02323 Introduction to Statistics

Week 5: Hypothesis Testing

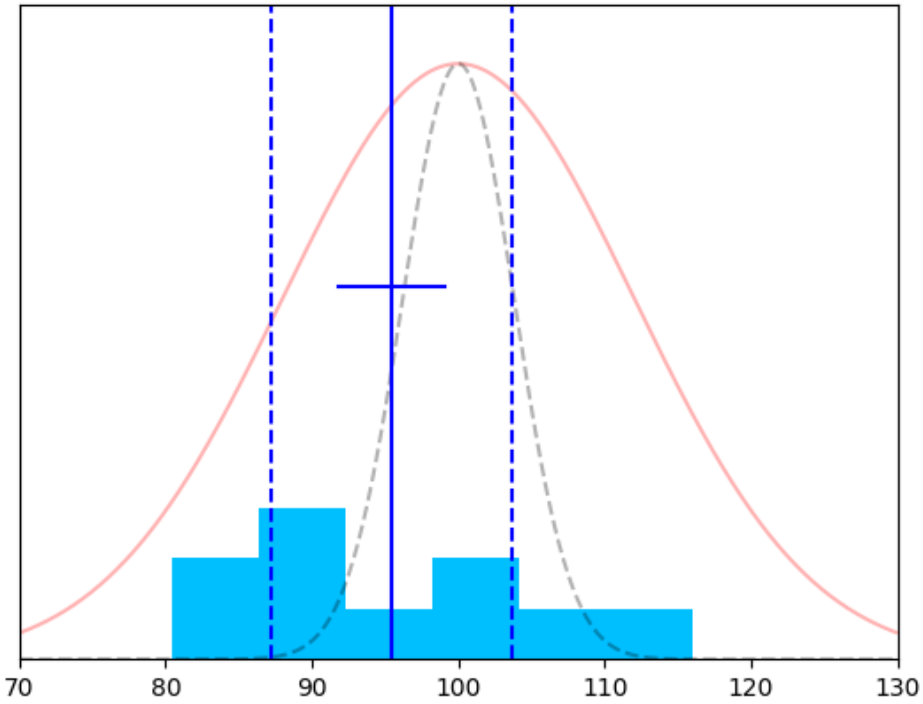
DTU Compute
Technical University of Denmark
2800 Kgs. Lyngby

Agenda

1. Summary from last week
2. Model Control
   - Q-Q plot
   - Transformation towards Normality
3. Hypothesis Testing
   - Null Hypothesis
   - t-test
   - p-value
   - One-sided vs Two-sided Tests
4. Type I and Type II Errors

Overview

From last week: Statistical inference and confidence intervals



The underlying distribution (red line) is unknown, but we can describe it with a theoretical model.

The sample mean $\bar{X}$ follows a normal distribution, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ (black dashed line), *if* the underlying (red) distribution is itself normal OR if n is large enough (Central Limit Theorem) (in that case, the red distribution can take any shape – the black will still be normally distributed).

From the sample data, we can estimate $\hat{\mu} = \bar{x}$ and calculate a corresponding confidence interval (blue dashed lines).

The confidence interval is a range within which we believe the "true" value of μ lies.

The width of the confidence interval depends on our choice of significance level (α).

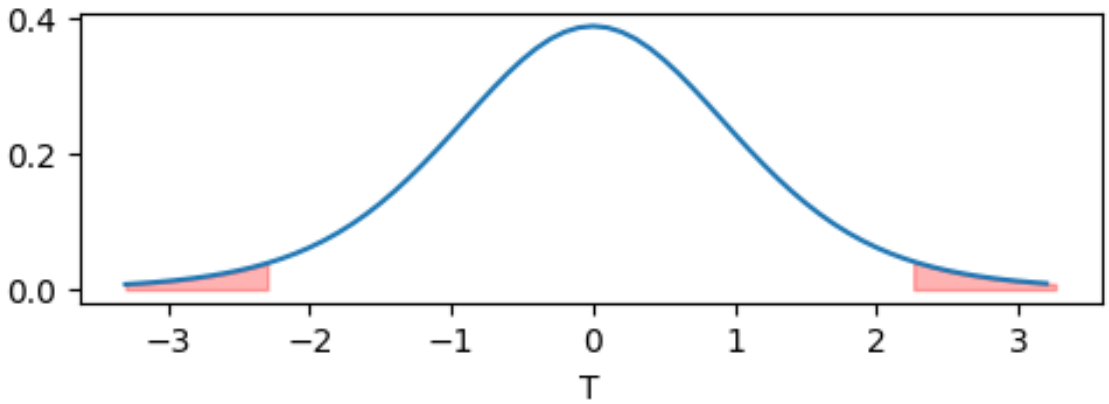
From last week

Confidence interval and significance level (1):

To calculate the confidence interval, we considered the statistic T:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

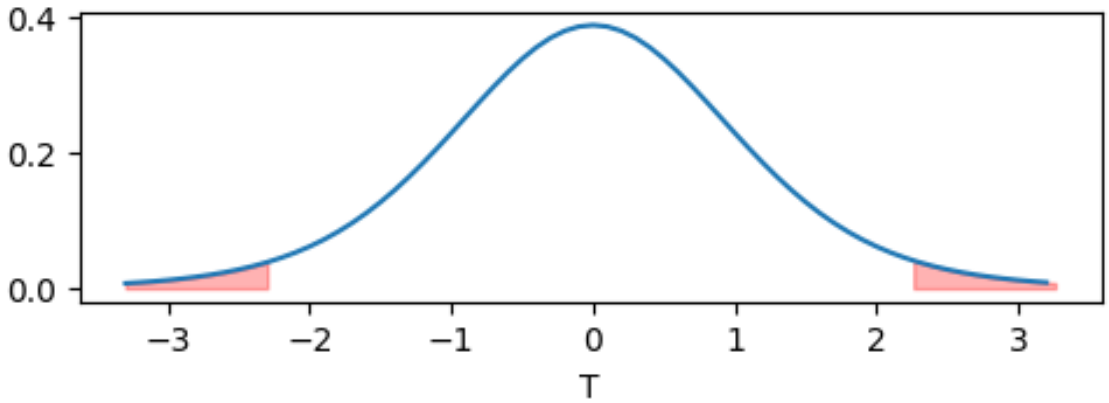
This statistic follows a t-distribution: $T \sim t(n-1)$.



It is most likely that T takes values close to zero ($\bar{x}$ close to μ).
"Extreme" values (large $|T|$) are less likely.

From last week

Confidence interval and significance level (2):



The values $\pm t_{1-\alpha/2}$ are the values of T that exclude the most **extreme values** of T and thus the values where $\bar{X}$ is farthest from μ.

