

Traitement et Analyse de Données

<https://github.com/rebdau/TADE>

Big green button

A screenshot of a GitHub repository page for "rebdau/TADE". The page shows a list of files: "main" (1 branch, 0 tags), "rebdau Add files via upload", "data", "notes", and "scripts". A red arrow points from the handwritten text "Big green button" down to the "Code" dropdown menu, which is highlighted with a red circle. Another red circle highlights the "Download ZIP" button, also indicated by a red arrow.

The GitHub interface includes a navigation bar with links for Product, Team, Enterprise, Explore, Marketplace, Pricing, Search, Sign in, and Sign up. Below the navigation bar is a header for the repository "rebdau / TADE (Public)". The main content area has tabs for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, and Insights. On the right side of the page, there is an "About" section with a detailed description of the repository, mentioning it's a course named "Traitement et Analyse de données expérimentales" taught at Université de Picardie Jules Verne (UPJV) in Amiens, France. It covers statistics, probability, linear regression, and ANOVA. The "About" section also lists 0 stars, 1 watching, and 0 forks. Below the "About" section are sections for Releases, Packages, and Languages, all currently empty or showing 100% R usage. The bottom of the screen shows the Windows taskbar with various pinned icons and the system tray.

<https://github.com/rebdau/TADE>

Program

- Statistics fundamentals
- Hypothesis testing
- Simple linear regression
- Multiple linear regression
- ANOVA

Analyses to investigate the relationship between variables

		Independent (predictor) variable(s)
Type of variables	X is continuous (interval or ratio)	X is nominal or ordinal
Dependent variable	Y is continuous (interval or ratio)	Scatterplots Correlation Regression Side-by-side dot plot Side-by-side box plot ANOVA or t-tests
Y is nominal or ordinal	Logistic regression	Chi-square tests

1. Statistics fundamentals

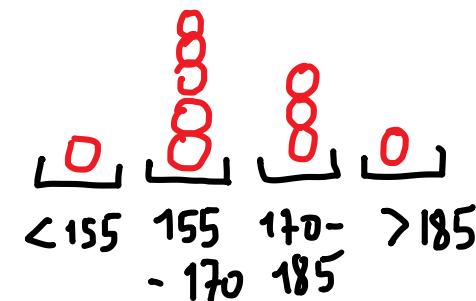
- ✓ Probability distributions
- ✓ Population parameters and Sample statistics
- ✓ Central Limit Theorem
- ✓ Student's t-distribution
- ✓ Confidence intervals for the mean
- ✓ Hypothesis testing

Probability distributions

Distributions

- Put measurements in bins → you get a histogram

Example: human height (cm)

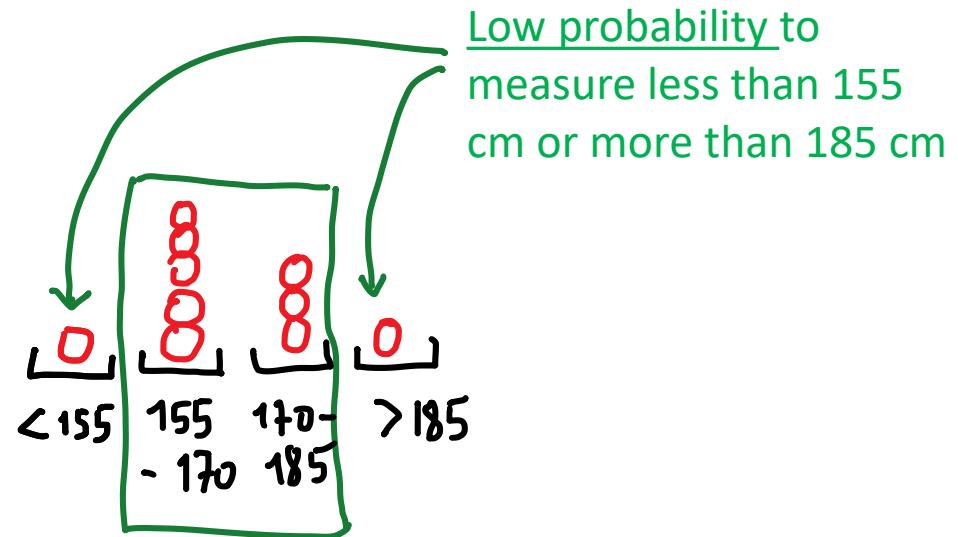


Distributions

- Put measurements in bins → you get a histogram

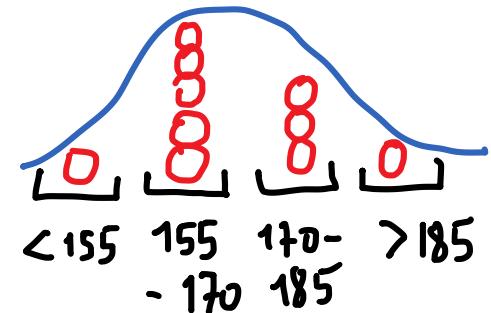
Example: human height (cm)

Smaller bins → more precise estimate of the distribution



Most measurements are between 155 and 185.
→ High probability

Distributions

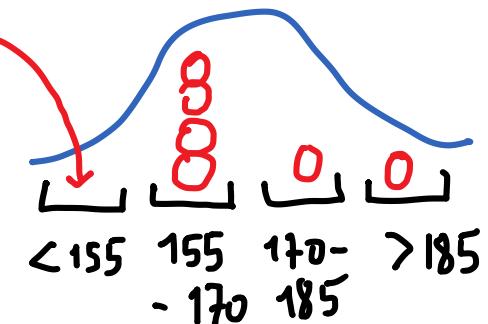


- Curve to approximate the histogram:

Low number of measurements (time or money are limiting!):

The approximate curve, based on the mean and standard deviation of the data we were able to collect, is usually good enough.

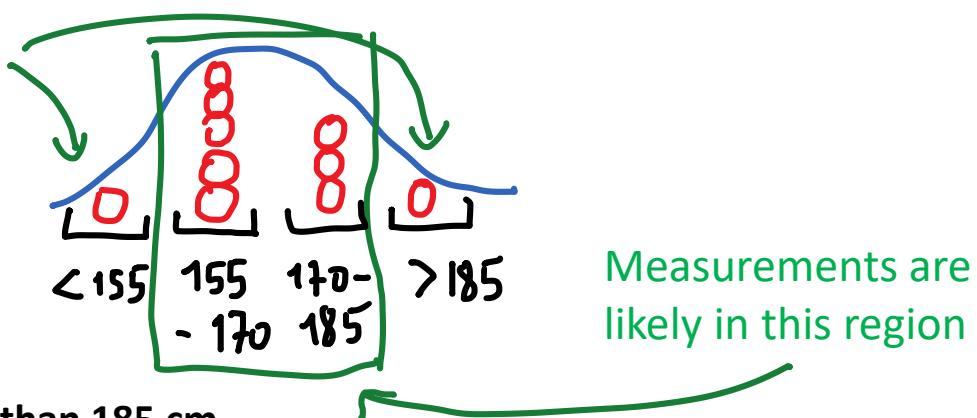
Not for all bins measurements
→ use curve to estimate probability



Distributions

- Histogram and Curve are probability distributions:
They show probabilities of how measurements are distributed

Measurements are less likely in these regions



Probability of humans being taller than 185 cm

(using histogram)

= number of humans taller than 185 cm / total number of humans

= 0.10

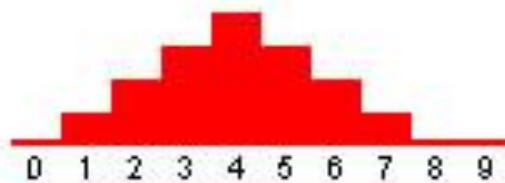
(using curve)

= area under the curve for all measurements above 185 cm / total area under the curve

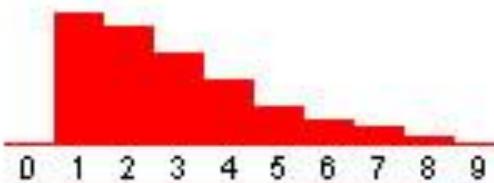
= 0.10 / 1

= 0.10

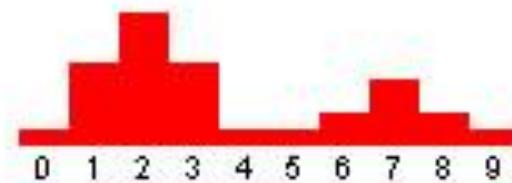
Distribution shapes



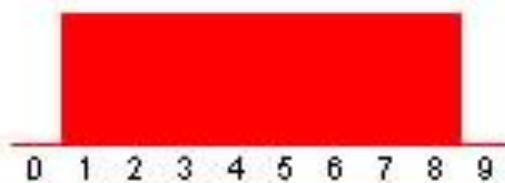
Symmetric, unimodal,
bell-shaped



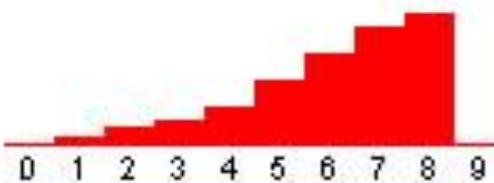
Skewed right



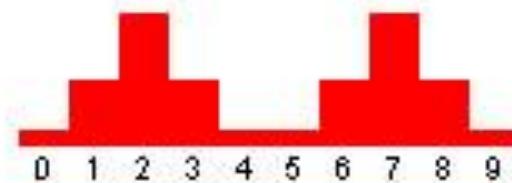
Non-symmetric, bimodal



Uniform



Skewed left



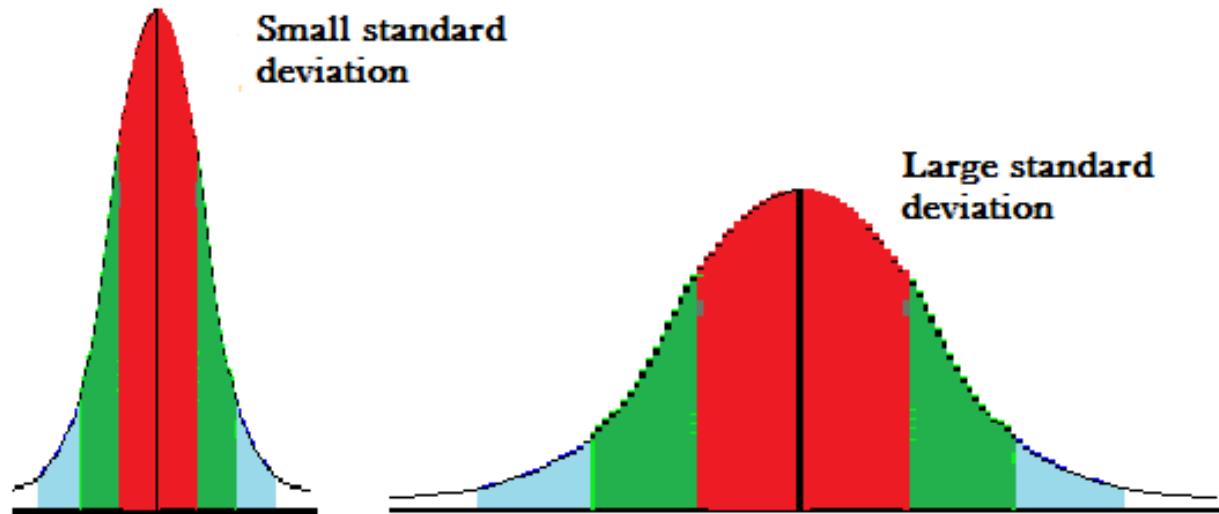
Symmetric, bimodal

Among all distributions, the normal distribution is the most common in nature.
Why?....The **CENTRAL LIMIT THEOREM!**

Normal distribution

- = Gaussian
- = Bell shaped curve
- Centered around the average value
- The width is defined by the standard deviation (SD)
- 95% of the measurements fall between $+/- 2$ SD around the mean
- Very common in nature.
Reason = « Central Limit Theorem » (*see later!*)

Normal distribution



Distance from
the Mean

Percentage of Values
Falling Within Distance

$$\mu \pm 1\sigma$$

68

$$\mu \pm 2\sigma$$

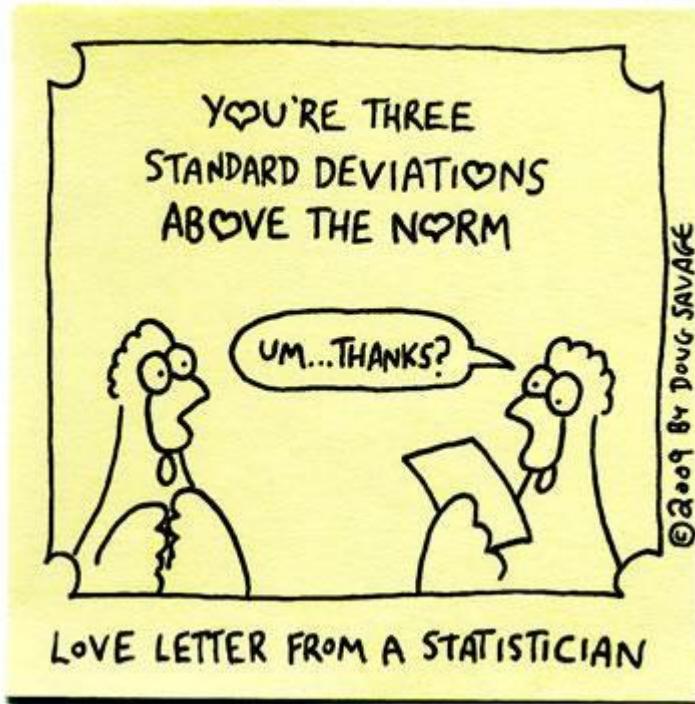
95

$$\mu \pm 3\sigma$$

99.7

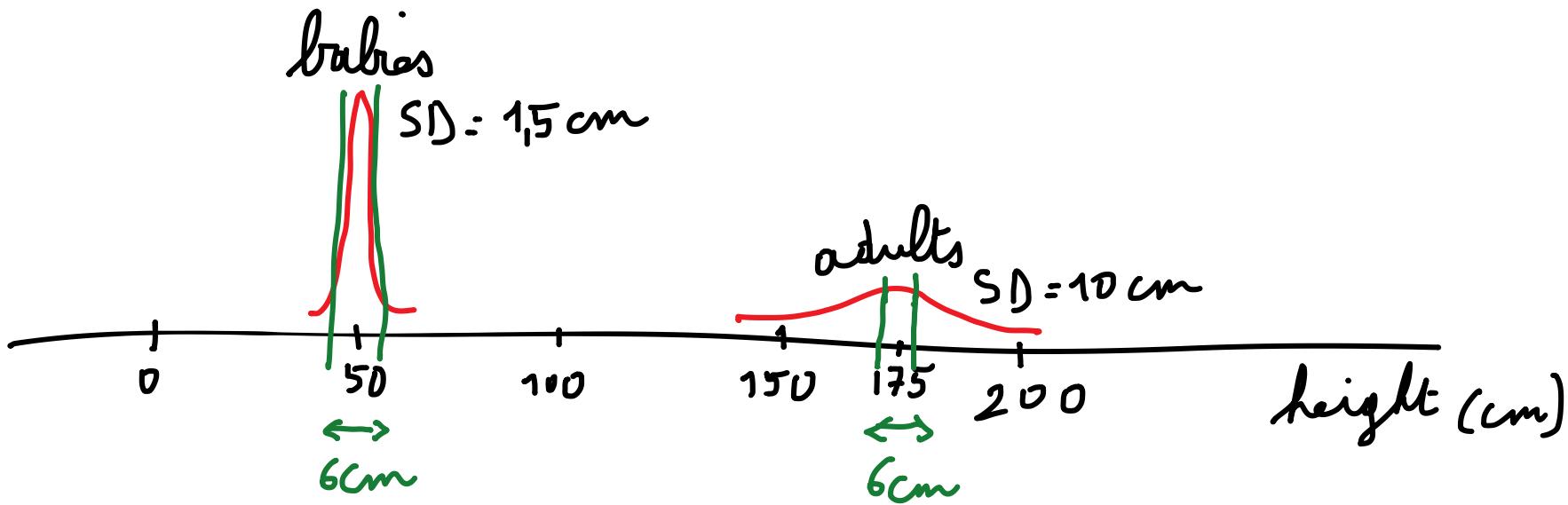
Savage Chickens

by Doug Savage



www.savagechickens.com

Normal distribution

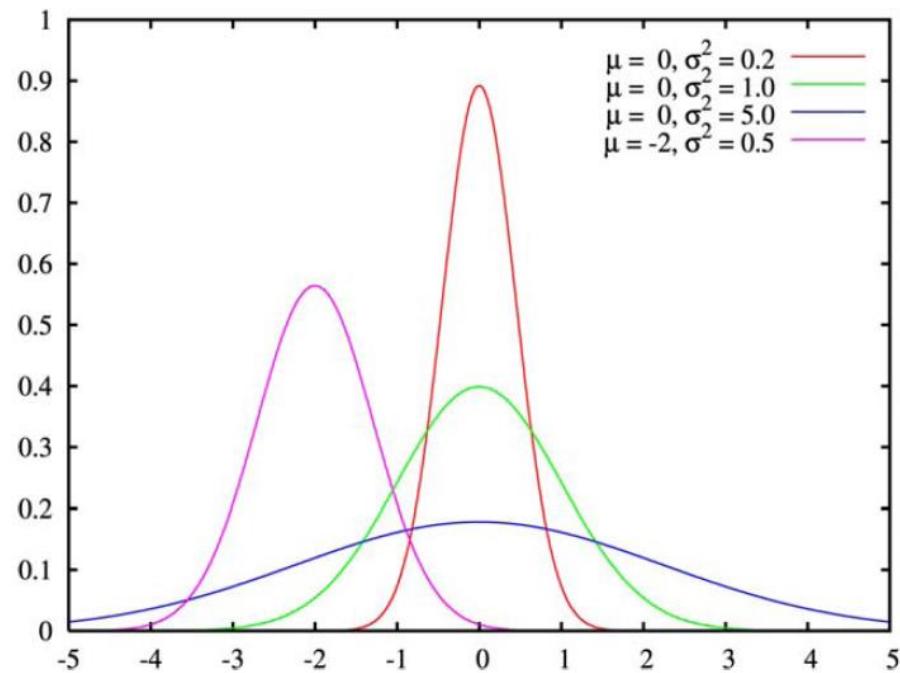


The width of the curve (SD) determines how tall it is:

- High probability to measure a baby within +/- 3 cm of the mean (95 %)
- Low probability to measure a man within +/- 3 cm of the mean (*because there are much more options for a man! → high SD*)

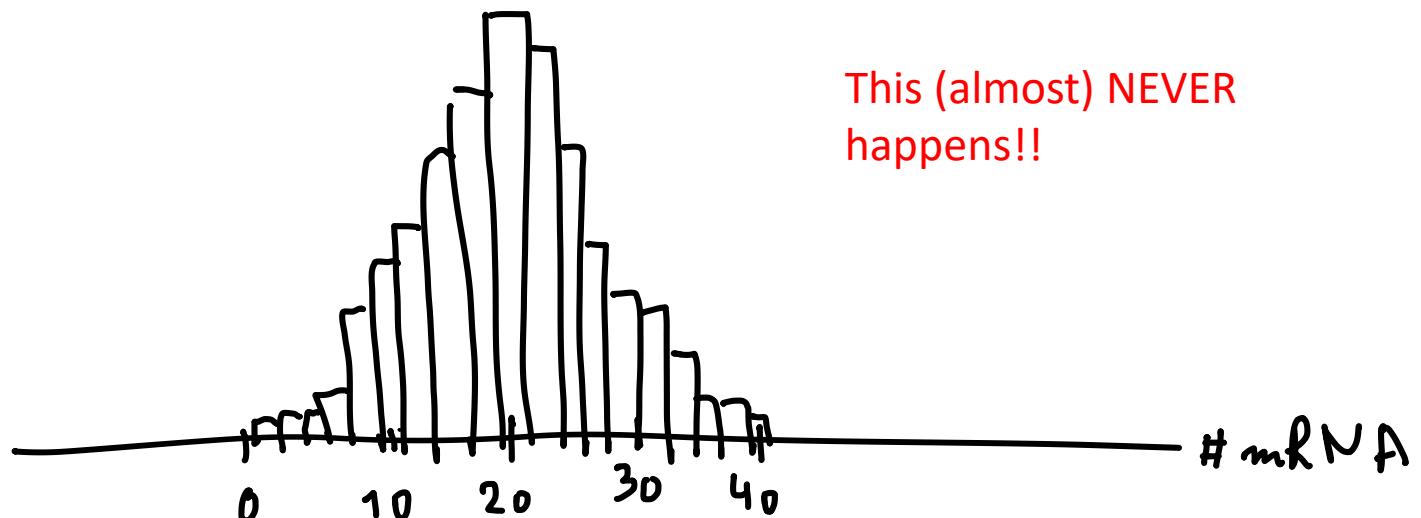
Normal distribution

- Symmetric around μ
- **Defined by μ and σ^2**
- If we know that a variable has a normal distribution, and we know these parameters, then we know the probability of getting any particular value for the variable.
- The area under the curve = probability
- Total area under the curve = 1



Population parameters

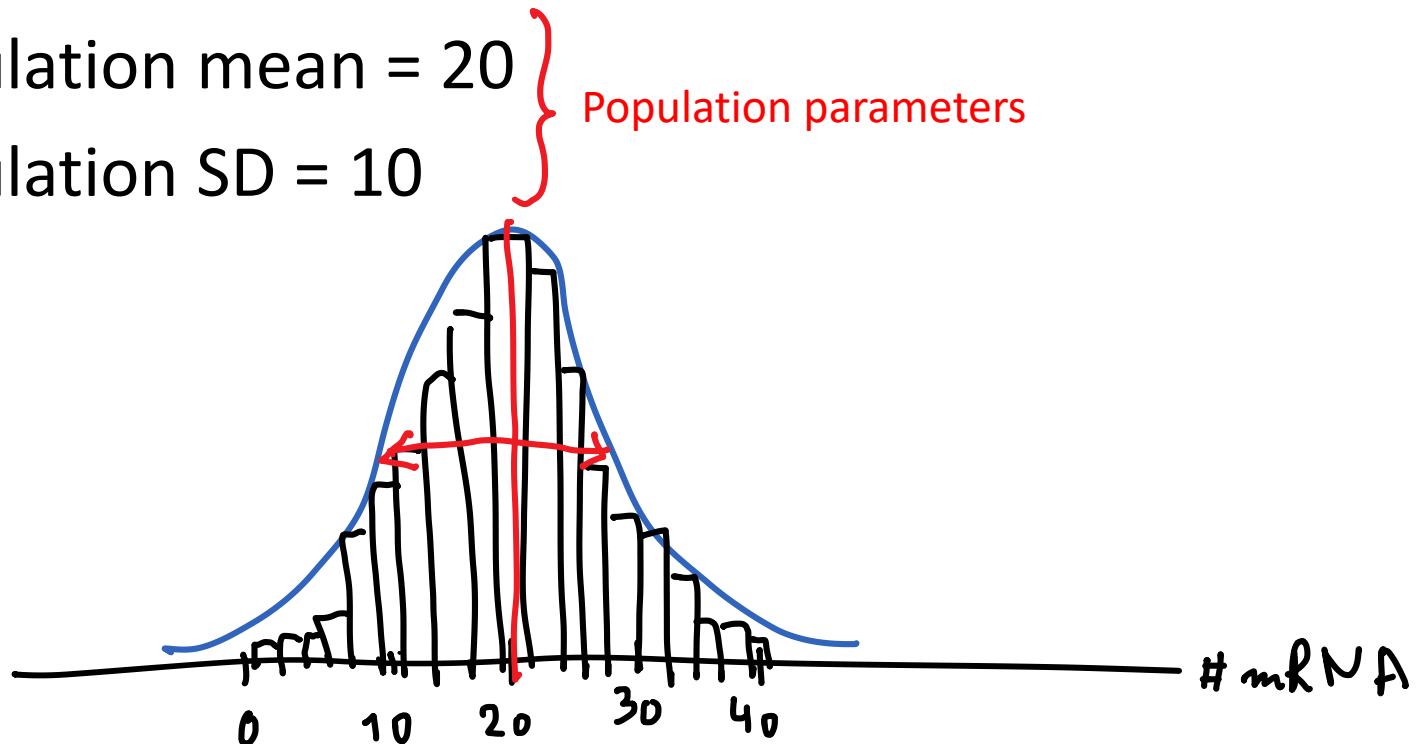
- Imagine: you had the tools, time and money to measure the number of mRNA transcripts of gene X in **ALL cells of the liver (240 000 000 cells) = POPULATION**



Population parameters

- (If the histogram looks like this,) we can fit a normal distribution to the data, with

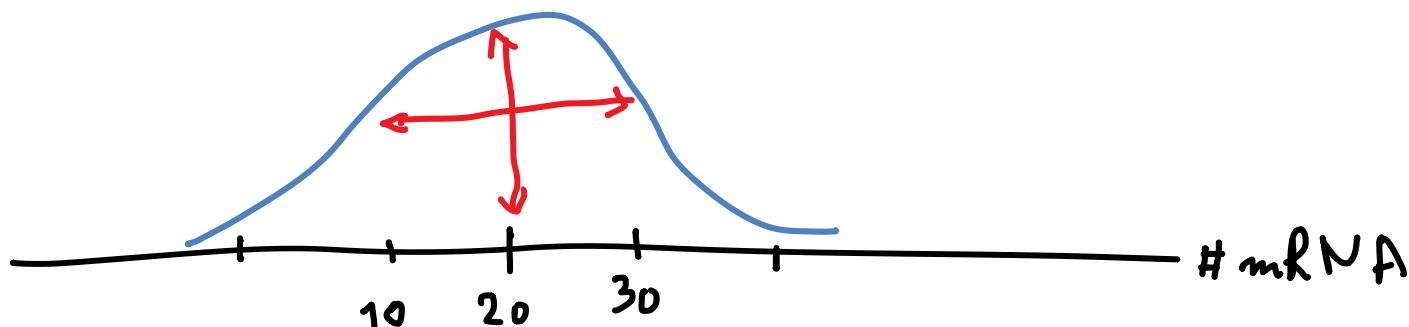
- population mean = 20
- population SD = 10



Population parameters

- (Almost) ALWAYS: we don't have these tools/time/money!!
- We estimate the population parameters using a relatively small number of measurements (= sample).

Population parameters



exp. 1: 5 cells



exp 2: 5 other cells



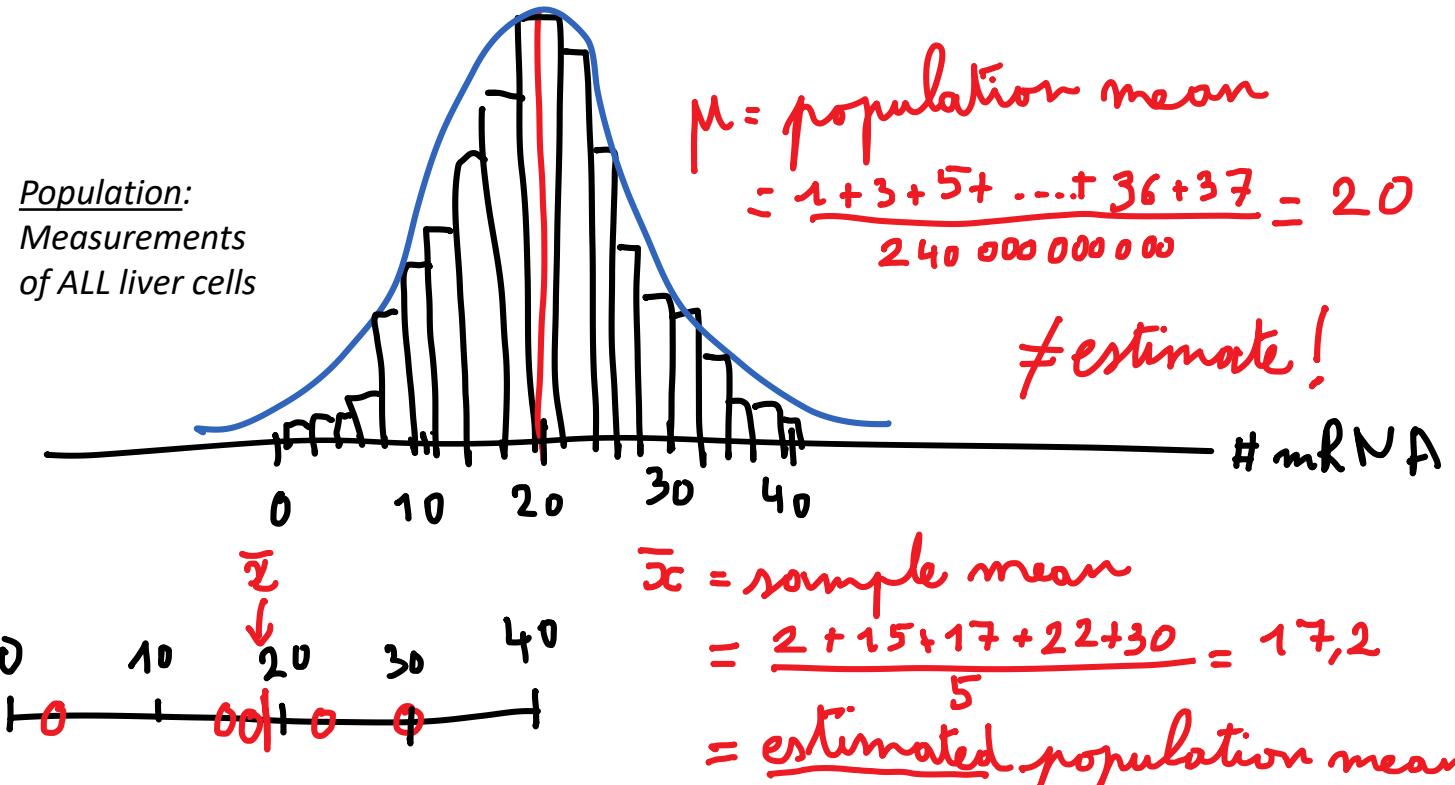
Instead of just describing the measurements of 5 cells (out of 240 000 000 000), we use them to estimate the population parameters (exp. 1)

New measurements in future experiments (exp. 2) will come from the same population!

We want to know the population parameters to ensure that the results drawn from our experiment are **reproducible**.

Mean

- In order to fit a normal curve to a histogram, we need to calculate **mean** and **variance** or **standard deviation (SD)** = Population parameters.
- Center normal curve over population mean
- $\bar{x} \neq \mu$ but with larger sample, \bar{x} gets closer to μ



Variance and SD

$$\sigma^2 = \text{Population variance} = \frac{\sum(x - \mu)^2}{N} = 100 \quad \begin{matrix} \leftarrow \text{to ensure that each difference} \\ \text{is positive!} \end{matrix}$$

σ^2 = average of the square differences between the data and μ

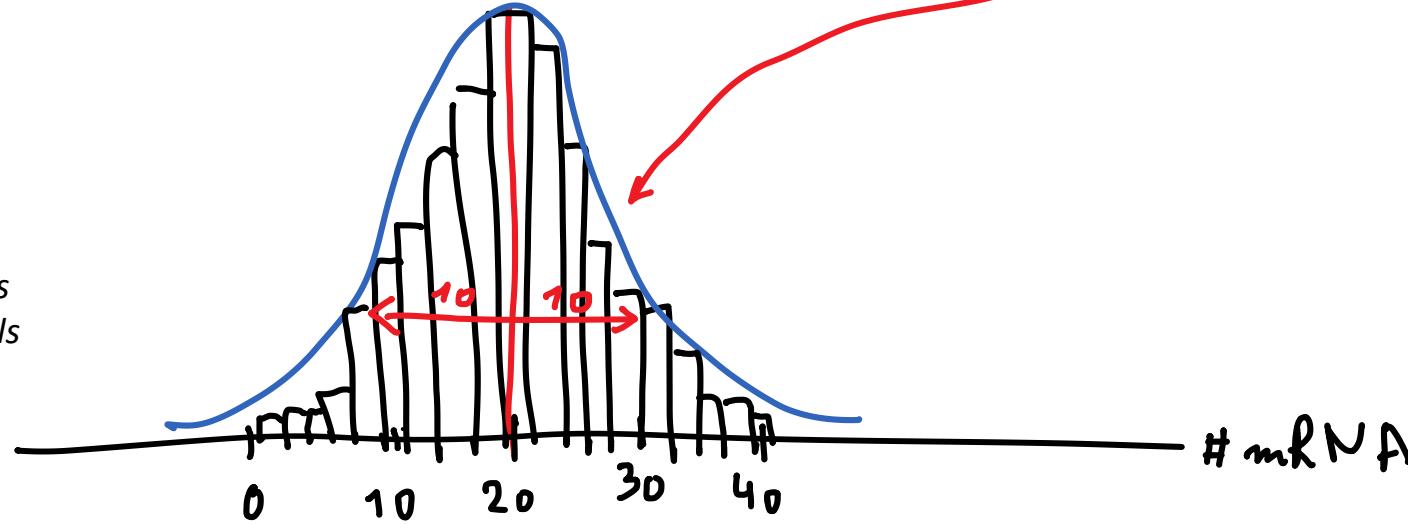
But the units for σ^2 , are mRNA²

→ Cannot plot σ^2 on X-axis

To solve this problem: take the square root :

$$\sigma = \text{Population Standard Deviation} = \sqrt{\frac{\sum(x - \mu)^2}{N}} = \sqrt{100} = 10$$

Population:
Measurements
of ALL liver cells



Variance and SD

estimate

$$\sigma^2 = \text{Population variance} = \frac{\sum(x - \mu)^2}{N} = 100$$

σ^2 = average of the square differences between the data and μ

But the units for σ^2 , are mRNA²

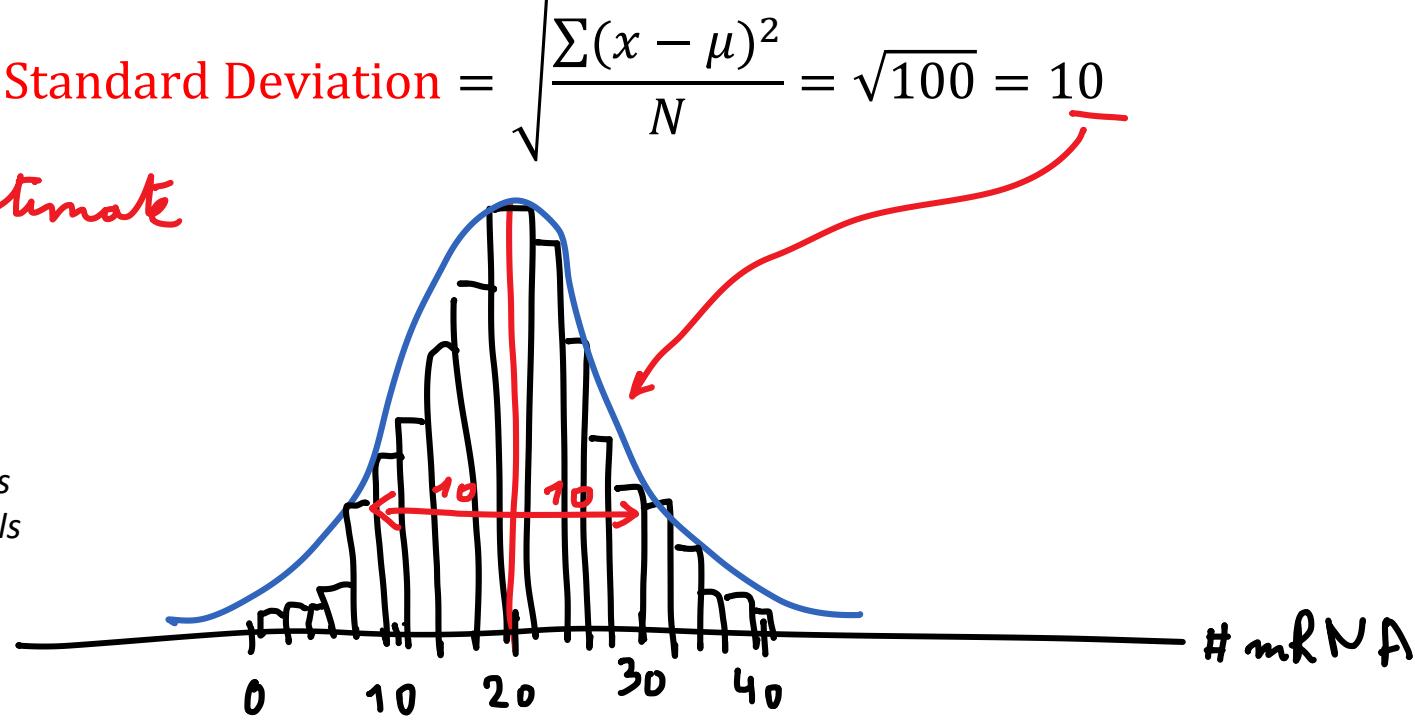
→ Cannot plot σ^2 on X-axis

To solve this problem: take the square root :

$$\sigma = \text{Population Standard Deviation} = \sqrt{\frac{\sum(x - \mu)^2}{N}} = \sqrt{100} = 10$$

estimate

Population:
Measurements
of ALL liver cells



Variance and SD

But we almost never have population data!

So we almost never calculate μ , σ^2 , and σ !

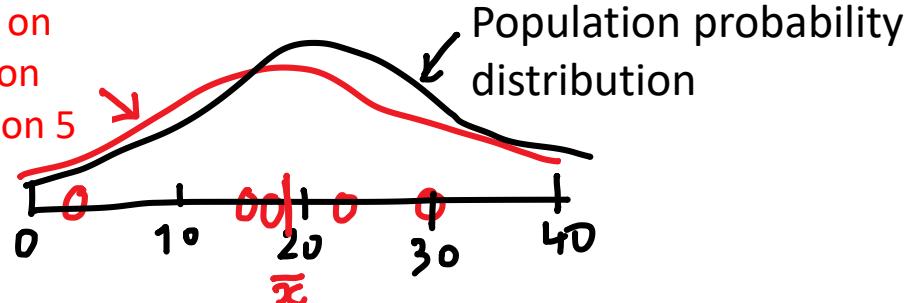
Instead, we estimate μ , σ^2 , and σ from a relatively small sample.

$$s^2 = \text{Estimated Population variance} = \frac{\sum(x - \bar{x})^2}{n - 1} = 105.7$$

$$s = \text{Estimated Population Standard deviation} = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = 10.3$$

Dividing by $n - 1$ instead of n , compensates for the fact that we are calculating differences from the sample mean instead of the population mean. Otherwise, we would consistently underestimate the variance around the population mean.

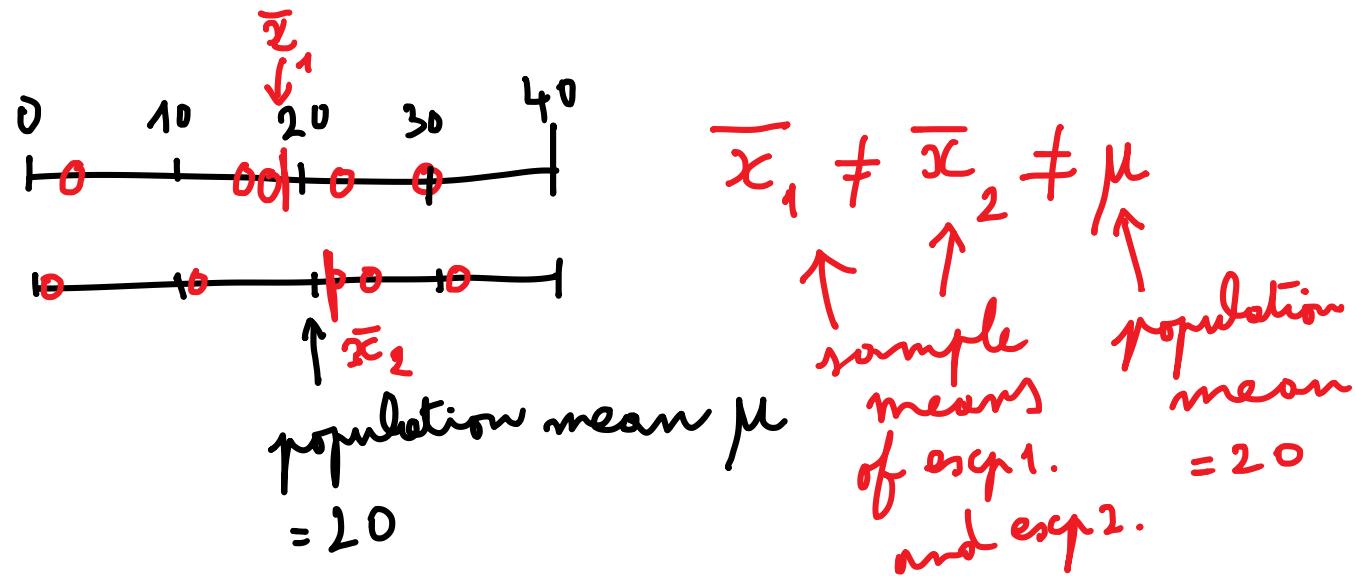
Estimated Probability
Distribution, based on
estimated population
parameters, based on 5
observations



$$\sqrt{\frac{\sum(x - \bar{x})^2}{n}} = 9.2$$

Population parameters

- Problem: at each repetition of an experiment, we get different estimates of the population parameters and all these estimates are different from the true population values!



Population parameters

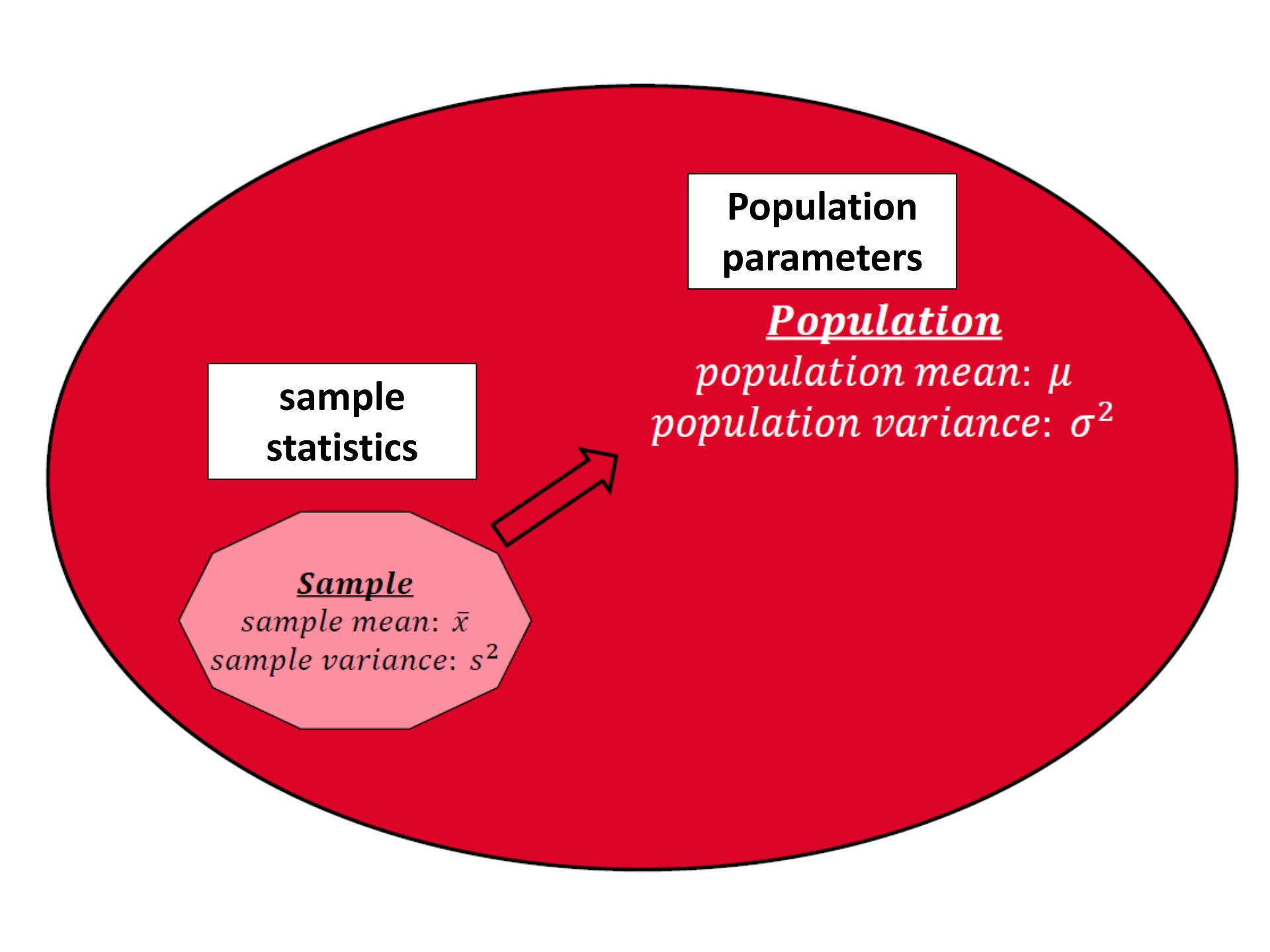
- The **more data** we have, the **more confidence** we can have in the accuracy of the estimates
- **GOAL of statistics:** quantifying how much confidence we can have in estimated population parameters

→ calculate p-values and confidence intervals

 → Tell if estimated population parameters of two experiments are significantly different

If estimated population parameters of two experiments are not significantly different, the measurements of the two experiments are samples of the same population

- **Descriptive Statistics:** summarize the sample data as means, variances, ranges, etc.
- **Inferential Statistics:** use the sample statistics to estimate the parameters of the population (confidence interval)



**Population
parameters**

Population

population mean: μ

population variance: σ^2

**sample
statistics**

Sample

sample mean: \bar{x}

sample variance: s^2

Sample Statistics to Estimate Population Parameters:

If simple random sampling (every observation has the same chance of being selected) is used to select n from N , then:

- Sample estimates are **unbiased estimates of their counterparts** (e.g., sample mean estimates the population mean), meaning that over all possible samples the sample statistics, averaged, would equal the population statistic.
- A particular sample value (e.g., sample mean) is called a “**point estimate**” – do not necessarily equal the population parameter for a given sample.
- Can calculate an interval where the true population parameter is likely to be, with a certain probability. This is a **Confidence Interval**, and can be obtained for any population parameter, IF the distribution of the sample statistic is known.

Parameters for populations

(→using Greek letters!)

MEASURES OF CENTRAL TENDENCY

1. Mean -- μ

e.g. for $N=5$

$$Y_1=24; Y_2=13; Y_3=19, Y_4=26, Y_5=11$$
$$\mu=18.6$$

2. Standard Deviation σ and Variance σ^2

$$\sigma^2 = \sum_{i=1}^N (y_i - \mu)^2 / N$$

$$\sigma = \sqrt{\sigma^2}$$

Statistics from the sample

(= estimates of the population parameters!)

