Descriptive statistics

Chrysafis Vogiatzis

Lecture 14

Learning objectives

After these lectures, we will be able to:

- Differentiate between populations and samples.
- Given a sample, calculate the sample mean, variance, range, and quartiles.
- Use graphical devices to present data, and more specifically:
  - histograms;
  - box plots;
  - scatter plots;
  - time series plots;
  - Q-Q plots.

Motivation: Summarizing information

We live in the era of big data. The size of the data we collect is doubling every 2 years (and this is a conservative estimate). When confronted with so much information, one way to make sense of it is to distill it in smaller, more manageable chunks. This is what we will be doing in this lecture.

Probabilities and statistics

In Lecture 3, we defined **probability** using the words "with every event, we associate a real number called probability to represent the likelihood of that event happening." We may use probability theory to help us address questions such as:

- How likely is it that we get a 6 and a 1 if we roll two dice?
- What is the probability that a patient survives a disease?

On the other hand, we define **statistics** as the field including all methods involved with collecting, describing, analyzing, interpreting data. We use statistics to answer questions such as:

- Are two dice fair?

- What is a good estimate for the mortality rate of a disease?

We show visually an example of the first question. Say we rolled dice multiple times and reported the average number we obtained for each series of rolls. We have two dice: a green and a blue one and their results are show in Figures 2 and 1.

Figure 1: The average number obtained by rolling the blue dice.

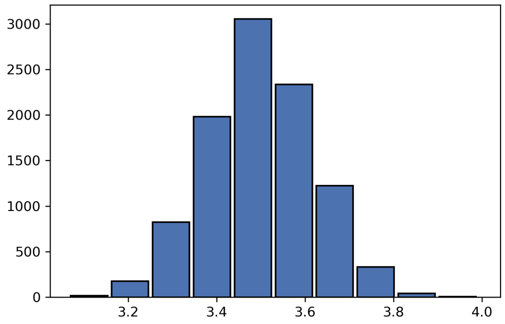
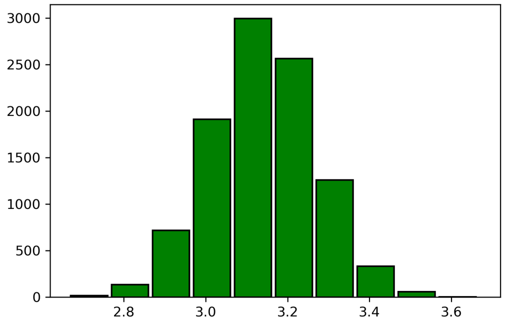


Figure 2: The average number obtained by rolling the green dice.



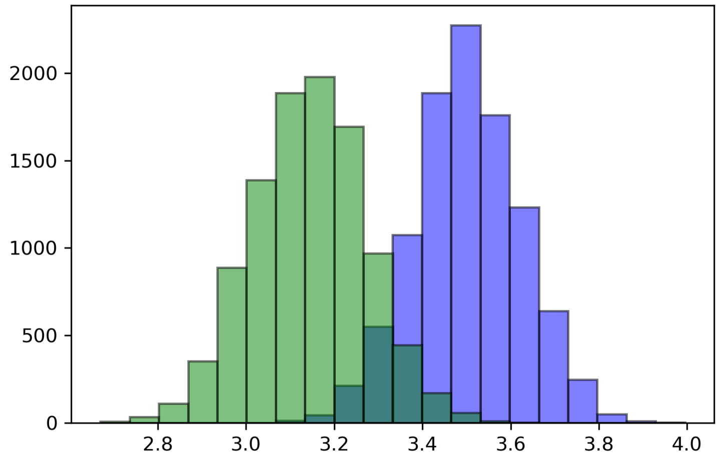
We want to make a decision about whether the two dice are fair. Let us plot both numbers together (see Figure 3). This makes it easy to compare them and deduce that the two dice do not look very similar. It appears the blue one is fair, with an average peak at 3.5 as expected, but the green one seems to favor smaller numbers, and thus unfair.

Statistical methods

In the remainder of the semester, we will be dealing with statistical methods. We differentiate methods in three very important categories:

1. **Descriptive statistics**: methods to *describe* and *present* data.

Figure 3: Plotting both dice average numbers at the same time to make comparisons easier.



2. **Inferential statistics**: methods to use observations in a smaller **sample** to *draw conclusions* for the larger **population**.

3. **Model building**: methods to build models to *predict* future data based on past observations.

The airline industry

The airline industry uses all three categories of statistical methods to help them guide decision-making based on available data. More specifically, example questions they use statistical methods include:

1. **Descriptive statistics**. Present the average delay for each route: this could be used to identify routes that are on average very late to depart from their origin or to arrive at their destination.

2. **Inferential statistics**. Select a subset of the routes to perform some prescriptive action: if the routes do indeed decrease their delays, can we claim that the action will work for all routes?

3. **Model building**. Build a model to predict delays: this is useful for identifying routes that are prone to be delayed and reroute passengers with connections that would be missed.

All three will be seen in subsequent lectures. However, for now, we will focus on **descriptive statistics** alone.

Descriptive statistics

This is the main part of today's lecture. We will specifically see two types of descriptive statistics: numerical and graphical.

What we will focus on in this lecture and in the worksheet is **descriptive statistics**. More specifically:

1. Numerical summaries of data.
   - sample mean, mode, median.
   - sample variance, standard deviation.
   - percentiles, quartiles, ranges.
2. Graphical displays of data.
   - Dot diagrams.
   - Histograms.
   - Stem-and-leaf diagrams.
   - Box plots.
   - Scatter diagrams.
   - Time series plots.
   - Q-Q plots.

Populations and samples

With the term **population** we refer to all possible observations we can collect. For example, a population could be the list of heights of every person in the world; or the SAT scores of every student in Illinois; or the time delays in all flights of a specific airline. The number of observations can grow to be very, very big and impractical to work with.

With the term **random sample** we refer to a subset of the observations selected from a population. For example, a sample could be the list of heights of 12 randomly selected people from our class; or the SAT scores of every student from a specific high school in Illinois; or the time delays in flights leaving ORD of a specific airline. This number of observations in the sample is expected to be significantly smaller than the population size, and hence, manageable to work with.

