CHAPTER

6

HYPOTHESIS TESTING

6.1 Introduction

Suppose that we are going to observe the value of a random vector $\mathbf{X}$. Let $\mathcal{X}$ denote the set of possible values that $\mathbf{X}$ can take and, for $\mathbf{x} \in \mathcal{X}$, let $g(\mathbf{x}|\theta)$ denote the probability that $\mathbf{X} = \mathbf{x}$ where the parameter θ is some unknown element of the set Θ.

Our assumptions specify g, Θ, and $\mathcal{X}$. A hypothesis specifies that θ belongs to some subset Θ_0 of Θ. The question arises as to whether the observed data $\mathbf{x}$ is consistent with the hypothesis that $\theta \in \Theta_0$, often written as $H_0 : \theta \in \Theta_0$. The hypothesis H_0 is usually referred to as the null hypothesis.

In a hypothesis testing situation, two types of error are possible.

- The first type of error is to reject the null hypothesis $H_0 : \theta \in \Theta_0$ as being inconsistent with the observed data $\mathbf{x}$ when, in fact, $\theta \in \Theta_0$ i.e. when, in fact, the null hypothesis happens to be true. This is referred to as type 1 error.
- The second type of error is to fail to reject the null hypothesis $H_0 : \theta \in \Theta_0$ as being inconsistent with the observed data $\mathbf{x}$ when, in fact, $\theta \notin \Theta_0$ i.e. when, in fact, the null hypothesis happens to be false. This is referred to as type 2 error.

Example 6.1.
Suppose the data consist of a random sample $X_1, X_2, \ldots, X_n$ from a $\mathcal{N}(\theta, 1)$ density. Let $\Theta = (-\infty, \infty)$ and $\Theta_0 = (-\infty, 0]$ and consider testing $H_0 : \theta \in \Theta_0$ or in other words $H_0 : \theta \leq 0$.

Solution of Example 6.1. *The standard estimate of θ for this example is $\bar{X}$. It would seem rational to consider that the bigger the value of $\bar{X}$ that we observe the stronger is the evidence against the null hypothesis that $\theta \leq 0$. How big does $\bar{X}$ have to be in order for us to reject H_0 ?*

Suppose that $n = 25$ *and we observe* $\bar{x} = 0.32$. *What are the chances of getting such a large value for* $\bar{x}$ *if, in fact,* $\theta \leq 0$ *? We know that* $\bar{X}$ *has a* $\mathcal{N}(\theta, \frac{1}{n})$ *density i.e. a* $\mathcal{N}(\theta, 0.04)$ *density. So the probability of getting a value for* $\bar{x}$ *as large as* 0.32 *is the area under a* $\mathcal{N}(\theta, 0.04)$ *curve between* 0.32 *and* ∞ *which is, in turn, equal to the area under a* $\mathcal{N}(0,1)$ *curve between* $\frac{0.32-\theta}{0.20}$ *and* ∞. *To evaluate the probability of getting a value for* $\bar{x}$ *as large as* 0.32 *if, in fact,* $\theta \leq 0$ *we need to find the value of* $\theta \leq 0$ *for which the area under a* $\mathcal{N}(0,1)$ *curve between* $\frac{0.32-\theta}{0.20}$ *and* ∞ *is maximised. Clearly this happens for* $\theta = 0$ *and the resulting maximum is the area under a* $\mathcal{N}(0,1)$ *curve between* $\frac{0.32}{0.20} = 1.60$ *and* ∞ *or* 0.0548. *This quantity is called the* p-value. *The p-value is used to measure the strength of the evidence against* $H_0 : \theta \leq 0$ *and* H_0 *is rejected if the p-value is less than some small number such as* 0.05. *You might like to try the* `R` *commands*

```
\texttt{1-pnorm(q=0.32,mean=0,sd=sqrt(.04))}
```

and

```
\texttt{1-pnorm(1.6)}
```

Example 6.2.
Consider the test statistic $T(\mathbf{X}) = \bar{X}$ *and suppose we observe* $T(\mathbf{x}) = t$. *We need to calculate* $p(t;\theta)$ *which is the probability that the random variable* $T(\mathbf{X})$ *exceeds* t *when* θ *is the true value of the parameter. If* θ *is the true value of the parameter* $T(\mathbf{X})$ *has a* $\mathcal{N}(\theta, 1/n)$ *density and so*

$$p(t;\theta) \;=\; P\left\{\mathcal{N}[\theta, 1/n] \geq t\right\} \;=\; P\left\{\mathcal{N}[0,1] \geq \sqrt{n}(t-\theta)\right\}$$

In order to calculate the p-value we need to find $\theta \leq 0$ *for which* $p(t;\theta)$ *is a maximum. Since* $p(t;\theta)$ *is maximised by making* $\sqrt{n}(t-\theta)$ *as small as possible the maximum over* $(-\infty, 0]$ *always occurs at* 0. *Hence we have that P-Value* $= P\{\ \mathcal{N}[0,1] \geq \sqrt{n}t\ \}$

Let us consider more concrete problems and explanations from [5]. Consider the following problems:

1. An engineer has to decide on the basis of sample data whether the true average lifetime of a certain kind of tyre is at least 22000 kilometres.

