HYPOTHESIS TESTING (Chapter 10)

Motivating scenario

We have a bent coin and there is interest in *p*, the probability of heads coming up on a single toss. In Chapters 8 and 9 we discussed techniques for estimating *p* (point and interval estimation, the method of moments and maximum likelihood estimation).

Now suppose that we are not really interested in the exact value of *p*, but wish to decide *only* whether the coin is *fair* or *unfair*. The theory of **hypothesis testing** (HT) provides a framework whereby such simple decisions can be made - and evaluated.

Example 1

The following is a procedure which could be used to decide whether a bent coin is fair or not. It illustrates the various components that make up a **hypothesis test** (HT).

1. H_0: $p = 1/2$ — This statement, that the coin is *fair*, is called the **null hypothesis** (NH).

2. H_a: $p \neq 1/2$ — This statement, that the coin is *unfair*, is called the the **alternative hypothesis** (AH). We may write the AH as H_1 rather than H_a.

3. Toss the coin $n = 10$ times, and record Y, the number of heads that come up. — This is the statistical experiment we intend to carry out for purposes of the test. Y is called the **test statistic** (TS). Note that $Y \sim Bin(10,0.5)$ if H_0 is true.

4. If $3 \leq Y \leq 7$ then accept H_0; otherwise reject H_0 in favour of H_a. — This defines the **acceptance region** (AR), {3,4,5,6,7}.

The AR is a set of values for the TS that are 'consistent' with H_0. The **rejection region** (RR) here is $S - \text{AR} = \{0,1,2,8,9,10\}$. This is a set of values for the TS that are 'inconsistent' with the NH. If the TS lies in the RR we will *reject* the NH. Otherwise, the TS will be in the AR, and we will *accept* the NH.

The above defines a hypothesis test, but that test is yet to be conducted.

Suppose that we now actually toss the coin 10 times and get two heads.

Then, because 2 is in the rejection region, we reject the null hypothesis and conclude that the coin is unfair (i.e., that the probability of heads coming up is not exactly 1/2).

But.... how confident can we be about this conclusion?

This question raises the need to study some *properties* of a hypothesis tests.

Definition

A **Type I error** occurs if H_0 is rejected when H_0 is true.

The probability of a Type I error is denoted α,
and may also be called the **significance level (SL)** of the test.

A **Type II error** occurs if H_a is rejected when H_a is true.

The probability of a Type II error is denoted β.

Example 2

Find α and β for the test in Example 1.

$$\begin{aligned}\alpha &= P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ true}) \\ &= P(Y \in \{0,1,2,8,9,10\}) \quad \text{where } Y \sim Bin(10,0.5) \\ &= 2P(Y \le 2) \quad \text{(by symmetry about 5)} \\ &= 2(0.055) \quad \text{by the binomial tables in the text (page 839)} \\ &= 0.11.\end{aligned}$$

Thus there is an 11% chance that we will conclude the coin is unfair if it is in fact fair.

$$\begin{aligned}\beta &= P(\text{Type II error}) = P(\text{Reject } H_a \mid H_a \text{ true}) \\ &= P(\text{Do not reject } H_0 \mid H_a \text{ true}) \\ &= P(Y \notin \{0,1,2,8,9,10\}) \quad \text{where } Y \sim \text{Bin}(10,p) \text{ and } p \neq 0.5. \\ &= P(3 \le Y \le 7).\end{aligned}$$

