CDSF04, CDSF05, CDSF06

Recap II

DS Academy

TOTVS

2018


PYTHON
is the new EXCEL

Pandas and SQL

Main Features - Pandas

I/O	Files	pd.read_csv(), pd.read_json(), pd.read_pickle(), ...
	DB's and DW's	pd.read_sql(), pd.read_gbq(), ...
	Python Objects	pd.DataFrame.from_dict(), pd.DataFrame.from_records()
Indexing	Integer, Range, Slice	df[start:end:step], df.iloc[], df.loc[]
	List	df[[1,2,4,5]], s.reindex([1,2,3])
	Boolean Array	df[(df['col_a']=='foo') & df['col_b']=='bar')]
Transform	Groupby	df.groupby('col').sum(), .apply(), .agg(), .transform(), ...
	Pivot, Stack, Unstack	df.pivot(), df.stack(), df.unstack()
Merge	Merge	df1.merge(df2, left_on='pk', right_on='fk', how='left')
	Join	df1.join(df2, how='inner')
	Concat	pd.concat([df1, df2, df3])

Table Of Contents

- What's New in 0.25.0
- Installation
- Getting started
- User Guide
  - IO tools (text, CSV, HDF5, …)
  - Indexing and selecting data
  - MultiIndex / advanced indexing
  - Merge, join, and concatenate
  - Reshaping and pivot tables
  - Working with text data
  - Working with missing data
  - Categorical data
  - Nullable integer data type
  - Visualization
  - Computational tools
  - Group By: split-apply-combine
  - Time series / date functionality
  - Time deltas
  - Styling
  - Options and settings
  - Enhancing performance
  - Sparse data structures
  - Frequently Asked Questions (FAQ)
  - Cookbook
- Pandas ecosystem
- API reference
- Development
- Release Notes

pandas.pydata.org/pandas-docs/stable/user_guide stack overflow

shift
+ tab

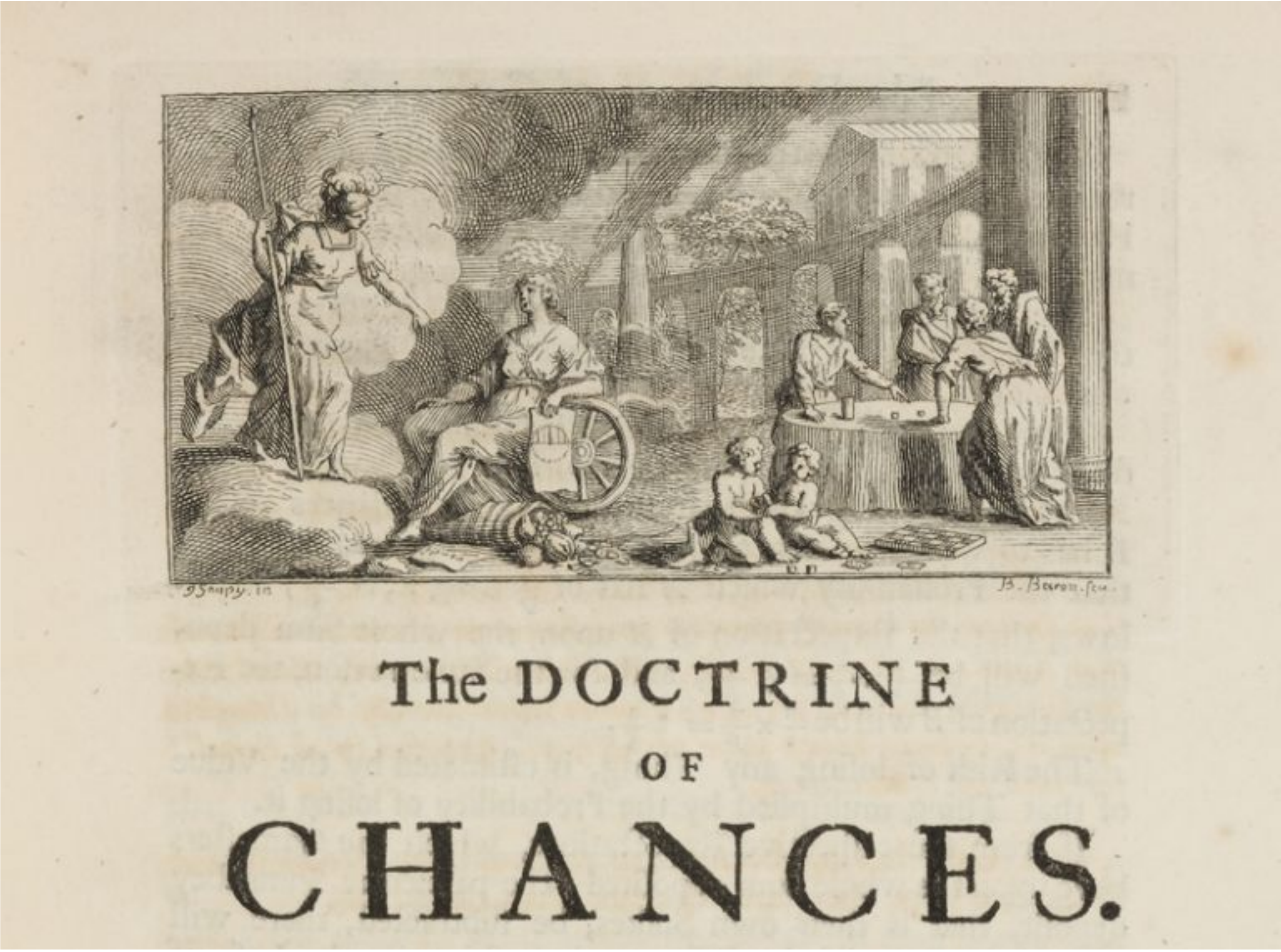
?pd.

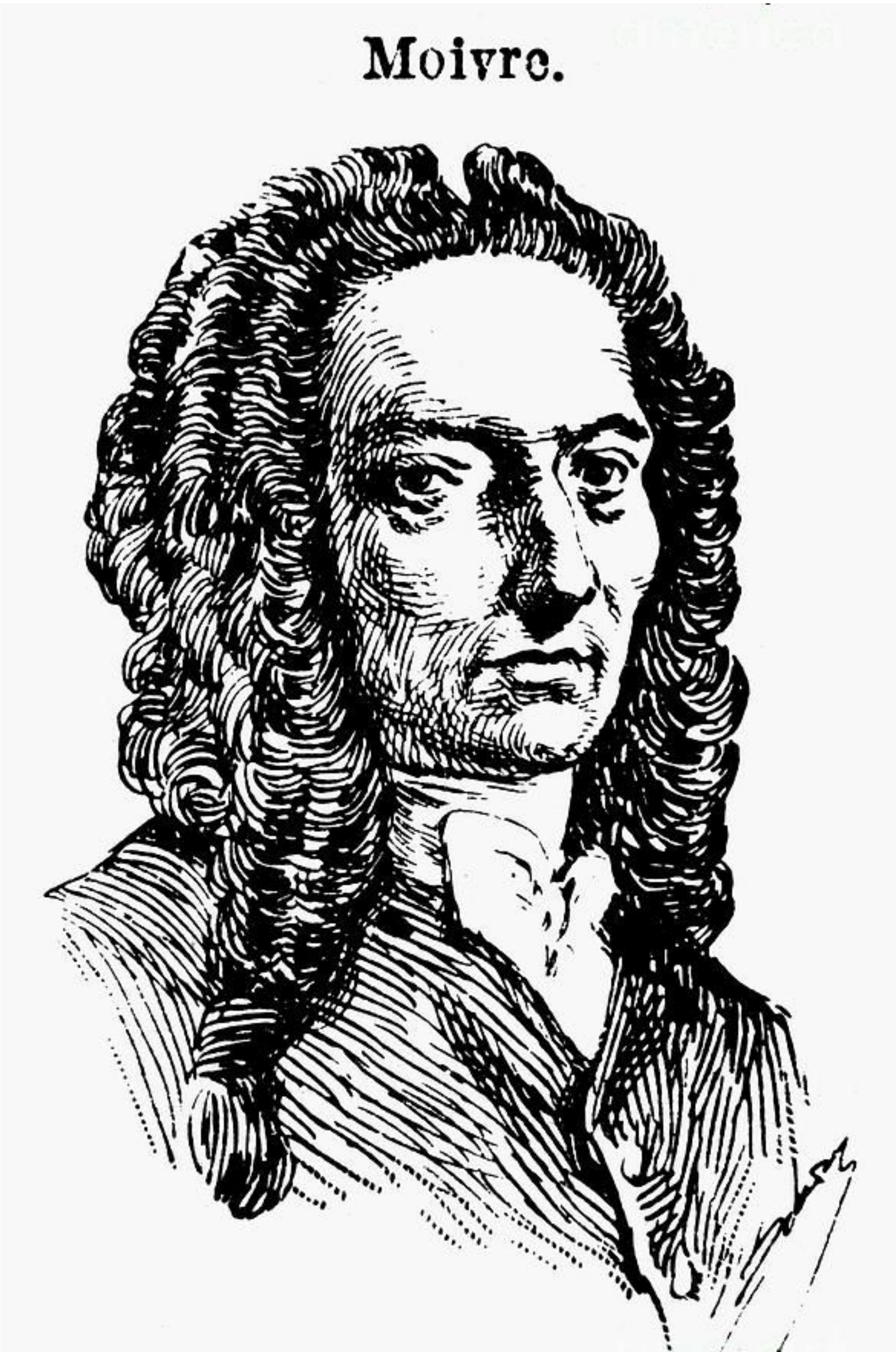
Probability Distributions and Frequentist Statistics

Agenda

- What is probability?
- Multiplication rule
- Independence
- Addition rule
- Mutually Exclusive
- PDF and CDF
- Mean, Median and Mode
- Standard error
- Variance
- The law of large numbers
- The Central Limit Theorem

What is Probability?





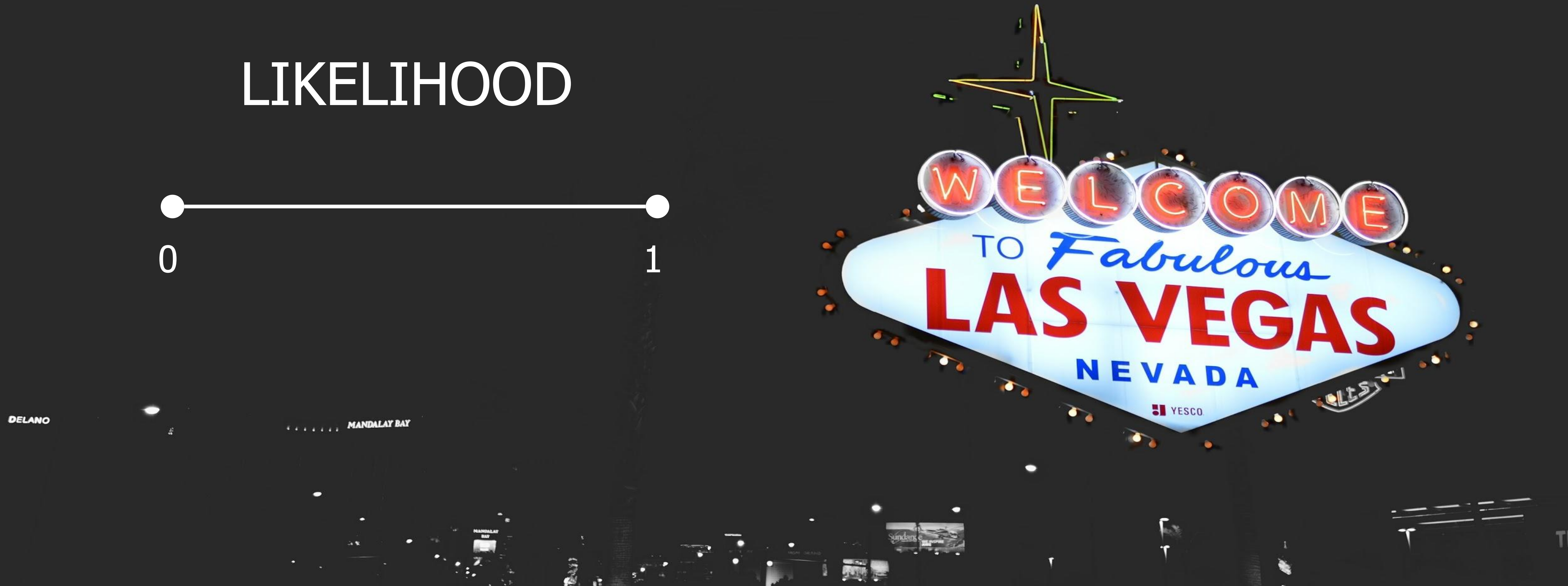
$

The frequency theory was originally designed to solve gambling problems

LIKELIHOOD

0

1



$$P(\text{event}) = \frac{\#\ \text{OCCURRENCES EVENT}}{\#\ \text{ALL POSSIBLE OUTCOMES}}$$

