Lesson 1
Review: Exploratory and Descriptive Measures

BIS 505b

Yale University
Department of Biostatistics

Date Modified: 2/7/2021

Goals for this Lesson

Addressing a Research Question

1. Identifying data types: Variable types
2. Describing the data: Numerical and graphical descriptions appropriate for each data type
   - *One* variable alone
   - Relationship between *two*+ variables

Contents

Progress this Unit

Biostatistics

- **Biostatistics** applies the principles of statistics to the biological and health-related problems
- What is statistics?

Definition

Statistics is the art and science of learning from data. It is concerned with:

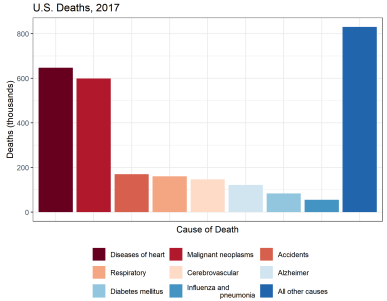
1. the **collection of data**,
2. their subsequent **description**, and their **analysis**,
3. which often leads to the **drawing of conclusions**.

Public Health Application: Framingham Heart Study

Public Health Application

In the U.S. in the 1940s and 50s, cardiovascular disease (CVD) became a major public health concern

- By 1950, 1 in 3 U.S. men developed CVD before age 60 and CVD was leading cause of death



Public Health Application: Framingham Heart Study

Public Health Application

In the U.S. in the 1940s and 50s, cardiovascular disease (CVD) became a major public health concern

- By 1950, 1 in 3 U.S. men developed CVD before age 60 and CVD was leading cause of death

Research focused on developing a preventive approach: identify preventable or modifiable predisposing factors

- How are those individuals who develop CVD different from those who do not?

Risk factors:

- Blood pressure, cholesterol, diabetes, smoking, and weight
- Smoking, being overweight/obese are top preventable causes of CVD death in U.S.
- Family history, nutrition, and physical activity are also important risk factors

Descriptive vs. Inferential Statistics

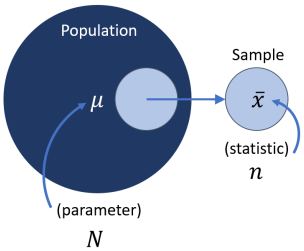
Table 1: Summary Statistics of FHS Data

Full Sample	Full Sample (N=4434)
Age (years)	
Missing	0
Mean (SD)	49.9 (8.7)
Median (Range)	49.0 (32.0, 70.0)
Body Mass Index	
Missing	19
Mean (SD)	25.8 (4.1)
Median (Range)	25.4 (15.5, 56.8)
Total Cholesterol	
Missing	52
Mean (SD)	237.0 (44.7)
Median (Range)	234.0 (107.0, 696.0)
Cigarettes Smoked Per Day	
Missing	32
0	2253 (51.2%)
1-19	907 (20.6%)
20-39	1077 (24.5%)
40+	165 (3.7%)

There are two main uses of statistics:

1. **Descriptive Statistics**
   - Numerical and graphical summaries of data
2. **Inferential Statistics**
   - The use of a sample of individuals to draw conclusions (make inferences) about the wider population of like individuals

Descriptive vs. Inferential Statistics



There are two main uses of statistics:

1. **Descriptive Statistics**
   - Numerical and graphical summaries of data
2. **Inferential Statistics**
   - The use of a sample of individuals to draw conclusions (make inferences) about the wider population of like individuals

Summarizing the Data

- Graphical and numerical summaries provide a strategy for organizing the data in a meaningful way

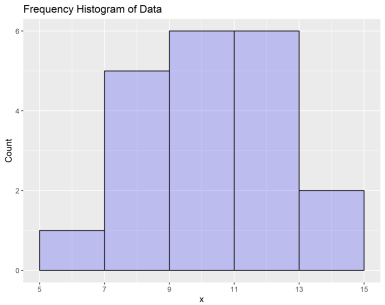
Data

```
> set.seed(6520)
> x <- round(rnorm(20, mean = 10, sd = 2), 1)
> x
 [1]  8.3  5.2 13.4 12.3 10.6 10.6  8.6 12.5
 [9] 10.3  9.7 12.5 13.2 10.2 12.0  8.2 11.7
[17]  8.5 12.6  9.2  8.5
```

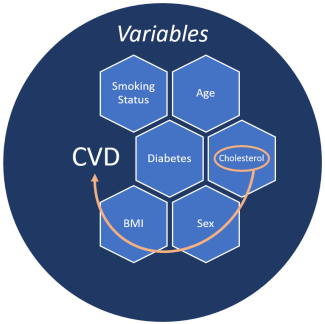
variability

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.200   8.575  10.450  10.405  12.350  13.400
```



Variables



- Objective of many studies is to learn about the variation of a variable in the population of interest
- Also interested in relationships between variables: learning of the effect of one or more variables on the variable(s) of interest

Response variable		Explanatory variables
Outcome variable		Predictor variables
Dependent variable	←	Independent variables
Primary endpoint		(Exposure variable, confounders)
Y		X

Exercise

Poll

- A recent study finds that people who feel enthusiastic and cheerful – what psychologists call '*positive affect*' – are less likely to experience memory decline as they age. This result adds to a growing body of research on positive affect's role in healthy aging.
- This study examined longitudinal associations between positive affect (i.e., feeling enthusiastic, attentive, proud, and active during the previous 30 days) and memory functioning (i.e., immediate- and delayed-recall performance) over 9 years using data from a large-scale national sample of middle-age and older adults in the United States. Models account (control) for age, gender, education, depression, negative affect, and extraversion.
- **Link to poll** (Lesson 01): *https://pollev.com/bis505b*

Link to article

Variable Types

- The key distinction for statistical analysis is between **categorical** and **quantitative** variables
- The type of variable(s) being analyzed determines the methods that should be used.
- This applies to...
  - exploratory and descriptive measures,
  - basic statistical analyses, and
  - regression modeling

Variable Types

- Variables can be classified as either:

Categorical	Quantitative
Nominal	Discrete
Ordinal	Continuous

1. Categorical variables - Each observation belongs to a set of categories
2. Quantitative variables - Take on numerical values

Variable Classification

Categorical

- Nominal data - Measurements whose values fall into categories that have no natural numerical value
  - Race, marital status, political affiliation, country of birth
- Ordinal data - Arise when measurements fall into categories that can be qualitatively ordered or ranked, but have no intrinsic numerical value
  - Pain scores (scale 1-5), stages of cancer, education level
- *Dichotomous* or *binary* - Two distinct values
  - Yes/no, no pain/some pain, sex (M/F)

Variable Classification

Quantitative

- Discrete data - Take on a finite number of possible values, usually integers $\{0, 1, 2, \ldots\}$; measured at *discrete* points on the scale
  - Count data: Number of side polyps removed, number of children, number of cigarettes
- Continuous data - Take on values in an interval scale, although sometimes we are limited in our ability to measure them
  - Age, BMI, viral load, total serum cholesterol, systolic blood pressure

Information Provided

Figure: Information Provided by Different Variable Types

Creating Variables

