Statistics!

Probability

- Statistics is all about probability
- What exactly is probability?
  - Well it’s how probable something is

Probability

- Statistics is all about probability
- What exactly is probability?
  - Well it’s how probable something is
- Let’s think about it a different way
  - Probability is how likely something is to occur
  - You can think of probability as the % chance something happens or is in a certain state

- Ok so if there is an event, with multiple outcomes
    - There’s a probability of each outcome happening?

- Ok so if there is an event, with multiple outcomes
    - There’s a probability of each outcome happening?

- Now let’s logic a little further
    - If we add the probability of each possible outcome, what should we get?

- Ok so if there is an event, with multiple outcomes
    - There’s a probability of each outcome happening?

- Now let’s logic a little further
    - If we add the probability of each possible outcome, what should we get?
    - $\sum_{\text{all outcomes}}$ Probability = 1

- Ok so if there is an event, with multiple outcomes
  - There’s a probability of each outcome happening?

- Now let’s logic a little further
  - If we add the probability of each possible outcome, what should we get?
  - $\sum_{\text{all outcomes}}$ Probability = 1
  - We just logic’ed out a fundamental theorem to probability theory
  - Law of total probability

Let’s keep going

- What’s the lowest the probability of something happening can be?

- The what’s the highest the probability of something happening can be?

Let’s keep going

- What’s the lowest the probability of something happening can be?

- The what’s the highest the probability of something happening can be?

The probability of something must be between 0 and 1!

And the sum of the probability of all outcomes must be 1

Let’s use this to solve a problem

- Say there’s a box full of different colored shirts, and the probability of pulling a red shirt out is 0.4
- What’s the probability of pulling out a shirt that’s not red?

- Say there’s a box full of different colored shirts, and the probability of pulling a red shirt out is 0.4
- What’s the probability of pulling out a shirt that’s not red?

We’re considering 2 possible outcomes,

Red and not red

- Say there's a box full of different colored shirts, and the probability of pulling a red shirt out is 0.4
- What's the probability of pulling out a shirt that's not red?

We're considering 2 possible outcomes,

Red and not red

P(red) = 0.4

- Say there’s a box full of different colored shirts, and the probability of pulling a red shirt out is 0.4
- What’s the probability of pulling out a shirt that’s not red?

We’re considering 2 possible outcomes,

Red and not red

P(red) = 0.4

P(red or not red) = P(red) + P(not red)

- Say there's a box full of different colored shirts, and the probability of pulling a red shirt out is 0.4
- What's the probability of pulling out a shirt that's not red?

We're considering 2 possible outcomes,

Red and not red

P(red) = 0.4

P(red or not red) = P(red) + P(not red)

P(red or not red) = 1 = P(red) + P(not red) **Law of total probability**

P(not red) = 1 - P(red) = 1 - 0.4 = 0.6

Probability of discrete outcomes

- Discrete outcomes are things that are specific
  - Red or not red
  - Heads or tails for a coin
  - 1-6 for a dice
  - Also, counting numbers, or integers

- The function that describes the probability of each outcome is called the Probability Mass Function

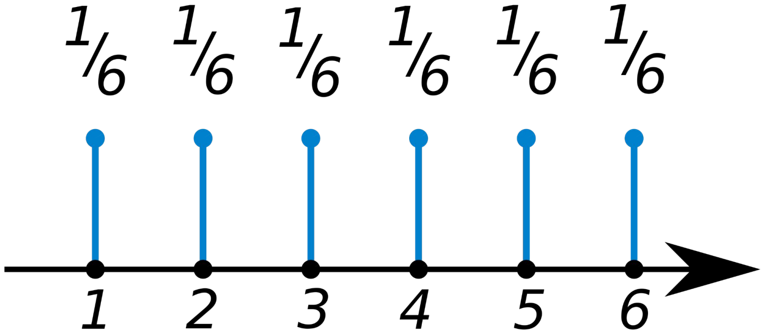
Probability Mass Function (PMF)

PMF for a coin flip

Heads = 1, tails = 0

$$p_X(x) = \begin{cases} \frac{1}{2}, & x = 0, \\ \frac{1}{2}, & x = 1, \\ 0, & x \notin \{0, 1\}. \end{cases}$$

PMF for a dice roll



What about outcomes that are not discrete?

Say I have a digital stopwatch that only goes to the number of seconds,

What is the probability that I could press stop and stop it exactly at 3 seconds?