Formal definitions

More formally, assume a population X where each of its element is distributed with the same distribution (assume mean μ and variance σ^2). Then, a random sample is a set of randomly selected elements from X referred to as $X_1, X_2, \ldots, X_n$. Each X_i is independently selected, and comes from the same population X with mean μ and variance σ^2. Hence, we have:

- $E[X_i] = E[X] = \mu$.
- $Var[X_i] = Var[X] = \sigma^2$.

Numerical summaries of data

Sample mode

Definition 1 (Sample mode) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample mode** is the value(s) x_i that appears most times.*

Small example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the sample mode is 60 as it appears twice.

Sample mean

Definition 2 (Sample mean/average) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample mean** or **average** is calculated as*

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Small example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the sample mean is

$$\frac{1}{5}(60 + 67 + 72 + 63 + 60) = 64.4.$$

Sample variance

Definition 3 (Sample variance) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample variance** is calculated as*

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} =$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{\sum\limits_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}.$$

*The sample standard deviation is denoted by $s = \sqrt{s^2}$. Furthermore, $n-1$ is also called the **degrees of freedom** of the sample.*

Small example

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60 with $\overline{x} = 64.4$. Then, the sample variance is

$$\frac{1}{4}\left(4.4^2 + 2.7^2 + 7.8^2 + 1.4^2 + 4.4^2\right) = \frac{108.81}{4} = 27.2025.$$

Population mean and variance

Percentiles and quartiles

Definition 4 (Percentile) *The number below which we can approximately find p% of the data in the sample is called the p-percentile.*

Based on the definition, we may calculate any p-percentile as follows:

1. Sort the data in increasing order.

2. Calculate $k = (n+1) \cdot \frac{p}{100}$.

3. The element at the k-th position in the sorted data is the p percentile.

Note that the calculation of $(n+1)p/100$ may well be fractional (i.e., the number has us search between two values). When this is the case, then we interpolate. [1]

[1] For example, say we calculate $k = 7.4$. Then the percentile is between the 7th and the 8th value. However, due to the .4 decimal part we would interpolate as: $0.6 \cdot x_7 + 0.4 \cdot x_8$.

Bigger example

Assume the heights of 9 people are 62, 64, 67, 58, 70, 61, 67, 65, 64. What is the 30% and the 67% percentile?

The ordered heights are 58, 61, 62, 64, 64, 65, 67, 67, 70.

30% percentile: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 30}{100} = 3$. The 3rd value is 62.

67% percentile: Plugging in the formula $\frac{(n+1)p}{100} = \frac{10 \cdot 67}{100} = 6.7$. The 6th value is 65 and the 7th is 67: interpolating, we get: $0.3 \cdot 65 + 0.7 \cdot 67 = 66.4$.

A special type of percentiles are the quartiles. They separate the data in four parts, each of which contains 25% of the data. Specifically, we have three quartiles, typically denoted as $Q1, Q2, Q3$:

- Q1: Splits the lower 25% from the rest of the data.
- Q2: Splits the lower 50% from the rest of the data.
- Q3: Splits the lower 75% from the rest of the data.

Q2 is also called the **median**.

Definition 5 (Sample median) *Given n observations $x_1, x_2, \ldots, x_n$ in a random sample, the **sample median** is the value below which (and above which) we find 50% of the observations. It is denoted by $\tilde{x}$ or $Q2$ (the second quartile).*

Bigger example

Earlier, we got the ordered 9 heights to be 58, 61, 62, 64, 64, 65, 67, 67, 70.

Q1: $\frac{(n+1)p}{100} = \frac{10 \cdot 25}{100} = 2.5$. So $Q1 = 61.5$.

Q2: $\frac{(n+1)p}{100} = \frac{10 \cdot 50}{100} = 5 \implies Q2 = \tilde{x} = 64$.

Q3: $\frac{(n+1)p}{100} = \frac{10 \cdot 75}{100} = 7.5 \implies Q3 = 67$.

Ranges and outliers

Definition 6 (Range) *The range of values in a sample or population is calculated as the difference of the maximum and the minimum value in the sample or population: $R = \max\{x_i\} - \min\{x_i\}$.*

By definition, the range of a population will always be greater than or equal to the range of a sample.

Definition 7 (Interquartile range) *The interquartile range is calculated as the difference of the third to the first quartile:* $IQR = Q3 - Q1$.

The IQR is in essence a measure of range but focusing on the middle part of the data considered.

Definition 8 (Outliers) *An outlier is a value that affects the range of our data but leaves the IQR unaffected. Specifically, we say that a data point is an outlier if it lies outside* $[Q1 - 1.5IQR, Q3 + 1.5IQR]$.

Describing aluminum-lithium specimens

A company has collected the following data for compressive strength (psi) of aluminum-lithium specimens: 105, 221, 183, 186, 121, 181, 180, 143, 97, 154, 153, 174, 120, 168, 167, 141, 245, 228, 174, 199, 181, 158, 176, 110, 163, 131, 154, 115, 160, 208, 158, 133, 207, 180, 190, 193, 194, 133, 156, 123, 134, 178, 76, 167, 184, 135, 229, 146, 218, 157, 101, 171, 165, 172, 158, 169, 199, 151, 142, 163, 145, 171, 148, 158, 160, 175, 149, 87, 160, 237, 150, 135, 196, 201, 200, 176, 150, 170, 118, 149.

What are the outliers?

We would first have to sort the data in increasing order, and then calculate $Q1, Q3$. Doing so gives us $Q3 = 181$, $Q1 = 144.5$ and $IQR = 36.5$. We may also calculate the minimum and maximum values as 76 and 245, respectively, leading to a range of 169.

Potential outliers would lie outside the range of $[Q1 - 1.5IQR, Q3 + 1.5IQR] = [89.75, 235.75]$. The only values satisfying this are: 245, 76, 87, 237.

Graphical devices of data

In this section, we discuss some visual tools to represent data.

