INTRO TO STATISTICS

PREPARING THE GROUND FOR A/B TESTING

WBS CODING SCHOOL

Populations & samples

Population & Samples

What would be the population and the sample for the A/B Test on Eniac's site?

My population

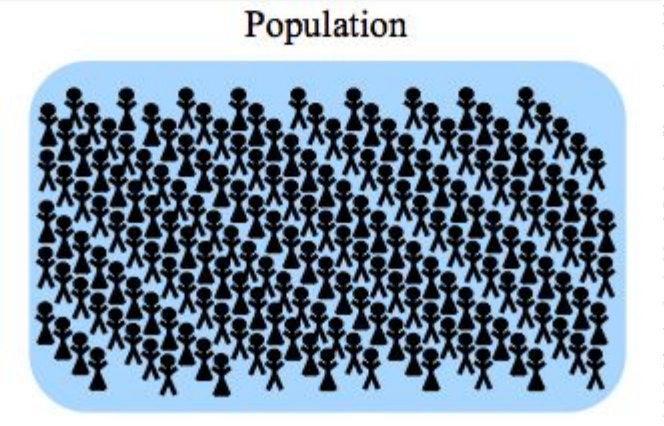
All my potential customers

N

A sample of the population

A group of actual visitors to my site during a certain period of time

n



Why sampling?

Impossible to survey the entire population

Population size is so big that the data cannot be properly explored

Computationally too expensive to use the whole population

Easier to ensure data quality (cleaning outliers, missing values…)

Replacement

In a sample **with replacement**, observations are “put back” into the population after being sampled, and they can be sampled again.

In a sample **without replacement**, observations are unavailable for future draws once selected.

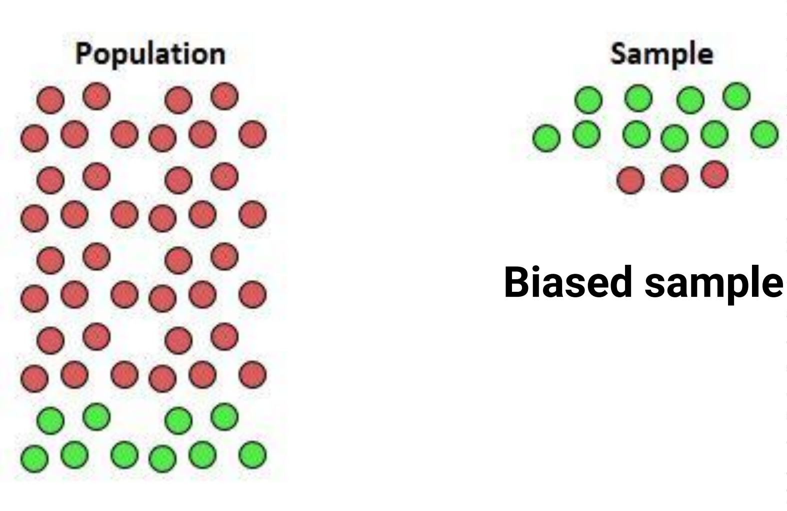
Bias

Biased sample

A sample with a **systematic error**: one or more parts of the population are favored over others to become part of the sample

