STAT 305: Chapter 6 - Part II

Hypothesis Testing

Amin Shirazi

ashirazist.github.io/stat305.github.io

Hypothesis Testing

Deciding What's True (Even If We're Just Guessing)

Let's Play A Game

A "Friendly" Introduction to Hypothesis Tests

My Game

The Rules

Let's Play A Game

The semester is getting a little intense! You are a livinLet's break the tension with a friendly game.

Here are the rules:

- I have a new deck of cards. 52 Cards, 26 with Suits that are Red, 26 with Suits that are Black
- You draw a red-suited card, you give me a dollar
- You draw a black-suited card, I give you two dollars

Quick Questions

What is the expected number of dollars you will win playing this game?

Would you play this game?

Are We Forgetting Something?

My Game

The Rules

The Assumptions

Be Careful About Your Assumptions

Pause for a minute and think about what you are assuming is true when you play this game. For instance,

- You assume I'm going to shuffle the cards fairly
- You assume there are 52 cards in the deck
- You assume the deck has 26 red-suited cards in it
- You assume the deck has **a** red-suited card in it

How can we make sure the assumptions are safe??

- Shuffling assumption: watch me shuffle, make sure I'm not doing magic tricks, etc
- 52 Cards assumption: count the cards
- Red-suit assumption: Count the number of red cards

Whew! We can actually make sure all of our assumptions are good!

One Problem

I Refuse to Show You The Cards

Do You Trust Me?

My Game

The Rules

The Assumptions

Our Assumptions

I'm not going to show you all the cards. In other words, I refuse to show you the *population of possible outcomes*. This is justified: we are in a statistics course after all.

So, let's start with our unverifiable assumption: Is it safe to assume that this is a fair game. Why would we make this assumption?

- You trust that I'm (basically) an honest person (*assumption of decency*)
- You trust that I'm getting paid enough that I wouldn't risk cheating students out of money (*assumption of practicality*)
- You saw the deck was new (*manufacturer trust assumption*)
- You want it to be an fair game because you would win lots of money if it was (*assumption in self-interest*)

Our Assumptions

In statistical terminology, we wrap all these assumptions up into one assumption: our "**null hypothesis**" is that the game is not rigged - that the probability of you winning is 0.5

> **Null Hypothesis**
> The assumptions we are operate under in normal circumstances (i.e., what we believe is true). We wrap these assumptions up into a statistical/mathematical statement, but we will accept them unless we have reason to doubt them. We use the notation H_0 to refer to the null hypothesis.

In this case, we could say that the probability of winning is p and that would make our null hypothesis

$$H_0 : p = 0.5$$

My Game

The Rules

The Assumptions

Our Assumptions

Of course our assumptions could be wrong. We call the other assumptions our "alternative hypothesis":

> **Alternative Hypothesis**
> The conditions that we do require proof to accept. We would have to change our beliefs based on evidence. We use the notation H_A (or sometimes, H_1) to refer to the alternative hypothesis.

In this case, we could say that our alternative to believing the game is "fair" is to believe the game is not fair, or that the probability of winning is not 0.5. We write:

$$H_A : p \neq 0.5$$

or $H_A : p > 0.5$

$H_A : p < 0.5$

or $H_A : p = 0.65$

A Compromise

I Won't Show You All The Cards

But I Will Let You Test The Game

My Game

The Rules

The Assumptions

The Test

Testing the Game

The test of whether or not the game is worth playing can be defined in term of whether or not our assumptions are true. In other words, we are going to test whether our null hypothesis is correct:

> **Hypothesis Tests**
> A **hypothesis test** is a way of checking if the outcomes of a random experiment are *statistically unusual* based on our assumptions. If we see really unusual results, then we have **statistically significant** evidence that allows us to **reject our null hypothesis**. If our assumptions lead to results that are not unusual, then we **fail to reject our null hypothesis**.

My Game

The Rules

The Assumptions

The Test

Testing the Game

So how can we test the game? What if we tried a single round of the game?

- What are the probabilities of the outcome of a single game?
- If we draw a single card do we have enough evidence that the game is fair?
- Do we have enough evidence that the game is rigged?

Based on a single round of the game, both of the possibel outcomes are pretty normal - that's not good enough.

If we draw a losing card, then we might be inclined to call the game unfair - even though a losing card is pretty common for a single round of the game

If we draw a winning card, then we might be inclined to call the game fair - even though a winning card may be common even when the game is not fair!

We can make lots of mistakes!!

The Mistakes We Might Make

We could of course be wrong: For instance, we could, just by random chance, see outcomes that are unusual for the assumptions we make and reject the assumptions even if (in reality they are true). This is called a "Type I Error"

> **Type I Error**
> When the results of a hypothesis test lead us to reject the assumptions, while the assumptions are actually true, we have committed a Type I Error.

My Game

The Rules

The Assumptions

The Test

The Errors

The Mistakes We Might Make

A common example of this is found in criminal court:

- We assume that a individual accused of a crime is innocent (our assumption)
- After examinig the evidence, we conclude that it is there is no reasonable doubt the person is not innocent (in other words, we reject the assumption because it is very unlikely to be true based on our evidence).
- If the person truly is innocent, then we have committed a Type I error (rejecting assumptions that were true).

The Mistakes We Might Make

We could also make a different error: we could choose not to reject the assumptions when in reality the assumptions are wrong.

> **Type II Error**
> When the results of a hypothesis test lead us to fail to reject the assumptions, while the assumptions are actually false, we have committed a Type II Error.

My Game

The Rules

The Assumptions

The Test

The Errors

The Mistakes We Might Make

Again, if we consider the example of criminal court:

- We assume that a individual accused of a crime is innocent (our assumption)
- After examinig the evidence, we conclude that it is there is **not** evidence beyond a reasonable doubt the person is not innocent (in other words, the evidence is not enough to reject our assumption because it is still reasonable to doubt the accused's guilt).
- If the person truly is not innocent, then we have committed a Type II error (failing to reject assumptions that were false).

In general, we want to make sure that a Type I error is unlikely. To take the example of court again,

- We commit a Type II error: a guilty person goes free
- We commit a Type I error: an innocent person goes to jail; the guiilty person is still free

The Mistakes We Might Make

Let's go back to my game: We assume I am an honest person (i.e., we assume that the probability of winning a single game is $p = 0.5$)

Type I Error: Rejecting True Assumptions

- We gather evidence
- Looking at our evidence, we decide that the game was not fair even though it was.
- Fallout: you slander me, you disparge me, we have a fight, BOOOM.

Type II Error: Failing to Reject False Assumptions

- We gather evidence
- Looking at our evidence, we decide that the game was fair even though it was not.
- Fallout: you play the game and lose some money.

Ideally, we won't make either error. However, we can only base our decision of our evidence we can gather - the truth is out of our grasp!

My Game

The Rules

The Assumptions

The Test

The Errors

The Evidence

Gathering Statistical Evidence

Okay, so we don't want to make either error - that means we need good evidence.

Like we talked about before, even if the game is fair one test round of the game would not be enough to make a good decision since drawing a red-suited card and drawing a black-suited card are both pretty normal for a single round of the game.

But what if we played the game 10 times in a row? After 10 rounds, do you think we would have enough evidence to make a decision about our assumption?

p-value

If we assume the null hypothesis, then we can make some assumptions about what results are likely and what results are unlikely. We describe the likelihood of the results that we actually get using a **p-value**

> **p-value**
> After gathering evidence (aka, data) we can determine the probability that we would have gotten the evidence we did if our assumptions were true. That probabiliity is called the p-value. If the p-value is really, really small that means that the assumptions we started with are pretty unlikely and we reject our assumptions. If the p-values is not small, then the evidence collected (aka, the data) is pretty normal for our assumptions and we fail to reject our assumptions.

p-value = P(observing some unusual events under H_0)

lets assume

p=0.5

My Game

The Rules

The Assumptions

The Test

The Errors

The Evidence

p-value

p-value

In other words, we collect evidence and determine a way to measure the whether or not our data was unusual *if our assumptions are true*.

If we have a very, very low chance of

- seeing both our results and
- having true assumptions then we reject the assumptions

Going along with the terminology we have introduced, if we have a small p-value then we reject our null hypothesis.

Gathering Statistical Evidence

In this game, if we assume that the game is fair, we have

- two outcomes: success (winning) and failure (losing)
- a constant chance of a successful outcome ($p = 0.5$), assuming the game is fair)
- independent rounds of the game (assuming fair shuffle, which we can check)

In other words, if we test the game 10 times we can model the number of successful outcomes as binomial: For X = the total number of wins,

$$P(X = x) = \frac{10!}{x!}(10 - x)!(0.5)^x(1 - 0.5)^{10-x}$$

This gives us a way of getting our p-value

Let's Test the Game

Gathering Statistical Evidence

We played the game. Let's figure out whether our results were unusual or not.

Again, we assume the game is fair and have decided that the number of times we win will follow a binomial distribution with probabiliity function

$$P(X = x) = \frac{10!}{x!}(10 - x)!(0.5)^x(1 - 0.5)^{10-x}$$

Now we need to make a conclusion: do we accept or reject our assumptions? What do we consider unusual? Is it fair to decide after we play?

$$P(X=0) = \binom{10}{0}(0.5)^0(0.5)^{10-0}$$

$$= (1/2)^{10} = 0.00097$$

Summary

- Sometimes we can know if something is true or not by examining the truth directly, but not always
- When we can't examine the truth, we need to test what we believe to be true
- A statistical test is a tool for testing our assumptions about what we believe
    - We state our assumed belief (generally our current beliefs, or the ethical beliefs, or the beliefs we hope are true, …)
    - We come up with a way of collecting data that could validate or invalidate our assumption
    - We measure how likely it was that we would have gathered the data we did if our assumptions were correct
    - We reject the assumptions if our data is very unlikely we are our current beliefs

Now let's make everything a little more formal

Section 6.3

Hypothesis Testing

Hypothesis Testing

Hypothesis testing

Last section illustrated how probability can enable confidence interval estimation. We can also use probability as a means to use data to quantitatively assess the plausibility of a trial value of a parameter.

Statistical inference is using data from the sample to draw conclusions about the population.

