1 Hypothesis Testing

Setting up and testing hypotheses is seen in most courses as an essential part of statistical inference. We think that such testing my loom too large in undergraduate courses but it is an important part of statistics and you need to be able to both understand and construct statistical tests. We begin with some ideas and definitions.

1.0.1 Hypotheses

We often make assertions and as in many cases we have incomplete information, the assertion is about a probability distribution. Such an assertion abut a distribution is called a *statistical hypothesis*. It may be a *simple hypothesis* if it completely specifies the distribution or *complex* if it is not simple.

1.0.2 Example

- $H_0 : f(x) = \theta \exp(-\theta x)$ is a *simple* hypothesis if we add $\theta = 3$.
- X is N(100,σ^2), for an unspecified σ, is *complex*.

Typically the hypothesis has been put forward, either because it is believed to be true or because it is to be used as a basis for argument.

1.0.3 Example

1. Suppose we have a coin which we wish to check is fair, that is P[Head]=1/2. If we assume the coin is fair we are assuming that the number of heads, X, is Binomial with $p = 1/2$.
2. We might be interested in the birth weight of babies born in Sheffield compared with those born in Brighton. Since we know (from the medics) that birth weights are Normal we can think of the hypothesis that the Sheffield and Brighton weights have Normal distributions with the same mean. This is *not* a simple hypothesis as the means and variances are not specified.

In each problem we shall consider, the question of interest is simplified into two competing hypotheses between which we have a choice;

- The null hypothesis, denoted H_0
- The alternative hypothesis, denoted H_1, which is the compliment (in the context of the problem) of the null.

These two competing hypotheses are not however treated on an equal basis, special consideration is given to the null hypothesis. Usually the experiment has been carried out in an attempt to prove or disprove a particular hypothesis, the null hypothesis. For example,

H_0: there is no difference in taste between coke and diet coke

against

H_1: there is a difference

Of the two hypotheses the null is almost always simple in that it completely specifies the underlying distribution, the alternative is often complex.

1.0.4 Example

1. H_0 : X is Binomial (100,1/2) i.e. p is specified
   H_1 : X is Binomial $(100, p)$ $p \leq 1/2$

2. H_0 : $X$is N(5, 20) i.e. μ and σ are specified
   H_1 : X is N(μ, 20) i.e. $\mu > 5$

1.1 Type I and II errors

As we are dealing with incomplete information errors are inevitable. The power of statistics lies in the fact that we accept these errors and in the way we deal with them.

If we have two competing hypotheses there are two kinds of errors that may arise and the following table gives a summary of possible results .

	Truth	
action	H_0	H_1
Accept H_0	ok	Type II error
Reject H_0	Type I error	ok

The two errors are traditionally called the type one and type two errors and we will be looking at their probabilities

1. $\alpha = P[$ Type I error$] = P[$reject $H_0 | H_0$ true$]$.
2. $\beta = P[$Type II error$] = P[$accept $H_0 | H_0$ false$]$

2 A general approach

Suppose we choose an acceptable value for the probability of a type I error,

$$\alpha = P[\text{ Type I error}] = P[\text{reject } H_0 | H_0 \text{ true}]$$

such as 0.05 or 0.01. Then take the sample space of outcomes $(x_1, x_2, \ldots, x_n)$ and split it into two parts C and R where $C \cap R$ is empty and $C \cup R$ is the whole space.

We choose C - *the critical region* - to be the set of unlikely points, that is the set of outcomes which are (under H_0) unlikely. Then if we observe a set of points in C we have two options

- Either H_0 is false
- Or we have observed an event of small probability.

In conventional testing we assume the second and say we reject the null hypothesis and accept the alternative.

By this we mean that it is rational on the evidence to believe that the null is not true.

Recall

- The probability of observing the unlikely event is α the probability of a type I error- **often referred to as the size of the test.**
- As we have seen the probability of a type II error $\beta = P[\text{accept } H_0 | H_1]$ is generally unknown and *needs to be calculated*. The alternative measure is the power $P[\text{reject } H_0 | H_1]$ but this also has to be computed.

If we do not reject the null hypothesis, *it may still be false* (a type II error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis). You should bear in mind that for any given set of data, the type I and type II errors are inversely related; so the smaller the risk of one, the higher the risk of the other.

