Hypothesis Testing

Case Study

Data Overview

The "AI-Powered Job Market Insights" dataset provides a snapshot of the modern job market, particularly focusing on the role of artificial intelligence (AI) and automation across various industries.

This dataset includes 500 unique job listings, each characterized by different factors like industry, company size, AI adoption level, automation risk, required skills, and job growth projections.

Data Preview

```
ai_jobs |>
  gt_preview() |>
  tab_header(title = "AI-Powered Job Market Insights") |>
 tab_source_note(source_note = "Source: Kaggle https://www.kaggle.com/datasets/
uom190346a/ai-powered-job-market-insights")
```

AI-Powered Job Market Insights

	job_title	industry	company_size	location	ai_adoption_level	automation_risk	salary_usd	required_skills	remote_friendly	job_growth_projection
1	Cybersecurity Analyst	Entertainment	Small	Dubai	Medium	High	111392.17	UX/UI Design	Yes	Growth
2	Marketing Specialist	Technology	Large	Singapore	Medium	High	73792.56	Marketing	No	Decline

Source: Kaggle https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights

AI-Powered Job Market Insights

	job_title	industry	company_size	location	ai_adoption_level	automation_risk	salary_usd	required_skills	remote_friendly	job_growth_projection
3	AI Researcher	Technology	Large	Singapore	Medium	Low	137170.26	UX/UI Design	Yes	Growth
4	Sales Manager	Retail	Small	Berlin	Low	Medium	83027.95	Project Management	No	Growth
5	Cybersecurity Analyst	Entertainment	Small	Tokyo	Low	Low	87552.92	Java, SQL	Yes	Decline
6..499										
500	HR Manager	Entertainment	Medium	Berlin	Medium	High	53764.38	Project Management	Yes	Decline

Source: Kaggle https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights

To simplify our later code, I have created a separate table which is already filtered for the groups we will be looking at:

```
ai_jobs_risk <- ai_jobs |>
  filter(automation_risk %in% c("Low", "High"))

ai_jobs_high <- ai_jobs_risk |>
  filter(automation_risk == "High")
```

```
ai_jobs_low <- ai_jobs_risk |>
  filter(automation_risk == "Low")
```

Dataset Features:

Source: Kaggle https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights

1. **Job_Title**:
   - **Description**: The title of the job role.
   - **Type**: Categorical
   - **Example Values**: “Data Scientist”, “Software Engineer”, “HR Manager”
2. **Industry**:
   - **Description**: The industry in which the job is located.
   - **Type**: Categorical
   - **Example Values**: “Technology”, “Healthcare”, “Finance”
3. **Company_Size**:
   - **Description**: The size of the company offering the job.
   - **Type**: Ordinal
   - **Categories**: “Small”, “Medium”, “Large”
4. **Location**:
   - **Description**: The geographic location of the job.
   - **Type**: Categorical
   - **Example Values**: “New York”, “San Francisco”, “London”
5. **AI_Adoption_Level**:
   - **Description**: The extent to which the company has adopted AI in its operations.
   - **Type**: Ordinal
   - **Categories**: “Low”, “Medium”, “High”
6. **Automation_Risk**:
   - **Description**: The estimated risk that the job could be automated within the next 10 years.
   - **Type**: Ordinal
   - **Categories**: “Low”, “Medium”, “High”
7. **Required_Skills**:
   - **Description**: The key skills required for the job role.
   - **Type**: Categorical
   - **Example Values**: “Python”, “Data Analysis”, “Project Management”
8. **Salary_USD**:
   - **Description**: The annual salary offered for the job in USD.
   - **Type**: Numerical
   - **Value Range**: $30,000 - $200,000
9. **Remote_Friendly**:
   - **Description**: Indicates whether the job can be performed remotely.
   - **Type**: Categorical

- **Categories**: “Yes”, “No”

10. **Job_Growth_Projection**:
    - **Description**: The projected growth or decline of the job role over the next five years.
    - **Type**: Categorical
    - **Categories**: “Decline”, “Stable”, “Growth”

Motivation

You have been tasked with examining the impact of AI skills and company AI adoption on the job market across the world. Think about what we might want to know about this sector.

