Introductory Statistics Lectures

Relative Standing

Descriptive Statistics IV

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:48:37 2009)

Contents

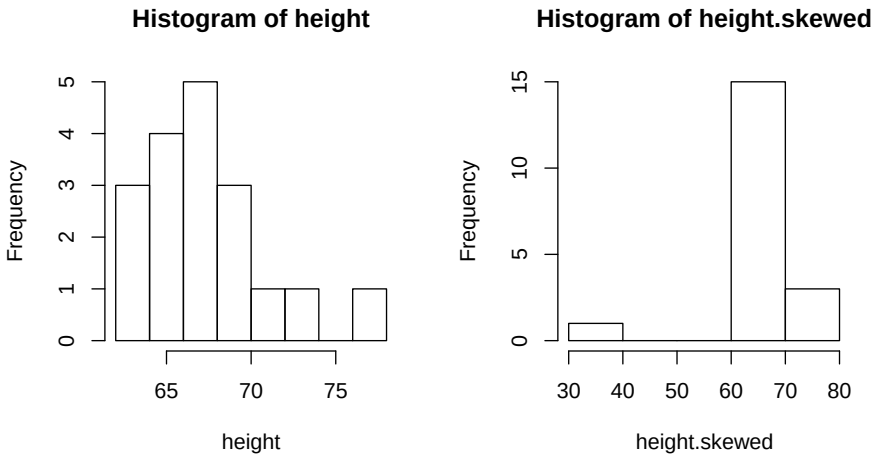
1 Relative Standing

R tip

If you want to keep a plot visible so you can compare it to another plot type `windows()` (on Windows) or `quartz()` (on Mac) to start a new plot window.

Let's setup or two variables of class height data:
Add Mini-Me (Verne Troyer) 2' 8" (32 inches) to make a skewed set of data.

```
R: load("ClassData.RData")
R: height = class.data$height
R: height.skewed = c(height, 32)
R: par(mfrow = c(1, 2))
R: hist(height)
R: hist(height.skewed)
```



Now we will look at methods for measuring the relationship of **individual** data points to the whole data set. Useful for comparing values from different data sets.

How unusual is a specific data point?
How does Mini-Me's height compare to the class as a whole? What is Mini-Me's relative standing?

1.1 Relative Standing

Z SCORES

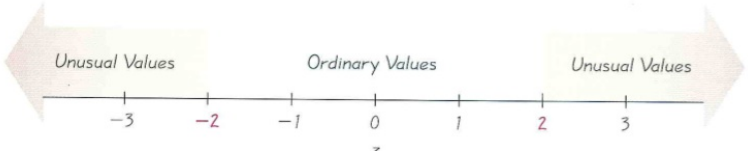
DEFINITION 1.1 z SCORE.

the **number of standard deviations** a given value x is away from the mean. (unit-less)

$$\text{population: } z = \frac{x - \mu}{\sigma} \tag{1}$$

$$\boxed{\text{sample: } z = \frac{x - \bar{x}}{s}} \tag{2}$$

Thus, **unusual value if** $|z| > 2$.



Treating our class with Mini-Me as a population, for the height data: $\mu = 66$ in, $\sigma = 9$ in.

Question 1. Find the z score for Mini-Me at 32 inches

Question 2. Is Mini-Me's height unusual? Why?

Illustration z score properties

Let's convert our class data to z scores:

```
R: z.height = (height - mean(height))/sd(height)
```

The z scores for heights are now: { -0.68, 0.1, 0.88, -0.42, 0.1, -0.68, -1.5, 0.1, 2.4, -1.5, 0.36, 0.62, -0.68, -1.2, -0.16, 1.4, 0.1, 0.62 }

```
R: mean(z.height)
[1] -8.265e-16
R: sd(z.height)
[1] 1
```

Thus, for any set of data converted to z scores:

Question 3. What is the mean of an arbitrary set of z scores?

Question 4. What is the standard deviation of an arbitrary set of z scores?

PERCENTILES

PERCENTILE P_k. DEFINITION 1.2

For a given data point x, the percentile of x is the **percent of data points less than** x.

The kth percentile is denoted as P_k.