> 1. **Interval estimation (confidence intervals):**
>   Estimates population parameters and specifying the degree of precision of the estimate.
>
> 1. **Hypothesis testing:**
>   Testing the validity of statements about the population that are formed in terms of parameters.

i.e: we test the validity of population parameter.

Hypothesis Testing

Null

Definition:

↳ In STAT 305, only μ (population mean).

Statistical **significance testing** is the use of data in the quantitative assessment of the plausibility of some trial value for a parameter (or function of one or more parameters).
Significance (or hypothesis) testing begins with the specification of a trial value (or **hypothesis**).

A **null hypothesis** is a statement of the form

$$\underbrace{\text{Parameter}}_{\mu} = \#$$

or

$$\text{Function of parameters} = \#$$

for some # that forms the basis of investigation in a significance test. A null hypothesis is usually formed to embody a status quo/"pre-data" view of the parameter. It is denoted H_0.

Hypothesis Testing

Null

Alternative

Definition:

An **alternative hypothesis** is a statement that stands in opposition to the null hypothesis. It specifies what forms of departure from the null hypothesis are of concern. An alternative hypothesis is denoted as H_a. It is of the form

$$\underbrace{\text{Parameter}}_{\mu} \neq \#$$

or

$$\text{Parameter} > \# \quad \text{or} \quad \text{Parameter} < \#$$

Examples (testing the true mean value):

$$\begin{array}{lll} \mathrm{H}_0 : \mu = \# & \mathrm{H}_0 : \mu = \# & \mathrm{H}_0 : \mu = \# \\ \mathrm{H}_a : \mu \neq \# & \mathrm{H}_a : \mu > \# & \mathrm{H}_a : \mu < \# \end{array}$$

composite alternative.

Often, the alternative hypothesis is based on an investigator's suspicions and/or hopes about th true state of affairs.

Hypothesis Testing

Null

Alternative

The **goal** is to use the data to debunk the null hypothesis in favor of the alternative.

1. Assume H_0.
2. Try to show that, under H_0, the data are preposterous. (using probability)
3. If the data are preposterous, reject H_0 and conclude H_a.

The outcomes of a hypothesis test consists of:

The ultimate decision in favor of:

Fixed prior to experiment

$P(\text{Type I error}) = \alpha$

The prob. of rejecting H_0 when H_0 is actually true!

H_0 is actually true, & we fail to reject H_0.

True state of experiment:

	H_0	H_a
H_0	OK	Type I error.
H_a	Type II error	OK

The prob. of accepting H_0 when it's actually wrong

Hypothesis Testing

Null

Alternative

Probability of type I error

H_a is actually true & we reject H_0 in favor of H_a.

It is not possible to reduce both type I and type II erros at the same time. The approach is then to fix one of them. & reduce the other one.

We then fix the **probability of type I error** and try to minimize the probability of type II error.

* > We define the probability of type I error to be α (the significance level)

recall: in CI: we have $(1-\alpha)$: the confidence level.

Hypothesis Testing

Null

Alternative

Example: [Fair coin]

Suppose we toss a coin $n = 25$ times, and the results are denoted by $X_1, X_2, \ldots, X_{25}$. We use 1 to denote the result of a head and 0 to denote the results of a tail. Then $X_1 \sim Binomial(1, \rho)$ where ρ denotes the chance of getting heads, so $\mathrm{E}(X_1) = \rho$, $\mathrm{Var}(X_1) = \rho(1-\rho)$. Given the result is you got all heads, do you think the coin is fair?

$= \mathrm{Bernoulli}(p)$

$$\text{Null hypothesis}: H_0: \text{the coin is fair or } H_0: \rho = 0.5$$
$$\text{Alternative hypothesis}: H_a: \rho \neq 0.5$$

* If H_0 was correct, then $P(\text{results are all heads}) = (1/2)^{25} < 0.000001$

observed 25 heads.

> I don't think this coin is fair (reject H_0 in favor of H_a)

In the real life, we may have data from many different kinds of distributions! Thus we need a universal framework to deal with these kinds of problems.

We have $n = 25 \geq 25$ iid trials $\Rightarrow$ By CLT we know if $H_0 : \rho = 0.5 (= \mathrm{E}(X))$ then

under H_0

$$\frac{\overline{X} - \rho}{\boxed{\sqrt{\rho(1-\rho)/n}}} \sim N(0,1)$$

$\to SD(\bar{X}) = \sqrt{\frac{Var(X)}{n}}$

all 25 obs. are heads.

$\to$ We obsrved $\boxed{\overline{X} = 1,}$ so

$\bar{x} = \frac{1}{25}(x_1 + \cdots + x_{25})$

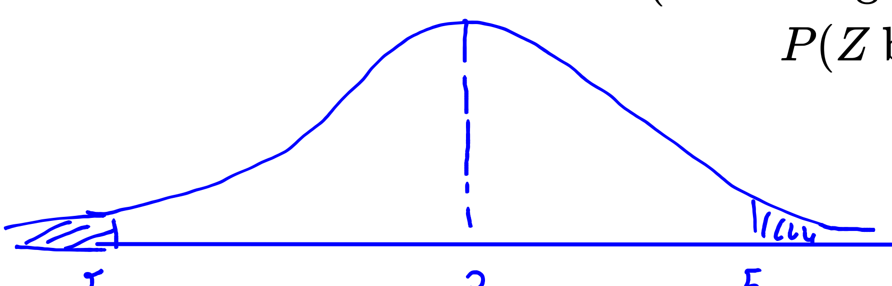
$= \frac{1}{25}(25) = 1$

test statistic

$$\frac{\overline{X} - 0.5}{\sqrt{0.5(1-0.5)/25}} = \frac{1 - 0.5}{\sqrt{0.5(1-0.5)/25}} = 5$$

Then the probability of seeing as *wierd or wierder* data is

$$P(\text{Observing something wierd or wierder}) = P(Z \text{ bigger than 5 or less than -5}) < 0.000001$$



Significance tests for a mean

Definition:
A **test statistic** is the particular form of numerical data summarization used in a significance test.

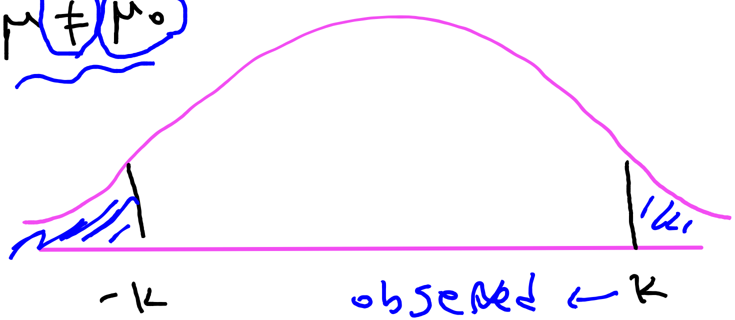
Definition:
A **reference (or null) distribution** for a test statistic is the probability distribution describing the test statistic, provided the null hypothesis is in fact true.

Definition:
The **observed level of significance or p-value** in a significance test is the probability that the reference distribution assigns to the set of possible values of the test statistic that are *at least as extreme as* the one actually observed.

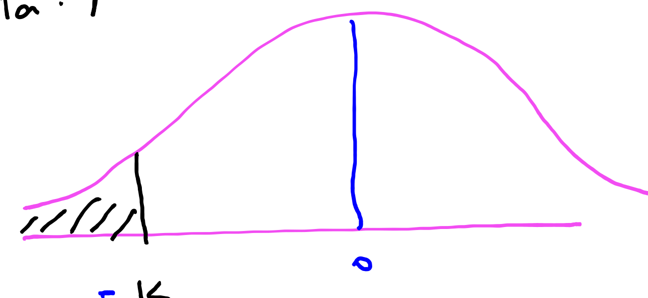
$H_0: \mu = \mu_0$

known

$H_a: \mu \neq \mu_0$



$H_a: \mu < \mu_0$



value of test
statistic.

Hypothesis Testing

Null

Alternative

P-value

Significance tests for a mean

In the previous example, the test statistic was

$$\frac{\overline{X} - \rho}{\sqrt{\rho(1-\rho)/n}} \sim N(0,1)$$

In the previous example, the null distribution was $N(0,1)$

In the previous example, the p-value was < 0.000001

because the p-value < α (significant level)

we reject the null Hypotesis.

i.e we have enough evidence to reject H_0 &
state that the coin is not fair ($\rho \neq 0.5$)

Significance tests for a mean

In other words:

Let K be the test statistics value based on the data

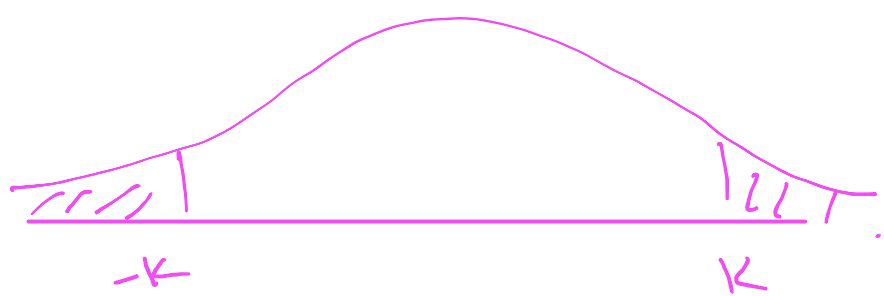
Say

know

$$\begin{cases} H_0 : \mu = \mu_0 \\ \boxed{H_a : \mu \neq \mu_0} \end{cases}$$

$$P(\text{observing data as or more extreme as K})$$
$$= P(Z < -K \; or \; Z > k)$$

is defined as the p-value



Significance tests for a mean

Based on our results from Section 6.2 of the notes, we can develop hypothesis tests for the true mean value of a distribution in various situations, given an iid sample $X_1, \ldots, X_n$ where $H_0 : \mu = \mu_0$.

