DataSci 520

lesson 1

data exploration

Dr. Mohamed Mneimneh

PROFESSIONAL &
CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Today's Agenda

- Class Introduction
- Class policy and expectations
- How to succeed in this course
- Assignments, quizzes and milestones
- Participation and grading
- Grading expectations and supplementary material
- Coding environment
- Overview of Data Exploration

Course Overview

- Lesson 01: Data Exploration
- Lesson 02: Effective Data Visualization
- Lesson 03: Combinatorics and Probability Distributions
- Lesson 04: Sampling Methods
- Lesson 05: Sampling and the Central Limit Theorem
- Lesson 06: Hypothesis Testing
- Lesson 07: Bayesian Statistics
- Lesson 08: Applications of Bayesian Methods
- Lesson 09: Linear Models, Part 1
- Lesson 10: Linear Models, Part 2

Instructional Team

- Mohamed Mneimneh, Lecturer (mneimneh@uw.edu)
- Prameela Gunturu, Instructional Assistant (prameelakarumanchi@gmail.com)

Student Introduction

- Name and current role
- Why enrolled in data science certificate
- Course and program expectations

Course Format

- On average, we spend about 40% on lecture and 60% on hands-on
- Let's have frequent discussions during the session and share diverse perspectives
- Let's take frequent short breaks to fit online format

How to succeed in this course

- Pay attention during class and ask questions
- Run notebooks alongside but don't let it distract you too much
- Learn from peers by participating in discussion boards and by sharing your code after assignment due dates
- Do your best not to fall behind, because lessons build on each other
- Basic to intermediate Python programming concepts should be second nature by now

Assignments, quizzes and milestones

- Assignments are due by 11:59 PM one week after lecture
- I will make no exceptions about assignment due dates
- Quizzes are taken in Canvas
- Assignments are graded by Prameela
- Questions about the assignments should be posted in EdDiscussion, where instructional team will answer them (or students if they wish)

How to submit a great assignment

- Your code should run from start to finish, so once you finish the assignment, restart the notebook and run again to make sure nothing breaks
- Each step in the assignment should include (1) one code cell with a few comments in the code to point out key parts, and (2) a Markdown cell explaining your reasoning and (3) a Markdown cell explaining your conclusions
- Someone who doesn't know Python could still be able to read through and understand the analysis
- When faced with ambiguity, you are free to make assumptions as long as you (1) clearly state your assumptions and (2) it is a reasonable assumption

let's take a tour through Canvas

Grading expectations

- Please use the discussion boards for questions on assignment
- DO NOT post your assignment code in discussion boards
- When necessary, it's generally OK to make an assumption about the assignment reqs - as long as you state your assumption
- More important to be on time than perfect:
- Meet the reqs with working code
- Write good comments for your code
- Write good explanations showing your line of thinking

Participation and grading

Activity	What you need to do	Grade
Participation	Be active in discussion boards	20%
Quizzes	Complete Short Quizzes	20%
Assignments	Submit by 11:59 PST the week after	60%

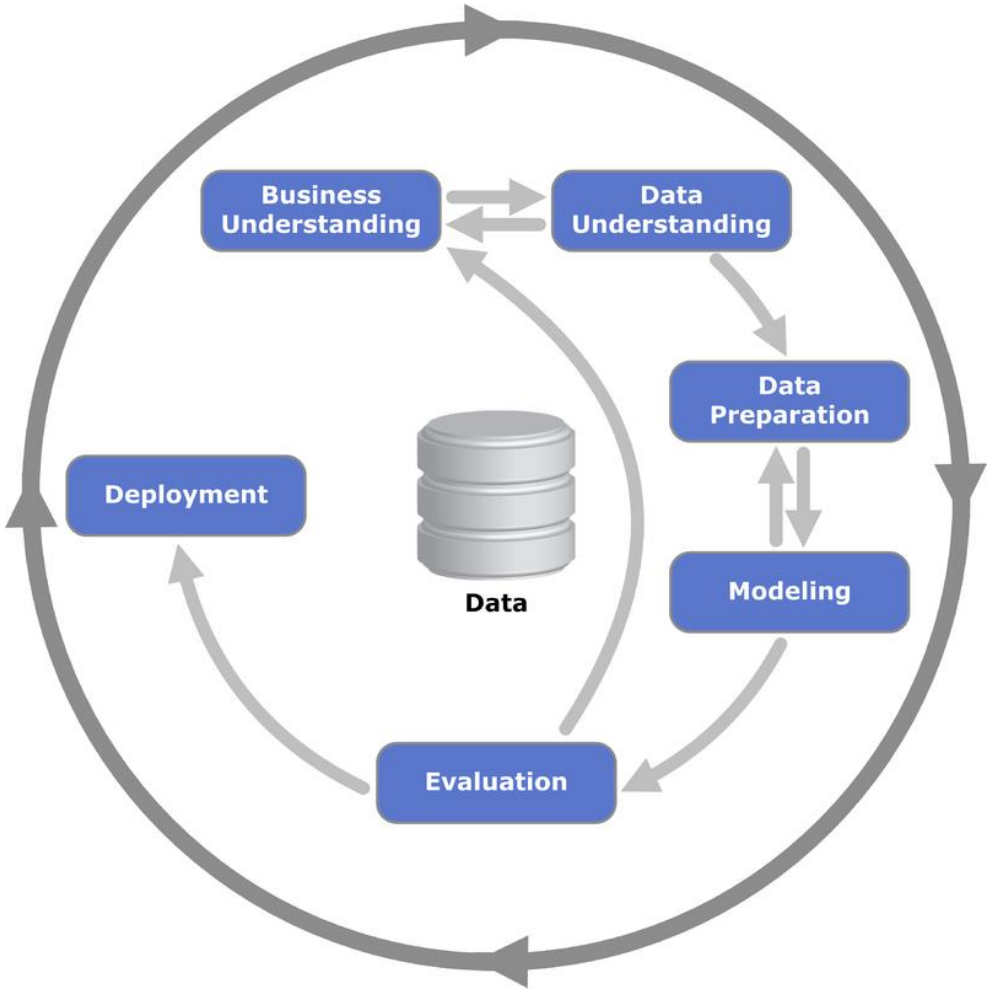
Coding Environment

- We will be using browser-based Jupyter notebooks as our python environment
- Basics of Jupyter notebooks:
  - Code cells and Markdown cells
  - Running and re-running code cells - order matters!
  - Magics are very useful shortcuts

