Statistical reasoning for humanities scholars

Lecture II: basics

Nalanda Academy Workshop 17-19 May 2025

Subodh Patil and Kapil Wankhede

Statistics in real life

Today, we begin the formal application lectures. **Please do view the recorded introductory lecture if you haven’t seen it already.**

- What follows will not use any advanced mathematics beyond what you may have learned up to Std. X.
- However, *this is not to say we won’t be covering advanced concepts!* You will have to work to understand the material.
- All important concepts will be highlighted in red at first appearance. Please do pay careful attention to them!

Statistics in real life

Important concepts covered in this lecture:

- *Frequentist* view of probability.
- *Estimators* and *bias*.
- *Sample variance.*
- *Normal, or Gaussian Probability Distribution Functions (PDFs).*
- *Independent* and *conditional probabilities.*
- *Statistical inference.*
- *Likelihood functions.*
- *Bayes theorem.*

Statistics in real life

Important concepts covered in this lecture:

- *Frequentist* view of probability.
- *Estimators* and *bias*.
- *Sample variance.*
- *Normal, or Gaussian Probability Distribution Functions (PDFs).*
- *Independent* and *conditional probabilities.*
- *Statistical inference.*
- *Likelihood functions.*
- *Bayes theorem.*

These are a lot of core concepts to digest, so do take the time to understand them!

Statistics in real life

We’re intuitively familiar with the concept of probabilities.

If the probability of something happening is p = 0.5, 0.3, 0.862 etc, we mean that ` on average’ the thing happens 50% of the time, 30% of the time, 86.2% of the time…

Implicit in this, is the idea of *frequency* of occurrence. A fancier word for this *interpretation* is known as the *frequentist* interpretation of probabilities. However, there are others!

Statistics in real life

Suppose we flip a ` fair’ coin. About half the time, it will land ` heads’, and the other half, ` tails’ – the origin of this terminology is European coinage, where the king or queen’s face is on one side of the coin…

Statistics in real life

Suppose we flip a ` fair’ coin. About half the time, it will land heads. But what if we wanted to *test* whether the coin is fair?

Let’s say every time a coin lands heads, we assign the outcome 1, and for tails, 0.

Statistics in real life

Suppose we flip a ` fair’ coin. About half the time, it will land heads. But what if we wanted to *test* this?

Ten coin flips – HTTHTTHTHT = 4, average # of heads = 4/10

Statistics in real life

Suppose we flip a ` fair’ coin. About half the time, it will land heads. But what if we wanted to *test* this?

Another ten flips – HTHHTHHTHT = 6, average # of heads = 6/10

Statistics in real life

Suppose we flip a ` fair’ coin. About half the time, it will land heads. But what if we wanted to *test* this?

Of course, if we did this for an *infinite* number of coin flips, we would expect the average to tend to 0.5... but in reality, we never have the luxury of an infinite number of trials.

Statistics in real life

Suppose we flip a ` fair’ coin. About half the time, it will land heads. But what if we wanted to *test* this?

We only ever deal with *finite* samples… for which there is an associated *sample variance*. This is the random fluctuations as we repeat the trials around the *expected* value of obtaining heads half the time.

