Statistical Inference

🎯 How can we guess the real value of a parameter based *only* on a limited sample of observations ?

1. Collect some observations of a parameter
2. Infer the true value of the parameter (leap of faith)
3. Estimate your level of confidence

Plan

1. Motivation
2. Probability Theory reminders
3. Sampling Distribution and Confidence Intervals
4. Hypothesis Testing (p-values)
5. t-tests
6. Bayesian Inference

1. Motivation

Recall our business problem

How to increase customer satisfaction while maintaining a healthy order volume?

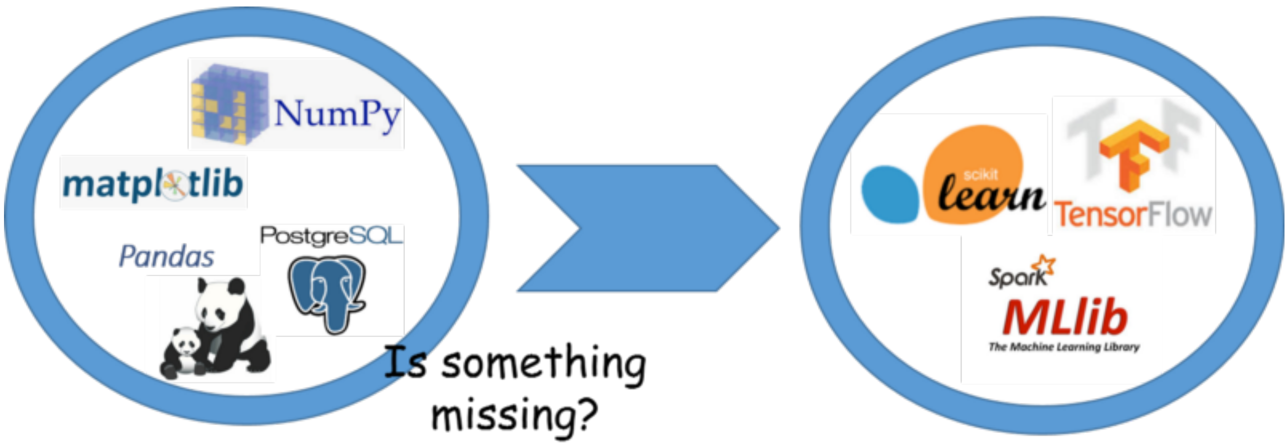
The customer satisfaction can be evaluted through the `review_score`

👉 We will investigate which features are the most impactful on `review_score`

Imagine we find `wait_time` to be strongly correlated with bad `review_score`

🤔 How can we be **confident** our findings (on historical orders) will **generalize** well ?

❌ We cannot wait years to prove that our findings were right or wrong!



Welcome Statistical Inference Analysis!

1️⃣ Train *linear ML* models to find correlations

2️⃣ Use stats (Central Limit Theorem!) to **quantify the statistical significance** of our findings

2. Probability Refreshers from the Maths module

Probability

Conditional Probability

$P(B|$

Bayes Theorem

$P(B|$

Random variable

X

= numerical outcome of a random experiment

Random process

$X = (X_k)_{0 \leq k \leq n}$

= repeated sequence of random experiments

Probability Distribution

$p(X) = p(\mu, \sigma, \dots)$

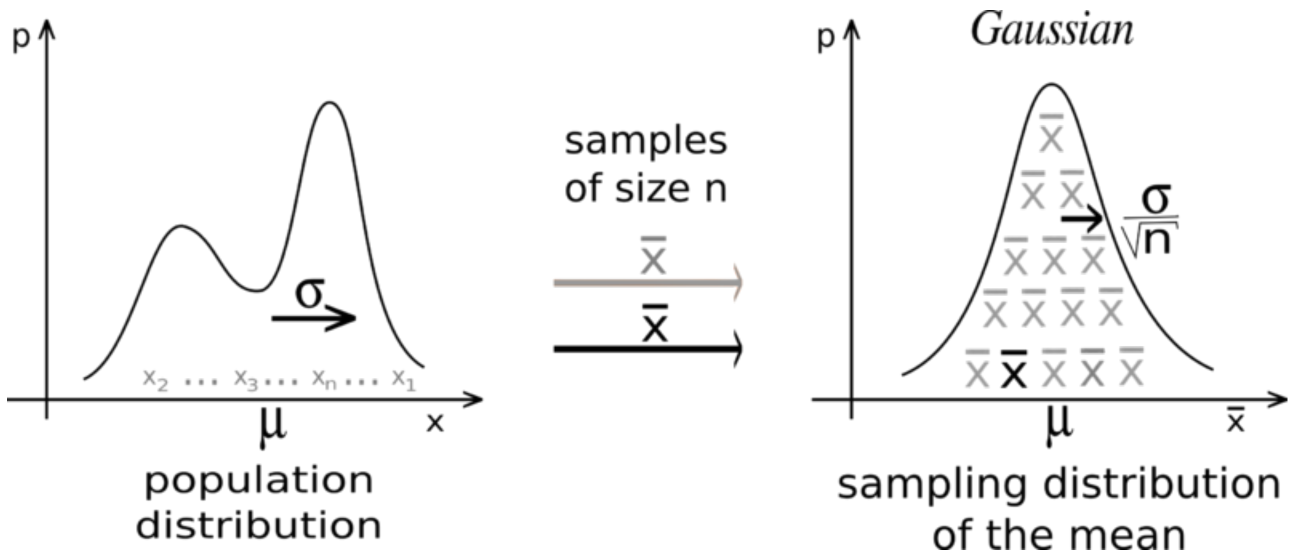
- Measures the underlying distribution of a random variable X
- The *mean* μ and *standard deviation* σ are called "statistics" that "describe" X
- Other statistics include kurtosis etc...