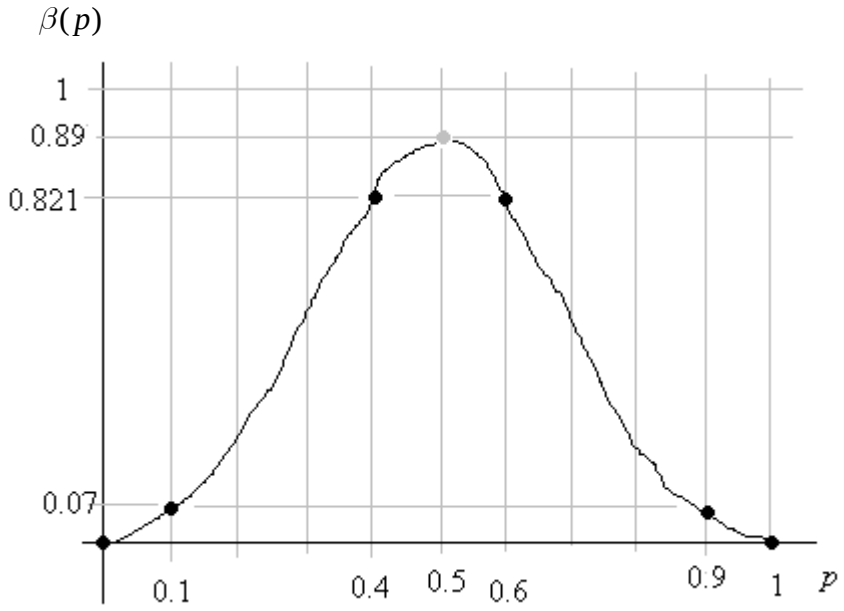
We see that β depends on p. Thus β is a function, and not a single value like α.

Explicitly, $\beta = \beta(p) = \sum_{y=3}^{7} \binom{10}{y} p^y (1-p)^{10-y}, \quad 0 \le p \le 1, \quad p \ne 0.5.$

For example, using binomial tables,

$$\beta(0.4) = P(Y \le 7) - P(Y \le 2)$$
$$= 0.988 - 0.167$$
$$= 0.821.$$

Similarly: $\beta(0.1) = 0.070$
$\beta(0.6) = 0.821$
$\beta(0.9) = 0.070,$
etc.



Note: $\beta(p) \approx 1 - \alpha = 0.89$ if $p \approx 0.5$.
In this case $\beta(p)$ is a symmetric function around $p = 0.5$.
$\beta(0) = \beta(1) = 0.$

Example 3

Consider Example 1 once again, but now suppose that the rejection region is {0,1,9,10} (a 'reduced version' of {0,1,2, 8,9,10}).

Find α and β in this case.

$$\alpha = P(\text{Type I error}) = P(Y \in \{0,1,9,10\}) \quad \text{where } Y \sim Bin(10,0.5)$$
$$= 2P(Y \leq 1) = 2(0.011) = 0.022.$$

Observe that α is *smaller* than before (0.11). This is *good*. There is now a *lower* chance of making a Type I error.

Also, $\beta(p) = P(2 \leq Y \leq 8)$ where $Y \sim Bin(10,p)$.

We find that: $\beta(0.1) = 0.264$ (> 0.070)
$\beta(0.4) = 0.952$ (> 0.821)
$\beta(0.6) = 0.952$ (> 0.821)
$\beta(0.9) = 0.264$ (> 0.070), etc.

Observe that $\beta(p)$ is now uniformly *larger* than before. This is *bad*. There is now a *higher* chance of making a Type II error, whatever the true value of *p*.

We see that when deciding upon the rejection region, a *trade-off* is involved. We can make α smaller, or we can make β smaller, but we cannot do both at the same time.

This assumes that we have the same amount of information available, in our case that *n* is fixed. If we were to toss the coin *n* = 50 times, instead of 10, we would be able to design a hypothesis test which has lower probabilities of *both* Type I and II errors.

If the coin were tossed *n* = 1,000,000 times (say), we would be able to design a HT with values of α and $\beta(p)$ all very small (except in a neighbourhood of 0.5, where $\beta(p) \approx 1-\alpha$). In the hypothetically ideal case $n = \infty$ we could get $\alpha = 0$ and $\beta(p) = 0,\ p \neq 0.5$.