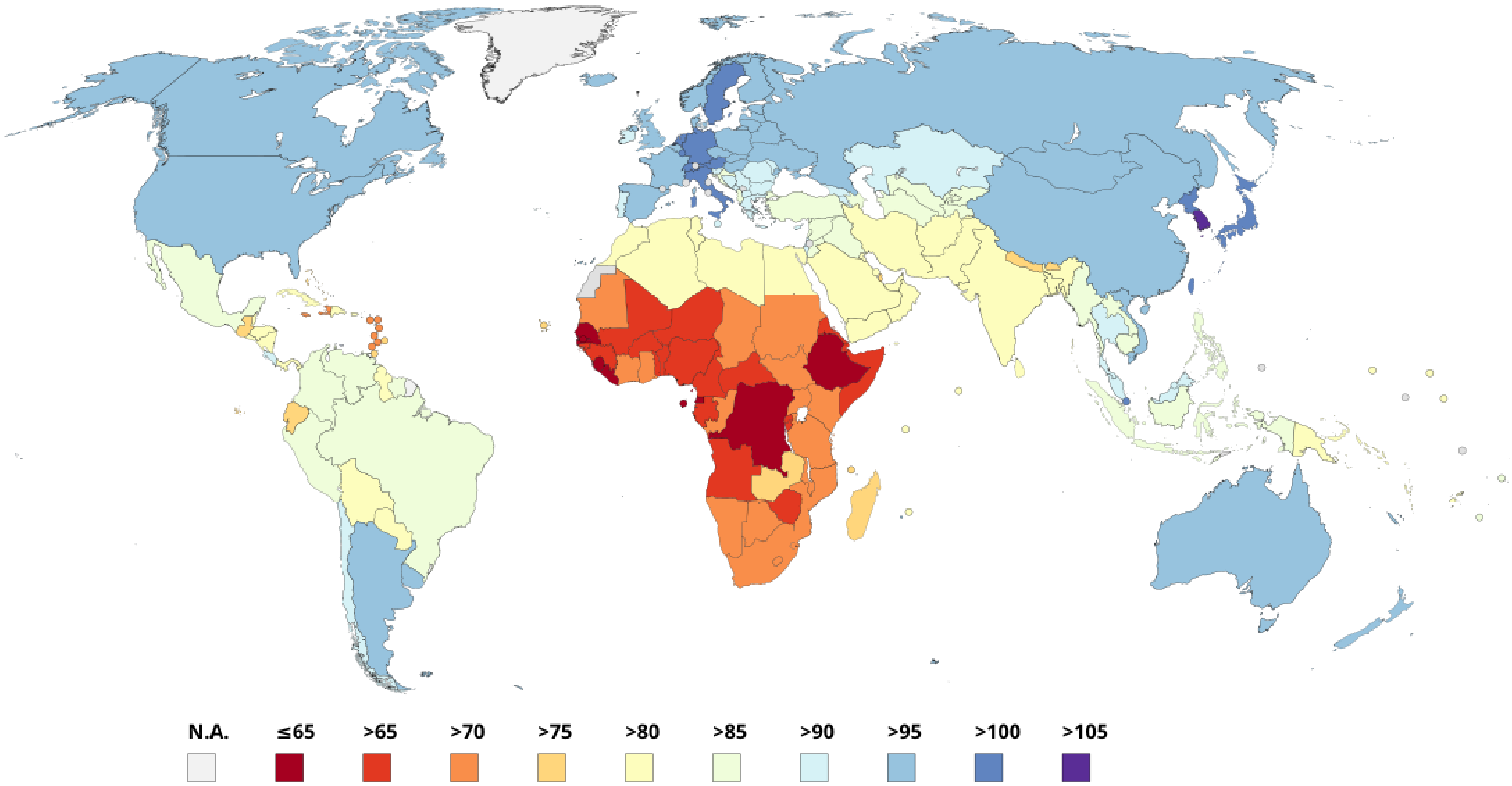
Statistics in real life

Suppose we flip a ` fair’ coin. About half the time, it will land heads. But what if we wanted to *test* this?

The quantity that averages the outcome by assigning 1 for every occurrence of a heads and 0 for every tail is known as an *estimator.* Estimators can either be *biased*, or *unbiased*, depending on whether they accurately approach the quantity we’re estimating.