QUANTILE x. DEFINITION 1.3

The data point x that corresponds to a specific percentile P_k is referred to as a quantile. Quantiles are data points taken at regular intervals from a cumulative distribution function of a random variable. Quantiles taken at quarter intervals are referred to as quartiles.

Example 1. If a student who scored 1100 on the SAT was in the 75th percentile (3rd quartile) then:

$$\overbrace{P_{75}}^{\text{percentile}=0.75} = \overbrace{1100}^{\text{quantile}=1100}$$

Finding the percentile of x

To find the percentile[1] of x_i:

$$\boxed{k = \frac{i - 0.5}{n} \cdot 100\%} \tag{3}$$

Data must be sorted first!

i position (sorted) of data point x.

n total number of data points

If a statistics exam had the following 10 scores:

$$x = \{43, 62, 73, 77, 83, 83, 85, 87, 94, 97\}$$

Question 5. If you scored 97% on the exam, what was your percentile?

Question 6. If you scored 83% on the exam, what was your percentile?

Finding the data point represented by P_k

Remember, k percent of values should be less than x (sorted).

$$P_k = x_i, \quad L = \frac{k}{100} \cdot n \tag{4}$$

- If L is an integer, $i = L + 0.5$ (hence average neighboring values)
- Otherwise $i = L$ rounded up to nearest integer.

If a statistics exam had the following 14 scores:

$$x = \{43, 62, 72, 73, 77, 83, 83, 84, 85, 87, 89, 94, 95, 97\}$$

Question 7. Find P_{75}?

QUARTILES

NITION 1.4

QUARTILES.

break data into four equal parts:

Q_1 P_{25}

Q_2 P_{50} (median)

Q_3 P_{75}

INTERQUARTILE RANGE: IQR. DEFINITION 1.5

$$IQR = Q_3 - Q_1 \tag{5}$$

Can also define terms such as: semi-IQR, midquartile.

If a statistics exam had the following 14 scores:

$$x = \{43, 62, 72, 73, 77, 83, 83, 84, 85, 87, 89, 94, 95, 97\}$$

Question 8. Find Q_1 and Q_3

Question 9. Find the IQR.

SIX NUMBER SUMMARY:
`summary(x)` R COMMAND

Where `x` is a vector. Returns min, P_{25}, P_{50}, $\bar{x}$, P_{75}, max.

Example 2. Using R to find the quartiles:

```
R: summary(height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   62.0    65.0    68.0    67.6    69.8    77.0
R: summary(height.skewed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   32.0    65.0    68.0    65.7    69.5    77.0
```

[1] There are about nine slightly different methods for calculating percentiles in common use.

1.2 Exploratory Data Analysis

OUTLIERS

DEFINITION 1.6 OUTLIER.

A data point very far from the rest of the data points.

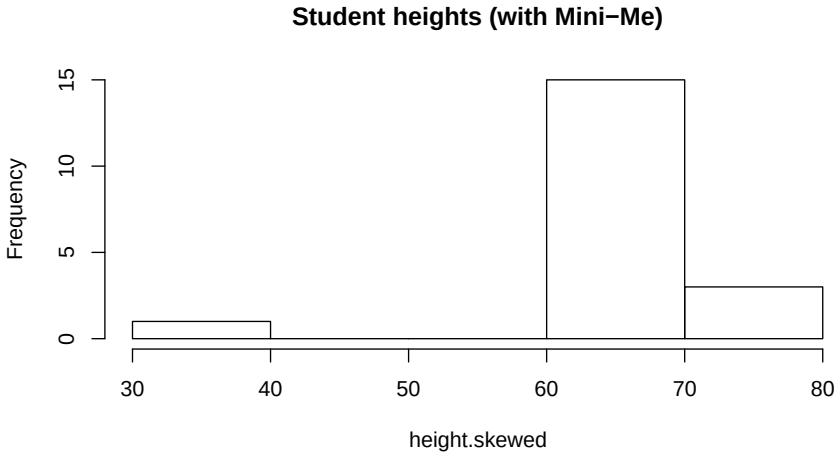
- Can reveal important information
- Strongly effect the mean and standard deviation
- Can distort histogram so most classes are empty.

An outlier is an unusual case or potentially a data entry error. You **cannot** discard the value unless you have proof that it is an error or not a member of the population under study.

