**Notes from Statistical Foundations for Data Science: Unit 1**

**Observational studies**

Observational studies are ones where researchers observe the effect of a risk factor, diagnostic test, treatment or other intervention without trying to change who is or isn’t exposed to it. Cohort studies and case control studies are two types of observational studies.

**Cohort study:** For research purposes, a cohort is any group of people who are linked in some way. For instance, a birth cohort includes all people born within a given time frame. Researchers compare what happens to members of the cohort that have been exposed to a particular variable to what happens to the other members who have not been exposed.

**Case control study:** Here researchers identify people with an existing health problem (“cases”) and a similar group without the problem (“controls”) and then compare them with respect to an exposure or exposures.

An observational study also has many benefits and weaknesses. Benefits include:

* Can often be performed for very little expense
* Can study issues that would otherwise have ethical issues

Weaknesses include:

* Very difficult to conclude causation; without control of dependent variables, it is hard to say that one variable is actually caused by the explanatory variable,
* Difficult to control biases and variables
* Results are often seen as weaker

**Experimental studies**

Experimental studies are ones where researchers introduce an intervention and study the effects. Experimental studies are usually randomized, meaning the subjects are grouped by chance.

**Randomized controlled trial (RCT):** Eligible people are randomly assigned to one of two or more groups. One group receives the intervention (such as a new drug) while the control group receives nothing or an inactive placebo. The researchers then study what happens to people in each group. Any difference in outcomes can then be linked to the intervention.

An experimental study has both benefits and weaknesses. The benefits of an experimental study include:

* Typically seen as stronger research
* Can more definitively draw conclusions
* Can eliminate some or most variables
* Easier to control
* Is often shorter in duration than observational studies
* Can be tailored to answer a specific question

The weaknesses of an experimental study include:

* Typically very expensive
* Limited in ability based on ethical issues
* When studying humans, often see high dropout rates

### **Randomized Experiments vs. Observational Studies**

* Statistical inferences of cause-and-effect relationships can be drawn from randomized experiments but not from observational studies.
* **Confounding variable:** a variable (measured or not) that is related to both group membership and to the outcome.
* Confounding variables make it hard to establish an outcome as the direct consequence of a group membership.
  + Can attempt to compensate for other variables but can't account for everything

### **Role for Observational Data**

* Causation may be established in other ways.
  + Evidence may allow us to decide if explanatory variable has effect
* Ethical considerations may rule out a randomized experiment.
  + Pregnant women’s use of Marijuana -would be unethical
* Establishing causation is not always the goal.
* Evidence toward causal theories may be gathered.
* Observations may suggest a direction for future research.

### **Using Observational Studies to Claim Cause and Effect**

* In the 1940s and 1950s, the medical profession saw a rise in both lung cancer rates and smoking.
* People wondered, "Are these related?"
* Observational studies:
  + Does the same relationship occur in different populations, at different times, in various circumstances?
  + Does an increase in the dose of the explanatory variable create a corresponding change in the response?
  + Is there a reasonable mechanism to explain why smoking might cause lung cancer

### **Smoking and Lung Cancer: Conclusions**

* Researchers found many relationships between smoking and lung cancer:
  + There was a reasonable, physical mechanism.
  + The rate of lung cancer increased with the amount smoked.
  + The relationship was present across a variety of populations.
  + Theoretical considerations ruled out almost any other cause.
* Observational studies count: smoking research was done mostly through observational studies.

**Ethics of Gathing information**

Informed consent is a process of communication between you and your health care provider that often leads to agreement or permission for care, treatment, or services. Every patient has the right to get information and ask questions before procedures and treatments. If adult patients are mentally able to make their own decisions, medical care cannot begin unless they give informed consent.

The informed consent process makes sure that your health care provider has given you information about your condition along with testing and treatment options before you decide what to do.