```
# Grouping BMI (ordinal)
fhs$BMIGRP[fhs$BMI < 18.5] = 1                    # <18.5 Underweight
fhs$BMIGRP[fhs$BMI >= 18.5 & fhs$BMI < 25] = 2  # 18.5-<25 Normal
fhs$BMIGRP[fhs$BMI >= 25 & fhs$BMI < 30] = 3    # 25-<30 Overweight
fhs$BMIGRP[fhs$BMI >= 30] = 4                     # >=30 Obese

# Coding Obese indicator (dichotomous)
fhs$OBESE <- ifelse(fhs$BMI >= 30, 1, 0) # if BMI>=30 is T, OBESE=1
                                         # if BMI>=30 is F, OBESE=0
```

Data

BMI	BMIGRP_factor	OBESE_factor
26.97	Overweight	No
28.73	Overweight	No
25.34	Overweight	No
28.58	Overweight	No
23.10	Normal	No
30.30	Obese	Yes

Progress this Unit

Population Parameters

- It is important to understand the type of variable you are analyzing

Variable Type	Population Parameter
Quantitative	Population mean, μ
Categorical	Population proportion, p

Frequency Table

- **Categorical variables** are summarized using tabular descriptions
- Frequency table is a listing of the values a variable can take and the number of observations in each category
- Relative frequency is the proportion of the total number of observations in each category. Can also be written as a percentage (%).

Relative Frequency (%)

$$\text{relative frequency}_j = \frac{\text{count in category } j}{\text{total number of observations}} \times 100$$

Frequency Table

R Code, Categorizing a Quantitative Variable

```
# Grouping cigarettes smoked/day (ordinal)
fhs$CIGPDAYGRP[fhs$CIGPDAY == 0] = 0
fhs$CIGPDAYGRP[fhs$CIGPDAY >= 1 & fhs$CIGPDAY < 20] = 1   # 1-19 cigarettes/day
fhs$CIGPDAYGRP[fhs$CIGPDAY >= 20 & fhs$CIGPDAY < 40] = 2  # 20-39 cigarettes/day
fhs$CIGPDAYGRP[fhs$CIGPDAY >= 40] = 3                     # 40+ cigarettes/day

# Option 1: Using mutate() function in dplyr package to create factor variable
library(dplyr)
fhs <- dplyr::mutate(fhs,
                     CIGPDAYGRP_factor = factor(CIGPDAYGRP,
                                                levels = c(0, 1, 2, 3),
                                                labels = c("0", "1-19", "20-39", "40+"),
                                                ordered = TRUE))    # Ordinal variable

# Option 2: Using traditional factor() function to create factor variable
fhs$CIGPDAYGRP_factor <- factor(fhs$CIGPDAYGRP,levels = c(0, 1, 2, 3),
                                labels = c("0", "1-19", "20-39", "40+"),
                                ordered = TRUE)
```

Frequency Table

R Code and Output, One-Way Frequency Table

```
> tab <- table(fhs$CIGPDAYGRP_factor, dnn = "Cigarettes Per Day")
> tab
Cigarettes Per Day
    0  1-19 20-39   40+
 2253   907  1077   165
> prop.table(tab)
Cigarettes Per Day
         0       1-19      20-39        40+
0.51181281 0.20604271 0.24466152 0.03748296
> table(fhs$CIGPDAYGRP_factor, dnn = "Cigarettes Per Day", useNA = "ifany")
Cigarettes Per Day
    0  1-19 20-39   40+  <NA>
 2253   907  1077   165    32
> prop.table(table(fhs$CIGPDAYGRP_factor, dnn = "Cigarettes Per Day", useNA = "ifany"))
Cigarettes Per Day
         0       1-19      20-39        40+       <NA>
0.50811908 0.20455571 0.24289581 0.03721245 0.00721696
```

Frequency Table

Table 2: Summary Statistics of Categorical FHS Data

	(N=4434)
Cigarettes Smoked Per Day	
Missing	32
0	2253 (51.2%)
1-19	907 (20.6%)
20-39	1077 (24.5%)
40+	165 (3.7%)
BMI Category	
Missing	19
Underweight	57 (1.3%)
Normal	1936 (43.9%)
Overweight	1845 (41.8%)
Obese	577 (13.1%)
Obese	
Missing	19
No	3838 (86.9%)
Yes	577 (13.1%)
Sex	
Male	1944 (43.8%)
Female	2490 (56.2%)
Current Use of Anti-Hypertensive Medication	
Missing	61
Yes	144 (3.3%)
No	4229 (96.7%)

Graphical Display: Bar Plot

- Graphical display used for displaying a **nominal categorical variable** is the bar plot
- A bar plot displays a vertical bar for each category

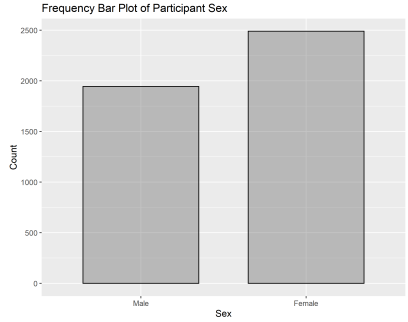
R Code, Bar Plot

```
library(dplyr)

# Frequency bar plot
ggplot(data = fhs, aes(x = SEX_factor, y = stat(count))) +
  geom_bar(col = "black", width = 0.7, alpha = 0.35) +
  labs(title = "Frequency Bar Plot of Participant Sex",
       x = "Sex", y = "Count")

