Chapter 16, Fourth Edition : Inference in Practice

Note : Here the 4-th and 5-th editions of the text have different chapters, but the material is the same.

The previous few Chapters introduce the idea of statistical inference for a population mean. There were a number of assumptions about the experiment and population distribution. Some of these are *critical* in the sense that inference will be invalid without these or some related property. Other assumptions were not *critical* in the sense that inferences can be made even when these assumptions are violated, although the techniques that are needed and used may be different.

Recall we made some assumptions Some Very Simple Conditions for Inferences about a Population Mean

1. A simple random sample is obtained from a population. There is no non-response or other practical difficulties with the data
2. the variable we measure has exactly a normal distribution $N(\mu,\ \text{sd}\ = \sigma)$
3. we do not know μ, but we know σ.

Before discussing these we summarize what are our goals with respect to these.

Some Very Simple Conditions for Inferences about a Population Mean

1. A simple random sample is obtained from a population. There is no non-response or other practical difficulties with the data

   - The simple random sample property, or a variant of it, is crucial to remove selection bias and to obtaining the sampling distribution of the statistic to be used for our inferences.

     Variants are the different types of sampling methods, for example the two stage sampling designs that involve stratification (stratified random sample or probability proportional to size) or randomized block designs. There are also some special models and tools for time series, but these are outside the scope of this course.

- There is no non-response or other practical difficulties with the data.

  This has to be taken into account, as without some special care the inferences can be invalid. The basic reason is that the respondents may represent a subpopulation that is quite different in population characteristics than the population that is of actual interest in the study.

2. the variable we measure has exactly a normal distribution $N(\mu,\ \text{sd} \ = \sigma)$

   This property can be removed and changed to other types of population distributions. The tools are typically different than that studied in this course, but these are well known and understood.

3. we do not know μ, but we know σ.

   If we restrict our attention to normal populations, the assumption that σ is known can be changed to σ is unknown and also needs to be estimated. This is studied in this course and involves the use of the so called Student's t distribution, which is tabulated in Table C of this text.

   The assumption about normality can also be changed, as was indicated in item 2.

Our main tool for statistical inference about a population mean is obtained from the sampling distribution of $\bar{x}$. In the simple case studied earlier this is based on the sampling distribution of

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

Under the assumptions listed at the beginning of this section (from Chapter 11), and if μ_0 represents the true population mean then

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

will have a standard normal distribution. It is typically called the Z statistic or Z score.

Based on this sampling distribution we also obtain confidence intervals for population means which are of the form

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

The term

$$z^* \frac{\sigma}{\sqrt{n}}$$

is often called the margin of error (or accuracy) for our confidence interval for the population parameter μ.

Where did the data come from?

The procedure that one uses assumes the data comes from a random sample. The procedure cannot correct for this if the data is for example from a convenience sample. It is always important to consider or ask how the data was obtained.

The text states that if your data is not from a random sample (or appropriate variation) then your conclusions may be challenged. What this means is that your conclusions my not be reproducible. If someone were to attempt to reproduce your experimental results from an independent experiment or study, they would follow the same protocol and not obtain a similar result, and possibly a result that is contradictory.

One of the main ideas behind this is that without an appropriate data selection method such as random samples there may be substantial selection bias. In a technical sense this means that the sampling distribution of Z given by equation (1) does not have a standard normal distribution.

Comments on p-values

How small is small enough?

Two factors come into play for this

- How plausible is H_0?
- What are the consequences of *rejecting* the null hypothesis H_0?

  These consequences may be expenses in switching from a current system (drug or medical treatment) to switching machinery for a factory or other engineering consequences. It may also be ethical considerations such as the potential of actual harming patients if we use a too easy hurdle to introduce a new procedure or use to much of a hurdle so that patients will loose future benefits of an improved treatment.

These properties must be balanced, and so it is somewhat subjective as to what constitutes a small p-value.

Generally a p-value much bigger than 0.05 is *not small*, and so does not present evidence against a null hypothesis. A p-value around 0.05 is typically considered small. A p-value of 0.01 or less is often considered as*strong* evidence against the null hypothesis and in favour of the alternative hypothesis.

