Lecture 3:
Descriptive Statistics
STAT 630, Fall 2021

Measures of Central Tendency

Let $x_1, x_2, \cdots, x_n$ be observations of a sample of size n. The **sample mean** is given by

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

Ex: The heights of 5 individuals: 63, 64, 66, 72, 62.

$$\bar{x} = \frac{63 + 64 + 66 + 72 + 62}{5} = 65.4$$

Measures of Central Tendency

The **sample median** of a set of observations is the middle value when values are ordered from smallest to largest.

Ex: (n odd) Find the median of $63, 64, 66, 72, 62$.

First, order the data: $62, 63, 64, 66, 72$
median $= 64$

Ex: (n even) Find the median of $63, 64, 66, 72, 62, 77$.

First, order the data: $62, 63, 64, 66, 72, 77$
median $= (64 + 66)/2 = 65$

Measures of Central Tendency

The median is resistant to outliers, while the mean is affected by outliers.

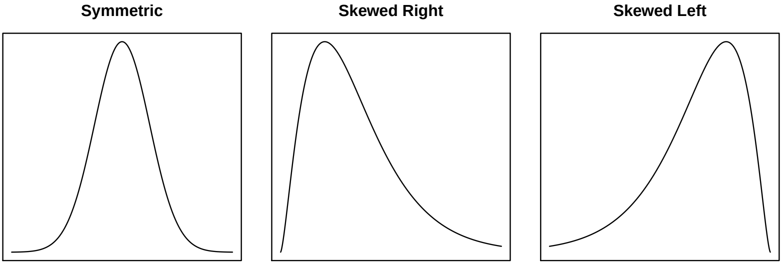
Ex: How do the mean and median compare for the sample: $62, 63, 64, 66, 72, 1000$?

```
> x <- c(62, 63, 64, 66, 72, 1000)
> mean(x)
[1] 221.1667
> median(x)
[1] 65
```

The sample mean is much larger than the median, since it is affected by the outlier. The median is a better measure of central tendency in this example.

Measures of Central Tendency

Compare the mean and median for distributions that are symmetric, skewed right, and skewed left.



Quartiles

- The **first quartile**, denoted by Q_1, is the value such that 25% of the data falls below, i.e., the 25^{th} percentile.
- The **third quartile**, denoted by Q_3, is the value such that 75% of the data falls below, i.e., the 75^{th} percentile.
- Note that the second quartile, Q_2, is the median.

A method for finding Q_1 and Q_3 by hand:

1. Order the data from smallest to largest
2. Divide the data into two sets using the median
3. Q_1 is the median of the first half, and Q_3 is the median of the second half

Quartiles

Ex: Find Q_1 and Q_3 for the following sample of $n = 10$ heights of individuals:
$68, 76, 66, 63, 70, 66, 71, 71, 64, 71$

Solution:
First, order the data: $63, 64, 66, 66, 68, 70, 71, 71, 71, 76$

median $= (68 + 70)/2 = 69$
$Q_1 = 66$
$Q_3 = 71$

Useful R commands:

```
> x <- c(68, 76, 66, 63, 70, 66, 71, 71, 64, 71)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   63.0    66.0    69.0    68.6    71.0    76.0 
> mean(x)
[1] 68.6
> median(x)
[1] 69
> min(x)
[1] 63
> max(x)
[1] 76
> sort(x)
 [1] 63 64 66 66 68 70 71 71 71 76
```

Percentiles

The more general $100 \cdot p^{th}$ **percentile**, where $0 \leq p \leq 1$, is the value such that $100 \cdot p\%$ of the data falls below. A related term is **quantile**; for example, the 0.3 quantile is the same as the 30^{th} percentile.

Ex: Use R to compute the 20^{th} and 80^{th} percentiles for the ages of the 20,000 individuals in the `cdc` data set (see lab 2).

```
> quantile(cdc$age, c(0.2, 0.8))
20% 80%
 29  61
```

Measures of Variation

- Range = Max - Min
- Interquartile range: IQR = Q_3 - Q_1
- Let $x_1, x_2, \cdots, x_n$ be a sample of n observations. The **sample variance** is defined as

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

and the **sample standard deviation** is defined as

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

Measures of Variation

- The sample variance can be thought of as the average of the squared deviations between the observations x_i and the sample mean $\bar{x}$. It measures how concentrated values are around the sample mean.
- The standard deviation is in the same units as the data (e.g., if the data are in ft, then s is in ft and s^2 is in ft^2).
- s^2, s, and the range are affected by outliers, while the IQR is resistant to outliers.

Measures of Variation

Ex: Calculate the variance and standard deviation of the following sample of $n = 5$ observation: $2, 5, 10, 15, 18$

$$\bar{x} = \frac{2+5+10+15+18}{5} = \frac{50}{5} = 10$$

$$\begin{aligned} s^2 &= \frac{1}{5-1}[(2-10)^2 + (5-10)^2 + (10-10)^2 + (15-10)^2 + (18-10)^2] \\ &= \frac{1}{4}(8^2 + 5^2 + 0^2 + 5^2 + 8^2) \\ &= \frac{178}{4} = 44.5 \\ s &= \sqrt{44.5} = \boxed{6.67} \end{aligned}$$

Useful R commands:

```
> x <- c(68, 76, 66, 63, 70, 66, 71, 71, 64, 71)
> var(x)
[1] 15.6
> sd(x)
[1] 3.949684
> max(x)-min(x) # range
[1] 13
> IQR(x)
[1] 5
```

Ex: Without doing any calculations, which of the following data sets do you think has the largest sample variance? Which has the smallest sample variance? Use R to verify.

