MATH9102

HYPOTHESIS TESTING

Sources used in creation of this lecture:
Statistics and Data Analysis, Peck, Olsen and Devore; Discovering Statistics Using R, Field, Miles and Field; Understanding Basic Statistics, Brase and Brase;SPSS Survival Manual, Julie Pallant

The General Linear Statistical Model

Concepts of interest (measured by their variables) are hypothesised to be related to each other in some way

The goal is to summarize or describe accurately what is happening in the data.

The General Linear Statistical Model

Easiest way to think about it is for bivariate data

- Looking to model the pattern in ordered pairs where each member of the pair is the value of one of the variables of interest
- Our statistical model takes the form of a mathematical equation
  - Outcome$_i$ = model + error$_{i\ (}$i = ith case)
  - Outcome in the data we observed = model we built + an error

The General Linear Statistical Model

In the linear model

- Trying to fit a line to it to model the pattern
- Using the equation of a line $y = b_0 + b_1x + e$
  - Where y is our dependent (or outcome) variable and x is the independent (or predictor)
  - b_0 is the intercept (value of y when x is 0)
  - e is the error term - the degree to which the line is in error in describing each data point

Before Building Predictive Models

We need to establish evidence to support going ahead with building a predictive model

If we are asserting a relationship between concepts

- E.g. as the value of x increases there is an equivalent, consistent pattern of increase/decrease in the value of y
- We need to investigate if there is any evidence of a relationship using the appropriate test and make a decision based on the results (strength, direction etc.)

If we are asserting a differential effect for different groups

- E.g. group a and group b experience x differently e.g. the mean of x for group a is significantly different to the mean of y for group b
- We need to investigate if there is any difference using the appropriate test and make a decision based on the result

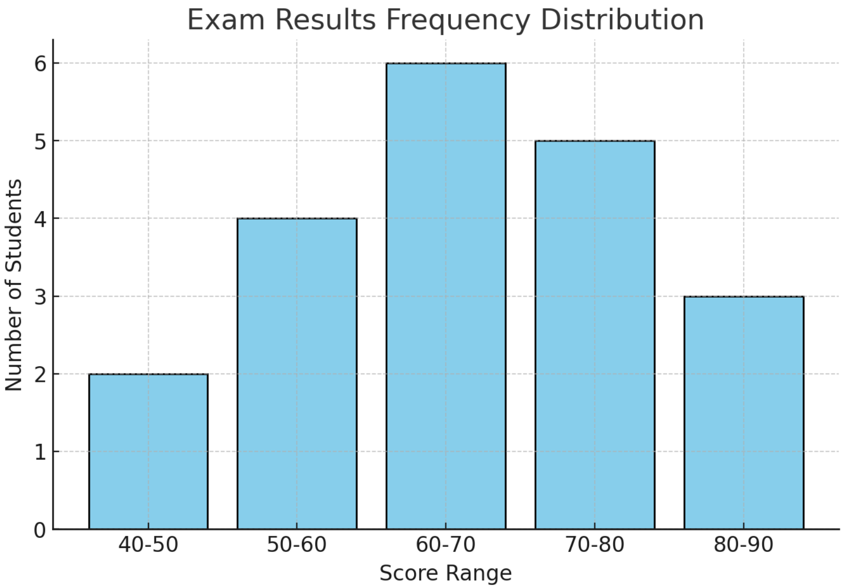
We do this using **hypothesis testing**.

What is a test of hypotheses?

A **test of hypotheses** is a method that uses sample data **to decide** between **two competing claims (hypotheses)** about the **population characteristic**.

We use statistical tests to generate the evidence to make the decision.