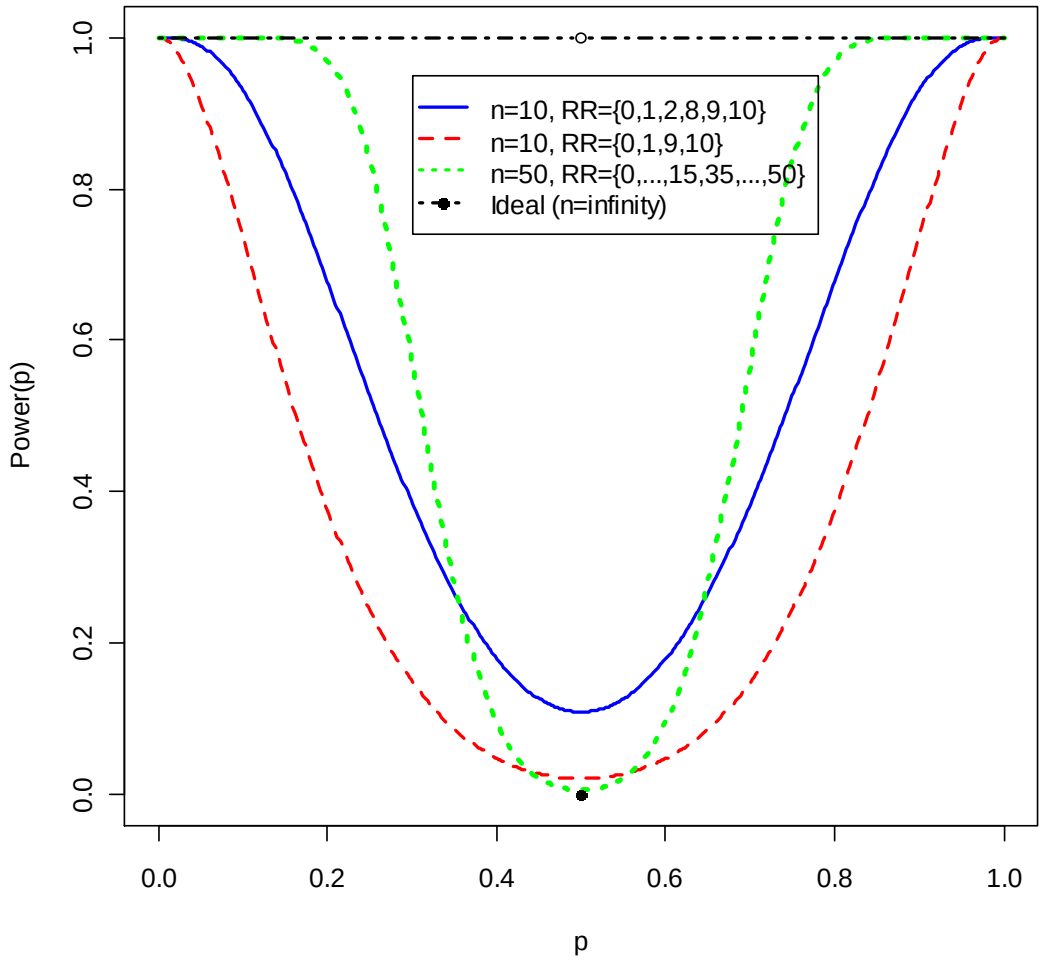
The power of a test

Another way to characterise a hypothesis test is in terms of the *power function*. This function combines all of the information provided by α and $\beta(p)$ above.

The power function, denoted $Power(\theta)$, is the probability of rejecting the null hypothesis when the true value of the parameter under consideration is θ.

Thus in Examples 2 and 3, $$Power(p) = \begin{cases} \alpha = 0.11, \ p = 0.5 \\ 1 - \beta(p), 0 \le p \le 1, p \ne 0.5 \end{cases}$$

This function is shown below for each of the rejection regions, {0,1,2, 8,9,10} and {0,1, 9,10}), when n = 10. Also shown is the power function for when n = 50 and RR = {0,...,15,35,...,50}, as well as the ideal power function corresponding to both types of error being impossible (and achieveable only in the limit as $n \to \infty$).



