STAT 390 A
Statistical Methods in Engineering and Science
Week 1 Lectures – Part 1 – Spring 2023
Overview and Descriptive Statistics

Alexander Giessing

Department of Statistics

University of Washington

March 26, 2023

Outline

Class Description and Related Information (1/3)

- Read the Syllabus! Website: https://canvas.uw.edu/courses/1635461
- **First half**: Introduction to Probability Theory for Statisticians
  - Day 1: Overview and Descriptive Statistics.
  - Week 1: Axiomatic Introduction to Probability.
  - Week 2: Discrete Random Variables.
  - Week 3: Continuous Random Variables.
  - Week 4: Joint Probability Distributions.
  - Week 5: Asymptotic Results.
- **Midterm Exam**: Wed, April 26, 2:30pm – 3:20pm (location TBA).
- **Second half**: see Syllabus.
- **Textbook:** Devore, J. L. (2015). *Probability and Statistics for Engineering and the Sciences*, 9th edition, Cengage Learning.

Class Description and Related Information (2/3)

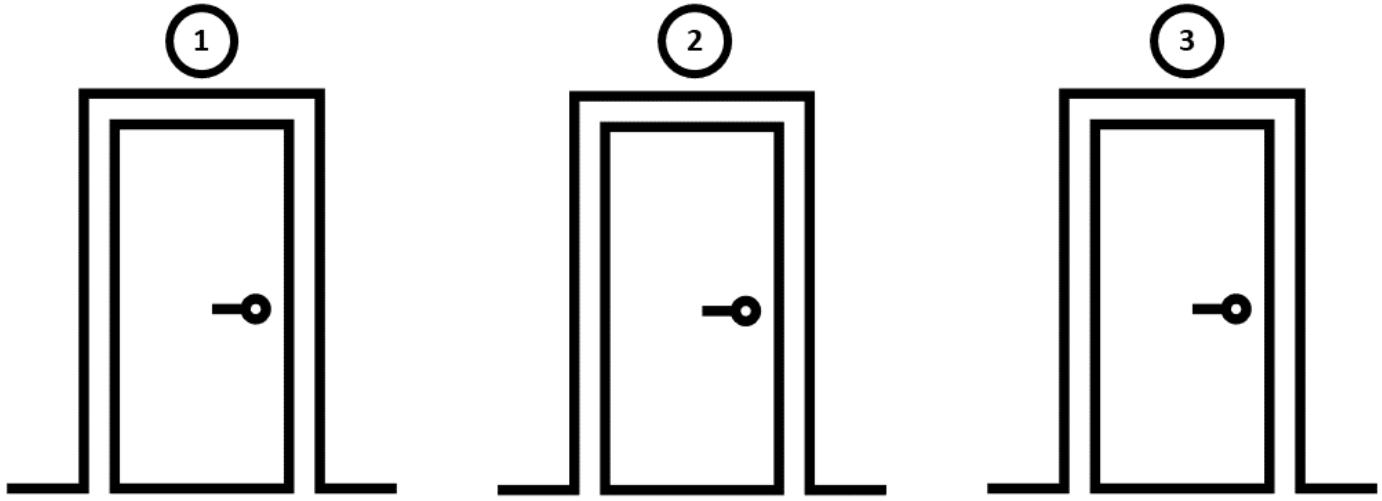
- My Office Hour: Mon, 11am – 12pm (B-308 PDL).
- Teaching Assistants:
  - Qiliang Chen (`qlchen@uw.edu`).
  - Dasha Petrov (`petrovd@uw.edu`).
  - Office hours TBA (see Canvas, landing page).
- Study Groups:
  - Discuss HW in study groups of 2–3 and on the Ed Discussion.
  - Submit individual solutions to homework, no verbatim copying.
- Homework:
  - Start early and finish on time!
  - Weekly problem sets; first PS is due Jan 13, 2023.
  - 35% of final grade; best 8 out of 9 problem sets.
- Quiz Sections: weekly, starting March 27, 2023; 5% of final grade.
- Midterm: 25% of final grade; Final: 35% of final grade.

Class Description and Related Information (3/3)

- Software/Coding:
  - `R` Base: https://cran.r-project.org/ (install prior to first lab)
  - `R` Studio: https://rstudio.com/ (install prior to first lab)
  - You are expected to learn R yourself with help of the TAs in the labs.
  - Homework will always contain a coding question.
  - Midterm and Final will test you on understanding of R code and the output of R code.
- Do you want to join the class but the class is full?
  - Please email Qiliang with Full Name, UW NETID, Grade in MATH 126, Class standing, and the Quiz Section which you'd like to join.
- Do you have any questions?
  - If not now, you can always send me an email (Tip: If you send me a message via Canvas, I will usually get back to you quicker.)

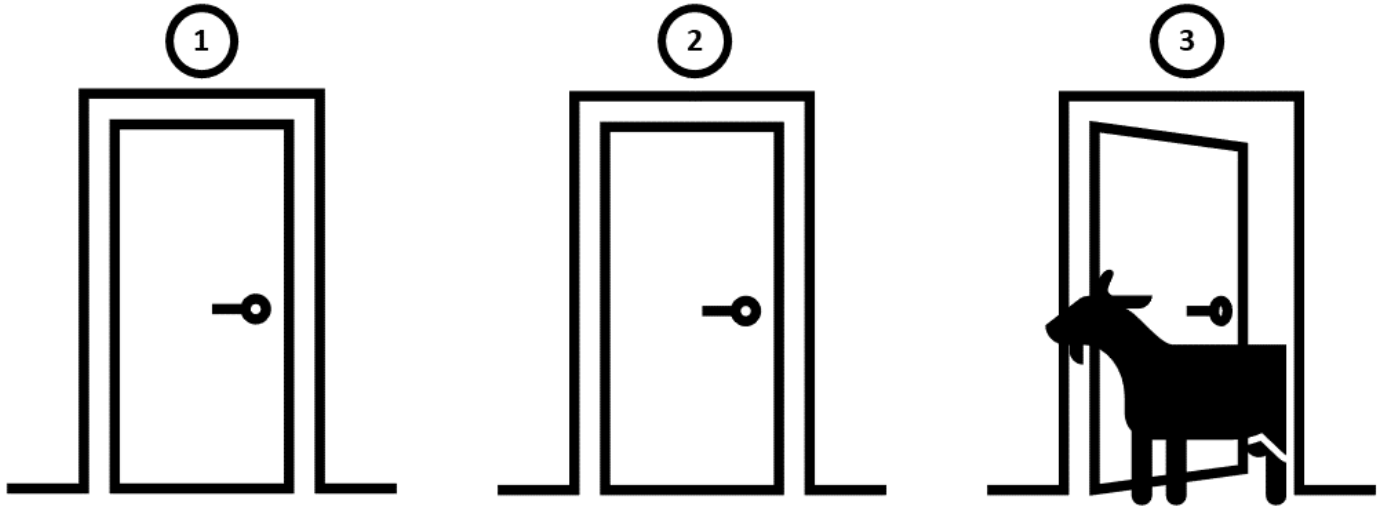
Outline

Cars and Goats: the Monty Hall Problem



Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? – Craig F. Whitaker.