Type I and Type II errors

Consider the idea of hypothesis testing. The population is Normal, with mean μ and standard deviation $\sigma =$. Consider the simplest case

$$H_0 : \mu = 0 \text{ versus } H_a : \mu = \mu_a = \frac{1}{2}$$

The sample size is n. The test statistic is

$$Z = \frac{\bar{x} - 0}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 0}{\frac{1}{\sqrt{n}}} = \sqrt{n}\bar{x}$$

Let us also suppose that the test is at significance level $\alpha = 0.025$, so that the critical value is $z^* = 1.96$ (we find the critical value from the Normal Table A).

When the alternative hypothesis is true, the *true* population mean is not 0, but is $\frac{1}{2}$. Therefore the sampling distribution of Z is no longer standard normal. It can be found, but that is not so important for us here.

In the Figures that follow, there are two regions. One corresponds to the decision that we do not reject H_0, which happens if the observed value of the Z statistic is smaller than z^*, equal to 1.96 in our case. The other region, shaded grey, corresponds to the case when we reject H_0 in favour of the alternative H_a, that is we decide $\mu = \frac{1}{2}$ instead of deciding $\mu = 0$.

These Figures show the sampling distribution IF H_0 is true (always standard normal) and the sampling distribution of Z IF H_a were true. This is done for various sample sizes, $n = 4, 10, 20$ and 40.

Notice that the white area gets smaller and the grey area gets bigger as n increases.

Notice in these plots we can can calculate the probability of rejecting IF H_a were true. This is obtained as

$$\begin{aligned} P_a(\frac{\bar{x} - 0}{\frac{\sigma}{\sqrt{n}}} > 1.96) &= P_a(\bar{x} > 1.96 \times \frac{\sigma}{\sqrt{n}}) \\ &= P_a(\bar{x} - \frac{1}{2} > 1.96 \times \frac{\sigma}{\sqrt{n}} - \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} &= P_a(\frac{\bar{x}-\frac{1}{2}}{\frac{\sigma}{\sqrt{n}}} > 1.96 - \frac{\frac{1}{2}}{\frac{\sigma}{\sqrt{n}}}) \\ &= P(Z > 1.96 - \frac{\frac{1}{2}}{\frac{\sigma}{\sqrt{n}}}) \end{aligned}$$

This uses the property that IF H_a were true, then $\bar{x}$ still has a normal distribution, but it is just not the standard normal distribution. Under H_a, we have $\bar{x}$ has a normal distribution with mean $\mu_1 = \frac{1}{2}$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{n}}$.

For example when $n = 10$ we have

$$\begin{aligned} P_a(\frac{\bar{x}-0}{\frac{1}{\sqrt{10}}} > 1.96) &= P_a(\bar{x} > 1.96 \times \frac{1}{\sqrt{10}}) \\ &= P_a(\bar{x} > 0.6198) \\ &= P_a(\bar{x} - \frac{1}{2} > 0.6198 - \frac{1}{2}) \\ &= P_a(\bar{x} - \frac{1}{2} > 0.1198) \\ &= P_a(\frac{\bar{x}-\frac{1}{2}}{\frac{1}{\sqrt{10}}} > \frac{0.1198}{\frac{1}{\sqrt{10}}}) \\ &= P(Z > 0.3788) \\ &= 1 - 0.648 = 0.352 \end{aligned}$$

This property of calculating the probability of rejecting H_0 when H_a is true (that is equivalent to making the correct decision when H_a is true) is known as a *power* calculation. The *power* of a statistical test is the probability of making the correct decision when H_a is true.

Remark : Earlier we calculated the sample size n required to obtain a confidence interval with a prescribed level of precision. In many settings, in particular pharmaceutical and medical settings, a sample size is determined so as to meet or satisfy a pre-specified power requirement. This involves solving a non-linear equation. For example, for a set value of β, for example $\beta = 0.8$, we could solve for the unknown n (but not in this course)