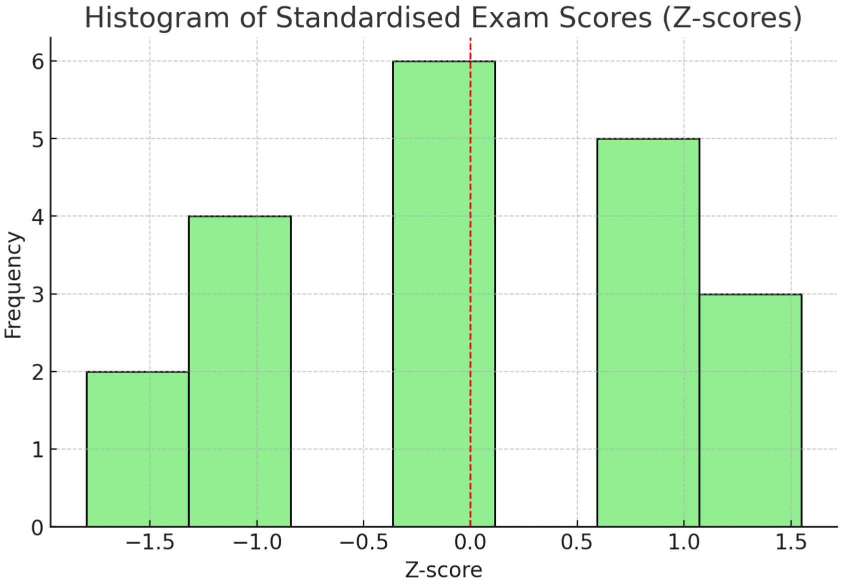
Revisiting Distributions



A frequency distribution shows how often values occur in a dataset (e.g., how many students scored between 50–60, 60–70, etc.).

To turn this into a probability distribution, you divide each frequency by the total number of observations.

- Example:
  - If 20 students took an exam and 5 scored between 70–80, then the probability of a randomly chosen student being in that range is:
    - $P(70 \leq X < 80) = 5/20 = 0.25$
  - This gives us a probability distribution: values (or ranges) with their associated probabilities.



Z Scores (standardised scores)

A z-score tells us how far (and in what direction) a value is from the mean, in units of standard deviation.

Formula:

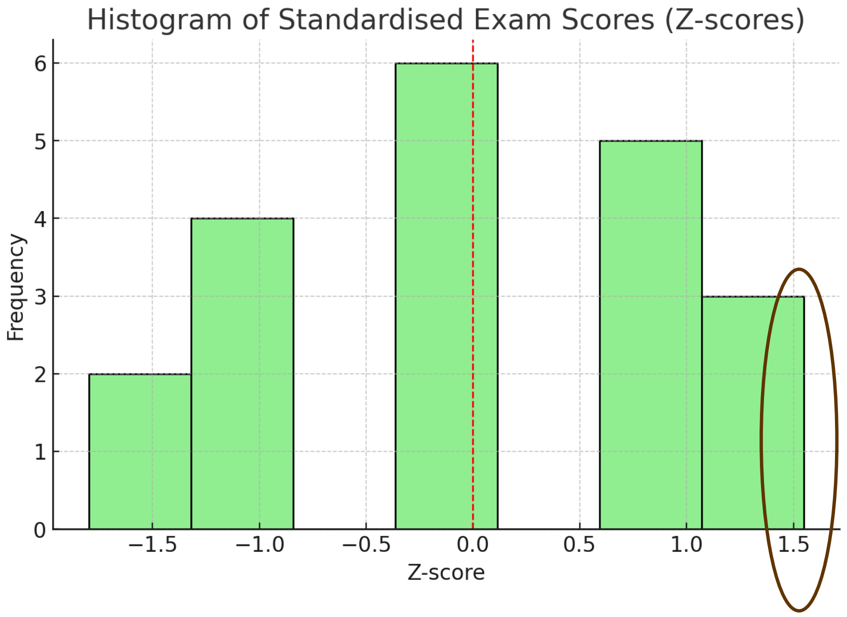
$$\frac{\times - \mu}{\sigma}$$

Where:

x = raw score

µ= mean

σ = standard deviation



Z Scores (standardised scores)

If exam scores have mean μ=65 and standard deviation σ= 10 then a student with x=80 has:

z=80−65/1.5

This means the student is **1.5 standard deviations above the mean**.