As dealing with high dimensional spaces is difficult we usually base our tests on a test statistic T computed from the observations. As above we find a critical region C defined by

$$P[T \in C | H_0] = \alpha$$

Thus the region C is those values of T which are unlikely when H_0 is true. Some workers prefer the *p-value*. The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone on the assumption that the null hypothesis H_0, is true. We think this is the wrong approach but being lazy will use it from time to time.

2.1 Power

The power of a statistical hypothesis test measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision. In other words, the power of

a hypothesis test is the probability of not committing a type II error that is $\gamma = 1 - \beta$. Usually statisticians think of the power as a function of the parameter, So if we have

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1$$

they would consider the power as being $\gamma(\theta_1)$ a function of possible alternatives.

2.2 Summary of the definitions

- A statistical hypothesis is an assertion about a probability distribution.
- A simple hypothesis completely specifies the distribution.
- $\alpha = P[$ Type I error$] = P[$reject $H_0 | H_0$ true$]$.
- $\beta = P[$ Type II error$] = P[$accept $H_0 | H_0$ false$]$.
- The critical region C is those values of T which are unlikely when H_0 is true.
- The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone on the assumption that the null hypothesis H_0, is true.
- The power of a statistical hypothesis test is the probability of rejecting the null hypothesis when it is false.

3 Constructing tests

While the apparatus above is reasonable it does not answer the questions as to how one might find a suitable statistic T and how we can be assured that T encapsulates the whole problem. Most statistics book give a list of recipes. The procedure is typically:

1. Set up H_0 and H_1
2. Pick a suitable test statistic T whose distribution is known under the assumptions of H_0.
3. Choose the size of the test $\alpha = P[$reject $H_0 | H_0$ true$]$
4. Find the critical region using the distribution in 2
5. Compute T.
6. If T lies in the critical region reject H_0.

Often we can derive such a method from insight into the problem, as we shall see.

3.1 A Binomial example

We toss a coin 12 times and observe

```
HH TH TT TH HT HH
```

We assume that X the number of heads in Binomial B(12,p) and our null hypothesis is

$$H_0 : P[\text{ Head }] = p = \frac{1}{2}$$

while the alternative is

$$H_1 : P[\text{ Head }] = p > \frac{1}{2}$$

1. One choice of statistic is just X the number of heads.
2. We choose α to be approximately 0.05.
3. To find a critical region C note

$$\alpha = P[X > c \quad | H_0 : p = 1/2]$$

From tables of the Binomial we have

c	P[$X > c$]
0	0.99976
1	0.99683
2	0.98071
3	0.92700
4	0.80615
5	0.61279
6	0.38721
7	0.19385
8	0.07300
9	**0.01929**
10	0.0032
11	0.0002
12	0.0000

Now an α of 0.05 is not possible but we can have 0.0193. So we choose a critical region of the form $X > 9$. In our sample we have $X = 7$ heads. Since this does not lie in the critical region we accept H_0 and conclude that $p = \frac{1}{2}$

Notice that if we had a simple alternative, say $H_1 : p = 0.7$ we could compute the type II error since $\beta = P[X \leq c | p = 0.7]$.

You might like compute this this yourself.

The power is just $\gamma(p) = P[X > 9|p]$ which is

P	$\gamma(p)$
0.5	0.01929
0.6	0.08344
0.7	0.25282
0.8	0.55835
0.9	0.88913

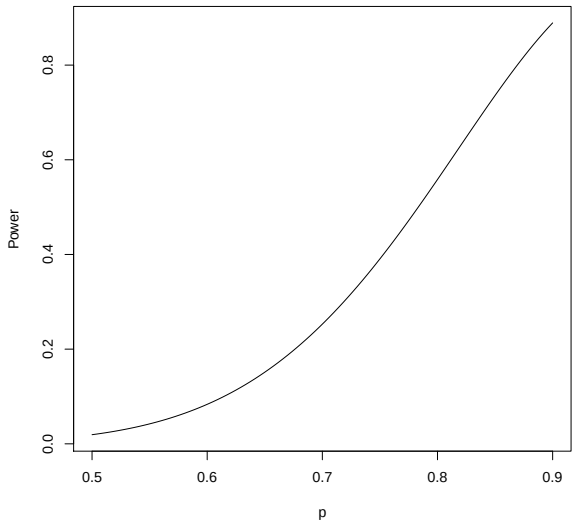


Figure 1: The power as being $\gamma(p)$ a function of possible alternatives with parameter $p = 0.5$ to 0.9 in 0.004 increments.

