upGrad
#LifeKoKaroLift

Inferential Statistics

upGrad

Module : Inferential Statistics

Instructor: Dr. Darshan Ingle

upGrad

What you'll learn

1 Probability

2 Random Variable

3 Binomial Distribution

4 Normal Distribution

5 PDF vs CDF

6 Standard Score

7 CLT and Baye's Theorem

Xpert: Dr. Darshan Ingle

Probability

upGrad

- Many events can't be predicted with total certainty. The best we can say is how **likely** they are to happen, using the idea of probability.

- Tossing a Coin

Head (H)
Tail (T)

$P(H) = \frac{1}{2}$ $P(T) = \frac{1}{2}$

- Throwing Dice

Single Die is Thrown : Outcomes 6

$\{1, 2, 3, 4, 5, 6\}$

$P(\text{any one of them}) = \frac{1}{6}$

Probability

upGrad

- In general:

$$\text{Probability of an event happening} = \frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}} \quad \frac{N}{D} = \frac{1}{6}$$

- Example: the chances of rolling a "4" with a die

$\therefore P(\text{rolling a } 4) = \frac{1}{6}$.

Probability

upGrad

- In general:

$$\text{Probability of an event happening} = \frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}} = \frac{4}{5} = 0.8$$

$= 80\%$

- Example: there are 5 marbles in a bag: 4 are blue, and 1 is red. What is the probability that a blue marble gets picked?

$\therefore$ There is 80% chance that a blue marble gets picked.

4 Blue

1 Red

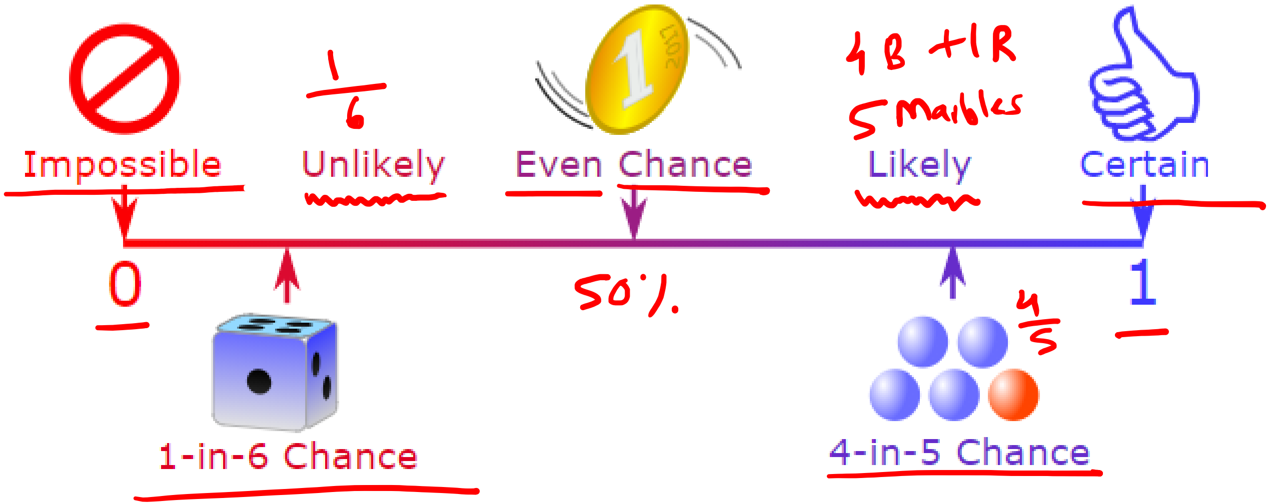
5 Marbles

Probability Line

upGrad

- Summary:

[0, 1] – Range of Probability



Probability is always between 0 and 1

Probability

upGrad

- **Experiment:** a repeatable procedure with a set of possible results.
- Example: Throwing dice

we can throw a dice again & again (Repeatable)

set of possible result = $\{1, 2, 3, 4, 5, 6\}$

Probability

- **Outcome:** A possible result of an experiment.
- Example: Getting a "6“

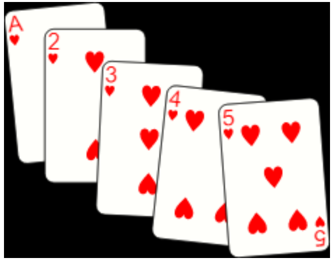
Numerator in our Probability Formula

eg: Throwing a dice.

Outcome: Getting a "6"!

Probability

upGrad



- **Sample Space:** all the possible outcomes of an experiment.
- Example: choosing a card from a deck

There are 52 cards in a deck (not including Jokers)

So the **Sample Space is all 52 possible cards**: {Ace of Hearts, 2 of Hearts, etc... }

- The Sample Space is made up of Sample Points:
- **Sample Point:** just one of the possible outcomes
- Example: Deck of Cards
  - the 5 of Clubs is a sample point
  - the King of Hearts is a sample point

"King" is not a sample point. There are 4 Kings, so that is 4 *different* sample points.

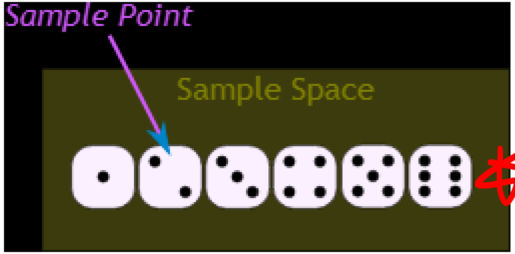
Probability

upGrad

- Example: Throwing dice Sample Space $= \{1, 2, 3, 4, 5, 6\}$

  s.p. is any one point in the S.S.

- There are 6 different sample points in the sample space.



Probability

upGrad

- **Event:** one **or more** outcomes of an experiment
- **Example Events:**
- An event can be just one outcome:
  - Getting a Tail when tossing a coin
  - Rolling a "5“

One Outcome

- An event can include more than one outcome:
  - Choosing a "King" from a deck of cards (any of the 4 Kings)
  - Rolling an "even number" (2, 4 or 6)

More than One Outcome

Probability

upGrad

- Example: Alex wants to see how many times a "double" comes up when throwing 2 dice.
- The **Sample Space** is all possible **Outcomes** (36 Sample Points):
  - {1,1} {1,2} {1,3} {1,4} ... {6,3} {6,4} {6,5} {6,6}

- The **Event** Alex is looking for is a "double", where both dice have the same number. It is made up of these **6 Sample Points**:
  - {1,1} {2,2} {3,3} {4,4} {5,5} and {6,6}

6 Sample Points

Descriptive vs. Inferential statistics

upGrad

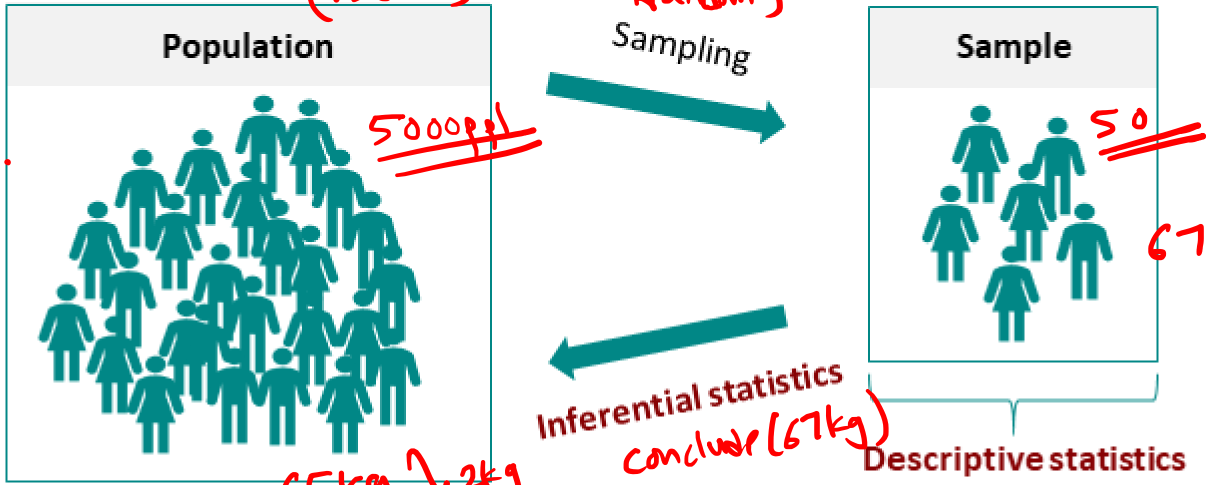
Team Sailcat! 50Lakh ppl

Find avg wt of an Indian?

⇓

Not possible

India (130 cr)



Pop = 65kg
Sample = 67kg
2kg error
(Tolerable)

Inferential : Sample

Descriptive : df.describe()
cnt()
min()
max()
mean()
std()
50%
x%

Saved! ① Time
② Storage
③ ~~Computation~~ Computational Resources reqd

Descriptive vs. Inferential statistics

upGrad

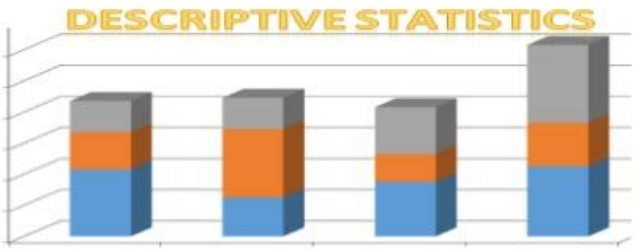
S. No	Descriptive Statistics	Inferential Statistics
1	Concerned with the describing the target population	Make inferences from the sample and generalize them to the population.
2	Organize, analyze and present the data in a meaningful manner	Compares, test and predicts future outcomes.
3	Final results are shown in form of charts, tables and Graphs	Final result is the probability scores.
4	Describes the data which is already known	Tries to make conclusions about the population that is beyond the data available.
5	Tools- Measures of central tendency (mean/median/ mode), Spread of data (range, standard deviation etc.)	Tools- hypothesis tests, Analysis of variance etc.

Descriptive vs. Inferential statistics

upGrad

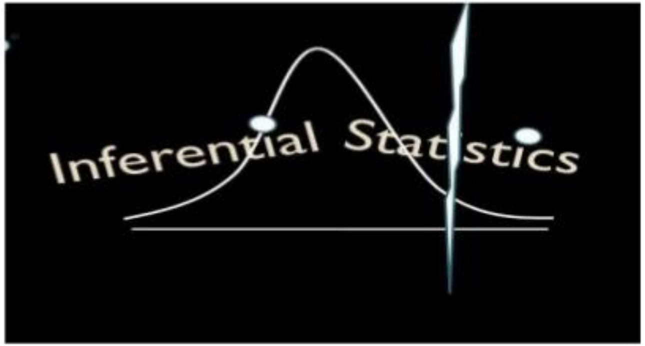
Descriptive Statistics

Descriptive statistics are numbers that are used to summarize and describe data. The word "data" refers to the information that has been collected from an experiment, a survey, a historical record, etc.



Inferential Statistics

The sample is a set of data taken from the population to represent the population. Probability distributions, hypothesis testing, correlation testing and regression analysis all fall under the category of **inferential statistics**.



Statistics vs Probability

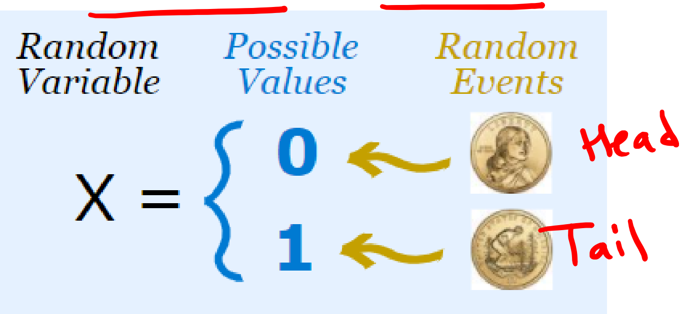
upGrad

Statistics	Probability
It deals with a sample of data.	It deals with a model of random data.
It infers information about a population or a random process that produced a sample.	It deduces information about random events or samples produced from a model.

Random Variables

upGrad

- A Random Variable is a set of **possible values** from a random experiment.
- Example: Tossing a coin: we could get Heads or Tails.
- Let's give them the values **Heads=0** and **Tails=1** and we have a Random Variable "X":



In short:

$X = \{0, 1\}$

Random Variables

upGrad

- **Not Like an Algebra Variable**

Var. like x, it is an unknown value

eg: $x + 22 = 93$

$\boxed{x = ?}$ ✓

Random variable is different.

eg: $X = \{0, 1, 2, 3\}$

X could be 0/1/2/3 randomly

Random Variables

upGrad

- **Sample Space**
- A Random Variable's set of values is the Sample Space.
- Example: Throw a die once
- Random Variable **X** = "The score shown on the top face".

X could be 1/2/3/4/5/6

$\therefore SS = \{1, 2, 3, 4, 5, 6\}$

Example: How many heads when we toss 3 coins?

X = "The number of Heads" is the Random Variable.

In this case, there could be 0 Heads (if all the coins land Tails up), 1 Head, 2 Heads or 3 Heads.

So the Sample Space = {0, 1, 2, 3}

But this time the outcomes are NOT all equally likely.

X= 0|1|2|3

The three coins can land in eight possible ways:

	X = "Number of Heads"
HHH	3
HHT	2
HTH	2
HTT	1
THH	2
THT	1
TTH	1
TTT	0

Looking at the table we see just 1 case of Three Heads, but 3 cases of Two Heads, 3 cases of One Head, and 1 case of Zero Heads. So:

- $P(X = 3) = 1/8$
- $P(X = 2) = 3/8$
- $P(X = 1) = 3/8$
- $P(X = 0) = 1/8$

Random Variables

upGrad

- Example: Two dice are tossed.
- The Random Variable is **X** = "The sum of the scores on the two dice".

1 2

Let's make a table of all possible values:

		1st Die					
		1	**2**	**3**	**4**	**5**	**6**
2nd Die	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

6 x 6 = 36 possible outcomes

There are **6 × 6 = 36** possible outcomes, and the Sample Space (which is the sum of the scores on the two dice) is {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

Let's count how often each value occurs, and work out the probabilities:

- 2 occurs just once, so P(X = 2) = 1/36
- 3 occurs twice, so P(X = 3) = 2/36 = 1/18
- 4 occurs three times, so P(X = 4) = 3/36 = 1/12
- 5 occurs four times, so P(X = 5) = 4/36 = 1/9
- 6 occurs five times, so P(X = 6) = 5/36
- 7 occurs six times, so P(X = 7) = 6/36 = 1/6
- 8 occurs five times, so P(X = 8) = 5/36
- 9 occurs four times, so P(X = 9) = 4/36 = 1/9
- 10 occurs three times, so P(X = 10) = 3/36 = 1/12
- 11 occurs twice, so P(X = 11) = 2/36 = 1/18
- 12 occurs just once, so P(X = 12) = 1/36

Not equally likely.

Probability Distribution

upGrad

- A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.
- Example to study relationship between random variables and probability distributions.
- Suppose you flip a coin two times.
- We get four possible outcomes: HH, HT, TH, and TT.
  2 1 1 0
- Now, let the variable X represent the number of Heads that result from this experiment.
- The variable X can take on the values 0, 1, or 2.
- In this example, X is a random variable; because its value is determined by the outcome of a statistical experiment.

$X = \#\text{Heads}$

$P(X=0) = \frac{1}{4} = 0.25$

$P(X=1) = \frac{2}{4} = 0.5$

$P(X=2) = \frac{1}{4} = 0.25$

Tabular form

Number of heads	Probability
0	0.25
1	0.50
2	0.25

Prob. Dist.

The table represents the probability distribution of the random variable X.

Cumulative Probability Distributions

upGrad

$X = 0/1/2$

- A **cumulative probability** refers to the probability that the value of a random variable falls within a specified range.
- Example: Coin flip experiment. If we flip a coin two times, we might ask:
- What is the probability that the coin flips would result in one or fewer heads?
- The answer would be a cumulative probability.

CPD:

$P(X \leq 1)$

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$$

- Like a probability distribution, a cumulative probability distribution can be represented by a table or an equation.

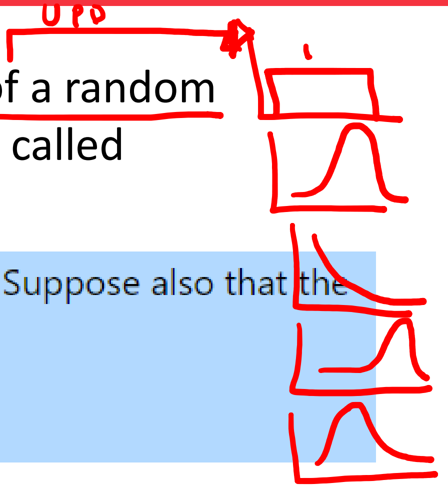
$P(X \leq 1) = 0.75$

$P(X \leq 2) = 1.00$

Number of heads: x	Probability: P(X = x)	Cumulative Probability: $P(X \leq x)$
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

Uniform Probability Distribution

upGrad



- The simplest probability distribution occurs when all of the values of a random variable occur with equal probability. This probability distribution is called the **uniform distribution**.

Uniform Distribution. Suppose the random variable X can assume k different values. Suppose also that the $P(X = x_k)$ is constant. Then,

$$\boxed{P(X = x_k) = 1/k}$$

Example 1

Suppose a die is tossed. What is the probability that the die will land on 5?

$S = \{1, 2, 3, 4, 5, 6\}$

Each possible outcome is a R.V. (X) & each is equally likely to occur

∴ we hv Uniform Distribution.

$P(X=5) = \frac{1}{6}$. $\boxed{P(X=1) \text{ or } P(X=2) \text{ or } \ldots P(X=6) = \frac{1}{6}}$

~~Uniform~~ Cumulative Probability Distribution

upGrad

Example 2

Suppose we repeat the dice tossing experiment described in Example 1. This time, we ask what is the probability that the die will land on a number that is smaller than 5?

$S = \{1, 2, 3, 4, 5, 6\}$ { each possible outcome is equally likely to occur.

$$\therefore P(X<5) = P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

Cumulative Probability

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

$$= \frac{4}{6}$$

$$\therefore \boxed{P(X<5) = \frac{2}{3}}$$

Expected Value (EV)

upGrad

- The expected value (EV) is an anticipated value for an investment at some point in the future.
- In statistics and probability analysis, the expected value is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all of those values.
- By calculating expected values, investors can choose the scenario most likely to give the desired outcome.

$$\boxed{EV = \sum P(X_i) \times X_i}$$

- **Example:** Consider a normal six-sided die. Once you roll the die, it has an equal one-sixth chance of landing on one, two, three, four, five, or six. Given this information, the calculation is straightforward:

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

$\vdots$

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

$$\left(\frac{1}{6} \times 1\right) + \left(\frac{1}{6} \times 2\right) + \left(\frac{1}{6} \times 3\right) + \left(\frac{1}{6} \times 4\right) + \left(\frac{1}{6} \times 5\right) + \left(\frac{1}{6} \times 6\right) = 3.5$$

(E.V)

- If you were to roll a six-sided die an infinite amount of times, you see the average value equals 3.5.

Standard Deviation of a Random Variable

upGrad

- The expected value of a variable may fail your analysis, as it works the best for large-sized data.
- The value can be affected by the presence of outliers; therefore, you need to have a measure that analyses the risk involved in the analysis. This measure is the standard deviation of the variable.
- The standard deviation of a random variable is simply the standard deviation of a sample as it goes to infinity.
- It is denoted by 'σ'.
- Some additional points related to the standard deviation of the variable are as follows:
- Variance = $Var(X) = \Sigma(X - \mu)^2 p(X)$
- In other words, the variance is the expected value of the squared difference of the random variable X and the expected value m.
- **Standard deviation (σ) = √Var(X)**

Standard Deviation of a Random Variable

upGrad

Robert's work schedule for next week will be released today. Robert will work either 45, 40, 25, or 20.8 hours. The probabilities for each possibility are listed below:

45 hours: 0.3
40 hours: 0.2
25 hours: 0.4
20.8 hours: 0.1

What is the standard deviation of the possible outcomes?

There are four steps to finding the standard deviation of random variables.

1. First, calculate the mean of the random variables.
2. Second, for each value in the group (45, 40, 25, and 20.8), subtract the mean from each and multiply the result by the probability of that outcome occurring.
3. Third, add the four results together.
4. Fourth, find the square root of the result.

Standard Deviation of a Random Variable

upGrad

Robert's work schedule for next week will be released today. Robert will work either 45, 40, 25, or 20.8 hours. The probabilities for each possibility are listed below:

45 hours: 0.3
40 hours: 0.2
25 hours: 0.4
20.8 hours: 0.1

What is the standard deviation of the possible outcomes?

$$\bar{x} = \frac{45 + 40 + 25 + 20.8}{4} = 32.7$$

$$(45 - 32.7)^2(0.3) = 45.38$$
$$(40 - 32.7)^2(0.2) = 10.66$$
$$(25 - 32.7)^2(0.4) = 23.72$$
$$(\cancel{12} - 32.7)^2(0.1) = 42.85$$

20.8

$$45.38 + 10.66 + 23.72 + 42.85 = 122.61$$

$$\sqrt{122.61} = 11.07$$

$$32.7 \pm 11.07$$

Discrete Probability Distributions

Addition Rules for Probability

upGrad

Addition Rule 1: When two events, A and B, are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event.

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

Addition Rules for Probability

upGrad

Experiment: A single 6-sided die is rolled.
What is the probability of rolling a 2 or a 5?

$$P(2) = \frac{1}{6}$$

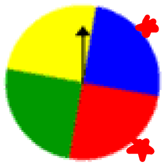
$$P(5) = \frac{1}{6}$$

$$P(2 \text{ or } 5) = P(2) + P(5)$$
$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}.$$

Addition Rules for Probability

upGrad

Experiment 2: A spinner has 4 equal sectors colored yellow, blue, green, and red.
What is the probability of landing on red or blue after spinning this spinner?



$P(\text{red}) = \frac{1}{4}$

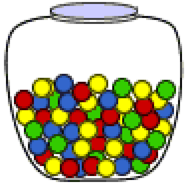
$P(\text{blue}) = \frac{1}{4}$

$P(\text{red or blue}) = P(\text{red}) + P(\text{blue})$

$= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

Addition Rules for Probability

Experiment 3: A glass jar contains 1 red, 3 green, 2 blue, and 4 yellow marbles.
If a single marble is chosen at random from the jar, what is the probability that it is yellow or green?



1R
3G
2B 4y

$1+3+2+4=10$

$P(y) = \frac{4}{10}$

$P(G) = \frac{3}{10}$

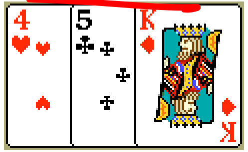
$P(y \text{ or } G) = \frac{4}{10} + \frac{3}{10} = \frac{7}{10}$.

Addition Rules for Probability

upGrad

- In each of the three experiments before, the events are mutually exclusive.
- ~~Let's look~~ at some experiments in which the events are non-mutually exclusive.

- **Experiment 4:** A single card is chosen at random from a standard deck of 52 playing cards. What is the probability of choosing a king or a club?



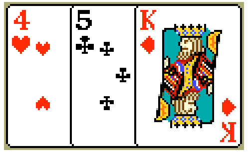
Additional Rule 2: When two events, A and B, are non-mutually exclusive, the probability that A or B will occur is:

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

Addition Rules for Probability

upGrad

- **Experiment 4:** A single card is chosen at random from a standard deck of 52 playing cards. What is the probability of choosing a king or a club?



$$P(K \text{ or } C) = P(K) + P(C) - P(K \text{ and } C)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$$

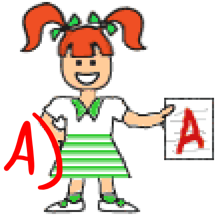
$$= \frac{4}{13}.$$

Addition Rules for Probability

upGrad

Denominator

- **Experiment 5:** In a math class of 30 students, 17 are boys and 13 are girls. On a unit test, 4 boys and 5 girls made an A grade.
- If a student is chosen at random from the class, what is the probability of choosing a girl or an A student?



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

$$= \frac{13}{30} + \frac{9}{30} - \frac{5}{30} = \frac{17}{30}$$

Independent Events

upGrad

- **Definition:** Two events, A and B, are **independent** if the fact that A occurs does not affect the probability of B occurring.

Eg:

- Landing on heads after tossing a coin AND rolling a 5 on a single 6-sided die.
- Choosing a marble from a jar AND landing on heads after tossing a coin.
- Choosing a 3 from a deck of cards, replacing it, AND then choosing an ace as the second card.
- Rolling a 4 on a single 6-sided die, AND then rolling a 1 on a second roll of the die.

Independent Events

upGrad

Multiplication Rule 1: When two events, A and B, are independent, the probability of both occurring is:

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

Blue
brown
red
white
black

red, brown

- **Experiment 1:** A dresser drawer contains one pair of socks with each of the following colors: blue, brown, red, white and black. Each pair is folded together in a matching set. You reach into the sock drawer and choose a pair of socks without looking. You replace this pair and then choose another pair of socks. What is the probability that you will choose the red pair of socks both times?

$P(\text{red}) = \frac{1}{5}$

$P(\text{red and red}) = \frac{1}{5} \cdot \frac{1}{5} = \frac{1}{25}$

Independent Events

upGrad

- **Experiment 2:** A coin is tossed and a single 6-sided die is rolled. Find the probability of landing on the head side of the coin and rolling a 3 on the die.

1 2

$$P(H) = \frac{1}{2} \qquad P(3) = \frac{1}{6}$$

$$P(H \text{ and } 3) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

Combinations Formula:

upGrad

$$C(n, r) = \frac{n!}{(r!(n-r)!)}$$

$${}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

For $n \geq r \geq 0$.

Combinations Formula:

$$C(n,r) = \frac{n!}{(r!(n-r)!)}$$

upGrad

Combination Problem 1
Choose 2 Prizes from a Set of 6 Prizes

$C(6,2) = {}^6C_2 = \frac{6!}{2!(6-2)!}$

You have won first place in a contest and are allowed to choose 2 prizes from a table that has 6 prizes numbered 1 through 6. How many different combinations of 2 prizes could you possibly choose?

- In this example, we are taking a subset of 2 prizes (r) from a larger set of 6 prizes (n). Looking at the formula, we must calculate "6 choose 2."
- C (6,2)= 6!/(2! * (6-2)!) = 6!/(2! * 4!) = 15 Possible Prize Combinations
- The 15 potential combinations are {1,2}, {1,3}, {1,4}, {1,5}, {1,6}, {2,3}, {2,4}, {2,5}, {2,6}, {3,4}, {3,5}, {3,6}, {4,5}, {4,6}, {5,6}

Combinations Formula:

$$C(n,r) = \frac{n!}{(r!(n-r)!)}$$

upGrad

Combination Problem 2
Choose 3 Students from a Class of 25

A teacher is going to choose 3 students from her class to compete in the spelling bee. She wants to figure out how many unique teams of 3 can be created from her class of 25.

In this example, we are taking a subset of 3 students (r) from a larger set of 25 students (n). Looking at the formula, we must calculate "25 choose 3."

C (25,3)= 25!/(3! * (25-3)!)= 2,300 Possible Teams

Binomial Probability Distribution

upGrad

A **binomial experiment** is a statistical experiment that has the following properties:

1. The experiment consists of *n* repeated trials.
2. Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
3. The probability of success, denoted by *P*, is the same on every trial.
4. The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

1. The experiment consists of repeated trials. We flip a coin 2 times.
2. Each trial can result in just two possible outcomes - heads or tails.
3. The probability of success is constant - 0.5 on every trial.
4. The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Binomial Probability Distribution

upGrad

Notation

The following notation is helpful, when we talk about binomial probability.

- *x*: The number of successes that result from the binomial experiment.
- *n*: The number of trials in the binomial experiment.
- *P*: The probability of success on an individual trial.
- *Q*: The probability of failure on an individual trial. (This is equal to 1 - *P*.)
- *n!*: The factorial of n (also known as n factorial).
- b(*x*; *n, P*): Binomial probability - the probability that an *n*-trial binomial experiment results in exactly *x* successes, when the probability of success on an individual trial is *P*.
- ${}_nC_r$: The number of combinations of *n* things, taken *r* at a time.

Binomial Probability Distribution

upGrad

Binomial Formula. Suppose a binomial experiment consists of *n* trials and results in *x* successes. If the probability of success on an individual trial is *P*, then the binomial probability is:

$$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n - x}$$

or

$$b(x; n, P) = \{ n! / [x! (n - x)!] \} * P^x * (1 - P)^{n - x}$$

Example 1

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

Solution: This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is 1/6 or about 0.167. Therefore, the binomial probability is:

$b(x, n, p) \quad {}^nC_x \cdot p^x \cdot (1-p)^{n-x}$

$$b(2; 5, 0.167) = {}_5C_2 * (0.167)^2 * (0.833)^3$$

$$b(2; 5, 0.167) = 0.161$$

Cumulative Binomial Probability

upGrad

A **cumulative binomial probability** refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit).

Example 2

The probability that a student is accepted to a prestigious college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?

Solution: To solve this problem, we compute 3 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

$$b(x \leq 2; 5, 0.3) = b(x = 0; 5, 0.3) + b(x = 1; 5, 0.3) + b(x = 2; 5, 0.3)$$

$$b(x \leq 2; 5, 0.3) = 0.1681 + 0.3601 + 0.3087$$

$$b(x \leq 2; 5, 0.3) = 0.8369$$

Continuous Probability Distributions

Continuous Random Variables

upGrad

Random Variables can be either Discrete or Continuous:

- **Discrete Data** can only take certain values (such as 1,2,3,4,5)

- **Continuous Data** can take any value within a range (such as a person's height)

PDF vs CDF

upGrad

PDF vs CDF

PDF vs CDF

Normal Distribution

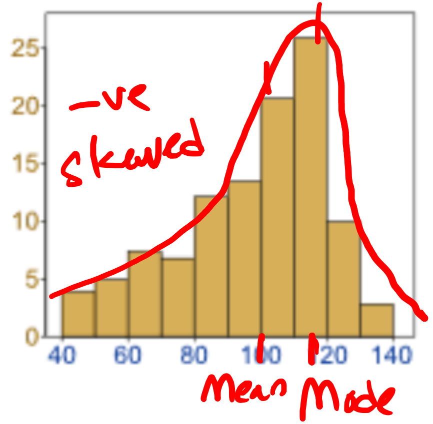
- Data can be "distributed" (spread out) in different ways.