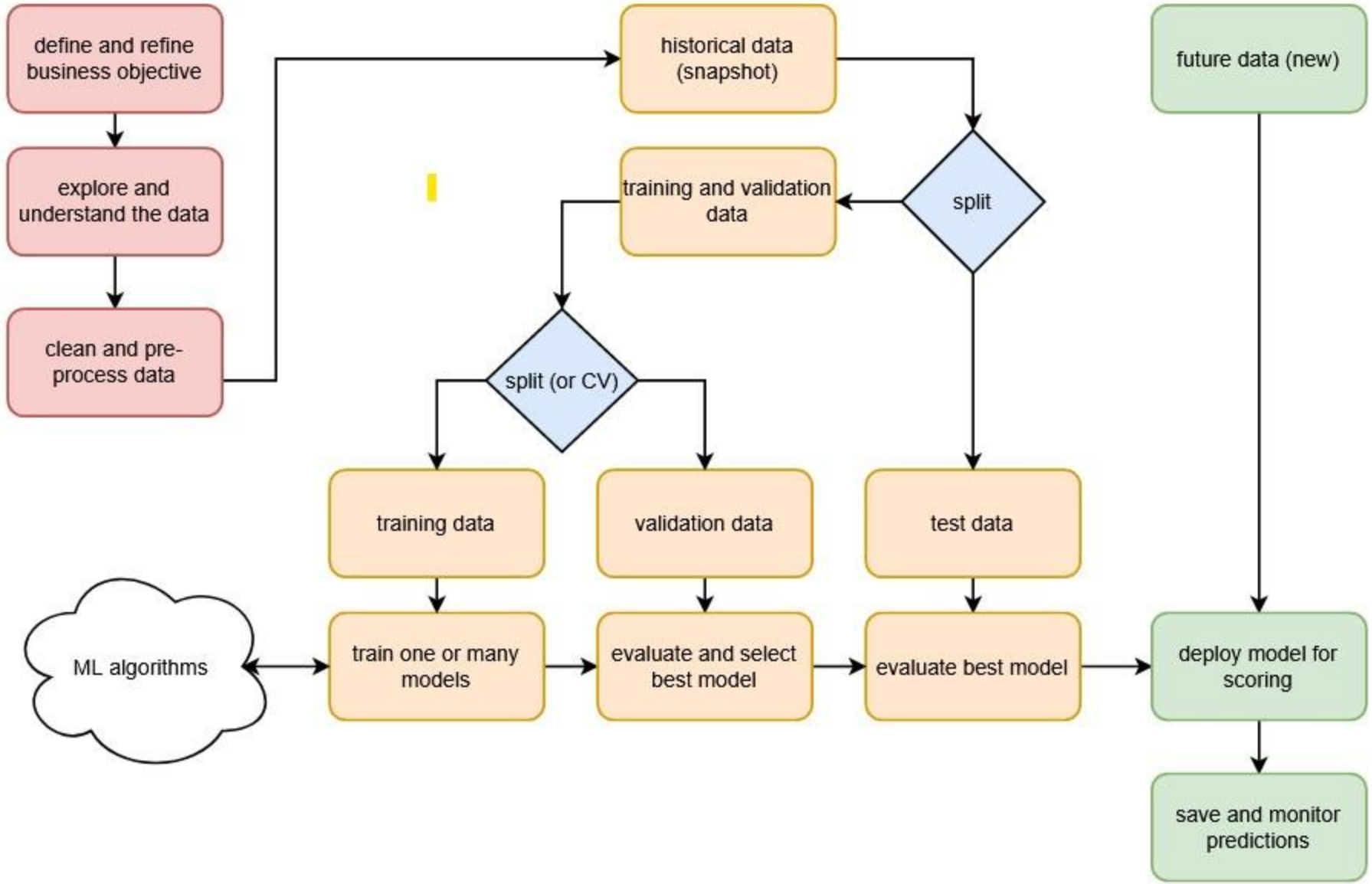
before we start

- in this lesson we are limiting mathematical formulas and derivations to a minimum
- the goal is to learn the intuition and without being bogged down by the math
- many of the concepts we learn (percentile, skewness, kurtosis) have multiple derivations with slight veriations between them
- we only include the important math and leave it to students to learn the rest

CRISP-DM



Source: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining


define and refine business objective
explore and understand the data
clean and pre-process data
historical data (snapshot)
training and validation data
split
future data (new)
split (or CV)
training data
validation data
test data
ML algorithms
train one or many models
evaluate and select best model
evaluate best model
deploy model for scoring
save and monitor predictions

what is the distribution of the data?

Statisticians often use the word **distribution** to refer to where the data points are and how far apart they are relative to each other

- if the space is 1D, we describe the distribution of points using **univariate measures and plots**, such as $\mathrm{mean}(x)$, $\mathrm{median}(x)$, $\mathrm{mode}(x)$, histograms, box plots, etc.
- if the space is > 1D, we describe the distribution of points
  - using **univariate measures** of *each of its dimensions* (variables): such as $\mathrm{mean}(x)$, $\mathrm{std}(y)$, histogram or box plot of each numeric variable, bar plots or one-way (contingency) tables for categorical variables etc.
  - using **bivariate measures** of any two of its dimensions: such as $\mathrm{corr}(x, y)$, scatter plot of any two variable, two-way (contingency) tables, etc.

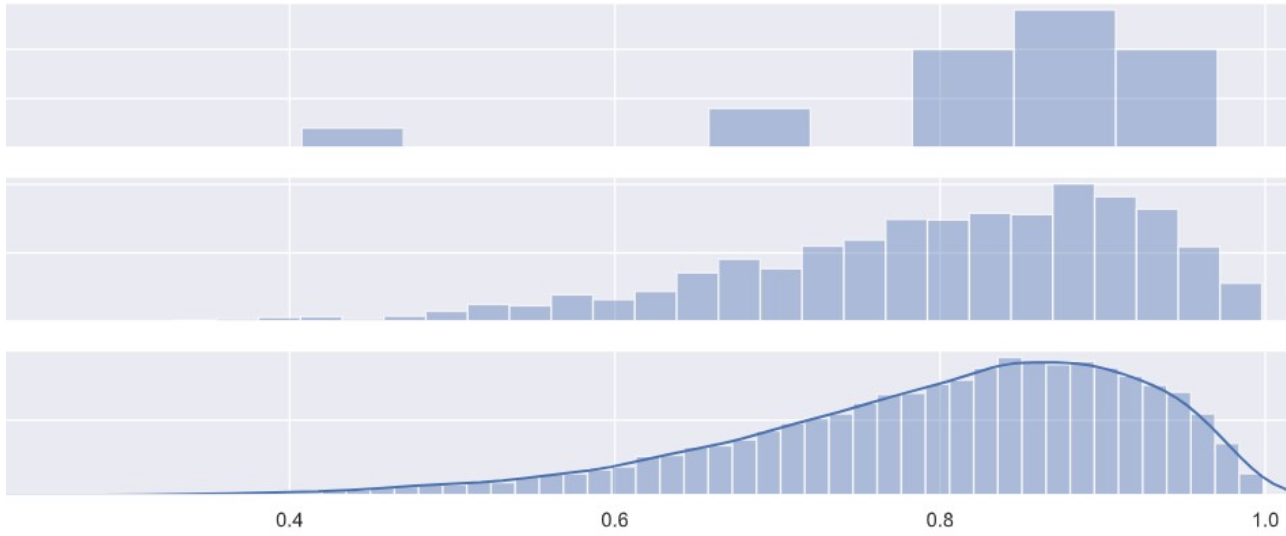
visualize a distribution

- **location**: mean, median, percentiles, box-plot, histogram, etc.
- **spread**: standard deviation, mean absolute deviation, IQR, range, box-plot, histogram, etc.
- **symmetry**: skewness (unitless), box-plot, histogram
- **centrality**: kurtosis (unitless), histogram

A histogram shows how many data points fall within equally-sized intervals (called **bins**), and all of the above properties show through in the shape of the histogram, assuming a large enough sample size, so a histogram is a great way to visualize a distribution

histograms

- the more data points we have, the more the histogram will look like the true distribution
- top to bottom: 20, 2,000, 20,000 samples drawn from a left-skewed distribution
- the blue line can be thought of as the "true distribution" the histogram is converging to



What Is a Histogram?

- A histogram is a bar graph that shows the distribution of data.
- A histogram is a bar graph that represents a frequency table.
- The horizontal axis represents the intervals.
- The vertical axis represents the frequency.
- The bars in a histogram have the same width and are drawn next to each other with no gaps.

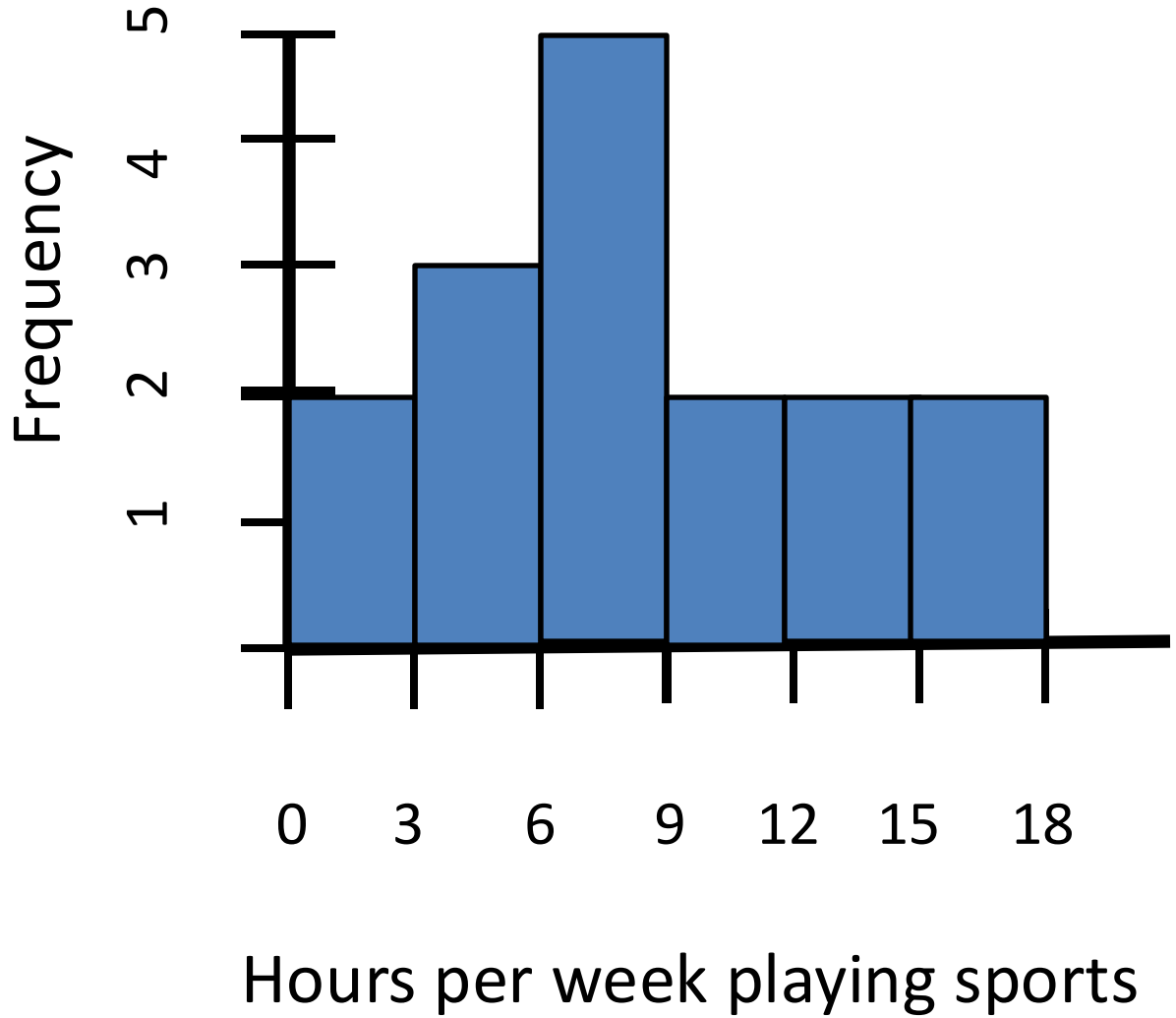
Constructing a Histogram

- ✓Step 1 - Count number of data points
- ✓Step 2 - Compute the range
- ✓Step 3 - Determine number of intervals (5-12)
- ✓Step 4 - Compute interval width
- ✓Step 5 - Determine interval starting & ending points
- ✓Step 6 - Summarize data on a frequency table
- ✓Step 7 - Graph the data