Let K be the value of the test statistic, $Z \sim N(0, 1)$, and $T \sim t_{n-1}$. Here is a table of p-values that you should use for each set of conditions and choice of H_a.

critical value

Situation	K	$H_a : \mu \neq \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$
$n \geq 25, \sigma$ known	$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$P(\lvert Z\rvert > K)$	$P(Z < K)$	$P(Z > K)$
$n \geq 25, \sigma$ unknown	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$P(\lvert Z\rvert > K)$	$P(Z < K)$	$P(Z > K)$
$n < 25, \sigma$ unknown	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$P(\lvert T\rvert > K)$	$P(T < K)$	$P(T > K)$

(data are iid $\sim N(0,1)$)

possible test statistic

Hypothesis Testing

Null

Alternative

P-value

Steps to perform a hypothesis test

1. State H_0 and H_1
2. State α, significance level, usually a small number (0.1, 0.05 or 0.01)
3. State form of the test statistic, its distribution under the null hypothesis, and all assumptions
4. Calculate the test statistic and p-value
5. Make a decision based on the p-value(if p-value < α, reject H_0 otherwise we fail to reject H_0)
6. Interpret the conclusion using the consept of the problem

I need all 6 steps on HW & exam.

Hypothesis Testing

Null

Alternative

P-value

Example:[Cylinders]

$n = 40$

The strengths of 40 steel cylinders were measured in MPa. The sample mean strength is 1.2 MPa with a sample standard deviation of 0.5 MPa. At significance level $\alpha = 0.01$, conduct a hypothesis test to determine if the cylinders meet the strength requirement of 0.8 MPa.

$\bar{X}$

S

μ mean

1, $H_0: \mu = 0.8 \ (= \mu_0)$

$H_a: \mu \neq 0.8$

2, $\alpha = 0.01$

3, $n \geq 25$ & σ (population variance) is unknown => The test statistic is

$$K = \frac{\bar{X} - \boxed{\mu_0}}{S/\sqrt{n}},$$ and K has a $N(0,1)$.

under H_0.

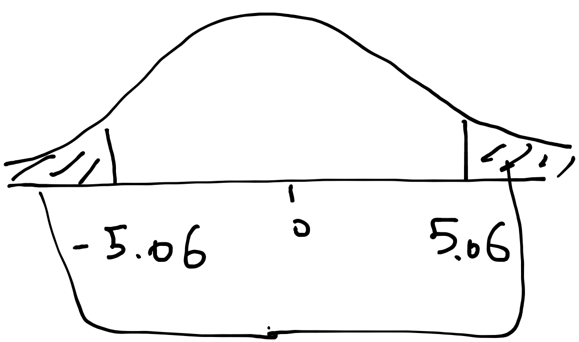
4, $k = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{1.2 - 0.8}{0.5/\sqrt{40}} = 5.06$

p-value: P(observing as or more extreme values than $k = 5.06$)

$$= P(|Z| > k) = P(|Z| > 5.06)$$

$$= P(Z > 5.06 \text{ or } Z < -5.06)$$

$$= P(Z > 5.06) + P(Z < -5.06)$$



same area

$$= 1 - \Phi(5.06) + \Phi(-5.06)$$

$$= 2\underbrace{\Phi(-5.06)}_{\approx 0} \approx 0$$

5, Since the P-value < 0.05, I reject H_0 in favor of H_a.

6, There's enough evidence to conclude that the mean cylinders doesn't meet the requirment of 0.8 MPa.

Hypothesis Testing

Null

Alternative

P-value

Example: [Concrete beams]

10 concrete beams were each measured for flexural strength (MPa). The data is as follows.

$n=10$ → [1] 8.2 8.7 7.8 9.7 7.4 7.8 7.7 11.6 11.3 11.8

The sample mean was $\overset{\bar{x}}{9.2}$ MPa and the sample variance s^2 was 3.0933 MPa. Conduct a hypothesis test to find out if the flexural strength is different from 9.0 MPa. at $\alpha=0.01$ level.

1. $H_0: \mu = \underbrace{9}_{\mu_0}$ vs. $H_a: \mu \neq 9$

2. $\alpha = 0.01$

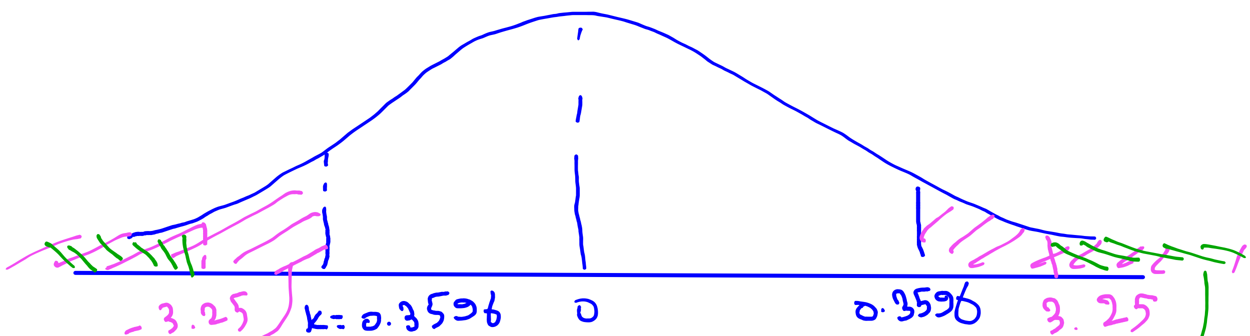
3. $n=10\ (<25)$ + σ is unknown. So,

$$K = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \quad \left(\text{we know that } K \sim t_{n-1}\right)$$

assumption for t-student also: $X_1, \ldots, X_{10} \overset{iid}{\sim} N(\mu, \sigma^2)$

4. $K = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{9.2 - 9}{\sqrt{\dfrac{3.0933}{10}}} = 0.3596$

$\text{P-value} = P(|T| > K) = P(|T| > 0.3596)$



$t_{9,\,1-\alpha/2} = t_{9,\,0.995} = 3.25$

by the t table

green shaded area

p-value

pink shaded area

by t-table $= \underbrace{P(|T| > 0.3596)}_{\text{p-value}} > \underbrace{P(|T| > t_{9,\, 0.995})}_{\alpha = 0.01}$

5. Since p-value is $> \alpha = 0.01$, we fail to reject the null Hypothesis (H_0)

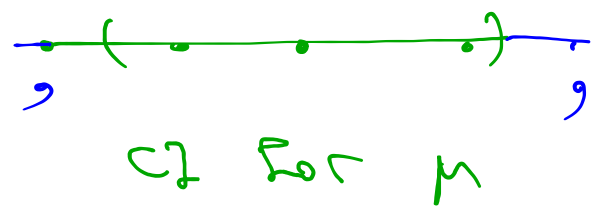
6. There is NOT enough evidence to conclude that the true mean flexural strength of the beams is different from 9 Mpa.

Hypothesis Testing Using Confidence Interval

Hypothesis testing using the CI

We can also use the $1 - \alpha$ confidence interval to perform hypothesis tests (instead of p-values). The confidence interval will contain μ_0 when there is little to no evidence against H_0 and will not contain μ_0 when there is strong evidence against H_0.

$$\begin{cases} H_0: \mu = 9 \\ H_a: \mu \neq 9 \end{cases}$$



Hypothesis Testing

Null

Alternative

P-value

CI method

Hypothesis testing using the CI

$H_0: \mu = \mu_0$

$H_a: \begin{cases} \mu \neq \mu_0 \\ \mu > \mu_0 \\ \mu < \mu_0 \end{cases}$

Steps to perform a hypothesis test using a confidence interval:

1. State H_0 and H_1
2. State α, significance level
3. State the form of $100\,(1-\alpha)$ % CI along with all assumptions necessary. (use one-sided CI for one-sided tests and two-sided CI for two sided tests)

   confidence level.
4. Calculate the CI
5. Based on $100\,(1-\alpha)$ % CI, either reject H_0 (if μ_0 is not in the interval) or fail to reject (if μ_0 is in the interval)
6. Interpret the conclusion in the content of the problem

Hypothesis Testing

Null

Alternative

P-value

CI method

Example:[Breaking strength of wire, cont'd]

Suppose you are a manufacturer of construction equipment. You make 0.0125 inch wire rope and need to determine how much weight it can hold before breaking so that you can label it clearly. You have breaking strengths, in kg, for 41 sample wires with sample mean breaking strength 91.85 kg and sample standard deviation 17.6 kg. Using the appropriate 95% confidence interval, conduct a hypothesis test to find out if the true mean breaking strength is above 85 kg.

$n=41$

>25

$\Downarrow$

CLT

Steps:

1- $H_0: \quad \mu = 85 \; vs. \quad H_1: \quad \mu > 85$

2- $\alpha = 0.05$

$100(1-\alpha)\% = 95\%$

$\Rightarrow \alpha = 0.05$

Significant level,

Hypothesis Testing

Null

Alternative

P-value

CI method

Example:[Breaking strength of wire, cont'd]

based on H_a

3- One-sided test and we care about the lower bound. So, we use $\left(\overline{X} - z_{1-\alpha}\frac{s}{\sqrt{n}}, +\infty\right)$. ✓

4- From the example in previous set of slides, the CI is $(87.3422, +\infty)$.

$H_0: 85$ 87.3422

5- Since $\mu_0 = 85$ is not in the CI, we **reject** H_0.

6- There is significant evidence to conclude that the true mean breaking strength of wire is greater than the 85kg. Hence the requirement is met.

$H_a: \mu > 85$

Example: [Concrete beams, cont'd]

10 concrete beams were each measured for flexural strength (MPa). The data is as follows.

$n = 10$

[1] 8.2 8.7 7.8 9.7 7.4 7.8 7.7 11.6 11.3 11.8 $\rightarrow \bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 9.2$

The sample mean was $\boxed{9.2}$ MPa and the sample variance was 3.0933 $(MPa)^2$. At $\alpha = 0.01$, test the hypothesis that the true mean flexural strength is 10 MPa using a confidence interval. Steps:

$s^2 =$

write down the hypothesis $\rightarrow$ 1- $H_0: \quad \mu = 10 \quad vs. \quad H_1: \quad \mu \neq 10$

2- $\alpha = 0.01$

$\rightarrow$ 3- This is two-sided test with $n = 10$ and 100 $(1-\alpha)$ % CI is

$n = 10$
σ unknown $\rightarrow$

$$(\bar{X} - t_{(n-1,1-\alpha/2)}\frac{s}{\sqrt{n}}, \bar{X} + t_{(n-1,1-\alpha/2)}\frac{s}{\sqrt{n}})$$

two-sided test $\equiv$ two-sided CI

one-sided test $\equiv$ one-sided CI

$t_{9,0.995} = 3.25$

Hypothesis Testing

Null

Alternative

P-value

CI method

Example:[Breaking strength of wire, cont'd]

4- Check that the CI is $(7.393, 11.007)$.