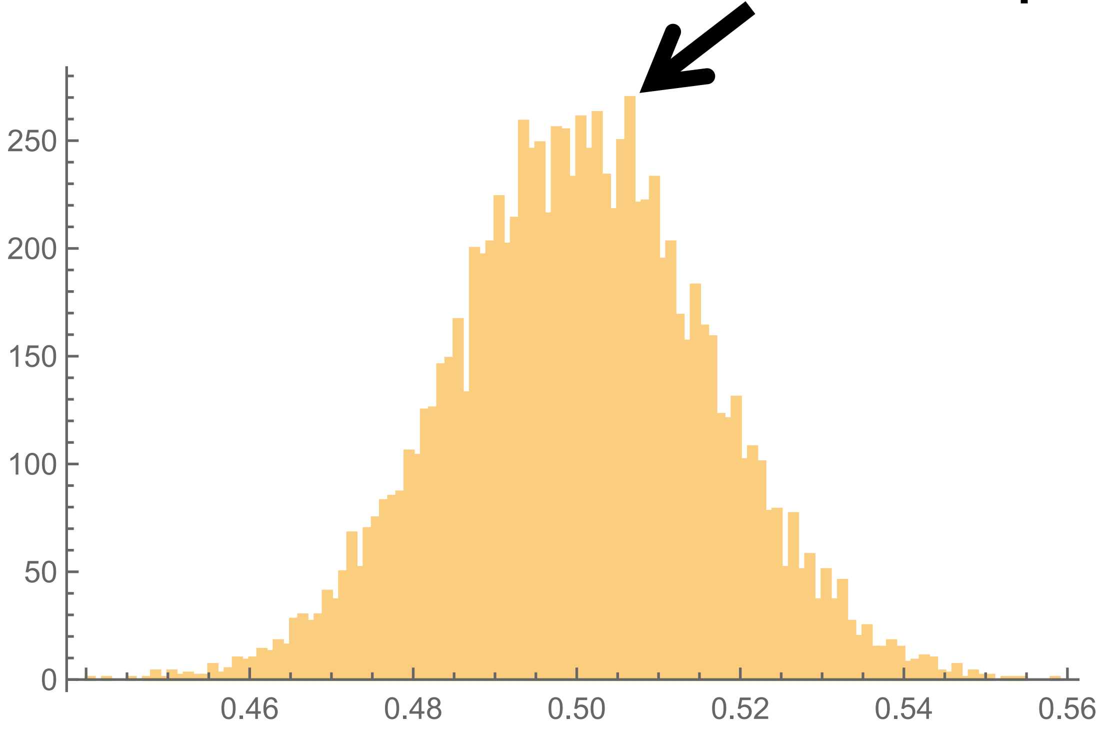
Statistics in real life

Biased/unbiased estimators: consider IQ international scores



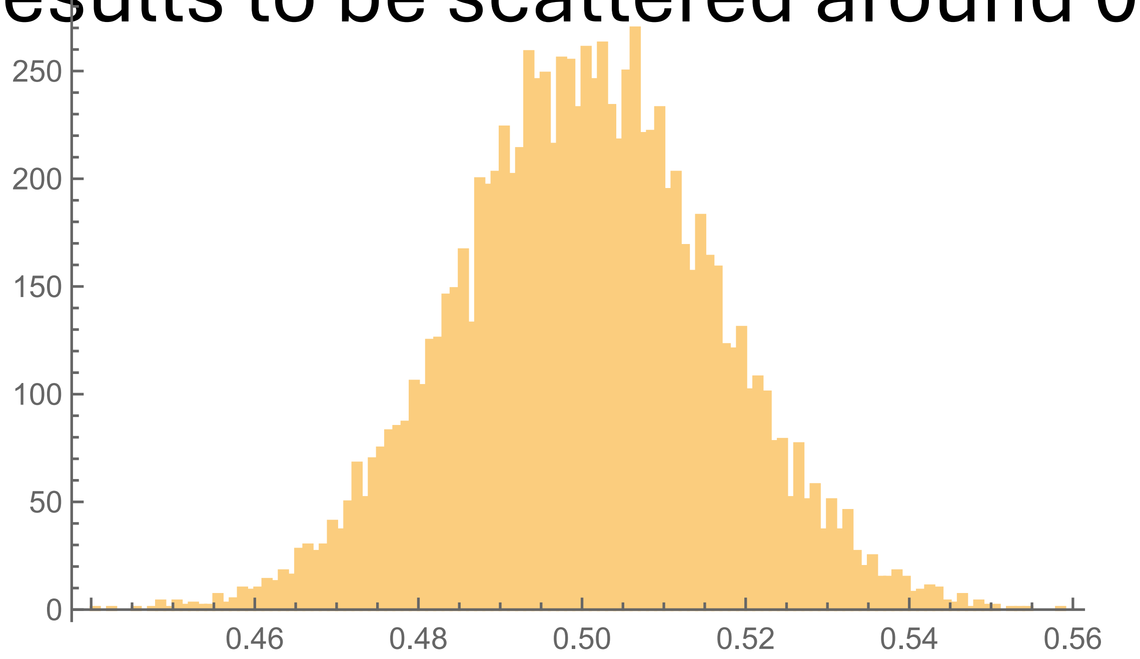
Statistics in real life

Let’s now do 10,000 trials of a 1,000 consecutive flips each... the result most often obtained in this experiment is an unfair coin!



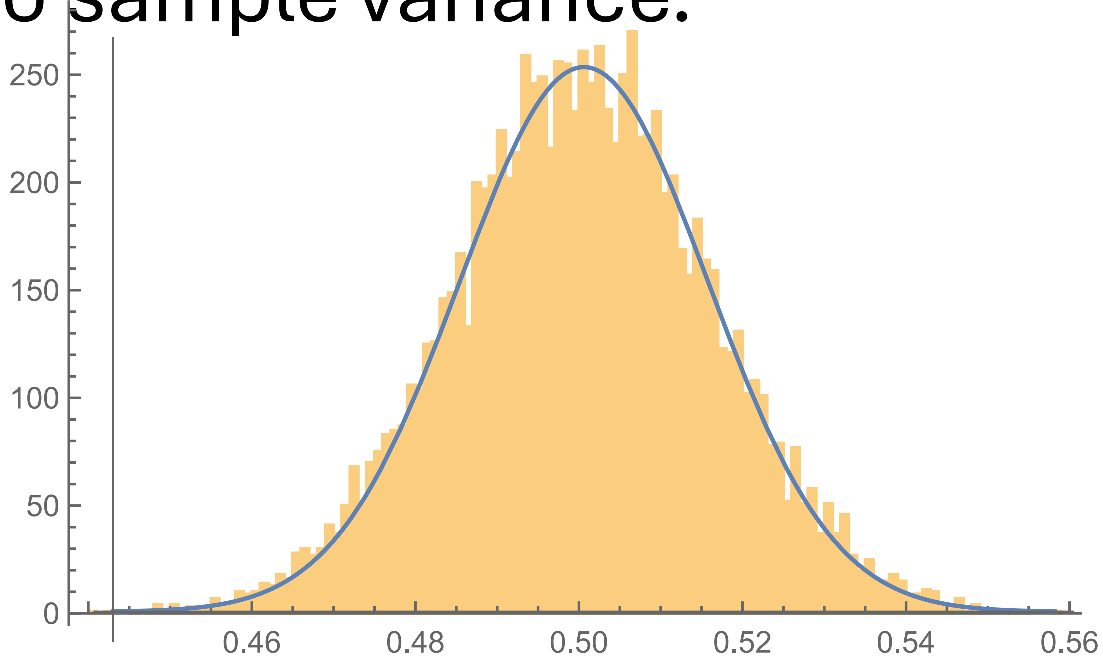
Statistics in real life

The difference between the outcomes observed for a finite number of trials is known as *sample variance*. For N trials, we expect the results to be scattered around 0.5 with a width of $\frac{1}{2\sqrt{N}}$