The normal distribution

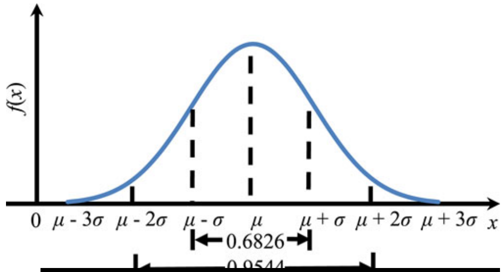
Many natural and social phenomena (heights, IQ scores, measurement errors) follow a normal distribution — a symmetric, bell-shaped curve.

Properties:

- Mean = Median = Mode at the centre.

Spread is controlled by standard deviation

- 68% of values lie within 1 standard deviation 95% within 2
- and 99.7% within 3 (the 68–95–99.7 empirical rule).



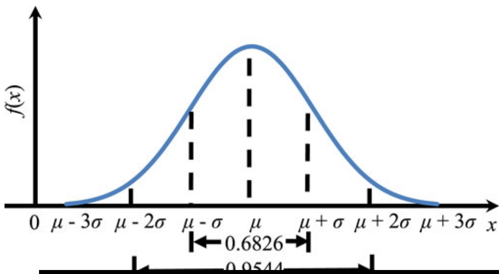
The standard normal distribution

By converting data to z-scores, we can use the standard normal distribution (mean = 0, standard deviation = 1).

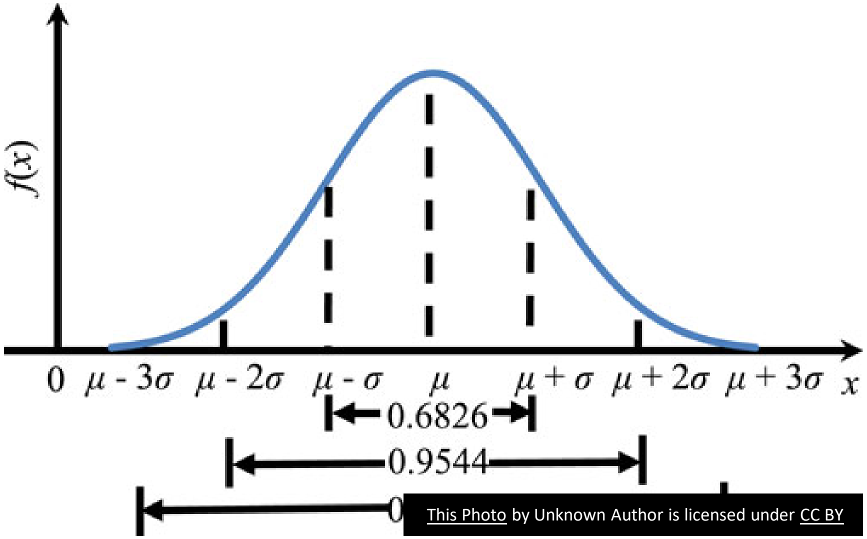
Then we can look up probabilities (areas under the curve) using z-tables or software.

Example:

- For the exam scores above, if we want the probability of scoring above 80:
- Convert to z: z=1.5z=1.5
- Look up $P(Z$>1.5) in the standard normal table → about 0.067 (6.7%).
- So about 7% of students are expected to score higher than 80.



The normal distribution



When we take a random sample from a population and compute a statistic (e.g., the sample mean), the Central Limit Theorem (CLT) tells us that this statistic will have an approximately normal distribution, even if the population itself is not perfectly normal (as long as the sample size is reasonably large).

How large is large enough?

Statistically the “n ≈ 30” rule is the **bare minimum** that makes the Central Limit Theorem work in theory.