The Gaussian Distribution (or Normal Distribution)

$\mathcal{N}$ $($

- is completely described by these two statistics only

Central Limit Theorem



When you consider **independent random variables** $X_1 \ldots X_n$ with a **common** underlying probability distribution $p(\mu, \sigma)$:

- Their mean $\overline{X}$ converges towards a Normal Distribution as n increases:
  - centered around the common mean $\mu_{\overline{X}} = \mu$
  - with a standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

z-score

- If x is an observation derived from a random variable $X(\mu, \sigma)$, we define its `z-score` as follows:

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

- z = value of x expressed in *number of standard deviations above/below the mean* μ

$Z = ($

3. Sampling Distribution

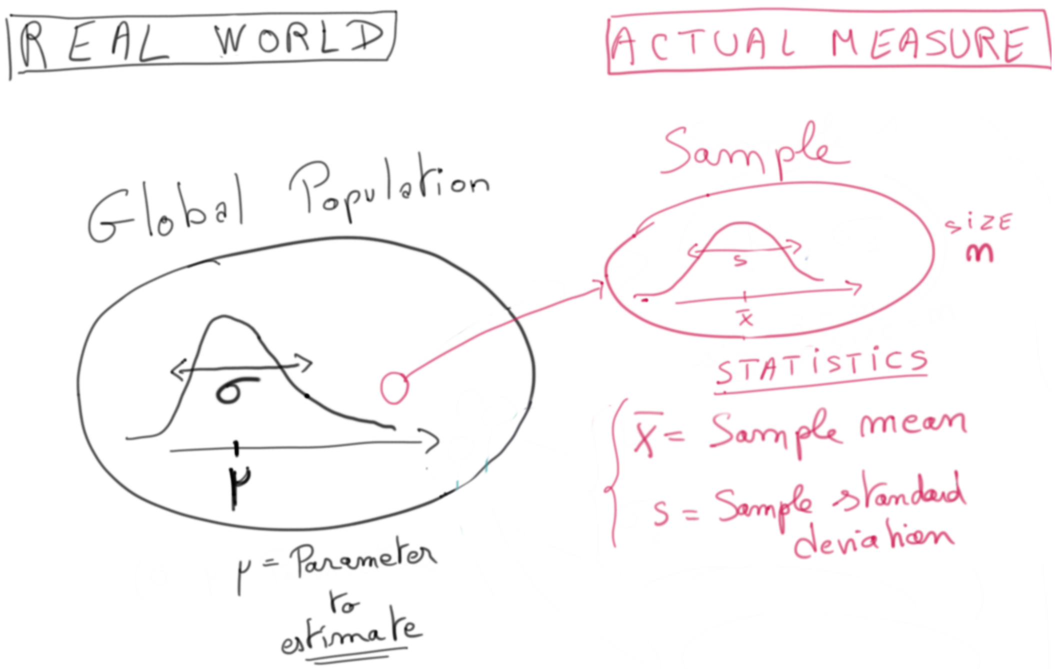
🥋 How to estimate the average height of US citizens ?

🎯 If my goal is to estimate the average height μ among the US citizens:

❌ I can't measure the entire US `population` (`N` = 331 M)

Random sampling method

- I randomly select a sample of size `n` = 1000 people from the population 🎲
- Based on these 1000 people, I can compute 🧮 :
    - the sample mean $\overline{X_n}$ = 170 cm
    - the sample standard deviation s = 20 cm



❓ What does it say about μ ❓

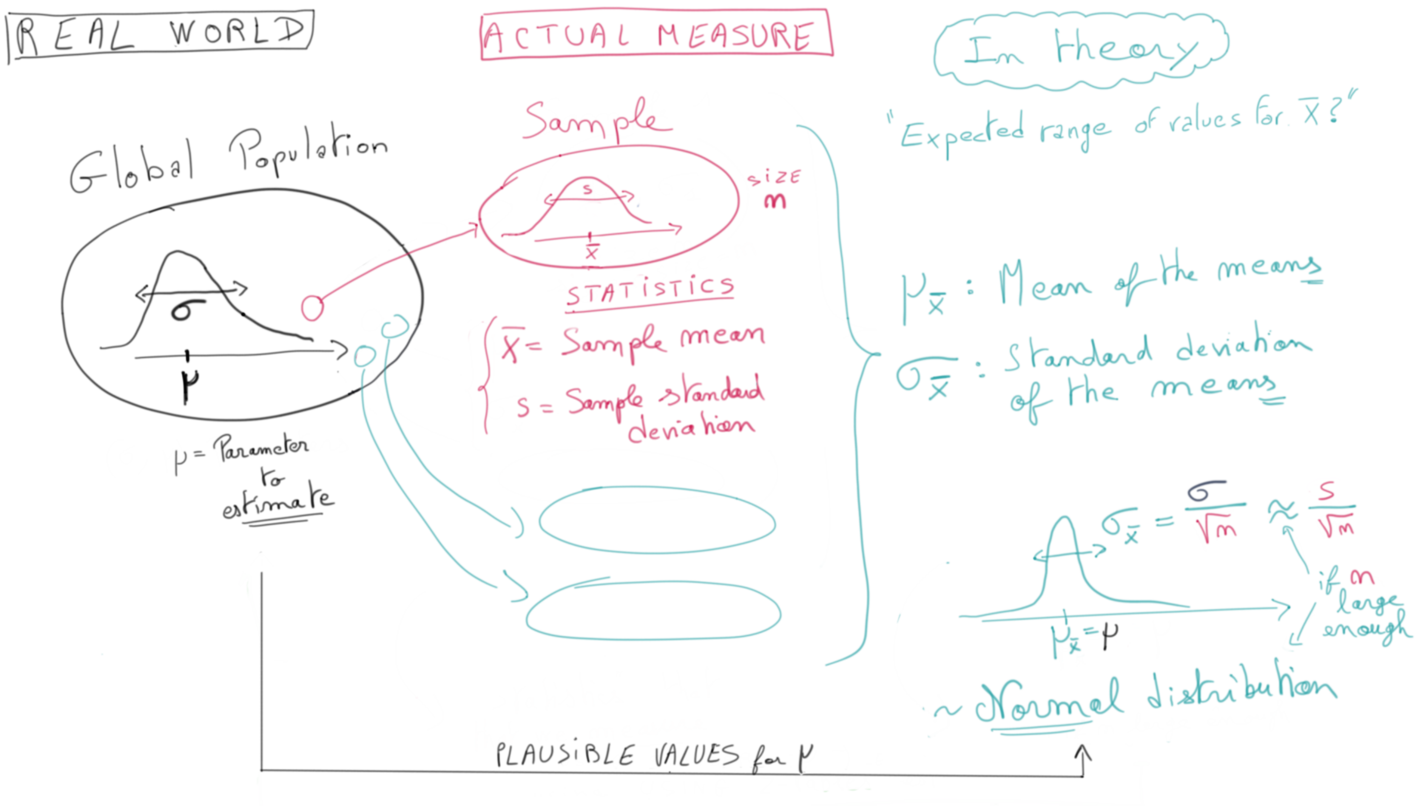
Best Guess

- Our best estimation for μ is $\overline{X_n}$ = 170 cm
- This intuitive fact is due to the Law of Large Numbers:

> When you consider **independent random variables** $X_1 \ldots X_n$ with a **common** underlying probability distribution $p(\mu, \sigma)$, their average $\overline{X_n}$ becomes a strong approximation of μ as the sample size n increases:
>
> $$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n} \xrightarrow[n \to \infty]{} \mu$$

Confidence Interval

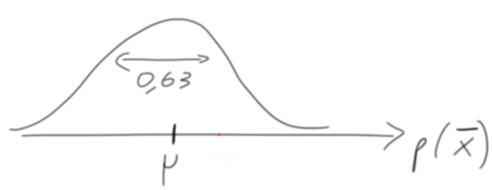
- We can also give a *distribution of plausible values* for μ 🎉
- Thanks to the Central Limit Theorem



Because
n
is large enough, and the citizens are randomly selected (CLT):

The distribution of sample mean**s**
$\overline{X_n}$
should follow the normal distribution:

$\overline{X_n} \approx \mathcal{N}$

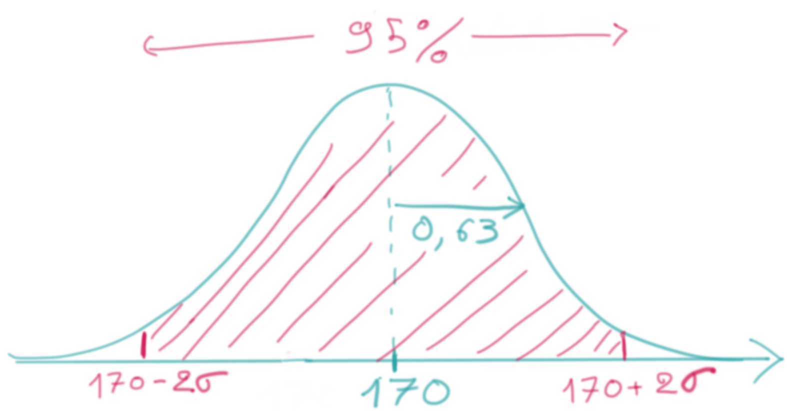


- So we know that $\overline{X_n}$ *should* be centered round $\mathcal{N}$ (
- And yet we *did* measure $\overline{X_n} = 170$ cm
- What distribution for μ is therefore the **most plausible / likely**?

$\mathcal{N}$ (

We say that $\overline{X_n}$ = 170 cm is the "**Maximum Likelihood Estimate (MLE)**" for μ

Estimated probability for μ :



👉 We read :

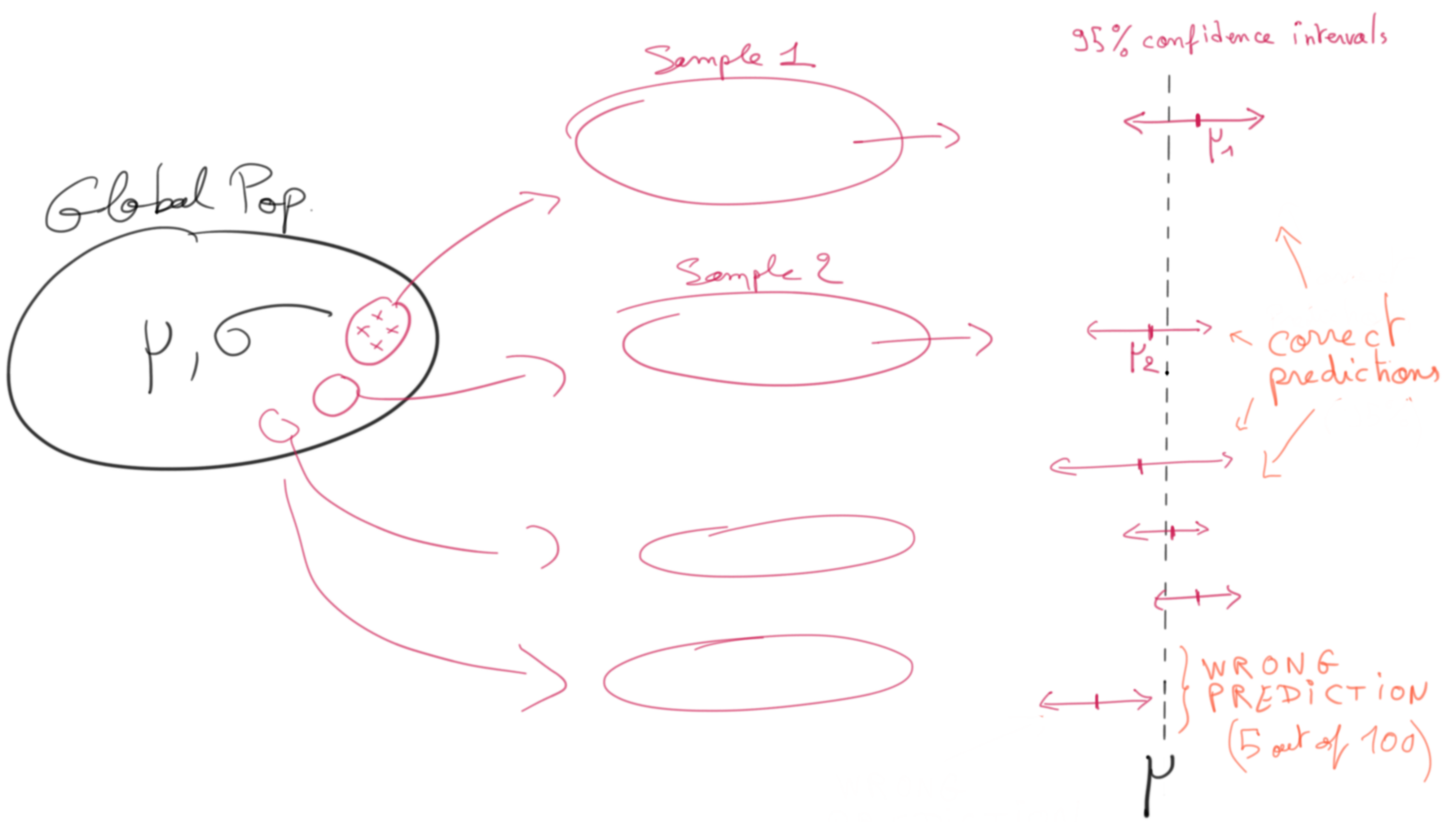
$\mu = 170 \pm 2 \times 0.63 \;\; [95\% \text{ confidence interval}]$

$\Leftrightarrow \mu = 170 \pm 1.26cm \;\; [95\% \text{ confidence interval}]$

$\Leftrightarrow \mu \text{ is between } 168.7 \text{ and } 171.2cm \;\; [95\% \text{ confidence interval}]$

Confidence Interval (interpretation)

✅ If we were to repeat this process and construct many other samples, 95% of the intervals produced will actually contain the true US mean pop height



✅ We're 95% confident that [168.7 - 171.2] captures the **true average height**.

