Week 2: Introduction to Statistics

Rabail Tahir

Recommended books

- Essential Math for Data Science

https://learning.oreilly.com/library/view/essential-math-for/9781098102920/

- Probability and Statistics for Computer Scientists, 2nd Edition

https://learning.oreilly.com/library/view/probability-and-statistics/9781439875919/

- The Statistics and Calculus with Python

https://learning.oreilly.com/library/view/the-statistics-and/9781800209763/cover.xhtml

What is Statistics?

- Statistics is the process of collecting and analyzing Data that is useful and can help in making decisions and predictions.
- Probability is an integral part of statistics that helps in estimating the likelihood of any event's occurrence.
- To understand the concepts of Big data, Data mining and Machine learning, we need to have a strong foundation in Statistics and Probability.

What is Data?

- Information? Facts?
- “Data is not important in itself. It’s the analysis of data (and how it is produced) that is the driver of all these innovations and solutions”
- Data consists of
  - Variables
  - Gaps
  - Bias

Can you tell a story of a family just by looking at one family photo?

“Data is actually just snapshots of a given time capturing only what it is aimed at.”

What questions should we ask about Data?

- How was the Data created?
- Who created it?
- What are the aspects which have not been captured?

"Does your data represent a ground truth that's verifiable and complete? Are the sensors and sources reliable and accurate? Or is the ground truth unknown?"

Descriptive vs Inferential Statistics

- Descriptive Statistics is used to summarize data
  - Mean
  - Mode
  - Median
  - Charts
  - Bell curve etc

Descriptive vs Inferential Statistics

- Inferential Statistics deals with making conclusions and inferences about a large population using samples taken from the population.
- It is not easy to get it right because it may end up in making wrong generalizations.

Applications of Statistics in Data Science

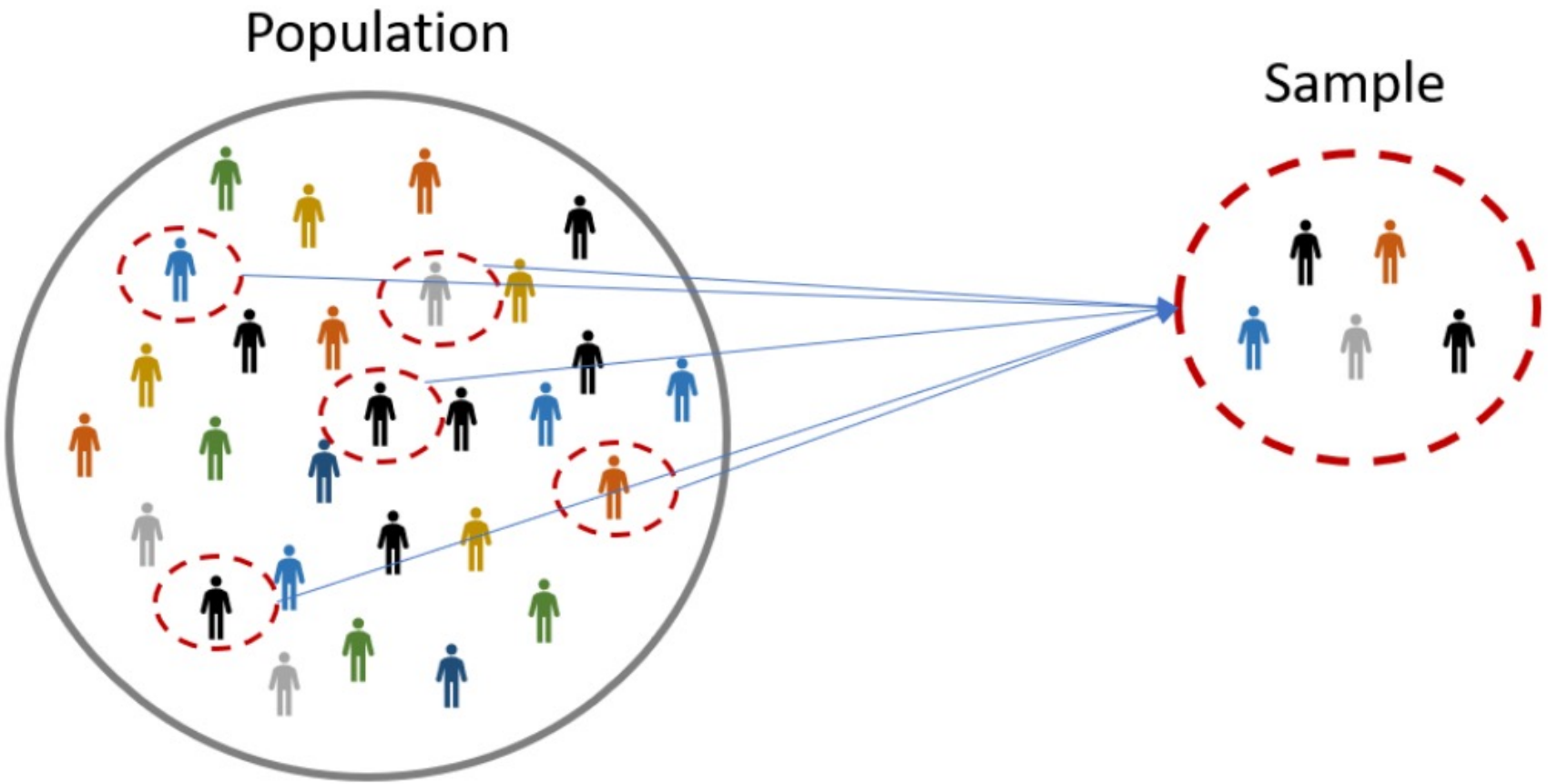
- Machine Learning
- Simulation
- Data Visualization
- Data Analytics
- Data Mining
- Data classification
- Logistic Regression

