Class 05

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-09-09

Request

Could you put “DATA1220:” at the beginning of the subject line of your emails?

I have 3 emails, and this will help me spot and respond to yours more quickly.

Reminder: Homework

Late policy: "This homework is due by 6:00pm on Monday, 9/9/24. No credit will be lost for assignments received by 7:00pm to account for issues with uploading. 10% of the points will be deducted from assignments received by 9:00am on Tuesday, 9/10/24. Assignments turned in after this point are only eligible for 50% credit, so it benefits you to turn in whatever you have completed by the due date."

Chapter 2 Pipeline

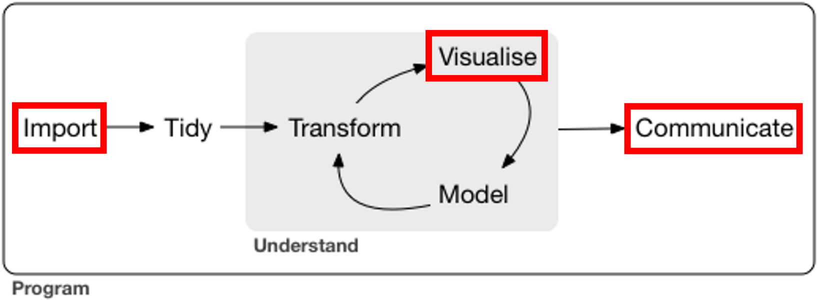


Figure 1: Data science pipeline priorities for Chapter 1

Chapter 2 Objectives: Numerical Data

- Describe the "shape" (i.e. distribution) of numerical variables
- Calculate means, medians, modes, variances, standard deviations, IQRs
- Learn the appropriate use of summary statistics (i.e. mean vs. median)
- Characterize the relationship between 2 numerical variables

Chapter 2 Objectives: Categorical Data

- Analyze contingency (e.g. 2x2) tables
- Summarizing categorical variables with proportions
- Comparison of numerical data between categorical groups

Chapter 2 Objectives: Visualizing Data

- Recognize common visualization techniques / plots
  - Numerical: Dot plots, histograms, density plots, QQ plots, box plots, violin plots
  - Categorical: bar plots, mosaic plots, tree map
- Build basic visualizations in `R` using `ggplot2`
- Data visualization do's and dont's

Load Packages for Today's Slides

```
# Contains the describe() function for comprehensive data s
library(Hmisc)
# For another describe() function with comprehensive data s
library(mosaic)
# Always load the tidyverse last
library(tidyverse)

# Set favorite ggplot2 theme for visualizations
theme_set(theme_bw())
```

Numerical Variables

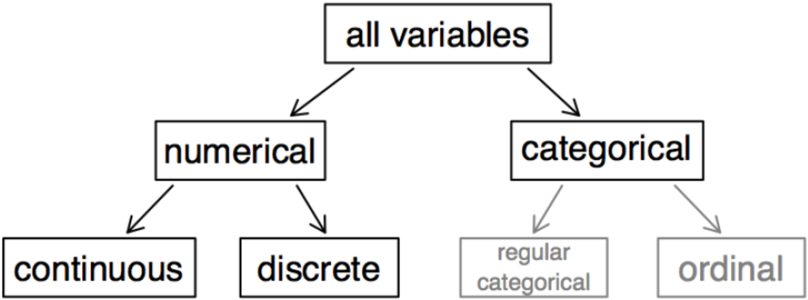


Figure 2: Numerical variables can be continuous or discrete.

Describing numerical distributions

The “shape” of numerical data is called its ***distribution***.

- ***Location:*** the “center” of the data
- ***Scale:*** the “spread” of the data

