a significant difference in mean birthweight by race.

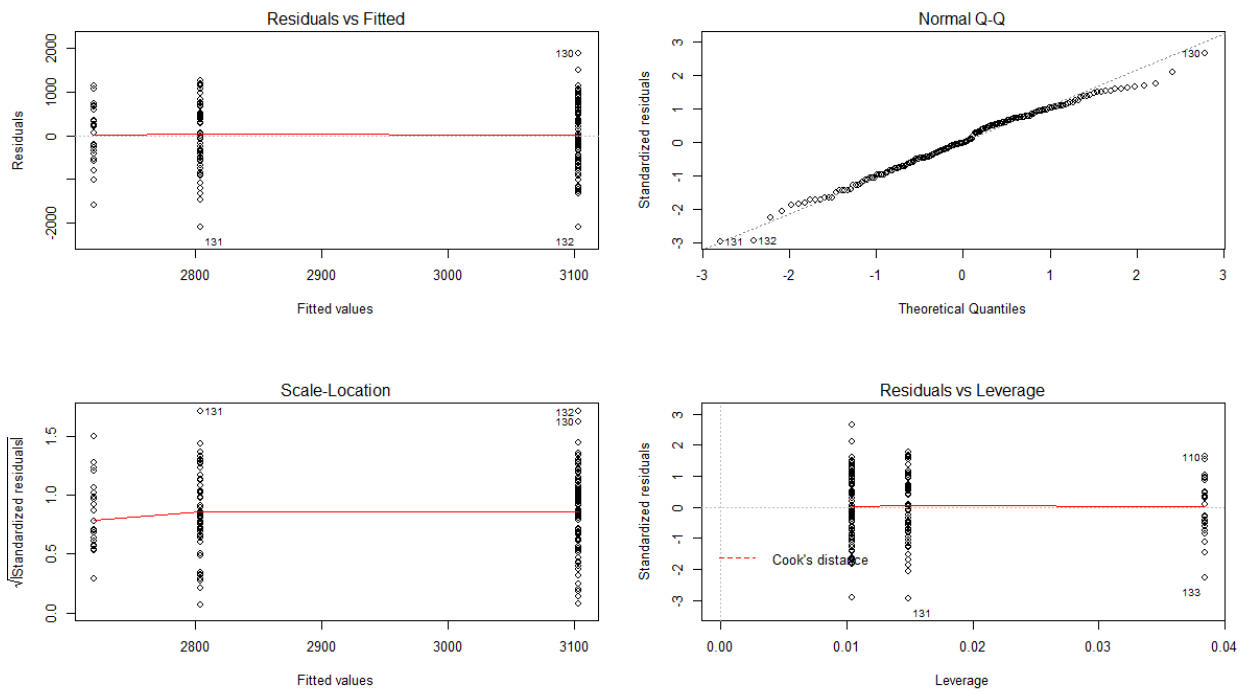
Next question: Where are these differences coming from? Try pairwise comparisons. Need to adjust for multiple comparisons (Tukey, Bonferroni, Scheffe, etc.). Topic covered in future courses!

Analysis of Variance (ANOVA)

Example: checking model assumptions – plot(example1)

Constant Variance

Normality of Residuals



Outliers/Influential Points

Categorical Data Analysis

- Categorical outcome (Y) with 2 levels (binary) or ≥3 levels (nominal or ordinal)
- Examples:
  - Nominal: race/ethnicity
  - Ordinal: clothing sizes (S, M, L, XL)
  - Binary: Disease/ No Disease; Republican/Democrat

- Predictor variables (X) can take on any form: binary, categorical, and/or continuous

Chi-Squared Test of Independence

- Use two categorical variables (row and column) to test whether they are independent or associated
- Hypotheses:

H_0: variables A and B are independent

vs

H_1: variables A and B are not independent

Test statistics: $\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$

Chi-Squared Test of Independence

- Create a (r X c) table
  - r represents the number of levels for the row variable
  - c represents the number of levels for the column variable
  - Most common example is a 2 x 2 table
- Use the observed and expected counts in each cell to calculate the chi-squared statistics
- If low expected cell counts (< 5), use Fisher's Exact test instead

Chi-Squared Test

- First you need to tabulate the two categorical variables and then apply chisq.test() to this table

R function:

table(mydata$row_var, mydata$col_var)

chisq.test(mydata$row_var, mydata$col_var)

Example: is there an association between history of uterine irritability and having a low birthweight baby?

```
> library(MASS)
> tbl <- table(low_birth_all$ui, low_birth_all$low)
> chisq.test(tbl)
```

Chi-Squared Test

Outcome: Categorical
Predictor: Categorical

Example: is there an association between smoking and having a low birthweight baby?

```
> table(lowbirth$smoke, lowbirth2$low)

     0  1
  0 86 29
  1 44 30
> chisq.test(lowbirth2$smoke, lowbirth2$low)

        Pearson's Chi-squared test with Yates' continuity correction

data:  lowbirth2$smoke and lowbirth2$low
X-squared = 4.2359, df = 1, p-value = 0.03958
```

Interpretation: At 0.05 significance level, we reject the null hypothesis (p-value=0.040) and conclude that there is a significant association between smoking and having a low birthweight baby.

Fisher's Exact Test

- Use instead of chi-squared test when low expected cell counts (<5)

Example: is there an association between smoking and having a low birthweight baby?

```
> fisher.test(lowbirth2$smoke, lowbirth2$low)

        Fisher's Exact Test for Count Data

data:  lowbirth2$smoke and lowbirth2$low
p-value = 0.03618
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.028780 3.964904
sample estimates:
odds ratio
  2.014137
```

Interpretation: At 0.05 significance level, we reject the null hypothesis (p-value=0.036) and conclude that there is a significant association between smoking and having a low birthweight baby.

Thank you!

Contact me: cmm2212@cumc.columbia.edu

Useful Resources:

http://p8105.com/topic_visualization_and_eda.html

https://stats.idre.ucla.edu/r/

Visit our BERD EDU website for additional resources:

http://irvinginstitute.columbia.edu/resources/biostat_educational_initiatives.html