Hacettepe University
Department of Industrial Engineering
Undergraduate Program
2023-2024 Fall

EMU 430 – Data Analytics
Week5
November 3, 2023

Instructor: Erdi Dasdemir

edasdemir@hacettepe.edu.tr
www.erdidasdemir.com

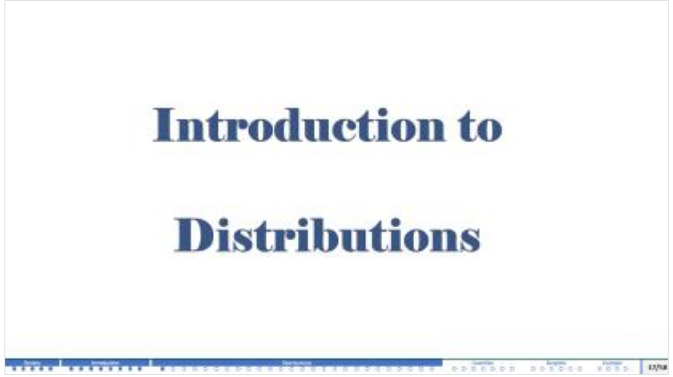
Outline

Previously on

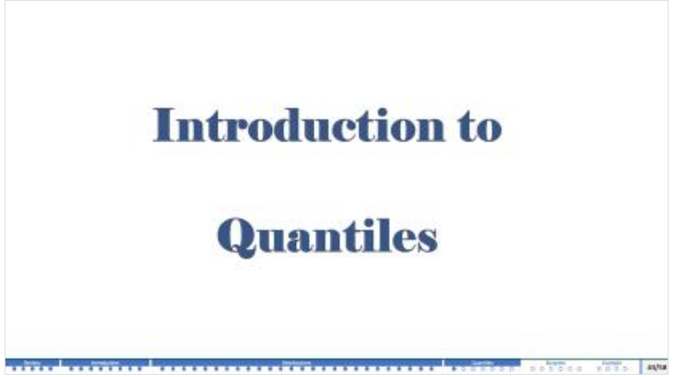
EMU430

Introduction to

Data Visualization

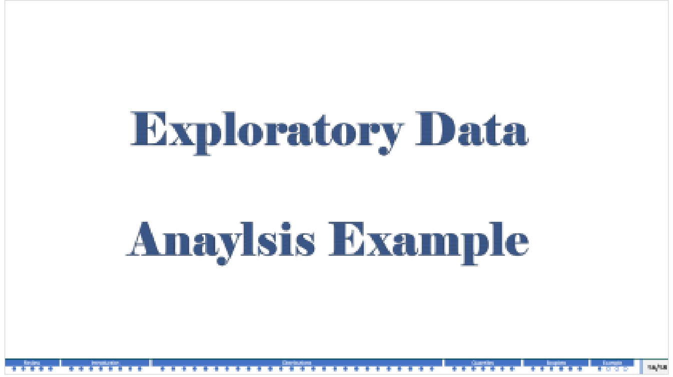

Introduction to

Distributions


Introduction to

Quantiles

Boxplots


Exploratory Data

Anaylsis Example

Acknowledgment

I drew inspiration primarily from Dr. Rafael Irizarry's "Introduction to Data Science" Book

and "Data Science" course by HarvardX on edX for the slides this week.

Previously on

EMU430

Productivity Tools

- Organizing files and document preparation

- **Principles:**

  1. **Be systematic when organizing your filesystem:** Minimize time spent looking for something

  2. **Automize when possible:** If you repeat the same task repeatedly, there is probably a way to automize it.

Productivity Tools

- The data analysis process is iterative and adaptive→ We are constantly editing our scripts and reports.
  - Version Control System
    - **Git:** a version control system; a powerful tool for keeping track of these changes.
    - **GitHub:** a service that permits you to host and share your code, facilitate collaborations
    - **Side effect:** Showcase your work to potential employers.
  - Write reports in Quarto (or Rmarkdown)
    - Incorporate text and code into a single document
    - Write reproducible and aesthetically pleasing reports by running the analysis and generating the report simultaneously.
    - Produce webpages, blogs, books etc.

Three reasons to use them:

1. Version control:
   - keep track of changes we make to our code
   - revert back to previous versions of our files
   - Git also permits us to create branches in which we can test out ideas and then decide if we can merge the new branch with the original.

2. Collaborating
   - a central repo, you can have multiple people make changes to the code and keep the version synced.
   - **Pull request:** anybody to suggest changes to your code. You can easily either accept or deny these request.

3. The third reason is to **share**, and this is the main one we use here. Even if we do not take advantage of the advanced and powerful version control functionality, we can still use Git and GitHub to share our code.

1. Install R
   - ➢ the programming language
2. Install RStudio
   - ➢ the integrated desktop environment
3. Install Git:
   - ➢ The version control system to keep track of changes made to our code and to sync local copies of code with copies hosted on GitHub.
   - ➢ For Windows users only, installing Git installs Git Bash, which emulates Unix on Windows Machines.
   - ○ Install Git → Install Git Bash, emulates Unix on Windows Machines
4. Open a GitHub account and sync it with RStudio
5. Download Quarto
   - An open-source scientific and technical publishing system for analyzing, sharing and reproducing

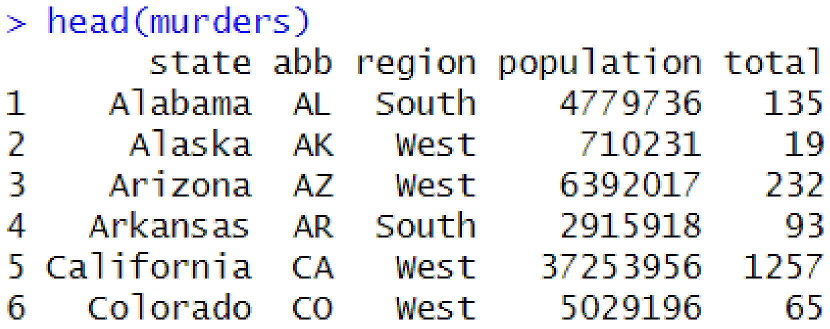
Introduction to

Data Visualization

Introduction to Data Visualization

- Numbers and character strings in a dataset is difficult to read and rarely useful.

- US murders data table.



```
> head(murders)
       state abb region population total
1    Alabama  AL  South    4779736   135
2     Alaska  AK   West     710231    19
3    Arizona  AZ   West    6392017   232
4   Arkansas  AR  South    2915918    93
5 California  CA   West   37253956  1257
6   Colorado  CO   West    5029196    65
```

- How quickly can you determine which states have the largest populations?

- Which states have the smallest?

- How large is a typical state?

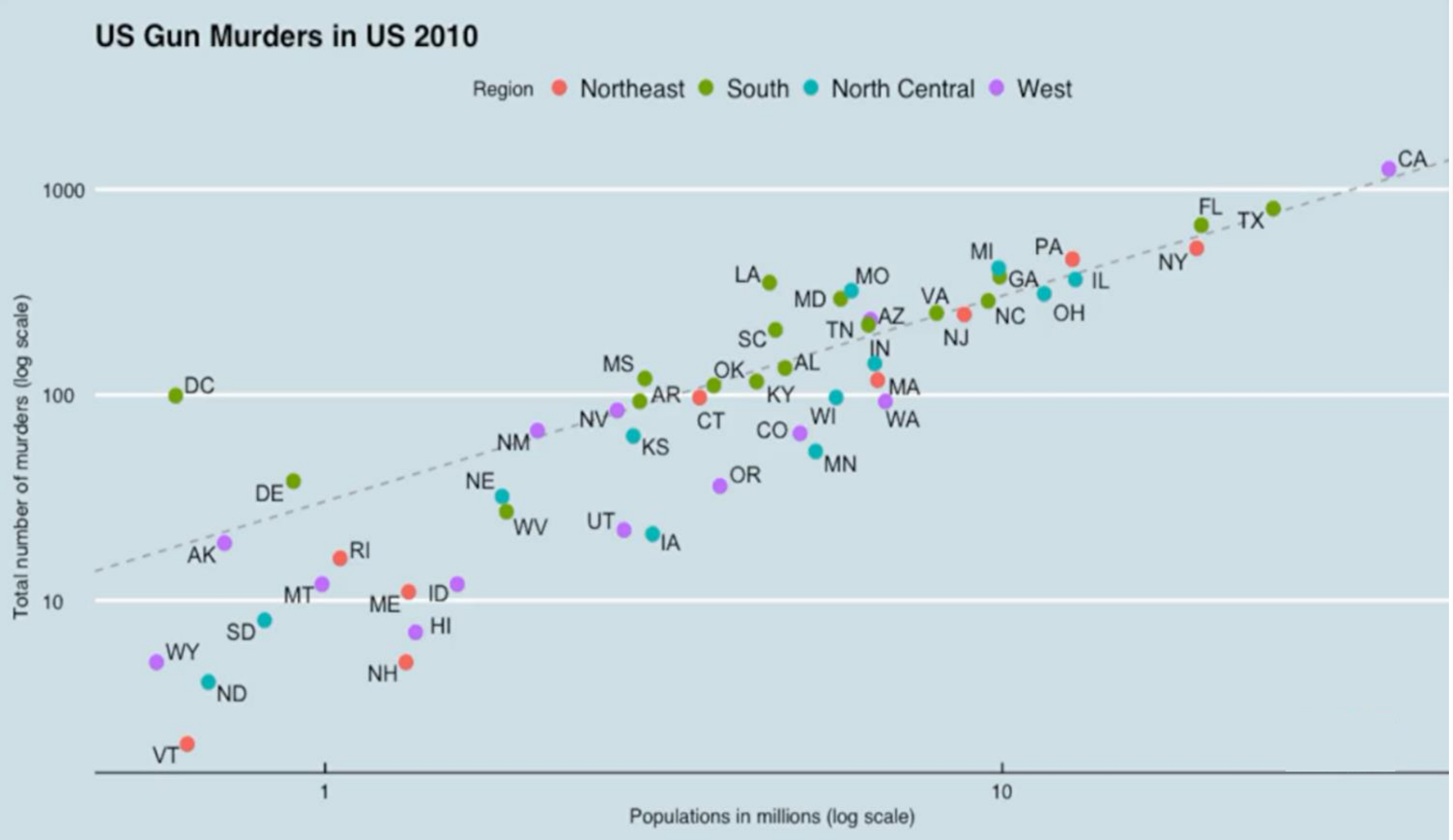
- Is there a relationship between population size and total murders?