Example

$$P(throw=6) = \frac{1}{6}$$

$$P(\text{throw not } 6) = \frac{5}{6}$$

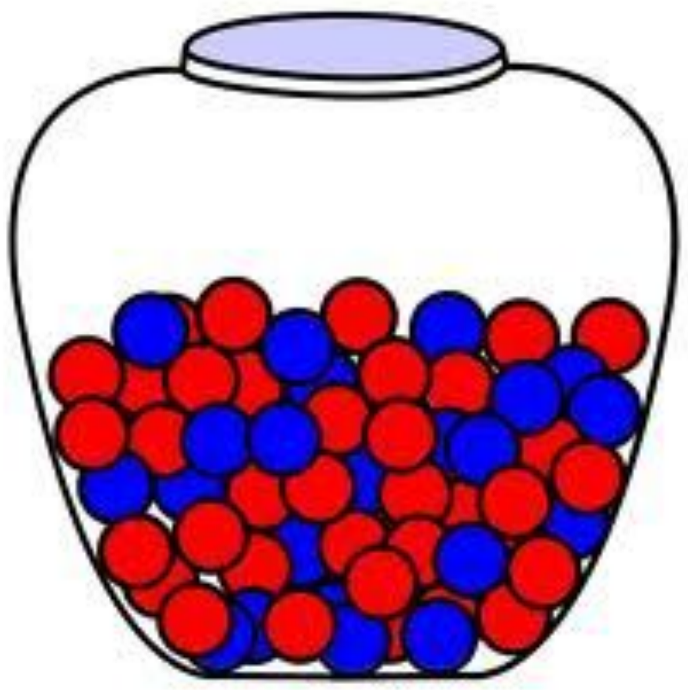
$$= 1 - P(\text{throw} = 6)$$

The chance that **both** two things will happen = The chance that the **first** thing will happen * The chance the **second** will happen

Two things are **independent** if the chances for the second, given the first are the same, no matter how the first one turns out. Otherwise, the two things are **dependent**.

WITH REPLACEMENT

INDEPENDENT



WITHOUT REPLACEMENT

DEPENDENT

Multiplication Rule

INDEPENDENDT

$$P(A \text{ and } B) = P(A) * P(B)$$

DEPENDENT

$$P(A \text{ and } B) = P(A) * P(B|A)$$

The chance that **both** two things will happen

The chance that the **first** thing will happen

The chance the **second** will happen

What is P(BBB) ?

WITH REPLACEMENT

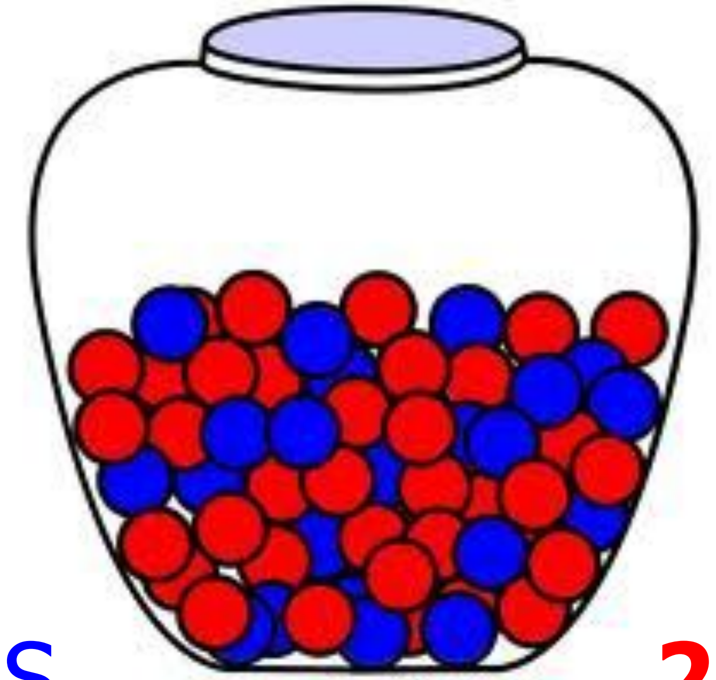
INDEPENDENT

30/55 * 30/55 * 30/55

WITHOUT REPLACEMENT

DEPENDENT

30/55 * 29/54 * 28/53



30 BLUES

25 RED

What is P(BRR) ?

WITH

REPLACEMENT

INDEPENDENT

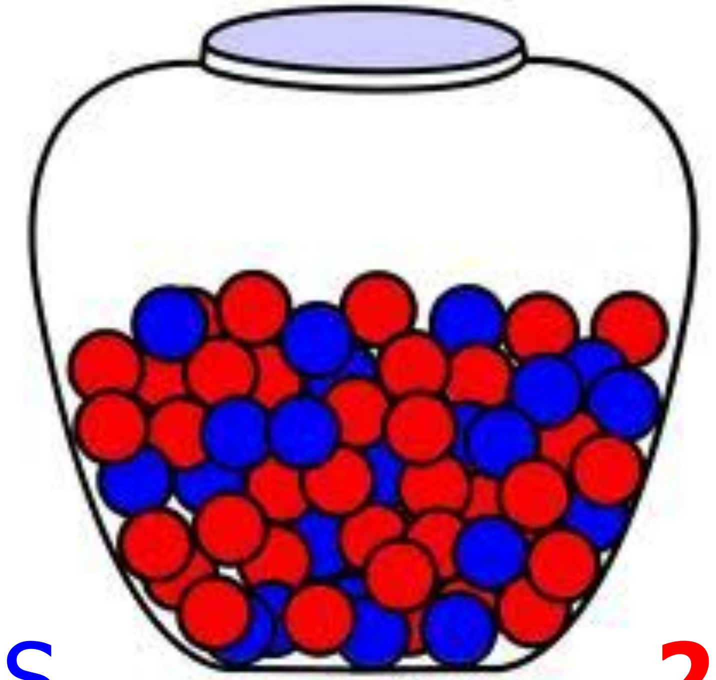
30/55 * 25/55 * 25/55

WITHOUT

REPLACEMENT

DEPENDENT

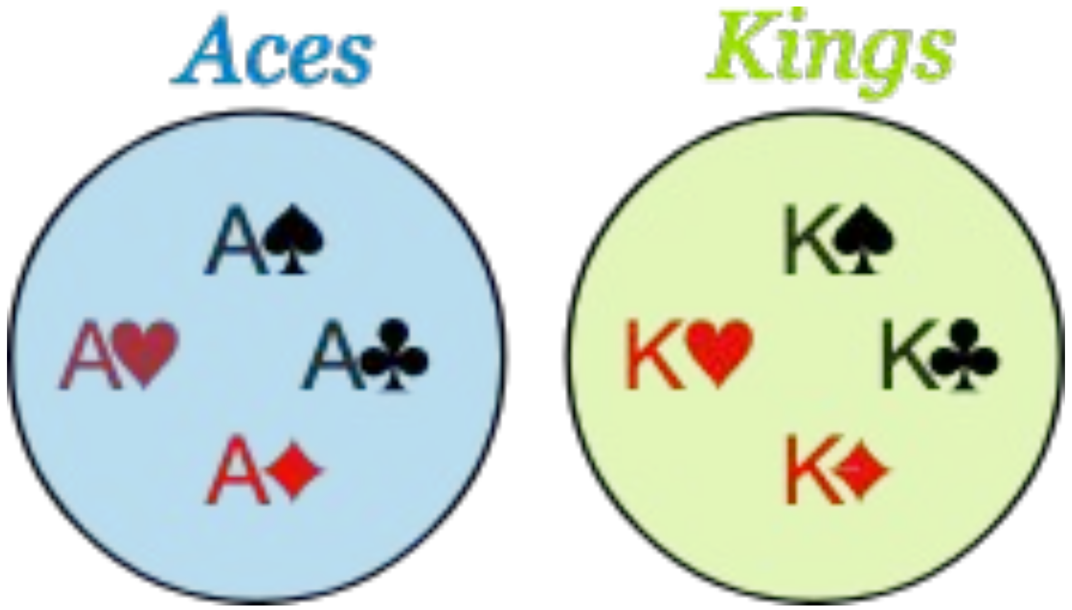
30/55 * 25/54 * 24/53

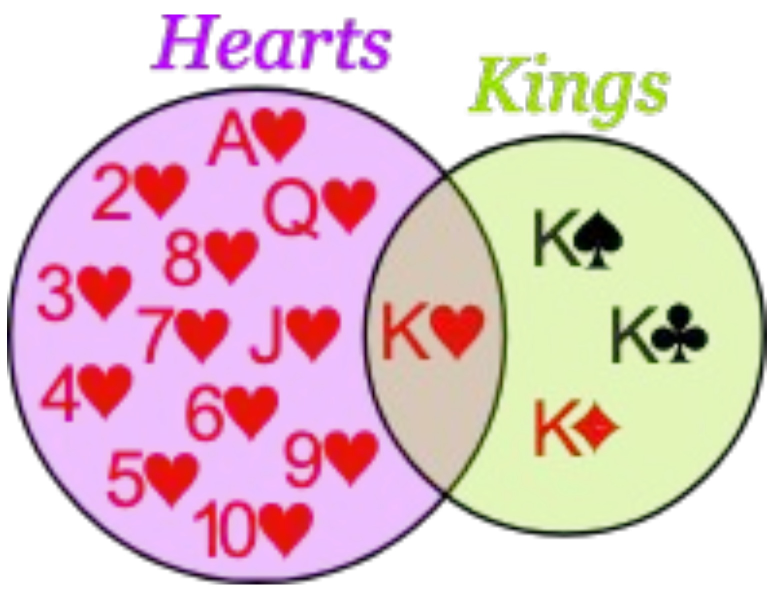


30 BLUES

25 RED

Two things are **mutually exclusive** if the occurrence of one prevents the occurrence of the other.