In practice, researchers usually prefer much larger samples, because:

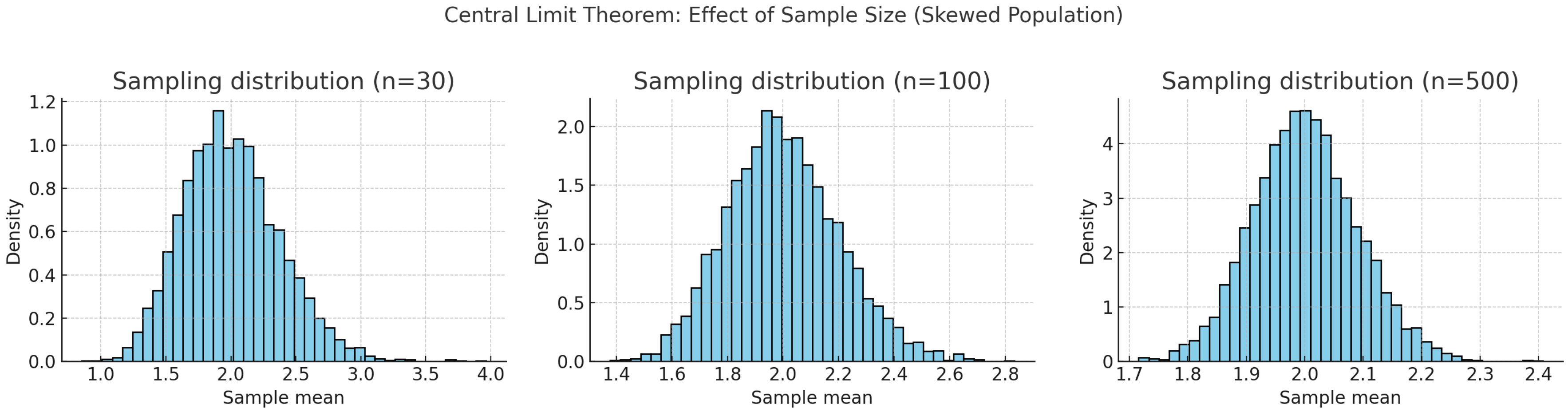
- They give **better normal approximation** (even if the population is skewed).
- They reduce **sampling variability** (your estimate gets more precise).
- They increase **power** for hypothesis tests.

Rules of thumb for practice

- **n = 100** → Usually good enough if the population isn’t extremely skewed.
- **n = 200–300** → Gives a very solid normal approximation in most cases.
- **n ≥ 500** → The Central Limit Theorem is essentially guaranteed to “kick in” even with strongly non-normal populations.
- **n in the thousands** → Common in modern data analysis (e.g., survey research, machine learning), sampling error becomes very small.

In summary: While a sample size of around 30 is often cited as sufficient for the CLT, in practice researchers typically aim for much larger samples (100–200 or more), particularly when dealing with skewed or heavy-tailed populations.

How large is large enough?



Each panel shows the sampling distribution of the **sample mean** based on 5,000 repeated samples of the same size n.

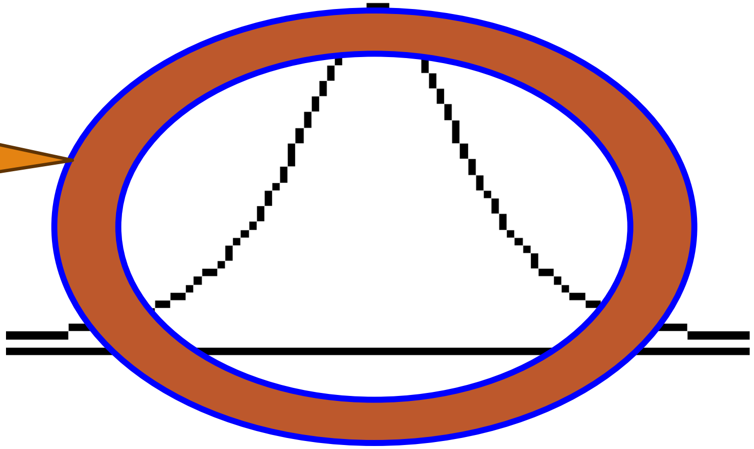
- If **n = 30** → Still skewed, not perfectly normal.
- If **n = 100** → Much closer to a bell shape.
- If **n = 500** → Very close to normal.

What is a test of hypotheses?