This information can include:

* The name of your condition
* The name of the procedure or treatment that the health care provider recommends
* Risks and benefits of the treatment or procedure
* Risks and benefits of other options, including not getting the treatment or procedure

Signing informed consent means

* You have received all the information about your treatment options from your health care provider.
* You understand the information and you have had a chance to ask questions.
* You use this information to decide if you want to receive the recommended treatment option(s) that have been explained to you. Sometimes, you may choose to receive only part of the recommended care. Talk to your health care provider about your options.
* If you agree to receive all or some of the treatment options, you give your consent (agree) by signing a consent form. The completed and signed form is a legal document that lets your doctor go ahead with the treatment plan.

Institutional review boards determine whether a study is ethical and if it meets the threshold for informed consent. The review board would looks at all the controls and conditions of the experiments.

Confidentiality and anonymity are often used interchangeably, but they have different meanings and implications in qualitative research. Confidentiality means that you protect the identity and information of your participants from unauthorized access or disclosure. Anonymity means that you remove any identifying information from your data and reports, such as names, locations, or characteristics. Depending on your research design, context, and purpose, you may need to apply one or both of these principles to your data collection and analysis.

**Anonymity:**Providing anonymity of information collected from research participants means that either the project does not collect identifying information of individual persons (e.g., name, address, email address, etc.), or the project cannot link individual responses with participants’ identities. A study should not collect identifying information of research participants unless it is essential to the study protocol. Anonymity cannot be guaranteed if any personally identifiable (PII) information will be collected.

**Confidentiality:**Maintaining confidentiality of information collected from research participants means that only the investigator(s) can identify the responses of individual participants. Regardless, the researchers must make every effort to prevent anyone outside of the project from connecting individual subjects with their responses.

[**Clinical trials**](https://clinicaltrials.gov/ct2/about-studies/learn#ClinicalTrials)are research studies that test a medical, surgical, or behavioral intervention in people. These trials are the primary way that researchers determine if a new form of treatment or prevention, such as a new drug, diet, or medical device (for example, a pacemaker), is safe and effective in people. Often, a clinical trial is designed to learn if a new treatment is more effective or has less harmful side effects than existing treatments.

Other aims of clinical research include:

* Testing ways to diagnose a disease early, sometimes before there are symptoms
* Finding approaches to prevent a health problem, including in people who are healthy but at increased risk of developing a disease
* Improving quality of life for people living with a life-threatening disease or chronic health problem
* Studying the role of caregivers or support groups

### **Statistical Sampling**

A yellow oval with black text

Description automatically generated

* Taking a small sample from a larger group of individuals
* **Population:** the entire group of individuals we are interested in but can't usually assess directly
* **Sample:** the part of the population we actually examine and for which we have tangible data
* **Parameter:** a number describing a characteristic of the population
* **Statistic:** a number describing a characteristic of a sample

### **Sample Design Differences**

* Glider study is a randomized experiment.
* August's child study is an observational study.
* Sample design is important:
  + Randomized experiments can lead to causal statements.
  + Observational studies may include confounding variables.
  + In observational studies, researcher is passive.
  + In randomized experiments, researcher is active.

### **Randomization**

* Allows for impersonal chance
* Allows for multiple treatment groups
* Simple random sample
  + Select n individuals from the sampling frame such that every set of n individuals has an equal chance to be in the sample.
* Stratified random sample
  + Divide the sampling frame into groups of similar individuals.
  + Choose a separate SRS in each stratum.
  + Combine all SRSs to form the full sample.
* Multistage sampling and other techniques are still more complicated

### **Representative Samples**

* Samples should be taken at random from the population, even in observational studies.
* Key goal: The sample should be representative of the population.
* Key tasks
  + Get the sample from the population in the correct fashion.
  + Further randomize the sample when necessary.