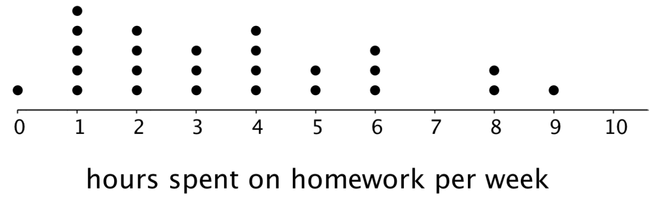
Dot diagrams Dot diagrams (as the name suggests) asks to place a dot on top of each data point. The mode and median are revealed pretty easily in a dot diagram: simply find the tallest set of dots for the mode, and the value below which 50% of the dots lie for the median. See Figure 4 for an example.

In the example, the mode was 1 hour, and the median is at 13 "dots" (for a total of 25 dots)[2] and is found at 3 hours.

It becomes clear from the way this is constructed that the dot diagram is only useful for smaller sized datasets.

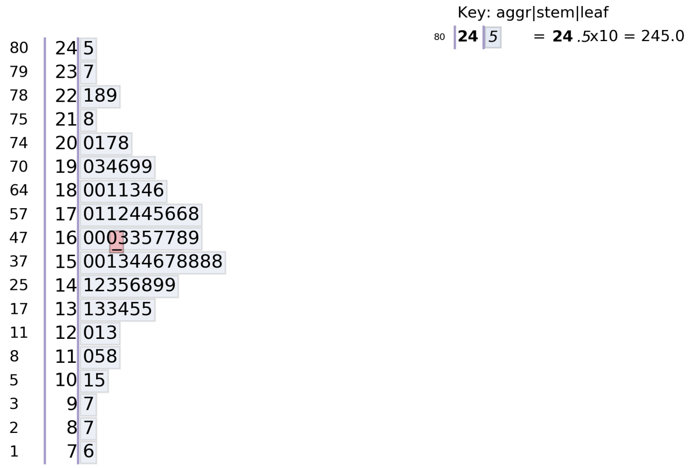
[2] Recall the median calculation is $(n+1)\,50/100 = 13$

Figure 4: An example of a dot diagram representing the amount of time each student spent on Homework assignment 1 (self-reported) during Fall 2019, rounded to the closest integer.



Stem-and-leaf plots A stem-and-leaf diagram only makes sense when all of the data consists of at least two digits. It is a striking visual tool to showcase frequency. The way it is constructed is simple: we pick a series of stems (the first, more important digits) and leaves (the least important digit). For example, the number 311, could be represented as a stem of 31 and a leaf of 1. The leaves are sorted in increasing order. An example is presented in Figure 5.

Figure 5: An example of a stem-and-leaf diagram representing the data from the aluminum-lithium specimens.

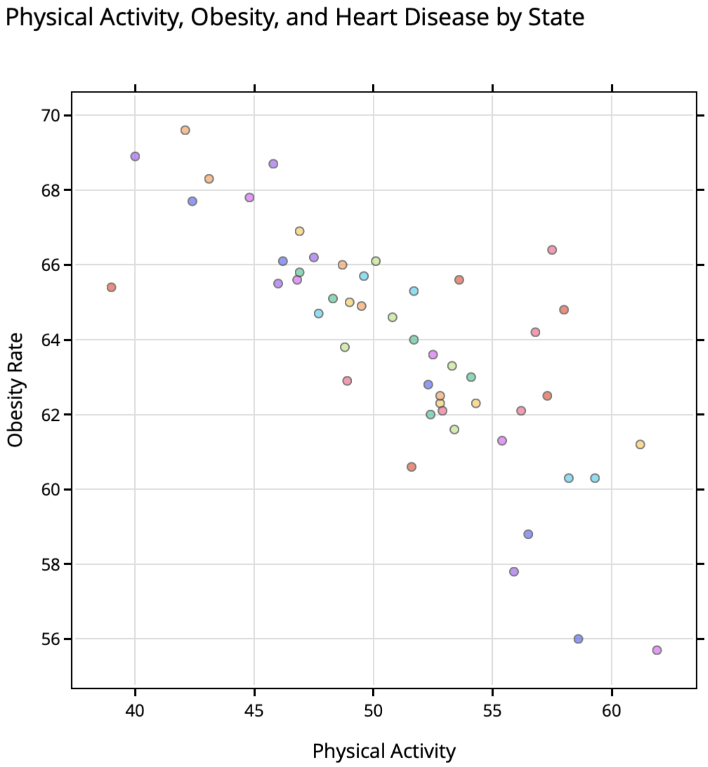


```
                                      Key: aggr|stem|leaf
                                   80 |24|5    = 24.5x10 = 245.0
80 | 24|5
79 | 23|7
78 | 22|189
75 | 21|8
74 | 20|0178
70 | 19|034699
64 | 18|0011346
57 | 17|0112445668
47 | 16|0003357789
37 | 15|001344678888
25 | 14|12356899
17 | 13|133455
11 | 12|013
8  | 11|058
5  | 10|15
3  |  9|7
2  |  8|7
1  |  7|6
```

Alongside the diagram, we typically present frequency (the cumulative number of observations up to and including a stem). In the example, we see that up to and including the stem of 10 we have five observations; on the other hand up to and including the stem of 18 we have 64 observations. This can be used to measure individual frequency: for example the stem 17 has 10 observations – we can tell because up to and including 17 we have 57 observations, whereas up to and including 16 we have 47 observations.

Scatter plots Scatter diagrams are particularly useful when we suspect that the data has some hidden relationship, either positive or negative. For example, what can you say about the following scatter plot of Figure 6 showing data points of activity and obesity in the US?

Figure 6: A scatter plot of the relationship between the rate of obesity cases and the average physical activity levels at each state.



Scatter plots may reveal a positive or negative relationship. They may also show that there seems to be no relationship between two variables. We reveal small, simple examples of each of the three cases in Figure 7.

Time series plots A time series plot is useful when the data are recorded in the order of time. For example, if we are given data that presents some number that changes every month, then it may be suitable to present in a plot where the x axis represents time, and the y axis the number of interest. Below we present two examples: from the city of Chicago for the number of reported crimes per month in Figure 8 (notice the huge drop every February, due to the fact that February has fewer days!) and from Champaign county on the number of COVID-19 cases every week in Figure 9.