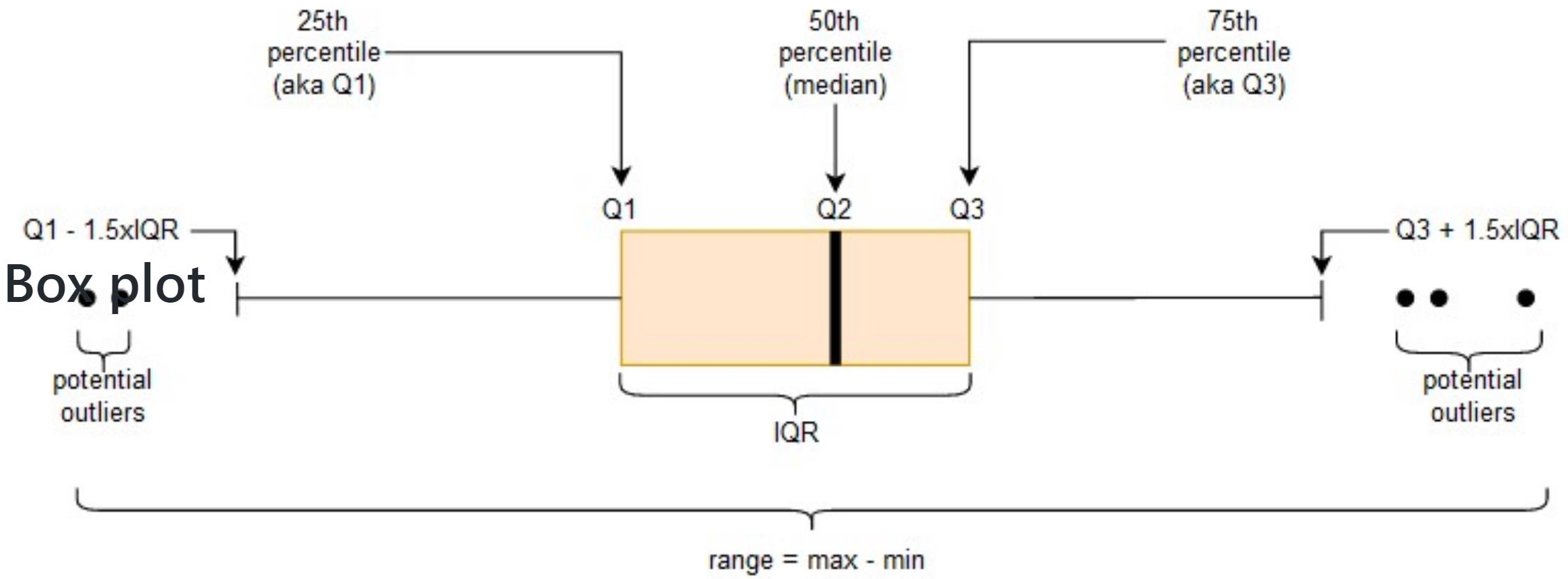
The data below shows the number of hours per week spent playing sports by a group of students.

2	7	17	9	6	13	8	4
5	12	3	11	1	8	15	6

1) **What is the minimum, maximum, & range?**
2) **Make a frequency tables using intervals you decide on.**
3) **Draw a histogram.**


Frequency
1
2
3
4
5
0
3
6
9
12
15
18
Hours per week playing sports

Box Plot


25th
percentile
(aka Q1)
50th
percentile
(median)
75th
percentile
(aka Q3)
Q1
Q2
Q3
Q1 - 1.5xIQR
Q3 + 1.5xIQR
Box plot
potential
outliers
IQR
potential
outliers
range = max - min

Finding First and Third Quartile

The **first quartile** lies one quarter of the way up the list. (25% of the data is below the first quartile.)

The **third quartile** lies three quarters of the way up the list. (75% of the data is below the third quartile.)

1.) Find the median to divide the data is half.
2.) Find the "median" of the lower half, this point will be the first quartile.
3.) Find the "median" of the upper half, this point will be the third quartile.

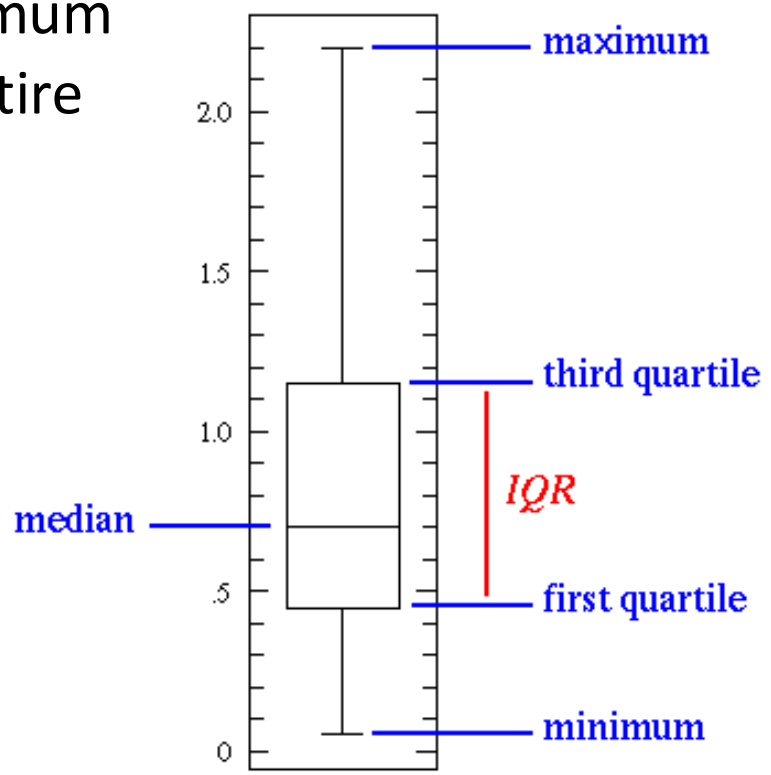
***In the case of two data points being either the first or third quartiles, then use same method as the median (add together and divide by 2)

Measures of Spread

How much do values typically vary from the center?

Range - is the difference of the maximum and minimum value - spread of the entire data set

Interquartile Range (IQR) - is the difference of the upper quartile (Q3) & the lower quartile (Q1) – spread of the middle 50% of the data



Find the minimum, maximum, median (Q2), lower quartile (Q1), and upper quartile (Q3) for the following sets of data and draw the Box and Whisker plot (or Boxplot)

1) 32, 40, 35, 29, 14, 32

2) 6, 1, 7, 6, 5, 5, 0, 1, 0, 8, 4

3) 121, 143, 98, 144, 165, 118

Checking for Outliers

An **outlier** is data point that is extremely far away from the rest of the data and may effect some of the measurements we take from that data.

