Measures of spread continued
Overview

Quick review:

- Statistics for central tendency: mean and median
- Outliers
- Standard deviations

Z-scores
Percentiles
Boxplots
How did Worksheet 2 go?

Questions?

The grinch doing yoga
Human computers
Descriptive statistics for one quantitative variable

We will be looking at:
• What is the general ‘shape’ of the data
• Where are the values centered (central tendency)
• How do the data vary (spread)

There are all properties of how the data is distributed
Central Tendency: sample and population mean

\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]

\( R: \text{mean}(x) \)

\( \mu \) \quad \text{parameter}

\( \bar{x} \) \quad \text{statistic}

(d) Symmetric but not bell-shaped
Histograms: plotting the shape of quantitative data

(a) Skewed to the right
(b) Skewed to the left
(c) Symmetric and bell-shaped
(d) Symmetric but not bell-shaped

\[ \mu \]

\[ \bar{x} \]
The median: another measure of central tendency

The **median** is a value that splits the data in half

- Sort the data and take the middle value (or the mean of the middle 2 values)

R: `median(v)`
Outliers

What is an outlier?
• A: An observed value that is notably distinct from the other values in a dataset

What are they problematic?
• A: can potentially have a large influence on the statistics you calculate

What should you do if you have an outlier in your data?
A: See if you can understand what is causing it!
• If it’s an error, delete the point
• If it’s a real value, make sure it is not having a big effect on your conclusions, and/or use resistant statistics

Is the mean and/or median resistant?
• A: The median is resistant when the mean is not
The standard deviation: a measure of spread

The **standard deviation** (for a quantitative variable) is a measure of the spread of the data.

It gives a rough estimate for a typical distance a point is from the center.
Population and sample standard deviation
The standard deviation

The standard deviation can be computed using the following formula:

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

**Part 1:** Calculate the mean and standard deviation for the number of hot dogs eaten!
The 95% rule for normal distributions

A normal distribution is a common distribution that is symmetric and bell shaped.

If a distribution of data is approximately normally distributed, about 95% of the data should fall within two standard deviations of the mean. i.e., 95% of the data is in the interval: $\bar{x} - 2s$ to $\bar{x} + 2s$
The 95% rule for *normal distributions*

A *normal distribution* is a common distribution that is symmetric and bell shaped.

If a distribution of data is approximately normally distributed, about 95% of the data should fall within two standard deviations of the mean.

i.e., 95% of the data is in the interval: $\bar{x} - 2s$ to $\bar{x} + 2s$

**Example:** IQ scores are normally distributed with a mean of 100 and a standard deviation of 15.
**Question:** what is the range of values that the middle 95% of IQ scores fall in?
**Answer:** (100 – 30) to (100 + 30), 95% of IQ scores are in the range 70 to 130.
z-Scores

The z-scores tells how many standard deviations a value is from the mean

- i.e., how far away a point \( x_i \) is from \( \bar{x} \) in a way that is independent of the units of measurement

\[
z\text{-score}(x_i) = \frac{x_i - \bar{x}}{s}
\]
Which Accomplishment is most impressive?

LeBron James is a basketball player who had the following statistics in 2011:
- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

The summary statistics of the NBA in 2011 are given below

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGPct</td>
<td>0.464</td>
<td>0.053</td>
</tr>
<tr>
<td>Points</td>
<td>994</td>
<td>414</td>
</tr>
<tr>
<td>Assists</td>
<td>220</td>
<td>170</td>
</tr>
<tr>
<td>Steals</td>
<td>68.2</td>
<td>31.5</td>
</tr>
</tbody>
</table>

Relative to his peers, which statistic is most and least impressive?
Which Accomplishment is most impressive?

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

The summary statistics of the NBA in 2011 are given below

\[
Z = \frac{(x - \bar{x})}{s}
\]

<table>
<thead>
<tr>
<th></th>
<th>Formula</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-score FGPct</td>
<td>((0.510 - 0.464)/0.053)</td>
<td>0.868</td>
</tr>
<tr>
<td>Z-score Points</td>
<td>((2111 - 994)/414)</td>
<td>2.698</td>
</tr>
<tr>
<td>Z-score Assists</td>
<td>((554 - 220)/170)</td>
<td>1.965</td>
</tr>
<tr>
<td>Z-score Steals</td>
<td>((124 - 68.2)/31.5)</td>
<td>1.771</td>
</tr>
</tbody>
</table>
Percentiles (quantiles)

The $P^{th}$ percentile is the value of a quantitative variable which is greater than $P$ percent of the data.

The income distribution is shown below. What are the 25$^{th}$ and 90$^{th}$ percentiles?