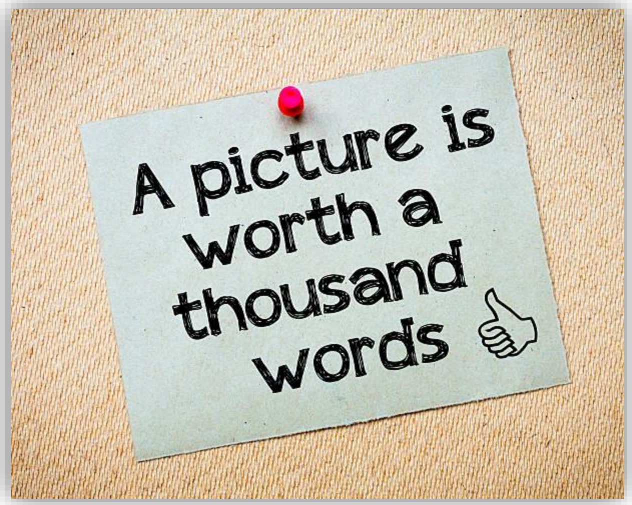
- How do the murder rates vary across regions of the country?

not easy for most human brains to answer quickly!

Introduction to Data Visualization

- the answer to all of these questions are readily available from examining this plot.



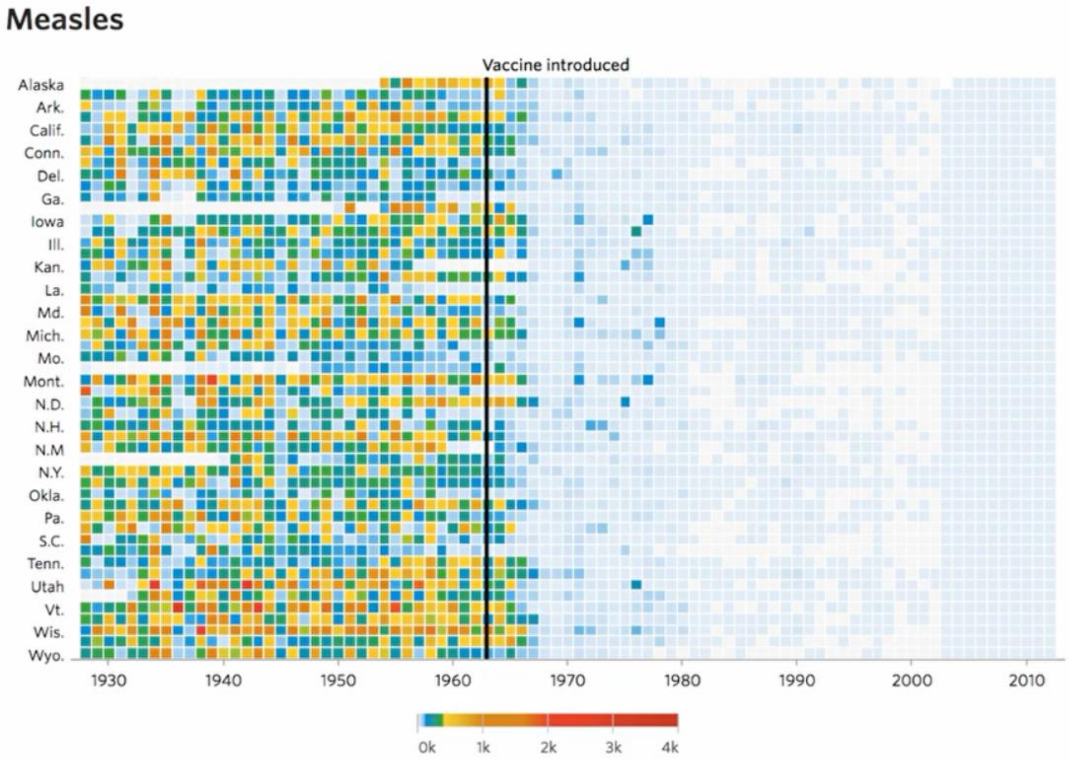


Introduction to Data Visualization

- The growing availability of informative data sets and software tools has lead to increased reliance on data visualization.

- A visible example is news organizations that started data journalism.

- Example: Wall Street Journal article showing data related to the impact of vaccines on battling infectious diseases.



Introduction to Data Visualization

○ Yhe New York Times

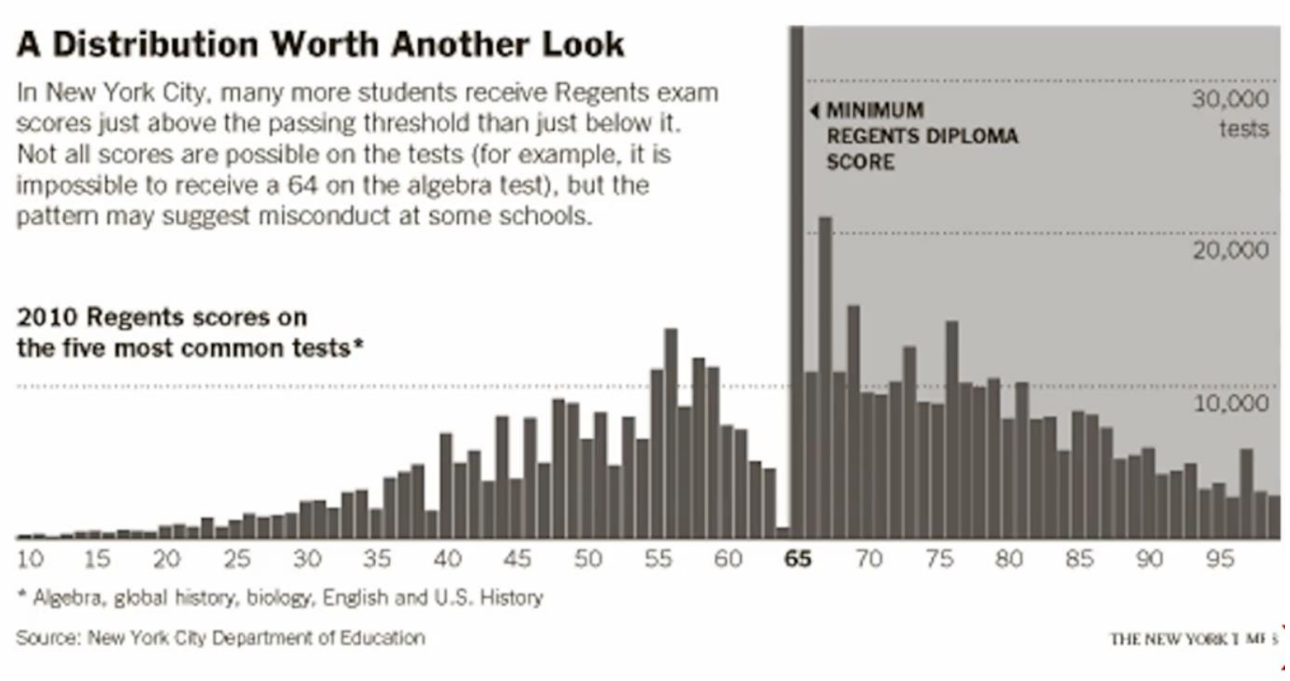
○ the New York City regents exam (**standardized examinations in core high school subjects)**.

○ In New York City, you need a 65 to pass.

○ What important problem this plot is showing?

BONUS

15 sec



A Distribution Worth Another Look

In New York City, many more students receive Regents exam scores just above the passing threshold than just below it. Not all scores are possible on the tests (for example, it is impossible to receive a 64 on the algebra test), but the pattern may suggest misconduct at some schools.

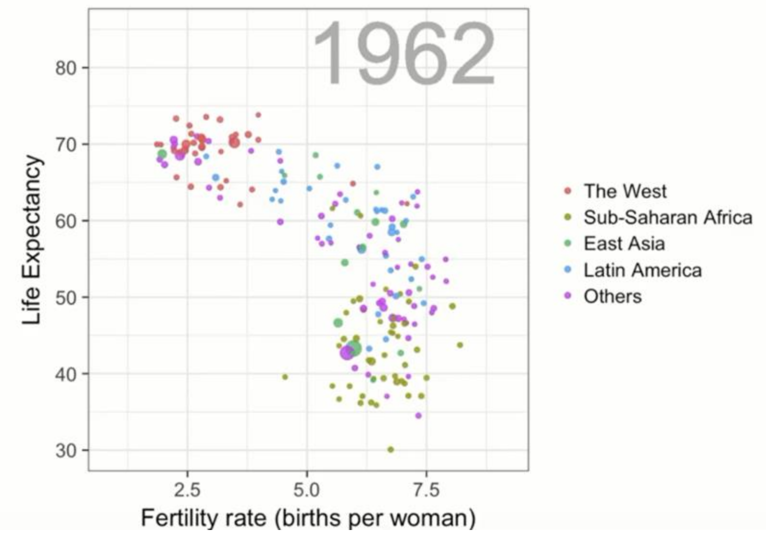
2010 Regents scores on the five most common tests*

* Algebra, global history, biology, English and U.S. History

Source: New York City Department of Education

Introduction to Data Visualization

- These are examples of how data visualization can lead to discoveries which would otherwise can be missed.
- Data visualization is the strongest tool of what we call **Exploratory Data Analysis.**
- "The greatest value of a picture is when it forces us to notice what we never expected to see." **John W. Tukey (considered as the father of exploratory data analysis)**
- Example: New Insights on Povertyand the Best Stats You've Ever Seen, Hans Roslings → forced us to notice the unexpected



Introduction to Data Visualization

- It is also important to note that
  - mistakes
  - biases
  - systematic errors

lead to data that should be handled with care.

Introduction to Data Visualization

- Our content is:
  - basics of data visualization and exploratory data analysis (EDA)
  - ggplot2 package
  - motivating examples
  - It is impossible to cover everything but we will make a good introduction.