$$\beta = P_a(\frac{\bar{x}-0}{\frac{\sigma}{\sqrt{n}}} > 1.96)$$

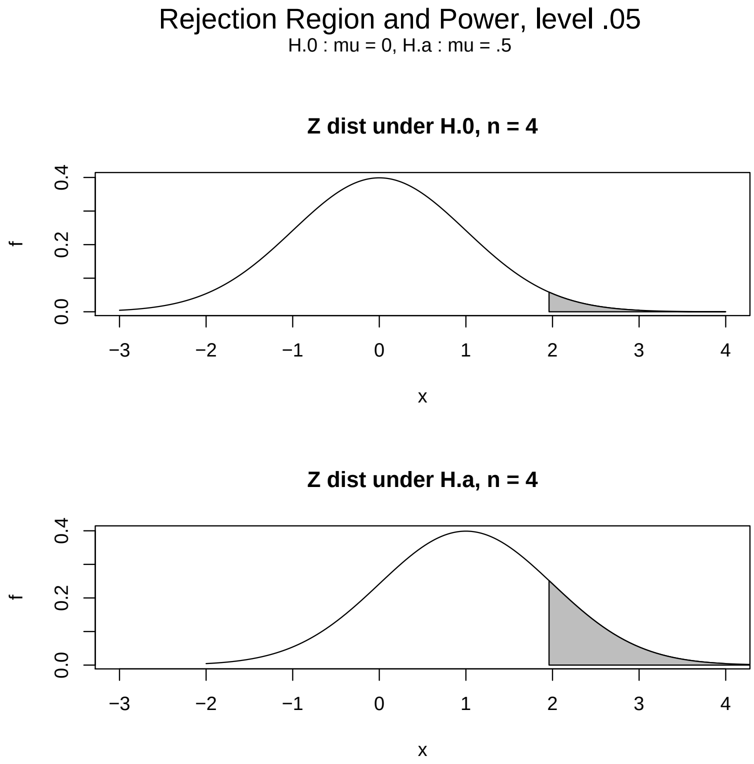

Figure 1: Power, n = 4

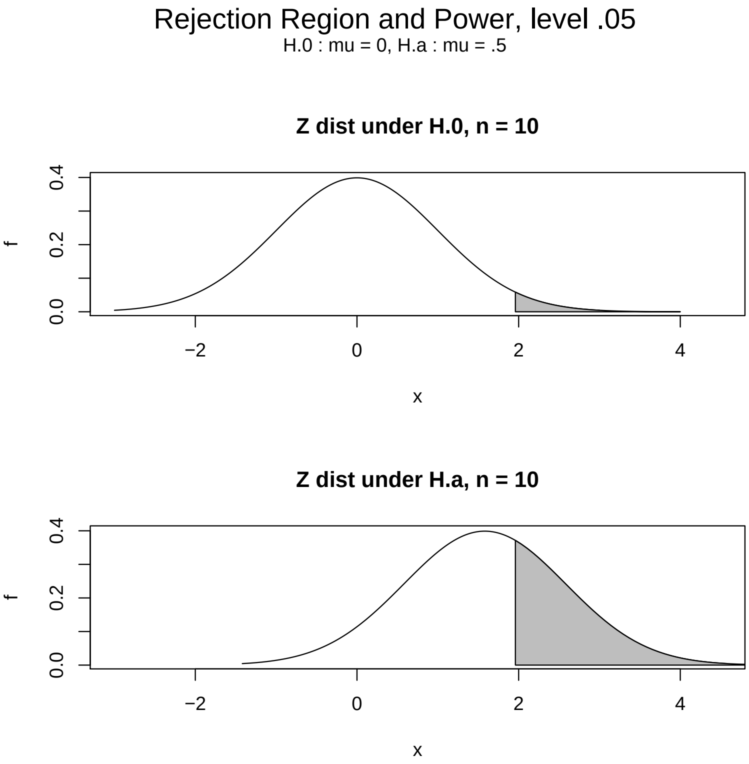

Figure 2: Power, n = 10

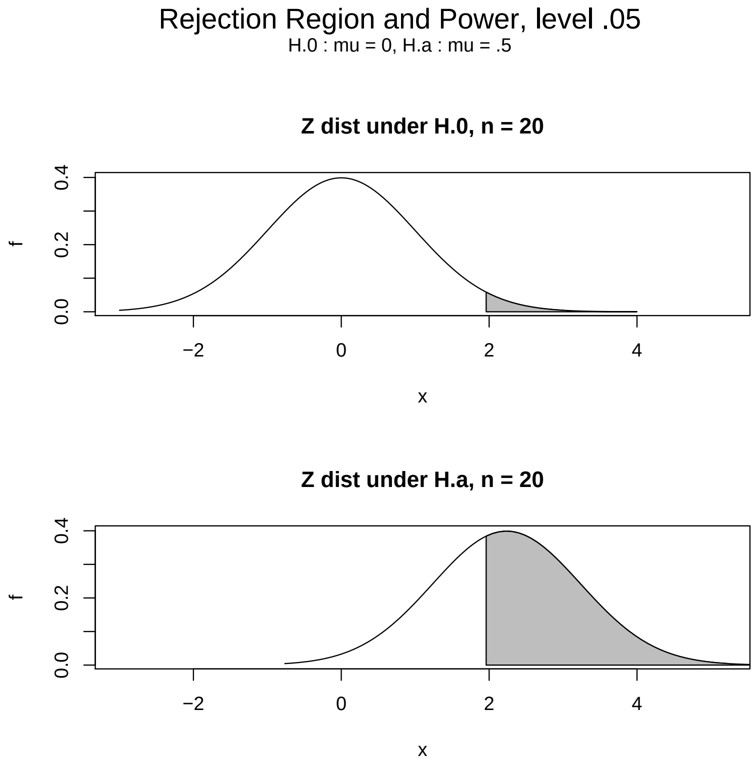

Figure 3: Power, n = 20

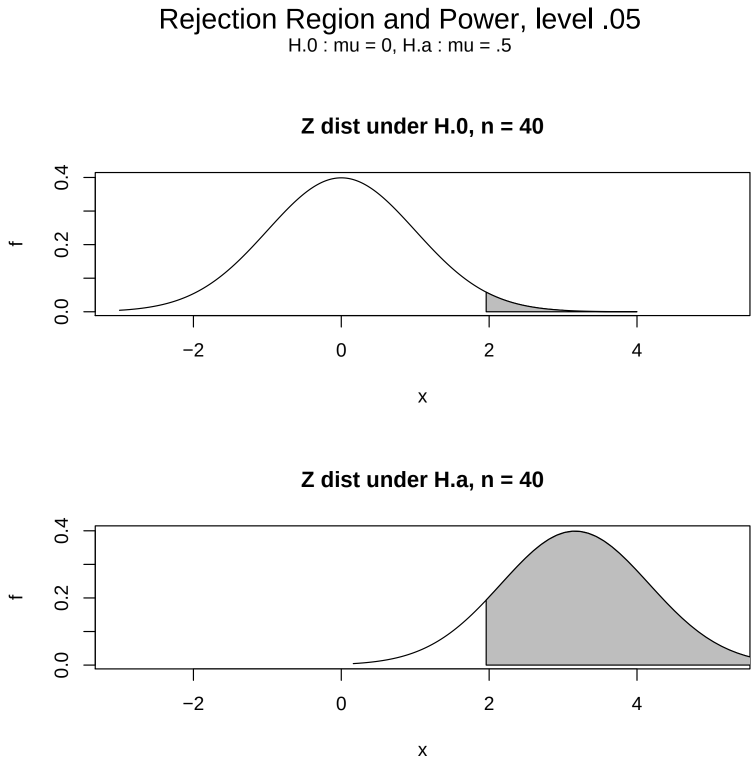

Figure 4: Power, n = 40

What do these Figures tell us?

There are various possibilities in our hypothesis testing.

1. H_0 is true ($\mu = 0$) and we decide H_0 is true. No error
2. H_0 is true ($\mu = 0$) and we decide H_a is true. Type I error
3. H_a is true ($\mu = \frac{1}{2}$) and we decide H_0 is true. Type II error
4. H_a is true ($\mu = \frac{1}{2}$) and we decide H_a is true. No error

In a table form these are

	decision	
true state (below)	decide in favour of H_0	decide in favour of H_a
H_0	no error	type I error
H_a	type II error	no error

Then the significance level then can be interpreted as the probability of rejecting H_0 when H_0 is true, that is

$$P(Z > 1) = 1 - P(Z \leq 1) = 1 - 0.842 = .158$$

Consider for example Figure 2. What would happen if we changed the critical value z^* to a value smaller than $z^* = 1.96$? For example suppose we change it to $z^* = 1$.