How can Statistics help us in Data Science?

- Organizes Data
- Helps in identifying trends
- Helps in estimation
- Provides with easier methods for Data visualization
- Helps reduce assumptions and bias
- Helps in measuring and analyzing multiple variables

Populations, Samples and Bias

- “A **population** is a particular group of interest we want to study.”
- For example, “all students studying at Roehampton”, “All people living in the Borough of Brent”.
- The boundary of a population can be wide and extensive and depends on factors such as geography and age.
- A **sample** is random and unbiased subset of population.
- Samples help us to infer attributes about a certain population when the population is too large and inaccessible.
- Sample should be as random as possible to avoid skewing the conclusion


Population
Sample

Sampling and non-sampling errors

- Sampling errors are caused when a very small portion of the population is used a sample. Sampling errors decrease when sampling size increases.
- Non-sampling errors are caused by inappropriate sampling schemes or wrong statistical techniques.

The problem of “Data Bias”. What do you think it means?

Types of Bias

- Confirmation Bias

"Confirmation bias is gathering only data that supports your belief, which can even be done unknowingly"

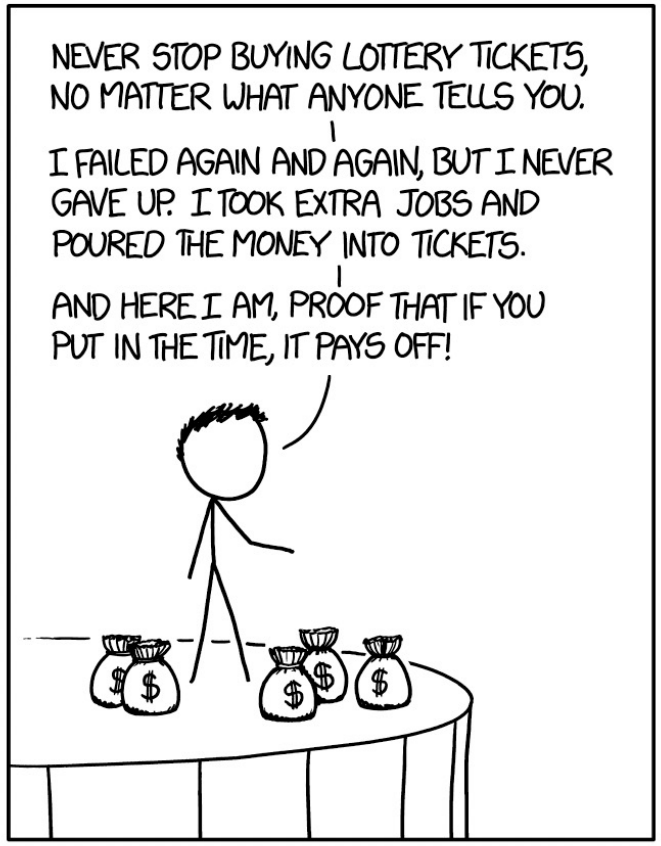
- Self selection Bias

"Which is when certain types of subjects are more likely to include themselves in the experiment"

- Survival Bias

"Captures only living and survived subjects, while the deceased ones are never accounted for"

Survivorship Bias



XKCD Comics

Descriptive Statistics
Mean

- The means is the average of a set of values.
- Sum of values divided by the number of values.

