Hypothesis testing

Another important use of sampling distributions is to test hypotheses about population parameters, e.g. mean, proportion, regression coefficients, etc. For example, it is possible to stipulate that the population mean is equal to some specified value and then use sample information to decide whether the hypothetical value can be rejected or not in the light of sample evidence. The decision will depend on (1) the size of the difference between the hypothetical population mean and the sample mean, (2) the size of the sampling error associated with the sample mean, and (3) the degree of certainty the decision-maker requires before rejecting the initial hypothesis.

Null and alternative hypotheses

First we set up what is known as the null hypothesis, H_0, about the population parameter, e.g. we may claim that the population mean μ is equal to some value μ_0, say. This is usually written as $H_0: \mu = \mu_0$. We then stipulate an alternative hypothesis, H_1, which may state, e.g., that the population mean is not equal to μ_0, $H_1: \mu \neq \mu_0$. The purpose of hypothesis testing is to see if we have sufficient evidence to reject the null hypothesis.

Typically, the null hypothesis says that there is nothing unusual or important about the data we are considering; for example, if we were looking at the average test scores of children who have received a particular teaching method, the null hypothesis would be that the mean is equal to the national average. If we are testing a new drug, and are looking at the proportion of people taking the drug whose condition improves, we would take as our null the proportion who improve with a placebo, or with a previous drug. If we are looking for a relationship between two variables, the null hypothesis is usually that there is no relationship, that is that the regression coefficient between them is 0.

The alternative hypothesis is thus that there is something interesting or different about the population – for example that the average test score from the new teaching method is not equal to the national average, or that the proportion who improve with the new drug is not equal to the previous rate, or that there is a relationship between the two variables, so that the regression coefficient is not equal to 0.

We treat H_0 as our "default position", and we usually require quite strong evidence to reject the null hypothesis – typically 90%, 95% or 99%, depending on the context.

Test statistic

Having set up our null and alternative hypotheses, we look for a suitable test statistic that will give us evidence for or against the two hypotheses. For example, if we are looking for evidence about the population mean (H_0:$\mu=\mu_0$ vs. H_1:$\mu\neq\mu_0$), we will most likely use a statistic based on the sample mean, $\overline{X}$. From our work in section 4, a suitable statistic (assuming we now the standard deviation σ of the population) is

$Z = \dfrac{\overline{X} - \mu_0}{(\sigma / \sqrt{n})}$ - that is, we measure $\overline{X}$ -μ_0 in terms of the Standard Error of $\overline{X}$ as an estimator for μ, which is equal to $\sigma/\sqrt{n}$. For large samples, $n\geqslant 30$, we know that the distribution of $\overline{X}$ is normal, so that Z will be a standard normal variable, that is $Z\rightarrow N(0,1)$.

The larger is ($\overline{X}$ -μ_0), the bigger is Z, and the less credible it is that H_0 is correct. So essentially what we are trying to do is to measure whether the sample mean, $\overline{X}$, is significantly different from μ_0.

Decision rule

We now have to decide how large Z must be for us to reject H_0. This is related to the risk we are prepared to take of an incorrect decision. In deciding whether to accept or reject a null hypothesis, there are two types of error we may make:

A Type 1 error is to reject the null hypothesis when it is correct.

A Type 2 error is to accept the null hypothesis when it is incorrect.

We usually specify our decision rule in terms of the probability of a type 1 error we are prepared to accept, denoted α. Depending on α, we can calculate critical values of the test statistic Z, so that if Z lies beyond the critical values, we reject H_0, while if Z lies within the critical values, we accept H_0.

Thus, in the case of the population mean, if our acceptable level of Type 1 error is $\alpha=0.05$, then the critical values of the test statistic will be

$Z=\pm 1.96$, since we know from section 4 that, *if H_0 is true and $\mu=\mu_0$*, then $P(-1.96<Z<1.96)=0.95$. Hence we know that, if $\mu=\mu_0$, there would be a less than 5% probability of obtaining a value of greater than 1.96 or less than -1.96, so that the probability of a type 1 error in rejecting H_0 is less than 5%. If we obtain a value of Z between the critical values, we conclude that we do not have sufficient evidence to reject H_0, so we accept it.

The acceptable probability of Type 1 error is also called the significance level of the test. If, say, $\alpha=5\%$, and we reject H_0, we will say that we reject H_0 at the 5% level of significance, or that $\bar{X}$ is significantly different from μ_0 at the 5% level of significance, etc.

Thus, we set up our decision rule to give H_0 the "benefit of the doubt". We require 95% confidence to reject it. Note again that if we reject the null hypothesis, we are not saying "there is a 95% probability that $\mu\neq\mu_0$". μ is a constant which either is equal to μ_0 or it isn't. What we are saying is that, if μ *were* equal to μ_0, there would be a 95% chance of obtaining a test statistic between the critical values. Only 5% of the time would we obtain a value for Z that would lead us to reject H_0. Hence

$P(\text{Reject } H_0|H_0 \text{ true})\leqslant 0.05$.