# Relative frequency bar plot
ggplot(data = fhs, aes(x = SEX_factor,
                       y = 100*(stat(count))/sum(stat(count)))) +
  geom_bar(col = "black", width = 0.7, alpha = 0.35) +
  labs(title = "Relative Frequency Bar Plot of Participant Sex",
       x = "Sex", y = "Relative Frequency (%)")
```

Figure: Bar Plot



Graphical Display: Bar Plot

- Graphical display used for displaying a **nominal categorical variable** is the bar plot
- A bar plot displays a vertical bar for each category

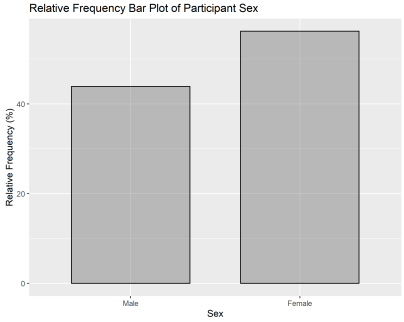
R Code, Bar Plot

```
library(dplyr)

# Frequency bar plot
ggplot(data = fhs, aes(x = SEX_factor, y = stat(count))) +
  geom_bar(col = "black", width = 0.7, alpha = 0.35) +
  labs(title = "Frequency Bar Plot of Participant Sex",
       x = "Sex", y = "Count")

# Relative frequency bar plot
ggplot(data = fhs, aes(x = SEX_factor,
                       y = 100*(stat(count))/sum(stat(count)))) +
  geom_bar(col = "black", width = 0.7, alpha = 0.35) +
  labs(title = "Relative Frequency Bar Plot of Participant Sex",
       x = "Sex", y = "Relative Frequency (%)")
```

Figure: Relative Frequency Bar Plot



Numerical Summaries

- **Quantitative variables** are summarized using measures of center and spread
- Measures of central tendency in the sample
  1. Mean - The average of all the observations

Sample Mean ($\bar{x}$)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

  2. Median - The middle data point, 50^{th} percentile or Q_2
     - Value in position $\frac{n+1}{2}$ of *ordered* data
  3. Mode - The most frequently occurring value

Numerical Summaries

R Code and Output, Measures of Center

```
> mean(fhs$BMI)
[1] NA
> sum(is.na(fhs$BMI))                    # number of missing values
[1] 19
> mean(fhs$BMI, na.rm = TRUE)
[1] 25.84616
> median(fhs$BMI, na.rm = TRUE)
[1] 25.45
> tab <- table(fhs$BMI)                  # number of occurrences for each unique value
> names(sort(tab, decreasing = TRUE)[1])   # mode
[1] "23.48"

> summary(fhs$BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  15.54   23.09   25.45   25.85   28.09   56.80      19
```

Numerical Summaries

- Measures of spread in the sample
  1. Range = Maximum − Minimum or (Minimum, Maximum)
  2. Inter-quartile range (IQR) $= Q_3 - Q_1$ or (Q_1, Q_3)
  3. Variance

Sample Variance (s^2)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1}$$

  4. Standard deviation $= s = +\sqrt{s^2}$

Numerical Summaries

R Code and Output, Measures of Spread

```
> min(fhs$BMI, na.rm = TRUE)
[1] 15.54
> max(fhs$BMI, na.rm = TRUE)
[1] 56.8
> range(fhs$BMI, na.rm = TRUE)
[1] 15.54 56.80
> quantile(fhs$BMI, na.rm = TRUE)      # quantile(fhs$BMI, probs=0.75, na.rm = TRUE) for
   0%   25%   50%   75%  100%          #                              specific percentile
15.54 23.09 25.45 28.09 56.80
> fivenum(fhs$BMI, na.rm = TRUE)       # min, Q1, median, Q3, max
[1] 15.54 23.09 25.45 28.09 56.80
> IQR(fhs$BMI, na.rm = TRUE)
[1] 5

> var(fhs$BMI, na.rm = TRUE)
[1] 16.82493
> sd(fhs$BMI, na.rm = TRUE)
[1] 4.101821
```

Frequency Table

- With **continuous** or **discrete data**, there could be many unique values

R Code and Output, Frequency Table

```
> table(fhs$BMI)
15.54 15.96 16.48 16.59 16.61 16.69 16.71 16.73 16.75 16.87
    1     1     1     2     1     1     1     1     1     1
16.92 16.98 17.11 17.17 17.23 17.32 17.38 17.44 17.48  17.5
    1     1     1     1     1     1     1     1     1     1
17.51 17.61 17.64 17.65 17.68 17.71 17.81 17.84 17.89 17.92
    1     2     1     1     1     1     2     1     1     1
[entries omitted]
```

Interval	Count	Relative Frequency
$[15, 20)$	215	4.87%
$[20, 25)$	1778	40.27%
$[25, 30)$	1845	41.79%
$[30, 35)$	447	10.12%
$[35, 40)$	101	2.29%
$[40, 45)$	25	0.57%
$[45, 50)$	2	0.05%
$[50, 55)$	1	0.02%
$[55, 60)$	1	0.02%
	4415	**100%**

- Summarize data: Group quantitative variable into categories or bins and count number of observations in each bin

Histogram

- Three primary graphical displays used for **quantitative variables** are the histogram, boxplot and the normal quantile-quantile (Q-Q) plot

1. A histogram depicts a frequency distribution for discrete, continuous data or ordinal categorical data
   - Horizontal axis displays the limits of the intervals after grouping data into bins
   - Vertical axis depicts either the frequency or the relative frequency of observations within each interval
   - To represent the continuity in the variable analyzed, histogram bars do not have gaps between them

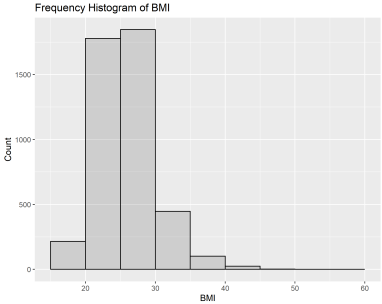
Histogram

R Code, Histogram

```
# Frequency histogram
ggplot(data = fhs, aes(x = BMI)) +
  geom_histogram(breaks = seq(15, 60, by = 5),
                 col = "black", alpha = 0.2,
                 closed = "left") +
  labs(title = "Frequency Histogram of BMI", x = "BMI",
       y = "Count")
```

- Gives idea of shape of distribution
- Bin width important

Figure: Histogram, Bin Width = 5



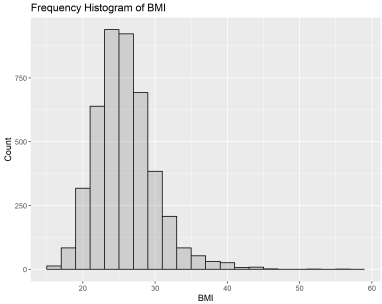
Histogram

R Code, Histogram

