Hypothesis Testing

Florian Oswald

October 3, 2013

Contents

1 Preface 2

2 Purpose 2

3 Setup 2

4 Procedure 3
- 4.1 Errors . . . 3

5 Examples 4
- 5.1 Defective Items: Testing a population proportion . . . 4
- 5.2 One-sided Hypothesis test about a population proportion . . . 5
- 5.3 Two-sided test about a proportion . . . 8
- 5.4 Confidence interval for a population proportion . . . 9
- 5.5 Type 2 Error and Power . . . 10
- 5.6 P-Value . . . 12

6 Tests in small samples 13
- 6.1 Test about population mean with unkown variance . . . 13
- 6.2 Test about population mean, population variance, and p-value . . . 15
  - 6.2.1 Testing the mean . . . 15
  - 6.2.2 P-value for test . . . 16
  - 6.2.3 Confidence interval for the population variance . . . 16
- 6.3 Testing the difference of two population means . . . 17
  - 6.3.1 $\sigma_X^2 \neq \sigma_Y^2$: unequal variances . . . 18
  - 6.3.2 $\sigma_X^2 = \sigma_Y^2$. equal variances. . . . 19

1 Preface

This is a hands-on tutorial about cassical hypothesis testing. It does not contain any rigorous proofs. Simple calculus is used throughout. We provide plenty of graphical illustrations and fully worked examples. Some examples use the statistical software R. Comments and suggestions are welcome.[1]

2 Purpose

With a sample of data at hand, we can compute statistics which summarize certain features of the data (such as the data's "center" (i.e. mean or median), or it's spread). Our aim is often to learn something about a population parameter, for which our statistic is an estimator (for example the sample mean is an estimator for the population mean). Hypothesis testing is a tool that allows us to assess whether a given population parameter lies in a certain range of values or not, given the information we obtained from our sample. This method is based on the sampling distributions of the statistics of interest, thus we make probabilistic statements. We acknowledge that our results might be wrong with a certain – and known – probability.

3 Setup

We know from previous sessions that an estimator θ (e.g. the sample mean) is a random variable. This becomes clear if we think of obtaining θ from a random sample of data $\{X_i\}_{i=1}^n$, for example by writing explicitly that $\theta = g(X_1, \ldots, X_n)$. Given this insight, we then also know that θ will have an associated probability distribution. Knowledge of this pdf will allow us to test hypothesis. See for example figure 1. Of course you recognize immediately that this pdf has the characteristic shape of a normal distribution. This is no coincidence: most of the times when testing hypothesis we either sample directly from a normal population distribution (in which case θ is normal), or we refer to results from large sample theory (such as the central limit theorem) to say that the sampling distribution of θ is *approximately* normal.

[1] f.oswald@ucl.ac.uk

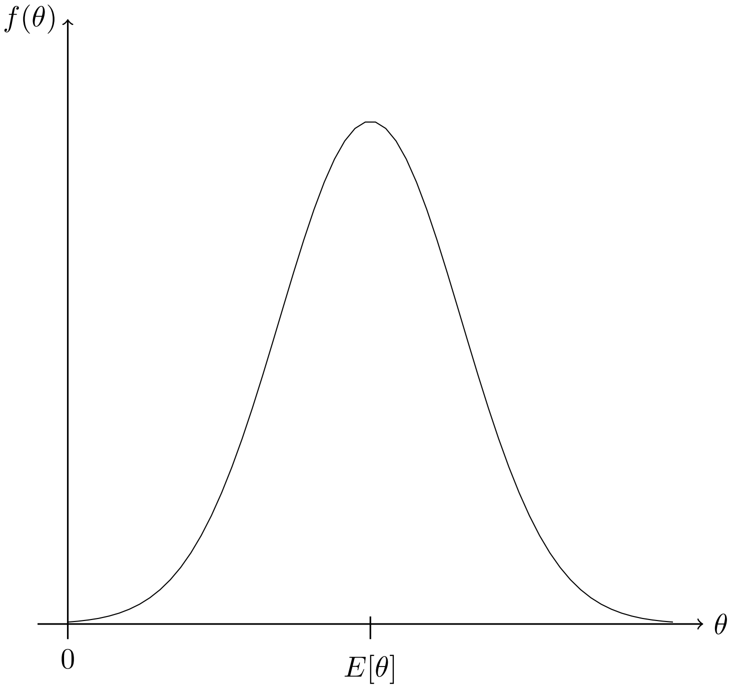

Figure 1: example of the sampling distribution of estimator θ

4 Procedure

Here is a recipe for a statistical test:

1. formulate hypothesis and specify significance level
2. define a suitable test statistic and the sampling distribution of this statistic under the assumption that H0 is true
3. define the rejection region, i.e. define what is meant by "far away from hypothesized value if H0 is true"
4. calculate the test statistic from a sample of data
5. decide

4.1 Errors

Let's repeat that with this approach we can be either right or wrong. In particular, we can make two types of error:

1. We make a mistake if we **reject** H0, even though H0 is **true**
2. We make a mistake if we **do not reject** H0, despite H0 being **false**

The first of these is accordingly called "type 1" error, or α error, the second on is called "type 2", or β error.
This is best done by example, so here goes:

5 Examples

5.1 Defective Items: Testing a population proportion

Suppose we deal with a supplier who claims that in his delivery comprising $N = 100.000$ lightbulbs, exactly 10% are defective. If we let θ denote the proportion of faulty items, he claims that $\theta = 0.1$. Suppose we have reason to doubt this claim, and state instead our believe that there are 20% defective items. In terms of hypothesis testing, we have two competing hypothesis:

- H0: $\theta = 0.1$
- HA: $\theta = 0.2$

where "A" is for alternative. Notice that there are several ways to express an alternative hypothesis, distinguishing between one-sided and two-sided tests, such as HA: $\theta > 0.1, \theta < 0.1, \theta \neq 0.1$.
The way this works is now to take a sample from the delivery we obtained, compute an estimator for θ, say $\hat{\theta}$, and decide whether or not we reject H0 based on this value. Our rationale is to reject the hypothesis if we observe a value for the estimator which is "far away" (in some sense) from the value we expected under H0. Of course the only way to be 100% certain would be to examine every single lightbulb, but this is costly in terms of time and effort. Obviously, this is the main reason for drawing samples to learn about populations in the first place. Let's remind ourselves that we can make two types of error:

1. Type 1 error (or α error). Imagine that $\theta = 0.1$, i.e. H0 is true. If we reject H0, we make a mistake. Based on the level of significance α we choose, we define a *rejection region*. If our sample produces a value for the estimator that falls inside this region, we reject H0 (based on the reasoning that such a value is "far away" from the hypothesized value. Too far away to qualify as sample variation at any rate. [any sample will produce a different value for the estimator.]) This is illustrated in figure 2. Given that we set a significance level α, which determines the size of this region, we can directly influence the size of this error:
$$\Pr(\text{type 1 error}) = \Pr(\text{reject H0}|\text{H0 true}) = \alpha$$
2. Type 2 error (or β error). Imagine $\theta = 0.2$, i.e. H0 is false. If we fail to reject H0, we make again a mistake. We would write
$$\Pr(\text{type 2 error}) = \Pr(\text{don't reject H0}|\text{H0 false})$$

Our aim will be to minimize the impact of both errors.

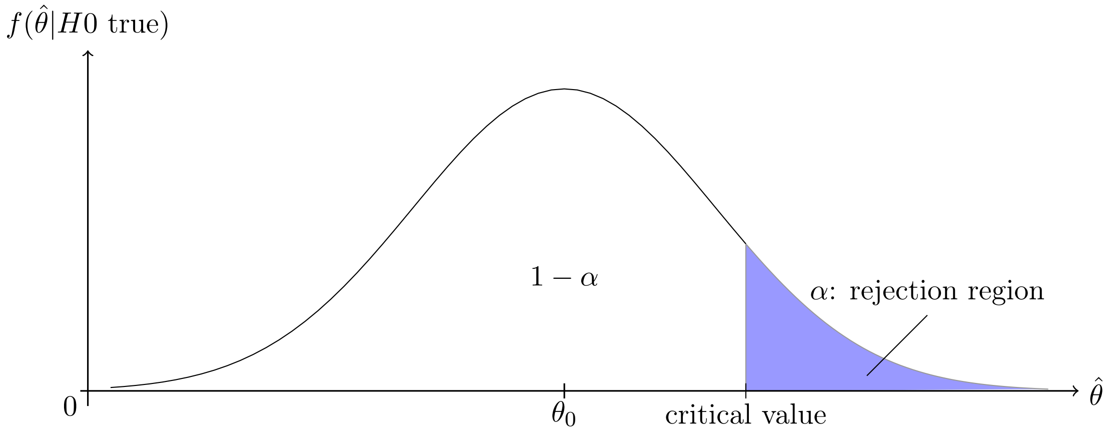

Figure 2: Hypothesis test setup. Notice how the pdf is specified specifically for the case that H_0 is true, i.e. we plot $f(\hat{\theta}|\theta_0)$. Under this assumption this pdf is indeed normal with mean θ_0. Under a different hypothesis H_1, this distribution would look different.

5.2 One-sided Hypothesis test about a population proportion

Let's continue with the above example.

Suppose we took a sample from our supplier's delivery with $n = 100$ and we found $Y = 13$ faulty items. At a level of significance $\alpha = 0.05$, does this mean that his claim that $\theta = 0.1$ is refuted? Well, we need to test this.

1. **formulate hypothesis** and specify significance level:

   H0: $\theta = \theta_0 = 0.1$

   HA: $\theta > \theta_0 = 0.1$

2. **define a suitable test statistic** and the sampling distribution of this statistic under the assumption that H0 is true:

   If H0 is true, then the population distribution is a Bernoulli with parameter θ_0, i.e. we have N RVs X who take two values, one for faulty and zero for not, and the probability that each of them is faulty is θ_0. Then $E(X) = Pr(X = 1)1 + Pr(X = 0)0 = \theta_0$. Our estimator is $\hat{\theta} = \frac{Y}{n} = 0.13$, with

$$\begin{aligned} E(\hat{\theta}) &= E(\frac{Y}{n}) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \theta_0 \end{aligned}$$

and

$$\begin{aligned} Var(\hat{\theta}) &= Var\left(\frac{Y}{n}\right) \\ &= \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}n\theta_0(1-\theta_0) \\ &= \frac{1}{n}\theta_0(1-\theta_0) \end{aligned}$$

Note that this should actually be $\frac{1}{n}\theta_0(1-\theta_0)\frac{N-n}{N-1}$, with N equal to the total number of items in the delivery. This is because we do sampling without replacement. This implies that we do not use the Binomial (for the sum of n Bernoullis), but the hypergeometric distribution. This distribution approximates the normal fairly well for values such that $n\theta_0(1-\theta_0) \geq 9$. Also, we assume that N is very big, so we can disregard the correction factor $\frac{N-n}{N-1}$. Therefore we get

$$\begin{aligned} Var(\hat{\theta}) &= \frac{\theta_0(1-\theta_0)}{n} \\ &= \frac{0.1(0.9)}{100} \\ &= 0.0009 \end{aligned}$$

and

$$\sigma_0 = 0.03$$

Next we need a test statistic: we will use

$$Z = \frac{\hat{\theta}-\theta_0}{\sigma_0} = \frac{\hat{\theta}-\theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \sim N(0,1)$$

and our sampling distribution is the standard normal. This is based on the above consideration that the Binomial (or hypergeometric) is approximated well by the normal distribution.