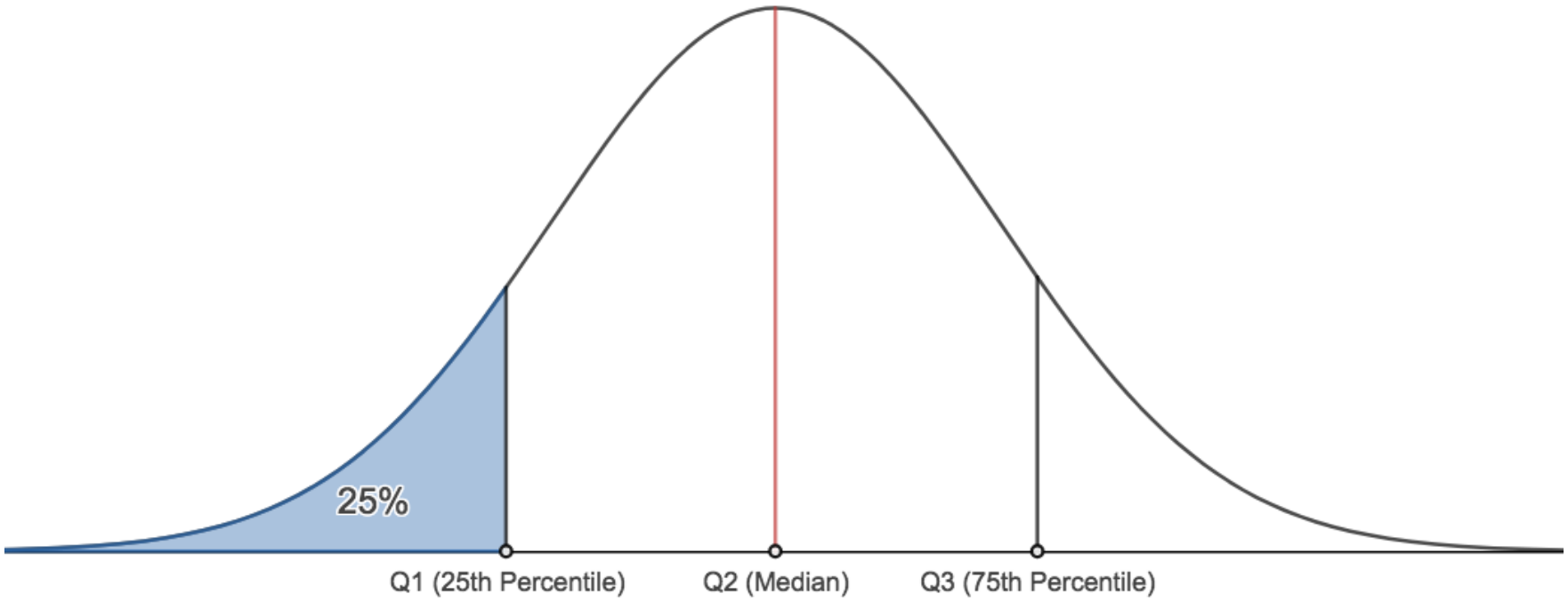
An **outlier** is any point that is farther than 1.5 x the IQR from the first or third quartile.

Now, check the previous data for outliers.

percentiles

- let p be a number between 0 and 1 (or between 0 and 100 if expressed as a percentage), so $p = 0.25$ or $p = 25\%$
- the p-th percentile data points $x_1, \ldots, x_n$ is some number $\tilde{x}_p$ so that p percent of data points are less than this number, i.e. $1 - p$ percentage of the data points are greater than this number
- to get percentiles we need to **sort the data** in ascending order first
- sorting the data is not always practical for large datasets, and **approximate methods** are available

Percentiles

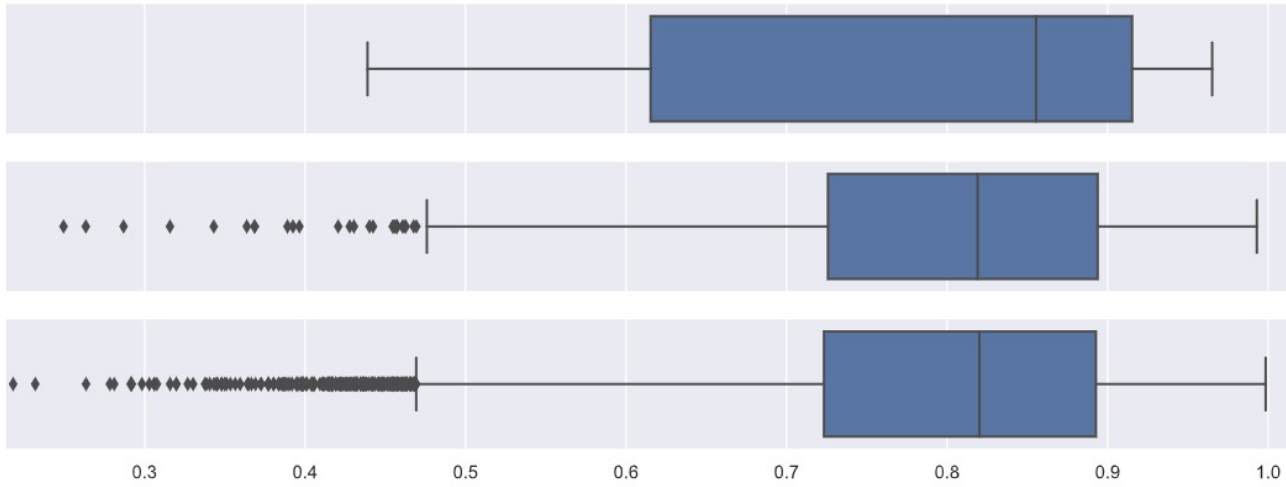


percentiles and ranks

- you can think of percentiles as **normalized ranks**
- the smallest x_i has rank 0, and it is the 0th percentile
- the largest x_i has rank n, and it is the 100th percentile
- the median has rank approx. $n/2$, and it is the 50th percentile
- the 25th and 75th percentiles are also known as $Q1$ and $Q3$ **quartiles**
- the 5th and 95th percentiles, or 1st and 99th percentiles are sometimes considered to be good cut-offs for **outliers** (but this is usually subjective and depends on the data; in fact finding outliers may become a much more sophisticated subject)

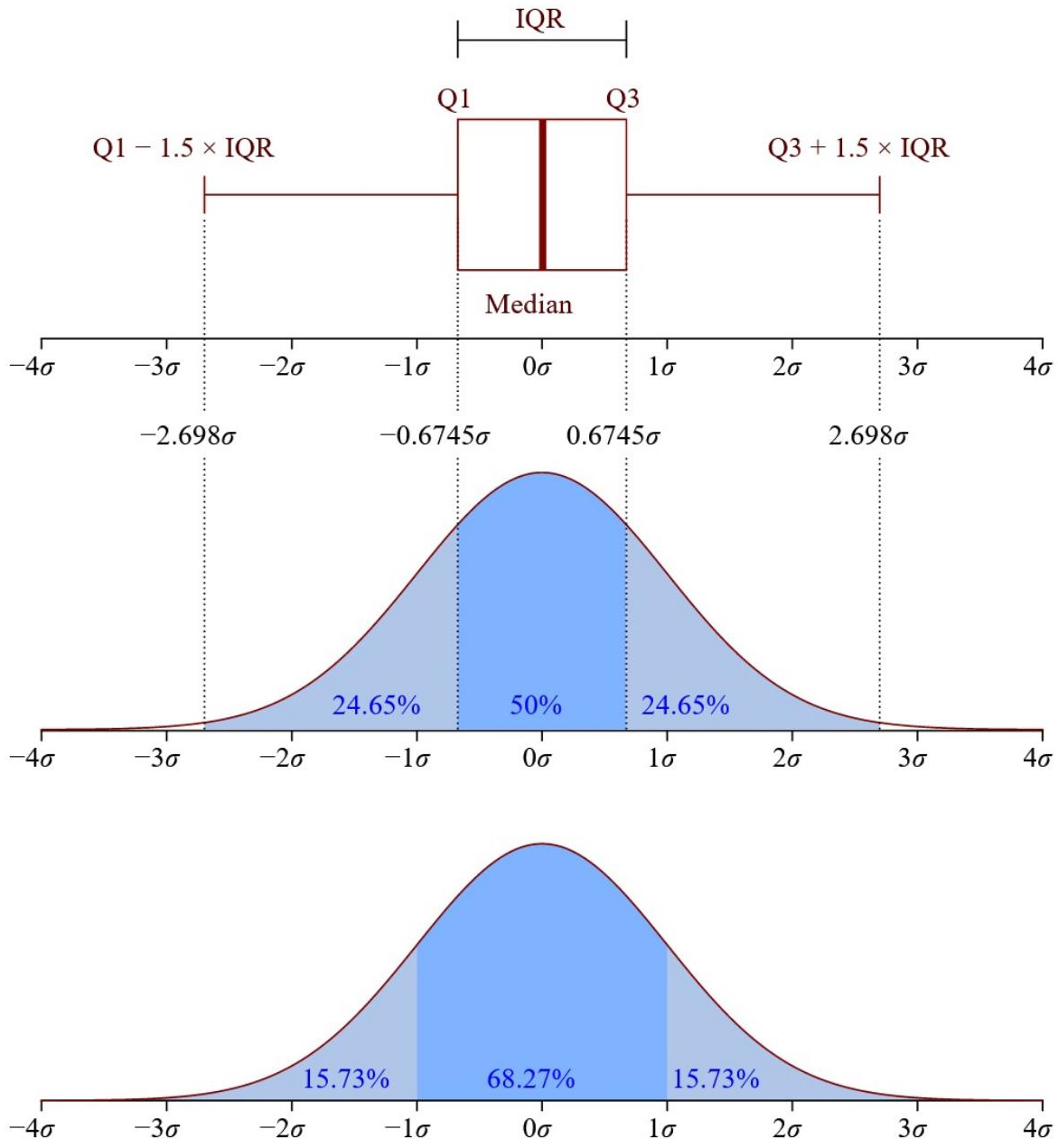
box plots

- box plots don't convey as much information as histograms, but they come close
- top to bottom: 20, 2,000, 20,000 samples drawn from a left-skewed distribution
- box plots are more helpful for detecting outliers or comparing groups



box plots vs density plots

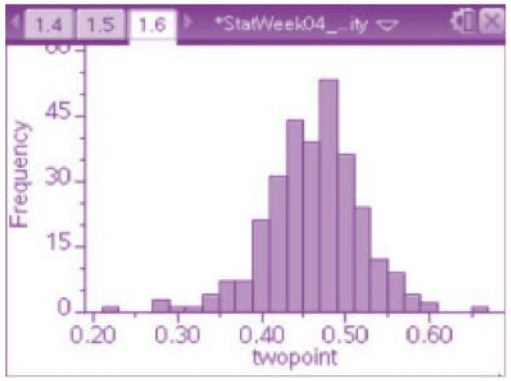
- here we're looking at the **standard normal distribution** (with mean 0 and standard deviation 1)
- the top distribution lines up with the box plot
- the bottom distribution lines up with moving one standard deviation away from the mean
- image source: wikipedia.org