Describing distribution shapes

Modality

Skewness

Figure 3: Commonly observed patterns in numerical distributions

Unimodal Distributions

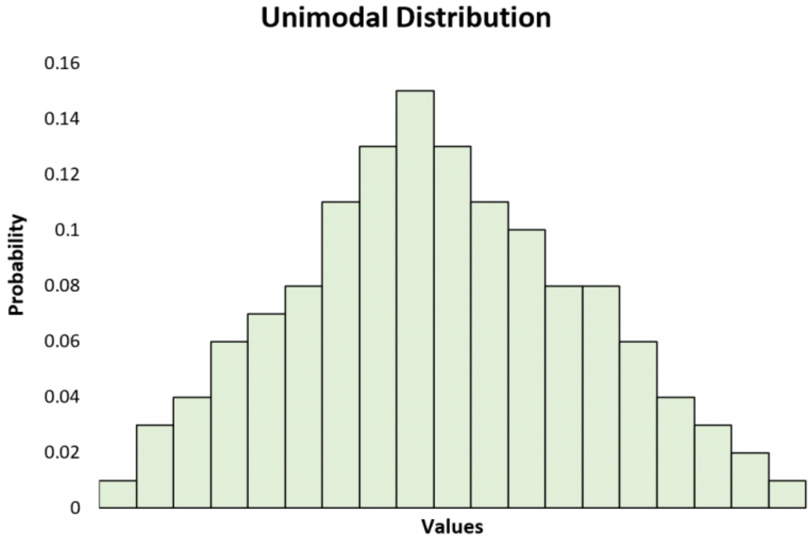


Figure 4: Unimodal distributions have one peak around which observations cluster

Bimodal Distributions

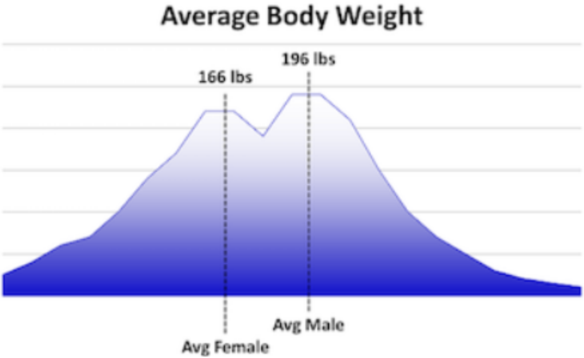


Figure 5: Bimodal distributions have 2 peaks around which observations cluster.

Trimodal Distributions

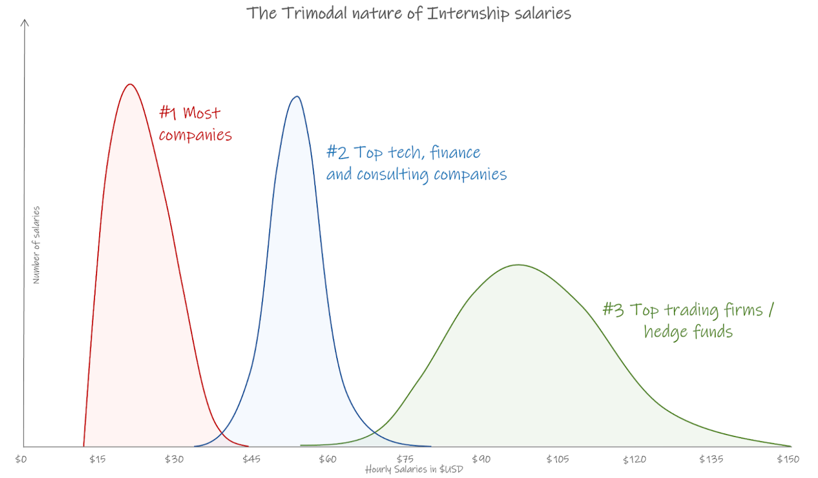
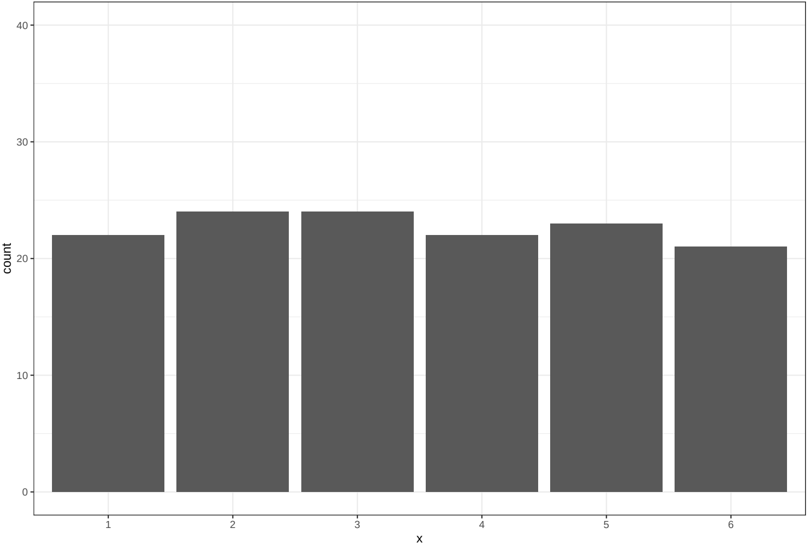