Figure 7: Positive relationship (left), negative relationship (right), and no discernible relationship (below).

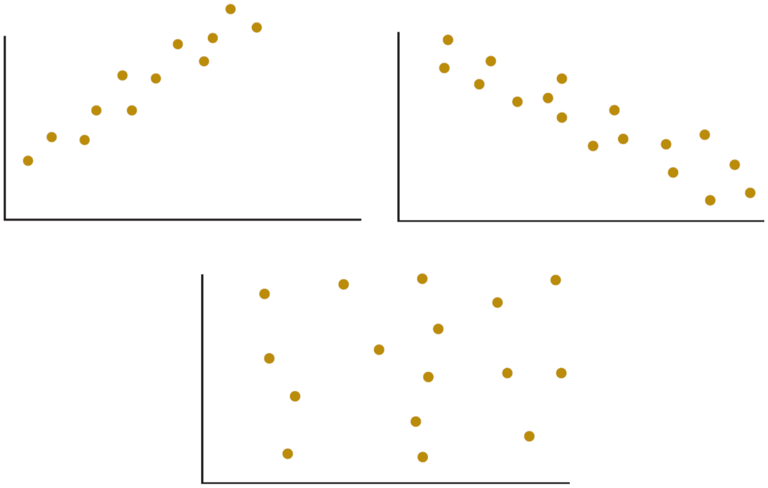


Figure 8: Reported crimes in Chicago by date. Data obtained by `https://data.cityofchicago.org` on October 8, 2019.

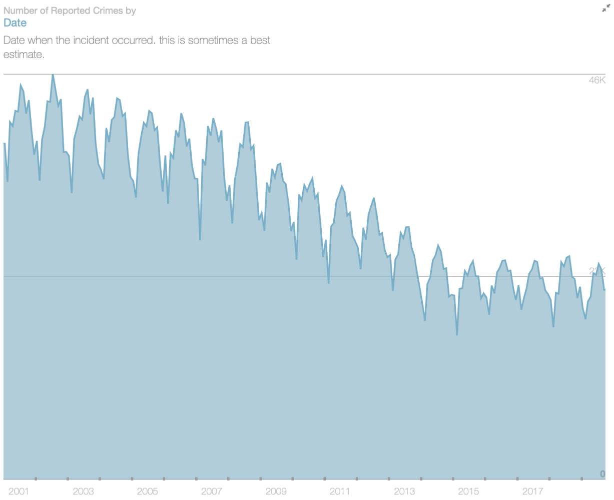
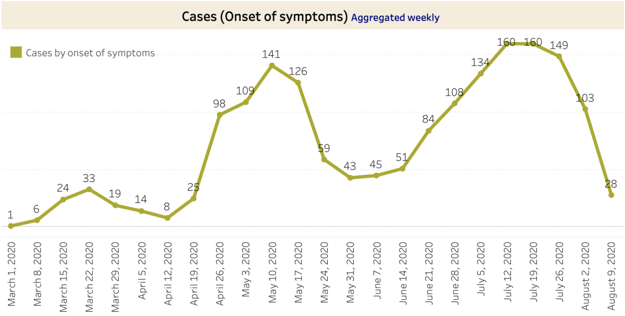


Figure 9: Number of cases (onset of symptoms) per week in Champaign county. Data obtained by `http:c-uphd.org` on August 14, 2020.



Box plots Box plots, sometimes also called box-and-whisker plots, are graphical devices built to reveal multiple interesting properties at once. Seeing a box plot reveals:

1. the center of the data;

2. the spread of the data;

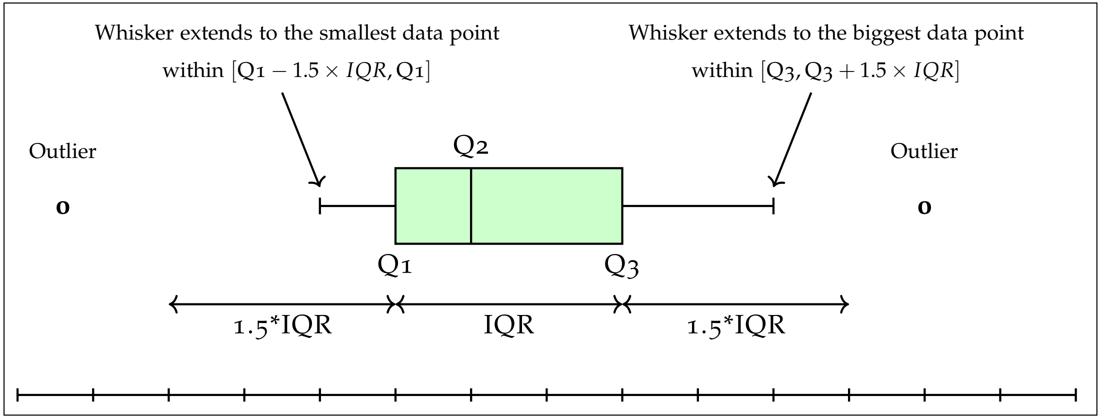
3. the shape of the data;

4. and the outliers in the data.

Seeing a box plot immediately shows the *min* value, the first quartile $Q1$, the median $Q2$, the third quartile $Q3$, the interquartile range IQR, and the *max* value of the data.

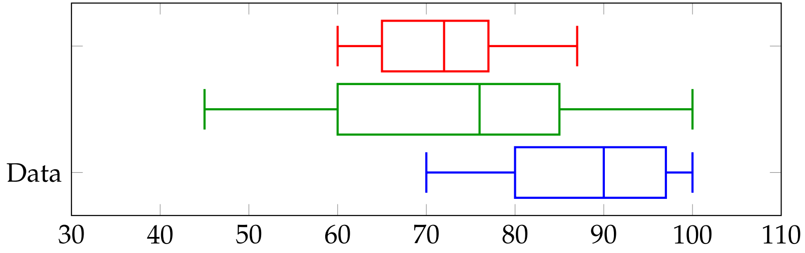
Figure 10 shows all the inner workings of constructing a box plot.

