Exploring Data with Statistics

By Joe Ganser

What is statistics?

What is probability?

Statistics versus probabilities

Descriptive vs inferential statistics

measures of central tendency

measures of variability

frequency distribution

What is statistics?

Statistics is the mathematical science of drawing conclusions about the world using data, where our conclusions have a level of certainty relative to the quantity/quality of our data and analytical methods.

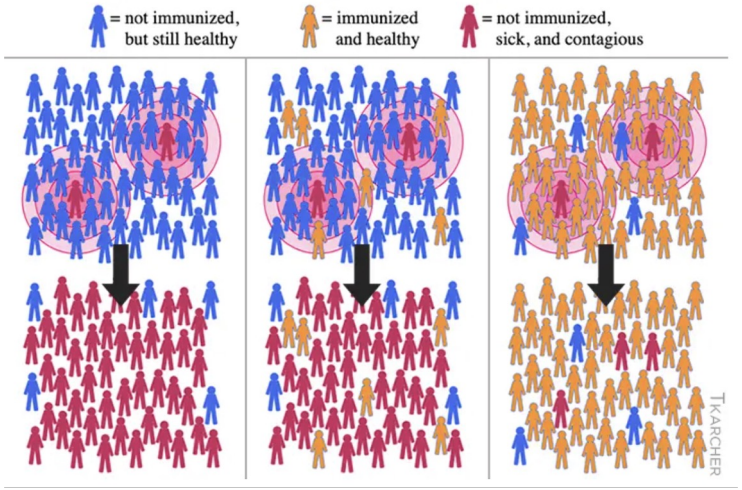
Statistics applications

- Describe the average age, income and purchase behavior of a group of customers

Source [3]

Statistics applications

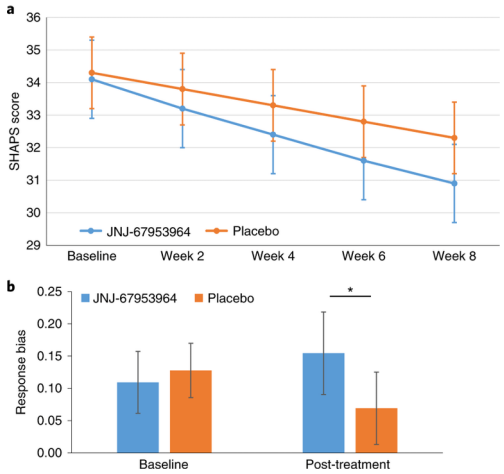
- Predict when herd immunity will occur from a pandemic



Source [4]]

Statistics applications

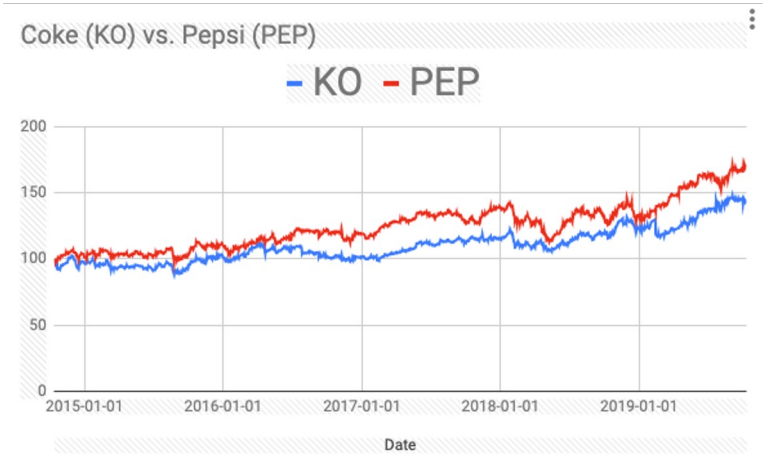
- Determine if a drug had any effect better than placebo



Source [5]

Statistics applications

- Identify stock trends and correlations - when one market rises, which one falls?



Source [6]

Statistics applications

- Take a random sample of people - use it to draw conclusions about their greater population

Goals of statistics

We use math to…

- Summarize properties of data (e.g. mean, number data points)

Goals of statistics

We use math to…

- Summarize properties of data
- Describe anomalies and patterns of data (e.g. find outliers)

Goals of statistics

We use math to…

- Summarize properties of data
- Describe anomalies and patterns of data
- Test hypothesis about data (did the drug lower blood pressure?)