1. Mean -- \bar{y} e.g. for $n=3$ and $y_1=5; y_2=6; y_3=7$, $\bar{y}=6$
2. Standard Deviation s and Variance s^2 (often called “mean square”)

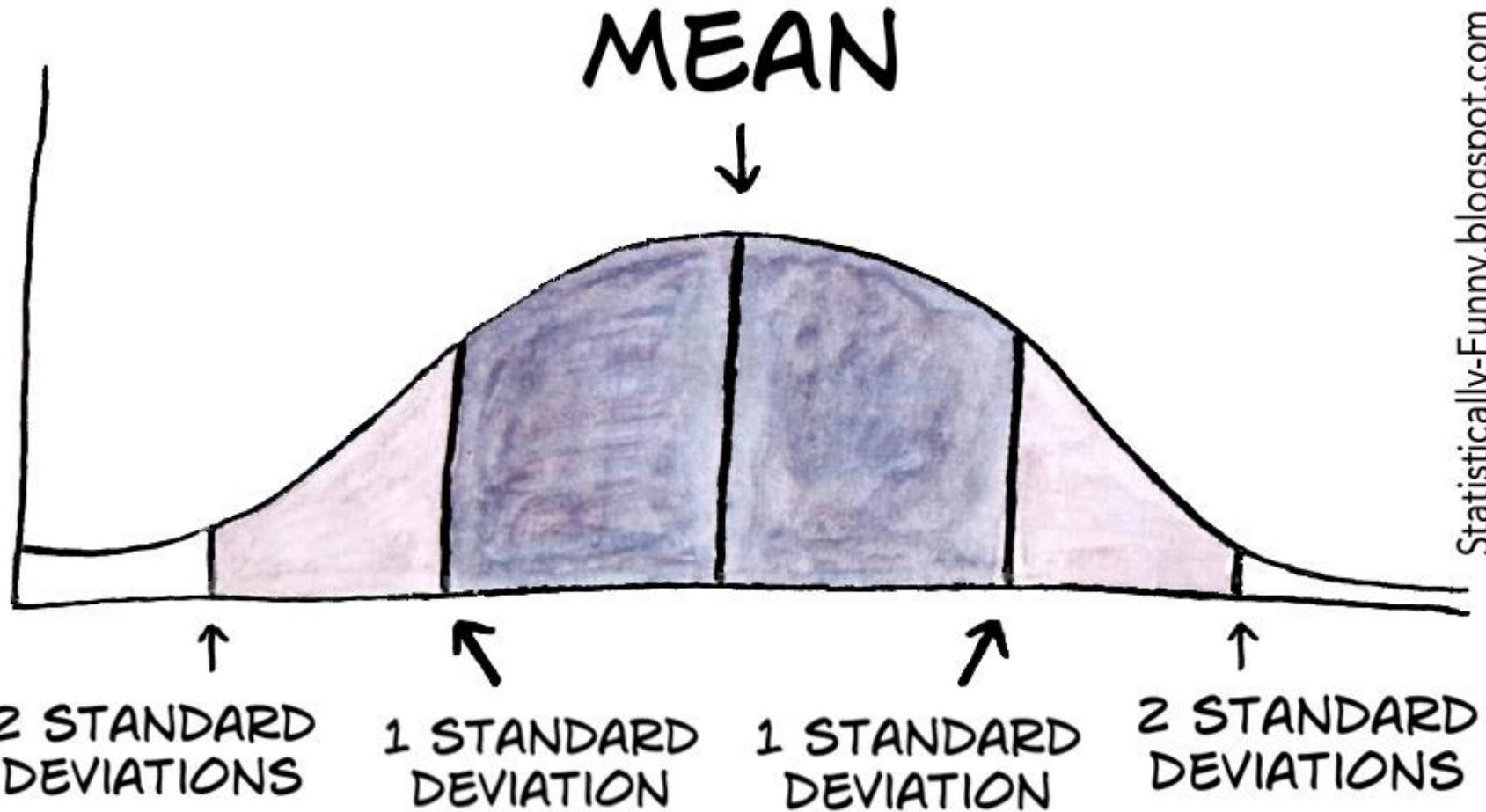
$$s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)$$

$$s = \sqrt{s^2}$$

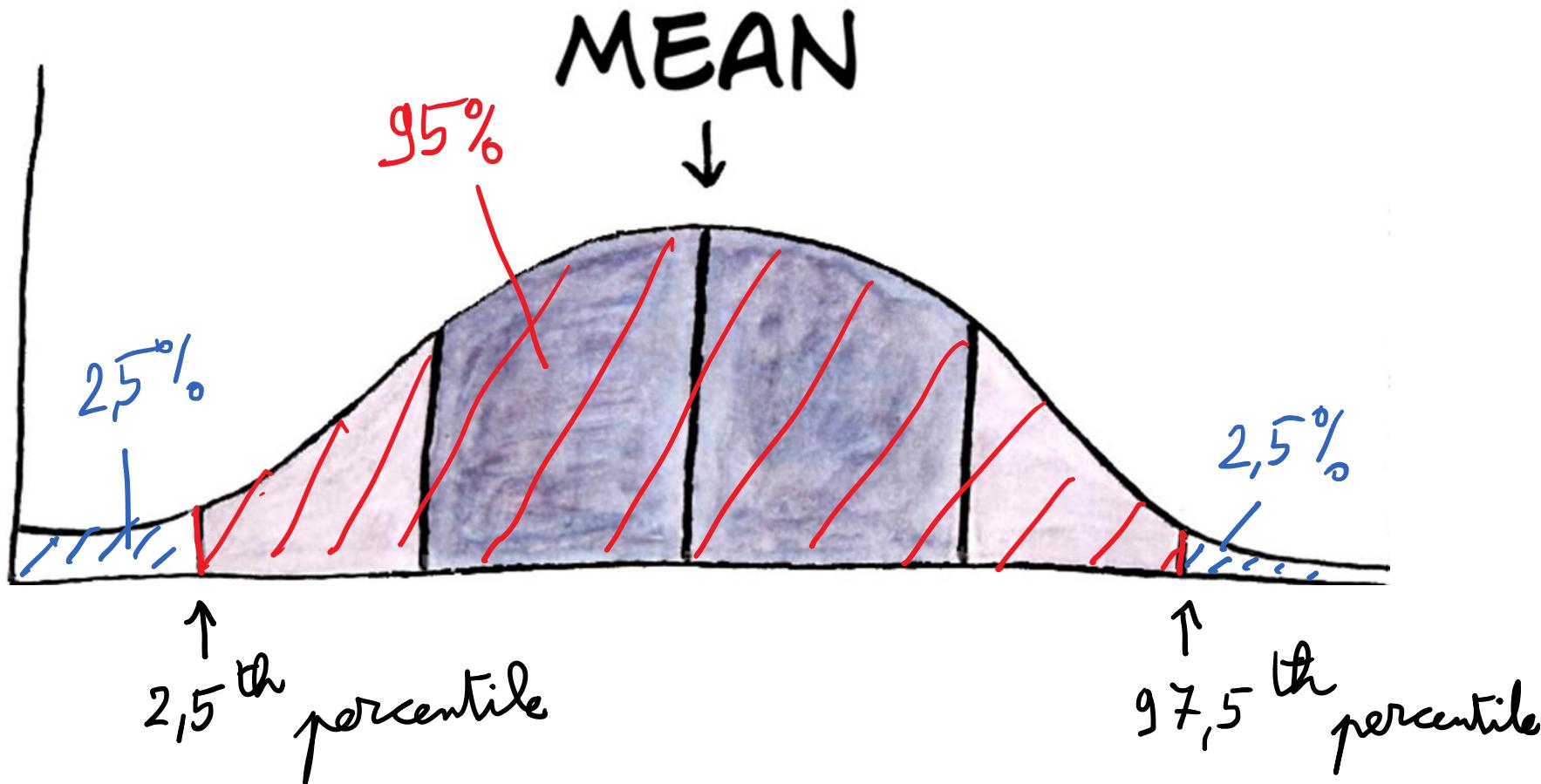
The fact that the sample mean is used instead of population mean makes the sum of squares of deviations smaller. You adjust for this by dividing by $n-1$ so that the sample variance would be an unbiased estimator of the population variance

A rough estimate of the standard deviation is

$$s \approx \frac{\text{range}}{4}$$



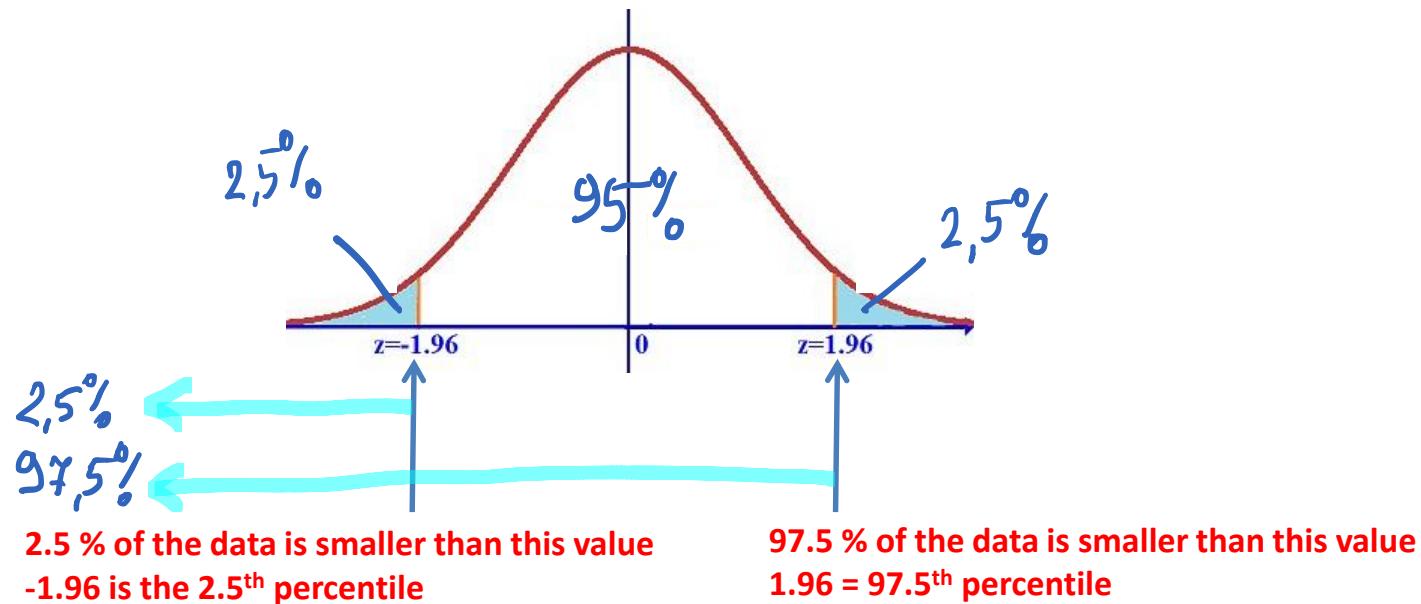
Percentiles



percentile

The n^{th} **percentile** of an observation variable is the value that cuts off the first n percent of the data values when it is sorted in ascending order.

(**quantile** is expressed in fractions vs. **percentile** is expressed in pct)



Normal distribution R functions

#random generation

rnorm(n, mean = 0, sd = 1)

```
> rnorm(5, mean = 0, sd = 1)
[1] -1.3986100  0.3028934 -0.9172586 -0.7261859 -1.6761790
```

#give quantile value based on probability

qnorm(p, mean = 0, sd = 1)

```
> qnorm(0.05/2 , mean = 0, sd = 1)
[1] -1.959964
```



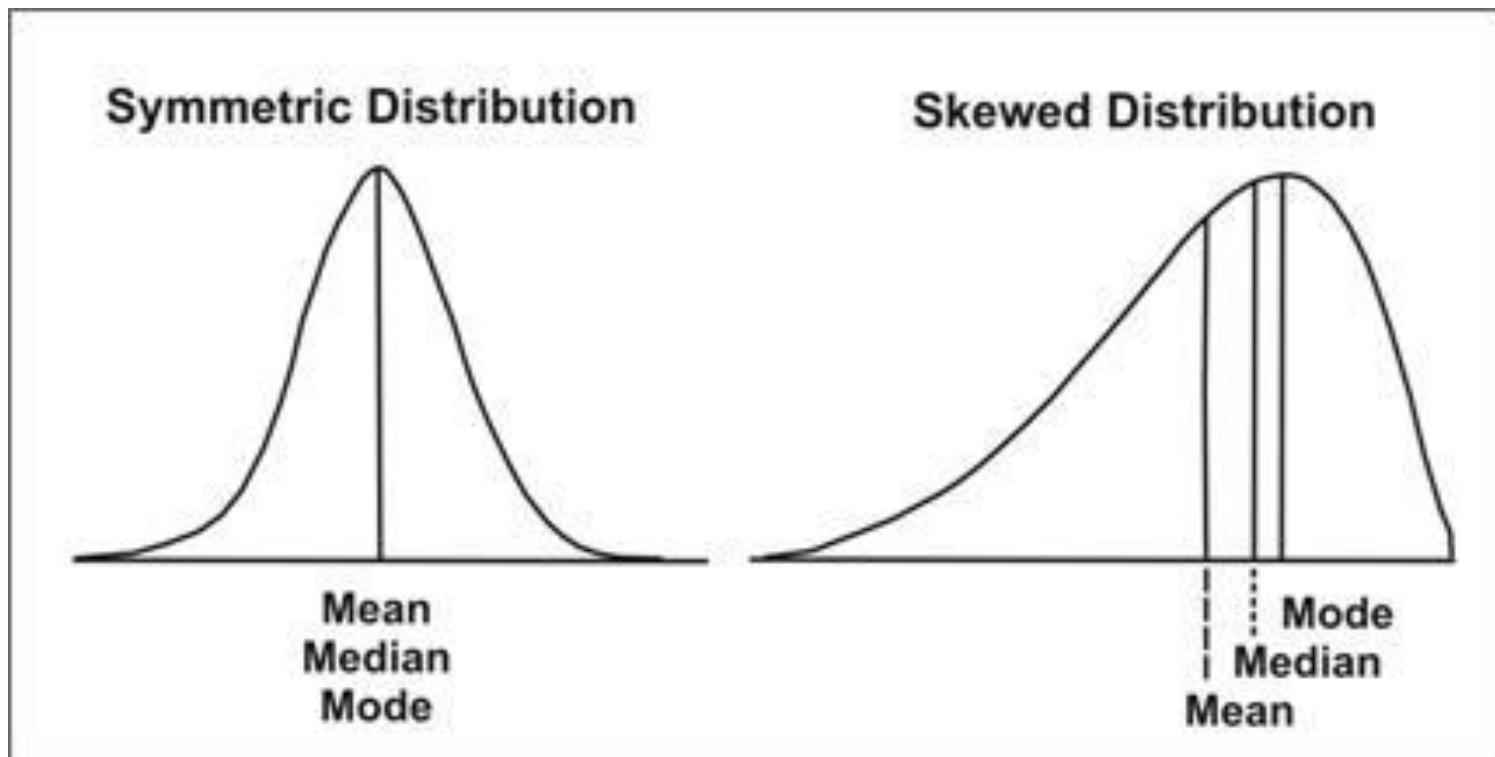
#give probability based on quantile value

pnorm (q, mean = 0, sd = 1)

```
> pnorm (-1.96, mean = 0, sd = 1)
[1] 0.0249979
```

Median -- value which divides the distribution (50% of N observations are above and 50% are below)
= 50th percentile

Unaffected by extremely large and extremely small values.



Median: Computational Procedure

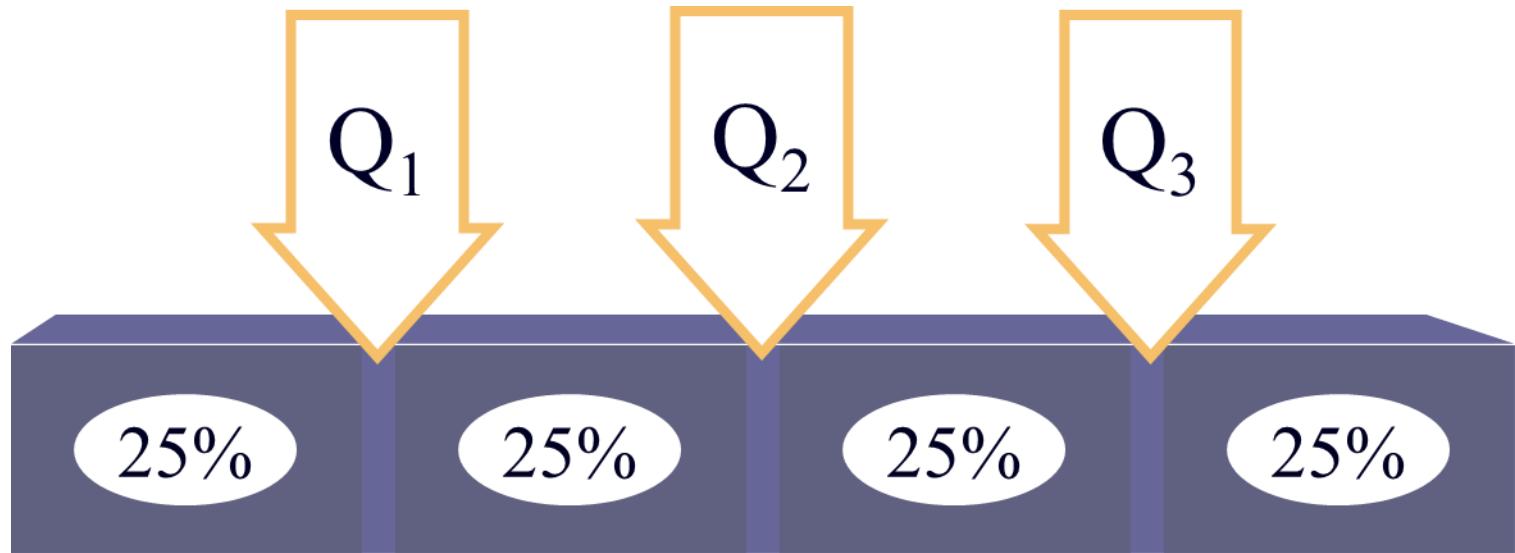
- First Procedure
 - Arrange observations in an ordered array.
 - If number of terms is odd, the median is the middle term of the ordered array.
 - If number of terms is even, the median is the average of the middle two terms.
- Second Procedure
 - The median's position in an ordered array is given by $(n+1)/2$.

DISCRETE VARIABLES

CONTINUOUS VARIABLES

Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of values is known		✓	✓	✓
“Counts,” aka “Frequency of Distribution”	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiply and divide values				✓
Has “true zero”				✓

Quartiles



Quartiles, *continued*

- Q_1 is equal to the 25th percentile
- Q_2 is located at 50th percentile and equals the median
- Q_3 is equal to the 75th percentile

Quartile values are not necessarily members of the data set

Quartiles: Example

- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129

- $$Q_1: i = \frac{25}{100}(8) = 2 \quad Q_1 = \frac{109+114}{2} = 111.5$$

- $$Q_2: i = \frac{50}{100}(8) = 4 \quad Q_2 = \frac{116+121}{2} = 118.5$$

- $$Q_3: i = \frac{75}{100}(8) = 6 \quad Q_3 = \frac{122+125}{2} = 123.5$$

R functions for sample statistics

`summary()`

`max()`, `min()`, `range()`

`mean()`, `median()`

`var()`,

`quantile()`, `IQR()`

→these functions work on **vectors** and full **matrices**

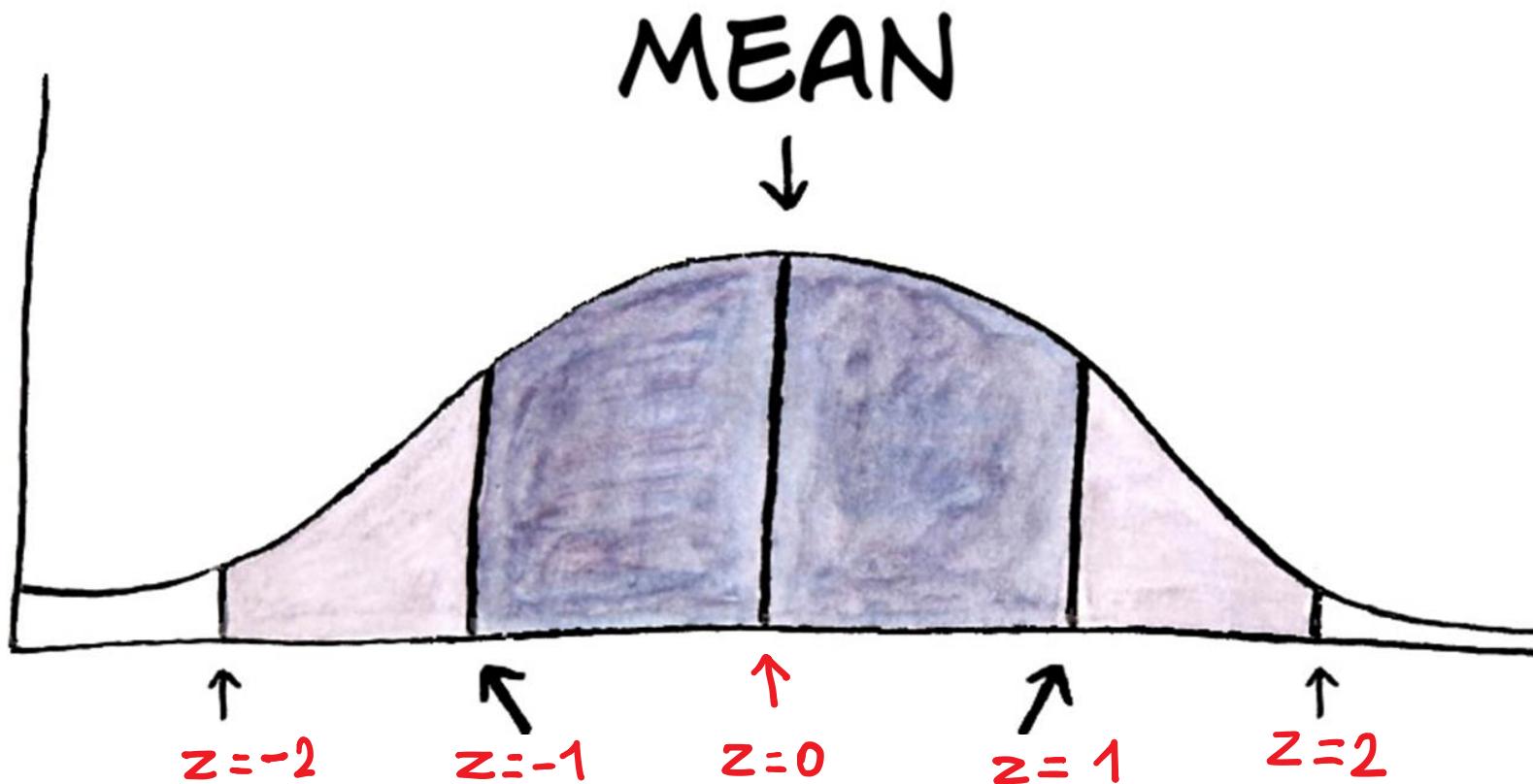
Exercise: In the "ufc_processed" dataset,

- what are the mean, median, standard deviation, minimum, maximum, and interquartile range of the diameter (in centimeter)?

script 1_confidenceinterval.R

Standardization

Z-score = number of standard deviations that a particular value is away from the mean



Z-score

- z-scores = standard score
- **Transformation to the z-score = standardization = scaling:**
scale the values for y by subtracting the mean, and dividing by the standard deviation.

$$z_i = \frac{y_i - \mu}{\sigma}$$

- z-score is the number of standard deviations that a particular value is away from the mean
- Z-score follows the standard normal distribution $N(0,1)$

E.g., for mean = 20, and standard deviation of 2 and $y = 10$,
 $Z = -5$ (an extreme value)

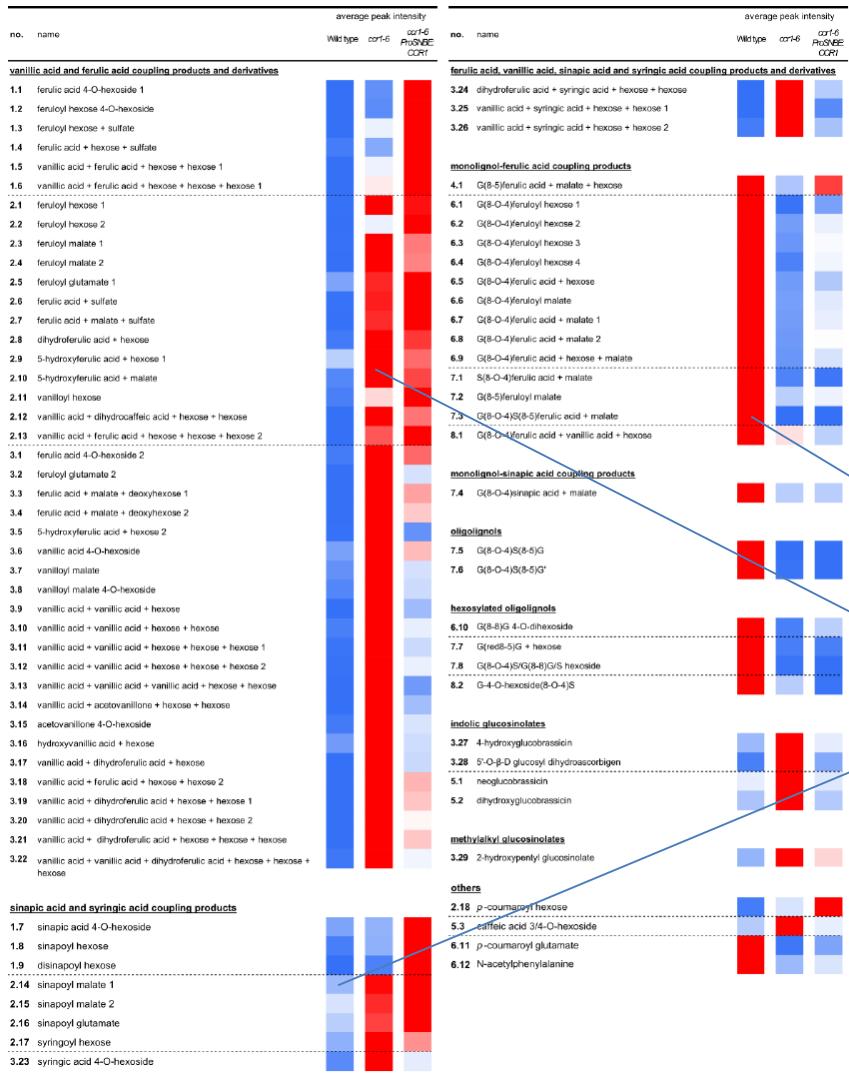


Figure 7. List of putatively structurally characterized metabolites with a different abundance in inflorescence stems of *ccr1-6 ProSNBE:CCR1* plants as compared with the wild type or *ccr1-6* mutants.

Eighty-two compounds were characterized, of which the abundance was significantly different between *ccr1-6 ProSNBE:CCR1* and the wild type, on the one hand, and/or *ccr1-6 ProSNBE:CCR1* and *ccr1-6*, on the other hand. The differentially accumulating metabolites were classified into eight different groups (Fig. 6B). Each putatively structurally characterized metabolite has a unique number that represents (1) the group that the metabolite belongs to and (2) the ranking within this group. The metabolites are listed per metabolic class, and the dashed lines separate the different groups present in a specific metabolic class. The average peak intensities are represented by a heat map, ranging from low values represented in blue to high values represented in red (wild type, n = 8; *ccr1-6*, n = 6; and *ccr1-6 ProSNBE:CCR1*, n = 13).

THIS IS WRONG!

43519
43352
5179

Different colors but same value
The blue value is higher than the red value

- Normal distribution for $\mu=0$ and $\sigma^2=1$
 $= N(0,1)$
= standard normal distribution
= Z-distribution
- Probability tables for $\mu = 0$ and $\sigma^2 = 1$, are often called **z-tables**.

- Example:

$$P(-1.96 < z < 1.96) = 0.95$$

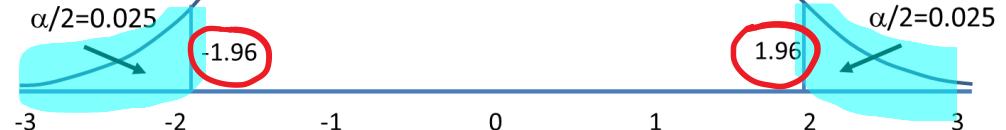
Notation:

For significance level $\alpha = 0.05$:

$$\text{Critical z-value} = z_{\alpha/2} = z_{0.025} = -1.96$$

$= 2,5^{\text{th}}$ percentile

95%

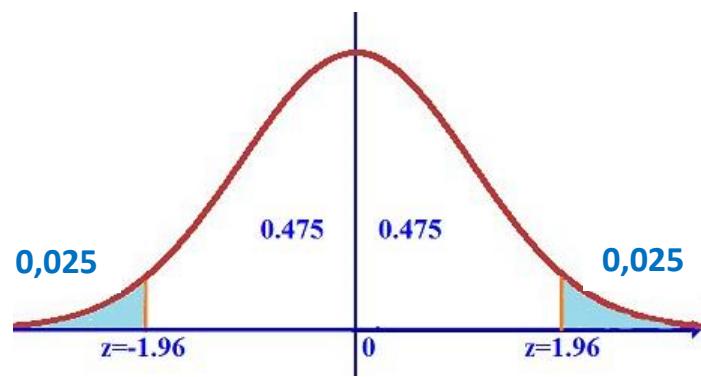
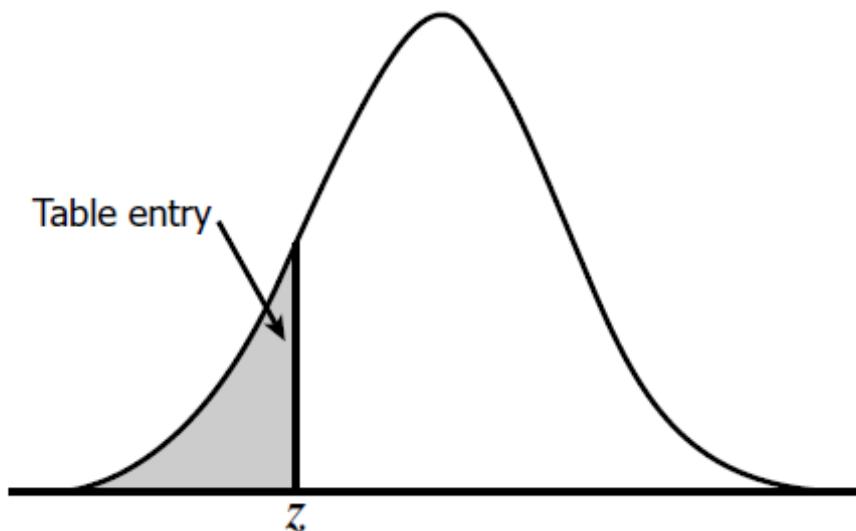


critical z-values for 95% confidence interval

Using the Z-table to find the critical z-value at significance level $\alpha=0.05$:

The z-table gives areas = probabilities

Look up the z-value corresponding to $\alpha/2 = 0.025$ in the table: 1.96!



Example

- **Problem**

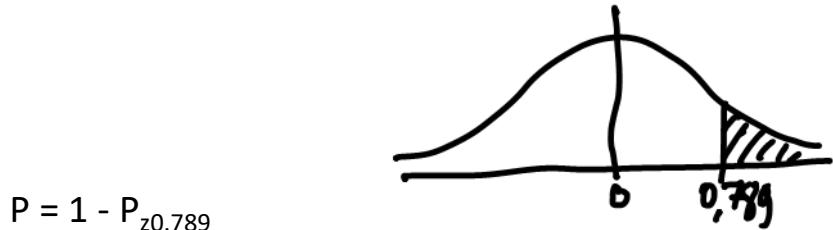
Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

- **Problem**
normal distribution; mean = 72; sd = 15.2
percentage above 84?
- **Solution using Probability tables**

We transform the value to the corresponding z-value

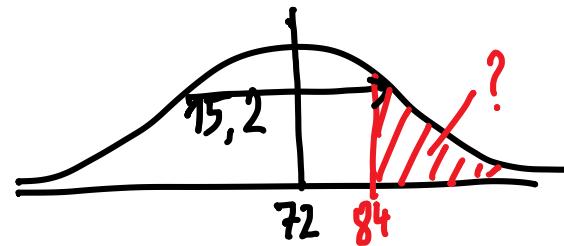
$$z = (84-72)/15.2 = 0.789$$

We draw the z-distribution (standard normal) and determine which area we are looking for

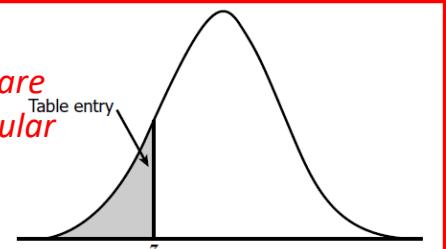


We look up the P-value for $z = -0.789$ in the table

$$P = 0.2148$$



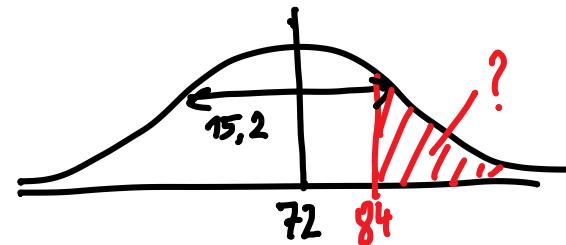
Check which probability values are given in this particular table!



Answer

The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

- **Problem**
normal distribution; mean = 72; sd = 15.2
percentage above 84?



- **Solution using R**

the function

```
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

gives the probability to be under (lower.tail) or above (lower.tail = FALSE) a certain percentile (quantile) value.

We don't need to transform to z-value: we can specify the mean and sd.

Here we are looking for the percentage of students scoring higher than 84, we are interested in the *upper tail* of the normal distribution.