❌ Don't say "*there is a 95% probability that* μ *is between ...*" because the real μ isn't random!

In [ ]:
```
# We can check these figure using a Cumulative Density Function `cdf`
from scipy import stats
mu_estim = stats.norm(170, 0.63)

# use the cdf to find the probabilities associated with height values
print('% confidence interval = ', round(mu_estim.cdf(171.2) - mu_estim.cdf(168.7),2))
```

```
% confidence interval =  0.95
```

💡 Actually, there is a formula to find the lower bound and the upper bound of any confidence interval

(ex: 99%)

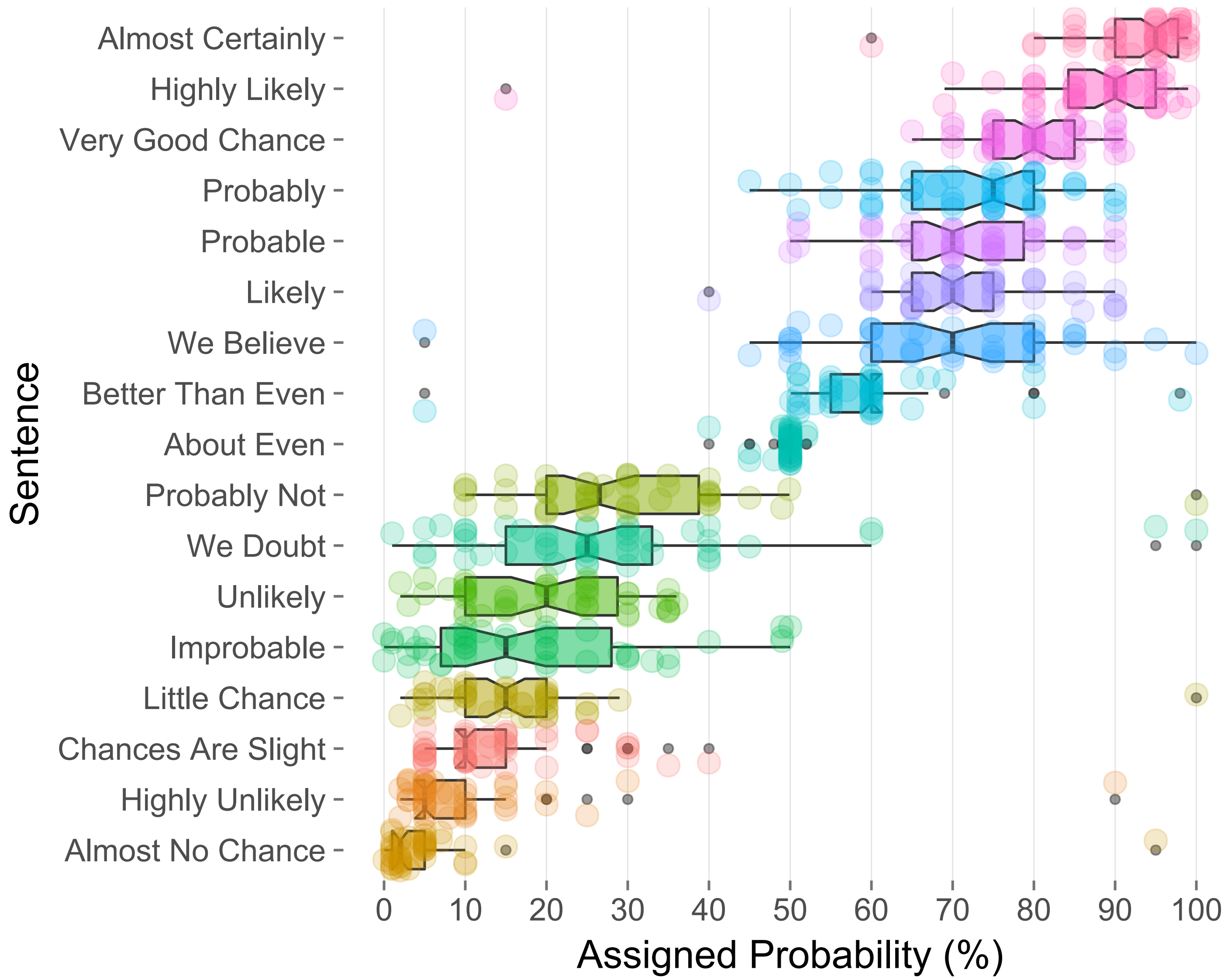
In [ ]:
```
confidence_interval = 0.99

sup_proba = (1 + confidence_interval)/2 # 99.5%
inf_proba = (1 - confidence_interval)/2 # 0.5%

mu_upper_bound = mu_estim.ppf(sup_proba)
mu_lower_bound = mu_estim.ppf(inf_proba)

# use the inverse of the cdf to find the heights associated with probabilities
print('mu_upper_bound: ', mu_upper_bound)
print('mu_lower_bound: ', mu_lower_bound)

print('% confidence interval = ', round(mu_estim.cdf(mu_upper_bound) - mu_estim.cdf(mu_lower_bound),2))
```

```
mu_upper_bound:  171.6227724612358
mu_lower_bound:  168.3772275387642
% confidence interval =  0.99
```

Human perception

Here are some usual English names for confidence intervals
_(according to the Intergovernmental Panel for Climate Change: [IPCC](https://archive.ipcc.ch/publications_and_data/ar4/wg1/en/ch1s1-6.html) (https://archive.ipcc.ch/publications_and_data/ar4/wg1/en/ch1s1-6.html) for instance)_

- 1-sigma (68%) `"likely"`
- 90% `"very likely"`
- 2-sigma (95%) `"extremely likely"`
- 3-sigma (99.7%) `"virtually certain"`
- 5-sigma: "proof" `threshold in theoretical physics`



[Source (https://mirkomazzoleni.github.io/blog/2016/perception_of_probability/)](https://mirkomazzoleni.github.io/blog/2016/perception_of_probability/)

Sample size `n` considerations

❓ When is n considered **large enough** for the CLT

Three cases:

- If $n > 30$ $\Rightarrow$ CLT applies **and** the sample std s can be used to approximate the true pop σ in z-statistics
- If $n > 10$ *and* observations are "non-skewed" and without outliers $\Rightarrow$ CLT still applies
- If the global population is known to be normally distributed $\Rightarrow$ CLT always applies even with an arbitrary small n, in a sense that we can use the Gaussian distribution

❓ When is n is **small enough** to consider each draw independent, even without replacement

- $n < 10\% \times N$

4. Hypothesis testing

🥋 Testing a new app feature

Imagine that I am the `PM (Product Manager)` for a Social Network Mobile App.