2. An agronomist has to decide on the basis of experiments whether fertilizer A produces a higher yield of soybeans than fertilizer B.

3. A manufacturer of pharmaceutical products has to decide on the basis of samples whether 90% of all patients given a new medication will recover from a certain disease.

These problems can be translated into the language of **statistical tests of hypotheses**.

1. The engineer has to test the assertion that if the lifetime of the tyre has pdf. $f(x) = \alpha e^{-\alpha x}, x > 0$, then the expected lifetime, $1/\alpha$, is at least 22000.

2. The agronomist has to decide whether $\mu_A > \mu_B$ where μ_A, μ_B are the means of 2 normal distributions.

3. The manufacturer has to decide whether p, the parameter of a binomial distribution is equal to .9.

In each case, it is assumed that the stated distribution correctly describes the experimental conditions, and that the hypothesis concerns the parameter(s) of that distribution. [A more general kind of hypothesis testing problem is where the form of the distribution is unknown.]

In many ways, the formal procedure for hypothesis testing is similar to the scientific method. The scientist formulates a theory, and then tests this theory against observation. In our context, the scientist poses a theory concerning the value of a parameter. He then samples the population and compares observation with theory. If the observations disagree strongly enough with the theory the scientist would probably reject his hypothesis. If not, the scientist concludes either that the theory is probably correct or that the sample he considered did not detect the difference between the actual and hypothesized values of the parameter.

Before putting hypothesis testing on a more formal basis, let us consider the following questions. What is the role of statistics in testing hypotheses? How do we decide whether the sample value disagrees with the scientist's hypothesis? When should we reject the hypothesis and when should we withhold judgement? What is the probability that we will make the wrong decision? What function of the sample measurements should be used to reach a decision? Answers to these questions form the basis of a study of statistical hypothesis testing.

6.2 Terminology and notation

6.2.1 Hypotheses

A statistical hypothesis is an assertion or conjecture about the distribution of a random variable. We assume that the form of the distribution is known so the hypothesis is a statement about the value of a parameter of a distribution.

Let X be a random variable with distribution function $F(x;\theta)$ where $\theta \in \Omega$. That is, Ω is the set of all possible values θ can take, and is called the parameter space. For example, for the binomial distribution, $\Omega = \{p : p \in (0,1)\}$. Let ω be a subset of Ω.

Then a statement such as "$\theta \in \omega$" is a statistical hypothesis and is denoted by H_0. Also, the statement "$\theta \in \overline{\omega}$" (where $\overline{\omega}$ is the complement of ω with respect to Ω) is called the **alternative** to H_0 and is denoted by H_1. We write

$$H_0 : \theta \in \omega \quad \text{and} \quad H_1 : \theta \in \overline{\omega} \text{ (or } \theta \notin \omega).$$

Often hypotheses arise in the form of a claim that a new product, technique, etc. is better than the existing one. In this context, H is a statement that nullifies the claim (or represents the *status quo*) and is sometimes called a **null hypothesis**, but we will refer to it as **the hypothesis**.

If ω contains only one point, that is, if $\omega = \{\theta : \theta = \theta_0\}$ then H_0 is called a **simple hypothesis**. We may write $H_0 : \theta = \theta_0$. Otherwise it is called **composite**. The same applies to alternatives.

6.2.2 Tests of hypotheses

A **test** of a statistical hypothesis is a procedure for deciding whether to "accept" or "reject" the hypothesis. If we use the term "accept" it is with reservation, because it implies stronger action than is really warranted. Alternative phrases such as "reserve judgement," "fail to reject"

perhaps convey the meaning better. A **test** is a rule, or decision function, based on a sample from the given distribution which divides the sample space into 2 regions, commonly called

1. the **rejection region** (or **critical** region), denoted by R;
2. the **acceptance region** (or region of indecision), denoted by $\overline{R}$ (complement of R).

If we compare two different ways of partitioning the sample space then we say we are comparing two tests (of the same hypothesis). For a sample of size n, the sample space is of course n-dimensional and rather than consider R as a subset of n-space, it's helpful to realize that we'll condense the information in the sample by using a statistic (for example $\overline{x}$), and consider the rejection region in terms of the range space of the random variable $\overline{X}$.

6.2.3 Size and power of tests

There are two types of errors that can occur. If we reject H when it is true, we commit a **Type I** error. If we fail to reject H when it is false, we commit a **Type II** error. You may like to think of this in tabular form.

		Our decision	
		do not reject H_0	reject H_0
Actual situation	H_0 is true	correct decision	Type I error
	H_0 is not true	Type II error	correct decision

Probabilities associated with the two incorrect decisions are denoted by

$$\alpha = P(H_0 \text{ is rejected when it is true}) = P(\text{Type I error}) \tag{6.2.1}$$

$$\beta = P(H_0 \text{ is not rejected when it is false}) = P(\text{Type II error}). \tag{6.2.2}$$

The probability α is sometimes referred to as the **size** of the critical region or the **significance level** of the test, and the probability $1 - \beta$ as the **power** of the test.