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \sum \frac{x_i}{n}$$

$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N} = \sum \frac{x_i}{N}$$

Note : X (“x-bar”) is used for sample and μ (“mu”) is used for population.

Weighted Mean

- Weighted Mean is when we assign a ”weight” with every value instead of giving equal importance to every value.

$$\text{weighted mean} = \frac{(x_1 \cdot w_1) + (x_2 \cdot w_2) + (x_3 \cdot w_3) + \ldots (x_n \cdot w_n)}{w_1 + w_2 + w_3 + \ldots + w_n}$$

Median

- Median is the centermost value in a set of ***ordered*** values.
- If the number of values is even, then the median will be the average of 2 centermost values.

Why Median is sometimes better than mean?

The median can be a helpful alternative to the mean when data is skewed by *outliers*, or values that are extremely large and small compared to the rest of the values. Here's an interesting anecdote to understand why. In 1986, the mean annual starting salary of geography graduates from the University of North Carolina at Chapel Hill was $250,000. Other universities averaged $22,000. Wow, UNC-CH must have an amazing geography program!

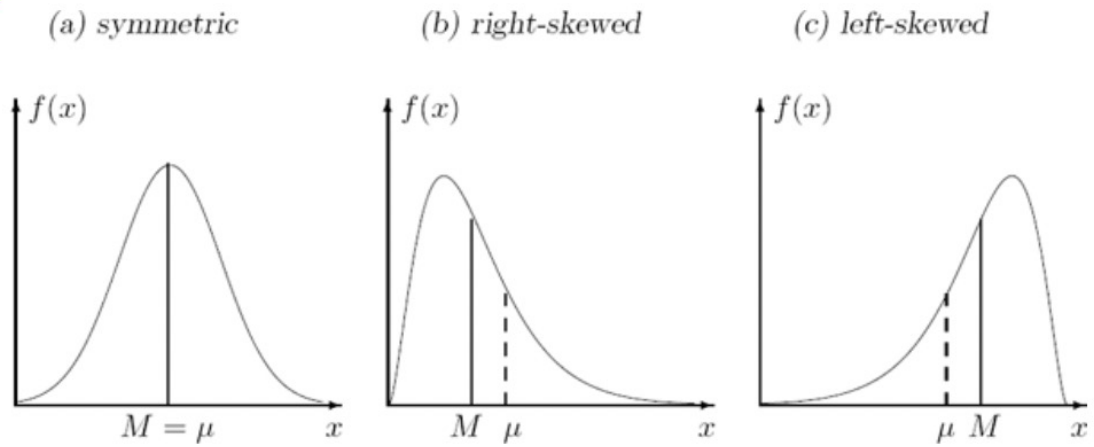
But in reality, what was so lucrative about UNC's geography program? Well... Michael Jordan was one of their graduates. One of the most famous NBA players of all time indeed graduated with a geography degree from UNC. However, he started his career playing basketball, not studying maps. Obviously, this is a confounding variable that has created a huge outlier, and it majorly skewed the income average.

This is why the median can be preferable in outlier-heavy situations (such as income-related data) over the mean. It is less sensitive to outliers and cuts data strictly down the middle based on their relative order, rather than where they fall exactly on a number line. When your median is very different from your mean, that means you have a skewed dataset with outliers.

Book excerpt taken from :
"Essential Math for Data Science"