👉 My `N = 1000 users` spend:

- on average μ = **300 seconds** per session
- with a standard deviation of σ = **50 seconds**.

💡 I have the intuition that changing the background color from a light to a dark mode would increase the time spent per session.

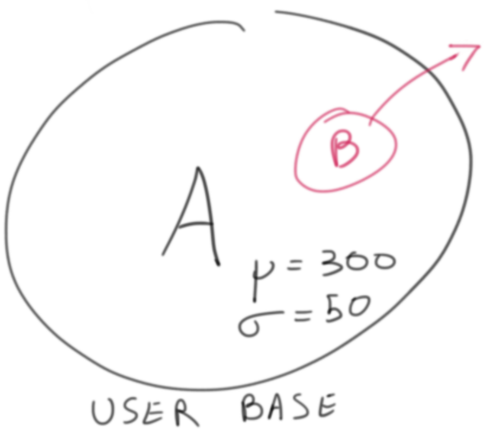
❓ **How could I test my hypothesis rigorously**, to convince my CTO to roll out the new feature ❓

Step ① : Create a "A/B Test" (the `Experiment Design`)

1. Develop the corresponding feature (dark mode)
2. Create two groups (`control group` vs. `treatment group`)
3. Randomly assign n = **100 users** to the treatment group
4. Deploy dark mode only to treatment group
5. Collect behavior statistics: We find a **sample mean** $\overline{X_n}$ = $\overline{X_{100}}$ **= 310 seconds**

Step ② : Test your hypothesis (= statistical analysis of the outcome)

① Choose
$$\begin{cases} H_0 : \text{DARK mode changes nothing} \\ H_a : \text{DARK is better} \\ \alpha = 0{,}05 \quad \text{significance level} \end{cases}$$



② Measure sample mean screen time $\bar{X} = 310$

③ $P(\bar{X} \text{ or more} \mid H_0) = \text{p-value}$

$> \alpha$ → Fail to reject H_0

$< \alpha$ → Reject H_0

④

Reality	Fail to reject H_0	Reject H_0 → GO DARK
H_0 true	correct	Type I error
H_0 false DARK = good	Type II error	correct

Step ② : Test your hypothesis (= `Statistical Analysis` of the outcome)

1. Create **`Null Hypothesis`** H_0 : μ *= 300 (unchanged) in dark mode*
2. Create **`Alternative Hypothesis`** H_a : μ *> 300 (increased) in dark mode*
3. Choose a **`Significance Level`** α for your experiment (ex: α = 5%)
4. Suppose that H_0 is true, and compute the probability of observing a sample mean $\overline{X_n} \geq 310$

👉 This probability is called the **`p-value`** = $P((\overline{X_n} \geq 310)|$

- If **p-value** < α , then we **reject** the null hypotheses H_0 in favor of the alternative hypothesis H_a
- If **p-value** > α , then we **fail to reject** the null hypothesis H_0 . This doesn't mean we accept H_0 ❗

Let's compute our `p-value`

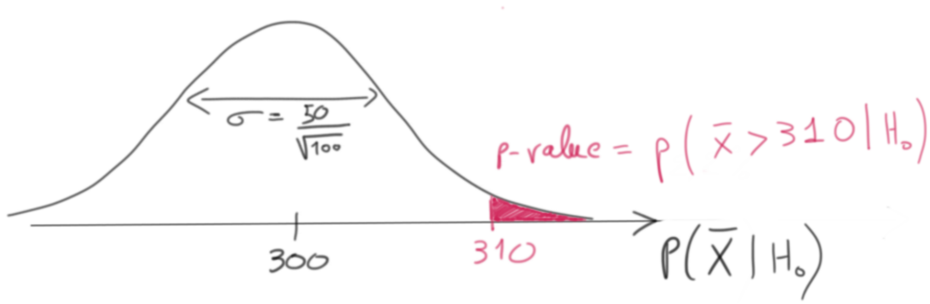
Since
$n = 100$
is large enough, the CLT applies and tells us that

👉 The distribution of sample means
$\overline{X_n}$
should follow the normal distribution:

$\overline{X_n} \approx \mathcal{N}$

Supposing
H_0
is true (unchanged behavior), then we know that
μ
= 300 and
σ
= 50

$\overline{X_{100}} \approx \mathcal{N}$

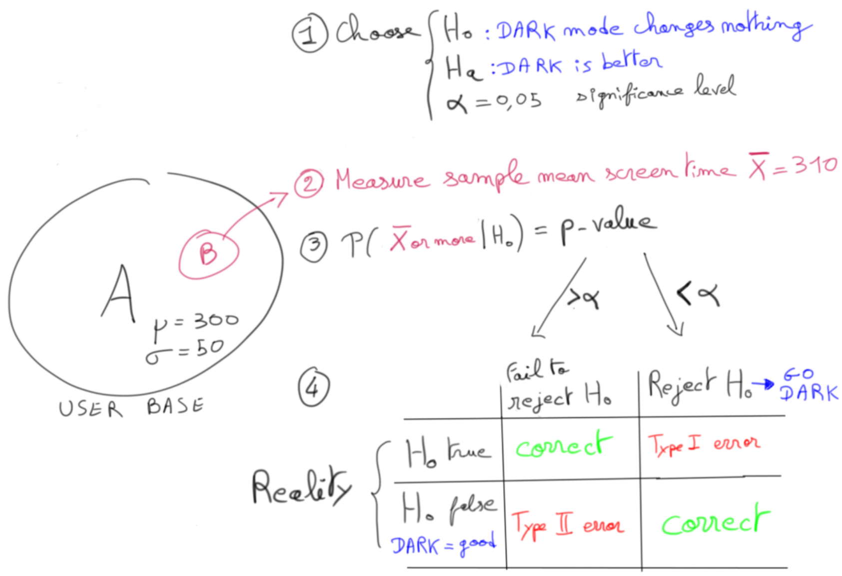


A standard z-table or a numerical computation would help up compute the `p-value`

In [ ]:
```python
from scipy.stats import norm
X = norm(300, 50/(100**0.5))
p_value = (1 - X.cdf(310));
round(p_value,2)
```

Out[ ]: 0.02

👍 `p-value` < 0.05
$\Rightarrow$
We can safely reject the Null Hypothesis
H_0
$\Rightarrow$
The dark mode is a real plus!