If your data does have a valid outlier, you may want to compute statistics including it and excluding it and show it's effect.

Outliers in class heights

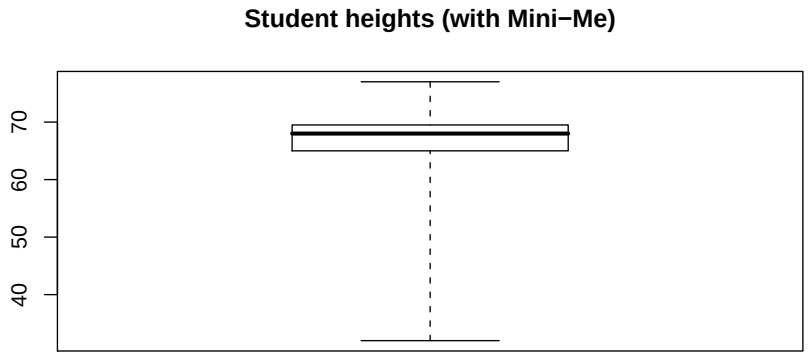
```
R: hist(height.skewed, main = "Student heights (with Mini-Me)")
```



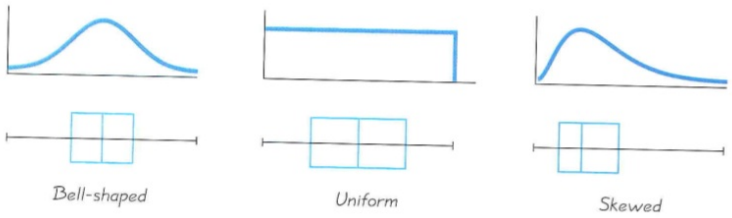
BOX PLOTS

DEFINITION 1.7 BOXPLOT.

A plot with: **whiskers**: min-max; **box**: IQR $Q_1 - Q_3$; **cent. line**: median Q_2



Box plots and distributions



Modified box plot. Definition 1.8

Boxplot where whiskers have a maximum length of $1.5 \cdot IQR$. Useful for identifying outliers.

Box plot: R Command

```
boxplot(x)
```

Where `x` is a vector of data. Plots a modified box plot.

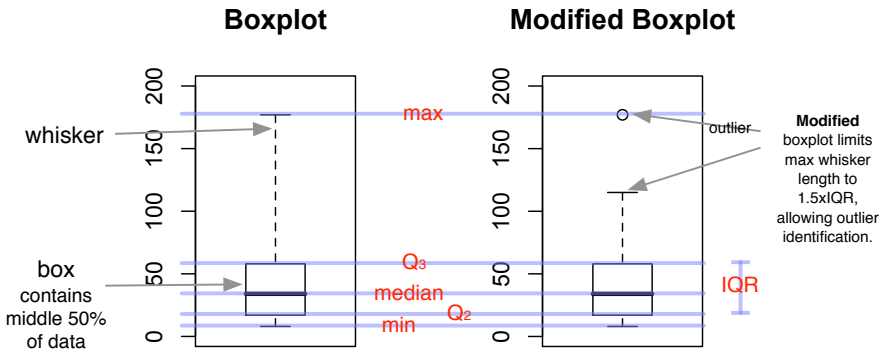
Box plot by categories: R Command

```
boxplot(x~c)
```

Where `x` and `c` are ordered vectors of data. The vector `c` is a category for each data point in `x`.