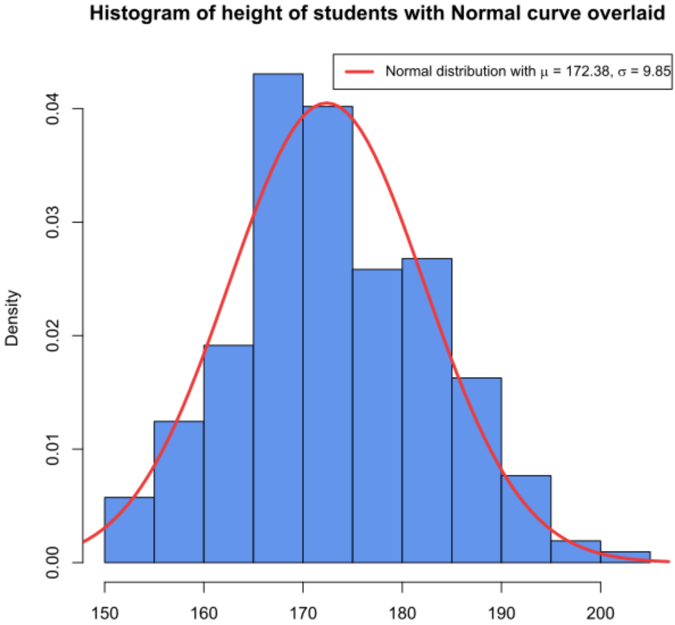
Figure 6: Trimodal distributions have 3 peaks around which observations cluster.