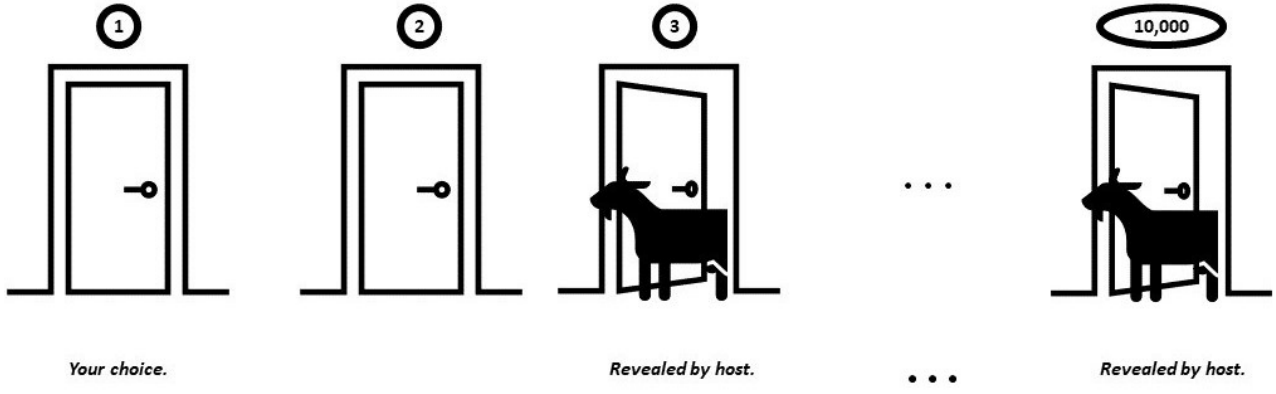
Cars and Goats: the Monty Hall Problem



Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? – Craig F. Whitaker.

Cars and Goats: the Monty Hall Problem

- The correct answer will most likely sound counter-intuitive at first!
- Suppose there are 10,000 doors, behind one is a car and behind the rest are goats. After you pick a door, the host will open 9,998 doors with goats. Will you switch now?



- To solve this and related problems rigorously, you will learn about ...
  - sample space and events,
  - (conditional) probabilities,
  - law of total probability,
  - Bayes' Rule.

The Space Shuttle *Challenger*

On January 28, 1986, the space shuttle *Challenger* exploded about one minute after taking off from the launch pad at Kennedy Space Center, Fla. The cause of the accident was a combustion gas leak through a joint in one of the booster rockets, which was sealed by a so-called O-ring.

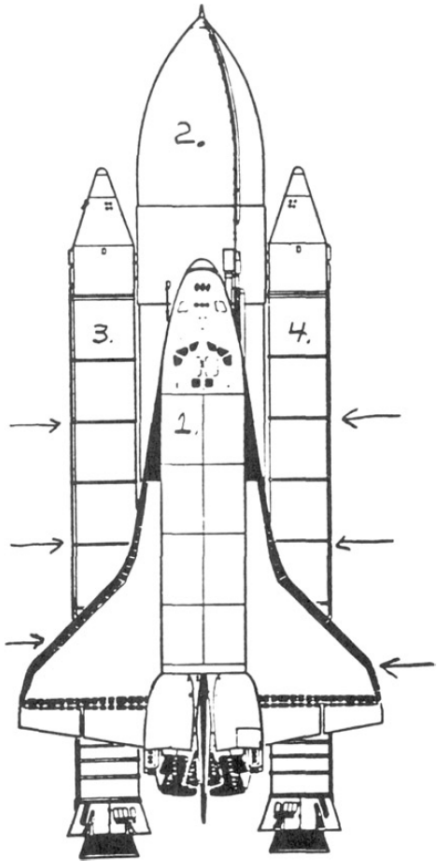


Figure 2. Space Shuttle: Orbiter, External Tank, Solid Rocket Motors, and Field Joints.

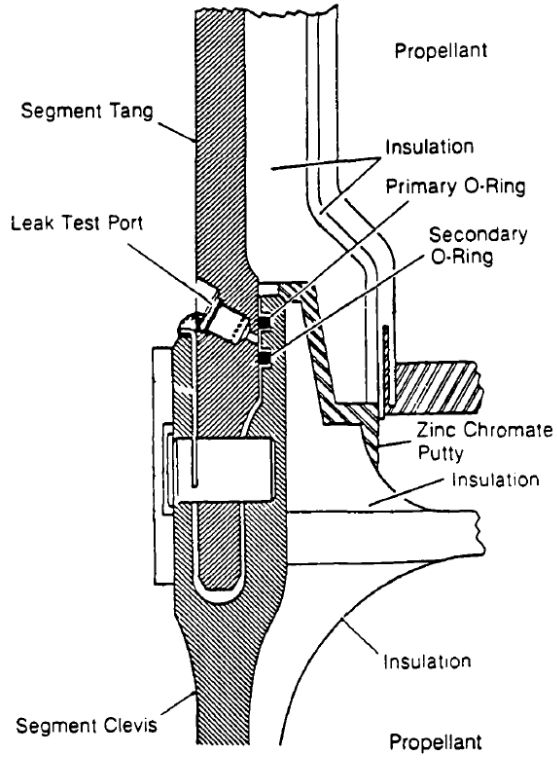


Figure 3. Solid Rocket Motor Cross Section: Tang, Clevis, and O-Rings.

The Space Shuttle *Challenger*

- On the evening of January 27, the decision to launch the next day was made, despite the fact that an unsually low temperatue of 31°F was predicted, well below the operating limit of 40°F of the booster rockets. Low temperatures were known to increase the likelihood of failure of O-rings.

- Was the management decision to overrule the engineers' recommendation to not launch justifiable? That is, was the risk of a failure of an O-ring so small that it would have been unreasonable to postpone the launch?

- To answer this question we need to ...
  - build a probabilistic model to assess the "risk of O-ring failures",
  - estimate the model using data,
  - use the estimated model to predict the risk at a temperature of 31°F.

The Space Shuttle *Challenger*

- To model the probability $p(t)$ of the failure of a single O-ring as a function of the temperature t, Dalal et al. (1989) propose a so-called logistic regression model:

$$p(t) = \frac{e^{a+b\cdot t}}{1+e^{a+b\cdot t}}.$$

- Since the number of failures of O-rings follows a binomial distribution, the expected number of failures of O-rings is $6 \cdot p(t)$.

- Dalal et al. (1989) use the method of maximum likelihood and historical data to obtain $a = 5.085$ and $b = -0.1156$.

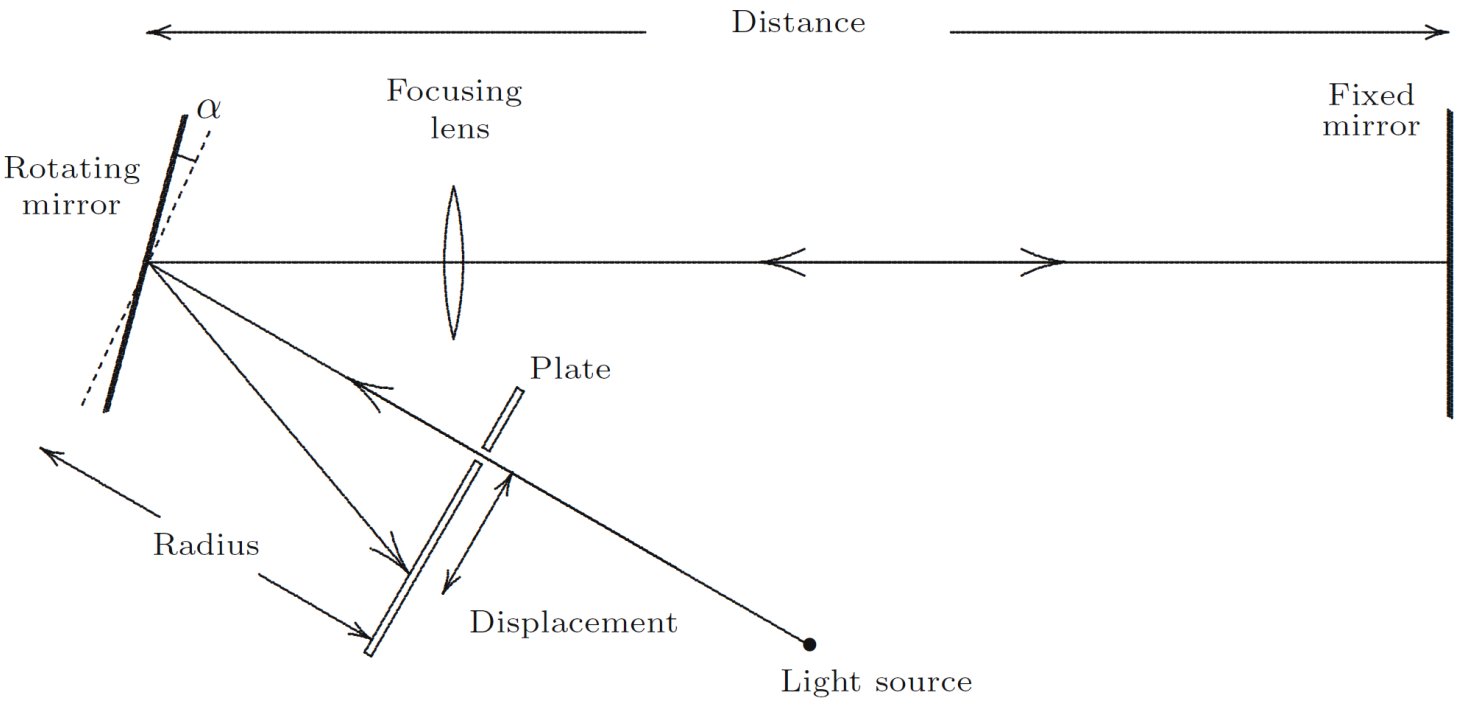
- Based on this model, the probability of an O-ring failure at 31°F is $p(31) = 0.8178 = 81.78\%$, the expected number of O-ring failures is $6 \cdot p(31) = 5.35$.