What might some be some interesting questions we could ask based on this data?

How would we answer them? How would we know whether our answer is *generalizable*?

We want to **generalize** the results of our analysis from the data we have to the situation we care about.

The Logic of Testing Hypotheses

A hypothesis is a statement about a population, or general pattern.

Testing a hypothesis amounts to gathering information (sampling) from a dataset and, **based on that information**, deciding whether that hypothesis is false or true **in the population**.

Two decisions are possible:

- Rejecting the hypothesis (if there is enough evidence against it)
- Not rejecting it (if there is not enough evidence against it)

Rejecting a hypothesis is a more conclusive decision than not rejecting it.

One such focused approach uses a statistic (e.g. a difference in two means) computed from our data to see whether its true value is equal to something we assume (e.g. the difference is zero). This is called hypothesis testing: using results in the data to see if we have enough evidence to tell whether a hypothesis (the two means are equal) is wrong (they are not equal) or whether we don’t have enough evidence (they may be equal.

Testing a hypothesis is a way of making an inference, with a focus on a specific statement. As with any kind of inference, we have to assess external validity too: the extent to which the population, or general pattern, represented by our data is the same as the population, or general pattern, we are truly interested in.

- Given a population (i.e. a distribution) with a parameter of interest (which could be the mean, variance, correlation, etc.), we would like to **decide between to complementary statements concerning the parameter**.
- These statements are called **statistical hypotheses**.
- The choice or decision between these hypotheses is to be **based on a sample** taken from the population of interest.
- The **goal** is to be able to choose which hypothesis is true in reality based on the sample data.

Steps of Analysis & Hypothesis Testing

1. Formulate research question

2. Specify hypotheses
- What statistic is appropriate to answer our question?
- Null and Alternative Hypotheses

3. Collect relevant data
- What information is needed to answer our question?
- What is our population? How do we sample from the population in a statistically valid way?

4. Compute test statistic
- Fit appropriate model
- Calculate test statistic
- Account for variability

5. Determine probability under the null hypothesis

6. Assess significance and meaningfulness

1. Formulate research question

This is where all analysis or research starts - what is it you want to know, and what will you do with that information?

For our AI Jobs dataset some research questions we might ask are:

- Which jobs are the highest paid?
- Which industries have seen the most AI adoption?
- **Are lower paid jobs at more risk of being automated?**

2. Specify hypotheses

Now we translate our general research questions into specific **and testable** hypotheses.

The actual test begins by considering two hypotheses. They are called the null hypothesis and the alternative hypothesis. These hypotheses contain opposing viewpoints.

- **The null hypothesis (**H_0**)** : It is often a statement of the accepted historical value or norm. This is your starting point that you must assume from the beginning in order to show an effect exists.
- **The alternative hypothesis (**H_a**)** : It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0

2. Specify Hypotheses

After you have determined which hypothesis the sample supports, you make a decision.

There are two options for a decision:

1. "reject H_0" if the sample information favors the alternative hypothesis, or

2. "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Mathematical symbols used in H0 and Ha:

Figure 6.12: Null and Alternative Hypotheses

H_0	H_a
equal (=)	not equal ($\neq$) **or** greater than ($>$) **or** less than ($<$)
greater than or equal to ($\geq$)	less than ($<$)
less than or equal to ($\leq$)	more than ($>$)

2. Specify Hypotheses

Decide on statistic of interest

Decide what statistic is appropriate to answer our question. **What do we need to calculate from our sample?**

- Difference in `mean(salary_usd)` for each level of `automation_risk`
- Since there are three automation risk levels (Low, Medium, High), the comparison we could make are:
  1. $\text{mean}_{\text{high risk}} - \text{mean}_{\text{medium risk}}$
  2. $\text{mean}_{\text{high}} - \text{mean}_{\text{low}}$
  3. $\text{mean}_{\text{medium}} - \text{mean}_{\text{low}}$

2. Specify Hypotheses

Decide on statistic of interest

For now, let's focus on just (2). Therefore:

$s = \overline{salary}_{\text{high risk}} - \overline{salary}_{\text{low risk}}$

s is our statistic of interest. We are interested in the true value of s in the population (s_{true}).