What about outcomes that are not discrete?

Say I have a digital stopwatch that only goes to the number of seconds,

What is the probability that I could press stop and stop it exactly at 3 seconds?

Probably ok

What about outcomes that are not discrete?

Say I have a digital stopwatch that only goes to the number of seconds,

What is the probability that I could press stop and stop it exactly at 3 seconds?

Probably ok

What if the stopwatch was instead analog

What is the probability that I could it stop at exactly 3 seconds?

What about outcomes that are not discrete?

Say I have a digital stopwatch that only goes to the number of seconds,

What is the probability that I could press stop and stop it exactly at 3 seconds?

Probably ok

What if the stopwatch was instead analog

What is the probability that I could it stop at exactly 3 seconds?

0

What about outcomes that are not discrete?

Continuous outcomes, are the opposite of discrete

Ex: all decimal numbers

There are infinite outcomes

To have a defined probability it would have to be over an interval

What’s the probability I can stop the analog watch, between 2 s and 4 s?

What about outcomes that are not discrete?

Continuous outcomes, are the opposite of discrete

Ex: all decimal numbers

There are infinite outcomes

To have a defined probability it would have to be over an interval

What’s the probability I can stop that analog watch, between 2 s and 4 s?

0

What about outcomes that are not discrete?

Continuous outcomes, are the opposite of discrete

Ex: all decimal numbers

There are infinite outcomes

To have a defined probability it would have to be over an interval

What's the probability I can stop that analog watch, between 2 s and 4 s?

0

It has no numbers, remember

But if it did have numbers then the probability would be > 0

If probability is only non-zero over intervals, how do we know how probable something is around some value?

- Probability density function (PDF)
    - $p(x) = dP/dx$
    - The derivative of the probability

If probability is only non-zero over intervals, how do we know how probable something is around some value?

- Probability density function (PDF)
    - $p(x) = dP/dx$
    - The derivative of the probability

The probability over an interval is then the definite integral

$$P(-2 \leq x \leq 2) = \int_{-2}^{2} p(x)\, dx$$

Statistics time

Statistics is the study of data

Data is a set of observations

Let’s say we have a sample of data,

$x_i \sim \{x_1, x_2, .. x_N\}$

What are some statistical measures you know?

Statistics time

Let's say we have a sample of data,

$x_i \sim \{x_1, x_2, .. x_N\}$

$$mean = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$x_i \sim \{1, 5, 3\}$

Mean = (1 + 5 + 3) / 3 = 3

$$Variance = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

Var = $((1-3)^2 + (5-3)^2 + (3-3)^2)$ / 3 = (4 + 4 + 0)/3 = 8/3

What is variance?

$$Variance = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

It’s the average deviation from the average squared

- The average of the value $(x - <x>)^2$

You may more often here the term standard deviation

standard deviation = σ = √(Variance)

This is the average distance to the mean

These are measures of how spread out your data is

Going one further there's also a measure called skewness

The average deviation from the mean cubed

$$skew = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

This is a measure of how asymmetric your data is

Moments

The mean, variance, and skew of your data are also known as

the first, second, and third moments

These give you the

- Location
- Spread
- Asymmetry

Of the underlying distribution of your data

Data taking from the perspective of probability theory

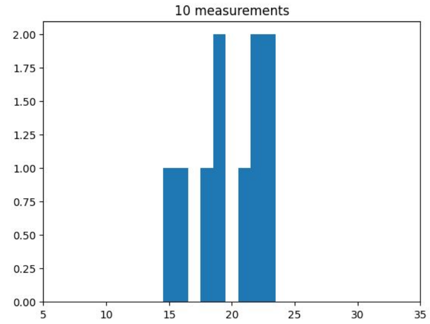
When you make a measurement for a data point you are sampling from an underlying distribution

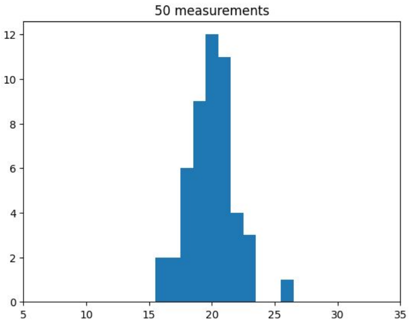
As you take more data the underlying distribution becomes more clear