Figure 10: A box plot. To construct it, we create a rectangle ranging from Q1 to Q3. We separate it into two parts drawing a line where the median Q2 is. Then, we extend two whiskers on the two sides all the way to the smallest and biggest value respectively so long as that value is less than $1.5 \times IQR$ away from the quartile. Finally, we note every point outside the whiskers as outliers with a "o".


Whisker extends to the smallest data point
within $[Q1 - 1.5 \times IQR, Q1]$
Whisker extends to the biggest data point
within $[Q3, Q3 + 1.5 \times IQR]$
Outlier
o
Q2
Outlier
o
Q1
Q3
1.5*IQR
IQR
1.5*IQR

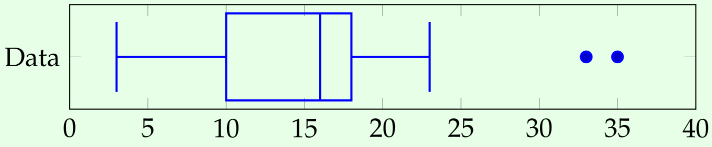
It is useful to compare box plots one next to the other. For example, see the box plot of Figure 11 containing information about the quality obtained in three different plants. We observe that the blue plant provides us with the highest quality index. The green one has better median quality index than the red one, but the red one has a narrower range of possible quality indices, making it more consistent.

Figure 11: The quality index obtained in three different plants.



A small example

We are given a set of data points, and we have calculated that $Q1 = 10, Q2 = 16, Q3 = 18$. The points outside the $[Q1, Q3]$ range are 3, 7, 8, 8, 9 from below and 19, 23, 33, and 35 from above. Draw the boxplot.



Histograms A histogram is a graphical construct that presents data by placing them in *bins*.

The bins could be numbers (in the case of Figure 12, the number of friends on fb.

The bins could represent age ranges (in the case of Figure 13, the age of Florida residents).

The bins could even represent letter grades (as you are probably used to seeing letter grade distributions after exams as in Figure 14!).

Histograms possess three important characteristics:

1. modality.
2. heavy/light tailedness.
3. skewness.

The **modality of a histogram** is concerned with the number of "noticeable peaks" in the data. Recall that a single peak would imply a single mode (most frequent value, or in a histogram's case most frequent range of values). A histogram can then be:

- unimodal (single mode).
- bimodal (two modes).
- multimodal (multiple modes).

Figure 12: The number of friends that a person has on Facebook.

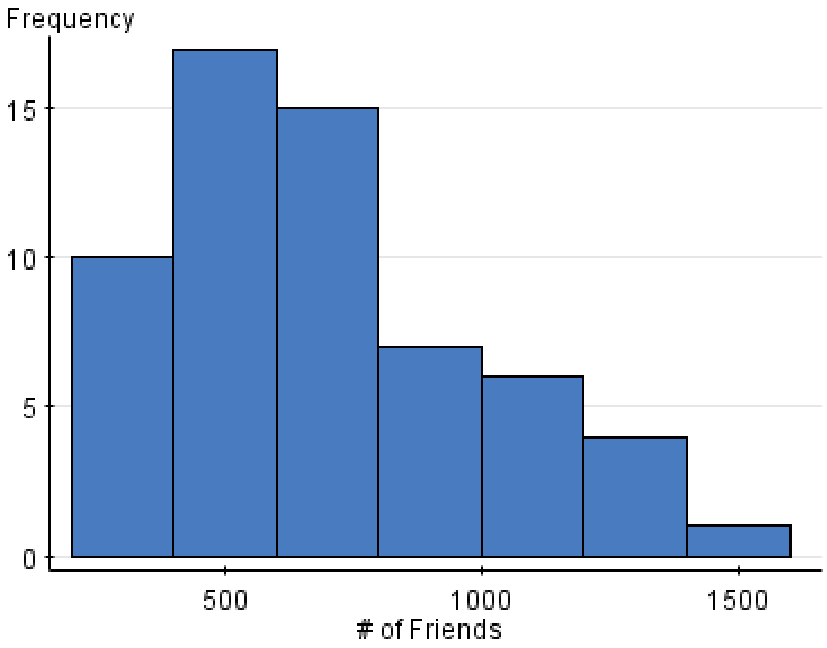


Figure 13: The age of Florida residents in 2018.

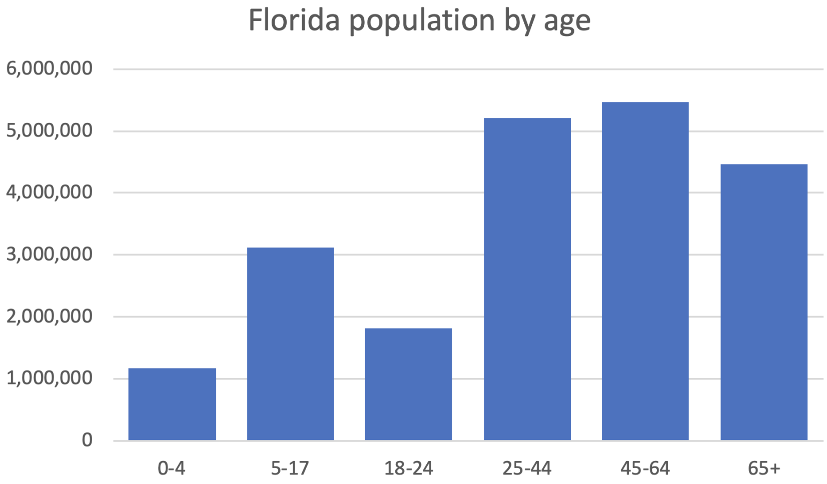
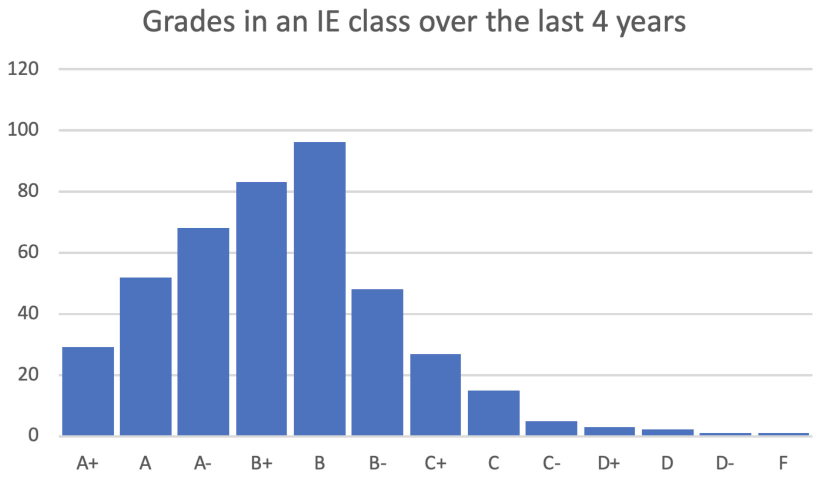