A **test of hypotheses** is a method that uses sample data **to decide** between **two competing claims (hypotheses)** about the **population characteristic**.

- **Is the value of the computed sample statistic a random occurrence due to natural variation?**

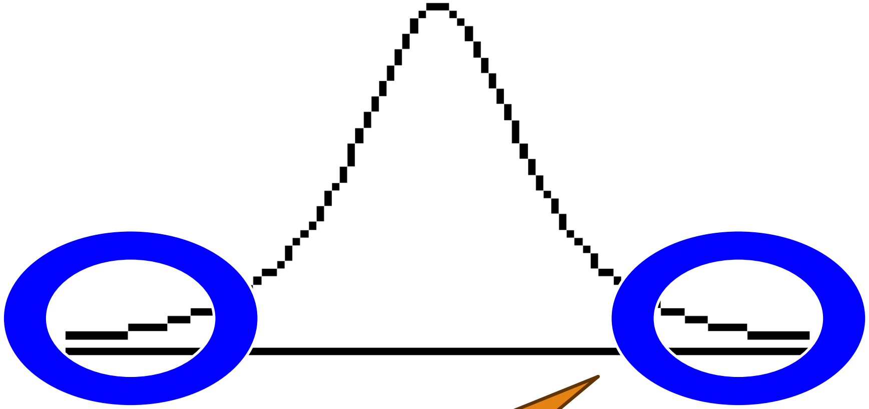
Is it one of the values of the sample statistic that are likely to occur?



What is a test of hypotheses?

A **test of hypotheses** is a method that uses sample data **to decide** between **two competing claims (hypotheses)** about the **population characteristic**.

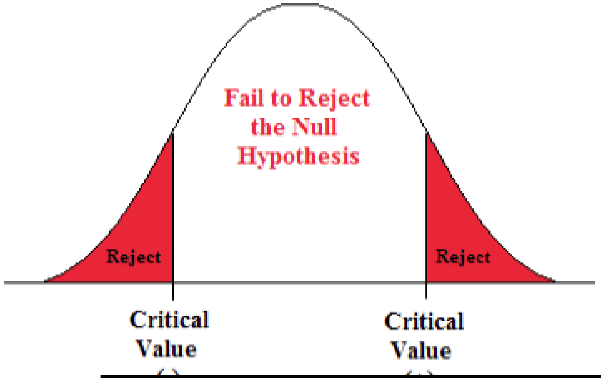
- **Or is the value of the sample statistic a value that would be considered surprising**



One that isn't likely to occur due to natural variation?

One that isn't likely to occur due to natural variation?

What is a test of hypotheses?



A **test of hypotheses** is a method that uses sample data **to decide** between **two competing claims (hypotheses)** about the **population characteristic**.

- How do we make that decision?
  - Based on our distribution, the statistic we generate and the number of pieces of information used to calculate it
    - We can compare our statistic to the **critical values** for an idealised distribution
    - We look it up in a table for the standardised version of this statistic that shows how extreme a statistic needs to be for a given number of degrees of freedom to be considered statistically significant.

Hypothesis Testing

Hypotheses may concern a **relationship** (effect) in the population or a **difference** between groups in a population.

The general goal of a hypotheses test is to rule out chance (sampling error) as a plausible explanation for the results from a research study.

- All hypothesis testing starts with the assumption that the null hypothesis holds:
  - that there is no effect or difference in the population.

Hypothesis Testing

Goal : Make statement(s) regarding unknown population parameter values based on sample data

Elements of a hypothesis test:

- Null hypothesis (H_0)
  - Statement regarding the value(s) of unknown parameter(s).
  - Typically, this will imply no association between independent and dependent variables in our theory (will always contain an equality)
- Alternative hypothesis (H_a)
  - Statement contradictory to the null hypothesis (will always contain an inequality)
  - Collect data and seek evidence against H_0 as a way of bolstering H_a (deduction
- Test statistic
  - Quantity based on sample data and null hypothesis which allows you to determine between null and alternative hypotheses.

