Summarizing Data with Statistics*

Alan T. Arnholt  
Department of Mathematical Sciences  
Appalachian State University  
arnholt@math.appstate.edu

Spring 2006 R Notes
Summary Measures
   Overview of Summary Measures

Using mean() and median()
   Using R for Basic Statistics

Quantiles
   Overview of Quantiles

Measures of Variability
   Range and $IQR$

The R Script
• Numerical summaries of the population are called **parameters**.
Parameters and Statistics

- Numerical summaries of the population are called parameters.
- Numerical summaries of the sample are called statistics.
Parameters that Measure the Center of a Population Distribution

- The **population mean**, $\mu$, measures the center of a distribution.
Parameters that Measure the Center of a Population Distribution

- The **population mean**, \( \mu \), measures the center of a distribution.
- The **population median**, \( \theta \), divides the distribution in half.
Statistics That Measure the Center of Data

• Given some numeric data $x_1, x_2, \ldots, x_n$, the sample mean is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$  (1)
Statistics That Measure the Center of Data

- Given some numeric data $x_1, x_2, \ldots, x_n$, the sample mean is defined as

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

- When a $p\%$ trimmed mean is computed, $p\%$ of the sorted data is deleted from each end of the distribution, and a mean is computed from the remaining values.
Statistics That Measure the Center of Data

- Given some numeric data $x_1, x_2, \ldots, x_n$, the sample mean is defined as

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

- When a $p\%$ trimmed mean is computed, $p\%$ of the sorted data is deleted from each end of the distribution, and a mean is computed from the remaining values.

- The sample median, $M$, of $x_1, x_2, \ldots, x_n$ is the $\left(\frac{n+1}{2}\right)^{st}$ observation of the sorted values. When $n$ is odd, $\frac{n+1}{2}$ is an integer, and finding the observation is straightforward. When $n$ is even, an average of the two middle observations is taken to find the median.
Example Using Data from BSDA

Use the data frame `Challeng` from the BSDA package to find the average recorded temperature, a 10% trimmed mean of the recorded temperature, and the median temperature of the shuttle.
Mean and Trimmed Mean

```r
> library(BSDA)
> attach(Challeng)
> SUM <- sum(temp)
> n <- length(temp)
> MEAN <- SUM/n
> MEAN
[1] 68.44
> mean(temp)
[1] 68.44
> p <- .1
> pd <- p*n
> pd
[1] 2.5
> ntd <- floor(pd)
> ntd
[1] 2
```
Mean and Trimmed Mean - Continued

> temp
[1]  66  70  69  80  68  67  72  73  70  57  63  70  78  67  53  67  75
[18]  70  81  76  79  75  76  58  31
> STEMP <- sort(temp)
> STEMP
[1]  31  53  57  58  63  66  67  67  67  68  69  70  70  70  70  72  73
[18]  75  75  76  76  78  79  80  81
> windsor <- STEMP[(ntd+1):(n-ntd)]
> windsor
[1]  57  58  63  66  67  67  67  68  69  70  70  70  70  72  73  75  75
[18]  76  76  78  79
> TM <- mean(windsor)
> TM
[1] 69.80952
> mean(temp,trim=.1)
[1] 69.80952
Median

\[
> MV <- \frac{n+1}{2}
\]
\[
> MV
\]
\[1\] 13

\[
> MEDIAN <- STEM[MV]
\]
\[
> MEDIAN
\]
\[1\] 70

\[
> \text{median}(\text{temp})
\]
\[1\] 70
Definition of Quantiles

The $p^{th}$ quantile, $0 \leq p \leq 1$, of a distribution is the $(p(n - 1) + 1)^{st}$ order statistic. When $p(n - 1) + 1$ is not an integer, linear interpolation is used between order statistics to arrive at the $p^{th}$ quantile. Given values $x_1, x_2, \ldots, x_n$, the $p^{th}$ quantile for the $k^{th}$ order statistic, $p(k)$, is

$$p(k) = \frac{(k - 1)}{(n - 1)}, \quad k \leq n.$$  

(2)
Quantiles and Median

By the previous definition of a quantile, it is seen that the 50% quantile (50\textsuperscript{th} percentile) is the median since

$$0.50 = \frac{k - 1}{n - 1} \implies k = \frac{n + 1}{2},$$

which by definition is the location of the order statistic that is the median.
Example Using Definition