Figure 14: The final grade distribution in an IE course over the last four years.



- uniform (no mode).

We present four examples to showcase each of the four types in Figures 15–18.

Figure 15: **Unimodal**: we note one observable peak at the "B" letter grade.

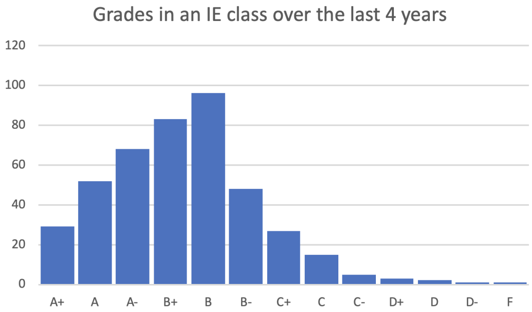
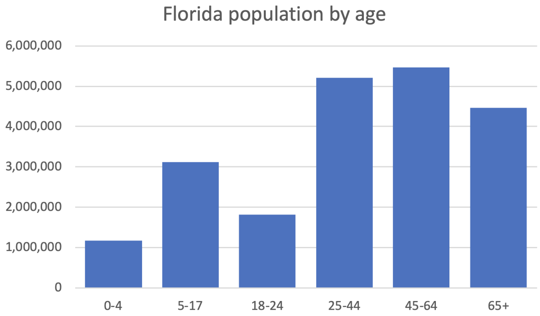


Figure 16: **Bimodal**: we note one observable peak at the "5-17" and the "45-65" age ranges.



A second histogram characteristic is whether it possesses a **heavy or light tail**. We say that a tail is "heavy" if it is "heavier" than the exponential distribution.

Figure 17: **Multimodal**: we can find four noticeable peaks here when observing the height of NBA players (in inches, all heights from the 2013 league).

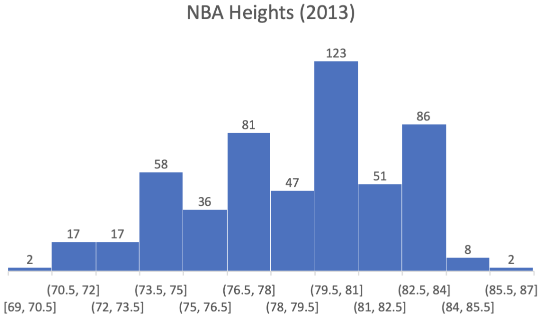


Figure 18: **Uniform**: when we roll multiple dice and report the outcomes.

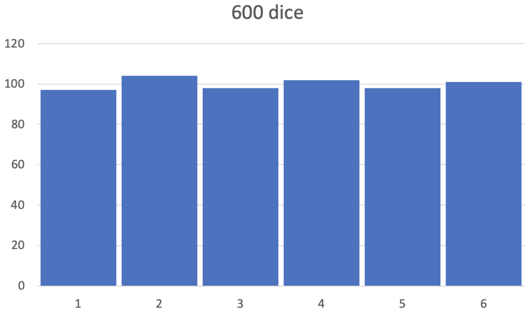
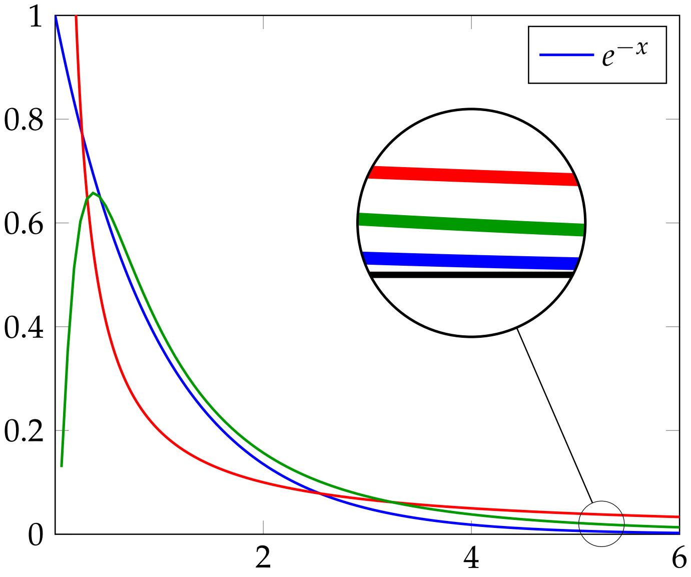
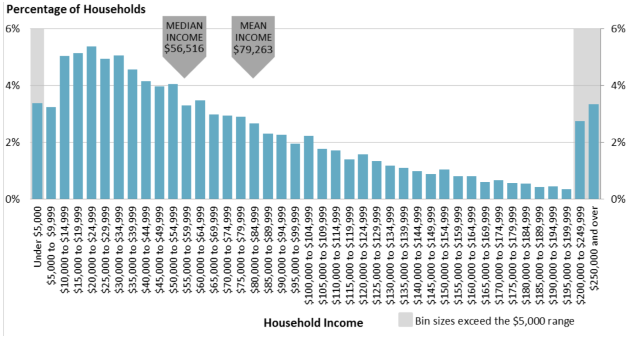


Figure 19: What heavier than an exponential distribution means. Here both the green and the red functions are located higher than the exponential for bigger values of x, so they would both be characterized as heavy-tailed.