Drawing Statistical conclusions:

Example of Test-Preparation Methods:

**Hypotheses Testing: Six Steps of One-Sample t-Test**

The t-test is one of the most common hypothesis tests in statistics. The t-test determines either whether the sample mean and the mean of the population differ or if two sample means differ statistically. The t-test distinguishes between

* one sample t-test
* [independent sample t-test](https://datatab.net/tutorial/unpaired-t-test)
* [paired samples t test](https://datatab.net/tutorial/paired-t-test)

A picture containing text, screenshot, font, design

Description automatically generated

The choice of which t-test to use depends on whether one or two samples are available. If two samples are available, a distinction is made between dependent and independent samples.

In this tutorial you will find everything about the **one-sample t-test**.

## Step 1: State your null and alternate hypothesis

After developing your initial [research hypothesis](https://www.scribbr.com/methodology/hypothesis/) (the prediction or assumption that you want to investigate), it is important to restate it as a null (Ho) and alternate (Ha) hypothesis so that you can test it mathematically. (H0 = *μ* n – *μ* t = 0)

* **Null hypothesis (H0):**There’s **no effect** in the [population](https://www.scribbr.com/methodology/population-vs-sample/). **An assumption that there is no difference in the mean.**
* **Alternative hypothesis (Haor H1):**There’s an **effect** in the population. There is a difference.

The **alternate hypothesis** is usually your initial hypothesis that predicts a relationship between variables. The [**null hypothesis**](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/) is ***a prediction of no relationship between the***[***variables***](https://www.scribbr.com/methodology/types-of-variables/) you are interested in. (Ha = *μ* n – *μ* t ≠ 0)

Hypothesis testing example:

You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis ***that men are, on average, taller than women.*** To test this hypothesis, you restate it as:

* H0: Men are, on average, not taller than women.  
  Ha: Men are, on average, taller than women.

**SMU Example:**

We would like to test the claim (Ha) that the true population *μ* mean is greater than 4. Therefore, the null hypothesis or the nullifying claim (H0) could be the population mean is less than or equal to 4. Let’s write it down.

**Step 1:** Identify the null (H0) and alternative (Ha) hypotheses.

|  |  |  |  |
| --- | --- | --- | --- |
| *μ* ≤ 4 – counter-argument | H0: *μ* ≤ 4 | The null hypothesis seeks to refute the claim or show the opposite of the claim and that there is no correlation between variables.  Note: Null hypothesis always has an equal sign. |  |
| *μ* > 4 – assumption | Ha: *μ* > 4 | This is the claim. |  |

**Step 2:** Draw and shade the model and find the ***critical value***.

First draw the model by looking at the alternative hypothesis and modeling it based on the mathematical sign (=, >, <) in that hypothesis. Next, choose either the [**one tailed T Distribution table**](https://www.statisticshowto.com/tables/t-distribution-table/) or [**two tailed T Distribution table**](https://www.statisticshowto.com/tables/t-distribution-table/#two)**.**

A picture containing diagram, line, text, plot

Description automatically generatedThe Alternative Hypothesis (Ha) can help you determine the direction of your tail or tails. Models are determined as follows:

* The true population mean is not equal to the null hypothesized value (Two-tailed). Ha: μ≠μ0

A diagram of a normal distribution

Description automatically generated with low confidence

* The true population mean is greater than hypothesized null value (Upper or right-tailed). Ha: μ>μ0

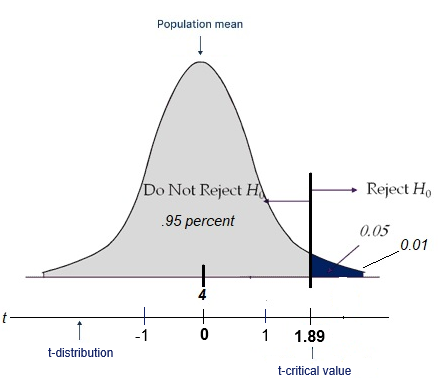
A diagram of a normal distribution

Description automatically generated with low confidence

* The true population mean is less than hypothesized null value (Lower or left-tailed). Ha: μ<μ0

**Step 2 Execution**: In our case, the alternative hypothesis (*Ha: μ* > 4) has a true population mean greater than 4, meaning a right-tailed test. Because we don’t know the standard deviation, we need to calculate a t-score. A t-score is found along the t-distribution line which is based on the curve of the model and is centered on the mean. The t-score is the number of standard deviations from the mean. We are using a t-score or the **critical value** because the sample is small and we don’t know the population variance.