Shape

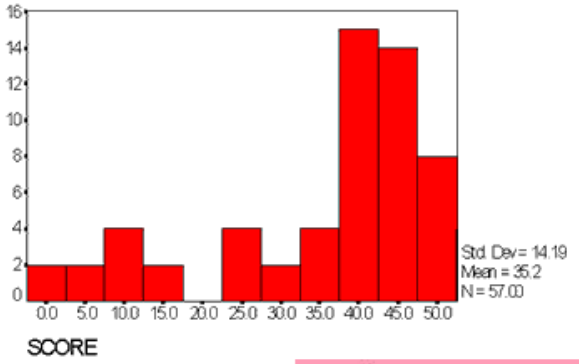
- "**Symmetrical/Normal**"

(mound shaped)



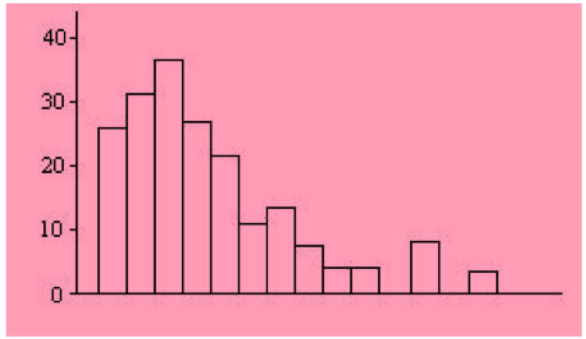
- "**Skewed Left**"

(extreme low values)

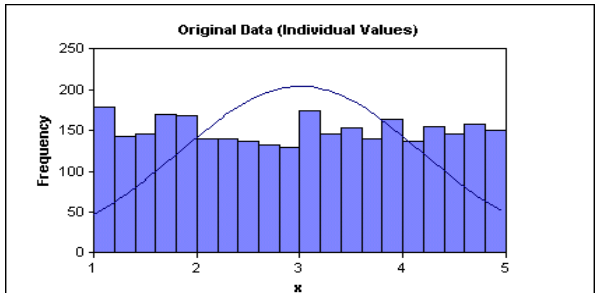


- "**Skewed Right**"

(extreme high values)

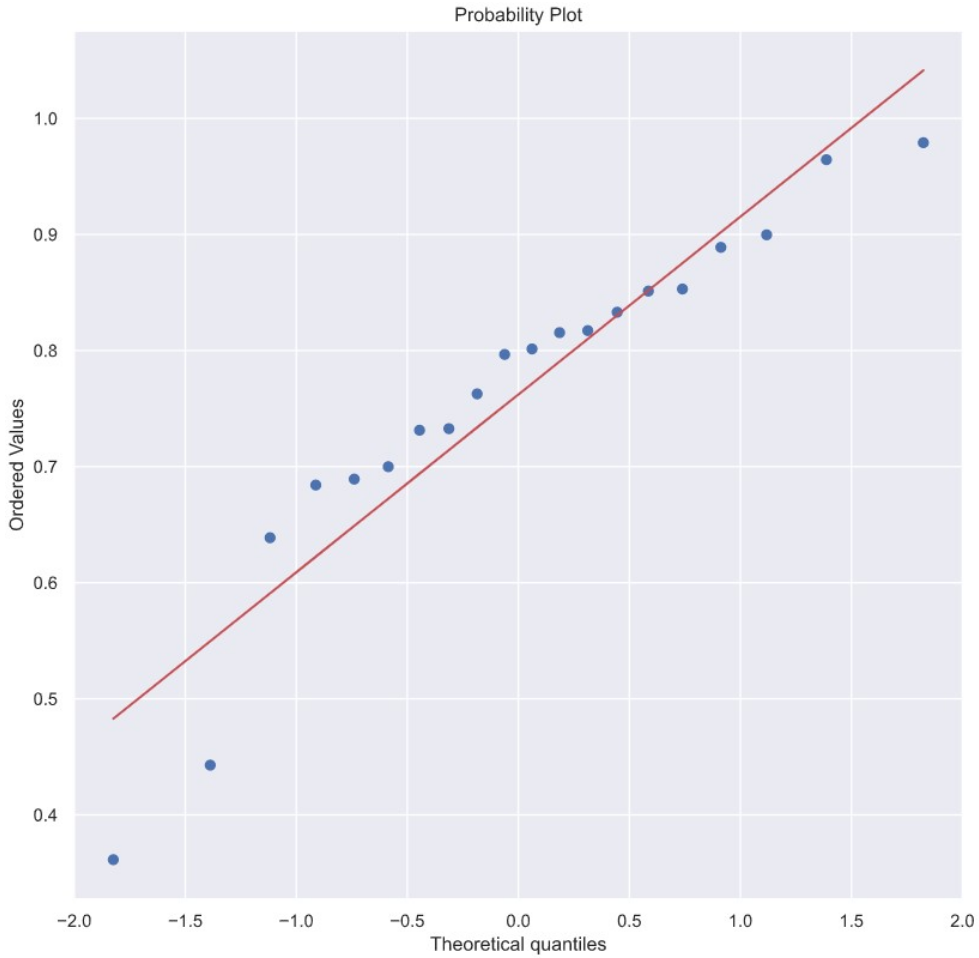


- "**Uniform**"



qq plot

- compares observed percentiles to theoretical percentiles (assuming a specific distribution)
- if observed and theoretical values are very close (follow 45 degree line), then the distributions are very close
- in this plot, we use the **normal distribution** for the theoretical distribution



