Week 3: Statistical Modelling

STA238: Probability, Statistics, and Data Analysis II

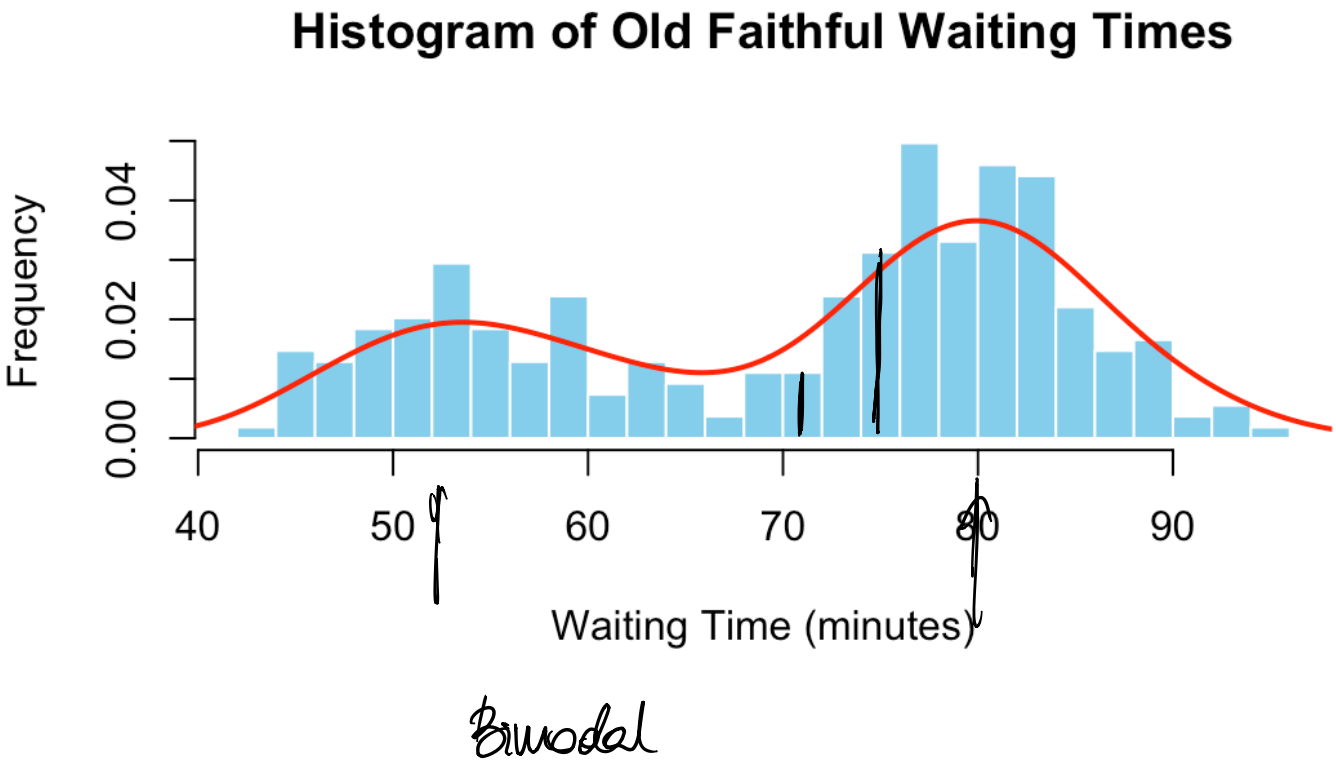
Luis Sierra Muntané

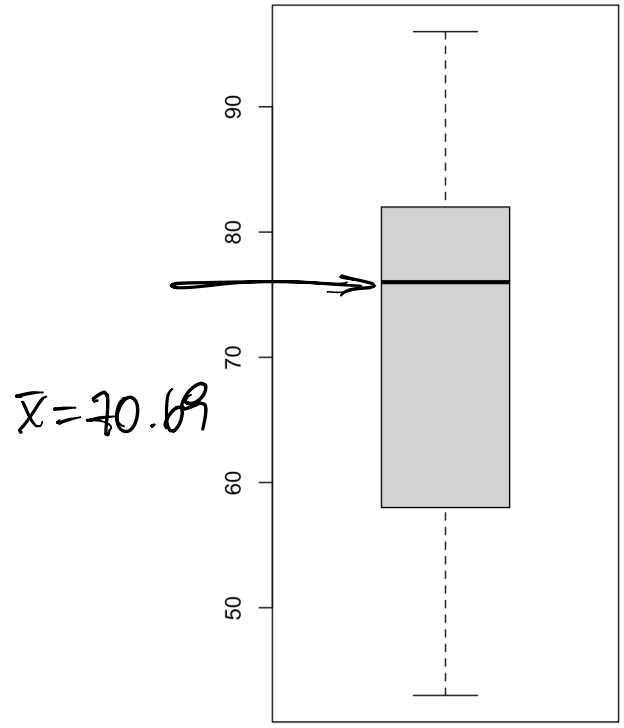
2026-01-20

Recap EDA:

Old Faithful geyser in Yellowstone National Park

Recap: EDA



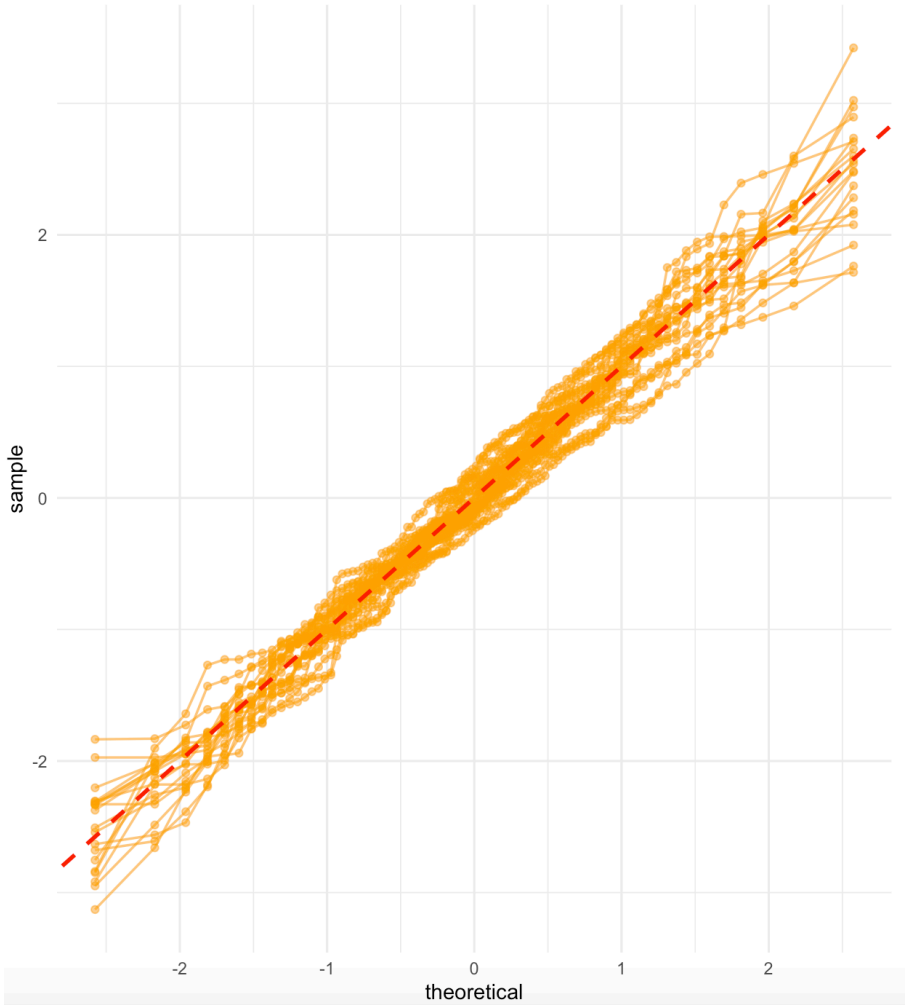


QQ Plot Recap:

Plots the quantiles of two chosen distributions (one of which is usually a standard normal).