3. **Define the critical region**

   For our one-sided test, we want to find value c such that the probability under the pdf of the test statistic (in this case, the standard normal $\phi()$) is equal to our chosen level of significance, i.e. α. In other words, find c such that

$$\begin{aligned}
\int_c^\infty \phi(x)dx &= \alpha \\
1-\Phi(c) &= \alpha \\
\Phi(c) &= 1-\alpha \\
c &= \Phi^{-1}(1-\alpha) = \Phi^{-1}(0.95) = 1.645
\end{aligned}$$

So for $\alpha = 0.05$ the cutoff is the 0.95 quantile of the standard normal cdf, 1.645. Notice that we can also think in terms of a critical value of the estimator, $\hat{\theta}_c$, beyond which we reject. This follows from the relationship

$$\begin{aligned}
c &= \frac{\hat{\theta}_c - \theta_0}{\sigma_0} \\
\hat{\theta}_c &= \theta_0 + c\sigma_0 \\
&= 0.1 + 1.645 \cdot 0.03 \\
&= 0.1494
\end{aligned}$$

You'll recognize that this looks very similar to the upper bound of a confidence interval around our estimator. If we were constructing a two-sided test, then this equivalence would be exact, i.e. we would get two critical values, say $\hat{\theta}_{\underline{c}}, \hat{\theta}_{\bar{c}}$ and they would correspond exactly to the bounds of a confidence interval for the population parameter θ_0.

4. **calculate the test statistic**

$$\begin{aligned}
z &= \frac{\hat{\theta} - \theta_0}{\sigma_0} \\
&= \frac{0.13 - 0.1}{0.03} \\
&= 1
\end{aligned}$$

Thus, we get a value for the test statistic outside the rejection region, which is the region $z = (1.645, \infty)$. The sample at hand does not provide strong enough evidence for us to reject the null hypothesis that 0.1 is the true proportion of faulty items.
Again, relating to the critical value of the test statistic calculated before, $\hat{\theta}_c = 0.1494$, in order to reject the null we would have had to get an estimate $\hat{\theta}' > \hat{\theta}_c$.

5. **Decide**: we cannot reject H0.

Have a look at picture 3 to get some intuition. The shaded area is the rejection region.

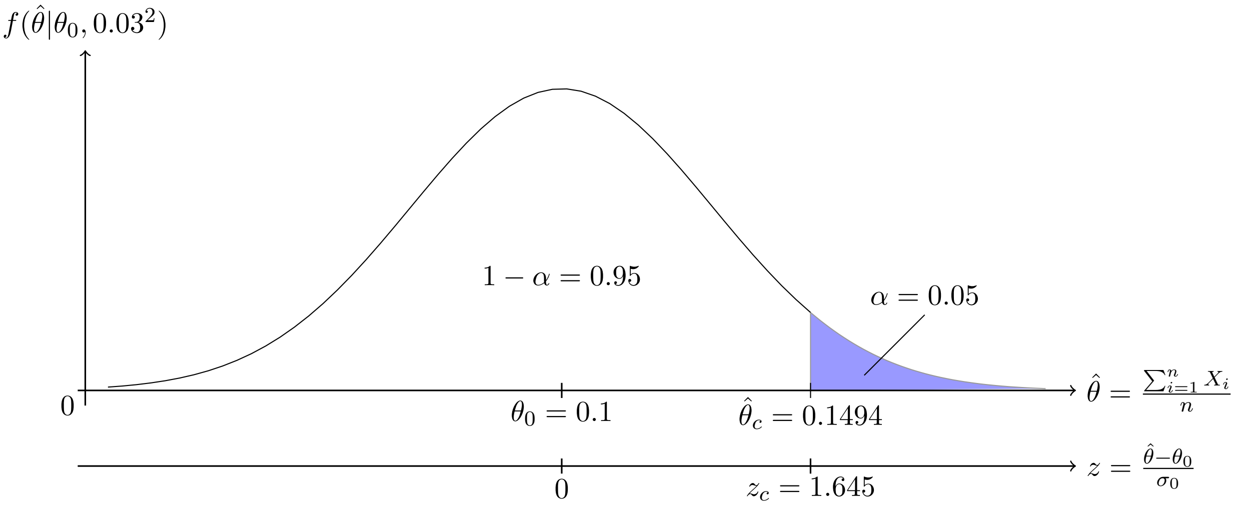

Figure 3: A one-sided test. Notice how we draw two x-axis to make the point that we work with *observed* values of the estimator, on the axis labelled $\hat{\theta}$, with mean $E[\theta]$, and with corresponding standardized values, i.e. axis labelled z and with mean 0. Also note again that the pdf is specified specifically for the case that H_0 is true, i.e. we plot $f(\hat{\theta}|\theta_0)$.

5.3 Two-sided test about a proportion

If we were to apply a two sided test, we would change our approach as follows:

1. H0: $\theta_0 = 0.1$
   HA: $\theta_0 \neq 0.1$

2. same test statististic

3. critical values: Choose c such that the are under the pdf of the test statistic is $1-\alpha$:

$$\begin{aligned}\int_{-c}^{c} \phi(x)dx &= 1-\alpha \qquad (1)\\ \Phi(c) - \Phi(-c) &= 1-\alpha \\ \Phi(c) - [1-\Phi(c)] &= 1-\alpha \\ 2\Phi(c) &= 2-\alpha \\ \Phi(c) &= 1.95/2 = 0.975 \\ c &= \Phi^{-1}(0.975) = 1.96\end{aligned}$$

such that $c = \pm 1.96$. Again, in terms of critical values of the estimator, we get

$$\begin{aligned}\hat{\theta}_{\underline{c}} &= \theta_0 - \sigma_0 c = 0.0412 \\ \hat{\theta}_{\overline{c}} &= \theta_0 + \sigma_0 c = 0.1588\end{aligned}$$

4. same value for test statistic: $z = 1$. Notice that sometimes we get a negative value, so it may be useful to check whether
$$|z| > |c|$$
5. same result: cannot reject H0

Some intuition again is provided in figure 4.