MUTUALLY EXCLUSIVE

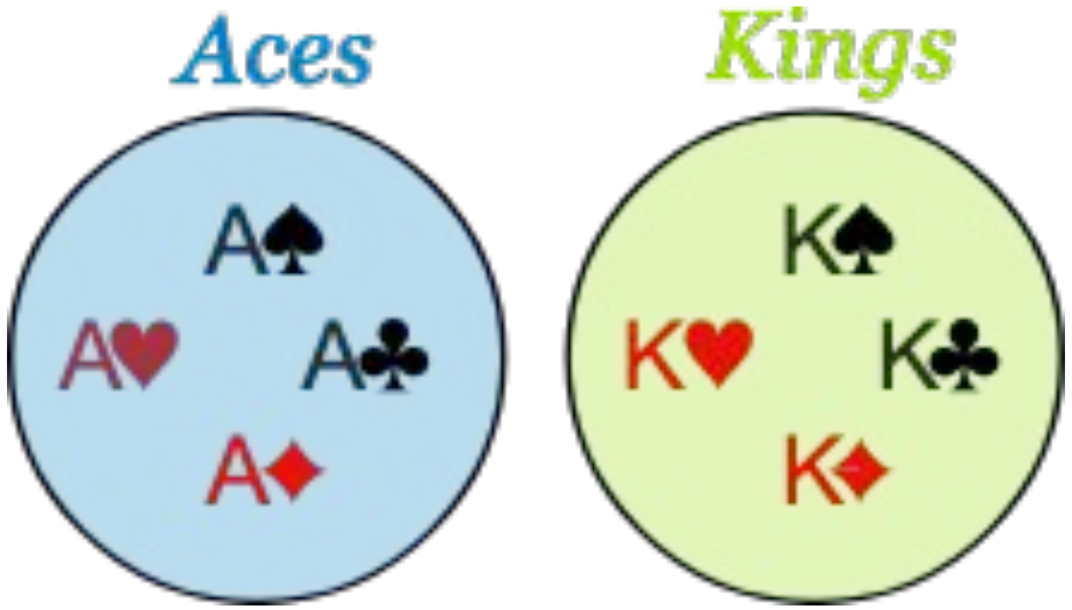
$$P(A \text{ or } B) = P(A) + P(B)$$

NOT MUTUALLY EXCLUSIVE

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

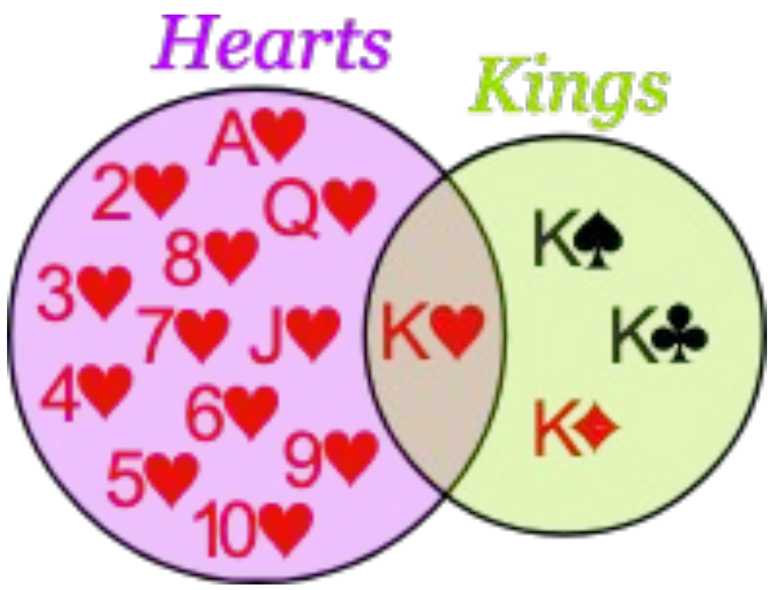
P(Aces or Kings) =

4/52 + 4/52 = 8/52



P(Hearts or Kings) =

4/52 + 13/52 - 1/52 = 16/52



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

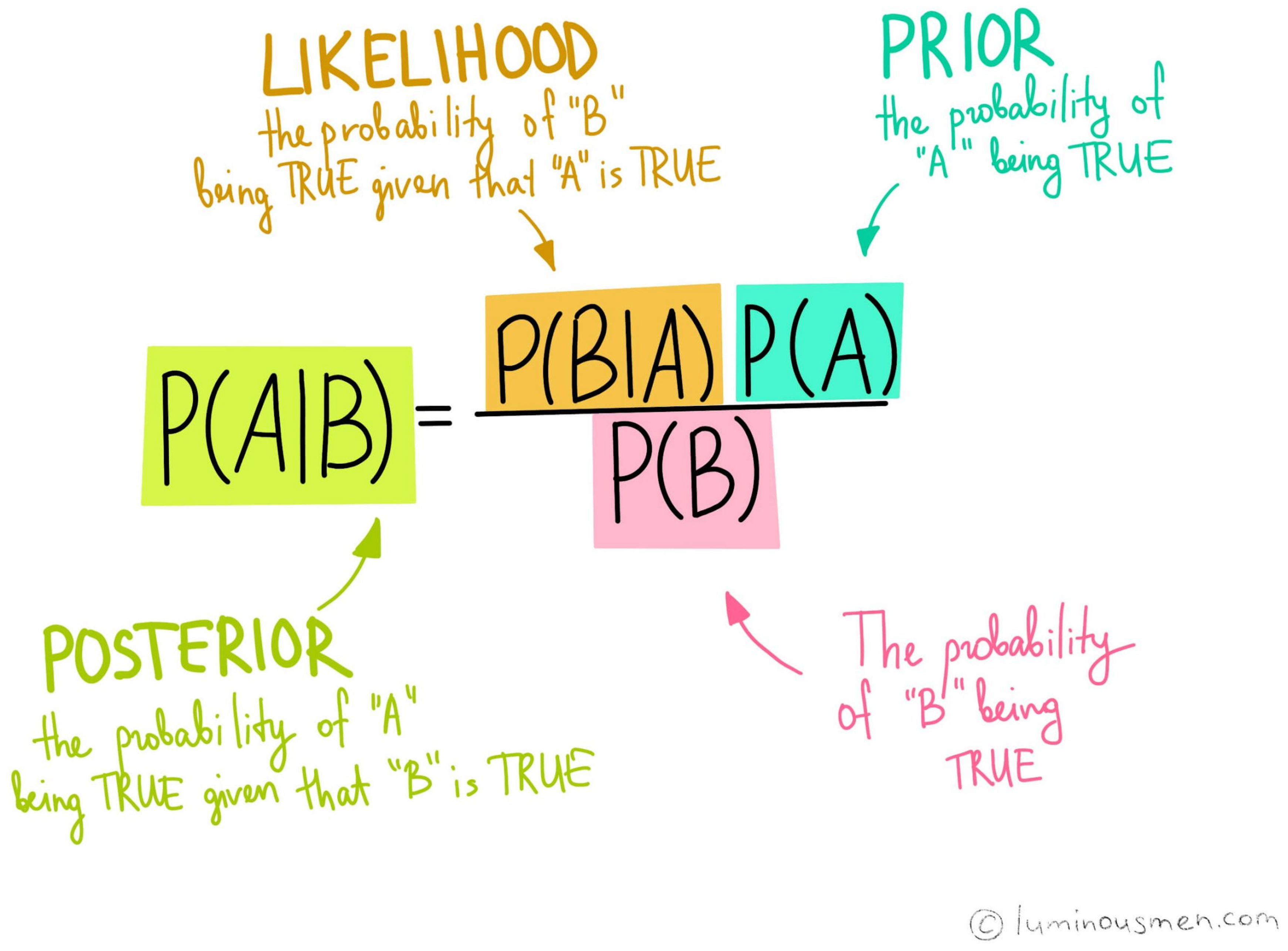
$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

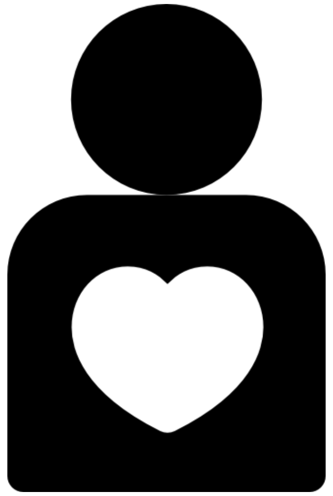
$$P(A)\,P(B \mid A) = P(A \cap B)$$


LIKELIHOOD
the probability of "B" being TRUE given that "A" is TRUE
PRIOR
the probability of "A" being TRUE
P(A|B) = P(B|A) P(A) / P(B)
POSTERIOR
the probability of "A" being TRUE given that "B" is TRUE
The probability of "B" being TRUE
© luminousmen.com

Example Bayes Theorem - Drug Testing

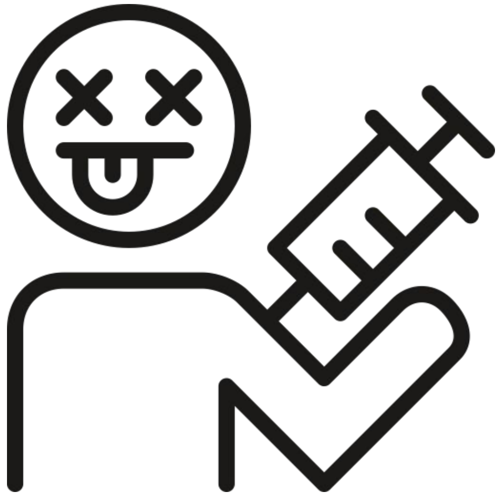
Test/ User	+	-
+	99	1
-	1	99

NON USERS



99.5%

USERS



0.5%

$$P(User \mid +) = \frac{P(+|User)P(User)}{P(+)}$$

$$P(+) = P(+|User)P(User) + P(+|NotUser)P(NotUser)$$

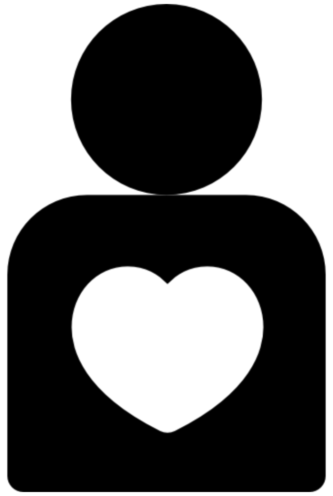
$$P(+) = .99*.005 + .01*.995$$

$$P(+) = 0.0149$$

Example Bayes Theorem - Drug Testing

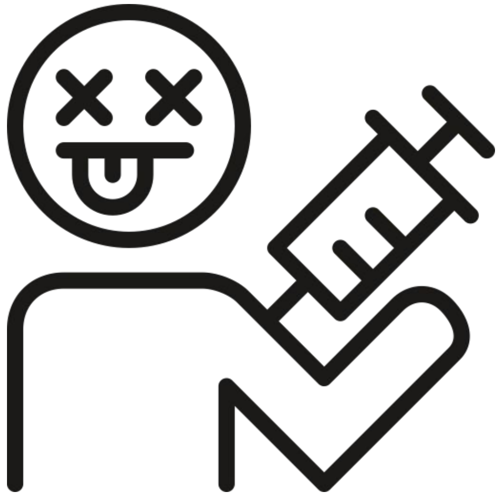
Test/ User	+	-
+	99	1
-	1	99

NON USERS



99.5%

USERS



0.5%

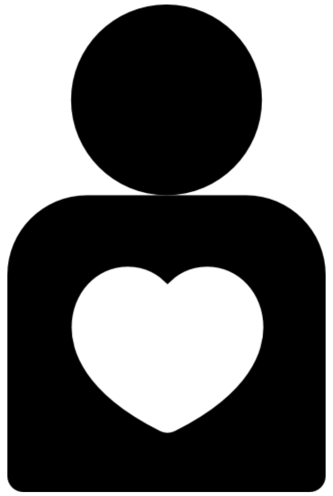
$$P(User \mid +) = \frac{P(+ \mid User)P(User)}{P(+)}$$