- X-axis: quantile of distribution A
- Y-axis: quantile of distribution B

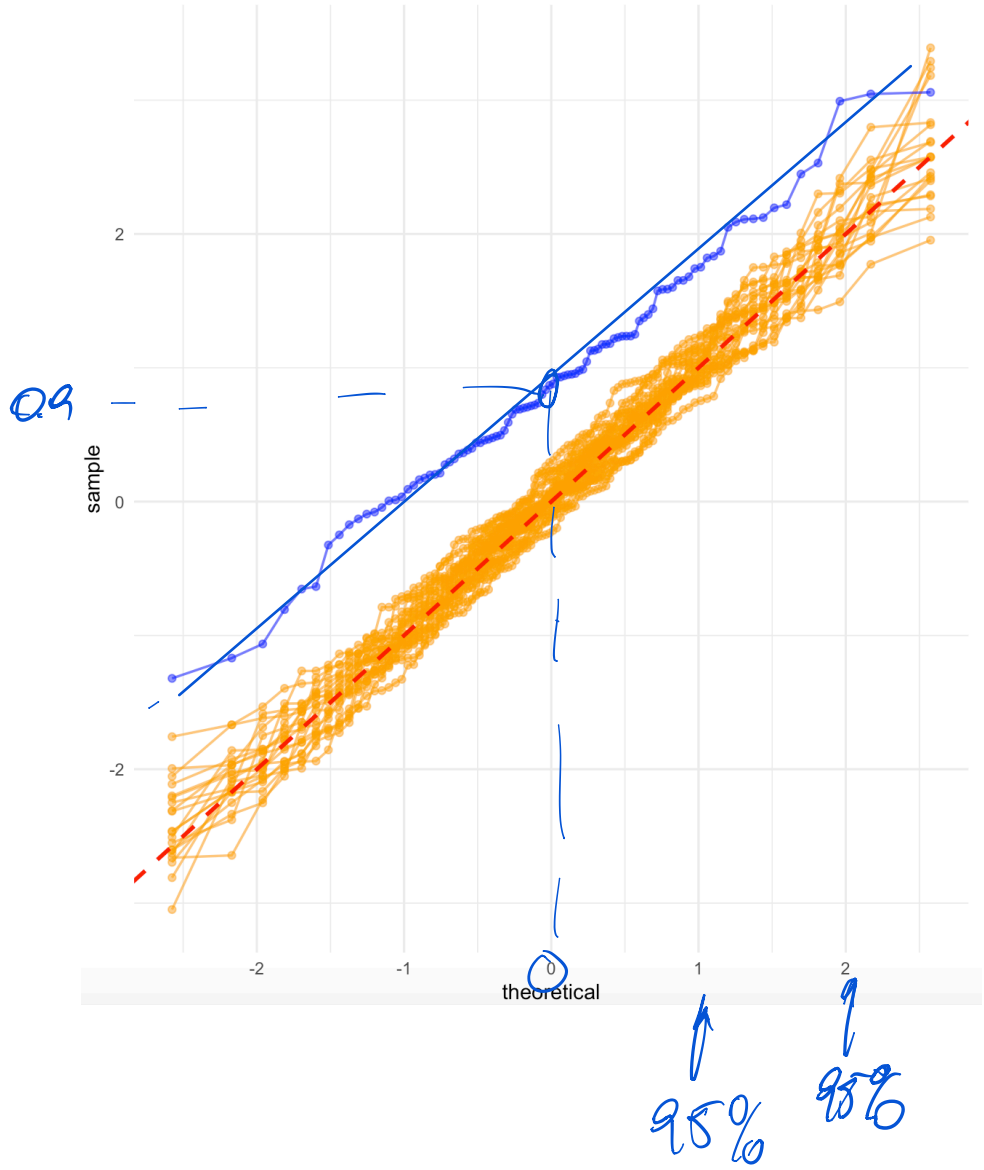
Points are then $(q_A(p_i), q_B(p_i))$, for a chosen set of values $p_i \in (0, 1)$.



QQ Plot Recap:

The distributions in orange are 20 separate sets of random samples from a standard Gaussian.

What is the distribution in blue?



Recap Quiz:

1. You are given this boxplot. What true statements you can make about the data?

a. There is only one mode.

b. Approximately 25% of the data is below 2.

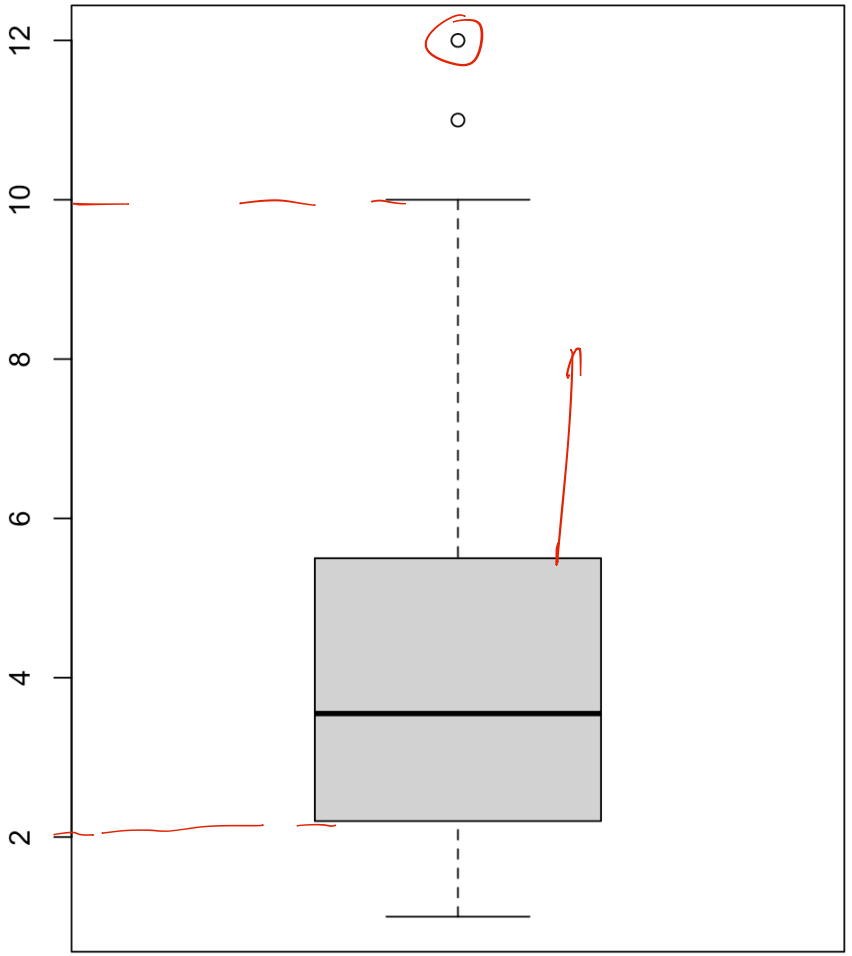
c. The maximum value in the data set is above 11.

d. The data is left-skewed.

e. Answers (a), (b), and (c) are true.

f. Answers (b), and (c) are true.

g. Answers (c), and (d) are true



Recap Quiz:

2. This density-scaled histogram was constructed on a sample of size = 200. What true statements you can make about the data?

a. Approximately $0.065 \times 2 = 13\%$ of the data is within $[4, 6)$.

b. Approximately $200 \times 0.2 = 40$ data points are in $[0, 2)$.

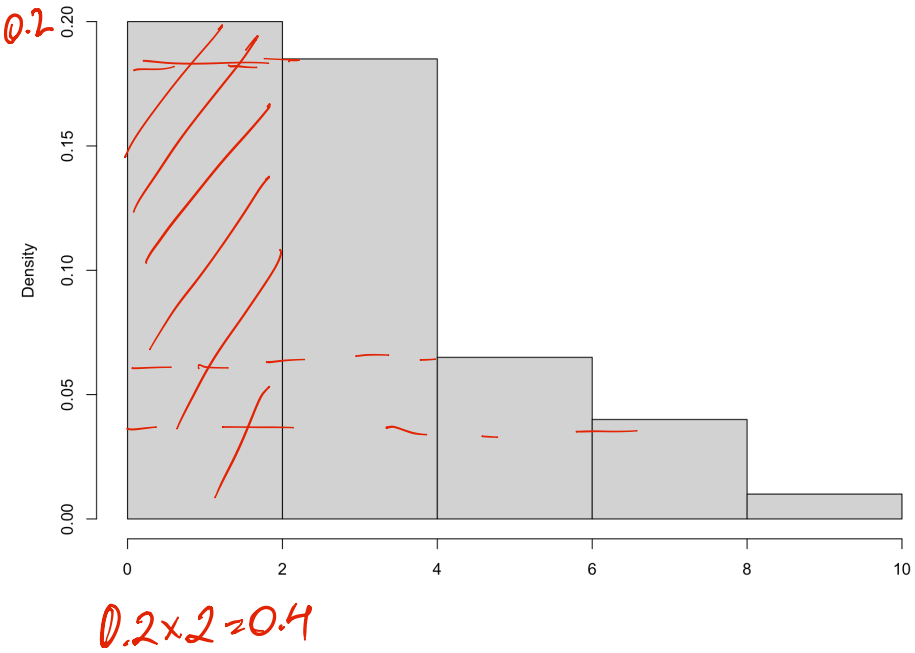
c. Approximately $200 \times 0.04 \times 2 = 16$ points are in $[6, 8)$.

d. Approximately 18% of the data is in the interval $[2, 4)$.

e. Both (a) and (b) are true.

f. Both (b) and (d) are true.

g. Both (a) and (c) are true.