Set 1: $100, 99, 98, 50, 2, 1, 0$
Set 2: $53, 52, 51, 50, 49, 48, 47$
Set 3: $51, 51, 51, 50, 49, 49, 49$

Solution: Set 1 has the largest variance since the values are most spread out around the mean. Set 3 has the smallest variance since the values are most concentrated around the mean. Note that $\bar{x} = 50$ for all three sets.

To verify in R:

```
> x1 = c(100, 99, 98, 50, 2, 1, 0); var(x1)
> x2 = c(53, 52, 51, 50, 49, 48, 47); var(x2)
> x3 = c(51, 51, 51, 50, 49, 49, 49); var(x3)
```

Shifting and Rescaling Data

- *Shifting*: Adding a constant to each data value affects measures of position (mean, median, quartiles), but not measures of variation (standard deviation, IQR)
- *Rescaling*: Multiplying each data value by a constant affects both measures of position (mean, median, quartiles) and measures of variation (standard deviation, IQR)

Theorem: Let $x_1, x_2, \cdots, x_n$ be observations of a sample of size n, and $\bar{x}$ and s_x the sample mean and standard deviation. For the transformation $y_i = ax_i + b$, where a and b are constants, show that $\bar{y} = a\bar{x} + b$ and $s_y = |a|s_x$.

Proof:

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\
&= \frac{a}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} b \\
&= a\bar{x} + \frac{1}{n}(nb) = a\bar{x} + b
\end{aligned}$$

$$
\begin{aligned}
s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[(ax_i+b)-(a\bar{x}+b)]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i-a\bar{x})^2 \\
&= \frac{a^2}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= a^2 s_x^2
\end{aligned}
$$

The standard deviation is the square root of the variance:

$$
s_y = \sqrt{a^2 s_x^2} = |a| s_x
$$

Ex: Consider the following temperature measurements in $^\circ F$: $72, 67, 73, 81, 75$.

(a) Calculate the mean and standard deviation. (You can use R for this)

```
> x <- c(72, 67, 73, 81, 75)
> mean(x)
[1] 73.6
> sd(x)
[1] 5.07937
```

(b) What is the mean and standard deviation if we convert from $^\circ F$ to $^\circ C$? The conversion formula is $^\circ C = \frac{5}{9}(^\circ F - 32)$

Mean in Celsius:

$$\frac{5}{9}(73.6 - 32) = 23.11^\circ C$$

Standard deviation in Celsius:

$$\frac{5}{9}(5.08) = 2.82^\circ C$$

Box Plot

A box plot useful way to display the distribution of data and identify outliers.

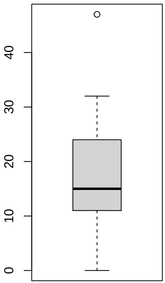
Upper Fence $= Q_3 + 1.5(IQR)$
Lower Fence $= Q_1 - 1.5(IQR)$

Values outside the fences are potential outliers.[1]

[1] Remark: Values falling above $Q_3 + 3(IQR)$ or below $Q_1 - 3(IQR)$ are "extreme" outliers.

Box Plot: Example

```
> x <- c(0, 18, 15, 32, 5, 22, 47, 15, 26, 13, 9)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   11.00   15.00   18.36   24.00   47.00
> boxplot(x)
```

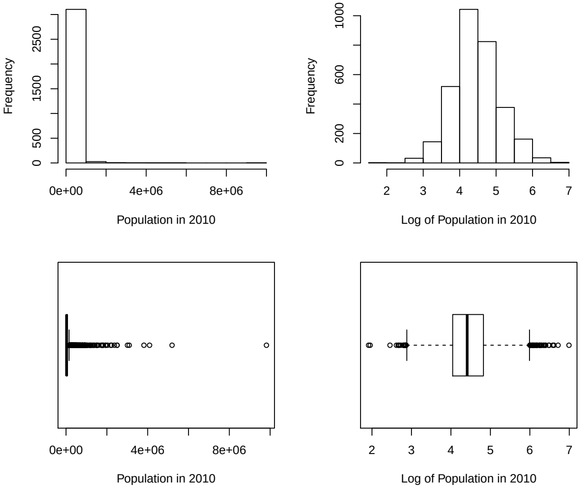


Data Transformations

- When distributions are heavily skewed and contain outliers it is often useful to transform the data.
- Common transformations include the logarithm, $\log(x)$; square root, $\sqrt{x}$; and reciprocal, $1/x$.
- The log-transformation is often used to make the data more symmetric and normal (bell-curve shaped), and to make any outliers far less extreme. Note that the log-transformation is only defined for strictly positive values, $x > 0$.

Data Transformations: Example

A histogram and box plot of the populations of all 3143 US counties in 2010 before and after taking the log-transformation.



Code used to create last plot:

```
> library(openintro) # load library to access data set
> par(mfrow=c(2,2)) # split plot into 4 panels
> hist(county$pop2010, xlab="Population in 2010", main="")
> hist(log10(county$pop2010), xlab="Log of Population in 2010", main="")
> boxplot(county$pop2010, xlab="Population in 2010", horizontal = TRUE)
> boxplot(log10(county$pop2010),
        xlab="Log of Population in 2010", horizontal = TRUE)
```