What we can actually calculate is $\hat{s}$ the value of s in our sample.

2. Specify Hypotheses

Express the hypotheses mathematically

Our hypothesis, plainly stated, is:

There is a relationship between job salary and the likelihood of automation. Expressed another way, the average salary of high risk jobs is different from that of low risk jobs.

What are the Null Hypothesis H_0 and Alternative Hypothesis H_A ?

2. Specify Hypotheses

Express the hypotheses mathematically

Null Hypothesis H_0

There is no difference in the average salary between those jobs at high risk of automation, and those at low risk:

$$H_0 : s_{true} = 0$$

Alternative Hypothesis H_a

The difference in the average salary of high risk jobs and low risk jobs is not zero:

$$H_A : s_{true} \neq 0$$

The null says that the true value of the statistic is zero; the alternative says that it's not zero. Together, these cover all logical possibilities for the true value of the statistic.

It may seem odd to have $H_0 : s_{true} = 0$ when, presumably, we analyze the data because we suspect that the true value of s is not zero. This seemingly twisted logic comes from the fact that testing a hypothesis amounts to seeing if there is enough evidence in our data to *reject the null*. It is sometimes said that the null is protected: it should not be too easy to reject it otherwise the conclusions of hypothesis testing would not be strong.

As we introduce the concept of hypothesis testing, it is helpful to relate its logic to the logic of a criminal court procedure. At court the task is to decide whether an accused person is guilty or innocent of a certain crime. In most modern societies the starting point is the assumption of innocence: the accused person should be judged guilty only if there is enough evidence against their innocence. This is so even though the accused person was brought before court presumably because there was a suspicion of their guilt. To translate this procedure to the language of hypothesis testing, H0 is that the person is innocent, and HA is that the person is guilty.

Medical tests are another instructive example. When testing whether a person has a certain medical condition, the null is that the person does not have the condition (healthy), and the alternative is that they have it (sick). The testing procedure amounts to gathering information to see if there is evidence to decide that the person has the condition.

The case when we test if HA: strue ≠ 0 is called a two-sided alternative as it allows for strue to be either greater than zero or less than zero. For instance, we focus on the difference in online and offline prices, with H0 being the equality. In such a case we are not really interested if the difference is positive or not, or whether it is negative or not.

The other case is working with a one-sided alternative, when we are indeed interested if a statistic is positive. The null and the alternative should be set up so that the hypothesis we are truly interested in is in the alternative set. So when we want to know if strue is positive, we want to put strue > 0 in the alternative thus, making the null strue ≤ 0:

$$H_0 : s_{true} \leq 0$$

$$H_A : s_{true} > 0$$

3. Using the Sample to Test the Null Hypothesis

Once you have defined your hypotheses the next step in the process, is to collect sample data.

In this case, we already have the data in `ai_jobs`. Before moving on to actually testing the hypothesis, let's take the naive approach - just calculate $\hat{s}$ the difference between the two groups.

```
mean_high <- mean(ai_jobs_high$salary_usd)
mean_low <- mean(ai_jobs_low$salary_usd)
s_hat <- mean_high - mean_low
```

$$\hat{s} = \overline{\text{salary}}_{\text{high risk}} - \overline{\text{salary}}_{lowrisk} = \$81.7k - \$99.7k = \textbf{-18.1k}$$

4. Compute the test statistic

There are several statistical tests used in Hypothesis Testing. Which one you use depends on what type of hypothesis you are testing and what kind of data you have.

For this example, where we are testing the difference in means of a numerical variable (`salary_usd`) across different groups (`automation_risk`), we use a test called the **t-test**.