```
# Frequency histogram
ggplot(data = fhs, aes(x = BMI)) +
  geom_histogram(breaks = seq(15, 60, by = 2),
                 col = "black", alpha = 0.2,
                 closed = "left") +
  labs(title = "Frequency Histogram of BMI", x = "BMI",
       y = "Count")
```

- Gives idea of shape of distribution
- Bin width important

Figure: Histogram, Bin Width = 2



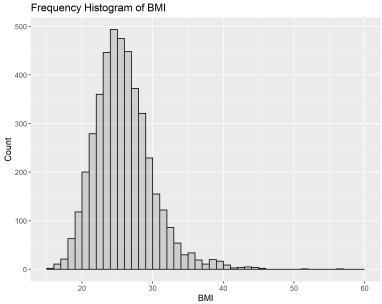
Histogram

R Code, Histogram

```
# Frequency histogram
ggplot(data = fhs, aes(x = BMI)) +
  geom_histogram(breaks = seq(15, 60, by = 1),
                 col = "black", alpha = 0.2,
                 closed = "left") +
  labs(title = "Frequency Histogram of BMI", x = "BMI",
       y = "Count")
```

- Gives idea of shape of distribution
- Bin width important

Figure: Histogram, Bin Width = 1



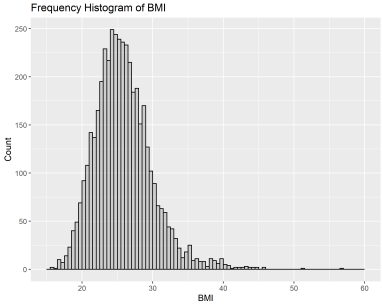
Histogram

R Code, Histogram

```
# Frequency histogram
ggplot(data = fhs, aes(x = BMI)) +
  geom_histogram(breaks = seq(15, 60, by = 0.5),
                 col = "black", alpha = 0.2,
                 closed = "left") +
  labs(title = "Frequency Histogram of BMI", x = "BMI",
       y = "Count")
```

- Gives idea of shape of distribution
- Bin width important

Figure: Histogram, Bin Width = 0.5



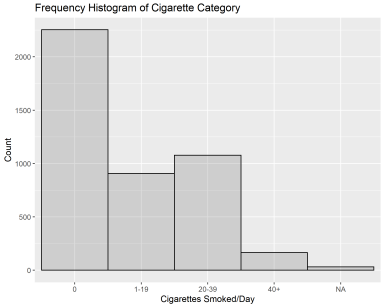
Histogram

R Code, Histogram with Ordinal Variable

```
# Frequency "histogram" of ordinal variable
ggplot(data = fhs, aes(x = CIGPDAYGRP_factor)) +
  geom_bar(col = "black", alpha = 0.2,
           width = 1) +           # increasing bar width
  labs(title = "Frequency Histogram of Cigarette Category",
       x = "Cigarettes Smoked/Day", y = "Count")

# Frequency "histogram" of ordinal variable - removing NAs
ggplot(data = subset(fhs, !is.na(CIGPDAYGRP_factor)),
       aes(x = CIGPDAYGRP_factor)) +
  geom_bar(col = "black", alpha = 0.2, width = 1) +
  labs(title = "Frequency Histogram of Cigarette Category",
       x = "Cigarettes Smoked/Day", y = "Count")
```

Figure: Histogram, Ordinal Variable



- Main difference between nominal and ordinal data is that ordinal data categories suggest a certain display order

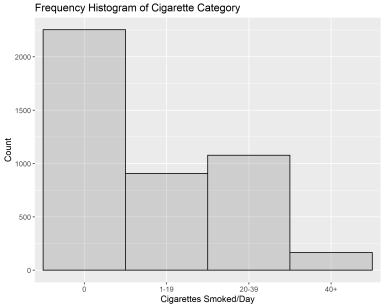
Histogram

R Code, Histogram with Ordinal Variable

```
# Frequency "histogram" of ordinal variable
ggplot(data = fhs, aes(x = CIGPDAYGRP_factor)) +
  geom_bar(col = "black", alpha = 0.2,
           width = 1) +           # increasing bar width
  labs(title = "Frequency Histogram of Cigarette Category",
       x = "Cigarettes Smoked/Day", y = "Count")

# Frequency "histogram" of ordinal variable - removing NAs
ggplot(data = subset(fhs, !is.na(CIGPDAYGRP_factor)),
       aes(x = CIGPDAYGRP_factor)) +
  geom_bar(col = "black", alpha = 0.2, width = 1) +
  labs(title = "Frequency Histogram of Cigarette Category",
       x = "Cigarettes Smoked/Day", y = "Count")
```

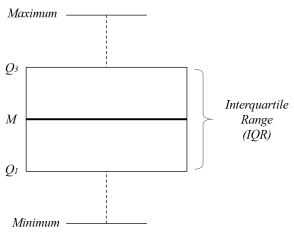
Figure: Histogram, Ordinal Variable



- Main difference between nominal and ordinal data is that ordinal data categories suggest a certain display order

Boxplot

2. A boxplot (box-and-whisker plot) depicts the five-number summary
   - Five-number summary - Summary statistic that includes the five most important sample percentiles:
     1. Maximum
     2. Q_3: 75^{th} percentile
     3. Median
     4. Q_1: 25^{th} percentile
     5. Minimum



- Note: A potential outlier is an observation more than $1.5 \times IQR$ below Q_1 or above Q_3

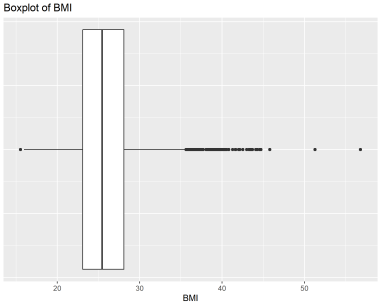
Boxplot

R Code, Boxplot