$$P(User \mid +) = \frac{.99 * .005}{0.0149}$$

Example Bayes Theorem - Drug Testing

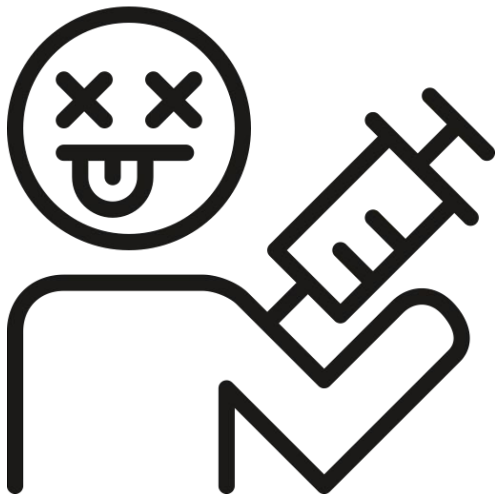
Test/ User	+	-
+	99	1
-	1	99

NON USERS



99.5%

USERS



0.5%

$$P(\text{User} \mid +) = \frac{P(+|\text{User})P(\text{User})}{P(+)}$$

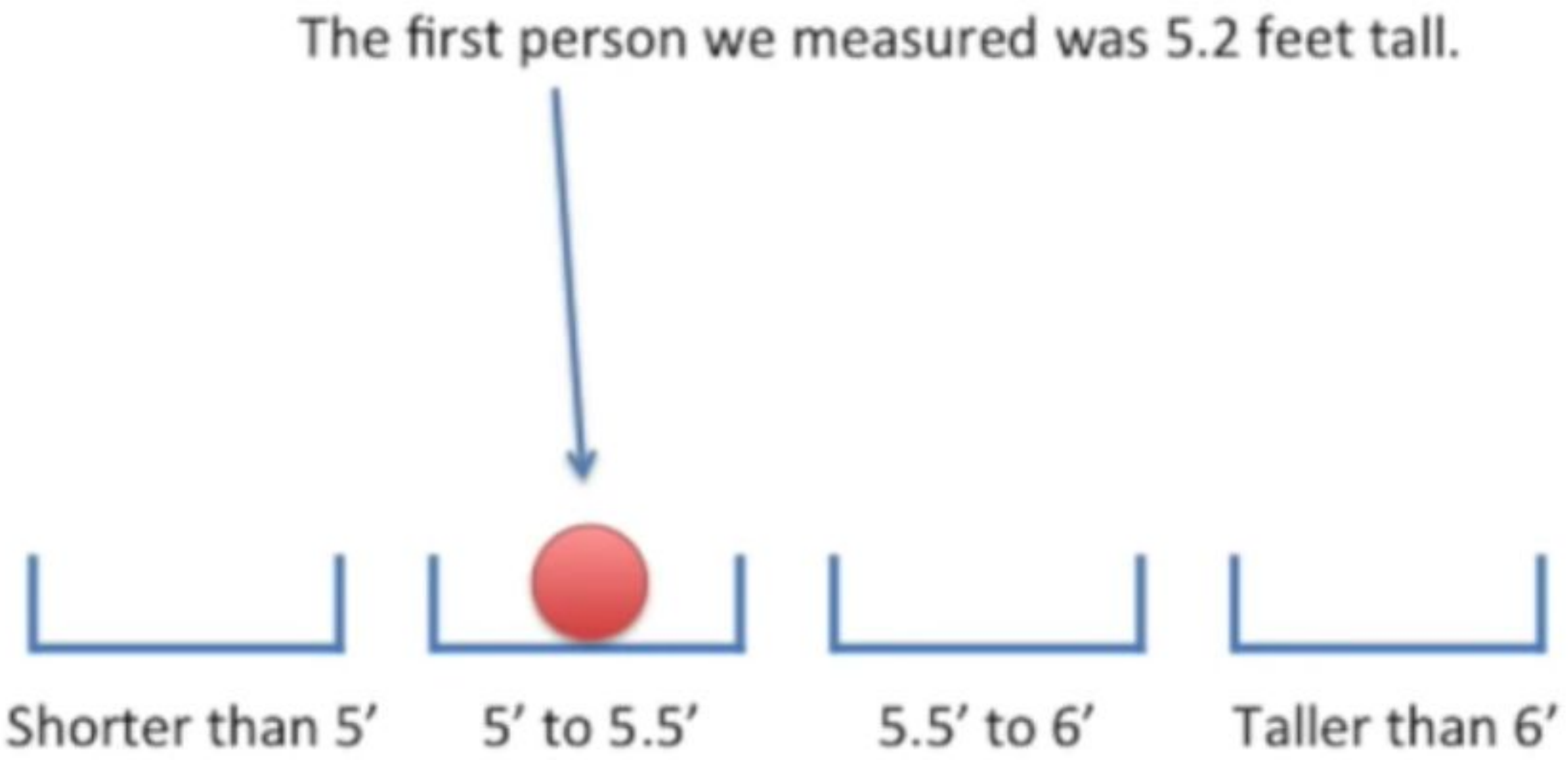
$P(\text{User} \mid +) =\sim$ 33.2%

Summary

Event	Probability
A	$P(A) \in [0,1]$
not A	$P(A^{\complement}) = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B) \qquad \text{if A and B are mutually exclusive}$
A and B	$P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$ $P(A \cap B) = P(A)P(B) \qquad \text{if A and B are independent}$
A given B	$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B\|A)P(A)}{P(B)}$

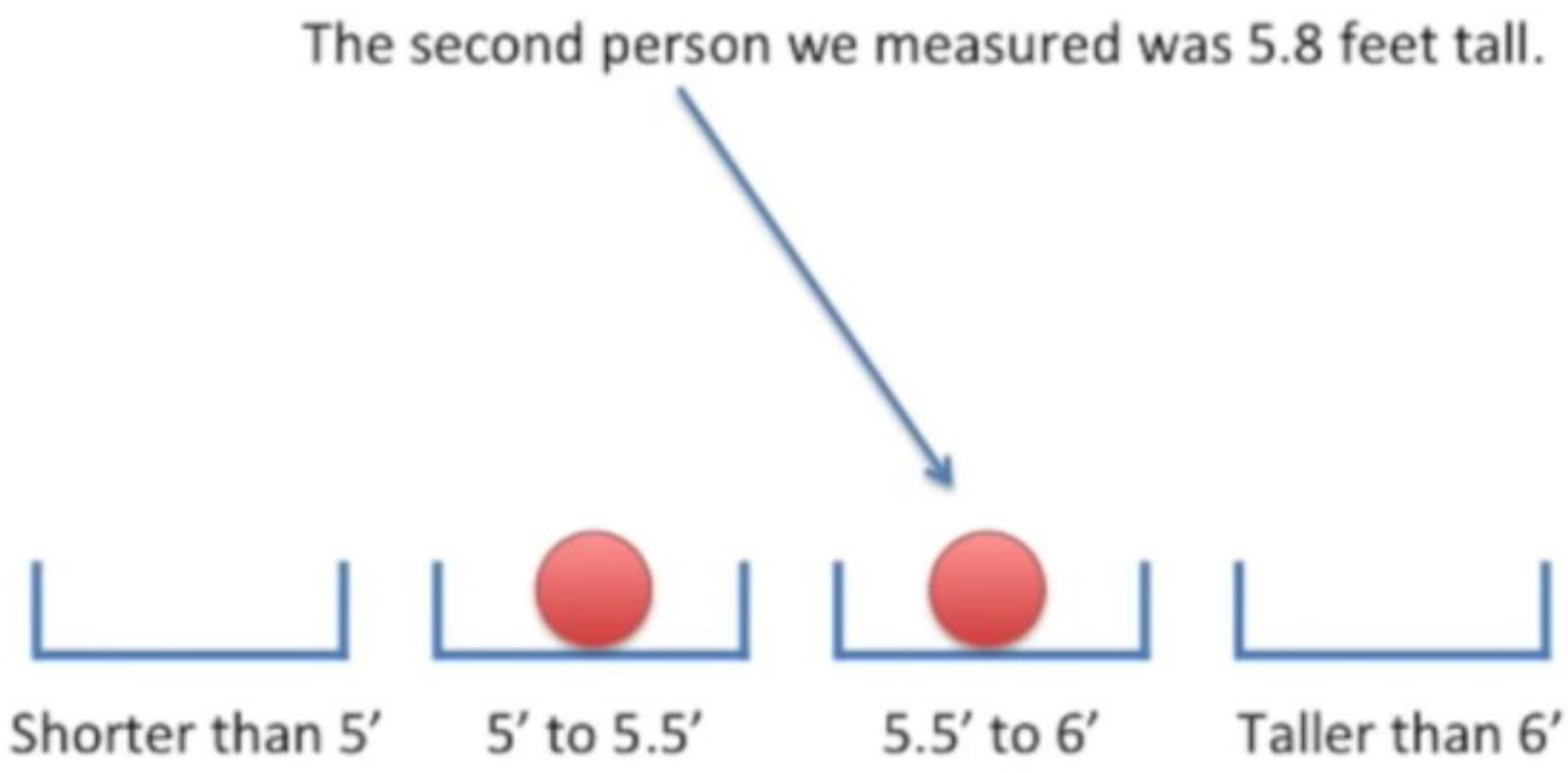
Imagine we measured the height of a lot of people.