```
> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
[1] 0.21492
```

- **Answer**

The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

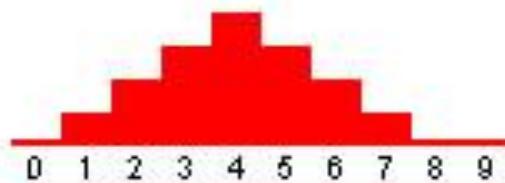
Sample Statistics to Estimate Population Parameters:

If simple random sampling (every observation has the same chance of being selected) is used to select n from N , then:

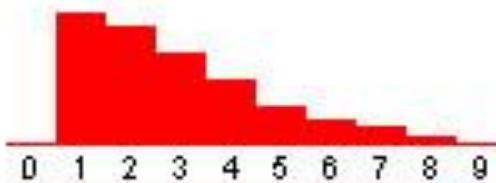
- Sample estimates are **unbiased estimates of their counterparts** (e.g., sample mean estimates the population mean), meaning that over all possible samples the sample statistics, averaged, would equal the population statistic.
- A particular sample value (e.g., sample mean) is called a “**point estimate**” – do not necessarily equal the population parameter for a given sample.
- Can calculate an interval where the true population parameter is likely to be, with a certain probability. This is a **Confidence Interval**, and can be obtained for any population parameter, **IF the distribution of the sample statistic is known.**

The Central Limit Theorem

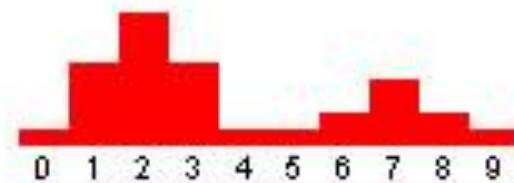
Distribution shapes



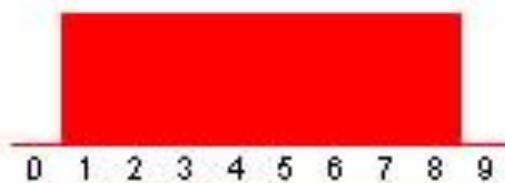
Symmetric, unimodal,
bell-shaped



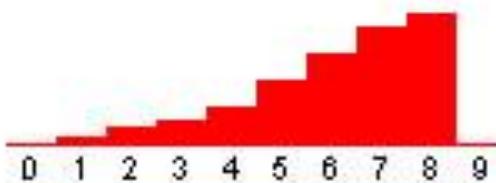
Skewed right



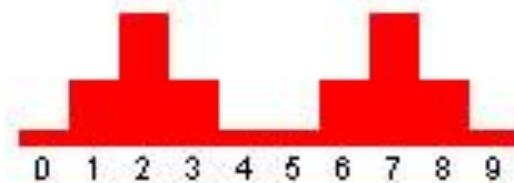
Non-symmetric, bimodal



Uniform



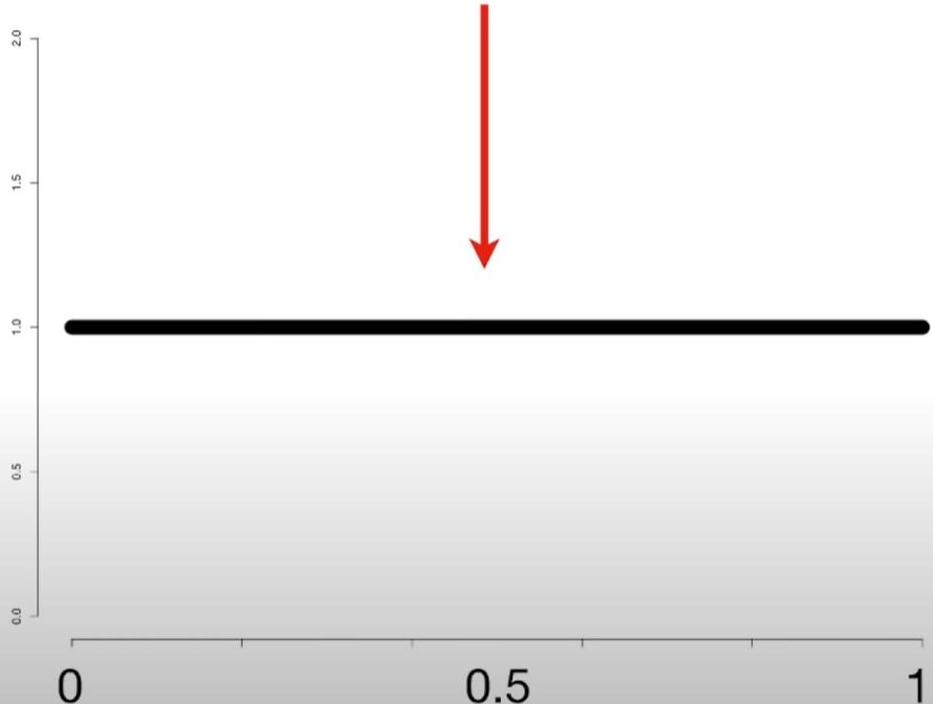
Skewed left



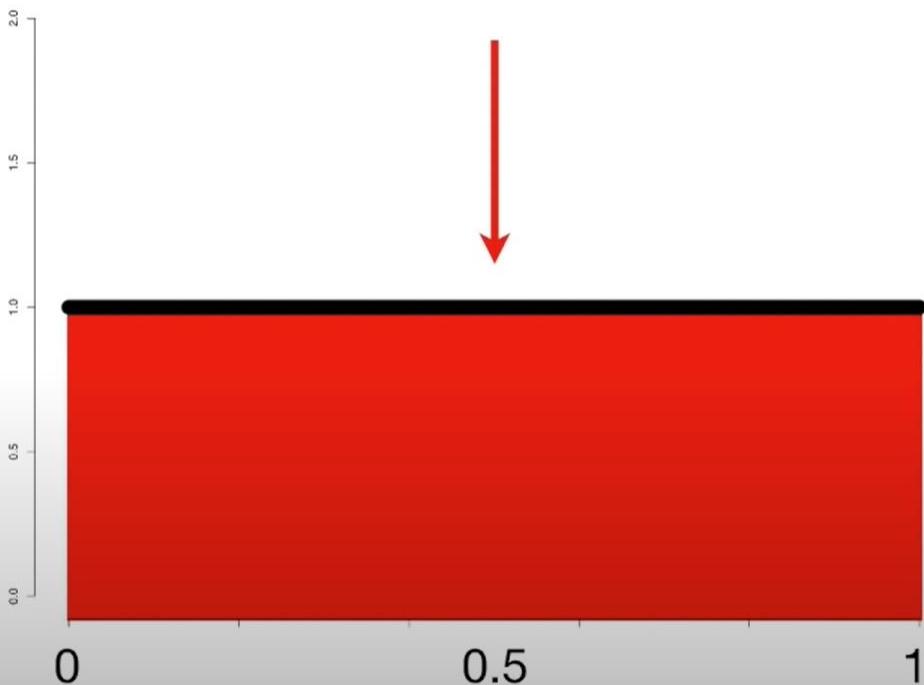
Symmetric, bimodal

Among all distributions, the normal distribution is the most common in nature.
Why?....The **CENTRAL LIMIT THEOREM!**

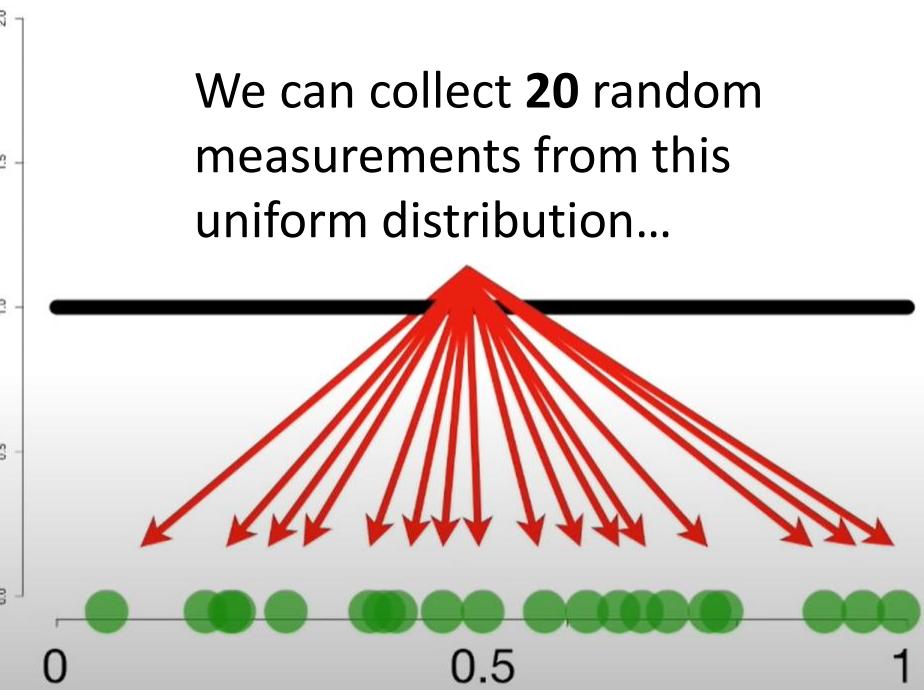
So let's start with a Uniform Distribution.



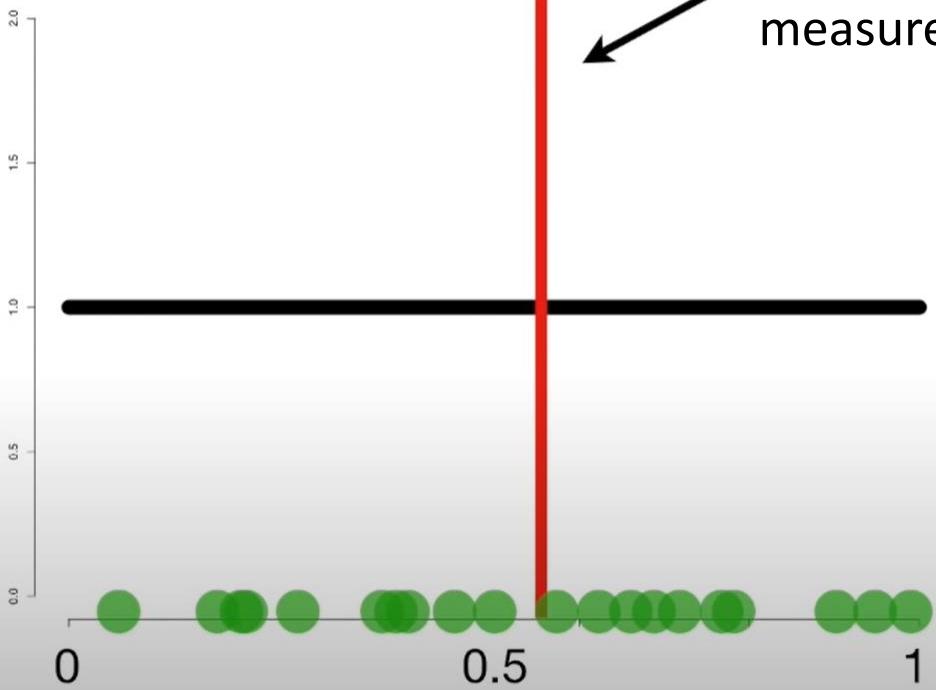
The probabilities are all equal,
and thus, are “uniform”.



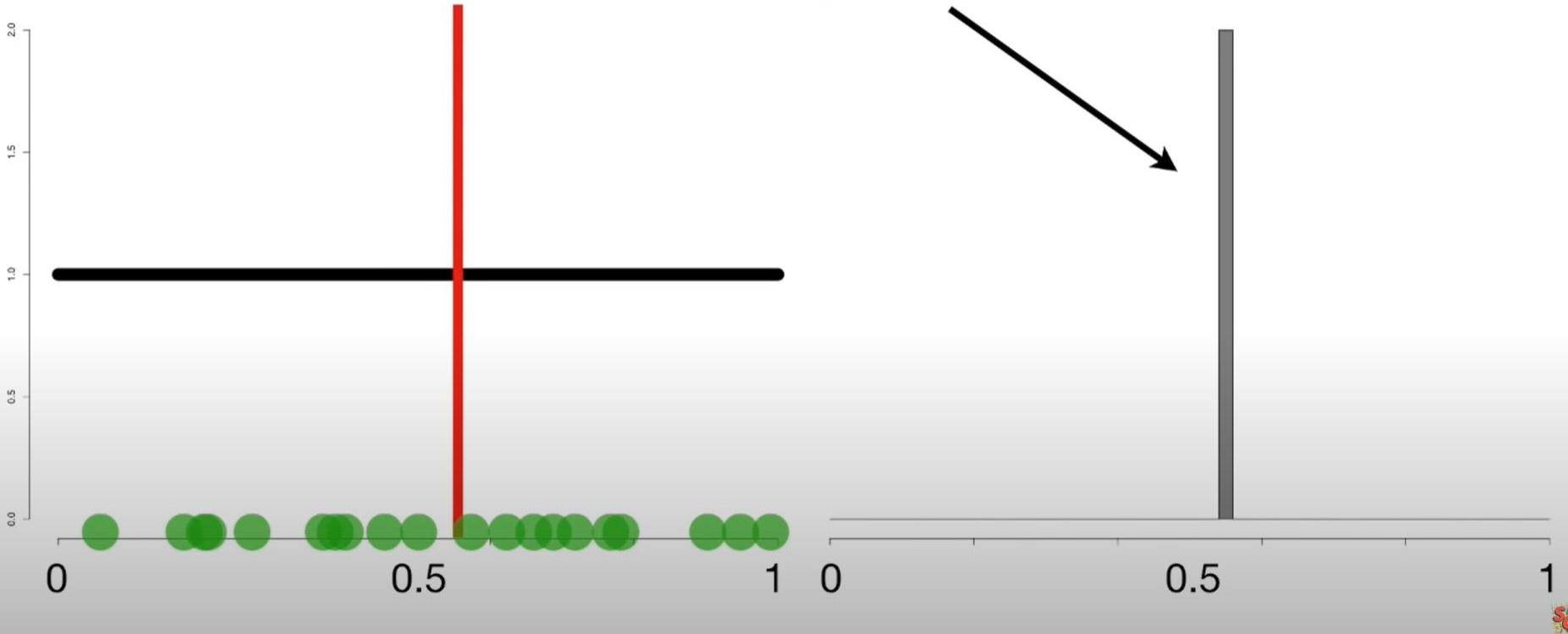
We can collect **20** random measurements from this uniform distribution...



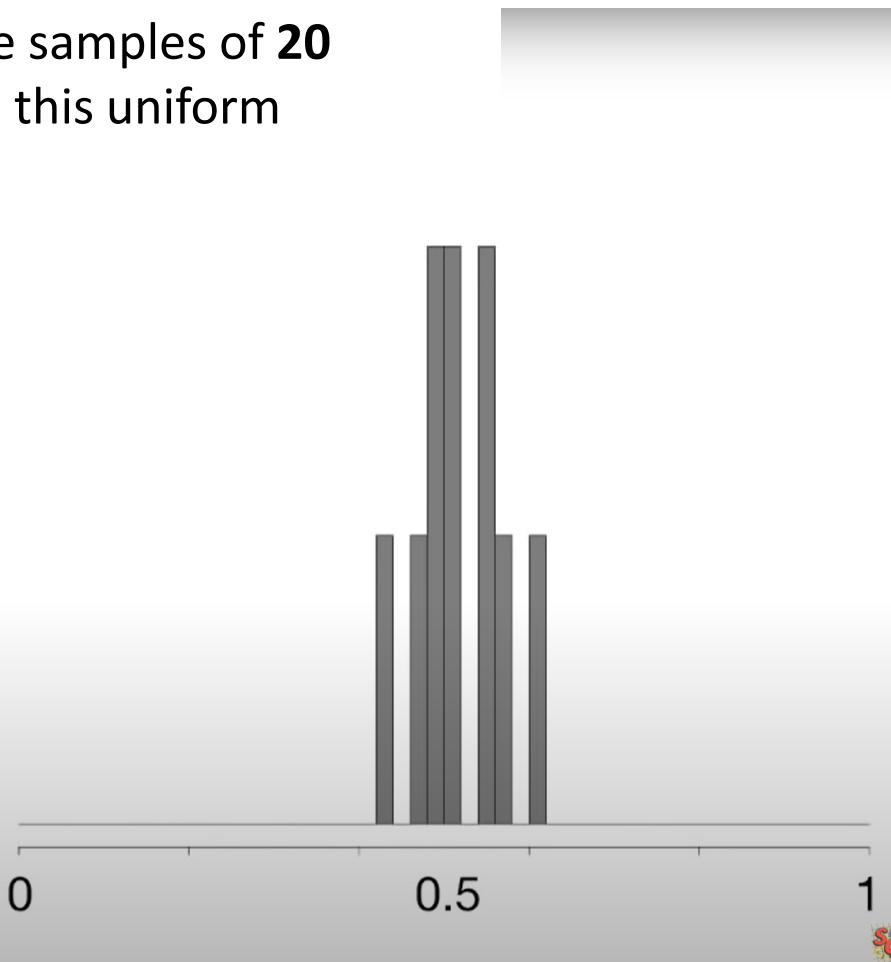
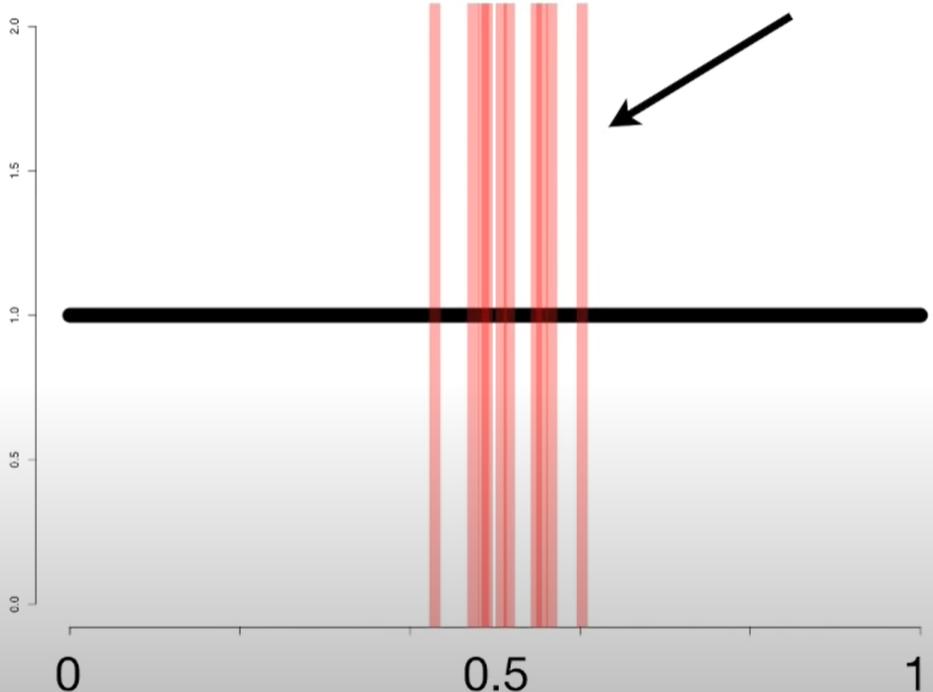
...and then calculate the
mean of the
measurements.



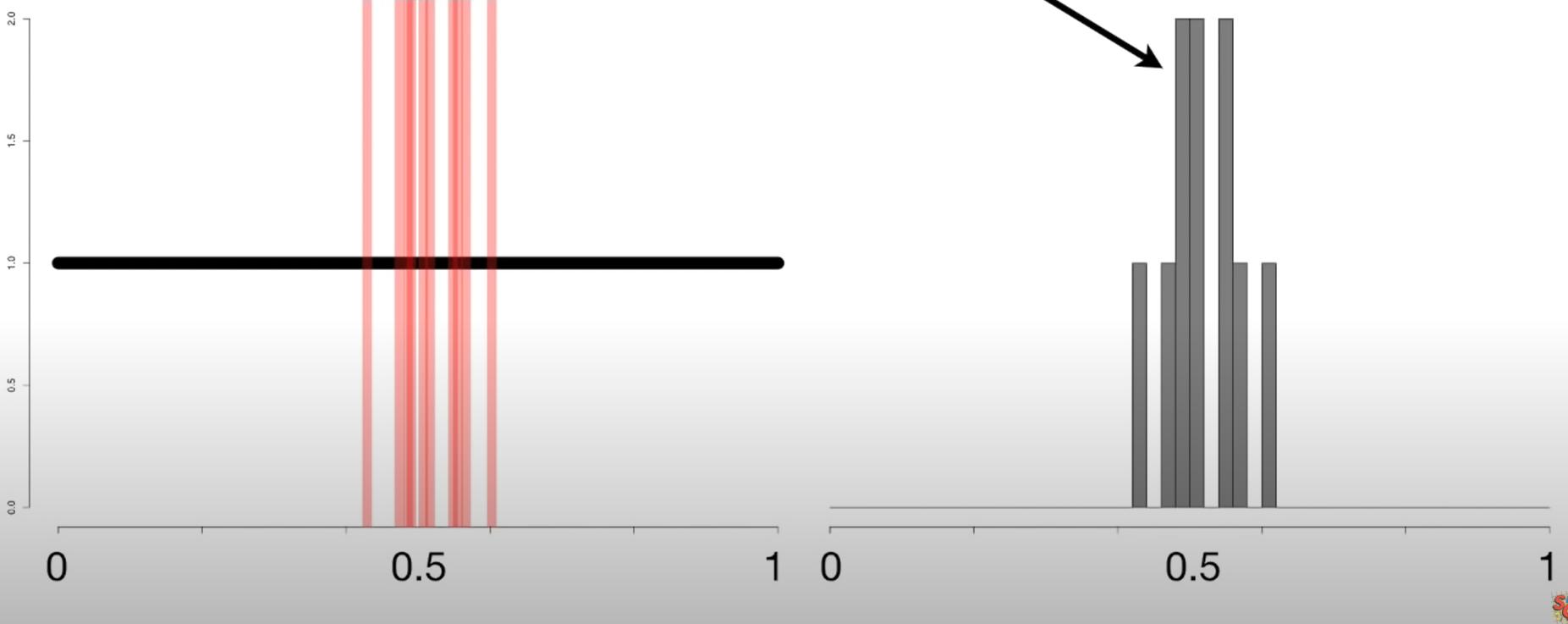
On the right, we can draw a histogram of the mean value.



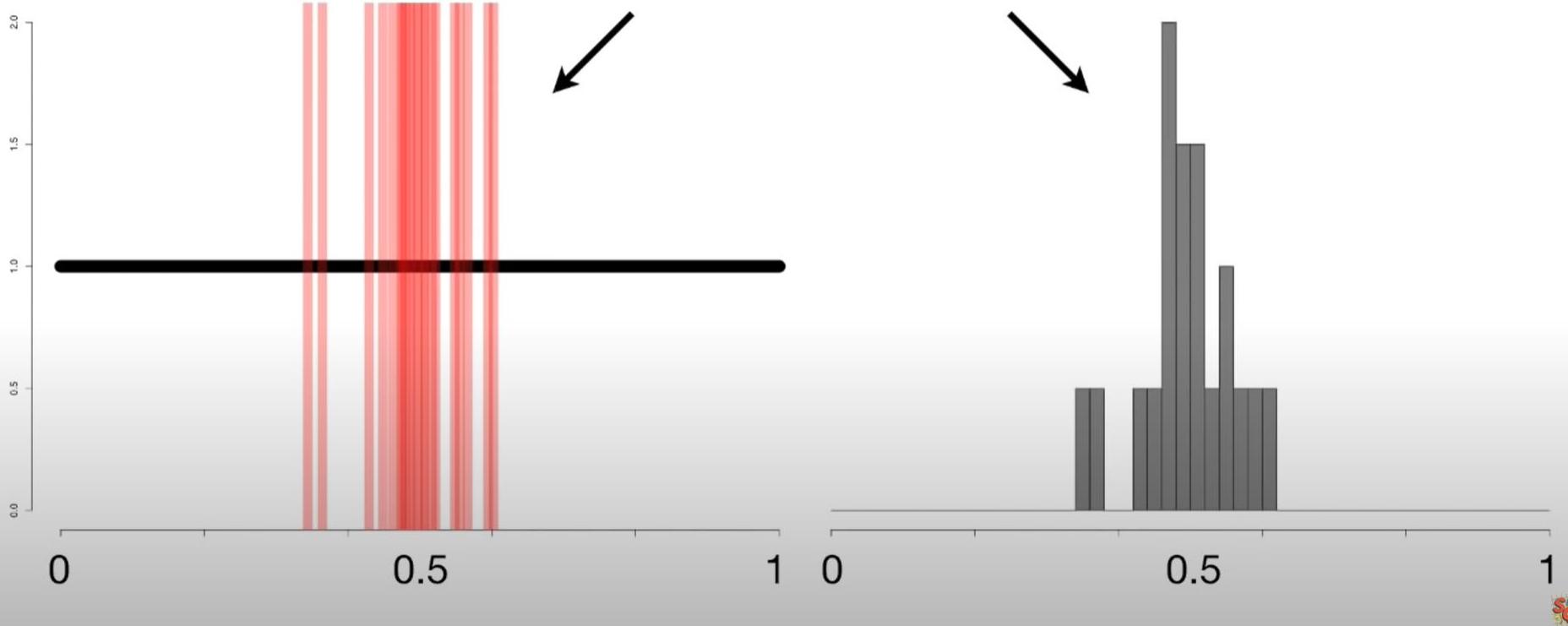
If we collect 10 more samples of **20** measurements from this uniform distribution...

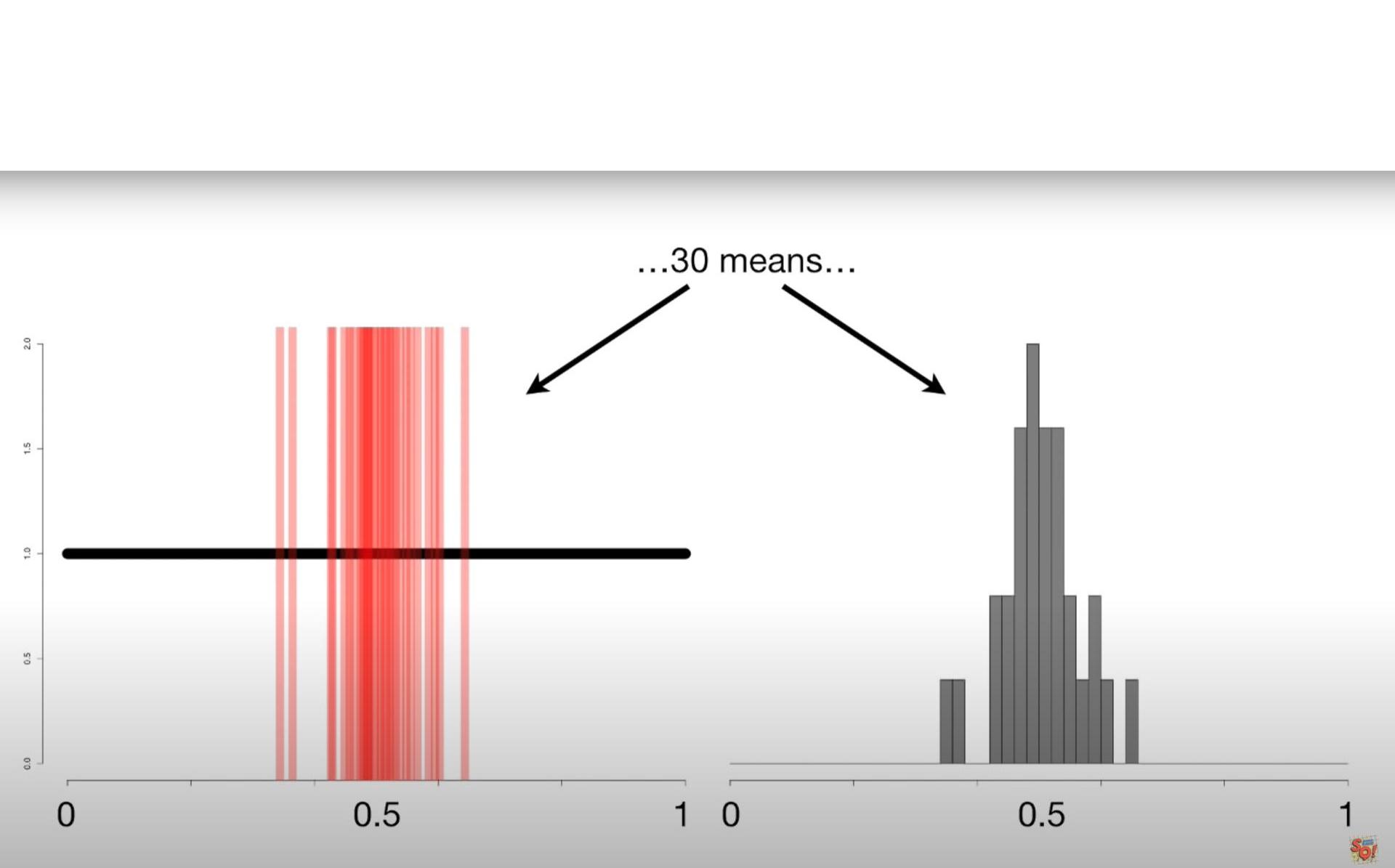


...the histogram starts to
look at little more
interesting.

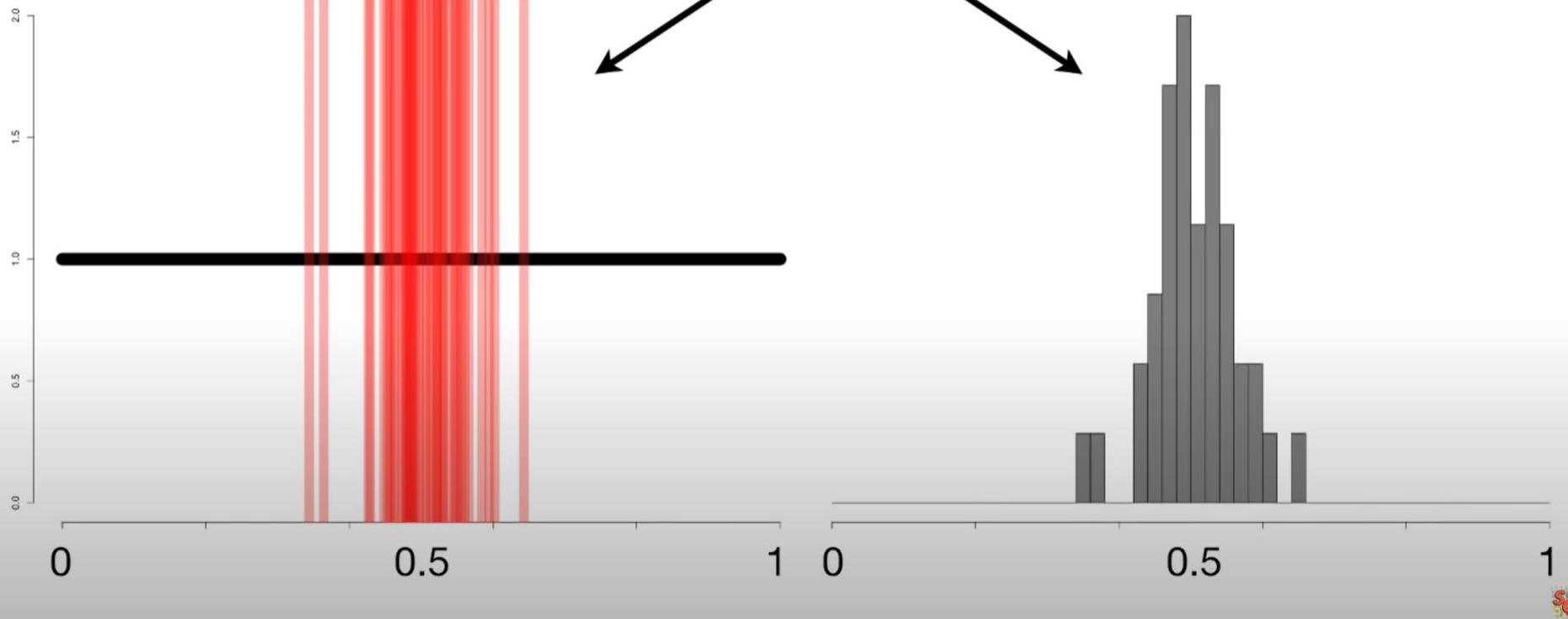


Here's the histogram after
collecting 20 samples and
calculating 20 means...

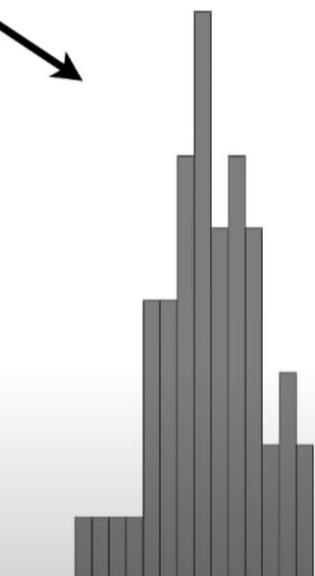
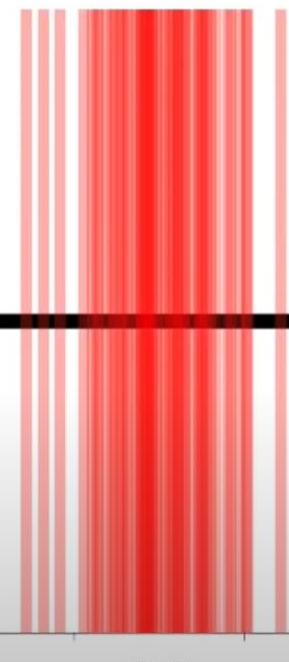


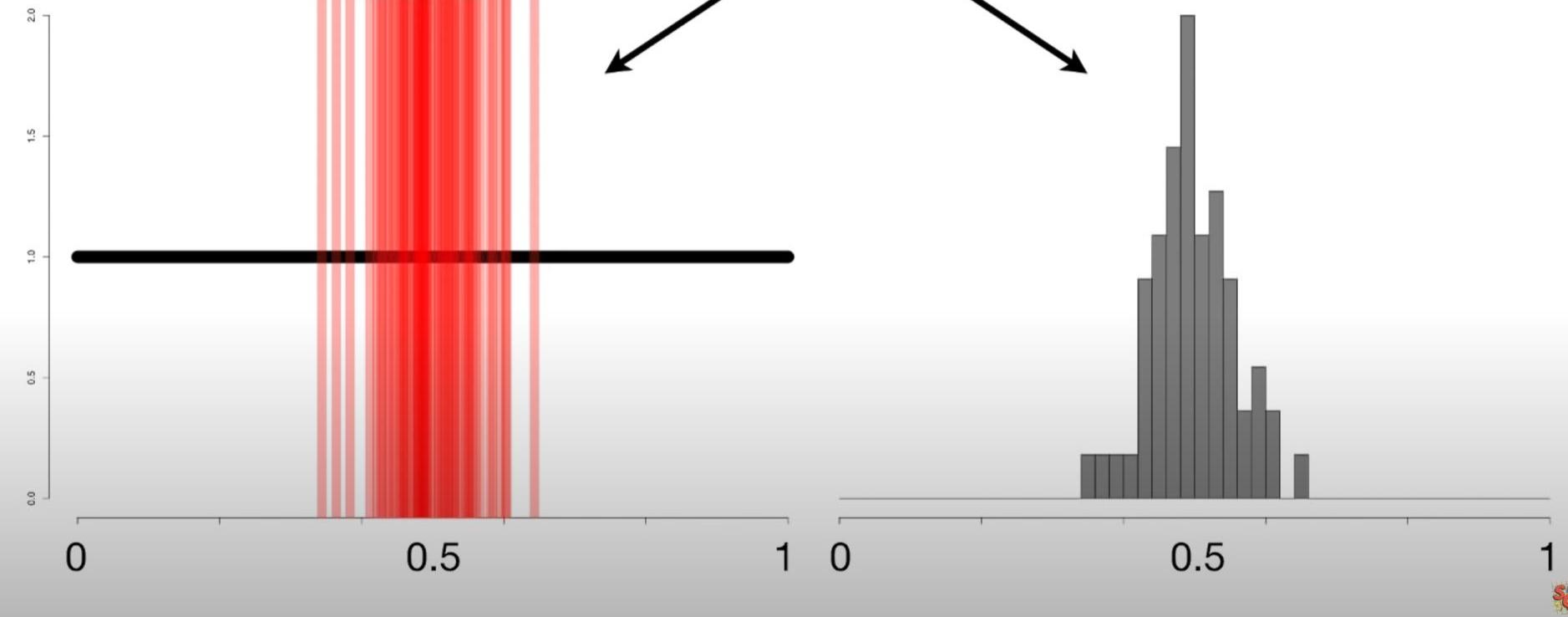


...40 means...

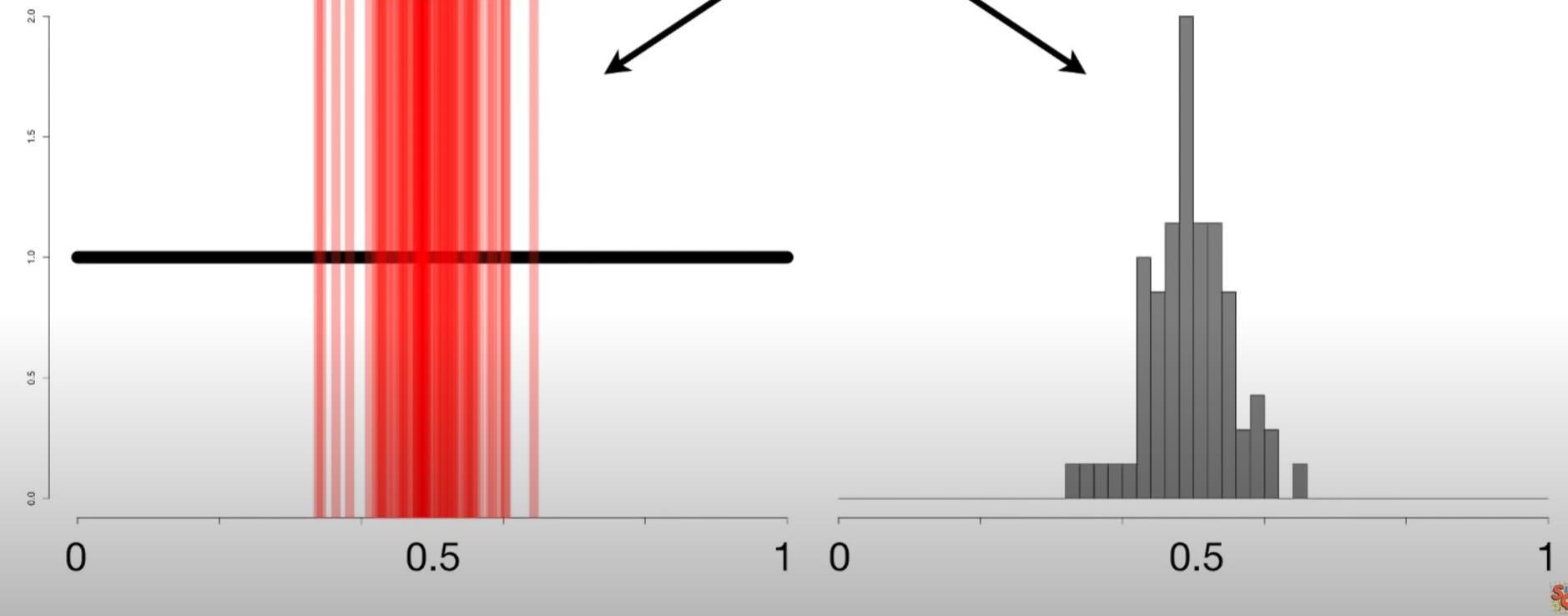


...50 means...

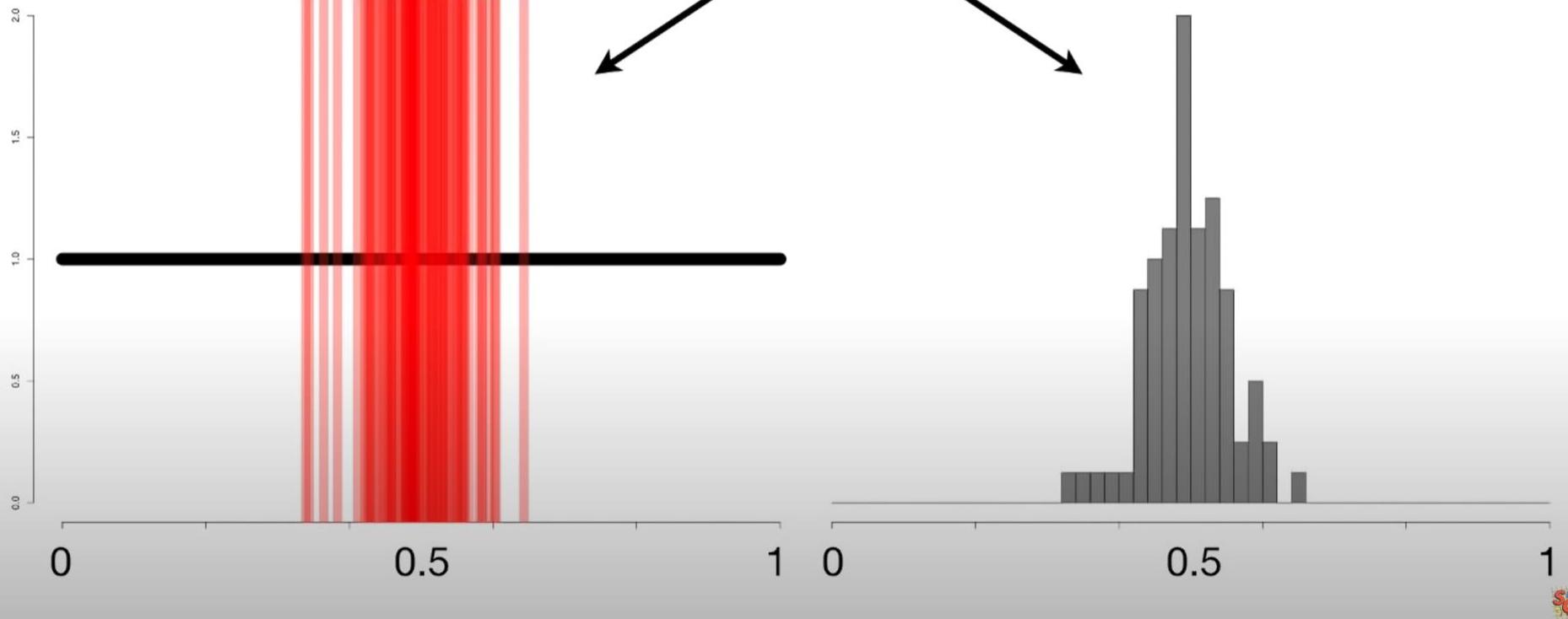


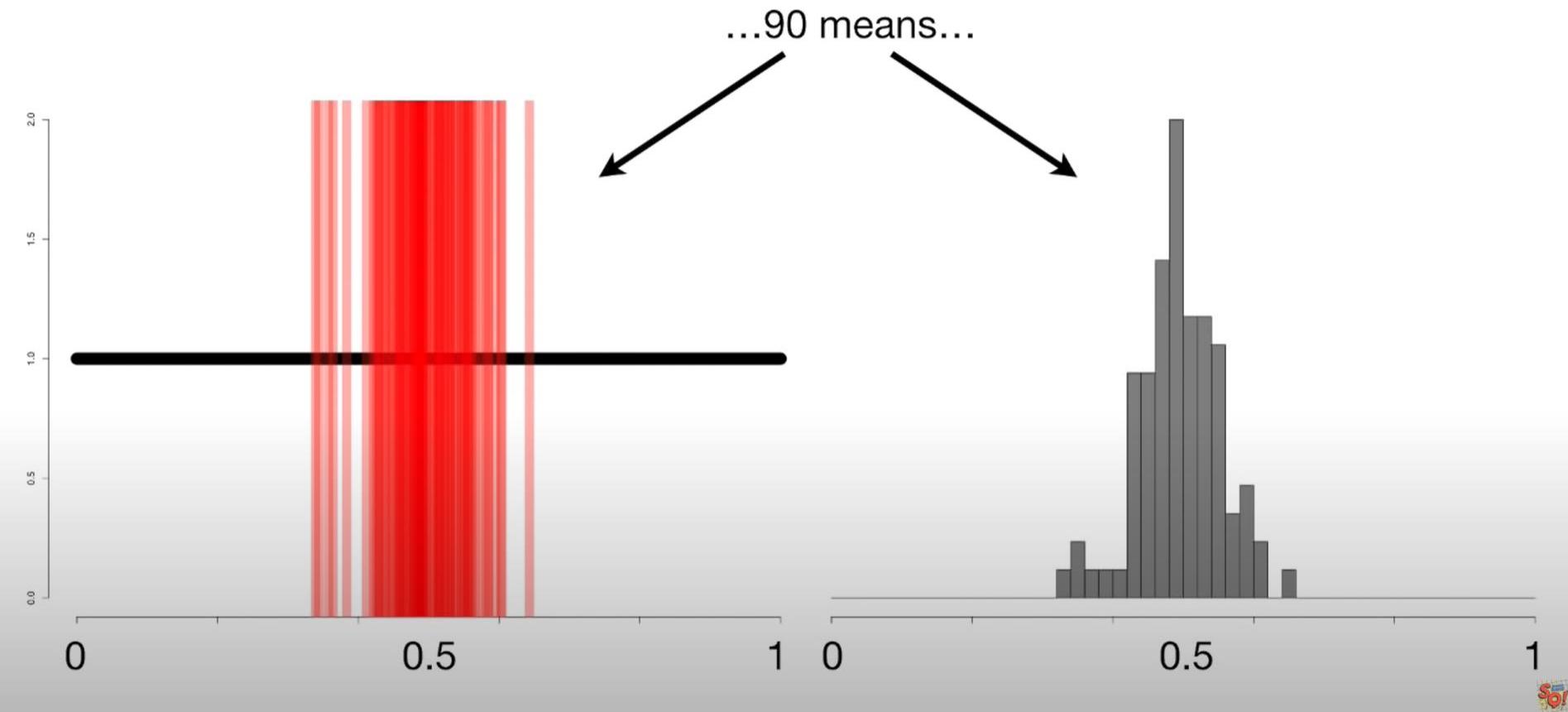


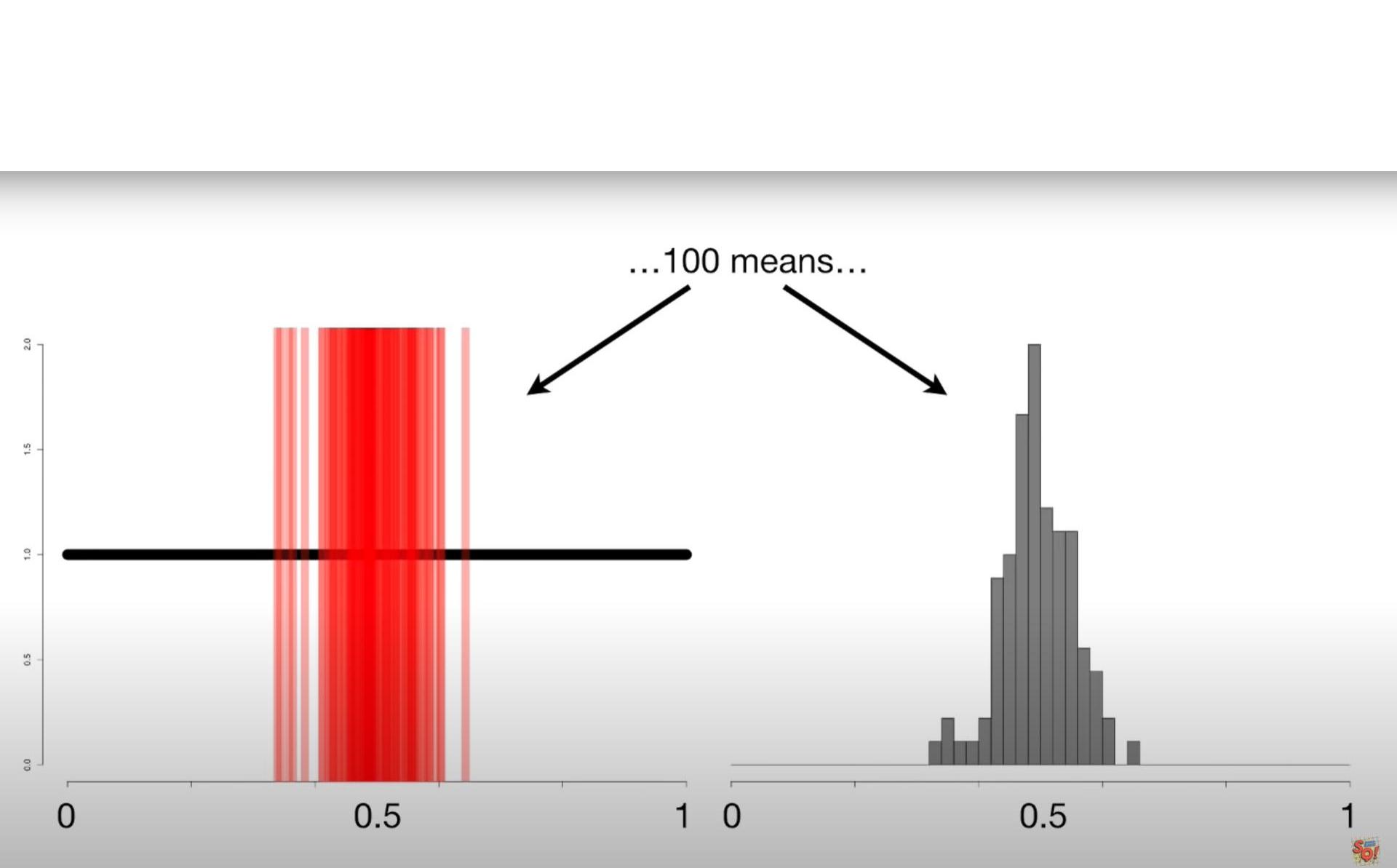
...70 means...



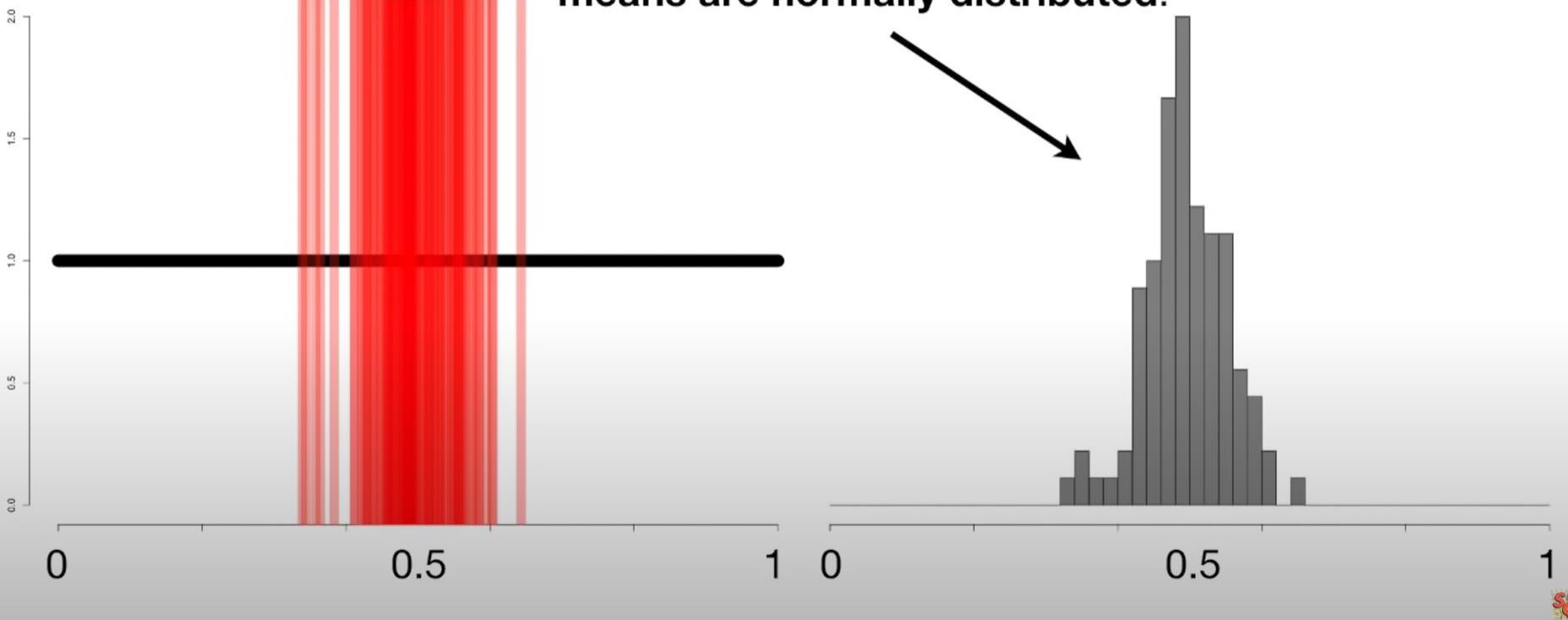
...80 means...



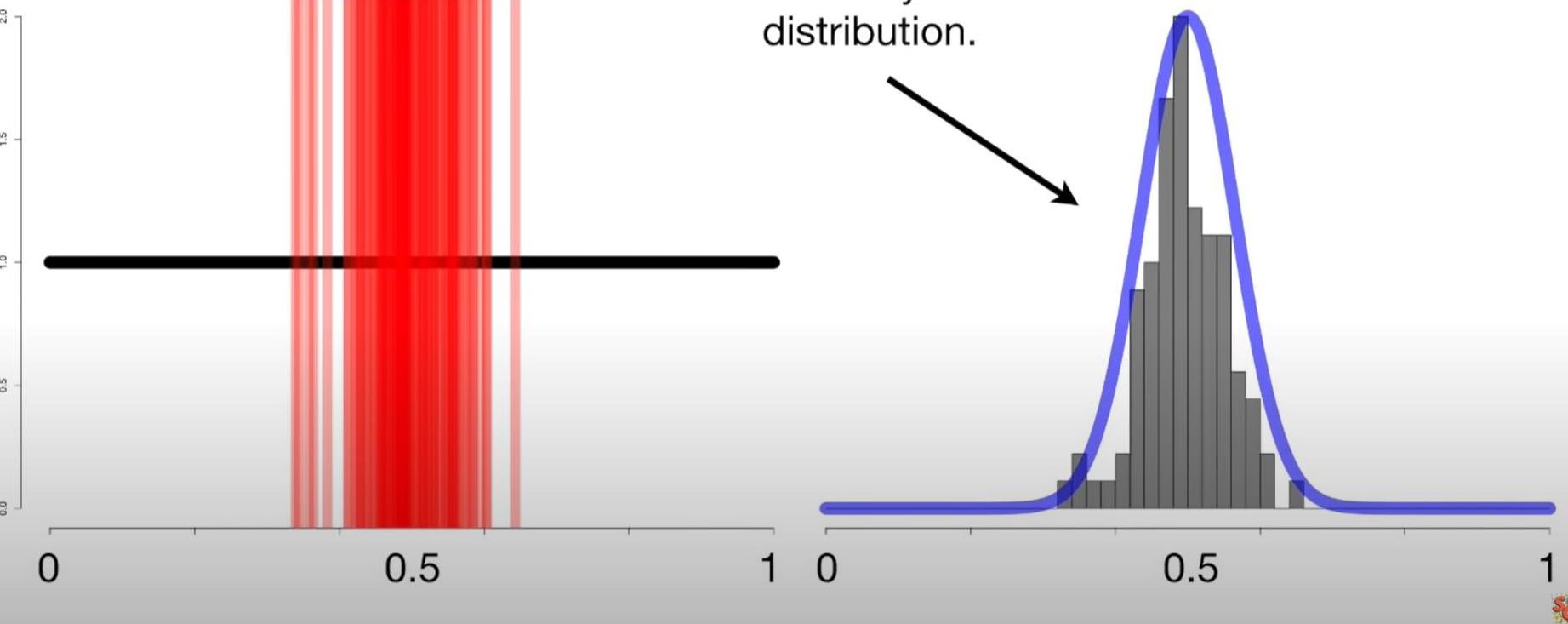




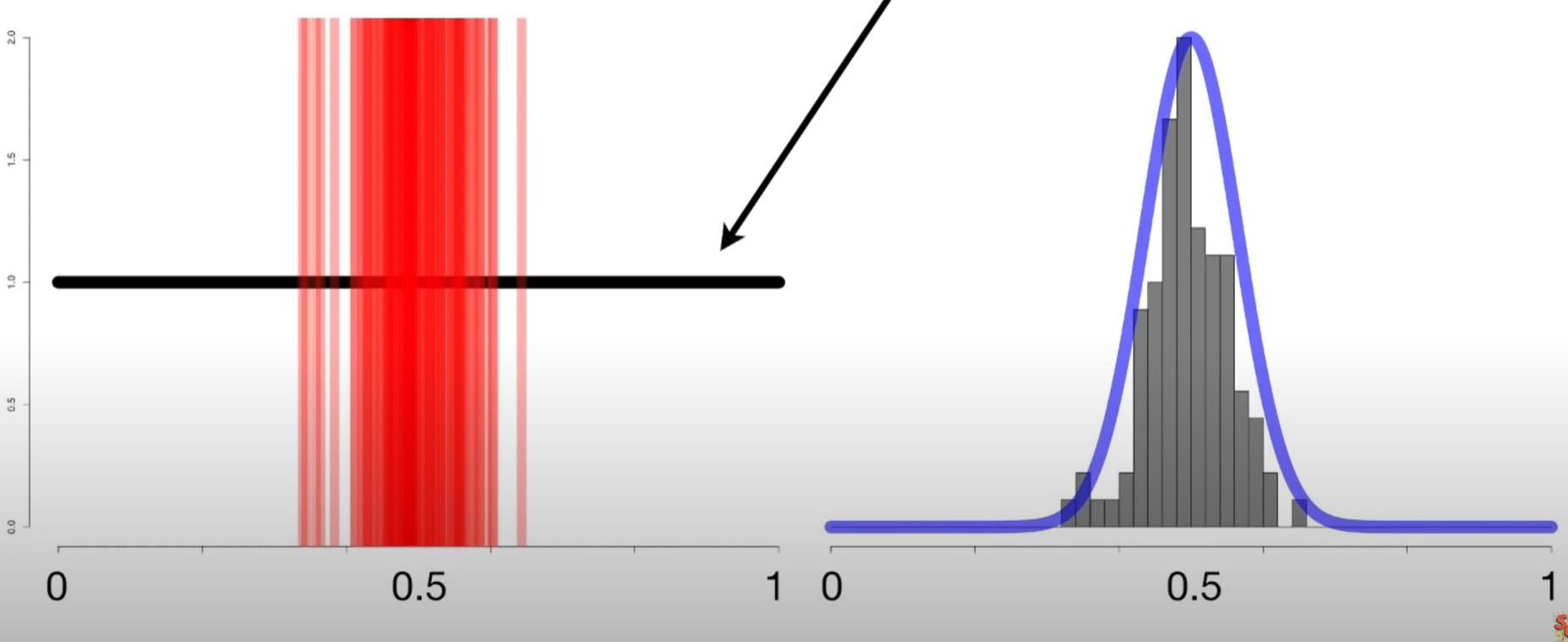
After adding 100 means to the histogram, it's pretty easy to see that these **means are normally distributed.**



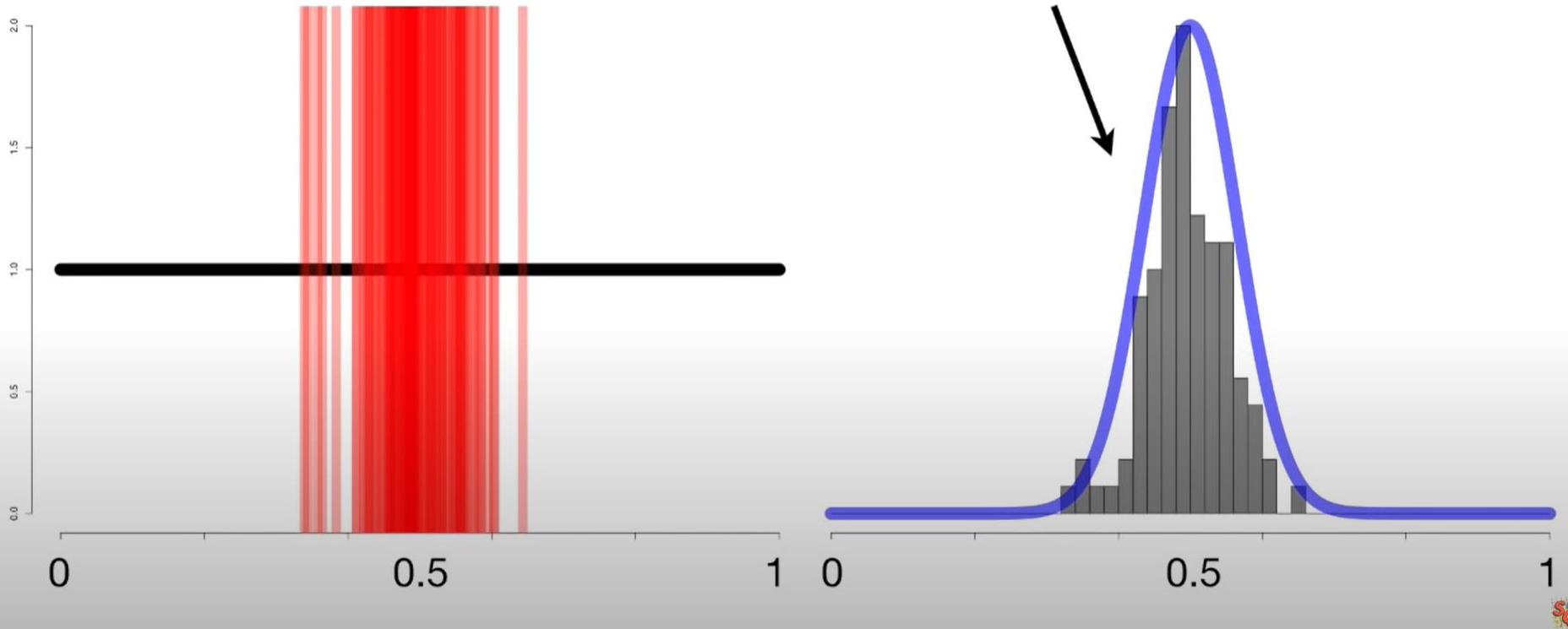
However, to make super easy to
see that the
means are normally distributed,
we can overlay a normal
distribution.



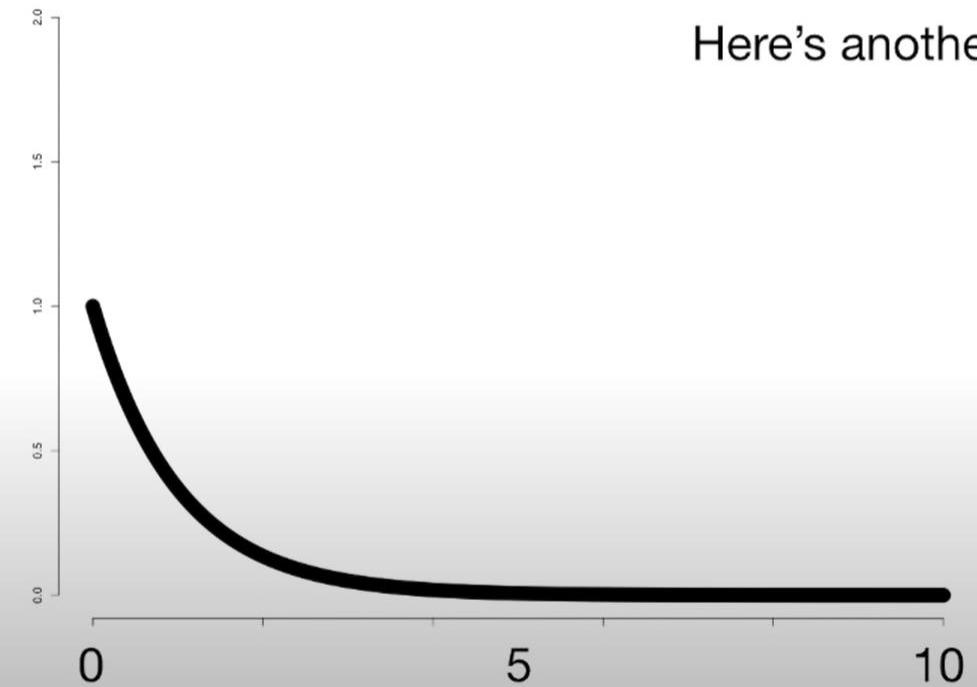
Even though these means
were calculated using data
from a uniform distribution...



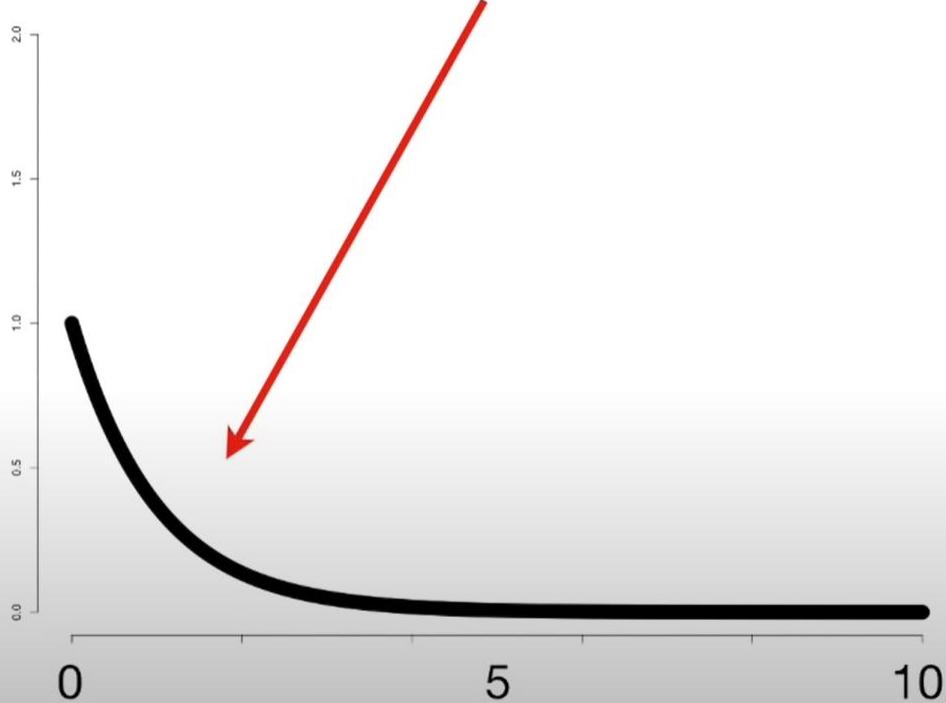
...the means themselves are not uniformly distributed. Instead, the **means are normally distributed.**



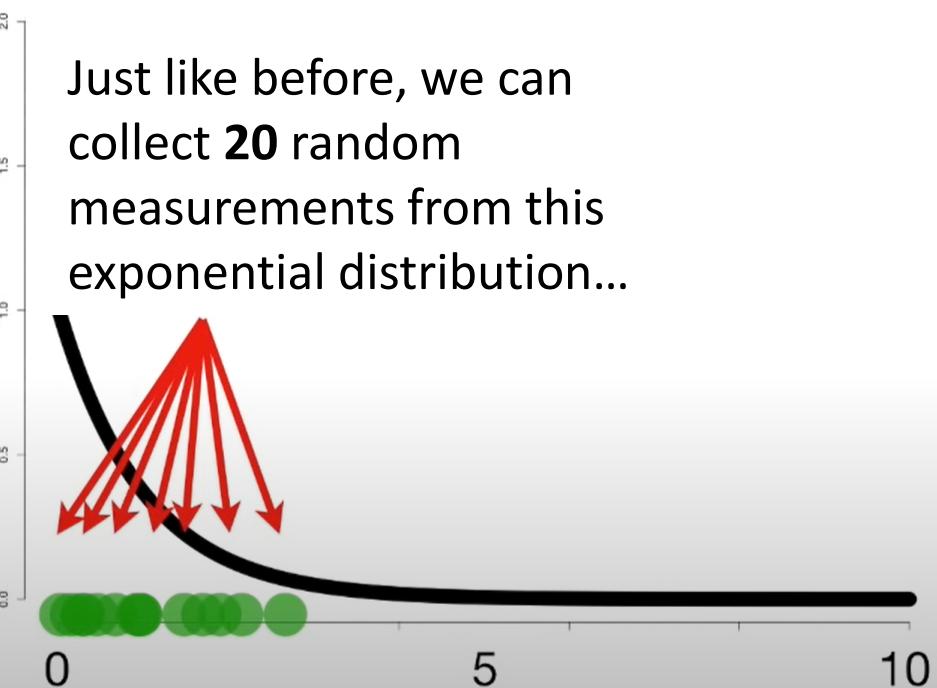
Here's another example...



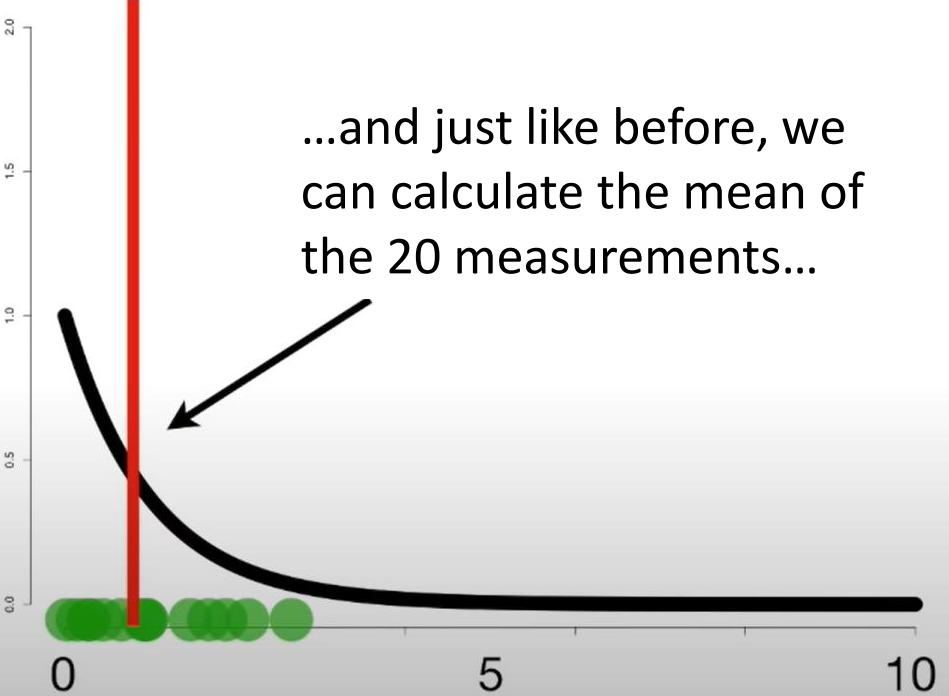
This time we'll start with an Exponential Distribution.



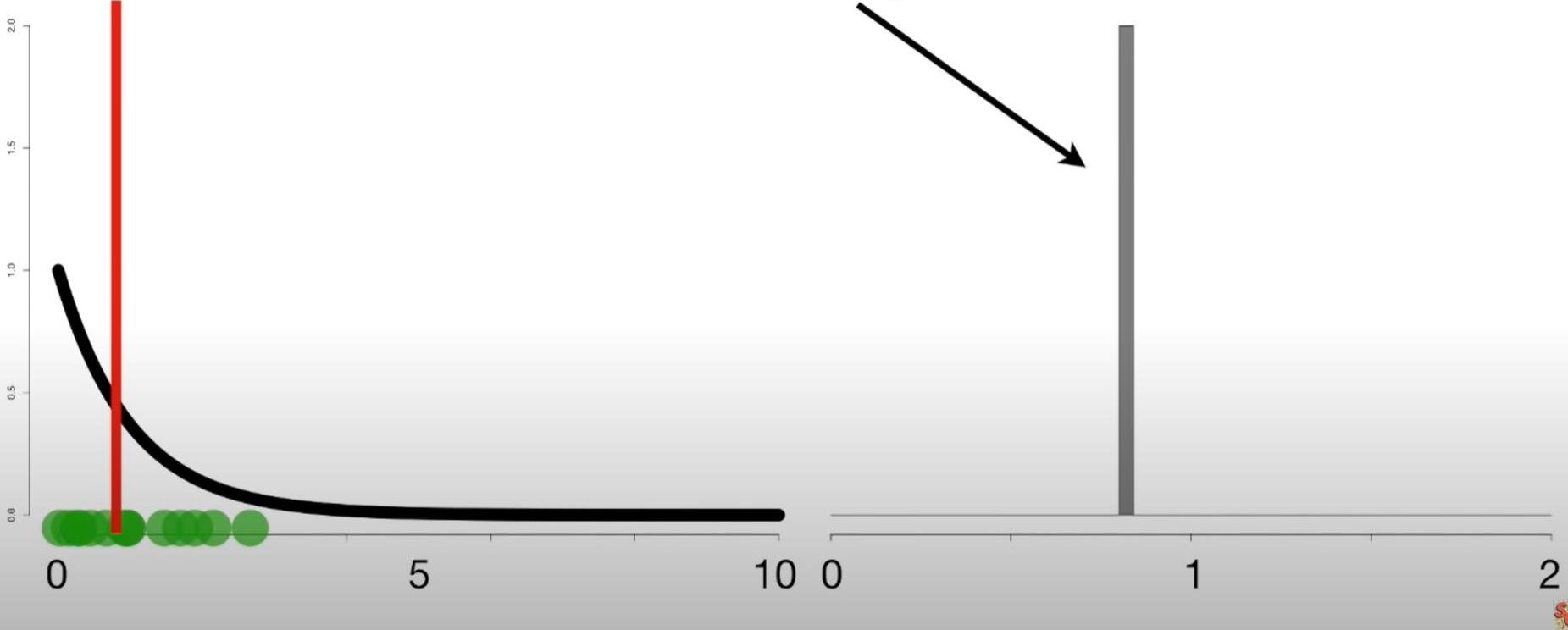
Just like before, we can collect **20** random measurements from this exponential distribution...

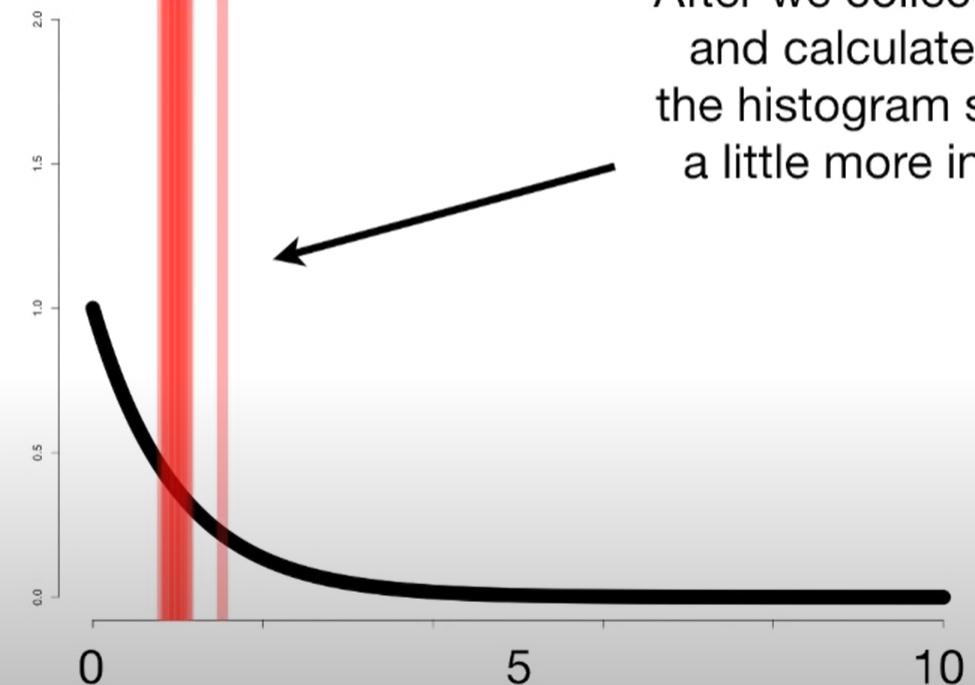


...and just like before, we
can calculate the mean of
the 20 measurements...

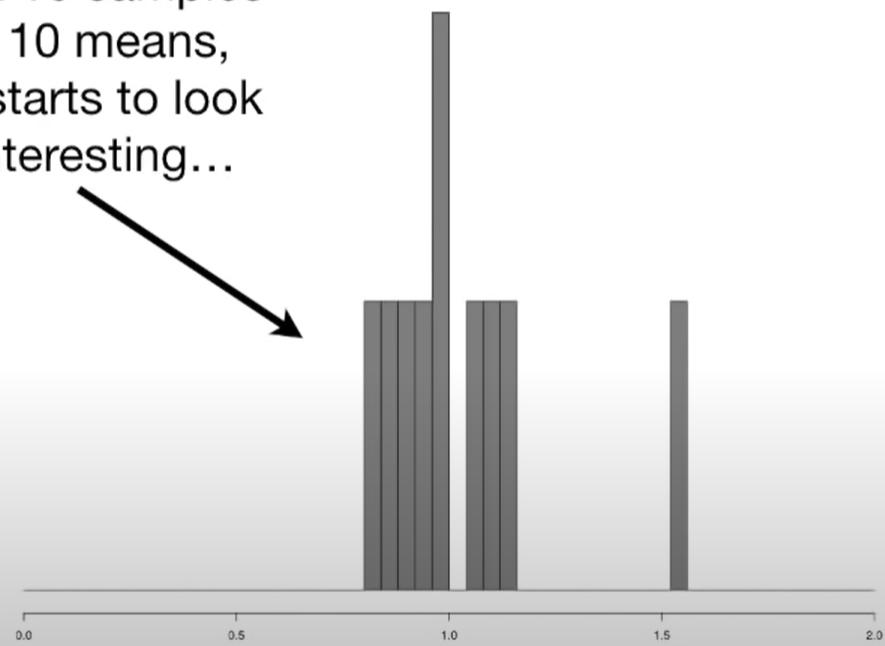


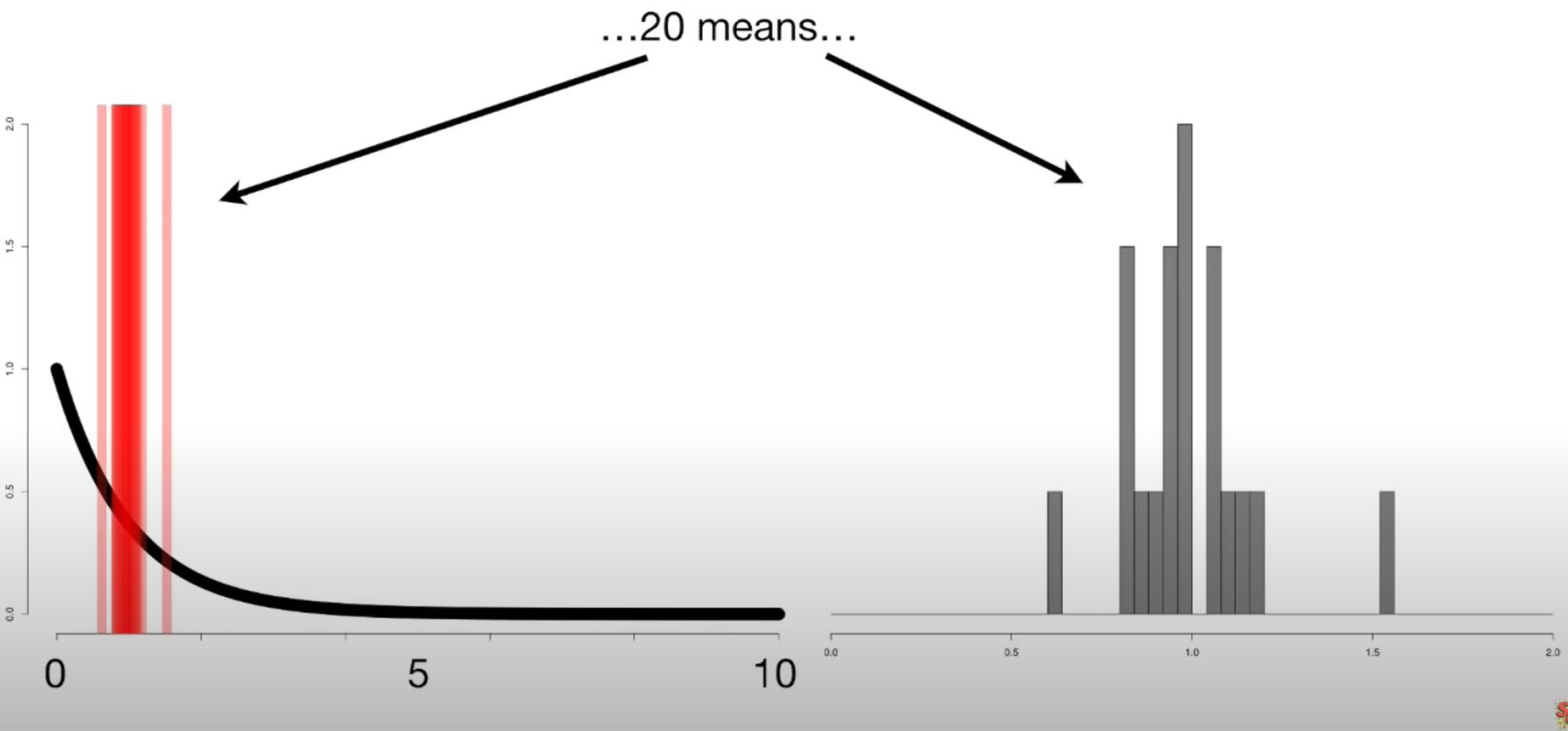
...and lastly, we can draw a histogram of that mean over here on the right.

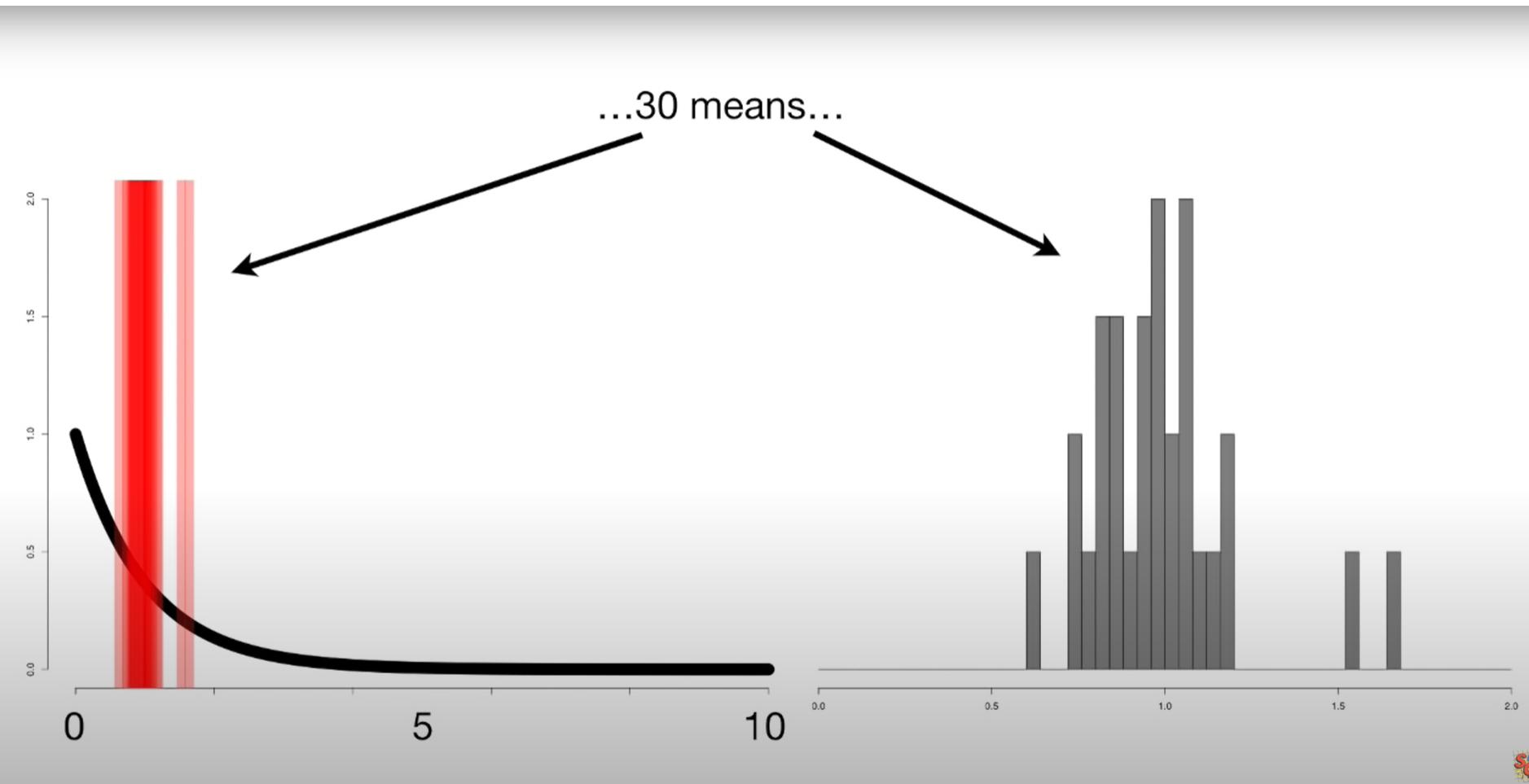




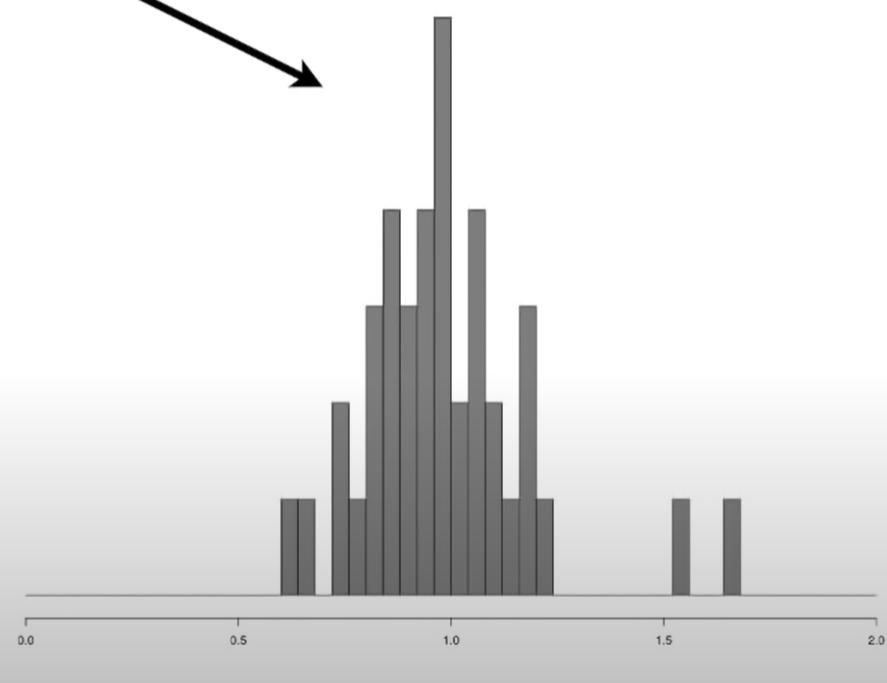
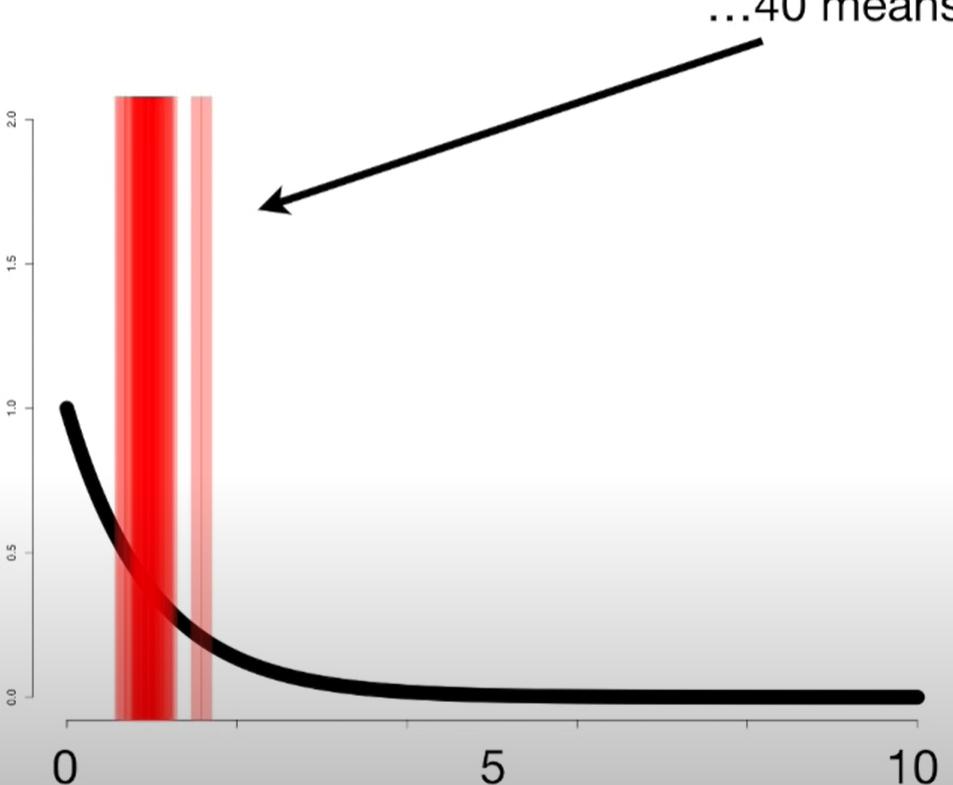
After we collect 10 samples
and calculate 10 means,
the histogram starts to look
a little more interesting...







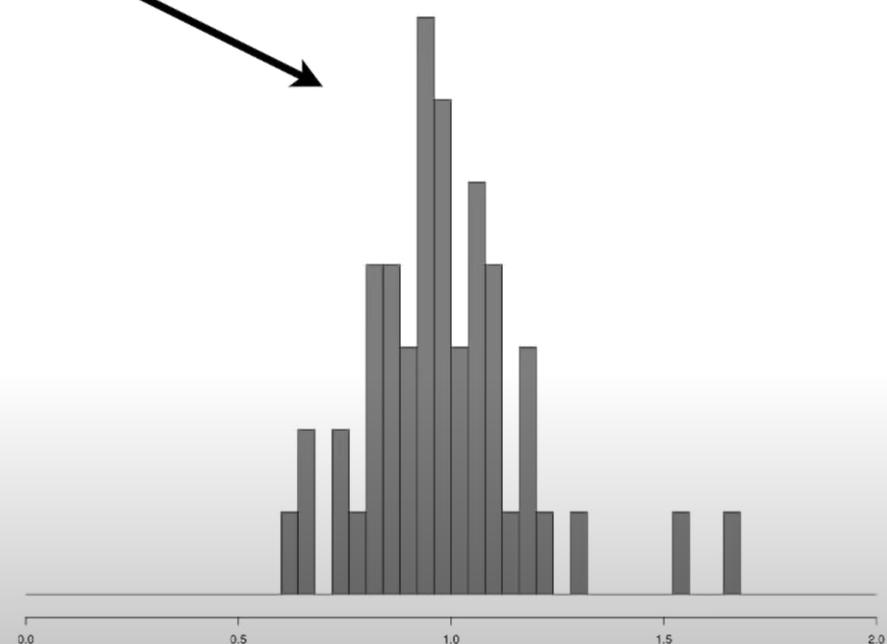
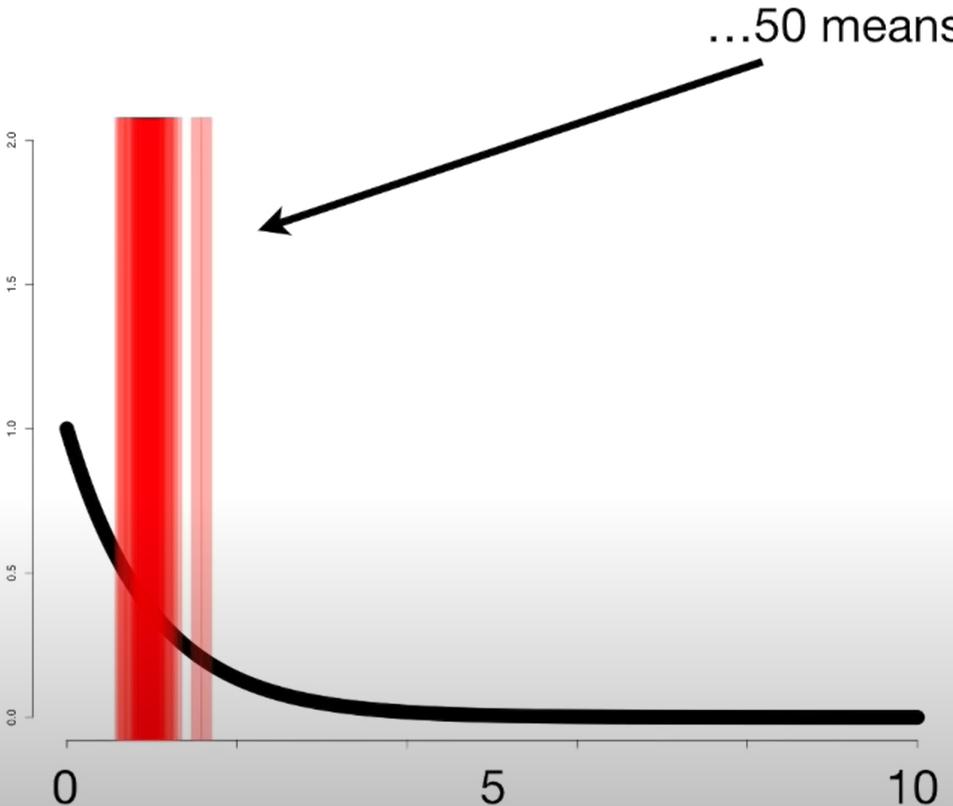
...40 means...



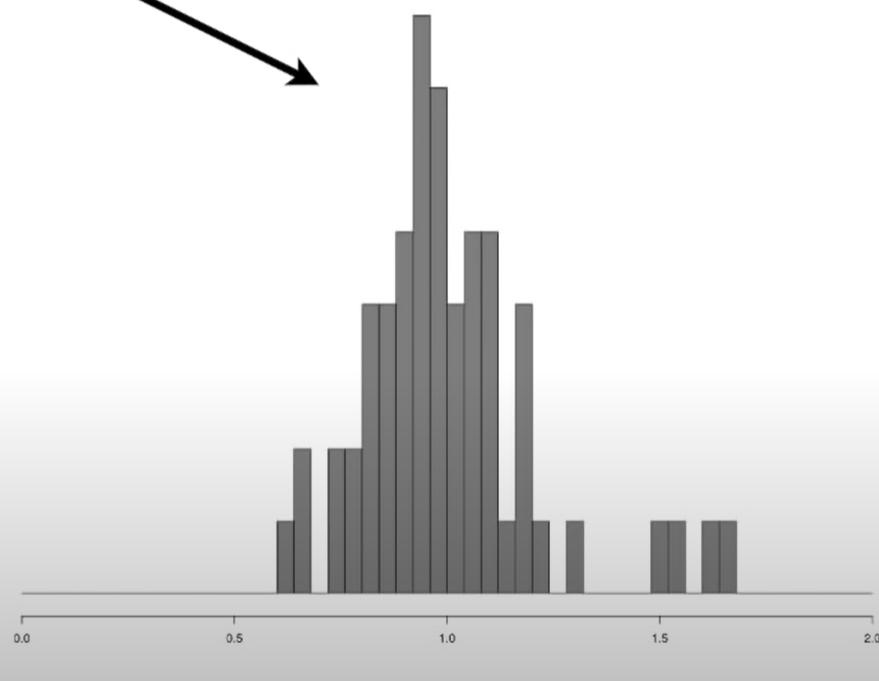
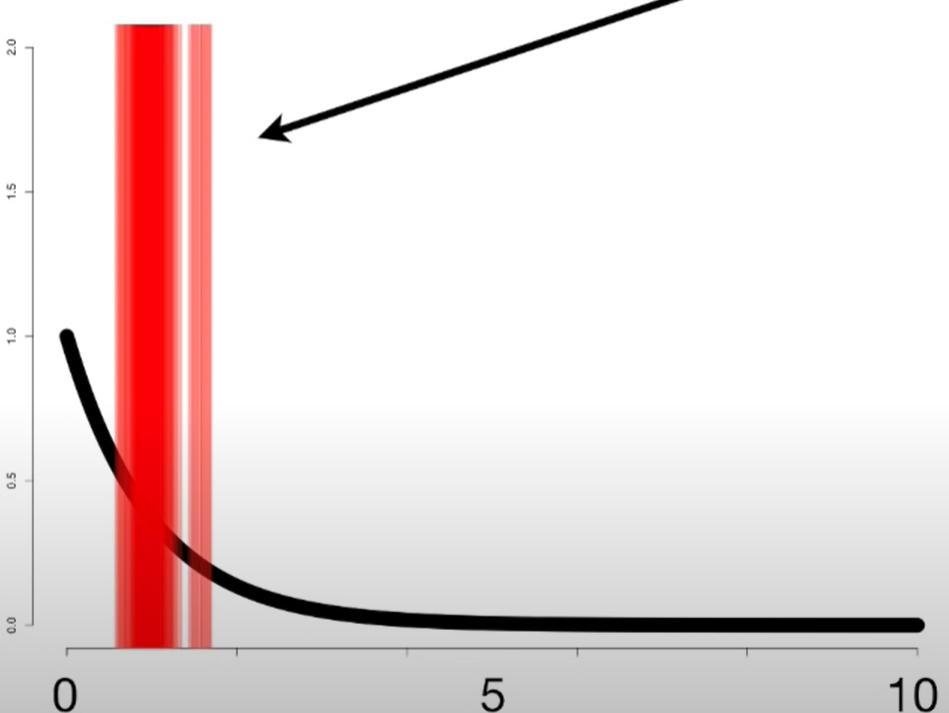
So!

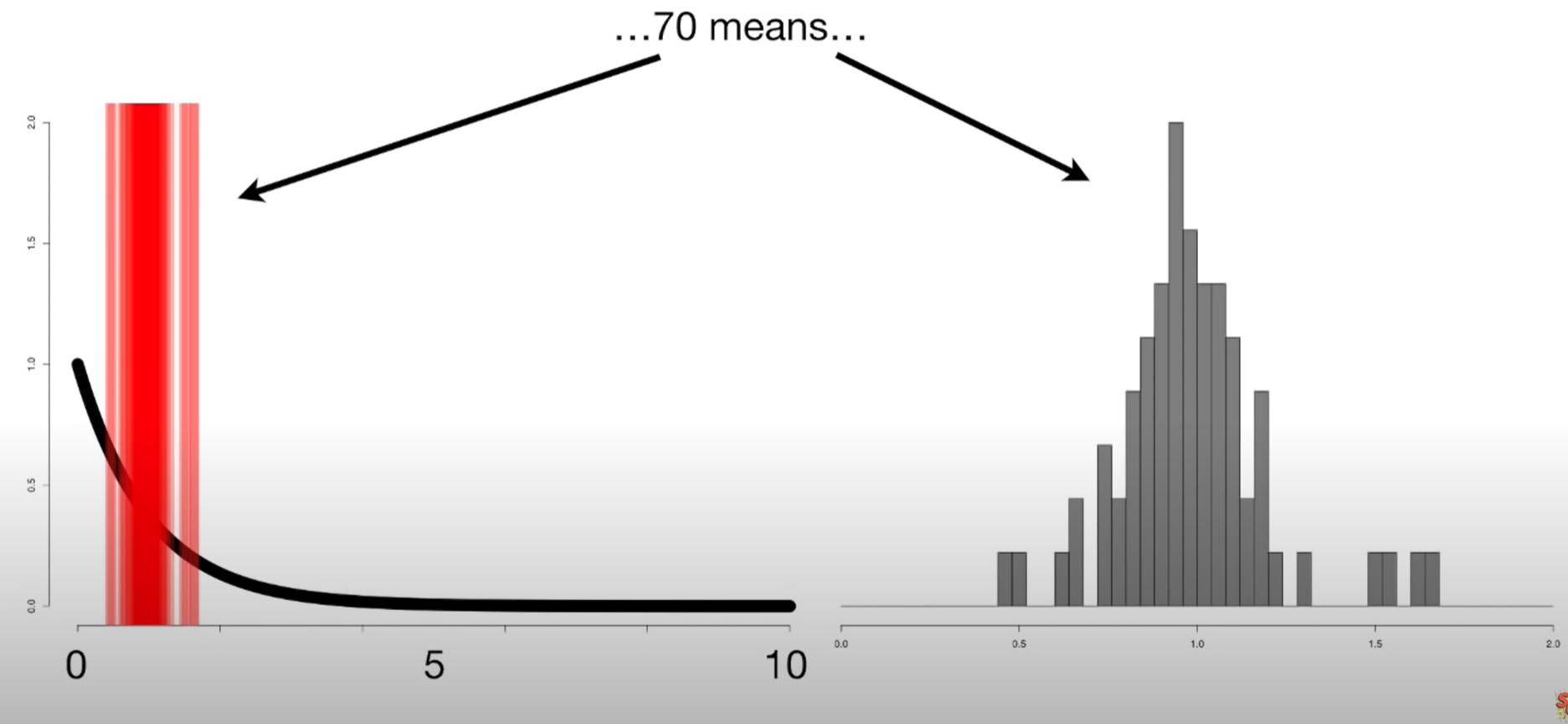


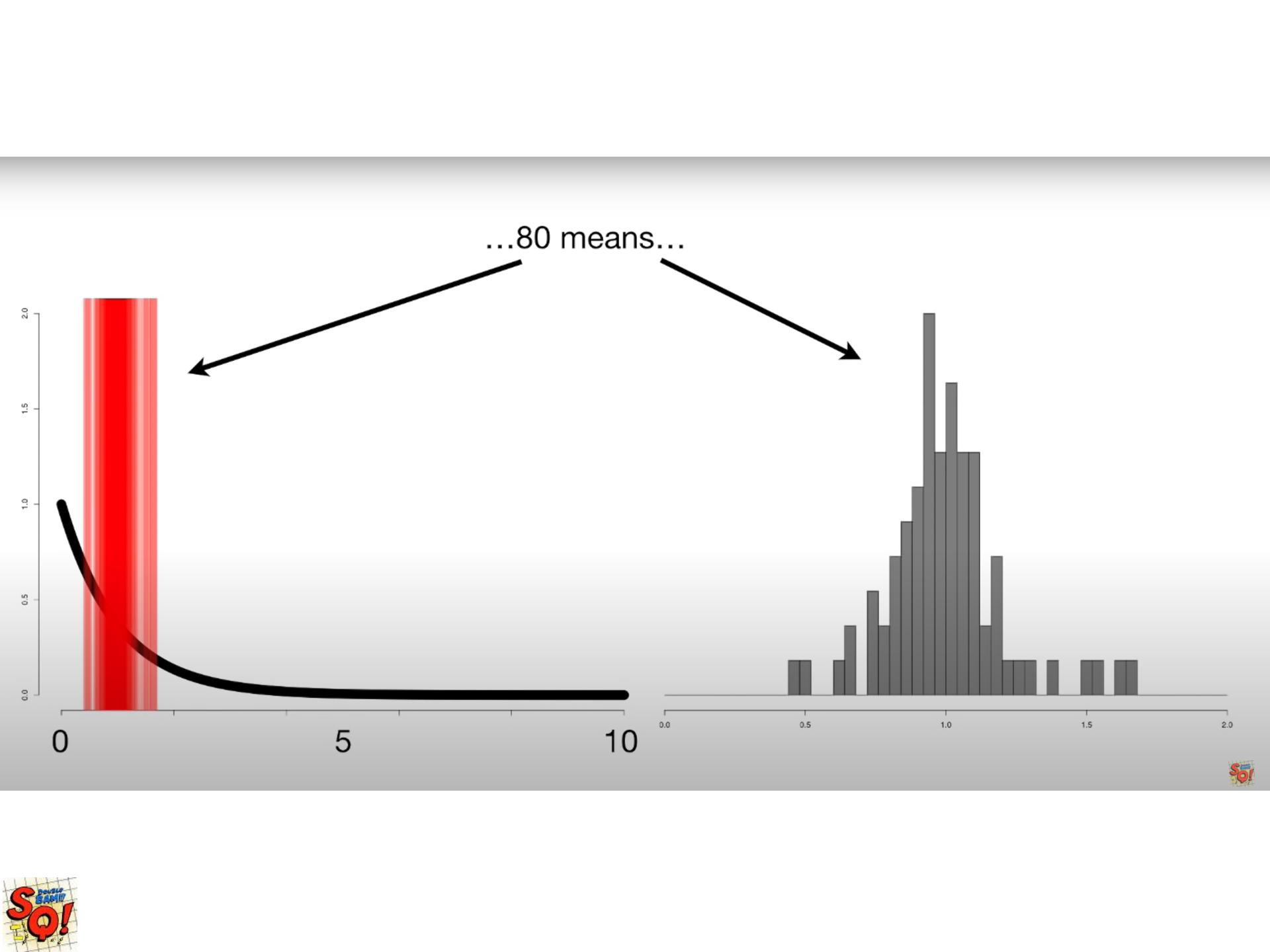
...50 means...



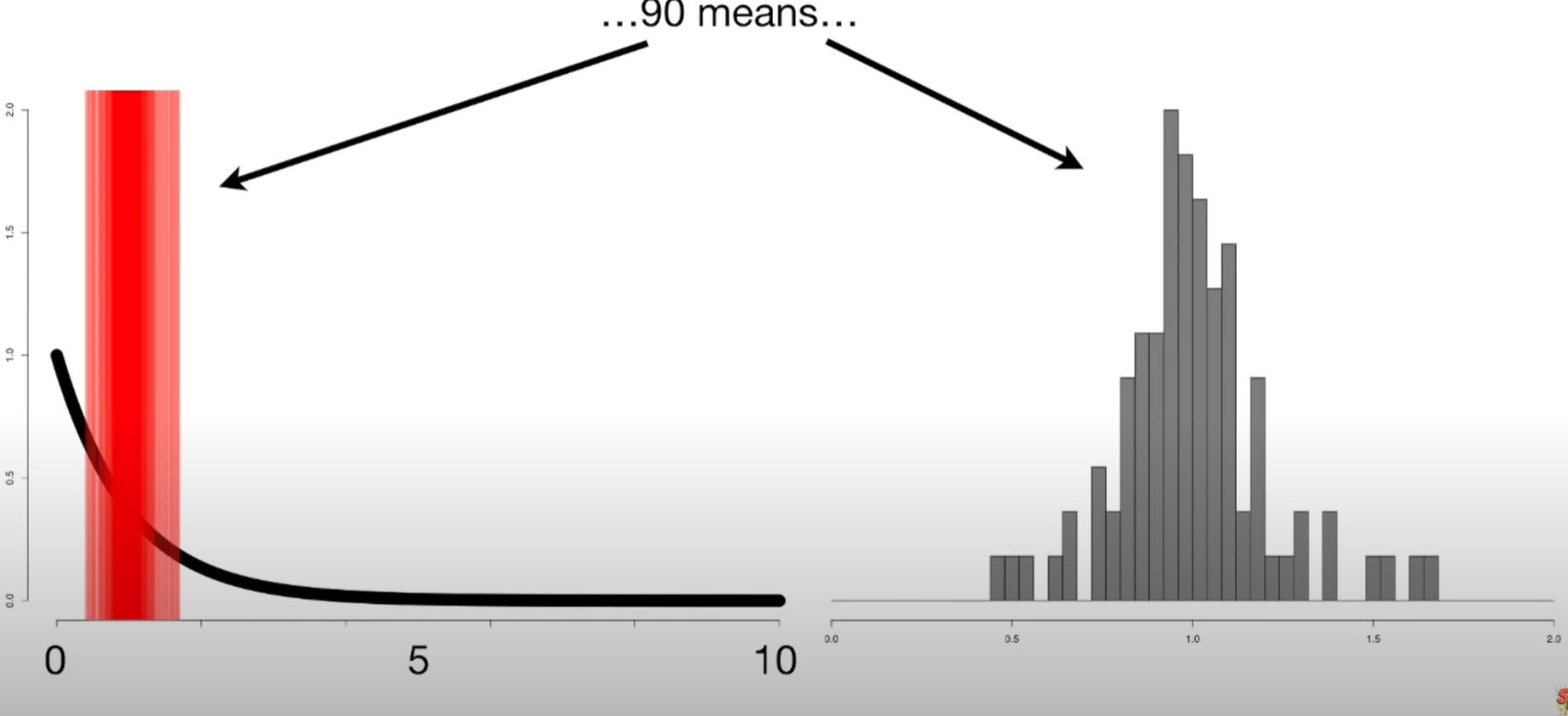
...60 means...

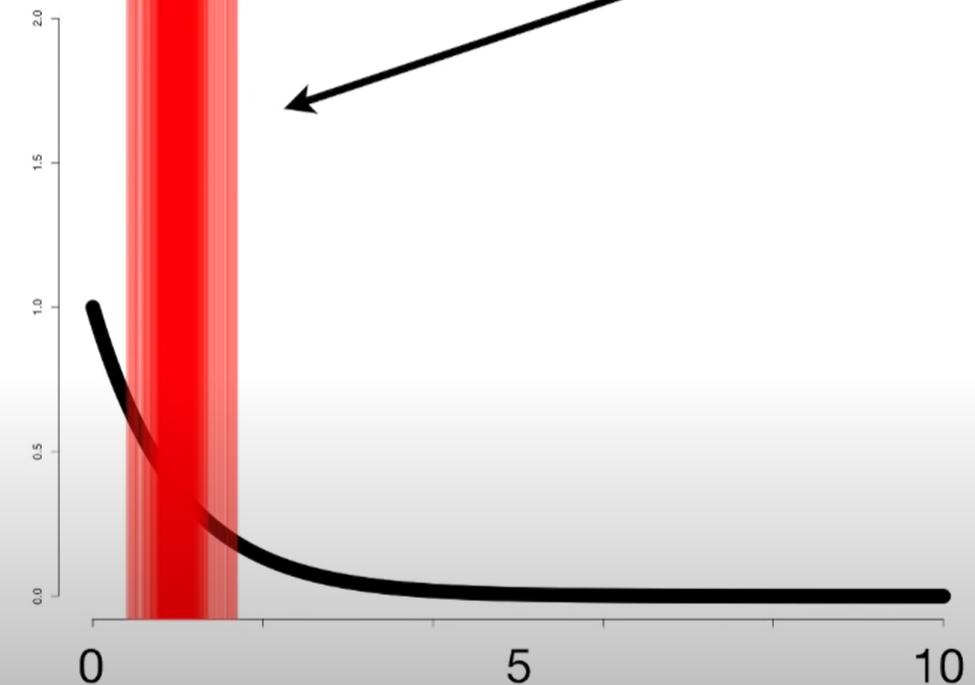




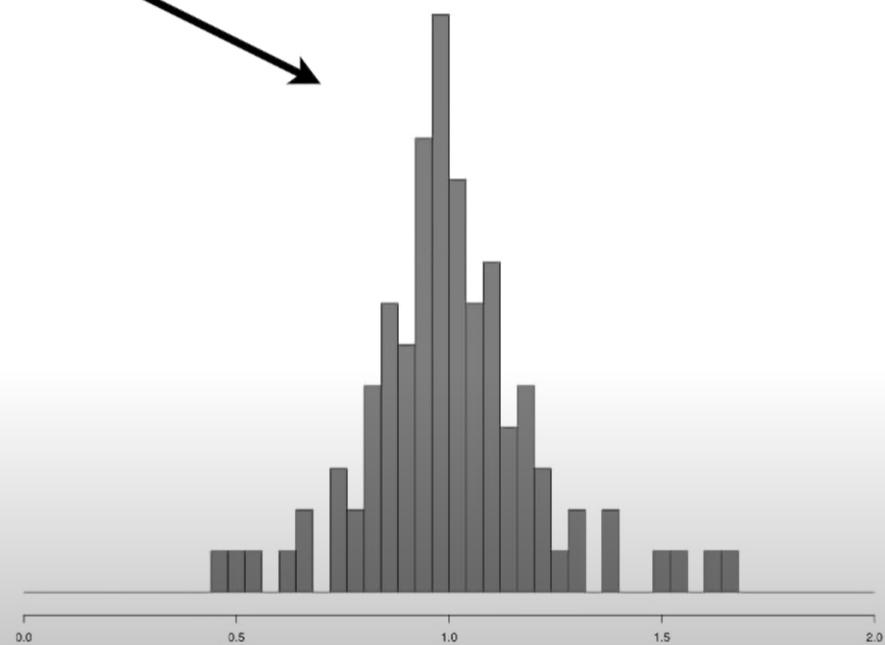


...90 means...





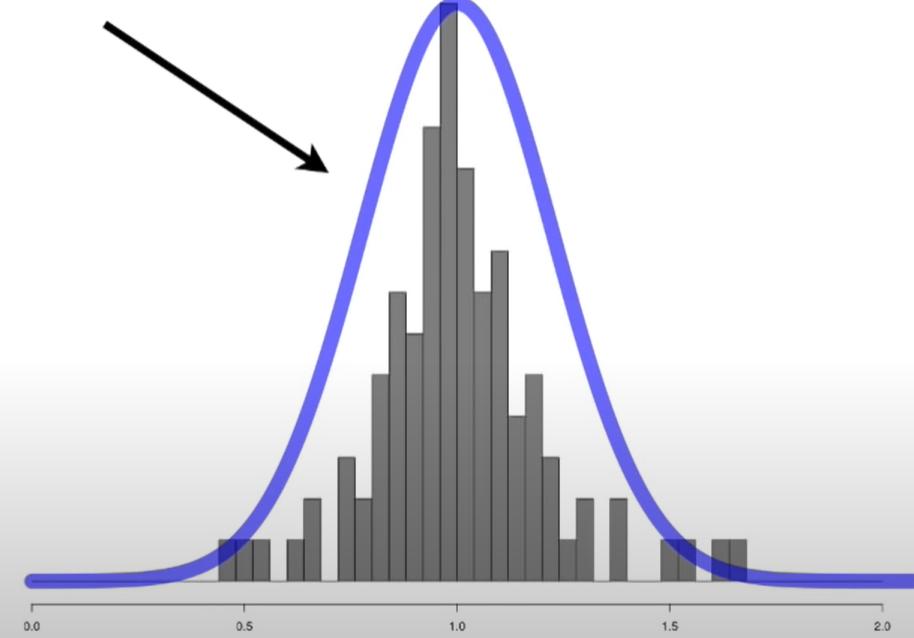
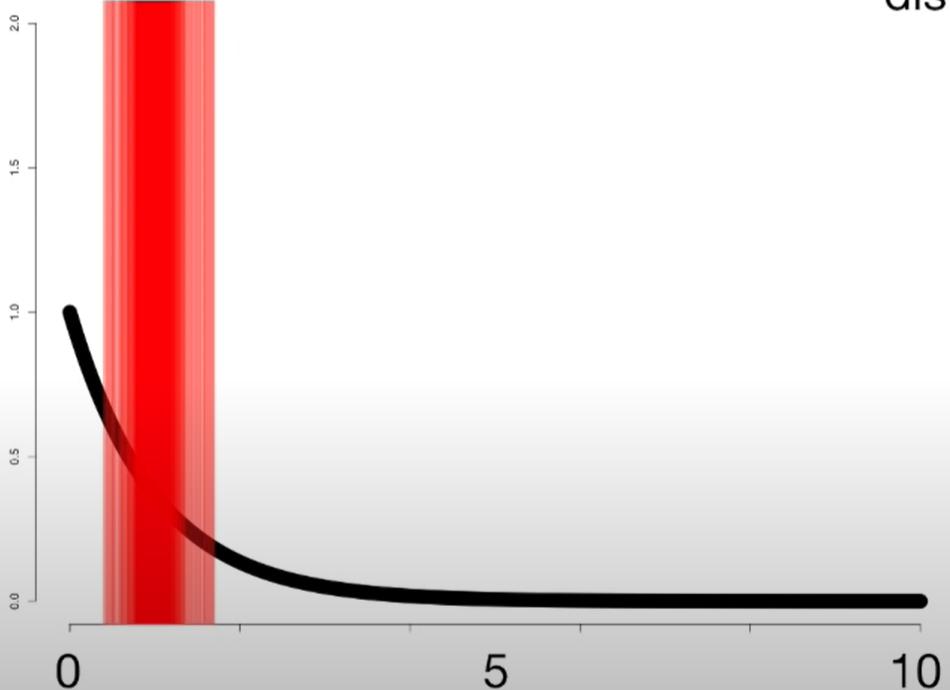
...100 means.

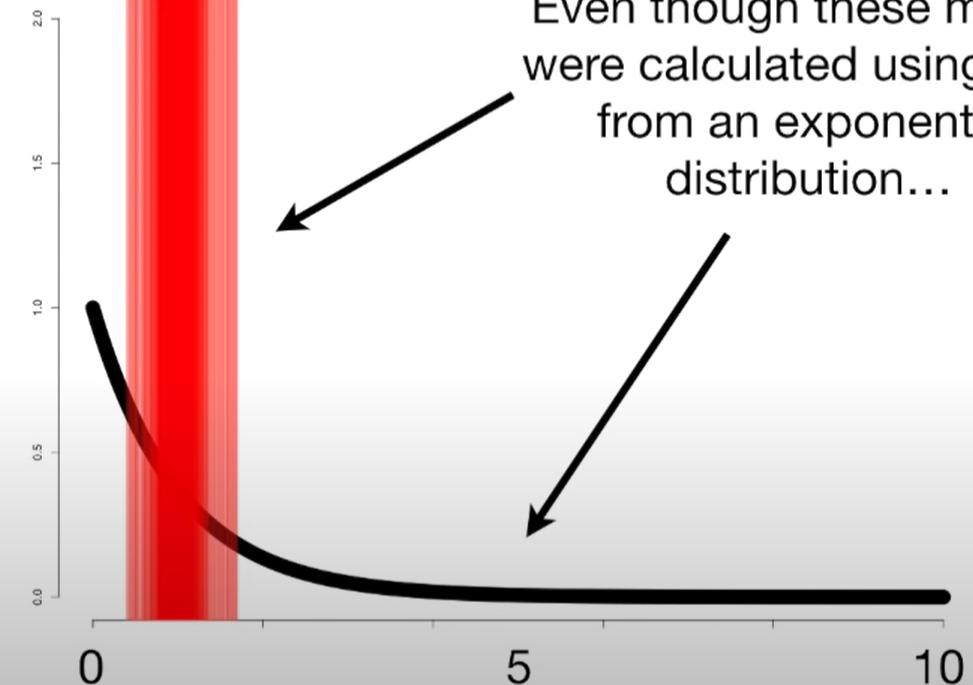


So!

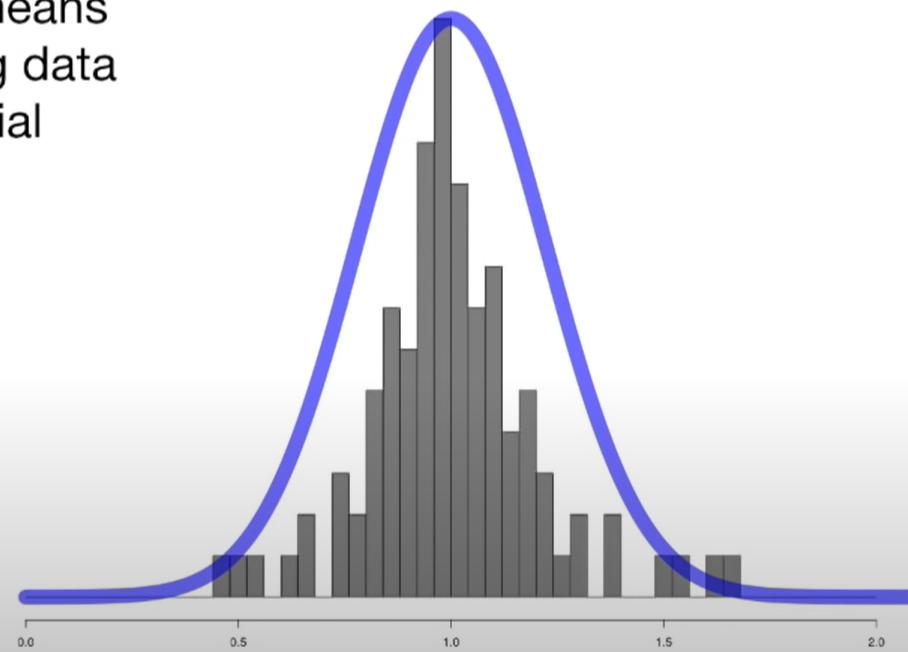


After adding 100 means to the histogram, we can see that they are normally distributed.

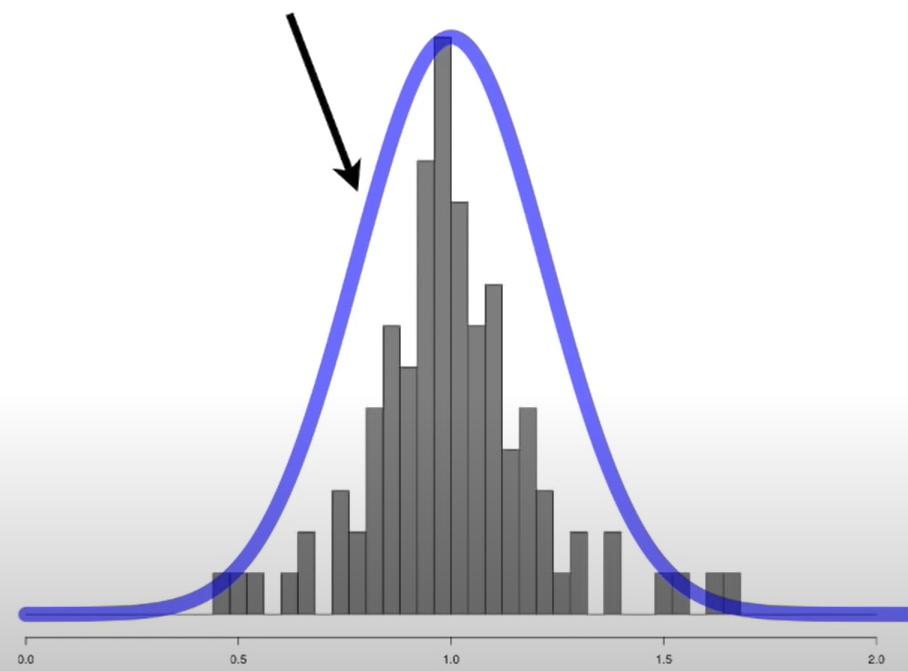
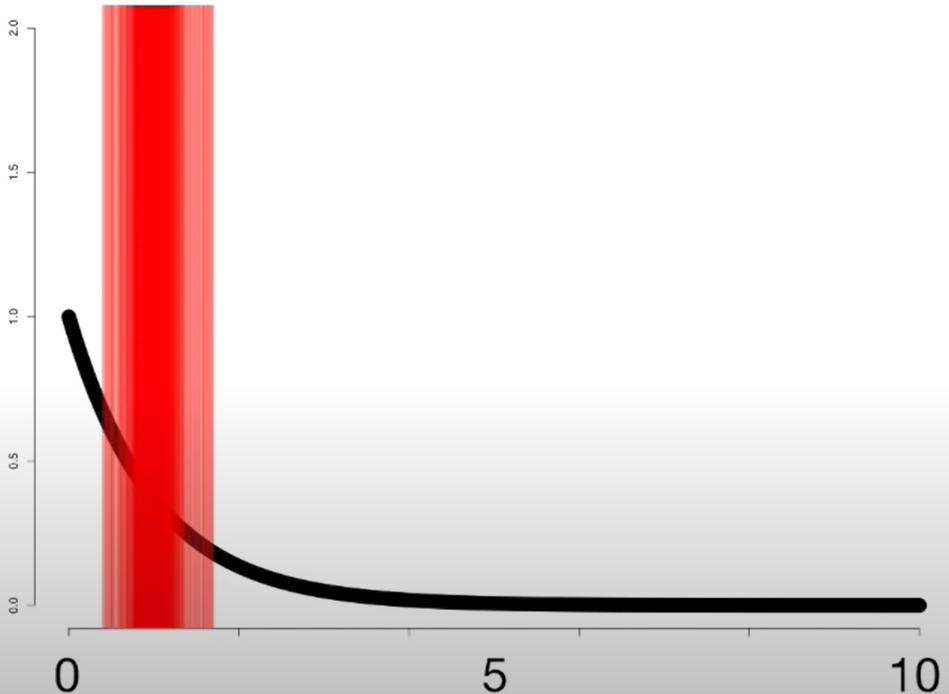


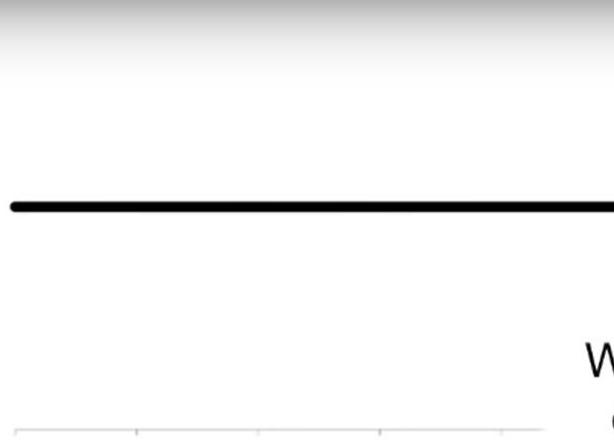


Even though these means
were calculated using data
from an exponential
distribution...

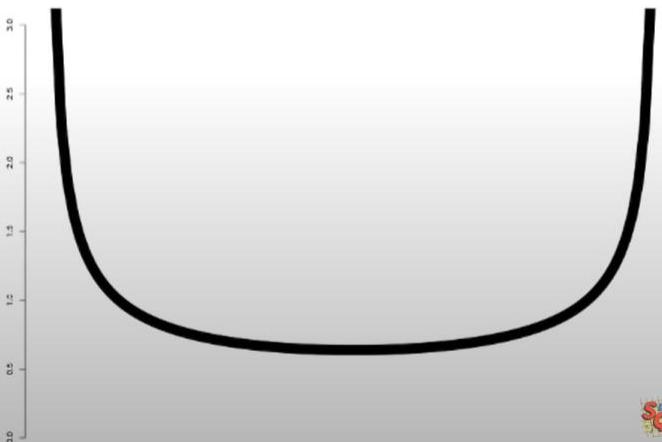
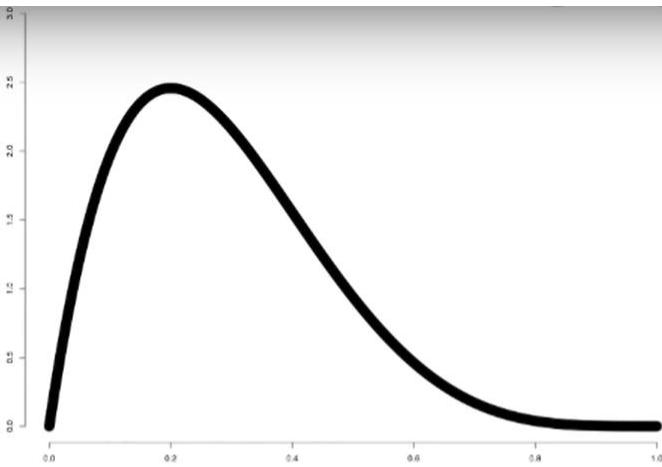


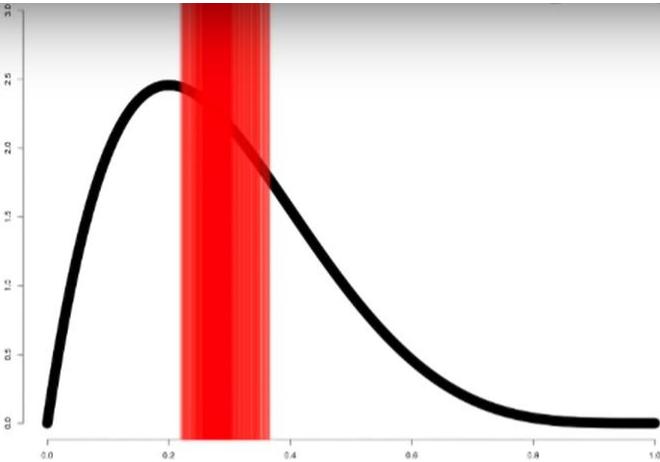
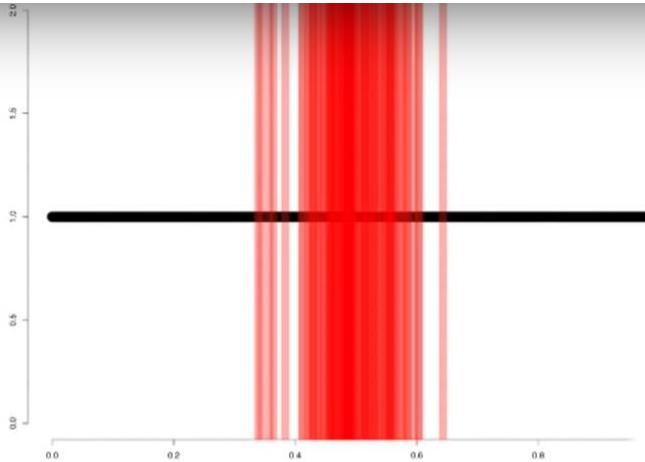
...the means themselves are not exponentially distributed. Instead, the **means are normally distributed.**



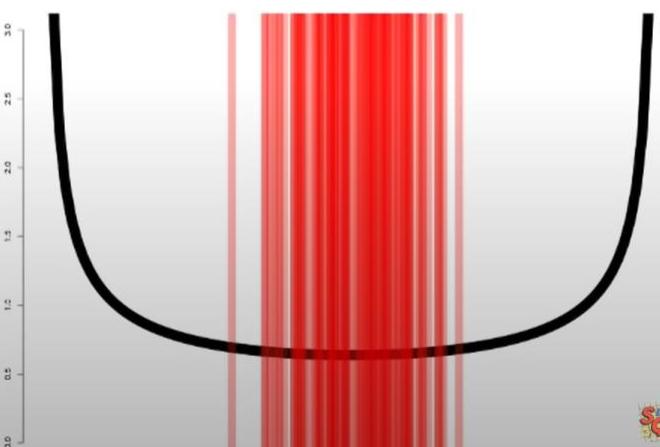
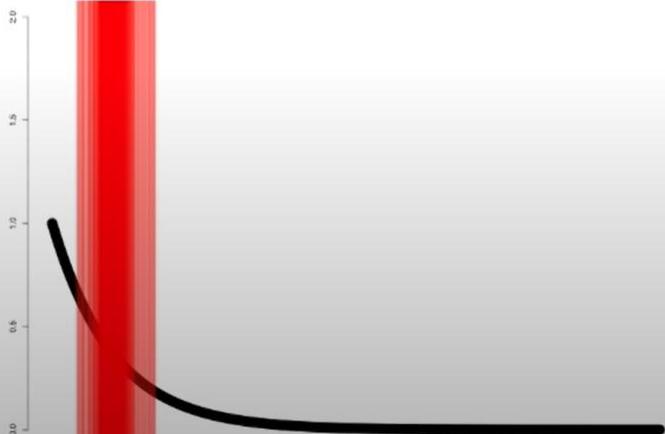


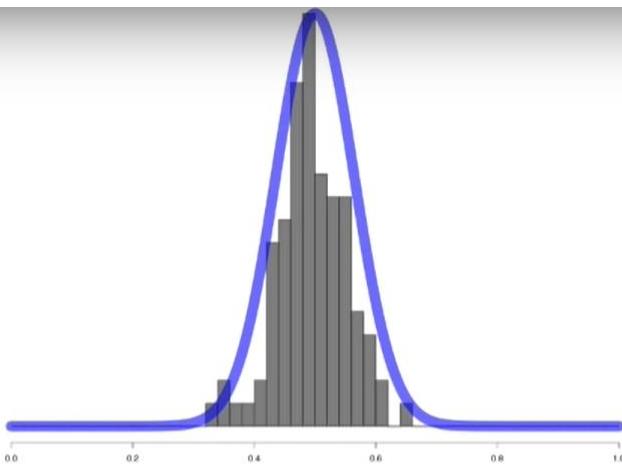
Well, it turns out that it
doesn't matter what
distribution you start
with...



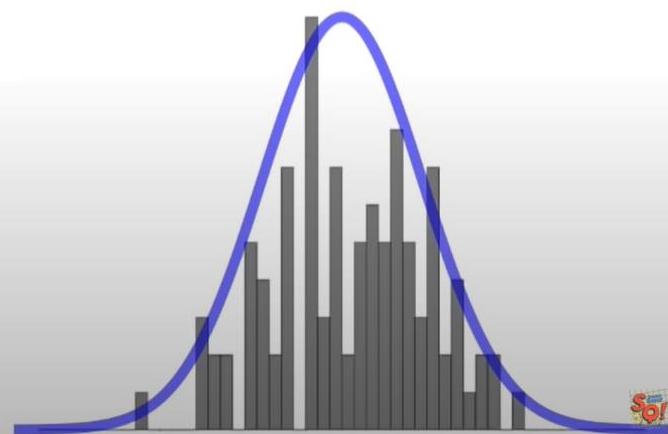
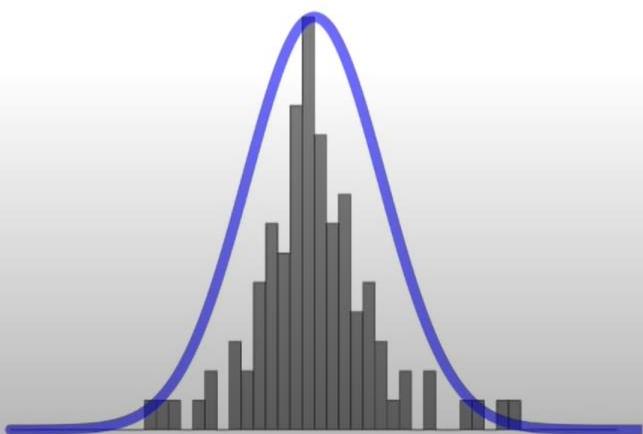
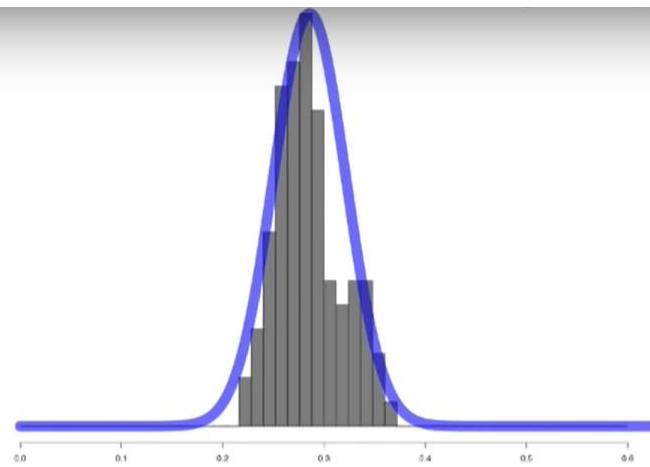


...if you collect
samples from those
distributions...

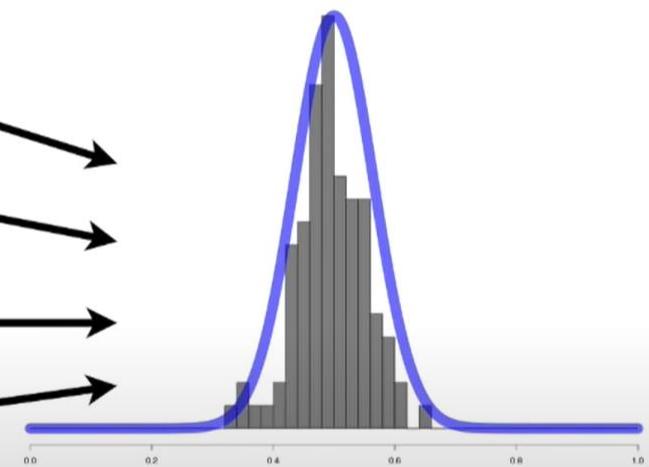
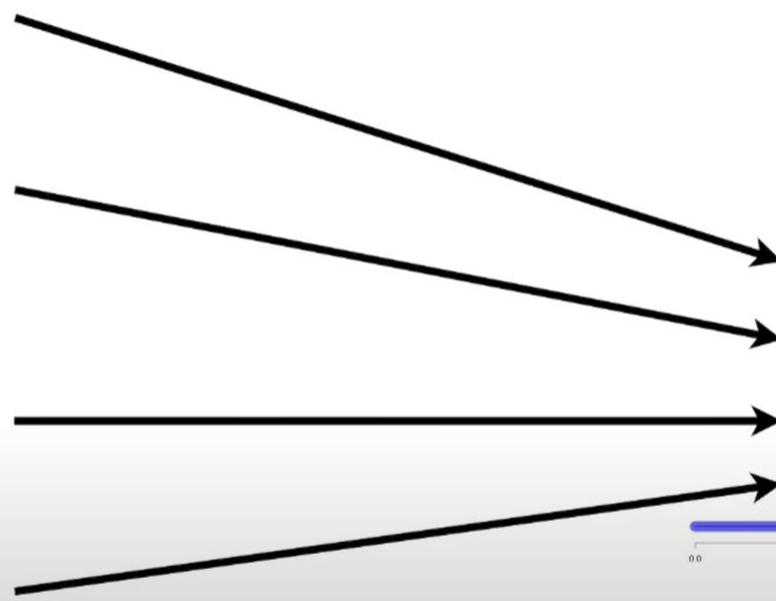
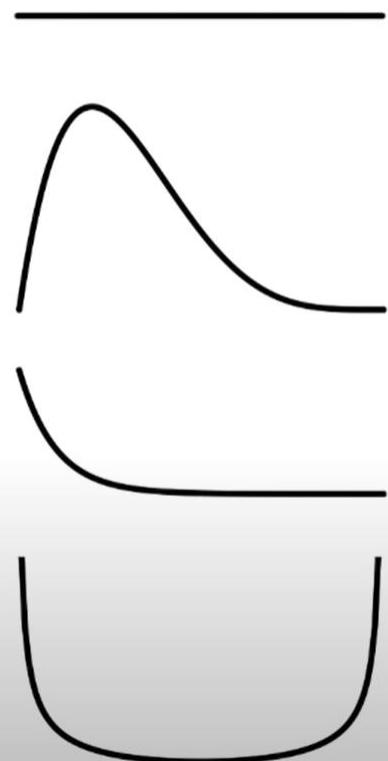


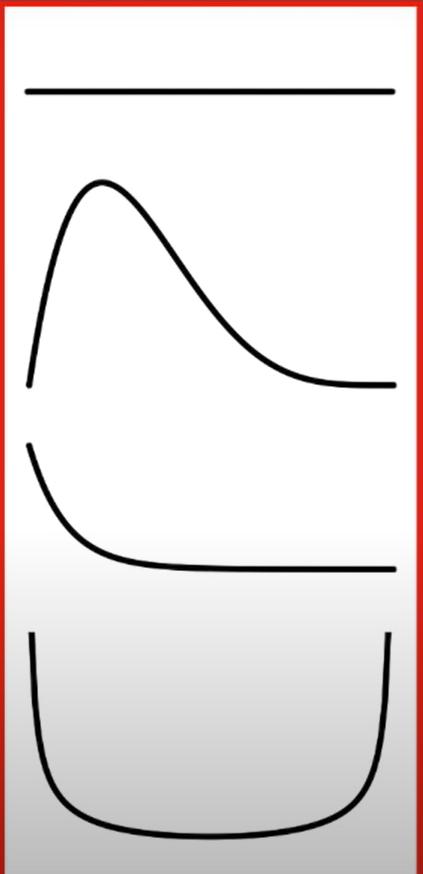


...then the means will
be normally
distributed.

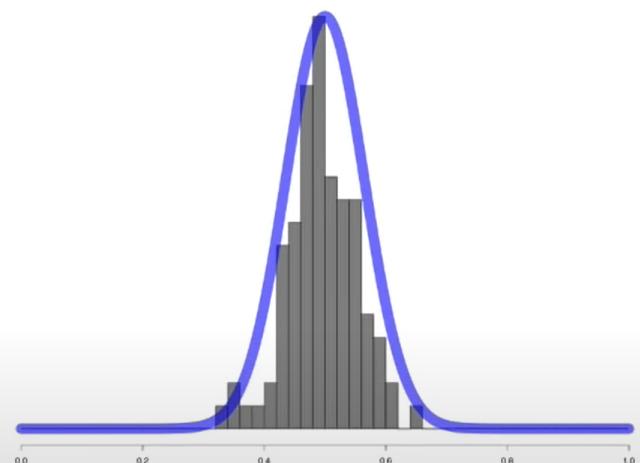


Because we know
that the sample
means are normally
distributed...

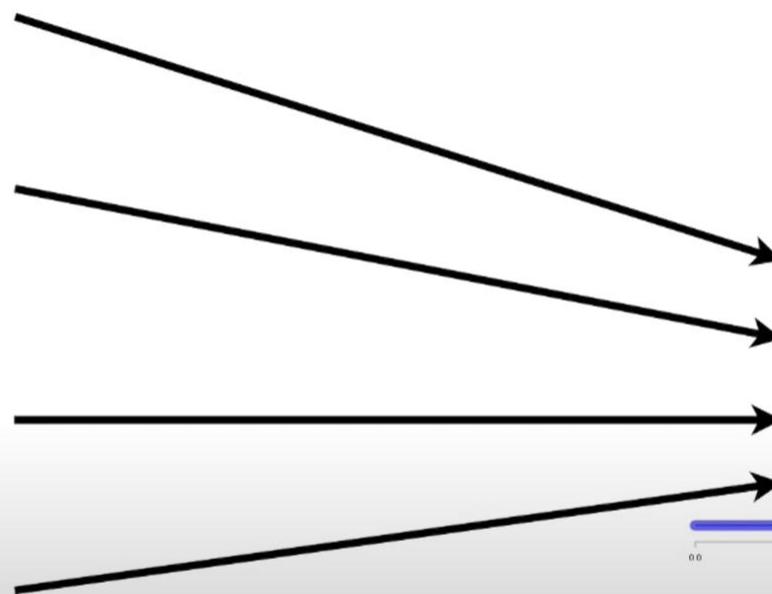
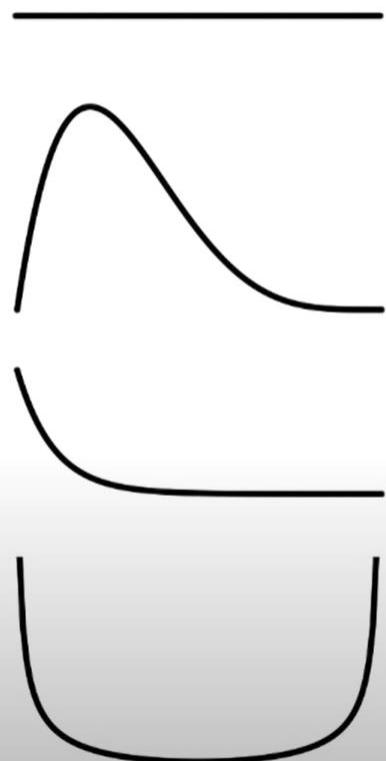
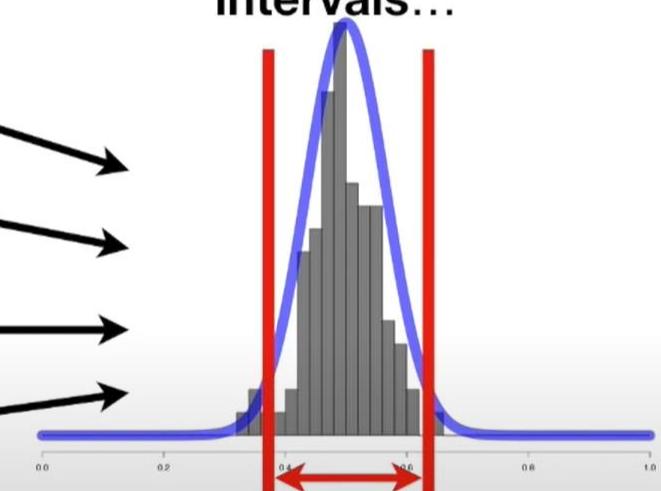




...we don't need to worry
too much about the
distribution that the
samples came from.

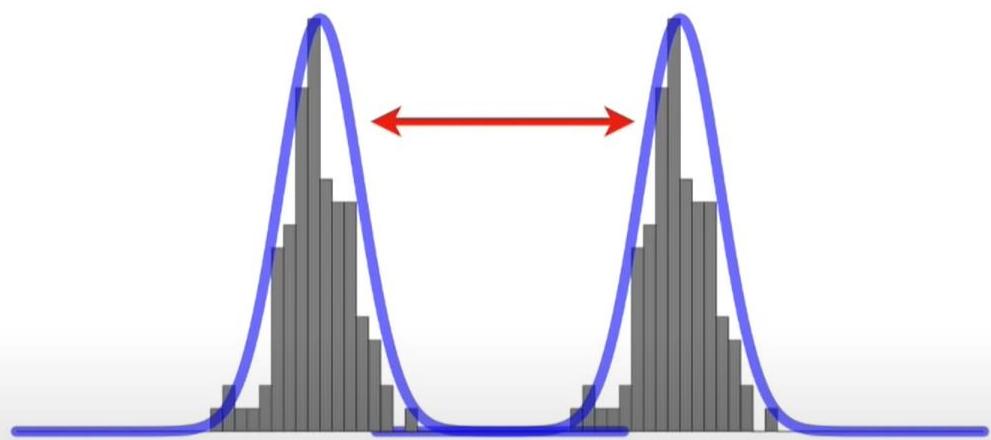


We can use the
mean's normal
distribution to make
**confidence
intervals...**

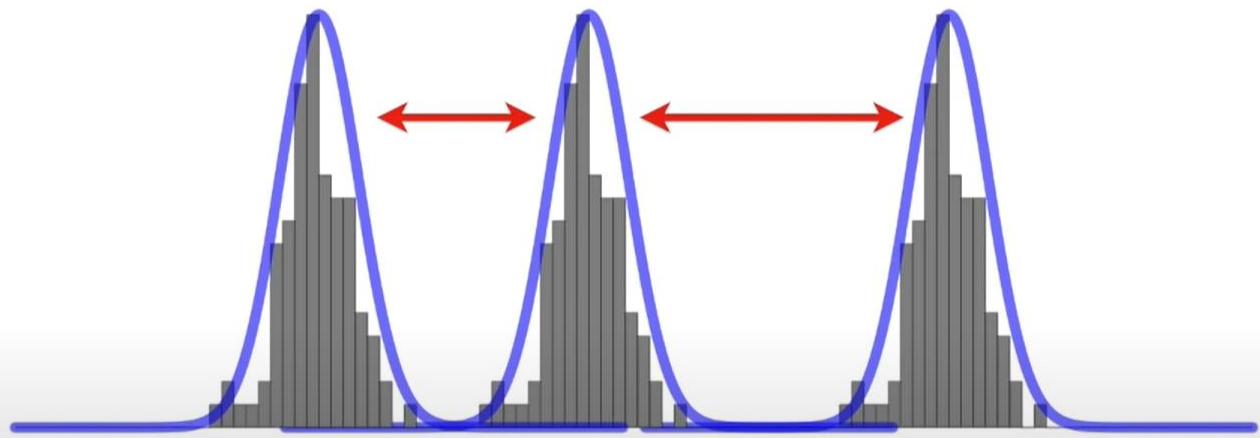




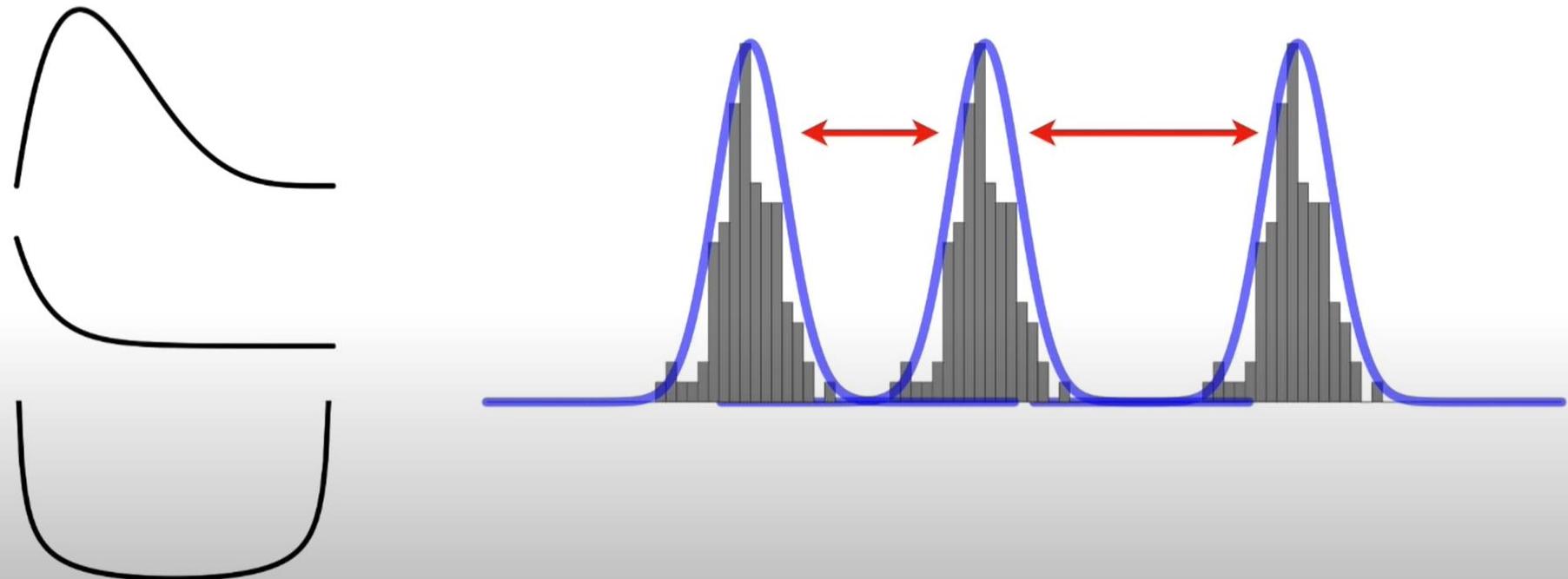
...do **t**-tests, where we ask if there is a difference between the means from two samples...

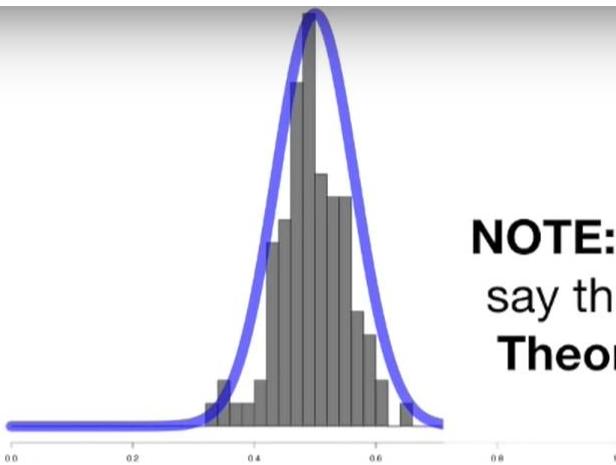


...and **ANOVA**, where we ask if there is a difference among the means from three or more samples...

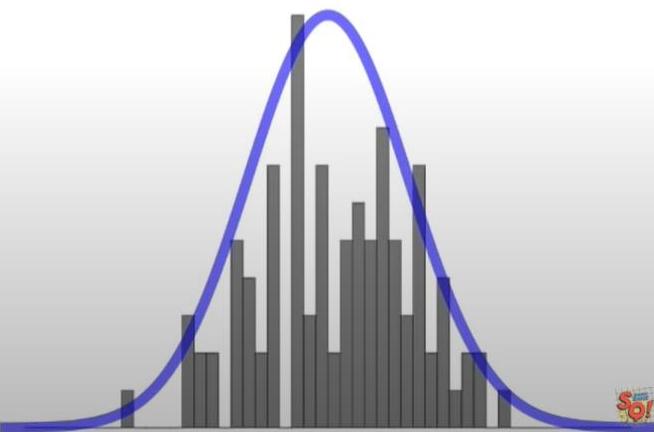
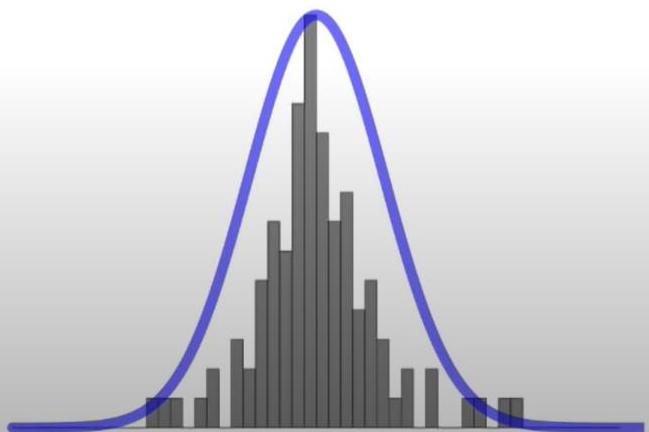
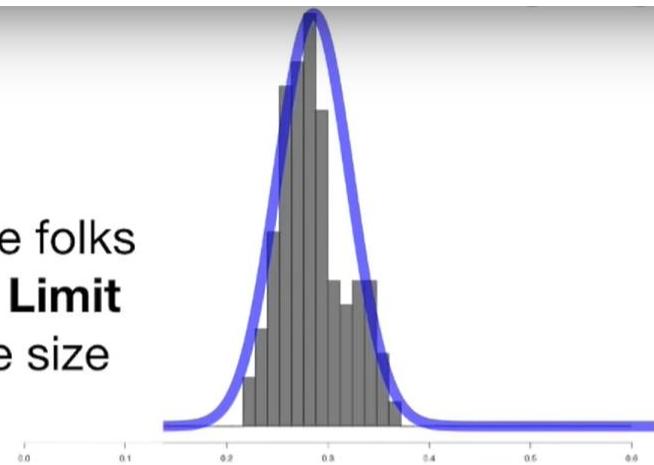


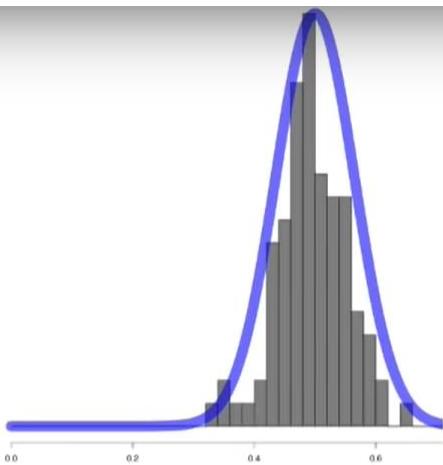
...and pretty much any statistical test
that uses the sample mean.



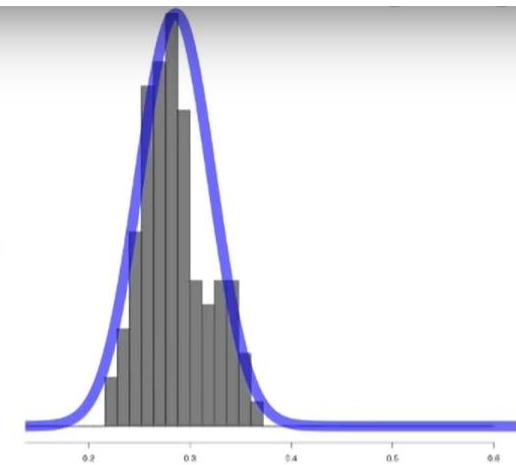


NOTE: Out there in the wild some folks say that in order for the **Central Limit Theorem** to be true, the sample size must be at least **30**.

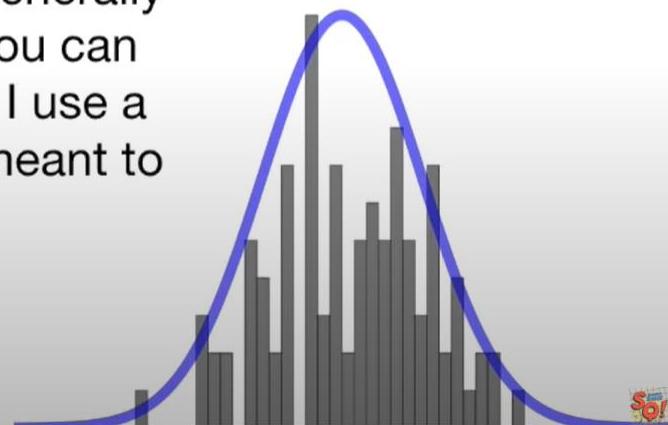
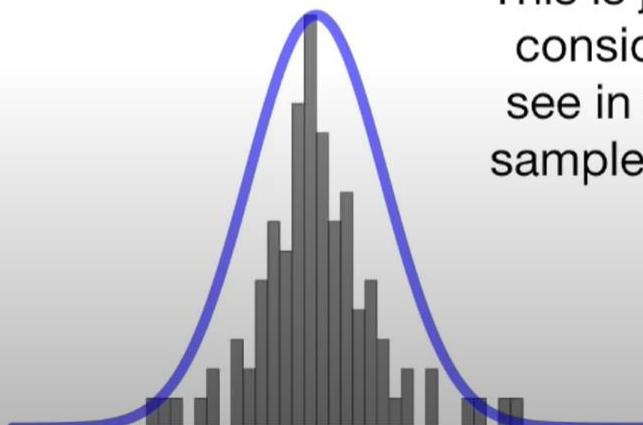




NOTE: Out there in the wild some folks say that in order for the **Central Limit Theorem** to be true, the sample size must be at least **30**.



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

- As n increases, the distribution of **sample means** will approach a normal distribution.
- More generally, as n increases, the distribution of the **sums** of n measurements will approach a normal distribution.

If you multiply something that is normally distributed (ex. means) by a constant value, you end up with something that is normally distributed:

mean = [sum of 30 measurements]/30

normally distributed

mean \times 30 = **[sum of 30 measurements]**

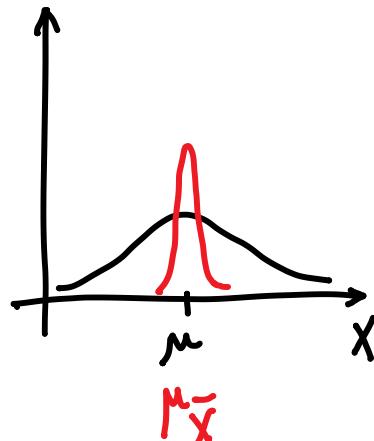
normally distributed

- In nature, things are often the sum of lots of little things! (example: human height)
→ **Normal distribution is common.**

Central Limit Theorem

When the original variable is normally distributed
OR sample size ≥ 30 ,

the distribution of the means of all possible samples of size n will be normally distributed with



$$\mu_{\bar{X}} = \mu$$

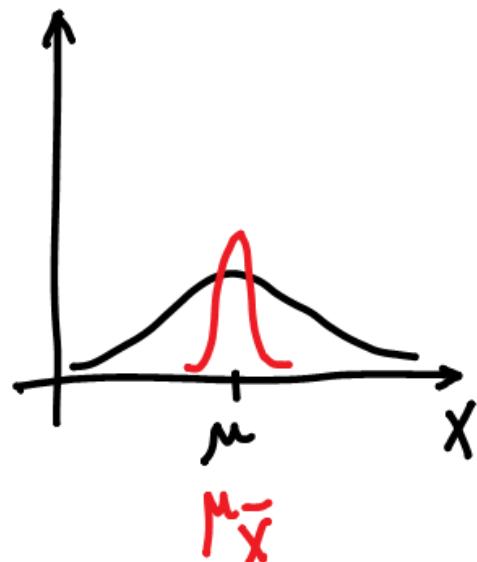
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Standard Error of the Mean
= Standard Error

Statistics from the sample

(= estimates of the population parameters!)

Standard Deviation of the sample means (also called the **Standard Error**, short for **Standard Error of the Mean**) and it's square, called the variance of the sample means, are estimated by:



$$s_{\bar{y}}^2 = s^2/n$$

$$s_{\bar{y}} = \sqrt{s^2/n} = \frac{s}{\sqrt{n}}$$

Standard Error of the Mean
= Standard Error

Other standard errors...

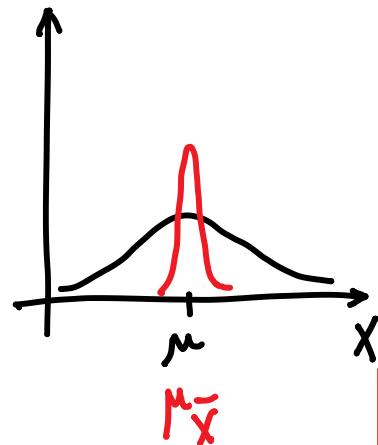
- We can calculate the standard deviation of any other statistic (standard deviation, median, mode, percentiles...) that we calculate for multiple samples.

→ This is the « **standard error** of ... » .

Central Limit Theorem

When the original variable is normally distributed
OR sample size ≥ 30 ,

the distribution of the means of all possible samples of size n will be normally distributed with



$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Standard Error of the Mean
= Standard Error

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
 follows a standard normal distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
 follows a standard normal distribution

The Student's t-distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ follows a standard normal distribution}$$

- We will use this notion to construct a confidence interval for the population mean μ .
- But σ is a population parameter
 - This value is unknown!
 - We will use the sample standard deviation s to estimate it.

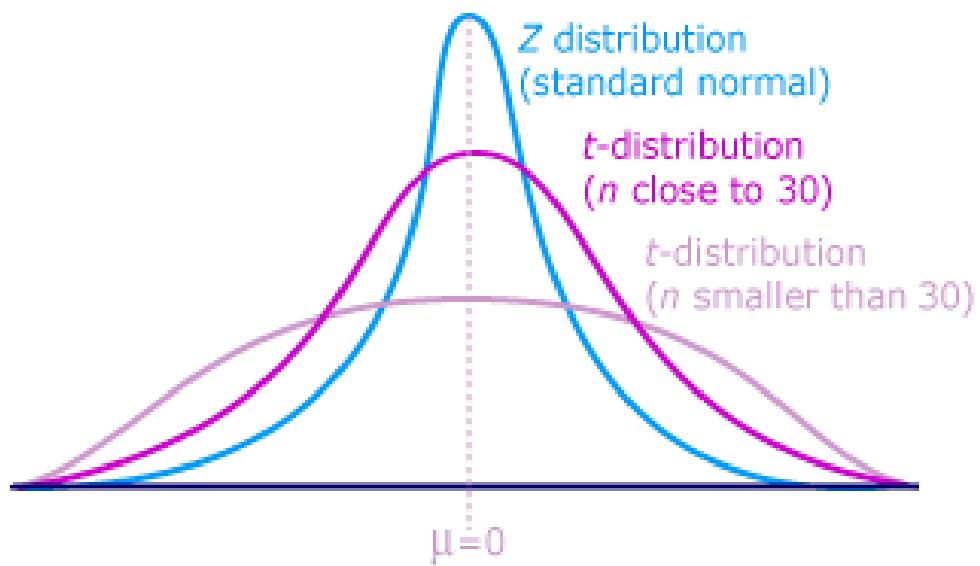
$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \text{ follows a t distribution with } n-1 \text{ degrees of freedom}$$

- s is a statistic
 - has a sample distribution; it will vary from sample to sample; t will be more variable than z
- The t-distribution looks like the Z-distribution but is wider: heavier tails and lower peak!
- The degrees of freedom of the t are related to the degrees of freedom of the sample variance s^2 (denominator = $n-1$)
- As the degrees of freedom increase (larger samples), the t-distribution looks more and more like the standard normal distribution (less variation in s!)

t-distribution

- Looks like standard normal distribution:
- Symmetric about the mean ($=0$)
- BUT
 - t-distribution is a family of curves with $\sigma^2 > 1$
- σ^2 depends on the degrees of freedom (d.f)
 - d.f. $\rightarrow +\infty$
 - $\sigma^2 \rightarrow 1$
 - t-distribution \rightarrow standard normal distribution $N(0,1)$
- Used for a normally distributed variable whenever the variance of that variable is not known.

t-distribution



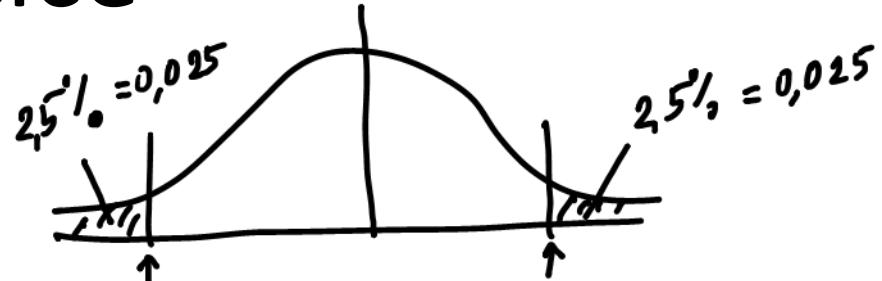
t-distribution R functions

#random generation
`rt(n, df)`

#give quantile value based on probability
`qt(p, df)`

#give probability based on quantile value
`pt (q, df)`

Exercice



- **Problem**

Find the 2.5th and 97.5th percentiles of the t-distribution with 5 degrees of freedom.

- **Solution using Probability tables**

We are looking for t-values based on area's

In the table we need the d.f., alpha, and to know whether we look at one or two tails.

Draw the t-distribution and the 2.5th and 97.5th percentiles

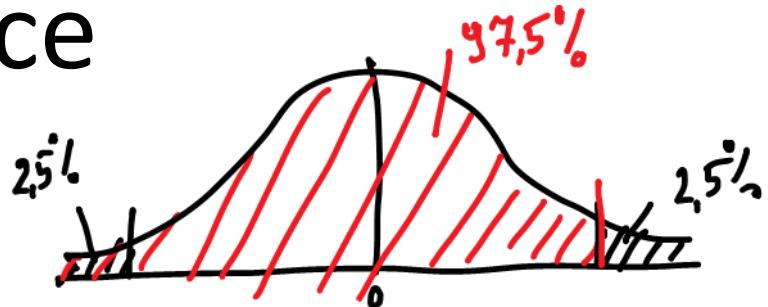
- we have two tails
- alpha is 5% (= 0.05) divided over two tails
- 5 d.f.

Correspondingly, in the table we find: $t = 2.571$

- **Answer**

The 2.5th and 97.5th percentiles of the Student t distribution with 5 degrees of freedom are -2.571 and 2.571 respectively.

Exercice



- **Problem**

Find the 2.5th and 97.5th percentiles of the t-distribution with 5 degrees of freedom.

- **Solution using R**

We are looking for t-values (quantiles, percentiles) based on area's (probabilities)

We apply the quantile function qt of the Student t distribution

`qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)`