Recap Quiz:

3. You have constructed an eCDF, $\hat{F}_{10}(x)$, based on a sample of size $n = 10$. You know that at a specific value $x = 3$, the eCDF evaluates to $\hat{F}_{10}(3) = 0.4$.You obtain two additional independent data points: $x_{11} = 2.5$ and $x_{12} = 3.5$. What is the updated value of the eCDF, $\hat{F}_{12}(3)$, based on the combined sample of 12 observations?

$F_{10}(3) = 0.4$, $\frac{4}{10}$ numbers ≤ 3

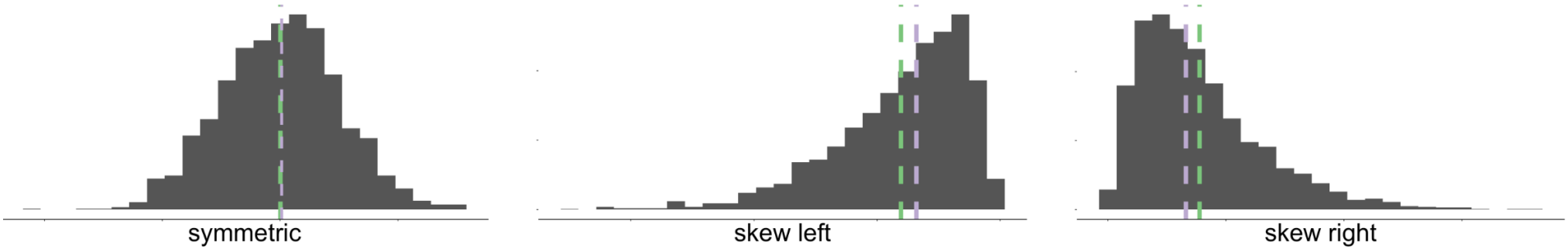
$F_{12}(3) = \frac{5}{12} \approx 0.42$

a. 0.33
b. 0.40
c. 0.42
d. 0.50
e. 0.58

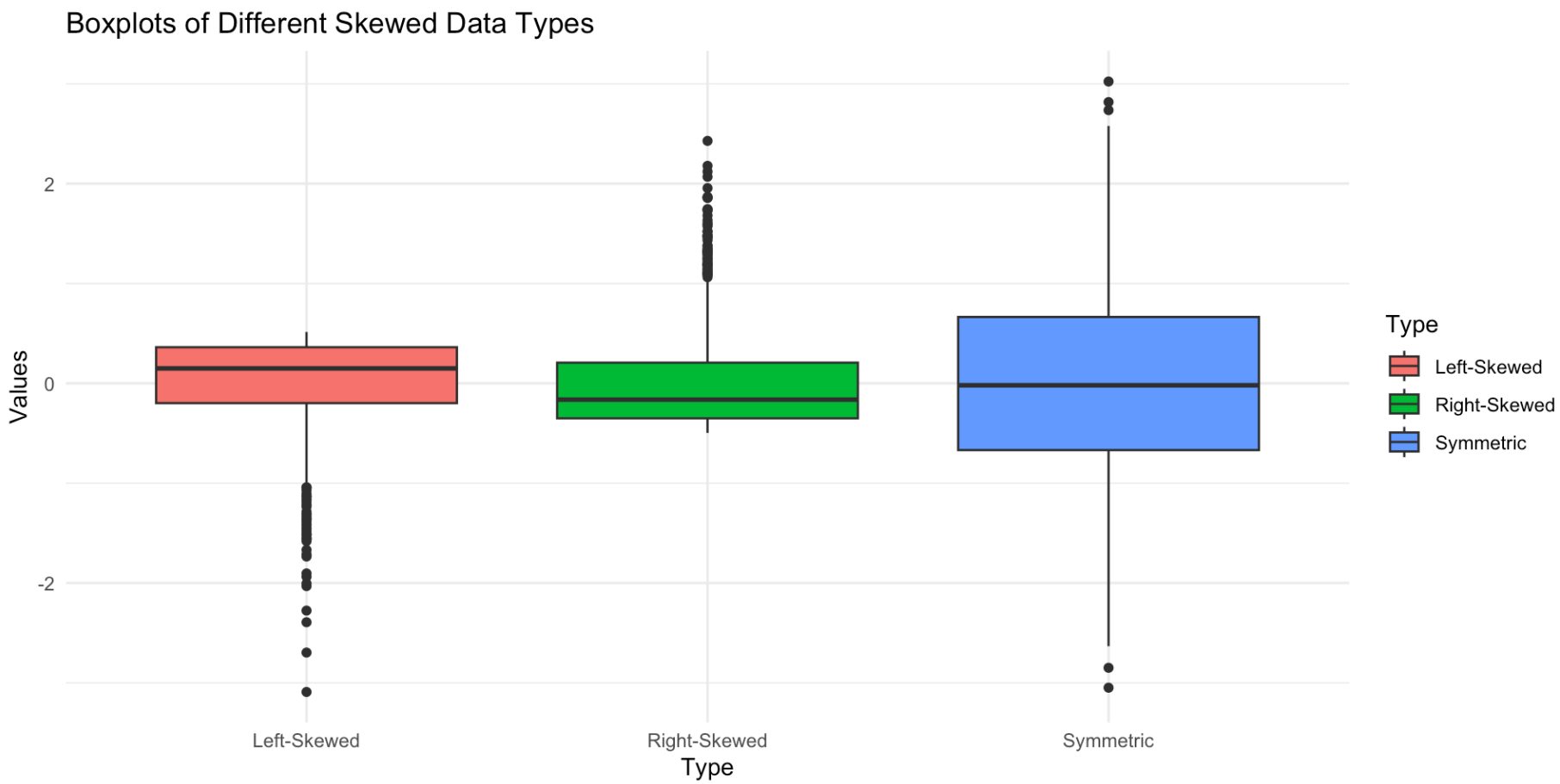
Skewness:

A measure of the asymmetry of a distribution. We often refer to it informally but there are ways of formalizing them.

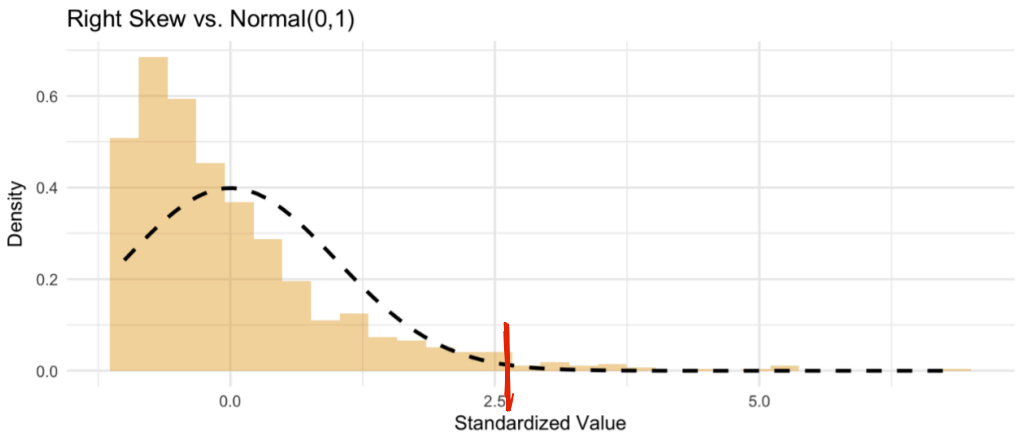
> One possible measure of skewness is
>
> $$\text{Sk}_n = \frac{1}{S^2} \sum_{i=1}^{n} (X_i - \bar{X})^3$$
>
> But there are other options, usually related to the third moment $\mathbb{E}[X^3]$.