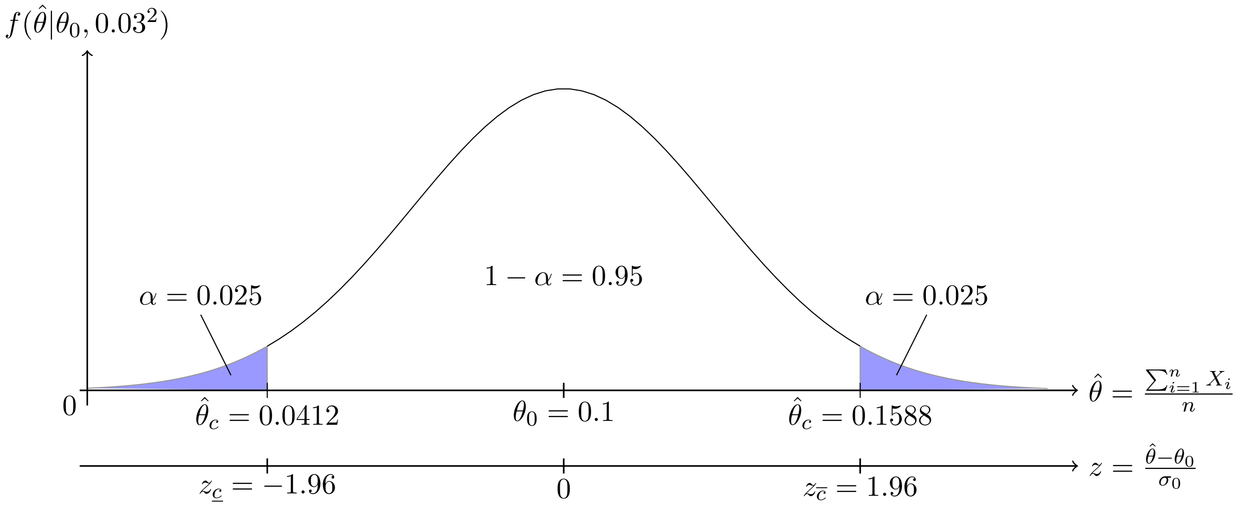

Figure 4: twosided test

Comparing both figures, nothing except the rejection regions changed. In particular, the sampling distribution is identical.

5.4 Confidence interval for a population proportion

Even though we present the confidence interval for a proportion, nothing substantial changes if the confidence interval is for another parameter like the population mean, variance etc. The standard errors of the according estimators need to be adjusted of course. In general, a $(1-\alpha) \times 100\%$ confidence interval of a population parameter specifies a region of the real line for which we say "the interval contains the true parameter with $(1-\alpha) \times 100\%$ probability". (for most of our cases parameters have only one dimension; for higher dimensions we have *confidence regions* in $\mathbb{R}^m$.) In figure 5 this region is shaded blue, and the interval thus described satisfies

$$\Pr\left[\theta - 1.96\sigma_{\hat{\theta}} \leq \hat{\theta} \leq \theta + 1.96\sigma_{\hat{\theta}}\right] \quad = \quad 1 - \alpha$$

in other words, the probability that the estimator lies in an interval $\pm 1.96\sigma_{\hat{\theta}}$ around the true value θ is $1-\alpha$. Notice that the values $\pm 1.96\sigma_{\hat{\theta}}$ correspond exactly to the values computed for c in expression (1).

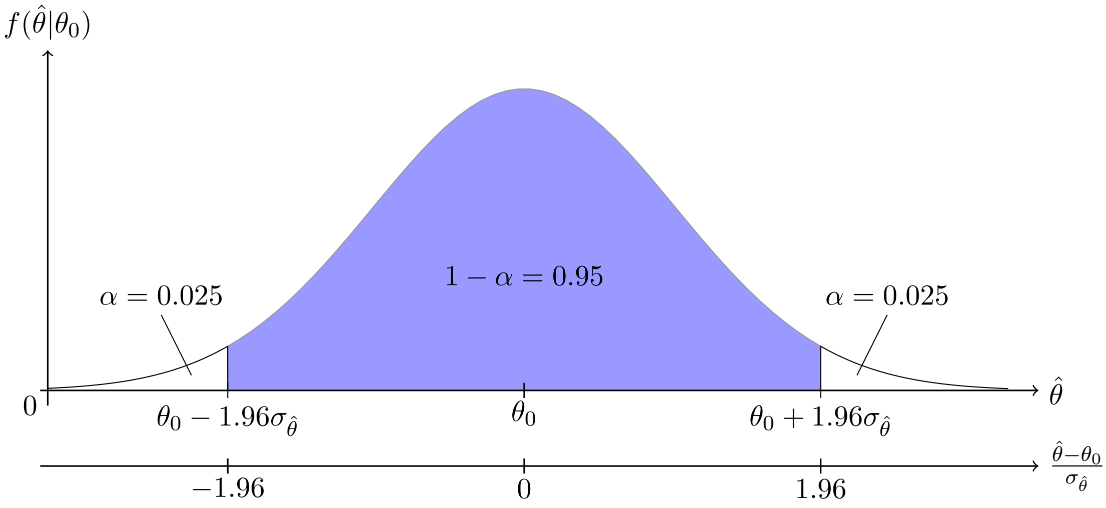

Figure 5: confidence interval

Let us denote the values ±1.96 as “critical values” c. Then some simple manipulations yield the more common form of writing the interval:

$$\begin{aligned}
\Pr\left[\theta - c\sigma_{\hat{\theta}} \leq \hat{\theta} \leq \theta + c\sigma_{\hat{\theta}}\right] &= 1-\alpha \\
\Pr\left[-c\sigma_{\hat{\theta}} \leq \hat{\theta} - \theta \leq +c\sigma_{\hat{\theta}}\right] &= 1-\alpha \\
\Pr\left[c\sigma_{\hat{\theta}} \geq \theta - \hat{\theta} \geq -c\sigma_{\hat{\theta}}\right] &= 1-\alpha \\
\Pr\left[\hat{\theta} + c\sigma_{\hat{\theta}} \geq \theta \geq \hat{\theta} - c\sigma_{\hat{\theta}}\right] &= 1-\alpha
\end{aligned}$$

and therefore

$$CI_\alpha = \left[\hat{\theta} - c\sigma_{\hat{\theta}}, \hat{\theta} + c\sigma_{\hat{\theta}}\right]$$

Notice that this is the exact same region as the one described in equation (1) above. In fact, the 95% confidence interval for the population proportion is described by $\left[\hat{\theta}_{\underline{c}}, \hat{\theta}_{\overline{c}}\right]$:

$$\begin{aligned}
\hat{\theta}_{\underline{c}} &= \theta_0 - \sigma_0 c = 0.0412 \\
\hat{\theta}_{\overline{c}} &= \theta_0 + \sigma_0 c = 0.1588
\end{aligned}$$

5.5 Type 2 Error and Power

Remember that

$$\Pr(\text{type 2 error}) = \Pr(\text{don't reject H0}|\text{H0 false})$$

In order to get at the size of this error, we need to choose a specific alternative hypothesis HA, such as

H0: $\theta_0 = 0.1$

HA: $\theta_A = 0.2$

In order to calculate $\beta = \Pr(\text{type 2 error})$, we need to assume a world where HA is true. In such a world $E(\hat{\theta}) = \theta_A = 0.2$, and accordingly, $\sigma_A^2 = \frac{0.2 \cdot 0.8}{100} = 0.0016$, thus $\sigma_A = 0.04$. In other words, we would face a different distribution.

Let's go back to the one-sided test from before. We obtained a critical value $\hat{\theta}_c = 0.1494$. We said that we would reject any H0 for which we observe an estimate larger than that. Now, the β error relates to values that are *below* this critical value, or *outside* the rejection region, but which really were drawn from the distribution under the assumption that $\theta = \theta_A$. We basically need to calculate the probability under the pdf given HA up to the point $\hat{\theta}_c = 0.1494$:

$$
\begin{aligned}
\beta &= \Pr(\text{don't reject H0}|\text{H0 false[i.e. HA is true]}) \\
&= \int_{-\infty}^{\hat{\theta}_c} f\left(\hat{\theta}|E\left[\hat{\theta}\right] = \theta_A = 0.2, Var\left[\hat{\theta}\right] = 0.0016\right) d\hat{\theta} \\
&= \int_{-\infty}^{\hat{\theta}_c} \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left\{-\frac{1}{2}\left(\frac{\hat{\theta} - \theta_A}{\sigma_A}\right)^2\right\} d\hat{\theta} \\
&= \Phi\left(\frac{\hat{\theta}_c - \theta_A}{\sigma_A}\right) \\
&= \Phi\left(\frac{0.1494 - 0.2}{0.04}\right) \\
&= \Phi(-1.265) \\
&= 1 - \Phi(1.265) \\
&= 1 - 0.8962 \\
&= 0.1038
\end{aligned}
$$

A related concept is called the power of a test. It tells us with which probability our test will be able to detect a false hypothesis. It is defined as

$$
\begin{aligned}
\Pr(\text{reject H0}|\text{H0 false[i.e. HA is true]}) &= 1 - \beta \\
&= 0.8962
\end{aligned}
$$

This is illustrated in figure 6. The important thing to note in the context of a test about proportions is that the variance changes automatically by assuming a different true value for the population. Also, notice the tradeoff between type 1 and type 2 error, and how our choice of α influences this tradeoff. Remember that by choosing a particular α we get a particular critical value for the test statistic.