Unbiased sample

A good sample is the one that is representative of the entire population

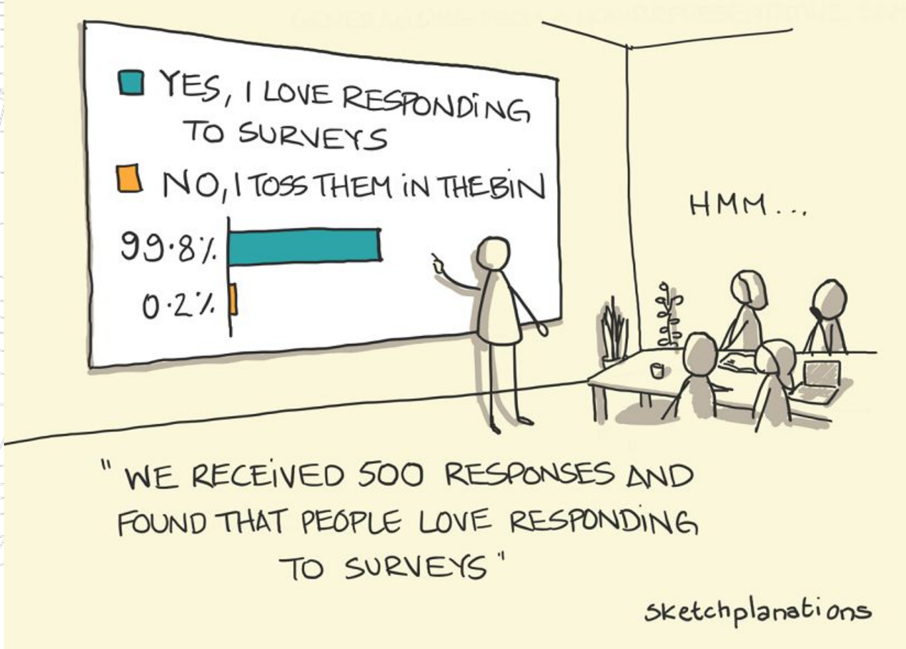


https://fourpillarfreedom.com/wp-content/uploads/2019/05/nonResponseBias.jpg

Non-response or self-selection bias

Let's say I send a survey to my customers, titled **¿Which color do you like the most?**

Certain people were more inclined to answer the survey: these users might have preferences that are not representative of the whole user base.



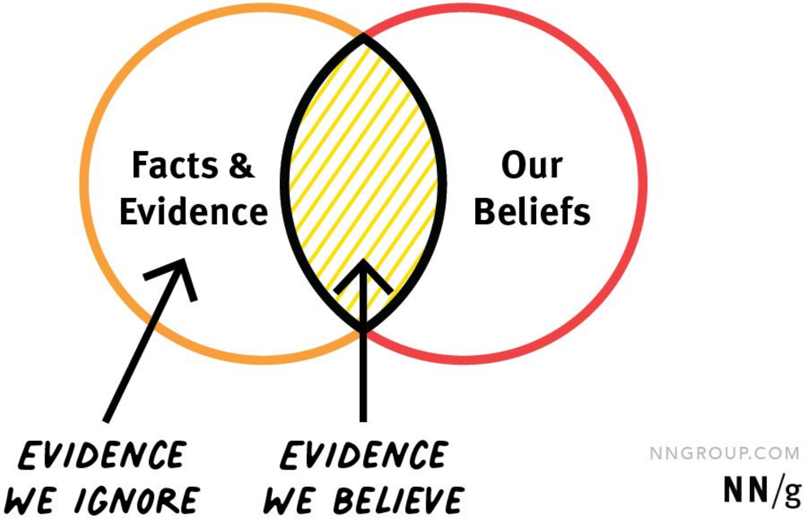
🟦 YES, I LOVE RESPONDING TO SURVEYS
🟧 NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"
sketchplanations

Anecdotal evidence

Let's say I directly ask feedback from 17 users and they all like color red.

While this information is true (evidence), it has been acquired in an informal way, relying purely on personal testimony. Anecdotal evidence is evidence collected in a casual or informal manner and relying heavily or entirely on personal testimony.

Whenever you hear someone saying something like "This is true! I have many friends who say that..." they are basing a claim on anecdotal evidence and you have the right to point it out.



Sample / Population in statistics notation

N	Population size
μ	(“mu”) - Population mean
σ	(“sigma”) Population standard deviation

n	Sample size
$\bar{x}$	“X-bar” - Sample mean
s	Sample standard deviation

Experiments

What is an experiment?

A procedure that can be infinitely repeated with defined possible outcomes.

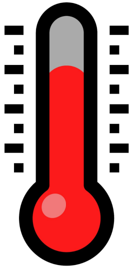
- Random → more than one outcome
- Deterministic → only one outcome

Used to make a discovery, test a hypothesis, or demonstrate a known fact.