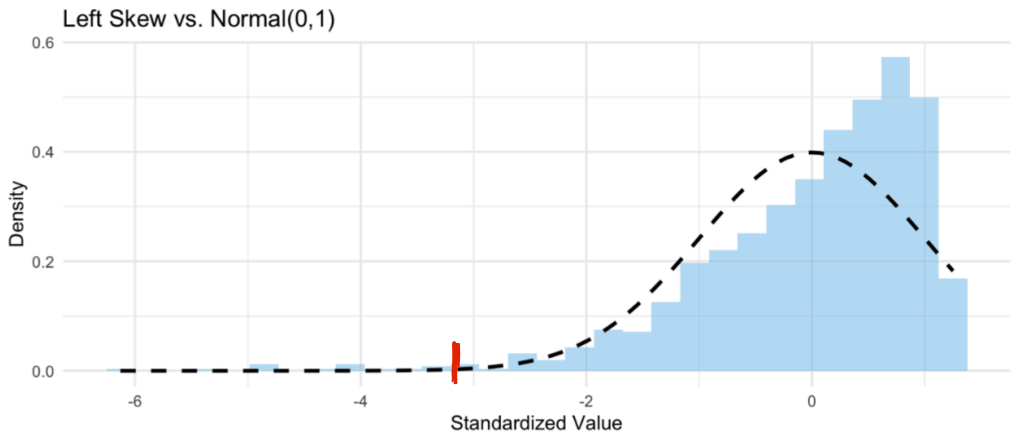


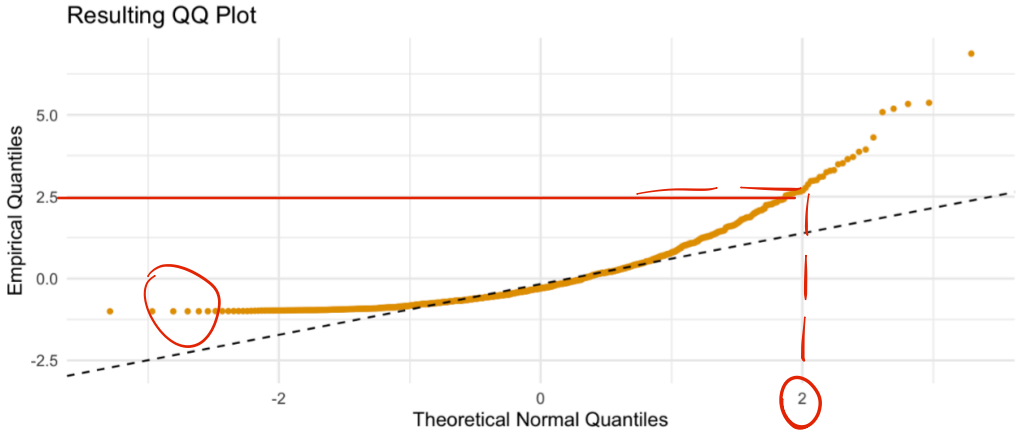
Skewness:

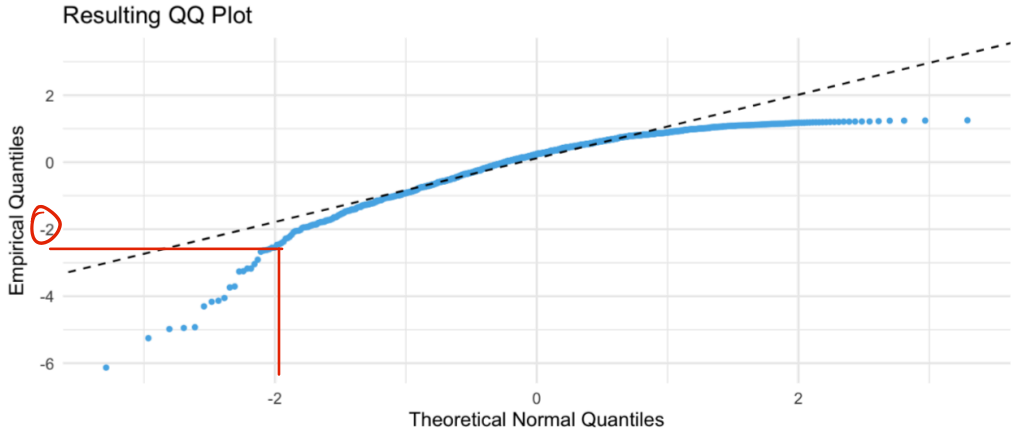


Skewness on a QQ-plot





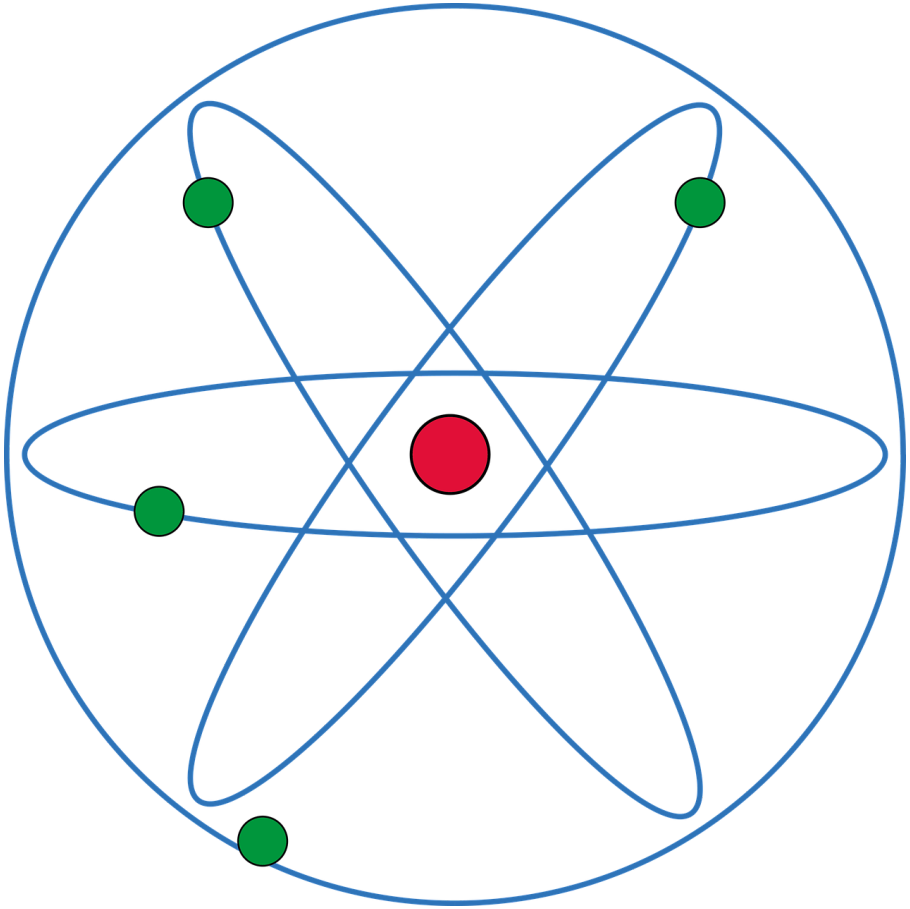




Statistical modelling and approximations

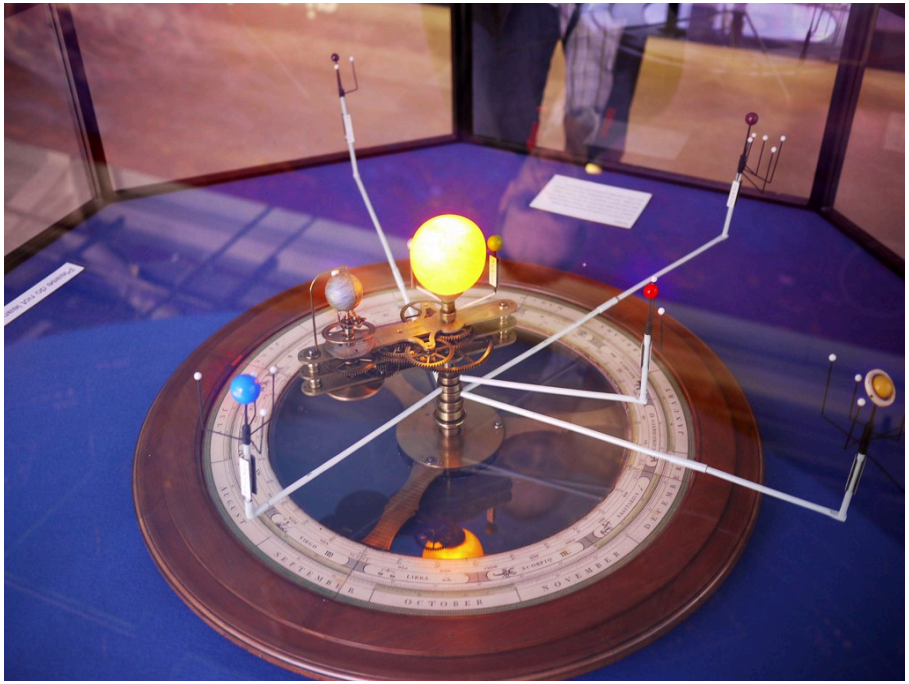
- MIPS Chapters 13 & 17
- What assumptions do we make about our data and why?
- The underlying (data-generating) distribution is hidden or partially hidden from us.
- We only observe realizations ($x_1, \ldots, x_n$)—this is our Data.
- We must use the data to make inferences about the unknown underlying distribution

What is modelling about



Rutherford model of the atom

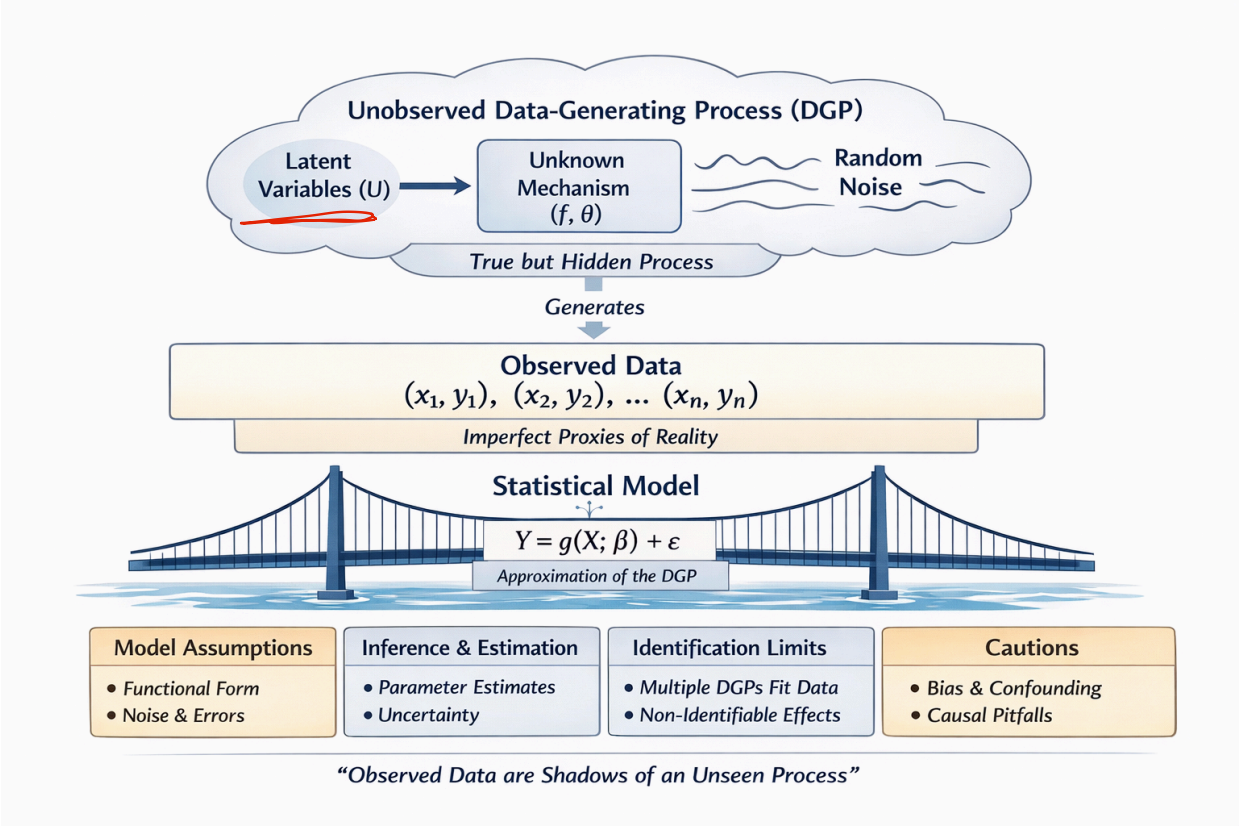
All electrons on the same energy level



Simplified planetary model

What is modelling about

> *"All models are wrong, but some are useful."* - George Box

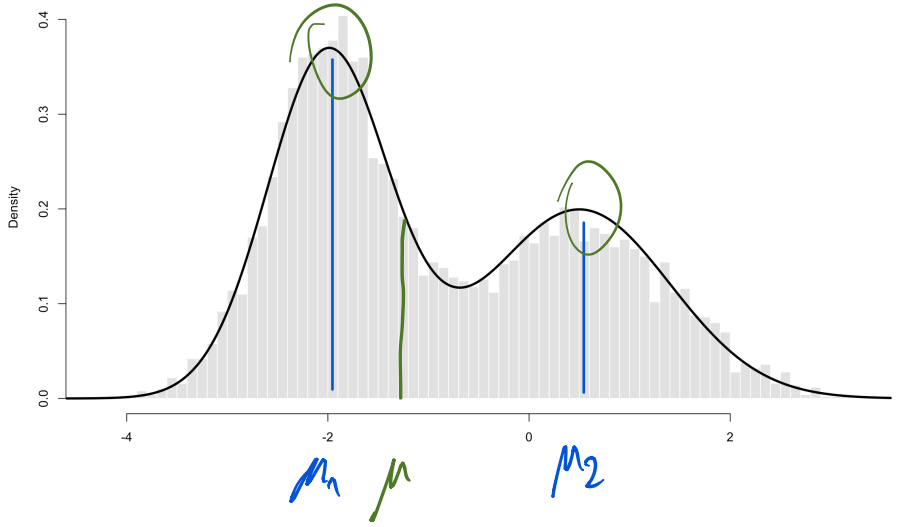


Parameters

These are values that govern the behavior of the DGP. They can be thought of as containing the information of the **population distribution**.

Consider the following example population distribution:

$$X_i \sim w_1 \mathcal{N}(\mu_1, \sigma_1^2) + w_2 \mathcal{N}(\mu_2, \sigma_2^2).$$



Population parameters:

- Weights: w_1, w_2.
- Means: μ_1, μ_2.
- Variances: σ_1^2, σ_2^2.

Is this *parametrization* unique?

Example: (MIPS)

We obtain a dataset of ten elements by tossing a coin ten times and recording the result of each toss. What is an appropriate statistical model and corresponding model distribution for this dataset?

What are reasonable assumptions we can make to build our model?