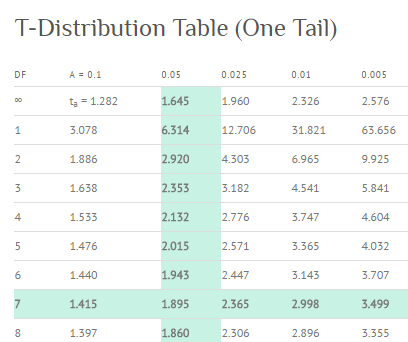
A **critical value**is a line on a graph that splits the graph into sections. One or two of the sections is the “[rejection region](https://www.statisticshowto.com/rejection-region/)“; if your test value falls into that region, then you [reject the null hypothesis](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/support-or-reject-null-hypothesis/).

****Note: The critical value of z is term linked to the area under the [standard normal model](https://www.statisticshowto.com/probability-and-statistics/normal-distributions/). Critical values can tell you what probability any particular data point will have. A critical value of z ([Z-score](https://www.statisticshowto.com/probability-and-statistics/z-score/)) is used when the [sampling distribution](https://www.statisticshowto.com/probability-and-statistics/sampling-in-statistics/sampling-distribution/) is normal, or close to normal. ***Z-scores are used when the population standard deviation is known or when you have larger***[***sample sizes***](https://www.statisticshowto.com/probability-and-statistics/find-sample-size/)***.*** While the z-score can also be used to calculate probability for unknown standard deviations and small [samples](https://www.statisticshowto.com/sample/), ***in real life you’ll probably use the***[***t distribution***](https://www.statisticshowto.com/probability-and-statistics/t-distribution/)***to calculate these probabilities***. That’s ***because you often don’t know the***[***population variance***](https://www.statisticshowto.com/population-variance/) (which is a requirement for using the[z test](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/z-test/)).

To find the t-score or the **critical value** we need to follow the steps below:

Step A: Subtract one from your [sample size.](https://www.statisticshowto.com/probability-and-statistics/find-sample-size/) This is your df, or [degrees of freedom](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/degrees-of-freedom/). For example, if the sample size is 8, then your df is 8 – 1 = 7.

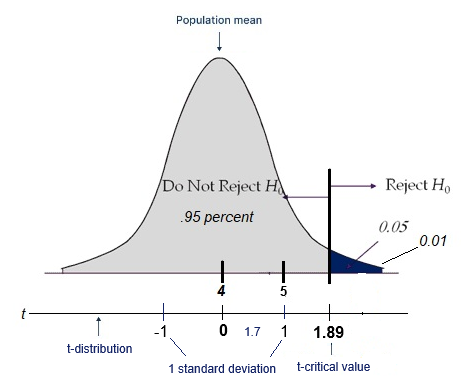
Step B: Choose an [alpha level.](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-an-alpha-level/) The alpha level is usually given to you in the question — the most common one is 5% (0.05).

[](https://www.statisticshowto.com/wp-content/uploads/2017/04/t-critical-value.png)Step C: Look up the df in the left hand (upper) side of the t-distribution table and the alpha level along the top row. Find the intersection of the row and column. For this example (7 df, α = .05,) the **t-critical value** is 1.895.