25$^{th}$ percentile = $27,794

95$^{th}$ percentile = $113,820

R: quantile(v, .95)
Quantiles: age of marijuana arrests in Toronto

> library(carData)  # load the data
> quantile(Arests$age, .2)  # get the 20th percentile value from a vector of ages of arrests

20th percentile value is 17

i.e., 20% of the arrests were of ages 17 or less
Quantiles: age of marijuana arrests in Toronto

Histogram of Ages of people arrested for marijuana use

> quantile(Arrests$age, c(.2, .6))  # get the 20th and 60th percentile values from a vector of ages of arrests

60th percentile value is 23
i.e., 60% of the arrests were of ages 23 or less
Five Number Summary

**Five Number Summary** = (minimum, $Q_1$, median, $Q_3$, maximum)

$Q_1 = 25^{\text{th}}$ percentile (also called $1^{\text{st}}$ quartile)

$Q_3 = 75^{\text{th}}$ percentile (also called $3^{\text{rd}}$ quartile)

Roughly divides the data into fourths
Range and Interquartile Range

**Range** = maximum – minimum

**Interquartile range (IQR)** = $Q_3 - Q_1$
Hot dog example

**Part 2:** For the hot dog data calculate:
- The 5 number summary
- The range
- Interquartile range

**Cheat sheet:**

**Five Number Summary** = (minimum, Q₁, median, Q₃, maximum)
**Range** = maximum − minimum
**Interquartile range (IQR)** = Q₃ − Q₁
Q₁ = 25th percentile, Q₃ = 75th percentile

Answer in R: `fivenum(v)`
Detecting of outliers

As a rule of thumb, we call a data value an **outlier** if it is:

- Smaller than: \( Q_1 - 1.5 \times \text{IQR} \)
- Larger than: \( Q_3 + 1.5 \times \text{IQR} \)

What is the range that a value would be called an outlier in the hot dog data?

Are there any outliers in the hot dog data?
Boxplots

A **boxplot** is a graphical display of the 5 number summary and consists of:

1. Drawing a box from $Q_1$ to $Q_3$

2. Dividing the box with a line (or dot) drawn at the median

3. Draw a line from each quartile to the most extreme data value that is not an outlier

4. Draw a dot/asterisk for each outlier data point.
Box plot of the number of hot dogs eaten by the men’s contest winners 1980 to 2010

R: `boxplot(v)`
Box plots extract key statistics from histograms
Box plots extract key statistics from histograms

**Question:** which Boxplot goes with which histogram?
Boxplots don’t capture everything
Comparing quantitative variables across categories

Often one wants to compare quantitative variables across categories.

**Side-by-Side** graphs are a way to visually compare quantitative variables across different categories.
Side-by-side boxplots

Side-By-Side (Comparative) Boxplots
Age of Best Actor/Actress Oscar Winners (1970-2001)
Side-by-size boxplots in R

```r
> boxplot(v1, v2, names = c("name 1", "name 2"),
         ylab = "y-axis name")
```

Try it yourself, create histograms and boxplots for this data:

```r
> load("/home/shared/intro_stats/cs26_data/diff_distribution_same_boxplot.Rda")
> boxplot(x1, x2, x3, x4)
```
Concepts/statistics for summarizing quantitative data

**z-scores** show how many standard deviations a point is from the mean

\[ z\text{-score}(x_i) = \frac{x_i - \bar{x}}{s} \]

**Quantiles** show the value \( x \), such that a fixed proportion of the data is less than \( x \)
- e.g., what is the value \( x \), such that 20% of the data is less than \( x \)

**Five Number Summary** give key summary statistics of a data sample
- minimum, \( Q_1 \), median, \( Q_3 \), maximum

A **boxplot** is a visualization of the five number summary
- Side-by-side boxplots allow you to compare key summary statistics
Summary of R

We can compute a z-score for a value \( x \), and a vector of values \( v \) using:

\[
\text{the\_mean} <- \text{mean}(v) \\
\text{the\_sd} <- \text{sd}(v) \\
\text{the\_zscore} <- (x - \text{the\_mean})/\text{the\_sd}
\]

We can compute quantiles using the \text{quantile()} \ function:

\[
\text{quantile}(v, .2) \quad \text{or} \quad \text{quantile}(v, c(.25, .4))
\]

We can compute a five number summary using the \text{fivenum()} \ function:

\[
\text{fivenum}(v)
\]

We can compute boxplots using the \text{boxplot()} \ function

\[
\text{boxplot}(v)
\]
Worksheet 3

Log on to asterius and type:

> source('/home/shared/intro_stats/cs206_functions.R')
> get_worksheet(3)

Due, Sunday September 30th at 11:59pm
Use Slack for questions – and answer questions posted by others

Worksheet is a little harder than the previous ones so get started early!