Introduction to

Distributions

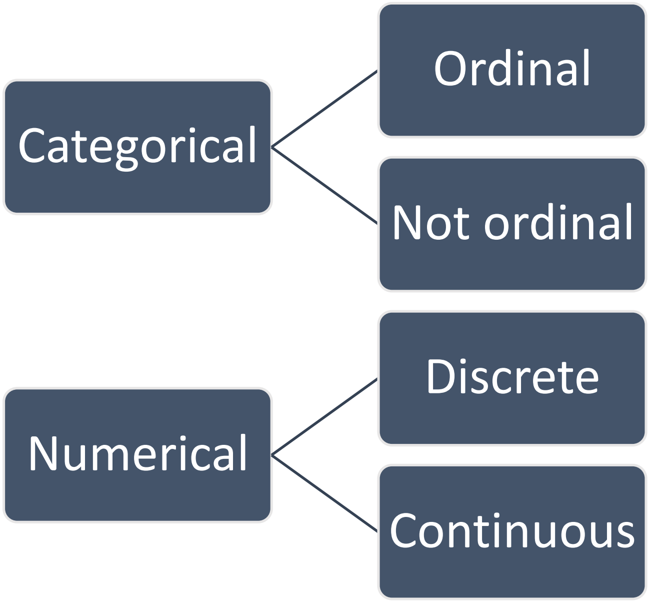
Introduction to Distributions

- Numerical data is often summarized with an **average value.**
- Occasionally, a second number is reported as well-- **the standard deviation**.
- For example, a report stating that scores at this high school were 680 ± 50.
- Is there any important information we're missing by only looking at this summary rather than the entire list?
- **We will learn to summarize lists of factors or numeric factors**
- **The most basic statistical summary of a list of objects or numbers ?????**
- **→ its distribution**

Introduction to Distributions

- Categorical: → a small number of groups
  - e.g. Regions: Northeast, South, North Central, West → **not ordinal**
  - Some categorical data can be ordered→ **ordinal data,** ex: spiciness: mild, medium, hot
- Numerical → Population size, murder rates, heights
  - **Continuous data:** can take any value, e.g. Heights
  - **Discrete:** population sizes (rounded numbers)

Variable types:



Introduction to Distributions | Categorical Data

Example Application

- We have to describe the heights of our classmates:

  1. Collect data: we ask students to report their heights

  2. Report their sex (male or female)

  3. Collect the data and save it in a data frame

```
library(dslabs)
data(murders)
head(heights)
```

```
         sex   height
1       Male       75
2       Male       70
3       Male       68
4       Male       74
5       Male       61
6     Female       65
```

- There are 1,050 heights.

- We can report a list of 1,050 heights. However, there are much effective ways to convey this information.

Introduction to Distributions › Categorical Data

- The most basic statistical summary of a list of objects or numbers is ????

  ➢ **its distribution.**

- For example, with categorical data, the distriburtion simply describes the proportions of each unique category.

- For example, the sex can be summarized by the proportions of categories:

```
prop.table(table(heights$sex))
Female          Male
0.2266667   0.7733333
```

Frequency table: the simplest form of a distribution

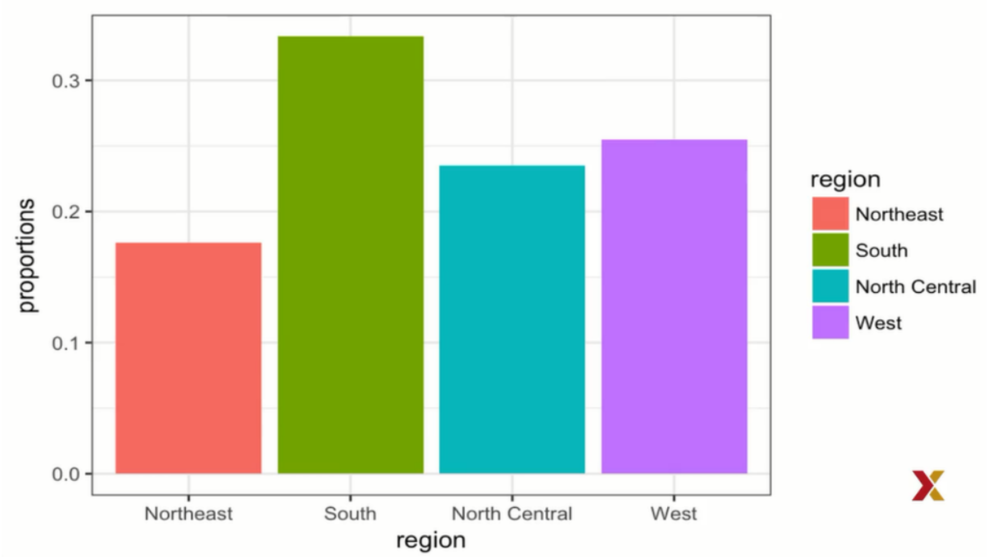
No need to see more, a number describes all story.

Here, 23% are females and the rest are male.

Introduction to Distributions | Categorical Data

- ○ When there are more categories, the simplest form is **barplot.**

- ○ **Convert a vector into a visualization that summarizes all the information in the vector.**



Introduction to Distributions | Numerical Data

- **Numerical data→** The task is much more challenging!

- We cannot report the frequency of each unique entry → as they are often unique

  - For example, students reported a height of 175 cm, but one student reported a height of 175.5039330007874.

- It is impossible to show frequency for each data in this way.

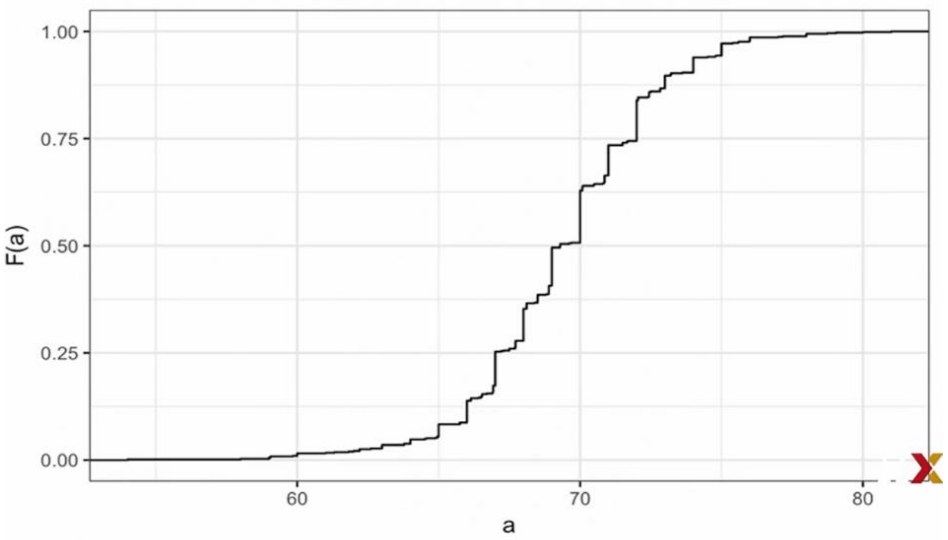
- Statistics textbooks teach us that a more useful way → **CDF: cumulative distribution function**

  - define a function that reports the proportion of the data below a value a for all possible values of A.

  - $F(a) = Pr(x \leq a)$

The proportion of value x less than or equal to a

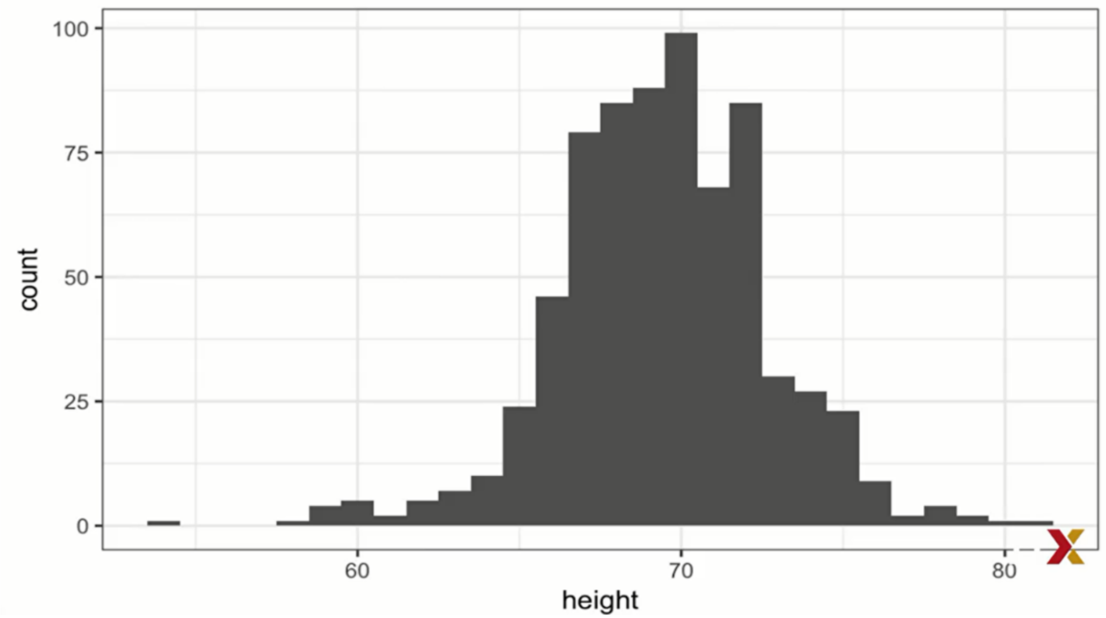
- Plot of function F ($F(a) = \Pr(x \leq a)$) for the height data:



- ✓ Distribution for numerical data.
- ✓ 14% of students in the class have a height below 66 inches.
- ✓ 84% of our students have heights below 72 inches.
- ✓ Popular in textbooks but it is not popular in practice??
  - ✓ It does not exactly convey characteristics of interest
  - ✓ For example, distribution center, symmetry etc.
- ✓ Histograms are much preferred as they can answer these questions

- The simplest way to make a histogram:
  - divide a span of our data into non-overlapping bins of the same size.
  - for each bin, count the number of values that fall in that interval.
  - the histogram plots these counts as bars with the base of the bar is the interval

→ split data into 1-inch intervals



→ What we can see??