Bellshaped Curve

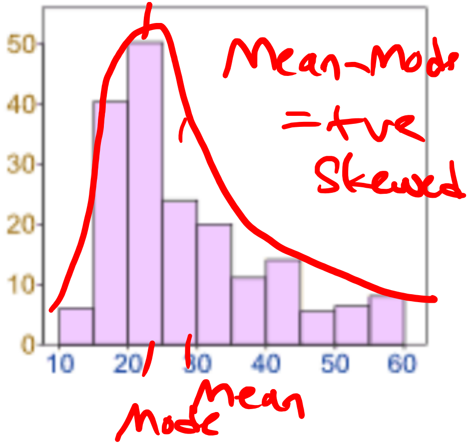
Histogram

Mean_Mode = -ve

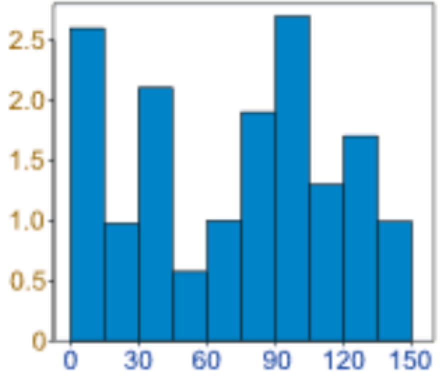
It can be spread out more on the left



Or more on the right

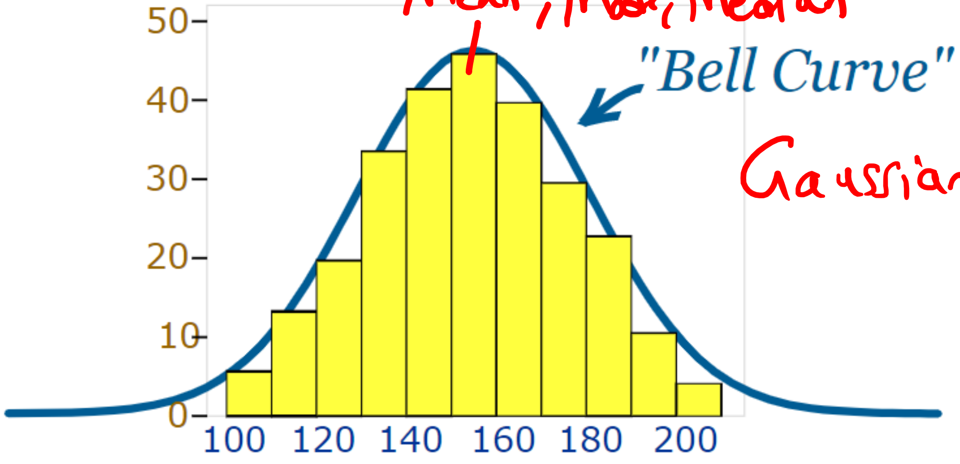


Or it can be all jumbled up



- But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:

Mean, Mode, Median



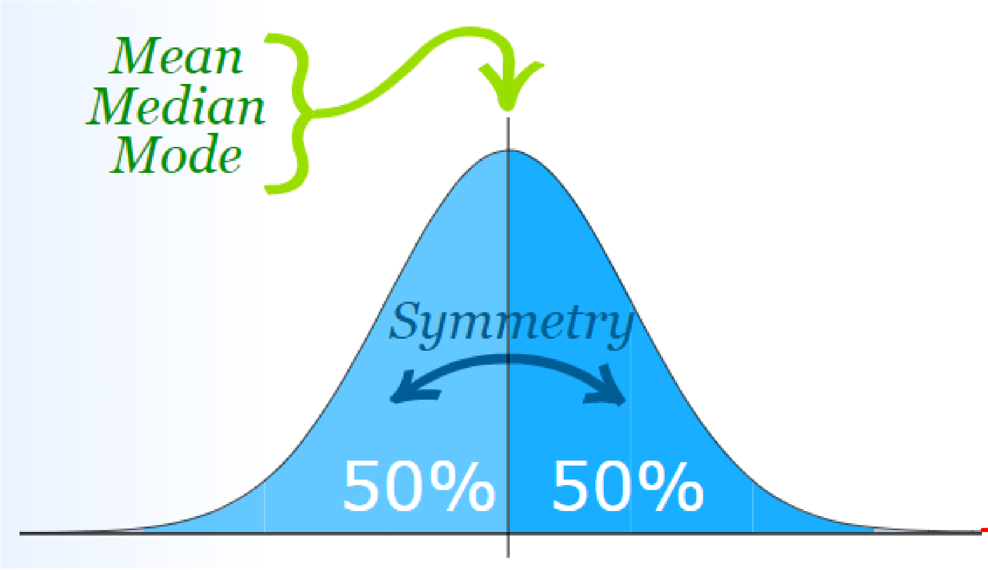
/ Normal Dist / Gaussian Dist.

A Normal Distribution

- It is often called a "Bell Curve" because it looks like a bell.
- Many things closely follow a Normal Distribution:

1. heights of people
2. size of things produced by machines
3. errors in measurements
4. blood pressure
5. marks on a test

- We say the data is "normally distributed":



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

Normal Distribution

- **Quincunx**
- You can see a normal distribution being created by random chance!
- It is called the Quincunx and it is an amazing machine.
- Have a play with it!
- https://www.youtube.com/watch?v=3m4bxse2JEQ
- https://www.transum.org/Maths/Activity/PhET/Plinko_Probability.asp

Standard Deviation

upGrad

- The Standard Deviation is a measure of how spread out numbers are.
- Its symbol is **σ** (the greek letter sigma)
- The formula is easy: it is the **square root** of the **Variance.** So now you ask, "What is the Variance?"

Variance

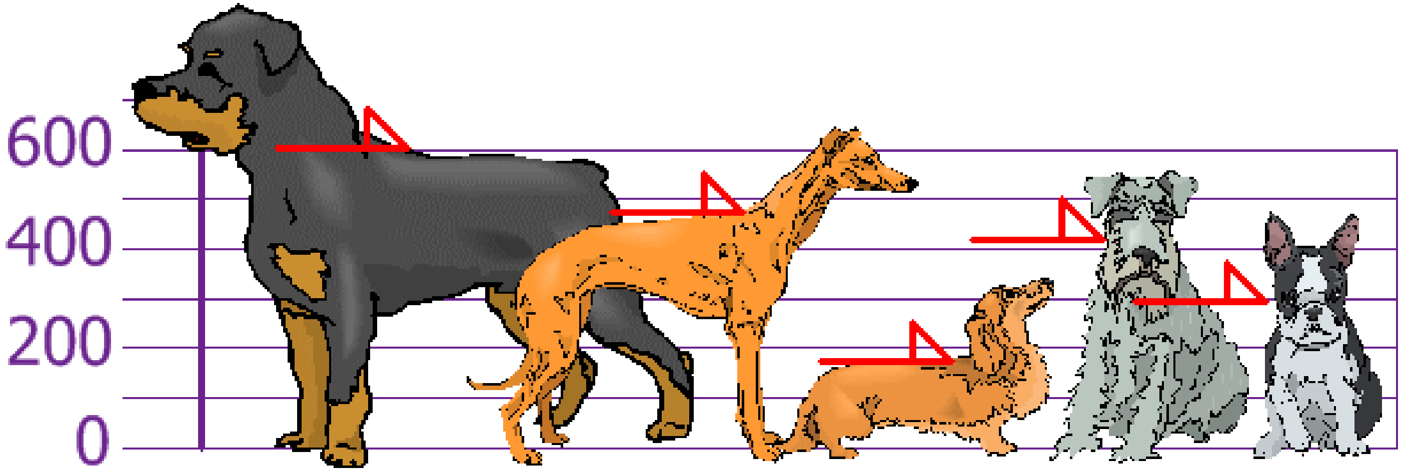
- The Variance is defined as:

  The average of the **squared** differences from the Mean.

- To calculate the variance follow these steps:
  - Work out the Mean
  - Then for each number: subtract the Mean and square the result
  - Then work out the average of those squared differences.

Example

- You and your friends have just measured the heights of your dogs (in millimeters):



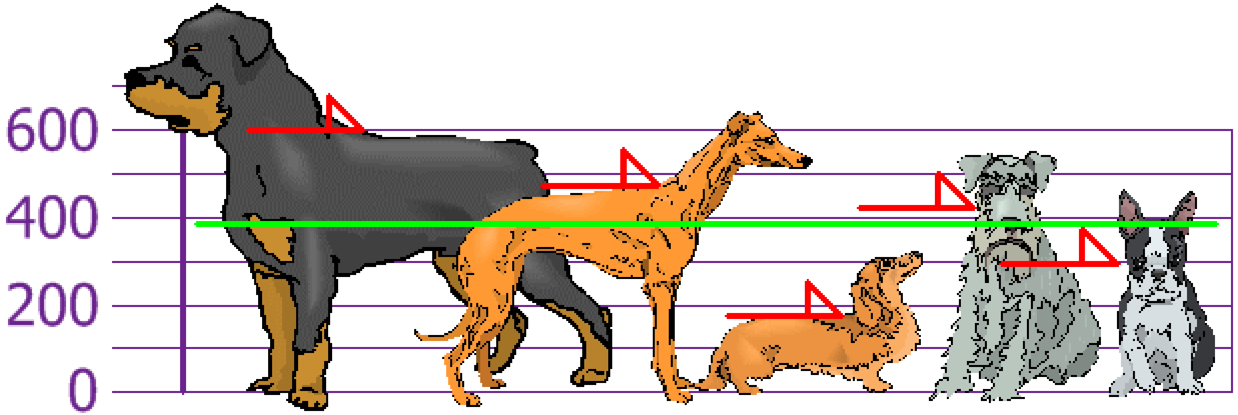
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.
- Find out the Mean, the Variance, and the Standard Deviation.

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

$$= \frac{1970}{5}$$

$$= 394$$

- so the mean (average) height is 394 mm. Let's plot this on the chart:



Example

Variance

$$
\begin{aligned}
\sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\
&= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} \\
&= \frac{108520}{5} \\
&= 21704
\end{aligned}
$$

- So the Variance is **21,704**

Example

upGrad

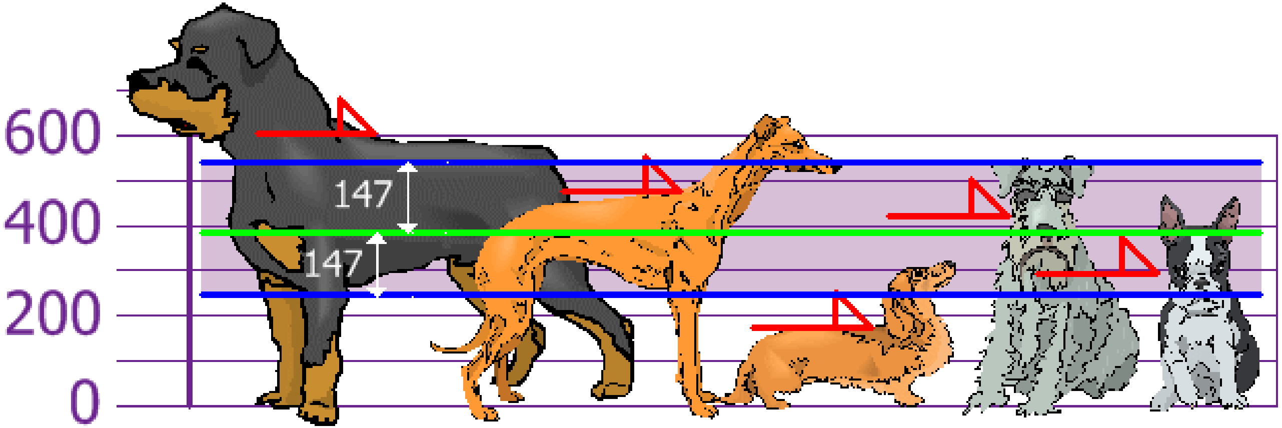
- And the Standard Deviation is just the square root of Variance, so:
- Standard Deviation

σ	=	$\sqrt{21704}$
	=	147.32...
	=	**147** *(to the nearest mm)*

Example

upGrad

- And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



- So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.
- Rottweilers **are** tall dogs. And Dachshunds **are** a bit short, right?

Formulaes

upGrad

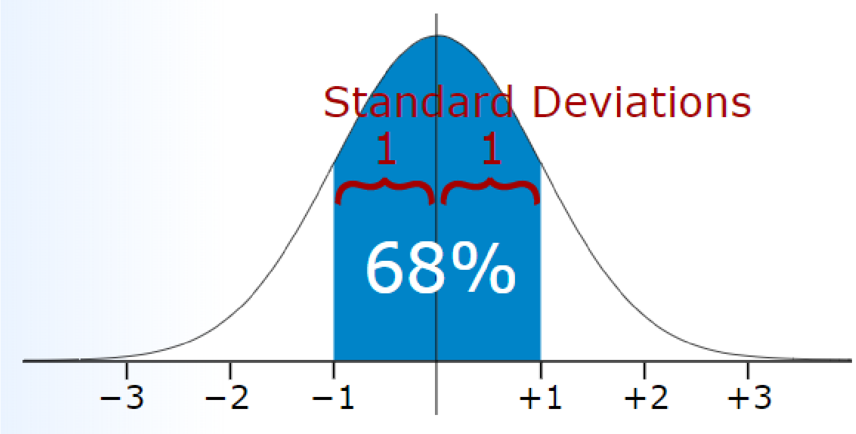
The "**Population** Standard Deviation": $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

The "**Sample** Standard Deviation**"**: $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$

- Looks complicated, but the important change is to divide by **N-1** (instead of **N**) when calculating a Sample Variance.

Range of Standard Deviation

- We can expect about 68% of values to be within plus-or-minus 1 standard deviation.

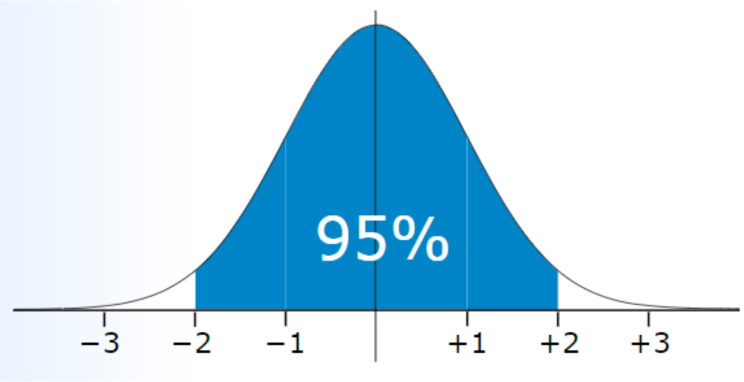


68% of values are within

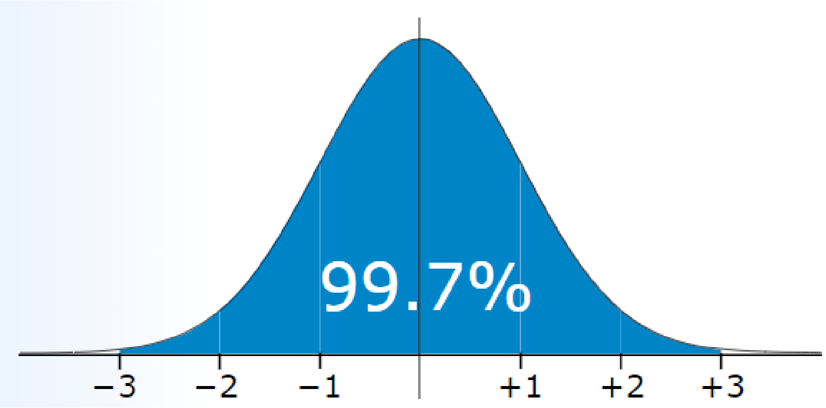
1 standard deviation of the mean

Range of Standard Deviation

upGrad

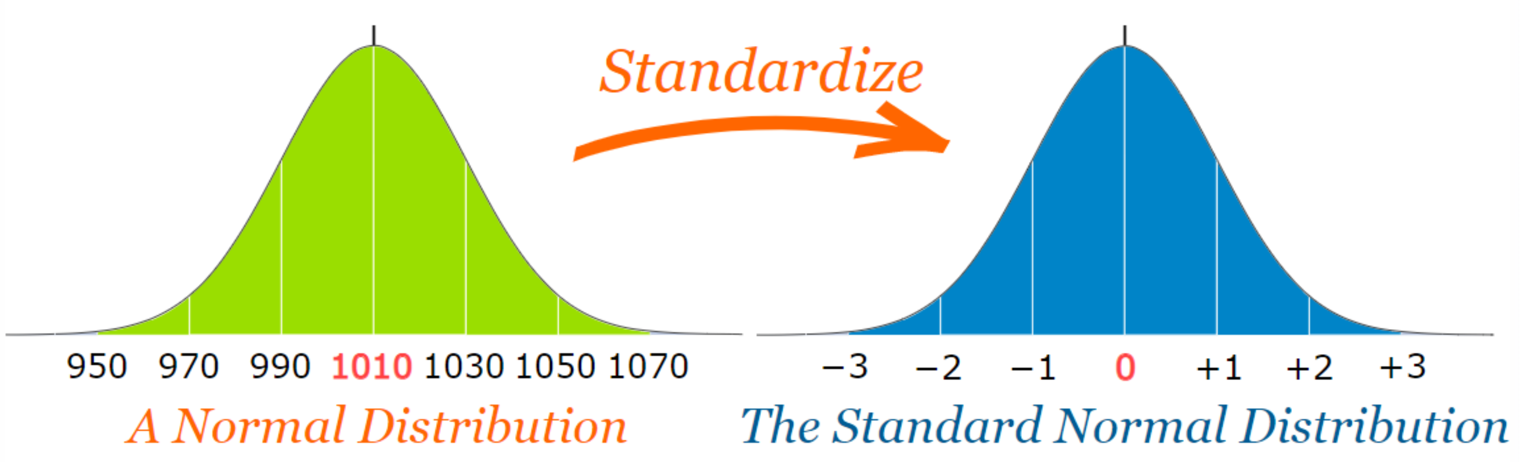


95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

- So to convert a value to a Standard Score ("z-score"):
  - first subtract the mean,
  - then divide by the Standard Deviation
- And doing that is called "Standardizing":



Standard Scores

upGrad

- We can take any Normal Distribution and convert it to The Standard Normal Distribution.
- **Example:** Travel Time
- A survey of daily travel time had these results (in minutes):
  - 26, 33, 65, 28, 34, 55, 25, 44, 50, 36, 26, 37, 43, 62, 35, 38, 45, 32, 28, 34
- The **Mean is 38.8 minutes**, and the **Standard Deviation is 11.4 minutes**
- Convert the values to z-scores ("standard scores").

- To convert **26**:
- first subtract the mean: $26 - 38.8 = -12.8$,
- then divide by the Standard Deviation: $-12.8/11.4 = -1.12$
- So **26** is **−1.12 Standard Deviations** from the Mean

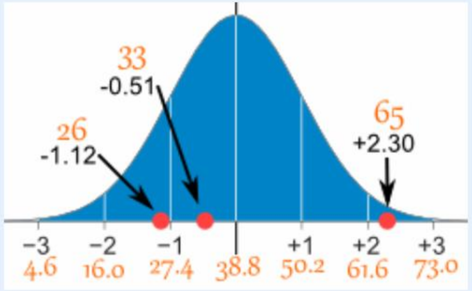
Standard Scores

upGrad

- Here are the first three conversions

Original Value	Calculation	Standard Score (z-score)
26	(26-38.8) / 11.4 =	**−1.12**
33	(33-38.8) / 11.4 =	**−0.51**
65	(65-38.8) / 11.4 =	**+2.30**
...	...	...

And here they are graphically:



Z-Score formula

upGrad

$$z = \frac{x - \mu}{\sigma}$$

- **z** is the "z-score" (Standard Score)
- **x** is the value to be standardized
- **μ** ('mu") is the mean
- **σ** ("sigma") is the standard deviation

Central Limit Theorem

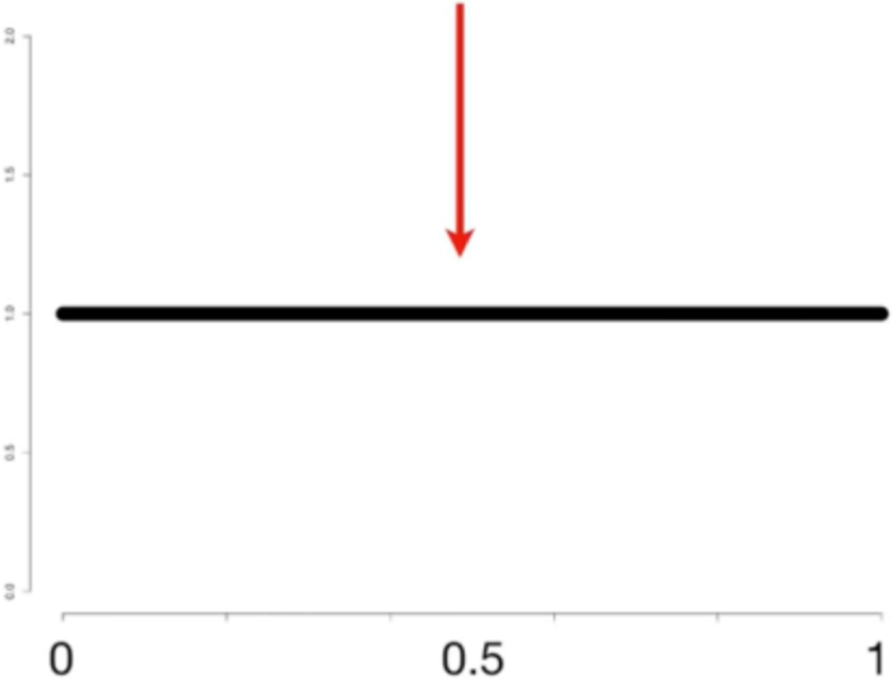
The Central Limit Theorem is the basis for a lot of statistics and the good news is that it is a pretty simple concept.

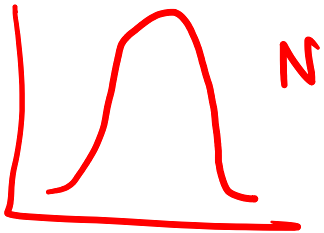
Like most things in statistics, I think The Central Limit Theorem is easiest to understand if we look at some examples.

Central Limit Theorem

upGrad

So let's start with a Uniform Distribution.

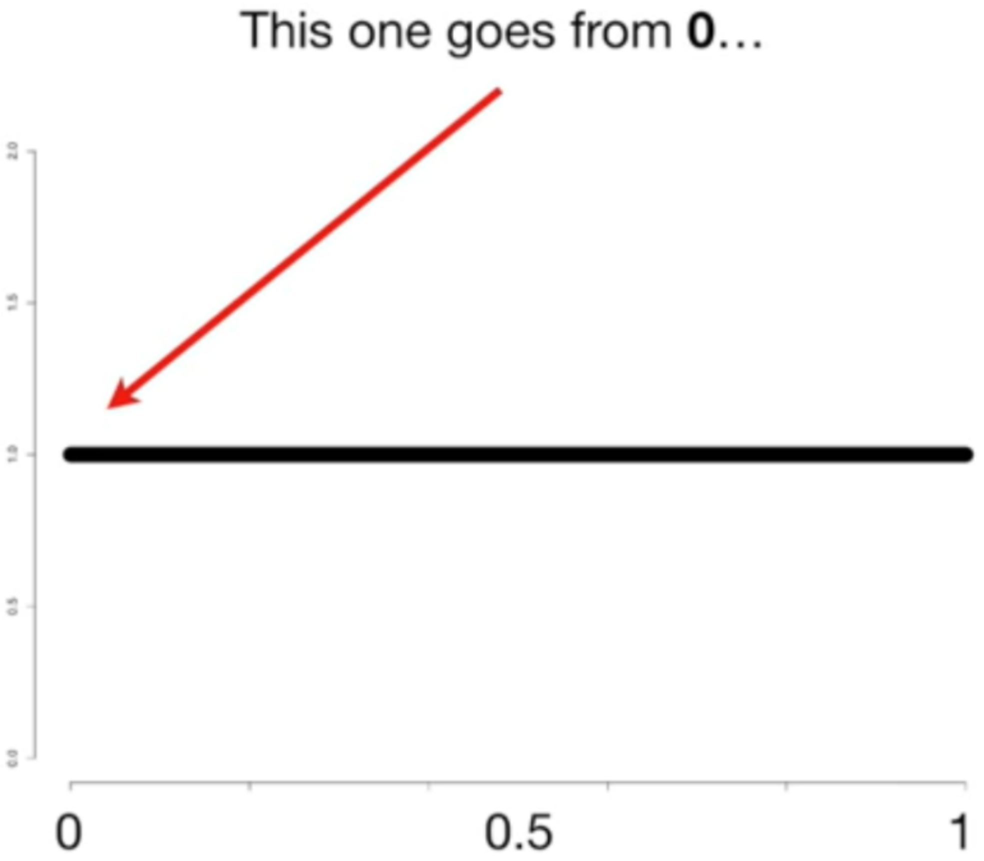




Normal Dist.

Central Limit Theorem

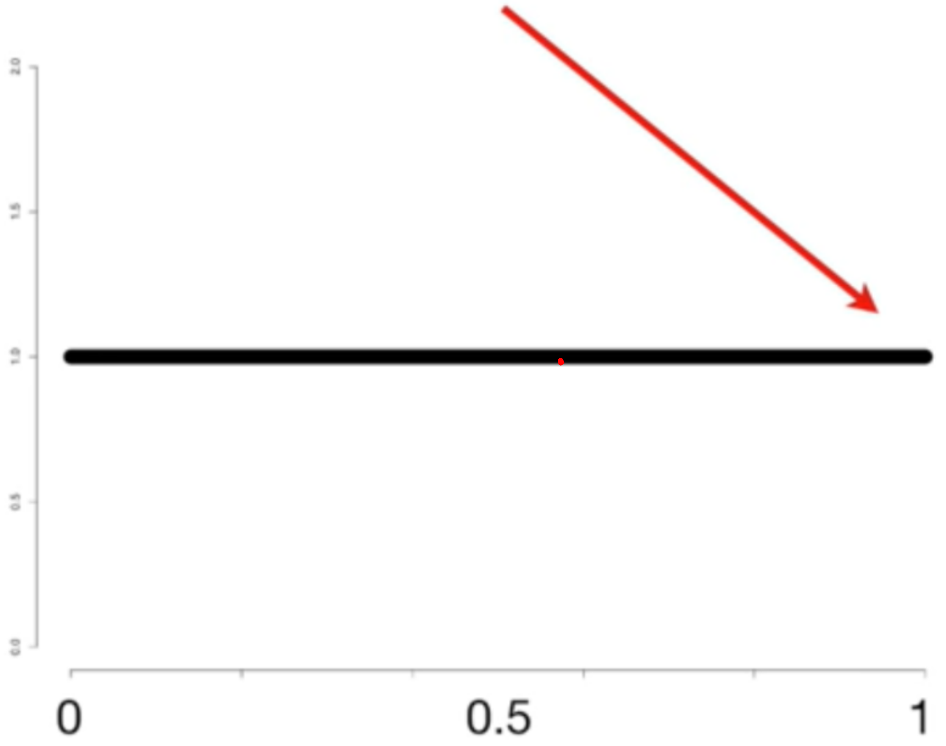
upGrad



Central Limit Theorem

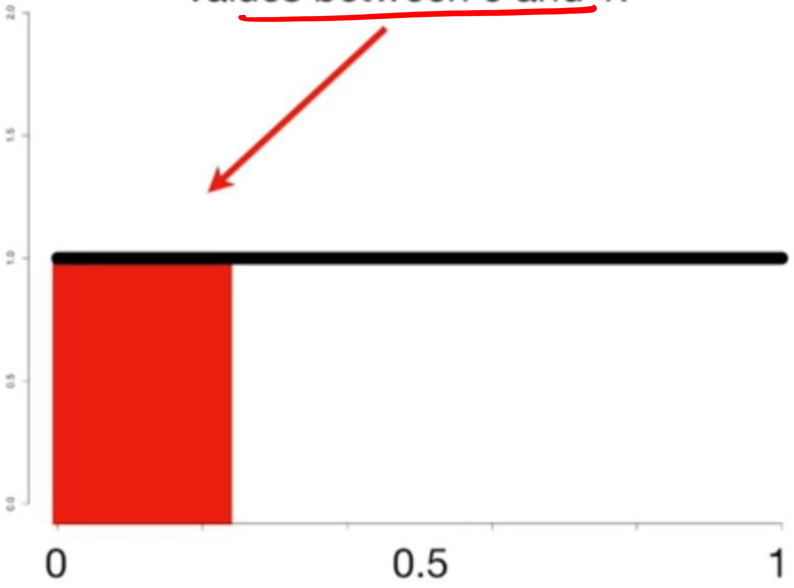
upGrad

...to **1**.



Central Limit Theorem

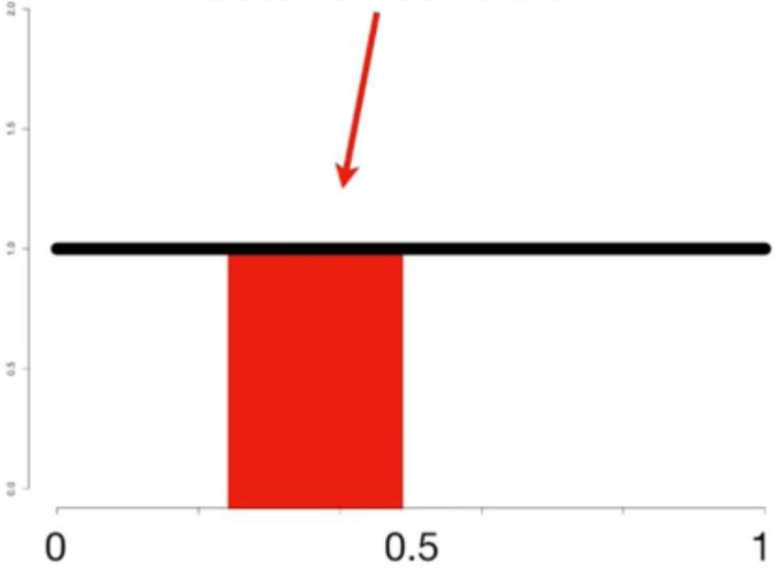
It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.