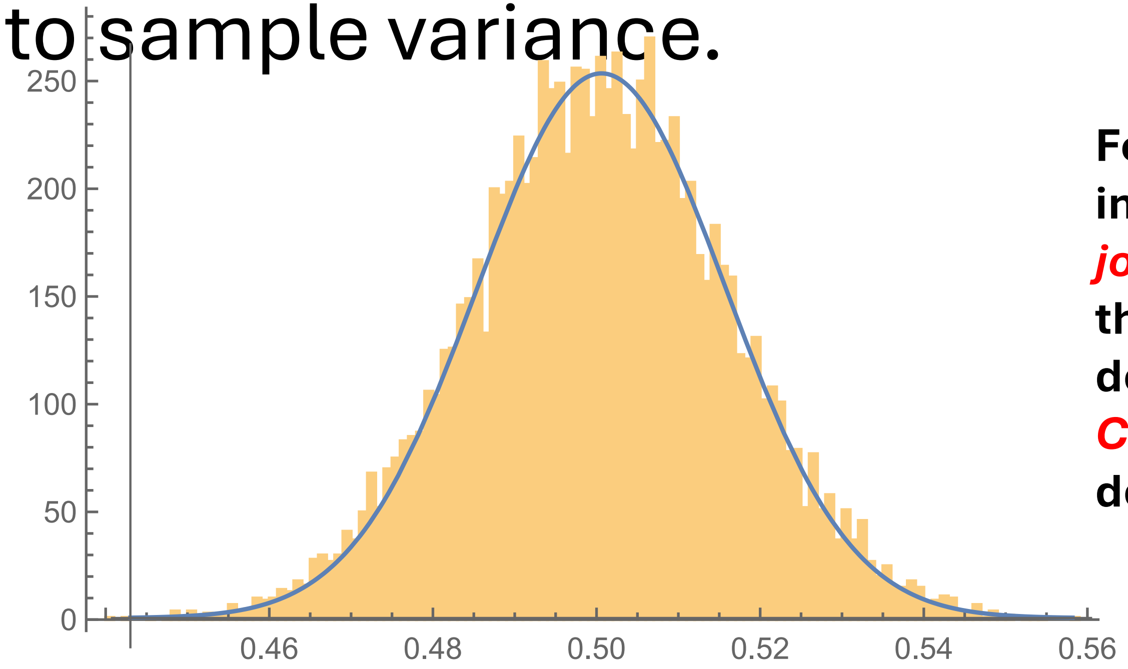
Statistics in real life

After an *infinite* number of trials, the outcomes will tend towards the *Normal*, or *Gaussian* distribution. The fluctuations around which for a finite number of trials correspond to sample variance.



Statistics in real life

After an *infinite* number of trials, the outcomes will tend towards the *Normal*, or *Gaussian* distribution. The fluctuations around which for a finite number of trials correspond to sample variance.



For a sufficiently large number of independent random variables, the *joint distribution* will always tend to the Normal distribution. This very deep fact about reality is called the *Central Limit Thoerem*, and will be derived in the advanced classes.

Statistics in real life

Through a simple example, we just touched upon several core concepts, which we will now explain in more detail.

Consider some *set* of outcomes {X} for some process. A set is simply a collection of objects (e.g. numbers).

- Coin toss: {X} = {H,T}
- Dice throw: {X} = {1,2,3,4,5,6}
- Heights of Std. X. children in meters: {X} = {1.32, 1.50, 1.48...}

Statistics in real life

If we consider x to be a *random variable* whose outcomes lie in the set {X}, then we can assign a probability for each outcome:

- Coin: {X} = {H,T}. For a fair coin: p(H) = ½, p(T) = ½
- Dice: {X} = {1,2,3,4,5,6}. For fair dice: p(1) = p(2) = … p(6) = 1/6

The assignment of a probability to each possible outcome is known as a *probability distribution function* or PDF for short.

Statistics in real life

Random variables, and their respective outcomes can either be *discrete* or *continuous*. We will mostly be dealing with discrete random variables in these lectures (*avoids having to use calculus, however subsequent advanced modules will cover this as well*).

- Coin: {X} = {H,T}. For a fair coin: p(H) = ½, p(T) = ½
- Dice: {X} = {1,2,3,4,5,6}. For fair dice: p(1) = p(2) = ... p(6) = 1/6

Statistics in real life

Given a random variable x, and a set of outcomes $\{X\}$, a PDF assigns some number (a probability) to each outcome $p: \{X\} \to [0,1]$

Suppose there are N possible outcomes (can be very big, even ∞!)

- Coin: $\{X\} = \{H,T\}$; $N = 2$
- Dice: $\{X\} = \{1,2,3,4,5,6\}$; $N = 6$
- Calls per hour in a call center: $\{X\} = \{1,2,3,4,\ldots\}$; $N = \infty$

Statistics in real life

Given a random variable x, and a set of outcomes {X}, a PDF assigns some number (a probability) to each outcome p: {X} -> [0,1]

Such that if $X = \{x_1, x_2, ..., x_N\}$, then $\sum_{i=0}^{N} p(x_i) = 1$

All possible outcomes will certainly happen, i.e. will occur with probability 1.