5- Since $\mu_0 = 10$ is within the CI, we **fail** to reject H_0.

6- There is not enough evidence to conclude that the true mean flexural strength is different from 10 Mpa.

Hypothesis Testing

Null

Alternative

P-value

CI method

Example:[Paint thickness, cont'd]

Consider the following sample of observations on coating thickness for low-viscosity paint.

[1] 0.83 0.88 0.88 1.04 1.09 1.12 1.29 1.31 1.48 1.49 1.59 1.62 1.65 1.71 ~~[15]~~ 1.76 1.83

$n=16$

Using $\boxed{\alpha = 0.1}$, test the hypothesis that the true mean paint thickness is 1.00 mm. Note, the 90\% confidence interval for the true mean paint thickness was calculated from before as $(1.201, 1.499)$.

1- $H_0: \quad \mu = 1 \quad vs. \quad H_1: \quad \mu \neq 1$

2- $\alpha = 0.1$

3- This is two-sided test with $n = 16$, σ unknown, so $100\,(1-\alpha)$ % CI is

$$(\overline{X} - t_{(n-1,1-\alpha/2)} \frac{s}{\sqrt{n}}, \overline{X} + t_{(n-1,1-\alpha/2)} \frac{s}{\sqrt{n}})$$

$n = 16 \, (< 25)$
$+$
σ unknown
$\Rightarrow$ use t dist.

$t_{15, 1-\frac{0.1}{2}} \equiv t_{15, .95}$

Hypothesis Testing

Null

Alternative

P-value

CI method

Example:[Breaking strength of wire, cont'd]

> 4- The CI is $(1.201, 1.499)$.
>
> 5- Since $\mu_0 = 1$ is not in the the CI, we **reject** H_0.
>
> H_0
>
> 6- There is enough evidence to conclude that the true mean paint thickness is not 1mm.
>
> $H_a: \mu \neq 1$

Section 6.4

Inference for matched pairs and two-sample data

Inference for matched pairs and two-sample data

An important type of application of confidence interval estimation and significance testing is when we either have *paired data* or *two-sample* data.

Recall: Matched pairs

Paired data is bivariate responses that consists of several determinations of basically the same characteristics

> **Example:**
>
> - Practice SAT scores *before* and *after* a preperation course
> - Severity of a disease *before* and *after* a treatment
> - Fuel economy of cars *before* and *after* testing new formulations of gasoline

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Inference for matched pairs and two-sample data

One simple method of investigating the possibility of a consistent difference between paired data is to

> 1. Reduce the measurements on each object to a single difference between them
>
> 2. Methods of confidence interval estimation and significance testing applied to differences (using Normal or t distributions when appropriate)

paired data

Example:[Fuel economy]

Twelve cars were equipped with radial tires and driven over a test course. Then **the same twelve cars** (with the same drivers) were equipped with regular belted tires and driven over the same course.

① After each run, the cars gas economy (in km/l) was measured. Using significance level $\alpha = 0.05$ and the method of critical values, test for a difference in fuel economy between the radial tires and belted tires.

② Construct a 95% confidence interval for true mean difference due to tire type. (i.e μ_d)

car	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0
① radial	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2
② belted	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.7	6.0	4.9

μ_d ↳ difference.

$\therefore d = \text{radial}_1 - \text{belted}_1$

Example:[Fuel economy]

	car	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0
d_1:	radial	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2
d_2:	belted	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.7	6.0	4.9
	d	0.1	-0.2	0.4	0.1	-0.1	0.1	0.0	0.2	0.5	0.2	0.1	0.3

Since we have paired data, the first thing to do is to find the differences of the paired data. ($d = d_1 - d_2$, where d_1 is associated with radial and d_2 is associated with belted tires.)

Then writing down the information available:

$$n = 12, \quad \bar{d} = 0.142, \quad s_d = 0.198$$

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

Then we just need to apply steps of hypothesis testing. Note that the null hypothesis here is that **there is no difference** between the gas economy recorded of the two different tires.(i.e $\mu_d = 0$)

$H_0: \mu_d = 0$

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Example:[Fuel economy]

1- $H_0: \quad \mu_d = 0 \; vs. \quad H_1: \quad \mu_d \neq 0$

2- $\alpha = 0.05$

μ_d

3- I will use the test statistics $K = \frac{\bar{d}-0}{s_d/\sqrt{n}}$ which has a $\boxed{t_{n-1}}$ distribution assuming that

n = 12
+
σ unknown
⇓
t-dist.

- H_0 is true and
- $d_1, d_2, \cdots, d_{12}$ are iid $N(\mu_d, \sigma_d^2)$

Example:[Breaking strength of wire, cont'd]

4- $K = \frac{0.421}{0.198/\sqrt{12}} = \boxed{2.48} \sim t_{(11,0.975)}$.

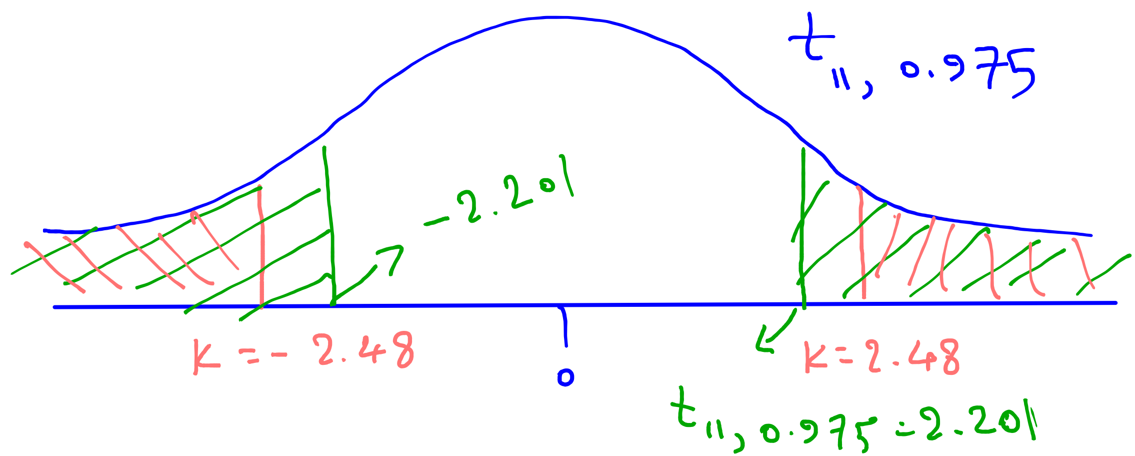
$$
\begin{aligned}
p - value = \boxed{P(|T| > K)} &= P(|T| > 2.48) \\
&= P(T > 2.48) + P(T < -2.48) \\
&= 1 - P(T < 2.48) + P(T < -2.48)
\end{aligned}
$$

software

(by ~~the t table~~) $= 1 - 0.9847 + 0.9694 = \boxed{0.03}$

5- Since p-value < 0.05, we **reject** H_0.

6- There is enough evidence to conclude that fuel economy differs between radial and belted tires.



$$\underbrace{P(|T| > 2.48)}_{\text{p-value}} < \underbrace{P(|T| > t_{11,0.975})}_{\alpha}$$

Hang on for a Second

Let's review slide 58 again

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Example:[Breaking strength of wire, cont'd]

$$
\begin{aligned}
p - value &= P(|T| > K) = P(|T| > 2.48) \\
&= P(T > 2.48) + P(T < -2.48) \\
&= 1 - P(T < 2.48) + P(T < -2.48) \\
(\text{by software}) &= 1 - 0.9847 + 0.9694 = 0.03
\end{aligned}
$$

We have seen t-student table

How do we get that p-value usin software!!!

What is happening?

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Unlike *standard Normal distribution* table which gives us probability under the standard Normal curve, t tables are quantile tables.

i.e We use the t table (Table B.4 in Vardeman and Jobe) to calculate quantiles.

To have exact probabilities, we need software.

Table B.4
t Distribution Quantiles

ν	$Q(.9)$	$Q(.95)$	$Q(.975)$	$Q(.99)$	$Q(.995)$	$Q(.999)$	$Q(.9995)$
1	3.078	6.314	12.706	31.821	63.657	318.317	636.607
2	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869

e.g. we use these quantiles to make confidence intervals.

$$\bar{x} \pm t_{(n-1, 1-\alpha/2)} \cdot s/\sqrt{n}$$

a quantile, not probability.

The approach in calculating p-value when t distribution is involved

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Two important points:

> P-value and α are both probabilities. (so $\in [0, 1]$).
>
> They are areas under the curve in tails under *null hypothesis*.