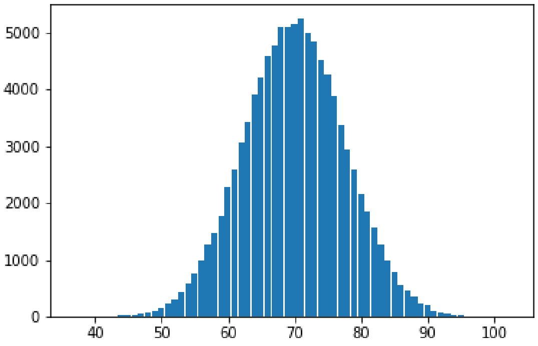


Let's turn our focus back to histograms. Here are two examples of how a heavy-tailed a light-tailed histogram would look like:

a) **Heavy-tailed**: the household income. Note how there is a heavy tail for some very high incomes.



b) **Light-tailed**: heights of a sample of the population in Denmark (restricted to people who self-identify as male).



Why should we care about this characteristic? Well, say we are devising a policy and we need to figure out whether the same policy should apply to all. When the feature we are looking at has a heavy tail, this might have us thinking twice before having the same policy, because many observations would lie far from the average or the bulk of our observations.

Finally, we discuss **skewness**, the third histogram characteristic. In essence, we want to answer the question of whether the histogram is symmetric or not. And, if not, is it right-skewed? Or left-skewed? How do we figure this out?

- If the tail is to the left, then it is left-skewed or has negative skewness.
- If the tail is to the right, then it is right-skewed or has positive skewness.
- Otherwise, it is symmetric.

Figure 20: An example of a left-skewed histogram, as the tail is to the left.

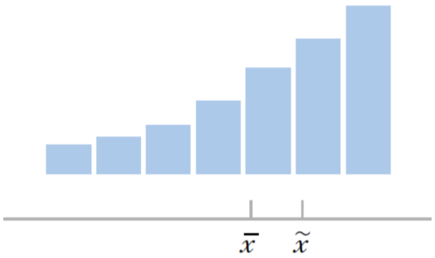


Figure 21: An example of a right-skewed histogram, as the tail is to the right.

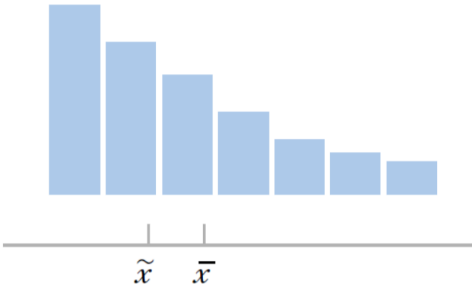
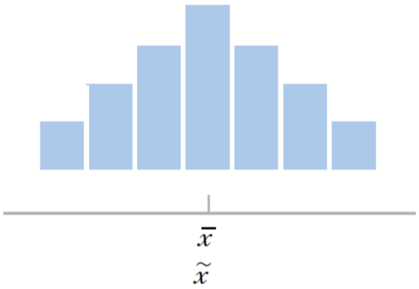


Figure 22: An example of a symmetric histogram with no discernible tail.

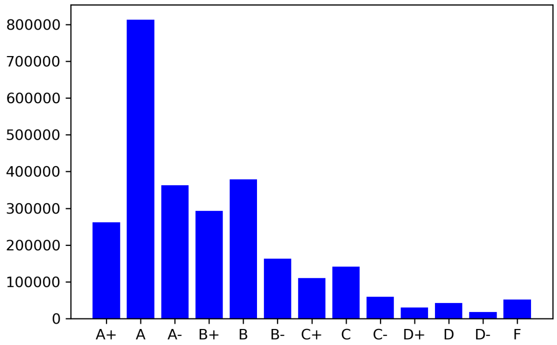


In all figures, $\overline{x}$ represents the average, and $\tilde{x}$ the median. Note that this helps us provide another definition for skewness. If the median is:

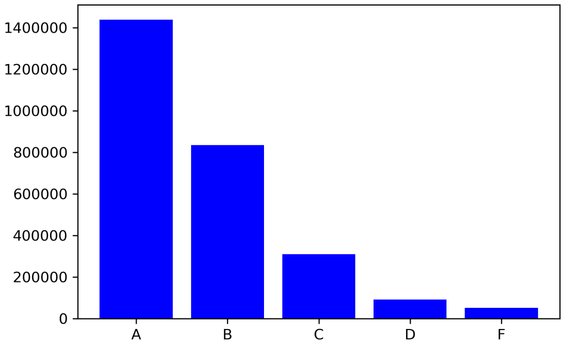
- to the right of the average, then the histogram is left-skewed.
- to the left of the average, then the histogram is right-skewed.
- in a similar location to the average, the the histogram is symmetric.

Finally, we address another important aspect. It is true that the same data can be presented in many, many different ways using histograms:

- Here we show every distinct letter grade that a student may receive:



- Whereas here we show only the main letter grades (for example, no “B+” or “B-”, instead we only have a “B” grade):



The more bins we introduce, the less width each bin has, and the more the shape resembles the actual distribution of the data (for larger amounts of data).

Q-Q plots Q stands for **quantile**. A Q-Q plot is useful when *comparing* two probability distributions or two samples. It is more "powerful" (as in easier to interpret) than comparing two histograms. It may also be used for "goodness of fit" to check whether our data follows a specific distribution. The most well-known Q-Q plot is the normal Q-Q plot that helps verify whether our data follows a normal distribution or not.

Before we introduce how Q-Q plots are built and read, we need to discuss quantile functions. What are quantile functions? As we saw earlier in the lecture, for any sample we have:

- p percentile: p% of the observations are below that value.
- Q1: first quartile, $p = 25$.
- Q2: second quartile, $p = 50$, also known as the median.
- Q3: third quartile, $p = 75$.