**Step 3: Find the test statistic:**

We would like to test the claim that the population mean is greater than 4. To do this, we take a sample of size *n* = 8 and find that sample mean is x¯ = 5.2. We also determine that the standard deviation is (sigma) *s* = 1.7.

|  |  |  |
| --- | --- | --- |
| **Step 1:** | Identify the null (H0) and alternative (Ha) hypotheses. | H0: *μ* = 4 Ha: *μ* > 4 |
| **Step 2:** | Draw and shade and find the critical value. | *α* = .05 (significance level) *df* = 8 − 1 = 7  t-critical value = 1.89 |

**Step 3 Execution:** To do this, we need to find the test-statistic value (t-value) (Note: that the t-value is different from the t-score or critical value). Find the **test-statistic value (t-value)** for the data and compare it to the t-score of the alternative hypotheses. To do this you use the following calculation to determine the value of t in the sample. Here we see that the t-value is 1.99, which is in the reject region of the model.

A picture containing screenshot, black, darkness, space

Description automatically generated

If the t-value in our data is greater than or has more standard deviations from the 1.89 critical value we reject the null hypothesis H0: μ≤μ0.

**Step 4: Quantify the difference with a p-value**

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

In most cases you will use the [p-value](https://www.scribbr.com/statistics/p-value/) generated by your statistical test to guide your decision. And in most cases, your predetermined [level of significance](https://www.scribbr.com/statistics/statistical-significance/) for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis ([Type I error](https://www.scribbr.com/statistics/type-i-and-type-ii-errors/)).

What is a p-value? The **p-value** is the probability of observing by random chance a result as extreme or more extreme than was what was actually observed under the assumption that the null hypothesis is true.

A diagram of a normal distribution

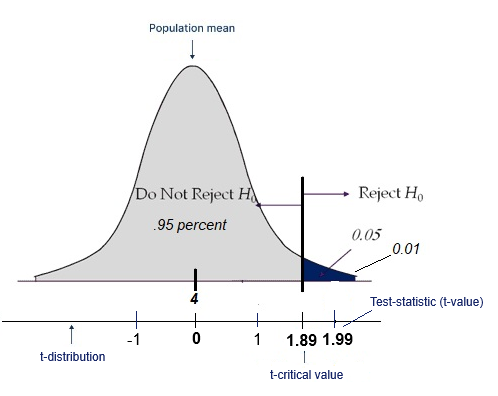
Description automatically generatedWe would like to test the claim that the population mean is greater than 4. To do this, we take a sample of size *n* = 8, find that x¯ = 5.2, *s* = 1.7, and test statistic (t-value) is 1.99.

|  |  |  |
| --- | --- | --- |
| **Step 1:** | Identify the null (H0) and alternative (Ha) hypotheses. | H0: *μ* = 4 Ha: *μ* > 4 |
| **Step 2:** | Draw and shade and find the critical value. | *α* = .05 (significance level) *df* = 8 − 1 = 7 |
| **Step 3:** | Find the test statistic (*t*-value) for the data. | Numbers on a black background  Description automatically generated with medium confidence |

**Step 4 execution:** Find the ***p*-value**: the probability of observing something as extreme or more extreme than what was observed, by random chance, under the assumption that the null hypothesis is true.

* The blue-shaded region
* Usually calculated by software: Used R to calculate (p-value = 0.04345)

**Step 5: Decide to reject or fail to reject (FTR) H0**

We would like to test the claim that the population mean is greater than 4. To do this, we take a sample of size *n* = 8 and find that x¯ = 5.2, *s* = 1.7 and the p-value = 0.04345.

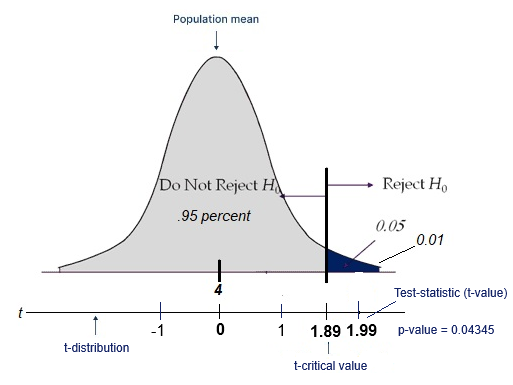
|  |  |  |
| --- | --- | --- |
| **Step 1:** | Identify the null (H0) and alternative (Ha) hypotheses. | H0: *μ* = 4 Ha: *μ* > 4 |
| **Step 2:** | Draw and shade and find the critical value. | *α* = .05 (significance level) *df* = 8 − 1 = 7 |
| **Step 3:** | Find the test statistic (*t*-value) for the data. | Numbers on a black background  Description automatically generated with medium confidence |
| **Step 4:** | Find the ***p*-value**: the blue-shaded region, usually calculated by software. | *Is the p*-value < .05  p-value = 0.04345  If so then the sample mean is in the reject area. |