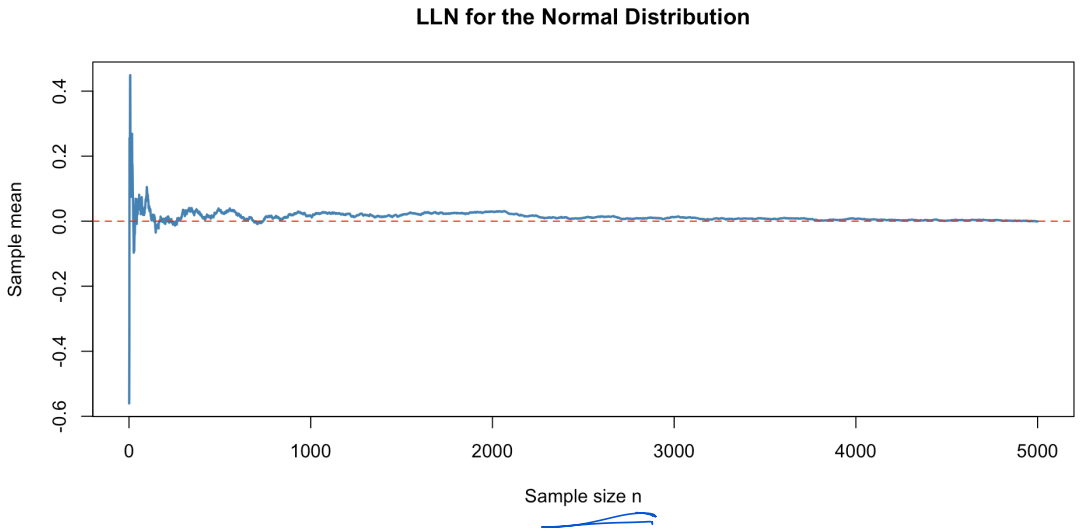
Fair toss
Independent tosses
2 outcomes
Identical trials
i.i.d.
$X \sim \text{Bin}(10, \phi)$

(Weak) Law of large numbers:

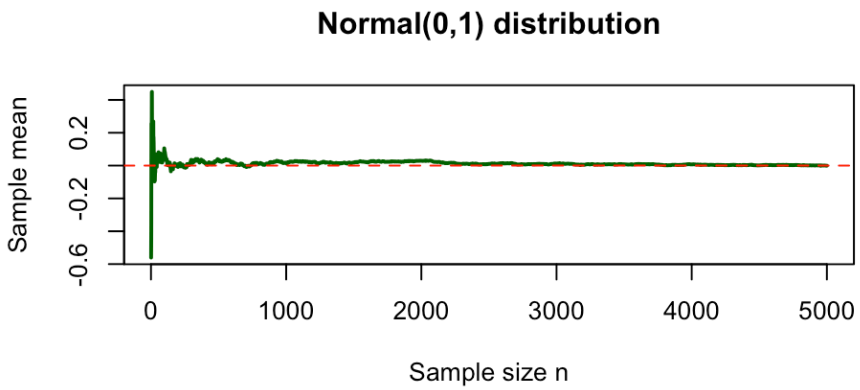
What happens as we collect more and more samples. Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}|X_1| < \infty$ and mean $\mathbb{E}[X_1] = \mu$. Then, as $n \to \infty$, the sample mean $\bar{X}_n$ satisfies

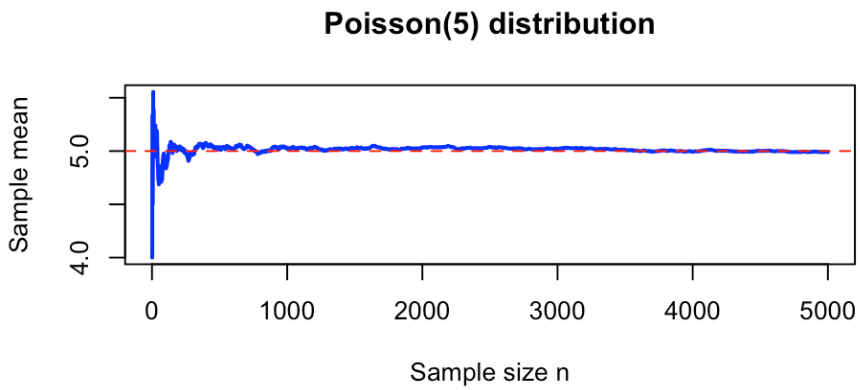
$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \xrightarrow{n\to\infty} 0. \quad \text{for any } \varepsilon > 0$$

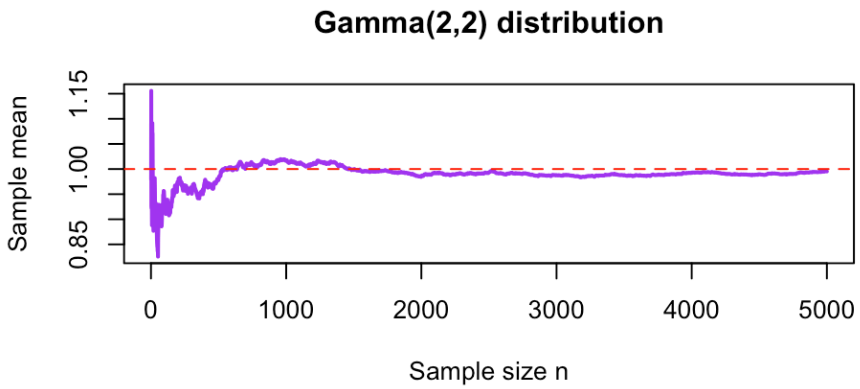
convergence in probability

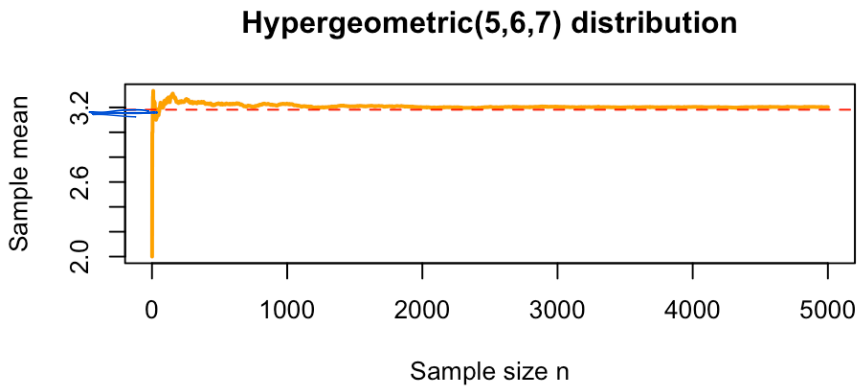


Example: LLN







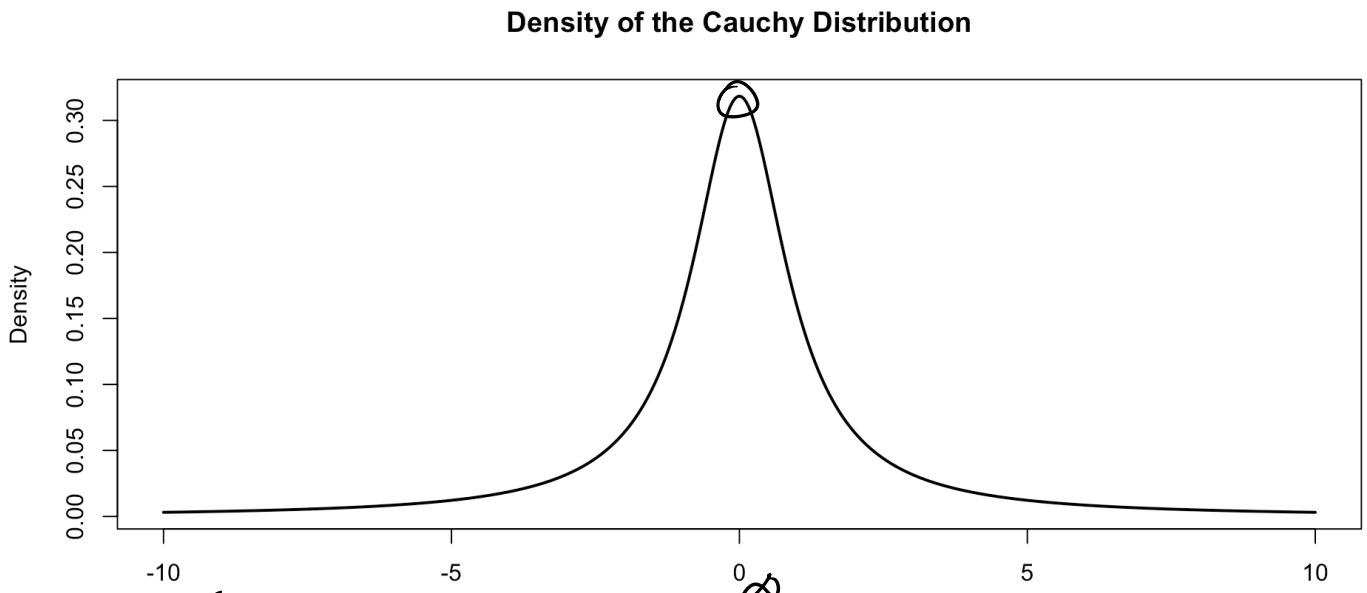


Example: Cauchy Distribution

Consider a random variable X following an absolutely continuous distribution with density defined for $x \in \mathbb{R}$ as $f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$.