Goals of statistics

We use math to…

- Summarize properties of data
- Describe anomalies and patterns of data
- Test hypothesis about data
- Identify patterns & relationships between variables

Goals of statistics

We use math to…

- Summarize properties of data
- Describe anomalies and patterns of data
- Test hypothesis about data
- Identify patterns & relationships between variables
- Draw population conclusions using a sample

Goals of statistics

We use math to…

- Summarize properties of data
- Describe anomalies and patterns of data
- Test hypothesis about data
- Identify patterns & relationships between variables
- Draw population conclusions using a sample
- Predict the long term results of a statistical trend

What is probability?

- Probability is a mathematical way of describing how likely some event is to happen

What is probability?

- Probability is a mathematical way of describing how likely some event is to happen
- Probability p; $0<=p<=1$

What is probability?

- Probability is a mathematical way of describing how likely some event is to happen
- Probability p; $0<=p<=1$
- We use probability to describe our statistical results

What is probability?

- Probability is a mathematical way of describing how likely some event is to happen
- Probability p; 0<=p<=1
- We use probability to describe our statistical results
- P(something specific) = #number of ways specific result can happen/ number of all possibilities

What’s the probability a single die rolls an even number?

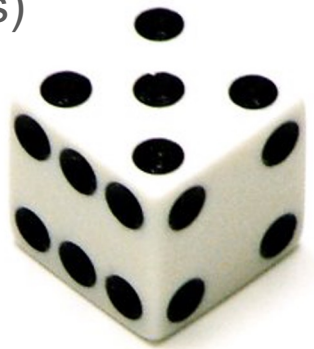
What we’re looking for = [2,4,6]

All possibilities = [1,2,3,4,5,6]

What's the probability a single die rolls an even number?

What we're looking for = [2,4,6] (3 possibilities)

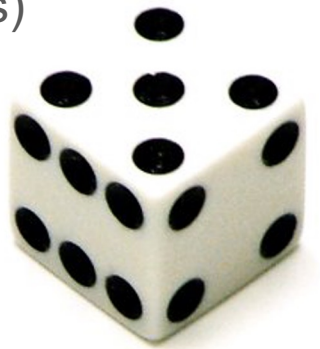
All possibilities = [1,2,3,4,5,6] (6 possibilities)



What’s the probability a single die rolls an even number?

What we’re looking for = [2,4,6] (3 possibilities)

All possibilities = [1,2,3,4,5,6] (6 possibilities)



P = 3/6 = 0.5

Probability & Statistics

We use probability to describe the results of statistics;

E.g.

- “95% chance the weight of a GMO fish in a special pond is between 1-1.3lbs”
- “1% chance we observed these measurements due to pure coincidence”

Descriptive & Inferential statistics

- Two main genres of statistics
- Descriptive: goal is to summarize and visualize the data
  - Describe a dataset
  - Relies more on metrics and graphs

- Inferential: goal is to make inferences/generalizations about a population using a sample
  - Draw conclusions about a population
  - Relies more on probability

Descriptive Statistics: some applications

- Average age of a customer group
- Median (middle number): median age people who received a medical treatment
- Mode (most frequent data point): blood type of people who tested positive for covid
- Range: the spectrum of possible investment returns
- Standard deviation: measuring the variability of blood pressure measurements in treatment groups
- Histograms (frequency plots): The count of each age group that signed up for a gym membership

Measures of central tendency

Suppose we weigh a bunch of athletes before and after a weight training program. How do we describe these numbers?

Data = [100kg, 77kg, 93kg, 93kg, 115kg]

- Mean, median, mode

Measures of central tendency: Mean

Data = [100kg, 77kg, 93kg, 93kg, 115kg]

$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

Measures of central tendency: Mean

Data = [100kg, 77kg, 93kg, 93kg, 115kg]

Mean = (100+77+93+93+115)/5 = 95.6

Measures of central tendency: Median

Data = [100kg, 77kg, 93kg, 93kg, 115kg]

First sort the data in increasing order, then use median formula

Data = [77kg, 93kg, 93kg,100kg,115kg]

$$\text{Median} = \begin{cases} \frac{(N+1)^{th}}{2} \text{ term; when N is odd} \\ \frac{\frac{N^{th}}{2} \text{ term} + \left(\frac{N}{2}+1\right)\text{term}}{2} \text{ ;when N is even} \end{cases}$$