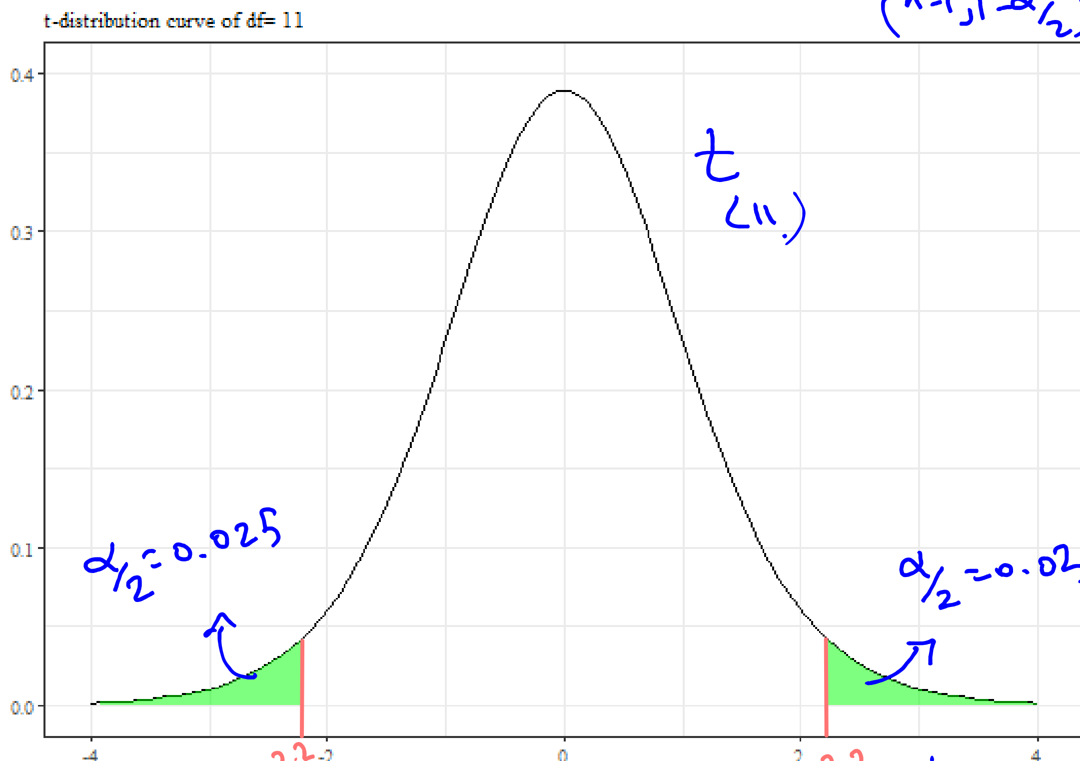
H_0

From example [breaking strength]

For a random variable with $\sim t_{(11,0.975)}$:

By the t table, the t quantile of $t_{(11,0.975)}$ is 2.2.

$t_{(n-1, 1-\alpha/2)}$



$-2.2 \rightarrow -t_{(11,0.975)}$

$2.2 = t_{(11,0.975)}$

Total shaded area is $\alpha/2 + \alpha/2 = \alpha = 0.05$ (The significance level)

For the critical value we calculated under the null hypothesis:

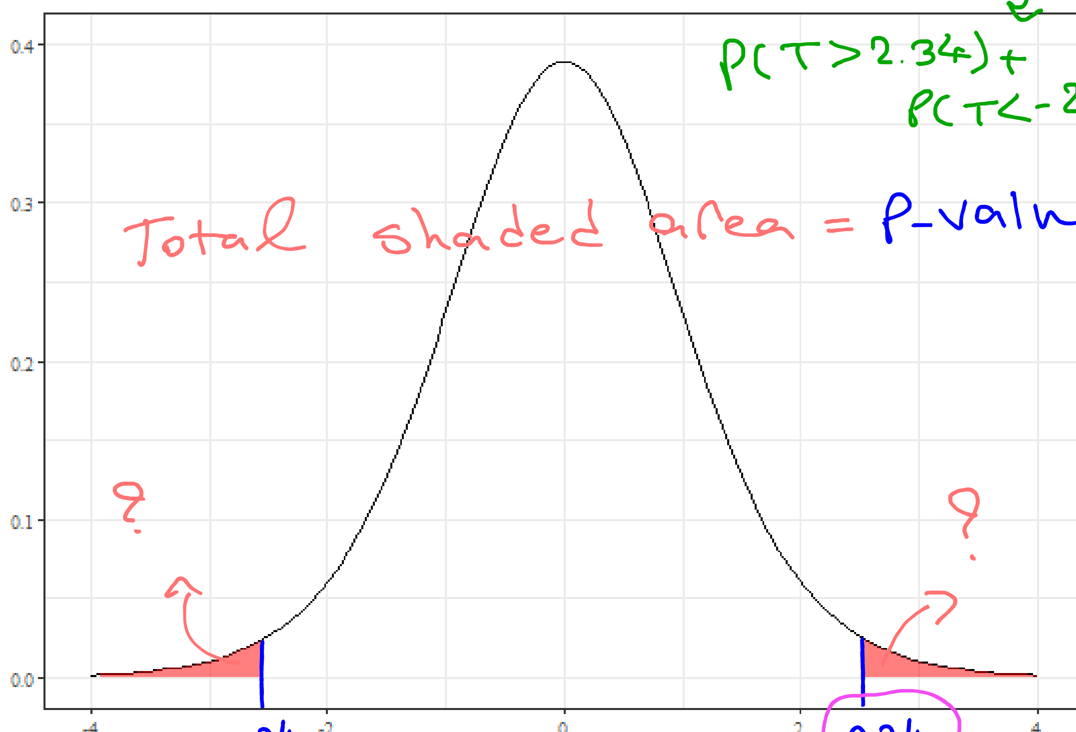
$T \sim t_{(n-1)}$

The critical value calculated is $K = 2.34$

Recall: P-value $= P(|T| > K) = P(|T| > 2.34)$

$= P(T > 2.34) + P(T < -2.34)$

t-distribution curve of df= 11 and K=2.34



Total shaded area = P-value

?

?

$-2.34 = K$

2.34

K

Here, we need software to find the total shaded area!

Hypothesis Testing

Null

Alternative

P-value

CI method

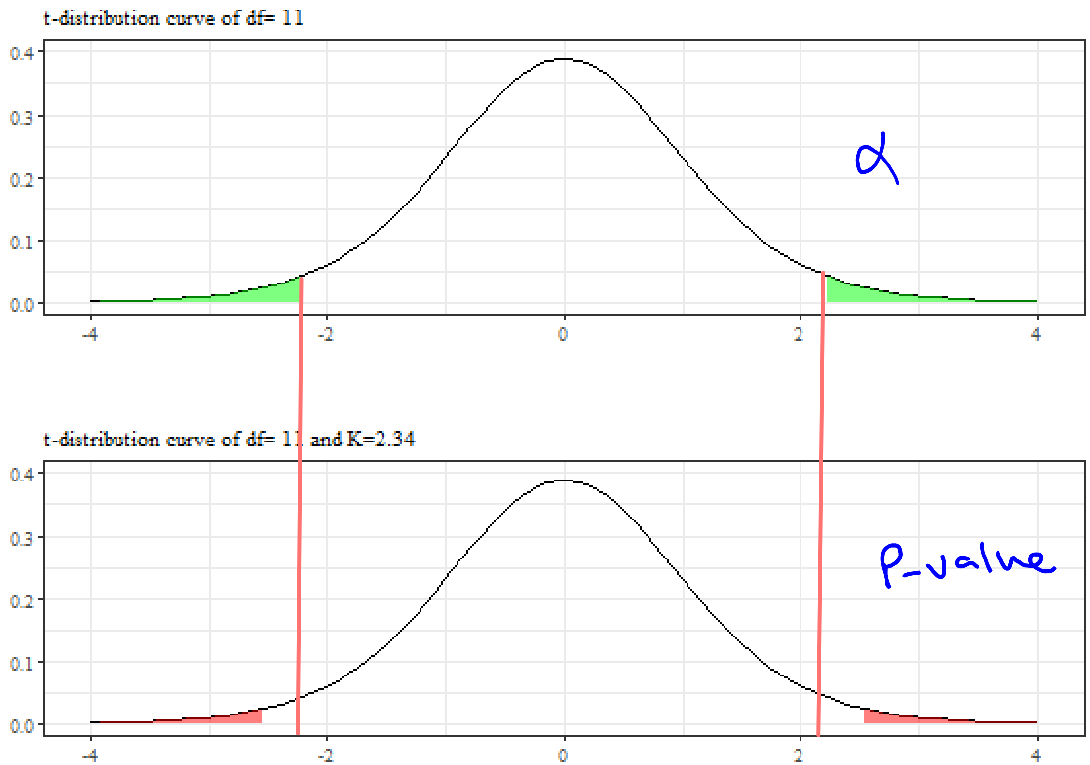
Matched Pairs

Two-sample

Both together

p-value

$K \gtrless t_{(n-1,\, 1-\alpha/2)}$ α



We reject the null if p-value < α.

> Remember p-value and α are areas under the curve

p-value < α

* Total green area : $\alpha = 0.05$

* Total red area : $P(|T| > k)$ = p-value = ?

(we don't need to find ? just need

→ to know if ? > α or ? < α)

① If the red area (p-value) < α (= 0.05 in this problem)

⇒ Reject H_0.

② If the red area (p-value) > α (= 0.05 in this problem)

⇒ Fail to reject H_0

The steps for p-value :

calculate p-value using table on slide 39.

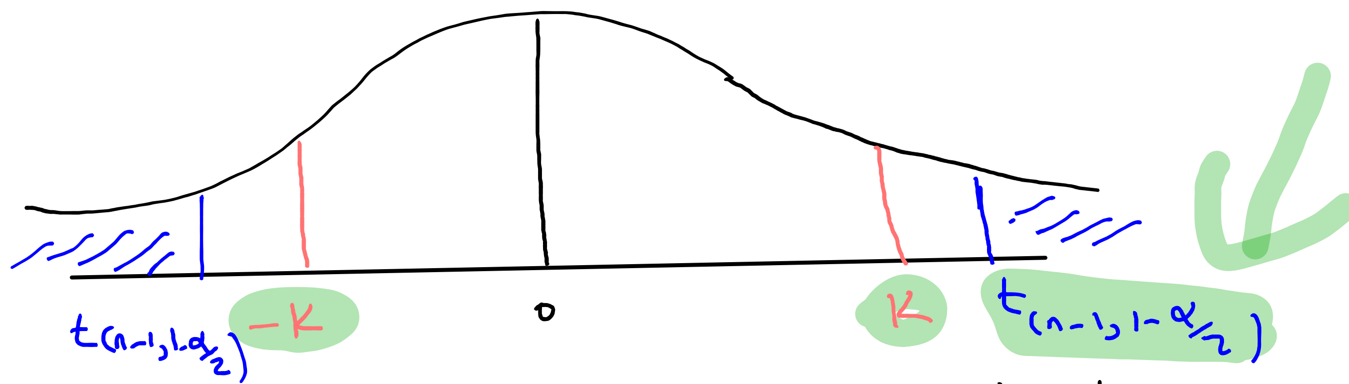
1 - If you use Z => use Normal table to find p-value

=> Then compare p-value & α (given in the problem) to reject or fail to reject H_0

2 - If you use T statistic, Find the quantiles

$\rightarrow t_{(n-1,\,1-\alpha/2)}$ (or $t_{(n-1,\,1-\alpha)}$ depending on the problem).