Measuring the Speed of Light

In theory, measuring the speed of light is a simple problem: since

$$\text{speed} = \frac{\text{distance}}{\text{time}},$$

we only need to accurately measure distance and time. In practice, this is a difficult problem. Take the experimental setup by Albert Michelson (1879):



Measuring the Speed of Light

- In Michelson's experiment, the elapsed time is determined by the revolutions per seconds (rps) of the mirror,

$$\text{time} = \frac{\alpha/2\pi}{\text{rps}},$$

  the angle α can be obtained from

$$\tan(2\alpha) = \frac{\text{displacement}}{\text{radius}},$$

  and the speed of light can be computed as

$$c = \frac{2 \times \text{distance}}{\text{time}}.$$

- Displacement and radius are relatively short and easy to measure.
  $\Longrightarrow$ time can be estimated accurately.

- The (long) distance is difficult to measure accurately.

Measuring the Speed of Light

- Michelson's solution:
  1. Measure the distance 5 times and compute c based on the average of these 5 measurements.
  2. Repeat Step 1 for 100 times.
  3. Estimate the speed of light by averaging over the 100 measures of c from Step 2.

- When does averaging 100 measurements make the final estimate more accurate? When less? Can we quantify the uncertainty of Michelson's estimate?

- To answer these questions, you will learn about ...
  - standard errors,
  - the central limit theorem,
  - confidence intervals.

Outline

Population and Sample

Population and Sample

A population is a well-defined collection of objects. A sample is a subset of a population.

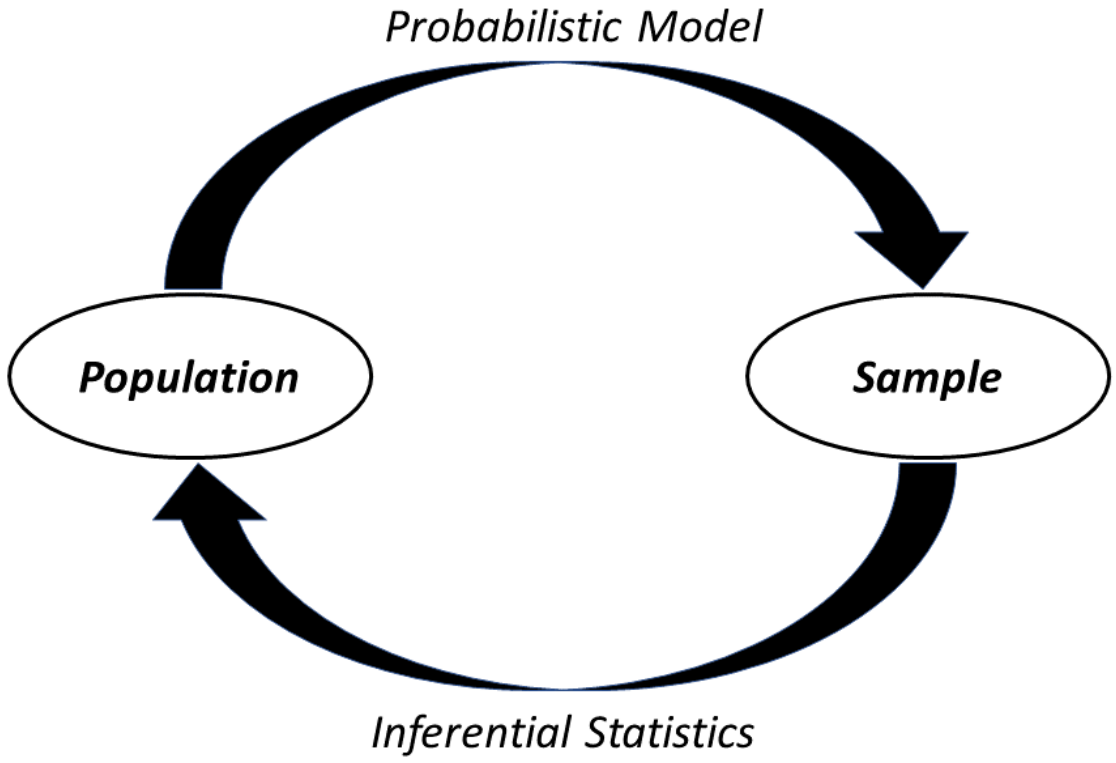
- Often we cannot measure every object in a population, instead we measure a small (representative) sample.

Simple Random Sample

A simple random sample is a sample which is constructed in such a way that each member of the population has equal chance of being part of the sample.

- Think of a lottery which assigns each member of the population a number, after which a subset of numbers are selected at random.

Probability and Statistics



- **Probabilistic models** describe how samples are drawn from a population.
- **Inferential statistics** uses samples and probabilistic models to learn about properties of a population.

Examples of Statistical Questions

- What is the typical pH value of human blood?
  - **Population**: all humans.
  - **Sample**: a few (randomly selected) participants whose blood pH (variable) is measured.
- Which images on social media attract more likes – cats or dogs?
  - **Population**: all images of dogs and cats on social media.
  - **Sample**: a few (randomly selected) images of cats and dogs and the corresponding counts of likes.
- Populations and samples can be abstract.
  - Population can be finite or infinite.
  - Measurements of interest could be discrete or continuous, or both.

Outline

Descriptive Statistics

- The following is a sample of 152 salaries (in thousands) of data scientists with a Master's degree:

122 127 130 131 132 133 134 134 135 135 135 136 137 138 139 140 143 124
127 130 132 132 133 134 134 135 135 136 136 137 138 139 140 143 124 128
131 132 133 133 134 134 135 135 136 136 137 138 139 141 143 125 128 131
132 133 133 134 134 135 135 136 136 137 138 139 141 143 126 129 131 132
133 133 134 134 135 135 136 137 137 138 139 141 144 126 129 131 132 133
133 134 134 135 135 136 137 138 138 140 141 144 126 129 131 132 133 133
134 134 135 135 136 137 138 138 140 142 144 127 129 131 132 133 133 134
134 135 135 136 137 138 138 140 143 147 127 130 131 132 133 134 134 135
135 135 136 137 138 139 140 143

- What are the salient features of this data?