The roles played by H_0 and H_1 are not at all symmetric. From consideration of potential losses due to wrong decisions, the decision-maker is somewhat conservative for holding the hypothesis as true unless there is overwhelming evidence from the data that it is false. He believes that the consequence of wrongly rejecting H is much more severe to him than of wrongly accepting it.

For example, suppose a pharmaceutical company is considering the marketing of a newly developed drug for treatment of a disease for which the best available drug on the market has a cure rate of 80%. On the basis of limited experimentation, the research division claims that the new drug is more effective. If in fact it fails to be more effective, or if it has harmful side-effects, the loss sustained by the company due to the existing drug becoming obsolete, decline of the company's image, etc., may be quite severe. On the other hand, failure to market a better product may not be considered as severe a loss. In this problem it would be appropriate to consider $H_0 : p = .8$ and $H_1 : p > .8$. Note that H_0 is simple and H_1 is composite.

Ideally, when devising a test, we should look for a decision function which makes probabilities of Type I and Type II errors as small as possible, but, as will be seen in a later example, these depend on one another. For a given sample size, altering the decision rule to decrease one error, results in the other being increased. So, recalling that the Type I error is more serious,

a possible procedure is to hold α fixed at a suitable level (say $\alpha = .05$ or $.01$) and then look for a decision function which minimizes β. The first solution for this was given by Neyman and Pearson for a simple hypothesis versus a simple alternative. It is often referred to as the Neyman-Pearson fundamental lemma.

Example 6.3 (The power function).
Suppose our rule is to reject $H_0 : \theta \leq 0$ *if the p-value is less than* 0.05. *In order for the p-value to be less than* 0.05 *we require* $\sqrt{n}t > 1.65$ *and so we reject* H_0 *if* $\bar{x} > 1.65/\sqrt{n}$. *What are the chances of rejecting* H_0 *if* $\theta = 0.2$ *? If* $\theta = 0.2$ *then* $\bar{x}$ *has a* $\mathcal{N}[0.2, 1/n]$ *density and so the probability of rejecting* H_0 *is*

$$P\left\{\mathcal{N}\left(0.2, \frac{1}{n}\right) \geq \frac{1.65}{\sqrt{n}}\right\} = P\left\{\mathcal{N}(0,1) \geq 1.65 - 0.2\sqrt{n}\right\}.$$

For $n = 25$ *this is given by* $P\{\mathcal{N}(0,1) \geq 0.65\} = 0.2578$. *This calculation can be verified using the* `R` *command* `1-pnorm(1.65-0.2*sqrt(25))`. *The following table gives the results of this calculation for* $n = 25$ *and various values of* θ.

θ:	-1.0	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	
Prob:	.000	.000	.000	.000	.000	.000	.000	.001	.004	.016	
θ:	0.00	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Prob:	0.50	.125	.258	.440	.637	.802	.912	.968	.991	.998	.999

This is called the power function of the test. The `R` *command* `Ns=seq(from=(-1),to=1, by=0.1)` *generates and stores the sequence* $-1.0, -0.9, \ldots, +1.0$ *and the probabilities in the table were calculated using* `1-pnorm(1.65-Ns*sqrt(25))`.

Example 6.4 (Sample size).
How large would n *have to be so that the probability of rejecting* H_0 *when* $\theta = 0.2$ *is* 0.90 *? We would require* $1.65 - 0.2\sqrt{n} = -1.28$ *which implies that* $\sqrt{n} = (1.65 + 1.28)/0.2$ *or* $n = 215$.

So the general plan for testing a hypothesis is clear: choose a test statistic T, observe the data, calculate the observed value t of the test statistic T, calculate the p-value as the maximum over all values of θ in Θ_0 of the probability of getting a value for T as large as t, and reject $H_0 : \theta \in \Theta_0$ if the p-value so obtained is too small.

6.3 Examples

Example 6.5.
Suppose that random variable X *has a normal distribution with mean* μ *and variance 4. Test the hypothesis that* $\mu = 1$ *against the alternative that* $\mu = 2$, *based on a sample of size 25.*

Solution of Example 6.2. *An unbiased estimate of* μ *is* $\overline{X}$ *and we know that* $\overline{X}$ *is distributed normally with mean* μ *and variance* σ^2/n *which in this example is 4/25. We note that values of* $\overline{x}$ *close to 1 support* H *whereas values of* $\overline{x}$ *close to 2 support* A. *We could make up a decision rule as follows:*

- *If* $\overline{x} > 1.6$ *claim that* $\mu = 2$,
- *If* $\overline{x} \leq 1.6$ *claim that* $\mu = 1$.