You will also often encounter the **power** of a statistical test (the larger the better):

- Power is the probability that we will **correctly reject the null hypothesis** (if it was correct to reject it)
- Power = Proba of *not missing out* a great feature in A/B testing
- Power = Proba of *not missing out* an effective new drug in clinical trial
- Power = P(not making a type II error)

📺 [StatQuest intuitive video (https://www.youtube.com/watch?v=Rsc5znwR5FA&list=PLblh5JKOoLUIcdlgu78MnlATeyx4cEVeR&index=112&t=0s)](https://www.youtube.com/watch?v=Rsc5znwR5FA&list=PLblh5JKOoLUIcdlgu78MnlATeyx4cEVeR&index=112&t=0s)

Choosing significance level α ?

$\alpha = 0.05$
is the standard significance level we generally start with.

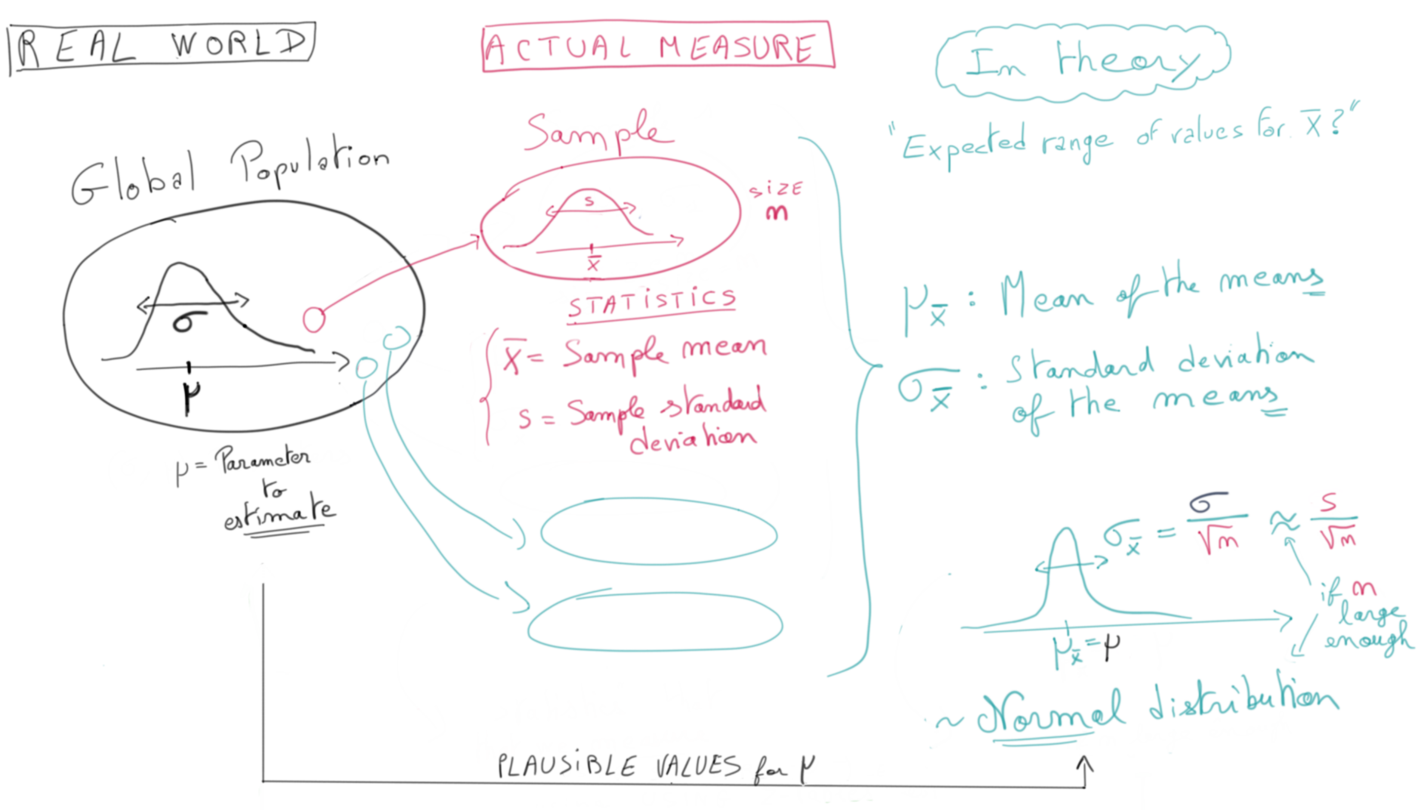
Notes:

1. It is the value usually used for clinical trials
2. It can vary from one industry to the other, from one experiment to the other, ...

❌ Never change
α
afterwards to reject / fail to reject your hypothesis to your own will...

✅ Choose your
α
beforehands, depending on your susceptibility to **`Type I errors` (False Positives)** *vs.* **`Type II errors` (False Negatives)**

5. t-tests (for small sample sizes)



🤔 What to do when the **sample size n is not large enough** and we don't know the true σ population ?

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- cannot be approximated by $\mathcal{N}$ (
- cannot be computed without knowing true σ of the population

💡 However we can always compute the `T-statistics` :

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

And fortunately we can prove that:
$T = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{T}_{n-1}$
is called a `Student distribution with n-1 degrees of freedom`

🙌 All good!