Measures of central tendency: Median

Data = [77kg, 93kg, 93kg, 100kg,115kg]

First sort the data in increasing order, then use median formula

Median = 93kg

Source [1]

Mode (most frequent value)

Data = [77kg, 93kg, 93kg, 100kg,115kg]

1. find a count of each value.
2. find the value with the highest count

Mode (most frequent value)

Data = [77kg, 93kg, 93kg, 100kg,115kg]

1. find a count of each value.
2. find the value with the highest count

Mode: 93kg

Data point count = {77:1, 93:2, 100:1, 115kg:1}

Central Tendency Measures: summary

$$\text{Mean } \bar{x} = \frac{\sum x_i}{N}$$

$$\text{Median} = \begin{cases} \frac{(N+1)^{th}}{2} \text{ term; when N is odd} \\ \frac{\frac{N^{th}}{2} \text{ term} + \left(\frac{N}{2} + 1\right) \text{term}}{2} \text{ ; when N is even} \end{cases}$$

$$\text{Mode} = \text{The value in the data set that occurs most frequently}$$

Source [1]

Variability Measures

Suppose our goal is to describe how our dataset varies. We can use the;

- Range
- Standard deviation/Variance

Variability Measures: Range

To find the range, we simply look at the min value, the max value and their difference.

Variability Measures: Range

To find the range, we simply look at the min value, the max value and their difference.

Data = [77kg, 93kg, 93kg, 100kg,115kg]

Variability Measures: Range

To find the range, we simply look at the min value, the max value and their difference.

Data = [77kg, 93kg, 93kg, 100kg,115kg]

Min = 77kg

Max = 115kg

range = [77kg, 115kg]

Diff = 38kg

Variability Measures: Standard deviation/Variance formula

	Sample	Population
Standard deviation	s	σ
Variance	s^2	σ^2
Datapoint	x_i	x_i
Average	$\bar{x}$	μ
Total dataset number	n	N
Standard deviation formula	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{\boldsymbol{n-1}}}$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
Variance formula	$s^2 = \frac{\sum(x_i - \bar{x})^2}{\boldsymbol{n-1}}$	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

Source [2]

Variance is just standard dev squared!

Variability Measures: Standard deviation/Variance

To find the variance, we

- subtract the mean from each value
- Square that new value
- Find the average of all these new data points

	Sample	Population
Standard deviation	s	σ
Variance	s^2	σ^2
Datapoint	x_i	x_i
Average	$\bar{x}$	μ
Total dataset number	n	N
Standard deviation formula	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{\boldsymbol{n-1}}}$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
Variance formula	$s^2 = \frac{\sum(x_i - \bar{x})^2}{\boldsymbol{n-1}}$	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

Variability Measures: Variance/standard dev

Find the variance

Data = [77kg, 93kg, 93kg, 100kg,115kg]

Mean = 95.6

Var = ((77-95.6)^2+(93-95.6)^2+(93-95.6)^2+(100-95.6)^2+(115-95.6)^2)/5

Var = 151.04

Std = Var^(0.5) = 12.29

Interquartile range

Data = [7, 9, 11, 12, 13, 14, 14, 15, 19]

We have an odd number of elements - what's our median?

[7, 9, 11, 12, 13, 14, 14, 15, 19]

Median is 13

Interquartile range

Data = [7, 9, 11, 12, 13, 14, 14, 15, 19]

[7, 9, 11, 12, 13, 14, 14, 15, 19]

Bottom half = [7,9,11,12]

Bottom median = (9+11)/2 = 10

Q1 = 10

Interquartile range

Data = [7, 9, 11, 12, 13, 14, 14, 15, 19]

[7, 9, 11, 12, 13, 14, 14, 15, 19]

Top half = [14, 14, 15, 19]

top median = (14+15)/2 = 14.5

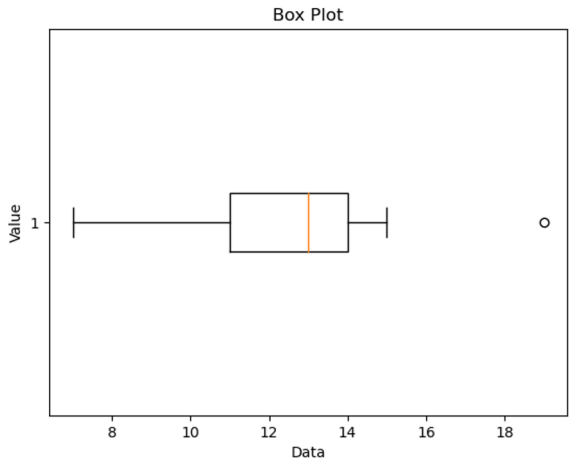
Q3 = 14.5

Interquartile range

Data = [7, 9, 11, 12, 13, 14, 14, 15, 19]

