Neyman-Pearson Decision Theory

Neyman-Pearson Decision Theory

In Neyman-Pearson decision theory, we consider two competing hypotheses, denoted H_0 and H_1.

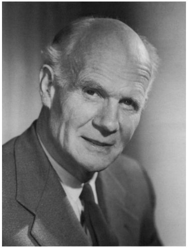
As before, we seek to ***reject H_0***, in which case we ***accept H_1***.

We say that

- H_0 is the ***null hypothesis***,
- H_1 is the ***research hypothesis*** or ***alternative hypothesis***.

The main difference to Fisher's approach is that we actually want to make a decision between two discrete possibilities instead of just finding evidence for or against H_0.

Neyman, Jerzy (1894-1981) Jerzy Neyman, Book of Proofs, https://www.bookofproofs.org/history/jerzy-neyman/



Egon Sharpe Pearson (1895-1980) Bartlett, M. S. Egon Sharpe Pearson. 11 August 1895-12 June 1980. Biographical Memoirs of Fellows of the Royal Society, vol. 27, 1981, pp. 425443. JSTOR

Example of Neyman-Pearson Decision Theory

15.1. Example. Let us revisit Example **??**. The mean burning rate for a rocket propellant is supposed to be $\mu_0 = 40\,\text{cm/s}$. It is known that the standard deviation is $\sigma = 2\,\text{cm/s}$. If the rocket propellant burns significantly too fast or too slowly, it can not be used. An experimenter sets out the two hypotheses

$$H_0\colon \mu = 40, \qquad\qquad H_1\colon |\mu - 40| \geq 1.$$

If there is evidence that H_1 is true, the rocket propellant must be discarded, otherwise it can be used.

The P-value in Fisher's test procedure represents a continuum of evidence against H_0, while in the Neyman-Pearson approach we will define a sharp cut-off point for our data. If the data lies beyond this cut-off point, H_0 is rejected and H_1 is accepted.

Accepting Hypotheses

The statistical test will end with either

- failing to reject H_0, therefore accepting H_0 or
- rejecting H_0, thereby accepting H_1.

If we accept H_0, we do not necessarily believe H_0 to be true; we simply decide to act as if it were true. The same is the case if we decide to accept H_1; we are not necessarily convinced that H_1 is true, we merely decide to assume that it is.

15.2. Example. In the situation described in Example 15.1,

- accepting H_0 means that we assume that the rocket propellant burns at a mean rate of $40\,\text{cm/s}$. It does not mean that we actually believe that the value is precisely 40 and not 39.993, for instance.
- accepting H_1 means that we assume that the rocket fuel burns at a rate different by more than $1\,\text{cm/s}$ from the nominal rate. It does not necessarily mean that we have evidence to support this, merely that we will assume that it is the case.

Type I and Type II Errors

Given a choice between H_0 and H_1, there are four possible outcomes of the decision-making process:

(i) We reject H_0 (and accept H_1) when H_0 is false.

(ii) We reject H_0 (accept H_1) even though H_0 is true ***(Type I error)***.

(iii) We fail to reject H_0 even though H_1 is true ***(Type II error)***.

(iv) We fail to reject H_0 when H_0 is true.

We will design a test to decide between rejecting or failing to reject H_0 based solely on the probability of committing Type I or Type II errors, which we want (of course) to keep as small as possible.

Power, Type I & Type II Error Probabilities

We define the probability of committing a Type I error,

$$\begin{aligned} \alpha := P[\text{Type I error}] &= P[\text{reject } H_0 \mid H_0 \text{ true}] \\ &= P[\text{accept } H_1 \mid H_0 \text{ true}]. \end{aligned}$$

The probability of committing a Type II error is denoted

$$\begin{aligned} \beta := P[\text{Type II error}] &= P[\text{fail to reject } H_0 \mid H_1 \text{ true}] \\ &= P[\text{accept } H_0 \mid H_1 \text{ true}]. \end{aligned}$$

Related to β is the ***power*** of the test, defined as

$$\begin{aligned} \text{Power} := 1 - \beta &= P[\text{reject } H_0 \mid H_1 \text{ true}] \\ &= P[\text{accept } H_1 \mid H_1 \text{ true}]. \end{aligned}$$

α and the Critical Region

To set up the test, we select a test statistic and determine a ***critical region*** for the test: if the value of the test statistic falls into the critical region, then we reject H_0. Our critical region is determined by the desire to keep α small, e.g., less than 5%.