Note that if we were prepared to accept a Type 1 error probability of 10%, we would set our critical values at $Z=\pm 1.645$, while if we were only prepared to accept a 1% Type 1 error, we would set critical values of $Z=\pm 2.58$.

Power of a test

The power of a hypothesis test is the probability β of a Type 2 error. Given two tests of a hypothesis H_0, we say that one test is more powerful than the other if, *given a specified level of Type 1 error*, it has a lower probability of Type 2 error.

Example

Suppose we know that average household income in the population is £300 p.w., with standard deviation £50 per week. We are trying to see whether households in a particular town have a higher or lower average income. We take a random sample of 100 households in the town, and find an average income of £285 p.w. We wish to test the hypothesis that

average household income in the town is equal to the national average, with a 5% level of significance.

Here H_0 is μ=£300, and H_1 is $\mu \neq$£300.

Our test statistic is $Z=\frac{\overline{X}-\mu_0}{(\sigma/\sqrt{n})}$, with μ_0=300, σ=50, and n=100. From the sample, $\overline{X}$ =285. Hence, Z=(285-300)/(50/√100) = -15/5 = -3.

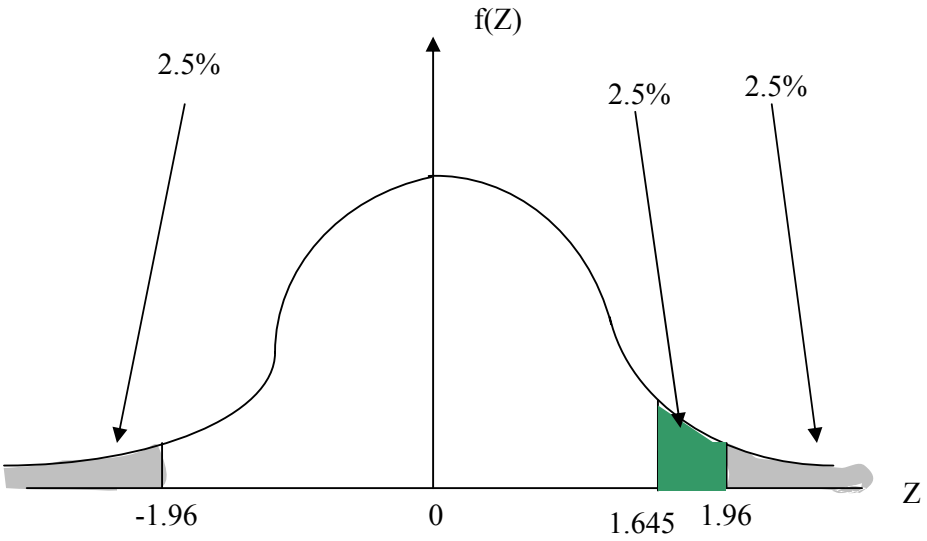
Given a 5% significance level, the critical values of the Z statistic are ±1.96. Our decision rule is to accept H_0 if -1.96<Z<1.96, and reject H_0 otherwise. Hence, we reject H_0, and conclude that $\mu \neq$£300. In fact, we may conclude that the average household income in this town is significantly less than the national average, at the 5% (or indeed at the 1%) level of significance.

Two-tailed and one-tailed tests

The example above involved a two-tailed test of significance – that is, we were trying to see if $\overline{X}$ was significantly higher *or* significantly lower than μ_0. That is, H_1 was specified as $\mu \neq \mu_0$. In a one-tailed test, the alternative hypothesis is H_1:$\mu > \mu_0$, or H$\mu < \mu_0$. This would be appropriate if we had some *a priori* reason to believe that we were likely to find a difference in a particular direction. For example, if we were trying to see if graduates have the same income as the rest of the population, we might use a 1-tailed test, as we would naturally assume that graduates tend to enjoy a higher income, so H_1 would be that $\mu > \mu_0$, where μ is graduate average income, and μ_0 is the average for the whole population.

When we use a 1-tailed test, the critical value of Z is different. For example, at the 5% level of significance, we would use a critical value for Z of 1.645, instead of ±1.96, since $P(Z>1.645|H_0)=5\%$. (Hence ±1.645 as the 10% critical value for a 2-tailed test, since $P(Z<-1.645|H_0)$ is also 5%, so we have 5% in each 'tail'.) If our alternative hypothesis were $\mu < \mu_0$, then our critical value would be Z=-1.645, rejecting H_0 if Z falls below this.

1-tailed vs. Two-tailed test



Proportions

The procedure and rationale for testing hypotheses about population proportions are similar to those used for means. They are based on the normal distribution and apply to large samples, n≥30. The null hypothesis is specified in terms of the population proportion P, and the sample proportion, p, and the standard error, SE(p)=(√P(1-P))/n are used in the test statistic. For example, suppose we wish to test the null hypothesis that the proportion of households in a certain town with at least one wage-earner is 0.85. We have a random sample of 100 households, and the proportion of the sample with at least one wage-earner is p=0.81. We have

H_0: $P=P_0=0.85$ H_1:$P\neq 0.85$.