Descriptive Statistics

Frequency Table

- Divide the range of the data into non-overlapping intervals.
- Count the number of observations in each interval, called frequencies.
- Compute relative frequencies as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{sample size}}.$$

Frequency table of salary data:

Intervals	122 – 124	124 – 126	126 – 128	128 – 130	$\cdots$	146 – 148
Frequency	3	4	6	7	$\cdots$	1
Rel. Freq.	0.0197	0.0263	0.03947	0.0461	$\cdots$	0.0066

Descriptive Statistics

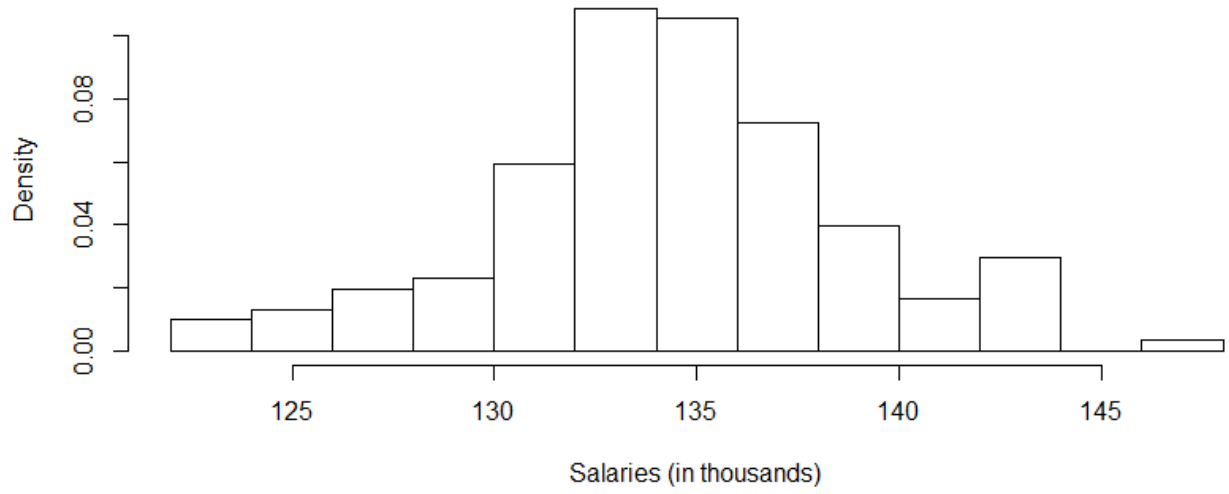
Histogram

- Purpose: Visualize the distribution of the data.
- Compute the so-called density of an interval as

$$\text{density} = \frac{\text{relative frequency of interval}}{\text{width of interval}}.$$

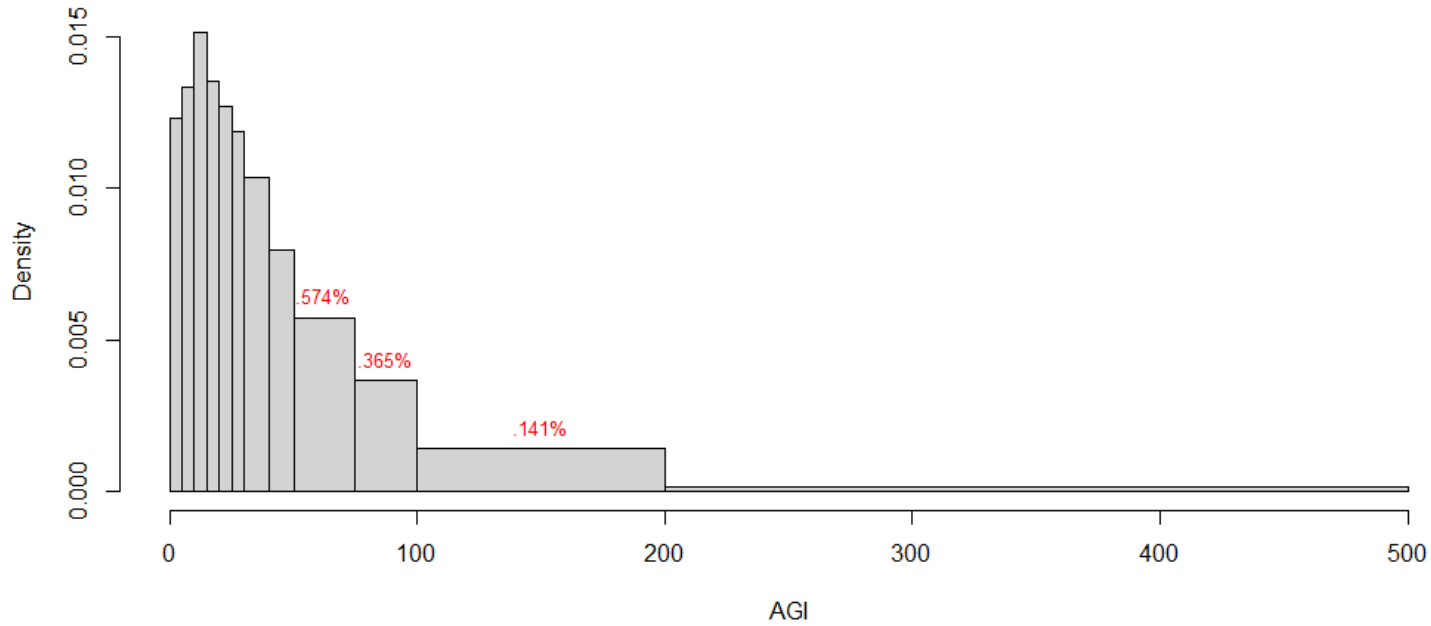
- Plot densities over corresponding intervals.

Histogram of salary data:



Example: Adjusted Gross Income 2020 (in thousands)

Intervals	0–5	5–10	10–15	15–20	20–25	25–30	30–40	40–50	50–75	75–100	100–200	200–500	> 500
rel. Freq.	0.0608	0.0659	0.0750	0.0669	0.0628	0.0588	0.1023	0.0790	0.1418	0.0902	0.1398	0.0456	0.0111

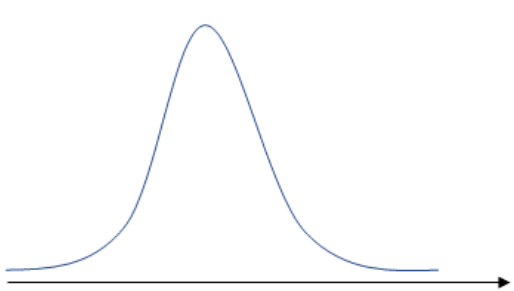


(https://www.irs.gov/statistics/soi-tax-stats-individual-statistical-tables-by-size-of-adjusted-gross-income)

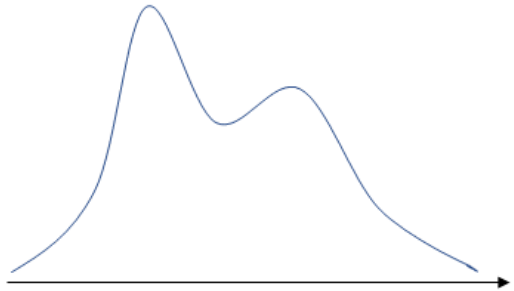
Example: Adjusted Gross Income 2020 (in thousands)

- What is the shape of the income distribution?
  - right/ positively skewed.
- Which interval is more crowded (dense): (10, 15) or (50, 75)?
  - the former.
- Which interval comprises more households: (10, 15) or (50, 75)?
  - the latter.
- Where is the mode of the distribution, i.e the most dense interval/ highest bar?
  - over the interval (10, 15).
- What is the density (height) over the interval (50, 75)?
  - $0.1418/25 = 0.574\%$.
- What is the percentage of households with income in (60, 120)?
  - $15 \times .574\% + 25 \times .365\% + 20 \times .0141\% = 18.02\%$.