Use the definition of the $p^{th}$ quantile to find the 25%, 50%, and 75% quantiles of the variable `temp` in the data frame `Challeng`.

```r
> k.25 <- 0.25*(n-1)+1 # Location Q1
> k.50 <- 0.50*(n-1)+1 # Location Q2
> k.75 <- 0.75*(n-1)+1 # Location Q3
> ks <- c(k.25,k.50,k.75) # Combine k.25, k.50, and k.75
> ks
[1]  7 13 19
> STEMP # Sorted temp values
 [1] 31 53 57 58 63 66 67 67 68 69 70 70 70 70 72 73 [18] 75 75 76 76 78 79 80 81
> STEMP[ks] # Extract values ks from STEMP
[1] 67 70 75
```
Using quantile() 

- `quantile(x)` finds the quantiles of `x`. 
Using quantile()

- quantile(x) finds the quantiles of x.
- x must be a numeric vector
Using `quantile()`

- `quantile(x)` finds the quantiles of `x`.
- `x` must be a numeric vector
- `probs` is a numeric vector with values between 0 and 1.
Using `quantile()`

- `quantile(x)` finds the quantiles of `x`.
- `x` must be a numeric vector.
- `probs` is a numeric vector with values between 0 and 1.
- **Note:** other definitions exist for computing quantiles. Type `?quantile` at the R prompt to read about other definitions.
Using quantile()

- `quantile(x)` finds the quantiles of `x`.
- `x` must be a numeric vector
- `probs` is a numeric vector with values between 0 and 1.
- **Note:** other definitions exist for computing quantiles. Type `?quantile` at the R prompt to read about other definitions.
- Using `type=6` computes quantiles with the same definition used by Minitab.
Example using `quantile(x)`

Find the 25%, 50%, and 75% quantiles of the variable `temp` stored in the data frame `Challeng` using the `quantile(x)` function.

```r
> quantile(temp)       # Note the default values
     0%   25%   50%   75%  100%
 31.00 67.00 70.00 75.00 81.00
> quantile(temp,probs=c(0,.25,.5,.75,1))
     0%   25%   50%   75%  100%
 31.00 67.00 70.00 75.00 81.00
> quantile(temp,type=6)  # Compare to values on pg. 37
     0%   25%   50%   75%  100%
31.00 66.50 70.00 75.50 81.00
```
Hinges and Five-Number Summary

An alternative method to calculating quartiles is to compute hinges. The idea behind both quartiles and hinges is to split the data into fourths. When a computer is not available, hinges are somewhat easier to compute by hand than are quartiles. The lower and upper hinges are the \( x(j) \) and \( x(n-j+1) \) order statistics where

\[
j = \left\lfloor \frac{n+1}{2} \right\rfloor + 1.
\]

A **five-number summary** for a data set consists of the smallest value, the lower hinge, the median, the upper hinge and the largest value all of which are computed with R’s function `fivenum()`.
Example Using `fivenum()`

```r
> j <- ( floor((n+1)/2) + 1 )/2
> j
[1] 7
> STEMP
[1]  31  53  57  58  63  66  67  67  67  68  69  70  70  70  70  72  73
[18]  75  75  76  76  78  79  80  81
> STEMP[j]
[1] 67
> STEMP[(n-j+1)]
[1] 75
> fivenum(temp)
[1]  31  67  70  75  81
```
range() and IQR()

- The **range** is the distance between the smallest and largest observations \((\text{max} - \text{min})\).
range() and IQR()

- The **range** is the distance between the smallest and largest observations (max - min).
- `range(x)` returns the values min and max.
The `range` is the distance between the smallest and largest observations (max - min).

- `range(x)` returns the values min and max.
- `x` must be a numeric vector.
The **range** is the distance between the smallest and largest observations ($\text{max} - \text{min}$).

- range(x) returns the values min and max.
- x must be a numeric vector.