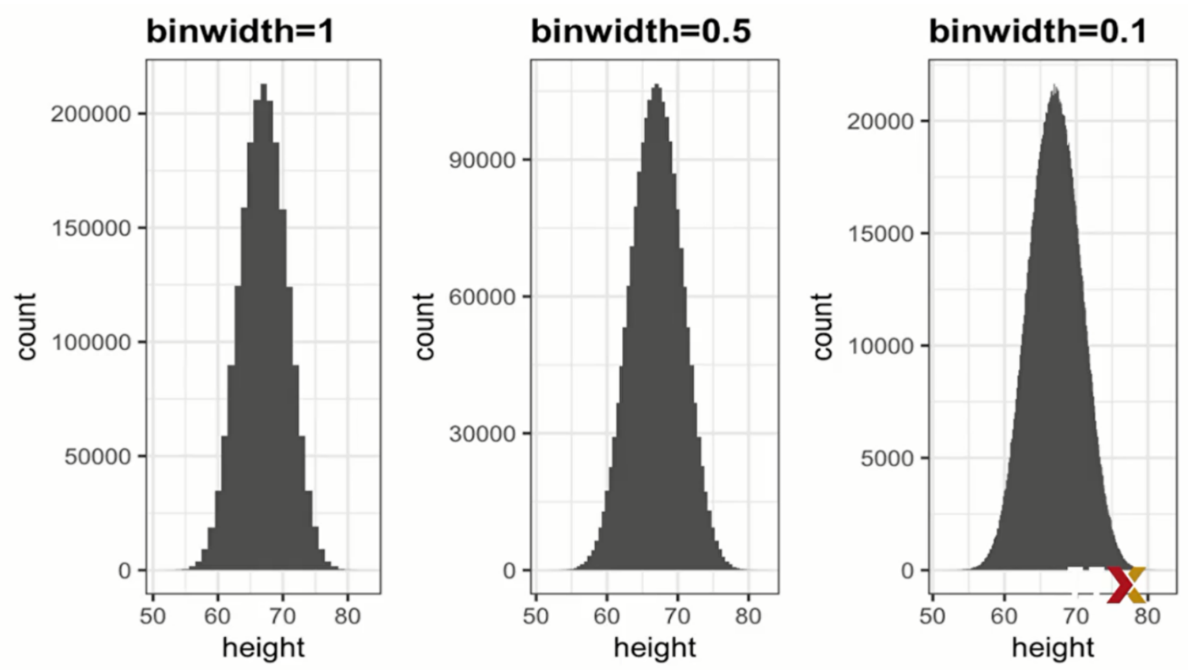
1. The range is from 55 to 81
2. More than 95% are between 63 and 75 inches.
3. Symmetric around 69 inches.
4. Proportion of data at any interval by adding up counts

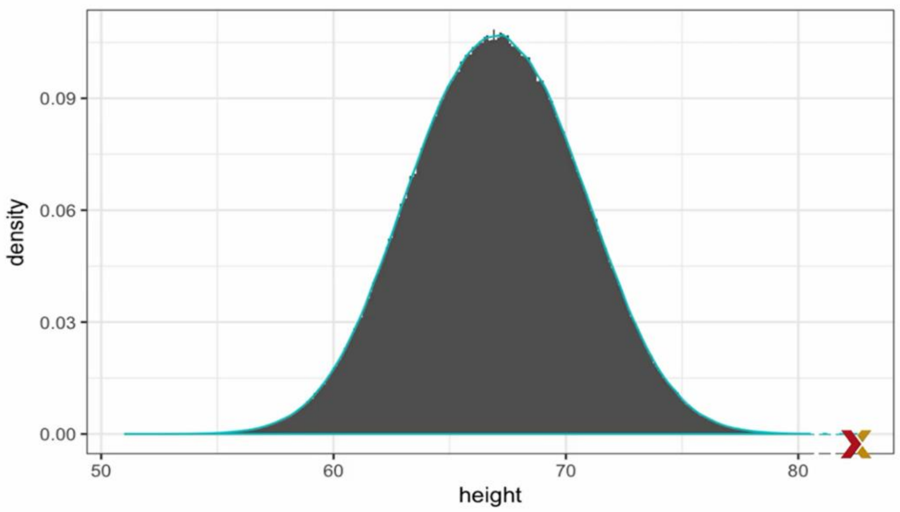
Almost all the information that we provided with CDF, and all the information in the raw data (708 heights)

- Histogram is an approximation, why??
- What information do we lose?
  - Each interval is assumed to be the same when computing the bin counts
  - For example, the histogram does not distinguish between 64, 64.1, and 64.2 inches.

Distributions Numerical Data Smooth Density Plots

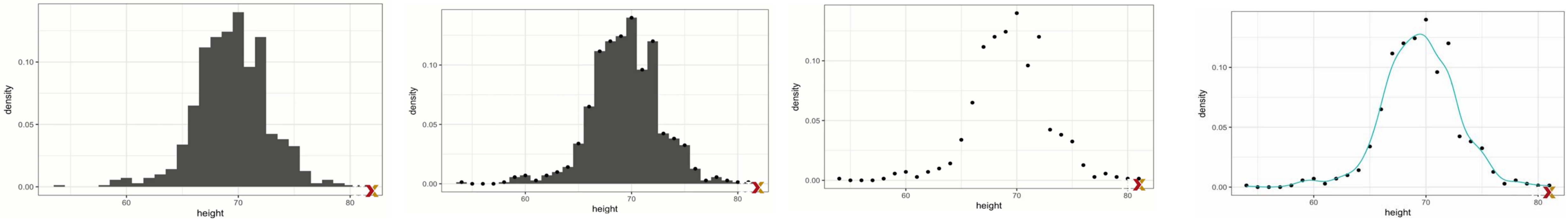
- Similar to histograms but aesthetically more appealing.
- **Male heights data:**
- Assume, hypothetically, we have 1 million heights data and we can make very small bins.



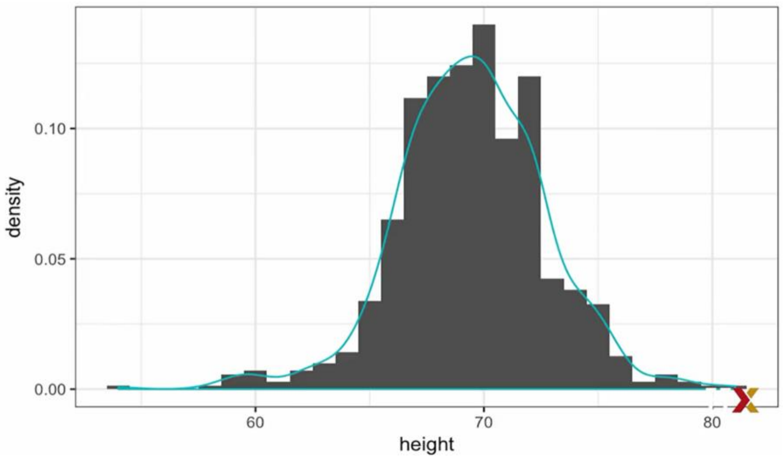


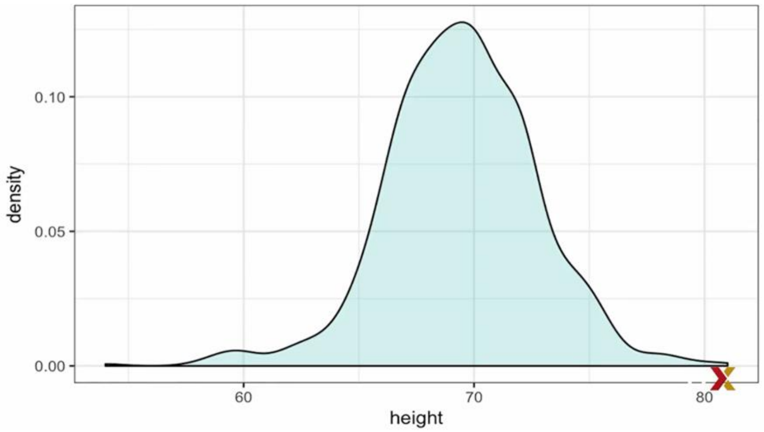
Distributions › Numerical Data › Smooth Density Plots

- To make to curve not depend on the hypothetical size of the hypothetical list, we compute the curve on the **frequency scale.**
- In our example, we have 708 measurements and we cannot make a histogram with very small bins.
- How can we estimate a hypothetical smooth curve?
  - 1. make a histogram with our data
  - 2. compute frequencies rather than counts
  - 3. choose an appropriate bin size



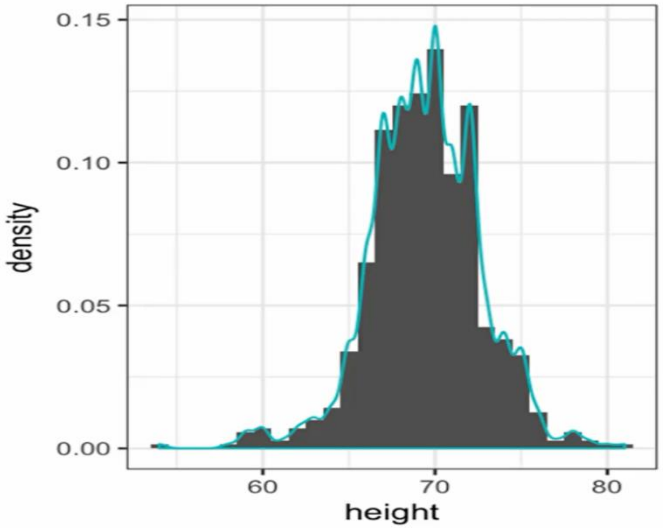
Distributions › Numerical Data › Smooth Density Plots