Taken from StatQuest



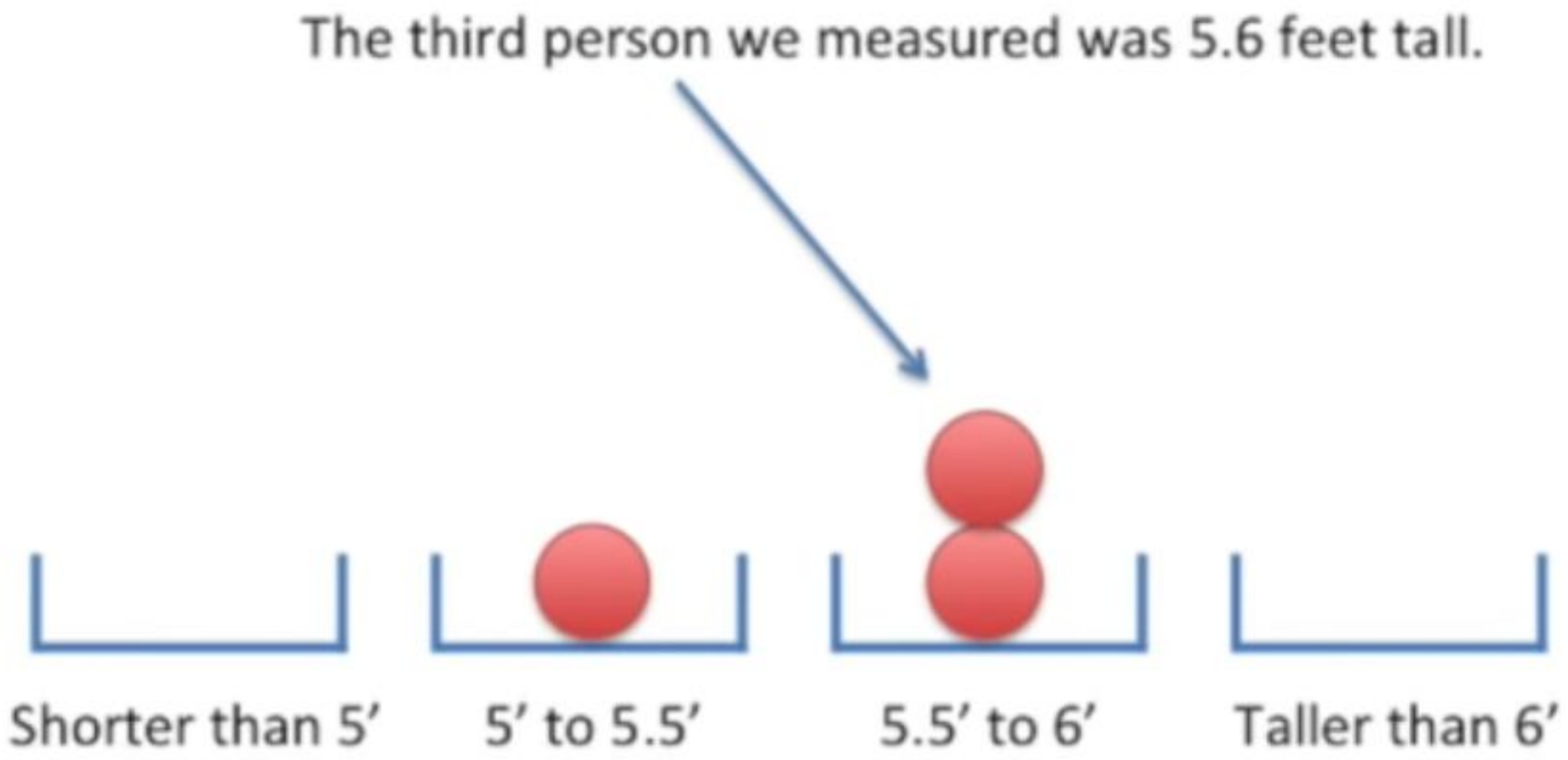
Taken from StatQuest

PDF and Distributions



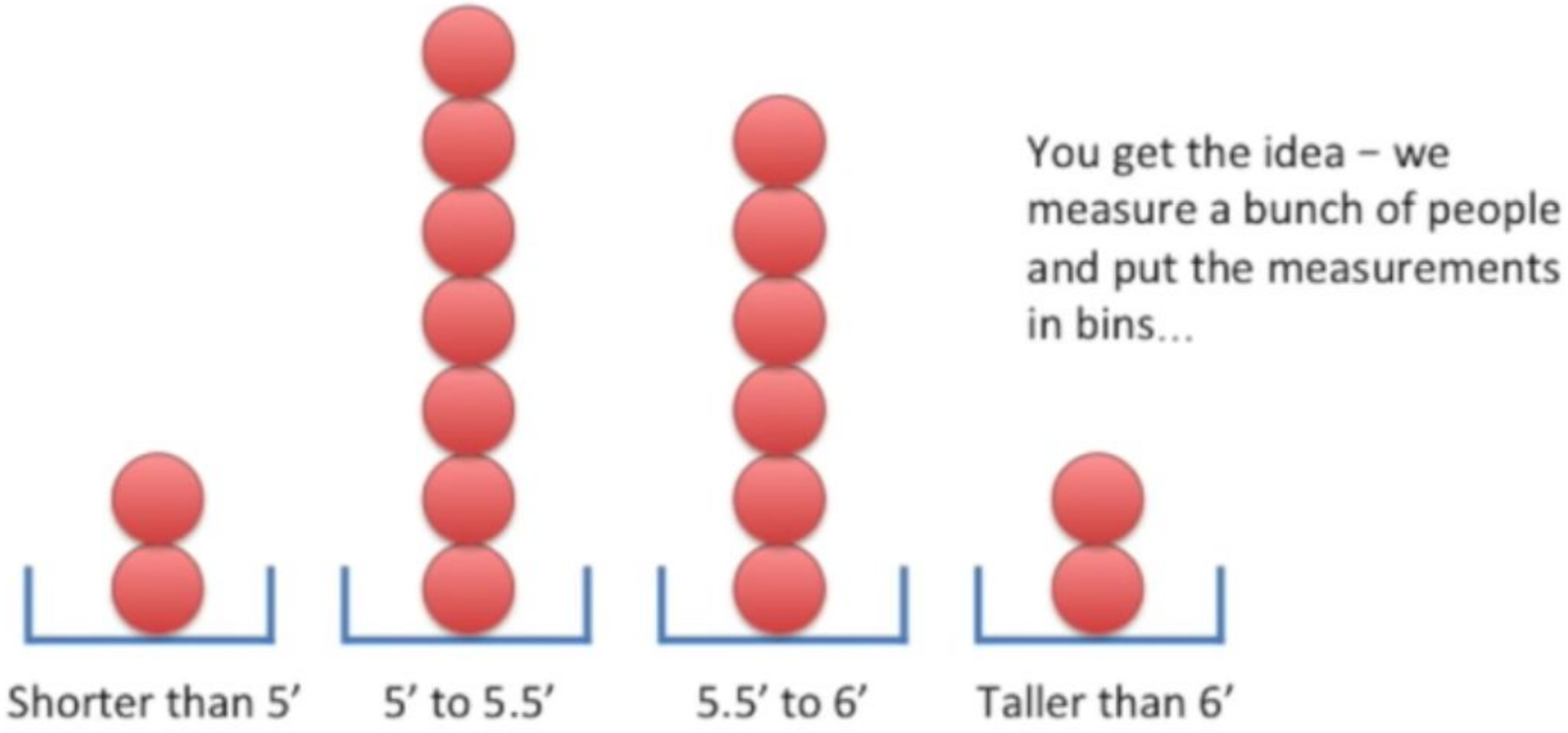
Taken from StatQuest

PDF and Distributions



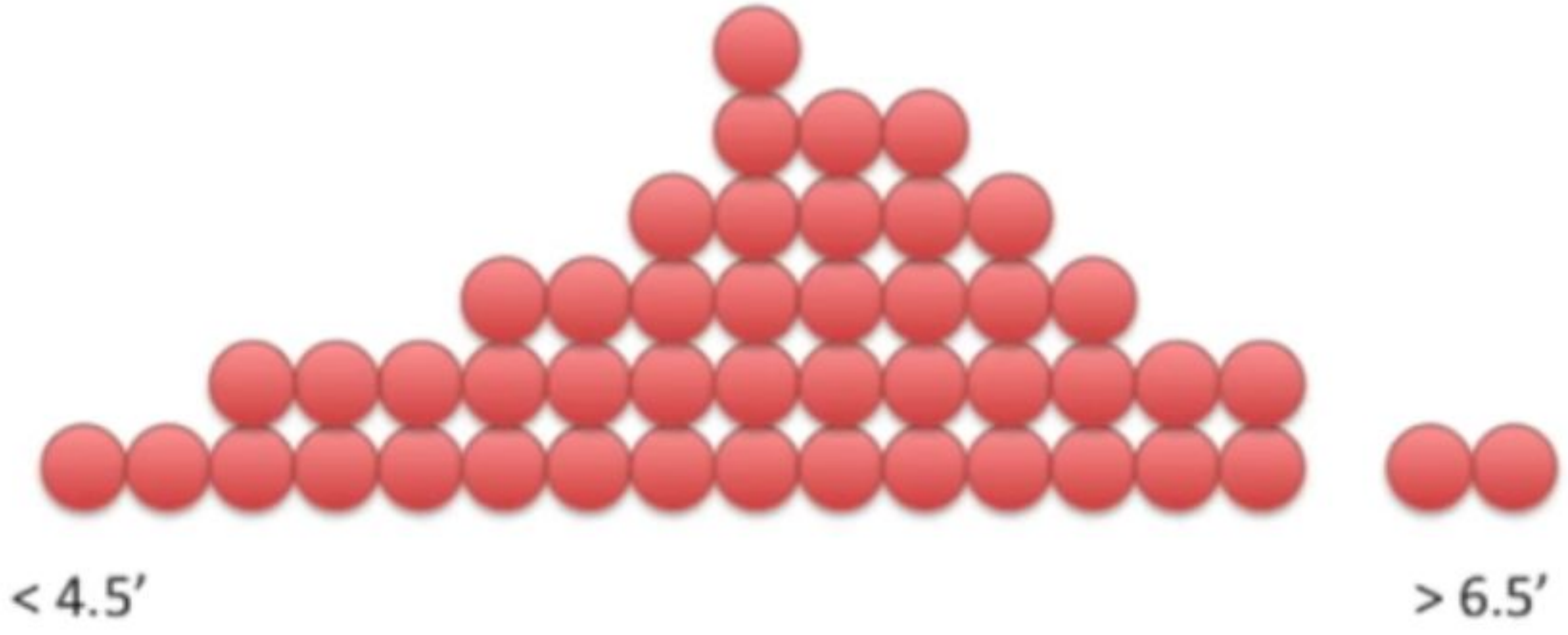
Taken from StatQuest

Histogram



Taken from StatQuest

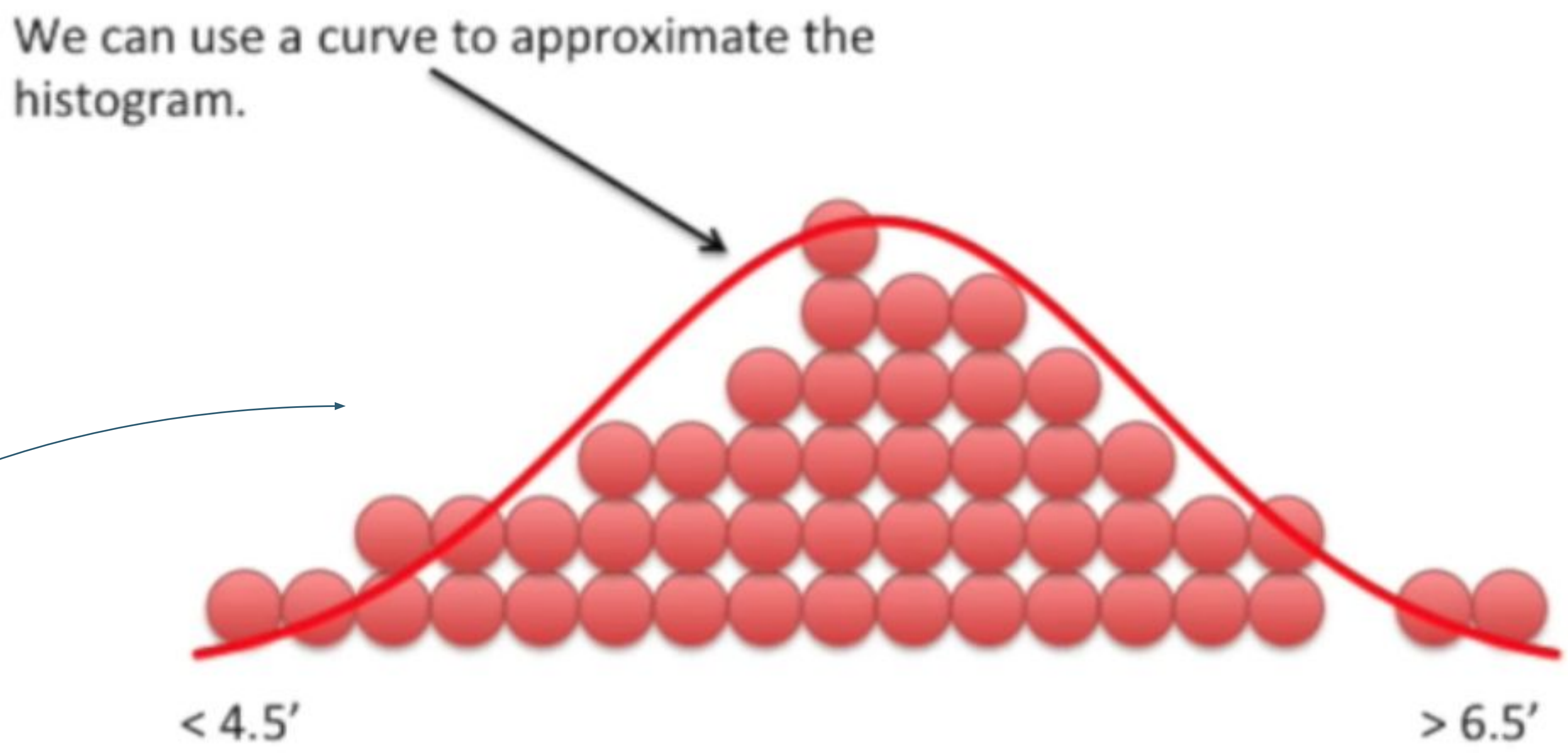
Histogram



By measuring more people and using smaller bins, we get a more accurate and more precise estimate of how heights are distributed.