Uniform Distributions



Uniform distributions have no peaks around which observations are clustered

Symmetric Distributions



Left-Skewed Distributions

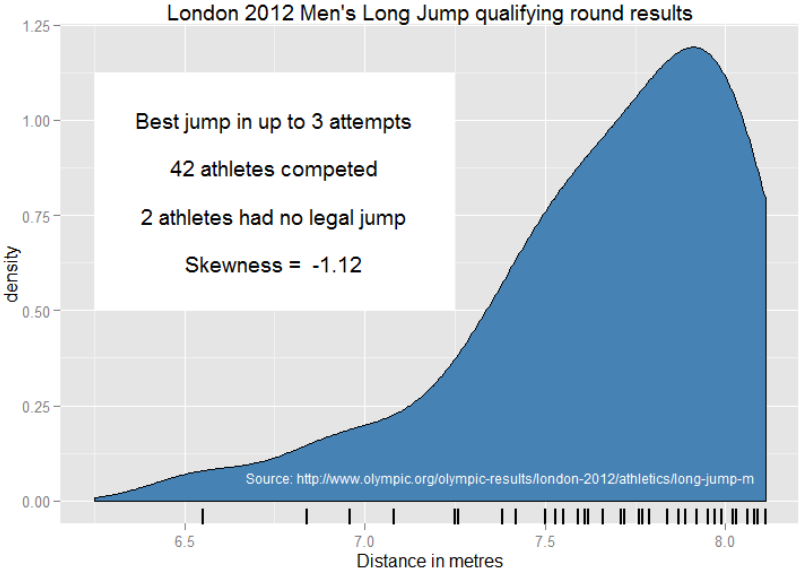


Figure 8: Left-skewed distributions have an excess of observations at the

Right-Skewed Distributions

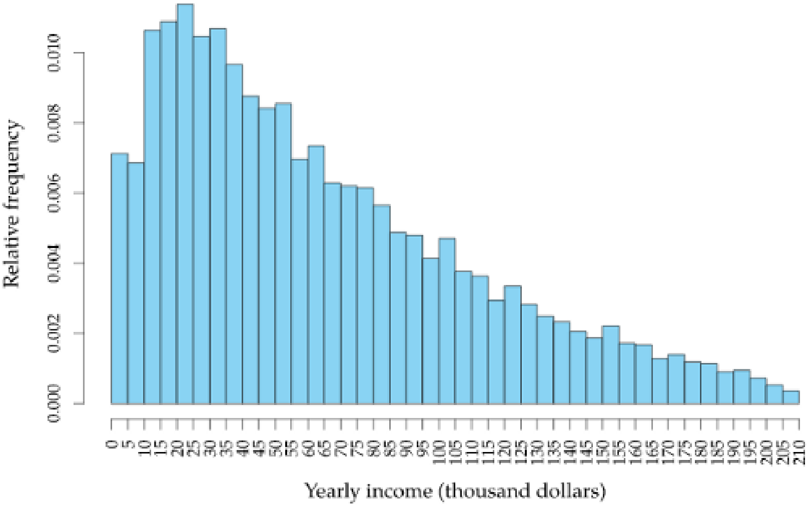


Figure 9: Right-skewed distributions have an excess of observations at the high end of the data range.

Describing a distribution's ***location***

The ***location*** of a numerical variable's distribution can be thought of as the "center" of the data, around which the bulk of the observations cluster.

- ***Mean:*** the sum of a values divided by the number of observations (i.e. "average")
- ***Median:*** the value in the exact middle of the data
- ***Mode:*** the most common value in the data (for discrete variables)

The Mean (Average)

Where are the bulk of observations concentrated?

The sample mean $\bar{x}$ is computed as the sum of all observed values $\sum_{i=1}^{n} x_i$, where i is the observation number, divided by the total number of observations n.

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

or

$$\bar{x} = \frac{\mathrm{sum}(x_1, x_2, ..., x_n)}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Calculating the mean in R

Consider the numerical variable x.

```
x
```

```
 [1] 4 5 8 5 5 8 6 2 4 4
```

```
length(x)
```

```
[1] 10
```

Calculating the mean in R

You can calculate the mean manually...

```
(4 + 5 + 8 + 5 + 5 + 8 + 6 + 2 + 4 + 4) / length(x)
```

```
[1] 5.1
```

Or you can use the `mean()` function.

```
mean(x,
     na.rm = T) # this parameter ignores missing values
```

```
[1] 5.1
```

Sample vs Population Mean

The ***sample mean*** is denoted as $\bar{x}$. The population mean is denoted μ. They are calculated the same way.

The ***sample mean*** is considered to be a good ***point estimate*** of the ***population mean*** if the sample population is *representative* of the study/target population.

What makes for a good sample?

The Median

The ***median*** is the middle value when the data are sorted in order.

- When the number of observations n is odd, this works as stated.
- When the number of observations n is even, the median is calculated as the mean of the 2 middle values.

Calculating the median in R

```
# Sort the data in order from least to most
sort(x)
```

```
 [1] 2 4 4 4 5 5 5 6 8 8
```

```
(5 + 5) / 2
```

```
[1] 5
```

```
median(x)
```

```
[1] 5
```

Describing a distribution's ***scale***

How far is each data value from the mean?

- ***Variance:*** s^2, the sum of the squared differences between each observation's value and the sample mean $\bar{x}$ divided by $n-1$
- ***Standard deviation:*** s, the square root of the variance
- ***Range:*** minimum to maximum
- ***Interquartile Range (IQR):*** 25th percentile to 75th percentile

The Variance

The ***deviance*** is how far each data value is from the mean. The ***variance***, denoted as s^2, is the squared sum of all observation ***deviations*** $\sum_{i=1}^{n}(y_i - \bar{y})^2$ where i is the observation number, divided by $n - 1$.

$$\text{Variance} = s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1}$$

Calculating the variance in R

```r
var(x)
```

```
[1] 3.433333
```

The Standard Deviation

The ***standard deviation*** is the square root of the variance, and is interpreted in the original unit of measurement for that variable.

$$\mathrm{StandardDeviation} = s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Calculating the standard deviation in R

```
sd(x, na.rm = T)
```

```
[1] 1.852926
```

The Range

The ***range*** of the data is the difference between the maximum value and the minimum value.

$$\text{Range} = \max(x) - \min(x)$$

Calculating the range in R

```
max(x) - min(x)
```

```
[1] 6
```

```
range(x)
```

```
[1] 2 8
```

The Interquartile Range (IQR)

- The 25th percentile of the data is called the ***first quartile*** or ***Q1***
- The 50th percentile of the data is called the ***median***
- The 75th percentile of the data is called the ***third quartile*** or ***Q3***
- The range between ***Q3*** and ***Q1*** is called the ***interquartile range*** or ***IQR***.

