1️⃣

Hypothesis Testing

Status	Done
Files	
Link	
Type	All Around
Course	Data Analysis

Descriptive vs Inferential Stats
Empirical vs Probability Distribution
Distributions
    PMF vs CDF vs PDF
    Bernoulli Distribution
    Binomial Distribution
    Normal/Gaussian Distribution
Normal Probability Plot
Z Score
Hypothesis Testing
    Overview
        Coin Example
    How to calculate p? Central Limit Theorem
    Errors
        P-Hacking
    Confidence Intervals
Two Sample Hypothesis Testing
Summary of which test to perform

Descriptive vs Inferential Stats

- Descriptive: summarise qualitatively

- Inferential: learn about the population, properties of the underlying probability distribution

Empirical vs Probability Distribution

- Empirical: the distribution you got after doing the survey (visualising the results)
- Probability: we construct based on the parameters from the empirical distribution

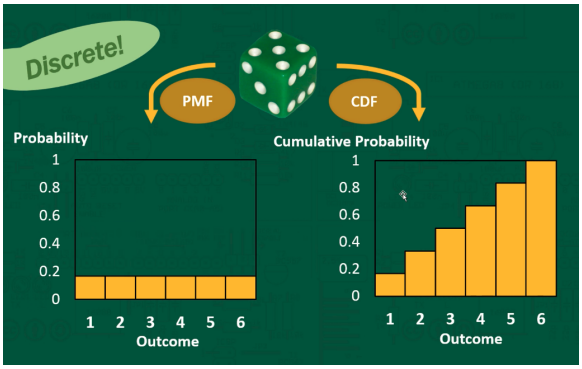
Distributions

- We look at
  - Mean, variance of the distribution
  - The possible values the individual values in the distribution can have,
  - and with what probability is a value x
  - things like probability mass function, cumulative distribution function

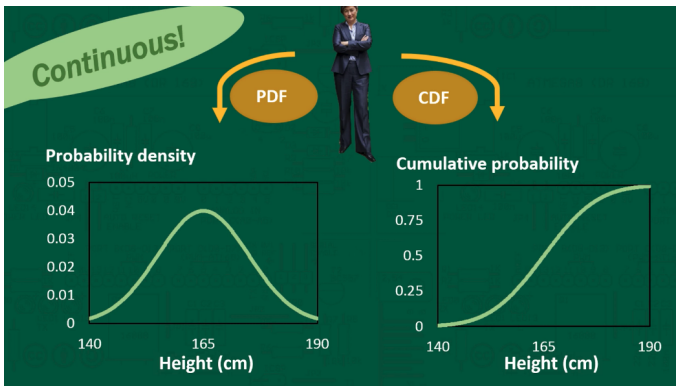
PMF vs CDF vs PDF

> Video for all these statistics terms:
> https://www.youtube.com/watch?v=YXLVjCKVP7U

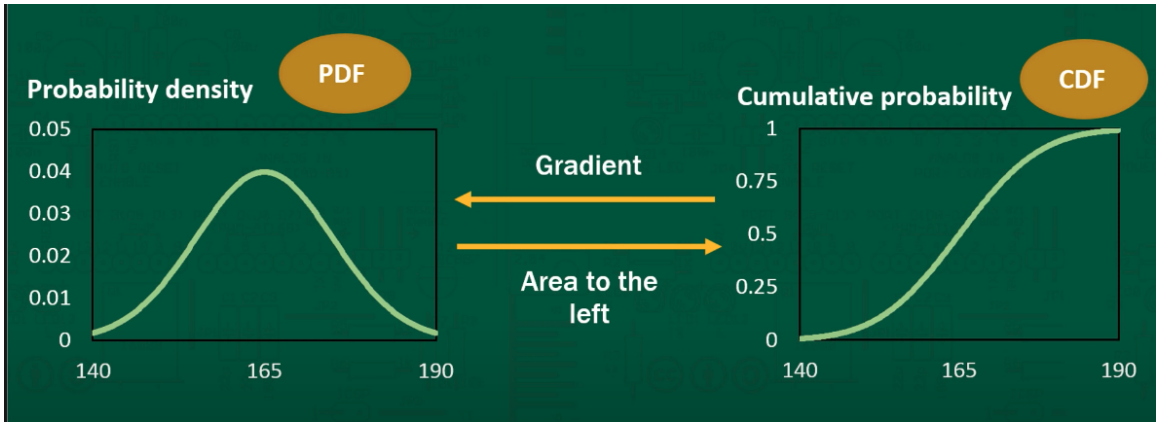
- PMF (probability **mass** function, for **discrete**) and PDF (probability **density** function, for **continuous**) are answering the question “what is the probability I get x?”
- **CDF** (**cumulative** distribution function) applies for both the continuous and discrete cases and is basically answering the question “what is the probability that I get x or less?”
- The gradient of the CDF gives you the PMF/PDF, the integral of the area of the variable and left side of PMF/PDF gives you the CDF
- Discrete