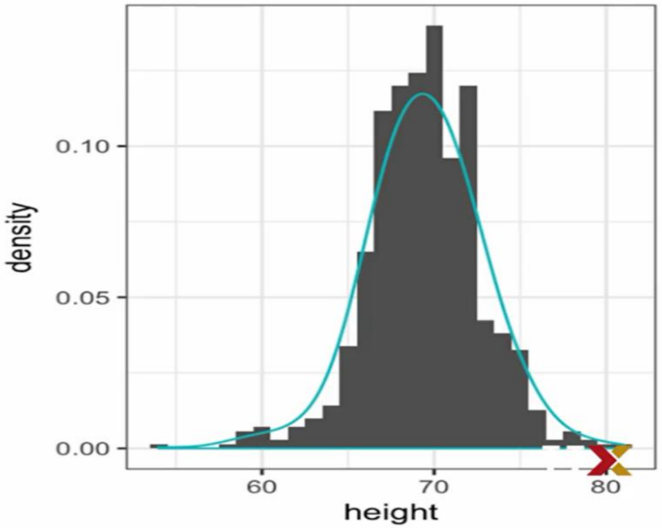


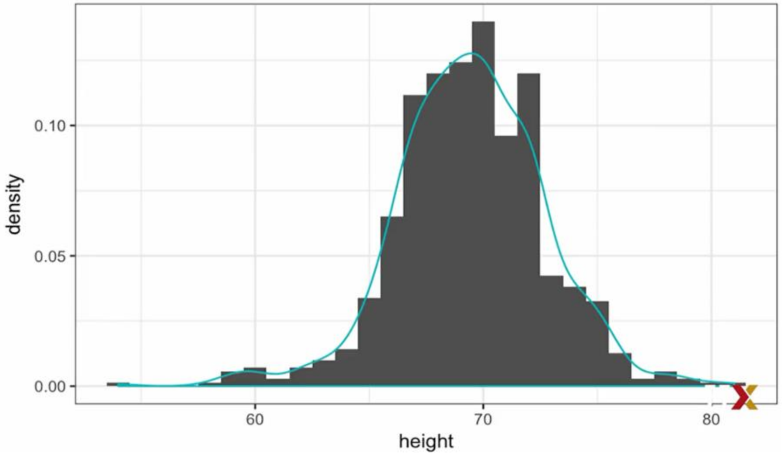


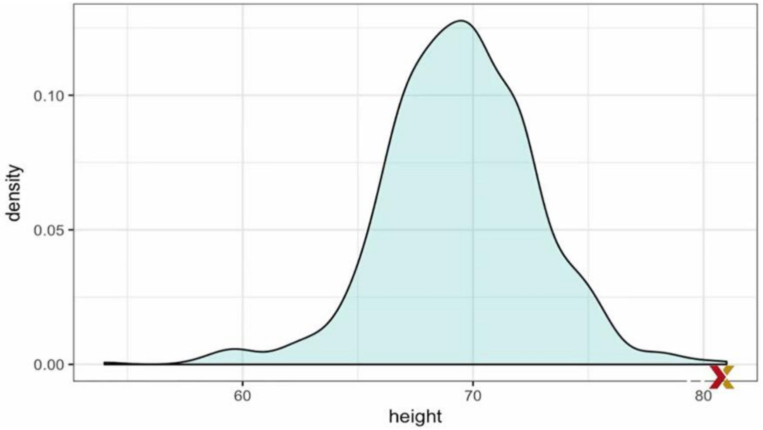
- Smoot is a relative term: → We can control the smoothness of the curve

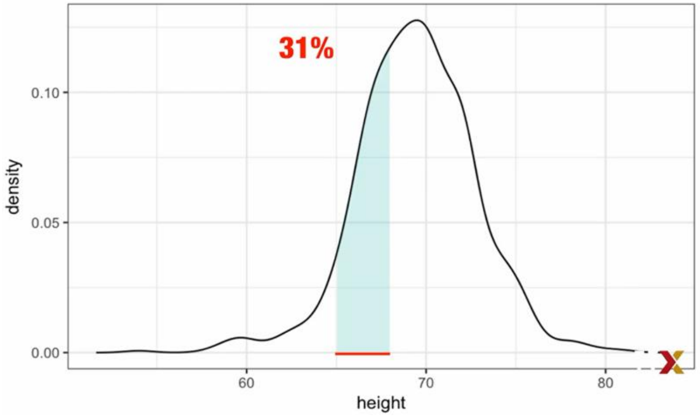
- We should select a degree of smoothness that we can defend as being representative of the underlying data.





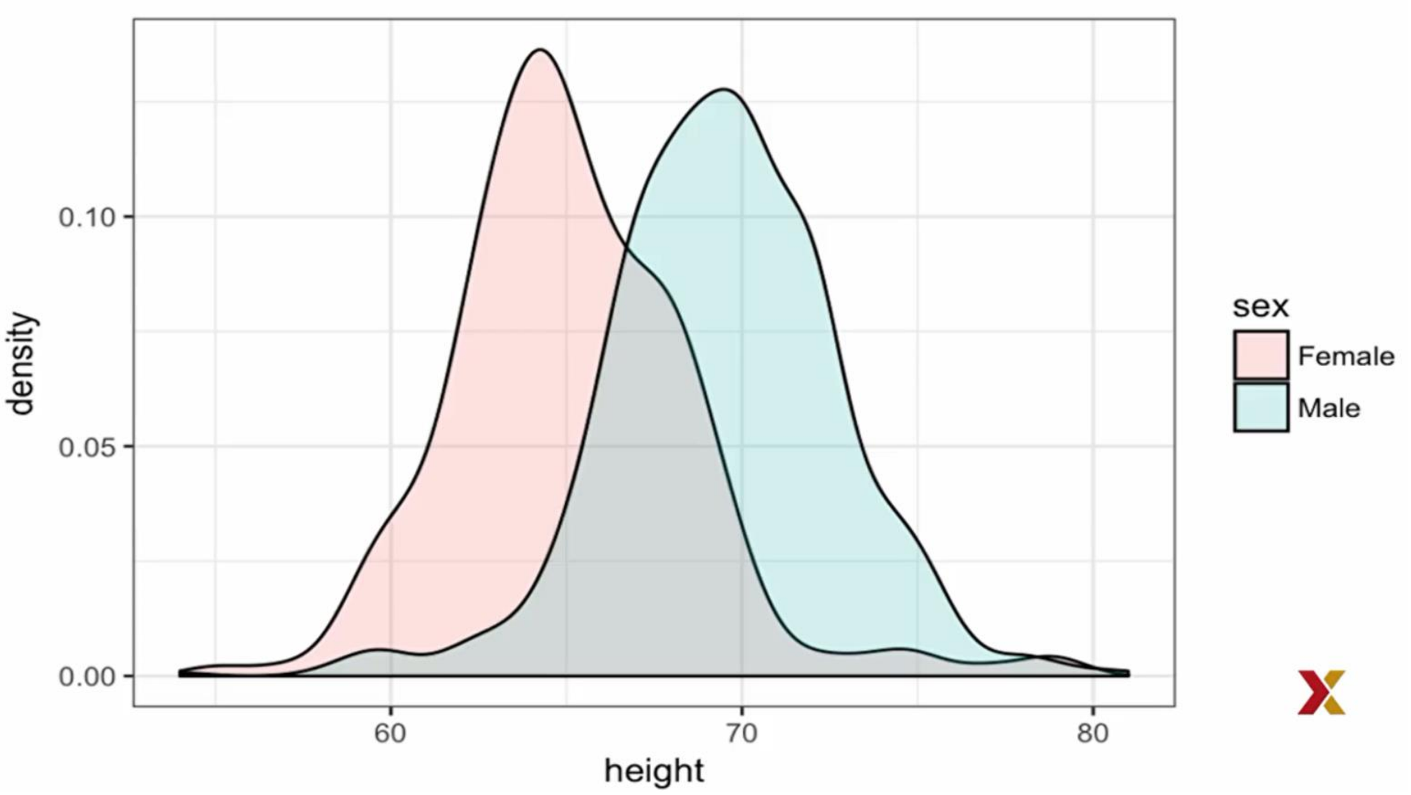






- y-axis?
  - not easy to understand.
  - It is scaled so that the area under the density curve adds up to 1.
  - Proportion of the total area contained in an interval.
  - For ex. 31% our values are between 65 and 68 inches.

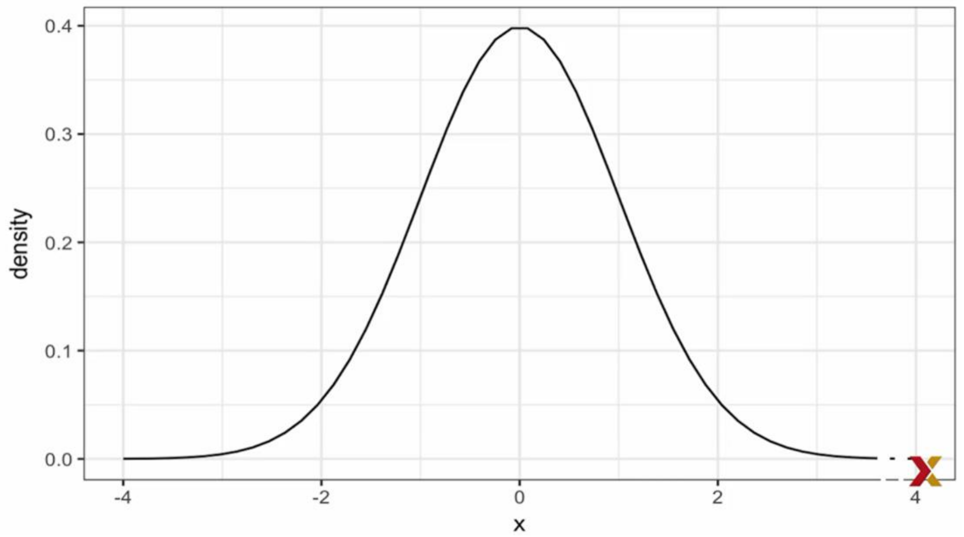
- Comparing two data sets are easier with smooth density plots compared to histograms.