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{.81 - .85}{\sqrt{\frac{.85 * .15}{100}}} = -.04/.0357 = -1.120.$$

Note that we use the standard error calculated from the population proportion based on the null hypothesis – this is because we are trying to say "If the null hypothesis were true, how likely would it be to get this

much difference between the sample proportion and population proportion?". So we consider the probability distribution of the test statistic that would apply if the null hypothesis were true.

As 1.120<1.96, the r% level of significance 2-tailed critical value of the Z statistic, we cannot reject H_0, in other words the sample proportion is not *significantly* different from 0.85 (at the 5% level). We therefore accept H_0.

Difference between two sample means

So far we have made inferences on a single sample. Now we shall make inferences from two samples. Typically we shall have two random samples from two populations and we shall be making inferences about the differences between the means of the two populations using the difference between the two sample means. For example, we may be interested in testing whether boys are achieving significantly different results in school than girls. To be able to answer such a question, we first need to study the sampling distribution of the difference between two sample means.

If a random sample of size n_1 is taken from one population with mean μ_1 and variance σ_1^2, and another random sample of size n_2 is taken from another population with mean μ_2 and variance σ_2^2, the difference between the two sample means is defined as

$d=(\overline{X}_1 - \overline{X}_2)$

where $\overline{X}_1$ and $\overline{X}_2$ are independent random variables because they will not vary from one set of two samples to another, and because changes in $\overline{X}_1$ are not influenced by changes in $\overline{X}_2$ and vice-versa.

$E(d) = E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2 = D.$

i.e. the sample difference (d) is an unbiased estimator of the population difference D.

$Var(d) = Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2) = (\sigma_1^2/n_1) + (\sigma+2^2/n_2)$

Since $\overline{X}_1$ and $\overline{X}_2$ are independent.

The standard error of d is given by SE(d)=$\sqrt{\frac{{\sigma_1}^2}{n}+\frac{{\sigma_2}^2}{n}}$ and shows that the larger are the two variances and the smaller the sample sizes, the larger will be the sampling error of d.

If X_1 and X_2 are normally distributed, then $\overline{X}_1$ and $\overline{X}_2$ are also normally distributed. Also, if both samples are large ($n_1,n_2 \geqslant 30$), then even if X and X_2 are not normally distributed, the Central Limit Theorem ensures that $\overline{X}_1$ and $\overline{X}_2$ will be approximately normally distributed. If either of these is true, then d will also be normally distributed, as the difference between two normal variables. Thus,

$$d=(\overline{X}_1-\overline{X}_2)\rightarrow N[(\mu_1-\mu_2),\ \frac{{\sigma_1}^2}{n}+\frac{{\sigma_2}^2}{n}]$$

The confidence interval for the difference between the population means can now be easily calculated. The 95% confidence interval is

$$(\mu_1-\mu_2)=(\overline{X}_1-\overline{X}_2)\pm 1.96\sqrt{\frac{{\sigma_1}^2}{n}+\frac{{\sigma_2}^2}{n}}$$

The calculated confidence interval will contain the true population difference in 95% of samples.

Hence, the <u>hypothesis test</u> for the population difference can also be performed in the usual manner. Let H_0: $\mu_1-\mu_2=0$, and H_1:$\mu_1-\mu_2\neq 0$. The test statistic is

$$Z=\frac{(\overline{X}_1-\overline{X}_2)-0}{\sqrt{\frac{{\sigma_1}^2}{n}+\frac{{\sigma_2}^2}{n}}},$$

and the decision rule, for a 5% significance level, will be to reject H_0 if $|Z| \geqslant 1.96$, otherwise accept H_0.

<u>Example</u>

A school wants to find out if there is a difference in test performance between boys and girls. A sample of test scores of 60 boys and 50 girls is

examined. It is found that the boys have sample mean $\overline{X}_1$=54 with standard deviation 14, and the girls have sample mean $\overline{X}_2$=60, with standard deviation 9. *NB: we shall ignore for now the problem of estimating the* population *standard deviations, and assume these figures are correct.*

We set up H_0: $\overline{X}_1 - \overline{X}_2 = 0$ H_1: $\overline{X}_1 - \overline{X}_2 \neq 0$.

Our test statistic is

$$\frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\frac{{\sigma_1}^2}{n} + \frac{{\sigma_2}^2}{n}}} = \frac{54 - 60}{\sqrt{\frac{14^2}{60} + \frac{9^2}{50}}} = -6/\sqrt{(4.68)} = -1.28.$$

As usual, for a 5% level of significance on a two-tailed test, our critical value for Z is ±1.96, so we do not have sufficient evidence to reject the null hypothesis. Girls are doing better, but not *significantly* better.

Difference between two sample proportions

This can be tested in a similar manner.

Exercise

Two different teaching methods are tried with different groups of students on the same course. In the first group, 47 out of 63 students pass. In the second group, 66 out of 78 pass. The department wants to work out whether one teaching method is significantly better than the other. Formulate suitable null and alternative hypotheses, and calculate a suitable test statistic, to test this.