```
> qt(c(.025, .975), df=5) # 5 degrees of freedom  
[1] -2.5706 2.5706
```

- **Answer**

The 2.5th and 97.5th percentiles of the Student t distribution with 5 degrees of freedom are -2.5706 and 2.5706 respectively.

T-distribution, example

$t_{n-1, 1-\alpha/2}$:

- $n-1$ is the degrees of freedom
- we are looking for the $1-\alpha/2$ percentile

For example, for $n = 5$ and $\alpha = 0.05$, we are looking for t with 4 degrees of freedom and the 0.975 percentile.

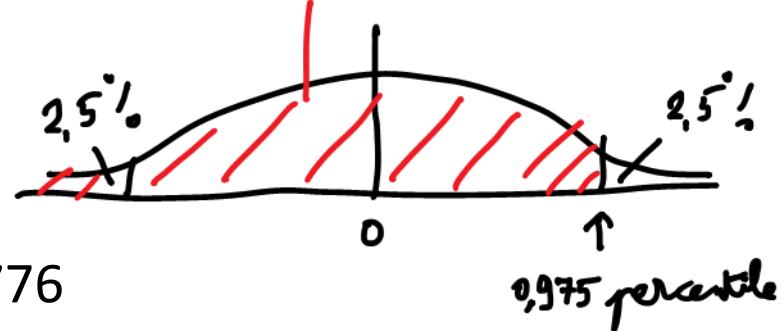
Table:

two tails; alpha = 0.05; d.f. = 4 $\rightarrow t=2.776$

Using R:

```
> qt(0.975, 4)  
[1] 2.776445
```

$$\begin{aligned} & 100\% - 2,5\% \\ & = 97,5\% \end{aligned}$$



Confidence intervals for the mean

Confidence intervals for the mean

- Collect data and get point estimates:
 - The sample mean \bar{y} to estimate the population mean μ .
 - The sample variance, s^2 to estimate the population variance σ^2
- Can calculate interval estimates of each point estimate
e.g. 95% confidence interval for the true mean
 - *If \bar{y} is normally distributed*
this is:
 - If the y 's are *normally distributed* *OR*
 - The sample size is large enough that the Central Limit Theorem holds -- \bar{y} *will be* normally distributed

A. Original variable normally distributed and σ known *OR*
sample size ≥ 30

- Distribution of the means is normal distribution
- Mean of sample means = population mean

$$\mu_{\bar{X}} = \mu$$

- Standard deviation of sample means (SE)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- In the case sample size ≥ 30 , s can substitute σ

A. Original variable normally distributed and σ known **(OR sample size ≥ 30)***

- Sample mean lays a number of standard deviations away from the mean (population mean)
- That number is the z-value (follows standard normal distribution $N(0,1)$)
- With $1 - \alpha$ percent certainty, the mean of the sample lays z standard errors away from the population mean μ (value of z depends on alpha)

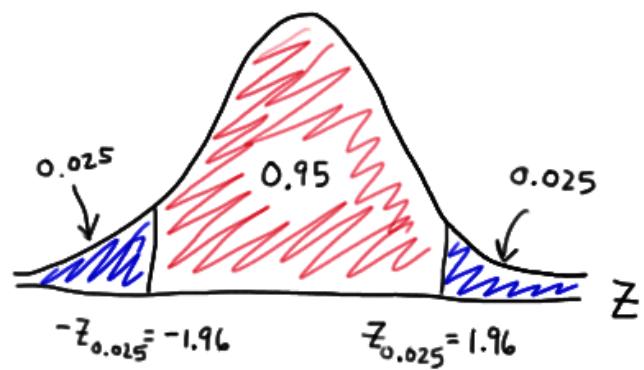
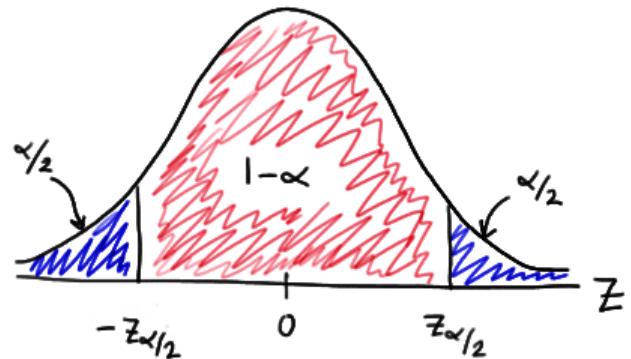
$$\mu - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \bar{X} < \mu + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

*depends on the points of view of
the statistician

- Based on standard normal distribution:
- Confidence interval for a specific α :
- For $\alpha = 0.05$
 \Leftrightarrow 95% confidence interval

The values are, with 95% confidence, how many standard deviations away from the mean? \rightarrow critical z-value

- Look up $100(1 - \alpha/2)$ percentile
 = 97.5th percentile
 = denoted $z_{\alpha/2}$
 = z value for $P = (0.95/2)$
 = z value for $P = (0.475)$
 = 1.96



A. Original variable normally distributed and σ known (OR sample size ≥ 30)*

- interval for the true mean of the population for a specific alpha (level of significance):

$$\bar{X} - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

- Intuitively: z values are actually the number of standard deviations that a value is away from the mean
- for $\alpha = 0.05$, 95% probability that real population mean is less than two standard errors away of \bar{X}
- Margin of Error:**

$$z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

*depends on the points of view of the statistician

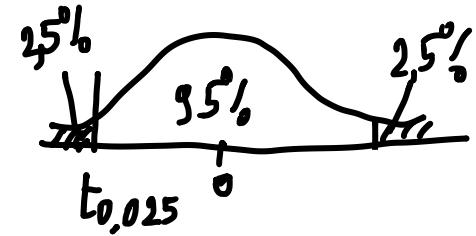
B. Original variable normally distributed but σ is unknown (and $n < 30$)*

- You can use the standard deviation of the sample in place of the standard deviation of the population BUT with another distribution: “**t-distribution**”
- interval for the true mean of the population for a specific alpha (level of significance):

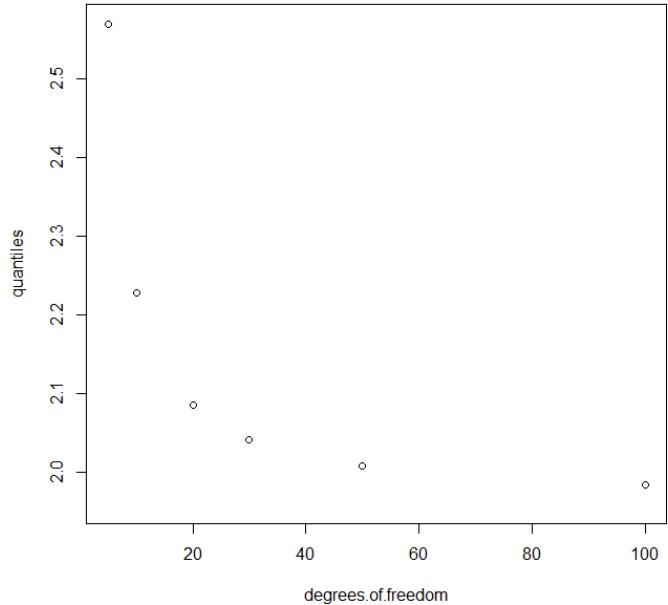
$$\bar{X} - t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) < \mu < \bar{X} + t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

*depends on the points of view of the statistician

Create your own point of view: calculate the $t_{.025}$ values (for 95% confidence intervals)



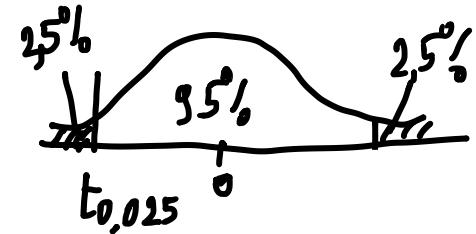
n	df	$t_{.025}$
6	5	
11	10	
21	20	
31	30	
51	50	
101	100	
∞	∞	$= z_{.025}$



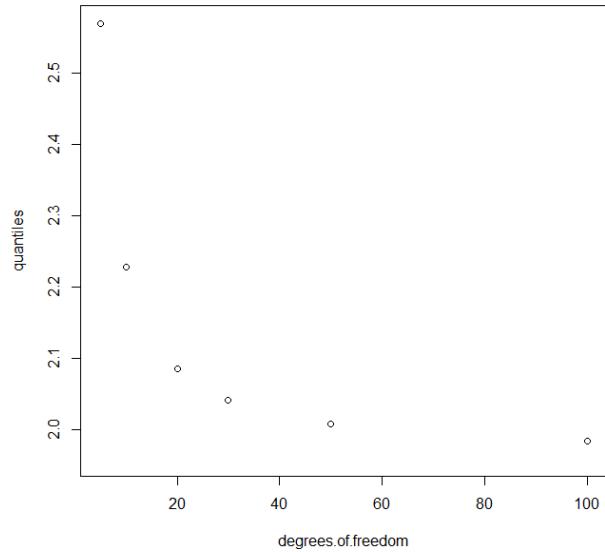
```
> degrees.of.freedom <- c(5,10,20,30,50,100,Inf)
> quantiles <- qt(p=0.025, df=degrees.of.freedom, lower.tail=FALSE)
> quantiles
[1] 2.570582 2.228139 2.085963 2.042272 2.008559 1.983972 1.959964
> plot(degrees.of.freedom, quantiles)
```

$= z_{.025}$

Create your own point of view: calculate the $t_{.025}$ values (for 95% confidence intervals)



n	df	$t_{.025}$
6	5	
11	10	
21	20	
31	30	
51	50	
101	100	
∞	∞	



If the samples are very small:

- the population mean μ is, with 95% confidence, in an interval of which the limits are MORE than two standard errors away from the mean.
- Our prediction of where the population mean should be is LESS PRECISE: a larger 95% confidence interval.
- The less data we have, the less confidence we can have in the accuracy of the estimates
- If we use the z-value, we underestimate the margin of error!

Example

- In the data set survey of the package “MASS”, the survey is performed on a sample of the student population. We will look at the student height data
- For convenience, we first filter out missing values in survey\$Height with the na.omit function, and save it in height.response.