4. Compute the test statistic

- Following the logic of hypothesis testing, we start from the assumption that the null (H_0) is true and thus $s_{true} = 0$.
- We look at the evidence to see if we want to reject this null or maintain our assumption that it's true.
- The evidence we look for is **how far** the estimated value $\hat{s}$ is from zero.
- We reject H_0 if the distance is large (i.e. $\hat{s}$ is sufficiently greater or lesser than 0)

How far is far enough?

The **test statistic** is the measure of how far the estimated value $\hat{s}$ is from what its true value would be if H_0 is true.

4. Compute the test statistic

t-statistic:

$$t = \frac{\hat{s}}{\text{SE}(\hat{s})} = \frac{\bar{x_A} - \bar{x_B}}{\text{SE}(\bar{x_A} - \bar{x_B})}$$

- The t-test is a procedure to decide whether we can reject the null H_0.
- The magnitude of the t-statistic t measures **the distance of $\hat{s}$ from what s_{true} would be if the null were true**.
  - The unit of distance is the standard error.
- The t-statistic transforms the original statistic of interest into a **standardized version**
  - For example: if $t = 1$ (or -1), it means $\hat{s}$ is exactly one standard error away from zero.

4. Compute the test statistic

R makes it very easy to apply a test such as the t-test. For most statistical tests, there exists a simple function to compute it.

For the t-test, we use the `t.test()` function:

```
t_res <- t.test(ai_jobs_high$salary_usd, ai_jobs_low$salary_usd)
```

$t = -6.54$

Making a Decision

- The following step is making a decision: either rejecting the null or not rejecting it.
- In hypothesis testing, this decision is based on a **clear rule specified in advance**.
- We specify in advance to avoid bias - before looking at the data, we state what it would take to reject the null hypothesis. We follow what the data says, whatever result that may be.

A clear rule also makes the decision transparent, which helps avoid biases in the decision. Unfortunately, we humans are often tempted to use evidence to support our pre-existing views or prejudices. If, for example, we think that jobs which are more highly valued by companies (i.e. have a higher salary) we may pay more attention to the evidence that supports that belief than to the evidence against it.

In partiular, we may be tempted to say that the estimated $\hat{s}$ difference is large enough to reject the null, because we believe that the null isn't true. Clear decision rules are designed to minimize the room for such temptations.

Once you have your test statistic there are two methods to use it to make your decision:

- Critical value method
- P-Value method – This is the preferred method we mostly will focus on.

Critical Values

We use a **critical value** to tell us whether the test statistic is large enough - is it far enough away from zero to reject the null?

To define the critical value, we need to decide how conservative we want to be with the evidence.

The larger we set the critical value, the harder it is to reject the null hypothesis.

Critical Values

The test sampling distribution

As with the sampling distribution for means we looked at earlier, our test-statistic t also has a sampling distribution. If we were to sample many times and calculate t for each sample, we would again get a distribution with a specific shape and parameters.