Taken from StatQuest

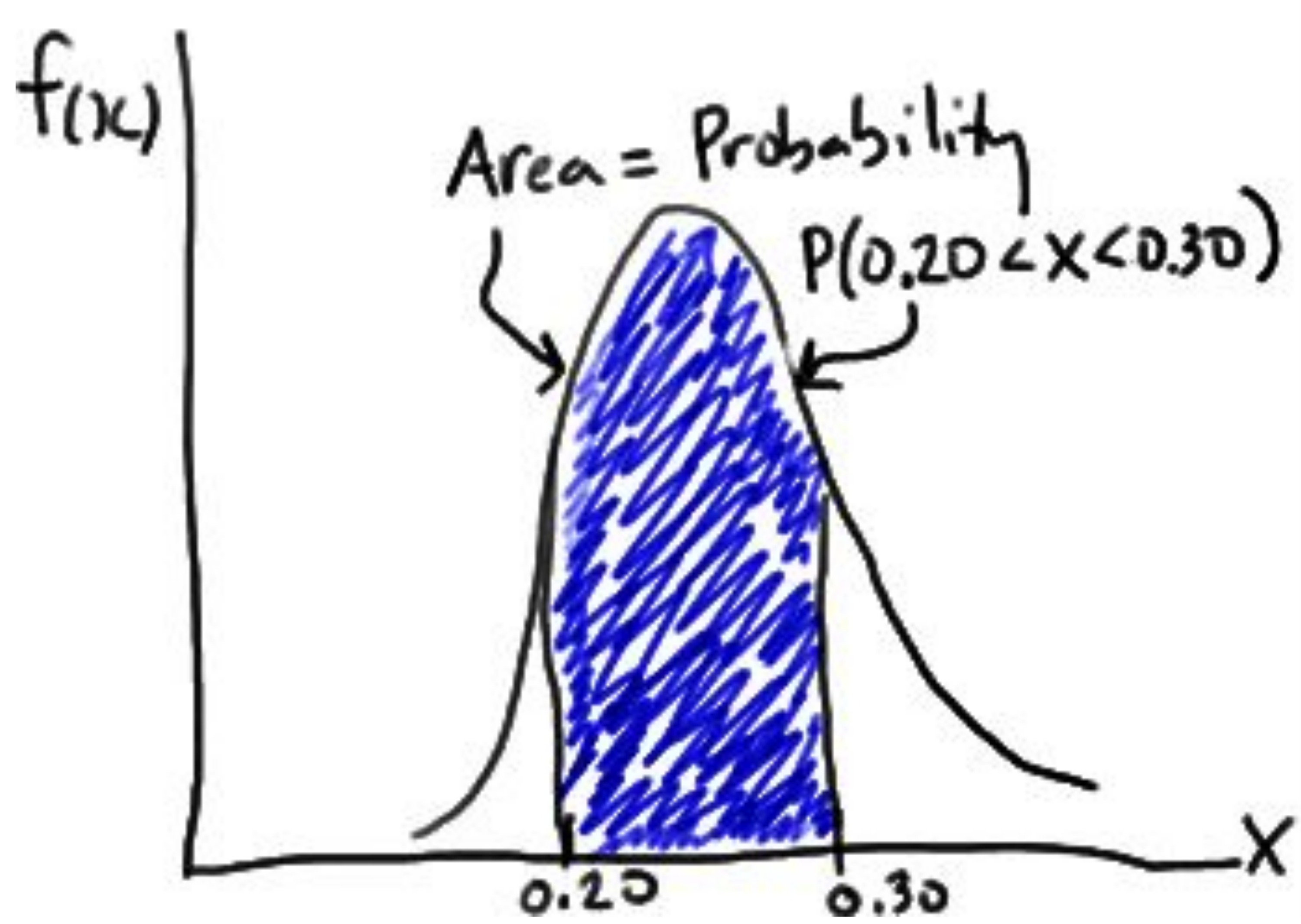
Histogram

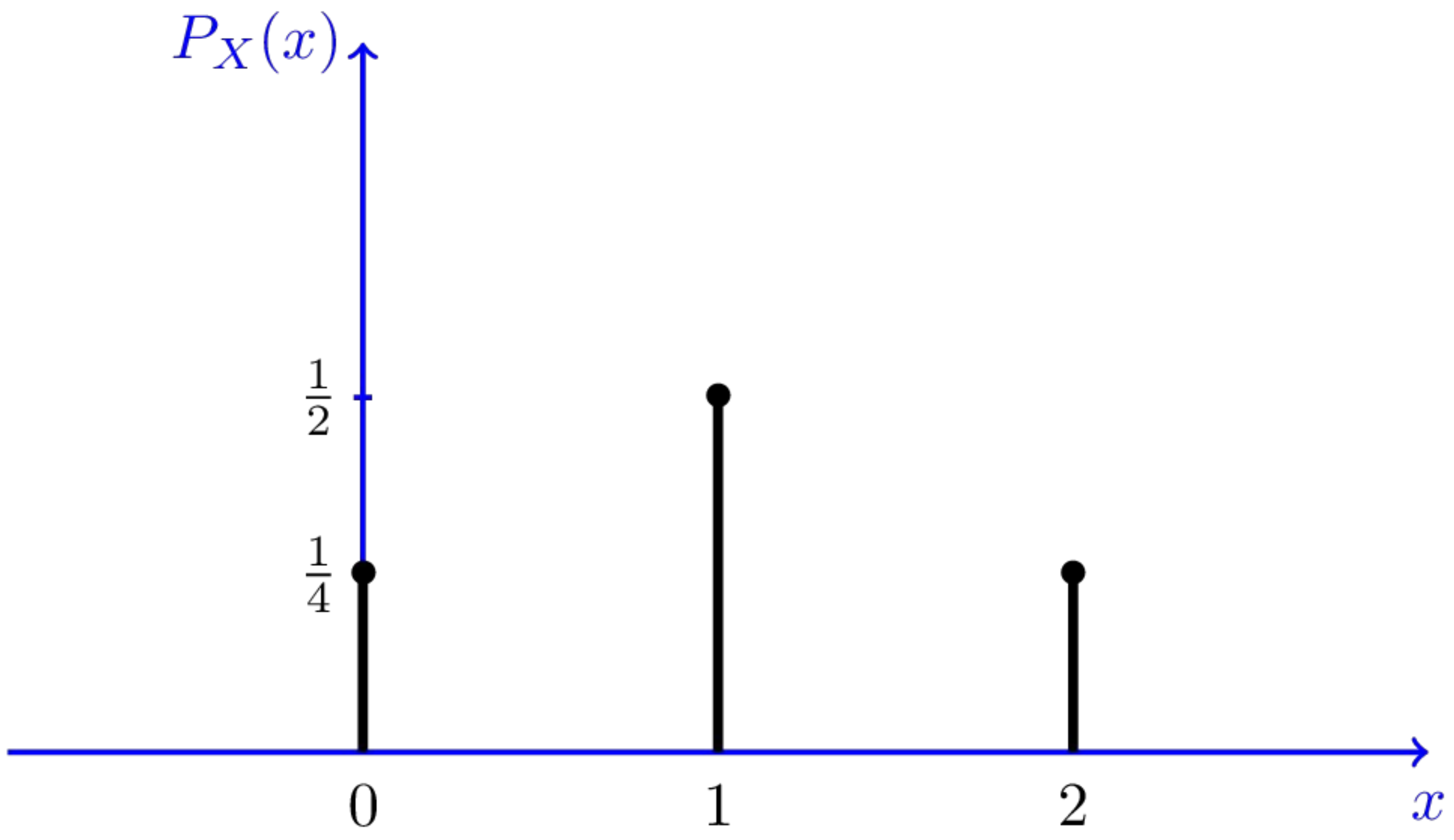


It is a function!