Hypothesis Testing

You start with the assumption (the 'null hypothesis' H_0) that there are no differences or relationships in the population as a whole.

You then state an alternative hypothesis (H_A) that there is a difference or a relationship.

You select a sample and find a difference/relationship in it.

You can then use a variety of tests for statistical significance to work out the probability of the difference/relationship you have found in your sample simply occurring by chance.

Hypothesis statements:

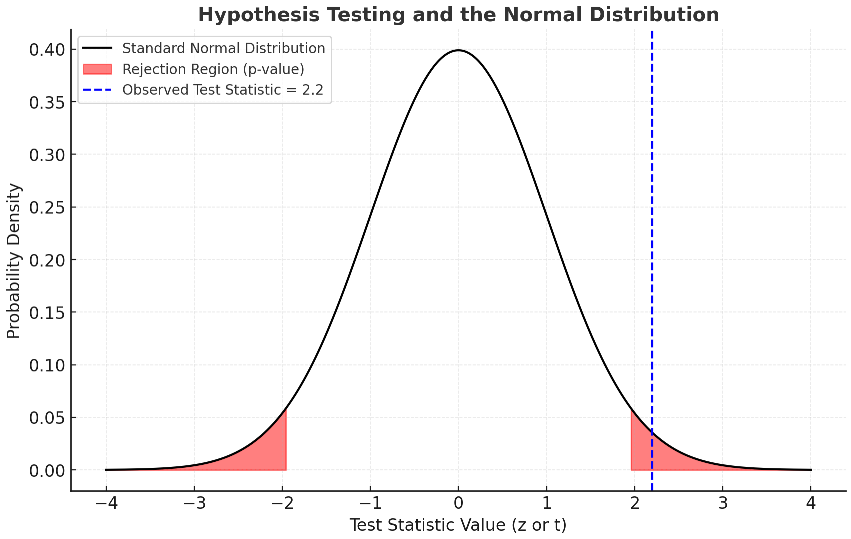
The **null hypothesis**, denoted by $\mathbf{H_0}$

- A claim about a population characteristic that is initially assumed to be hold.
- Null – nothing.
- Presumption of status quo or no change.

The **alternative hypothesis**, denoted by $\mathbf{H_a}$

- The competing claim.
- You are usually trying to determine if this claim is believable.
- Should state what you expect the data to show, based on your theory.
- You need to keep your research objectives in mind when stating this.

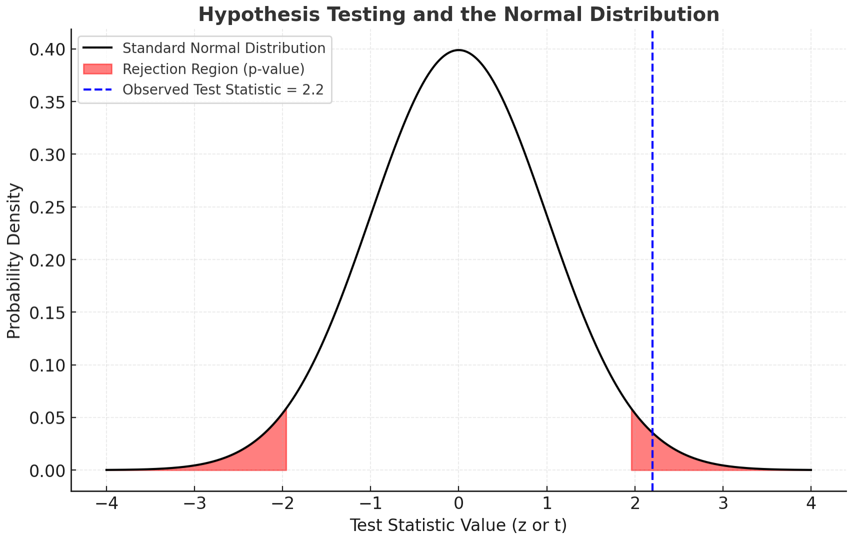
Hypothesis Testing



Hypothesis testing depends on calculating a **p-value:**

- **The probability of observing a test statistic as extreme as the one we obtained, assuming the null hypothesis is true.**

Hypothesis Testing



Hypothesis testing depends on calculating a **p-value:**

- **The probability of observing a test statistic as extreme as the one we obtained, assuming the null hypothesis is true.**

These probabilities are computed using the normal distribution (or distributions derived from it: t, F, χ^2).