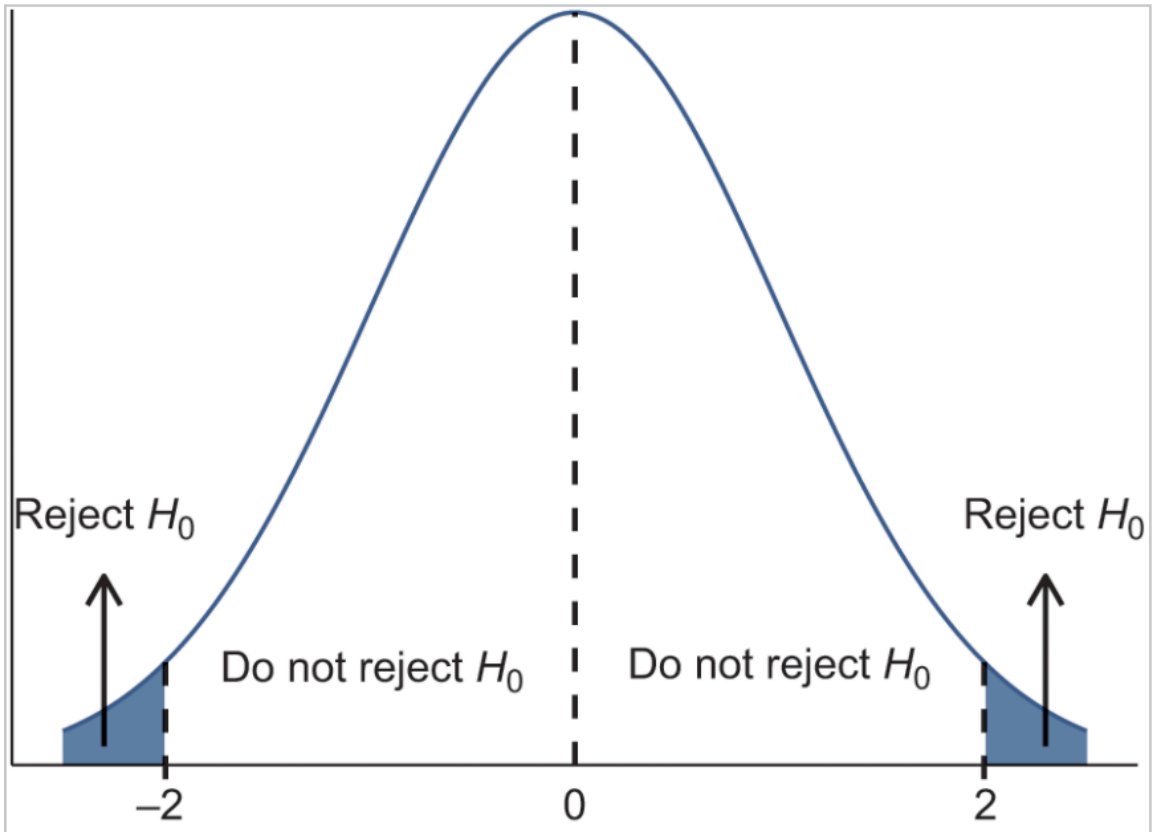


Figure 1: The sampling distribution of the test statistic when the null is true.

Recall: Approximately 95% of values fall **within two standard deviations of the distribution**.