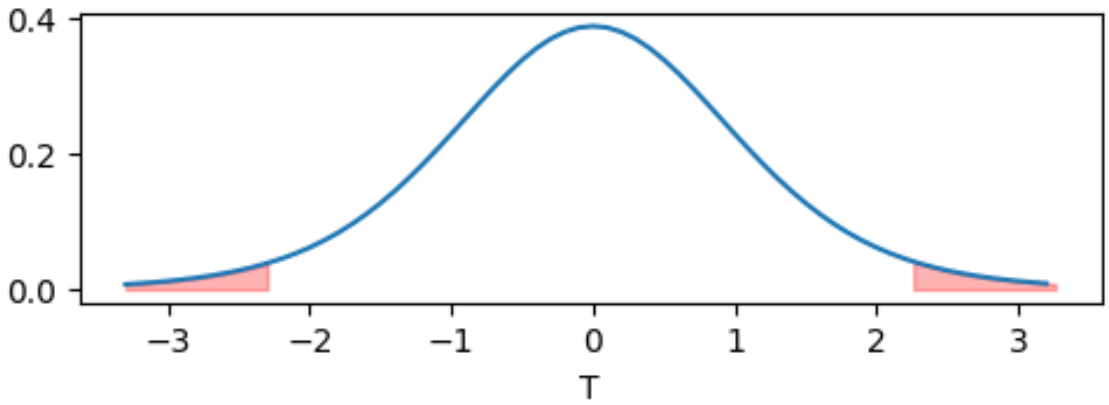
If $\alpha = 0.05$, then 95% of the *least extreme* values of T fall within the boundaries $\pm t_{1-\alpha/2}$, and 5% of the *most extreme* values fall outside $\pm t_{1-\alpha/2}$ (the red region in the plot).

Kahoot!

(x2)

From last week

Confidence interval and significance level (3):



To calculate the confidence interval, we choose a **significance level** (α) and then calculate:

$$\left[\overline{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

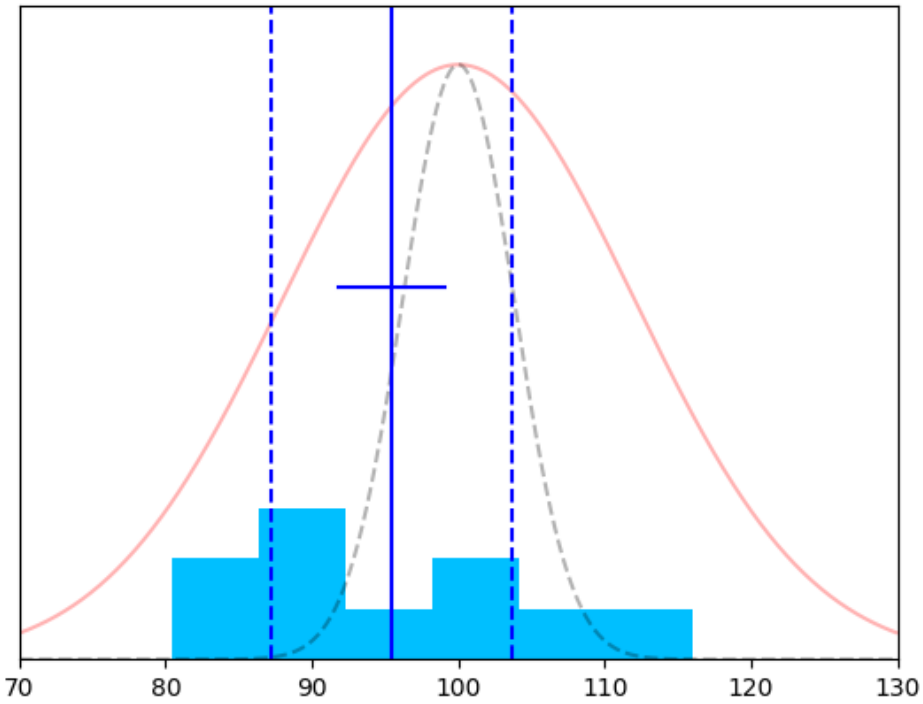
Larger confidence level (smaller α) $\rightarrow$ wider confidence interval.

Kahoot!

Overview

1. Summary from last week
2. Model Control
   - Q-Q plot
   - Transformation towards Normality
3. Hypothesis Testing
   - Null Hypothesis
   - t-test
   - p-value
   - One-sided vs Two-sided Tests
4. Type I and Type II Errors

Model Control



Normal distribution assumption:

If the sample size is small, our confidence interval calculation relies on the assumption that the underlying distribution (the red line) is normally distributed. In such cases, we should verify whether this assumption holds.

We can get an idea of what the underlying distribution looks like, for instance, from a histogram of our sample data.

Python: Sample from underlying normal distribution

- Go to today's Python notebook in VS Code
  - "Simulation: Sample from normal distribution"

Model Control

It can be difficult to determine from a **histogram** whether the data is normally distributed—especially if the sample size is small. Additionally, the shape of a histogram depends on the choice of bin size.

Another option is to plot the **empirical cumulative distribution function** (ECDF).

You can also use a **Q-Q plot** (quantile-quantile plot). This plot compares the sorted observations $x_{(1)}, \ldots, x_{(n)}$ against the theoretical quantiles of the normal distribution.
Different definitions of quantiles exist:

- For $n > 10$, the preferred quantiles are: $p_i = \frac{i-0.5}{n}$, $i = 1, \ldots, n$
- For $n \leq 10$, the preferred quantiles are: $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, \ldots, n$