Comparison of Box Plots to Histogram

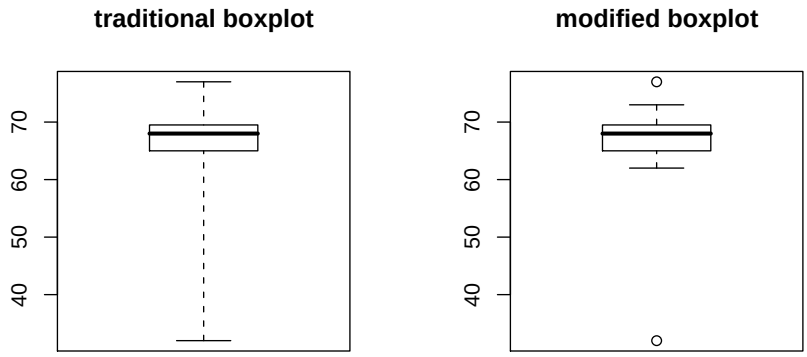
Anatomy of a Box Plot

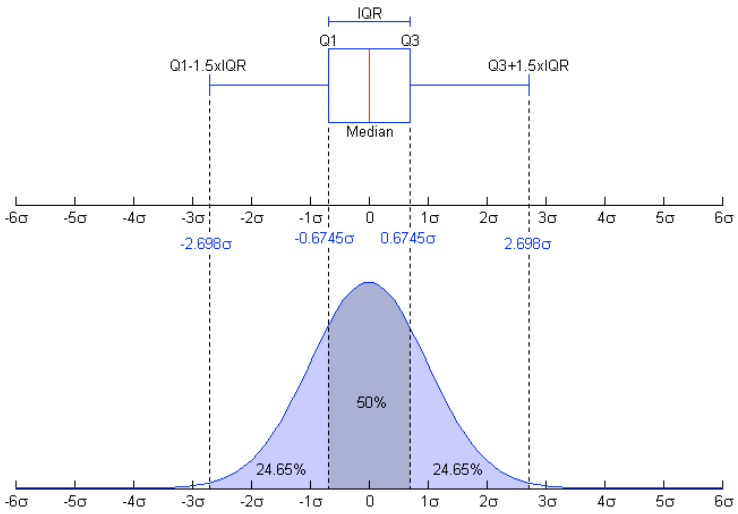


The box represents where the middle 50% of the data lies. The bottom of the box is Q1 and the top of the box is Q3. The height of the box is the IQR. The bold line in the middle of the box is the median. The dotted lines — the whiskers — extend out to the min and the max of the data. A modified box plot limits the length of the whiskers to 1.5×IQR. If data points lie beyond this, we stop the whisker and make a dot to indicate outliers.

Box plots with R

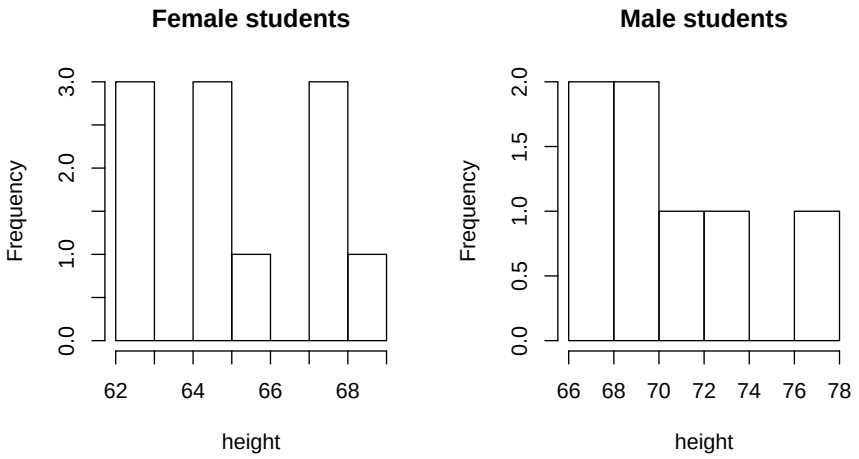
```
R: par(mfrow = c(1, 2))
R: boxplot(height.skewed, range = 0, main = "traditional boxplot"↩
    )
R: boxplot(height.skewed, main = "modified boxplot")
```