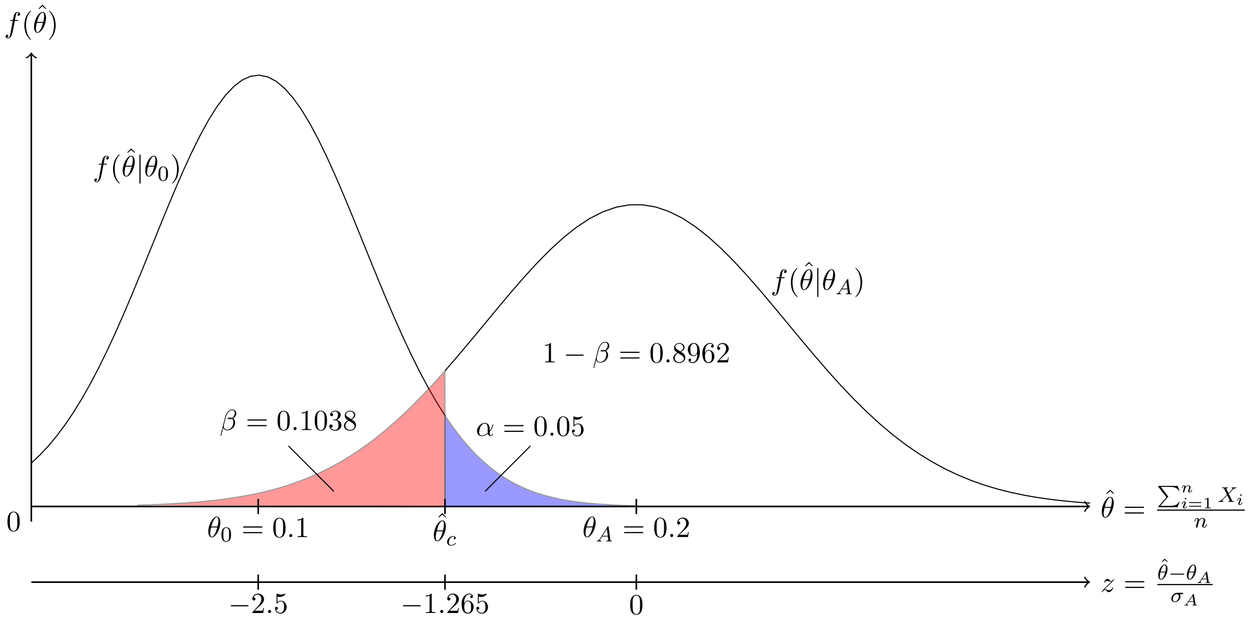

Figure 6: Type 2 error and power of a test.

5.6 P-Value

Let's go back to figure 3. In this example we said that we reject H0 if we get a test statistic such that $z > 1.645$, or a value for the estimator s.t. $\hat{\theta} > 0.1494$. One limitation of the hypothesis-test terminology is that we are constrained to use only two outcomes: reject, or not reject. In other words, for the outcome of our test it doesn't matter if we obtained $z_1 = 1.646$ or $z_2 = 6.5$. Both values of the test statistic would have led to rejection of H0. However, z_2 **provides much stronger evidence than** z_1 (why?). To deal with this shortcoming, we compute p-values.

The p-value of a test is the significance level α^* up to which all H0's would be rejected. In symbols, for any given value of the test z', the p-value is

$$\begin{array}{rcl} \text{"reject if"} & & \\ \text{LHS} & \geq & \text{RHS} \\ z' & \geq & \Phi^{-1}(1-\alpha^*) \\ \alpha^* & \geq & 1-\Phi(z') \end{array}$$

(for a one sided test. for two-sided the condition is $|z'| \geq \Phi^{-1}(1-\frac{\alpha^*}{2})$).

In terms of our example, we reject H0 if

$$\begin{array}{rcl} \alpha^* & \geq & 1-\Phi(1) \\ & = & 1-0.8413 \\ & = & 0.1587 \end{array}$$

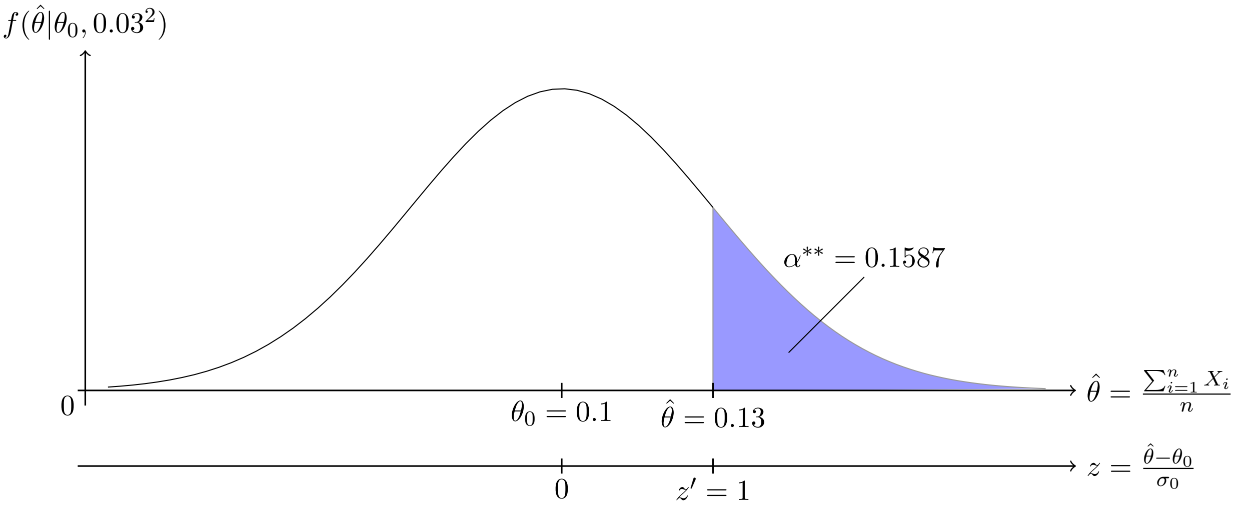

Figure 7: P-value for one sided test given an observed test statistic $z' = 1$.

i.e. we would say that the p-value for this test and this data is 0.1587. This means that we would be able to reject the H0 for any significance level greater or equal 0.1587, and we would not be able to reject it for any smaller significance level. Notice that $\alpha^{**} = 0.1587$ is the smallest of all those levels α^{*} at which we could reject H0, so typically the p-value refers to α^{**}. We would say "test value $z' = 1$ is just significant at the level of significance $\alpha^{**} = 0.1587$ ". Have a look at this in figure 7.

6 Tests in small samples

Nothing substantial changes in the setup of our procedure if we have small samples. What changes, though, are the sampling distributions which we have to consider. I'll go here over a couple examples.