The critical region depends on the alternative - try finding the critical region for testing

$$H_0 : p = \frac{1}{2} \text{ against the alternative } H_1 : p > \frac{1}{2}$$

3.2 A two tailed test!

Suppose we wish to test

$$H_0 : p = \frac{1}{2} \text{ against the alternative } H_1 : p \neq \frac{1}{2}$$

then the critical region will consist of small values of X and large values of X. Hence we have a region of the form $X < c_1$ and $X > c_2$. We split our α between the two segments so

$$P[X < c_1|H_0 : p = 1/2] = P[X > c_2|H_0 : p = 1/2]$$

Extending the table above gives

c	P[$X > c$]	P[$X \leq c$]
0	0.99976	0.00024
1	0.99683	0.00317
2	0.98071	0.01929
3	0.92700	0.07300
4	0.80615	0.19385
5	0.61279	0.38721
6	0.38721	0.61279
7	0.19385	0.80615
8	0.07300	0.92700
9	0.01929	0.98071
10	0.00317	0.99683
11	0.00024	0.99976
12	0.00000	1.00000

Now if we choose $c_2 = 9$ as before with $c_1 = 3$ then $\alpha = 2 \times 0.019287$ or about 0.0386.

Notice the critical region was in both tails of the distribution. Tests with regions of this for are often called *two tailed tests*.

3.2.1 A Normal example

Suppose we have a sample of size 100 from a Normal distribution. We wish to test

$$H_0 : \mu = 68 \text{ against } H_1 : \mu \neq 68$$

To simplify matters we assume that the standard deviation of the population is $\sigma = 16$.

A possible statistic is $\bar{X}$ which is Normal N($\mu, \sigma^2/n$). Rather simpler is the standardized random variable

$$z = \frac{\bar{X} - 68}{\sigma/\sqrt{n}}$$

which we know is standard Normal *when* H_0 *is true*.

Now if the true mean is **not** 68 then z will be very different from zero. The distribution of z is shown in the Figure 2 (c) and we see that the two areas in the tails must total α. A little inspection gives us the critical regions as

$$z < -z_{1-\alpha/2} \text{ and } z > z_{1-\alpha/2}$$

Now if we choose $\alpha = 0.05$ then we will reject H_0 when $z < -1.96$ or $z > 1.96$

In our case $\bar{X} = 68.04$. We know that $\sigma = 16$ so $z = 0.025$. This is not in the critical region so we accept H_0

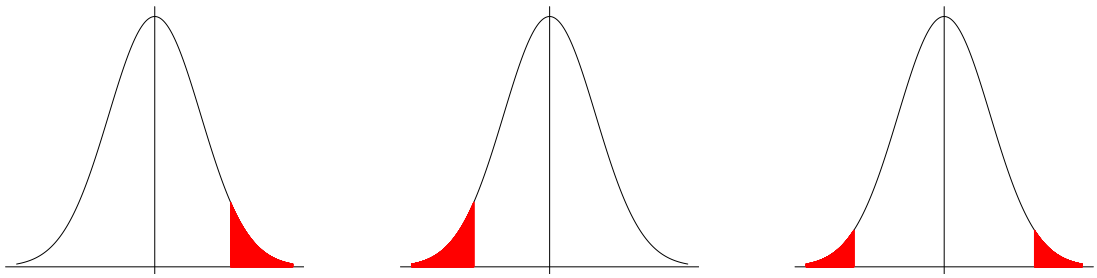
This is somewhat unrealistic since if we are unsure of μ is is most unlikely that we know σ!

Here we have a large sample and for for large samples (exceeding 50) we can use an estimate. Here $\hat{\sigma} = 13.724$ so

$$T = \frac{\bar{X} - 68}{\sigma/\sqrt{n}} = 0.029$$

and in consequence we accept H_0.

What is the critical region for the alternative $H_1 : \mu > 68$?



(a) $z > z_{1-\alpha}$ (b) $z < -z_{1-\alpha}$ (c) $|z| > z_{1-\alpha/2}$

Figure 2: Critical regions for testing $H_0 : \mu = \mu_0$ against (a) $H_1 : \mu_1 > \mu_0$, (b) $H_1 : \mu_1 < \mu_0$, and (c) $H_1 : \mu_1 \neq \mu_0$, based on $z = \frac{\bar{x}-\mu_0}{\sqrt{s^2/n}} \sim N(0,1)$.

3.3 Normal small sample case

For a small sample we would then have to find the distribution of

$$t = \frac{\bar{X} - 68}{\hat{\sigma}/\sqrt{n}}$$

in order to compute the size of the critical region.