The diagram in Fig. 6.3.1 shows the sample space of $\overline{x}$ *partitioned into*

1. *the critical region,* $R = \{\overline{x} : \overline{x} > 1.6\}$,
2. *the acceptance region,* $\overline{R} = \{\overline{x} : \overline{x} \leq 1.6\}$.

Here, 1.6 is the critical value of $\overline{x}$.

We will find the probability of Type I and Type II error,

$$P\left(\overline{X} > 1.6 | \mu = 1, \sigma = \frac{2}{5}\right) = .0668$$

with

```
pnorm(q=1.6,mean=1,sd=0.4,lower.tail=F)
```

This is

$$P(H_0 \text{ is rejected} | H_0 \text{ is true}) = P(\text{Type I error}) = \alpha$$

Also

$$\begin{aligned}\beta = P(\text{Type II error}) &= P(H_0 \text{ is not rejected} | H_0 \text{ is false}) \\ &= P\left(\overline{X} \leq 1.6 | \mu = 2, \sigma = \frac{2}{5}\right) \\ &= .1587\end{aligned}$$

with

```
(pnorm(q=1.6,mean=2,sd=0.4,lower.tail=T))
```

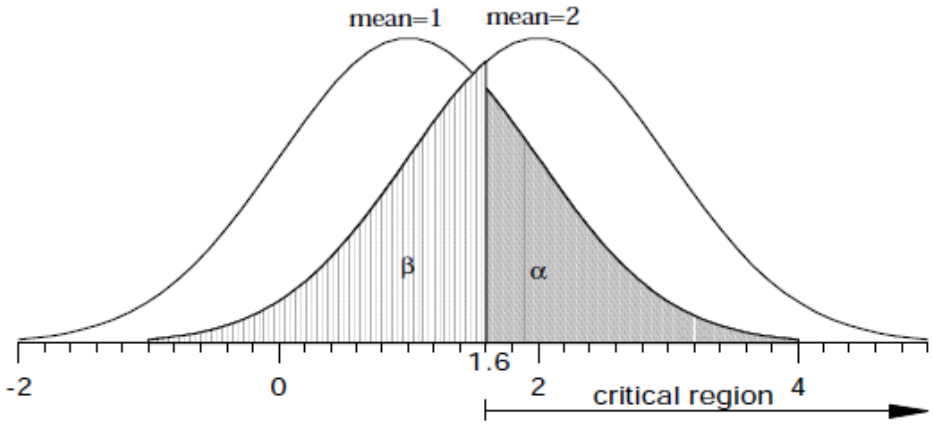


Figure 6.3.1: Critical Region – Upper Tail

To see how the decision rule could be altered so that $\alpha = .05$, *let the critical value be* c. *We require*

$$P\left(\overline{X} > c | \mu = 1, \sigma = \frac{2}{5}\right) = 0.05$$

$$\Rightarrow c = 1.658 \quad (\texttt{qnorm(p=0.05,mean=1,sd=0.4,lower.tail=T)})$$

$$P\left(\overline{X} < c | \mu = 2, \sigma = \frac{2}{5}\right) = 0.196 \quad (\texttt{pnorm(q=1.658,mean=2,sd=0.4,lower.tail=T)})$$

This value of c *gives an* α *of* 0.05 *and a* β *of* 0.196 *illustrating that as one type of error* (α) *decreases the other* (β) *increases.*

Example 6.6.
Suppose that we have a random sample of size n from a $N(\mu, 4)$ distribution and wish to test $H_0 : \mu = 10$ against $H_1 : \mu = 8$. The decision rule is to reject H_0 if $\overline{x} < c$. We wish to find n and c so that $\alpha = 0.05$ and $\beta \approx 1$.

Solution of Example 6.3. *In Fig. 6.3.2 below, the left curve is $f(\overline{x}|H_1)$ and the right curve is $f(\overline{x}|H_0)$. The critical region is $\{\overline{x} : \overline{x} < c\}$, so α is the left shaded area and β is the right shaded area. Now*

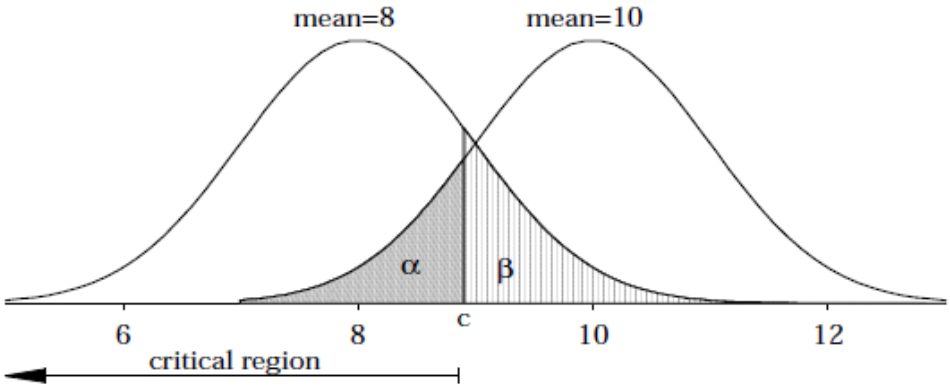


Figure 6.3.2: Critical Region – Lower tail

$$\alpha = 0.05 = P\left(\overline{X} < c|\mu = 10, \sigma = \frac{2}{\sqrt{n}}\right) \tag{6.3.1}$$

$$\beta = 0.1 = P\left(\overline{X} \geq c|\mu = 8, \sigma = \frac{2}{\sqrt{n}}\right) \tag{6.3.2}$$

