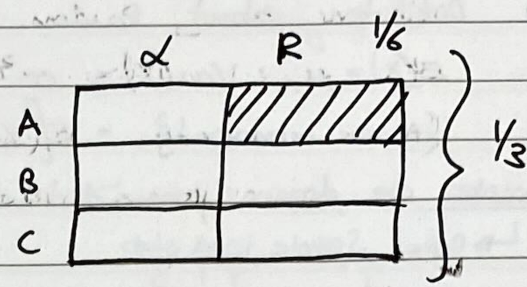
Theory Recap:

① Conditional Probability + Bayes (Short)



- A is an event, B is another event

① P(A), P(B). $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$1/6 / 1/2 = 1/3$

② $P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$

(Never bother to learn Bayes Rule!)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Bigg\} \text{ Bayes Rule}$$

- Rare events question ↓ Unintuitive answer!

③
- why? Interesting? Data Science Interview (3 simple steps to these)
- ① Define A and B. A: cloudy, B rain.
  ② Write down exactly what you have P(A), P(B), P(A|B)
  ③ Plug it all into Bayes Rule simple but really fiddly!

~ pause ~

---

② Random Variables and Distributions

"take different values due to randomness"

① • what is a random variable? An object or quantity depending on randomness.

① X number when you roll a dice. Can be {1, 2, ..., 6}

② ② X the height of an individual picked at random

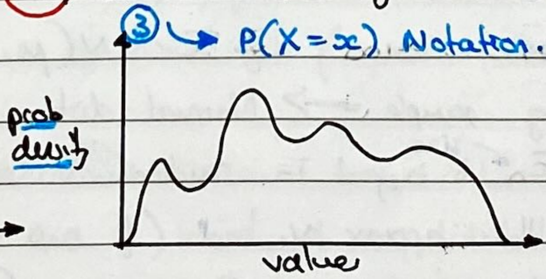
• Random variables can be discrete or continuous → Someone to define each

Difference between them? x a realisation of X is either from a predefined list or not

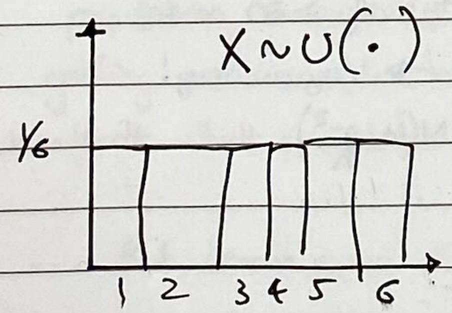
X, $x = 3$

③ → $P(X = x)$. Notation. $\sum_{i=1}^{n} p(X = x_i) = 1$ notation

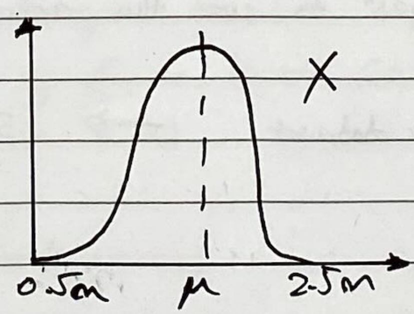
④ X follow distribution →



Distributions



$x = 1, 2 \ldots$



... Also written μ_X, σ_X^2

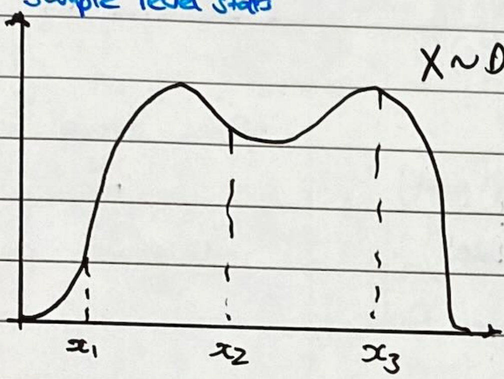
⊙ Distributions about Random variable have properties.

- $E(X) = \mu$, $Var(X) = \sigma^2$ } Population Moments, Population mean

  {Never memorise!} $= E[(X-\mu)^2] = E(X^2) - E(X)^2$ Population sample

<u>Samples</u> are drawn from distributions.