Central Limit Theorem

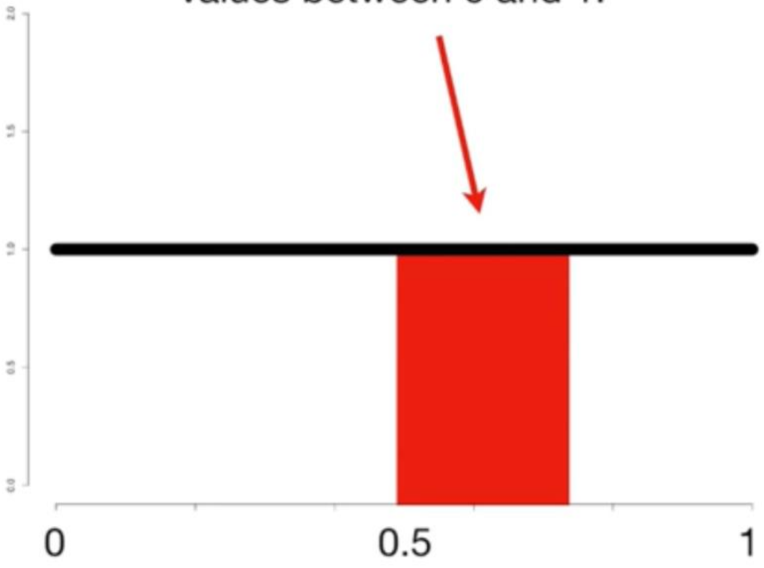
upGrad

It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.



Central Limit Theorem

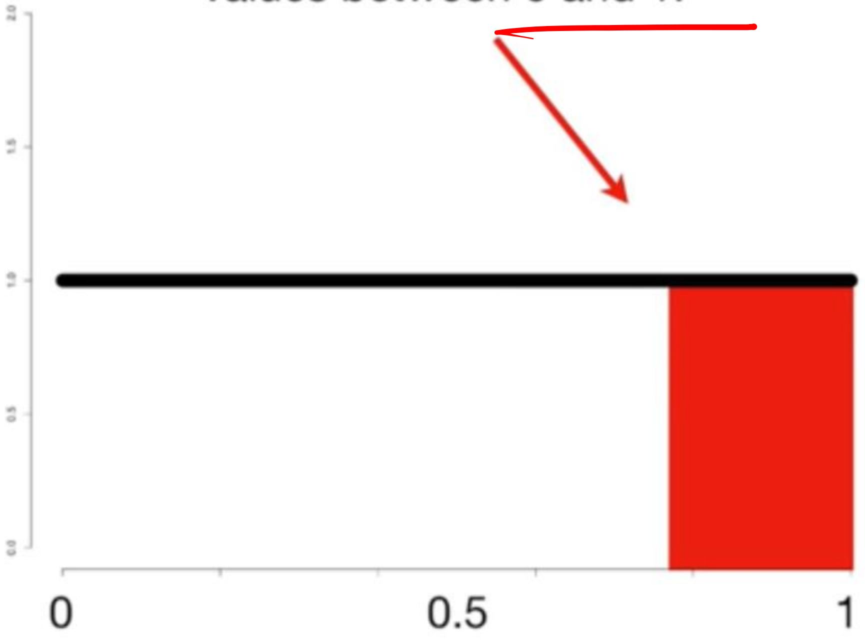
It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.



Central Limit Theorem

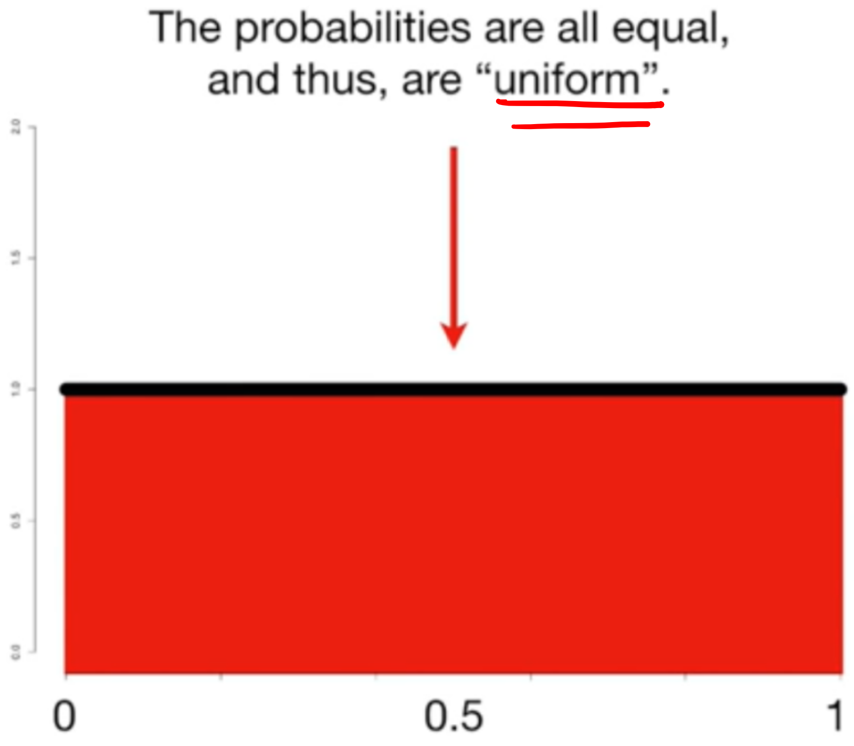
upGrad

It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.



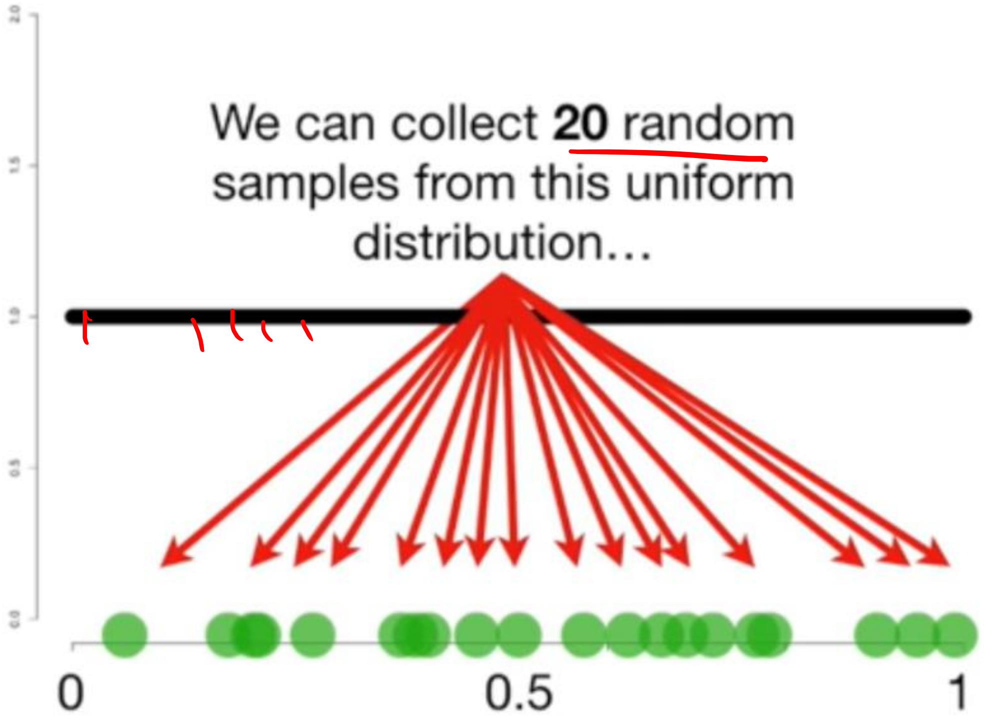
Central Limit Theorem

upGrad

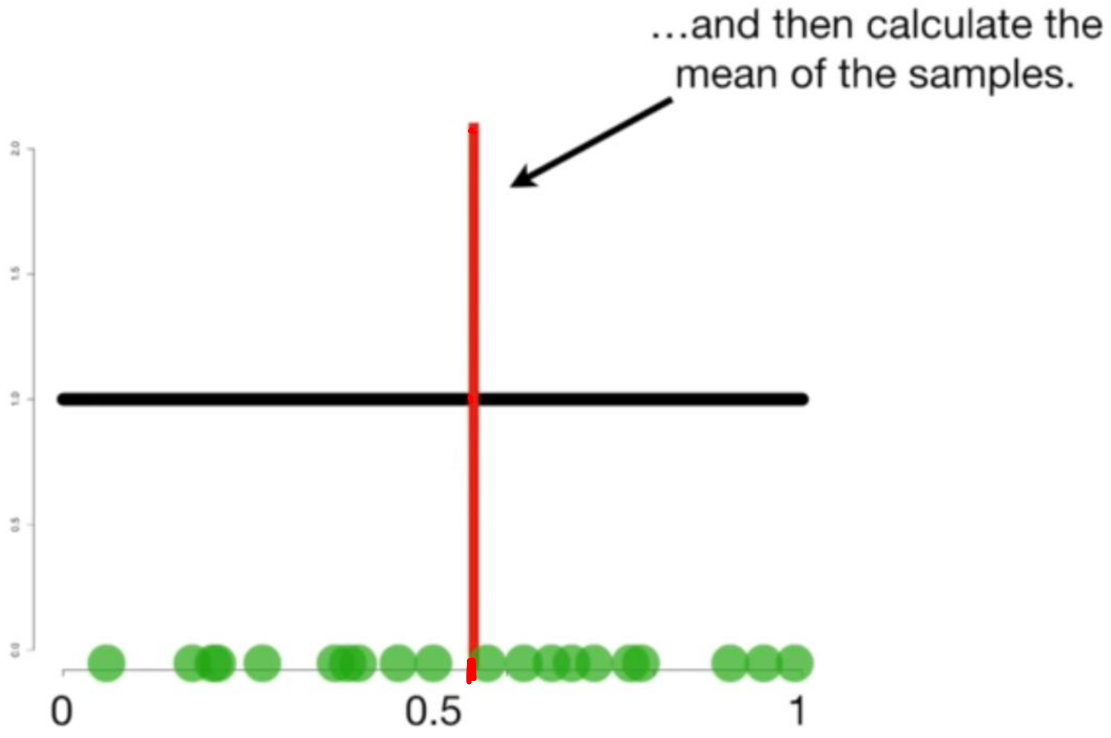


Central Limit Theorem

upGrad



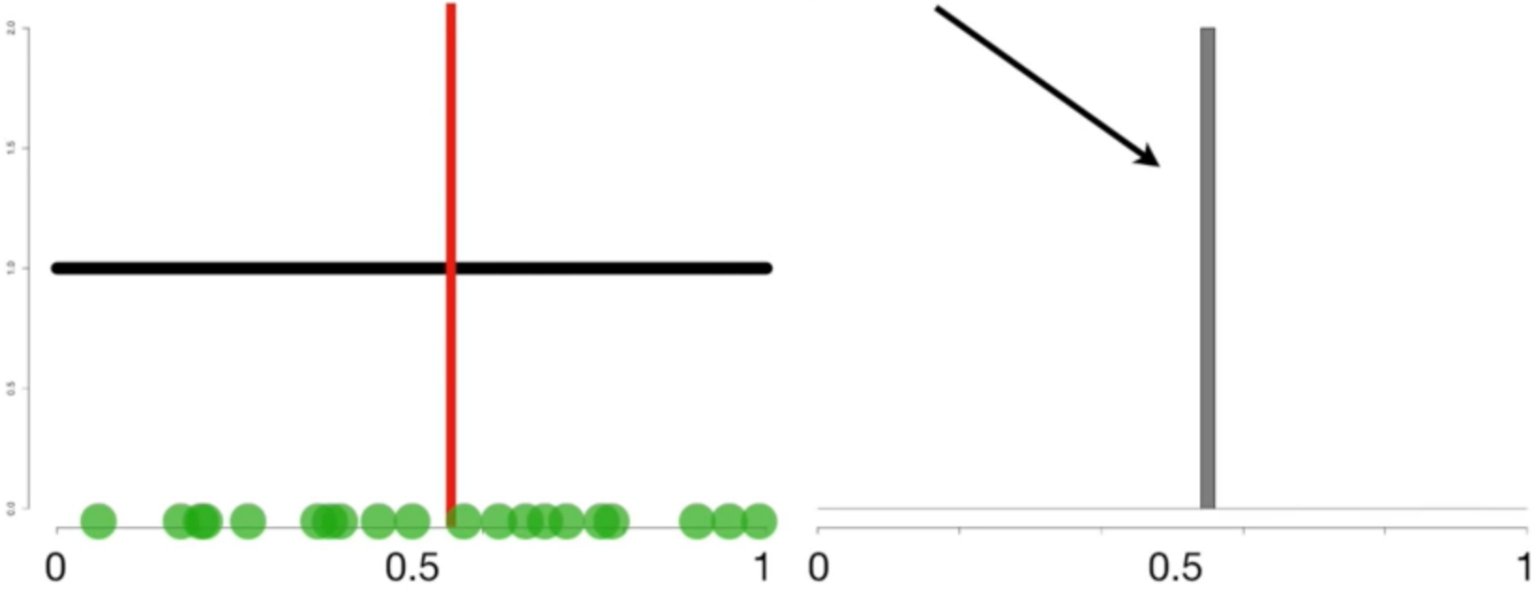
upGrad



Central Limit Theorem

upGrad

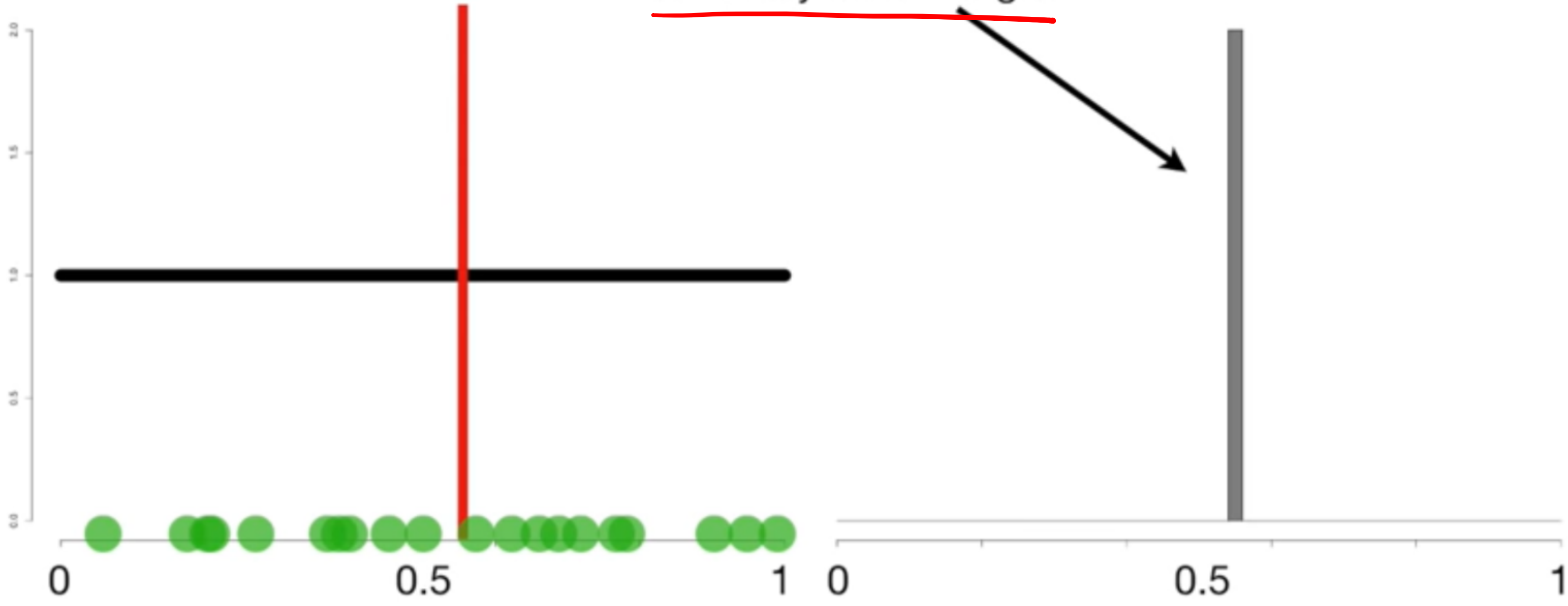
On the right, we can draw a histogram of the mean value.



Central Limit Theorem

upGrad

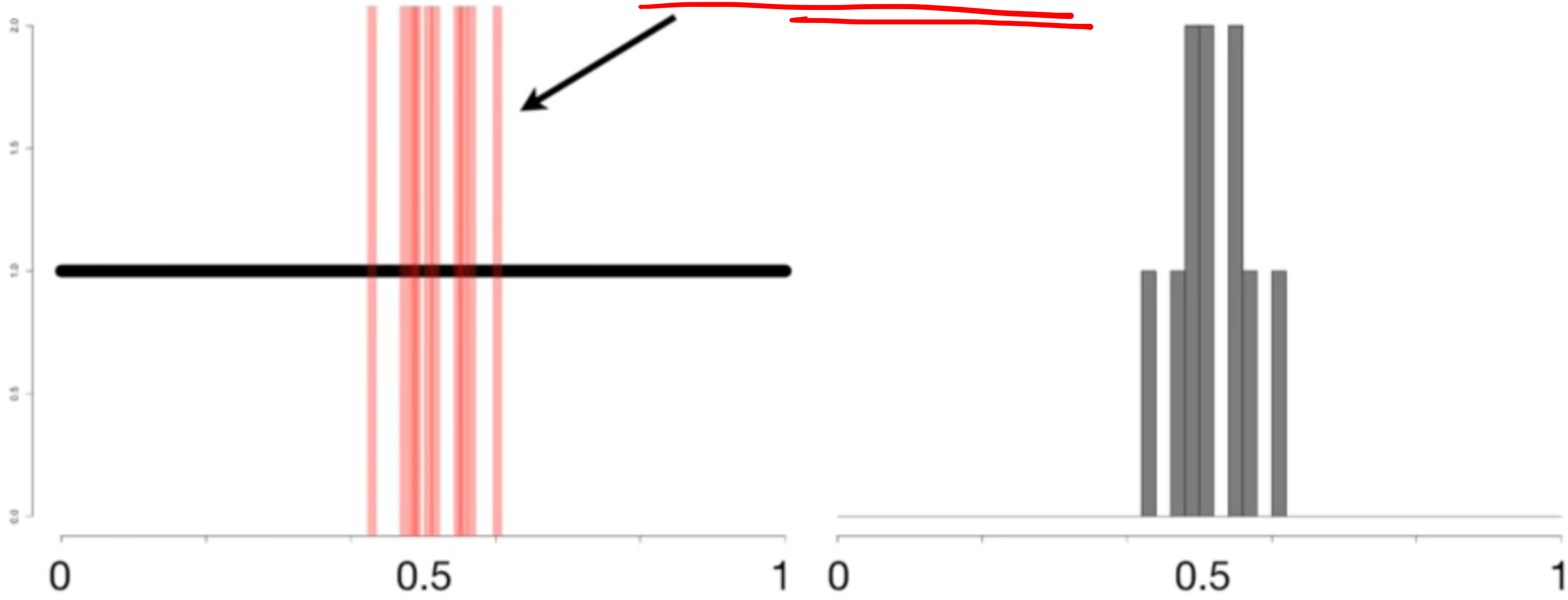
Since we only have one mean value, the histogram isn't very interesting…



Central Limit Theorem

upGrad

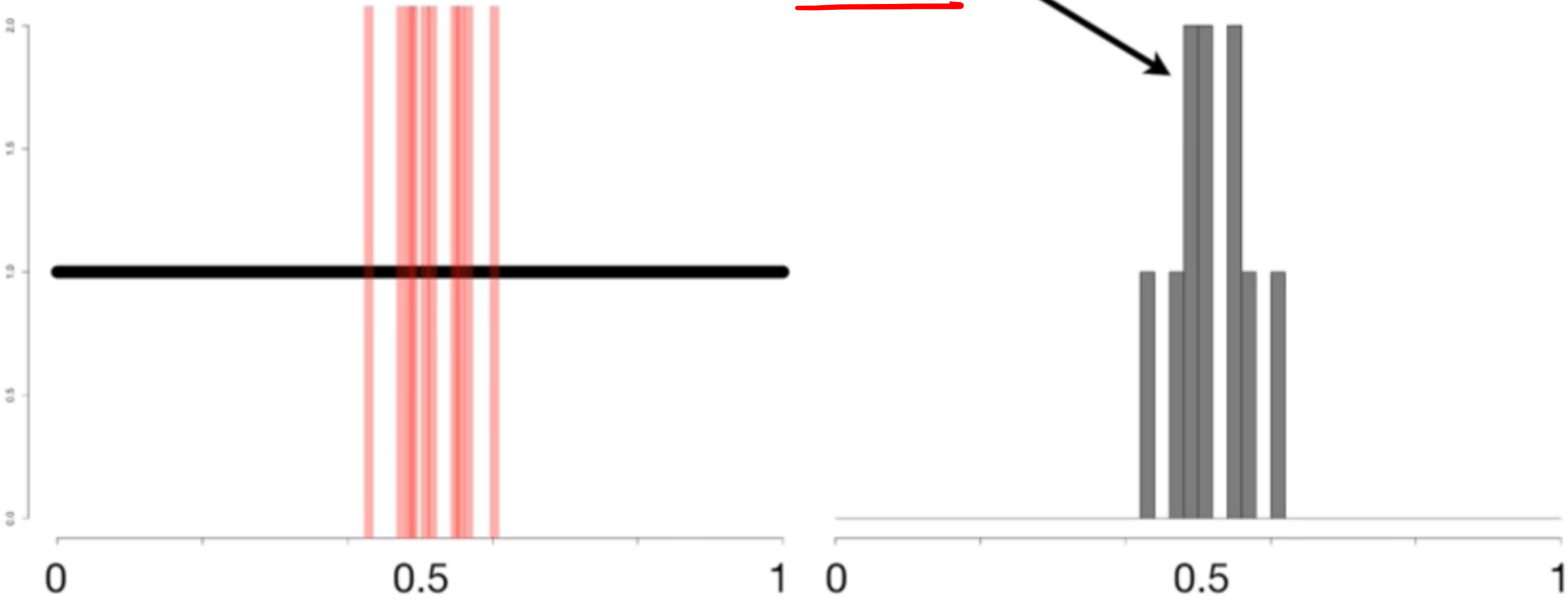
...but after we collect 10 more samples and calculate 10 more means...



Central Limit Theorem

upGrad

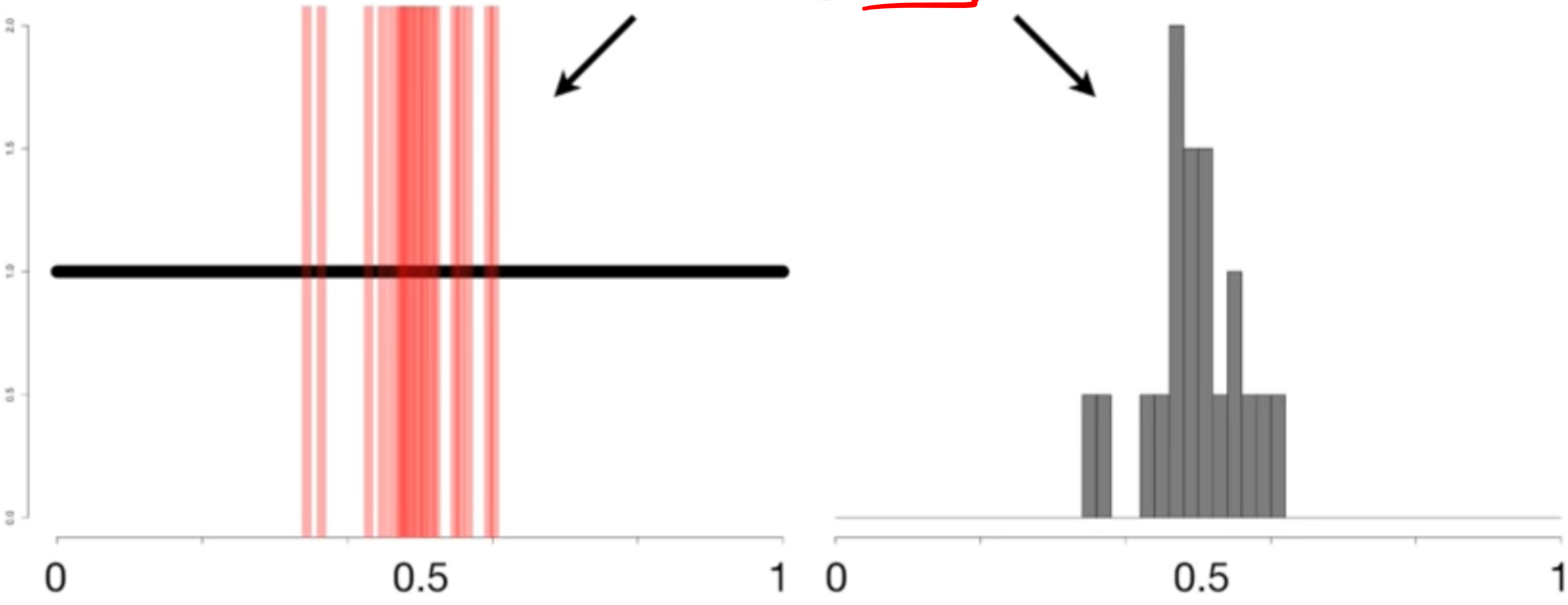
...the histogram starts to look at little more interesting.



