

02-680 Module 20

Essentials of Mathematics and Statistics

Aidan Jan

December 4, 2025

Hypothesis Testing

Hypothesis Testing is

- A technique to evaluate if a model fit matches our assumptions about the data.
- Allows us to assign a numerical value (e.g., p -value) to assess this match.

There are two main types of hypothesis:

- **Null Hypothesis** (H_0): The assumption that there is no effect or no difference.
- **Alternate Hypothesis** (H_1): The assumption that there is an effect or a difference.

Hypothesis Test Outcome

The test helps us decide to:

- Reject $H_0 \rightarrow$ Evidence supports H_1 .
- Retain $H_0 \rightarrow$ Not enough evidence to support H_1 .

Example context: Testing whether a drug impacts cholesterol:

- H_0 : Cholesterol stays the same. (No effect.)
- H_1 : Cholesterol level changes.

Errors

- Type I Error (False Positive): Rejecting the null hypothesis when H_0 is actually true.
- Type II Error (False Negative): Retaining the null hypothesis when H_0 is actually false.

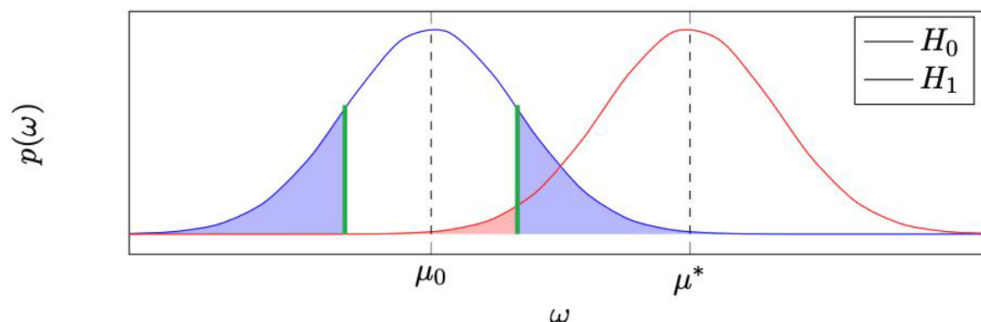
Truth	Hypothesis Test Result	
	Retain H_0	Reject H_0
H_0	Correct	Type I Error
H_1	Type II Error	Correct

We say the Type I Error Rate is $p(\text{reject } H_0 \mid H_0 \text{ is true})$, and Type II Error Rate is $p(\text{retain } H_0, H_1 \text{ is true})$, and Statistical Power is $1 - \text{Type II Error Rate}$. The last point means that the higher power tests have a stronger ability to detect signals for H_1 .

Example

Let's look at it visually, first for what we call a two-sided test, that is

$$H_0 : \mu = x \text{ and } H_1 : \mu \neq x$$



In the figure above, when we pick some boundary around our desired x (the green lines) we will have some probability of Type I Error (blue shaded regions) and Type II Error (red shaded region).

Defining Errors

In both cases we can choose the cutoff (the thick green lines) of where to make the distinction between H_0 and H_1 , but there is a **tradeoff**: as Type I error decreases, Type II error increases, and the power decreases. Similarly, as μ^* and μ_0 become further apart, both errors will decrease, and the signal becomes easier to detect.

Performing Tests

A test can be performed in three steps:

1. Compute a test statistic (a function of the data) that is appropriate for the distribution: $T = r(X_1, X_2, \dots, X_n)$.
2. Compute a p -value
3. For a desired significance level β and the p -value, decide whether to retain or reject H_0

Example

Suppose we are testing the efficacy of a drug for high cholesterol, and that we know the typical variance of cholesterol among humans (σ) and that this is not going to change between the conditions.