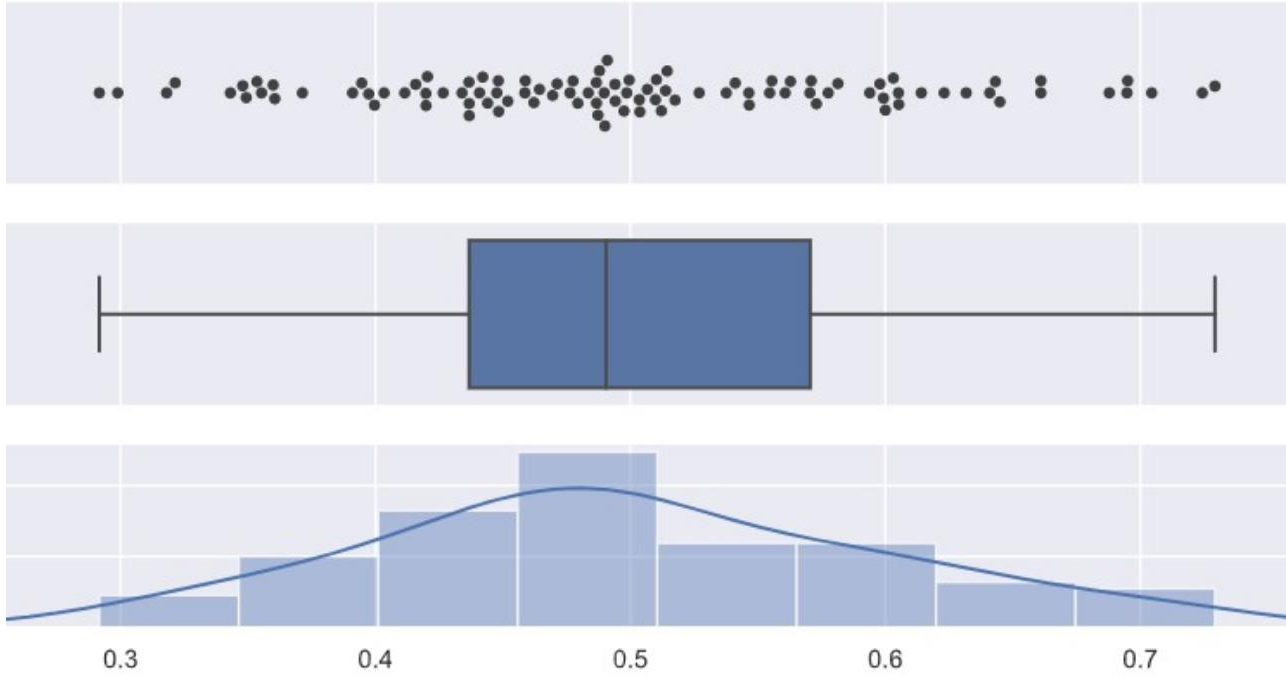
example: bell-shaped distribution

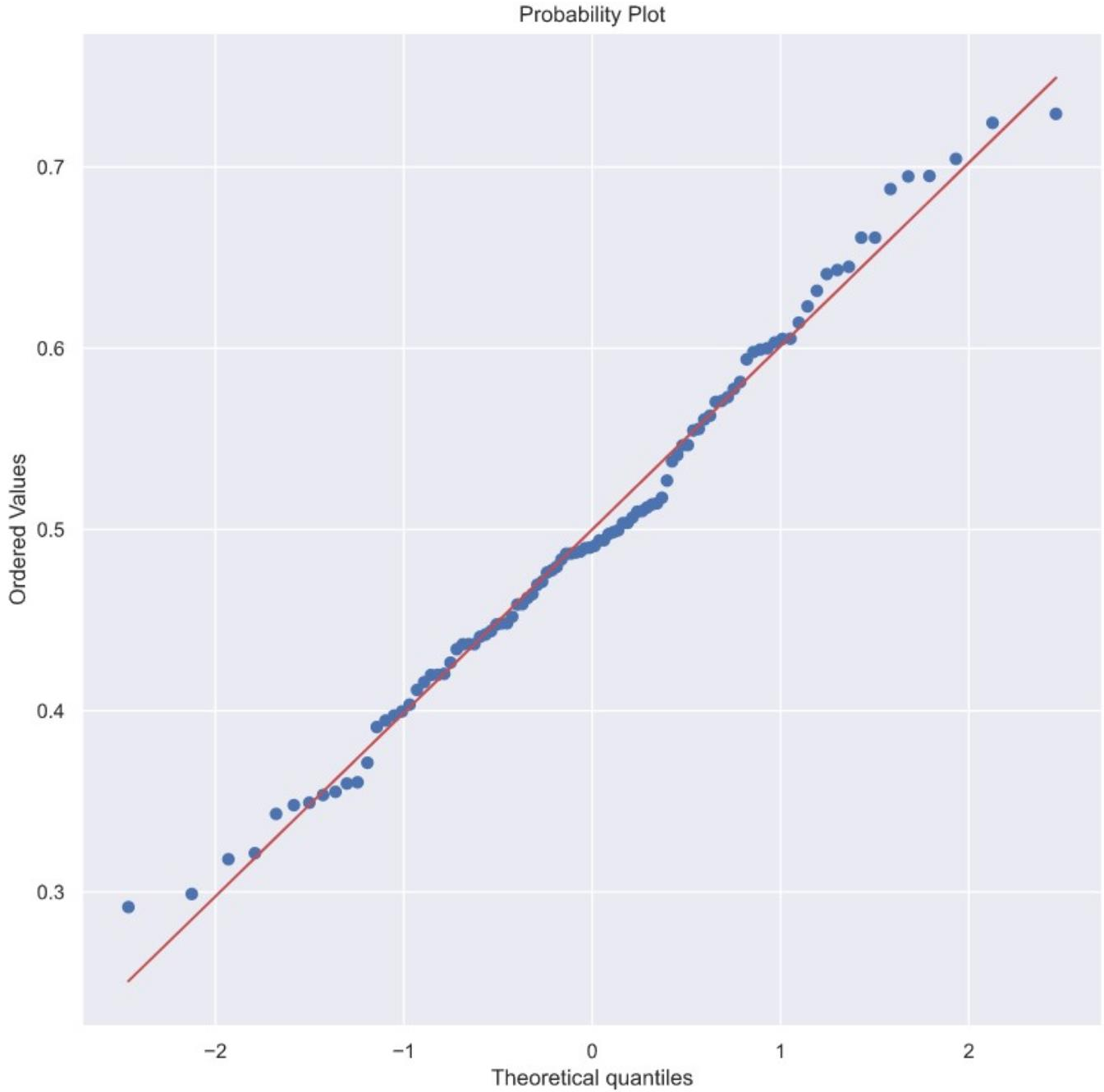
mean = 0.48

standard deviation = 0.13

skewness = 0.02

kurtosis = -0.45




Probability Plot
Ordered Values
0.7
0.6
0.5
0.4
0.3
−2
−1
0
1
2
Theoretical quantiles