Central Limit Theorem

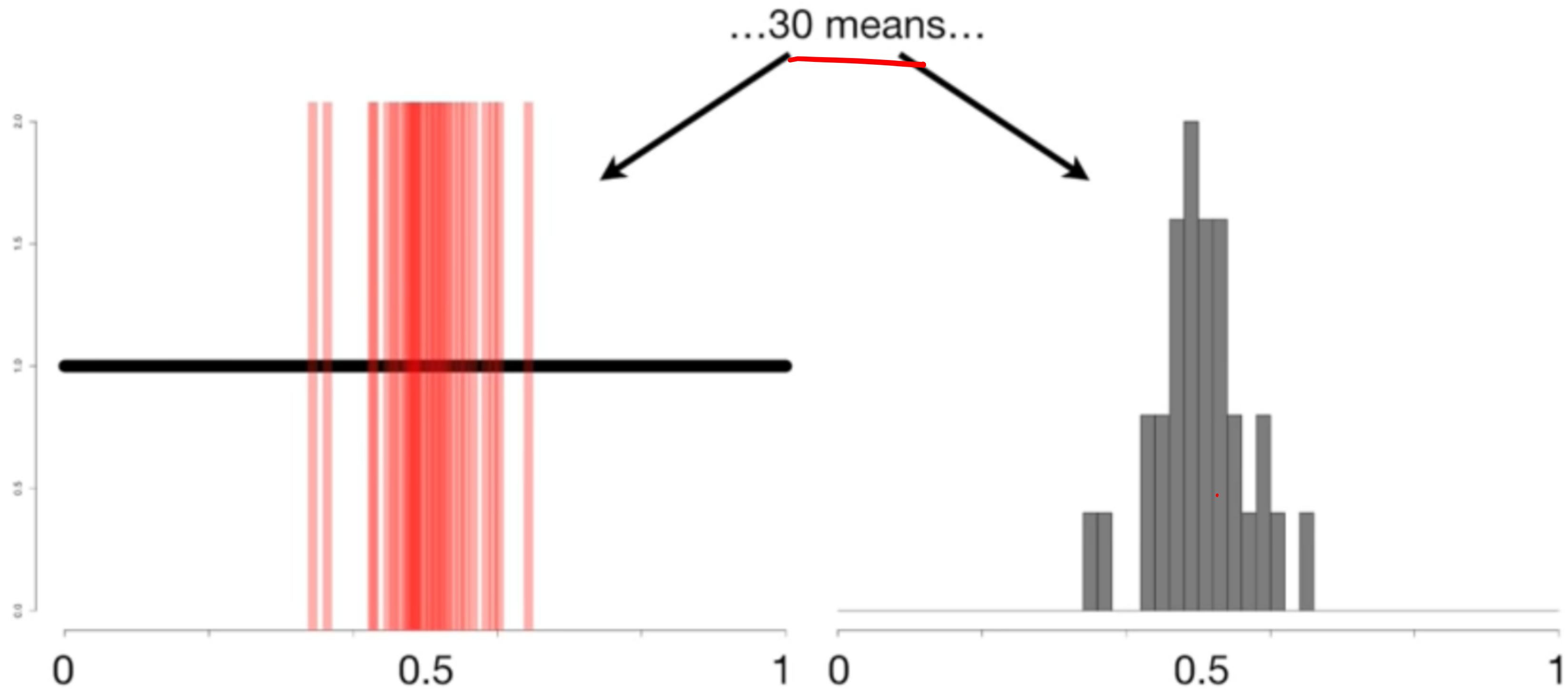
upGrad

Here's the histogram after collecting 20 samples and calculating 20 means…



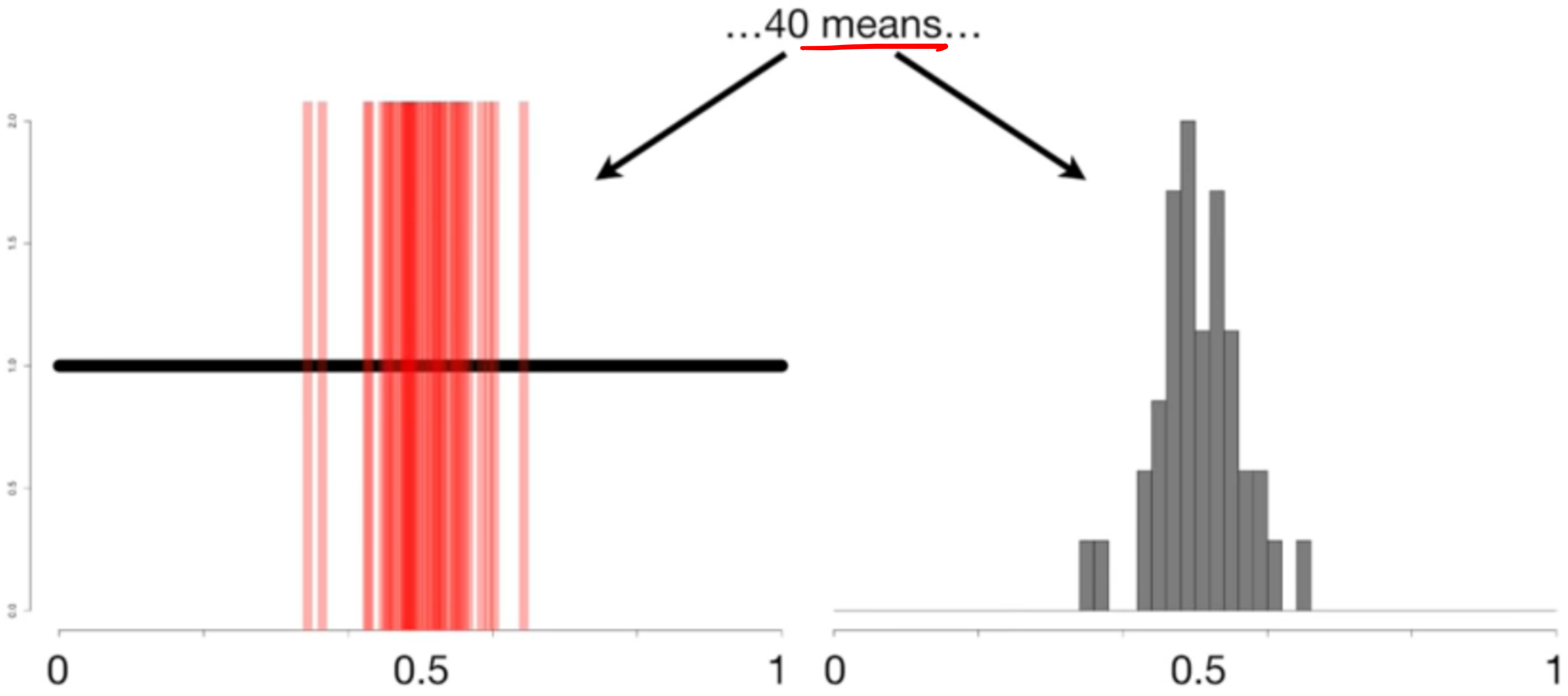
Central Limit Theorem

upGrad

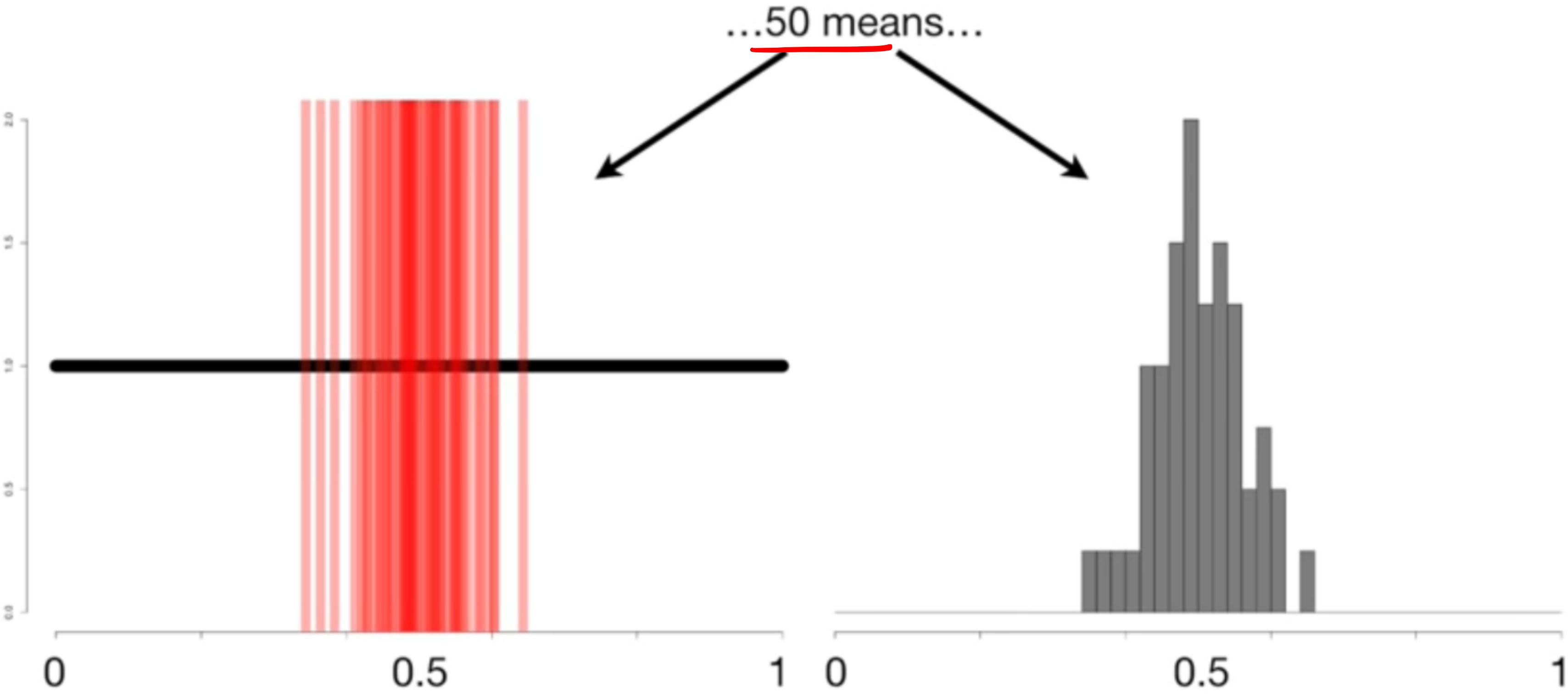


Central Limit Theorem

upGrad

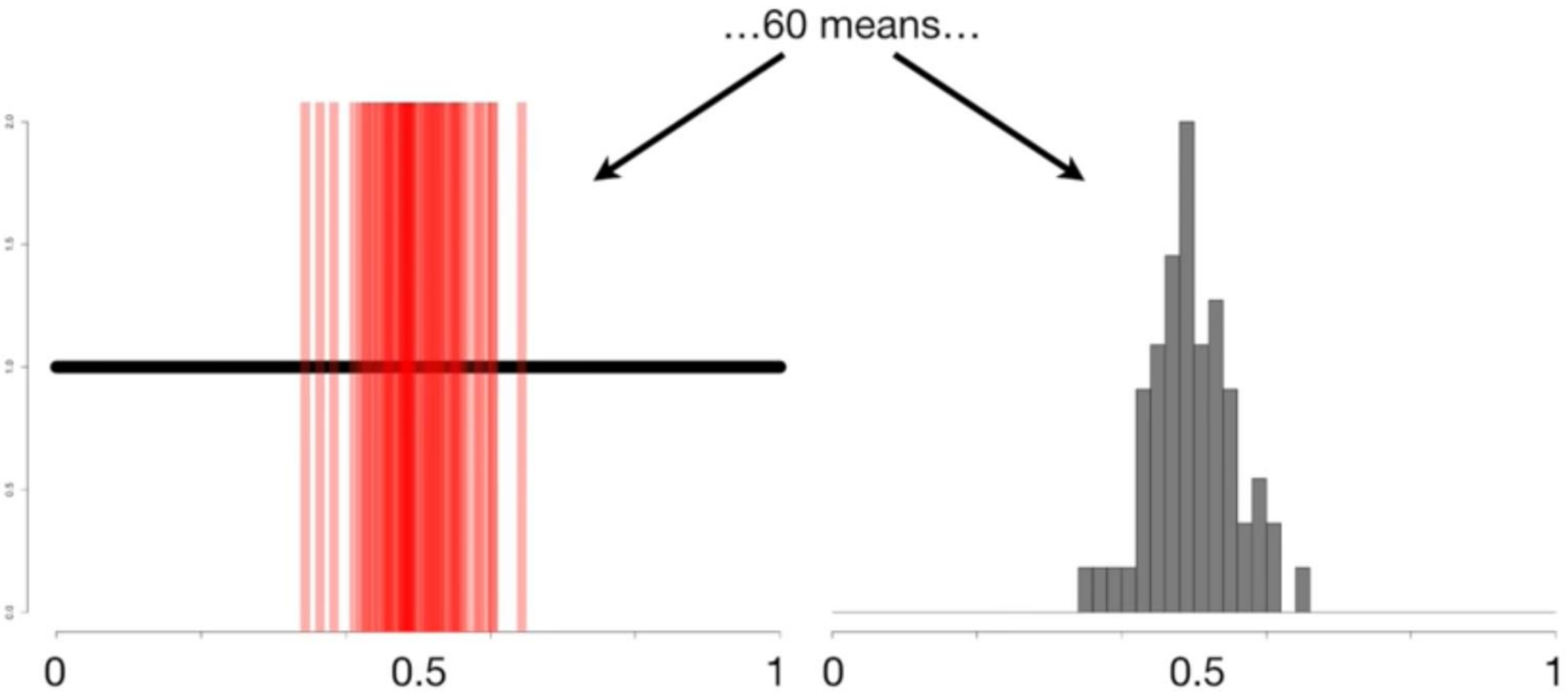


Central Limit Theorem

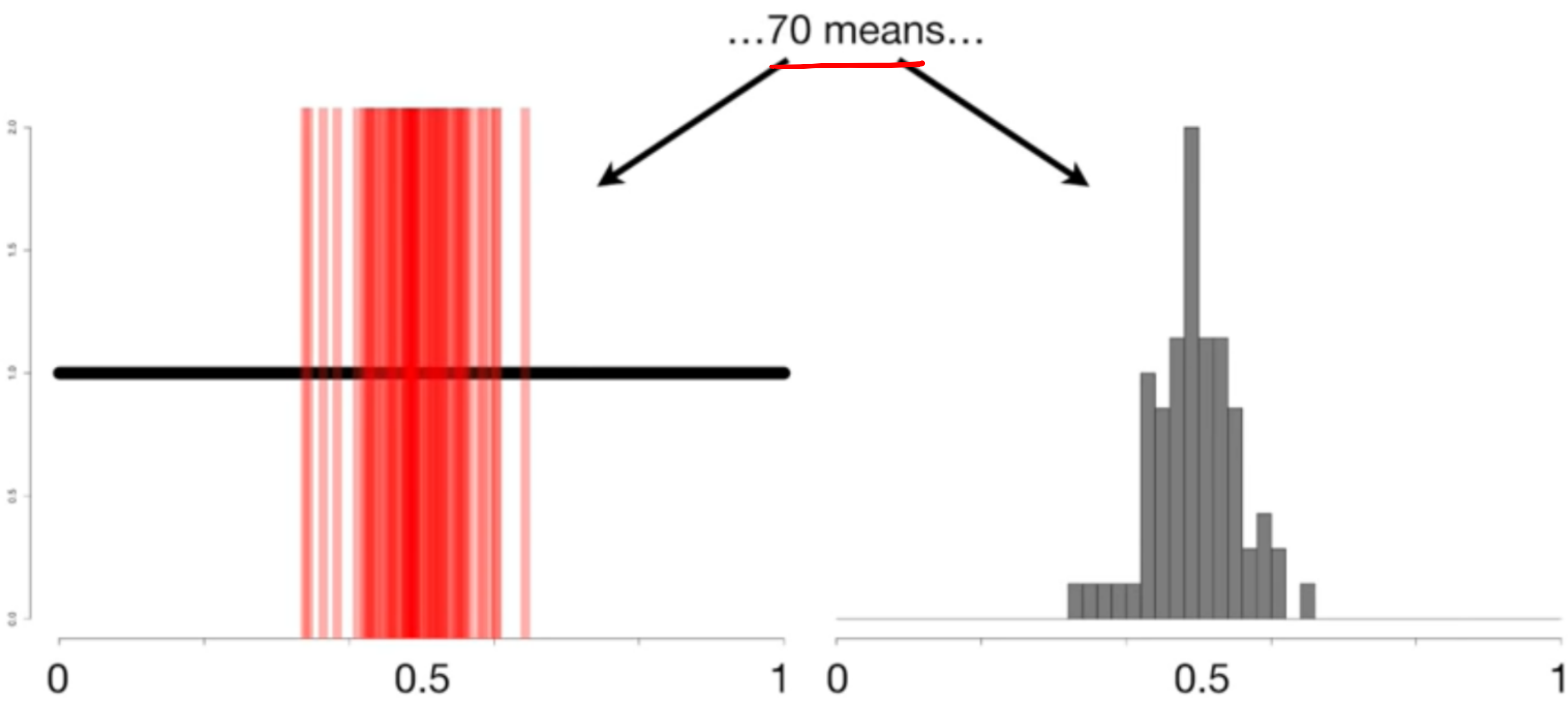


Central Limit Theorem

upGrad



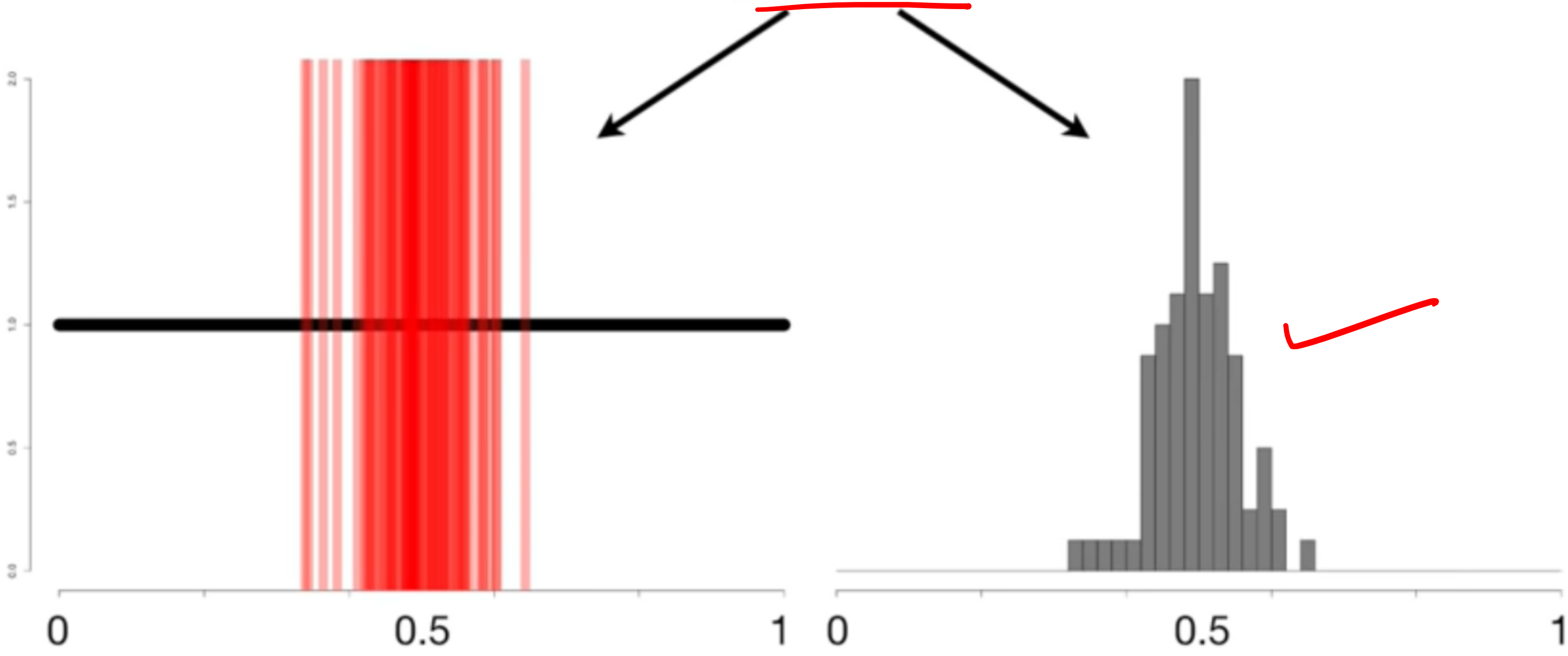
Central Limit Theorem



Central Limit Theorem

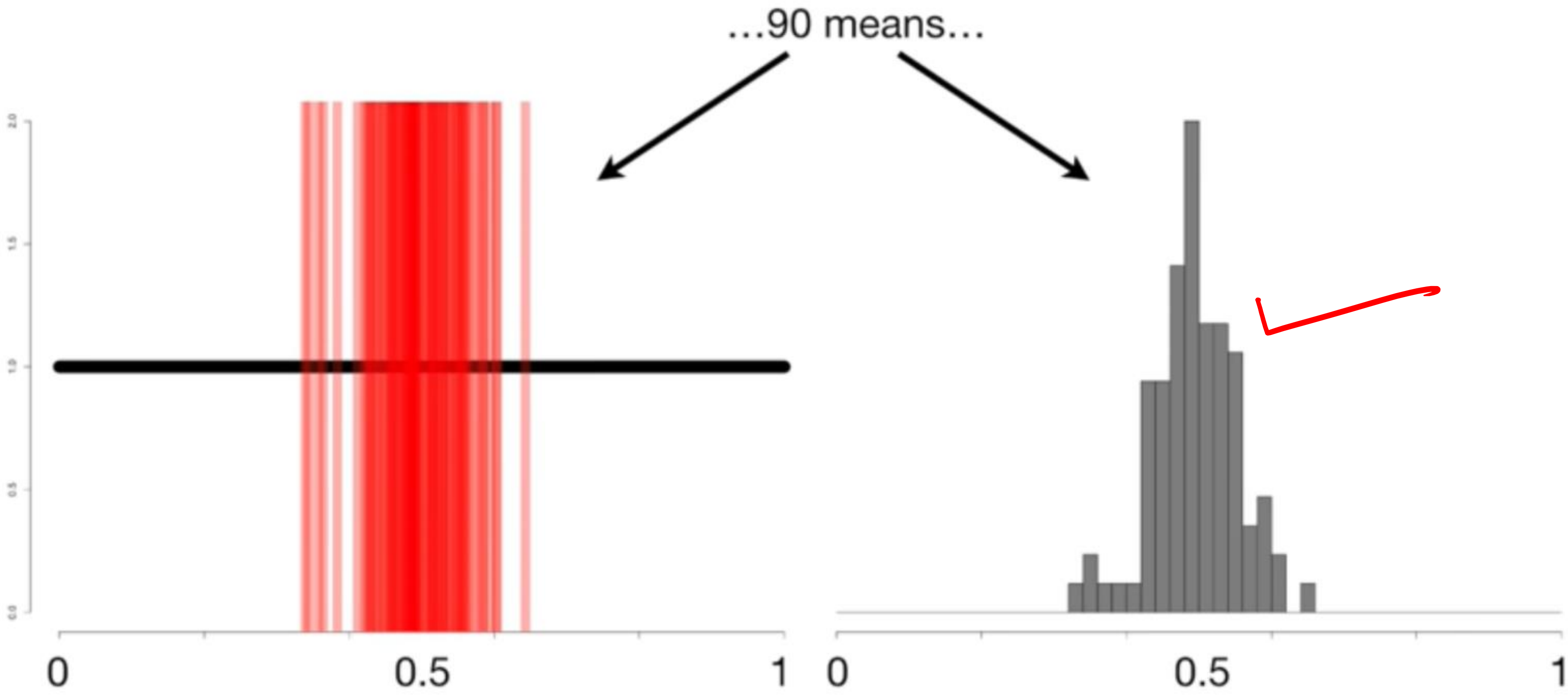
upGrad

...80 means...



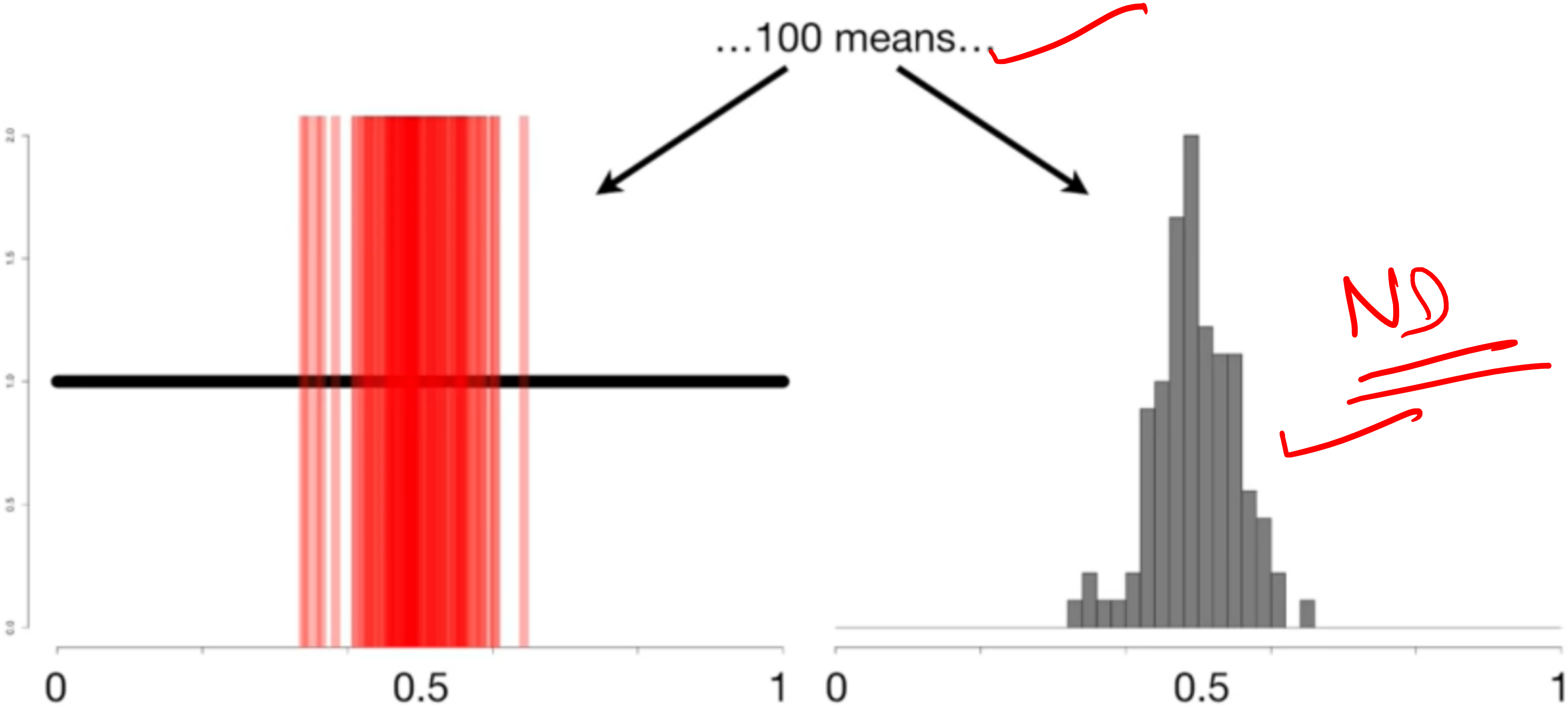
Central Limit Theorem

upGrad



Central Limit Theorem

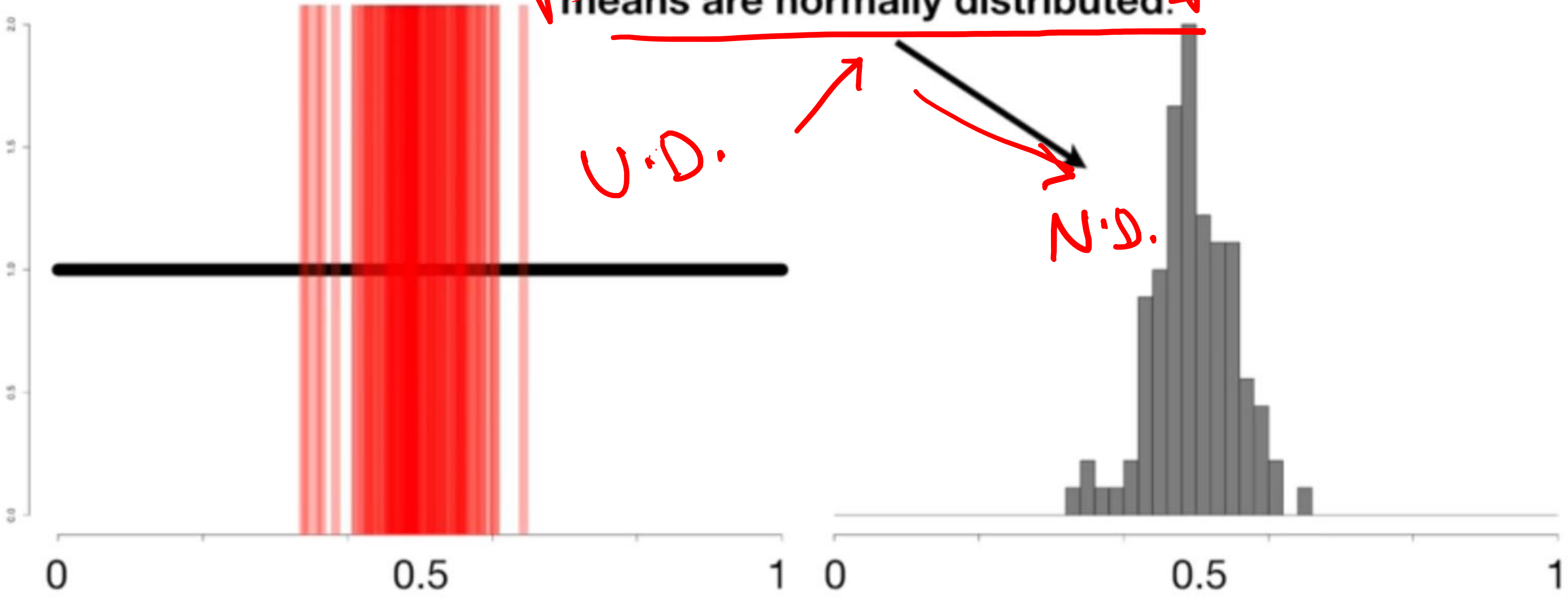
upGrad



Central Limit Theorem

upGrad

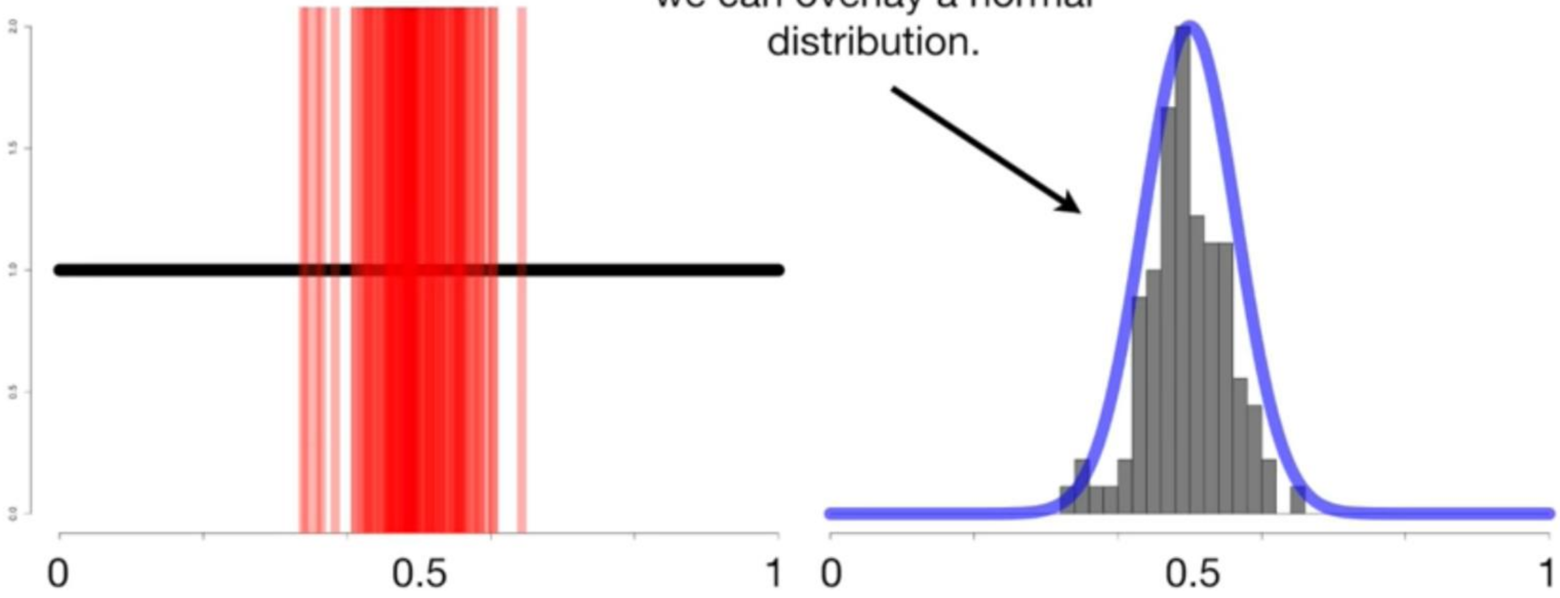
After adding 100 means to the histogram, it's pretty easy to see that these **means are normally distributed.**



Central Limit Theorem

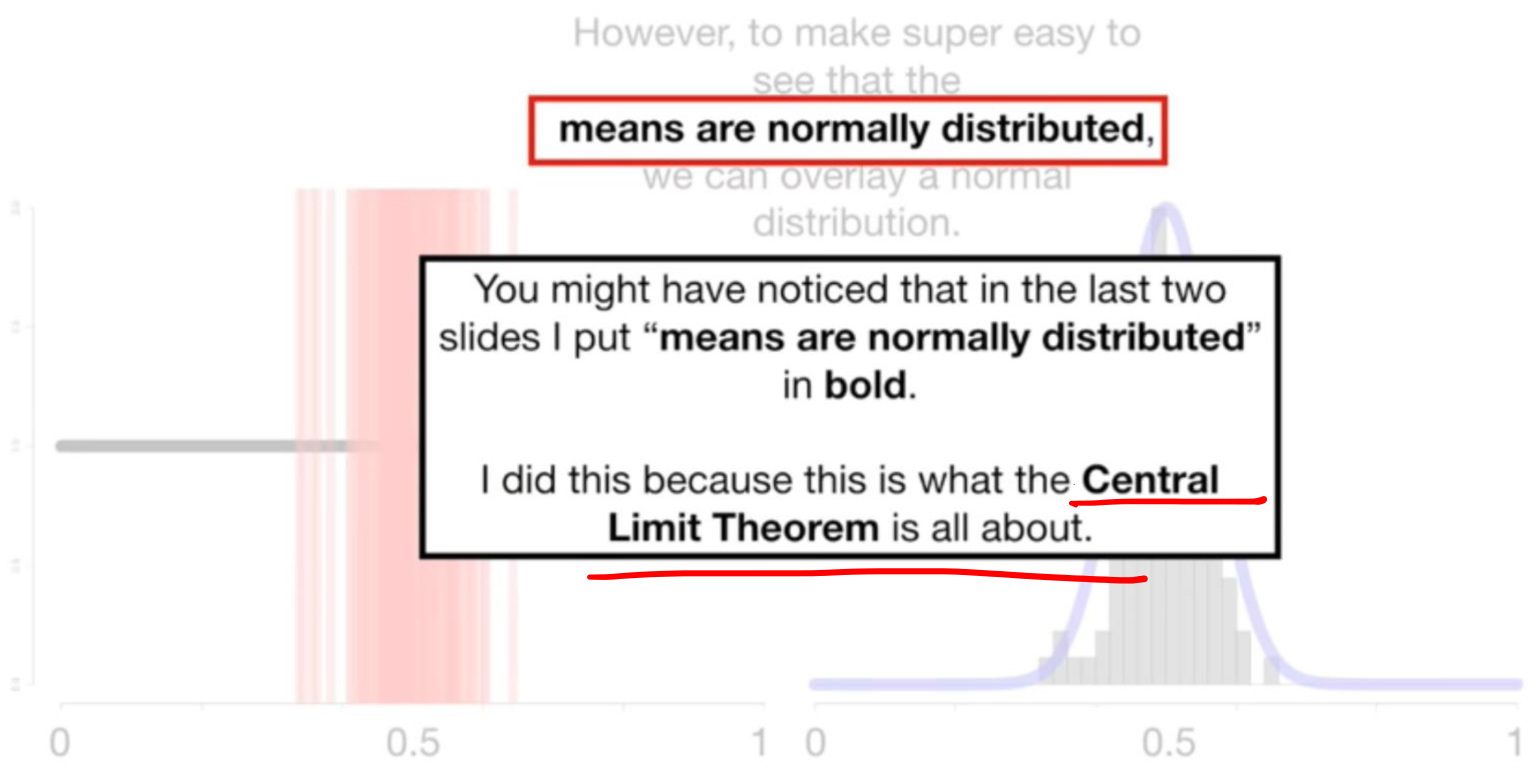
upGrad

However, to make super easy to see that the **means are normally distributed**, we can overlay a normal distribution.