Controlling

How will the sun affect the plant growth?

25ºC



Once per week

Controlling means to establish a **common environment** to all participants in an experiment to prevent extraneous variables (variables that are not the focus of the study) to affect the outcome.

Randomization

Randomizing the observations that go to the control group is a way of dealing with variables that cannot be controlled.

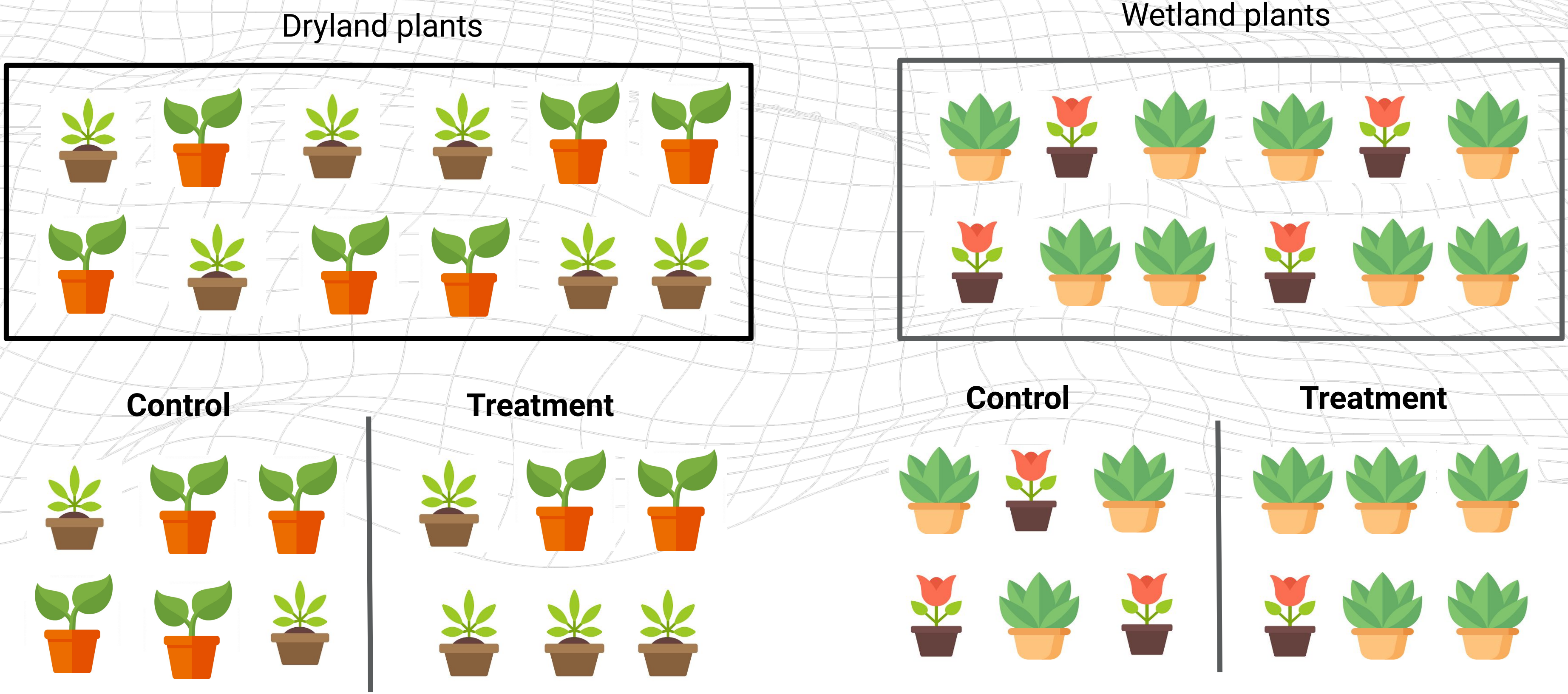
Replication

n =20

n =6

Can I repeat the experiment with a similar result?