We need to solve these two equations simultaneously for n as shown in Fig. 6.3.3. The R code for the above diagram is

```
n <- 3:12
alpha <- 0.05
beta <- 0.1
Acrit <- qnorm(mean=10,sd=2/sqrt(n),p=alpha)
Bcrit <- qnorm(mean=8,sd=2/sqrt(n),p=beta,lower.tail=F)

plot(Acrit ~ n,type="",xlab="sample size",ylab="Critical value",las=1,ylim=c(7,10) ,lwd=
lines(n,Bcrit,lty=2,lwd=2)
```

A sample size $n = 9$ and critical value $c = 8.9$ gives $\alpha \approx 0.05$ and $\beta \approx 0.1$.

6.4 One-sided and two-sided Tests

Consider the problem where the random variable X has a binomial distribution with $P(\text{Success}) = p$. How do we test the hypothesis $p = 0.5$. Firstly, note that we have an experiment where the outcome on an individual trial is *success* or *failure* with probabilitites p and q respectively. Let us repeat the experiment n times and observe the number of successes.

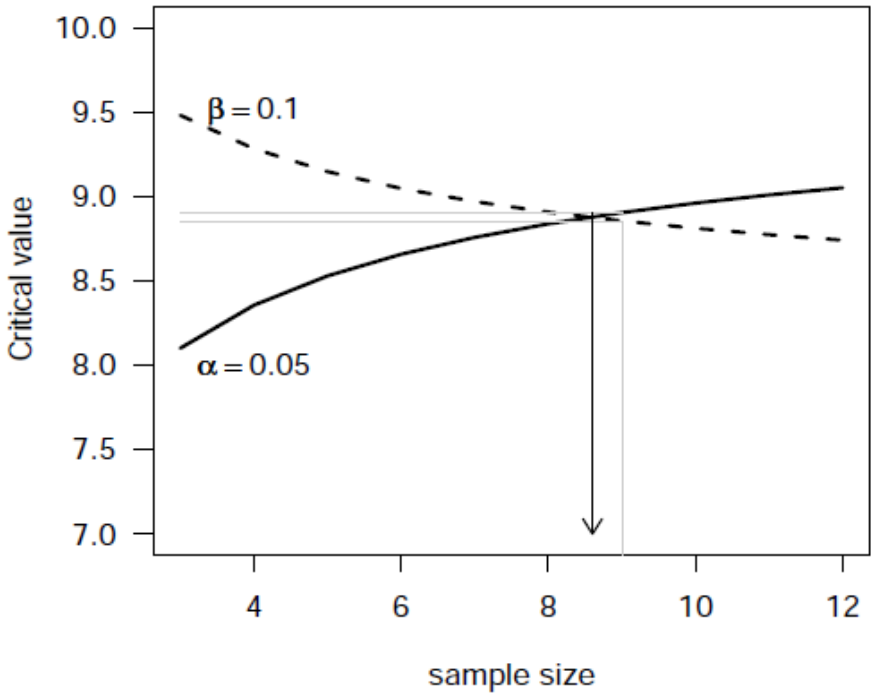


Figure 6.3.3: Solution for size and power of test

Before continuing with this example it is useful to note that in most hypothesis testing problems we will deal with, H_0 is simple, but H_1 on the other hand, is composite, indicating that the parameter can assume a range of values. Examples 1 and 2 were more straightforward in the sense that H_1 was simple also.

If the range of possible parameter values lies entirely on the one side of the hypothesized value, the aternative is said to be **one-sided**. For example, $H_1 : p > .5$ is one-sided but $H_1 : p \neq .5$ is two-sided. In a real-life problem, the decision of whether to make the alternative one-sided or two-sided is not always clear cut. As a general rule-of-thumb, if parameter values in only one direction are physically meaningful, or are the only ones that are possible, the alternative should be one-sided. Otherwise, H_1 should be two-sided. Not all statisticians would agree with this rule.