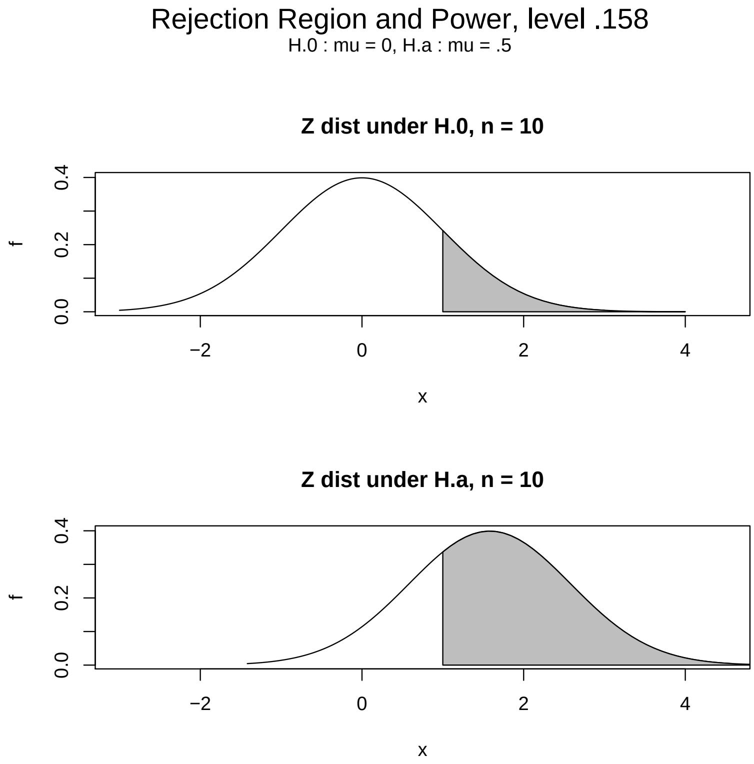

Figure 5: Power, n = 10

If on the other hand we increase the critical value from $z^* = 1.96$ we decrease the probability of Type I error, but at the same time we decrease the change of making the correct decision IF IN FACT H_a were true. In other words by decreasing Type I error, we increase Type II error, and vice versa. The two decision possibilities have to be balanced.

Remark : This is the same as in our legal analogy. If we set the requirement for demonstrating *guilty* to high, then innocent people do not get convicted (good) but at the same time the guilty as not convicted as often as the should be (bad). On the other hand if we the threshold for demonstrating guilt too low then innocent people at convicted too often (bad) but the guilty are convicted more often (good).

Remark : Again consider a medical or pharmaceutical setting. We have an existing drug treatment, that is good. A new drug therapy is being considered. Should it be used and replace the current one? Heree we need to trade off two very important features. H_0 will play the role of the *current drug therapy is as good (or better than) as the new treatment.* H_a will play the role of *the new drug treatment is strictly better than the current drug treatment.* Of course in our statistical problem we need to translate this into a property in terms of the parameters of the statistical model. For our purposes we interpret this as the population mean μ_0 = mean effectiveness of patients treated with current drug treatment is better than the population mean μ_a = mean effectiveness of patients treated with new or proposed drug treatment.

How should our decision procedure behave, at least in terms of *controlling* type I and type II errors? We do not want to replace the current good treatment if the new or proposed drug treatment or therapy is *not better*, so we want this probability of rejecting H_0 in favour of H_a to be small. On the other hand, *if the new treatment is better than the current treatment* we do want to replace the current drug treatment with new one; that is we want to reject H_0 in favour of H_a. This means that we want to accept H_a with large probability *when H_a is true*, that is have *large power.* This is equivalent to having a small probability of type II error when H_a is true. Here is where the tradeoff come in to play. If we make the cut off or critical value of our statistical test so that type I error is small, it forces the type II error probability to be large. If we make the probability of type II error small, then we are forced to make the probability of type I error large.