↳ Ogive! Sample level stats



$X \sim D(\cdot)$

x_1, x_2, x_3

$(x_i : i = 1, \dots, n)$

{ Statistics is about sample to distribution!

$x_1, x_2, \dots, x_n \longrightarrow \underbrace{X \sim D(\cdot)}_{?}$ ⟹ <u>Claims about the world</u>

<u>Problems</u> ① Don't know the distribution

② Most distributions are messy so <u>difficult to do inference</u>

↳ Hard to do sample → population!

③ <u>Central Limit Theorem</u>:

✱ Arguably the most beautiful thing in mathematics/statistics... ✱

⚠ – <u>Whatever</u> the underlying distribution, you can still do inference because the sample mean will be normally distributed as n gets large

– <u>Gateway to all statistics</u>

"Approximately distributed"

- Sample: $x_1, x_2, \dots, x_{100}$ } $\bar{x} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$ } Explain this visually
- Lots and lots of sample ⟶ Normal dist
  - Sample mean $\bar{x}_n$ size is in fact a random variable with its own distribution
- If n is small then approx v. bad (If n is 1 then just the underlying dist)
- n must be "large" to use the approx (typically > 30 or 40)

  ↳ Very arbitrary!

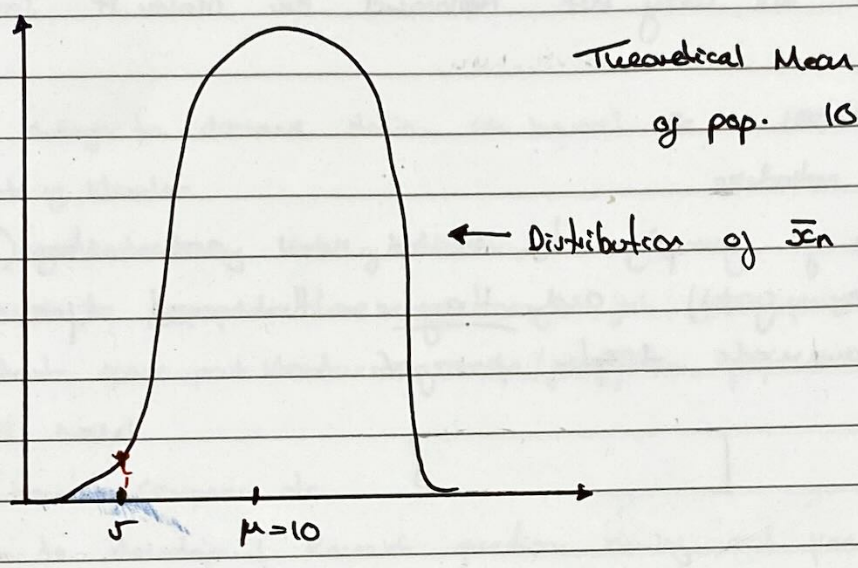
$x_1, x_2, \dots, x_n$ } Your dataset ⟶ $\bar{x}_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$ think it should be

actually collected

Idea of hypothesis testing ...

Only of time

One example $\bar{x}_n \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n})$



- $x_1, x_2, \dots, x_{1000}$ - $\bar{x}_{1000}$. 5 . 5 is different to 10
- Can we say our sample is different?
- CLT allows us to work out the probability of us getting $\bar{x} = 5$ given we think $\mu=10$. If its really low → our sample different...

Hypothesis test which you'll see next time

Build the intuition for hypothesis testing! I think the population mean is 10, I collect loads of data (big sample) and the sample mean is 5. This seems lower than 10 so I suspect the original hypothesis that the population mean is 10 might be wrong. But how can I conclude this (inferential statistics...)

... well given the Central Limit Theorem I know the distribution of $\bar{x}$, the sample mean, and I know it is distributed around the population mean. I can therefore calculate the probability of getting a sample mean of 5 or lower, GIVEN, my hypothesis that the population mean is 10

... But more on this next week!