Hence, we determine the critical region in such a way that if H_0 is true, then the probability of the test statistic's values falling into the critical region is not more than α.

15.3. Example. In the situation described in Example 15.1, we may use $\overline{X}$ as a test statistic. The experimenter tests a sample of $n = 25$ specimen.

If H_0 is true, $\overline{X}$ will follow a normal distribution with mean $\mu = 40$ and $\sigma/\sqrt{n} = 2/5$, i.e.,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

follows a standard normal distribution.

α and the Critical Region

Hence, with a probability of $1-\alpha$,

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}.$$

If H_0 is true, then the probability that

$$\frac{|\overline{X}-\mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

is equal to α. Therefore, the critical region is determined by

$$\overline{x} \neq \mu_0 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{15.1}$$

α and the Critical Region

Suppose the experimenter would like to limit α, the probability of committing a Type I error if she rejects H_0, to 5%. This corresponds to $z_{\alpha/2} = 1.96$ and inserting the values for μ_0, σ and n, we find with probability $1 - \alpha$,

$$39.216 < \overline{X} < 40.784.$$

Hence the ***critical region*** is determined by

$$|\overline{X} - 40| > 0.784. \tag{15.2}$$

If $\overline{X}$ falls into the range of values satisfying (15.2), the experimenter will reject H_0, knowing that this decision will be wrong with a probability of at most 5%.

α and the Critical Region

15.4. Remarks.

(i) In this scheme, The decision whether to reject H_0 or not is not driven by the probability of H_0 being true or not, but solely by the probability of committing an error if H_0 is falsely rejected.

(ii) Only H_0 plays a role in the calculation of the critical region. H_1 does not enter into the discussion at all.

(iii) Rejecting H_0 (when the data falls into the critical region) does not actually mean that there is proof that H_1 is true; in the example above, H_0 can be rejected even if $|\overline{X} - 40| < 1$.

α and the Critical Region

If the experimenter in the previous example had wanted to reduce the probability of making a wrong decision when rejecting H_0, she could have set a higher bar for rejection: to achieve $\alpha = 1\%$, she would require

$$\left| \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2} = 2.575.$$

This would lead to a critical region of

$$|\overline{X} - 40| > 1.03.$$

If H_0 were then rejected because the sample mean fell into the critical region, the chance of this being in error would only be 1%. The trade-off is that it becomes less likely that the data will allow rejection of H_0 in the first place.

In this context, it is important to note:

> *In order for the statistical procedure to be valid, the critical region must be fixed* ***before data are obtained****.*

β and the Sample Size

The second type of error concerns failing to reject H_0 even though H_1 is true. We calculate this probability in the case of

$$H_0\colon \mu = \mu_0, \qquad\qquad H_1\colon |\mu - \mu_0| > \delta$$

as follows. Suppose that the true value of the mean is $\mu = \mu_0 + \delta$, $\delta > 0$. The test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

will then follow a normal distribution with unit variance and mean $\delta\sqrt{n}/\sigma$. Supposing that the critical region and α have been fixed, we will fail to reject H_0 if

$$-z_{\alpha/2} \le Z \le z_{\alpha/2}.$$

Calculating β for the Normal Distribution

Using the density of the normal distribution, we then find

$$\begin{aligned} &P[\text{fail to reject } H_0 \mid \mu = \mu_0 + \delta] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-z_{\alpha/2}}^{z_{\alpha/2}} e^{-(t-\delta\sqrt{n}/\sigma)^2/2}\, dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-z_{\alpha/2}-\delta\sqrt{n}/\sigma}^{z_{\alpha/2}-\delta\sqrt{n}/\sigma} e^{-t^2/2}\, dt \qquad (15.3) \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2}-\delta\sqrt{n}/\sigma} e^{-t^2/2}\, dt. \end{aligned}$$