✅ Everything applies as before, but replace $\mathcal{N}$ with $\mathcal{T}$:

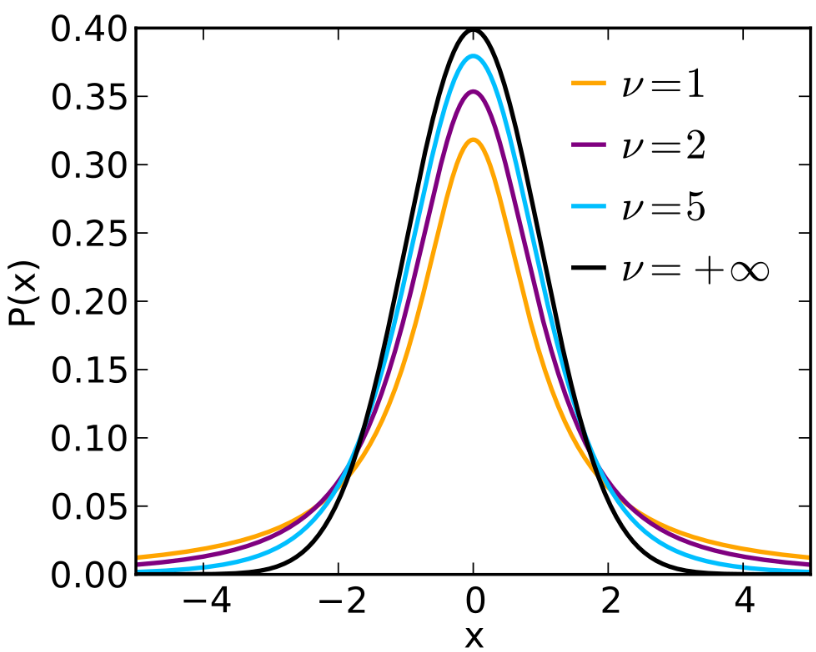
- Use `t-tests` instead of `z-tests`
- Compute **confidence interval** using the `c.d.f.` of a Student distribution
  - Choose the correct number of degrees of freedom!
- **Test Hypothesis** (compute `p-value` with a significance level α)

❗ Still requires *independent* and *random* sampling

Student t-distribution

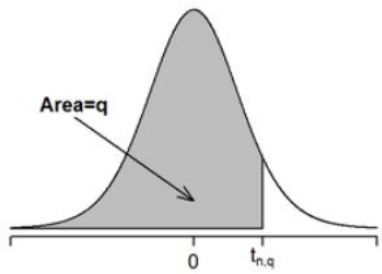
$\mathcal{T}_\nu(\mu, \sigma)$

- One distribution per **degree of freedom** ν
  - 📚 [Statistics By Jim - Hypothesis Testing - Degrees of Freedom=](https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/) (https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/)
- Accessible via `scipy.stats.t` or via `t-table`
- "Fatter" tails compared to a $\mathcal{N}$ ormal distribution
- $\mathcal{T}_\nu \xrightarrow[\nu \to \infty]{} \mathcal{N}$



Quartiles of the t Distribution

The table gives the value if $t_{n;q}$ - the qth quantile of the t distribution for n degrees of freedom



	q = 0.6	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
n = 1	0.3249	1.0000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.2887	0.8165	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.2767	0.7649	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.2707	0.7407	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.2672	0.7267	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.2648	0.7176	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.2632	0.7111	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.2619	0.7064	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.2610	0.7027	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.2602	0.6998	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.2596	0.6974	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.2590	0.6955	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.2586	0.6938	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.2582	0.6924	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140

> If we were to measure 14 people randomly and to find an average height value with a `t-score` of 3, then this measured average height would extremely highly (99.5%) improbable!

Central Limit Theorem (generalized)

- $X_1 \ldots X_n$ independent random variables sampled from a global pop with mean μ and std σ
- $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ the sample mean (also referred to as the *empirical mean*)
- $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$ the sample standard deviation

👉 For n large enough:

$$T \sim Z = \left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \right.$$

👉 For any n in $\mathbb{N}$:

$$T = \left(\frac{\overline{X} - \mu}{\mathbf{s} / \sqrt{n}} \right.$$

🧐 Why (n-1)? Cf. [Bessel's Correction (https://www.statisticshowto.com/bessels-correction/)](https://www.statisticshowto.com/bessels-correction/)

6. Bayesian Interpretations

🥋 Let's sample a coin by flipping it
n
times to measure its **fairness** (i.e, the probability
$p(H) = p(\mu, \sigma)$
of landing on *Head*).

🤔 Is
μ
equal to 0.5 ? Is the coin fair ?

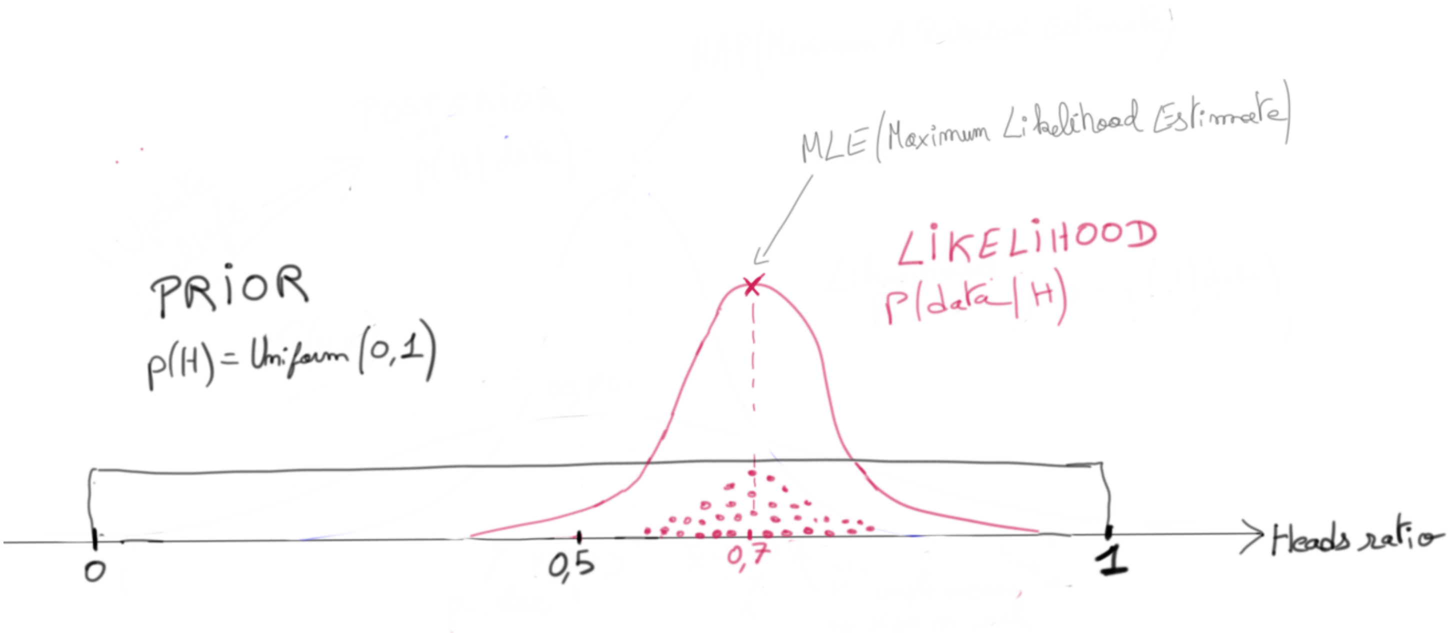
1. Prior to the experiment, we may have an opinion about the coin fairness