R Code (non-assessable)

```
pvec = seq(0,1,0.002); m = length(pvec)
powvec1 = rep(NA,m); powvec2 = rep(NA,m); powvec3 = rep(NA,m)
for(i in 1:m){   pval = pvec[i]
                 powvec1[i] = pbinom(2,10,pval)  +  1-pbinom(7,10, pval)
                 powvec2[i] = pbinom(1,10, pval)  + 1-pbinom(8,10, pval)
                 powvec3[i] = pbinom(15,50, pval)  + 1-pbinom(34,50, pval)   }
plot(c(0,1),c(0,1),type="n",xlab="p",ylab="Power(p)")
lines(pvec,powvec1,lty=1,lwd=2,col="blue")
lines(pvec,powvec2,lty=2,lwd=2,col="red")
lines(pvec,powvec3,lty=3,lwd=3,col="green")
lines(c(0,0.49),c(1,1),lty=4,lwd=2,col="black")
lines(c(0.51,1),c(1,1),lty=4,lwd=2,col="black")
points(c(0.5,0.5),c(0,1),pch=c(16,1),col="black")

legend(0.3,0.95,    c(   "n=10, RR={0,1,2,8,9,10}",
                         "n=10, RR={0,1,9,10}",
                         "n=50, RR={0,...,15,35,...,50}",
                         "Ideal (n=infinity)"),
          lty=c(1,2,3,4),lwd=c(2,2,2,2),pch=c(NA,NA,NA,16),
         col=c("blue","red","green","black"))
```

Example 4

A coin is to be tossed $n = 100$ times to test whether or not it is fair. Determine a rejection region for Y, the number of heads which come up, so that the probability of making a Type I error is 5%.

If the coin is fair then we'd expect about 50 heads to come up.
So let the rejection region be of the form $RR = \{0,...,50 - k,\quad 50 + k,...,100\}$.
(This is a set of values furthest away from 50.)

We now need to determine k such that

$$0.05 = P(Y \in RR),$$

where $Y \sim Bin(100,0.5)$.

Thus $0.025 = P(Y \geq 50+k) = P\left(\dfrac{Y-50}{5} \geq \dfrac{50+k-50}{5}\right)$

$\approx P(U > k/5)$ where $U \sim N(0,1)$

by the central limit theorem.

But 0.025 = $P(U > 1.96)$.
So we equate 1.96 = k/5, and get k = 9.8.

So we should take the rejection region as RR = {0,...,40, 60,...,100}.
This will result in a significance level of approximately 5%.

For example, suppose that y = 38 heads come up.
Then since $y \in RR$, we reject H_0 and conclude that the coin is unfair.

Note: Using a computer it can be shown that $P(Y \in RR \mid H_0) = 0.057$ (exactly), which is slightly *higher* than the intended significance 0.05.

If we slightly reduce the rejection region to {0,...,39, 61,...,100}, then $P(Y \in RR) = 0.0352$, which is *lower* than the intended 0.05.

In general, it is impossible to achieve the desired significance level *exactly* when performing a hypothesis test involving a discrete test statistic (like Y).

Another way to write the rejection criterion

The rejection criterion $Y \in \{0,...,40, 60,...,100\}$ can be written in other ways:

$|Y-50| > 9.8$;

$\left|\dfrac{Y-50}{5}\right| > 1.96$;

$\left|\dfrac{Y-50}{\sqrt{100(0.5)(1-0.5)}}\right| > 1.96$;

$\left|\dfrac{\frac{Y}{100} - \frac{50}{100}}{\sqrt{\frac{0.5(1-0.5)}{100}}}\right| > 1.96$;

$\left|\dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right| > z_{\alpha/2}$, where $\hat{p} = Y/n$, $p_0 = 0.5$, $\alpha = 0.05$;

$|Z| > z_{\alpha/2}$, where $Z = \dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

These observations lead us to the 'standard form' for expressing a hypothesis test concerning a single binomial proportion.

_Z_-test for a binomial proportion

Context: $Y \sim Bin(n,p)$, where n is large (> 30, say)
$\hat{p} = Y/n$ (the proportion of successes)

H_0: $p = p_0$

H_a: $p \neq p_0$

TS: $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1)$ (if H_0 is true)

RR: $|Z| > z_{\alpha/2}$

In our example, where 38 heads come up on 100 tosses, the realised value of Z is

$$z = \frac{0.38 - 0.5}{\sqrt{\dfrac{0.5(1-0.5)}{100}}} = -2.4 .$$

Since the absolute value of $z = -2.4$ is greater than $z_{0.025} = 1.96$, we reject $H_0: p = 1/2$ at the 5% level and conclude (as before) that the coin is unfair.