The **interquartile range** ($IQR$) is the distance between the first and third sample quartiles, $Q_3 - Q_1$. 

range() and IQR()

- The **range** is the distance between the smallest and largest observations ($\text{max} - \text{min}$).
- range(x) returns the values min and max.
- x must be a numeric vector.
- The **interquartile range** ($IQR$) is the distance between the first and third sample quartiles, $Q_3 - Q_1$.
- IQR(x) computes the $IQR$ of a numeric vector x based on quantiles.
Example Computing $IQR$ and Range

Compute the range and $IQR$ for the variable temp stored in the Challeng data frame.

```
> sort(temp)
[1]  31  53  57  58  63  66  67  67  67  68  69  70  70  70  70  72  73
[18]  75  75  76  76  78  79  80  81
> range(temp)
[1]  31  81
> RANGE <- range(temp)[2] - range(temp)[1]
> RANGE
[1]  50
```

```
> quantile(temp,c(.25,.75))
     25%    75%
67.000 75.000
```

```
> IQR(temp)
[1]  8
```
Sample Variance and Sample Standard Deviation

The **sample variance**, $s^2$, can be thought of as the average squared distance of the sample values from the sample mean. It is not quite the average because we divide by $n - 1$ instead of $n$ in the formula

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n - 1}.$$  

When the positive square root of the sample variance is taken, we have the **sample standard deviation**, $s$. It is often preferable to report the sample standard deviation instead of the variance since the units of measurement for the sample standard deviation are the same as those of the individual data points in the sample.
• The function \texttt{var(x)} computes the sample variance of a numeric vector \texttt{x}. 
The function `var(x)` computes the sample variance of a numeric vector `x`.

The function `sd(x)` computes the sample standard deviation of a numeric vector `x`. 
The function \texttt{var(x)} computes the sample variance of a numeric vector \texttt{x}.

The function \texttt{sd(x)} computes the sample standard deviation of a numeric vector \texttt{x}.

\textbf{Note:} If your data contains missing values, see the help files for \texttt{var()} and \texttt{sd()} for methods to compute the sample variance and sample standard deviation respectively.
Computing the Sample Variance and Sample Standard Deviation

Find the sample variance and sample standard deviation of the following five observations: 31, 55, 75, 78, and 81.

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \]  

\[ > \quad x \leftarrow c(31, 55, 75, 78, 81) \]
\[ > \quad n \leftarrow \text{length}(x) \]
\[ > \quad m.x \leftarrow \text{mean}(x) \]
\[ > \quad m.x \]
\[ \quad [1] \quad 64 \]
\[ > \quad x.mx \leftarrow (x - m.x) \]
\[ > \quad x.mx2 \leftarrow x.mx^2 \]
\[ > \quad \text{values} \leftarrow \text{cbind}(x, x.mx, x.mx2) \]
Continuation of Example

> values
  x  x.mx  x.mx2
[1,] 31   -33  1089
[2,] 55   -9   81
[3,] 75    11  121
[4,] 78   14  196
[5,] 81    17  289
> sum(x.mx2)/4
[1] 444
> s2 <- (sum((x-mean(x))^2))/(n-1)
> s2
[1] 444
> var(x)
[1] 444
Continuation of Example

```r
> s <- sqrt(s2)
> s
[1] 21.07131
> sd(x)
[1] 21.07131
```
Empirical Rule

If a stem-and-leaf plot, histogram, or similar graph reveals a bell-shaped appearance, then:

1. Approximately 68% of the measurements will fall within one standard deviation of the mean.
Empirical Rule

If a stem-and-leaf plot, histogram, or similar graph reveals a bell-shaped appearance, then:

1. Approximately 68% of the measurements will fall within one standard deviation of the mean.

2. Approximately 95% of the measurements will fall within two standard deviations of the mean.
Empirical Rule

If a stem-and-leaf plot, histogram, or similar graph reveals a bell-shaped appearance, then:

1. Approximately 68% of the measurements will fall within one standard deviation of the mean.

2. Approximately 95% of the measurements will fall within two standard deviations of the mean.

3. Approximately 99.7% of the measurements will fall within three standard deviations of the mean.
Normal Areas

The area between 85 and 115 is 0.6827

\[ X \sim \text{Normal} \left( \mu = 100, \sigma = 15 \right) \]

The area between 70 and 130 is 0.9545

\[ X \sim \text{Normal} \left( \mu = 100, \sigma = 15 \right) \]

The area between 55 and 145 is 0.9973

\[ X \sim \text{Normal} \left( \mu = 100, \sigma = 15 \right) \]
Link to the R Script

- Go to my web page
  - **Script for Summarizing Data with Statistics**
- Homework: problems 1.61 - 1.77 odd
- **See me if you need help!**