```
> library(MASS)          # load the MASS package  
> height.response <- na.omit(survey$Height)
```

script 1_confidenceinterval.R

Example

- **Problem 1**

Assume the population standard deviation σ of the student height in survey is 9.48. Find the best point estimate of the population mean and the margin of error and interval estimate at 95% confidence level.

- **Problem 2: the real world!**

Without assuming the population standard deviation of the student height in survey, find the margin of error and interval estimate at 95% confidence level.

script 1_confidenceinterval.R

Exercise 1

- Find the 95% confidence interval of the population mean of all numeric data in “ufc_processed.csv”

script TADE_exercise_1.R

Hypothesis testing

Hypothesis Tests:

- Can hypothesize what the true value of any population parameter might be, and state this as **null hypothesis (H_0 :**)
- We also state an **alternate hypothesis (H_1 : or H_a :**)
- that it is a) not equal to this value; b) greater than this value; or c) less than this value
- Collect **sample data** to test this hypothesis
- From the sample data, we calculate a sample statistic as a **point estimate of this population parameter and an estimated variance of the sample statistic.**
- We calculate a “**test-statistic**” using the sample estimates
- **Under H_0 , this test-statistic will follow a known distribution.**
- If the test-statistic is very unusual, compared to the tabular values for the known distribution, then the H_0 is very unlikely and we conclude H_1

Errors of hypothesis testing

	H0 True	H0 False
Accept	$1-\alpha$	β (Type II error)
Reject	α (Type I error)	$1-\beta$

- **Type I Error:** Reject H0 when it was true.

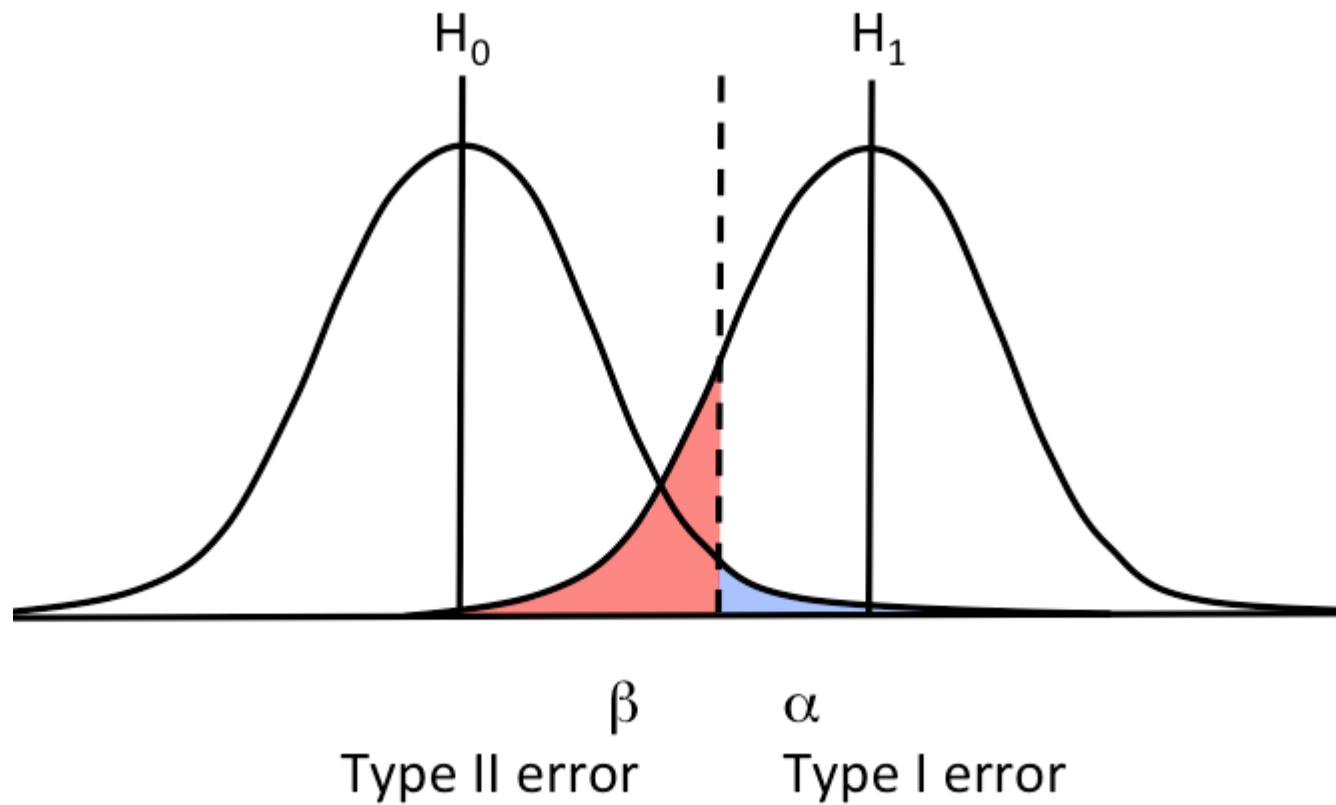
Probability of this happening is α

- **Type II Error:** Accept H0 when it is false.

Probability of this happening is β

- **Power of the test:** Reject H0 when it is false.

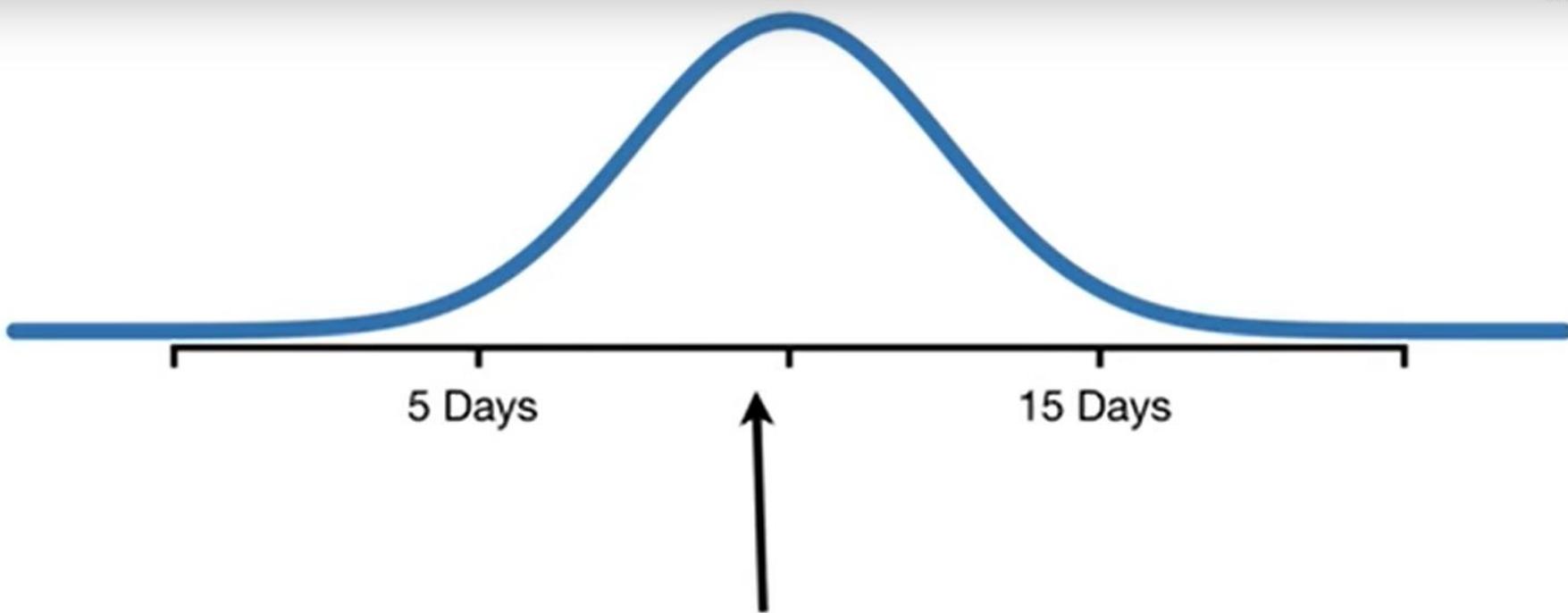
Probability of this is $1-\beta$



- **Level of significance**
 - = maximum probability of committing a type I error
 - = alpha

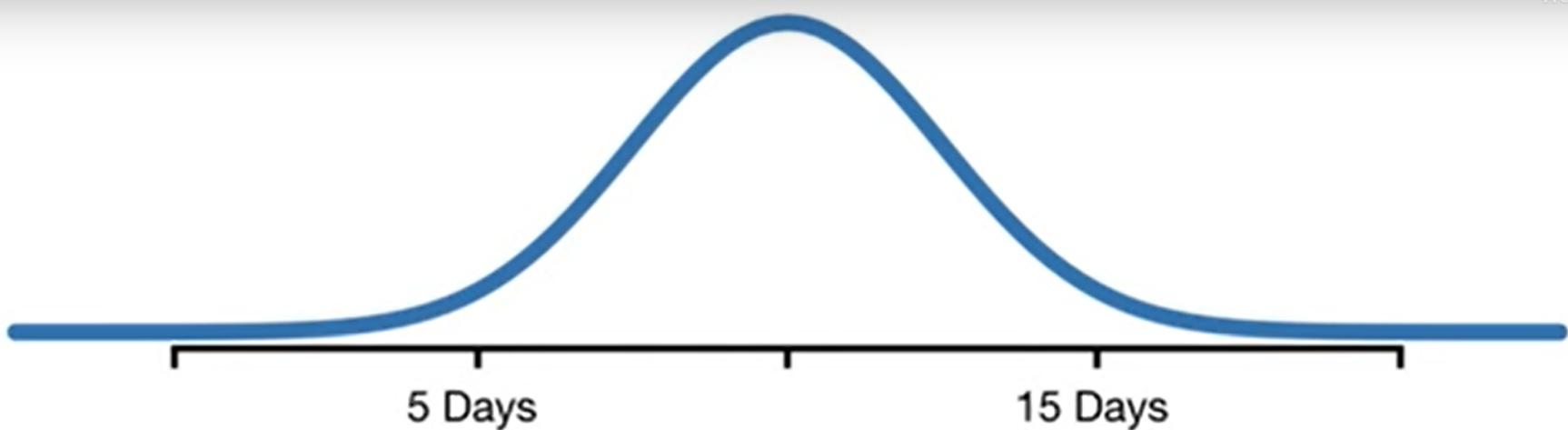
Chosen by the researcher before the statistical test is conducted
 - **P-value**
 - = Probability of getting a sample statistic,
OR equally extreme but different,
OR a more extreme statistic in the direction of the alternative hypothesis
when the null hypothesis is true
 - Computed after sample statistic is found
 - $P\text{-value} \leq \alpha \Leftrightarrow \text{reject } H_0$
 - $P\text{-value} > \alpha \Leftrightarrow \text{accept } H_0$
- } Same for one-tailed and two-tailed tests

One- or two-tailed tests?

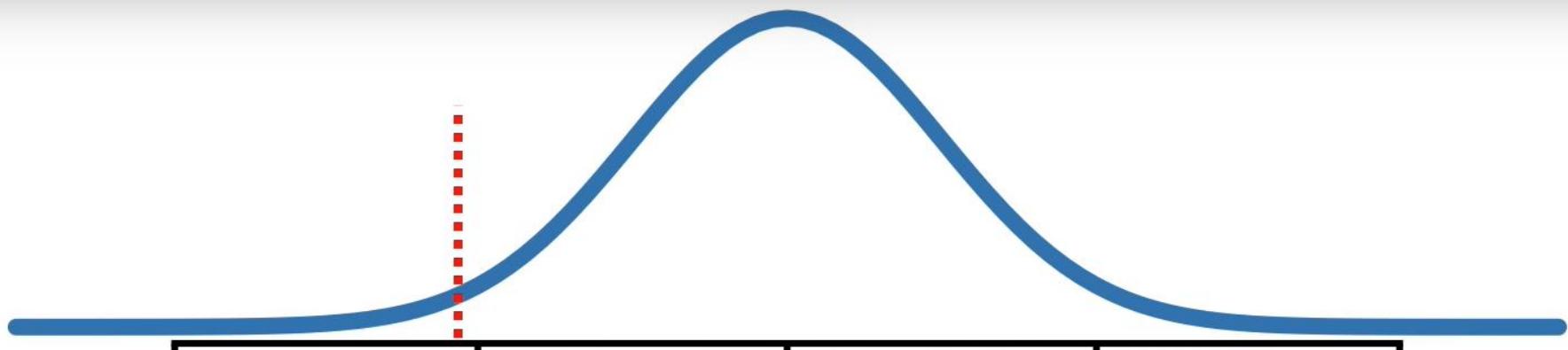


Imagine we measured how long it took a bunch of people to recover from an illness.

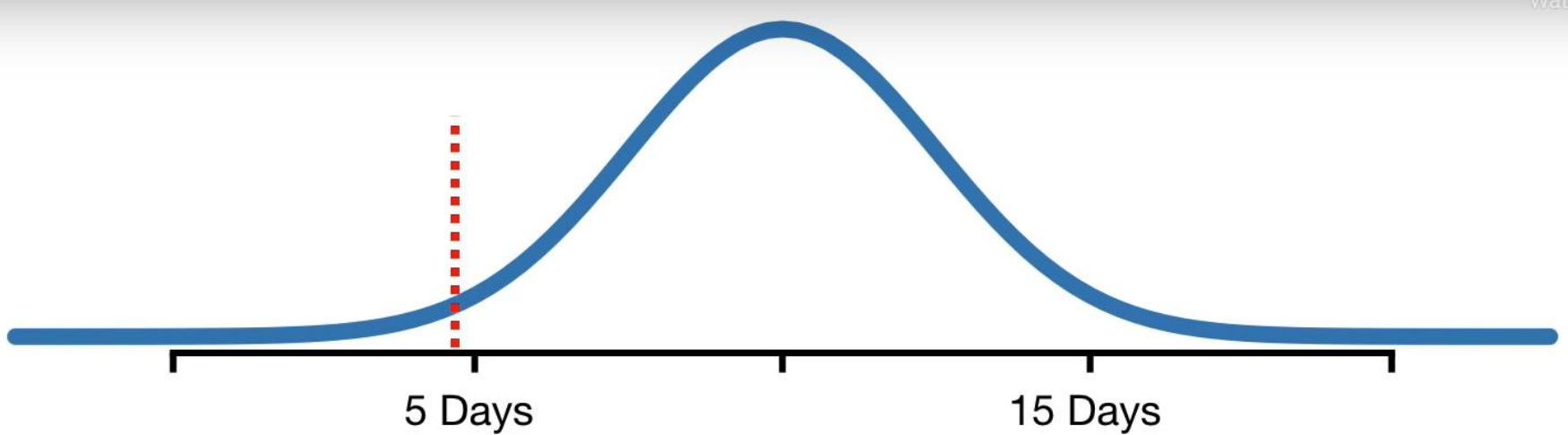




Now imagine we created a new drug,
SuperDrug, and wanted to see if it
helped people recover in fewer days.

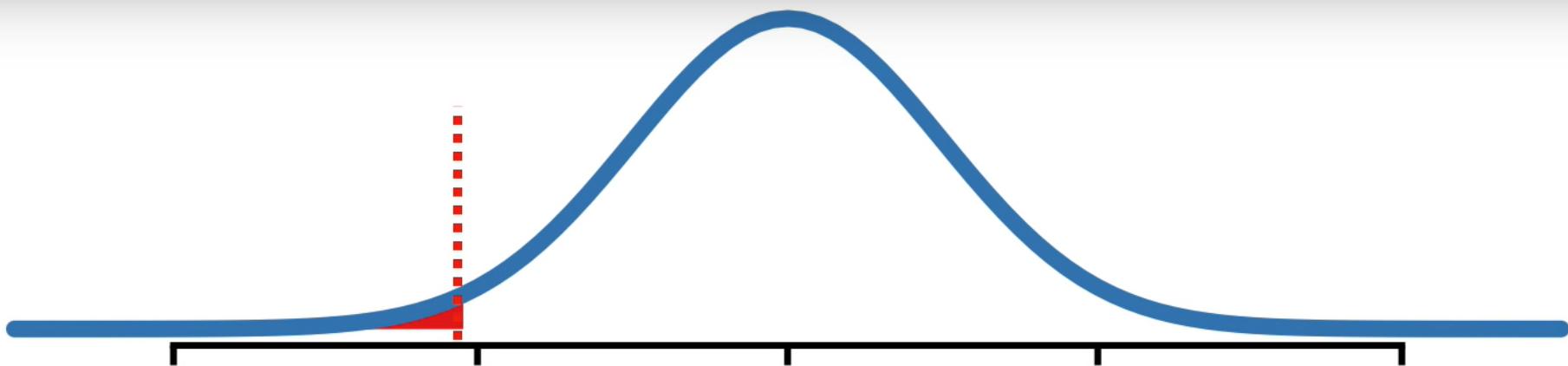


If we gave **SuperDrug** to a bunch of people
and the average recovery was **4.5** days...



...then a **Two-Sided p-value**, like the ones we've been computing all along, would be...



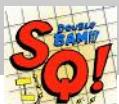


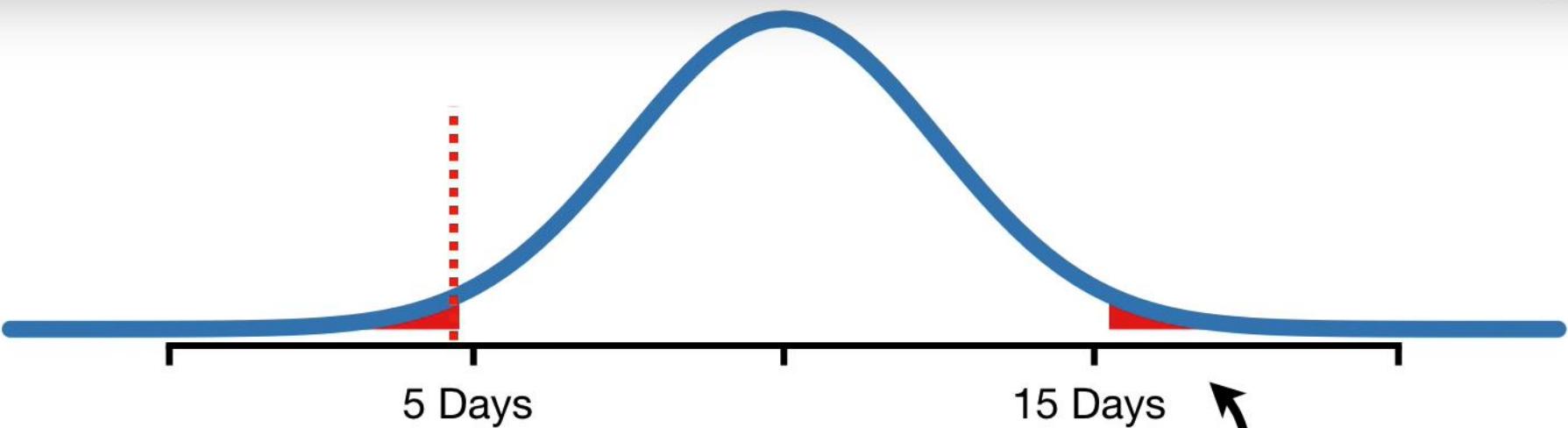
5 Days

15 Days

...the sum of *this* area under
the curve, **0.016...**

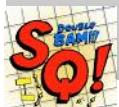
Two-Sided p-value for 4.5 days = 0.016

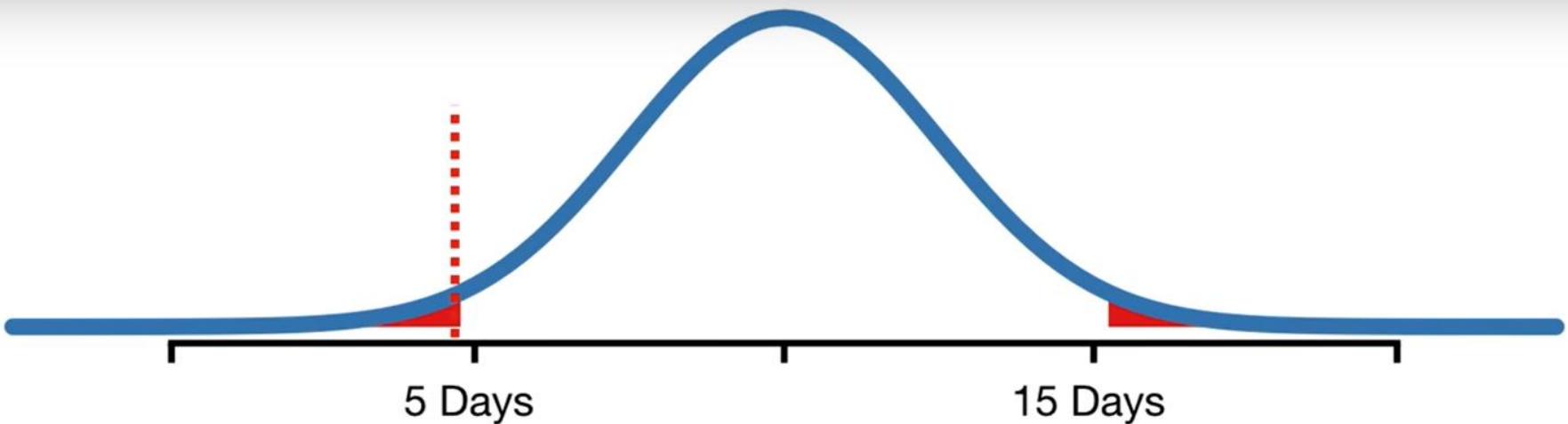




...plus *this* area under the curve, 0.016...

Two-Sided p-value for 4.5 days = $0.016 + 0.016$

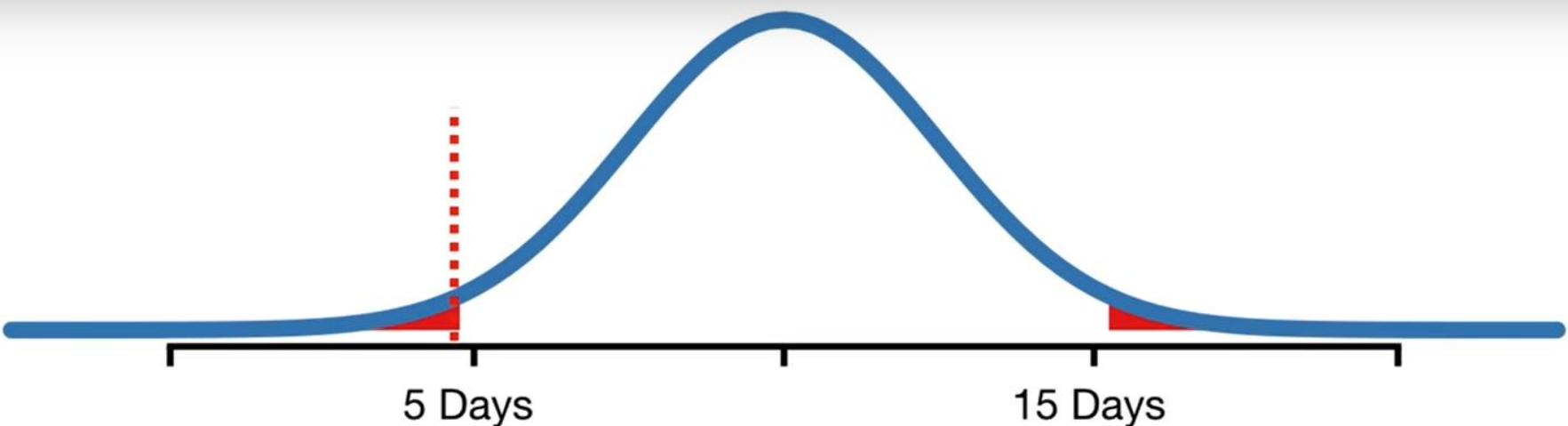




...and the total is **0.03**.



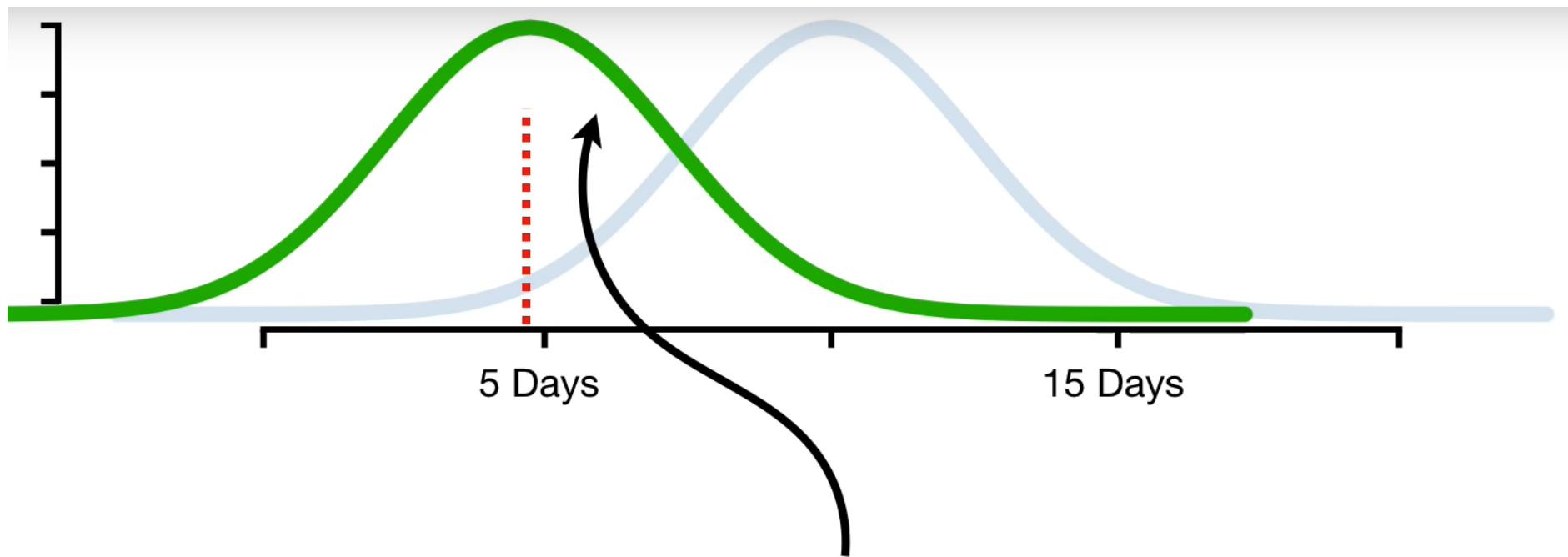
Two-Sided p-value for 4.5 days = $0.016 + 0.016 = 0.03$



And since **0.03 < 0.05**, the **Two-Sided p-value** tells us that, given this distribution of recovery times, **SuperDrug** did something unusual.

Two-Sided p-value for 4.5 days = $0.016 + 0.016 = 0.03$

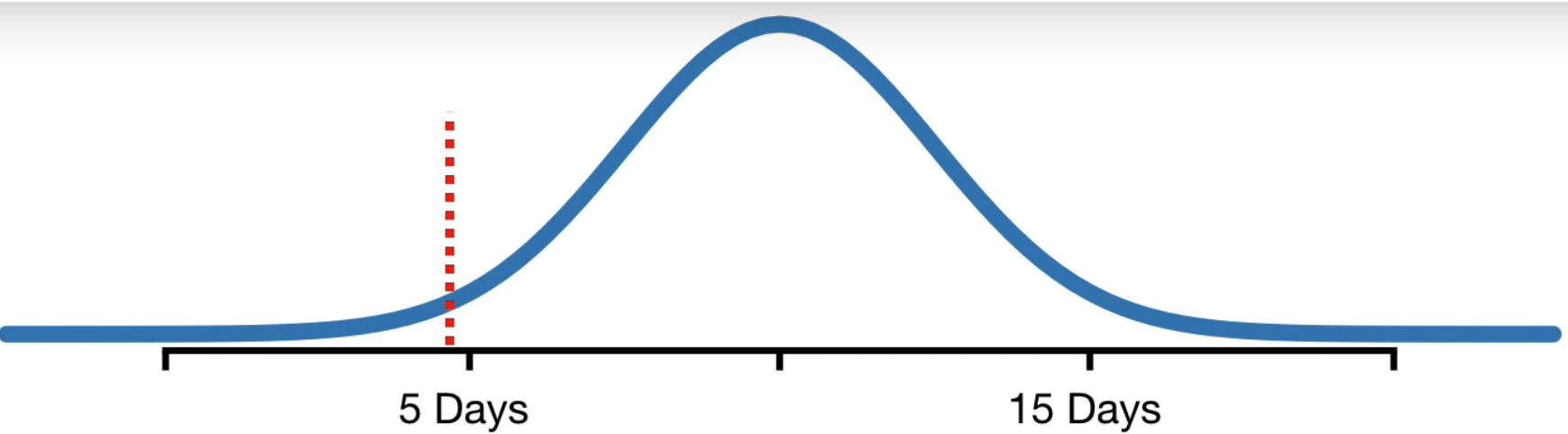




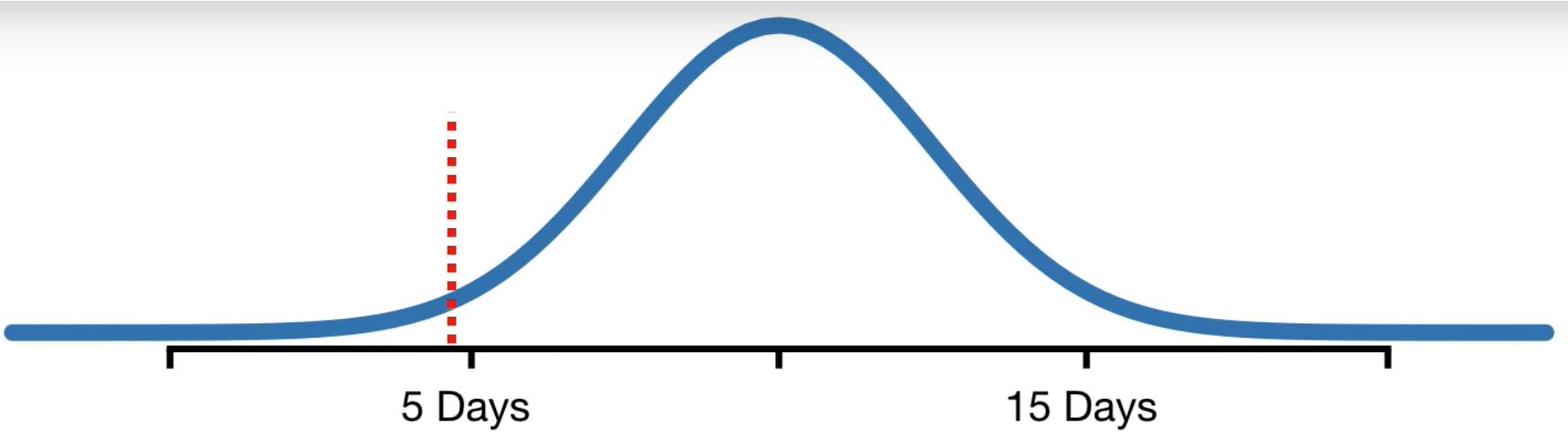
...and that suggests that some other distribution does a better job explaining the data.

Two-Sided p-value for 4.5 days = $0.016 + 0.016 = 0.03$



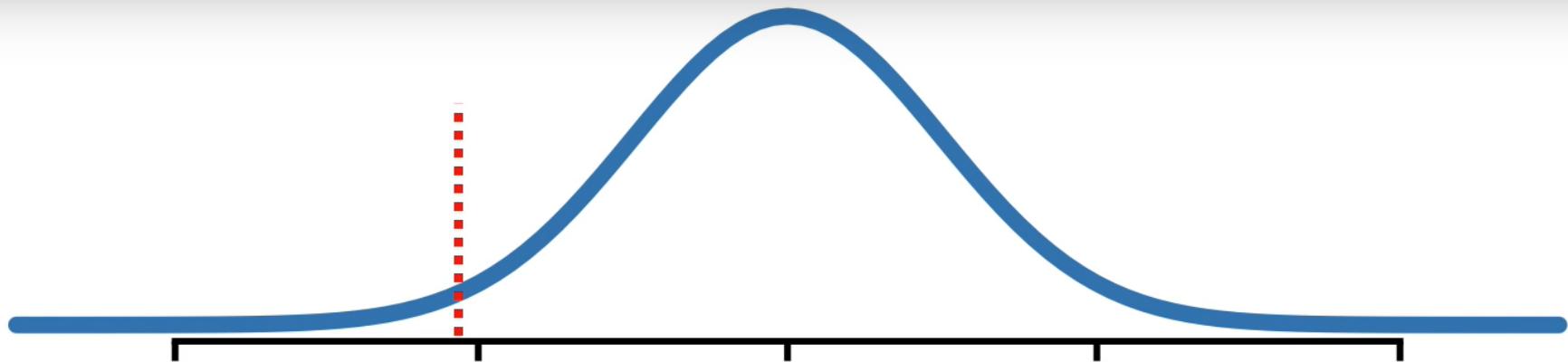


For a **One-Sided p-value**, the first thing we do is decide which direction we want to see change in.



In this case, we'd like **SuperDrug** to shorten the time it takes to recover from the illness...



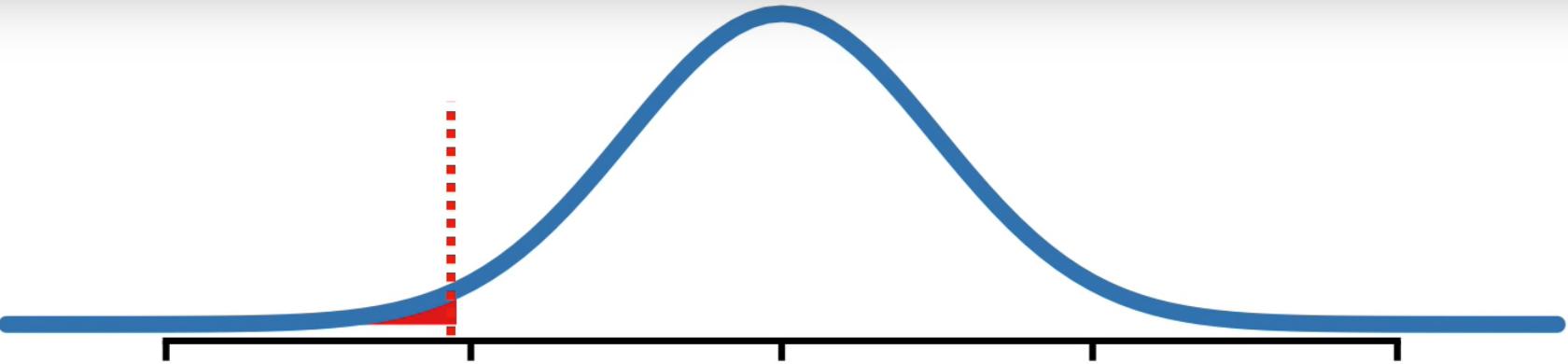


Shorter Recovery Times

5 Days

15 Days

...so that means we want to see if recovery times are shorter.

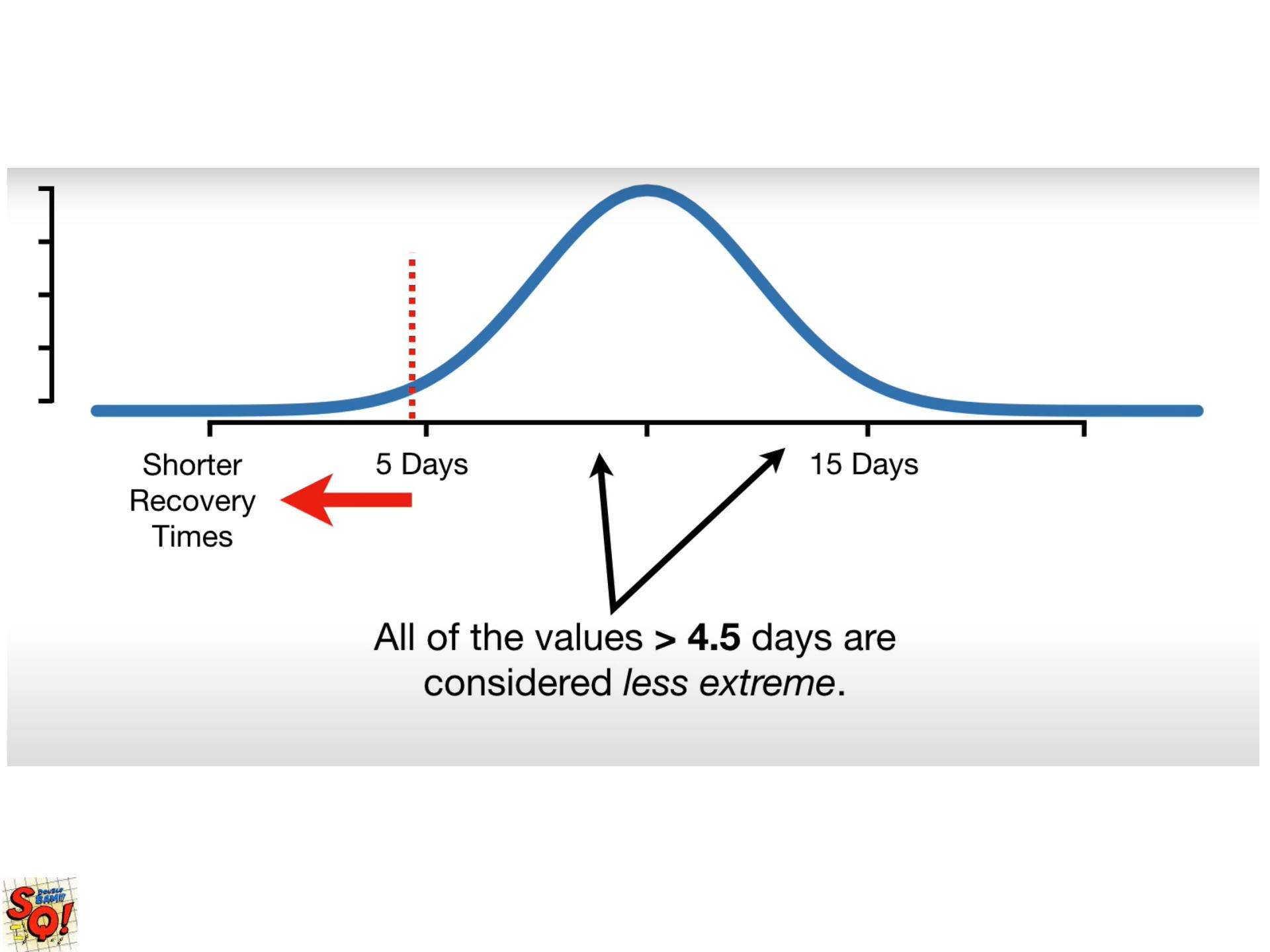


Shorter
Recovery
Times

5 Days

15 Days

Because we want to see change in
this direction, the only **more**
extreme values are < 4.5 days.

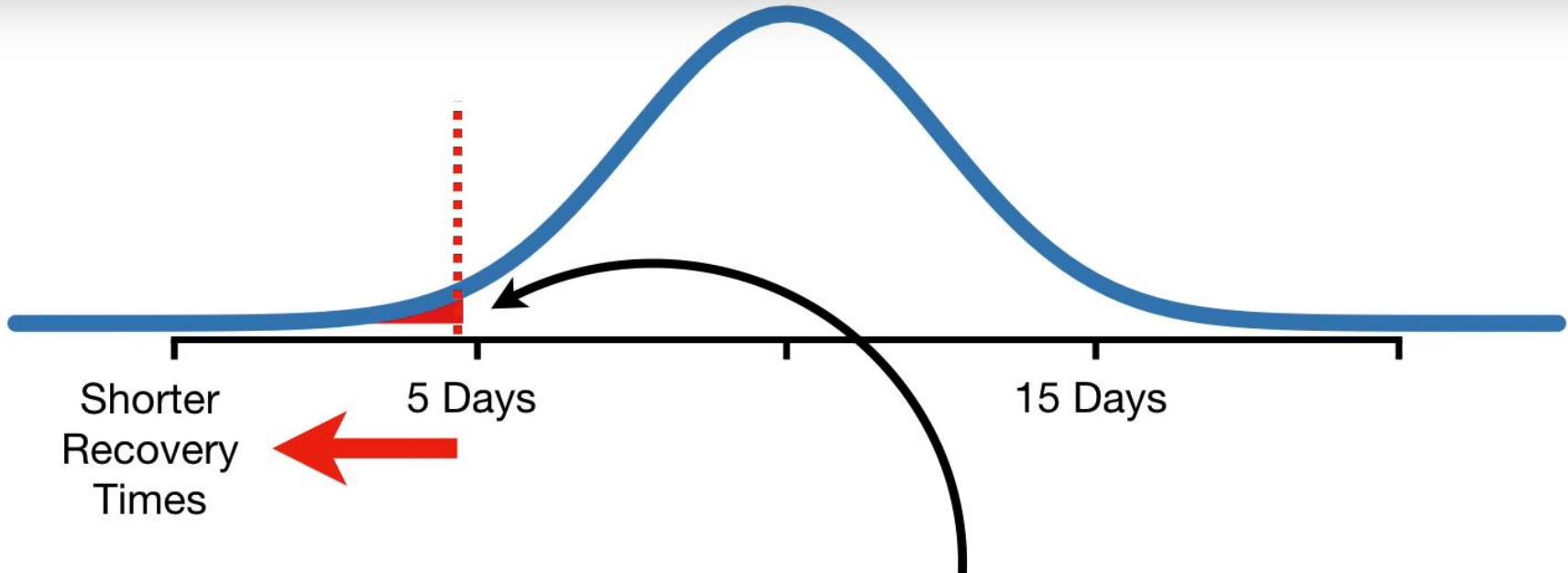


Shorter Recovery Times

5 Days

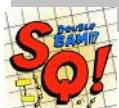
15 Days

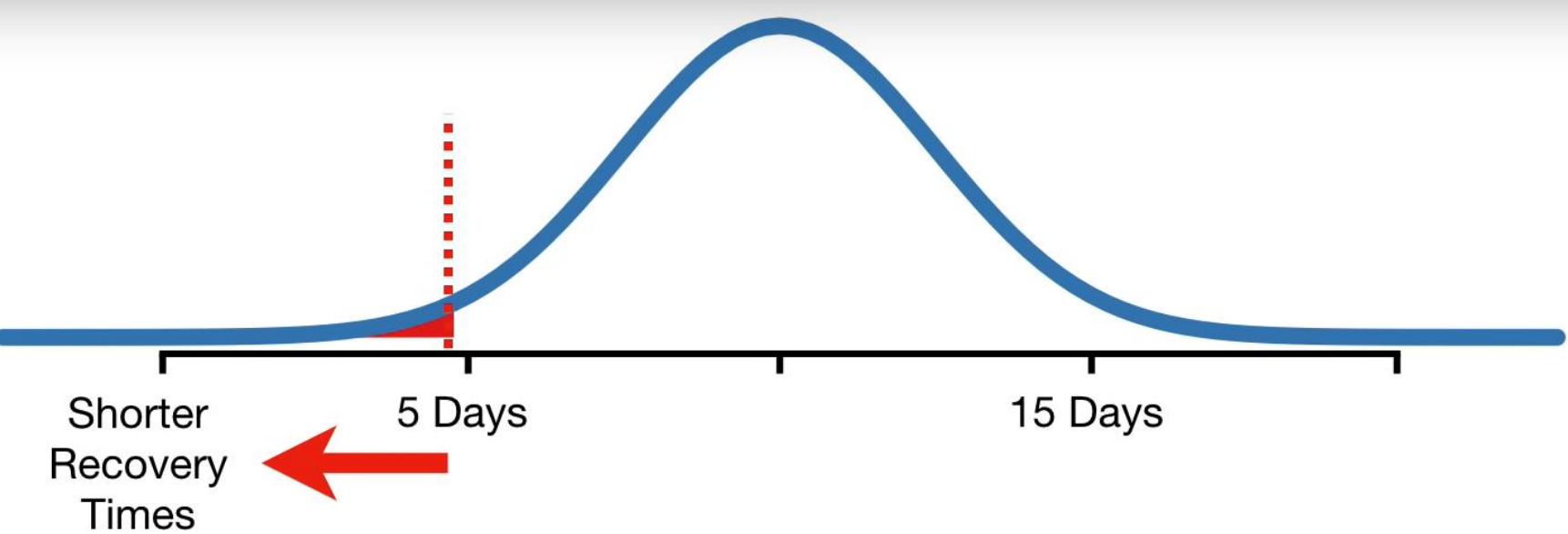
All of the values **> 4.5 days** are considered *less extreme*.



So, when we calculate a **One-Sided p-value**,
we only use the area that is in the direction
we want to see change, **0.016**.

One-Sided p-value for 4.5 days = 0.016

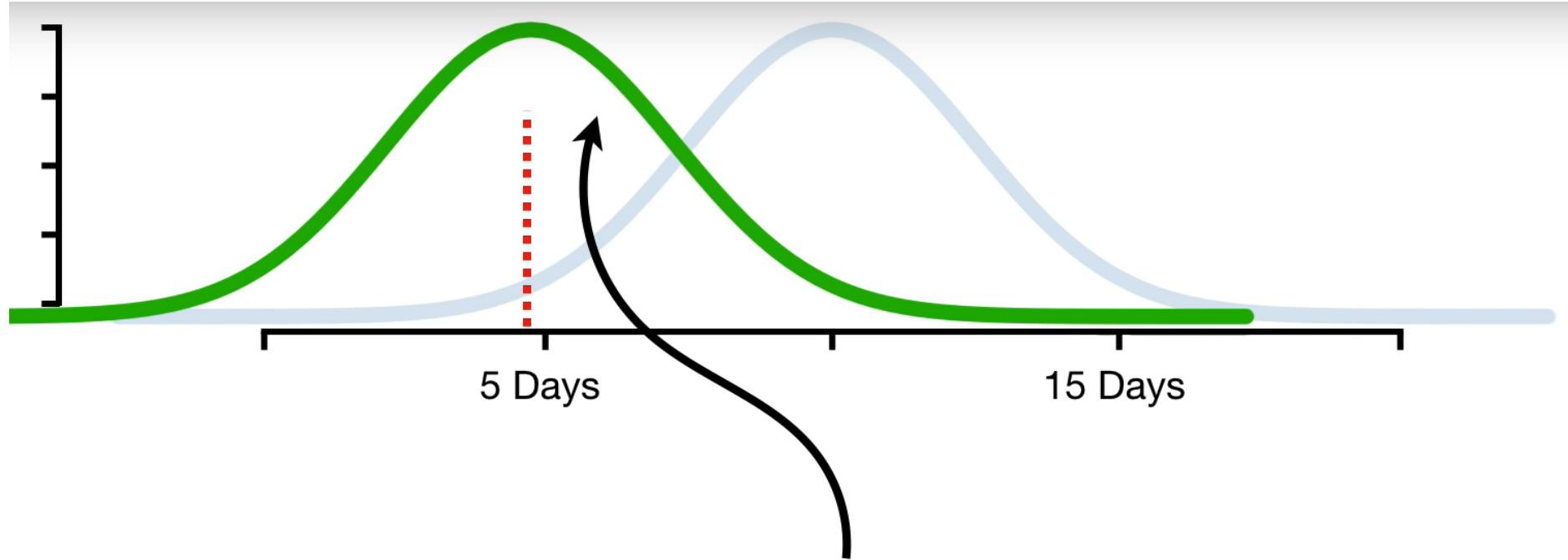




Again, since $0.016 < 0.05$, the **One-Sided p-value** would tell us that, given this distribution, **SuperDrug did something unusual...**

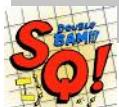
One-Sided p-value for 4.5 days = 0.016

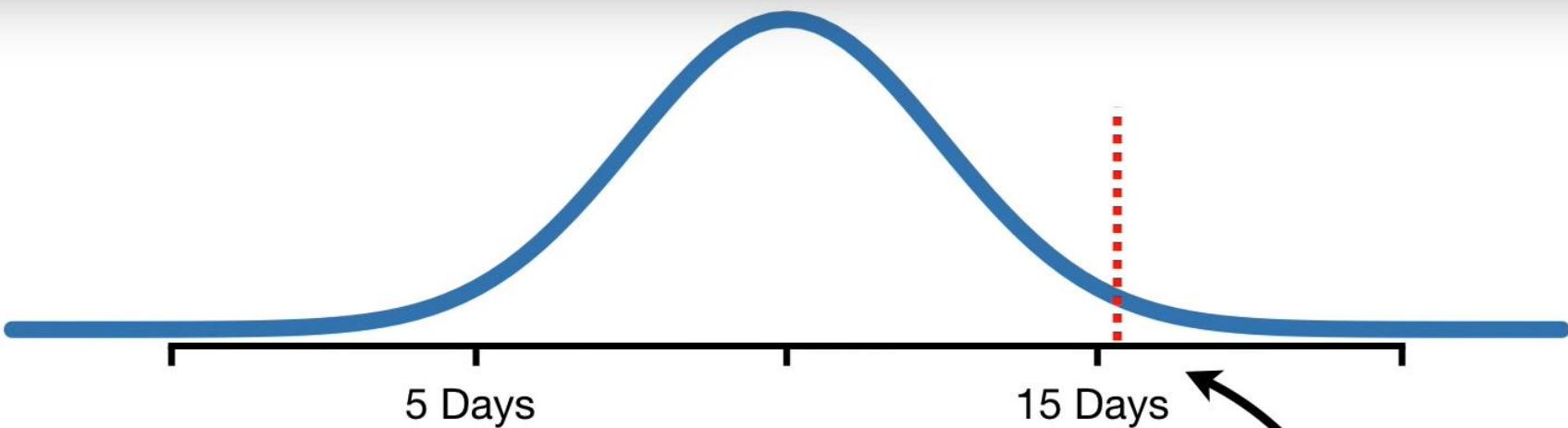




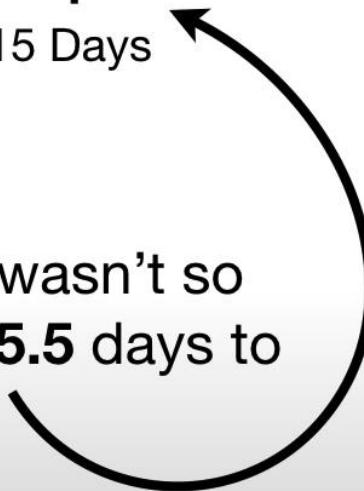
...and that some other distribution
makes more sense.

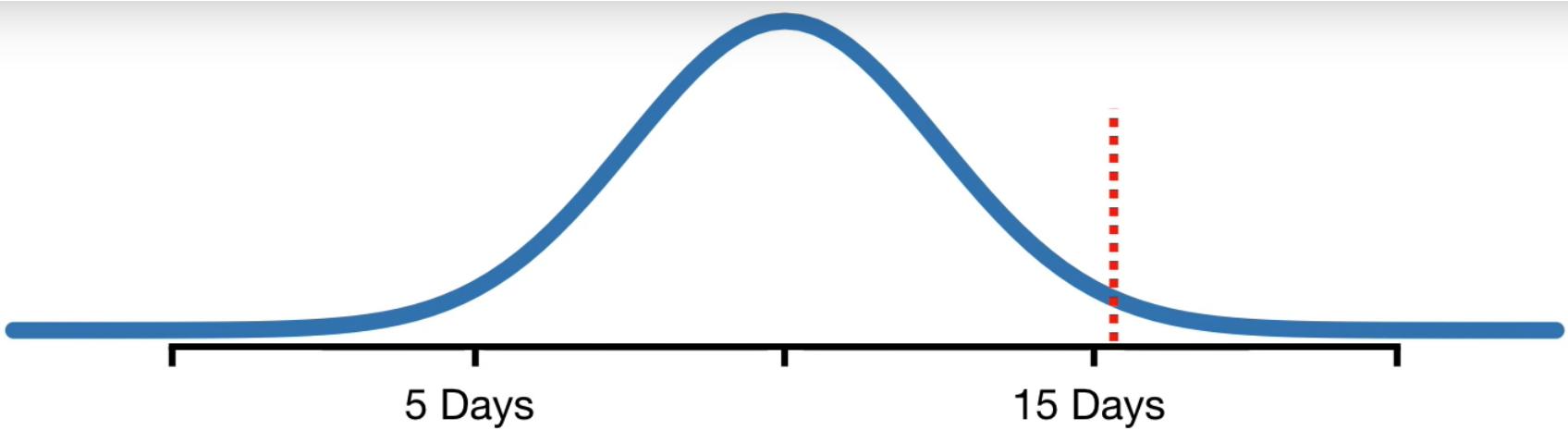
One-Sided p-value for 4.5 days = 0.016





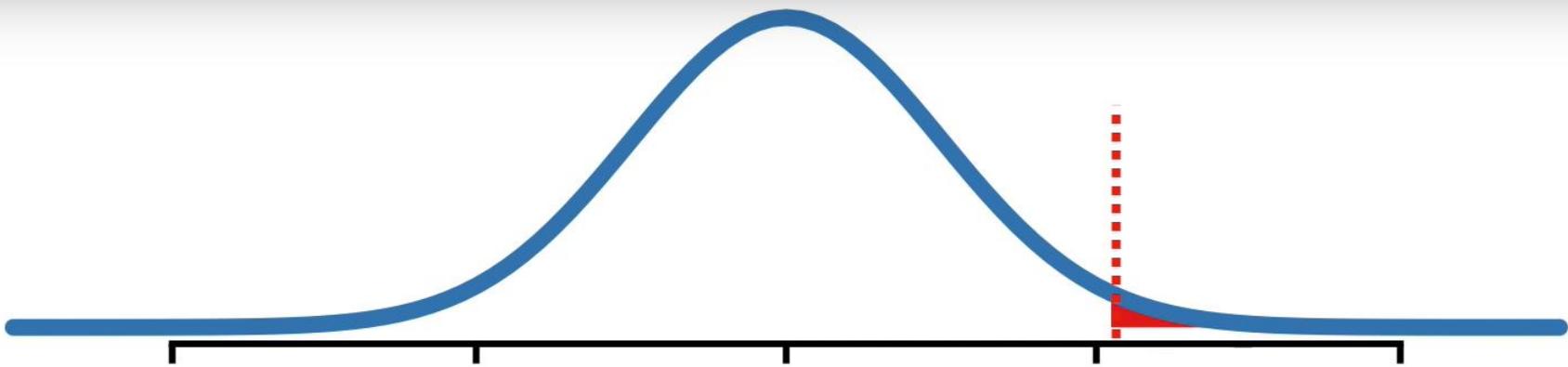
Now, imagine that **SuperDrug** wasn't so super, and, on average, it took **15.5** days to recover.





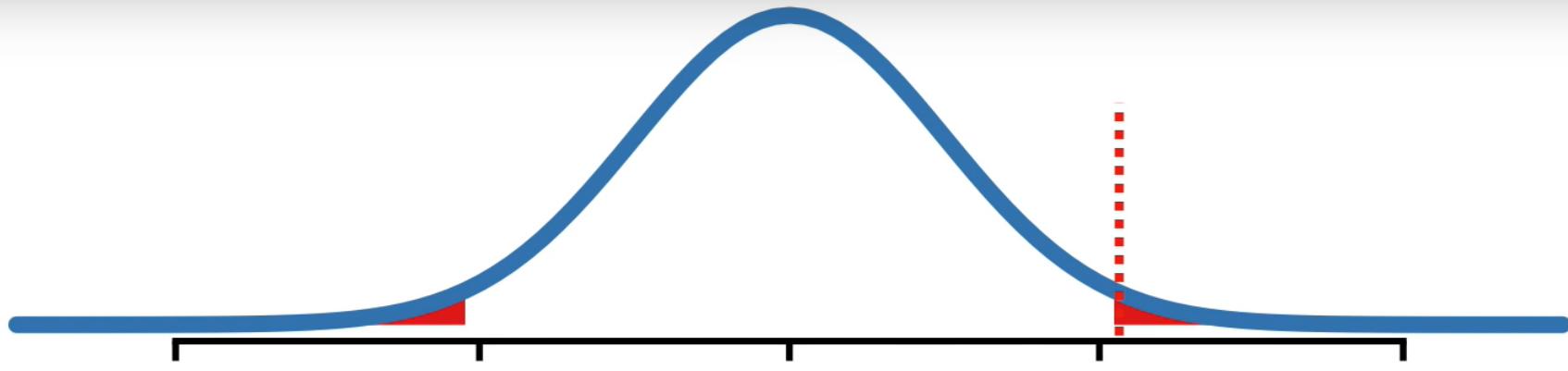
Just like before, the **Two-Sided p-value**
would be...





...the sum of *this* area under
the curve, **0.016...**

Two-Sided p-value for 15.5 days = 0.016

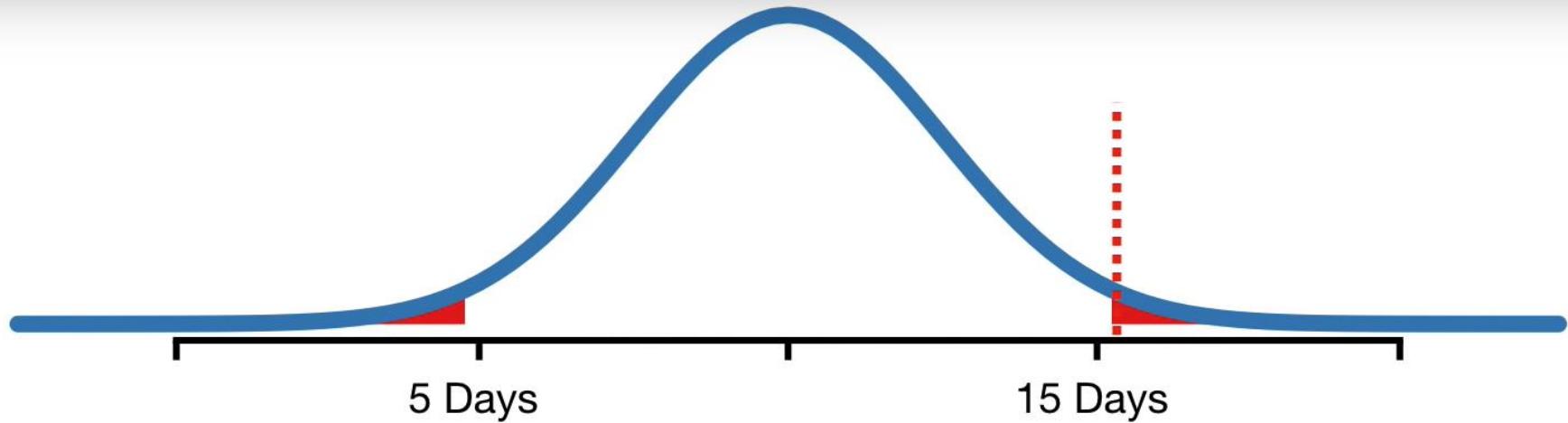


5 Days

15 Days

...plus *this* area under
the curve, 0.016...

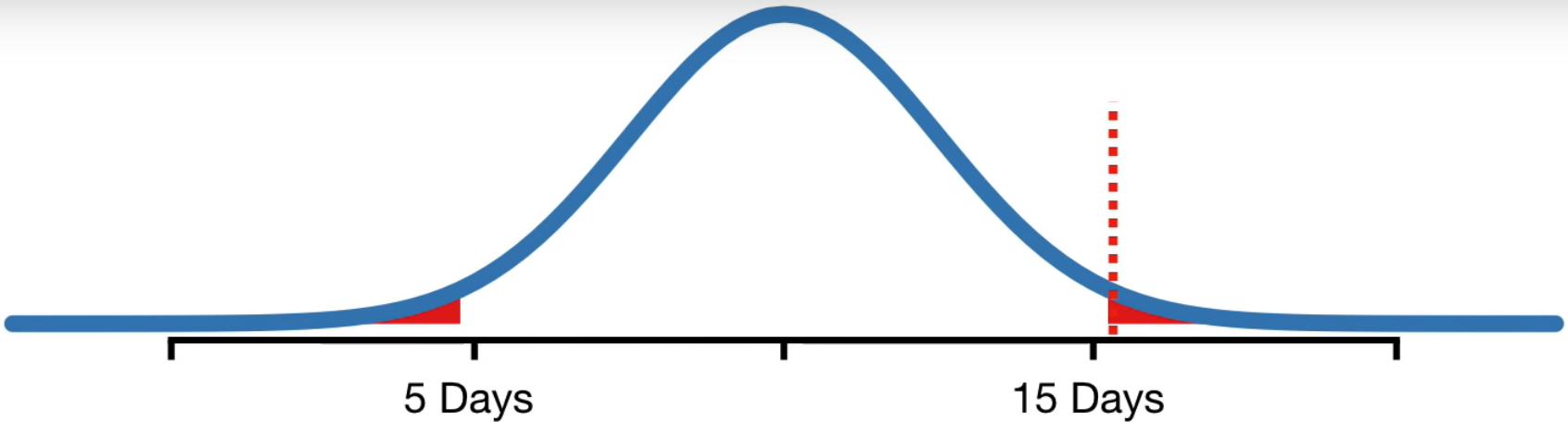
Two-Sided p-value for 15.5 days = $0.016 + 0.016$



...and the total is **0.03**.

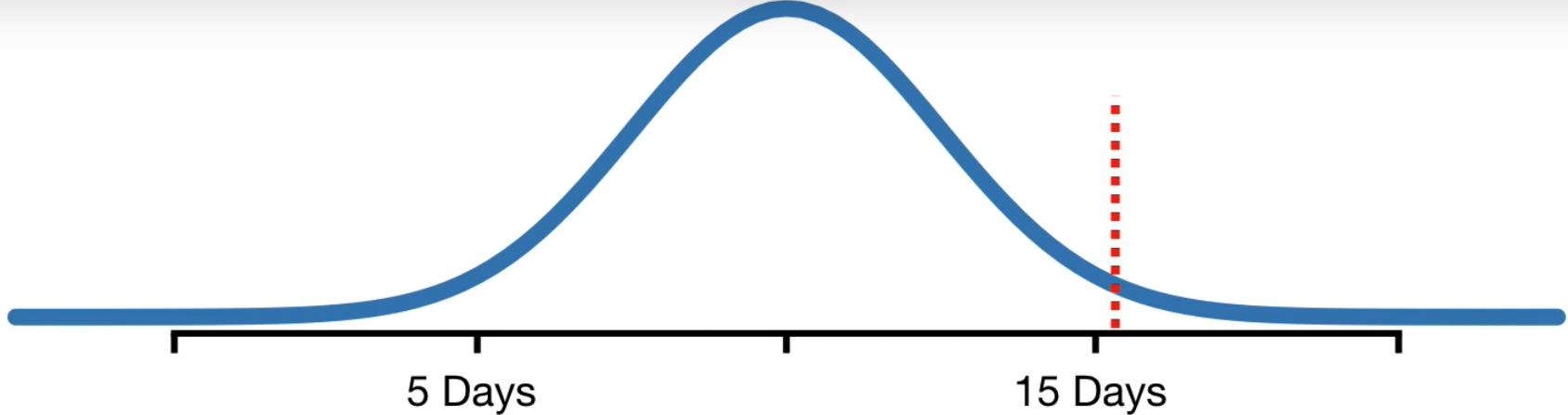
Two-Sided p-value for **15.5** days = $0.016 + 0.016 = 0.03$





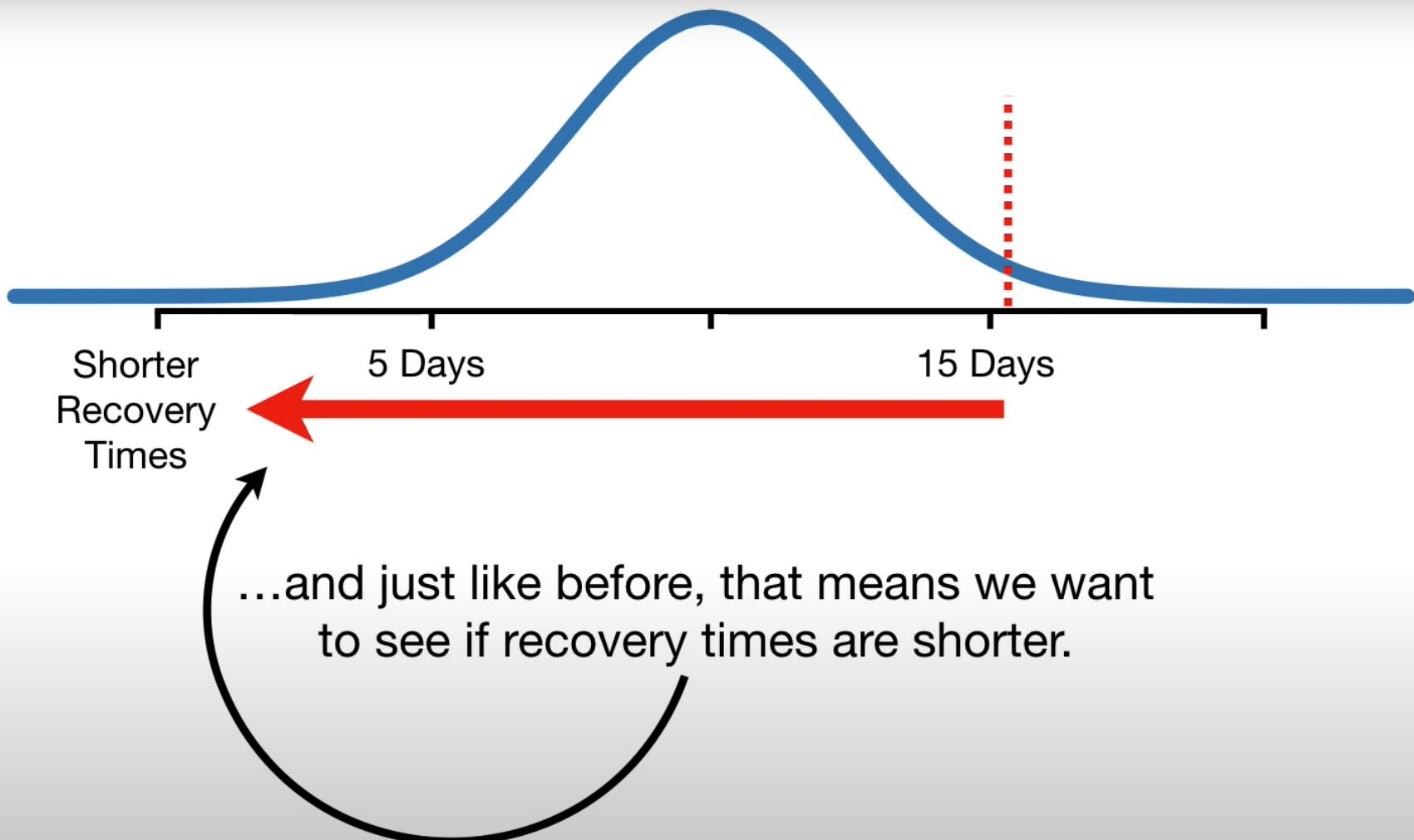
In other words, regardless of whether **SuperDrug** is super and makes things better, or if it is not so super and makes things worse, a **Two-Sided p-value** will detect something unusual happened.

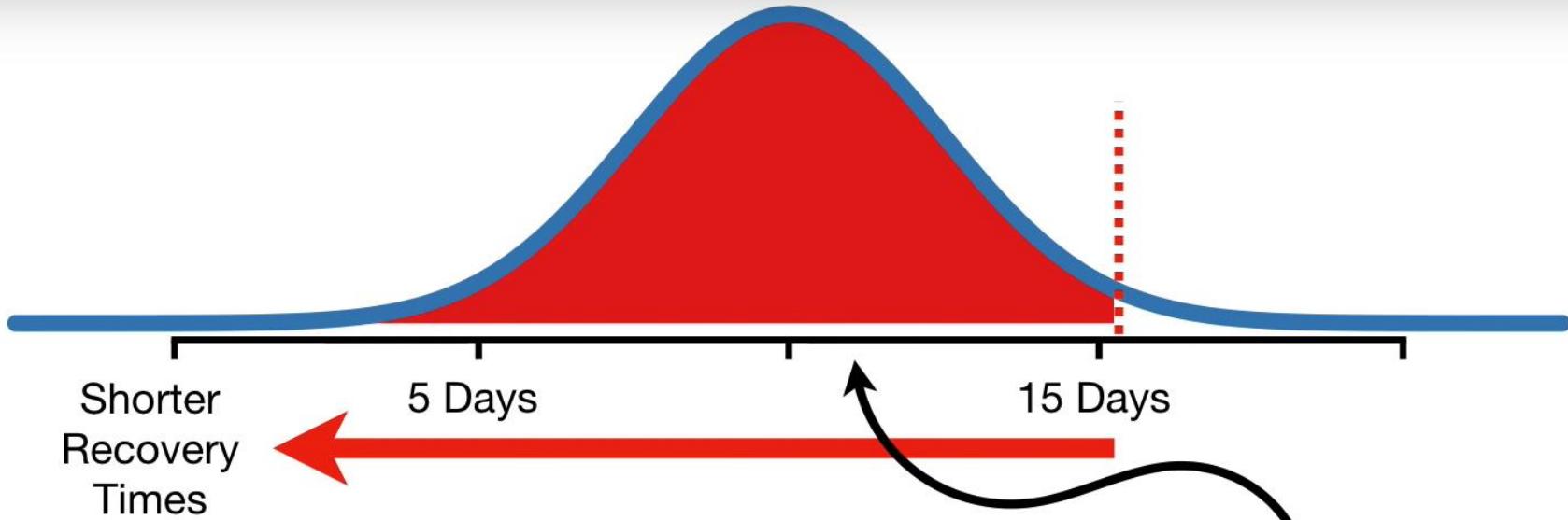




For a **One-Sided p-value**, the first thing we do is decide which direction we want to see change in.

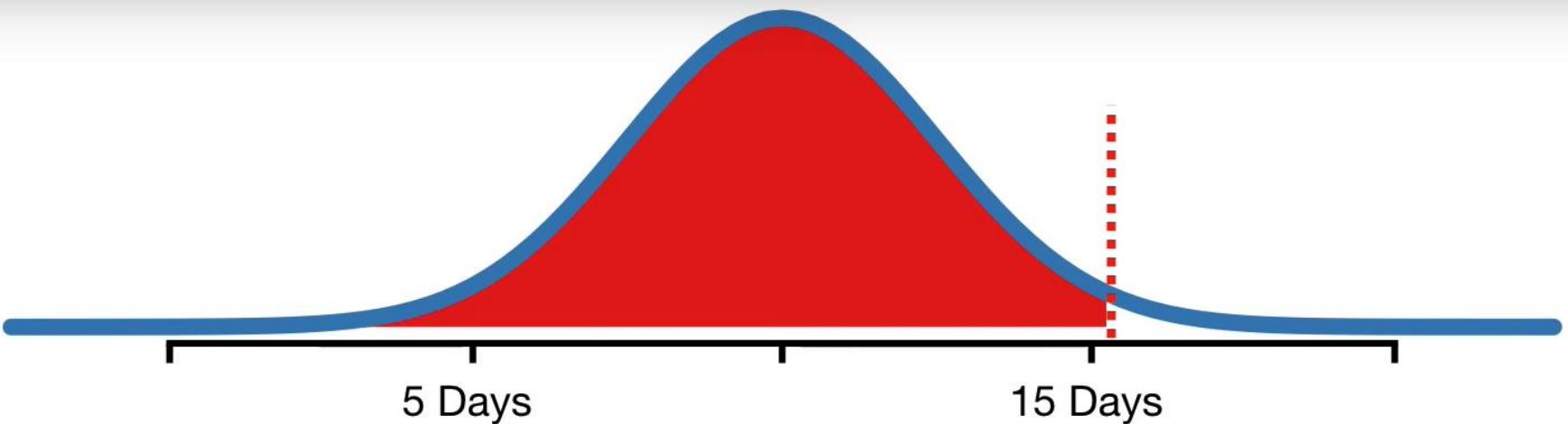






So the **One-Sided p-value** is this huge area, **0.98**, because it is **more extreme** in the direction we want to see change.

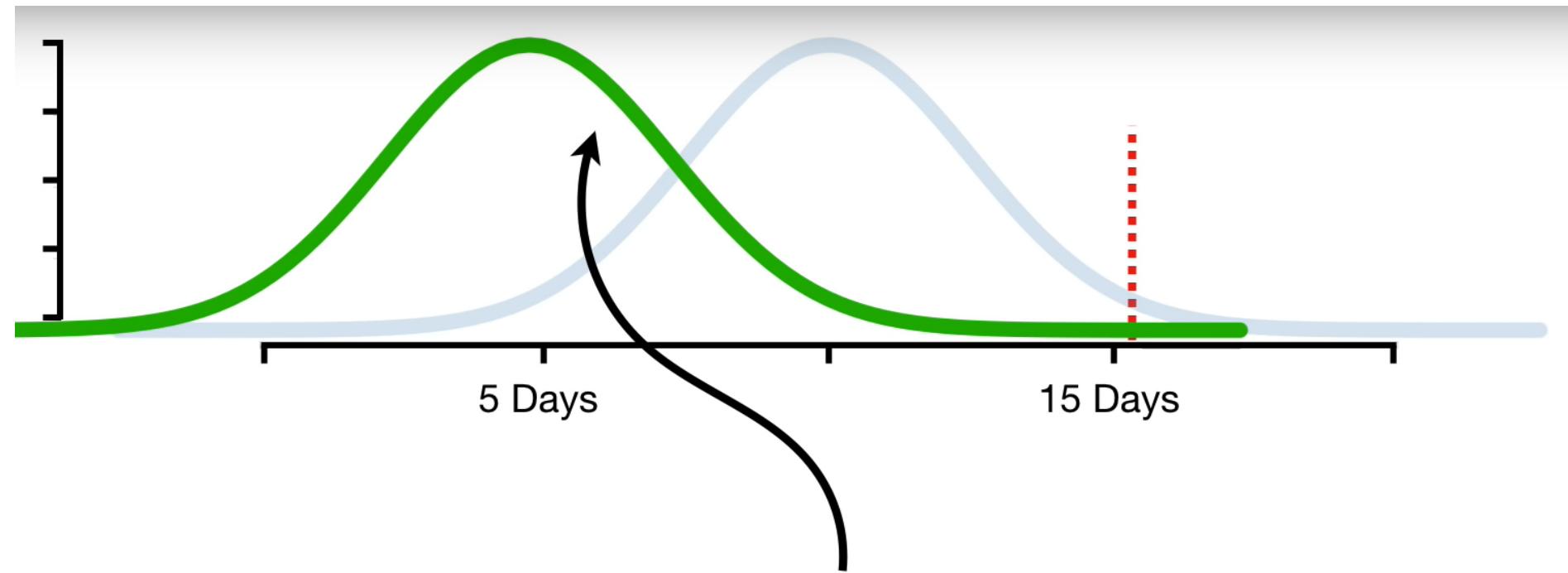
One-Sided p-value for 15.5 days = 0.98



And since **0.98 > 0.05**, the **One-Sided p-value** would not detect that **SuperDrug** was doing anything unusual.

One-Sided p-value for 15.5 days = 0.98

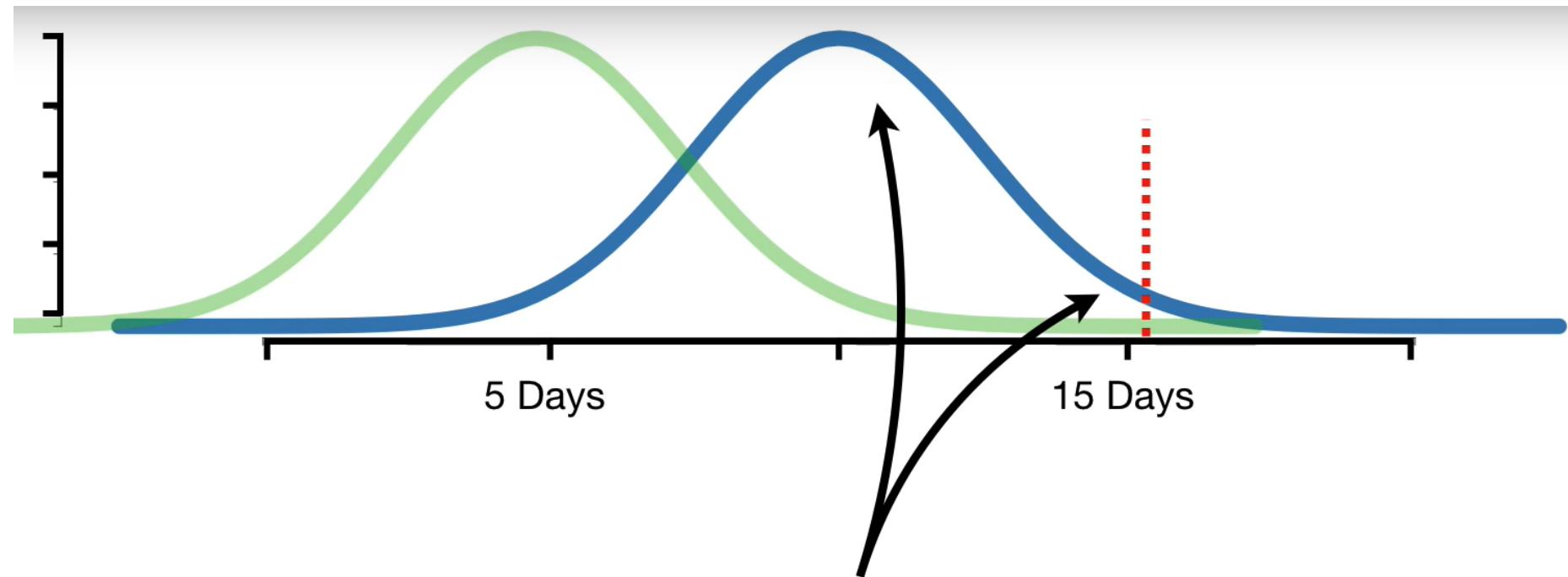




In other words, the **One-Sided p-value** is only looking to see if a distribution to the left of the original mean makes more sense...

One-Sided p-value for 15.5 days = 0.98

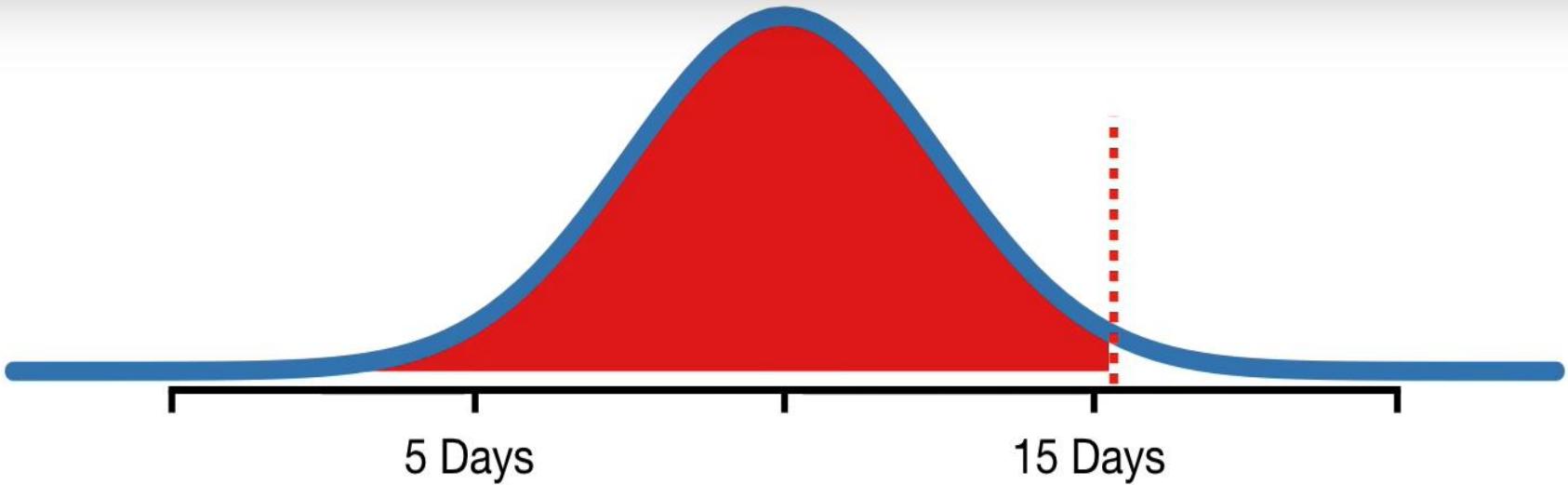




...and since the observation is on the right side of the mean, we fail to reject the hypothesis that the original distribution makes sense.

One-Sided p-value for 15.5 days = 0.98





Since failing to detect that **SuperDrug** is making things worse would be bad, **One-Sided p-values** are tricky and should be avoided....*or you should really know what you are doing.*



Example for a single mean

- **Problem 1:**

We believe that the average “normal” weight of ravens in Yukon is 1kg. In a sample of 10 ravens, the average bird weight is 0.8 kg, and the sample standard deviation was 0.15 kg. At 0.05 significance level, can we reject the null hypothesis that the mean raven weight does not differ from the “normal” weight?

script 2_hypothesistests.R

Example for a single mean

- **Solution**

$$H_0: \mu = 1$$

$$H_1: \mu \neq 1$$

we can use a t-test. (why not a z-test?)

Mean: $0.8 = \bar{y}$

sample standard deviation: $s = 0.15$

Standard Error of the Mean:

$$SEM = s_{\bar{y}} = \frac{s}{\sqrt{n}} = \frac{0.15}{\sqrt{10}} = 0.0474$$

Example for a single mean

Test -statistic: t-distribution

$$t = \frac{\bar{y} - \mu_0}{\sigma_{\bar{y}}} = \frac{0,8 - 1}{0,0474} = -4,2$$

Under H0: this will follow a t-distribution with df = n-1

The idea is:

The sample means are normally distributed with

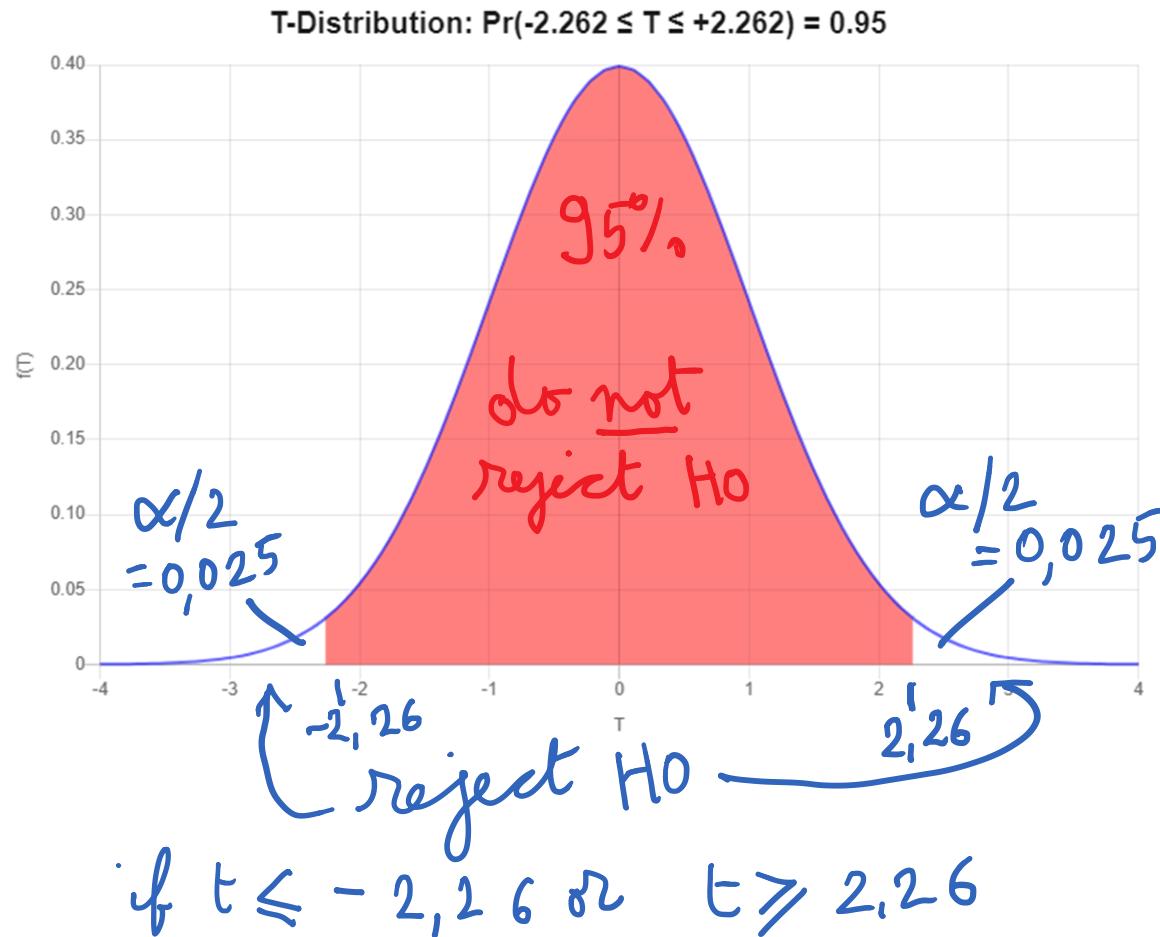
- *mean = population mean (Under H0: population mean = 1)*
- *sd = standard error of the mean*

The standardized sample means (= the number of standard deviations that the sample means are away from the population mean) will follow a z-distribution or, if the population variance is unknown, a t-distribution

For what values of t will we reject H0?

Find the critical t-value for $\alpha=0.05$
from t-table and compare:

$$t_{n-1, 1-\frac{\alpha}{2}} = t_{9, 0.975} = 2.262$$



Example for a single mean

Conclude:

$|t\text{-statistic}| > \text{critical value}$

→ it is not a usual value

The test statistic -4.2 is not between the critical values -2.262, and 2.262. Hence, at 0.05 significance level, we **reject the null hypothesis** that the mean bird weight does not differ from the assumed “normal” weight of 1 kg.

We reject H₀ with 5% chance that it was true (type I error)

R code