```
# Horizontal boxplot
ggplot(data = fhs, aes(x = BMI)) +
  geom_boxplot() +
  theme(axis.title.y = element_blank(),   # remove y-axis title
        axis.text.y = element_blank(),    # remove labels
        axis.ticks.y = element_blank()) + # remove tick marks
  labs(title = "Boxplot of BMI", x = "BMI", y = "")
```

- Information displayed by boxplot:
  - Location (median)
  - Spread (IQR and range)
  - Presence of outliers
  - Some information about shape

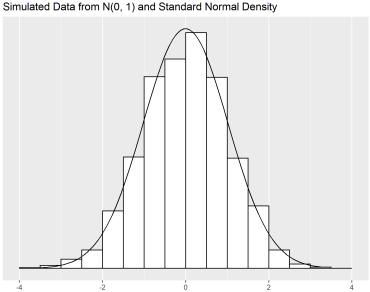
Figure: Boxplot



Q-Q Plot

3. In general, a Q–Q plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other; linearity of the Q-Q plot suggests that the two samples come from a common population distribution

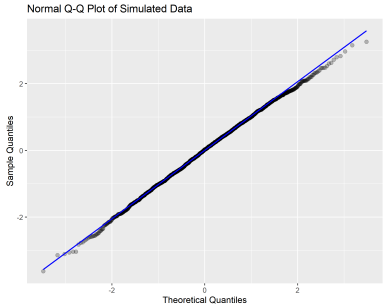
Figure: Standard Normal Distribution



- Typically used to assess if a sample comes from a Normal distribution (i.e., Normal Q-Q plot)
- Many statistical inference procedures assume normality

Normal Q-Q Plot

Figure: Normal Q-Q Plot



- Plot ordered observed sample values (sample quantiles) versus the expected theoretical quantiles based on the Normal distribution
- If the data are approximately normally distributed, the plot should be roughly a straight line

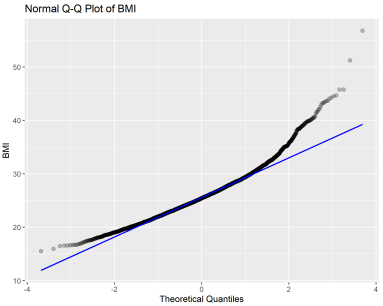
Normal Q-Q Plot

R Code, Normal Q-Q Plot

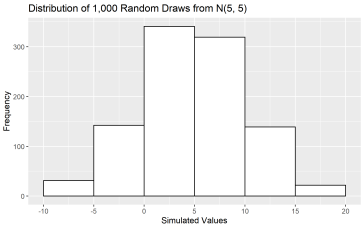
```
ggplot(data = fhs, aes(sample = BMI)) +
  stat_qq(size = 2, alpha = 0.25) +
  stat_qq_line(size = 0.75, color = "blue") +
  labs(title = "Normal Q-Q Plot", x = "Theoretical Quantiles",
       y = "BMI")
```

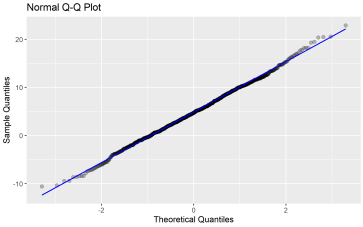
- BMI variable does not appear normally distributed
- Largest values in our sample are larger than would be expected if they came from a normally distributed population (heavy right tail)

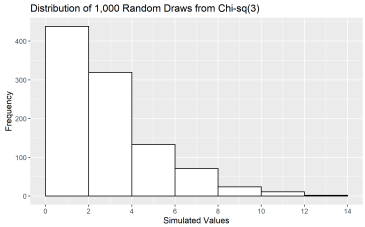
Figure: Normal Q-Q Plot

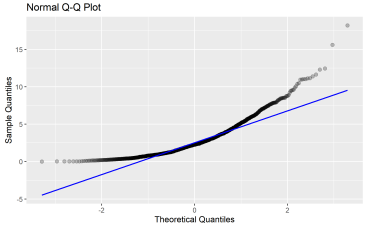


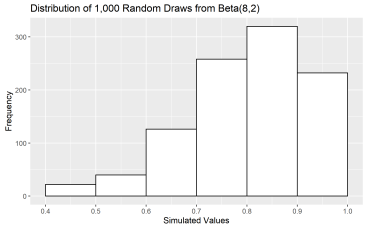
Normal Q-Q Plot

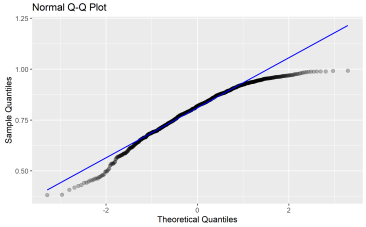












Progress this Unit

Relationships among Variables

- In this course, we are primarily interested in examining whether there is a relationship among variables
- Identify whether a variable is being predicted by the remaining variables or whether it is being used to make the prediction

Response variable	←	Explanatory variables
Outcome variable		Predictor variables
Dependent variable		Independent variables
Primary endpoint		(Exposure variable, confounders)
Y		X

- Methods of description and analysis are driven by the type of outcome variable

Public Health Application: FDA Guidance on Multiple Endpoints

Public Health Application

Because most diseases have more than one consequence, many trials are designed to examine the effect of a drug on more than one endpoint. . . . When the rate of occurrence of a single event is expected to be low, it is common to combine several events (e.g., death/ICU admission within 30 days in hospitalized COVID patients) in a "composite event endpoint" where the occurrence of any of the events would constitute an "endpoint event."

When there are many endpoints prespecified in a clinical trial, they are usually classified into three families: primary, secondary, and exploratory. Secondary endpoints may be selected to demonstrate additional effects after success on the primary endpoint. For instance, a drug may demonstrate effectiveness on the primary endpoint of survival, after which the data regarding an effect on a secondary endpoint, such as functional status, would be tested.

Link to FDA guidance

Mean or Proportion Confusion

Poll

- In a population of hypertensive individuals, suppose we want to compare the efficacy of an antihypertensive drug vs. placebo in a 12-week RCT

Response variable ← Explanatory variable

- Possible endpoints:
  - Change in systolic blood pressure over study period
  - Achievement of $\geq 20\%$ reduction in systolic BP from BL by 12 weeks
  - Achievement of target systolic BP of 120 mmHg by 12 weeks
  - Stroke, MI, or death (would require longer study)
- **Link to poll** (Lesson 01): *https://pollev.com/bis505b*

Cross-tabulation/Contingency Table

- A cross-tabulation specifies the joint frequency distribution of **two categorical variables**
- An association between two dichotomous variables can be displayed in a 2×2 contingency table:

	Success (D)	Failure ($\bar{D}$)	Row Sum
Exposed (E)	a	b	$a+b$
Unexposed ($\bar{E}$)	c	d	$c+d$
	$a+c$	$b+d$	N

- $\hat{p}_1 = \hat{P}(D|E) = \dfrac{a}{a+b}$
- $\hat{p}_2 = \hat{P}(D|\bar{E}) = \dfrac{c}{c+d}$

} Row proportions

Example: 2×2 Table

- Example: Results of study investigating the effectiveness of antihypertensive drug in reducing systolic BP. 150 individuals followed over 12 weeks.

	$\geq 20\%$ Reduction Achieved		
	Yes	No	Row Sum
Drug	33	42	75
Placebo	15	60	75
	48	102	150

- $\hat{p}_1 = \hat{P}(\text{Reduction}|\text{Drug}) =$
- $\hat{p}_2 = \hat{P}(\text{Reduction}|\text{Placebo}) =$

2×2 Table

R Code and Output, Two-Way Frequency Table