Skewness

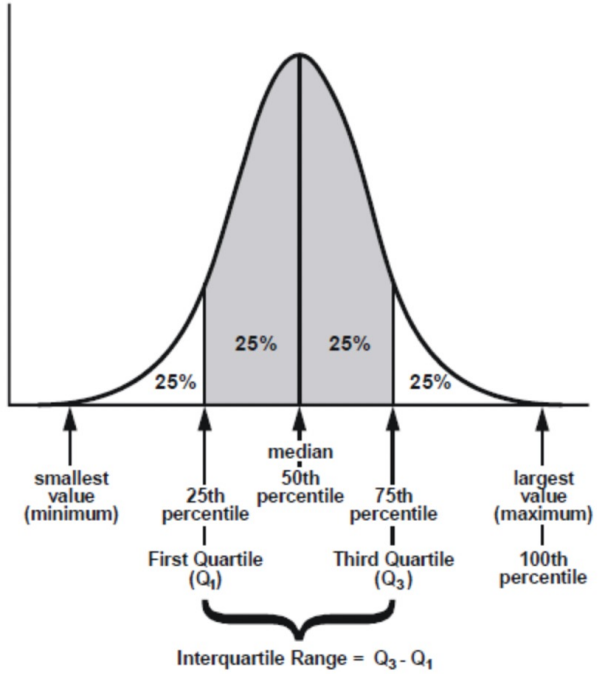
- Comparing the means and median will help us understand and analyze the distributions.



A mean μ and a median M for distributions of different shapes.

Quantile and Quartile

- Cutting data in places other than the middle.
- Median is 50% quantile.
- Quartile is when you cut data in 25% increments.



Mode

- Mode is the most frequently occurring set of values.
- It is useful when data is repetitive.
- When 2 values occur with equal frequency. That dataset is called ***bimodal.***

Population Variance and Standard deviation

- Variance is the measure of how spread out the data is.

$$\text{population variance} = \frac{(x_1 - mean)^2 + (x_2 - mean)^2 + \ldots + (x_n - mean)^2}{N}$$

More formally, here is the variance for a population:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Standard Deviation

- Square root of variance gives us standard deviation.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