Critical Values

Picking a critical value

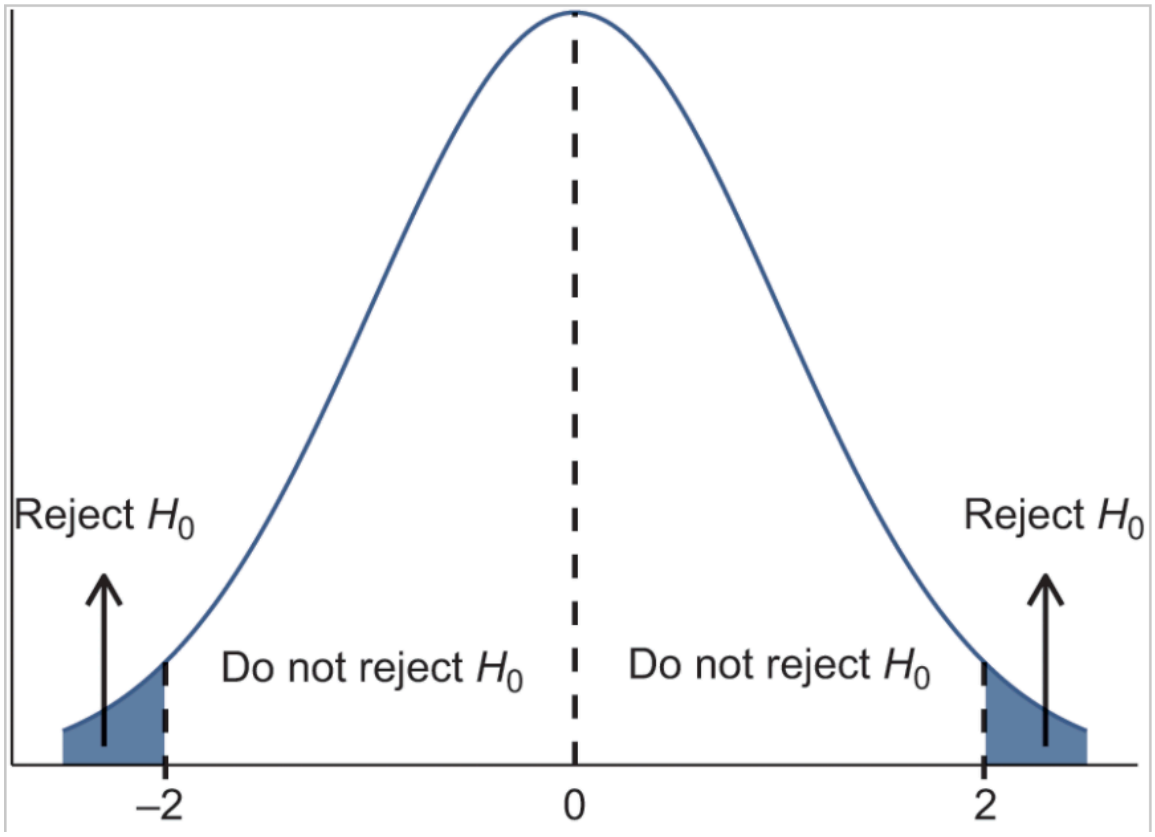


Figure 2: The sampling distribution of the test statistic when the null is true.

Recall: Approximately 95% of values fall **within two standard deviations of the distribution**.

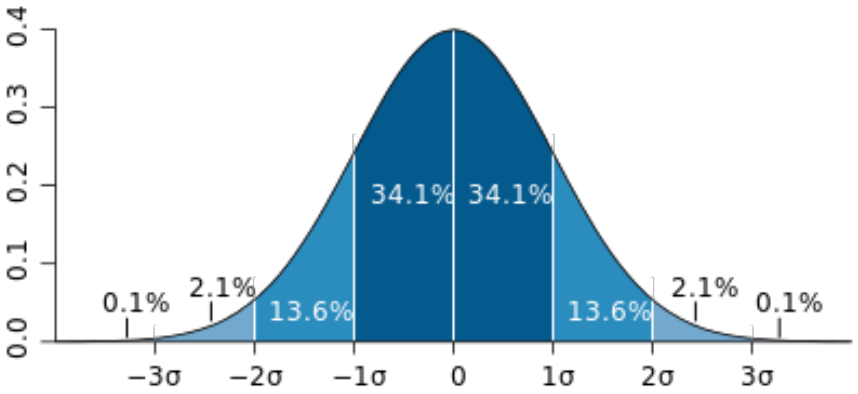
Since 95% of values fall within 2 SD, if we want to reject the null hypothesis with 95% confidence, then we say that our test statistic must fall outside of 2 SD.

In other words, since the units of $\hat{s}$ are standard deviations: $\hat{s} \geq \pm 2$

Critical Values

A critical value of 2 is standard. However, it is ultimately just a convention. We could choose to set other critical values that correspond to different probabilities. There is not anything inherently special about setting our threshold at 95% vs 90%.

If we make the critical value ± 1.6, the chance of a false positive is 10%.



Different fields have different standards for evidence - for instance, a critical value of 5 (99.994%) is standard in particle physics (referred to as 5σ).

Interpret our results

Since our test statistic $t = -7.5 < -2$, at a confidence level of 95%, we would have sufficient evidence to reject H_0

Therefore, we would say:

1. The average salary of jobs at high risk of automation **is not the same** as the average salary of jobs at low risk.
2. We have evidence that higher salary jobs are at less risk of automation than low salary jobs.

Interpret our results

Important!

This does not inherently mean we **accept** the alternative hypothesis. We are narrowing the realm of possible answers, but very rarely (perhaps never) are we able to statistically **prove** a single explanation in one go.

We have increased our reasons to believe our hypothesis, but several other possibilities exist.

Science is then the process of continually investigating our hypothesis and pitting it against new null hypotheses and rejecting them as well.