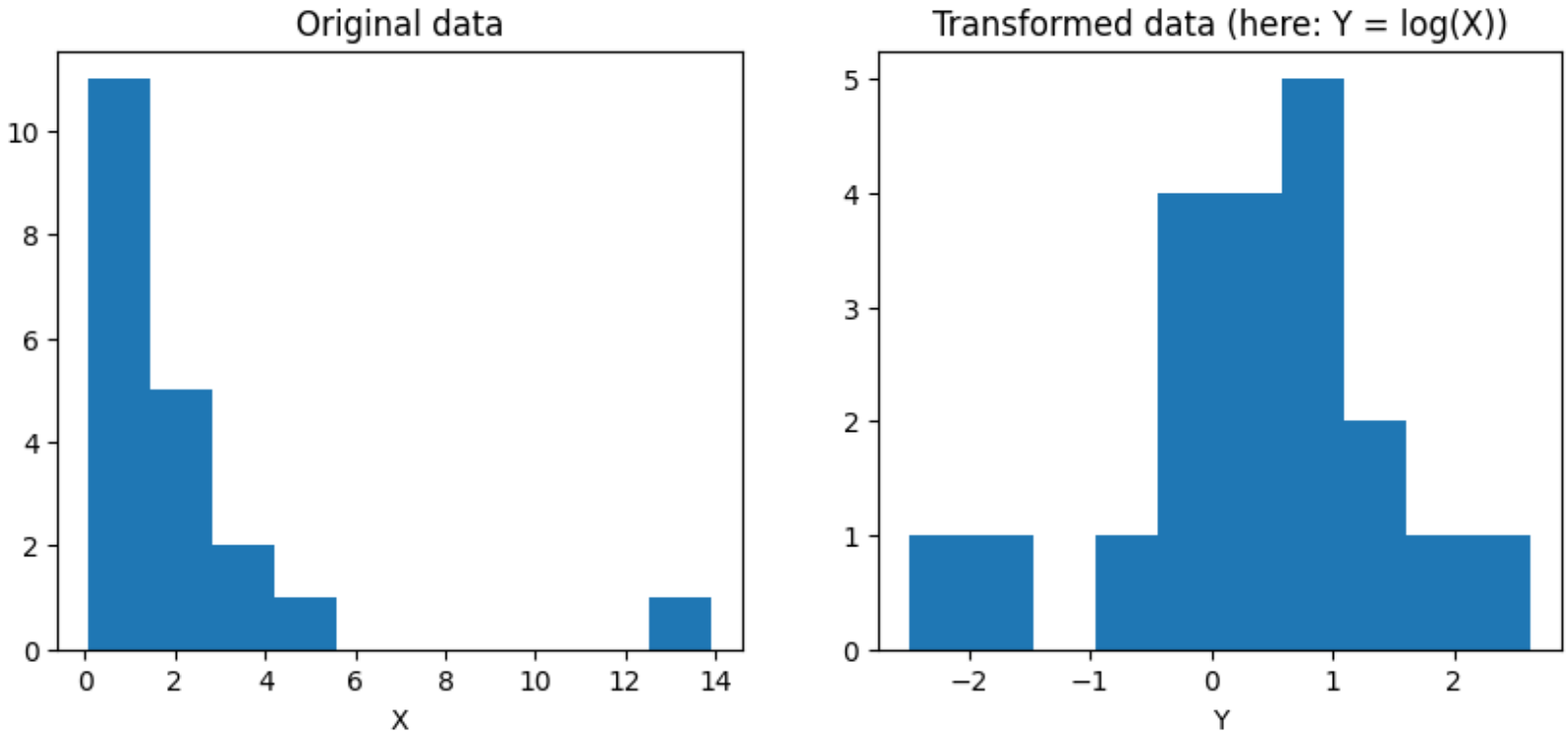
Python: ECDF and Q-Q plot

- Go to today's Python notebook in VS Code
  - "Simulation: ECDF and Q-Q plot"

Transformation towards Normality

If data is *not* normally distributed, one option is to *transform* the data in the hope that the transformed data will better follow a normal distribution.

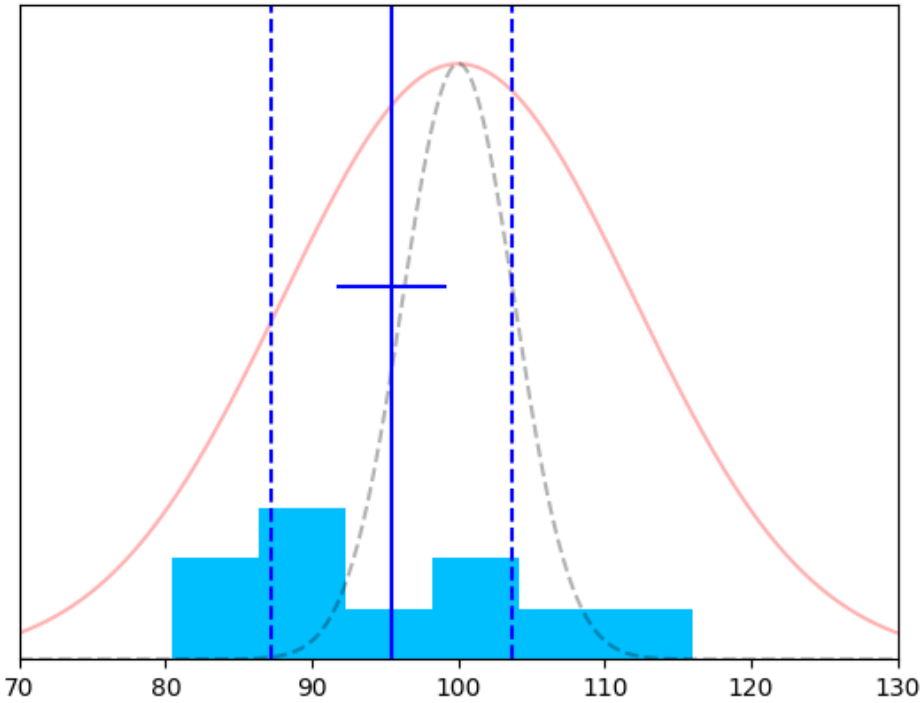
In such cases, calculations will be performed on the transformed data. It may be necessary to transform the results back to the original scale at the end.



Overview

1. Summary from last week
2. Model Control
   - Q-Q plot
   - Transformation towards Normality
3. Hypothesis Testing
   - Null Hypothesis
   - t-test
   - p-value
   - One-sided vs Two-sided Tests
4. Type I and Type II Errors

Hypotheses



Let us assume that we believe the necessary assumptions hold, and we have a sample from which we estimate $\hat{\mu}$ with an accompanying confidence interval.

The plot shows a 95% confidence interval (blue dashed lines).

We do not know the underlying distribution (red) – it could be shifted more to the right or left, or it could be wider or narrower.

Today, we will discuss **hypotheses** and **tests**.

For example, one might hypothesize that $\mu = 120$. Does that seem like a reasonable hypothesis? Why or why not?

What about the hypothesis $\mu = 90$?

Which hypotheses would we "accept"? And which hypotheses would we "reject"?

Example – Light Rail (Letbane)

Voltage drop measurements in light rail

During the construction of a light rail, voltage drop is measured at a location where it is expected to be zero (or very small). Ten independent measurements are made, resulting in the following voltages (in volts):

Sample:

Measurement	Voltage drop
1	0.75
2	-0.85
3	4.23
4	2.12
5	3.04
6	0.53
7	-0.35
8	1.69
9	1.52
10	-0.42

Example – Light Rail (Letbane)

Voltage drop measurements in light rail

During the construction of a light rail, voltage drop is measured at a location where it is expected to be zero (or very small). Ten independent measurements are made, resulting in the following voltages (in volts):

Sample:

Measurement	Voltage drop
1	0.75
2	-0.85
3	4.23
4	2.12
5	3.04
6	0.53
7	-0.35
8	1.69
9	1.52
10	-0.42