```
> samplemean <- 0.8# sample mean
> mu0 <- 1           # hypothesized population mean
> s <- 0.15          # sample standard deviation
> n <- 10            # sample size
> sem <- s/sqrt(n)
> sem    #standard error of the mean
[1] 0.04743416
> t <- (samplemean-mu0)/sem
> t                  # test statistic
[1] -4.21637
> alpha <- 0.05
> t.half.alpha <- qt(1-alpha/2, df=n-1)
> c(-t.half.alpha, t.half.alpha) #critical values
[1] -2.262157 2.262157
```

P-value

Alternatively:

Find the **probability** of getting this test statistic -4.2 :

In the table:

look across the row with $df = 9$: you see that the absolute value of the test statistic is higher than 3.25, corresponding to alpha 0.01 in a two-tailed test. Hence, P-value < 0.01.

We reject the null hypothesis that $\mu = 1$.

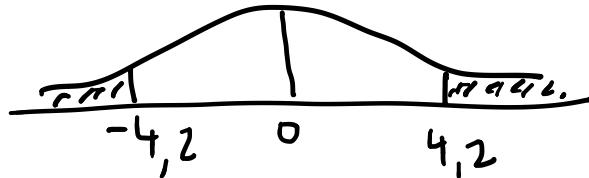
P-value

Alternatively:

Find the probability of getting this test statistic -4.2

OR something else, equally extreme

OR something else more extreme :



In R:

We apply the `pt()` function to compute the two-tailed **p-value** of the test statistic. We need to double the *lower* tail p-value to know what is the chance of getting such extreme value. Since the p-value turns out to be smaller than the 0.05 significance level, we reject the null hypothesis that $\mu = 1$.