- Continuous



- Relation between PMF/PDF and CDF



Bernoulli Distribution

- 1 (with probability *p*) or 0 (with probability *q = 1-p*)
- Mean is *p*, variance is *p(1-p)*

$$\text{Prob}(X = k) = f(k) = \begin{cases} p & k = 1 \\ 1 - pk = 0 \end{cases}$$

Binomial Distribution

- Values of [0, n]
- Mean is *np*, variance is *np(1-p)*

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

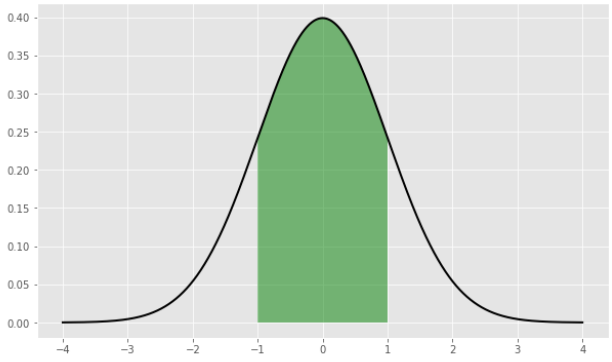
Here, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of ways to arrange the

Normal/Gaussian Distribution

- Random variables, bell curve centred at mean μ and standard deviation σ (variance)

$$\text{Prob}(X \in [a, b]) = \int_a^b f(x)dx.$$

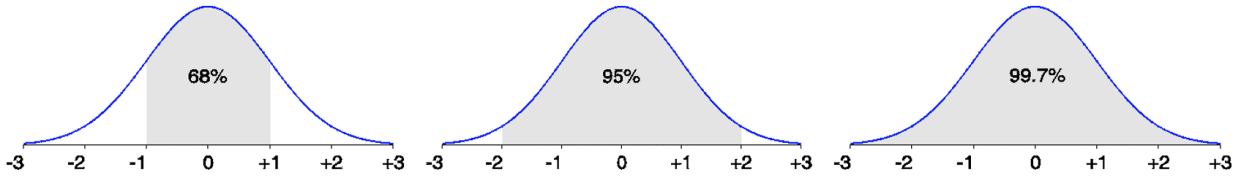
This is just the area under the curve for this interval. For $a = \mu - \sigma$ and $b = \mu + \sigma$, we plot this below.