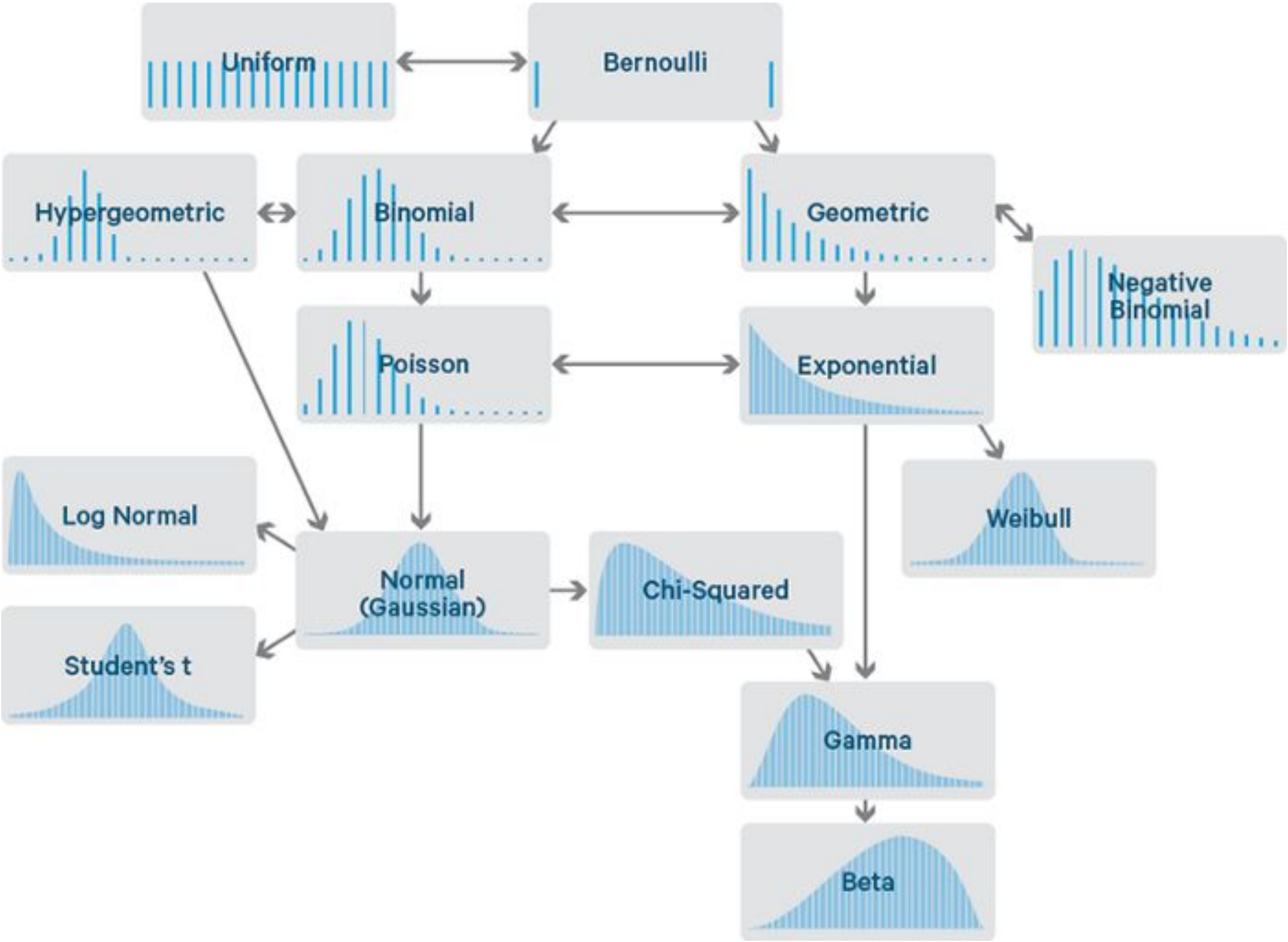
Taken from StatQuest

PDF - Density Histogram

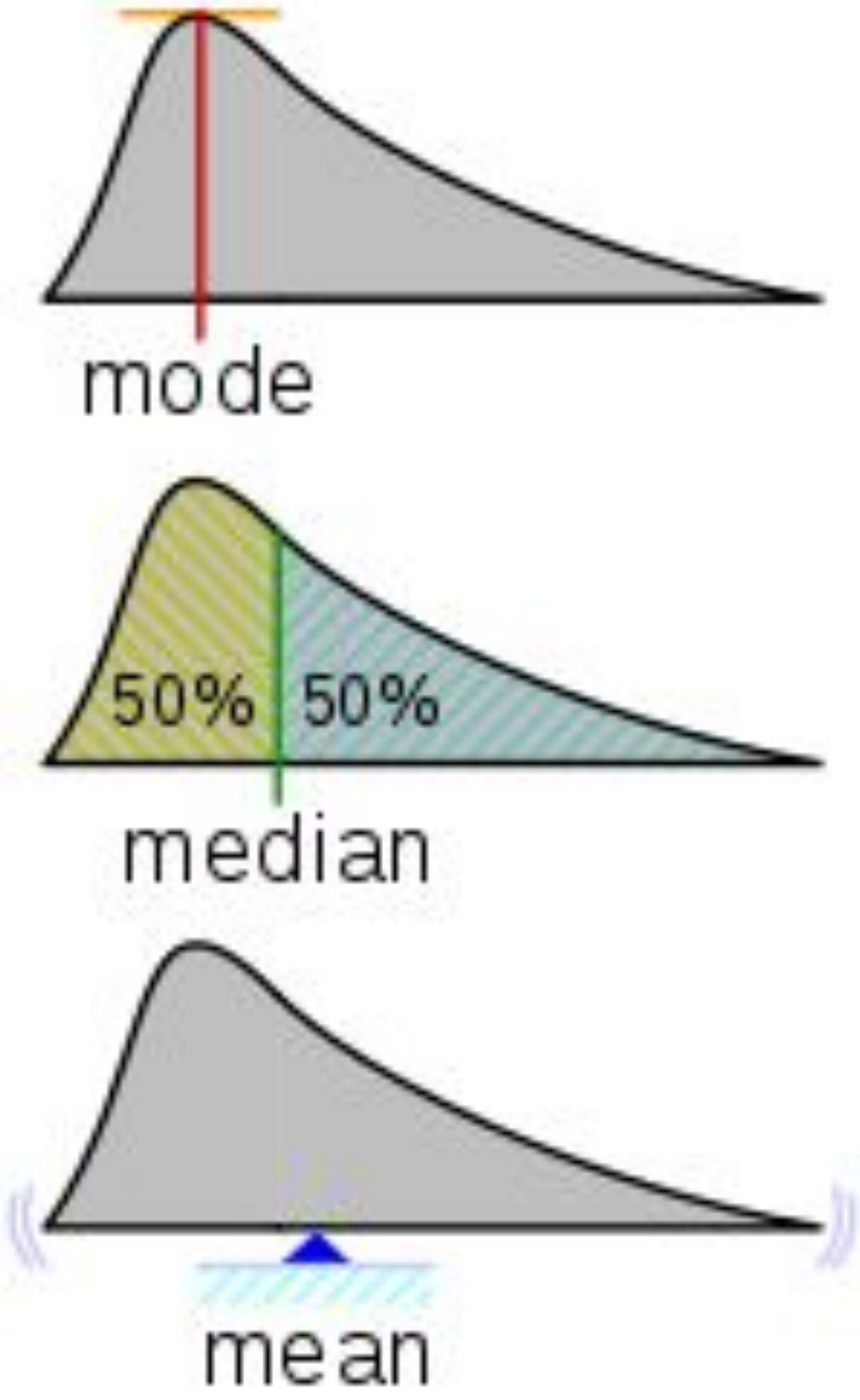




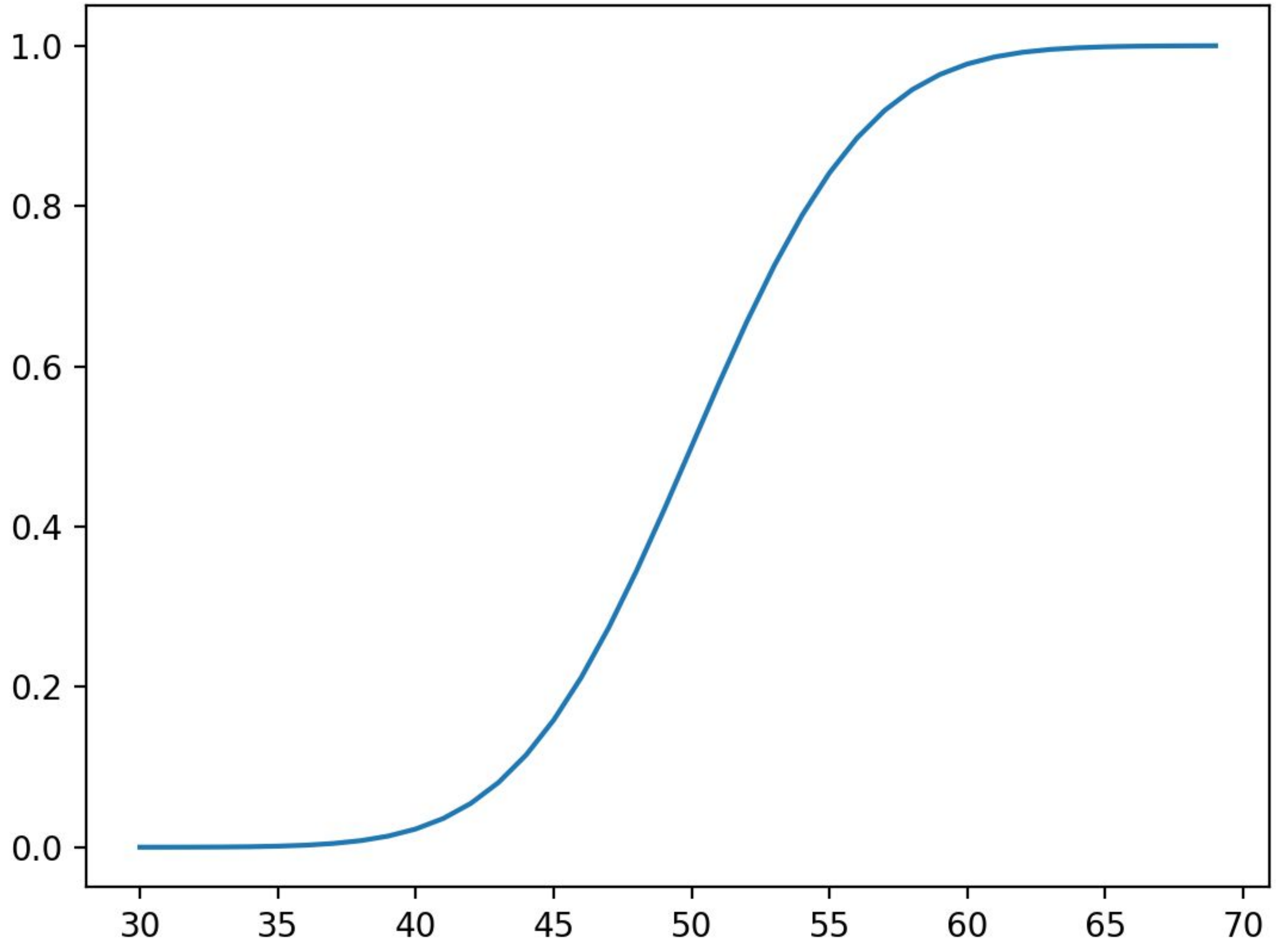
Distributions



PDF and Descriptive Parameters



CDF



STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\Sigma\,(X - \overline{X})^2}{n - 1}}$$