[7, 9, 11, 12, 13, 14, 14, 15, 19]

Interquartile range = Q3 - Q1 = 14.5 - 10 = 4.5



Histograms and Frequency distributions

We previously learned that the mode of a dataset is simply the most frequently observed number

Data = [1,2,1,1,1,4,5,6]

Observations = {1:4,2:1,4:1,5:1,6:1} {number: count}

Mode = 1

Histograms and Frequency distributions

Histograms are convenient ways of representing the count of each data point, or of each data group

Consider a (sorted) dataset:

Data = [2, 9, 14, 25, 27, 28, 35, 37, 46, 47, 47, 49, 52, 54, 63, 76, 87, 87, 91, 98]

Histograms and Frequency distributions

Data = [2, 9, 14, 25, 27, 28, 35, 37, 46, 47, 47, 49, 52, 54, 63, 76, 87, 87, 91, 98]

Now suppose we made bins;

bin	count
(2,26)	4
(26,50)	8
(50,74)	3
(74,98)	5

Histograms and Frequency distributions

Data = [2, 9, 14, 25, 27, 28, 35, 37, 46, 47, 47, 49, 52, 54, 63, 76, 87, 87, 91, 98]

And we can also make percentages

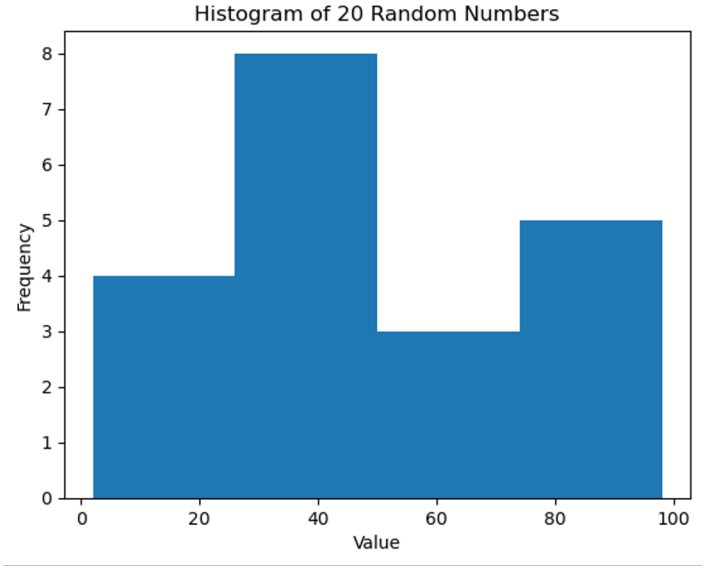
bin	count	%
(2,26)	4	20
(26,50)	8	40
(50,74)	3	15
(74,98)	5	25

Histograms and Frequency distributions

Data = [2, 9, 14, 25, 27, 28, 35, 37, 46, 47, 47, 49, 52, 54, 63, 76, 87, 87, 91, 98]

Now we use this to make a histogram

bin	count	%
(2,26)	4	20
(26,50)	8	40
(50,74)	3	15
(74,98)	5	25



Sources

1. https://dataanalyze.wordpress.com/a/
2. https://s4be.cochrane.org/blog/2022/02/24/using-measures-of-variability-to-inspect-homogeneity-of-a-sample-part-1/
3. https://fashinza.com/textile/tips-for-fashion-brands/demographics-and-psychographics-how-to-know-your-customers-inside-out/
4. https://www.medpagetoday.com/infectiousdisease/covid19/88401
5. https://www.researchgate.net/figure/SHAPS-and-PRT-results-a-Effects-of-study-drug-versus-placebo-on-mean-SHAPS-score-ITT_fig4_340287074
6. https://www.ngpf.org/blog/question-of-the-day/qod-coke-vs-pepsi-which-companys-stock-has-performed-better-over-the-past-five-years/
7. https://datatab.net/tutorial/descriptive-inferential-statistics