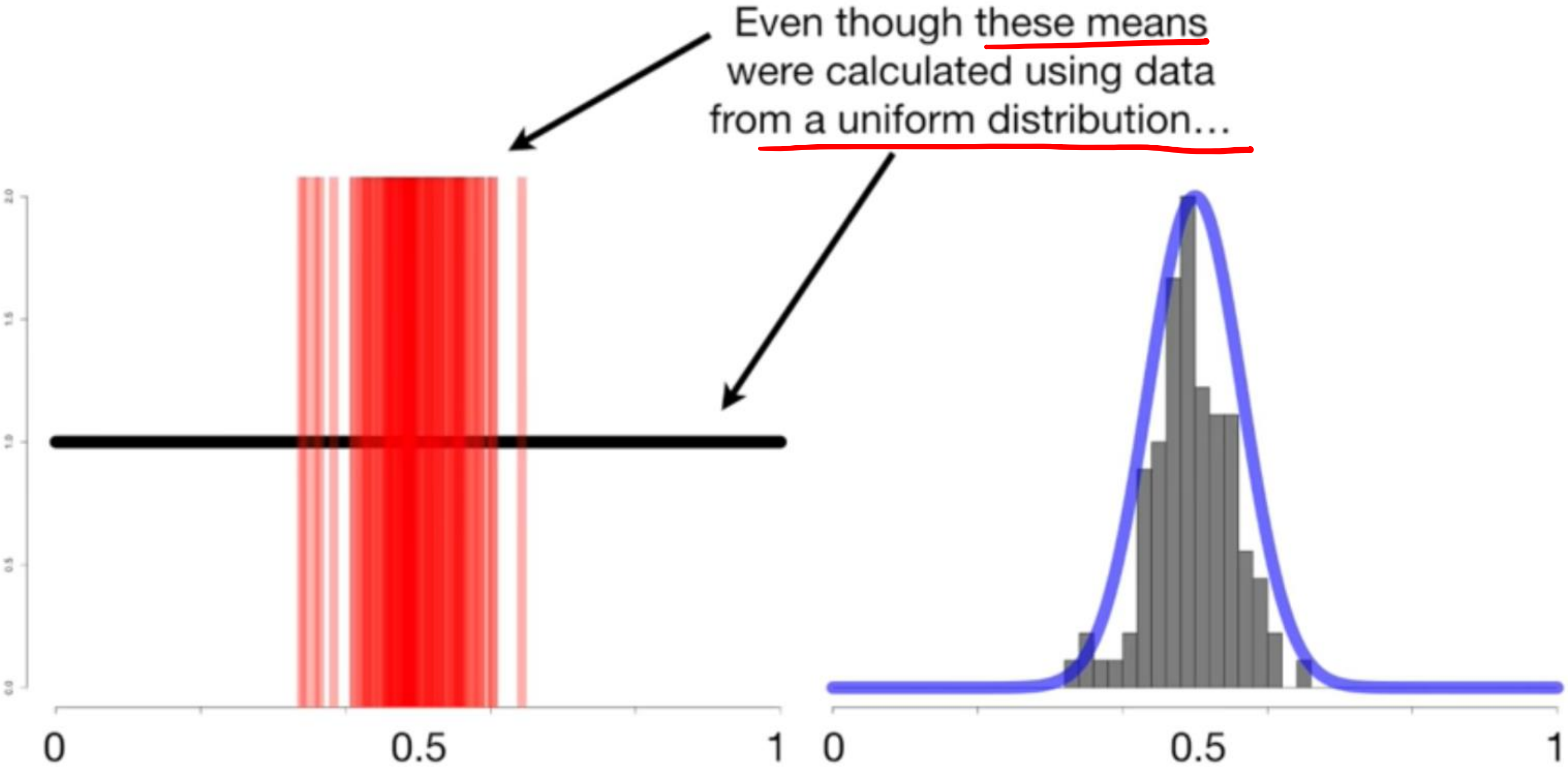
Central Limit Theorem

upGrad



Central Limit Theorem

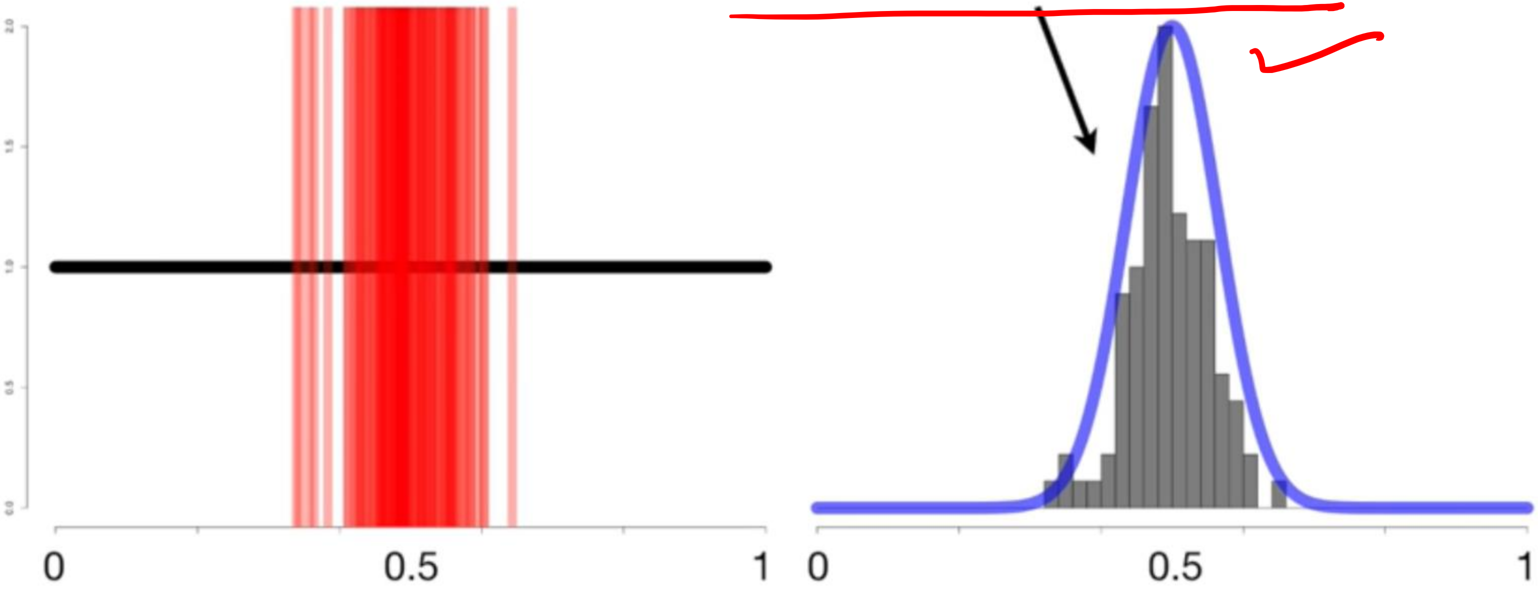
upGrad



Central Limit Theorem

upGrad

...the means themselves are not uniformly distributed. Instead, the **means are normally distributed**.



Central Limit Theorem

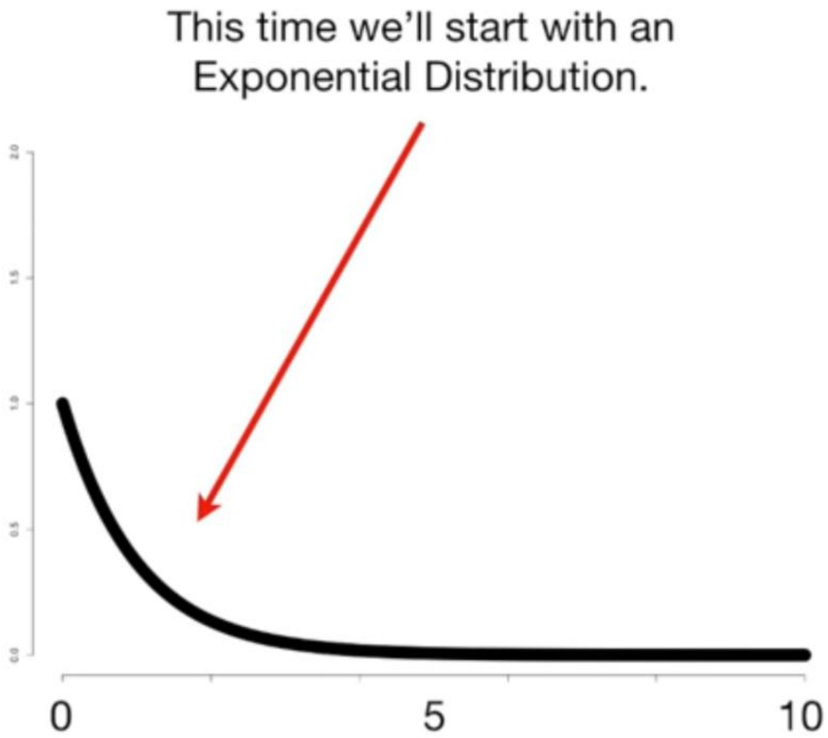
upGrad

exponential

Here’s another example…

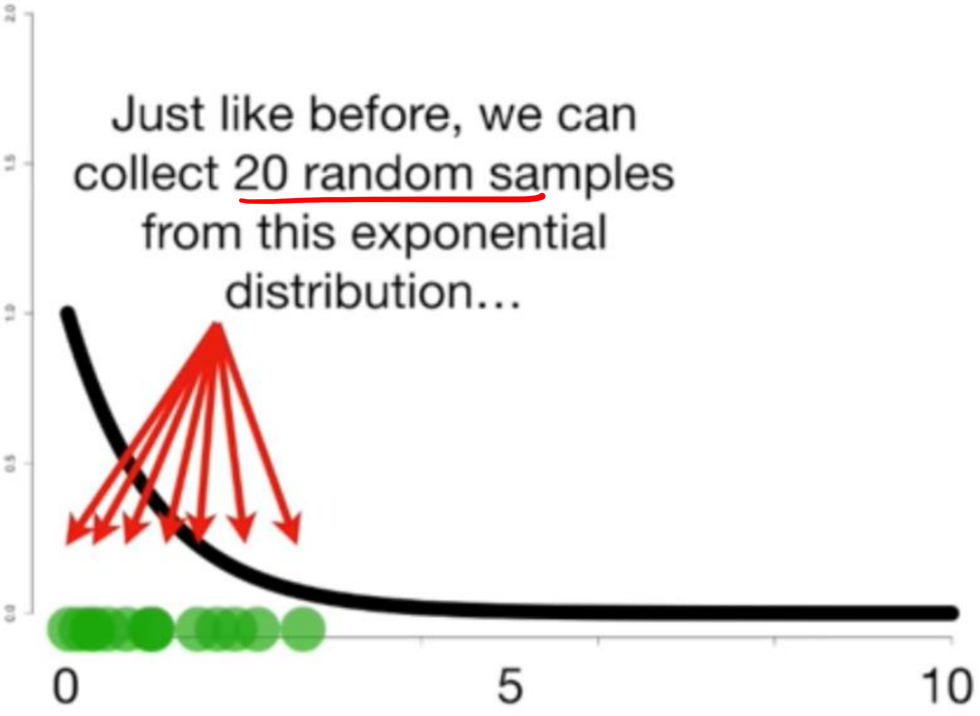
Central Limit Theorem

upGrad



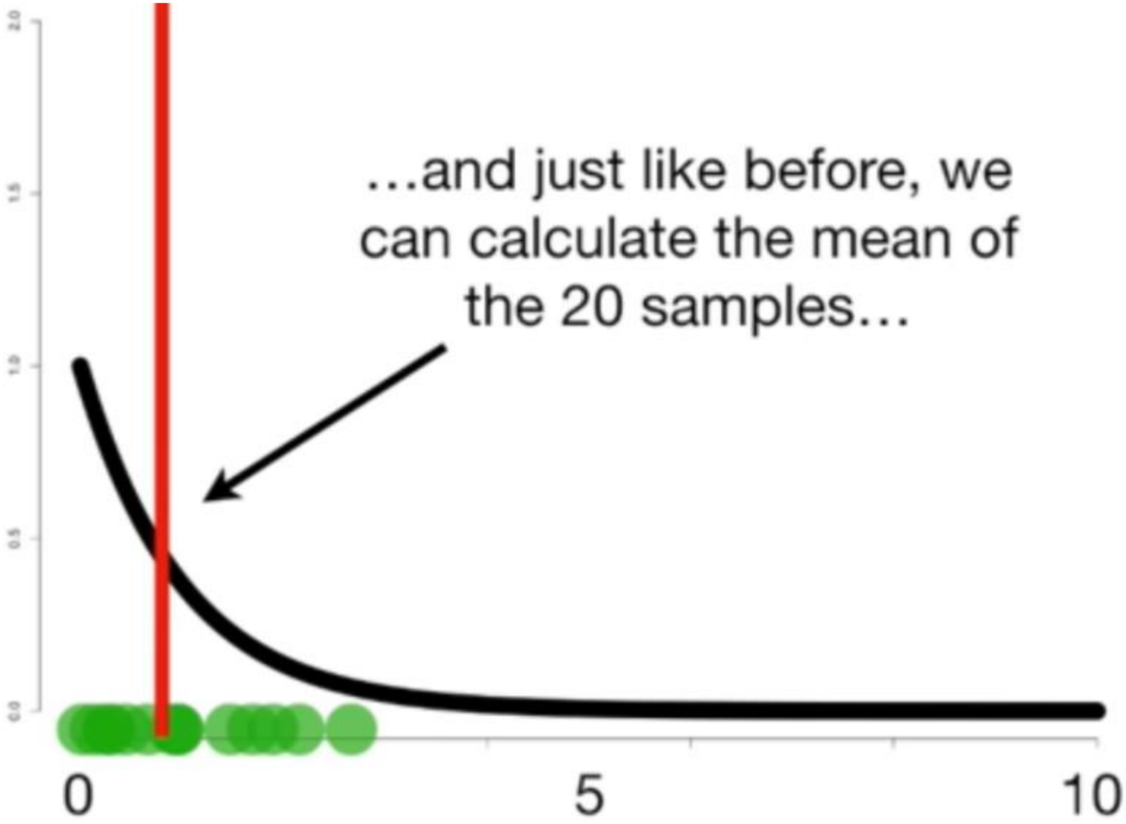
Central Limit Theorem

upGrad



Central Limit Theorem

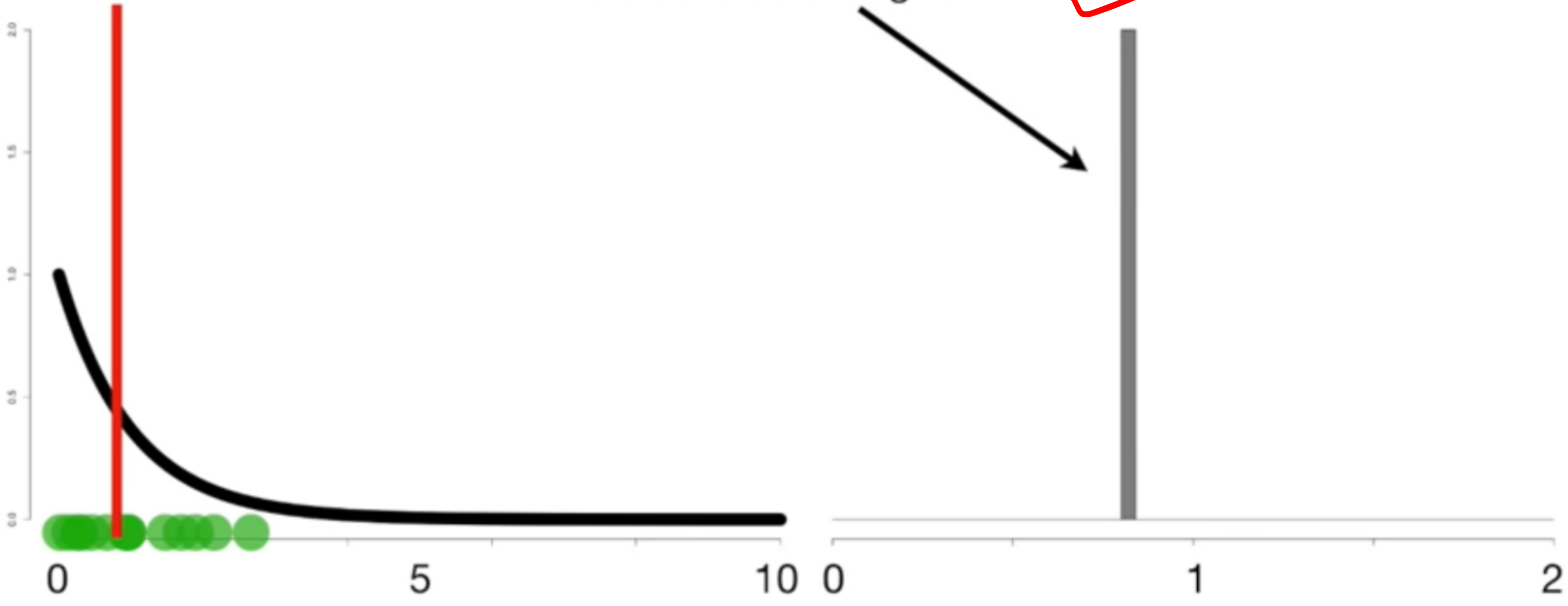
upGrad



Central Limit Theorem

upGrad

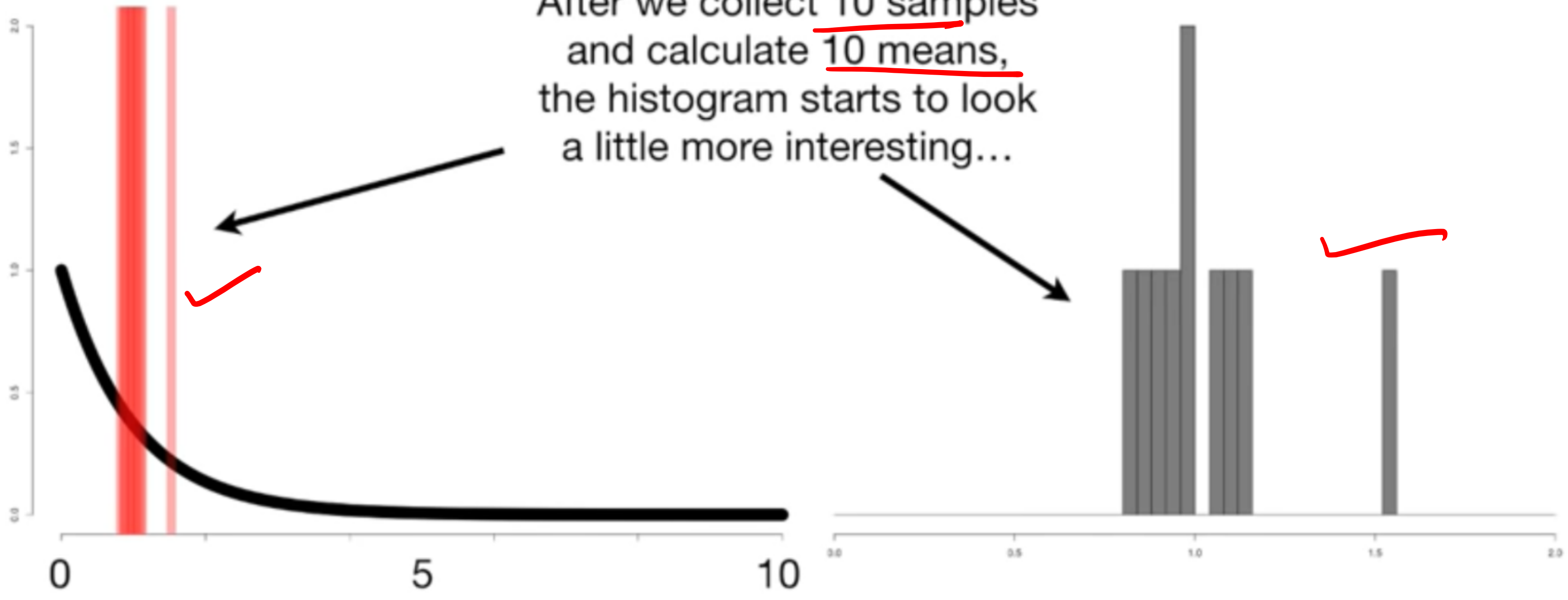
...and lastly, we can draw a histogram of that mean over here on the right.



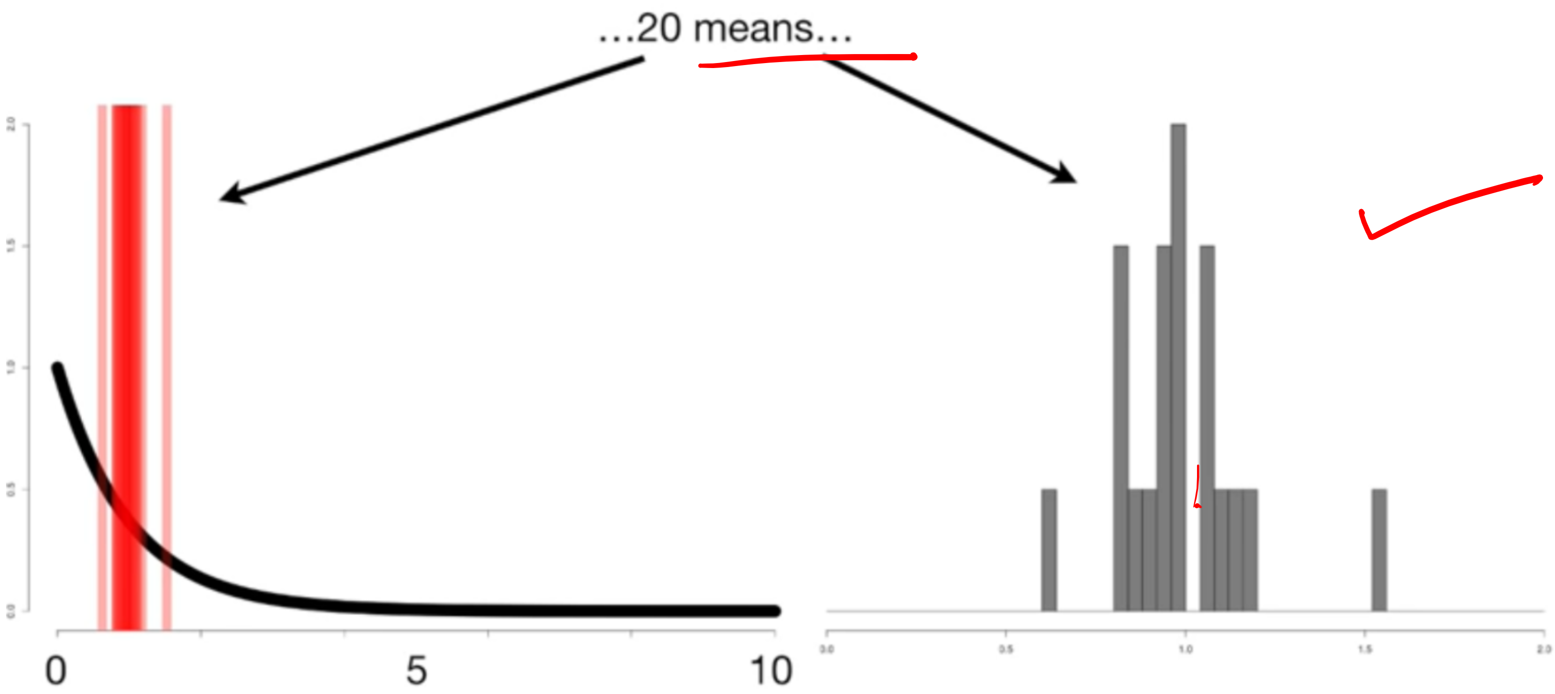
Central Limit Theorem

upGrad

After we collect 10 samples and calculate 10 means, the histogram starts to look a little more interesting…