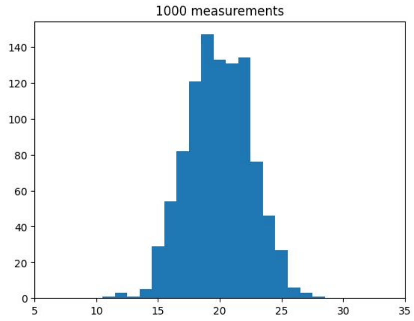
Data taking from the perspective of probability theory

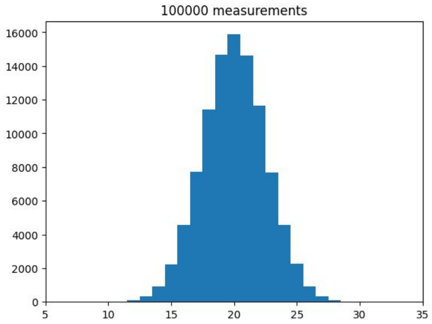
When you make a measurement for a data point you are sampling from an underlying distribution

As you take more data the underlying distribution becomes more clear









Histograms

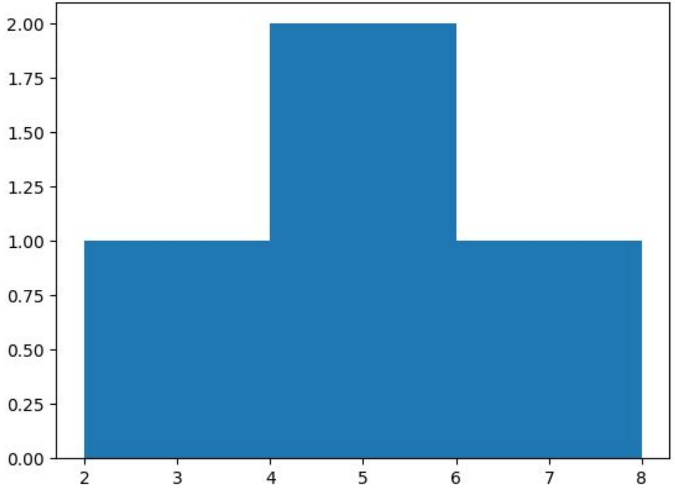
- Histograms are a key and simple tool in statistics
- Given a list of values and a set of bins,
- A histogram is the number of values in each bin

Histograms

- Histograms are a key and simple tool in statistics
- Given a list of values and a set of bins,
- A histogram is the number of values in each bin

x = [2.3, 4.5, 5.5, 7.1]

Bins = [2, 4, 6, 8]

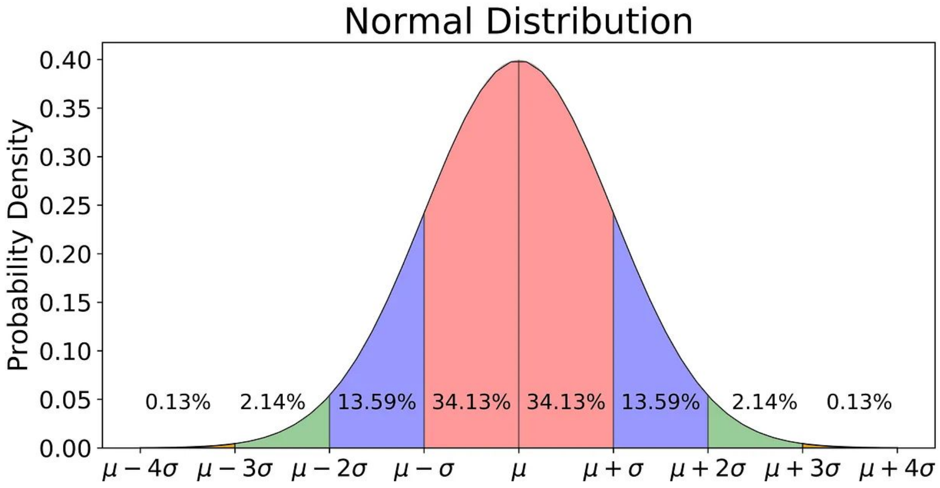


Normal distribution

One of the most common probability distributions is a Normal distribution

Sometimes also known as a Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Normal distribution

Normal distributions are so common due to something called the

Central Limit Theorem -

If you take the average of many data samples all from the same underlying distribution, those averages will follow a Normal distribution

Let’s move on to the tutorial to see an example of this