6.1 Test about population mean with unkown variance

Similarly to the setup for confidence intervals about the mean of a normal distribution, with unknown variance σ^2, we use the sample variance S^2 to estimate it and get a test statistic

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1} \tag{2}$$

Example:

A machine is set up to produce metal plates with thickness $\mu = 0.25$ cm. There is some variation in the machine's output, and we assume that this is normally distributed. Taking a random sample of $n = 10$ pieces, we find a sample mean $\bar{x} = 0.253$ cm and a standard deviation of $s = 0.003$ cm. We want to test whether the machine is working according to our specifications or not at a significance level of $\alpha = 0.05$. In short we want to test the following hypothesis:

1. Setup hypothesis and significance level:
   H0: $\mu = 0.25$ cm
   HA: $\mu \neq 0.25$ cm
   level: $\alpha = 0.05$

2. Test statistic and test distribution
   $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_9$

3. Rejection region
   We have a two sided test, thus the same setup as in section 4. We are looking for the the 0.975 quantile of the t-distribution with 9 d.f. or for critical values $\pm c = \pm T_9^{-1}(0.975) = \pm 2.262$. Therefore for a test statistic $|t| < |c|$ we cannot reject H0.

4. calculate test statistic
   $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{0.253-0.25}{\frac{0.003}{\sqrt{10}}} = 3.162$

5. decide: given that $|t| > |c|$, we can reject H0. The machine is not working according to specification according to this data and test.

This test is illustrated in figure 8.

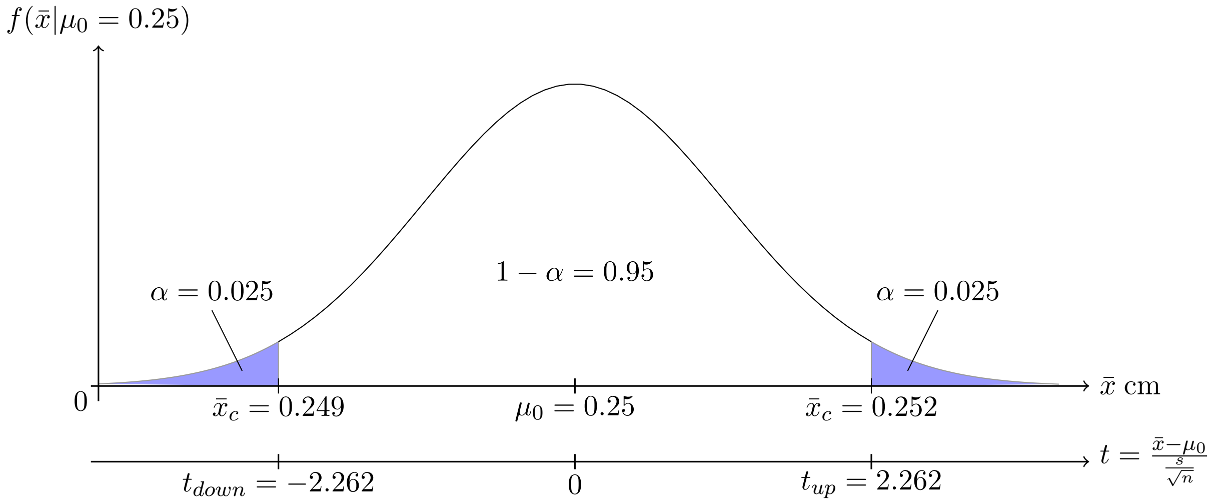

Figure 8: twosided test about a population mean.

Suppose we have $N = 9$ observations drawn from $X_i \sim N(\mu, \sigma^2)$. This *is* a small sample, where the difference between normal and student t-distribution is important. We perform some computations with **R** to illustrate this.[2]

[2] R project website. R. If you want to learn R checkout my site at UCL wiki

```
x <- c(14.2, 18.1, 12.4, 19.6, 16.7, 15.4, 13.5, 10.1, 21.3)  # manually input the 9 values
x  # print them to screen

## [1] 14.2 18.1 12.4 19.6 16.7 15.4 13.5 10.1 21.3
```

6.2 Test about population mean, population variance, and p-value

We are asked to perform a test about whether $\mu = 18$ at significance level 5%.
Given this is a small sample, our test statistic is going to be t-distributed with $N - 1 = 8$ degrees of freedom, identical to the statistic in equation (2). Formulate the hypothesis as

$$\begin{aligned} H_0 : \mu &= 18 \\ H_1 : \mu &\neq 18 \end{aligned}$$

and compute the test statistic

$$t^* = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{N}}} \sim t_8$$

6.2.1 Testing the mean

Doing this in R:

```
N     <- length(x)  # number of obs
xbar  <- mean(x) # sample mean
mu.h0 <- 18      # hypothesized value under H0
S     <- 1/(N-1) * sum( (x-xbar)^2 )  # sample variance
tstar <- (xbar - mu.h0) / sqrt(S / N) # test statistic
```

gives us S^2 = 12.845 and a value of the test statistic t^* = -1.9252. Now, as outlined in previous examples above, we reject the Null in a two-sided test if the statistic is smaller than the critical value in absolute terms, or reject if

$$t^* < t_c$$

where t_c is the corresponding critical value from the t distribution. In our case this is to the left of zero, so we want $t^* < t_c$:

```
crit.val <- qt(p = 0.025, df = 8)  # qt() is the quantile function of the t-dist
crit.val

## [1] -2.306

(tstar < crit.val)  # is tstar smaller than crit.val?

## [1] FALSE
```

So $t^* < t_c$ is false and we cannot reject.