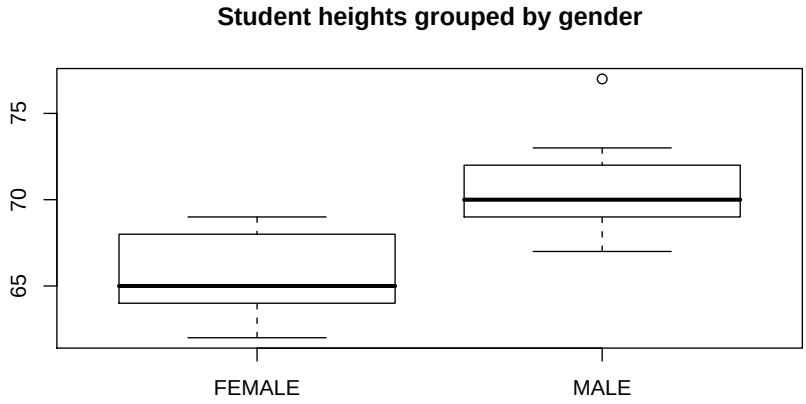
Example 3. Comparison of student heights versus gender using histograms.

```
R: par(mfrow = c(1, 2))
R: hist(class.data$height[class.data$gender == "FEMALE"],
+     main = "Female students", xlab = "height")
R: hist(class.data$height[class.data$gender == "MALE"],
+     main = "Male students", xlab = "height")
```



Example 4. Comparison of student heights versus gender using a boxplot. **Boxplots make comparing distributions easier and more informative.**.

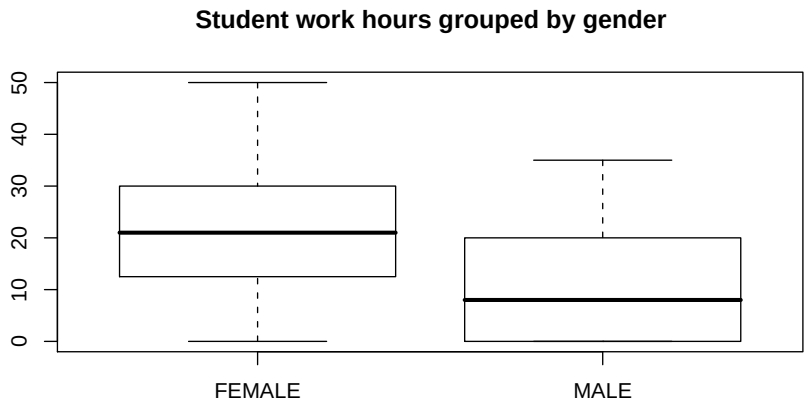
```
R: boxplot(class.data$height ~ class.data$gender,
+      main = "Student heights grouped by gender")
```



Question 10. What differences do you observe?

Example 5. Comparison of student work hours versus gender.

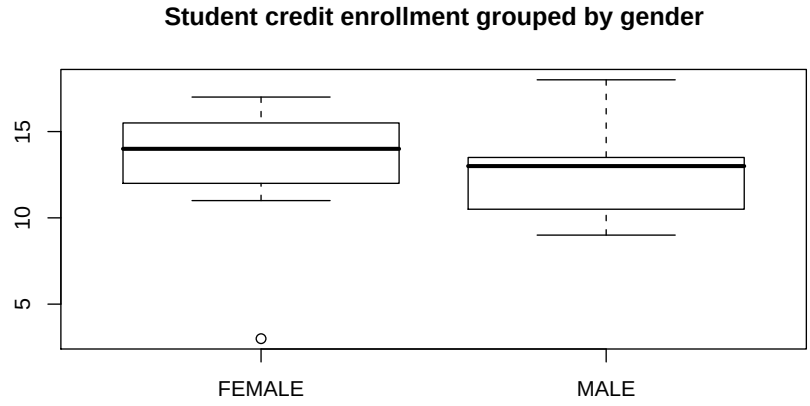
```
R: boxplot(class.data$work_hours ~ class.data$gender,
+     main = "Student work hours grouped by gender")
```



Question 11. What differences do you observe?

Example 6. Comparison of number of credits students are enrolled in versus gender.

```
R: boxplot(class.data$credits ~ class.data$gender,
+     main = "Student credit enrollment grouped by gender")
```



Question 12. What differences do you observe?

1.3 Summary

Measures of relative standing for a specific data point:

1. $z - scores$: number standard deviations from the mean.

$$z = \frac{x - \mu}{\sigma}$$

2. Usual values: $|z| \leq 2$
3. percentiles: percent of data less than x
4. quartiles: percentiles at 25%, 50%, 75%. R: `summary(x)`

Box plots: `boxplot(x)`

- Encapsulates histogram data in graphic for comparisons.
- Displays outliers, range, IQR, median.
- Useful for looking at quantitative data vs. categorical data.