(I suggest quickly plot $t_{(n-1, 1-\alpha/2)}$ to understand better)



• write values of $\underline{\underline{K}}$ (critical value) and the corresponding $t_{(n-1, 1-\alpha/2)}$

$H_0: \mu = \#$

$H_a: \mu \neq \#$

$$K = \frac{\bar{x} - \#}{s/\sqrt{n}}$$

• Note: area under the curve corresponding to $t_{(n-1, 1-\alpha/2)}$ is equal to α. (e.g 0.05)

· Note: area under the curve corresponding to k (critical value) is p-value (which we don't need the exact value)

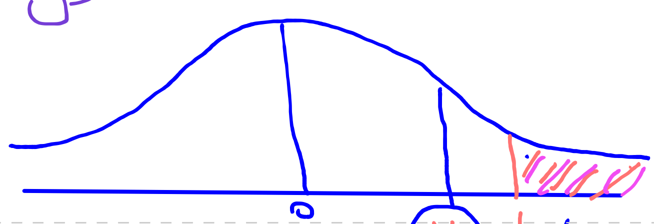
- Now compare the areas under the curve.

if p-value $< \alpha$ $\Rightarrow$ reject H_0

if p-value $> \alpha$ $\Rightarrow$ fail to reject H_0

Note Note Note:

* α is NOT always 0.05 (be careful)
* The test is NOT always two sided!



P-value (R) $t_{(n-1,\,1-\alpha)}$

- Hypothesis Testing
- Null
- Alternative
- P-value
- CI method
- Matched Pairs
- **Two-sample**

Example:[End-cut router]

Consider the operation of an end-cut router in the manufacture of a company's wood product. Both a leading-edge and a trailing-edge measurement were made on each wooden piece to come off the router.

Paired data

Is the leading-edge measurement different from the trailing-edge measurement for a typical wood piece?

(Let's see if there's any difference between the measurement

Do a hypothesis test at $\alpha = 0.05$ to find out. Make a two-sided 95% confidence interval for the true mean of the difference between the measurements.

piece	1.000	2.000	3.000	4.000	5.000
leading_edge	0.168	0.170	0.165	0.165	0.170
trailing_edge	0.169	0.168	0.168	0.168	0.169
Difference: d_i	-.001	.002	-.003	-.003	.001

(-)

$n = 5$ $\quad \bar{d} = \frac{1}{5}(-.001 + .002 - .003 - 0.003 + .001)$, $\quad S_d = \frac{1}{n-1}\sum_{i=1}^{5}(d_i - \bar{d})^2$

$= -8\times10^{-4}$ $\quad\quad = 0.0023$

steps:

(1) $H_0: \mu_d = 0$ (There's No difference between measurements)

$H_a: \mu_d \neq 0$ (There's difference between measurements)

(2) $\alpha = 0.05$

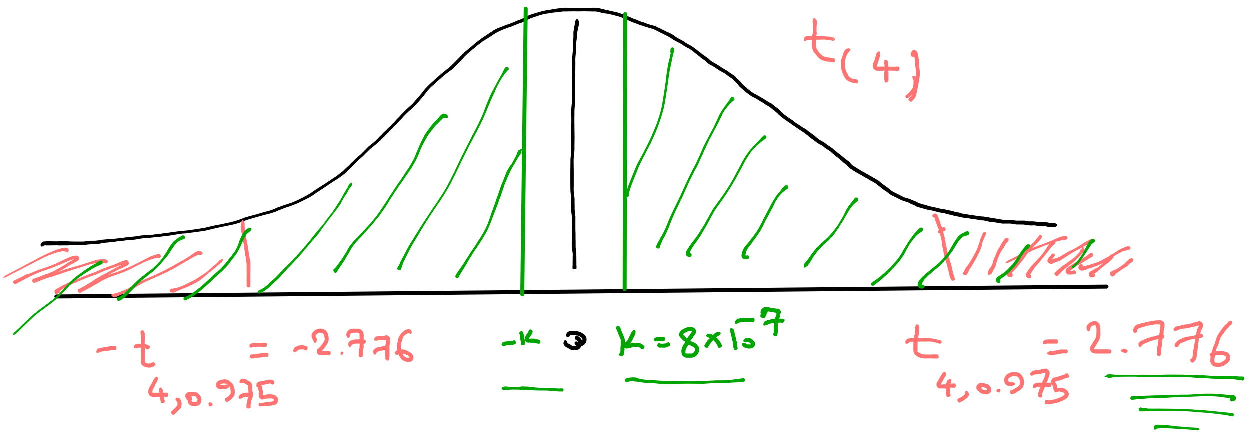
(3) since σ_d is unknown & $n = 5 < 25$, I will use

(4) $$K = \frac{\mu_d - 0}{\frac{S_d}{\sqrt{n}}} = \frac{-8\times10^{-4} - 0}{\frac{0.0023}{\sqrt{5}}} = \boxed{-8\times10^{-7}}$$

p-value $= P(|T| > |K|) = P(|T| > |-8\times10^{-7}|)$

$= P(|T| > 8\times10^{-7})$

by The table : $t_{(n-1, 1-\alpha/2)} = t_{(5-1, 0.975)} = 2.776$



(obviously area under the curve corresponding to p-value > area under the curve corresponding to $t_{4,0.975}$ $(= \alpha)$)

(5) since p-value $> \alpha$, we fail to reject H_0.

(6) there is not enough evidence to conclude that

There's significant difference between the leading edge and trailing edge on average

---

Using 95% CI method:

$H_0: \mu_d = 0$

$H_a: \mu_d \neq 0$

$$\rightarrow \bar{d} \pm \boxed{t_{(n-1,\, 1-\alpha/2)}} \, S_d/\sqrt{n}$$

$$= -8\times10^{-4} \pm t_{(4,\, 0.975)} \cdot \frac{0.0023}{\sqrt{5}}$$

$$= -8\times10^{-4} \pm 2.776\,(0.001)$$

$$\rightarrow = (-0.00358,\ 0.00198)$$

Since the 95% CI contains zero, we fail to reject H_0. There's not enough evidence to conclude that the leading edge measurement is significantly different from the trailing edge measurement.

Two-Sample Data

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Two-sample data

Paired differences provide inference methods of a special kind for comparison. Methods that can be used to compare two means where two different *unrelated* samples will be discussed next.

> SAT score of high school A vs. high school B
>
> Severity of a disease in men vs. women
>
> Height of Liverpool soccerr players vs. Man United soccer players
>
> Fuel economy of gas formula type A vs. formula type B

Two-sample data

Notations:

	Sample 1	Sample 2
Sample size	n_1	n_2
* true means	$\boxed{\mu_1}$	$\boxed{\mu_2}$
Sample means	$\bar{x}_1$	$\bar{x}_2$
* true variance	σ_1^2	σ_2^2
Sample variance	s_1^2	s_2^2

$$H_0: \mu_1 = \mu_2 \equiv \mu_1 - \mu_2 = 0$$

$H_a: \mu_1 - \mu_2 \neq 0$ ✓

Large Samples

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

(want to compare two true means μ_1, μ_2)

Large samples $(n_1 \geq 25, n_2 \geq 25)$

$H_0: \mu_1 = \mu_2 \Rightarrow \boxed{\mu_1 - \mu_2 = 0}$ vs. $H_a: \boxed{\mu_1 - \mu_2 \neq 0}$

The difference in sample means $\overline{x}_1 - \overline{x}_2$ is a natural statistic to use in comparing μ_1 and μ_2.

$\mu_1 - \mu_2 < 10$

or $\mu_1 - \mu_2 > 15$

i.e

$$\mathrm{E}(\overline{X}_1) = \mu_1 \quad \mathrm{E}(\overline{X}_2) = \mu_2 \quad \mathrm{Var}(\overline{X}_1) = \frac{\sigma_1^2}{n_1} \quad \mathrm{Var}(\overline{X}_2) = \frac{\sigma_2^2}{n_2}$$

If σ_1 and σ_2 are **known**, then we have

$$\mathrm{E}(\overline{X}_1 - \overline{X}_2) = \mathrm{E}(\overline{X}_1) - \mathrm{E}(\overline{X}_2) = \mu_1 - \mu_2$$

$$\mathrm{Var}(\overline{X}_1 - \overline{X}_2) = \mathrm{Var}(\overline{X}_1) + \mathrm{Var}(\overline{X}_2) = \boxed{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$= \mathrm{Var}(\overline{X}_1) + (-1)^2 \mathrm{Var}(\overline{X}_2)$

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Large samples $(n_1 \geq 25, n_2 \geq 25)$

If, in addition, n_1 and n_2 are large,

$\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ is independent of $\overline{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$ (by CLT).

So that $\overline{X}_1 - \overline{X}_2$ is approximately Normal (trust me)

test statistic

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \overbrace{(\mu_1 - \mu_2)}^{\#}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Previously: $Z = \frac{\bar{X} - \mu^{\#}}{\sqrt{\sigma^2/n}} \sim N(0,1)$

(one sample)

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Large samples $(n_1 \geq 25, n_2 \geq 25)$

So, if we want to test $\boxed{H_0 : \mu_1 - \mu_2 = \#}$ with some alternative hypothesis, σ_1 and σ_2 are known, and $n_1 \geq 25, n_2 \geq 25$, then we use the statistic

$$\boxed{K = \frac{\overline{X}_1 - \overline{X}_2 - (\#)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}} \leftarrow *$$

which has a $N(0, 1)$ distribution if

1. H_0 is true
2. The sample 1 points are iid with mean μ_1 and variance σ_1^2, and the sample 2 points are iid with mean μ_2 and variance σ_2^2.