Blocking



Dryland plants

Wetland plants

Control

Treatment

Control

Treatment

Measures of Central Tendency

The mean, the median and the mode

[1, 2, 2, 4, 6, 8, 50, 1000]

Mean = sum(1, 2, 2, 4, 6, 8, 50, 1000) / 8 = $\boxed{130}$

Median [1, 2, 2, $\boxed{4, 6}$, 8, 50, 1000] = 5

Mode [1, $\boxed{2, 2,}$ 4, 6, 8, 50] = 2

Measures of Dispersion

Key terms

Deviations (errors, residuals): Difference between observed values and an estimate.

Variance: The sum of squared deviations from the mean divided by n - 1.

Standard deviation: Square root of the variance.

Range: Difference between the largest and the smallest value in a data set.

Percentile: The value such that P percent of the values take on this value or less.

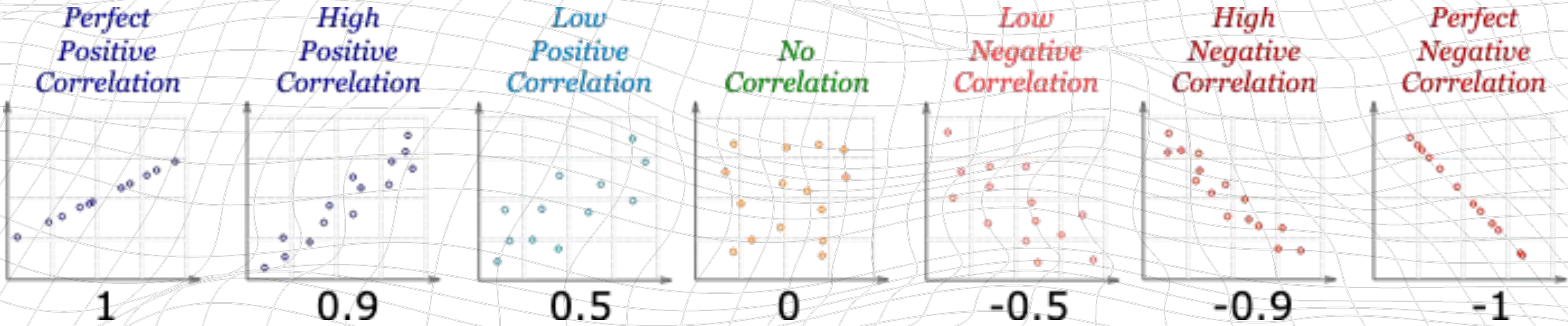
Interquartile range: The difference between the 75th percentile and the 25th percentile.

Relationships between variables

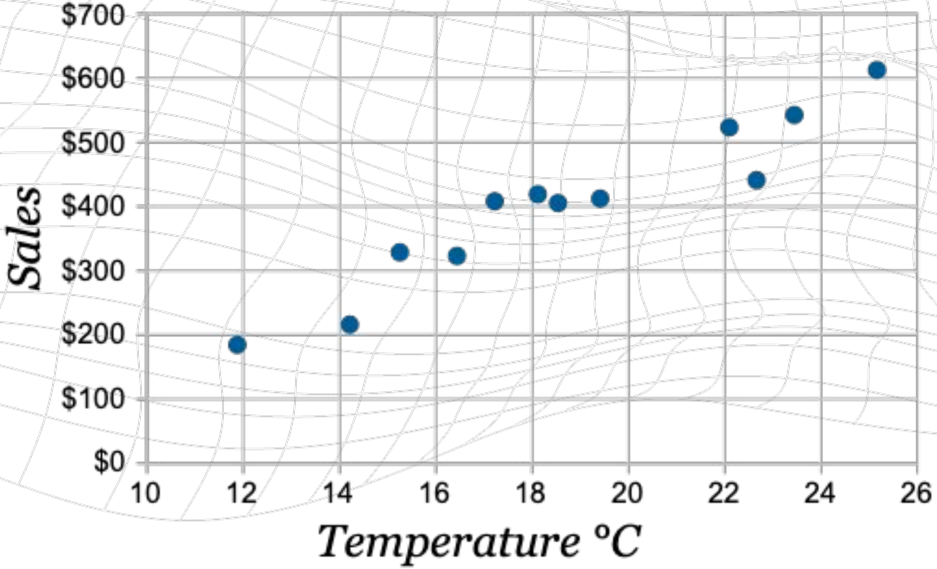
Linear Correlation

Two variables are **positively correlated** when they both increase together. If when one variable increases, the other decreases, they are **negatively correlated.**