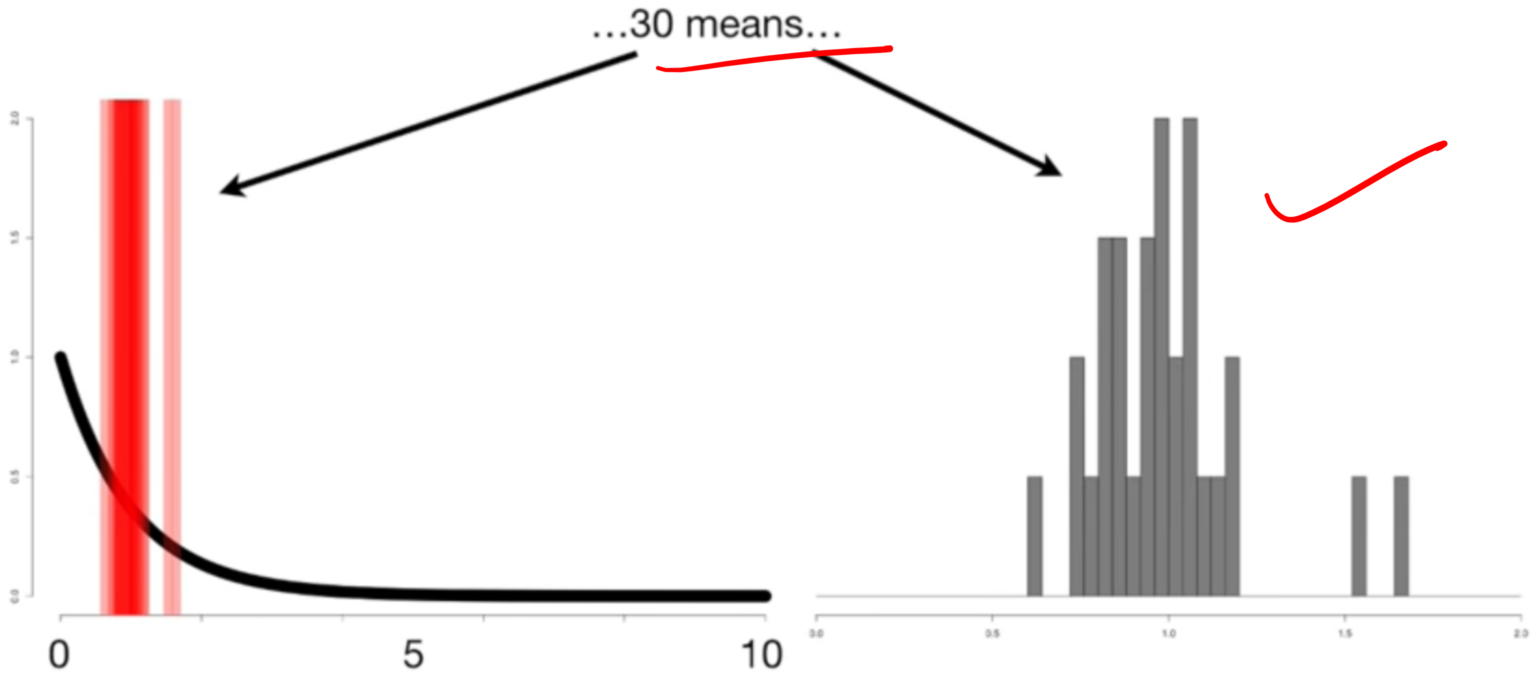


Central Limit Theorem



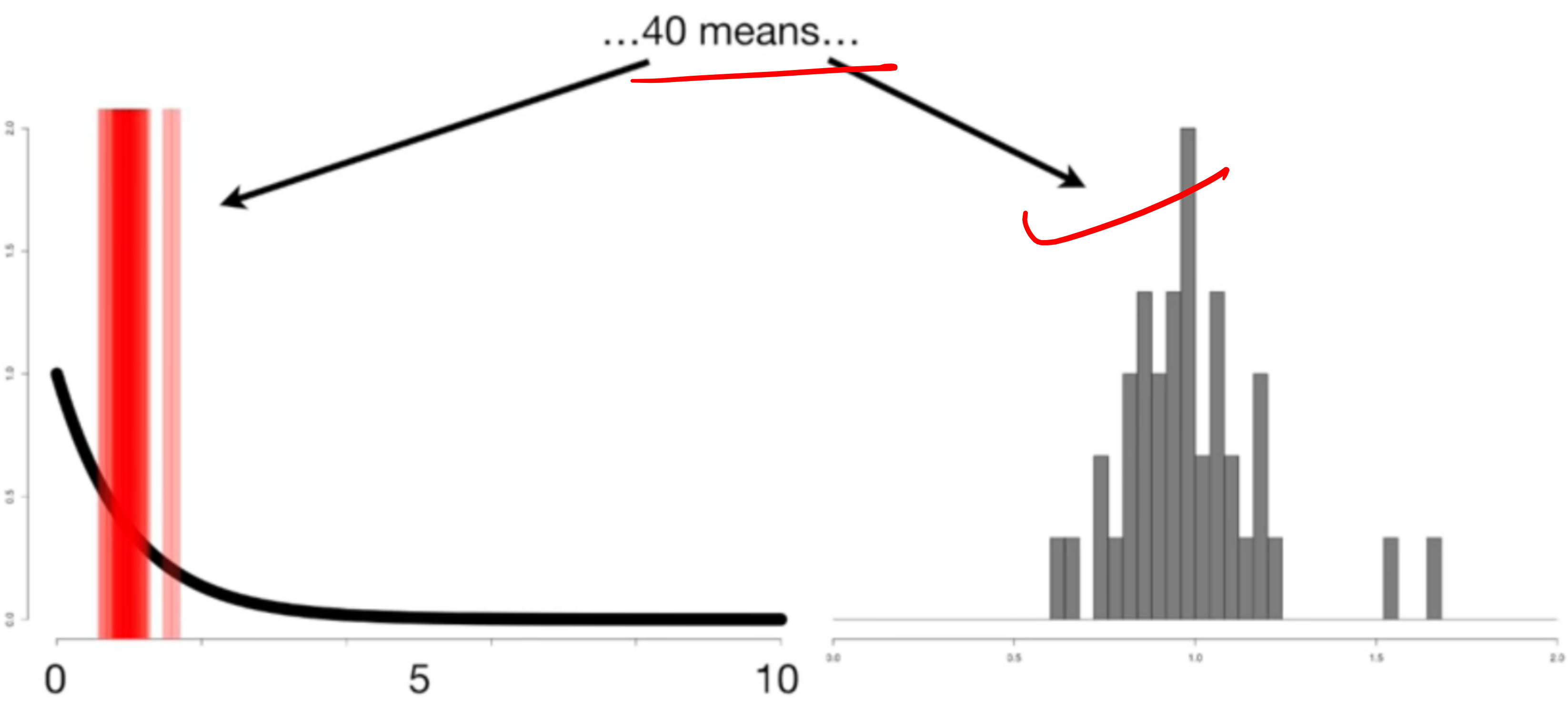
Central Limit Theorem

upGrad



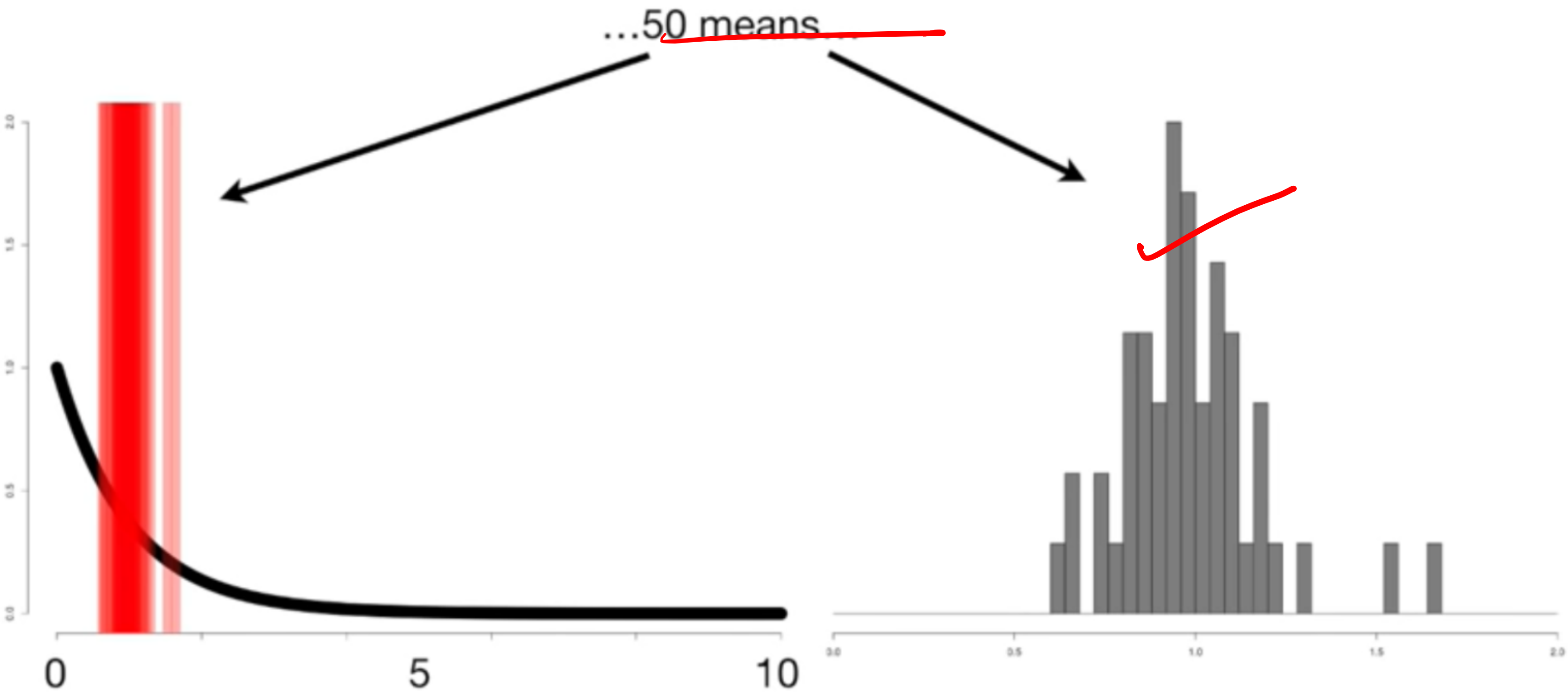
Central Limit Theorem

upGrad



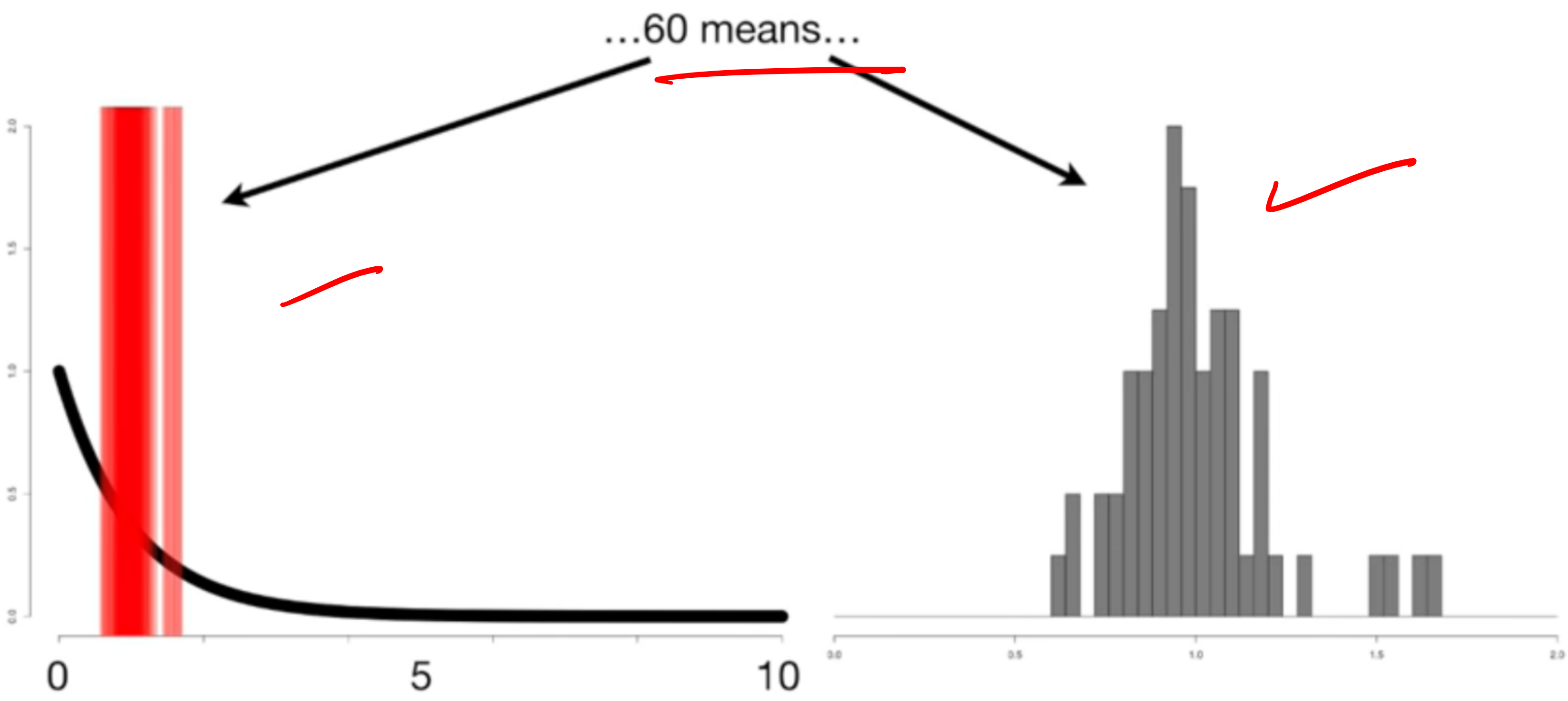
Central Limit Theorem

upGrad



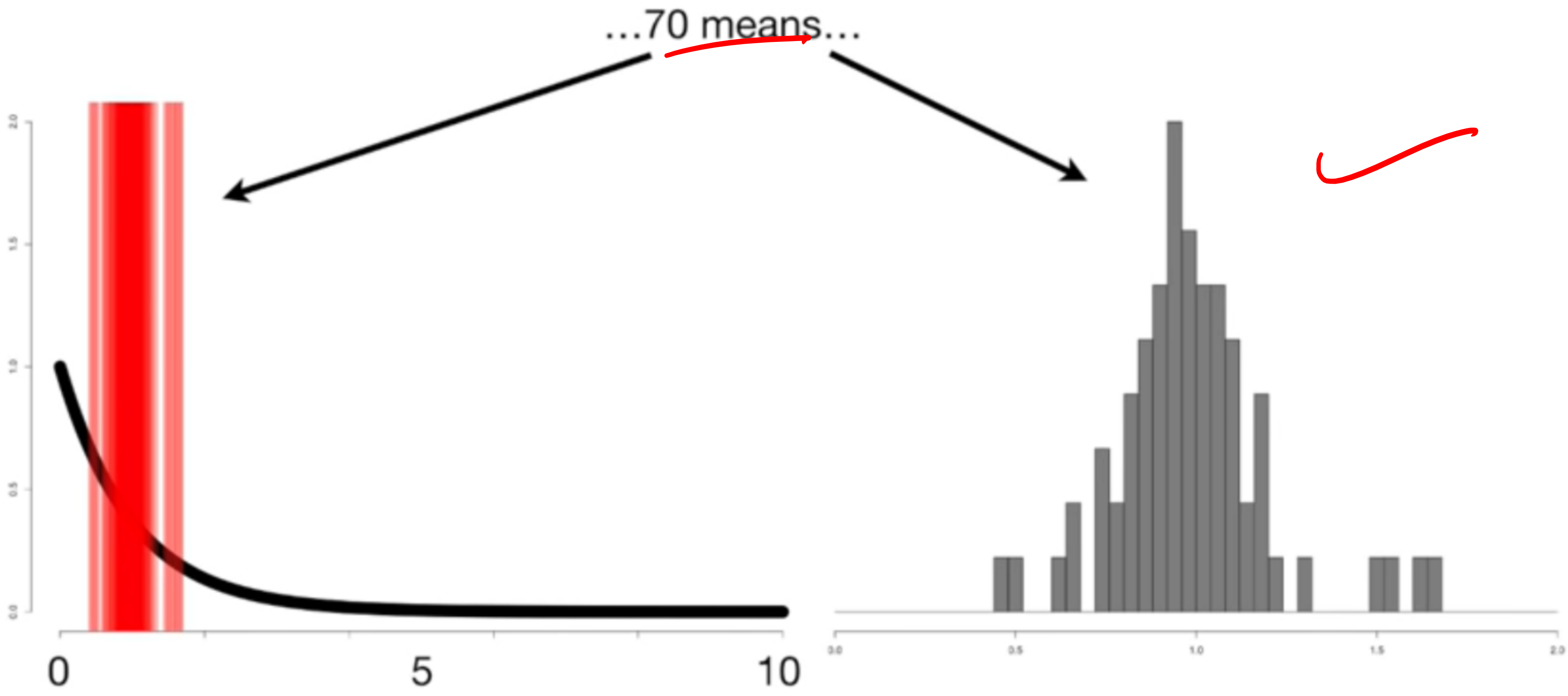
Central Limit Theorem

upGrad



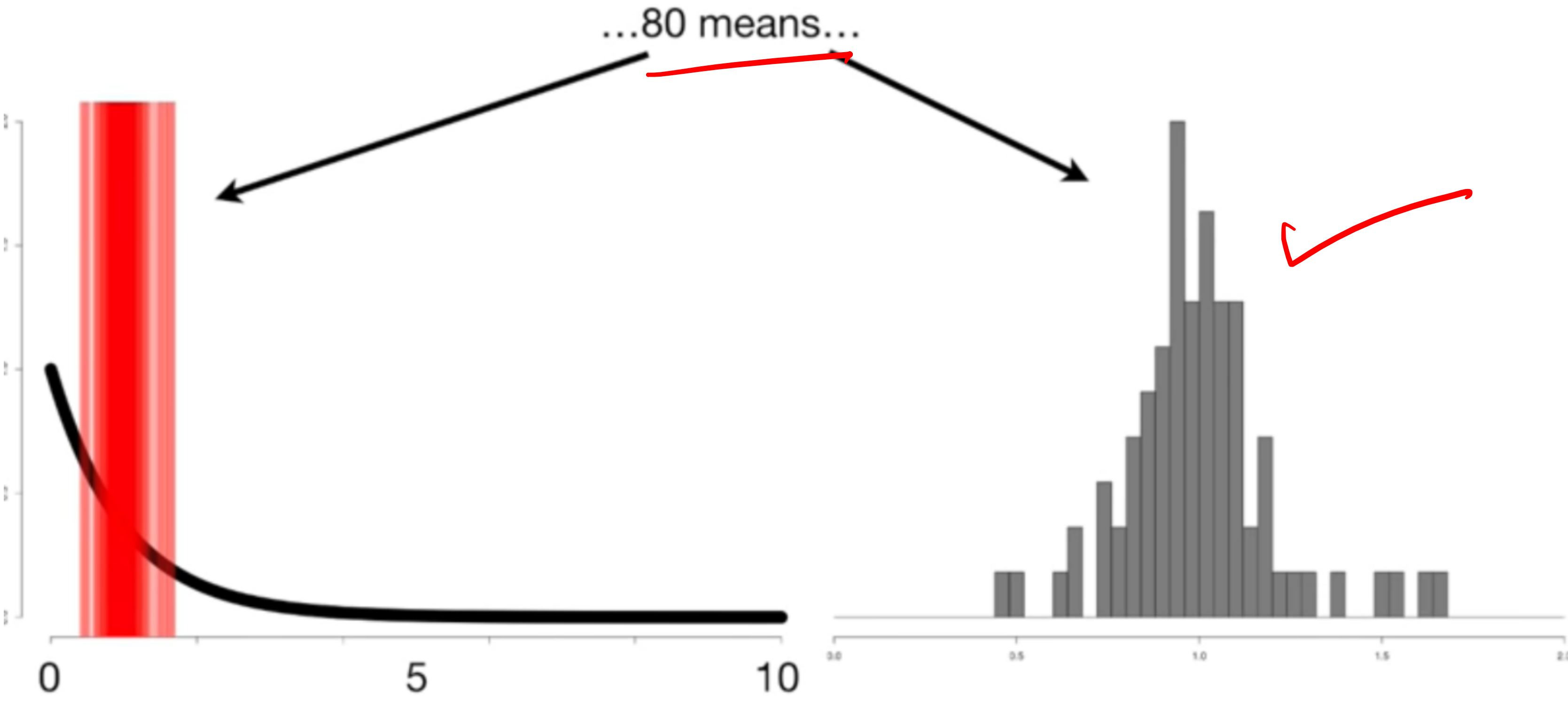
Central Limit Theorem

upGrad



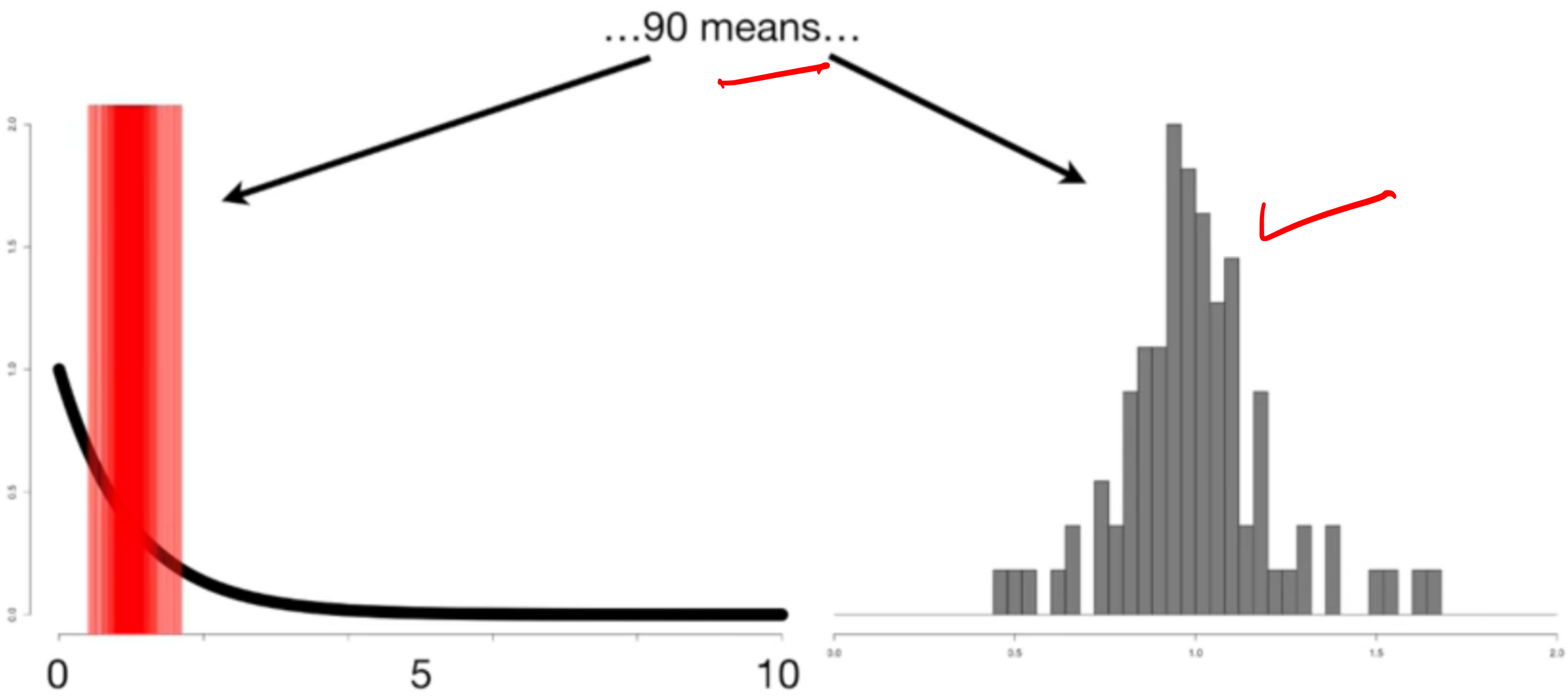
Central Limit Theorem

upGrad



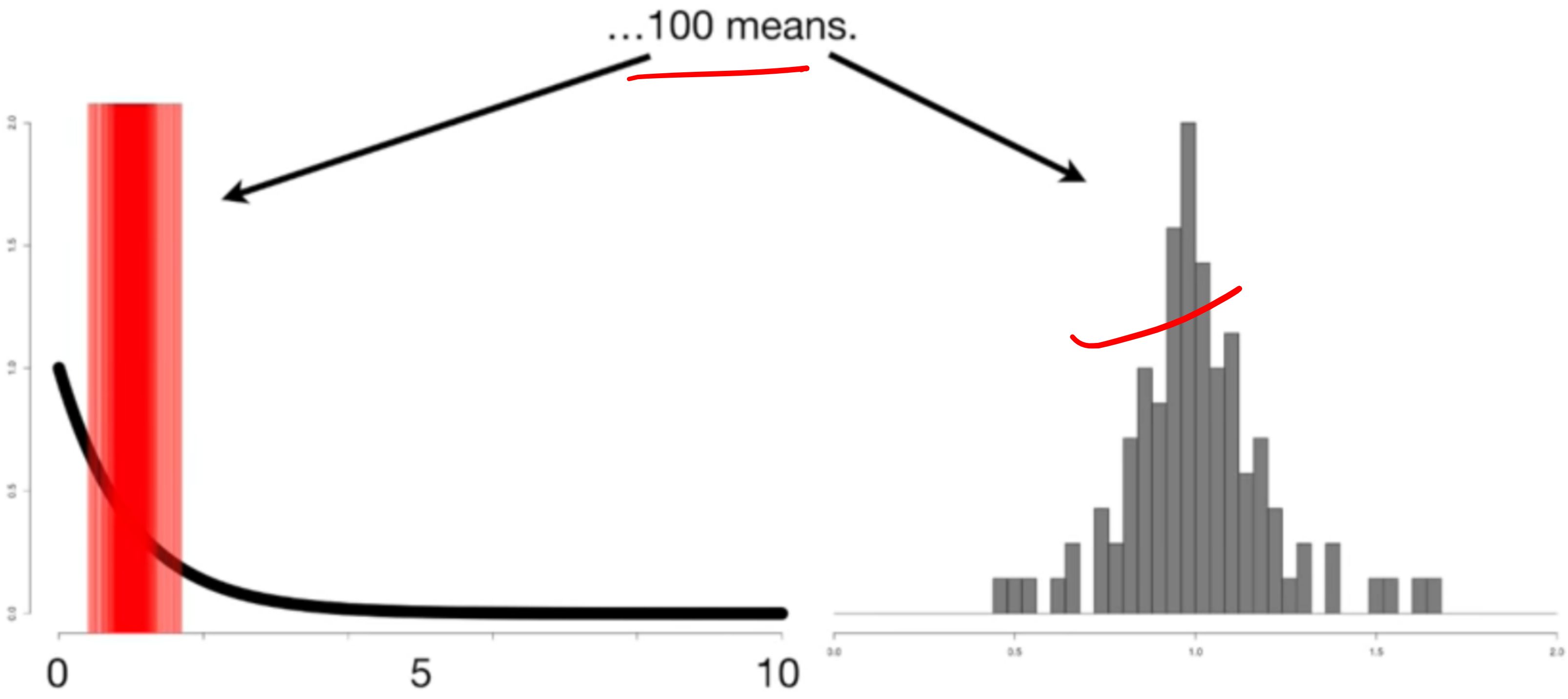
Central Limit Theorem

upGrad



Central Limit Theorem

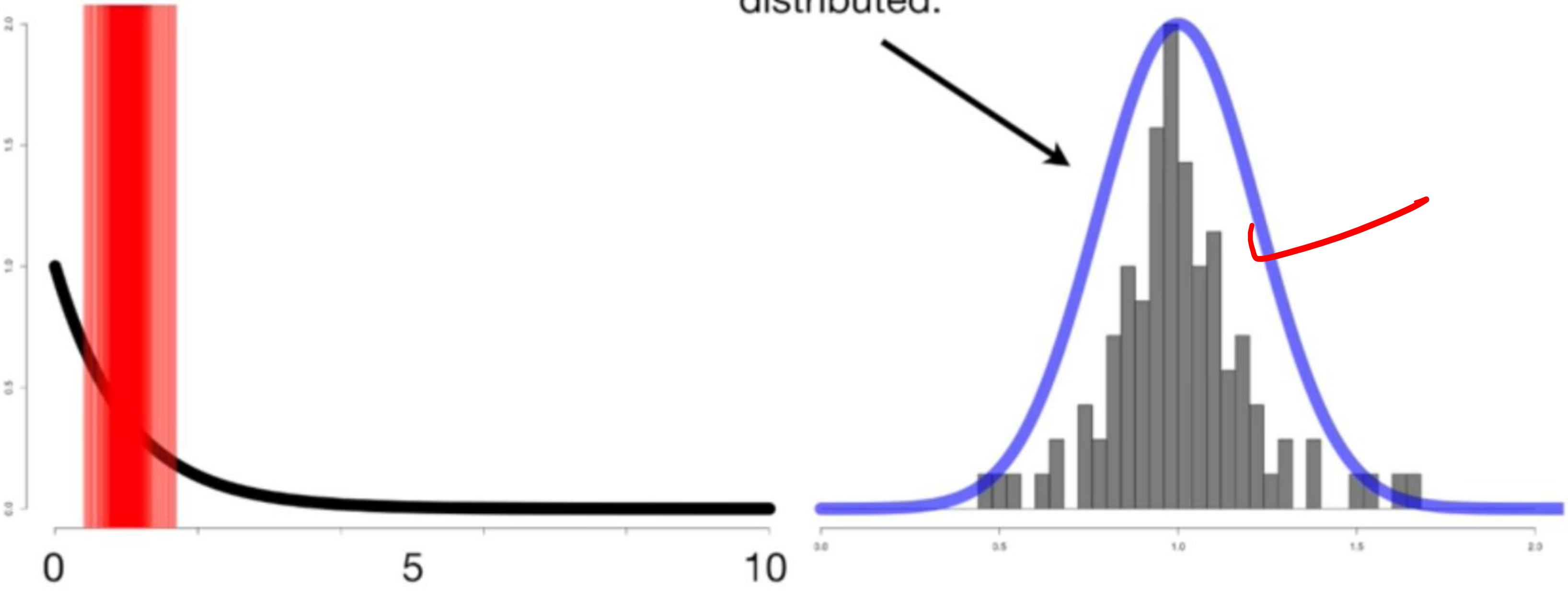
upGrad



Central Limit Theorem

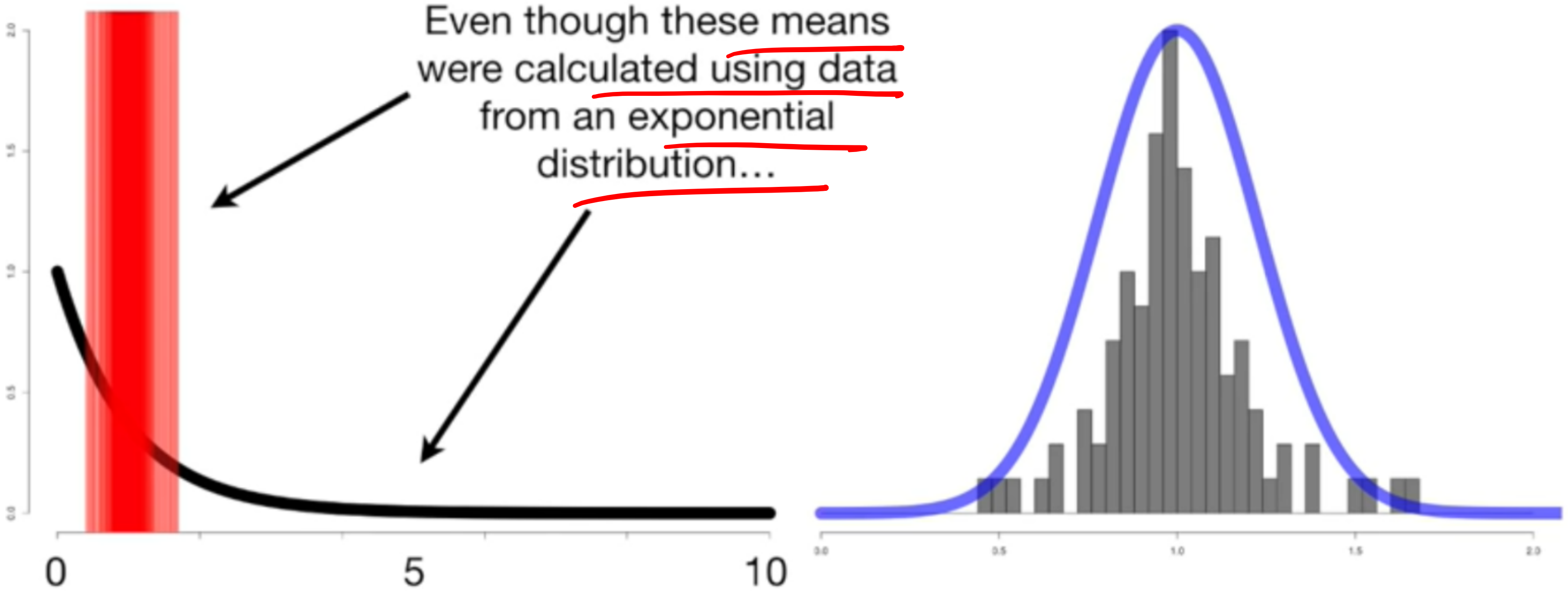
upGrad

After adding 100 means to the histogram, we can see that they are normally distributed.