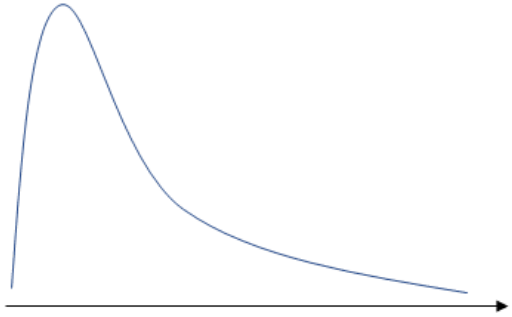
Common shapes of histograms



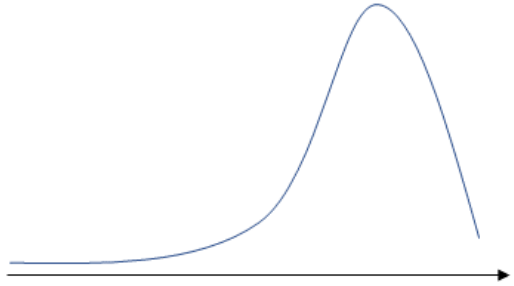
symmetrical unimodal



bimodal



Positively skewed/
skewed to the right



negatively skewed/
skewed to the left

Summary Statistics of Location

Sample Mean

Let $x_1, \ldots, x_n$ be a sample. The sample mean is defined as

$$\bar{x} := \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Sample Median

Let $x_{(1)}, \ldots, x_{(n)}$ be an ordered sample, i.e. $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. The sample median is defined as

$$x_{\text{med}} := \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{if } n \text{ is even.} \end{cases}$$

- The sample median is the middle value after ordering the observations from smallest to largest.

Example: What To Report – Mean or Median?

- The survival times (in days) of 6 patients after heart transplants in a hospital are

$$15, 3, 46, 623, 126, 64.$$

  - The sample mean is

$$\bar{x} = \frac{15 + 3 + 46 + 623 + 126 + 64}{6} = 146.2 \text{ days.}$$

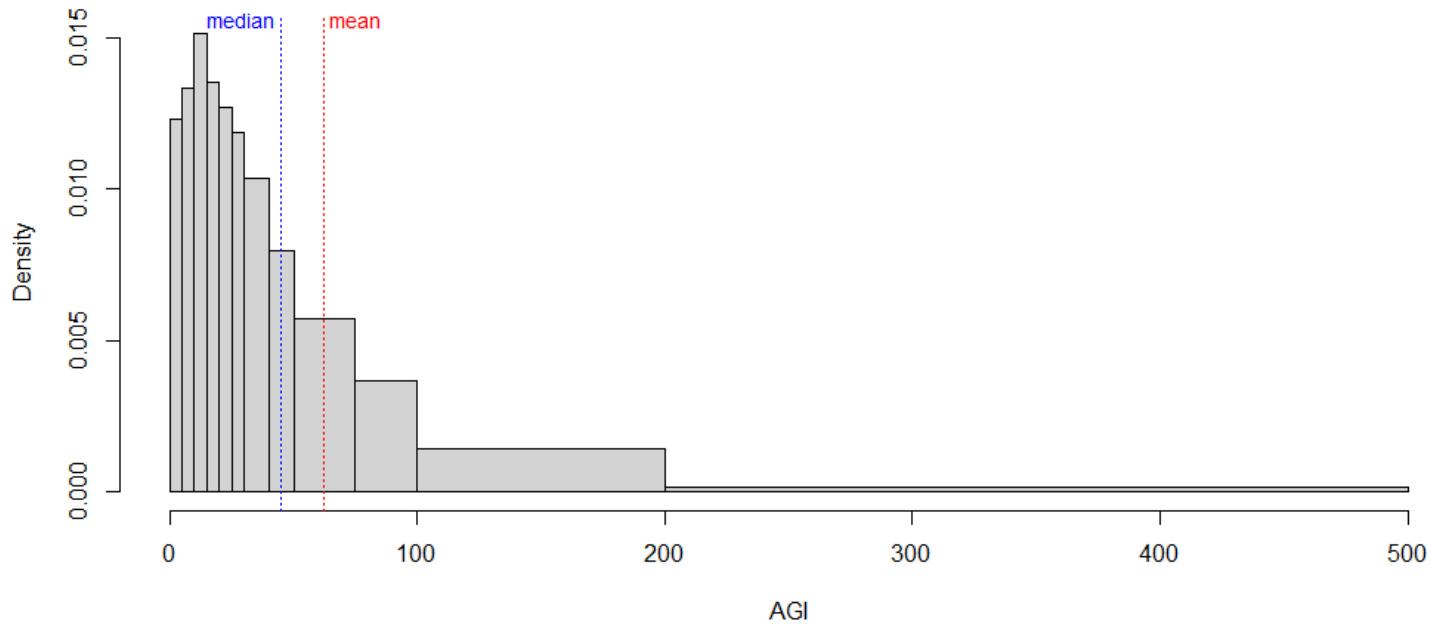
  - The sample median is

$$x_{\text{med}} = \frac{46 + 64}{2} = 55 \text{ days.}$$

- Only 1 out of 6 patients survived longer than the mean. Clearly, in this example, the median summarizes the data better than the mean.

Relation between Histogram, Mean, and Median

- The sample median is robust to outliers (i.e. unusually large or small observations), whereas the sample mean is not.
- The sample mean is the center of gravity of the corresponding histogram. A histogram balances when supported at the sample mean.
- The sample median divides the corresponding histogram so that half of the area is to its left and half to its right.



Descriptive Statistics

Order Statistics

Let $x_1, \ldots, x_n$ be a sample. The order statistics consist of the same elements $x_1, \ldots, x_n$, but arranged in ascending order. If we denote by $x_{(k)}$ the kth element in the ordered list, then

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Empirical Quantiles

The pth empirical quantile is the smallest number $q_n(p)$ which divides a sample in two parts in such a way that at least a proportion p of the sample is less than $q_n(p)$ and at most a proportion $1 - p$ is greater than $q_n(p)$.

Example: Finding Empirical Quantiles

- The following data show the grades of 30 students:

  25 34 55 59 63 63 65 71 73 74 75 77 78 80 81
  81 82 84 85 85 86 86 87 88 90 91 92 95 98 99

- 0.14th quantile is 63.
  - $0.14 \times 30 = 4.2 \implies q_n(0.14) = x_{(5)} = 63.$
- 0.5th quantile is 81.
  - $0.5 \times 30 = 15 \implies q_n(0.5) = x_{(15)} = 81.$
- 0.2th quantile is 63.
- 0.75th quantile is 87.