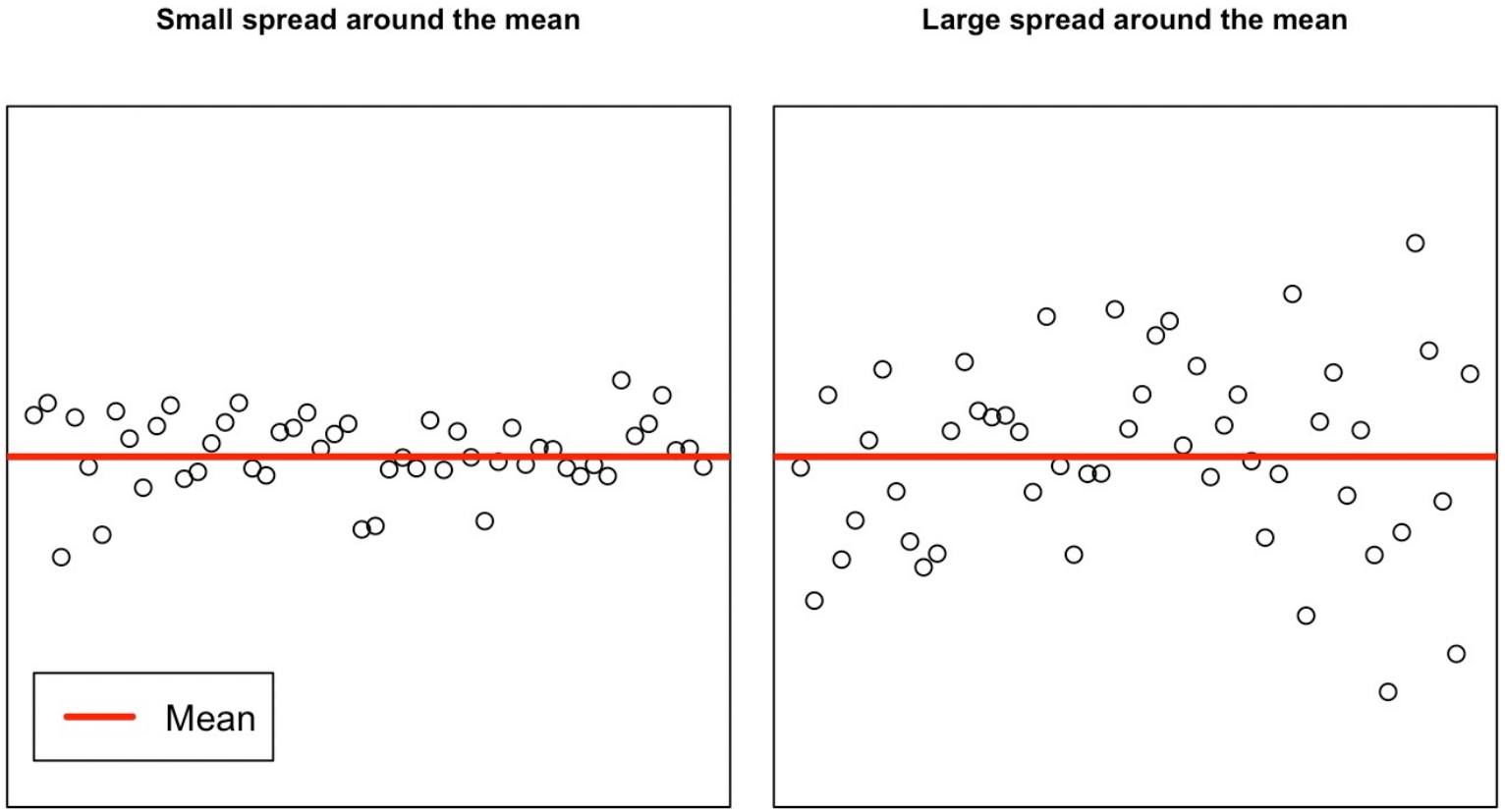
Sample Variance and Standard Deviation

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

Can you spot the difference between Population Variance and SD and Sample Variance and SD?

What does Variance and Standard Deviation imply?


Small spread around the mean
Large spread around the mean
Mean

Graphical Statistics

Before you do anything with a data set,
look at it!

Graphical Statistics

By first looking at the data, we can find out:

- What probability distribution to use ?
- Statistical methods suitable for the given data
- Presence or absence of outliers;
- Presence or absence of heterogeneity;
- Existence of time trends and other patterns;
- Relation between two or several variables.

Ways to visualize Data

- Histograms,
- Stem-and-leaf plots,
- Boxplots,
- Time plots, and
- Scatter plots.

Histogram

- The area of each bar is proportional to the frequency in the interval.
- To construct a histogram, we split the range of data into equal intervals, “bins,” and count how many observations fall into each bin.
- A frequency histogram consists of columns, one for each bin, whose height is determined by the number of observations in the bin.
- A relative frequency histogram has the same shape but a different vertical scale. Its column heights represent the proportion of all data that appeared in each bin.

Histogram Example – CPU times

9	15	19	22	24	25	30	**34**	35	35
36	36	37	38	42	43	46	48	54	55
56	56	**59**	62	69	70	82	82	89	139

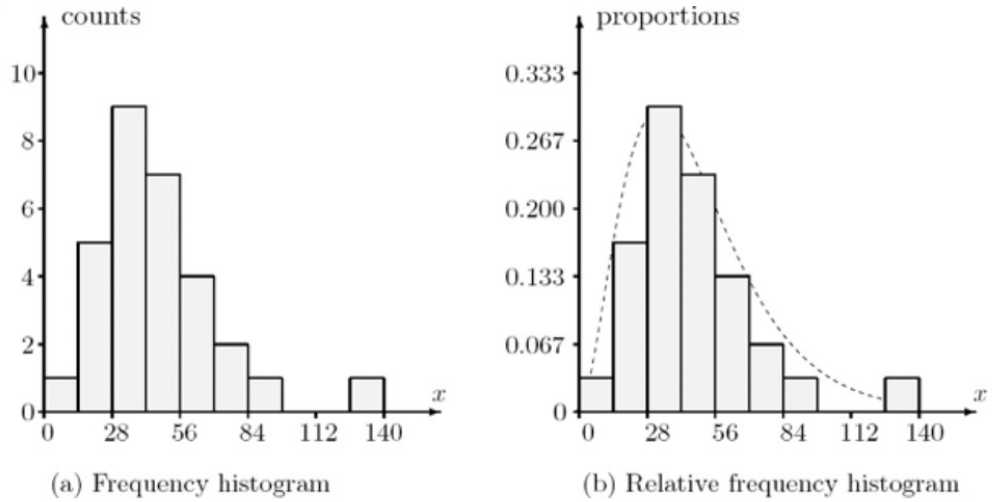
Histogram Example

Sample of CPU times

1	observation	between	0	and	14
5	observations	〃	14	〃	28
9	〃	〃	28	〃	42
7	〃	〃	42	〃	56
4	〃	〃	56	〃	70

. . . . . .

Histogram Example



(a) Frequency histogram

(b) Relative frequency histogram

What information can we get from a Histogram?