First, since we are assuming that the samples are coming from a Gaussian distribution ($X_1, X_2, \dots, X_n \sim N(\mu_0, \sigma^2)$), the test statistic is the mean (\bar{X}_n). Assume we are running a two-sided test (we don't know if the impact is going to lower or raise cholesterol, we just want to know if it changes). Thus, $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

Under the null hypothesis,

$$\bar{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n)$$

or, we can also write

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

To compute the p -value, which is essentially the probability that we would see the test statistic under the null hypothesis, $T \sim N(0, 1)$:

$$p \left(|T| > \left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right| \right)$$

Depending on how strict we want to be, we accept or reject H_0 based on the p -value. Some general rules on how to choose α :

p-value	interpretation
< 0.01	very strong evidence against H_0
0.01 – 0.05	strong evidence against H_0
0.05 – 0.1	weak evidence against H_0
> 0.1	little to no evidence against H_0

For a one-sided test, that is one where for example, $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$. The only thing that changes is that in Step 2:

$$p \left(T > \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right)$$

Notice that usually one-sided tests can be more powerful as they are only integrating over one region.

Summary

When performing tests, the main thing we need to do is to determine:

- the distribution we think the data came from
- the test statistic that is appropriate for those samples
- the distribution that applies to the test statistic (it may be different than the one for the data).

t -tests: When σ is Unknown

Let's assume again that $X_1, X_2, \dots, X_n \sim N(\mu_0, \sigma^2)$, but this time we don't know σ . We are going to define two statistics:

$$\bar{X}_n = \frac{1}{n} \sum X_i$$

$$\bar{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$$

We're then going to say the following:

$$\frac{\bar{X}_n - \mu_0}{\bar{\sigma}/\sqrt{n}} \sim t_{n-1}$$

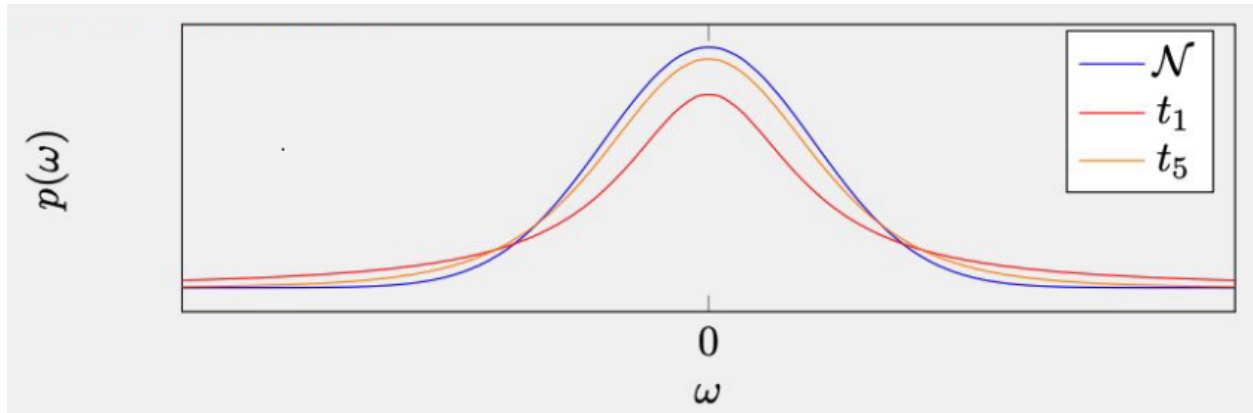
The t distribution

The t distribution has one parameter, ν , which is the degrees of freedom.

$$X \sim t_\nu$$

$$p(X = x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu} \cdot \Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu} \right)^{-\frac{\nu+1}{2}}$$

The shape of the t distribution is similar to the shape of the normal distribution but t distribution has a thicker tail.



As $\nu \rightarrow \infty$, the distribution above becomes a normal distribution.

So, we have a way to compute something that should be in t_{n-1} , that is, we assume there is one less degree of freedom than there are elements in the observation.

Since the probability of the t distribution is difficult to calculate, we typically have lookup tables for it. (Google these for reference).

- To use one of these tables, find your degrees of freedom in the left column and use that row to find the column with the next smaller number from your statistic.
- Read the probability in the top row.
- Since your t will probability be a little bit bigger than the value in the table, your p will be smaller, e.g., $p < 0.01$.
- If your t is to the right of all numbers, then p is less than the right-most probability.

Paired Data: Paired t -tests

Many times the data we have is a set of samples measured in two different conditions. As an example, a set of patients measured before and after a treatment. In that case we want to know if the treatment made a consistent change across the population. We also don't know where each of the individuals sits compared with some (unknown) μ .