Terminology for Empirical Quantiles

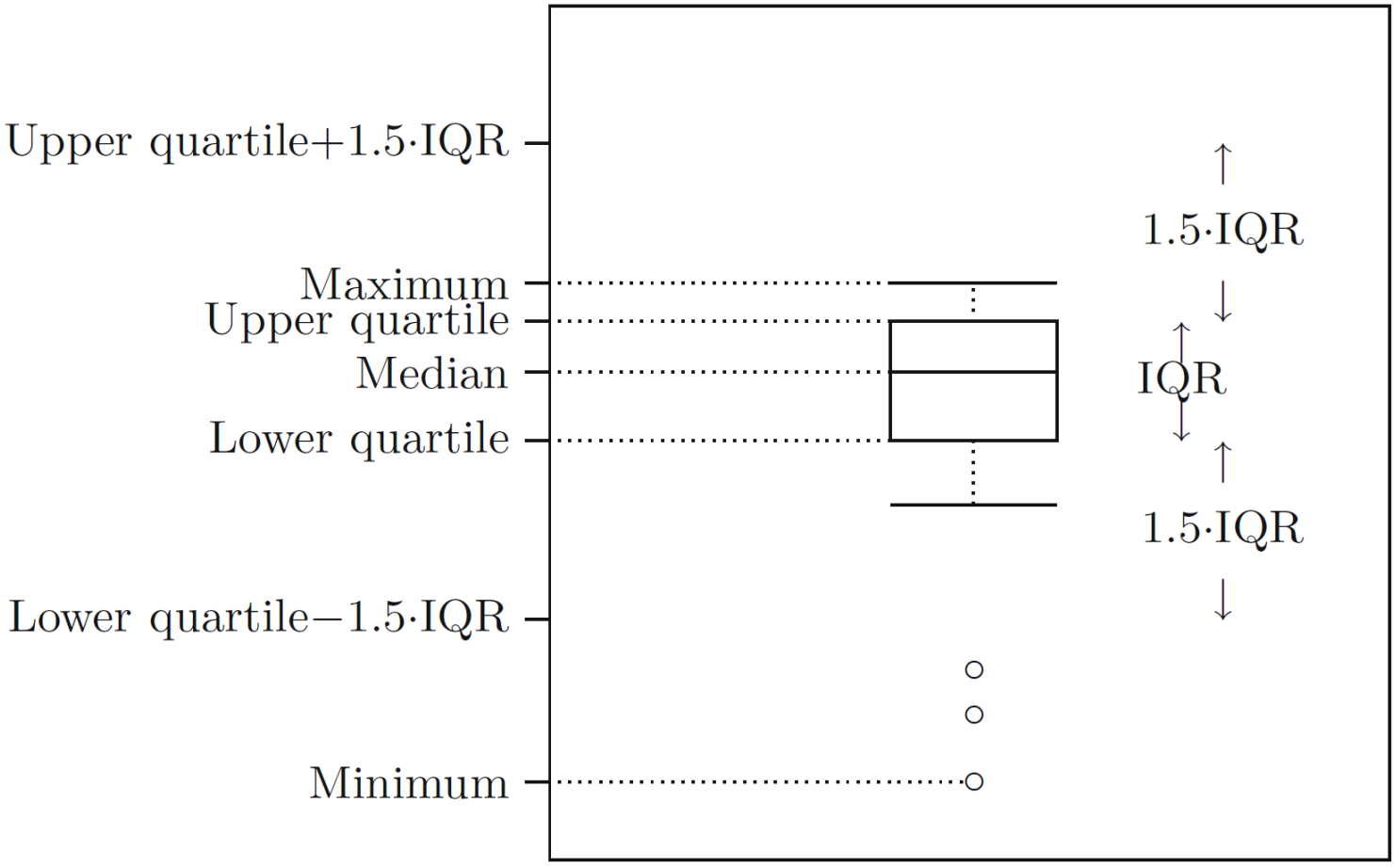
- $100p$th **empirical percentile** = pth emp. quantile
  = $100p\%$ largest value.
- **Lower (First) Quartile:** $Q_1 = 25$th percentile.
- **Second Quartile (Median):** $Q_2 = 50$th percentile.
- **Upper (Third) Quartile:** $Q_3 = 75$th percentile.
- **Inter Quartile Range:** $\text{IQR} = Q_3 - Q_1$.

Descriptive Statistics

Boxplot

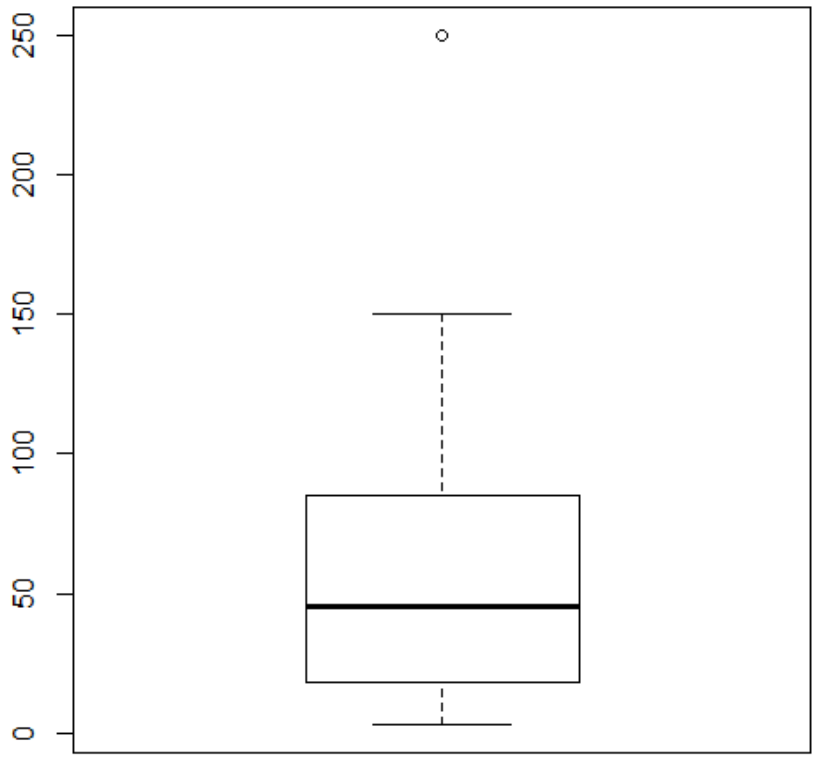
- Purpose: Visualize the skewness and variability of a sample.
- Data is on the vertical axis, width of the horizontal box is irrelevant.
- Box: spans from lower to upper quartile.
- Location of two whiskers:
  - largest observation that lies within $1.5 \times$ IQR of the upper quartile.
  - smallest observation that lies within $1.5 \times$ IQR of the lower quartile.
- Outliers: observations that lie beyond the whiskers; denoted by $\circ$.

Schema of a Boxplot



Example: Adjusted Gross Income 2018 (in thousands)

Intervals	0–5	5–10	10–15	15–20	20–25	25–30	30–40	40–50	50–75	75–100	100–200	200–500	> 500
rel. Freq.	0.0608	0.0659	0.0750	0.0669	0.0628	0.0588	0.1023	0.0790	0.1418	0.0902	0.1398	0.0456	0.0111



- We infer that the distribution is skewed to the right.
- We cannot tell whether the distribution is uni-, bi-, or multi-modal.

Summary Statistics of Variability

- Sample mean and median provide a numerical summary of the center of a sample. How to summarize the variability in a sample?
- Deviations from the mean: primary measure of variability.
- For a sample $x_1, x_2, \cdots, x_n$ the deviations from the mean are

$$x_1 - \bar{x}, \;\; x_2 - \bar{x}, \;\; \ldots, \;\; x_n - \bar{x}.$$

- We want to combine these deviations into a meaningful quantity.
  - Squared average:

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

  - Absolute average:

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

- For reasons that will be discussed in Week 6 Lectures, we use slightly different measures of variability (defined on the next slide).

Summary Statistics of Variability

Sample Standard Deviation

Let $x_1, \ldots, x_n$ be a sample. The sample standard deviation is defined as

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

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ is the sample mean.

- Shortcut formula: $s^2 = \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) / (n-1)$.

Sample Median Absolute Deviation

Let $x_1, \ldots, x_n$ be a sample. The sample standard deviation is defined as

$$\text{MAD} := \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{\text{med}}|,$$

where x_{med} is the sample median.

- A robust alternative to the sample standard deviation.

Example with R Code: SP500 index and IBM stock

We consider the adjusted closing prices of the SP500 index and the IBM stock from Jan. 1, 2000 to Sep 8, 2016. The following questions guide you through a comparison of the distributions of their log-returns 8 years before and after the 2008 financial crisis. The data sets are available on Canvas (Files → Data Sets).