- Skewness
- Outliers
- Type of curve

Stem and leaf plot

- “Stem and Leaf diagrams give the shape of a distribution in the same way as a histogram with equal intervals , but also shows and keeps the details.”
- Stem and leaf plot can help you find the median and quartiles very easily.
- To construct a stem-and-leaf plot, we need to draw a stem and a leaf. The first one or several digits form a stem, and the next digit forms a leaf. Other digits are dropped; in other words, the numbers get rounded.

Stem and leaf plot for CPU times

Leaf unit = 1

0	9
1	5 9
2	2 4 5
3	0 4 5 5 6 6 7 8
4	2 3 6 8
5	4 5 6 6 9
6	2 9
7	0
8	2 2 9
9	
10	
11	
12	
13	9

Boxplot

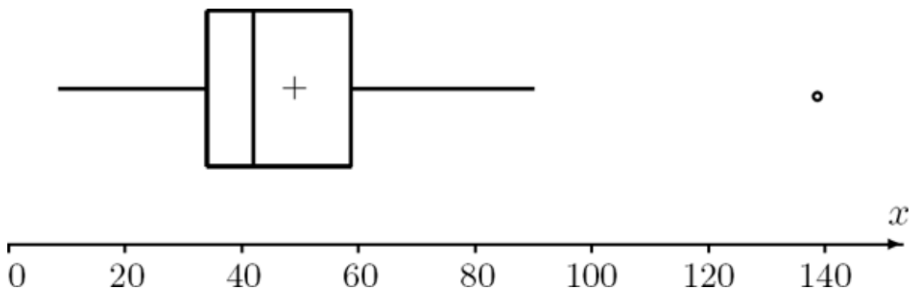
- Boxplot shows 5 points from the distribution – the top, and bottom of the range, the median and the upper and lower quartiles (UQ and LQ)
- To construct a boxplot, we draw a box between the first and the third quartiles, a line inside a box for a median, and extend whiskers to the smallest and the largest observations, thus representing a so-called five-point summary:

$$\text{five} - \text{point summary} = \left(\min X_i, \widehat{Q}_1, \widehat{M}, \widehat{Q}_3, \max X_i\right).$$

Boxplot Example – CPU times

- Mean and 5-point summary of the CPU times is given by:

$$\overline{X} = 48.2333; \min X_i = 9, \widehat{Q}_1 = 34, \widehat{M} = 42.5, \widehat{Q}_3 = 59, \max X_i = 139.$$

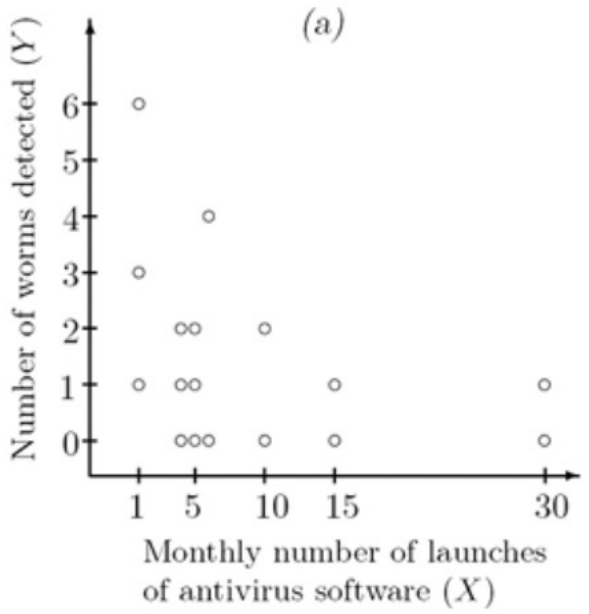


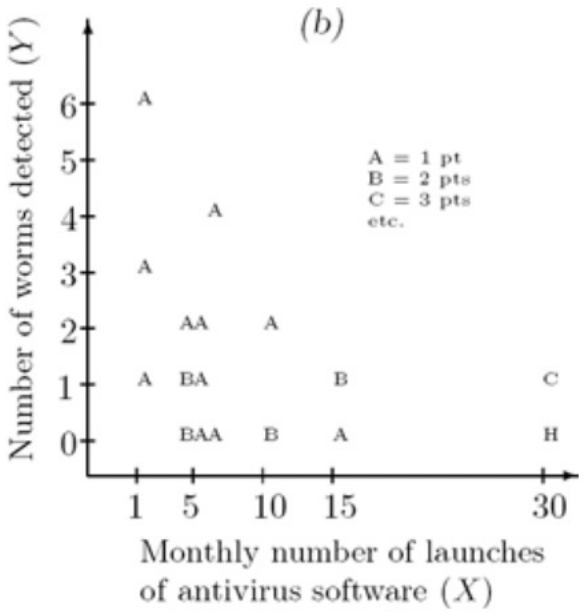
Boxplot of CPU time data.