Now, for any random variable X with CDF $F(x)$, we define the **quantile function** as:

$$Q(p) = \inf\{x : F(x) \leq p\}, \quad \text{for } 0 \leq p \leq 1.$$

In English: look for the smallest value of x such that $F(x)$ is smaller than or equal to the given probability p. An interesting note: $Q(p) = F^{-1}(x)$.

So how do we build a Q-Q plot? We have a sample of n observations, and we have a theoretical distribution we believe our sample follows (e.g., exponential, normal, etc.). With this information at hand, we follow the procedure:

1. First, identify some quantile levels of interest: $0 < p_1 < p_2 < \ldots < p_n$.
   - Typically, we choose $p_i = \frac{i}{n+1}$, for n observations.
   - We could also choose $p_i = \frac{i-0.5}{n}$.
2. Then, we compute the *sample*'s quantiles. Let them be $X_1, X_2, \ldots, X_n$.
3. Now, we compute the *theoretical* quantiles, based on the $F(x)$ selected. Let them be $F^{-1}(p_1)$, $F^{-1}(p_2)$, …, $F^{-1}(p_n)$.
4. Finally, plot the sample quantiles against the theoretical quantiles in the same plot. See Figure 23 for an example.

Figure 23: An example of a Q-Q plot. Here the x axis shows the theoretical quantiles and the y axis the sample quantiles.

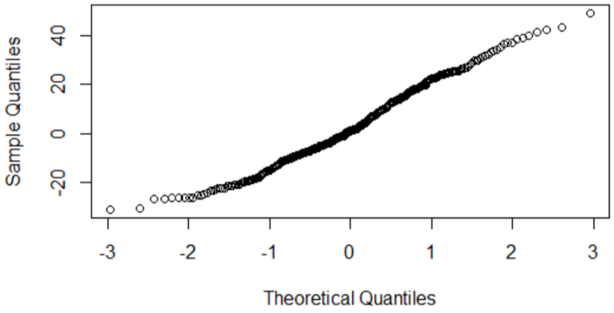
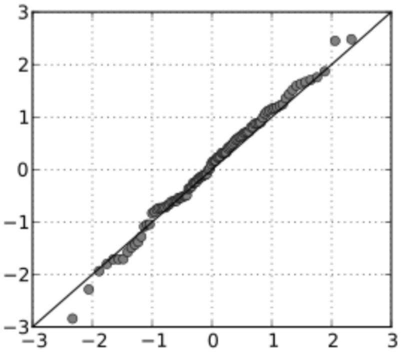


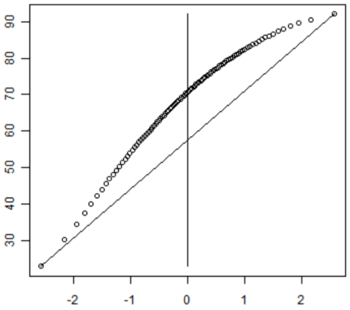
Figure 24: Here we assumed the sample follows a normal distribution so we plotted the sample's quantiles to the normal quantiles. We get a straight line, meaning our data comes indeed from the normal distribution!



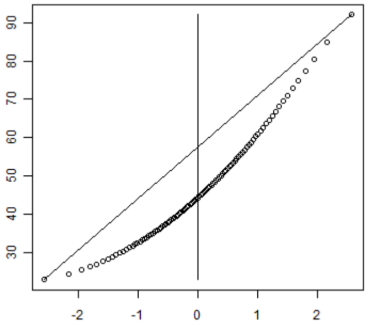
After we have a Q-Q plot, how do we use it? How do we read it? Well, the nice thing is that if indeed the sample follows that (theorized) distribution, then the Q-Q plot will look like a straight (45°) line, as in Figure 24!

We can also tell whether the distribution is left or right skewed.

a) Left-skewed:

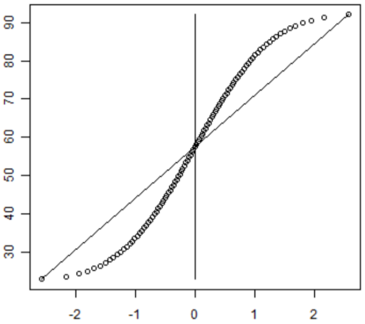


b) Right-skewed:

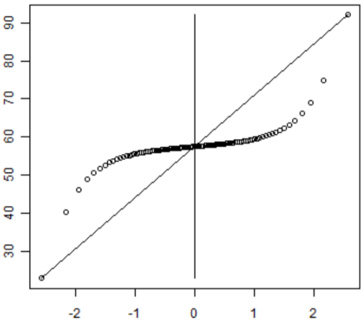


Finally, we can tell if it light- or heavy-tailed.

a) Light-tailed:



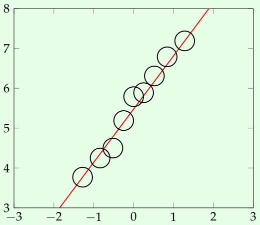
b) Heavy-tailed:



Constructing a Q-Q plot

Assume we collected the following observations from some population: $3.77, 4.25, 4.50, 5.19, 5.89, 5.79, 6.31, 6.79, 7.19$. Do the observations seem to come from a normal distribution? Let us construct a Q-Q plot to prove or disprove this.

We have $n = 9$ observations, so we can get $p_1 = 10\%, p_2 = 20\%, \ldots, p_9 = 90\%$. Find the z-values corresponding to the 9 percentage: -1.28, -0.84, -0.52, -0.25, 0, 0.25, 0.52, 0.84, 1.28. Finally, we plot them and see if they appear to form a 45° line or not.



Constructing a Q-Q plot

Assume we collected the following observations from some population: $3.77, 4.25, 4.50, 5.19, 5.89, 5.79, 6.31, 6.79, 7.19$. Do the observations seem to come from a normal distribution? Let us construct a Q-Q plot to prove or disprove this.

We have $n = 9$ observations, so we can get $p_1 = 10\%, p_2 = 20\%, \ldots, p_9 = 90\%$. Find the z-values corresponding to the 9 percentage: -1.28, -0.84, -0.52, -0.25, 0, 0.25, 0.52, 0.84, 1.28. Finally, we plot them and see if they appear to form a 45° line or not.