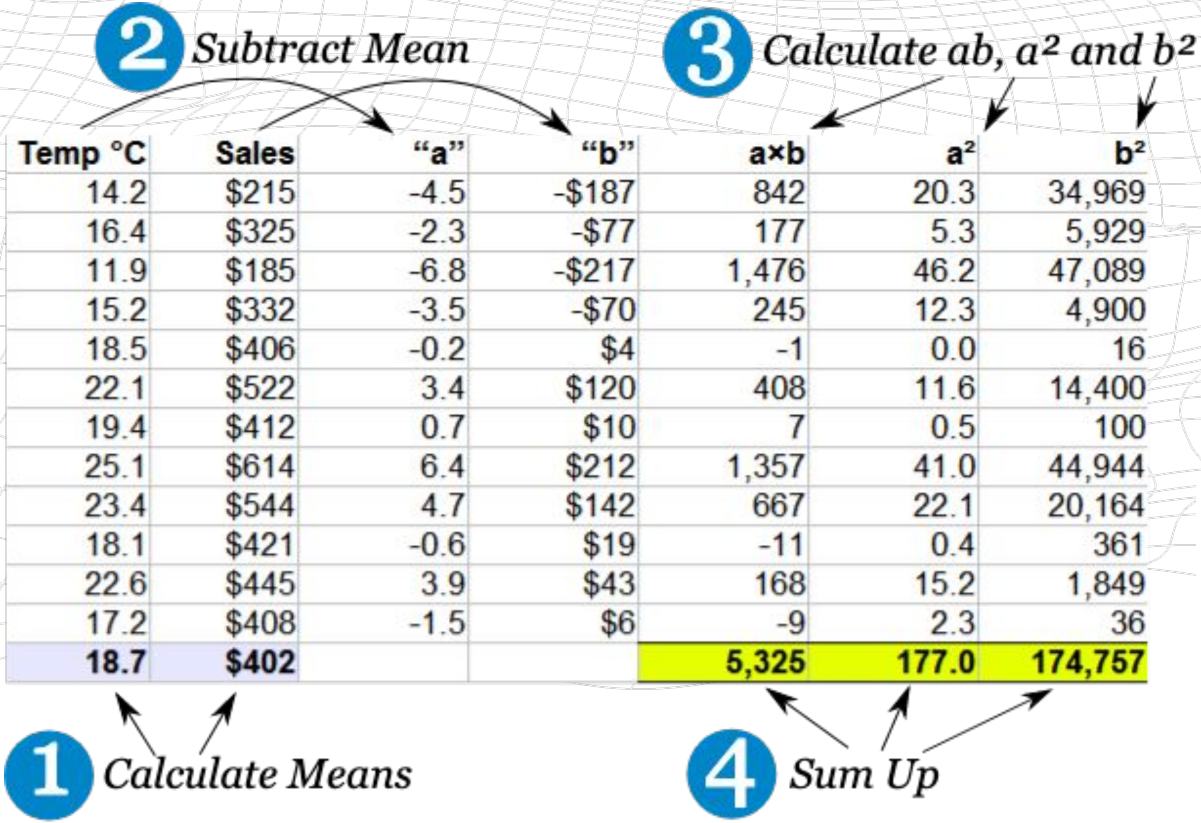
Linear Correlation



Ice-cream sales & temperature



5 $\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$



2 *Subtract Mean*

3 *Calculate ab, a² and b²*

Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

1 *Calculate Means*

4 *Sum Up*

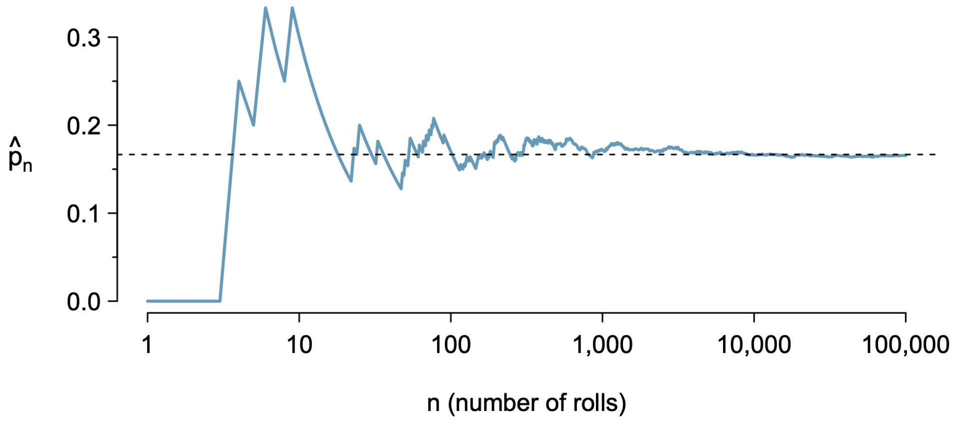
Pearson's correlation coefficient

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

Probability

Events and their probability

For 1 die.The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times. It is expressed as a number between 0 and 1.