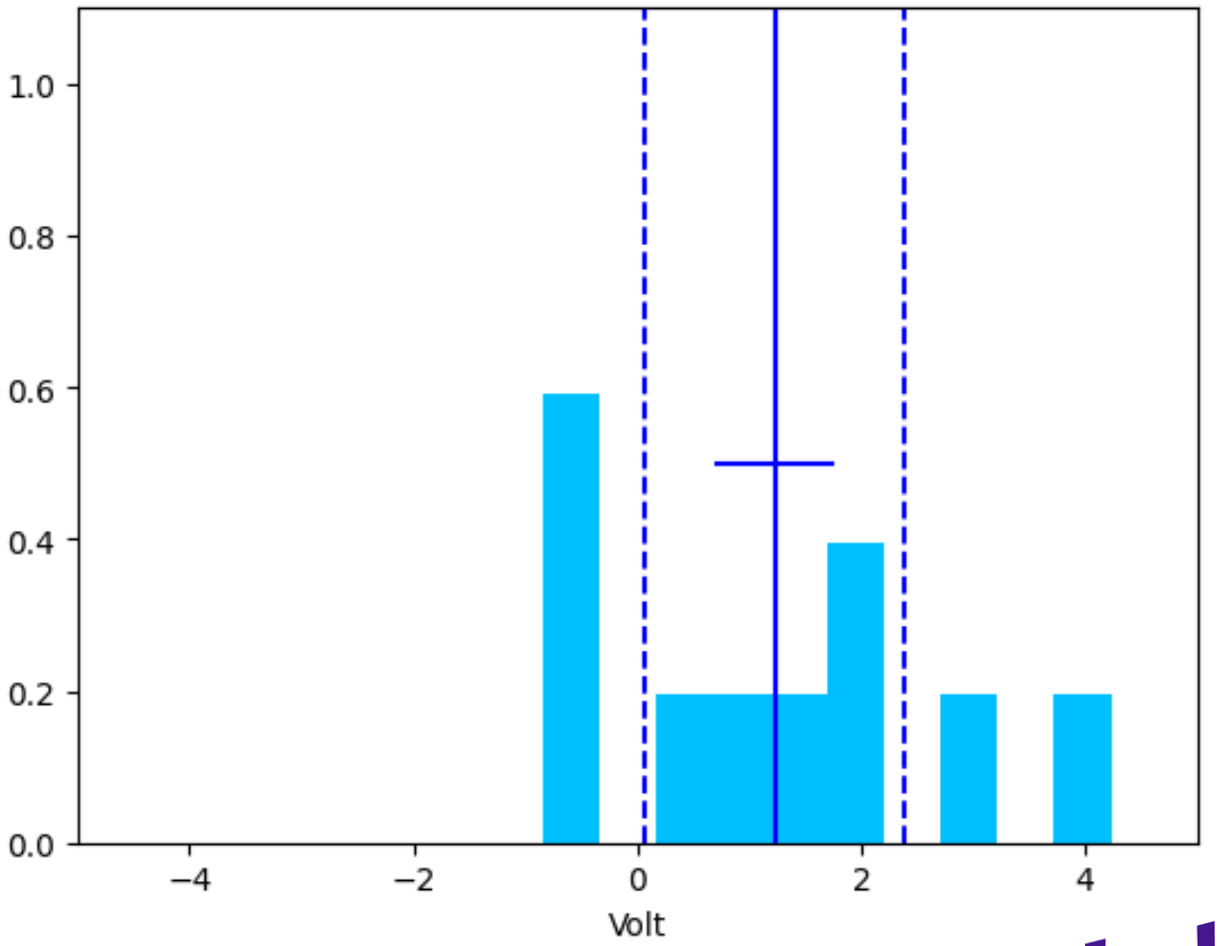
$x_i =$ voltage drop in measurement i

$\bar{x} = 1.23$ (sample mean)
$s = 1.62$ (sample standard deviation)

$\hat{\mu} = 1.23$

95%-confidence interval for $\hat{\mu}$:
$[0.07; 2.9]$

Example - Light Rail



How could the underlying distribution look?

Could the "true" μ be 0?

Kahoot!

(x2)

Null Hypothesis

Null Hypothesis:

We assume that μ takes a specific value

$$H_0 : \ \mu = \mu_0$$

where μ is the true population mean.

An example could be assuming $\mu_0 = 0$. This would often correspond to the null hypothesis of "no effect."

Null Hypothesis

Null Hypothesis:

We assume that μ takes a specific value

$$H_0: \ \mu = \mu_0$$

where μ is the true population mean.

An example could be assuming $\mu_0 = 0$. This would often correspond to the null hypothesis of "no effect."

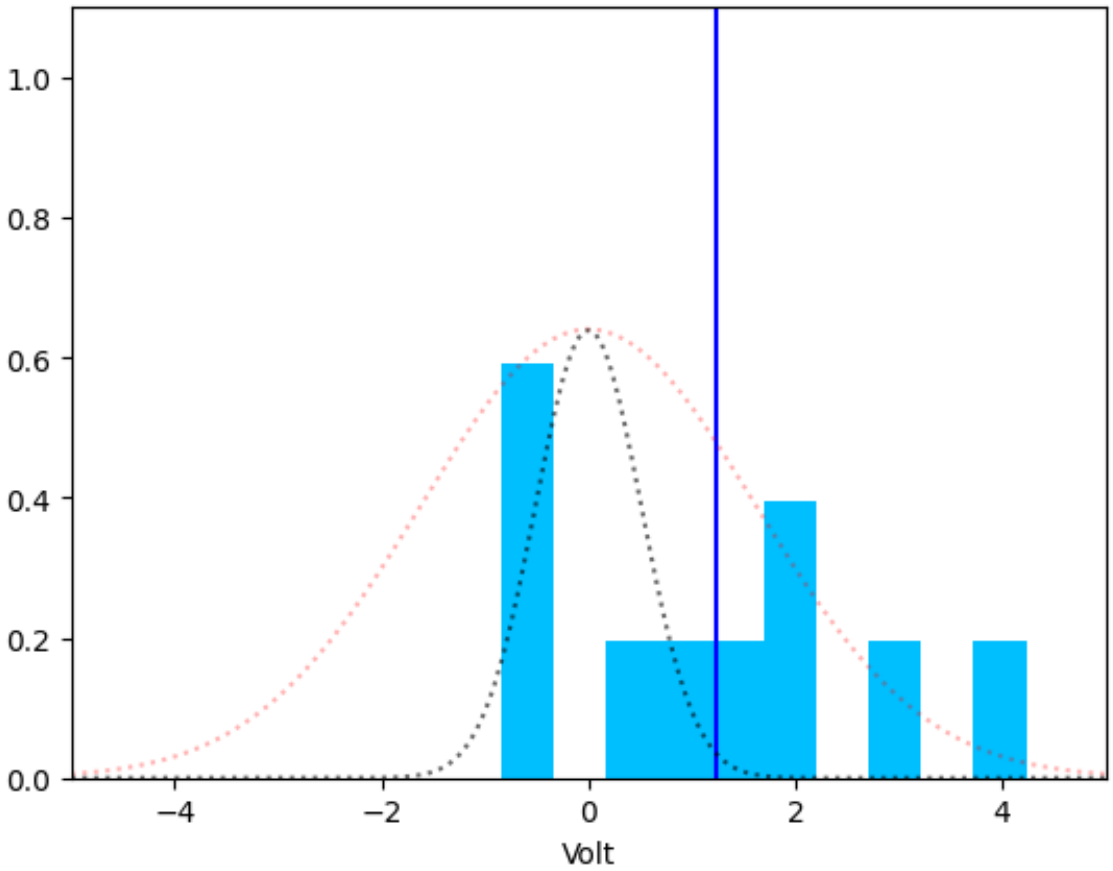
Data from the light rail example:

$$\bar{x} = 1.23 = \hat{\mu}$$
$$s = 1.62 = \hat{\sigma}$$

Is the data consistent with the null hypothesis H_0?

Data: $\bar{x} = 1.23$, $H_0: \ \mu = 0$

Example - Visualization of Null Hypothesis $\mu_0 = 0$



This plot shows a normal distribution $N(0, s)$ (red)
and a normal distribution $N(0, \frac{s}{\sqrt{n}})$ (black)
$(s = 1.62$ and $n = 10)$

Test Statistic: t_{obs}

Now we will find a way to "test" our null hypothesis:

We consider the statistic T under the **assumption that the null hypothesis is true**:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

This statistic follows a t-distribution: $T \sim t(n-1)$.

Based on the observed data, we calculate:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

t_{obs} is our observation of the random variable T (under the assumption that the null hypothesis is true).