Notation: $\sum_{i=1}^{N} f(x_i) = f(x_1) + f(x_2) + f(x_3)... + f(x_N)$

Statistics in real life

Given a random variable x, and a set of outcomes {X}, a PDF assigns some number (a probability) to each outcome p: {X} -> [0,1]

Such that if $X = \{x_1, x_2, ..., x_N\}$, then $\sum_{i=0}^{N} p(x_i) = 1$

Average (angle bracket notation): $\langle x \rangle = \sum_{i=1}^{N} x_i p(x_i)$

Dice: $\langle x \rangle = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$

Statistics in real life

Given a random variable x, and a set of outcomes {X}, a PDF assigns some number (a probability) to each outcome p: {X} -> [0,1]

Such that if $X = \{x_1, x_2, ..., x_N\}$, then $\sum_{i=0}^{N} p(x_i) = 1$

Average (angle bracket notation): $\langle x \rangle = \sum_{i=1}^{N} x_i p(x_i)$

Dice: $\langle x \rangle = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$

Statistics in real life

But we might be interested in more complicated quantities, such as the average of the *square* (or any other power) of the dice roll, or perhaps whether the roll of the dice is an even number etc.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\} = \{1, 2, 3, 4, 5, 6\}; \quad p(x_i) = \frac{1}{6}$$

$$\langle x^2 \rangle = \sum_{i=1}^{6} x_i^2 p(x_i) = \sum_{i=1}^{6} \frac{x_i^2}{6} = \frac{91}{6}$$

$$\langle x^4 \rangle = \sum_{i=1}^{6} x_i^4 p(x_i) = \sum_{i=1}^{6} \frac{x_i^4}{6} = \frac{2275}{6}$$

etc. $\langle x^n \rangle$ is referred to as the n'th *moment* of the distribution.

Statistics in real life

One particular quantity we're interested in is how certain quantities differ from the average. That is, its *deviation* from the average: $x - \langle x \rangle$

... but the average of this vanishes!

$$\begin{aligned} \langle x - \langle x \rangle \rangle = \textstyle\sum_{i=1}^{6} (x_i - \langle x \rangle) p(x_i) &= \textstyle\sum_{i=1}^{6} x_i p(x_i) - \langle x \rangle \sum_{i=1}^{6} p(x_i) \\ &= \langle x \rangle - \langle x \rangle \textstyle\sum_{i=1}^{6} p(x_i) \\ &= \langle x \rangle - \langle x \rangle \\ &= 0 \end{aligned}$$

Statistics in real life

One particular quantity we're interested in is how certain quantities differ from the average. That is, its *deviation* from the average: $x - \langle x \rangle$

... but the average of this vanishes!
How about the average of $(x - \langle x \rangle)^2$ instead?
Can show that (exercise): $\langle (x - \langle x \rangle)^2 \rangle \;\; = \langle x^2 \rangle - \langle x \rangle^2$

This is known as the *variance* of x, its square root (the root mean of the squares [rms]) is known as the *standard deviation*.

Statistics in real life

Quick recap:

- Random variable x, set of outcomes {X}, a PDF assigns a probability to each outcome p: {X} -> [0,1]

- If $X = \{x_1, x_2, ..., x_N\}$, then $\sum_{i=0}^{N} p(x_i) = 1$

- Angle bracket notation $\langle f(x) \rangle = \sum_{i=1}^{N} f(x_i) p(x_i)$

- The n’th moment of PDF is given by $\langle x^n \rangle = \sum_{i=1}^{N} x_i^n p(x_i)$

Statistics in real life

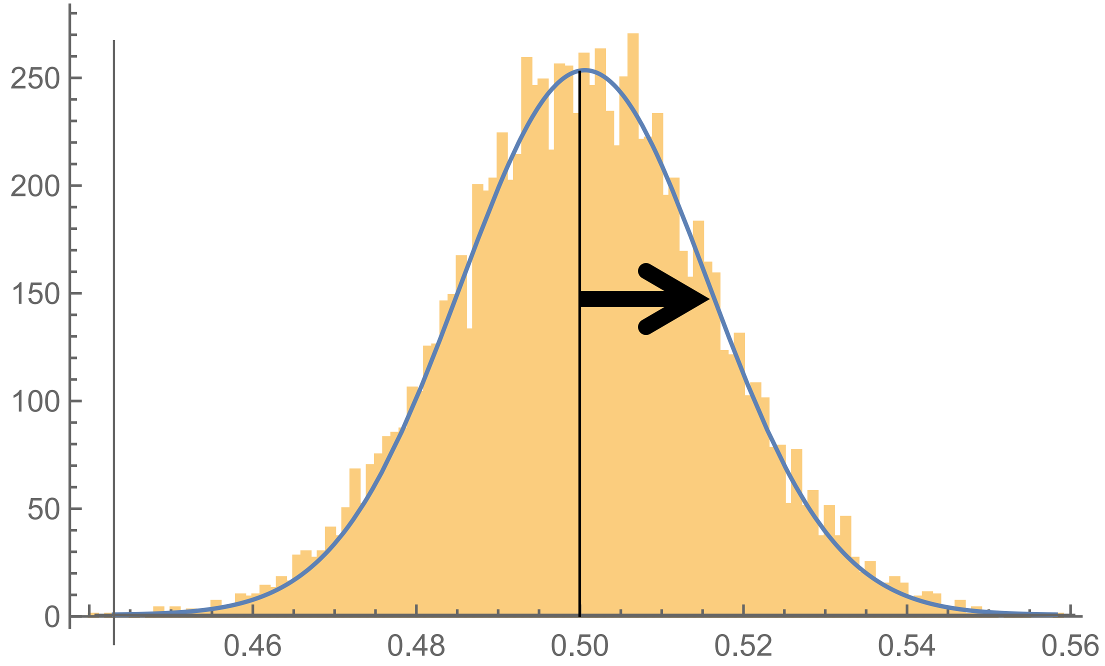
Quick recap:

- Of particular interest are the average and the variance
- Average: $\langle x \rangle = \sum_{i=1}^{N} x_i p(x_i)$
- Variance: $\langle (x - \langle x \rangle)^2 \rangle = \sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)$
- Standard deviation = $\sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)}$

Statistics in real life

Example: limiting case of infinite repeats of the average of 1000 consecutive coin tosses:

- Average = 0.5
- Standard deviation = 0.0158



Statistics in real life

This is all fine in *theory*… but what about in practice, where we may not know anything about the underlying PDF? Suppose for instance, you’re asked to *test* whether a die or a coin is fair?

We then have to *estimate* the underlying moments of the PDF. From this, we can reasonably reconstruct the full PDF if we know enough of the moments. This process is known as *statistical inference*.

Statistics in real life

In order to estimate the moments of the PDF we're dealing with, we make use of *estimators:*

For example, if we have *N* trials of a given random process, then it seems reasonable to estimate the average by

$$\langle x\rangle_E = \frac{1}{N}\sum_{i=1}^{N} x_i$$

As $N \to \infty$, whatever the underlying PDF $\langle x\rangle_E \to \langle x\rangle$

The *estimator* of the average tends to the *true average*. It is unbiased.

Statistics in real life

In order to estimate the moments of the PDF we're dealing with, we make use of *estimators:*

Now consider an estimator for the mean square:

$$\langle (x - \langle x \rangle)^2 \rangle_E = \frac{1}{N} \sum_{i=1}^{N} (x_i - \langle x \rangle_E)^2$$

Where $\langle x \rangle_E$ is our estimator for the average. It turns out that as $N \to \infty$ this does *not* tend to the *true variance*. It is said to be biased.

Statistics in real life

In order to estimate the moments of the PDF we’re dealing with, we make use of *estimators:*

However, in the advanced lectures, we can *prove* that

$$\langle (x - \langle x \rangle)^2 \rangle_E = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \langle x \rangle_E)^2$$

is an unbiased estimator. As $N \to \infty$, it tends to the true variance.

Statistics in real life

Another important concept in statistical inference is the so-called *likelihood function.*

Suppose someone handed you a coin, but didn't tell you if it was fair. That is, its probability of landing a heads is some unknown value between 0 and 1: $p(H) = \theta, \quad 0 \leq \theta \leq 1$
And so therefore $p(T) = 1 - \theta$
We rewrite this as $p(H;\theta) = \theta,\ p(T;\theta) = 1 - \theta$ where$\theta = \frac{1}{2}$for the case of a fair coin.

Statistics in real life

Another important concept in statistical inference is the so-called *likelihood function.*

The likelihood function $\mathcal{L}$ views the *outcome* of the trial as a fixed given, but allows the probability of the outcome to be a function of the parameter of the PDF:

$$\mathcal{L}(\theta|X) = p(X;\theta)$$

Statistics in real life

Another important concept in statistical inference is the so-called *likelihood function.*

Imagine we did 8 coin flips, and found the outcome HTHTHHHH

$$\mathcal{L}(\theta | HTHTHHHH) = \theta^6 (1-\theta)^2$$

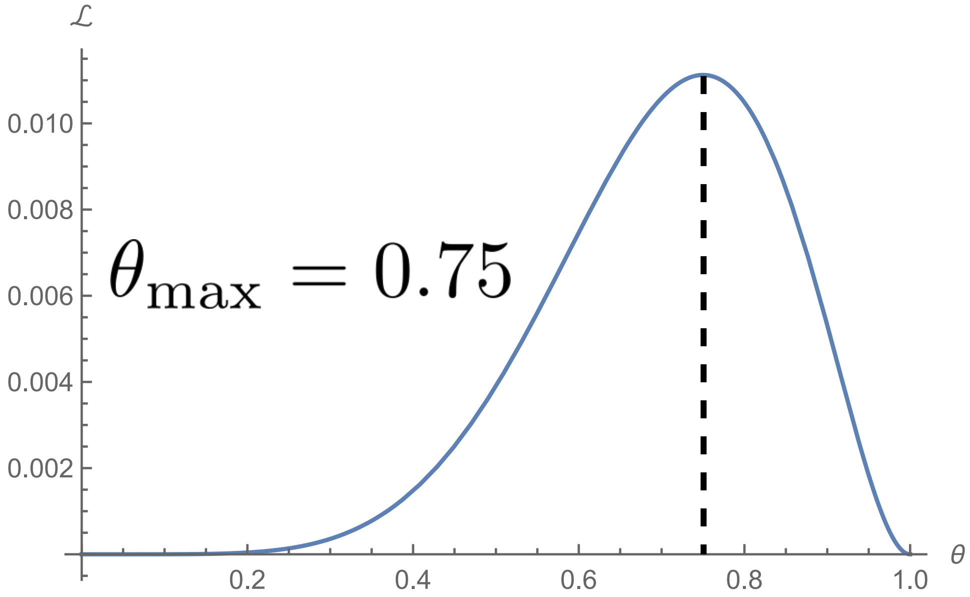
If the coin is fair,$\theta = 1/2$ so that $\mathcal{L} = \frac{1}{2}^2 \left(1 - \frac{1}{2}\right)^6 = \frac{1}{2^8} = \frac{1}{256}$

… *but we don’t know*θ. Suppose we had to estimate it.