VARIANCE

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

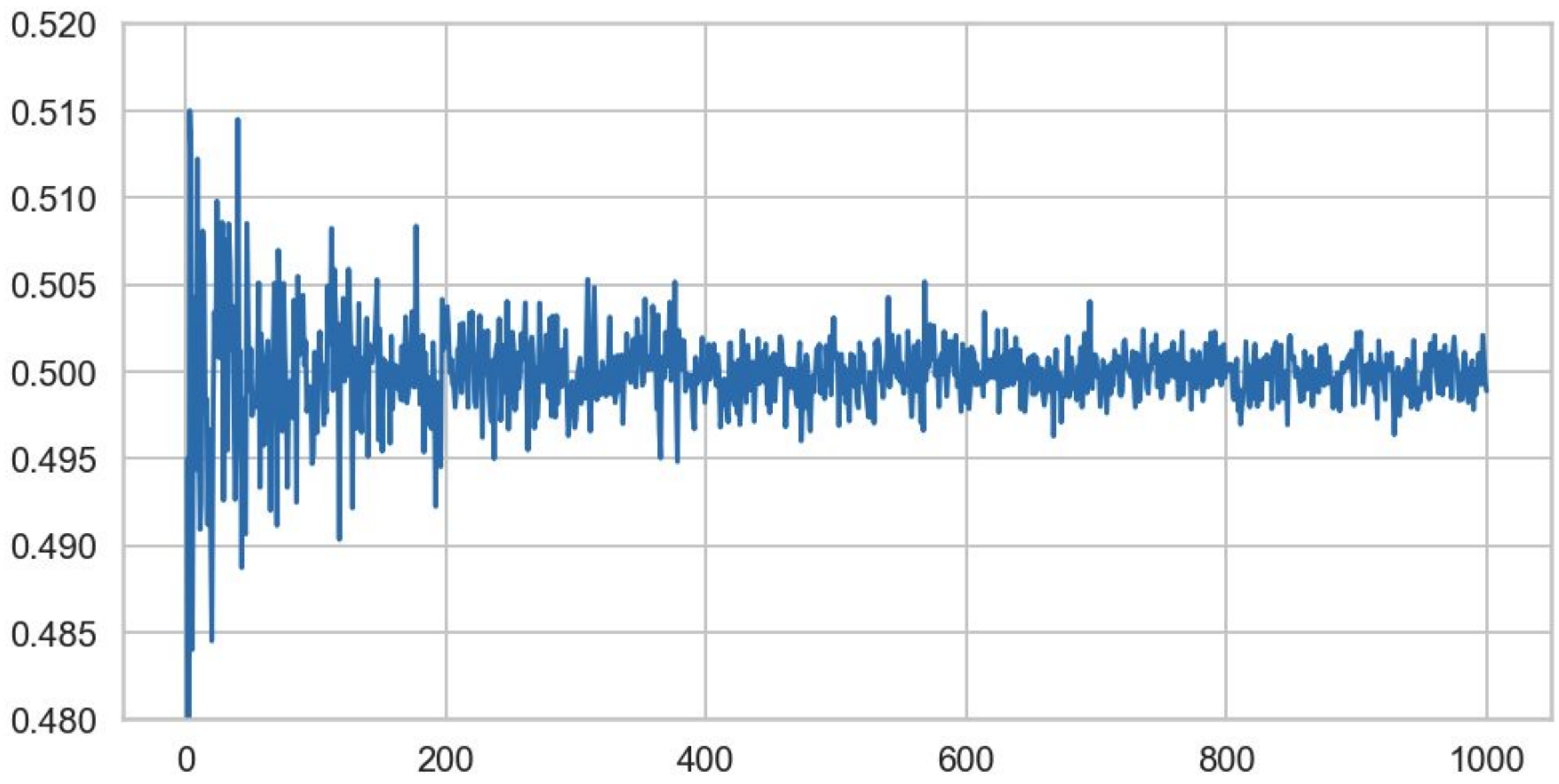
Standard Error

$$SE = \frac{\sigma}{\sqrt{n}}$$

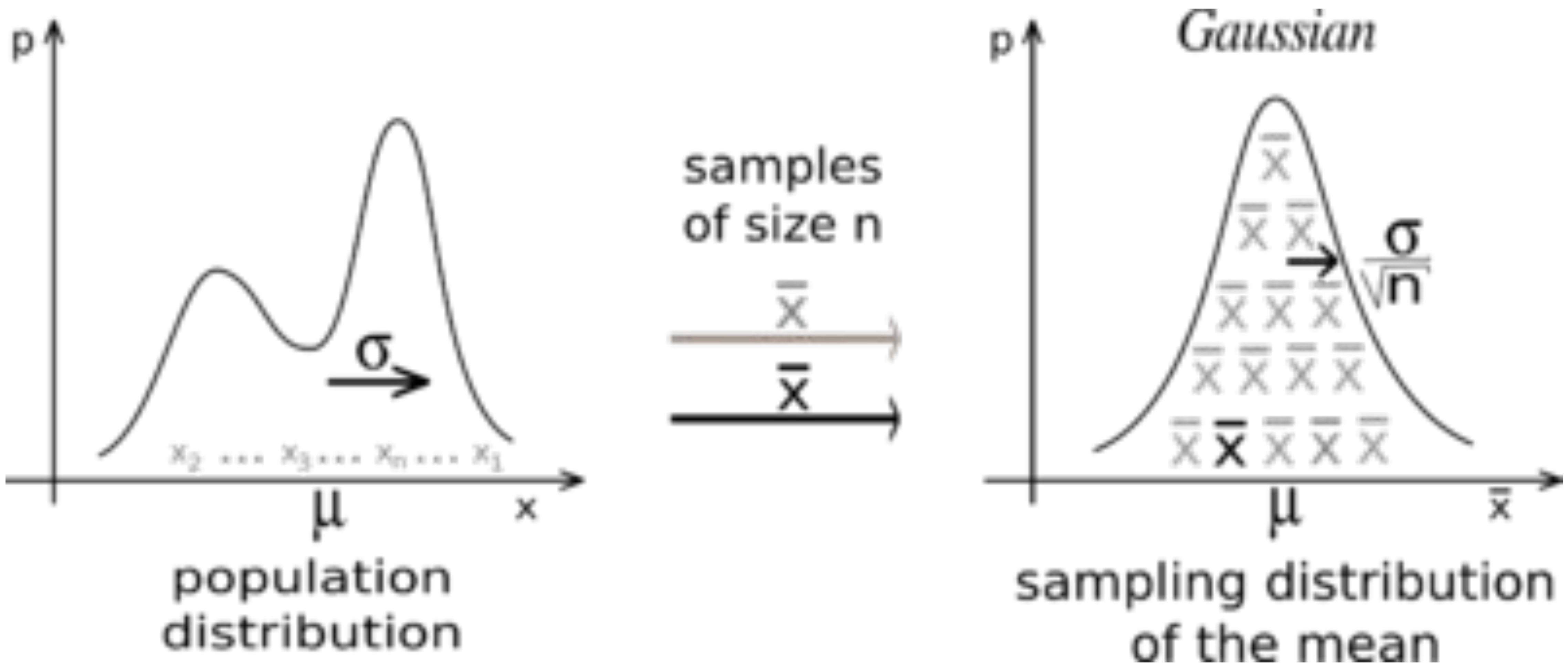
σ ← Standard deviation

n ← Number of samples

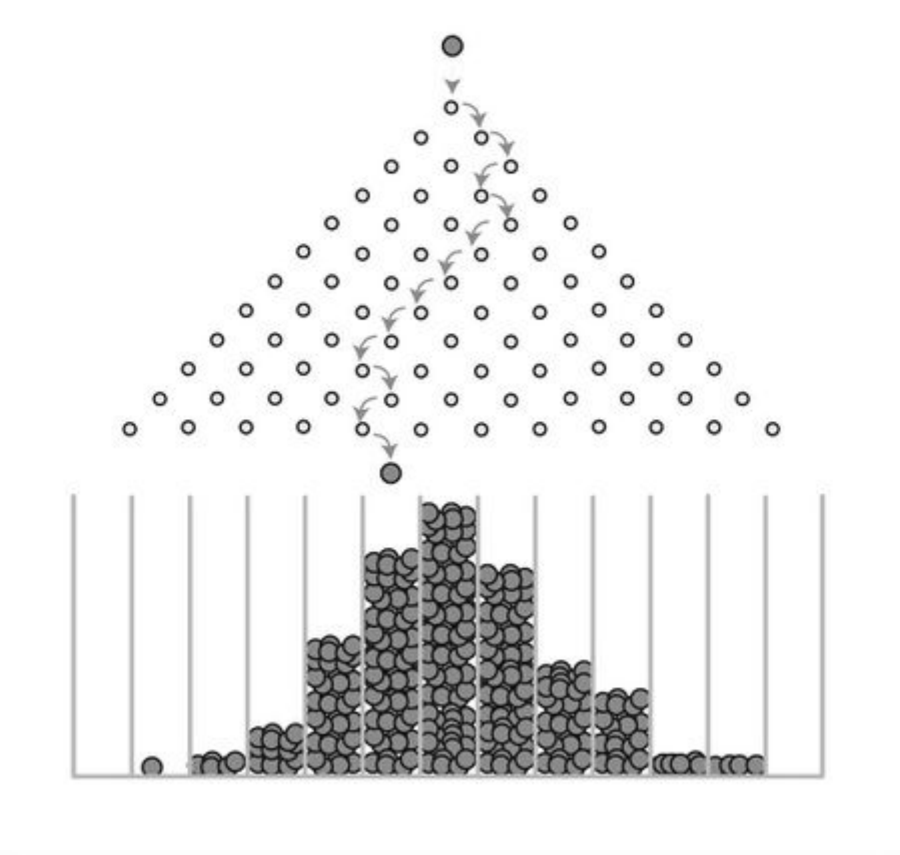
The Law of Large Numbers



The Central Limit Theorem



PROOF



Statistical Models

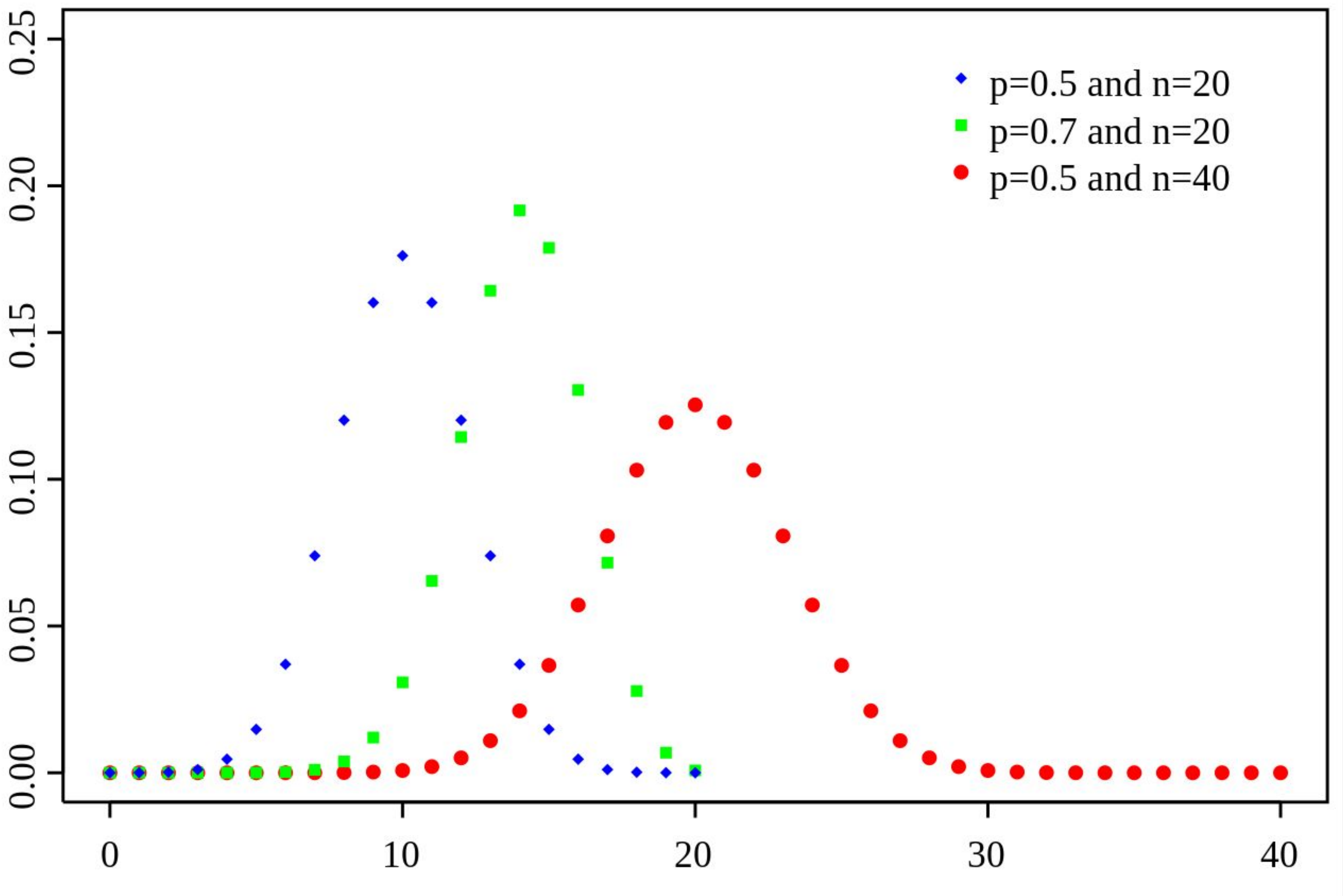
Binomial Distribution

Bernoulli n=1

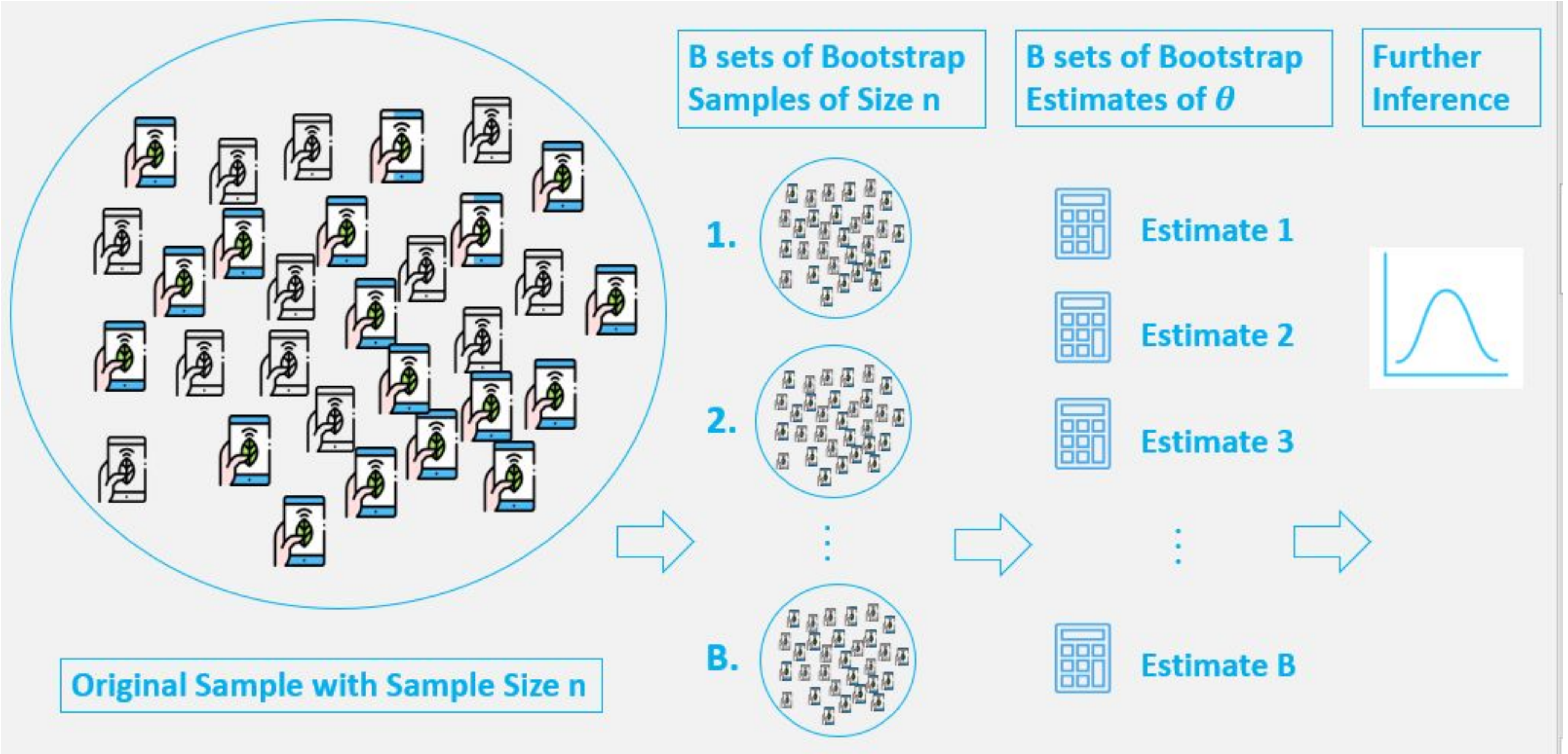
$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$



Bootstrapping

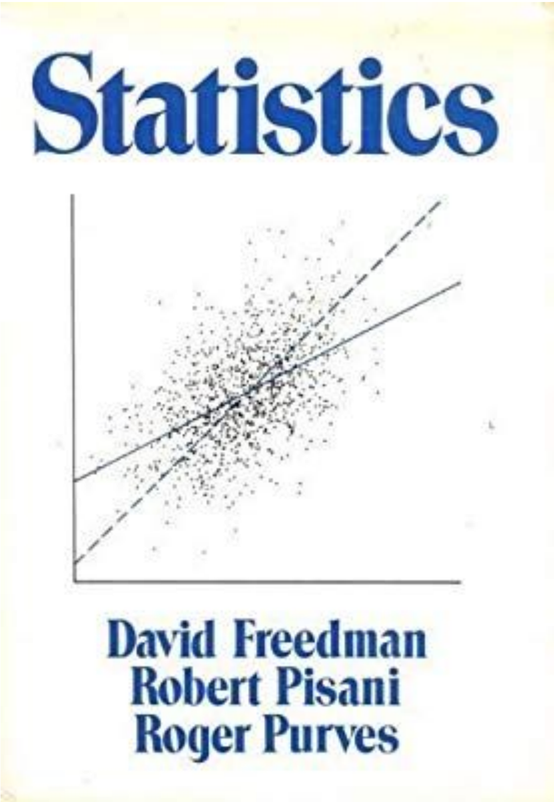


Taken from TowardsDataScience

Useful Resources

Khan Academy

Stat Quest



THANK YOU

GABRIEL MAGALHÃES

Data Scientist

gabriel.magalhaes@totvs.com.br

totvs.com

company/totvs

@totvs

fluig.com

Tecnologia + Conhecimento são nosso DNA.

O sucesso do cliente é o nosso sucesso.

Valorizamos gente boa que é boa gente.

#SOMOSTOTVERS