The next question is what test statistic we use to base our decision on. In the above problem, since X/n is an unbiased estimator of p, that would be a possibility. We could even use X itself. In fact the latter is more suitable since its distribution is known. Recall that, the principle of hypothesis testing is that we will assume H_0 is correct, and our position will change only if the data show **beyond all reasonable doubt** that H_1 is true. The problem then is to define in quantitative terms what reasonable doubt means. Let us suppose that $n = 18$ in our problem above. Then the range space for X is $R_X = \{0, 1, \ldots, 18\}$ and $E(X) = np = 9$ **if** H_0 is true. If the observed number of successes is close to 9 we would be obliged to think that H was true. On the other hand, if the observed value of X was 0 or 18 we would be fairly sure that H_0 was not true. Now **reasonable doubt** does not have to be as extreme as 18 cases out of 18. Somewhere between x-values of 9 and 18 (or 9 and 0), there is a point, c say, when for all practical purposes

the credulity of H_0 ends and reasonable doubt begins. This point is called the **critical value** and it completely determines the decision-making process. We could make up a decision rule

- If $x \geq c$, reject H_0.
- If $x < c$, conclude that H_0 is probably correct.

In this case, $\{x : x \geq c\}$ is the rejection region.

We will consider appropriate tests for both one- and two-sided alternatives in the problem above.

6.4.1 Case (a): Alternative is one-sided

In the above problem, suppose that the alternative is $H_1 : p > .5$. Only values of x much **larger** than 9 would support this alternative and a decision rule such as the one we just mentioned would be appropriate. The actual value of c is chosen to make α, the size of the critical region, suitably small. For example, if $c = 11$, then $P(X \geq 11) = .24$ and this of course is too large. Clearly we should look for a value closer to 18. If $c = 15$, $P(X \geq 15) = \sum_{x=15}^{18} \binom{18}{x}(.5)^{18} = 0.004$, on calculation. We may now have gone too far in the other extreme. Requiring 15 or more successes out of 18 before we reject $H_0 : p = 0.5$ means that only 4 times in a thousand would we reject H_0 wrongly. Over the years, a reasonable consensus has been reached as to how much evidence against H_0 is enough evidence. In many situations we define the beginning of **reasonable doubt** as the value of the test statistic that is equalled or exceeded by chance 5% of the time when H_0 is true. According to this criterion, c should be chosen so that $P(X \geq c|H_0 \text{ is true}) = 0.05$ That is c should satisfy

$$P(X \geq c|p = 0.5) = 0.05 = \sum_{x=c}^{1} 8\binom{18}{x}(0.5)^{18}.$$

A little trial and error shows that $c = 13$ is the appropriate value. Of course because of the discrete nature of X it will not be possible to obtain an α of exactly 0.05.

Defining the critical region in terms of the x-value that is exceeded only 5% of the time when H_0 is true is the most common way to quantify reasonable doubt, but there are others. The figure 1% is frequently used and if the critical value is exceeded only 1% of the time we say there is **strong evidence** against H_0. If the critical value is only exceeded .1% of the time we may say that there is **very strong evidence** against H_0.

So far we have considered a one-sided alternative. Now we'll consider the other case where the alternative is two-sided.

6.4.2 Case (b): Two-sided Alternative

Consider now the alternative $H_1 : p \neq 0.5$. Values of x too large or too small would support this alternative. In this case there are two critical regions (or more correctly, the critical region consists of two disjoint sets), one in each 'tail' of the distribution of X. For a 5% critical region, there would be two critical values c_1 and c_2 such that

$$P(X \leq c_1|H_0 \text{ is true}) \approx 0.025 \quad \text{and} \quad P(X \geq c_2|H_0 \text{ is true}) \approx 0.025.$$

This can be seen in Fig. 6.4.1 below, where the graph is of the distribution of X when H_0 is true. (It can be shown that $c_1 = 4$ and $c_2 = 14$ are the critical values in this case.)

Tests with a one-sided critical region are called **one-tailed** tests, whereas those with a two-sided critical region are called **two-tailed tests**.

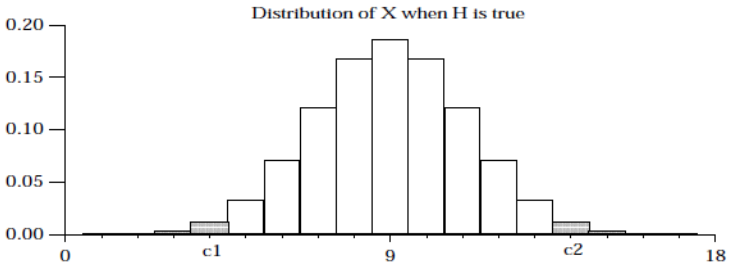


Figure 6.4.1: Critical Region – Twosided Alternative

Computer Exercise 6.1. *Use a simulation approach to estimate a value for c in*

- *If $x \geq c$, reject H_0.*
- *If $x < c$, conclude that H_0 is probably correct.*

Solution of Computer Exercise 6.1. *Use the commands*

```
#Generate 1000 random variables fron a bin(18,0.5) distribution.
rb <- rbinom(n=1000,size=18,p=0.5)
table(rb) #Tabulate the results
1 2 3  4  5  6   7   8   9  10  11 12 13 14 15
1 1 3 11 28 77 125 174 187 166 126 63 22 11  5
```