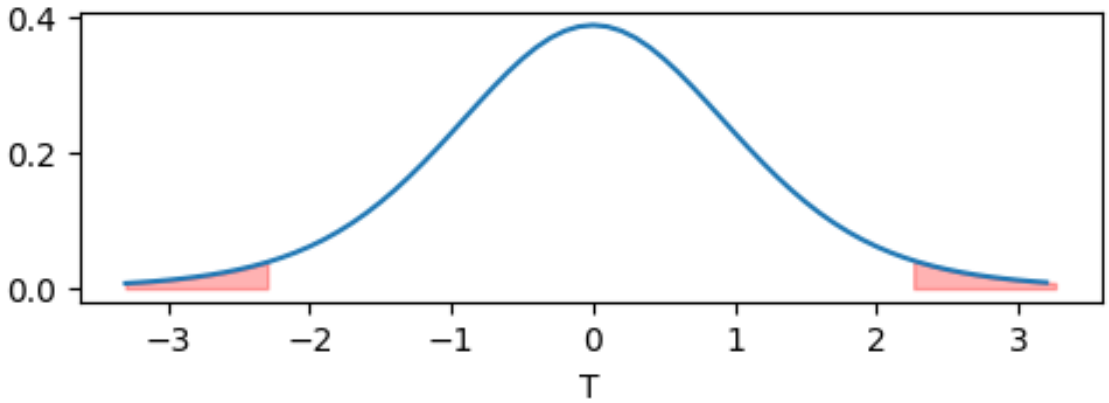
The question is whether the observed value t_{obs} is "too extreme"?

Student's t-test, or simply t-test

Hypothesis test: t-test

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

To answer whether the observed value t_{obs} is too "extreme," we **test** if t_{obs} lies within the interval $\pm t_{1-\alpha/2}$ (which depends on our chosen significance level). The test result is either Yes/No, along with the chosen significance level.



If $|t_{\text{obs}}| > t_{1-\alpha/2}$, we conclude that $|t_{\text{obs}}|$ is too large to believe in the null hypothesis — and we say that we **reject the null hypothesis**.

Example – Light Rail

Hypothesis of voltage drop $= 0$:

$$H_0: \ \mu = 0 \ \ (= "\mu_0")$$

where μ is the average voltage drop.

Example – Light Rail

Hypothesis of voltage drop = 0:

$$H_0: \ \mu = 0 \ \ (= "\mu_0")$$

where μ is the average voltage drop.

Calculate the test statistic:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.23 - 0}{1.62/\sqrt{10}} = 2.39$$

Example – Light Rail

Hypothesis of voltage drop = 0:

$$H_0: \ \mu = 0 \ \ (= "\mu_0")$$

where μ is the average voltage drop.

Calculate the test statistic:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.23 - 0}{1.62/\sqrt{10}} = 2.39$$

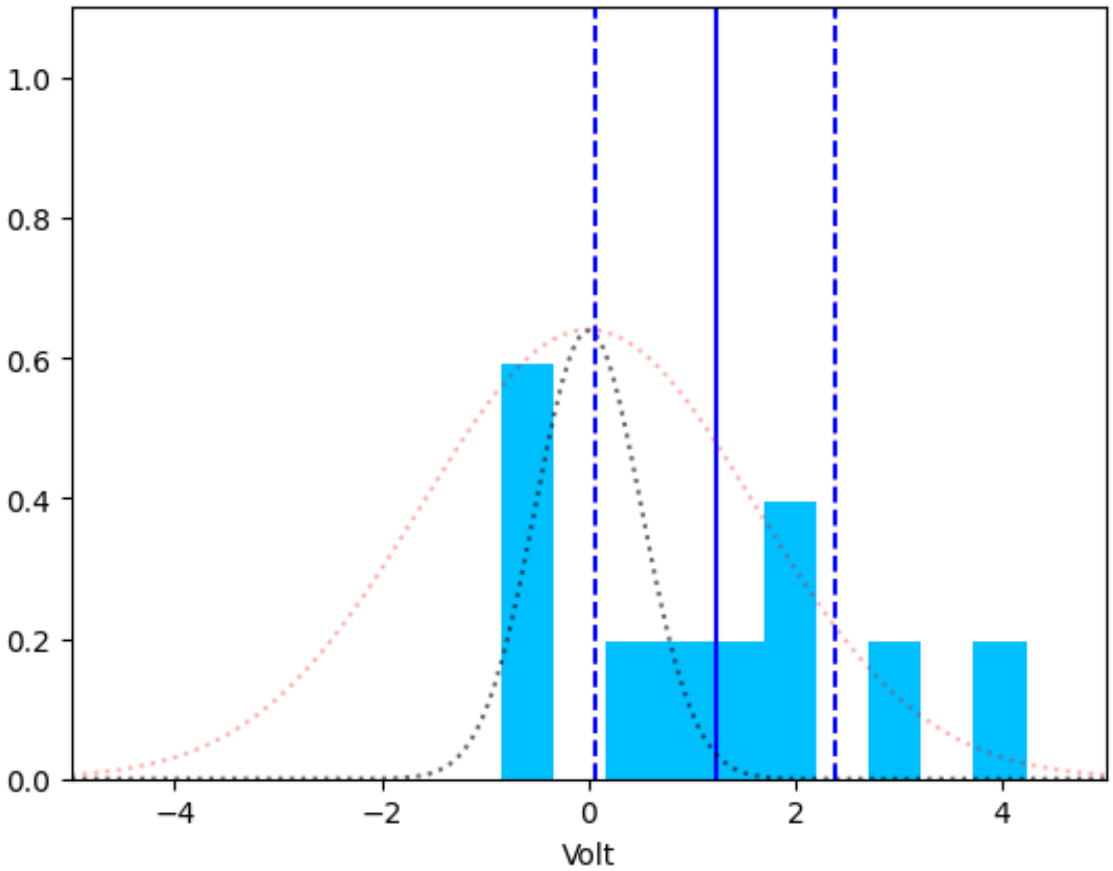
Compare with $t_{1-\alpha/2}$:

$$t_{1-\alpha/2} = 2.26 \ \ (\alpha = 0.05)$$

Conclusion:

$|t_{\text{obs}}| > t_{1-\alpha/2}$, and therefore we reject the null hypothesis (at the significance level of 0.05).