```
> pval <- 2 * pt(-4.2, df = 9) # lower tail  
> pval # two-tailed p-value  
[1] 0.002306693
```

$$H_0: \mu = 1$$

$$H_1: \mu \neq 1$$

$$\alpha = 0,05$$

SEM

$$t = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$s = 0,15 \quad n = 10$$

$$\frac{s}{\sqrt{n}} = 0,0474$$

s/\sqrt{n} SEM

reject H_0 if $t \leq -2,262$ or $t \geq 2,262$

| **For what values of \bar{Y} will we reject H_0 ?**

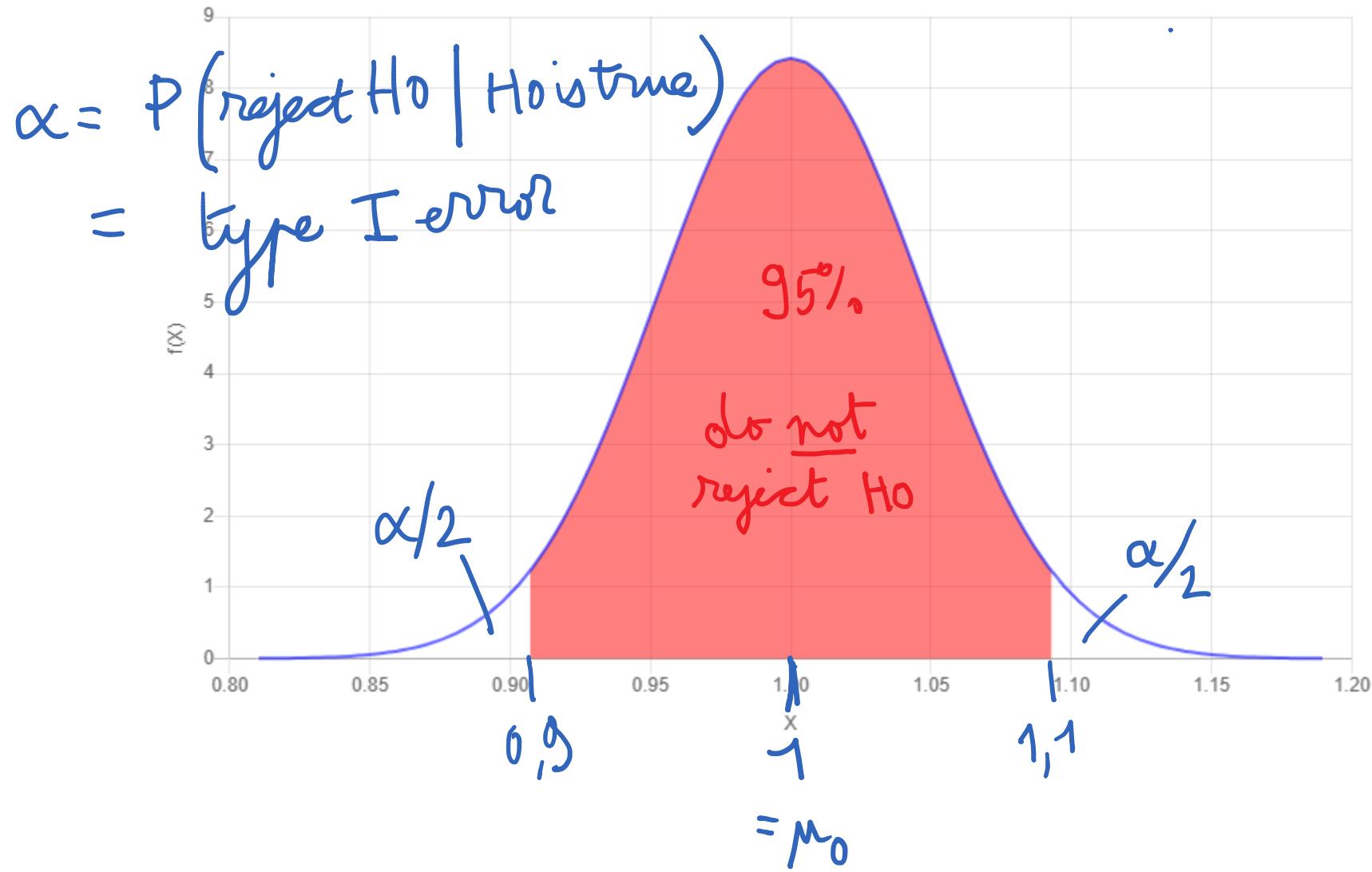
$$\bar{y} = \mu_0 + \frac{s}{\sqrt{n}} t \quad t_{n-1, 1-\frac{\alpha}{2}} = t_{9, 0.975} = 2.262$$

$$\text{upper limit of } \bar{y} = 1 + (0,0474 \times 2,262) = 1,1$$

$$\text{lower limit of } \bar{y} = 1 - (0,0474 \times 2,262) = 0,9$$

reject H_0 if $\bar{y} \geq 1,1$ or $\bar{y} \leq 0,9$

Normal Distribution: $\Pr(0.9070977 < X < 1.092902) = 0.95$



Calculating the probability of committing a type II error = β

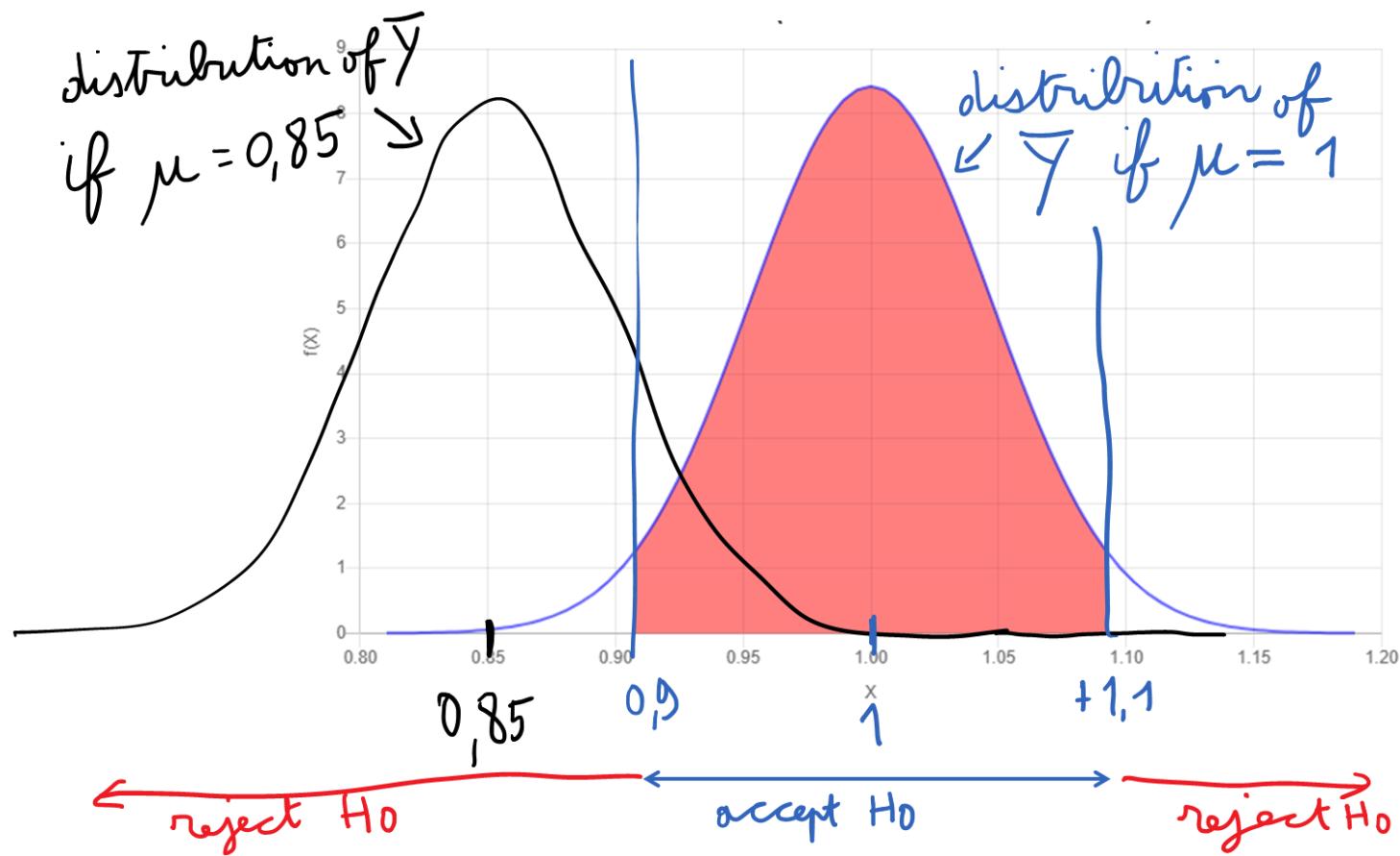
- If the sample of 10 ravens with sample mean 0.8 kg came from *another population*, of which the *true mean weight is 0.85*, what is then the probability that we commit a type II error?

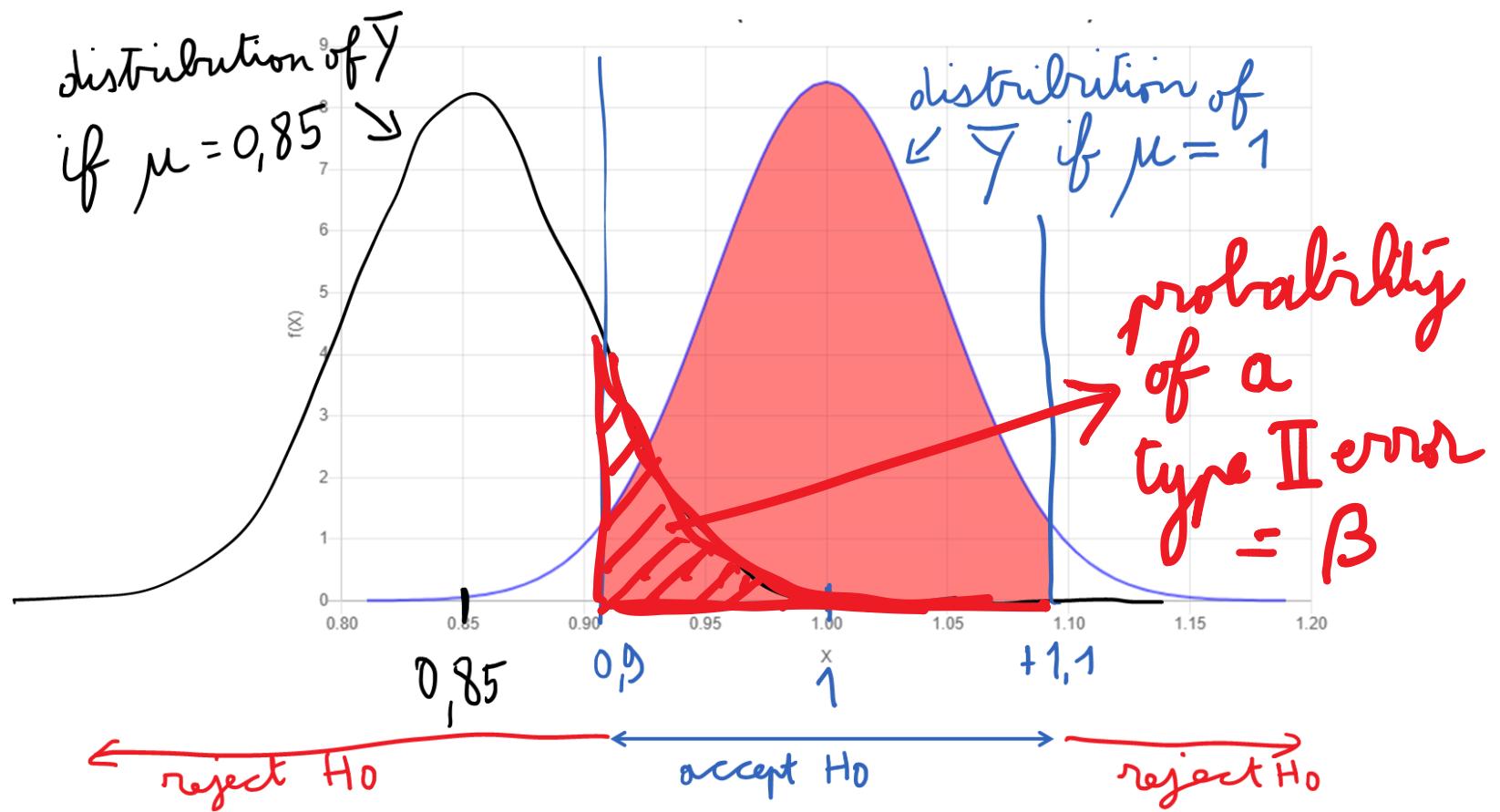
$$H_0 : \mu = 1$$

$$H_1 : \mu \neq 1$$

$$\beta = P(\text{Do not reject } H_0 \mid \mu = 0,85)$$

$$= P(0,9 < \bar{y} < 1,1 \mid \mu = 0,85)$$





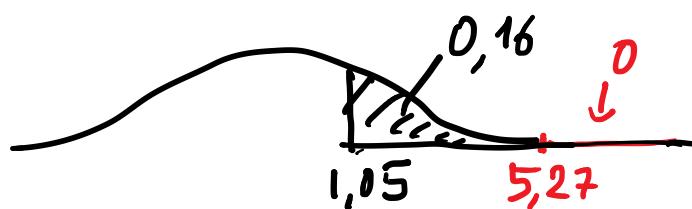
$$\beta = P(\bar{Y} > 0,9 \mid \mu = 0,85) - P(\bar{Y} > 1,1 \mid \mu = 0,85)$$

$$\begin{aligned}
 \beta &= P(\bar{Y} > 0,9 | \mu = 0,85) - P(\bar{Y} > 1,1 | \mu = 0,85) \\
 &= P\left(t > \frac{0,9 - 0,85}{0,0474}\right) - P\left(t > \frac{1,1 - 0,85}{0,0474}\right) \\
 &= P(t > 1,05) - P(t > 5,27)
 \end{aligned}$$

```

> pt(q=1.05, df=9, lower.tail=FALSE)
[1] 0.160547
> pt(q=5.27, df=9, lower.tail=FALSE)
[1] 0.0002569349

```

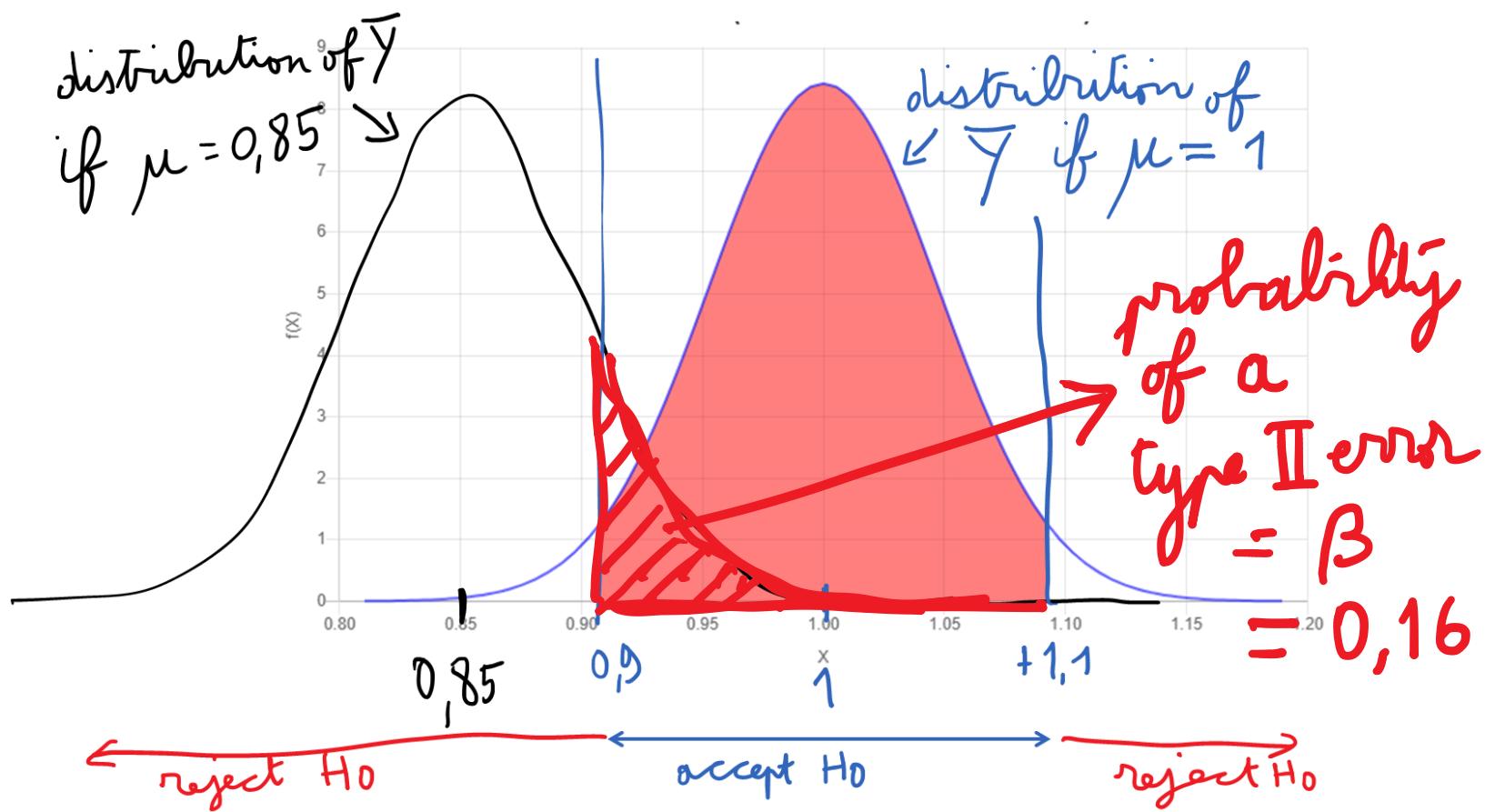


$$= 0,16 - 0,00$$

$= 0,16$ = probability of Type II error

$$\text{Power} = 1 - \beta = 1 - 0,16 = 0,84$$

= probability of rejecting H_0 if H_0 is false



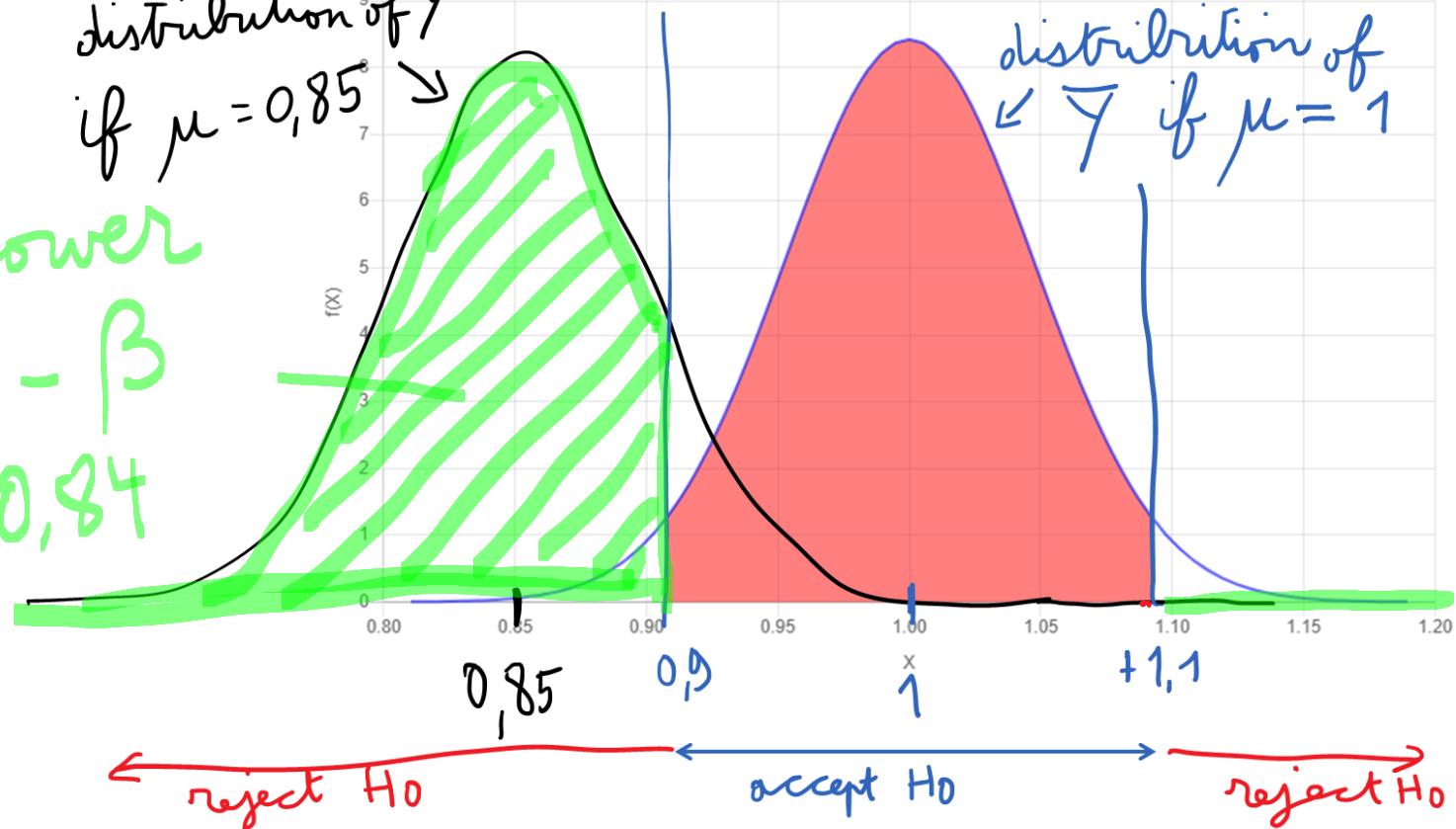
distribution of \bar{Y}
if $\mu = 0,85$

= Power

= $1 - \beta$

= 0,84

distribution of
 \bar{Y} if $\mu = 1$



- If the sample size $n = 20$ ravens, and $\mu = 0.85$, what is the power of the test?

$$H_0 : \mu = 1$$

$$H_1 : \mu \neq 1$$

$$\text{power} = P(\text{Reject } H_0 \mid \mu = 0,85)$$

$$= P(\bar{Y} < 0,9 \mid \mu = 0,85) + P(\bar{Y} > 1,1 \mid \mu = 0,85)$$

$$t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}}$$

SEM

$$s = 0,15 \quad n = 20$$

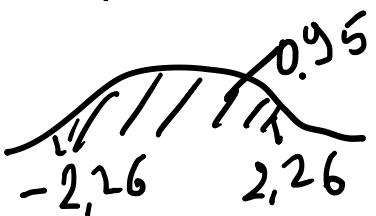
SEM

$$\frac{s}{\sqrt{n}} = \frac{0,15}{\sqrt{20}} \approx 0,03$$

$$n = 10$$

reject H_0 if

$$|t| \geq 2,26$$



$$SEM$$

$$0,05$$

$$\text{lower limit } \bar{y} = 1 - (0,05 \times 2,26) = 0,89$$

$$\text{upper limit } \bar{y} = 1 + (0,05 \times 2,26) = 1,11$$

$$\text{power} = P(\text{reject } H_0 \mid \mu = 0,85)$$

$$= P(\bar{y} < 0,89 \mid \mu = 0,85)$$

$$+ P(\bar{y} > 1,11 \mid \mu = 0,85)$$

$$= 0,84$$

$$n = 20$$

$$|t| \geq 2,09$$



$$0,03$$

$$1 - (0,03 \times 2,09) = 0,93$$

$$1 + (0,03 \times 2,09) = 1,07$$

$$= P(\bar{y} < 0,93 \mid \mu = 0,85)$$

$$+ P(\bar{y} > 1,07 \mid \mu = 0,85)$$

$$= \dots$$

$n=20$

$SEM = 0,03$

$$H_0 : \mu = 1 \quad H_1 : \mu \neq 1$$

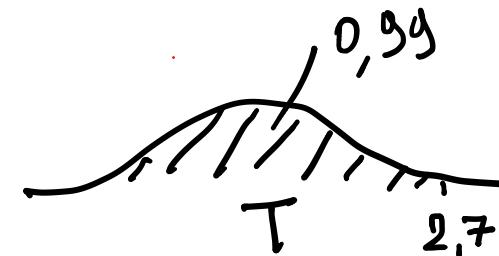
$$\begin{aligned} \text{power} &= P(\text{Reject } H_0 \mid \mu = 0,85) \\ &= P(\bar{Y} < 0,93 \mid \mu = 0,85) + P(\bar{Y} > 1,07 \mid \mu = 0,85) \\ &= P\left(T < \frac{0,93 - 0,85}{0,03}\right) + P\left(T > \frac{1,07 - 0,85}{0,03}\right) \\ &= P(T < 2,7) + P(T > 7,3) \end{aligned}$$

```
> pt(q=2.7, df=19, lower.tail=TRUE)
```

```
[1] 0.9929054
```

```
> pt(q=7.3, df=19, lower.tail=FALSE)
```

```
[1] 3.175725e-07
```



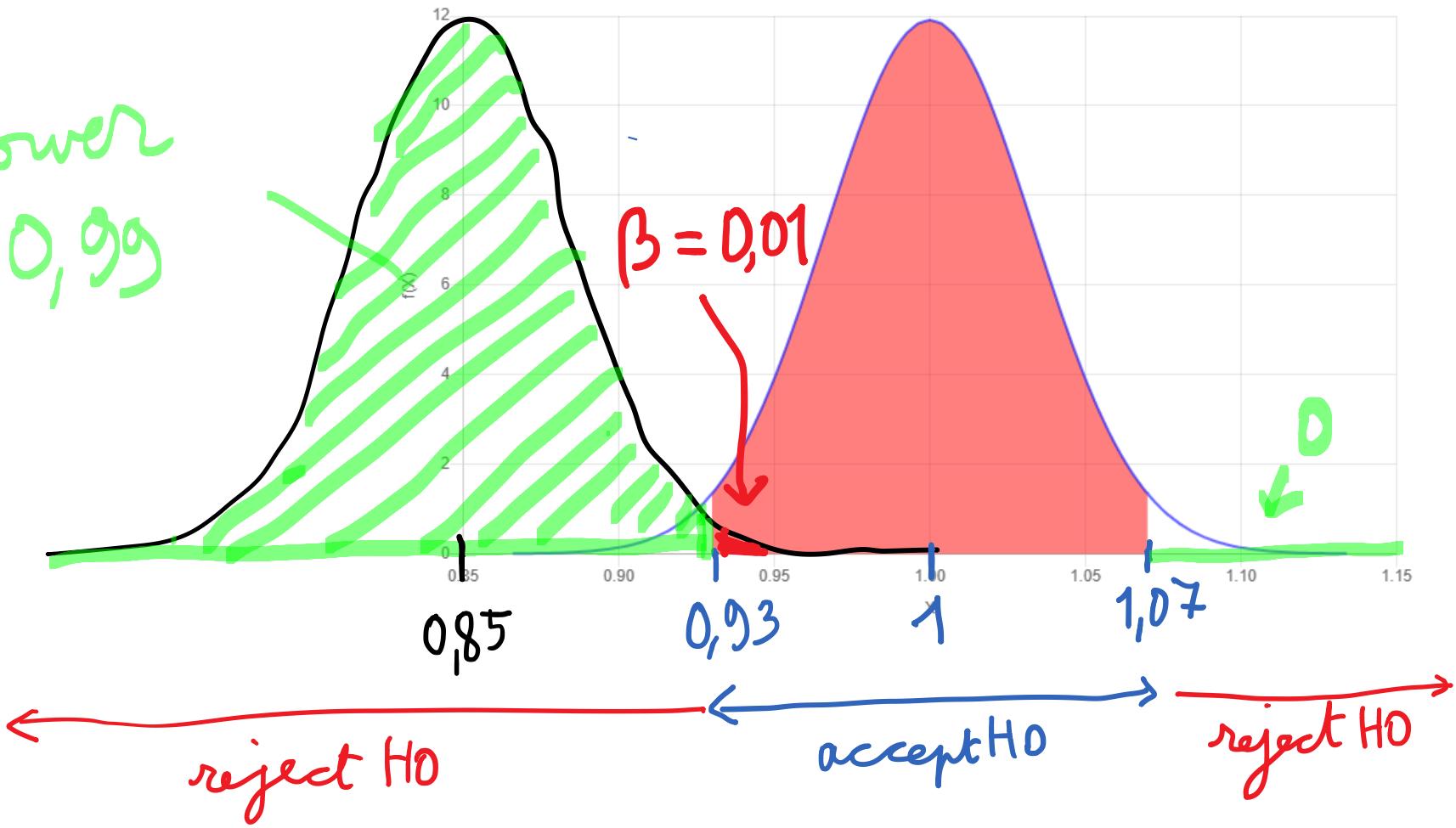
$$= 0,99 + 0 = 0,99$$

$$\boxed{n = 20}$$

distribution of \bar{y}
if $\mu = 0,85$
↓

distribution of \bar{y}
if $\mu = 1$
↓

power
 $= 0,99$



$$n = 10$$

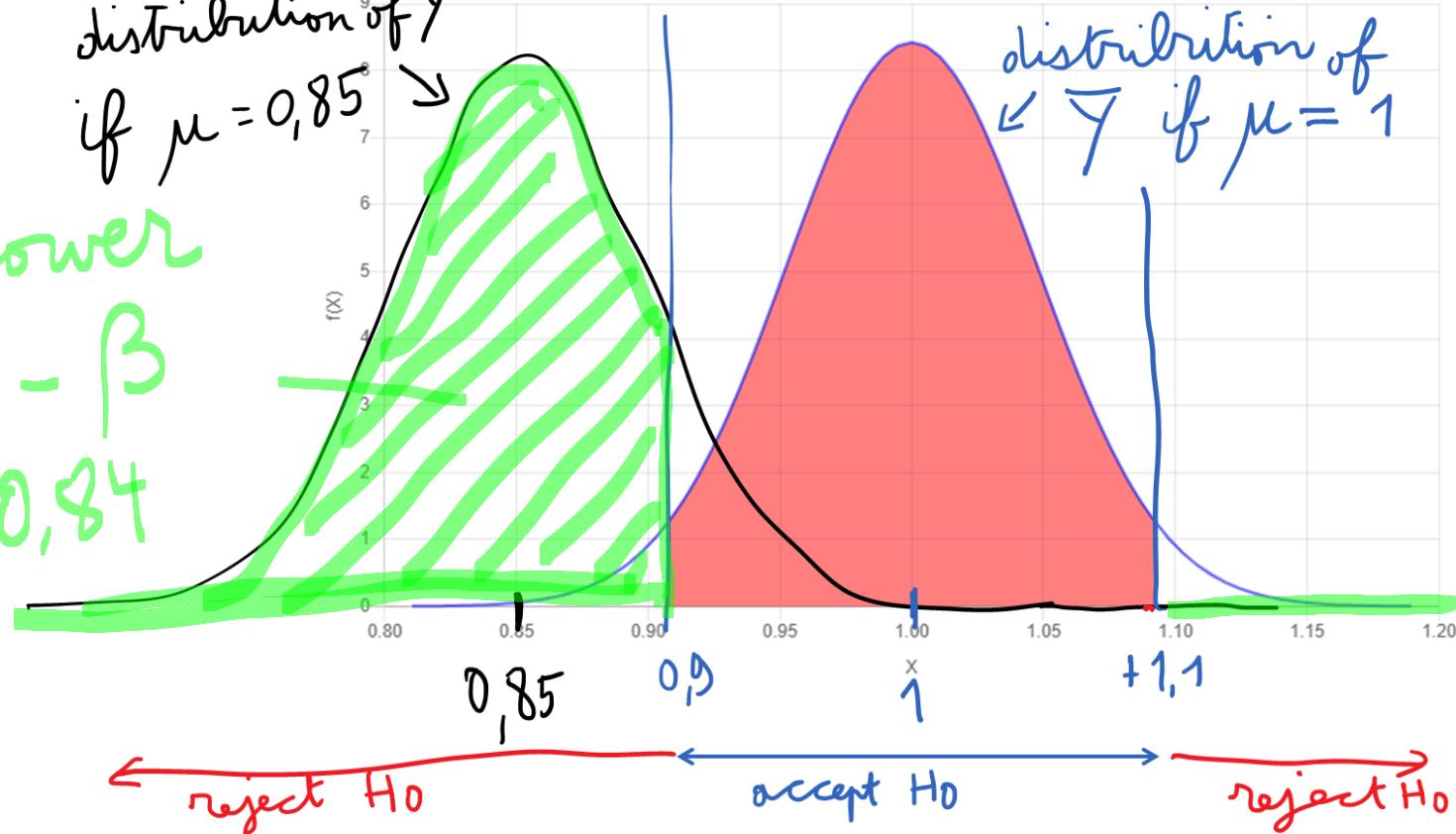
distribution of \bar{Y}
if $\mu = 0,85$

= Power

= $1 - \beta$

= 0,84

distribution of
 \bar{Y} if $\mu = 1$



Example for a single mean (2)

- If you have the sample data:

```
>t.test(sampledata, mu, alternative="two.sided")
```

- **Problem 2**

We assume that students are on average 20 years old. Using the sample of the student population in the data set survey of the package “MASS”, at 0.05 significance level, can we reject the null hypothesis that the mean student age is 20 years?

script 2_hypothesistests.R

Example for a single mean (2)

Solution:

```
> t.test(age, mu=20, alternative="two.sided")

One Sample t-test

data: age
t = 0.8905, df = 236, p-value = 0.3741
alternative hypothesis: true mean is not equal to 20
95 percent confidence interval:
 19.54600 21.20303
sample estimates:
mean of x
 20.37451
```

Example for a single mean (2)

Solution:

t.test() says: t-statistic = 0.89 p-value = 0.37

Look in the t-distribution table:

When the degrees of freedom are 29 or more, the row ∞ is used. Note that the values in this row are the same as the values for the z-distribution, since, as the sample size increases, the t-distribution approaches the z-distribution. When the sample size is 30 or more, the two distributions are considered identical

the test statistic lies between 0.674 and 1.282,
corresponding to alpha 0.50 and 0.20 in a **two-tailed test**.

Hence, $0.2 < P\text{-value} < 0.5$.

R was right.

The chance of getting this t-statistic is not extreme and
therefore we do not reject the null hypothesis

Example for comparing two means

- **Problem 3:**

We believe that the average weight of male ravens differs from female ravens.

$$H_0: \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 \neq \mu_2 \text{ or } \mu_1 - \mu_2 \neq 0$$

A sample of 20 birds is taken and each bird is weighed and released. 12 birds were males with an average weight of 1.2 kg and a standard deviation of 0.02 kg. 8 birds were females with an average weight of 0.8 and a standard deviation of 0.01 kg.

```
script 2_hypothesistests.R
```

Example for comparing two means

Test -statistic: t-distribution

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{s_{\bar{y}_1 - \bar{y}_2}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

Under H0: this will follow a t-distribution with df = (n₁ + n₂ - 2)

The idea is:

The sample means are normally distributed and thus also the difference between sample means is normally distributed, with

- *mean = population mean (Under H0: population mean = 0)*
- *sd = standard error of the difference between the means*

The standardized difference between sample means will follow a z-distribution or, if the population variance is unknown, a t-distribution

Example for comparing two means

Find the critical t-values and
compare to the t-statistic, or use
the p-value

Example for comparing two means

- In R: can split the dataframe by a factor column and do t-test on the column pairs
`t.test(numericvariable ~factor)`
- **Problem 4:**
Using the sample of the student population in the data set survey of the package “MASS”, at 0.05 significance level, can we reject the null hypothesis that the mean height of male and female students is equal?

script 2_hypothesistests.R

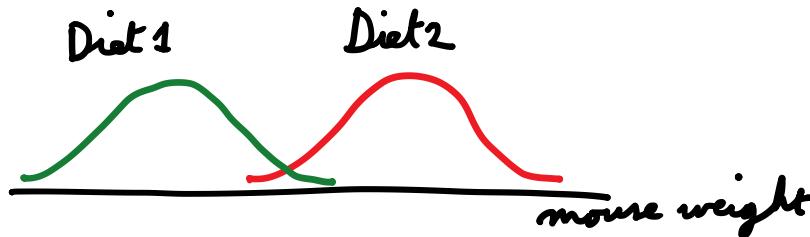
Exercise 2

- In the Survey dataset of the MASS package, is there a significant difference between males and females for the following variables: pulse, height and age.
- do all the tests together (NOT variable by variable)
- create a table representing the mean values per category (male and female), the t-statistic and the p-value for all variables

script TADE_exercise_2.R

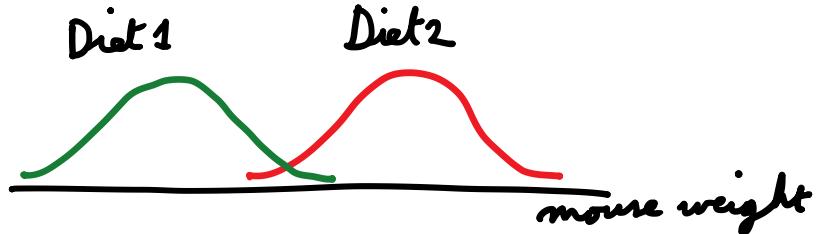
Statistical Power and Power calculations

(A)



Two distinct distributions
with small overlap

- Since there is only a little overlap, most of the mice on Diet 1 weigh less than the mice on Diet 2.
- If we collect a **small set of measurements** from Diet 1 and another small set of measurements from Diet 2
 - we get a $p\text{-value} < 0.05$
 - which causes us to correctly reject the Null hypothesis that both sets came from the same distribution
- If we repeat the experiment several times
 - There is a high probability that each statistical test will give correctly a small p-value



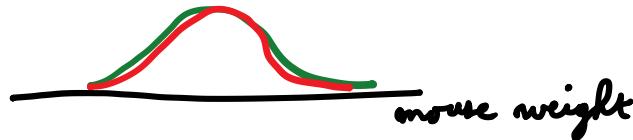
Two distinct distributions
with small overlap

- But, IF we sample all data from the overlapping zone (this will be rarely happening here)
 - we will get a p-value > 0.05
 - in that case, we cannot correctly reject the null hypothesis that all data comes from the same distribution

Power = probability of correctly getting a small p-value and rejecting H_0

- If overlap is small, Power is high

(B)



Diet 1 and Diet 2 share
the same distribution

- Here, H_0 is true
 - There is no such thing as « correctly rejecting H_0 »
 - The concept of power doesn't apply in this situation

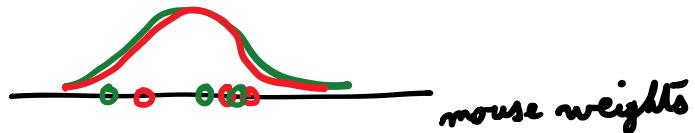
(C)



Diets make a small difference.
Two distinct distributions with large overlap

- If we collect a **small set of measurements** from Diet 1 and another small set of measurements from Diet 2
 - we **rarely** get a p-value < 0.05
 - which causes us to **correctly reject** H_0 that both sets came from the same distribution
 - we **often** get a p-value > 0.05
 - which causes us to **fail to reject** H_0 that both sets came from the same distribution
- **Large overlap between distributions AND small sample size: low Power**

p-hacking



- T-test comparing two means
- If the p-value is little above 0.05 (ex.: 0.06)
- There is a surprisingly high probability that **just adding 1 new measurement** to both groups will result in a **FALSE POSITIVE**
- *Reason: SEM decreases and T-distribution becomes narrower*
- = **p-hacking**

p-hacking: NOT recommended!

In order to avoid p-hacking, need to determine the proper sample size before doing the experiment:

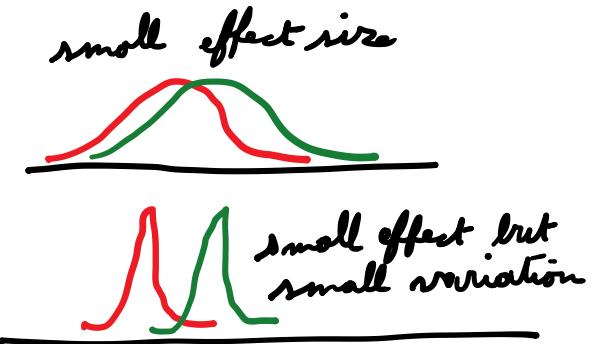
- Preliminary data: Power calculation before you do the experiment to determine how many samples you need
- If you didn't do it before: use the present data to calculate which effect size you can detect with sufficient power.

4 things affect power:

1. The effect size



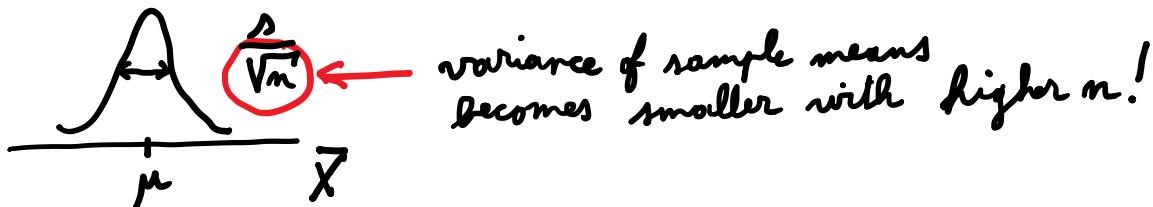
2. The variation



3. Sample size

A large sample size higheres the probability of having a good estimate (of the mean)

→ can compensate for a small effect size and high variation

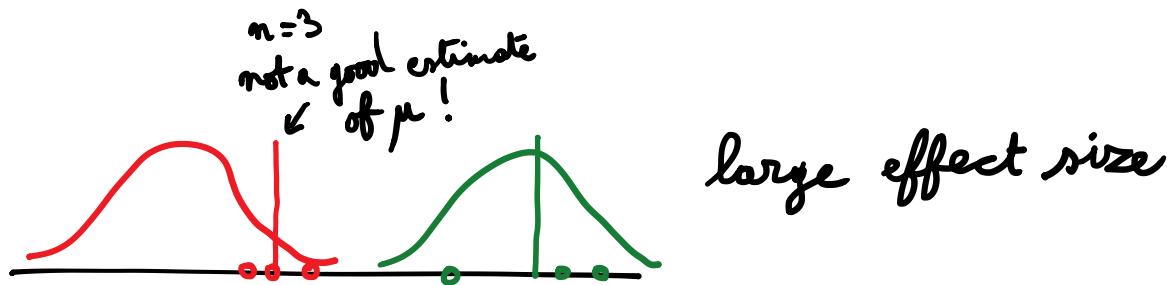


4. The statistical test you use (T-test is a surprisingly powerful test)

Power analysis

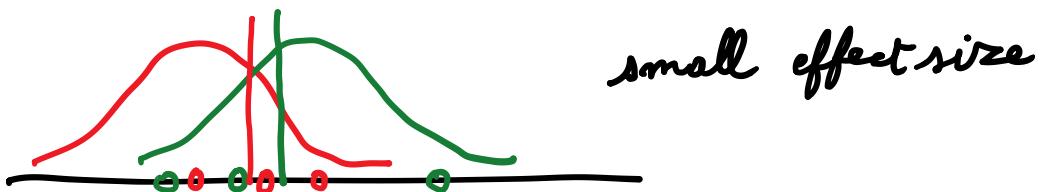
- We can increase the power by increasing the number of measurements we collect.
- Power Analysis will tell us how many measurements we need to collect to have a good amount of power

Large effect size



- With a small sample size, the 95% confidence interval for the population mean is wide
- Even with a small sample size, we have a high probability of getting a small p-value
- probability of correctly getting a small p-value = POWER

Small effect size

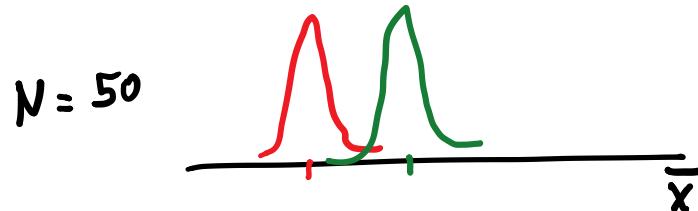
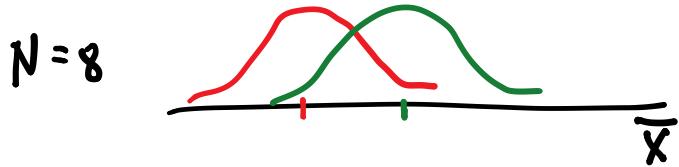
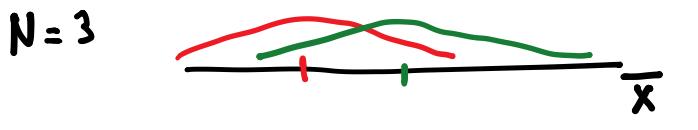


- We need to have a lot of confidence in the accuracy of the estimated mean in order to correctly detect a difference
 - ↔ narrow 95% confidence interval of the population mean
 - ↔ small SEM

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

For a constant distribution of a variable, with a given variation (standard deviation σ),

the standard deviation of the sample means (SEM) depends on the sample size:



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

The variation in the distributions determines how much power will increase when you increase the sample size

- $\sigma = \text{large} \rightarrow \text{need large } n \text{ to have a reliable estimate of the mean}$
low $\frac{\sigma}{\sqrt{n}}$
- $\underline{\sigma = \text{small}} \rightarrow \text{don't need large } n \text{ to have a reliable estimate of the mean}$
low $\frac{\sigma}{\sqrt{n}}$
- Power calculations boil down to solving for “n” in the SEM: $\frac{\sigma}{\sqrt{n}}$

Power calculation

To do Power calculation, you need two of these three variables:

1. Effect size
2. Variation in the data
3. Sample size N

a) You have preliminary data

Estimate what the distributions look like:

- Estimate means -> estimate effect size
 - Estimate Standard deviation
- > can calculate N

Power calculation

b) You have NO preliminary data

1. Start by assuming you will do 3 replicates
2. Assume you are looking for a two-fold difference
3. Calculate how much **variation** will still give you good power
4. If the variation is unrealistically small, increase sample size or increase effect size

Power calculation

Ex. RNA-seq: variation has already been worked out for human data and laboratory animal or plant species

- 1) Plug in known variation
- 2) Plug in an **effect size** you want to detect and determine **sample size**
OR Plug in **sample size** and determine which **effect size** you can detect

Tools:

<http://powerandsamplesize.com>

Mac/Windows Program: Gpower

R package: pwr