Central Limit Theorem

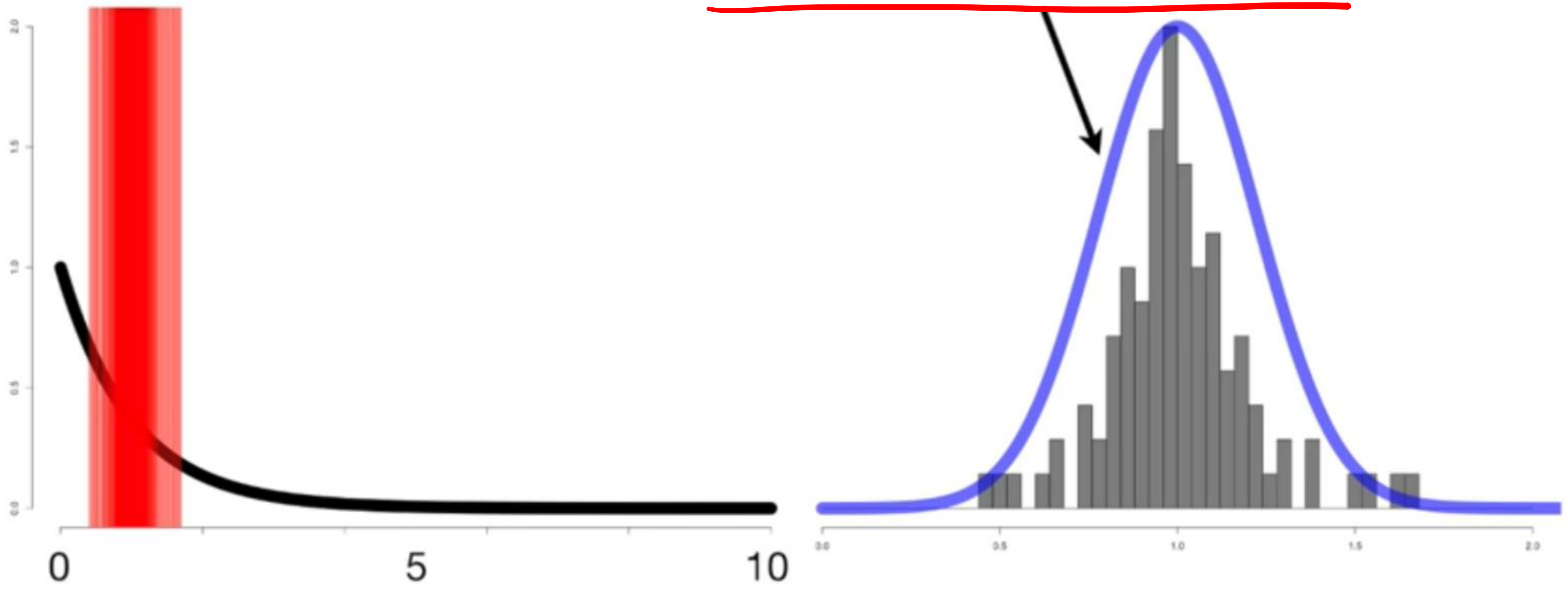
upGrad



Central Limit Theorem

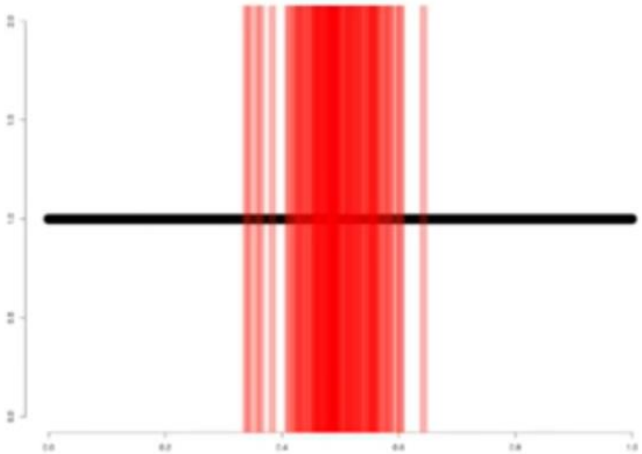
upGrad

…the means themselves are not exponentially distributed. Instead, the **means are normally distributed.**



Central Limit Theorem

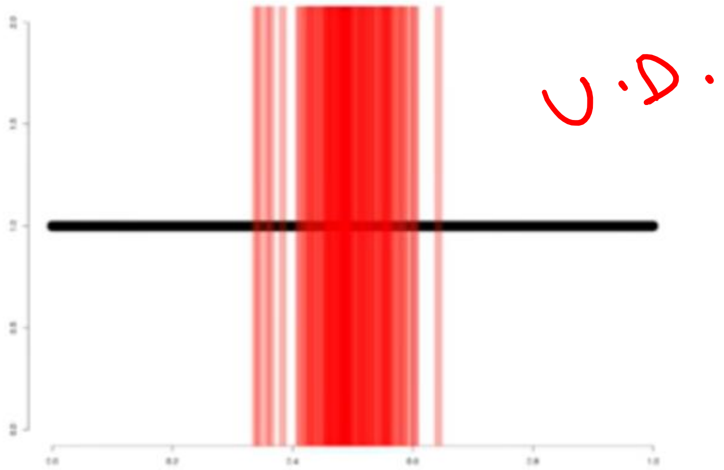
upGrad

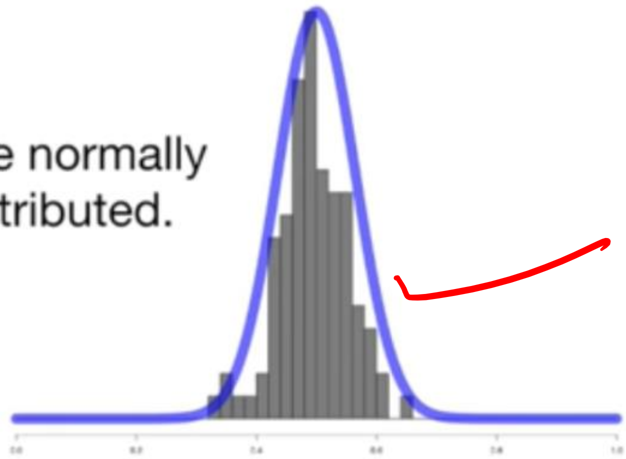


So far we have seen that means calculated from samples taken from a uniform distribution…

Central Limit Theorem

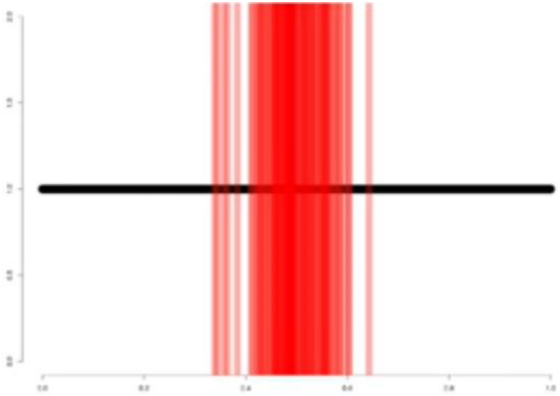
upGrad

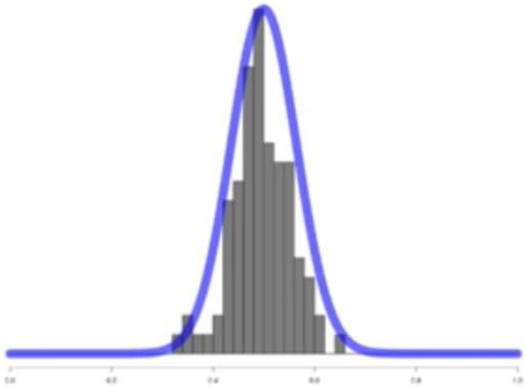




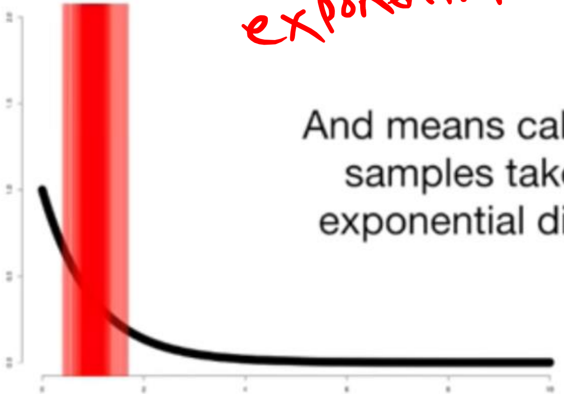
Central Limit Theorem

upGrad



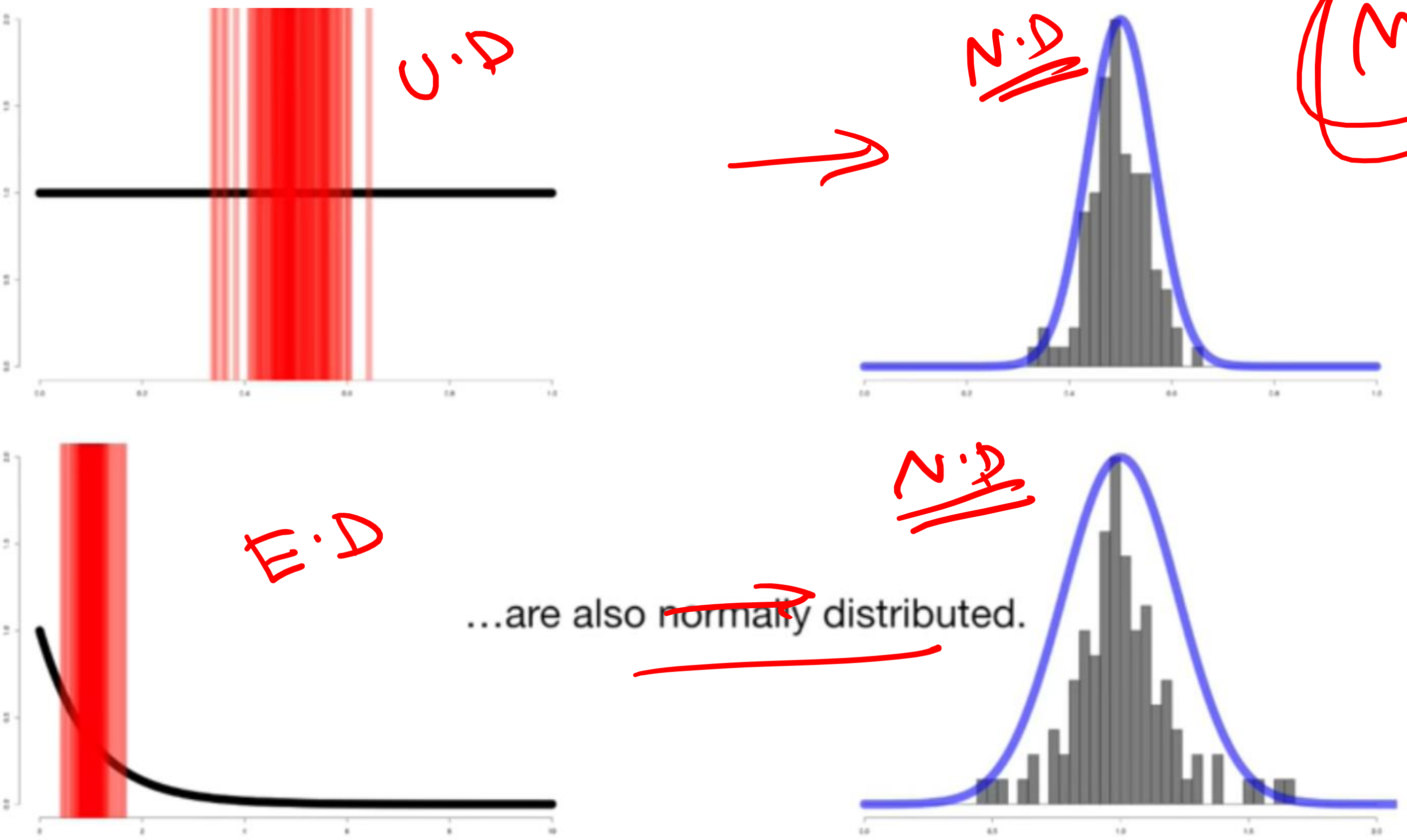


exponential



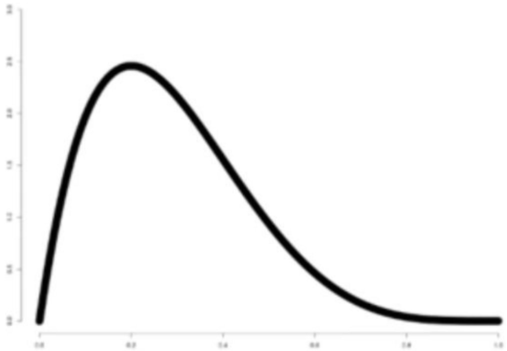
And means calculated from samples taken from an exponential distribution…

Central Limit Theorem

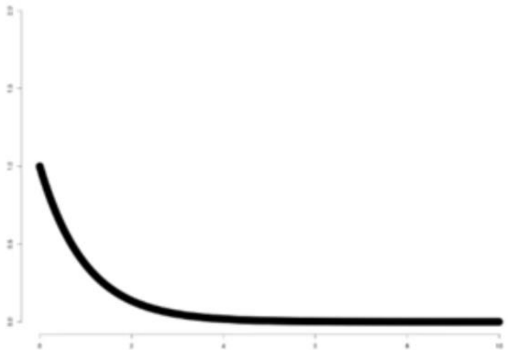


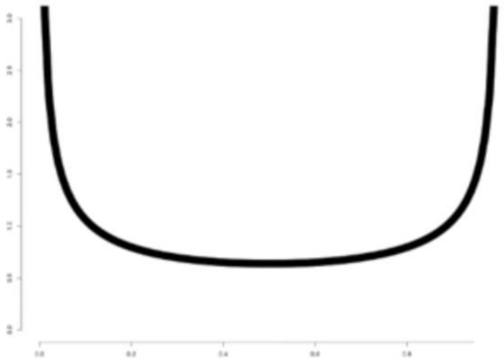
Central Limit Theorem

upGrad



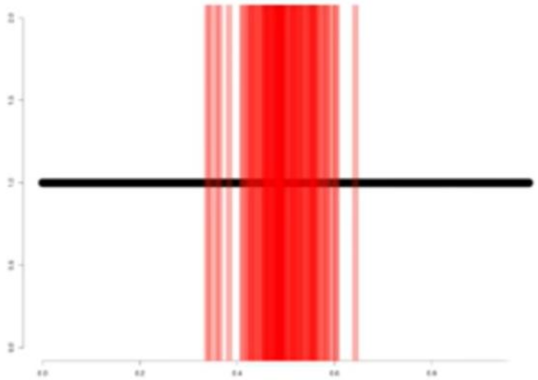
Well, it turns out that it doesn't matter what distribution you start with…

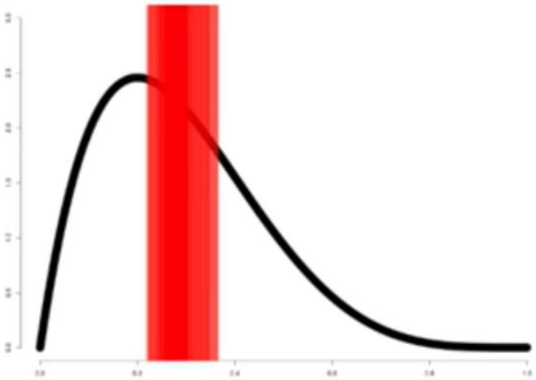




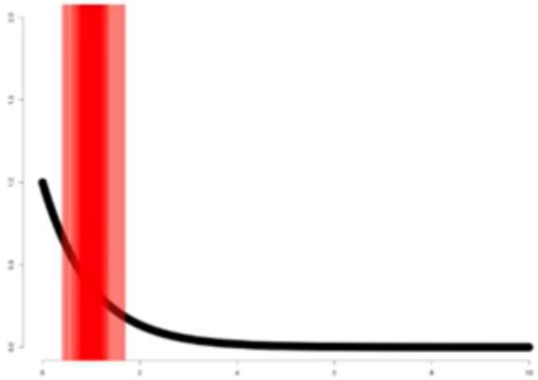
Central Limit Theorem

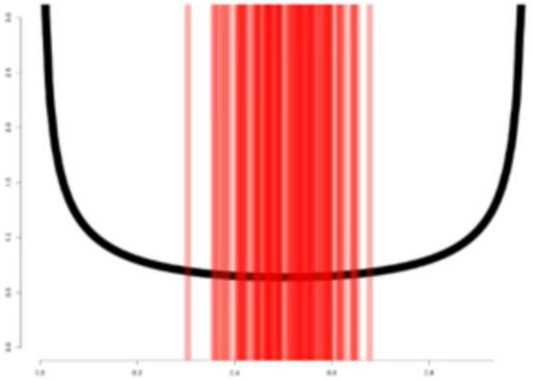
upGrad





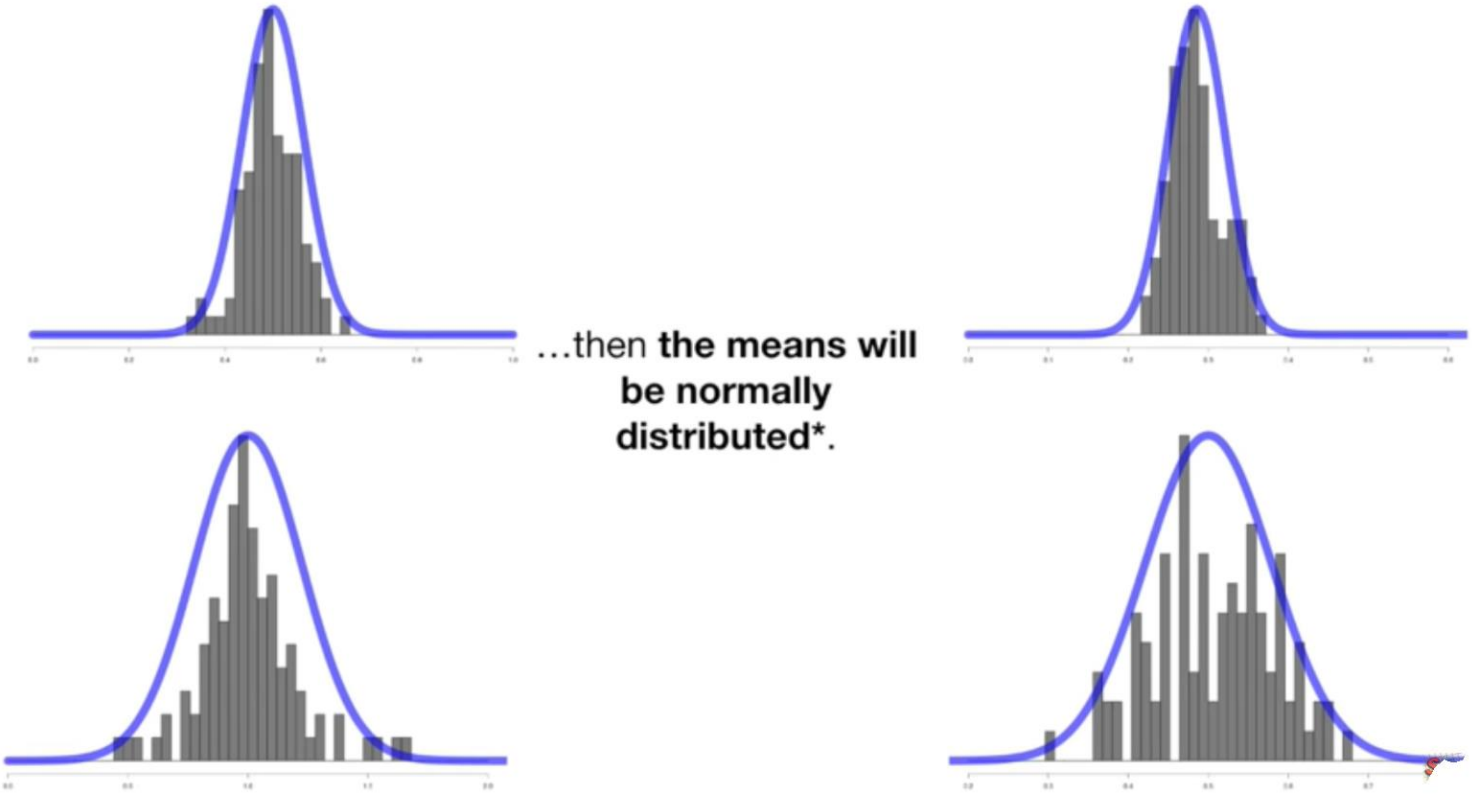
...if you collect samples from those distributions...





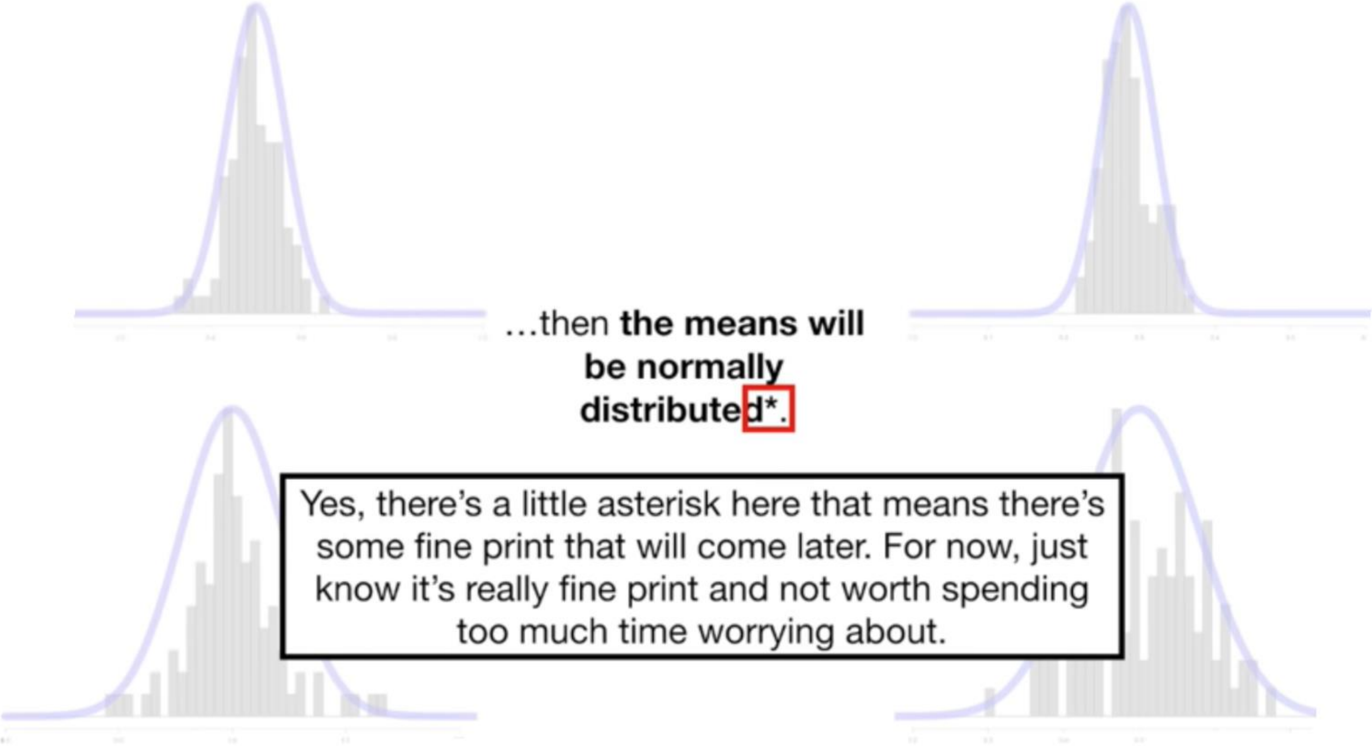
Central Limit Theorem

upGrad



...then **the means will be normally distributed***.

Central Limit Theorem



Central Limit Theorem

upGrad

Cool!

But what are the practical implications of knowing that means are normally distributed?

Central Limit Theorem

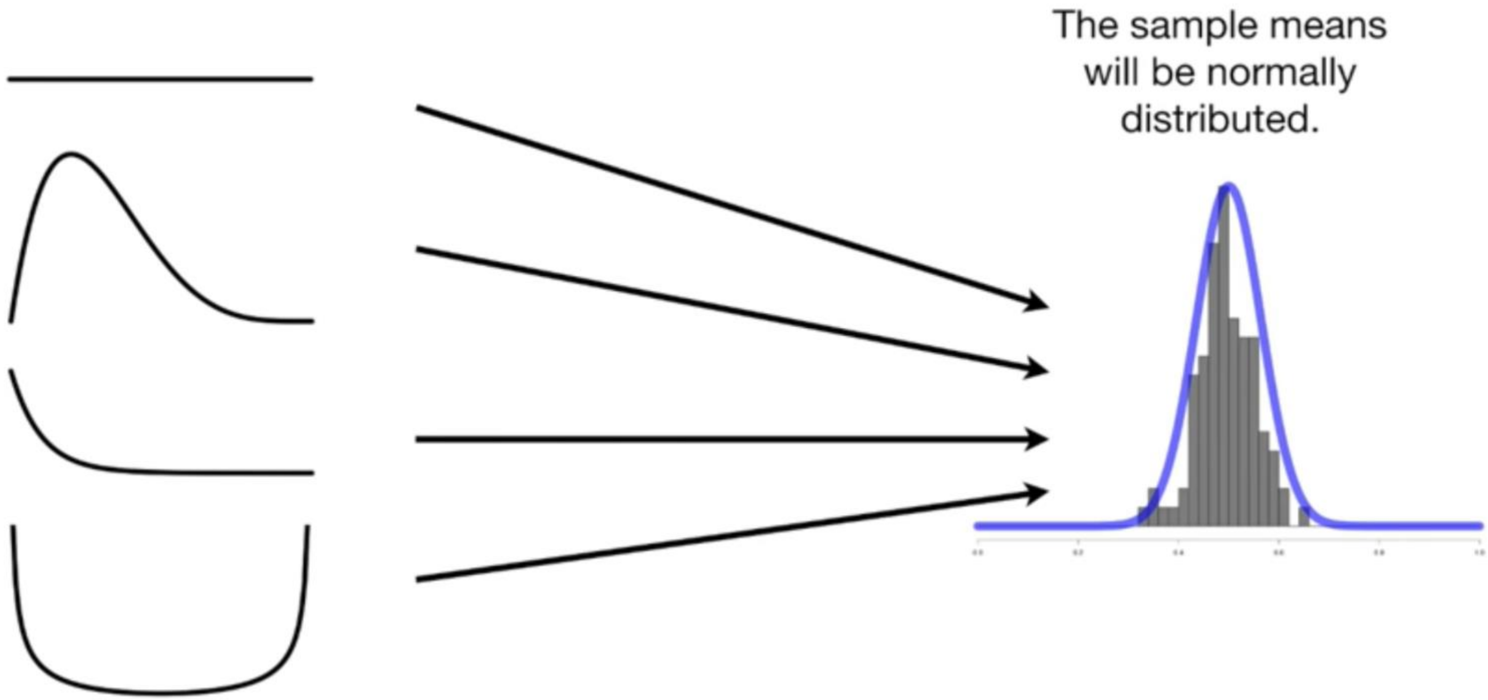
upGrad

When we do an experiment, we don't always know what distribution our data comes from.