Example - Light Rail



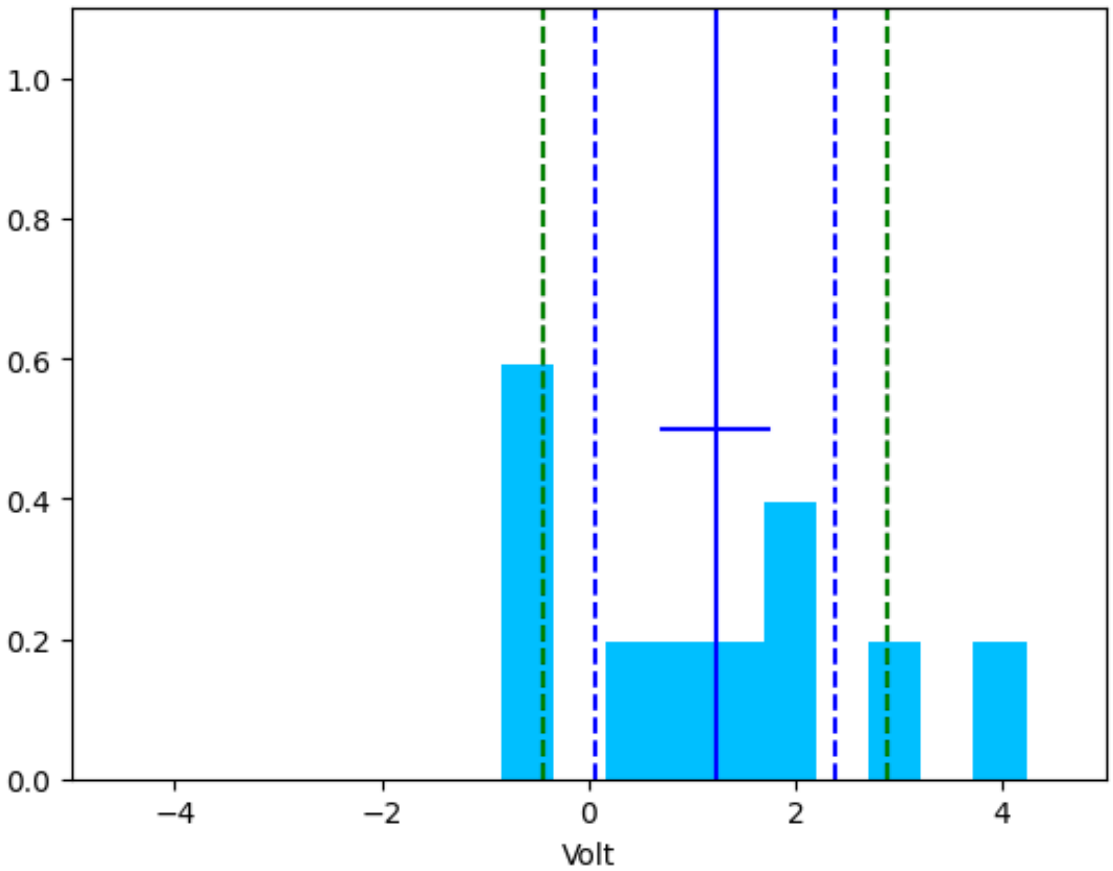
In this case, we reject the null hypothesis.

Note also that μ_0 is just outside the confidence interval (the 95% confidence interval for μ was $[0.07; 2.9]$).

Kahoot!

(x2)

Example - Light Rail



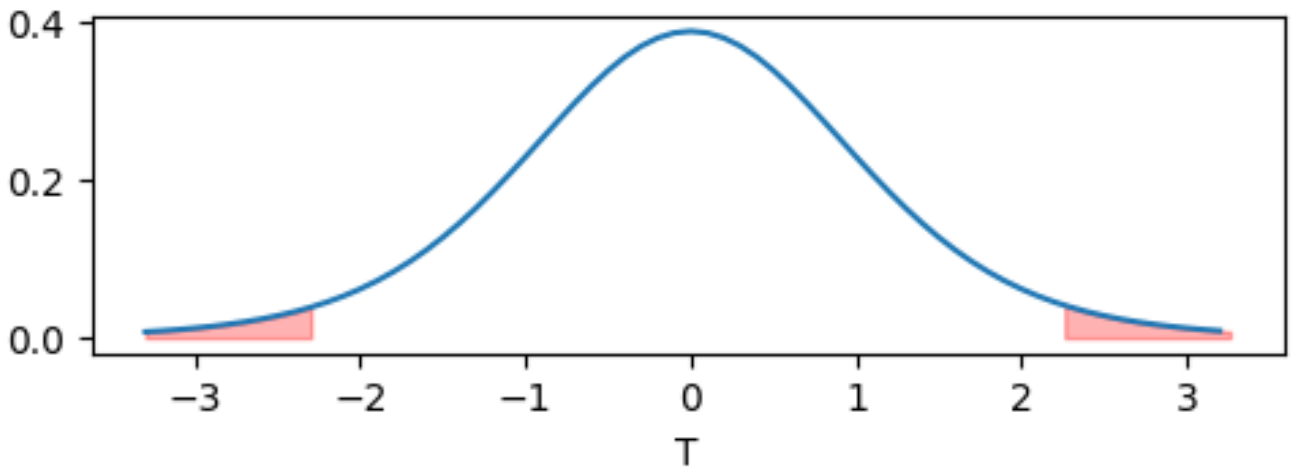
We could also ask - what **significance level** should we have chosen if the value $\mu = 0$ were to be inside the confidence interval?
95% (blue), 99% (green) ... ?

p-value

The next step in the t-test is calculating the p-value:

What significance level should we choose for t_{obs} to fall within our range of accepted T-values?

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



We now choose the boundaries precisely at $T = \pm t_{\text{obs}}$

That way, t_{obs} just barely fits within the interval, while values of T that are "more extreme" than t_{obs} fall outside the interval.

p-value

Definition of p-value:

For a (quantitative) situation with **one sample**, the **p-value** is given by:

$$\begin{aligned} p\text{-value} &= 2 \cdot P(T > |t_{\text{obs}}|) \\ &= P(T < -|t_{\text{obs}}|) + P(T > |t_{\text{obs}}|) \end{aligned}$$

where T follows a t-distribution with $(n-1)$ degrees of freedom.

The p-value is compared with the chosen significance level:

The desired significance level is expressed by α
(e.g., $\alpha = 0.05$).

We now check if the p-value is less than α.

Example – Light Rail

Hypothesis of no voltage drop:

$$H_0: \ \mu = 0 \ \ (= "\mu_0")$$

where μ is the average voltage drop.

Example – Light Rail

Hypothesis of no voltage drop:

$$H_0: \ \mu = 0 \ \ (= "\mu_0")$$

where μ is the average voltage drop.

Calculate the test statistic:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.23 - 0}{1.62/\sqrt{10}} = 2.39$$

Example – Light Rail

Hypothesis of no voltage drop:

$$H_0: \ \mu = 0 \ \ (= "\mu_0")$$

where μ is the average voltage drop.

Calculate the test statistic:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.23 - 0}{1.62/\sqrt{10}} = 2.39$$

Calculate the p-value:

$$2P(T > 2.39) = 0.0404$$

Conclusion:

If we choose $\alpha = 0.05$, we get $p < \alpha \rightarrow$ We reject the null hypothesis.

However, if we had chosen $\alpha = 0.04$, we would have accepted the null hypothesis.

Definition and Interpretation of p-value (General)

The p-value expresses *evidence* against the null hypothesis – Table 3.1:

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Very weak or no evidence against H_0

Definition 3.22 of p-value:

The p-value is the probability of observing a test statistic that is **at least as extreme** as the observed test statistic. This probability is calculated under the assumption that the null hypothesis is true.

Python

- Go to today's Python notebook in VS Code
  - "Example: Voltage drop"

Statistical Significance and Terminology

We say that we are *performing a hypothesis test* when we decide to reject or accept a null hypothesis based on data.

A null hypothesis is *rejected* if p-value $< \alpha$ (α is chosen in advance).

Otherwise, the null hypothesis is said to be *'accepted'*. It is more accurate (and preferable) to say that the null hypothesis cannot be rejected.

Statistical Significance: An *effect* is said to be *(statistically) significant* if the p-value is less than the significance level α.
This terminology is most meaningful when the null hypothesis is: $H_0: \ \mu = \mu_0 = 0$, which translates to the null hypothesis of "no effect."

It can also be said that μ is *significantly different from* μ_0.

Sometimes, we also say that we 'accept' the *alternative hypothesis*: $\mathrm{H}_A : \mu \neq \mu_0$

Steps in a t-test – An Overview

In general, a t-test consists of the following steps:

1. Formulate the null hypothesis (and alternative hypothesis) and choose a significance level α (set the "risk level").
2. Calculate the test statistic value from the observed data.
3. Compute the p-value based on the test statistic compared to the appropriate distribution.
4. Compare the p-value with the significance level α and conclude.