- This initial belief is called the **prior probability**
  $p(H)$
- If we have no opinion, we model the
  $p(H)$
  as a uniform distribution over [0,1]

2. Toss the coin n = 10 times

- We find sample mean of 0.7 and sample deviation
  s

❓ What is our new best guess for H, after having seen the new data
❓ i.e What is
$p(H|$
= posterior proba



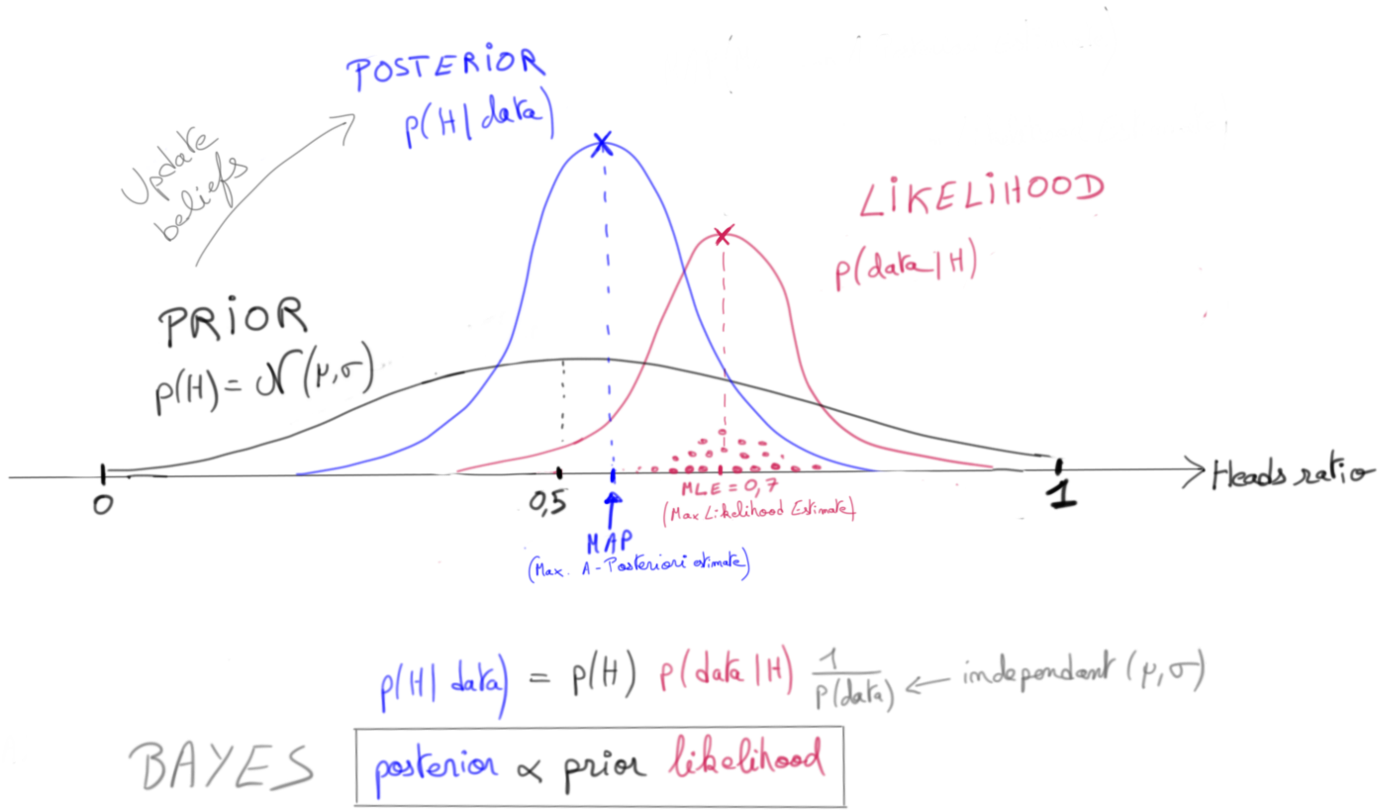
- In the absence of more **prior beliefs p(H)**, we can rely only on our observation
  - Our most likely estimate for μ is now 0.7 (maximum likelihood estimate MLE)
  - Our new best guess (**posterior proba**) of the coin fairness $p(H|$ is equal to $\mathcal{N}$ (
- Indeed, the CLT tells us that the most likely distribution from which such a mean of 0.7 may have been drawn is $\mathcal{N}$ (. (called the **likelihood** of observing data)

Now, imagine **we do have a prior belief** about the coin's fairness

- Extreme values close to 0 and 1 seem very unlikely to us (we can see it visually)
- Most coins are usually fair, so we think the most probable value is 0.5

👉 We will model our prior belief $p(H) = \mathcal{N}$

❓ What is our new posterior proba estimate
$p(H|$
= ❓



👉 Use **Bayes** to update our prior belief
$p(H)$
into our **posterior belief**
$p(H|$

Bibliography and Videos

- 📺 3Blue1Brown - Bayesian Updating and Probability Density Functions (https://www.youtube.com/watch?time_continue=3&v=rhuMH8A5t8s&feature=emb_logo)
- 📚 Towards Data Science - Jonny Brooks-Bartlett - Bayesian Inference for Parameter Estimation (https://medium.com/towards-data-science/probability-concepts-explained-bayesian-inference-for-parameter-estimation-90e8930e5348)
- 📺 Khan Academy - Statistics and Probability (https://www.khanacademy.org/math/statistics-probability) (~ 20h)
- 📚 Miguel A. Hernán and James M. Robins - Causal Inference - What if (https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2019/10/ci_hernanrobins_26oct19.pdf) (300-page M.Sc.level textbook)

🚀 Your turn

🗓️ Now:

- Creation of the "Orders" training set
- Quick analysis of the training set with a simple Linear Regression

🗓️ Next session:

- In-depth analysis of the "Orders" dataset with a Multivariate Linear Regression