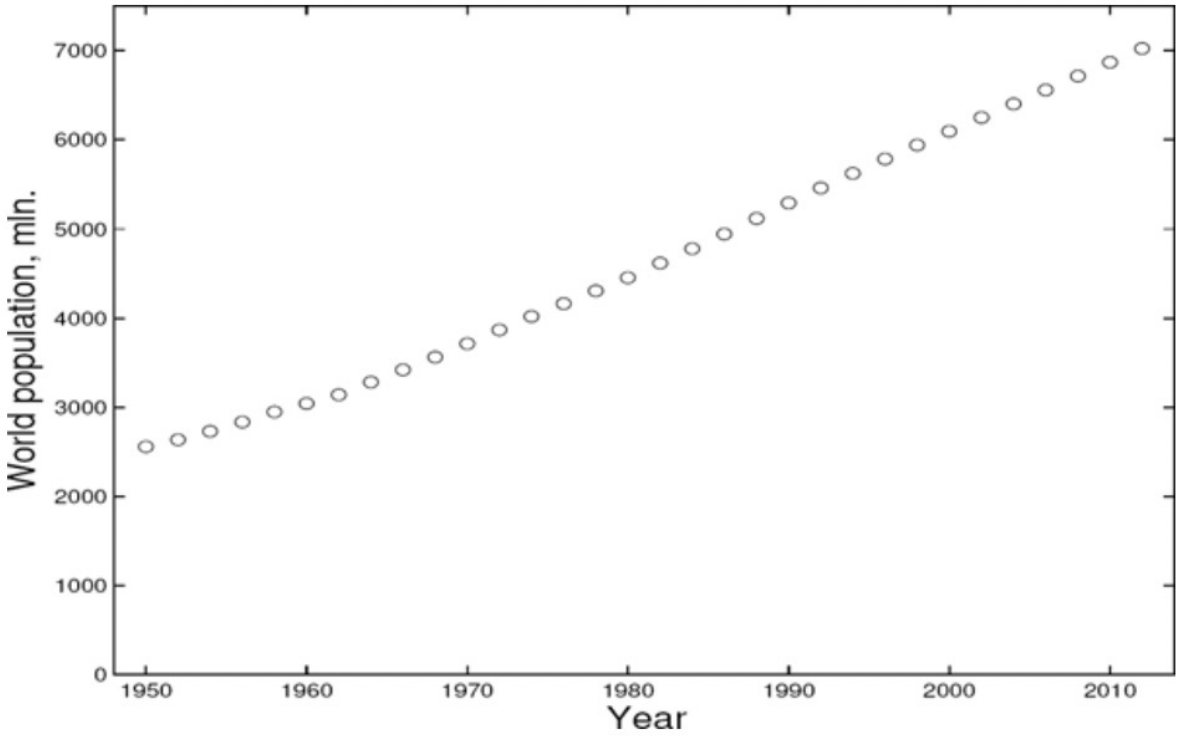
- From this boxplot, one can conclude:
- The distribution of CPU times is right skewed because (1) the mean exceeds the median, and (2) the right half of the box is larger than the left half.
- Each half of a box and each whisker represents approximately 25% of the population. For example, we expect about 25% of all CPU times to fall between 42.5 and 59 seconds.

Scatter plot and Time plot

- A **scatter plot** consists of n points on an (x, y)-plane, with x- and y-coordinates representing the two recorded variables.
- It is used to check the relationship between 2 variables.
- When we study time trends and development of variables over time, we use **time plots**. These are scatter plots with x-variable representing time.


(a)
Number of worms detected (Y)
0
1
2
3
4
5
6
1
5
10
15
30
Monthly number of launches of antivirus software (X)


(b)
Number of worms detected (Y)
A = 1 pt
B = 2 pts
C = 3 pts
etc.
A
A
AA
A
A
BA
B
C
BAA
B
A
H
1
5
10
15
30
Monthly number of launches of antivirus software (X)



Time plot of the world population in 1950–2012.

Any questions?