***P(test statistic | H0) > alpha (5%): failed to reject H0 , H0 is true and Not statistically significant***

***P(test statistic | H0) ≤ alpha (5%): reject H0 , H0 is rejected and statistically significant***

**Step 5 execution:** Reject or fail to reject (FTR) H0. This is dependent on the p-value. If the p-value is less than alpha (.05) then you will reject H0. If it is greater than alpha you will fail to reject the H0. In this case: we reject H0 because the p-value < .05 and in the blue area.

**Step 6 After rejecting or failing to reject the hypothesis, you write up a conclusion.**

We would like to test the claim that the population mean is greater than 4. To do this, we take a sample of size *n* = 8 and find that x¯ = 5.2 and *s* = 1.7.

|  |  |  |
| --- | --- | --- |
| **Step 1:** | Identify the null (H0) and alternative (Ha) hypotheses. | H0: *μ* ≤ 4 Ha: *μ* > 4 |
| **Step 2:** | Draw and shade and find the critical value. | *α* = .05 (significance level) *df* = 8 − 1 = 7 |
| **Step 3:** | Find the test statistic (*t*-value) for the data. | Numbers on a black background  Description automatically generated with medium confidence |
| **Step 4:** | Find the ***p*-value**: the blue-shaded region, usually calculated by software. | *p*-value = 0.04345  p-value < .05 |
| **Step 5:** | Reject or fail to reject (FTR) H0. | Reject H0 |

**Step 6:** Conclusion: There is enough evidence to suggest that the true mean is greater than 4 (*μ* > *μ0).* (Because *p*-value = 0.04345 < .05 or .04 is a smaller percentage than .05. H0 is rejected).

* Always phrase the conclusion in the language of the problem.
* The smaller the *p*-value, the stronger the evidence is to reject H0.
* Note: the scale of blue is from .05 to .01 to .0 as it moves to the right and .95 percent of the values lay under the curve which would count upward to the left.

### Communicating Results

* What is the research question?
  + E.g., "How does intrinsic or extrinsic motivation affect the scores of students?"
* What were the treatment groups?
  + (1) Intrinsically motivated
  + (2) Extrinsically motivated
* Need to know:
  + How motivations were determined
  + What was measured
  + How responses were obtained
* Where possible: additional details about the measurement tool

### Descriptive Statistics

* Mean
* Standard deviation
* Other measurements
* Outliers and possible explanations
* Important differences
* The p-value and whether it supports the H0
* What rejecting (or accepting) the H0 means
  + E.g., do intrinsically motivated students seem to score better?
* Difficulties
* Confounding variables
* Test environment
* Other questions like:
  + How can we generalize the results?
  + Can we generalize the results?

### Summary

* Report more than a few numbers and whether the H0 was rejected
* Share the context and help your audience make its own conclusions
* Share enough details to let a reader duplicate your findings

Variance is the square of the standard deviation –

Scope of Inference – Looks to establish a causal relationship in the study design. (oral exam)

-----------------------------------------------------------------------------------------------------

Null Hypothesis /Assumption – playing fair, but the probability is small

* Reject the null – not enough evidence to support
* Fail to reject null – sufficient evidence to suggest

Alternative Hypothesis – not playing fair –

Creativity study - The null hypothesis is rejected at the 95% confidence level.

What is the probability that the experiment would happen again

Permutation test – if the null is true then that subject would have got the same score in the other group. So scrambling the labels it would not really matter.