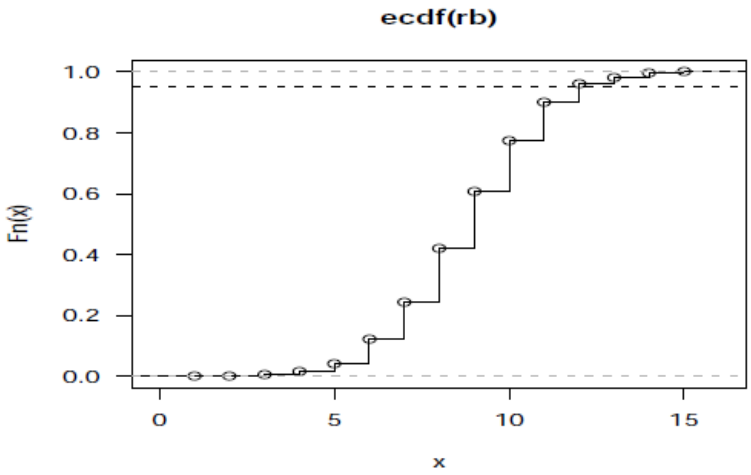


Figure 6.4.2: Empirical cumulative distribution function for binomial `rv`'s

This would indicate the onesided critcal value should be $c = 13$ as the estimate of $P(X \geq 13)$ is 0.038. *For a two-sided test the estimated critical values are $c_1 = 4$ and $c_2 = 13$.*

These results from simulation are in close agreement with theoretical results obtained in the two preceding subsections.

6.4.3 Two approaches to hypothesis testing

It is worthwhile considering a definite procedure for hypothesis testing problems. There are two possible approaches.

1. See how the observed value of the statistic compares with that expected **if H_0 is true**. Find the probability, assuming H_0 to be true, of this event or others more extreme, that is, further still from the expected value. For a two-tailed test this will involve considering extreme values *in either direction*. If this probability is small (say, < 0.05), the event is an unlikely one if H_0 is true. So if such an event has occurred, doubt would be cast on the hypothesis.

2. Make up a decision rule by partitioning the sample space of the statistic into a critical region, R, and its complement $\overline{R}$, choosing the critical value (or two critical values in the case of a two- tailed test) c, in such a way that $\alpha = 0.05$. We then note whether or not the observed value lies in this critical region, and draw the corresponding conclusion.

Example 6.7.
Suppose we want to know whether a given die is biased towards 5 or 6 or whether it is "true." To examine this problem the die is tossed 9000 times and it is observed that on 3164 occasions the outcome was 5 or 6.

Solution of Example 6.4. *Let X be the number of successes (5's or 6's) in 9000 trials. Then if $p = P(S)$, X is distributed* $\text{bin}(9000, p)$. *As is usual in hypothesis testing problems, we set up H_0 as the hypothesis we wish to "disprove." In this case, it is that the die is "true," that is, $p = 1/3$. If H_0 is not true, the alternative we wish to claim is that the die is biased towards 5 or 6, that is $p > 1/3$. In practice, one decides on this alternative before the experiment is carried out. We will consider the 2 approaches mentioned above.*

Approach (i), probabilities *If $p = 1/3$ and $N = 9000$ then $E(X) = np = 3000$ and $Var(X) = npq = 2000$. The observed number of successes, 3164, was greater than expected if H_0 were true. So, assuming $p = 1/3$, the probability of the observed event together with others more extreme (that is, further still from expectation) is*

$$P_B(X \geq 3164 | p = 1/3) = 0.0001$$

as

```
pbinom(q=3164,size=9000,prob=1/3,lower.tail=F)
```

This probability is small, so the event $X \geq 3164$ is an unlikely one if the assumption we've made ($p = 1/3$) is correct. Only about 1 times in 10000 would we expect such an occurence. Hence, if such an event did occur, we'd doubt the hypothesis and conclure that there is evidence that $p > 1/3$.

Approach (ii), quantiles *Clearly, large values of X support H_1, so we'd want a critical region of the form $x \geq c$ where c is chosen to give the desired significance level, α. That is, for $\alpha = 0.05$, say, the upper tail 5% quantile of the binomial distributio with $p = 1/3$ and $N = 9000$ is 3074 as*

```
qbinom(size=N,prob=px,p=0.05,lower.tail=F)
```

The observed value 3164 exceeds this and thus lies in the critical region $[c, \infty]$. *So we reject* H_0 ***at the 5% significance level***. *That is, we will come to the conclusion that* $p > 1/3$, *but in so doing, we'll recognize the fact that the probability could be as large as 0.05 that we've rejected* H_0 *wrongly.*

The 2 methods are really the same thing. Figure 6.4.3 with the observed quantile 3164 and associated with it is $P(X > 3164)$. *The dashed lines show the upper* $\alpha = 0.05$ *probability and the quantile* $C_{1-\alpha}$. *The event that* $X > C_{1-\alpha}$ *has a probability* $p < \alpha$.

The rejection region can be defined either by the probabilities or the quantiles.