6.2.2 P-value for test

Remember from section 5.6 that the p-value is defined as the tail probability of a certain t^* under H0, or

$$p = \Pr\left(t_8 < t^* | \mu = 18\right)$$

thus we are looking for the probability that a t-distributed variable with 8 degrees of freedom comes to lie below -1.9252 (we are looking for the value of the distribution function at t^*). This is found as

```
tail.prob <- pt(q = tstar, df = (N - 1))  # pt() gives the probability under the t-dist at quant
tail.prob

## [1] 0.04519

(0.025 > tail.prob)  # reject if true

## [1] FALSE
```

i.e. we would reject if the p-value is smaller than the cutoff, which is 0.025 for a two-sided test, and it's not the case here.

6.2.3 Confidence interval for the population variance

The task is to find the 90% confidence interval for the population variance.
Given X is normal, we know that

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{N-1}$$

A 90% confidence interval means for numbers $a < b$ that

$$\Pr[a < \sigma^2 < b] = 0.9$$

If we apply the same operations on both sides of the inequalities on the LHS of this, we obtain

$$\begin{aligned}
\Pr[a < \sigma^2 < b] &= 0.9 \\
\Pr\left[\frac{1}{a} > \frac{1}{\sigma^2} > \frac{1}{b}\right] &= 0.9 \\
\Pr\left[\frac{(N-1)S^2}{a} > \frac{(N-1)S^2}{\sigma^2} > \frac{(N-1)S^2}{b}\right] &= 0.9
\end{aligned}$$

Given the stampling distribution is symmetric about it's mean, this implies for a, b that the probability of falling outside the interval, i.e. 10% needs to be split equally on both sides of the center:

$$\begin{aligned}
\Pr\left[\frac{(N-1)S^2}{a} > \frac{(N-1)S^2}{\sigma^2}\right] &= 0.05 \\
\Pr\left[\frac{(N-1)S^2}{a} > \chi\right] &= 0.05 \\
1 - \Pr\left[\frac{(N-1)S^2}{a} > \chi\right] &= 1 - 0.05 \\
\Pr\left[\frac{(N-1)S^2}{a} < \chi\right] &= 0.95 \\
\frac{(N-1)S^2}{a} &= q_\chi(0.95) \\
a &= \frac{(N-1)S^2}{q_\chi(0.95)}
\end{aligned}$$

and similarly for b

$$\begin{aligned}
\Pr\left[\frac{(N-1)S^2}{b} < \frac{(N-1)S^2}{\sigma^2}\right] &= 0.05 \\
\Pr\left[\frac{(N-1)S^2}{b} < \chi\right] &= 0.05 \\
\frac{(N-1)S^2}{b} &= q_\chi(0.05) \\
b &= \frac{(N-1)S^2}{q_\chi(0.05)}
\end{aligned}$$

such that we get

```
a <- (N - 1) * S/qchisq(0.95, 8)
b <- (N - 1) * S/qchisq(0.05, 8)
```

and the resulting confidence interval is

$$CI = [6.6266, 37.6047]$$

6.3 Testing the difference of two population means

This is the problem if we want to test whether two normal distributions have the same mean. To be precise, suppose have two iid samples $\{X_i\}_{i=1}^n, \{Y_i\}_{i=1}^m$, with $X_i \sim N(\mu_X, \sigma_X^2)$ and $Y_i \sim N(\mu_Y, \sigma_Y^2)$, where both all population parameters are unknown. The problem is then, upon observing our samples, to decide whether $\mu_X = \mu_Y$. There are two cases, depending on whether we assume that $\sigma_X^2 = \sigma_Y^2$, or not.

6.3.1 $\sigma_X^2 \neq \sigma_Y^2$: unequal variances

Here we will assume that both sample sizes n, m are large enough s.t. the CLT can be invoked. Under our iid assumption on the samples, the variable Z is

$$Z = \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

Given our large sample assumption, we may use the sample variances $S_J^2 = \frac{1}{n-1}\sum_{i=1}^n \left(J_i - \bar{J}\right)^2, J = X, Y$ to estimate the variances and obtain

$$Z = \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim N(0,1)$$

Example: Two separate groups of students with $n = 40, m = 50$ are given the same exam paper. the mean score in group X is $\bar{x} = 74$ with a standard deviation of $s_X = 8$ points, and in group Y it's $\bar{y} = 78, s_Y = 7$. We want to test whether the means of the underlying population distributions are different at a 5% level of significance.

1. Setup hypothesis:
   H0: $\mu_X = \mu_Y$
   HA:$\mu_X \neq \mu_Y$
2. test statistic and sampling distibution under H0 (ie equal means!):
   $Z = \frac{\left(\bar{X}-\bar{Y}\right)-\left(\mu_X-\mu_Y\right)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}} = \frac{\left(\bar{X}-\bar{Y}\right)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}} \overset{a}{\sim} N(0,1)$
3. Rejection Region:
   We are using a two sided test. As seen earlier, with a standard normal test distribution and $\alpha = 0.05$ one obtains $z_c = 1.96$. I.e. if we find that $|z| > 1.96$, we reject H0.
4. compute test statistic:

$$\begin{aligned} z &= \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \\ &= \frac{(74 - 78)}{\sqrt{\frac{64}{40} + \frac{49}{50}}} \\ &= -2.49 \end{aligned}$$

5. decide: given $|z| > 1.96$, we reject H0

In terms of figures, this is identical to the situation in figure 4.

6.3.2 $\sigma_X^2 = \sigma_Y^2$. equal variances.

With equal variances it can be shown that the relevant test-statistic is given by

$$T = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}}$$

which is t distributed with $n + m - 2$ degrees of freedom.

References

Bleymüller, J., G. Gehlert, and H. Gülicher (1994): *Statistik für Wirtschaftswissenschaftler.* Vahlen, 14 edn.

DeGroot, M., and M. Schervish (2001): *Probability and Statistics-International Edition.* Addison-Wesley. Publishing. Company., Reading, Massachusetts.

R Development Core Team (2011): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.