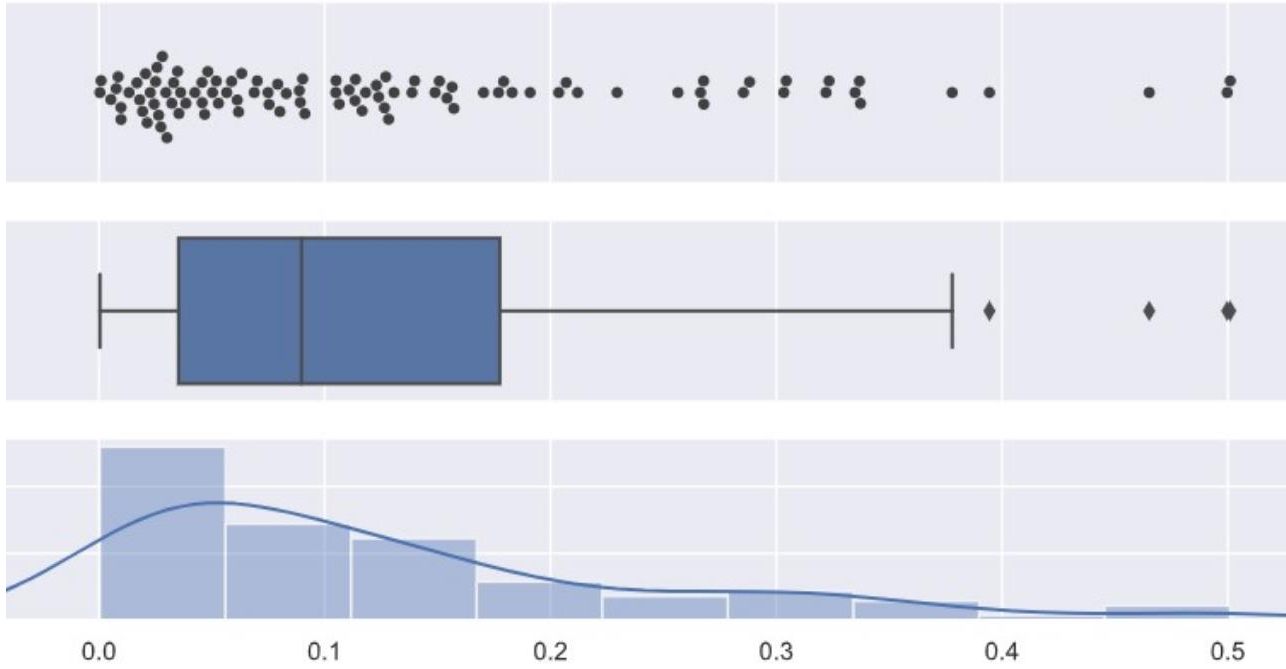
example: right-skewed distribution

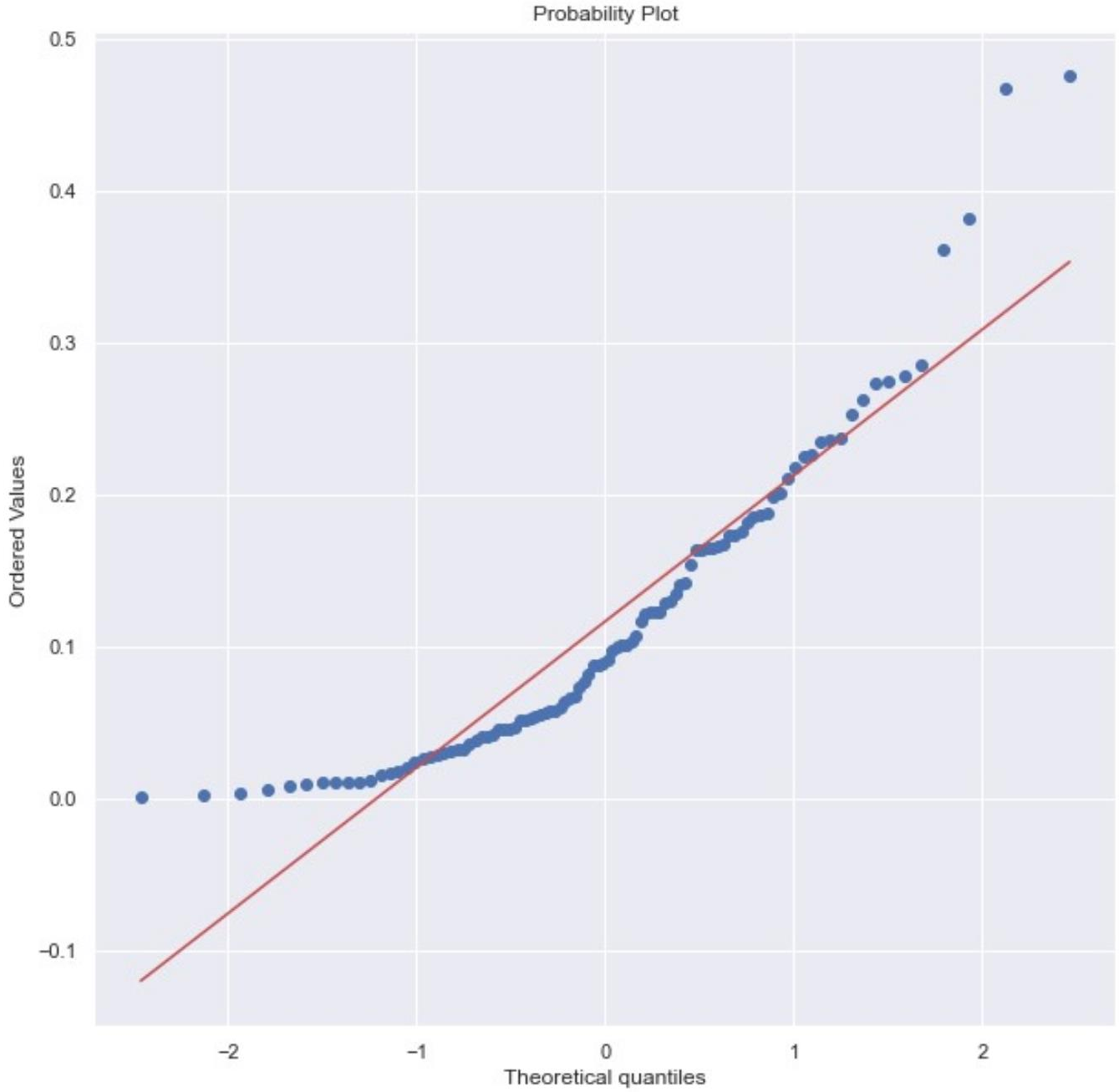
mean = 0.11

standard deviation = 0.11

skewness = 1.36

kurtosis = 1.19




Probability Plot
Ordered Values
Theoretical quantiles
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−2
−1
0
1
2

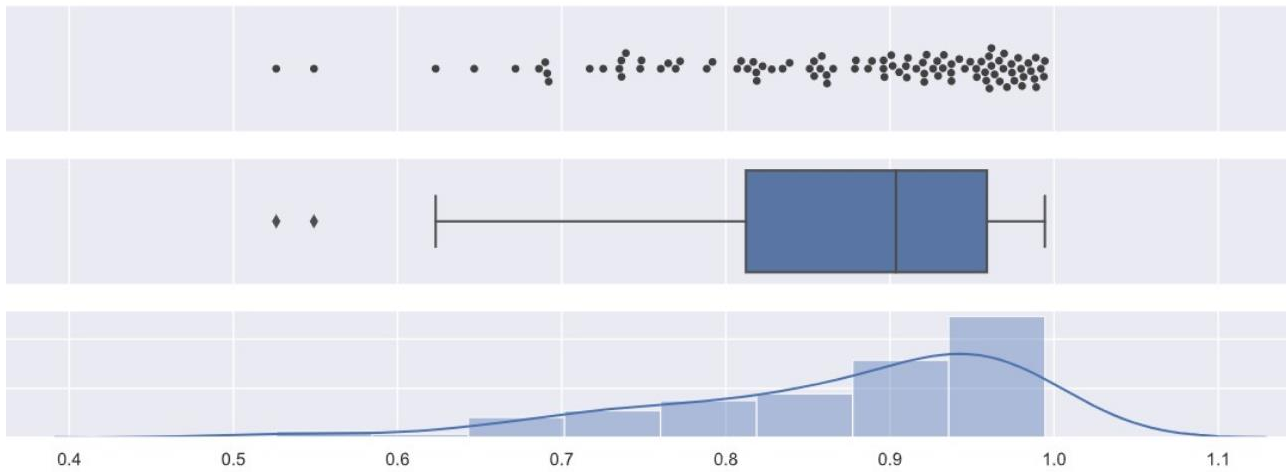
example: left-skewed distribution

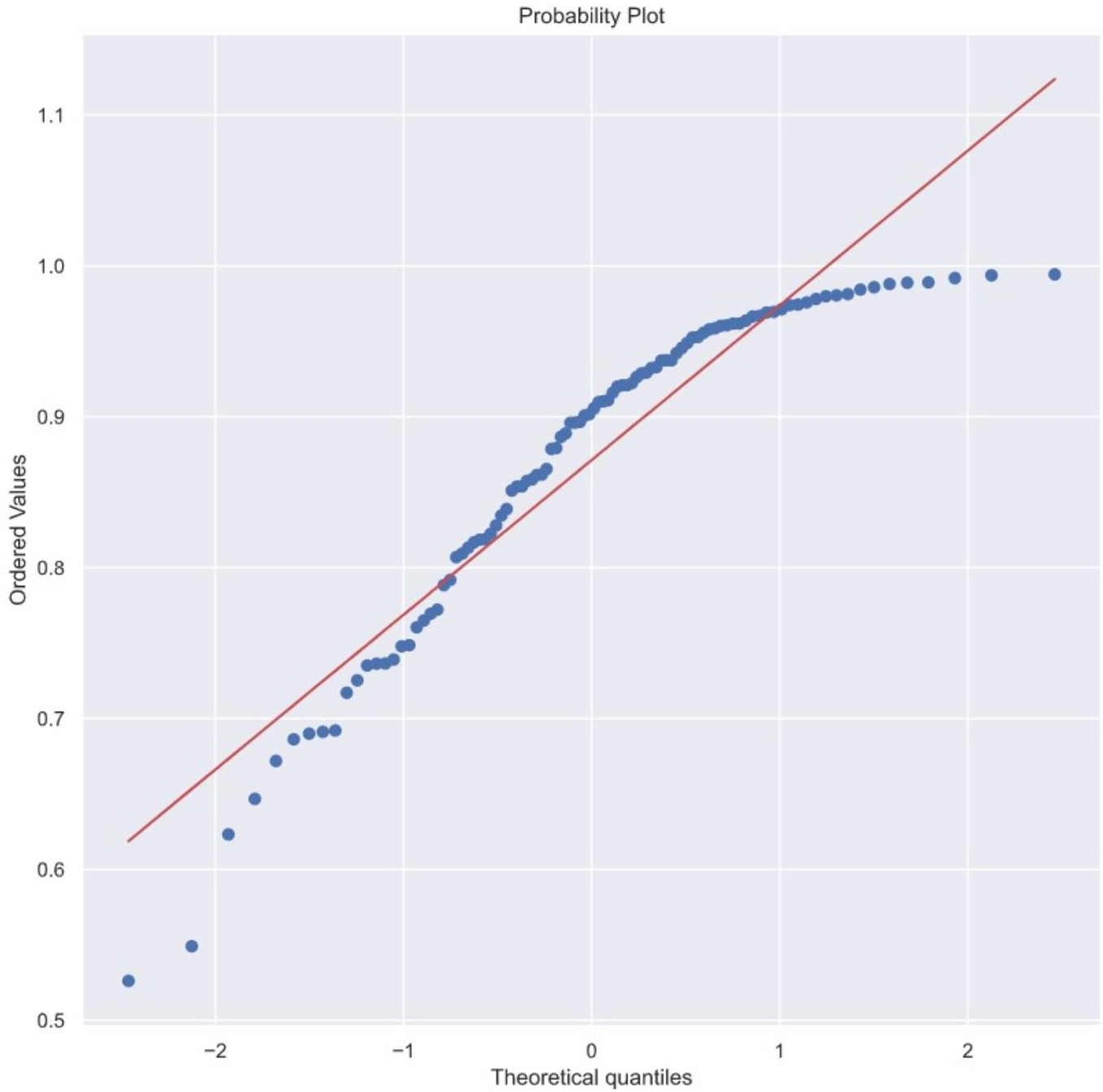
mean = 0.88

standard deviation = 0.11

skewness = -1.34

kurtosis = 1.87




Probability Plot
Ordered Values
Theoretical quantiles
1.1
1.0
0.9
0.8
0.7
0.6
0.5
−2
−1
0
1
2