Statistics in real life

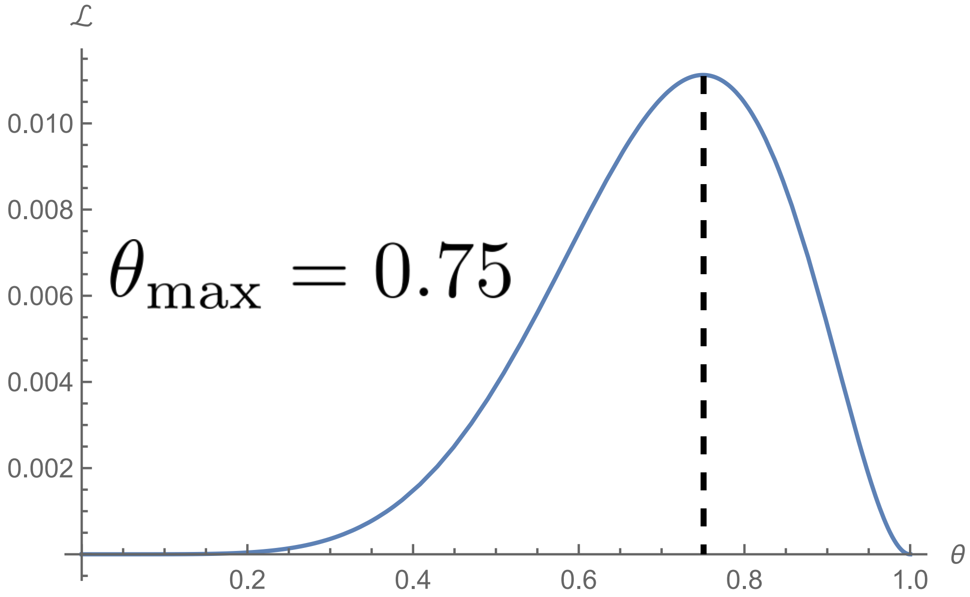
We can imagine inferring θ by finding the value that maximizes the likelihood function. This is known as a *maximum likelihood estimate:*

$$\mathcal{L}(\theta|HTHTHHHH) = \theta^6(1-\theta)^2$$



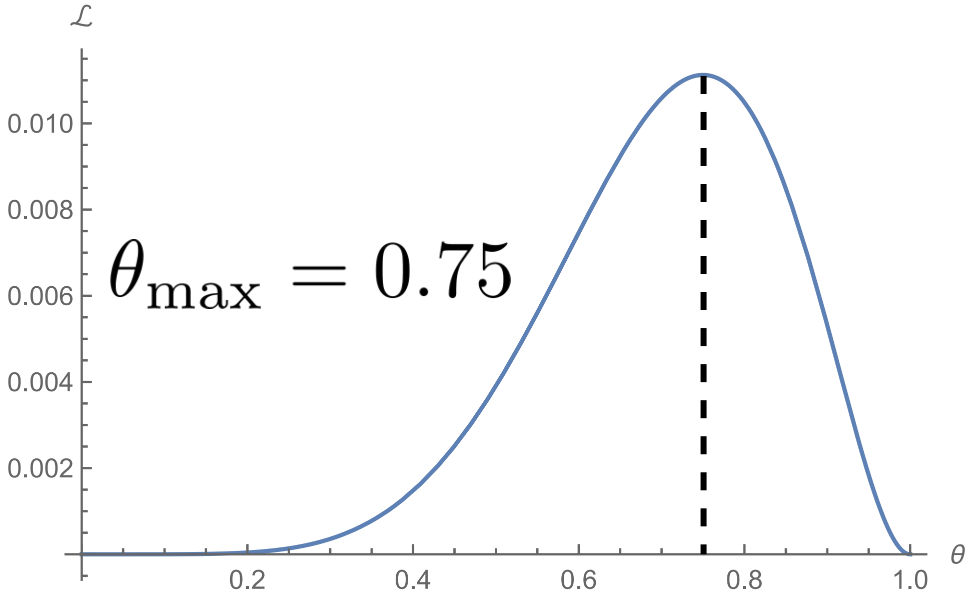
Statistics in real life

The value for θ that maximizes the *likelihood* i.e. probability of the observed outcome of six heads and two tails is$\theta = 0.75$… that is, *if all we have is this one outcome, we're liable to conclude the coin is unfair.*