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

Calculating the interquartile range in R

```r
c(quantile(x, 0.25), quantile(x, 0.75))
```

```
 25%  75%
4.00 5.75
```

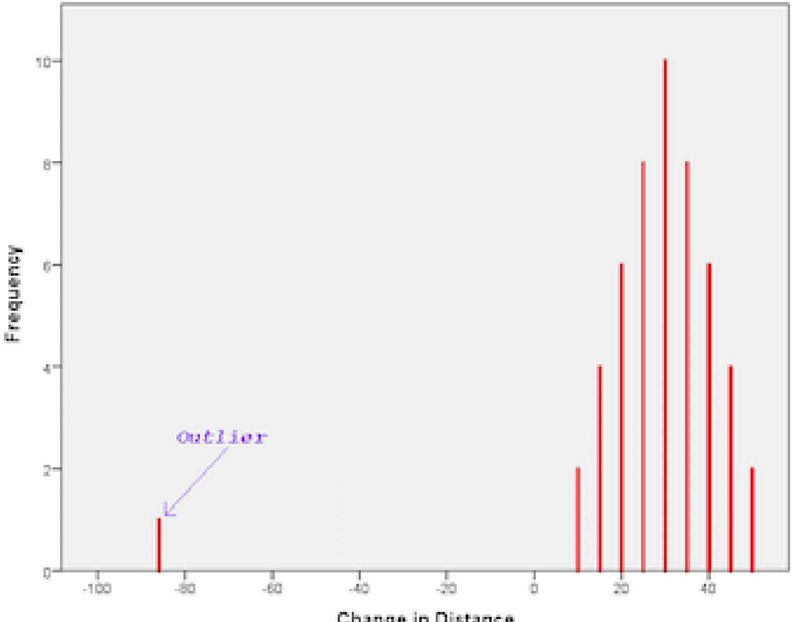
```r
quantile(x, 0.75) - quantile(x, 0.25)
```

```
 75%
1.75
```

```r
IQR(x)
```

```
[1] 1.75
```

Outliers



Frequency
10
8
6
4
2
0
Outlier
-100
-80
-60
-40
-20
0
20
40
Change in Distance

How do skew and outliers affect numerical summary statistics?

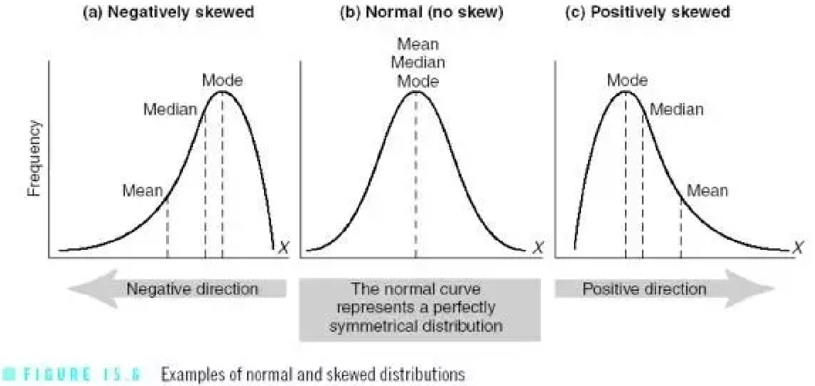


Figure 10: The presence of outliers and/or skew in a numerical variable's distribution affects how well summary statistics describe a distribution's location.

Robust statistics

The ***median*** and ***interquartile range*** are considered to be ***robust statistics*** for the numerical summary of data because they are less sensitive to ***skew*** and ***outliers*** than the ***mean***, ***variance***, and ***standard deviation***.