MATH 1051H S61: Lecture #04b (Live)

Examining Numerical Data

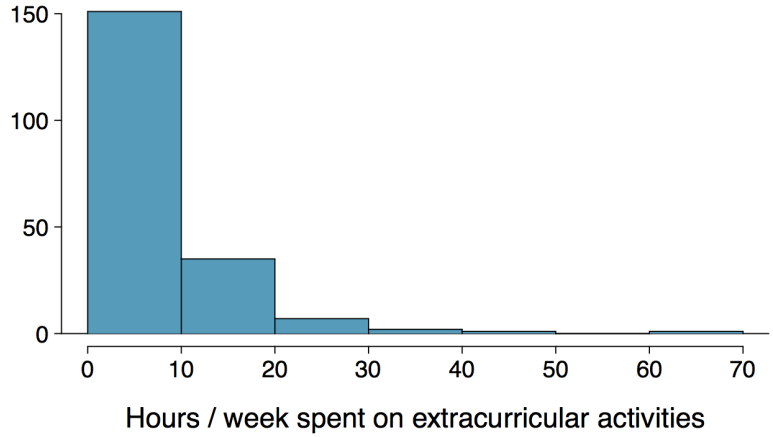
Scatterplots

Let's look at Gapminder briefly (link on Blackboard) and chat about scatterplots.

Gapminder Link

Histograms — Extracurricular Hours

- Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the **shape** of the data distribution.
- The chosen **bin width** can alter the story the histogram is telling.



Essay on Histograms

Let’s look at an interactive essay on histograms made by a colleague in Minnesota.

Exploring Histograms

Are you typical?

YouTube Link

Statistics

1. Sample Mean

The average, you've done it before.

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

2. Sample Median

The median is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th **percentile**.

3. Variance

Variance is roughly the average squared deviation from the mean.

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

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

4. Standard Deviation

The **standard deviation** is the square root of the variance, and has the same units as the data.

$$s = \sqrt{s^2}$$

Percentile

A **percentile** is the the smallest value from an ordered list of numbers which is greater than or equal to that percentage of list elements.

Example: The 42^{nd} percentile of the numbers $\{1, 2, 3, \cdots, 99, 100\}$ is 42.

It can become quite complicated when there aren't an even multiple of 100 items!

5-7. Q1, Q3 and IQR

- The 25th percentile is also called the first quartile, **Q1**.
- The 50th percentile is also called the median.
- The 75th percentile is also called the third quartile, **Q3**.

Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the **interquartile range**, or the IQR.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Calculating

I will never expect you to compute a variance or SD without a calculator. In practice, we do this using R always. For example,

```
x <- sample(1:100, size = 20, replace = TRUE)
mean(x)
```

```
## [1] 49
```

```
median(x)
```

```
## [1] 52.5
```

Calculating

```
var(x)
```

```
## [1] 1052
```

```
sd(x)
```

```
## [1] 32.43455
```

```
sqrt(var(x))
```

```
## [1] 32.43455
```

Calculating

```
quantile(x, probs = c(0.25, 0.50, 0.75))
```

```
##  25%  50%  75%
## 14.5 52.5 73.5
```

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    14.5    52.5    49.0    73.5    96.0
```

Plotting as Numerical Summary

Most Important Plot: Boxplot