In fact we know that the distribution of t has a Student's t distribution with $n-1$ degrees of freedom so all we need to do is to find the critical region using tables of t rather than Normal tables.

A sample of 10 batteries are randomly selected from a production batch and their lifetimes found. The mean lifetime is 30.3 and the estimated variance is 16.08456. The manufacturer claims a lifetime of 36 months. Suppose we assume a a normal population with mean μ and test

$$H_0 : \mu = 36 \text{ against } H_1 : \mu < 36$$

Then

$$t = \frac{30.3 - 36}{\sqrt{16.08/10}} = -4.49$$

Now we find the critical region using the t distribution with 10-1=9 degrees of freedom. The critical value is -1.833, for a test size of 0.05 (check this !) so we have a value of t in the critical region and we reject H_0

3.3.1 Summary

We have deduced the procedure given in statistical recipe books. Suppose we are given a random sample $X_1, X_2, \cdots, X_n$ from a $\mathrm{N}(\mu, \sigma^2)$ distribution and we wish to test

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu_1 > \mu_0, \quad H_1 : \mu_1 < \mu_0, \quad H_1 : \mu_1 \neq \mu_0$$

then we compute

- For large samples ($n > 50$) we compute

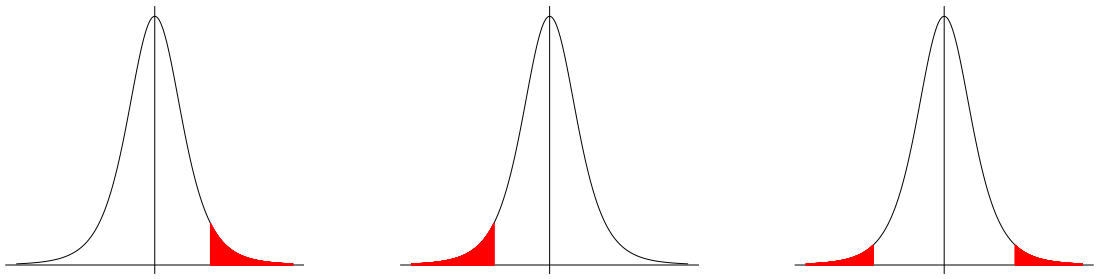
  $$z = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  and we reject H_0 when $z > z_{1-\alpha}$, $z < -z_{1-\alpha}$, $|z| > z_{1-\alpha/2}$, see Figure 2.

- When the sample size is small, $n < 50$, we use the t distribution with $n-1$ degrees of freedom and compute

  $$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  and we reject H_0 when $t > t_{1-\alpha}$, $t < -t_{1-\alpha}$, $|t| > t_{1-\alpha/2}$ using Student's t with $n-1$ degrees of freedom, see Figure 3.



(a) $t > t_{1-\alpha}$ (b) $t < -t_{1-\alpha}$ (c) $|t| > t_{1-\alpha/2}$

Figure 3: Critical regions for testing $H_0 : \mu = \mu_0$ against (a) $H_1 : \mu_1 > \mu_0$, (b) $H_1 : \mu_1 < \mu_0$, and (c) $H_1 : \mu_1 \neq \mu_0$, based on $t = \frac{\bar{x}-\mu_0}{\sqrt{s^2/n}} \sim t_{n-1}$.

3.3.2 Variances

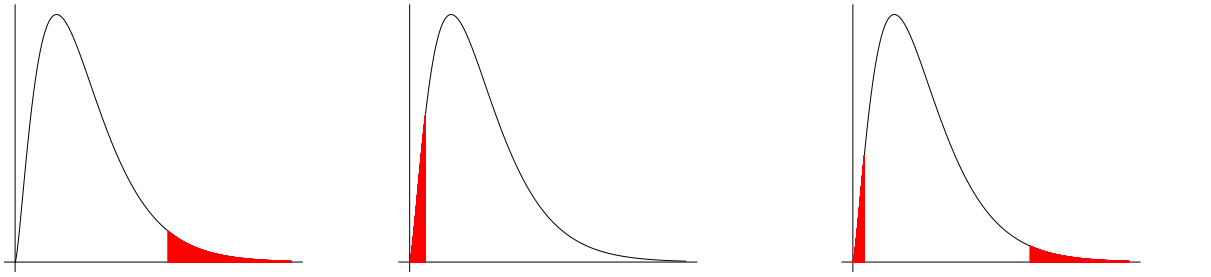
Suppose we are given a random sample $X_1, X_2, \cdots, X_n$ from a $N(\mu, \sigma^2)$ distribution and we wish to test