Calculating β for the Normal Distribution

Let us suppose H_1 is true, i.e., $|\mu - \mu_0| > \delta$. Then

$$\begin{aligned}\beta &= P[\text{fail to reject } H_0 \mid H_1 \text{ true}] \\ &\leq P[\text{fail to reject } H_0 \mid \mu = \mu_0 + \delta]\end{aligned}$$

and we have (to good approximation)

$$\beta \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2} - \delta\sqrt{n}/\sigma} e^{-t^2/2}\, dt.$$

Adapting the notation from (13.2), we use the number $z_\beta \in \mathbb{R}$ to indicate

$$\beta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z_\beta} e^{-t^2/2}\, dt.$$

Calculating β for the Normal Distribution

Then the relationship between δ, α, β and n with σ known is given by

$$-z_\beta \approx z_{\alpha/2} - \delta\sqrt{n}/\sigma$$

or

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}. \tag{15.4}$$

In this way a desired (small) β can be attained by choosing an appropriate sample size n.

Designing an Experiment for Desired α and β

15.5. Example. Revisiting Example 15.1, the experimenter would like to test the hypotheses

$$H_0 \colon \mu = 40, \qquad H_1 \colon |\mu - 40| \geq 1.$$

in such a way that $\alpha = 5\%$ and $\beta = 10\%$, i.e, if H_0 is rejected, there is a 5% chance of this being in error, and if H_0 is not rejected (H_1 is accepted) there is a 10% chance of this being in error.

The critical region is set as before and the necessary sample size is calculated from (15.4) using $\beta = 0.10$, $\alpha = 0.05$, $\sigma = 2\,\text{cm/s}$ and $\delta = 1\,\text{cm/s}$. Then

$$n \approx 42,$$

so the sample size should be at least 42 to ensure $\beta \leq 0.10$.

Power

Another way to think about β is in terms of ***power***, defined as $1 - \beta$ and formally given by

$$1 - \beta = P[\text{accept } H_1 \mid H_1 \text{ true}].$$

A given experiment is set up so that we either reject H_0 or we don't. Generally, we would like the probability of rejecting H_0 if the alternative hypothesis is true to be high, i.e., β to be small. Choosing a sufficiently large sample size ensures that the data gathered is powerful enough to actually reject H_0, assuming H_1 is true.

One says that an experiment has ***high power*** if rejection of H_0 is likely, assuming H_1 is true.Generally speaking, a given test is more powerful than another if it requires a smaller sample size to attain the same β.

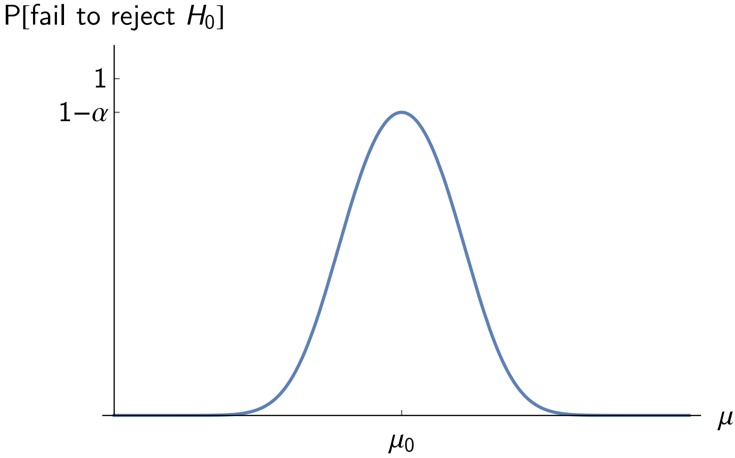
Operating Characteristic (OC) Curves

In (15.3) we calculated the probability of failing to reject H_0 as an integral. In practice, it may be difficult to perform such a calculations for non-normal distributions and evaluating the resulting integral may be impractical. For this reason, it is possible to refer to so-called ***operating characteristic curves***, known also as ***OC curves***.