Probability of rolling a 1 → P(rolling a 1) → P(1)

Random variables

$x + 3 = 7$ → Not a random variable

$X = \begin{cases} 1 \text{ if it rains} \\ 0 \text{ if it doesn't rain} \end{cases}$ → Random variable: an outcome expressed numerically.

Types of random variables

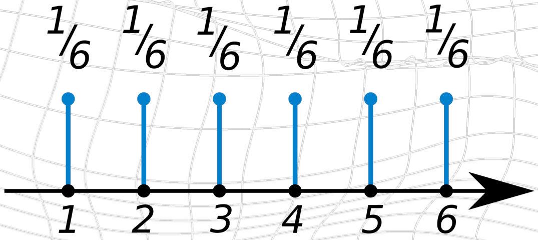
Discrete variables	Number of clicks
Continuous variables	Viewing time of a movie Amount of money spent

Probability distributions

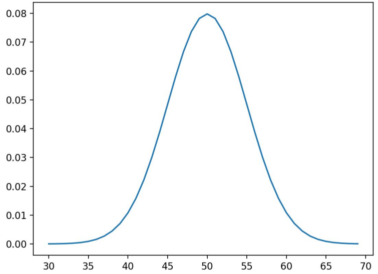
Probability distributions

- Mathematical function that gives the probabilities of occurrence of different possible outcomes