3.Sample I is independent of sample II

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Large samples $(n_1 \geq 25, n_2 \geq 25)$

The confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$ are:

- *Two-sided* $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$

$$(\overline{x}_1 - \overline{x}_2) \pm z_{1-\alpha/2} * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- *One-sided* $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ with a upper confidence bound

$$(-\infty\ ,\ (\overline{x}_1 - \overline{x}_2) \pm z_{1-\alpha} * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

- *One-sided* $100(1-\alpha)\%$ confidence interval for μ with a lower confidence bound

$$((\overline{x}_1 - \overline{x}_2) \pm z_{1-\alpha} * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})\ ,\ \ +\infty)$$

Large samples $(n_1 \geq 25, n_2 \geq 25)$

If σ_1 and σ_2 are **unknown**, and $n_1 \geq 25, n_2 \geq 25$, then we use the statistic

replace σ_1^2 by s_1^2
and σ_2^2 by s_2^2

$$K = \frac{\overline{X}_1 - \overline{X}_2 - (\#)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

and confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$:

- *Two-sided* $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$

$$(\overline{x}_1 - \overline{x}_2) \pm z_{1-\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Large samples $(n_1 \geq 25, n_2 \geq 25)$

- *One-sided* $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ with a upper confidence bound

$$(-\infty\ ,\ (\overline{x}_1 - \overline{x}_2) \pm z_{1-\alpha} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$$

- *One-sided* $100(1-\alpha)\%$ confidence interval for μ with a lower confidence bound

$$((\overline{x}_1 - \overline{x}_2) \pm z_{1-\alpha} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})\ ,\ +\infty)$$

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Example:[Anchor bolts]

An experiment carried out to study various characteristics of anchor bolts resulted in 78 observations on shear strength (kip) of 3/8-in. diameter bolts and 88 observations on strength of 1/2-in. diameter bolts.

Let Sample 1 be the 1/2 in diameter bolts and Sample 2 be the 3/8 indiameter bolts.

Using a significance level of $\alpha = 0.01$, find out if the 1/2 in bolts are more than 2 kip stronger (in shear strength) than the 3/8 in bolts. Calculate and interpret the appropriate 99% confidence interval to support the analysis.

given info.

- $n_1 = 88, n_2 = 78$
- $\overline{x}_1 = 7.14, \overline{x}_2 = 4.25$
- $s_1 = 1.68, s_2 = 1.3$

① $H_0: \mu_1 - \mu_2 = 2$ vs. $H_a: \mu_1 - \mu_2 > 2$

② $\alpha = 0.01$

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Example:[Anchor bolts]

- $n_1 = 88, n_2 = 78$
- $\overline{x}_1 = 7.14, \overline{x}_2 = 4.25$
- $s_1 = 1.68, s_2 = 1.3$

3 Since $n_1, n_2 > 25$, I'll use

$$\boxed{K = \frac{(\bar{x}_1 - \bar{x}_2) - 2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}}$$

If we assume H_0 is true, sample I is iid with mean μ_1, variance σ_1^2 independent of sample II iid with mean μ_2 and variance σ_2^2, $K \sim N(0,1)$

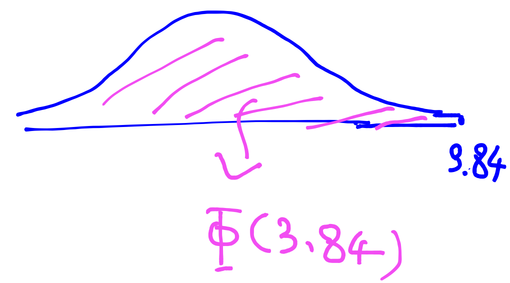
④ $K = \dfrac{(\bar{x}_1-\bar{x}_2)-2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \dfrac{(7.14-4.25)-2}{\sqrt{\frac{1.68^2}{88}+\frac{1.3^2}{78}}} = \boxed{3.84}$

p-value $= P(Z > K) = P(Z > 3.84)$

$H_a: \mu_1 - \mu_2 > 2$

$= 1 - \Phi(3.84)$ ≈ 1

$\approx 1 - 1 \approx 0$



⑤ with a p-value $\approx 0 < \alpha = 0.01$, we reject H_0 in favor of H_a.

⑥ There's enough evidence that the $\frac{1}{2}$ in bolts are more than 2 kip stronger than $\frac{3}{8}$ in bolts on average

99% lower bound CI (since $H_a: \mu_1 - \mu_2 > 2$)

$$\left(\bar{X}_1 - \bar{X}_2 - Z_{1-\alpha}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, +\infty\right)$$

$$= \left((7.14 - 4.25) - Z_{1-0.01}\sqrt{\frac{1.68^2}{88} + \frac{1.3^2}{78}}, +\infty\right)$$

$$Z_{.99}$$

$$= \left(2.89 - \boxed{2.33}(0.232), +\infty\right)$$

$$= (2.35, +\infty)$$

We're 99% confident that the true mean strength of the $\frac{1}{2}$ in bolts is at least 2.35 kip stronger than the true mean strength of the $\frac{3}{8}$ in bolts.

Small Samples

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Small samples

If $n_1 < 25$ or $n_2 < 25$, then we need some **other assumptions** to hold in order to complete inference on two-sample data.

> We need two independent samples to be iid Normally distributed and $\sigma_1^2 \approx \sigma_2^2$

A test statistic to test $\mathrm{H}_0 : \mu_1 - \mu_2 = \#$ against some alternative is

$$* \quad K = \frac{\boxed{\overline{X}_1 - \overline{X}_2 - (\#)}}{S_p\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where S_p^2 is called **pooled sample variance** and is defined as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Small samples

Also assuming

- 1 H$_0$ is true,
- 2 The sample 1 points are iid $N(\mu_1, \sigma_1^2)$, the 3 sample 2 points are iid $N(\mu_2, \sigma_2^2)$,
- 4 and the sample 1 points are independent of the sample 2 points and $\boxed{\sigma_1^2 \approx \sigma_2^2.}$ 5

Then

5 assumptions to use

$$K = \frac{\overline{X}_1 - \overline{X}_2 - (\#)}{S_p\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \sim \boxed{t_{(n_1+n_2-2)}}$$

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Small samples

$1-\alpha$ confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1-\mu_2$ under these assumptions are of the form:

(let $\boxed{\nu = n_1+n_2-2}$)

- *Two-sided* $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$

$$(\bar{x}_1-\bar{x}_2) \pm \boxed{t_{(\nu,1-\alpha/2)}} * \boxed{S_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$$

- *One-sided* $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ with a upper confidence bound

$$(-\infty\ ,\ (\bar{x}_1-\bar{x}_2)+t_{(\nu,1-\alpha)} * S_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}$$

n_1+n_2-2

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Small samples

- *One-sided* $100(1-\alpha)\%$ confidence interval for μ with a lower confidence bound

$$((\overline{x}_1 - \overline{x}_2) - t_{(\nu, 1-\alpha)} * S_p \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}\ ,\ +\infty)$$

$\nu = n_1 + n_2 - 2$

In general: CI:

estimate $\pm$ Population quantiles $\times$ SD(estimate)

$\mu = 0$

$\mu \neq 0$: $\bar{x} \pm Z_{1-\alpha/2}\, S/\sqrt{n}$

large sample

Hypothesis Testing

Null

Alternative

P-value

CI method

Matched Pairs

Two-sample

Small samples

Example:[Springs]

The data of W. Armstrong on spring lifetimes (appearing in the book by Cox and Oakes) not only concern spring longevity at a 950 N/ mm^2 stress level but also longevity at a 900 N/ mm^2 stress level.

Let sample 1 be the 900 N/ mm^2 stress group and sample 2 be the 950 N/ mm^2 stress group.

900 N/mm2 Stress	950 N/mm2 Stress
216, 162, 153, 216, 225, 216, 306, 225, 243, 189	225, 171, 198, 189, 189, 135, 162, 135, 117, 162

$n_1 = 10$ $n_2 = 10$

Hypothesis Testing

Null

Alternative

P-value

CI method

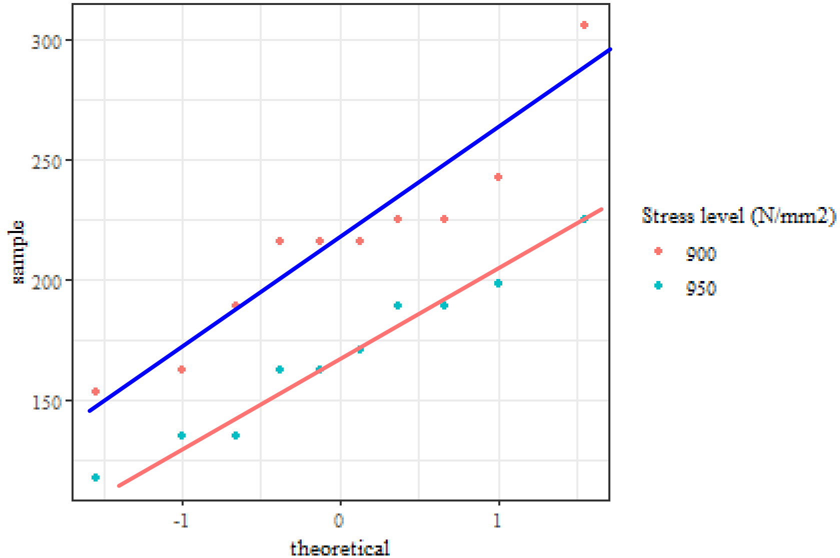
Matched Pairs

Two-sample

Small samples

Normal plots of spring lifetime under two different stress level.

Example:[Springs]



Let's do a hypothesis test to see if the sample 1 springs lasted significantly longer than the sample 2 springs. (For $\alpha = 0.05$)

Note: however the sample sizes are small, the data look pretty Normal.