- Mean and standard deviation → normal distribution (bell curve, Gaussian distribution)
- one of the most famous mathematical concepts → approximately normal distributions occur in many situations.
- Our focus will be on **how the normal distribution helps us summarize data.**
- Proportion of values in the interval (a, b) is computed using:
  - (no need to remember this as we will have R code that computes it)
  - two parameters: m (mean) and s (standard deviation) → **enough the describe all the dataset**

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2} dx$$

- The distribution is
  - Symmetric
  - Centered at the average
  - Most values (95%) are within two standard deviations from the average

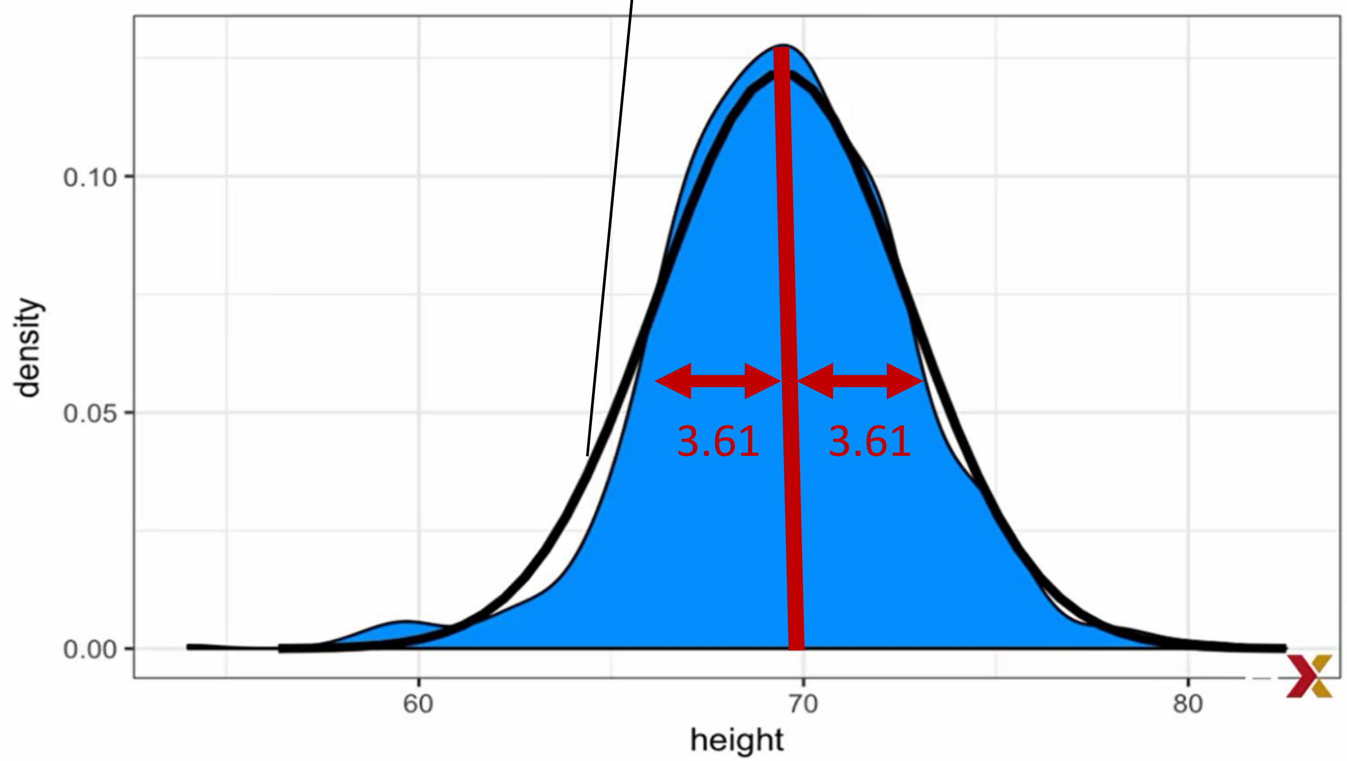


Mean → 0, standard deviation → 1

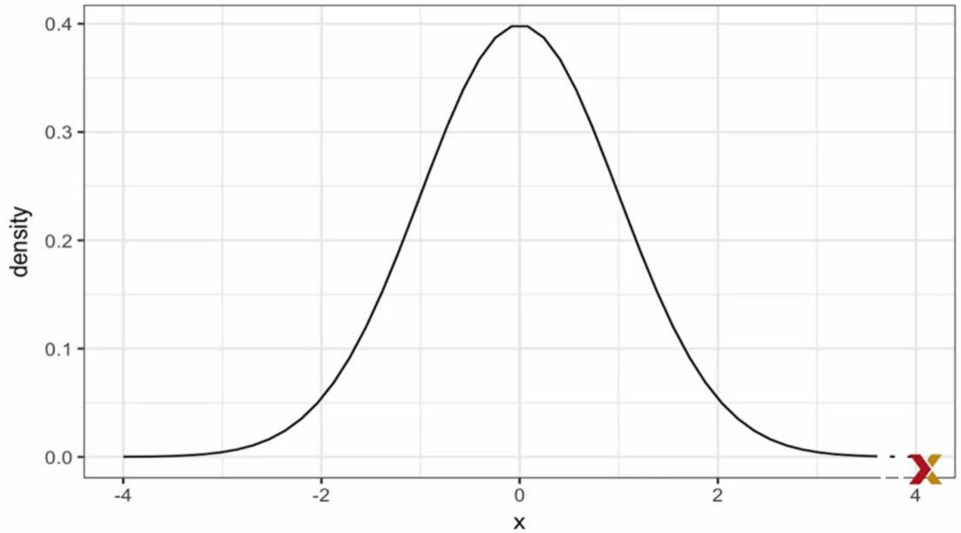
Distributions › Numerical Data › Normal Distribution

```
# assume x is a vector
average <- sum(x) / length(x)
sd <- sqrt( sum( (x-average)^2) / length(x))

library(dslabs)
data("heights")
index <- heights$sex == "Male“
x <- heights$height[index]
average <- mean(x)
sd <- sd(x)

c(average = average, sd=sd)
  average        sd
    69.31      3.61
```

Normal distribution approximates the distribution of our male heights very well



- The distribution is
  - Symmetric
  - Centered at the average
  - Most values (95%) are within two standard deviations from the average



Mean → 0, standard deviation → 1

- **Standard units**
- For data that is approximately normal, we can think in terms of standard units.
- Standard unit of value tells us how many standard deviations away from the average this value is.
- Specifically, for value x,

$$z = (x - average)/sd$$

- If we convert normally distributed data into standard units, we can quickly know more information about the value. For example,
  - $z = 0$ → person is about average height
  - $z = 2$ → person is tall
  - $z = -2$ → person is short
  - $z > 3$ → extremely tall
  - $z < -3$ → extremely short

It does not matter what the original units are!!!

In R, we can obtain standard units

```
z <- scale (x)
```

- How many men are within 2 standard deviations from the average?

```
# number of z's that are less than 2 and bigger than negative 2
sum(abs(z) < 2)
[1] 771
```

```
mean(abs(z) < 2)
[1] 0.95
```

95% ???

BONUS

15 sec

%95 of the men have heights that are within 2 standard deviations from the average

- If we can assume that the data is approximately normal,
  - We can predict the proportion without actually looking at the data
  - We simply know that 95% of the data, for normally distributed data, is between - 2 and 2.

- Cumulative distribution for the normal distribution → R has a build in function, `pnorm()`

- $\mathrm{F(a) = pnorm(a, avg, s)}$

- If we use normal distribution approximation, we don't need to see entire data set.
  - What is the probability that a randomly selected student is taller than 70.5 inches.

```
pnorm(70.5, mean(x), sd(x))
[1] 0.628631
```

- The normal distribution **is derived mathematically.**
  - We only need data to calculate the **mean** and **standard deviation**. We do not need anything else.
- The normal distribution is defined for c**ontinuous variables.**
- With continuous distributions, the probability of a singular value is not even defined.
  - It does not make sense to ask what is the probability that a normally distributed value is 70.
  - We use intervals → **what is the probability that some is between 69.99 and 70.01**
  - For datasets where the data is rounded (for example heights), the normal approximation is particularly useful when **intervals include exactly one round number.**

```
mean(x <= 68.5) - mean(x <= 67.5)
[1] 0.114532
mean(x <= 69.5) - mean(x <= 68.5)
[1] 0.1194581
mean(x <= 70.5) - mean(x <= 69.5)
[1] 0.1219212
```

```
pnorm(68.5, mean(x), sd(x)) - pnorm(67.5, mean(x), sd(x))
[1] 0.1031077
pnorm(69.5, mean(x), sd(x)) - pnorm(68.5, mean(x), sd(x))
[1] 0.1097121
> pnorm(70.5, mean(x), sd(x)) - pnorm(69.5, mean(x), sd(x))
[1] 0.1081743
```

```
mean(x <= 70.9) - mean(x <= 70.1)

[1] 0.02216749
```

```
pnorm(70.9, mean(x), sd(x)) - pnorm(70.1, mean(x), sd(x))

[1] 0.08359562
```

- ✓ This situation is called discretization.
- ✓ Although the true height distribution is continuous, the reported heights tend to be at discrete values.
- ✓ This is because of rounding.

Introduction to Quantiles

Qantiles

- We can use quantile-quantile, or q-q plots, to check how good is the normal approximation.
- $p = 0.05, 0.1, 0.15, \dots .0.95$
- For each p, we determine q, so that the proportion of the values in the data below q is p.
- q's are referred as quantiles.