$$H_0 : \sigma^2 = \sigma_0^2 \text{ against } H_1 : \sigma^2 > \sigma_0^2, \quad H_1 : \sigma^2 < \sigma_0^2, \quad H_1 : \sigma^2 \neq \sigma_0^2,$$

our critical regions will be respectively

$$X^2 > \chi^2_{1-\alpha,n-1}, \quad X^2 < \chi^2_{\alpha,n-1}, \quad X^2 > \chi^2_{1-\alpha/2,n-1} \text{ or } X^2 < \chi^2_{\alpha/2,n-1},$$

where $X^2 = \frac{(n-1)s^2}{\sigma_0^2}$ has a Chi-squared distribution with $n-1$ degrees of freedom under H_0.



(a) $X^2 > \chi^2_{1-\alpha,n-1}$ (b) $X^2 < \chi^2_{\alpha,n-1}$ (c) $X^2 > \chi^2_{1-\alpha/2,n-1}$ or $X^2 < \chi^2_{\alpha/2,n-1}$

Figure 4: Rejection regions for testing $H_0 : \sigma^2 = \sigma_0^2$ against (a) $H_1 : \sigma^2 > \sigma_0^2$, (b) $H_1 : \sigma^2 < \sigma_0^2$, and (c) $H_1 : \sigma^2 \neq \sigma_0^2$ based on $X^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim X^2_{n-1}$.

3.3.3 Example

We are given a random sample with $n = 10$ from a $N(\mu, \sigma^2)$, where $s^2 = 12.6$. For $\alpha = 0.05$ we test

$$H_0 : \sigma^2 = 9 \text{ against } H_1 : \sigma^2 > 9.$$

The statistic is $X^2 = \frac{(n-1)s^2}{\sigma_0^2}$ with critical region $X^2 > \chi^2_{1-\alpha,n-1}$. Then $X^2 = \frac{9 \times 12.6}{9} = 12.6$ and $\chi^2_{0.95,9} = 16.919$. So we have a value of not in the critical region and we cannot reject H_0.

3.3.4 A Binomial/Normal example

We can use for large samples the normal approximation to the Binomial, to turn what are essentially Binomial problems in to Normal ones. The binomial distribution is the discrete

probability distribution of the number of successes in a sequence of n independent experiments, each of which yields success with probability p. If $X \sim B(n,p)$ (that is, X is a binomially distributed random variable), then the expected value of X is $E(X) = np$ and the variance is $Var(X) = np(1-p)$. Therefore if n is large ($n > 50$) we can approximate the binomial distribution with the normal distribution $N(\mu = np, \sigma^2 = npq)$ or the standard normal $Z = (X - np)/\sqrt{(npq)}$. According to this approximation we test

$$H_0 : p = p_0 \text{ against } H_1 : p_1 > p_0, \quad H_1 : p_1 < p_0, \quad H_1 : p_1 \neq p_0$$

by computing

$$z = \frac{\bar{x} - np_0}{\sqrt{np_0(1-p_0)}}$$

and we reject H_0 when $z > z_{1-\alpha}, \quad z < -z_{1-\alpha}, \quad |z| > z_{1-\alpha/2}$.

3.4 Sample size choice

In many situations we can use our definitions of the type I and type II probabilities to specify the sample size we require.

3.4.1 A crossover trial

Suppose we wish to test the effects of two kinds of medication A and B on reducing blood pressure in males. We intend to teat n patients for five weeks on A and five weeks on B. The order of application will be randomized. The response is the average blood pressure in the third week of each treatment.

If we pick $\alpha = 0.05$ and $\beta = 0.1$ how large a sample do we need?

We assume that the responses $A_i, B_i, i = 1, 2, \cdots, n$ are normal and look at the differences $D_i = A_i - B_i, i = 1, 2, \cdots, n$. The test is

$$H_0 : \mu_D = 0 \text{ against } H_1 : \mu_D > 0$$

Now the background to this experiment is that A is the current treatment and we expect to see a change from A to B whose size is about 1/2 a standard deviation.