```
# Input "levels" as "1" then "0" for correct table setup
> rawdata <- dplyr::mutate(rawdata,
                           drug_factor = factor(drug,
                                                levels = c(1, 0),
                                                labels = c("Drug", "Placebo")),
                           endpoint_factor = factor(endpoint,
                                                    levels = c(1, 0),
                                                    labels = c("Yes", "No")))

# Contingency table (row variable: exposure, column variable: disease)
> tab <- table(rawdata$drug_factor, rawdata$endpoint_factor,
               dnn = c("Treatment", ">= 20% Reduction"))
> tab
         >= 20% Reduction
Treatment Yes No
  Drug     33 42
  Placebo  15 60
```

2×2 Table

R Code and Output, Counts

```
# 2x2 table
> tab
         >= 20% Reduction
Treatment Yes No
  Drug     33 42
  Placebo  15 60

# Row totals
> rowSums(tab)
   Drug Placebo
     75      75

# Column totals
> colSums(tab)
Yes  No
 48 102
```

R Code and Output, Proportions

```
# Cell proportions
> prop.table(tab)
         >= 20% Reduction
Treatment  Yes   No
  Drug    0.22 0.28
  Placebo 0.10 0.40

# Row proportions
> prop.table(tab, margin = 1)
         >= 20% Reduction
Treatment  Yes   No
  Drug    0.44 0.56
  Placebo 0.20 0.80

# Column proportions
> prop.table(tab, margin = 2)
         >= 20% Reduction
Treatment       Yes        No
  Drug    0.6875000 0.4117647
  Placebo 0.3125000 0.5882353
```

Numerical Measure of Association: Odds Ratio

- Odds ratio: Measure of association commonly used between exposure and outcome
  - $\text{OR} = \dfrac{\text{Odds of outcome in exposed}}{\text{Odds of outcome in unexposed}} = \dfrac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}} = \dfrac{ad}{bc}$
  - > 1: Exposure increases risk of outcome
  - < 1: Exposure protective against outcome
  - $= 1$: Exposure has no effect on outcome

	$\geq 20\%$ Reduction Achieved	
	Yes	No
Drug	33 (44%)	42 (56%)
Placebo	15 (20%)	60 (80%)

$\hat{\text{OR}} =$

Odds Ratio

R Code and Output, Two-Way Frequency Table and Odds Ratio

```
> library(epiR)
> epi <- epi.2by2(tab, method = "cohort.count", units = 1) # units=1 for proportions
> epi                                                      # print results
             Outcome +    Outcome -       Total        Inc risk *        Odds
Exposed +           33           42          75              0.44       0.786
Exposed -           15           60          75              0.20       0.250
Total               48          102         150              0.32       0.471

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                2.20 (1.31, 3.70)
Odds ratio                                    3.14 (1.52, 6.50)
Attrib risk *                                 0.24 (0.10, 0.38)
Attrib risk in population *                   0.12 (0.00, 0.24)
Attrib fraction in exposed (%)               54.55 (23.57, 72.97)
Attrib fraction in population (%)            37.50 (10.21, 56.49)
-------------------------------------------------------------------
 Test that OR = 1: chi2(1) = 9.926 Pr>chi2 = 0.00
 Wald confidence limits      CI: confidence interval      * Outcomes per population unit
```

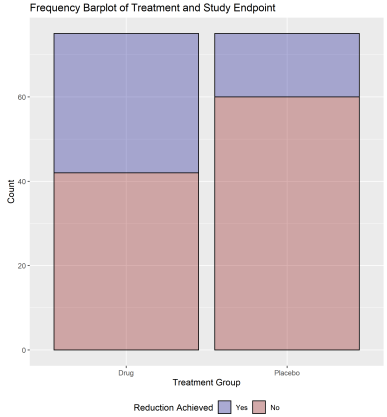
Barplots by Group

R Code, Barplots by Group

```
# Stacked frequency barplot by group
ggplot(data = rawdata, aes(x = drug_factor, y = stat(count),
                           fill = endpoint_factor)) +
  geom_bar(col = "black", alpha = 0.3) +
  labs(title = "Frequency Barplot of Treatment and Study
       Endpoint", x = "Treatment Group", y = "Count",
       fill = "Reduction Achieved") +
  scale_fill_manual(values = c("darkblue", "darkred")) +
  theme(legend.position = "bottom")

# Side-by-side frequency barplot
# Code identical to above except geom_bar() function
# includes position = position_dodge() option
#  ... geom_bar(col = "black", alpha = 0.3,
#           position = position_dodge()) + ...
```

Figure: Barplots of Number Responders within Treatment



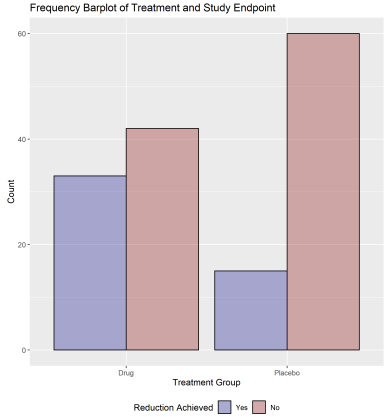
Barplots by Group

R Code, Barplots by Group

```
# Stacked frequency barplot by group
ggplot(data = rawdata, aes(x = drug_factor, y = stat(count),
                           fill = endpoint_factor)) +
  geom_bar(col = "black", alpha = 0.3) +
  labs(title = "Frequency Barplot of Treatment and Study
       Endpoint", x = "Treatment Group", y = "Count",
       fill = "Reduction Achieved") +
  scale_fill_manual(values = c("darkblue", "darkred")) +
  theme(legend.position = "bottom")