- This property holds for any normal distribution:

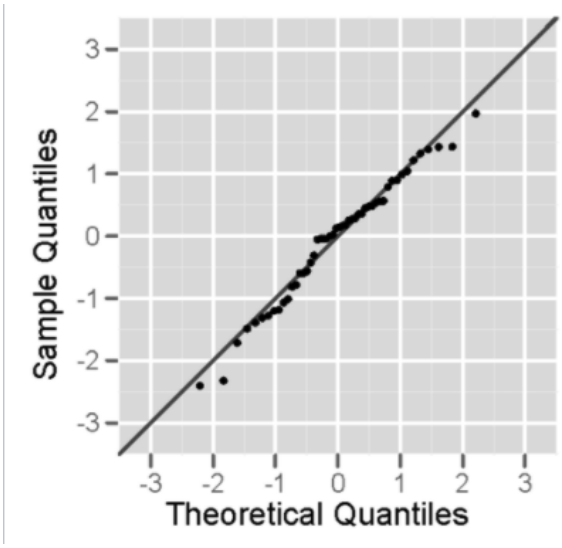
This means that 68% of the time, this normal random variable will have values between $\mu - \sigma$ and $\mu + \sigma$.

Similarly, 95% of the time the normal random variable will have values between $\mu - 2\sigma$ and $\mu + 2\sigma$ and 99.7% of the time between $\mu - 3\sigma$ and $\mu + 3\sigma$.



Normal Probability Plot

- Used to check if a sequence of random variables actually came from a random probability distribution, the closer to the line all the points it is the more likely it is that it came from a normal distribution, but its fine if there are some deviations it also heavily depends on the number of points, if its few points you should be more lenient
- He does not dive into details how the theory behind it works or how to do it manually, but it has to do with the scale of the axis changing and one of the scales is theoretical so it does not mean anything and then you just put your sample data on the other axis on the line, its really amazing how simple it is



Z Score

- Compares the points value against the mean
  - 0 → equal to mean

    - +1 → 1 standard deviation above
    - -2 → 2 standard deviations below
- Formula

$$z_i = \frac{x_i - \bar{x}}{s}$$

- He has some questions on calculating it, so its probably important to know how to calculate it

Hypothesis Testing

Finally

Overview

- We formulate a null hypothesis H_0 and alternative hypothesis H_a (typically H_a **is the hypothesis the researcher wants to validate**)
- Then you do the experiment and measure the data → **what is probability of the observed results happening assuming H_0 is true**

> ❗ **Warning:**
> A p-value is not the probability that the null hypothesis is true or false. It is the probability of observing an event as extreme as the one we did, assuming the null hypothesis to be true.

- Significance level α has to be chosen, traditionally 1% or 5%, many studies on which makes more sense in what case, but no conclusive results

Coin Example

- H_0 coin is fair (p=0.5)

- H_a coin is not fair (p≠0.5)
- a we choose to be 1%
- Run experiment: 1000 coin flips, 545 heads, what is the probability of that happening? Assuming that the coin is fair? 0.49% (we care about how to calculate this number in the next section) so the experiment p value is 0.49%
- p<α therefore we reject the null hypothesis as unlikely to be true, therefore the alpha hypothesis is true and the coin is not fair

How to calculate p? Central Limit Theorem

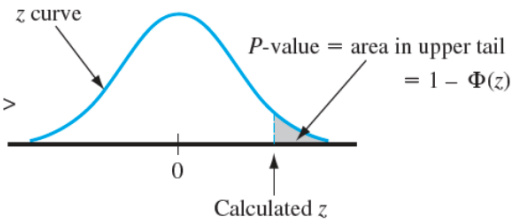
- Again, *p* is the probability of observing such results assuming the hypothesis H_0 is true
- Basically we need a generic way to calculate it → we want a generic method that does not care about what the underlying distribution is
- Central Limit Theorem (CLT) why is it so amazing?
  - Even if you are not normal ..... the average/sum ... is normal (https://youtu.be/YAlJCEDH2uY)
  - Essentially the idea is that if you have a distribution that is not normal, if you take *enough* samples and *enough* means, the sum of the samples and the means will be normally distributed, enough is case-specific and distribution dependent (rule of thumb online n≥30)
- More formally:

Central Limit Theorem. Let $\{X_1, \ldots, X_n\}$ be a sample of n random variables chosen identically and independently from a distribution with mean μ and finite variance σ^2. If n is 'large', then