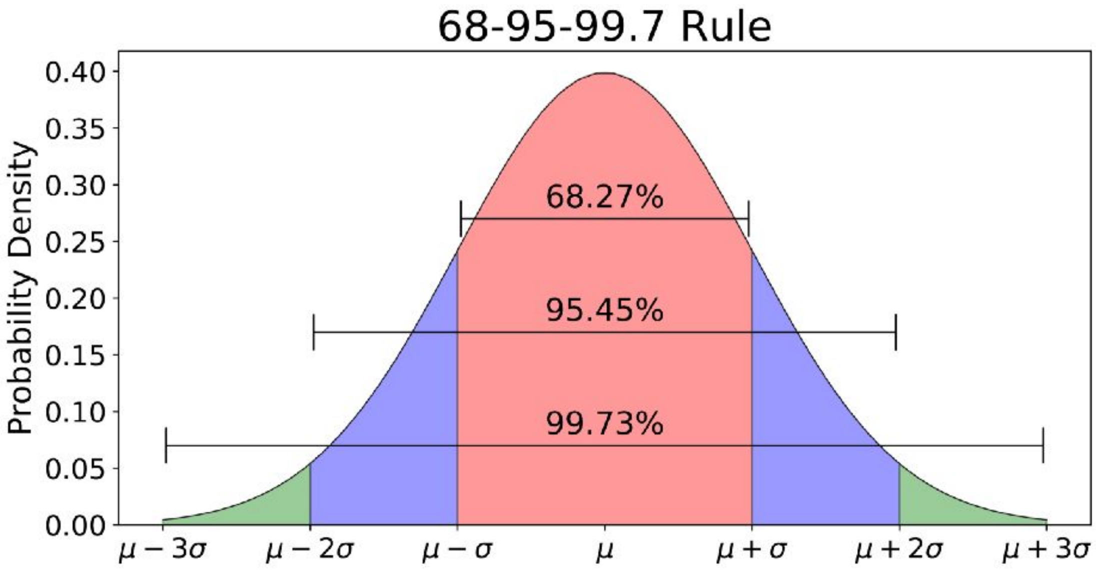
Discrete →



Continuous →



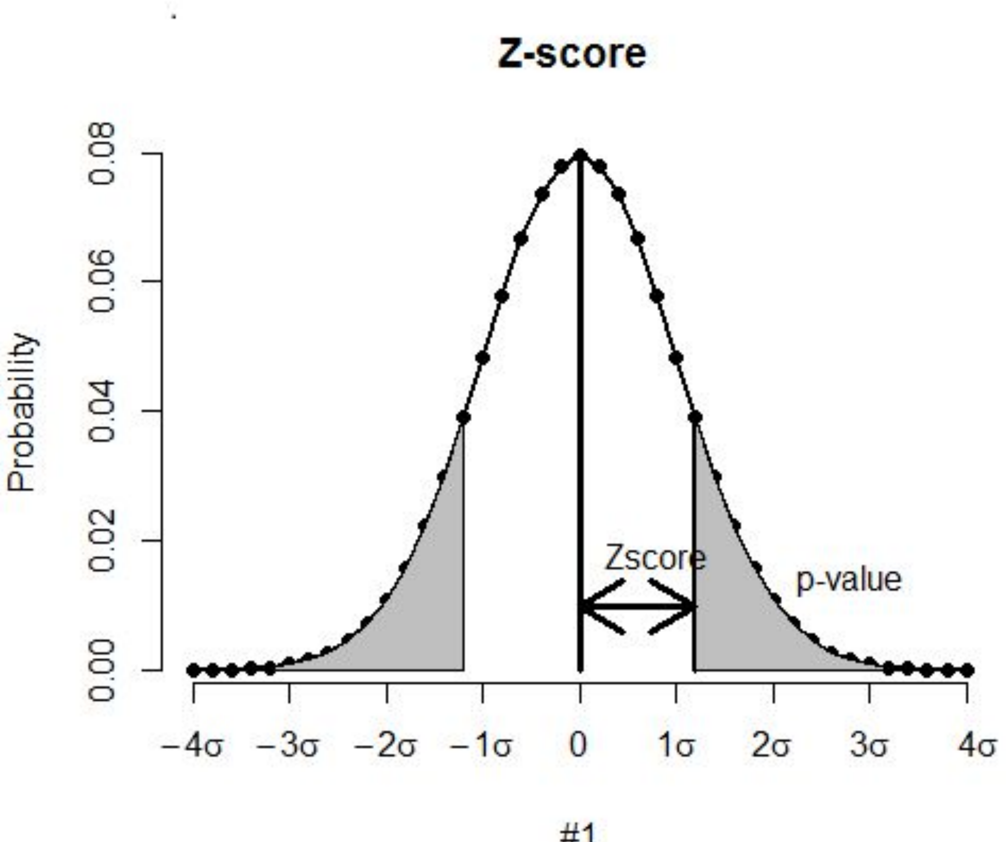
The normal distribution



The normal distribution: Z values

How many standard deviations does A differ from the mean?

$$Z = \frac{X - \mu}{\sigma}$$


Z-score
Probability
Zscore
p-value
−4σ
−3σ
−2σ
−1σ
1σ
2σ
3σ
4σ
#1