Without the normal distribution framework, we couldn’t assign these probabilities reliably.

One v Two Tailed Hypotheses

One-tailed tests:

- Allow for the possibility of an effect or difference in one direction.
- checks for an effect or difference in one specific direction (either greater than or less than a certain value)

Two-tailed tests:

- Allow for the possibility of an effect or difference in two directions—positive and negative.

One Tailed Hypotheses

Suppose you run a factory that produces light bulbs.

You want to test if the average lifetime of the bulbs is **greater than** 1,000 hours (your null hypothesis is that it's 1,000 hours).

You're not concerned if the lifetime is shorter, only if it's **longer**.

- **Null Hypothesis (H_0):** The average lifetime is ≤ 1,000 hours.
- **Alternative Hypothesis (H_A):** The average lifetime is > 1,000 hours.

This is a **one-tailed test**, because you're only testing if the lifetime is **greater** than 1,000 hours

Two Tailed Hypotheses

Suppose you run a factory that produces light bulbs.

You want to test whether the average lifetime of your bulbs is different from 1,000 hours.

You're interested in finding out if it's either greater than or less than 1,000 hours.

- **Null Hypothesis (H_0):** The average lifetime is = 1,000 hours.
- **Alternative Hypothesis (H_A):** The average lifetime is ≠ 1,000 hours.

This is a **two-tailed** test, because you're testing for any difference, in either direction (both higher and lower).

Derive Hypotheses Pairs

The mean number of years Americans work before retiring is 34.

Is this a null hypothesis/alternate hypothesis?

Complete the hypothesis pair

Derive Hypotheses Pairs

The mean number of years Americans work before retiring is 34.

null hypothesis

Alternate: The mean number of years Americans work before retiring $\neq 34$

Derive Hypotheses Pairs

The average high temperature in July in Dublin historically has been 19C.

Perhaps it is higher now.

State the hypothesis pair.

Derive Hypotheses Pairs

The average high temperature in July in Dublin historically has been 19C. Perhaps it is higher now.

Null hypothesis: The average high temperature in July in Dublin is 19C.

Alternate: The average high temperature in July in Dublin is > 19C.

Is this one-tailed or two-tailed?

Derive Hypotheses Pairs

Express H0 and HA for each of the following:

- The mean annual starting salary for computer science majors is greater than £40,000
- The proportion of people that suffer from diabetes in Ireland is less than 9%

Identify whether they are one or two tailed hypotheses.

P-value

The **threshold of probability** or **level of significance** (denoted as **α**) is the level of significance is the threshold at which you decide whether to reject the null hypothesis.

It represents the probability of making the error of incorrectly rejecting a null hypothesis.

Common levels of significance are:

5% (0.05):

- This means that you are willing to accept a 5% chance of incorrectly rejecting the null hypothesis.

1% (0.01):

- This means you are willing to accept only a 1% chance of making that error.

0.1% (0.001):

- This means you are willing to accept only a 0.1% chance of making that error.

P-value and Confidence Interval

A confidence interval is a range of values that is likely to contain the true population parameter (e.g., a mean or proportion) with a certain degree of confidence.

The level of confidence is typically $1-\alpha$:

- If α is 0.05 (5%), the confidence interval would be 95%.
  - A 95% confidence interval means that if you were to repeat the experiment 100 times, in 95 of those experiments, the calculated interval would contain the true population parameter.
- For α=0.01, the confidence interval would be 99%.

P-value

Decide in advance your cut-off value (0.05, 0.01)

P-value >= cut-off

- No evidence to reject null hypothesis
- Report decision and report the actual p-value

P-value < cut-off

- Need to also consider the strength of your statistic
- Have evidence to reject null hypothesis in favour of the alternate
- Report p < your cut-off
  - Note: convention is if p value is < .000 then report it as < 0.001 even if working at a level of 0.05.