- Compare time series the stock prices of the SP500 index and IBM.
- Compare the skew of the distributions of SP500 index and IBM stock.
- Which asset is riskier, i.e. has higher volatility?
- Did the distribution of log-returns of the SP500 change before/ after the 2008 financial crises (structural break)?
- Compute the 95% empirical percentile of the losses of the SP500 before/ after the 2008 financial crises. *(The so-called "Value-at-Risk at* $100\alpha\%$*" measures the maximal loss that an investor might incur with probability* $1-\alpha$*.)*

Example with R Code: SP500 index and IBM stock

- First, read data and take a look at the structure....

```
> IBM <- read.csv("IBM.csv", header=T)      # read data
> IBM[1:3,]                                 # display first 3 rows
      Date   Open   High    Low  Close  Volume Adj.Close
1 9/8/2016 160.55 161.21 158.76 159.00 3919300    159.00
2 9/7/2016 160.19 161.76 160.00 161.64 2867300    161.64
3 9/6/2016 159.88 160.86 159.11 160.35 2994100    160.35

> SP500 <-  read.csv("SP500.csv", header=T)
> SP500[1:3,]
      Date    Open    High     Low   Close     Volume Adj.Close
1 9/8/2016 2182.76 2184.94 2177.49 2181.30 3727840000   2181.30
2 9/7/2016 2185.17 2187.87 2179.07 2186.16 3319420000   2186.16
3 9/6/2016 2181.61 2186.57 2175.10 2186.48 3447650000   2186.48
```

- We are interested in the 7th column, i.e. `Adj.Close`.
- Most recent observation is on top of the data set, oldest at the bottom.

Example with R Code: SP500 index and IBM stock

- Get adjusted closing prices and convert them into log-returns.

```
> pSP500 <- SP500[,7] #take adj close price column
> pSP500 <- rev(pSP500) #reverse the time order
> rSP500 <- diff(log(pSP500))*100 #percentage of returns

> pIBM <- rev(IBM[,7]) #Closing prices of IBM
> rIBM <- diff(log(pIBM))*100 #percentage of log-returns
```

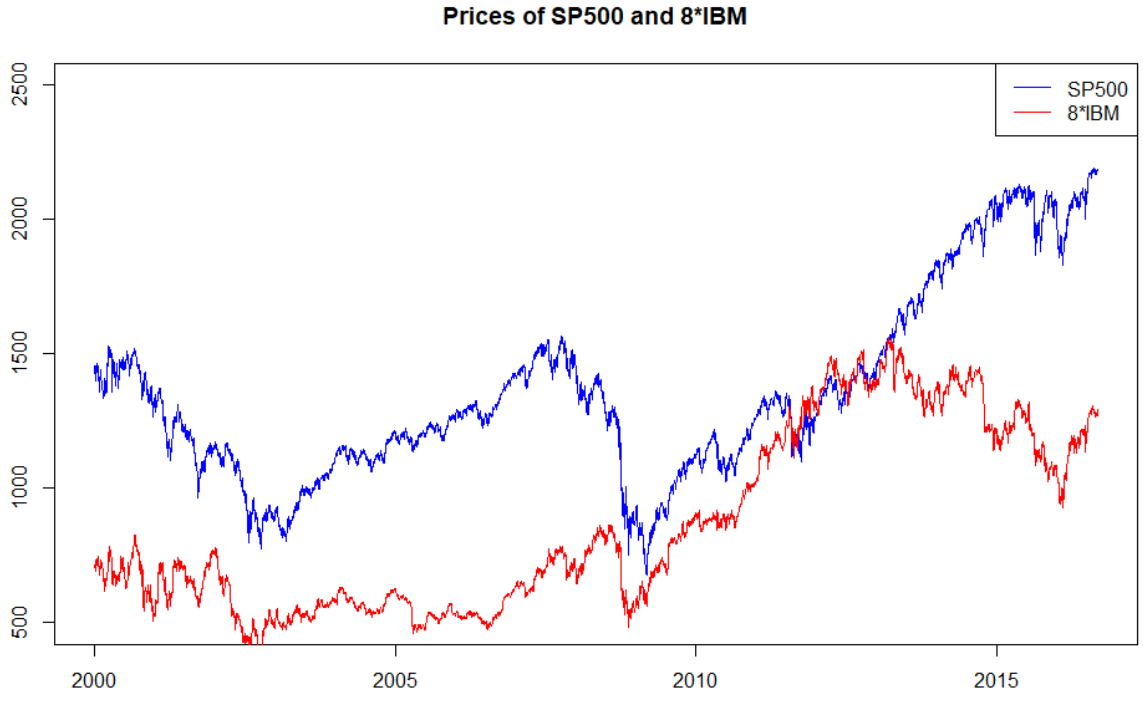
- Re-order data so that oldest observation is on top (o/w time series plot will be reversed – try it out yourself!)

```
> Dates <- as.vector(IBM[,1]) # dates of Data
> Dates <- strptime(Dates, "%m/%d/%Y") # convert to POSIXlt
                                       # (a date class)
> Dates <- rev(Dates) # time from past to future
```

Example with R Code: SP500 index and IBM stock

- Plot time series of adj. closing prices of SP500 index and IBM stock. Note that we scale the IBM stock price by 8 for a more compact plot.

```
> plot(Dates, pSP500, ylim=c(500, 2500), col=4, type="l") # creat plot
> lines(Dates, 8*pIBM, col=2) # add lines
> title("Prices of SP500 and 8*IBM") # add title
> legend(x = "topright", legend = c("SP500", "8*IBM"),
+ lty = c(1, 1), col = c("blue", "red"), lwd = 1) # add legend
```