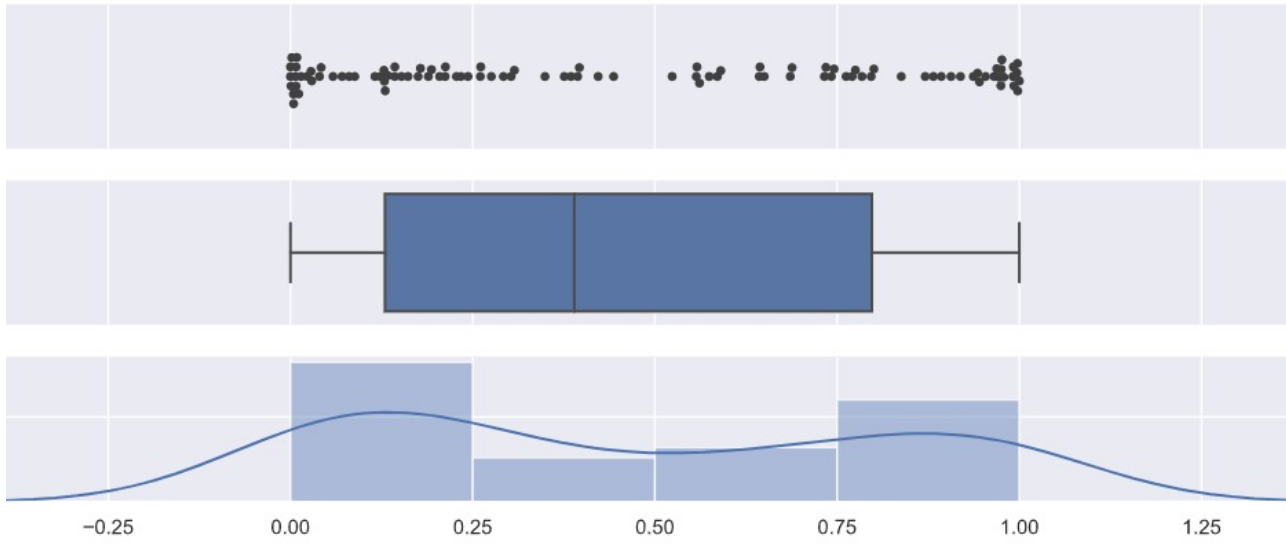
example: bimodal distribution

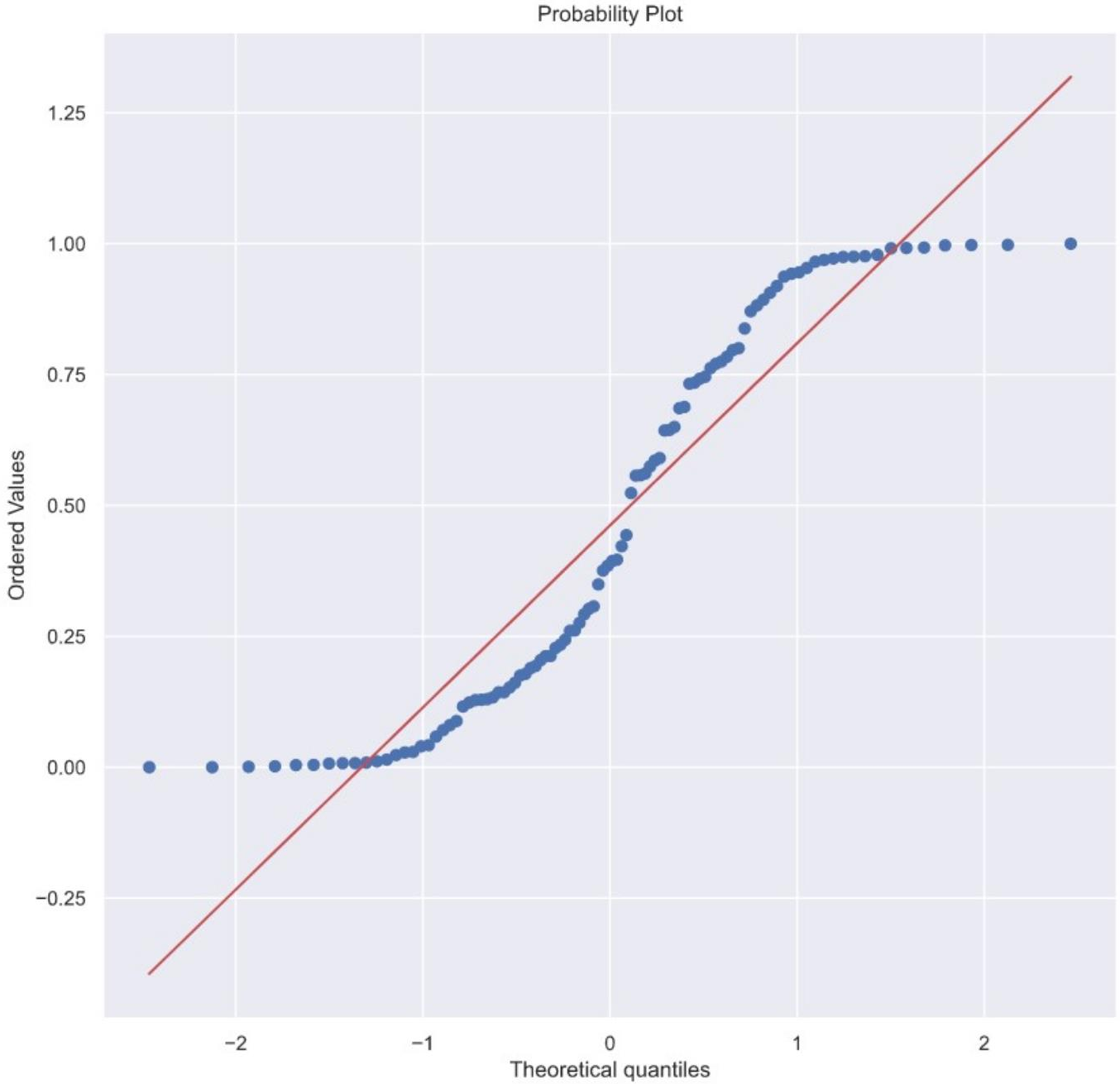
mean = 0.46

standard deviation = 0.36

skewness = 0.17

kurtosis = -1.45




Probability Plot
Ordered Values
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−2
−1
0
1
2
Theoretical quantiles

mean and standard deviation

- **central tendency** refers to the location where data points tend to gather
  the **mean** or simple average is the most common measure of central tendency

$$\text{mean}(x) := \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{mean}(2, 4, 5, 9, 200) = 44.0$$

- **spread** refers to how far apart data points are from each other on average
  the **standard deviation** is the most common example of measures of spread

$$\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{std}(2, 4, 5, 9, 200) \approx 78.03$$

  standard deviation is the square root of variance

mean vs median

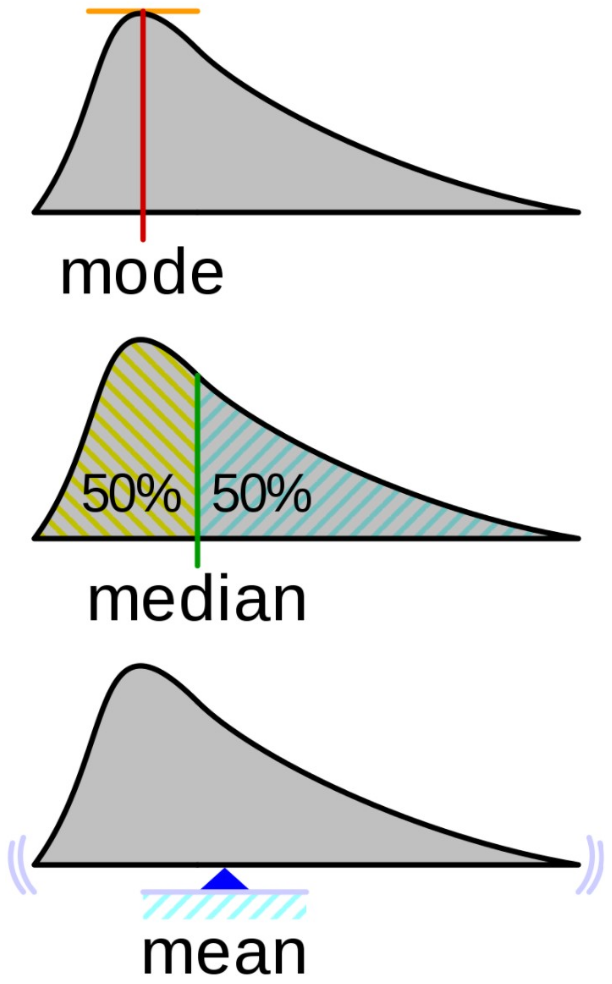
- the **median** is the "middle value" if data is sorted:

$$\text{median}(2, 4, 5, 9, 200) = 5 \qquad \text{mean}(2, 4, 5, 200) = 4.5$$

- notice how the mean is affected by outliers, but the median is not
- we say the median is **robust to outliers**
- there are other ways to achieve robustness against outliers, such as using a **trimmed mean**
- generally robust methods refer to methods that are less affected by outliers

mean vs median vs mode

- here we're looking at the **standard normal distribution** (with mean 0 and standard deviation 1)
- image source: wikipedia.org



notebook time

we return to the lecture later

the end