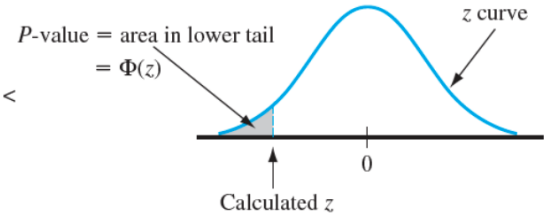
- the sum of the variables $\sum_{i=1}^{n} X_i$ is also a random variable and is approximately **normally** distributed with mean $n\mu$ and variance $n\sigma^2$ and
- the mean of the variables $\frac{1}{n}\sum_{i=1}^{n} X_i$ is also a random variable and is approximately **normally** distributed with mean μ and variance $\frac{\sigma^2}{n}$.

- Once we use the CLT to make a normal distribution for whatever the data is, we can easily calculate p using the CDF of the normal distribution we described in a previous section
- How to determine which sided testing to do

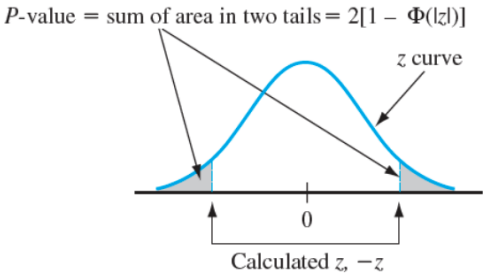
1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <



3. **Two-tailed test**
   H_a contains the inequality ≠



- We can only use CLT if we have enough samples, what if there are not enough? If the samples are drawn from a dis Use student t-distribution

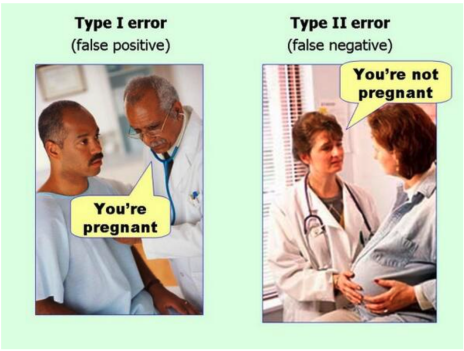
Errors

- Probability of error type 1 is α

Examples

(1) In vaccine testing, we take the null hypothesis (H0): "This vaccine has no effect on COVID19." A type I error detects an effect (the vaccine is effective against the disease) that is not present. A type II error fails to detect an effect (the vaccine is effective) that is present.

(2) In a trial, we take the null hypothesis (H0): "This man is innocent." A type I error convicts an innocent person. A type II error lets a guilty person go free.



P-Hacking

- If you do enough hypothesis tests, eventually you'll have a Type I (false positive) error
- Data dredging.
- This is a real problem in a world with tons of data in which it is easy to do many, many hypothesis tests automatically.
- One method to avoid this is called *cross-validation*, which we'll discuss later

Confidence Intervals

- Estimate of an unknown parameter, for example mean
- The 95% confidence interval *does not* mean that with probability 95%, the true value of u lies within the interval.
- A 95% confidence interval means that if we were to repeat the same experiment many times, and compute the confidence interval using the same formula, 95% of the time it would contain the true value of u.

Two Sample Hypothesis Testing

If the variable of interest is **quantitative**, we compare the **means** μ_A, μ_B of the two populations. For example:

$$H_0 : \mu_A = \mu_B$$
$$H_A : \mu_A \neq \mu_B$$

If the variable of interest is **categorical** with 2 categories, we compare the **proportions** p_A, p_B of the two populations. For example:

$$H_0 : p_A = p_B$$
$$H_A : p_A \neq p_B$$

Summary of which test to perform

8. Most Common Statistical Tests (Summary) ¶

Variable type	# Groups	Classic approach
Quantitative	2	t-test
>>	3+	ANOVA
Binary (proportions)	2	z-test
>>	3+	χ^2 test
Categorical	2+	χ^2 test