A single OC curve plots the probability of failing to reject H_0 in a one-sided or two-sided test as a function of the parameter θ. A single such curve represents a choice of test parameters α and n. Other parameters of the distribution are also incorporated into the graph.

Operating Characteristic (OC) Curves

The figure below shows an OC curve for a two-sided test of the null hypothesis $H_0\colon \mu = \mu_0$ performed at fixed level α and fixed sample size n.



Effect of α on an OC Curve

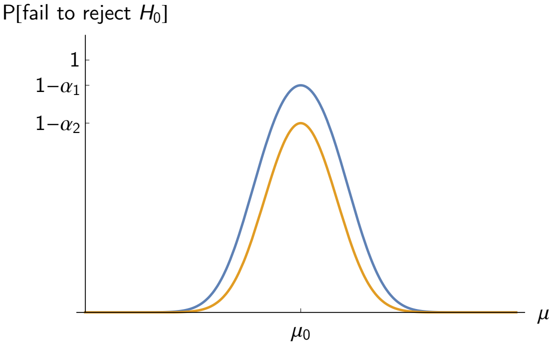
Note that

$$P[\text{fail to reject } H_0 \mid \mu = \mu_0] = 1 - \alpha,$$

since

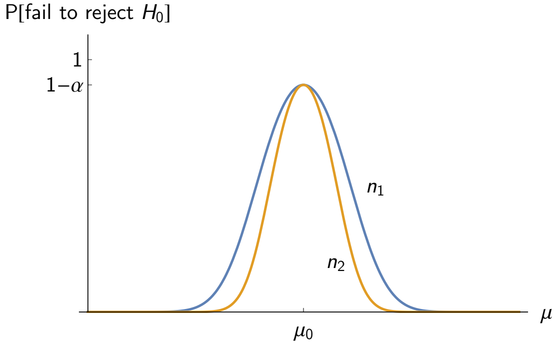
$$P[\text{reject } H_0 \mid \mu = \mu_0] = P[\text{reject } H_0 \mid H_0 \text{ true}] = \alpha,$$

by the construction of the test. For different values of α, the curves scale correspondingly:



Effect of the Sample Size on an OC Curve

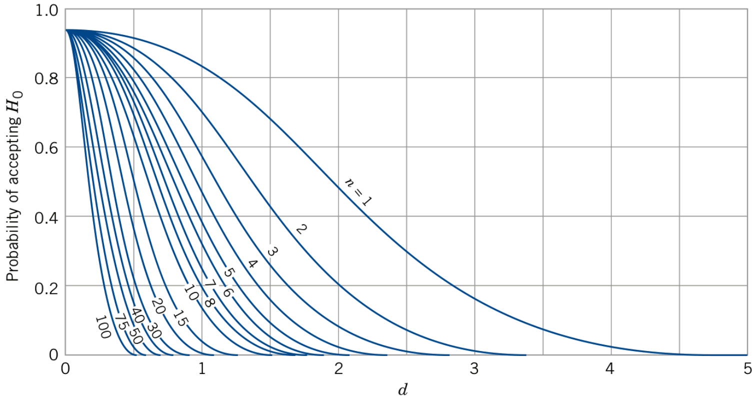
The sample size affects an OC curve as shown below for $n_2 > n_1$:



A typical graph will show OC curves for various values of n. Furthermore, for two-sided tests, only the right-hand half of the curve is shown to save space.

Using OC Curves to Relate Sample Sizes with β

15.6. Example. Continuing from Example **??**, suppose that the analyst is concerned about the probability of a Type II error if the true mean burning rate is $\mu = 41\,\text{cm/s}$. We may use the following operating characteristic curve (specific to $\alpha = 0.05$) to find β:



Using OC Curves to Relate Sample Sizes with β

In this graph,

$$d := \frac{|\mu - \mu_0|}{\sigma} = \frac{41 - 40}{2} = \frac{1}{2}.$$

Since in our example $n = 25$ we can read off $\beta \approx 0.30$.