P-Value Method

- Hopefully, the critical value is fairly intuitive to you now. However, it is not the typical way that statistical results are presented.
- Instead, you will typically see something called a **p-value**.

p-value: The probability than an event will occur, assuming the null hypothesis is true.

The p-value essentially flips the critical value statement:

- Instead of saying a test statistic value > 2 falls outside the 95% bound, we calculate where out test statistic falls in the distribution

The p-value is the probability that the test statistic will be as large, or larger, than we calculate from the data, **if the null hypothesis is true.** i.e. $P(data \mid H_0)$.

P-value

$p = P(|t| > \text{critical value})$

Because the p-value tells us the smallest level of significance at which we can reject the null hypothesis, **it summarizes all the information we need to make the decision**.

This is why the p-value is used - rather than needing to set a critical value and calculate the test statistic, **we can instead use just the p-value**.

Interpreting the P-value

- Like with the critical value, we should set our desired **significance level** before carrying out the analysis.
- We then compare our calculated p-value with the significance level. If it is less, we reject the null hypothesis.
- The **significance level** (α) is the probability that a true null hypothesis will be rejected.
- A typical significance level is $\alpha < 0.05$, which corresponds with a critical value of 2, or a probability of 5%.

Interpreting the P-value

R output

Again, R will provide us with the p-value. Let's now look at the full output from our `t.test()`:

```
t.test(ai_jobs_high$salary_usd, ai_jobs_low$salary_usd)
```

```
    Welch Two Sample t-test

data:  ai_jobs_high$salary_usd and ai_jobs_low$salary_usd
t = -6.5366, df = 288.07, p-value = 2.854e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23513.51 -12630.27
sample estimates:
mean of x mean of y
 81673.57  99745.46
```

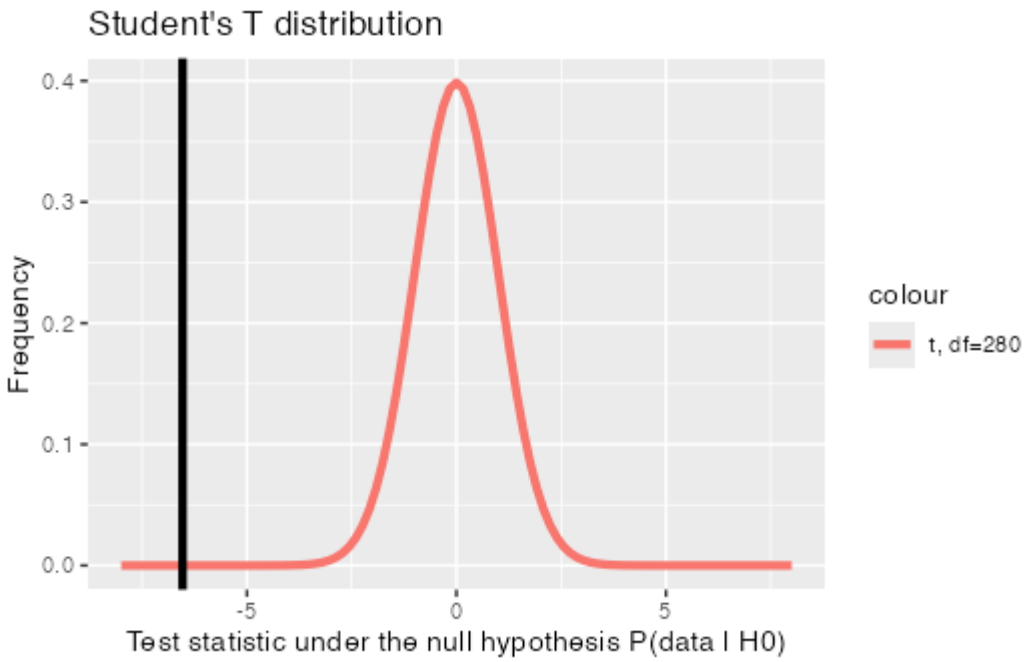
- If our p-value is smaller than our pre-set significance level (α), we reject the null hypothesis and can say the result is "statistically significant" at $p < 0.05$.