# Side-by-side frequency barplot
# Code identical to above except geom_bar() function
# includes position = position_dodge() option
#  ... geom_bar(col = "black", alpha = 0.3,
#           position = position_dodge()) + ...
```

Figure: Barplots of Number Responders within Treatment

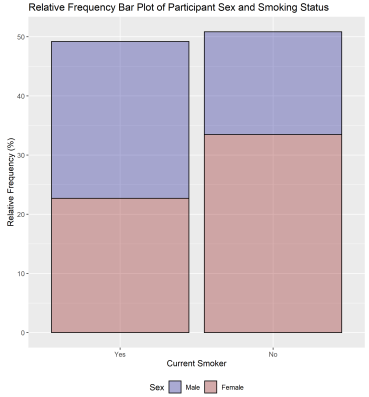


Barplots by Group

R Code and Output, Two-Way Frequency Table

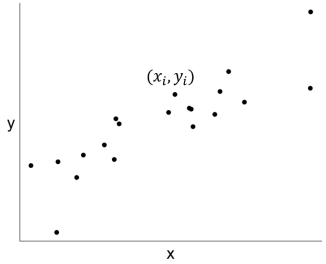
```
# FHS example: Smoking and Sex
> tab <- table(fhs$SEX_factor, fhs$CURSMOKE_factor,
               dnn = c("Sex", "Smoker"))
> tab
        Smoker
Sex       Yes   No
  Male   1175  769
  Female 1006 1484
> prop.table(tab)
        Smoker
Sex              Yes        No
  Male   0.2649977 0.1734326
  Female 0.2268832 0.3346865
> colSums(prop.table(tab))     # proportion smokers and non-smokers
      Yes        No
0.4918809 0.5081191
> prop.table(tab, margin = 2)  # column proportions P(M|S=Y), etc.
        Smoker
Sex              Yes        No
  Male   0.5387437 0.3413227
  Female 0.4612563 0.6586773
```

Figure: `ggplot(data=fhs, aes(x=CURSMOKE_factor, y=100*(stat(count))/sum(stat(count)), fill=SEX_factor)) + ...`



Scatterplot

- A two-way scatterplot visually examines the relationship between **two quantitative variables**
- Each point on the graph represents a combination of values (x_i, y_i)



- Explanatory variable plotted on the x (horizontal) axis
- Response variable plotted on the y (vertical) axis
- If there is no explanatory-response distinction, either variable can be displayed on the either axis

Scatterplot

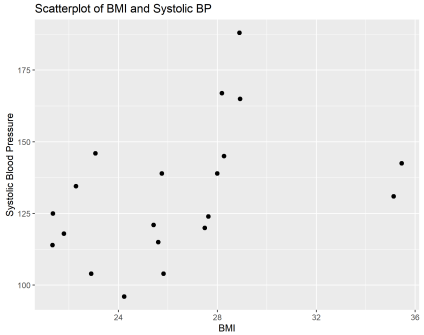
R Code, Scatterplot

```
# Select random sample of 20 observations from fhs
fhs20 <- fhs[sample(nrow(fhs), 20), ]

ggplot(fhs20, aes(x = BMI, y = SYSBP)) +
  geom_point(size = 2, shape = 19) +
  labs(title = "Scatterplot of BMI and Systolic BP",
       x = "BMI", y = "Systolic Blood Pressure")
```

- There is a positive association between BMI and systolic blood pressure

Figure: Scatterplot

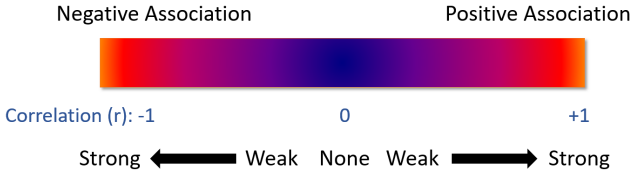


Interpreting Scatterplots

- After plotting two variables on a scatterplot, we describe the overall pattern of the relationship
- You can describe the overall pattern of a scatterplot by the direction, form, and strength of the relationship
  - **Form**: linear, non-linear, clusters, no pattern
  - **Direction**: positive, negative, no direction
  - **Strength**: how closely the points fit the "form"

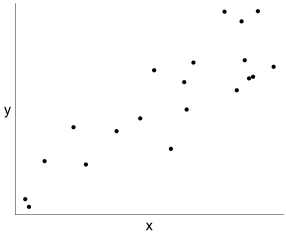
Pearson Correlation

- Pearson correlation (r): Numerical summary of the linear relationship between x and y in the sample
  - *Strength* of the linear association (strong, weak)
  - *Direction* of the association (positive, negative)



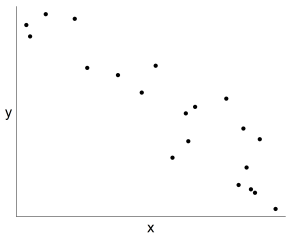
Direction

Positive Association



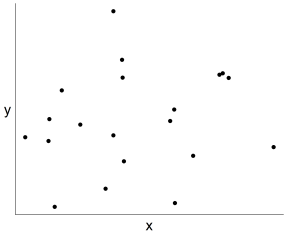
$r > 0$: $X \uparrow, Y \uparrow$ and $X \downarrow, Y \downarrow$

Negative Association



$r < 0$: $X \downarrow, Y \uparrow$ and $X \uparrow, Y \downarrow$

No Linear Association



$r = 0$: No linear relationship

Pearson Correlation

R Code, Pearson Correlation

```
# Select subset of variables
> fhsselect <- fhs20 %>% select(BMI, SYSBP)
> cor(x = fhsselect, method = "pearson", use = "pairwise.complete.obs")
          BMI   SYSBP
BMI   1.00000 0.41819
SYSBP 0.41819 1.00000
```

- The correlation between BMI and systolic blood pressure is equal to 0.42: moderate positive linear association
- Each variable is perfectly correlated with itself ($r = 1$)
- Correlation matrix is symmetric (i.e., the [1,2] entry = [2,1] entry)

Associations between Quantitative and Categorical Variables

- When looking for associations between a **quantitative variable** and a **categorical variable**, one-variable description strategies of the quantitative variable are presented by level of the categorical variable
  - Summary statistics by group
  - Histograms by group
  - Boxplots by group

Table 3: Summarizing Some FHS Variables by Sex

Sex	Male (N=1944)	Female (N=2490)
Age (years)		
Missing	0	0
Mean (SD)	49.8 (8.7)	50.0 (8.6)
Median (Range)	49.0 (33.0, 69.0)	49.0 (32.0, 70.0)
Body Mass Index		
Missing	5	14
Mean (SD)	26.2 (3.4)	25.6 (4.6)
Median (Range)	26.1 (15.5, 40.4)	24.8 (16.0, 56.8)
Total Cholesterol		
Missing	7	45
Mean (SD)	233.6 (42.4)	239.7 (46.2)
Median (Range)	231.0 (113.0, 696.0)	237.0 (107.0, 600.0)
Systolic Blood Pressure		
Missing	0	0
Mean (SD)	131.7 (19.4)	133.8 (24.5)
Median (Range)	129.0 (83.5, 235.0)	128.5 (83.5, 295.0)
Diastolic Blood Pressure		
Missing	0	0
Mean (SD)	83.7 (11.4)	82.6 (12.5)
Median (Range)	82.0 (48.0, 136.0)	81.0 (50.0, 142.5)

Summary Statistics by Group

R Code, Means by Group