- Male height data:
- 50% of the data is below 69.5 inches

```
mean(x <= 69.5)

[1] 0.515
```

- If $p\ = 0.5$, then $q\ = 69.5$

Qantiles

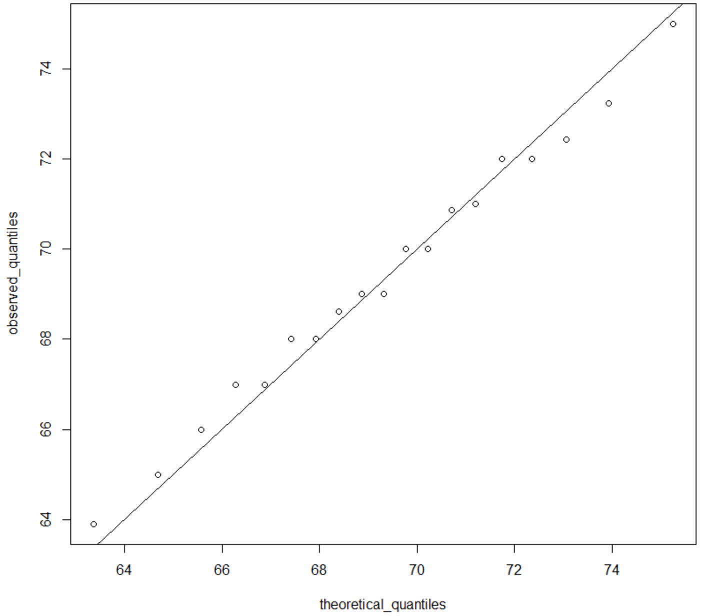
- If the quantiles for the data match the quantiles for the normal distribution → data is approximated by a normal distribution.
- Observed quantiles:

```
observed_quantiles <- quantile(x, p) # observed quantiles
```

- Theoretical normal distribution quantiles:

```
theoretical_quantiles <- qnorm(p, mean = mean(x), sd = sd(x)) # theoretical quantiles
```

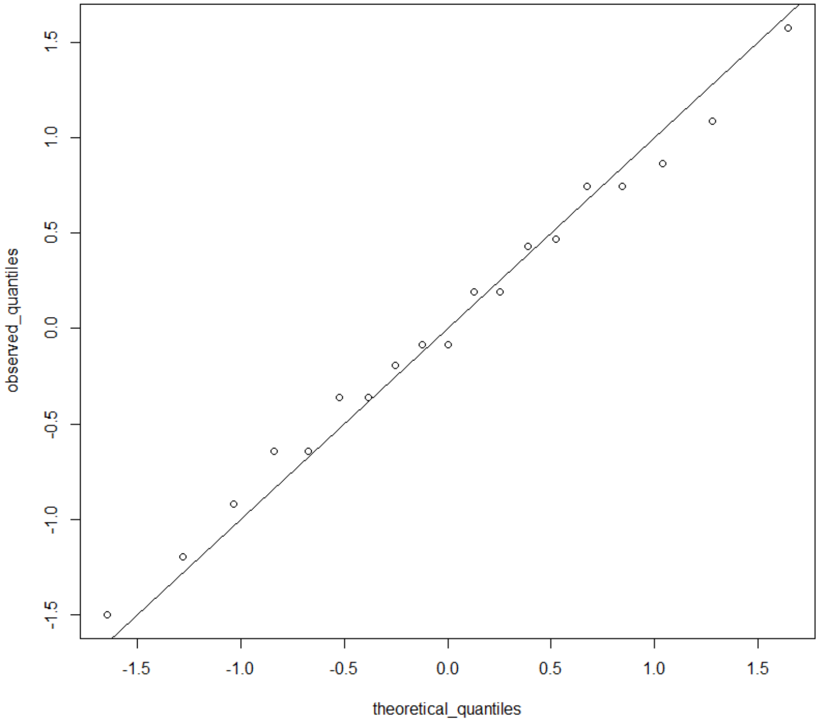
```
p <- seq(0.05, 0.95, 0.05)
observed_quantiles <- quantile(x, p)
theoretical_quantiles <- qnorm(p, mean = mean(x), sd = sd(x))
plot(theoretical_quantiles, observed_quantiles)
abline(0, 1)
```



Qantiles

- It is much easier if we use standard units → we do not need to define the mean and standard deviation

```
p <- seq(0.05, 0.95, 0.05)
z <- scale(x)
observed_quantiles <- quantile(z, p)
theoretical_quantiles <- qnorm(p)
plot(theoretical_quantiles, observed_quantiles)
abline(0, 1)
```



Qantiles

- For the male height data,
  - Histograms
  - Density plots
  - q-q plots

→

- The data is well approximated by normal distribution

- **Reporting a summary as a data scientist**
  - ➢ normal distribution, average = 69.44 inches, sd = 3.27 inches
  - This will be enough to describe everything!

Percentiles

- **Percentiles are special cases of quantiles**
- Quantiles that you obtain when you define $p = 0.01, 0.02, \dots, 0.99\ (1\%, 2\%, \dots .99\%)$
- The famous percentile is 50th → median
- For the normal distribution, the median and the average are the same.

Quartiles

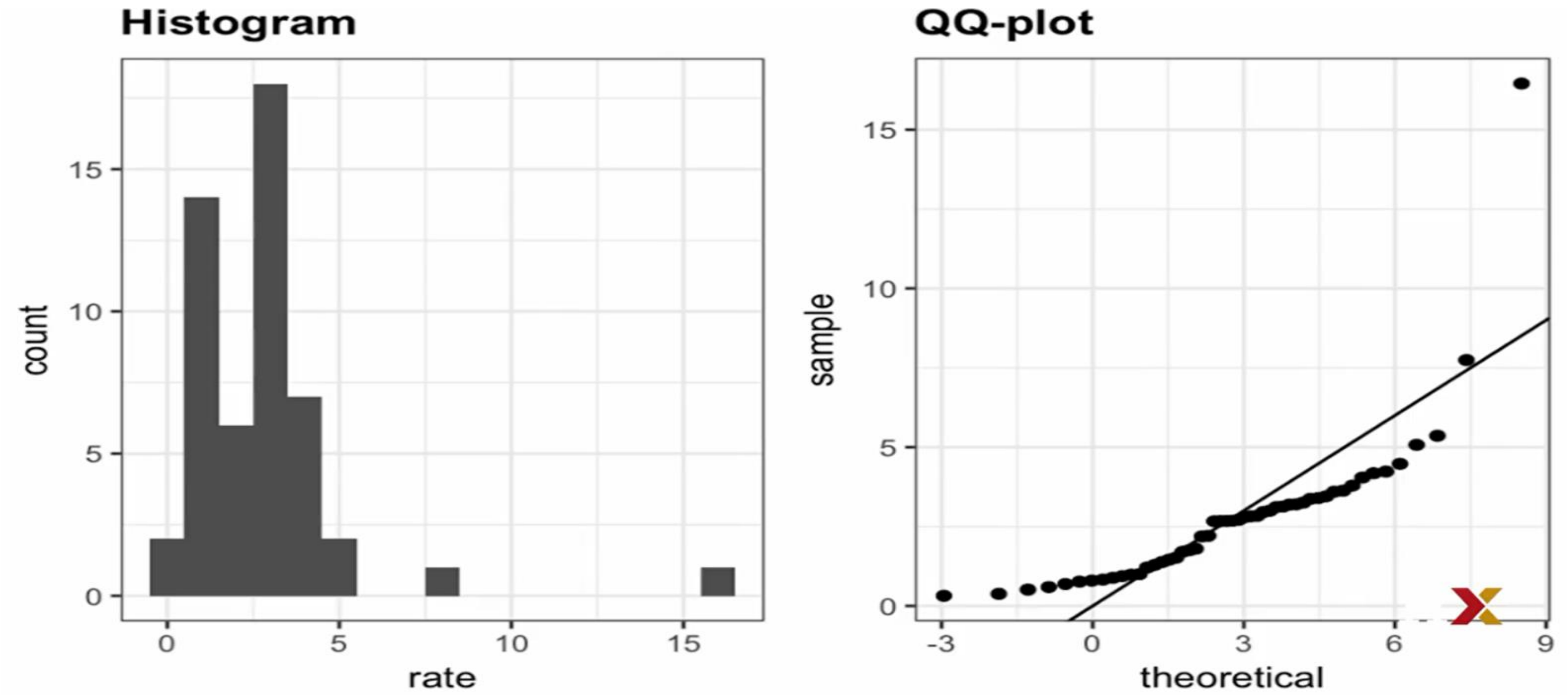
Quartiles:

Quantiles that you obtain when you define $\boldsymbol{p} = \mathbf{0.25, 0.50, 0.75}$

Boxplots

Boxplots

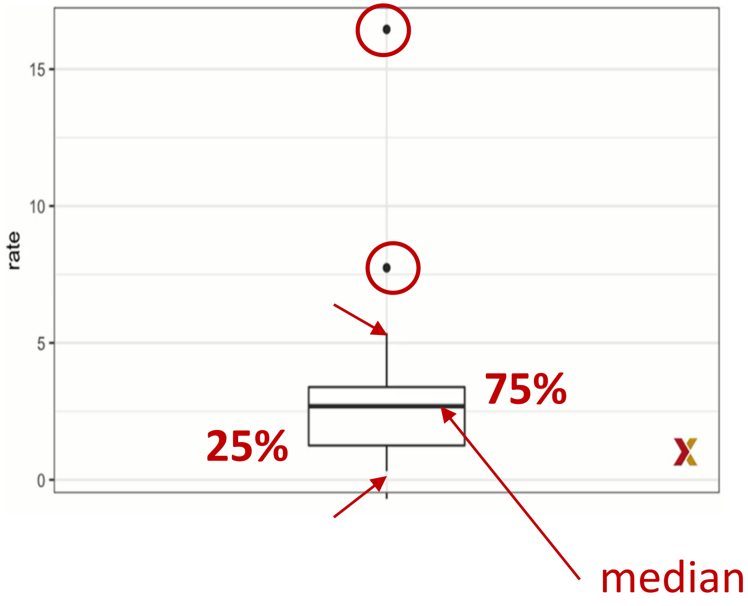
- Go back to our *muder data set* and summarize the *murder rate distribution*



Normal approximation???