One-sided vs Two-sided Tests

So far, it has been implied that the test is two-sided (non-directional):

The alternative to $H_0: \ \mu = \mu_0$ is $H_A: \ \mu \neq \mu_0$.

One-sided vs Two-sided Tests

So far, it has been implied that the test is two-sided (non-directional):

The alternative to $H_0: \ \mu = \mu_0$ is $H_A: \ \mu \neq \mu_0$.

There may be other situations, such as one-sided (= directional) alternative hypotheses:

The alternative to $H_0: \ \mu = \mu_0$ is $H_A: \ \mu < \mu_0$.

One-sided vs Two-sided Tests

So far, it has been implied that the test is two-sided (non-directional):

The alternative to $H_0: \ \mu = \mu_0$ is $H_A: \ \mu \neq \mu_0$.

There may be other situations, such as one-sided (= directional) alternative hypotheses:

The alternative to $H_0: \ \mu = \mu_0$ is $H_A: \ \mu < \mu_0$.

In this course, we stick to the two-sided test (non-directional)!

Example with t-test

A t-test is often used to test if there is a **difference** between two things.

Example:
In a study, we want to examine whether there is a significant difference in calorie intake throughout women's menstrual cycle.

We measure the **difference in calorie intake** in 11 women before and after menstruation.

The data (difference in calories) is as follows: 1350, 1250, 1755, 1020, 745, 1835, 1540, 1540, 725, 1330, 1435

What is a relevant null hypothesis?

Do we believe there is a significant difference?

How should we conduct our statistical analysis of this data?

(x4)

Python:

- Go to today's Python notebook in VS Code
  - "Example: Difference in calorie intake"

Overview

Type I and Type II Errors

There are two types of errors (only one at a time)

Type I: Rejecting H_0 when H_0 is true.
Type II: Accepting (not rejecting) H_0 when H_1 is true.

Type I error is called a false positive.
Type II error is called a false negative.
In this terminology, $H_0 =$ "negative" ("no effect")
and $H_1 =$ "positive" (there is an "effect")

The risk of the two types of errors is usually referred to as:

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

$$P(\text{Type II error}) = \beta$$

Courtroom Analogy

A person is brought before a court:

A person is being prosecuted under a specific charge.
The null and alternative hypotheses are:

H_0 : The person is innocent.

H_1 : The person is guilty.

Courtroom Analogy

A person is brought before a court:

A person is being prosecuted under a specific charge.
The null and alternative hypotheses are:

H_0 : The person is innocent.

H_1 : The person is guilty.

Not being proven guilty is not the same as being proven innocent:

In other words:
Accepting a null hypothesis is not a statistical proof that the null hypothesis is true!

Errors in Hypothesis Testing

Theorem 3.39: The significance level is the risk of making a Type I error

The significance level α in hypothesis tests is the risk of a Type I error:

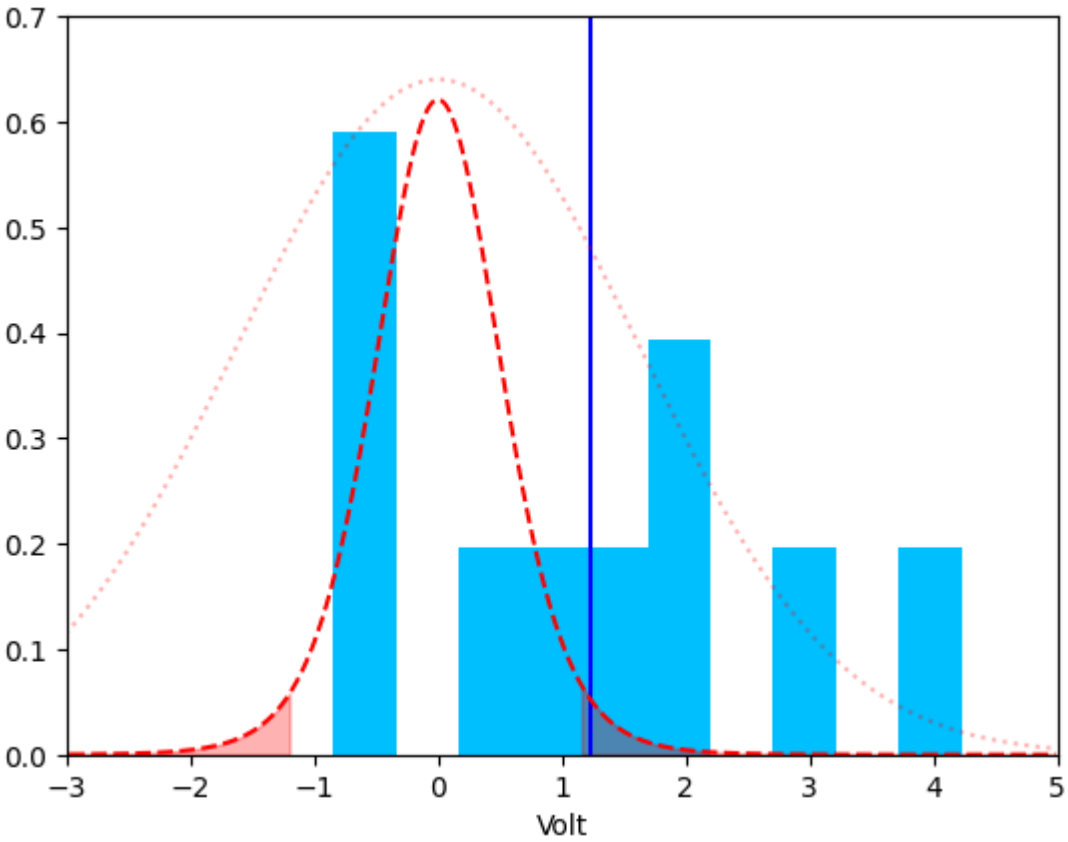
$$P(\text{Type I error}) = P(\text{Rejecting } H_0 \text{ when } H_0 \text{ is true}) = \alpha$$

Errors in Hypothesis Testing

Theorem 3.39: The significance level is the risk of making a Type I error

The significance level α in hypothesis tests is the risk of a Type I error:

$$P(\text{Type I error}) = P(\text{Rejecting } H_0 \text{ when } H_0 \text{ is true}) = \alpha$$