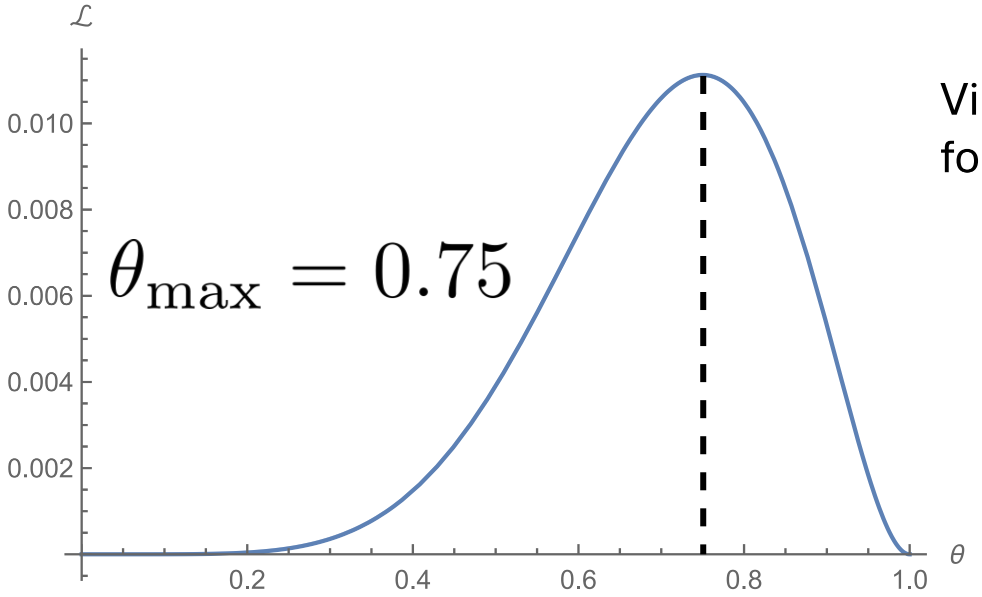
Statistics in real life

Here we again come up against the perils of doing statistics with a finite sample. Is the coin really unfair, or is it an artifact of only doing a small number of trials?



Statistics in real life

Exercise to try at home: Pull out a coin from your pocket, and repeat this exercise for 5 coin flips, and do a maximum likelihood estimation. Now do the same for 10 coin flips, the 20. What do you notice?



Visit https://www.wolframalpha.com/ for an online resource to plot such figures.

Statistics in real life

Maximum likelihood estimation is at the core of statistical inference. We often *observe* the outcomes of some random process, and we’re trying to deduce what the parameters of the random process are.

- Average heights and standard deviations from a sample.
- How many infected individuals in a locality from the results of random testing?
- What are the parameters of some physical model given the outcomes of some experiments or observations?

Statistics in real life

There are also other types of questions in statistical inference, such as testing whether a particular hypothesis is true. E.g. does a particular chemical *cause* cancer? Typically, one proposes a hypothesis, and looks for evidence as to whether it’s true. A very important tool in this endeavour is *Bayesian hypothesis testing*. For which we need to introduce a few more basic concepts…

Statistics in real life

Often, we are interested in situations involving more than one random variable. For example, let’s say we have multiple dice, or one dice and one coin, etc:

$p(x, y)$ = probability that both x and y occur

e.g. one dice rolls a 6 and another dice rolls a 2, or the probability that you roll a six and then flip a heads.

Statistics in real life

Both of these examples involve *independent variables*. The probability of both events occurring is just the probability of one occurring times the probability of the other:

$$p(x, y) = p(x)p(y)$$

e.g. one dice rolls a 6 and another dice rolls a 2 = 1/6 x 1/6 = 1/36

Probability that you roll a six and then flip a heads = 1/6 x ½ = 1/12

Statistics in real life

But sometimes this isn’t the case. Consider a deck of cards: what is the probability of drawing a queen followed by a 4 of any suite?

$$p(Q,4) = \frac{4}{52}\frac{4}{51} \neq \frac{4}{52} \times \frac{4}{52}$$

$$p(x,y) \neq p(x)p(y)$$

Statistics in real life

But sometimes this isn’t the case. Consider a deck of cards: what is the probability of drawing two consecutive queens of any suite?

$$p(Q_1, Q_2) = \frac{4}{52}\frac{3}{51} \neq \frac{4}{52} \times \frac{4}{52}$$

$$p(x, y) \neq p(x)p(y)$$

Statistics in real life

In this case, we talk of *conditional probabilities*... that is, what is the probability that x has happened *given* y, or vice-versa.

$$p(x, y) = p(x|y)p(y)$$

In words, the probability that *both* x and y occur, is the probability that x occurs *given* y, times the probability that y occurs.

Statistics in real life

In this case, we talk of *conditional probabilities*... that is, what is the probability that x has happened *given* y, or vice-versa.

$$p(x, y) = p(y|x)p(x)$$

In words, the probability that *both* x and y occur, is the probability that y occurs *given* x, times the probability that x occurs.

Statistics in real life

Q) What is the probability that you are infected with Covid *given* that you've tested positive?

$$\begin{aligned} p(x, y) &= p(y|x)p(x) \\ &= p(x|y)p(y) \end{aligned}$$

It helps to say this out aloud to yourself to understand the *tautology* that these equations imply. Because what follows is at once one of the most profound, yet perhaps not obvious results in statistics.

Statistics in real life

$$\begin{aligned} p(x,y) &= p(y|x)p(x) \\ &= p(x|y)p(y) \end{aligned}$$

so:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

This is known as Bayes theorem, and it’s implications are profound!