To this, **The Central Limit Theorem** says, **"Who Cares???"**

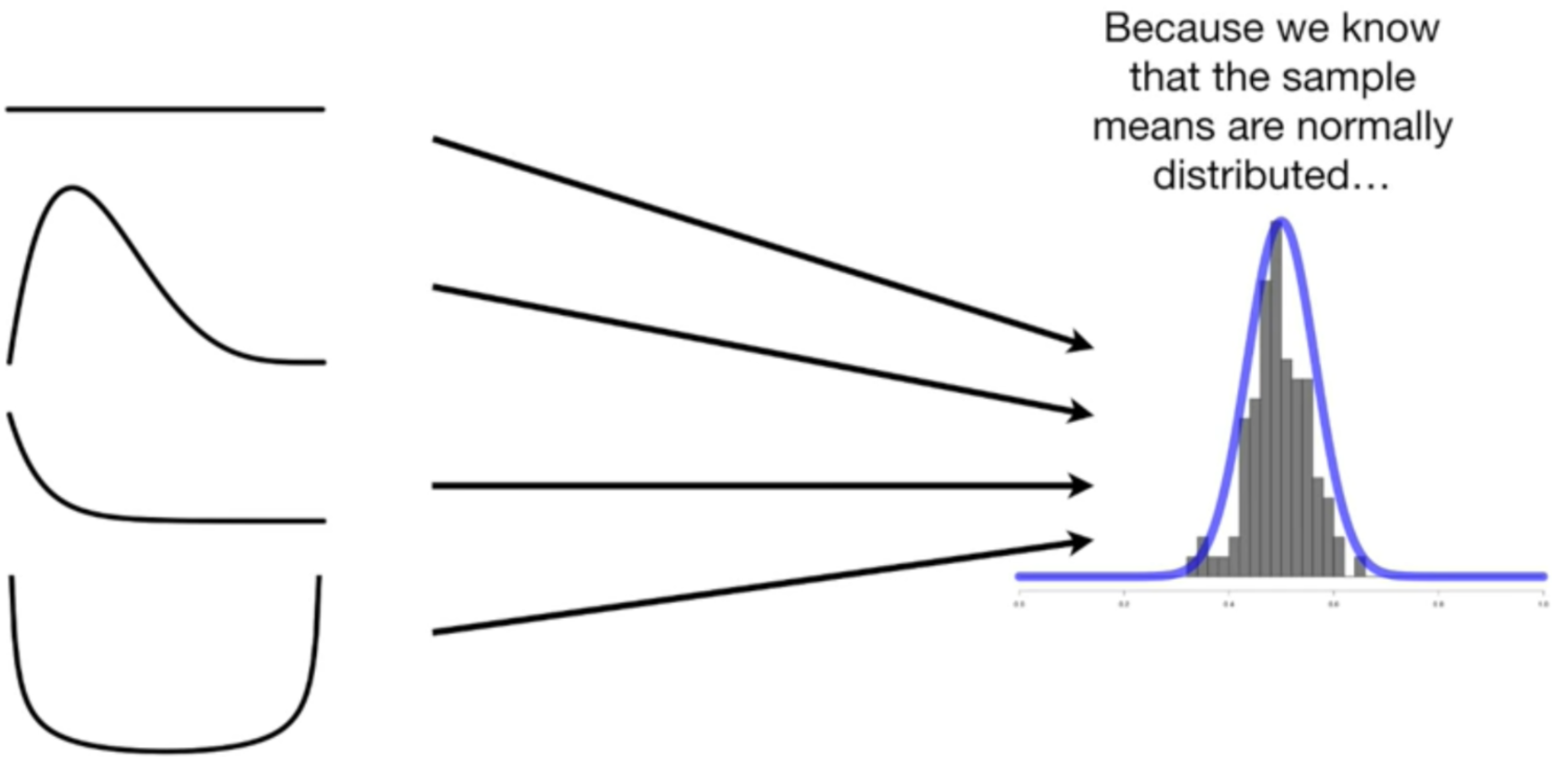
Central Limit Theorem

upGrad



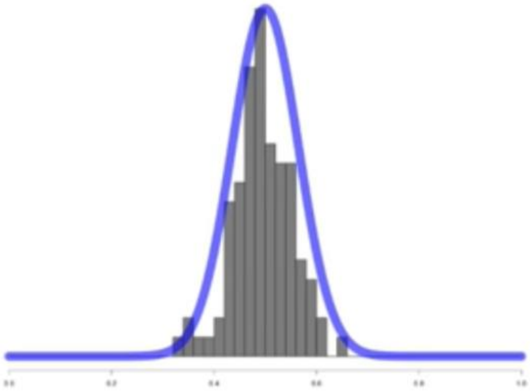
Central Limit Theorem

upGrad



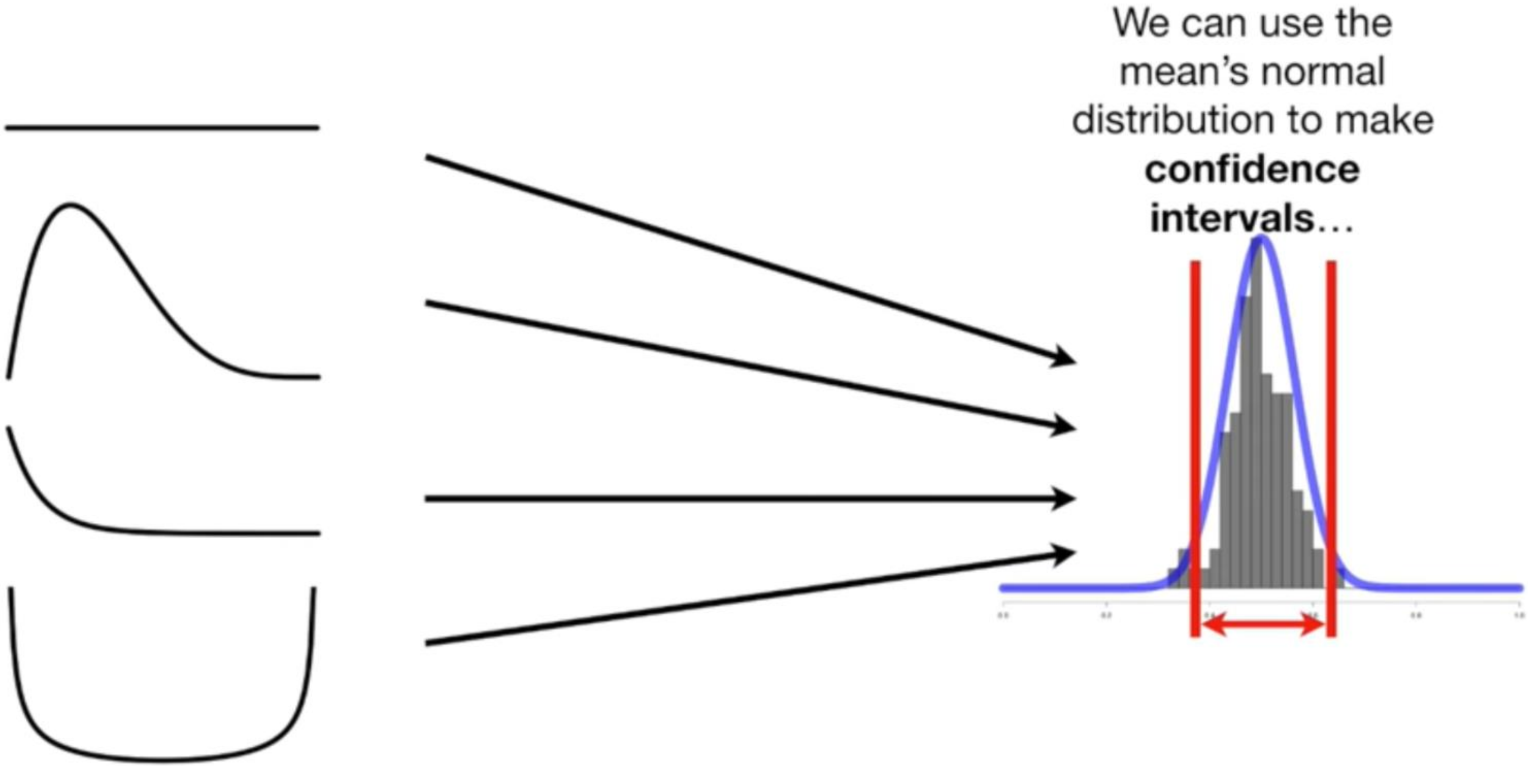
Central Limit Theorem

…we don't need to worry too much about the distribution that the samples came from.

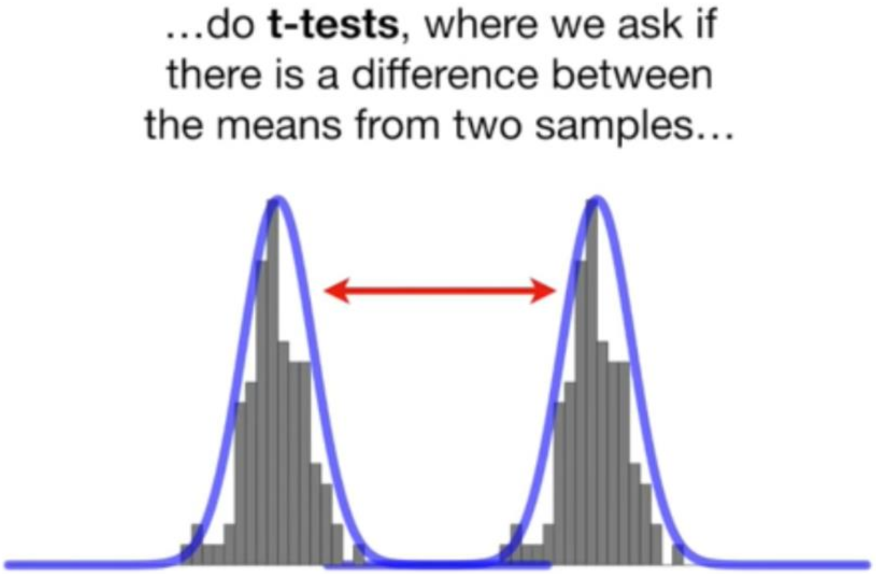


Central Limit Theorem

upGrad



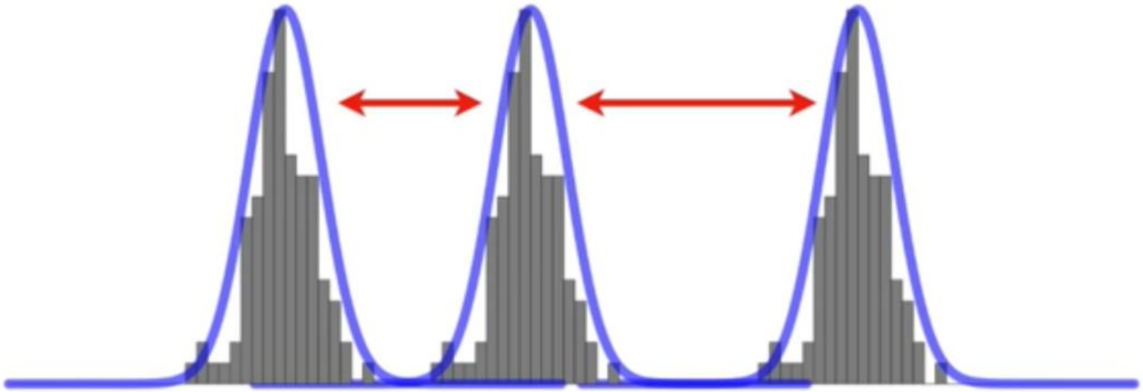
Central Limit Theorem



Central Limit Theorem

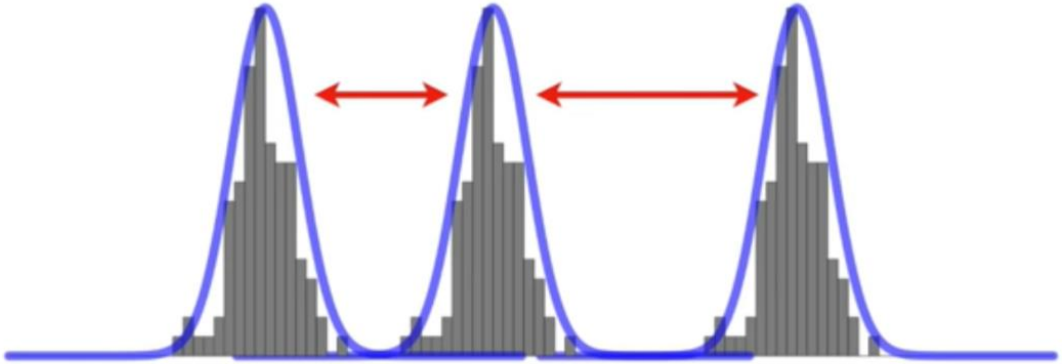
upGrad

…and **ANOVA**, where we ask if there is a difference among the means from three or more samples…



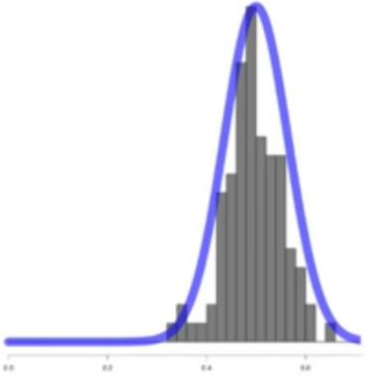
Central Limit Theorem

...and pretty much any statistical test that uses the sample mean.

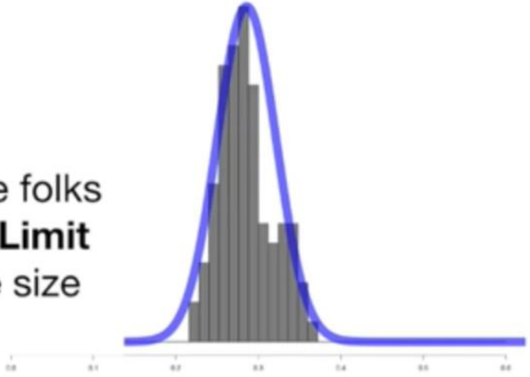


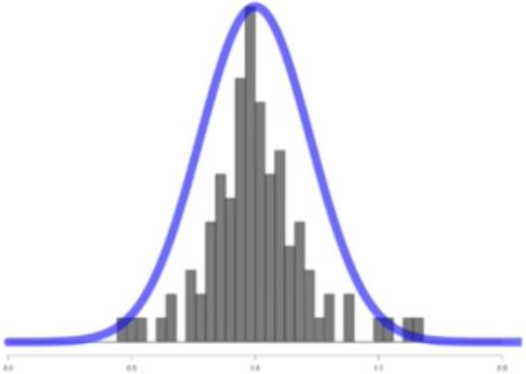
Central Limit Theorem

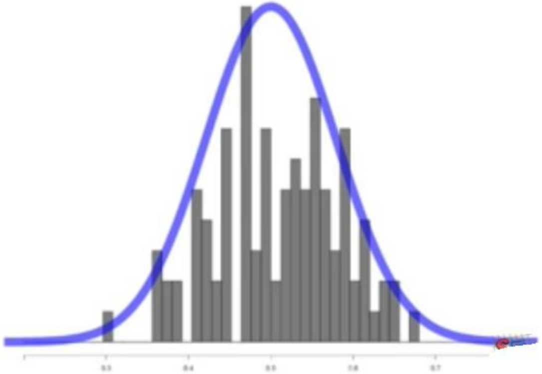
upGrad



NOTE: Out there in the wild some folks say that in order for the **Central Limit Theorem** to be true, the sample size must be at least **30**.

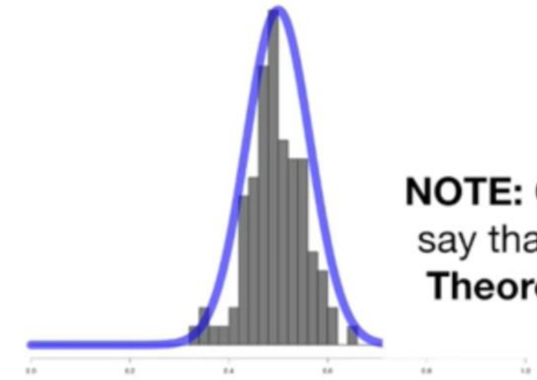


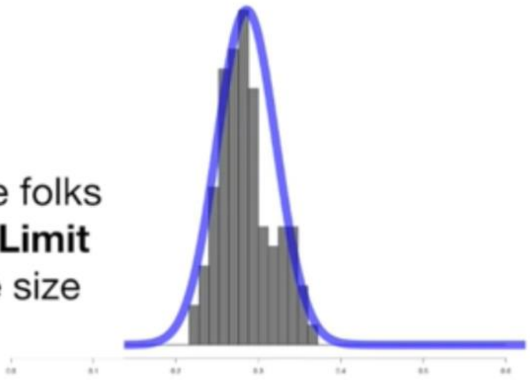




Central Limit Theorem

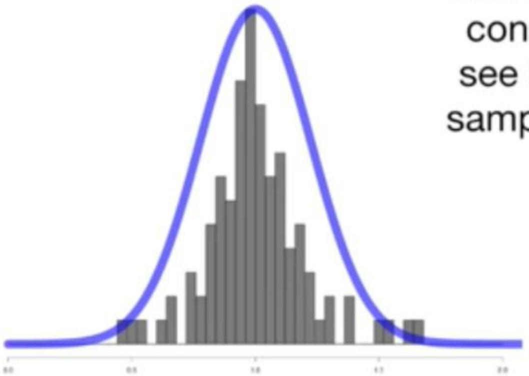
upGrad

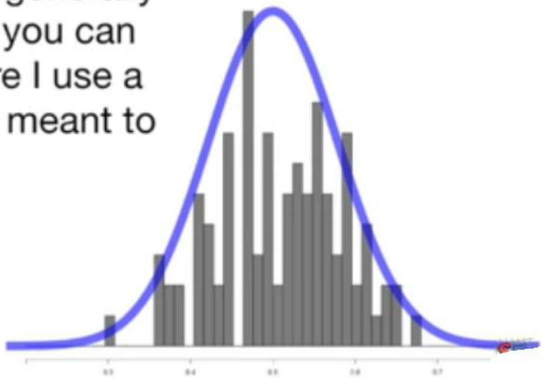




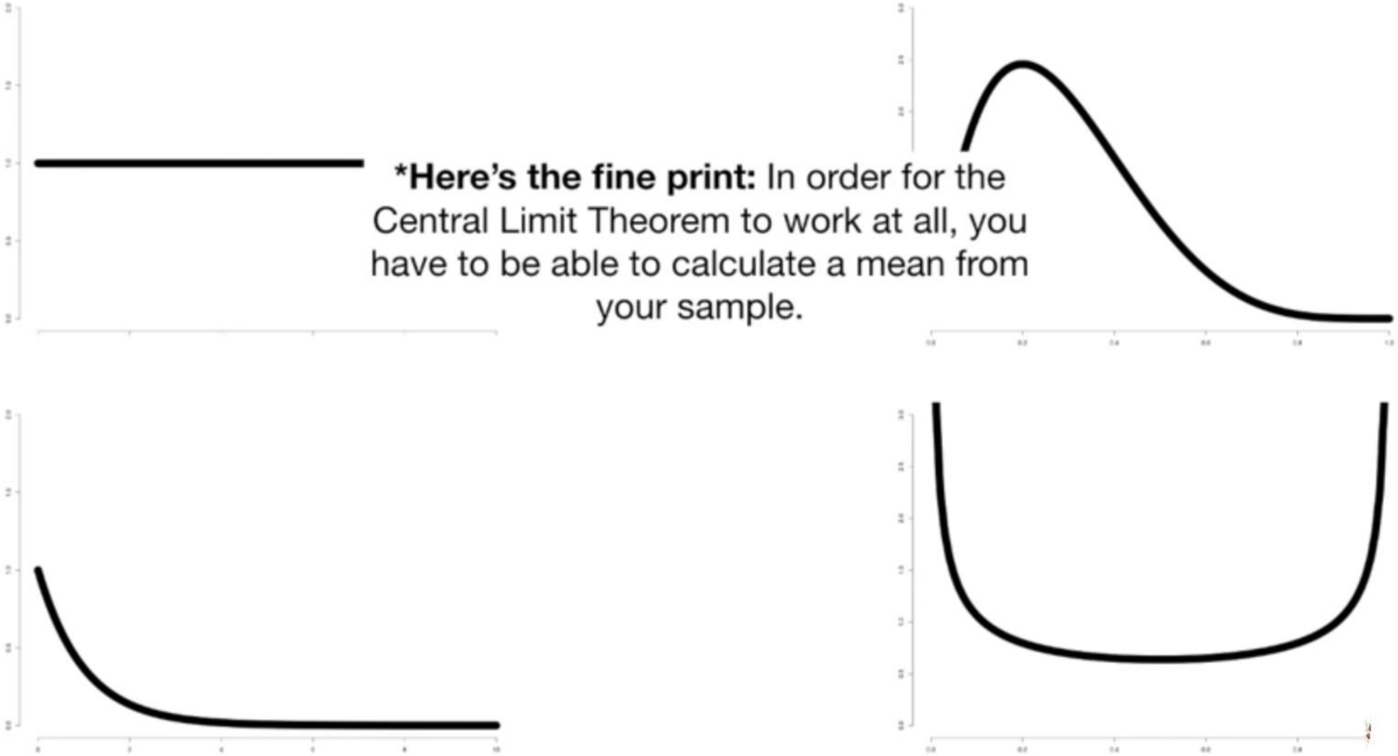
NOTE: Out there in the wild some folks say that in order for the **Central Limit Theorem** to be true, the sample size must be at least **30**.

This is just a rule of thumb and generally considered safe. However, as you can see in the examples here where I use a sample size of **20**, the rule was meant to be broken.





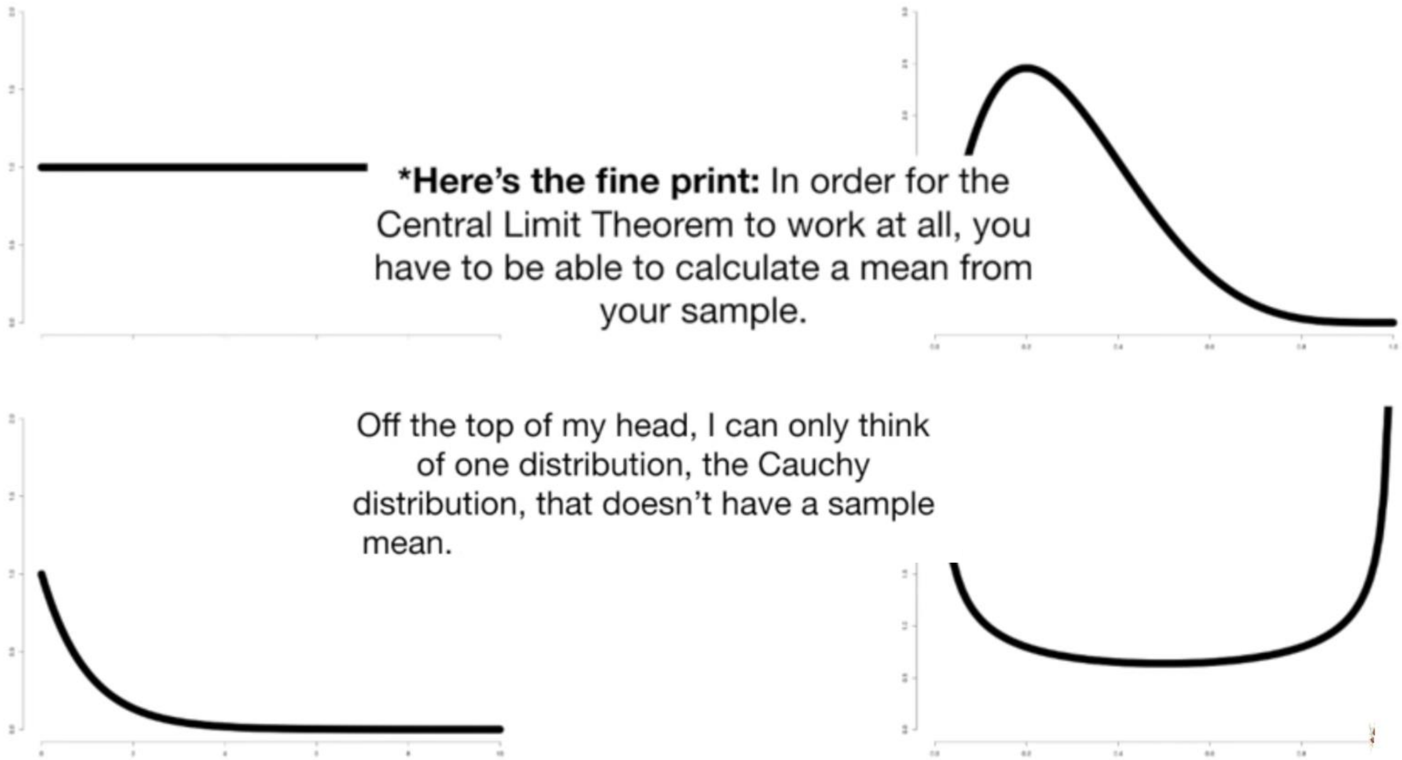
Central Limit Theorem



***Here's the fine print:** In order for the Central Limit Theorem to work at all, you have to be able to calculate a mean from your sample.

Central Limit Theorem

upGrad

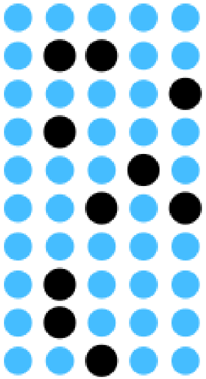


***Here's the fine print:** In order for the Central Limit Theorem to work at all, you have to be able to calculate a mean from your sample.

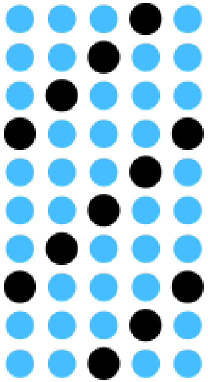
Off the top of my head, I can only think of one distribution, the Cauchy distribution, that doesn't have a sample mean.

Sampling Methods

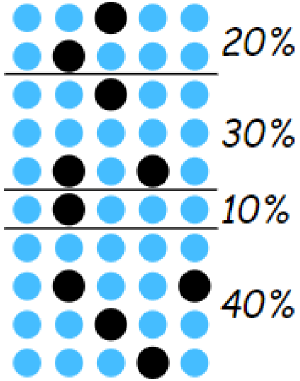
upGrad



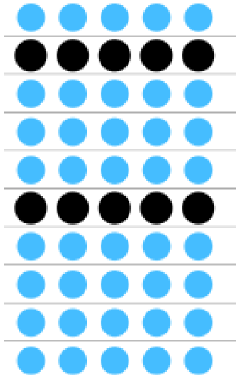
Random Sample
(pick randomly from list)



Systematic Sample
(such as every 4th)



Stratified Sample
(randomly, but in ratio to group size)



Cluster Sample
(choose whole groups randomly)

- **Volunteer sampling**
- **Opportunity sampling**

False Positives and False Negatives

upGrad

When you have a test that can say "Yes" or "No" (such as a medical test), you have to think:

- It could be **wrong** when it says "Yes".
- It could be **wrong** when it says "No".

It is like being told you **did** something when you **didn't**!

Or you didn't do it when you really did.

False Positives and False Negatives

upGrad

- They each have a special name: **"False Positive"** and **"False Negative"**:

	They say you **did**	They say you **didn't**
You really did	*They are right!*	**"False Negative"**
You really didn't	**"False Positive"**	*They are right!*

Example

upGrad

- Hunter says she is itchy. There is a test for Allergy to Cats, but this test is not always right:
  - For people that **really do** have the allergy, the test says "Yes" **80%** of the time
  - For people that **do not** have the allergy, the test says "Yes" **10%** of the time ("false positive")

	Test says "Yes"	Test says "No"
Have allergy	**80%**	**20% "False Negative"**
Don't have it	**10% "False Positive"**	**90%**

Question:

upGrad

- <u>If 1% of the population have the allergy, and Hunter's test says "Yes", what are the chances that Hunter really has the allergy?</u>
- There are three different ways to solve this:
  - "Imagine a 1000",
  - "Tree Diagrams" or
  - "Bayes' Theorem"

Solution 1: Imagining A Thousand People

upGrad

When trying to understand questions like this, just imagine a large group (say 1000) and play with the numbers:

- Of 1000 people, only **10** really have the allergy (1% of 1000 is 10)
- The test is 80% right for people who **have** the allergy, so it will get **8 of those 10 right**.
- But 990 **do not** have the allergy, and the test will say "Yes" to 10% of them, which is **99 people** it says "Yes" to **wrongly** (false positive)
- So out of 1000 people the test says "**Yes**" to (8+99) **= 107 people**

	1% have it	Test says "Yes"	Test says "No"
Have allergy	10	**8**	**2**
Don't have it	990	**99**	**891**
	1000	107	893

Solution 1: Imagining A Thousand People

	1% have it	Test says "Yes"	Test says "No"
Have allergy	10	**8**	**2**
Don't have it	990	**99**	**891**
	1000	107	893

So 107 people get a "Yes" but only 8 of those really have the allergy:

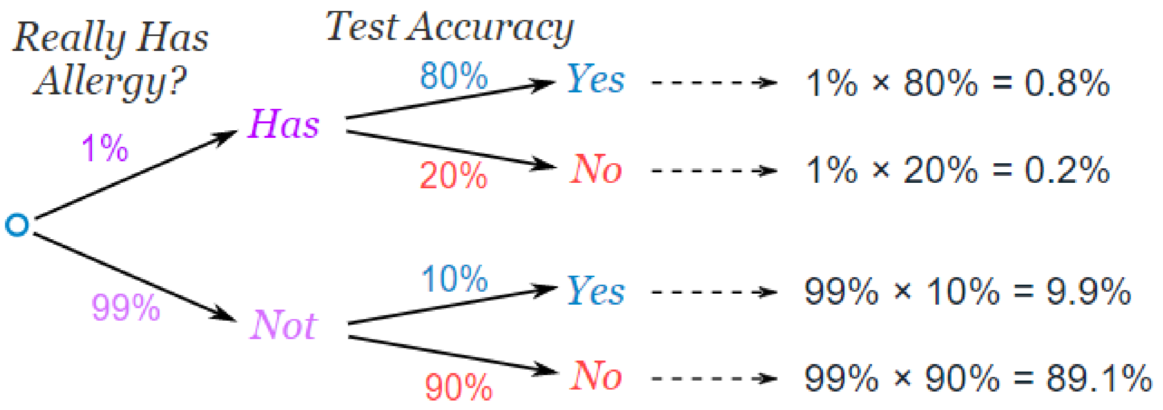
$$8 / 107 = \text{about } 7\%$$

So, even though Hunter's test said "Yes", it is still only **7% likely** that Hunter has a Cat Allergy.

Why so small? Well, the allergy is so rare that those who actually have it are greatly **outnumbered** by those with a false positive.

Solution 2: Tree Diagram

upGrad



First of all, let's check that all the percentages add up:

0.8% + 0.2% + 9.9% + 89.1% = **100%** (good!)

And the two "Yes" answers add up to 0.8% + 9.9% = **10.7%**, but only 0.8% are correct.

0.8/10.7 = **7%** (same answer as above)

Solution 3: Bayes' Theorem

upGrad

- Bayes' Theorem has a special formula for this kind of thing:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)}$$

where:

- P means "Probability of"
- | means "given that"
- A in this case is "actually has the allergy"
- B in this case is "test says Yes"

- So:
- **P(A|B)** means "The probability that Hunter actually has the allergy given that the test says Yes"
- **P(B|A)** means "The probability that the test says Yes given that Hunter actually has the allergy"

Solution 3: Bayes' Theorem

upGrad

- **P(A|B)** means "The probability that Hunter actually has the allergy given that the test says Yes"
- **P(B|A)** means "The probability that the test says Yes given that Hunter actually has the allergy"

To be clearer, let's change A to **has** (actually has allergy) and B to **Yes** (test says yes):

$$P(\text{has}|\text{Yes}) = \frac{P(\text{has})P(\text{Yes}|\text{has})}{P(\text{has})P(\text{Yes}|\text{has}) + P(\text{not has})P(\text{Yes}|\text{not has})}$$

And put in the numbers:

$$P(\text{has}|\text{yes}) = \frac{0.01 \times 0.8}{0.01 \times 0.8 + 0.99 \times 0.1}$$
$$= 0.0748...$$

Which is about **7%**

upGrad
#LifeKoKaroLift

Thank You!

Connect me: https://www.linkedin.com/in/dr-darshan-ingle-corporate-trainer/