In this case, our hypotheses are:

$$H_0 : X_1 = X_2 \quad H_1 : X_1 \neq X_2$$

where X_1 and X_2 are the two experimental conditions. Another way to say this is to define some $\delta = X_1 - X_2$, and let the hypotheses be

$$H_0 : \delta = 0 \quad H_1 : \delta \neq 0$$

As we did before we can compute $\bar{\delta}_n$ and $\bar{\sigma}_\delta$ from the data and let our statistic be

$$\frac{\bar{\delta}_n - \mu_0}{\bar{\sigma}_\delta / \sqrt{n}} \sim t_{n-1}$$

Testing Categorical Data: χ^2 tests

Sometimes we have data where we have some underlying conceptual probabilities for a group of categories, and want to know how well what we observed fits this concept. Specifically, assume we have some underlying assumption that we will see a set of n categories with the following probabilities:

$$\dot{p} = (\dot{p}_1, \dot{p}_2, \dots, \dot{p}_n)$$

And a set of observations, which we converted to probabilities

$$p = (p_1, p_2, \dots, p_n)$$

We then want to test

$$H_0 : \dot{p} = p \quad H_1 : \dot{p} \neq p$$

(in both cases we assume $\sum \dot{p}_i = 1$ and $\sum p_i = 1$.)

Example

A good example of this is Mendel's pea experiment. Mendel's hypothesis was that the proportion of round/yellow peas, wrinkled/yellow peas, round/green peas, and wrinkled/green peas is given as

$$\dot{p} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right)$$

Let's assume the observations were as follows:

	round/yellow	wrinkled/yellow	round/green	wrinkled/green	total
count	315	101	108	32	556
expected counts	312.75	104.25	104.25	34.75	556

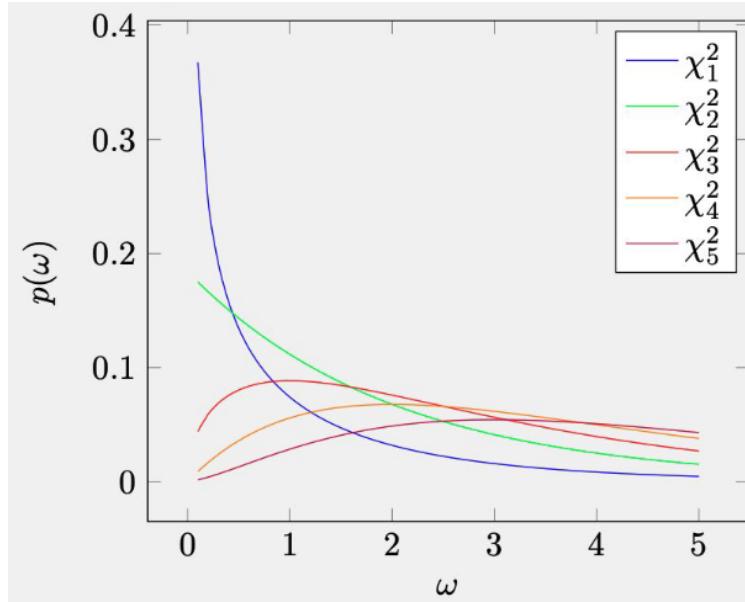
In this case, the test statistic will be

$$\sum \frac{(c_i - n \cdot \dot{p}_i)^2}{\dot{p}_i} \sim \chi_{n-1}^2$$

Here, p is the number of degrees of freedom, and in all cases, $x > 0$.

$$X \sim \chi_p^2$$

$$p(X = x) = \frac{x^{\frac{p}{2}-1}}{\Gamma\left(\frac{p}{2}\right) 2^{\frac{p}{2}} e^{-\frac{x}{2}}}$$



If Z_1, Z_2, \dots, Z_p are independent standard normal random variables, then

$$\sum Z_i^2 \sim \chi_p^2$$

For the example with Mendel's pea plants,

$$\sum_{i=1}^4 \frac{(c_i - n \cdot \dot{p}_i)^2}{\dot{p}_i}$$

$$= \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} \approx 0.47$$

Just like with t -tests we typically look up the p -value for χ^2 tests in a table. (Google this). We have four categories, which corresponds to 3 degrees of freedom. The p -value is 0.9524, thus we cannot reject the null hypothesis.