Example 5

A die was tossed 1000 times and 196 sixes came up. Conduct an appropriate hypothesis test at the 1% level to decide whether or not the die is fair.

H_0: $p = 1/6$ (the die is fair) (p = pr of 6 coming up on a single roll)

H_a: $p \neq 1/6$ (the die is unfair)

TS: $Z = \dfrac{\hat{p} - 1/6}{\sqrt{\dfrac{(1/6)(1-1/6)}{1000}}} \dot{\sim} N(0,1)$ if H_0 is true

RR: $|Z| > z_{0.005} = 2.576$ (thus $RR = (-\infty, -2.576) \cup (2.576, \infty)$)

$\hat{p} = y/n = 196/1000 = 0.196$ (an unbiased estimate of p)

$z = \dfrac{0.196 - 1/6}{\sqrt{\dfrac{(1/6)(1-1/6)}{1000}}} = 2.489$ (the realised value of the test statistic Z)

|2.489| < 2.576. Thus $z \notin RR$. So do not reject H_0.

We conclude that the die is fair.

Note: 2.489 is quite close to 2.576. So there is 'moderately strong' evidence for the conjecture that the die is unfair. But that evidence is not quite strong enough to reject the NH at the 1% significance level.

Suppose that the significance level were larger, for example $\alpha = 0.05$. Then the rejection region would be given by $|Z| > z_{0.025} = 1.96$. Since 2.489 > 1.96, we now reject H_0 and conclude that the die is unfair.

Ideally we should decide on the test and significance level *first* and only *then* carry out the experiment. Otherwise unwanted biases may affect the hypothesis test. But this is not commonly done in practice.

In many situations there will be interest in the relationship between *two* binomial proportions. Then we may wish to conduct the following hypothesis test.

Z-test for the difference between two binomial proportions

Context: $X \sim \text{Bin}(n,p)$; $\hat{p} = X/n$

$Y \sim \text{Bin}(m,q)$; $\hat{q} = Y/m$

n and m are both large

$X \perp Y$

H_0: $p - q = \delta$

H_a: $p - q \neq \delta$

TS: Use $Z = \dfrac{(\hat{p}-\hat{q})-\delta}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}+\dfrac{\hat{q}(1-\hat{q})}{m}}} \overset{\cdot}{\sim} N(0,1)$ under H_0 if $\delta \neq 0$

Use $Z = \dfrac{\hat{p}-\hat{q}}{\sqrt{\hat{r}(1-\hat{r})\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}} \overset{\cdot}{\sim} N(0,1)$ under H_0 if $\delta = 0$,

where $\hat{r} = \dfrac{X+Y}{n+m}$

RR: $|Z| > z_{\alpha/2}$

Example 6

100 people were sampled in Sydney, and 63 of these were found to be Liberals.
200 people were sampled in Melbourne, and 102 of these were found to be Liberals.

We're interested in whether the proportion of Liberals in Sydney is the same as the proportion of Liberals in Melbourne.

Carry out an appropriate hypothesis test at the 5% level.

Here: $\hat{p}$ = 63/100 = 0.63, $\hat{q}$ = 102/200 = 0.51

$\hat{r} = (63+102)/(100+200) = 0.55$, α = 0.05

$z_{\alpha/2} = z_{0.025}$ = 1.96, δ = 0.

So the required test is as follows:

$H_0: \quad p - q = 0 \qquad H_a: \quad p - q \neq 0$

RR: $|Z| > 1.96$

$$z = \frac{0.63 - 0.51}{\sqrt{0.55(0.45)\left(\frac{1}{100} + \frac{1}{200}\right)}} = 1.97 \quad \in RR \Rightarrow \text{ reject } H_0.$$

So at the 5% level of statistical significance, the proportions of Liberals in the two cities are different.