15.7. Example. In Examples 15.5 we used a formula to find the sample size necessary to reject H_0 if H_1 is actually true. We can also read the result directly from the OC curve as follows:

We want to have $\beta \le 0.1$ if

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\mu - 40|}{2} \ge \frac{1}{2}.$$

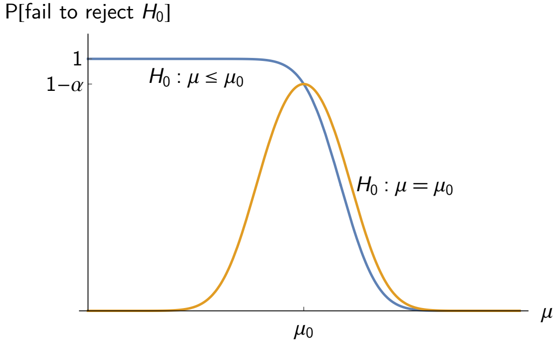
We see that the point $(d, \beta) = (0.5, 0.1)$ is intersected by the OC curve for $n = 40$ and that the curve remains below 0.1 for $d > 1/2$. Thus the test should involve a sample size of $n = 40$ or more.

OC Curves for One-Tailed Tests

Given a one-sided null hypothesis of the form

$$H_0\colon \theta \leq \theta_0, \qquad \text{or} \qquad H_0\colon \theta \geq \theta_0$$

an analogous calculation the probability of failing to reject H_0 may be performed, leading to an OC curve as shown below:



Summary of Neyman-Pearson Decision Theory

(i) Select appropriate hypotheses H_1 and H_0 and a test statistic;

(ii) Fix α and the critical region for the test;

(iii) Fix β and the sample size for the test;

(iv) Obtain the sample statistic; if the test statistic falls into the critical region, reject H_0 at significance level α and accept H_1. Otherwise, accept H_0.

Comparison of Fisher and Neyman-Pearson Tests

Superficially, Fisher's test of H_0 and the Neyman-Pearson test are related as follows:

> *If the P-value in Fisher's test is no greater than the value of α in Neyman-Pearson's decision process, then H_0 is rejected and H_1 accepted. Otherwise, H_0 is not rejected.*

However, this ignores the different philosophies of the approaches: Fisher is concerned about gathering evidence against H_0, without necessarily coming to an outright rejection, while Neyman-Pearson desire a definite decision for either H_1 or H_0.

Relationship to Confidence Intervals

We have seen in (15.1) that the two-tailed null hypothesis $H_0 \colon \mu = \mu_0$ is rejected if

$$\overline{x} \neq \mu_0 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

This is equivalent to

$$\mu_0 \neq \overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Hence, we have the following relationship to hypothesis tests:

- Neyman-Pearson: $\overline{x}$ lies in the critical region for α if and only if the null value μ_0 does not lie in a $100(1-\alpha)\%$ two-sided confidence interval for μ.
- Fisher: H_0 is rejected at significance level α if and only if the null value μ_0 does not lie in a $100(1-\alpha)\%$ two-sided confidence interval for μ.

This generalizes to one-sided tests and is also true for other (non-normal) distributions.

Interpretation of the Neyman-Pearson Decision

Suppose that you are performing a Neyman-Pearson test for a population mean with

$$H_0 \colon \mu \leq \mu_0, \qquad\qquad H_1 \colon \mu > \mu_1$$

where $\mu_0 < \mu_1$. The test has been designed so that $\alpha = 1\%$, $\beta = 5\%$.

Finally, H_0 is not rejected, i.e., H_0 is accepted. Then

(1) There is at most a 5% chance that H_1 is true.

(2) There is a 99% chance that H_0 is true.

(3) There is a 95% chance of this conclusion being correct.

(4) If H_1 is in fact true, the chance of reaching this conclusion is at most 5%.