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} dx = \frac{1}{\pi} \arctan \Big|_{-\infty}^{\infty}$$

$$= \frac{1}{\pi} \left(\frac{\pi}{2} - \left(-\frac{\pi}{2} \right) \right) = 1$$



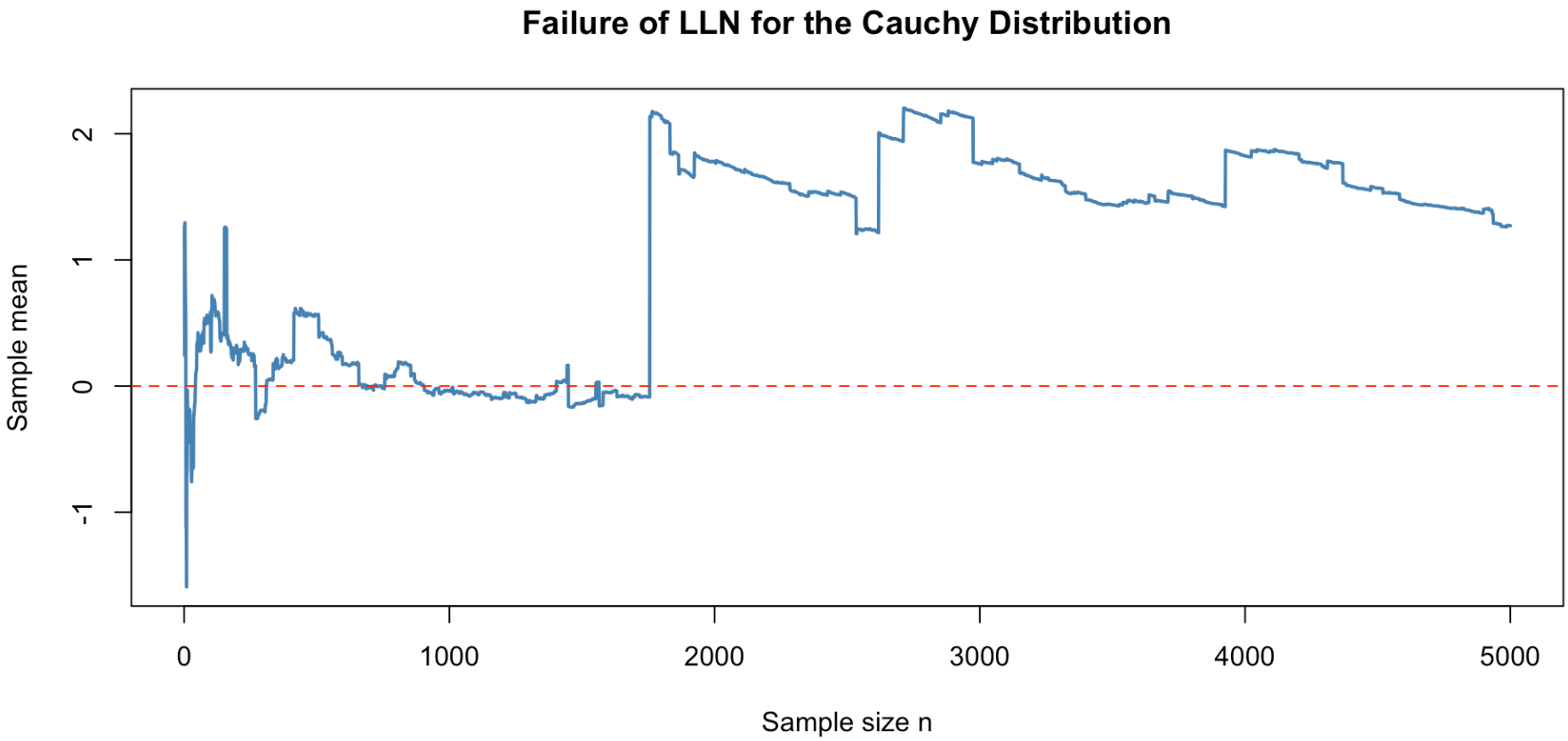
$\mathbb{E}X$ does not exist

$\int x \cdot f(x) dx$

$$\mathbb{E}X = \int_{-\infty}^{\infty} x \cdot \frac{1}{\pi} \cdot \frac{1}{1+x^2} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2x}{1+x^2} dx = \frac{1}{2\pi} \log(1+x^2) \Big|_{-\infty}^{\infty} \quad \text{“}= \text{”} \quad \frac{1}{2\pi}(\infty - \infty)$$

$$\mathbb{E}|X| = \int_{-\infty}^{\infty} |x| \cdot f(x) dx \longrightarrow \infty$$

Example: Cauchy Distribution



Heavy tails:

Experiment 1: Imagine a stadium filled with 50,000 randomly selected people. We record everyone's height and use that to calculate the average height.

- Now, the tallest person in history (Robert Wadlow, 8 ft 11 in) walks in.
- Does the average change? No. It barely moves a millimeter.

Experiment 2: Imagine the same stadium, but we record the net worth of the randomly selected people there.

We have a mix of students, teachers, and professionals. We calculte the average net worth.

- Now, Elon Musk walks in.
- The average ~~income~~ wealth of the stadium instantly skyrockets to over $4 billion per person.

"Heavy tails"

Heavy tails:

Gaussian pdf. $\frac{1}{2\pi} e^{-\frac{1}{2}x^2}$

A mathematical defnition: a random variable X has a right heavy tail if, for all $\lambda > 0$:

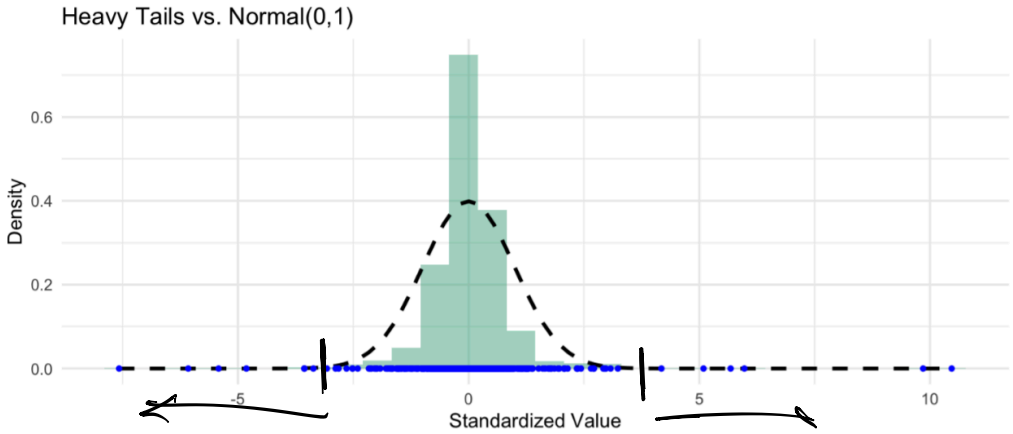
$$\lim_{x \to \infty} e^{\lambda x} P(X > x) = \infty,$$

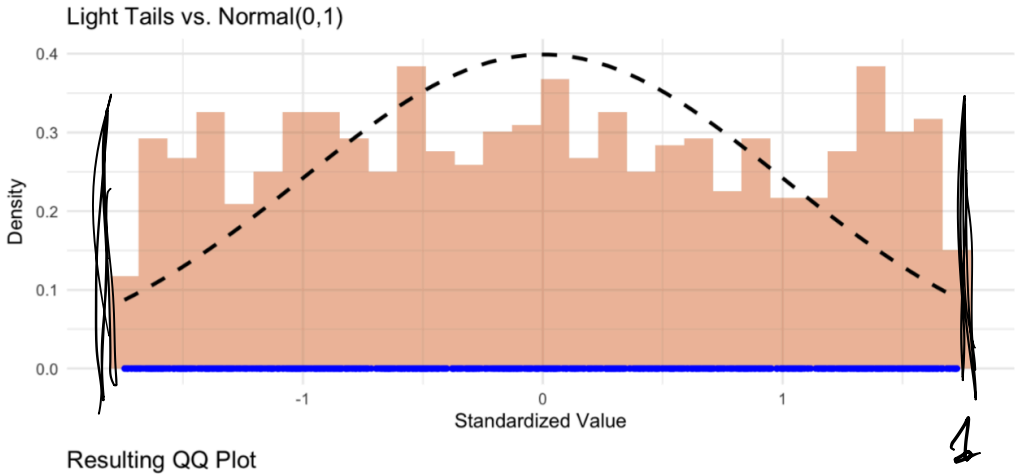
"Tails contain more probability than a normal distribution."