T distribution

```
xpos <- seq(-5, 5, by = 0.01)
```

```r
degree <- 280
ypos <- dt(xpos, df = degree)

ggplot() +
  xlim(-8, 8) +
   geom_function(aes(colour = "t, df=280"), fun = dt, args = list(df = 280),
linewidth = 1.5) +
  geom_vline(xintercept = t_res$statistic, color = "black", linewidth = 1.5) +
  xlab("Test statistic under the null hypothesis P(data | H0)") +
  ylab("Frequency") +
  ggtitle("Student's T distribution")
```



Decision and conclusion

The preset α is **the probability of a False Positive error** (called a Type I error) - rejecting the null hypothesis when the null hypothesis is true.

Back to our two possible decisions:

- **If** $\alpha >$ p-value**, reject** H_0.
  - The results of the sample are statistically significant.
  - We can say there is sufficient evidence to conclude that H_0 is an incorrect believe and that the alternative hypothesis, H_A **may be** correct.
- **If** $\alpha <$ p-value**, fail to reject** H_0.
  - The results of the sample are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_A may be correct.

NOTE: When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Closing - What does a statistically significant result mean?

Does it mean our result is meaningful or practically important?

Effect size

No. There is an essential distinction between **statistical significance** and **practical significance**.

Let's say we performed an experiment to examine the effect of a particular diet on body weight, which gives a *statistically significant* effect at $p < 0.05$. This doesn't tell us how much weight was lost, which we refer to as the **effect size**.

Would the loss of 20 grams (i.e. the weight of a few potato chips) be practically significant, even if it were statistically significant?

Whether a result is practically significant depends on the effect size and the context of the research question. It's up to the researcher to know whether it is meaningful.

Sample size

As with the standard error (and a direct result of it), the p-value depends on the sample size. A very large sample size will give a *statistically significant* result in many cases, even with a very small effect size.

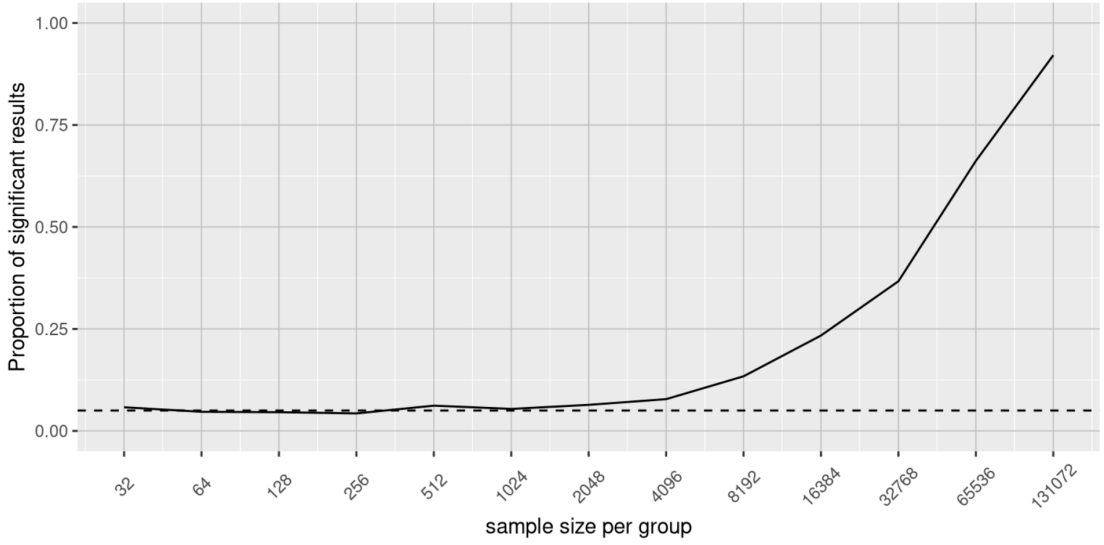


Figure 3: The proportion of significant results for a very small change (~20g which is about 0.001 standard deviations) as a function of sample size

Bibliography