Boxplots

- John Tukey:
  - Provide a five-number summary composed of the **range** along with the quartiles (25th, 50th, 75th percintiles)
  - Ignore outliers when computing the range, plot them independently
  - Plot this as a box with whiskers



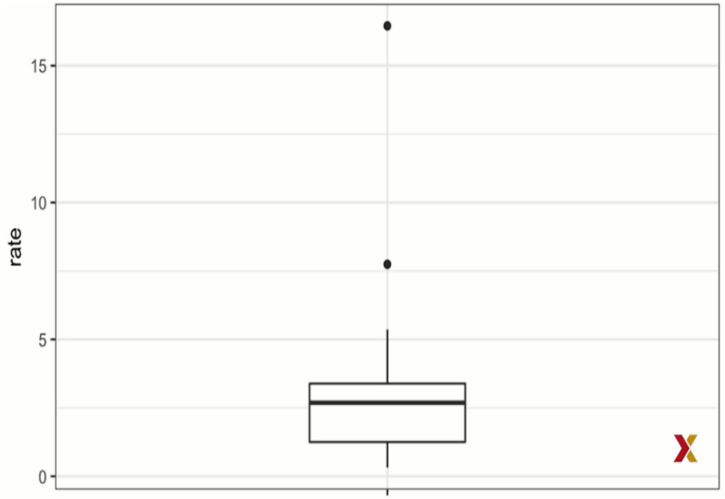
75%

25%

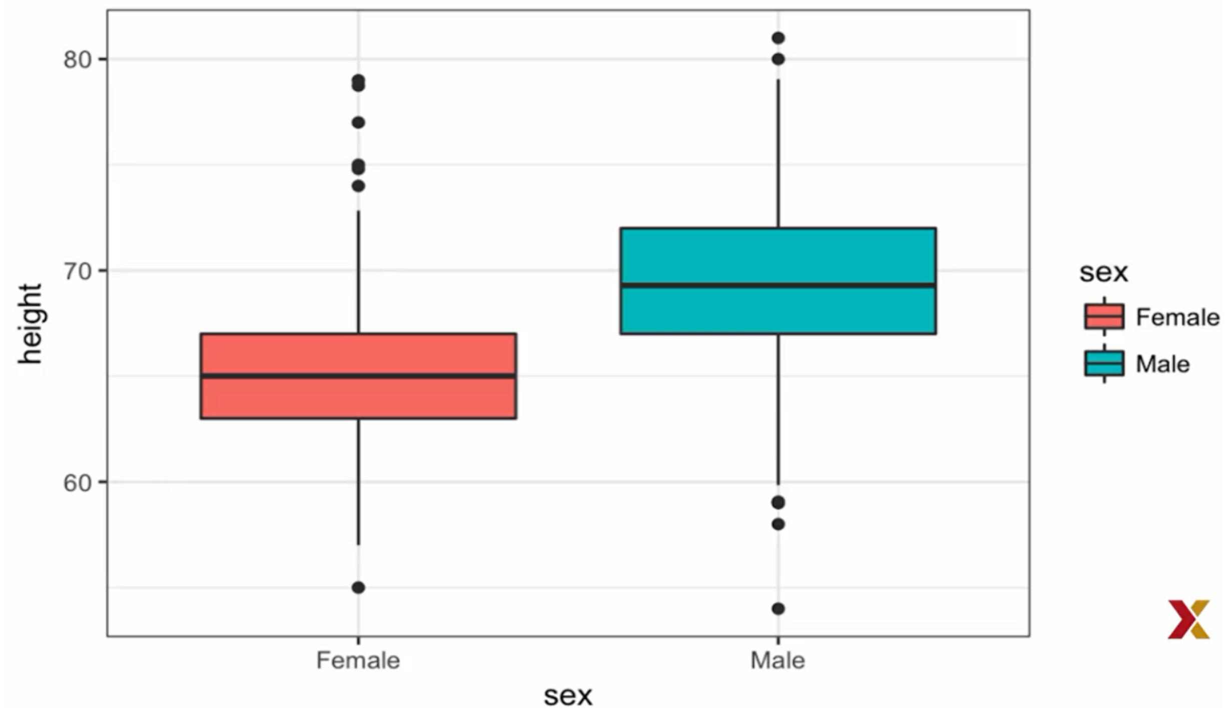
median

Boxplots

- Summary:
  - Median is about 2.5
  - Distribution is not symmetric
  - The range is 0 to 5, for the great majority of states with two exceptions



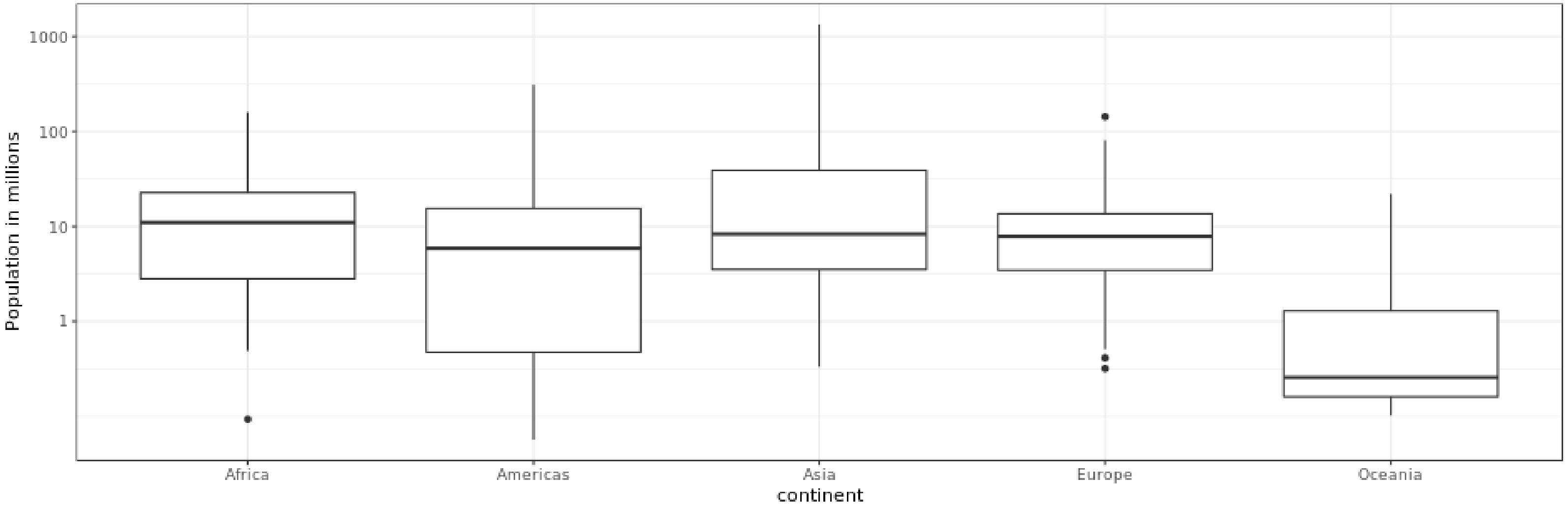
Boxplots



Boxplots

15
sec

Exercise



1- Which continent has the country with the largest population size?

2- Which continent has the largest median population?

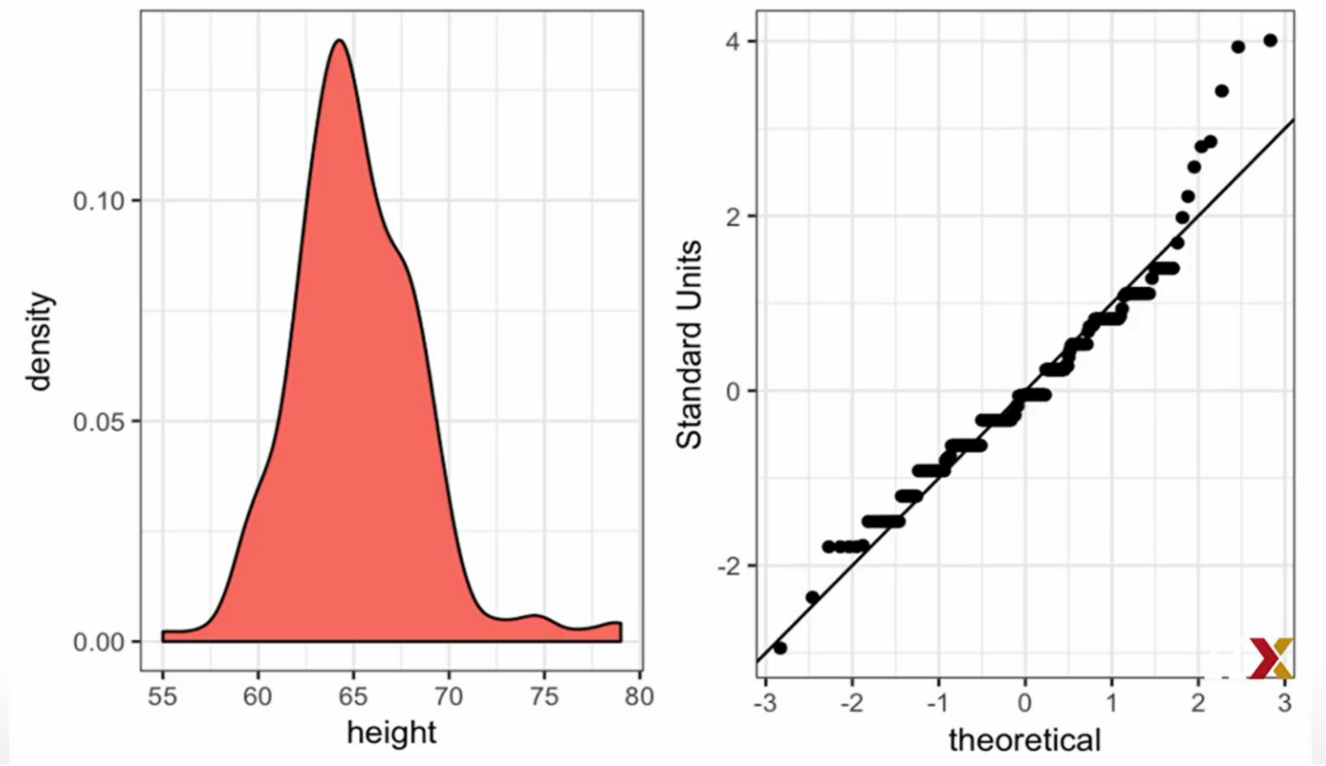
3- What proportion of countries in Europe have populations below 14 million? (0.75, 0.50, 0.25, 0.01)

Exploratory Data Anaylsis Example

An Example Story

- Assume that now your task is describing the heights of **female** students.

- We start by checking normal distribution.



- **Is this normally distributed?**

BONUS

15 sec

An Example Story

- What would we do?
    - We can not report just the average and the standard deviation.
    - We can provide a histogram
- A good data scientist should do more: “we noticed something that we did not expect to see”
    - **“The greatest value of a picture is when it forces us to notice what we never expected to see.” John W. Tukey)**

An Example Story

- If we look at other sources on web to see about other female height distributions, we find they are well approximated by the normal distribution.
- Why do we have a second hump?
- Why do we see so many outliers, so many taller than expected women? What would be the reasons?
  - Is our female population from a basketball team?
  - Are they claiming to be taller than they are?
  - **Answer: In the form students enter their height, the default sex is "Female". Some males may have entered their heights, but they forgot to change their sex variables. → Females in our data set are actually males.**

Any
Questions