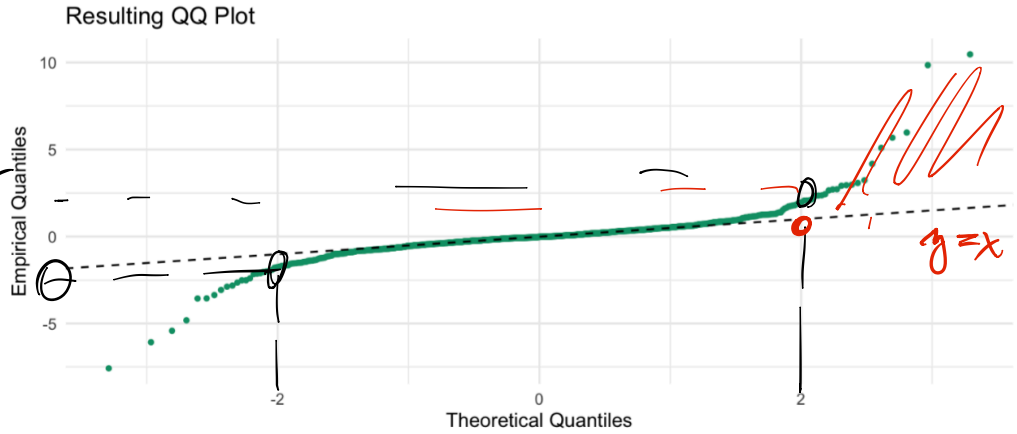
Many real-world data exhibit heavy-tail behaviour:

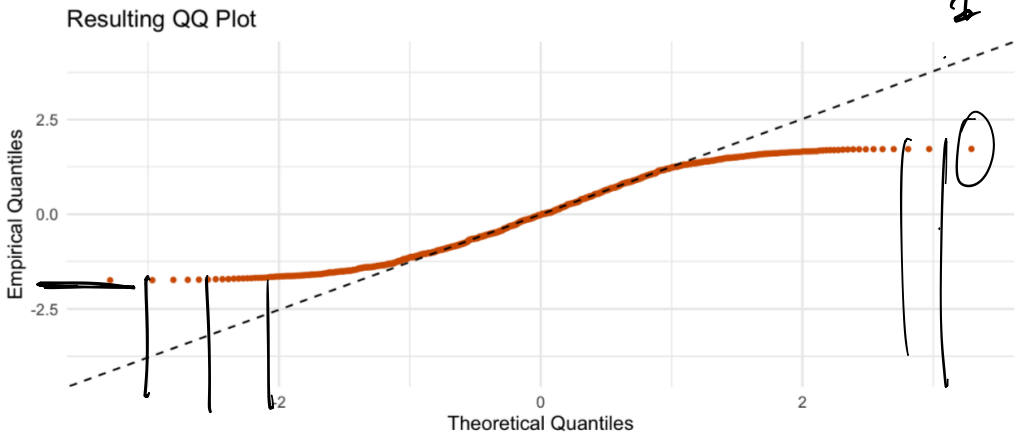
- Stock market returns
- Population of cities
- File sizes
- Earthquake magnitudes
- Book sales...

Tail behaviour on a QQ-plot









S shape

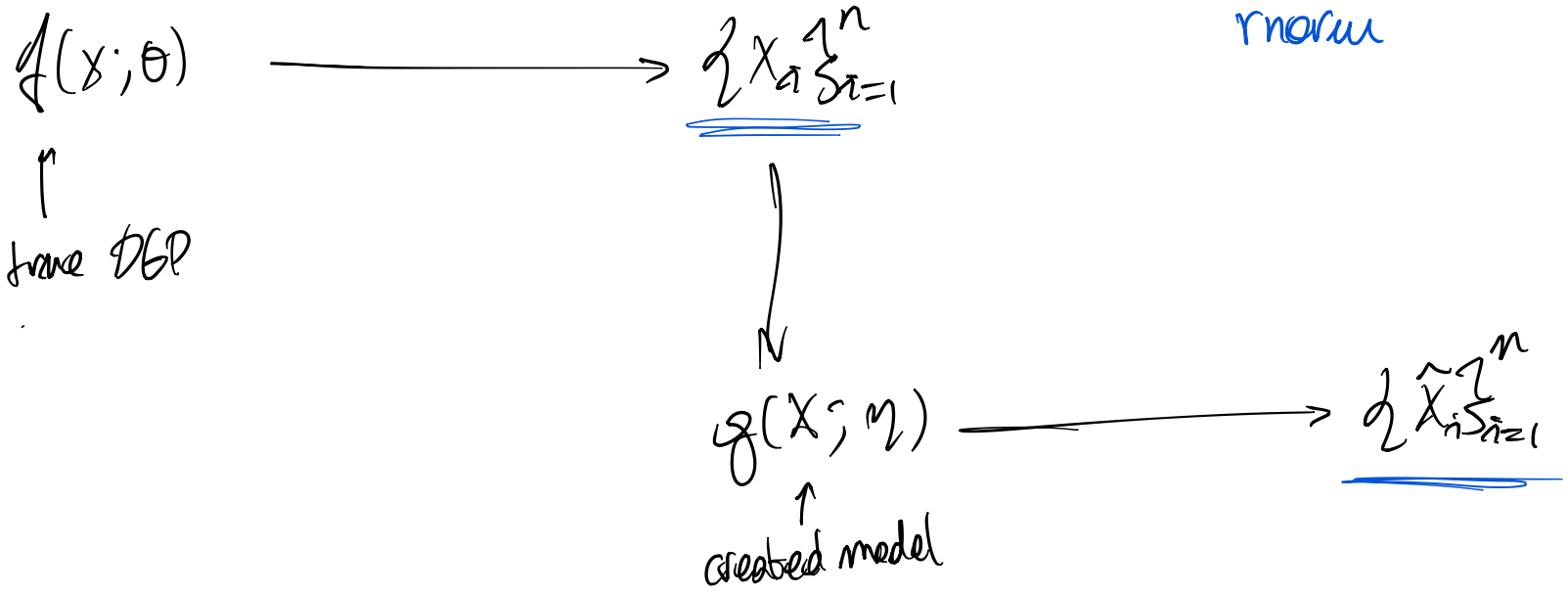
Simulations

How can we validate whether our models satisfy the assumptions we want?

Sometimes, explicit calculations can be too complicated — simulations offer an experimental way of checking results.

MIPS Chapter 6: *"Stochastic simulation is an alternative approach: values are generated for the random variables and inserted into the model, thus mimicking outcomes for the whole system."*

For it, we need to be able to generate random numbers (RNGs). R has in-built functions that generate realizations of random variables.



Example: sampling with the CDF

How to produce samples from a distribution with an invertible cdf F, from uniform samples.

1. $Z_i \sim \mathbf{Unif}(0,1)$
2. Transform each uniform draw via the inverse CDF: $X_i = F^{-1}(U_i)$
3. Output $X_1, \ldots, X_n$.

Checking the distribution of X_i:

$$\mathbb{P}(X_i \leq x) = \mathbb{P}(F^{-1}(U_i) \leq x) = \mathbb{P}(U_i \leq F(x)) = F(x).$$

Note: Bayesian inference

Classically, we aim to recover the parameters of a model by looking at more and more iid. data, expecting that the LLN will give us greater confidence about our estimations. Very often, this is not possible:

- What will the weather be like tomorrow?
- Who will win the superbowl?
- Will I pass my next exam?

$$p(\theta|D) = \frac{P(D|\theta)\,p(\theta)}{P(D)}$$ ← prior

Estimate $P(X = \text{heads})$

$$p(p=1 \mid H) = \frac{p(H \mid p=1)\cdot 1/2}{p(D)} = 1/2$$

Summary

- Statistical Modelling is an ill-posed problem, but careful inspection can help us incorporate assumptions into our models.
- Parameters are often unknown yet they govern the behavior of a distribution.
- The (Weak) Law of Large Numbers justifies many results in statistical inference.
- However it can’t be universally applied.
- Bayesian inference changes the way we interpret parameters in a useful way under uncertainty or lack of repetition.