Back to definitions - in general

$$\alpha = P(z > z_{1-a} | H_0 : \mu = \mu_0) \Longleftrightarrow 0.05 = P(z > z_{0.95} | H_0 : \mu = 0) \Longleftrightarrow P(z \leq z_{0.95} | H_0 : \mu = 0) = 0.95.$$

From the statistical tables (inverse normal) one can see that $z_{0.95} = 1.645$. Therefore the critical region is

$$z > 1.645 \Longleftrightarrow \frac{\bar{D} - \mu_0}{\sigma/\sqrt{n}} > 1.645 \Longleftrightarrow \bar{D} > \mu_0 + 1.645\sigma/\sqrt{n}$$

Now in addition we require

$$\beta = 0.1 = P[\bar{D} \leq \mu_0 + 1.645\sigma/\sqrt{n} | H_1 : \mu = \mu_1 = \mu_0 + \frac{\sigma}{2}] \Longleftrightarrow$$

$$0.1 = P[\frac{\bar{D}-\mu_1}{\sigma/\sqrt{n}} \leq \frac{\mu_0 + 1.645\sigma/\sqrt{n} - \mu_0 - \sigma/2}{\sigma/\sqrt{n}} | H_1 : \mu = \mu_1 = \mu_0 + \frac{\sigma}{2}] \Longleftrightarrow$$

$$0.1 = P[z \leq \frac{1.645\sigma/\sqrt{n} - \sigma/2}{\sigma/\sqrt{n}} | H_1 : \mu = \frac{\sigma}{2}] \Longleftrightarrow$$

$$0.1 = P[z \leq 1.645 - \sqrt{n}/2 | H_1 : \mu = \frac{\sigma}{2}].$$

From the statistical tables (inverse normal) one can see that $z_{0.1} = -1.2816$. So

$$1.645 - \sqrt{n}/2 = -1.2816 \Longleftrightarrow n = 35.$$

3.4.2 A Binomial/Normal case

A medic knows that of patients admitted to hospital for cardiac problem 60% will be readmitted on an emergency basis within 2 months. She believes that treatment with X will reduce this readmission level by half, that is to 30%. To check her theory she must experiment on patients and to gain permission to do so she must show that her experiment has a reasonable chance of detecting a change and uses the minimum number of patients.

This could be framed as a Binomial, with p the probability of readmission. Suppose we take n patients and administer the drug to them. We set up the hypotheses

$$H_0 : p = 0.6 \text{ against } H_1 : p < 0.6$$

We observe R of our n treated patients readmitted. We will assume that R is Binomial and that we can approximate R by a Normal distribution.

Now how to specify the sensitivity of our procedure. We ask that the type II error probability be

$$\beta = P[R \geq k | p = 0.3] = 0.1$$

You should check that you understand this!

I am going to specify the test size as $\alpha = 0.05$

The critical region is of the form $R \leq k$ and assuming normality we have the definition of test size

$$\alpha = P[R \leq k | p = 0.6] = 0.05$$

Using the normal approximation

$$0.05 = \alpha = P[R \leq k | p = 0.6] = P[z = \frac{R - np}{\sqrt{np(1-p)}} < -z_{1-a} | p = 0.6] = P[z = \frac{R - n \times 0.6}{\sqrt{n \times 0.6 \times 0.4}} < z_a | p = 0.6].$$

From the statistical tables (inverse normal) one can see that $z_{0.05} = -1.645$. Therefore the critical region is

$$R < n \times 0.6 - 1.645 \times \sqrt{n \times 0.6 \times 0.4}$$

For the type II error

$$0.1 = \beta = P[R \geq n \times 0.6 - 1.645 \times \sqrt{n \times 0.6 \times 0.4} | p = 0.3] \Longleftrightarrow$$

$$0.1 = P\Big[z = \frac{R - np}{\sqrt{np(1-p)}} \geq \frac{n \times 0.6 - 1.645\sqrt{n \times 0.6 \times 0.4} - n \times 0.3}{\sqrt{n \times 0.7 \times 0.3}} | p = 0.3\Big] \Longleftrightarrow$$

$$0.9 = P\Big[z < \frac{n \times 0.6 - 1.645\sqrt{n \times 0.6 \times 0.4} - n \times 0.3}{\sqrt{n \times 0.7 \times 0.3}}\Big].$$

From the statistical tables (inverse normal) one can see that $z_{0.9} = 1.2816$. Therefore

$$1.2816 = \frac{n \times 0.6 - 1.645\sqrt{n \times 0.6 \times 0.4} - n \times 0.3}{\sqrt{n \times 0.7 \times 0.3}} \Longleftrightarrow \cdots \Longleftrightarrow n = 22.$$