When you perform a hypothesis test you make a decision:

Decision:

- **reject** H0

OR

- ***fail* to reject H0**

Each could possibly be a wrong decision;

- Therefore, there are two types of errors:
  - Type I
  - Type II

Type I error

The error of ***rejecting H0*** **when** ***H0 actually holds***

The probability of making a Type I error is denoted by**α**.

- **α** *is called the* ***significance level*** *of the test*

Key Slide

Type II error

The error of ***failing to reject H0*** **when *H0 is does not hold***

The probability of making a Type II error is denoted by**β**

Key
Slide

This is the lower-case Greek letter “beta”.

Power of a statistical test

The power of a statistical test is the probability that it correctly rejects the null hypothesis when the alternative hypothesis holds.

In other words, it’s the test's ability to detect a true effect.

The power of a test is calculated as:

Power$=1-\beta$

Key Slide

P-value, α – statistical significance

Key Slide

A probability measure of evidence about $H_{0.}$

$H_{0\,:}$ the assumption that null hypothesis holds.

- There is no effect or no difference.
- It represents the default state or a baseline assumption about the population

Statistical significance in simple terms:

- Given our presumption that the null hypothesis holds
  - α is the probability that the results could have been obtained purely because of chance alone.

P-value, α – statistical significance

Key Slide

The probability (under presumption that H_0 holds) the test statistic equals observed value or value even more extreme predicted by H_a

The ***P*-value** allows us to answer the question:

- Do our sample results allow us to reject H_0 in favour of H_a?
- If that probability (p-value) is small, it suggests the observed result cannot be easily explained by chance.

Key Slide

Statistical Significance

Working with random samples can never have 100% certainty that findings we derive from the sample will reflect real differences in the population as a whole.

Convention is that (for your field of study) there is an accepted level of probability such that it is considered so small that the finding from your sample is unlikely to have occurred by chance or sampling error.

- Normally, that line is drawn at $p=0.05$ or $p=0.01$.
  - In other words, when a statistical test tells us that the finding has less than a 5% or 1% chance of occurring due to sampling error then we tend to conclude that we can be sufficiently confident that the finding is therefore likely to reflect a ‘real’ characteristic of the population as a whole.
- When this occurs, you can say that your finding is **statistically significant**.

Hypothesis Testing

You start with the assumption (the 'null hypothesis' H_0) that there are no differences or relationships in the population as a whole.

You then state an alternative hypothesis (H_A) that there is a difference or a relationship.

You select a sample and find a difference/relationship in it.

You can then use a variety of tests for statistical significance to work out the probability of the difference/relationship you have found in your sample simply occurring by chance.

Hypothesis Testing

Using the standard level accepted by your domain (e.g. $p \leq 0.05$ or $p \leq 0.01$)

If the probability less than this value then you reject the null hypothesis and thus accept the alternative hypothesis and you can state that your findings are 'statistically significant'.

If the probability is greater this value then you conclude that there is no evidence to reject the null hypothesis and your findings are not ‘statistically significant’.

- N.B. This is different from concluding that you have evidence to accept the null hypothesis.
- In these cases, your findings are said to be ‘not significant’.

Hypothesis Testing

Caveat

- If we get a p-value of 0.051 should we accept the null hypothesis?
- Should we reject the null hypothesis if we get a p-value of 0.049?
- Need to allow for some flexibility in interpretation

Accepting and Rejecting Hypotheses

A non-statistically significant test result **does not** mean that the null hypothesis is true or that you have proved the null hypothesis

- It means you do not have sufficient statistically significant evidence to reject the null in favour of the alternate

Key Slide

Accepting and Rejecting Hypotheses

A significant result **does not** mean that the null hypothesis is false or that you have proved the alternate hypothesis

- It means you have sufficient statistically significant evidence to reject the null in favour of the alternate

Key Slide