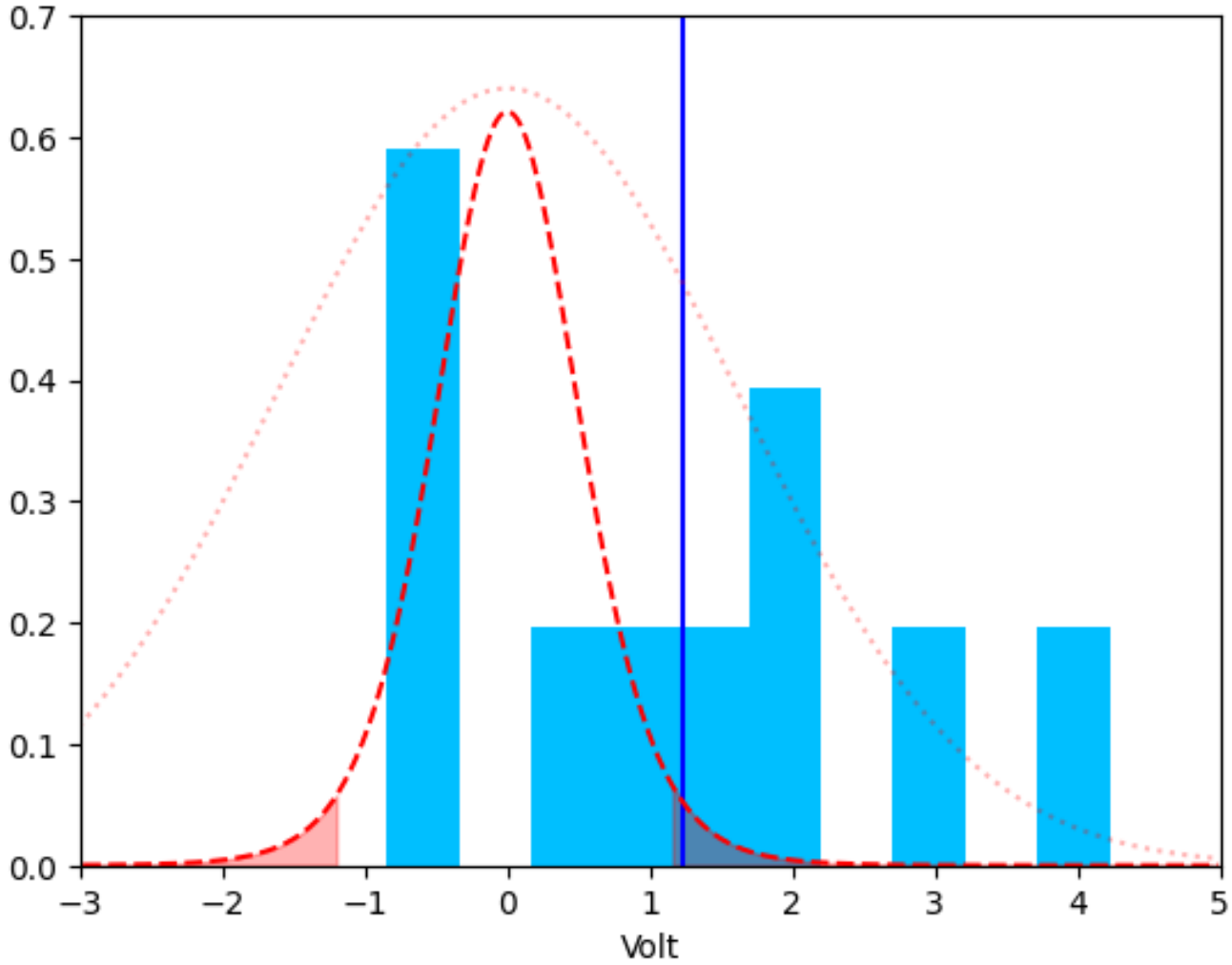


Errors in Hypothesis Testing

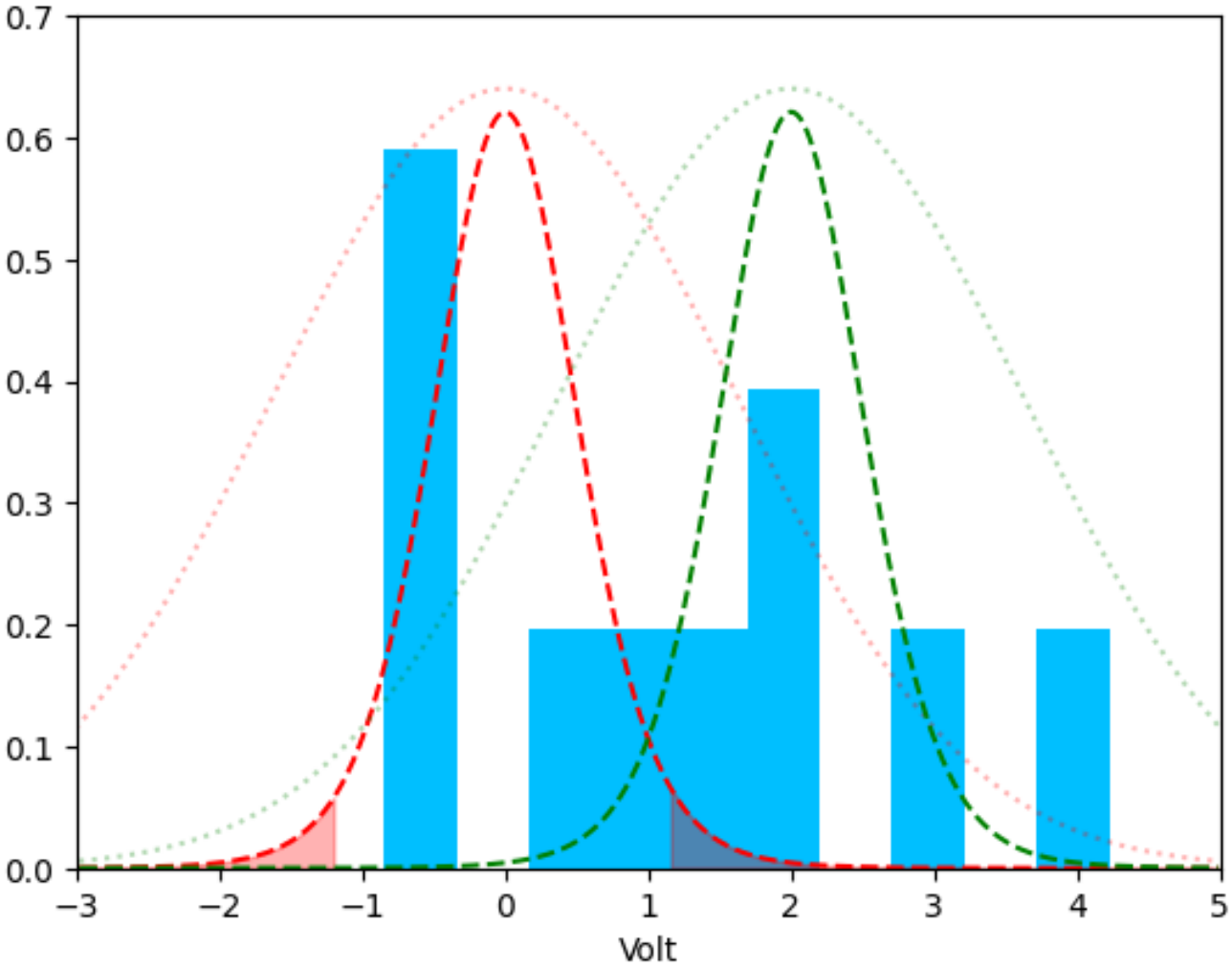
Two possible truths against two possible conclusions:

	Rejecting H_0	Not rejecting H_0
H_0 is true	Type I error (α)	Correct acceptance of H_0
H_0 is false	Correct rejection of H_0	Type II error (β)

Type I Error



Type I and Type II Errors



Smaller α = Larger β (and vice versa)

Agenda

1. Summary from last week
2. Model Control
   - Q-Q plot
   - Transformation towards Normality
3. Hypothesis Testing
   - Null Hypothesis
   - t-test
   - p-value
   - One-sided vs Two-sided Tests
4. Type I and Type II Errors