```
# aggregate() function
> aggregate(x = list(meanage = fhs$AGE, meanbmi = fhs$BMI, meantotchol = fhs$TOTCHOL,
                     meansysbp = fhs$SYSBP, meandiabp = fhs$DIABP),
            by = list(sex = fhs$SEX_factor), FUN = mean, na.rm = TRUE)
     sex  meanage  meanbmi meantotchol meansysbp meandiabp
1   Male 49.78652 26.16958    233.5798  131.7369  83.70885
2 Female 50.03454 25.59288    239.6814  133.8219  82.59538

# Can also apply our own functions in aggregate()
> miss <- function(x){       # count number of missing values
    sum(is.na(x))
  }

> aggregate(x = list(meanage = fhs$AGE, meanbmi = fhs$BMI, meantotchol = fhs$TOTCHOL,
                     meansysbp = fhs$SYSBP, meandiabp = fhs$DIABP),
            by = list(sex = fhs$SEX_factor), FUN = miss)
     sex meanage meanbmi meantotchol meansysbp meandiabp
1   Male       0       5           7         0         0
2 Female       0      14          45         0         0
```

Summary Statistics by Group

R Code, Means by Group

```
# Apply functions to subsets of individuals (e.g., males, females separately)
> mean(fhs$AGE[which(fhs$SEX_factor == "Male")], na.rm = TRUE)
[1] 49.78652
> mean(fhs$AGE[which(fhs$SEX_factor == "Female")], na.rm = TRUE)
[1] 50.03454

# Using subset() function
> mean(subset(fhs, SEX_factor == "Male")$AGE, na.rm = TRUE)
[1] 49.78652
> mean(subset(fhs, SEX_factor == "Female")$AGE, na.rm = TRUE)
[1] 50.03454
```

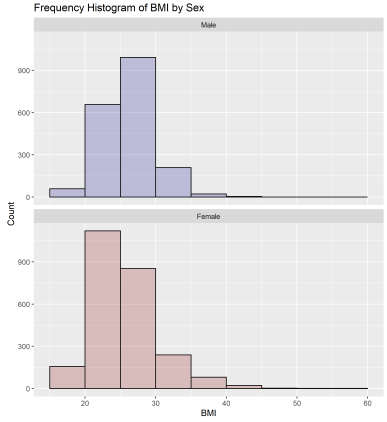
Histograms by Group

R Code, Histograms by Group

```
# Histograms by group
h <- ggplot(fhs, aes(x = BMI, fill = SEX_factor)) +
  geom_histogram(breaks = seq(15, 60, by = 5),
                 col = "black",
                 alpha = 0.2,
                 closed = "left") +
  labs(title = "Frequency Histogram of BMI by Sex",
       x = "BMI", y = "Count", fill = "Sex") +
  scale_fill_manual(values = c("darkblue", "darkred")) +
  theme(legend.position = "none")

# Panel plots, 1 column with 1 row for each SEX
h + facet_wrap(~ SEX_factor, ncol = 1)
```

Figure: Histograms by Sex

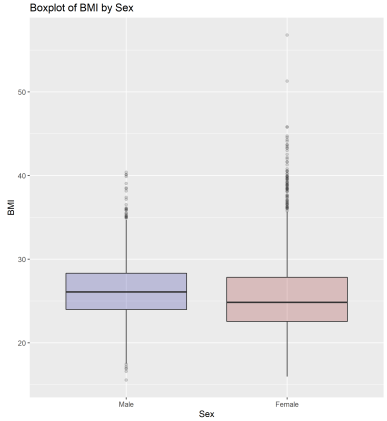


Boxplots by Group

R Code, Boxplots by Group

```
# Boxplots by group
ggplot(data = fhs, aes(x = SEX_factor, y = BMI,
                       fill = SEX_factor)) +
  geom_boxplot(alpha = 0.2) +
  labs(title = "Boxplot of BMI by Sex",
       x = "Sex", y = "BMI", fill = "Sex") +
  scale_fill_manual(values = c("darkblue", "darkred")) +
  theme(legend.position = "none")
```

Figure: Boxplots by Sex



Multivariable Descriptions

- Description of more than 2-3 variables simultaneously becomes difficult
- One approach is to look at pairwise associations
  - **Categorical variables**: Series of two-way tables
  - **Quantitative variables**: Correlation matrix or scatterplot matrix

Correlation Matrix and Scatterplot Matrix

R Code and Output, Correlation and Scatterplot Matrix

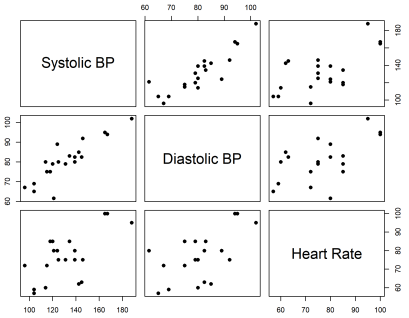
```
# Select random sample of 20 observations from fhs
> fhs20 <- fhs[sample(nrow(fhs), 20), ]

# Select subset of variables
> fhsselect2 <- fhs20 %>% select(SYSBP, DIABP, HEARTRTE)

> cor(x = fhsselect2, method = "pearson",
      use = "pairwise.complete.obs")
             SYSBP     DIABP  HEARTRTE
SYSBP     1.0000000 0.8730936 0.6470337
DIABP     0.8730936 1.0000000 0.5829046
HEARTRTE 0.6470337 0.5829046 1.0000000

# Scatterplot matrix
> pairs(fhsselect2, pch = 19, labels = c("Systolic BP",
                        "Diastolic BP", "Heart Rate"))

# Equivalently,
> pairs(~ SYSBP + DIABP + HEARTRTE, data = fhs20, pch = 19)
```



Lesson Summary

- Exploratory and descriptive measures (summary statistics and graphs) uncover properties of the data
- The key distinction for statistical analysis is between **categorical** and **quantitative** variables
- The type of variable(s) being analyzed determines the methods that should be used. Applies to...
  - exploratory and descriptive measures,
  - basic statistical analyses, and
  - regression modeling