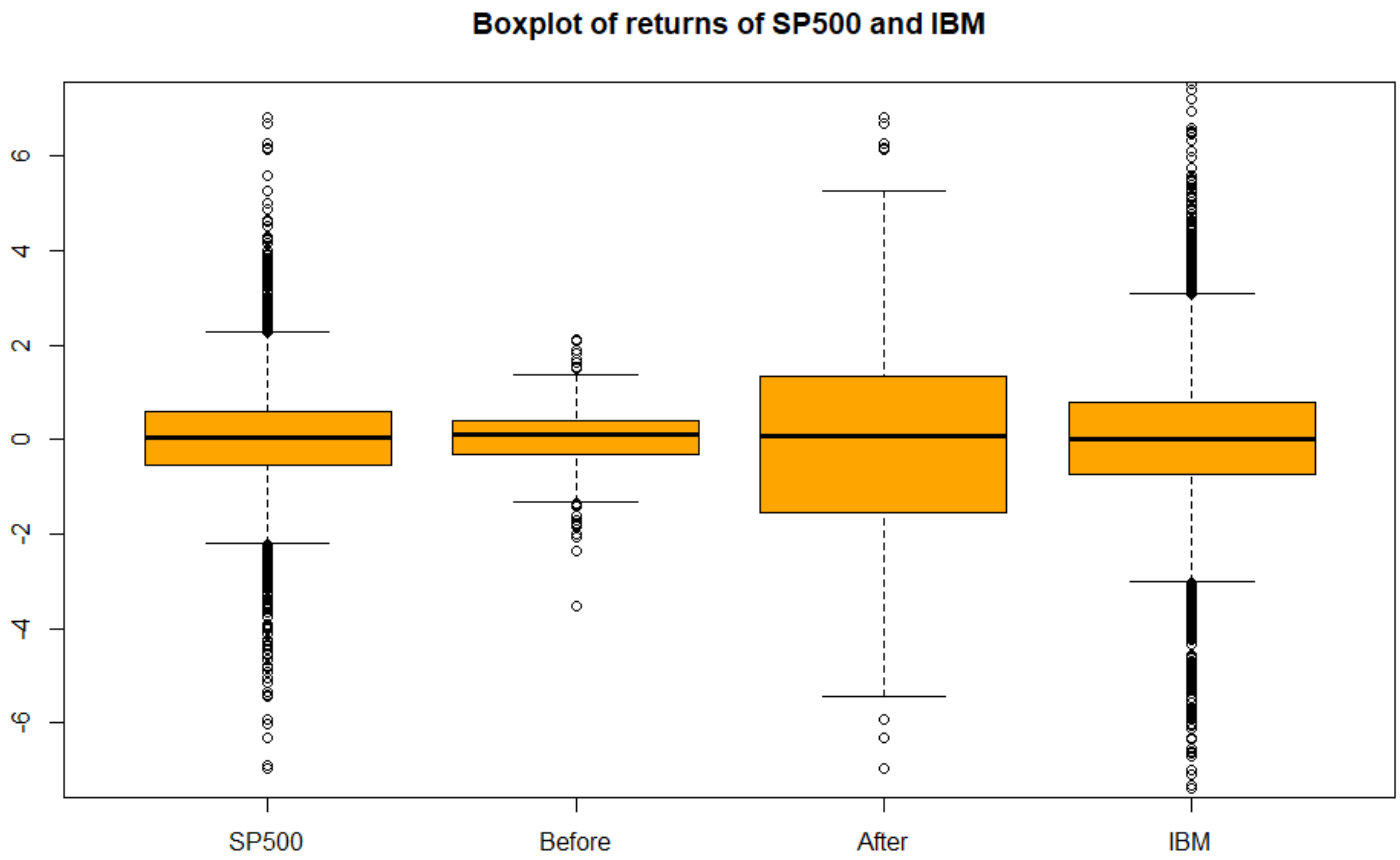


Example with R Code: SP500 index and IBM stock

- Compare the skew of the distributions.
- Which asset is riskier, i.e. has higher volatility?
- Did the distribution of log-returns of the SP500 change before/ after the 2008 financial crises (structural break)?

```
> rSP500a <- rSP500[2136:2387] # returns from 7/1/08 -- 6/30/09
                               # (after financial crisis)
> rSP500b <- rSP500[1509:1904] # returns from 01/03/06 -- 07/31/07
                               # (before financial crisis)
> boxplot(list(rSP500, rSP500b, rSP500a, rIBM),
> names=c("SP500", "Before", "After", "IBM"),
+ col="Orange", ylim=c(-7,7), xlab="") # side-by-side boxplots
> title("Boxplot of returns of SP500 and IBM")
```

Example with R Code: SP500 index and IBM stock

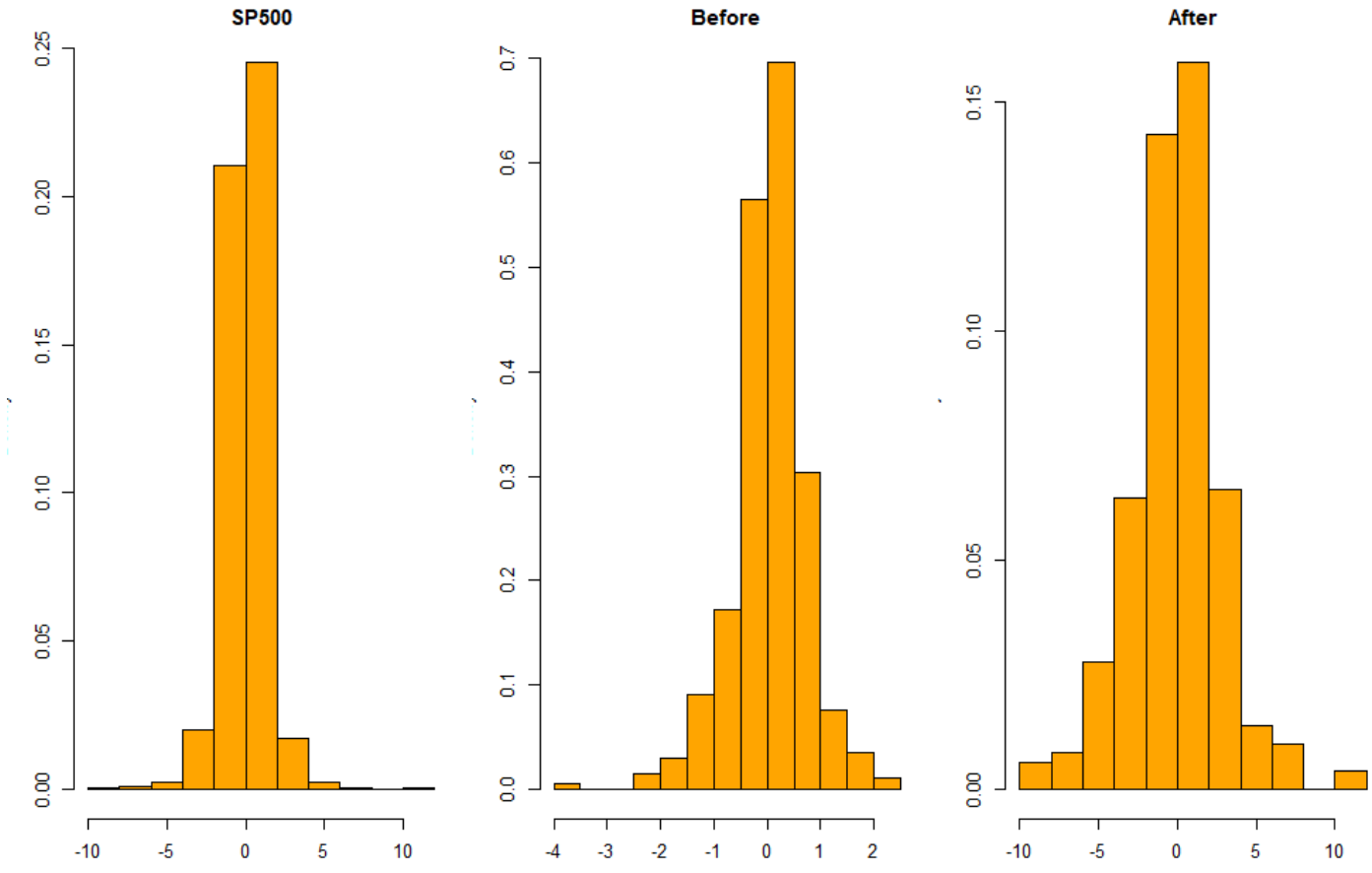


- IBM stock was more risky than the SP500 index.
- The SP500 was more volatile right after the financial crisis 2008 than right before.

Example with R Code: SP500 index and IBM stock

- We reach a similar conclusion by plotting side-by-side histograms.

```
> par(mfrow = c(1,3), cex=0.8) #1x3 subplots
> hist(rSP500, col="orange", prob=T, main="SP500", xlab="")
> hist(rSP500b, col="orange", prob=T, main="Before", xlab="")
> hist(rSP500a, col="orange", prob=T, main="After", xlab="")
```



Example with R Code: SP500 index and IBM stock

- A more quantitative analysis can be done as follows:

```
> summary(rSP500)
Min.    1st Qu.    Median      Mean   3rd Qu.      Max.
-9.469512 -0.536993  0.052386  0.009644  0.588574 10.957197
> summary(rSP500b)
Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
-3.53427 -0.29520  0.09643  0.03645  0.38891  2.13358
> summary(rSP500a)
Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
-9.46951 -1.53979  0.06895 -0.13113  1.34223 10.95720
> summary(rIBM)
 Min.    1st Qu.    Median      Mean   3rd Qu.      Max.
-16.89162  -0.73623   0.01864   0.01376   0.79594  11.35364
```

- The functions for the inter quartile range and the sample standard deviation are `IQR(...)` and `sd(...)`, respectively.

Example with R Code: SP500 index and IBM stock

- Compute the 95% quantile of the losses of the SP500 before/ after the 2008 financial crises.

```
> quantile(-rSP500b, 0.95)  # losses are negative returns!
    95%
1.15855
> quantile(-rSP500a, 0.95)
    95%
4.922432
```

- The 95% empirical percentile of the losses increased by more than 400% right after the financial crises compared to right before.

Getting ready for your first Lab (March 28, 2023)

- Install `R` Base and `R` Studio.
- Read STAT 390 Introduction to Programming in R.
  - Available on Canvas (Files → Lab Notes).
  - Read Intro 1 **before** for your Lab on Tuesday, Jan 10, 2023!
  - Read Intros 2 and 3 at a later time.
- The best and the most effective way to learn `R`: use it!
  - Hands-on experience is the most illuminating.
  - Experiment with the code from today's lecture.