1, $\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \boxed{\mu_1 - \mu_2 > 0} \end{cases}$

2, $\alpha = 0.05$

3, If we assume that ① H_0 is true & ② sample 1 is iid $N(\mu_1, \sigma_1^2)$ ③ independent from sample 2 ④ iid $N(\mu_2, \sigma_2^2)$ and ⑤ $\sigma_1^2 \approx \sigma_2^2$, then the test statistic is

$$K = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \;\dot\sim\; t_{(n_1 + n_2 - 2)}$$

4, calculates $\left\{ \begin{array}{l} \bar{x}_1 = 215.1, \quad S_1 = 42.9, \quad S_1^2 = 1840.41 \\ \bar{x}_2 = 168.3, \quad S_2 = 33.1, \quad S_2^2 = 1095.61 \end{array} \right.$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{(10-1)1840.41 + (10-1)1095.61}{(10+10-2)}}$$

$$\Rightarrow S_p = 38.3$$

Then $K = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\boxed{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}} = \dfrac{(215.1 - 163.3)}{38.3\sqrt{\frac{1}{10}+\frac{1}{10}}} = 2.7$

Recall:

We need to calculate p-value now.

* Corresponding to α: $t_{(\underline{n_1+n_2-2},\,1-\alpha)} = t_{\underline{(18,\,0.95)}}$

by table $= \underline{1.73}$

(By the methods we learned about p-value and α, we just need to decide if p-value $\gtrless \alpha$)

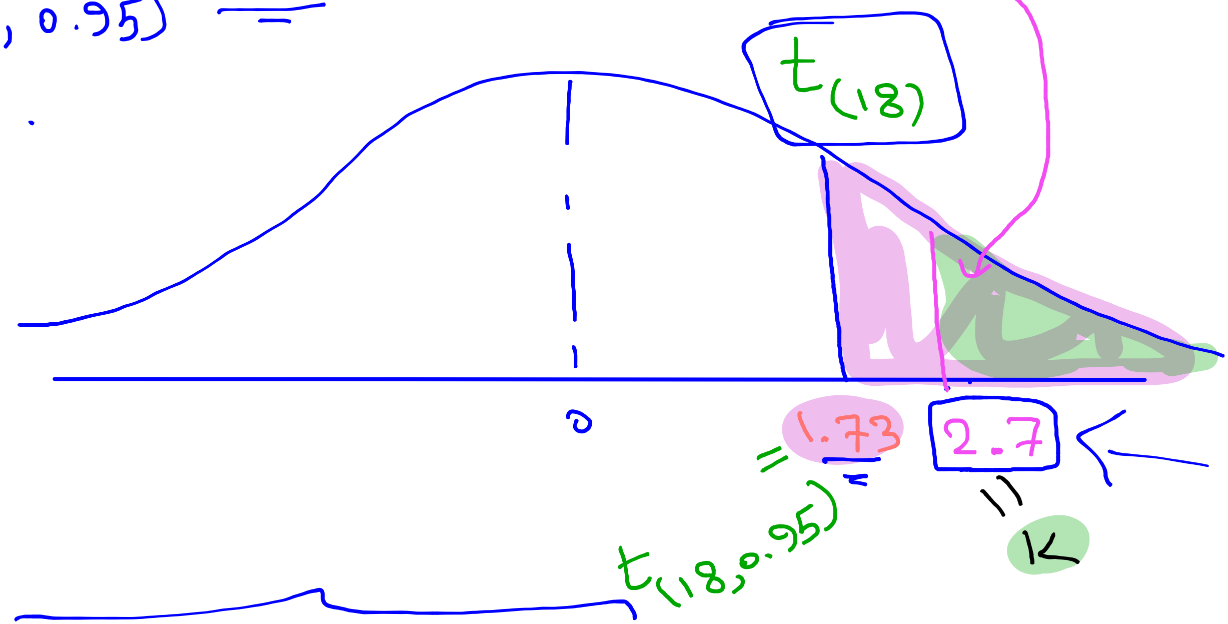
→ * Corresponding to p-value $\equiv k = \underline{2.7}$

Now:

one sided test

$$\text{p-value} = P(T > K) = P(T > 2.7)$$

$$t_{(18,\,0.95)} = 1.73$$



5/ Since $K > 1.73 \; (= t_{(18,0.95)})$

(area under the curve for $P(T>K)$ is smaller than α)

$\Rightarrow$ p-value $< \alpha$ $\Rightarrow$ Reject H_0

6, There is enough evidence to conclude that springs on average last longer subjected to $\underline{900}$ N/mm^2 of stress than $\underline{\underline{950}}$ N/mm^2 of stress.

Small samples

Example:[Stopping distance]

Suppose μ_1 and μ_2 are true mean stopping distances (in meters) at 50 mph for cars of a certain type equipped with two different types of breaking systems.

Suppose $n_1 = n_2 = 6, \overline{x}_1 = 115.7, \overline{x}_2 = 129.3, s_1 = 5.08$, and $s_2 = 5.38$.

Use significance level $\boxed{\alpha = 0.01}$ to test
$\mathrm{H}_0 : \mu_1 - \mu_2 = -10$ vs. $\mathrm{H}_A : \mu_1 - \mu_2 < -10.$

Construct a 2-sided 99% confidence interval for the true difference in stopping distances.

① $\begin{cases} H_0: \mu_1 - \mu_2 = -10 \\ H_a: \mu_1 - \mu_2 < -10 \end{cases}$ ⟵

② $\alpha = 0.01$

③ Under the assumptions that ① H_0 is true and

② sample 1 is iid $N(\mu_1, \sigma_1^2)$ ③ independent of

④ sample 2 iid $N(\mu_2, \sigma_2^2)$ and ⑤ $\sigma_1^2 \approx \sigma_2^2$ ✓

we use test statistic

$$K = \frac{\bar{x}_1 - \bar{x}_2 - (-10)}{S_p \sqrt{1/n_1 + 1/n_2}}$$

where $K^* \sim t_{(n_1+n_2-2)}$

(4) $S_p = \sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\dfrac{(6-1)5.08^2+(6-1)5.38^2}{6+6-2}}$

$\rightarrow S_p = 5.23$

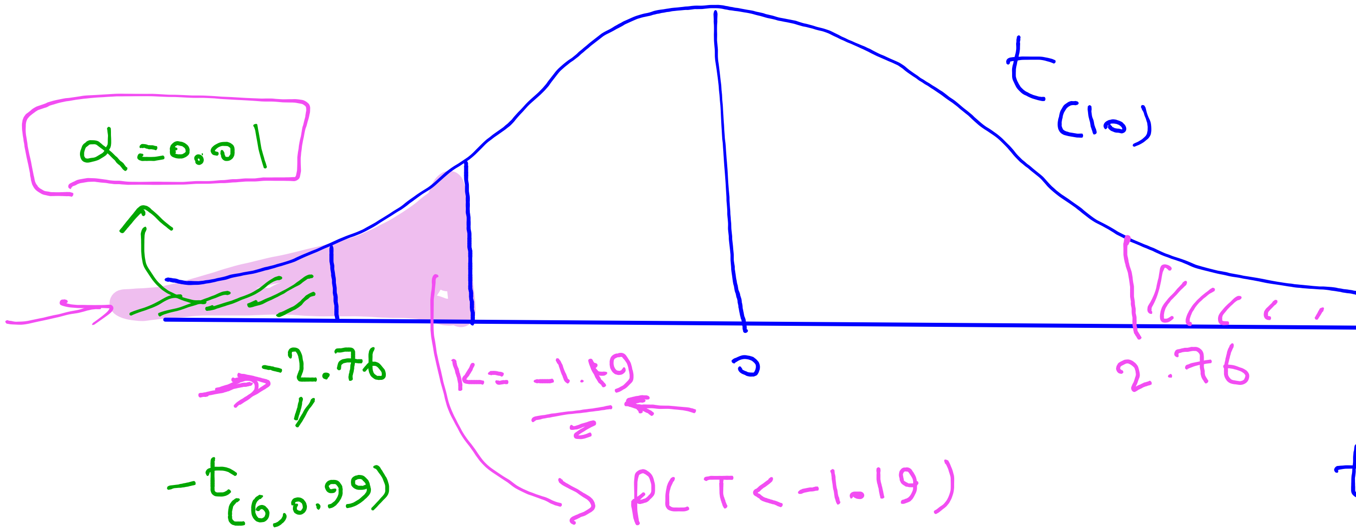
$\rightarrow K = \dfrac{(\bar{x}_1-\bar{x}_2)-(-10)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \dfrac{(115.7-129.3)+10}{5.23\sqrt{\frac{1}{6}+\frac{1}{6}}} = -1.19$

$\begin{cases} H_0: \mu_1-\mu_2 = -10 \\ H_a: \mu_1-\mu_2 < -10 \end{cases}$

$T \sim t_{(n_1+n_2-2)}$

$\rightarrow$ p-value $= P(T < K) = P(T < -1.19)$

by the table: $t_{(6+6-2,\ 1-\alpha)} = t_{(10,\ 0.99)} = \boxed{2.76}$



$$\underbrace{P(T<K) = P(T<-1.19)}_{\text{p-value}} > \underbrace{P(T<-2.76)}_{\alpha}$$

Since the p-value is greater than α, where $P(T<-2.76)$ is marked as $-t_{(6,\,0.99)}$:

we fail to reject H_0.

(5) Since p-value > α, we fail to reject H_0

(6) There is Not enough evidence to conclude that stopping distances for breaking system 1 are on average less than those of breaking system 2 by over 10 m.

Construct $\boxed{\text{2-sided}}$ 99% CI for $\mu_1 - \mu_2$ (the true difference of mean stopping distance):

using formula:

$$\rightarrow \left((\bar{x}_1 - \bar{x}_2) - t_{(n_1+n_2-2,\, 1-\alpha/2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \, , \, (\bar{x}_1 - \bar{x}_2) + t_{(n_1+n_2-2,\, 1-\alpha/2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

$$\underbrace{\quad}_{10} \qquad \underbrace{\quad}_{1-\frac{0.01}{2} = 0.995}$$

$$= \left((115.7 - 129.3) - 3.17 \cdot (5.23)\sqrt{\frac{1}{6}+\frac{1}{6}} \, , (115.7 - 129.3) + 3.17 (5.23)\sqrt{\frac{1}{6}+\frac{1}{6}} \right)$$

$$= \boxed{(-23.17, -4.03)}$$

we are 99% confident that the true mean stopping distance of system 1 is anywhere between 23.17 m to 4.03 m less than that of system 2.