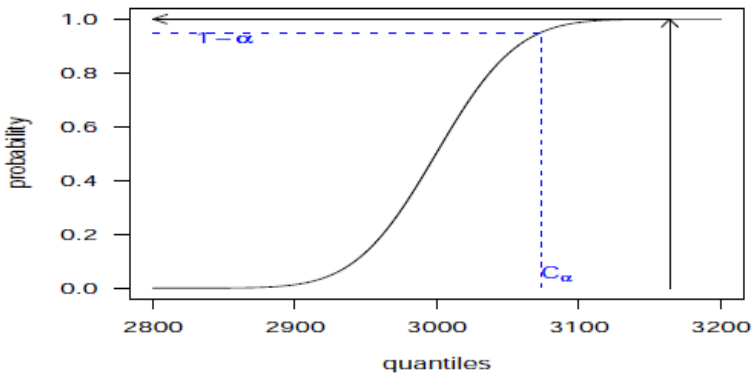


Figure 6.4.3: using either quantiles or probability to test the null hypothesis

In doing this sort of problem it helps to draw a diagram, or at least try to visualize the partitioning of the sample space as suggested in Figure 6.4.4.

If $x \in R$ *it seems much more likely that the actual distribution of* X *is given by a curve similar to the one on the right hand side, with mean somewhat greater than* 3000.

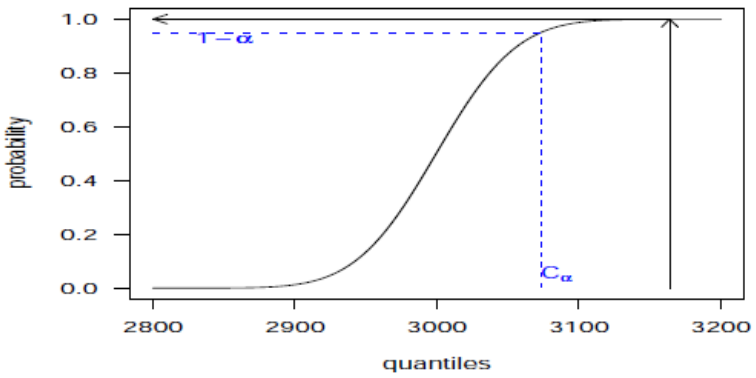


Figure 6.4.4: One Sided Alternative – Binomial

Computer Exercise 6.2. *The following random sample was drawn from a normal distribution with* $\sigma = 5$. *Test the hypothesis that* $\mu = 23$.

Solution of Computer Exercise 6.2.

18	14	23	23	18
21	22	16	21	28
12	19	22	15	18
28	24	22	18	13
18	16	24	26	35

```
x <- c(18,14,23,23,18,21,22,16,21,28,12,19,22,15,18,28,24,22,18,13,18,16,24,26,35)
xbar <- mean(x)
n <- length(x)
> xbar
[1] 20.56
pnorm(q=xbar, mean=23,sd=5/sqrt(n))
[1] 0.007
qnorm(p=0.05,mean=23,sd=5/sqrt(n))
[1] 21
```

We can now use approach (i). For a two sided alternative calculated probability is $P = 0.015 = 2 \times 0.00734$ so that the hypothesis is unlikely to be true.

For approach (ii) with $\alpha = 0.05$ the critical value is 21. The conclusion reached would therefore be the same by both approaches.

For testing $\mu = 23$ against the one-sided alternative $\mu < 23$, $P = 0.0073$.

Example 6.8 (One Gaussian sample).
Suppose that we have data $X_1, X_2, \ldots, X_n$ which are iid observations from a $\mathcal{N}(\mu, \sigma^2)$ density where both μ and σ are unknown. Here $\theta = (\mu, \sigma)$ and $\Theta = \{(\mu, \sigma) : -\infty < \mu < \infty, 0 < \sigma < \infty\}$. Define

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{and} \qquad s^2 = \frac{\sum_{i=1}^{\mathcal{N}} (X_i - \bar{X})^2}{n-1}.$$

(a) Suppose $\Theta_0 = \{(\mu, \sigma) : -\infty < \mu \leq A, 0 < \sigma < \infty\}$. Define $T = \bar{X}$. Let t denote the observed value of T. Then

$$\begin{aligned} p(t;\theta) &= P[\bar{X} \geq t] \\ &= P\left[\frac{\sqrt{n}(\bar{X} - \mu)}{s} \geq \frac{\sqrt{n}(t-\mu)}{s}\right] \\ &= P\left[t_{n-1} \geq \frac{\sqrt{n}(t-\mu)}{s}\right]. \end{aligned}$$

To maximize this we choose μ in $(-\infty, A]$ as large as possible which clearly means choosing $\mu = A$. Hence the p-value is

$$P\left[t_{n-1} \geq \frac{\sqrt{n}(\bar{x} - A)}{s}\right].$$

(b) Suppose $\Theta_0 = \{(\mu, \sigma) : A \leq \mu < \infty, 0 < \sigma < \infty\}$. Define $T = -\bar{X}$. Let t denote the observed value of T. Then

$$\begin{aligned} p(t;\theta) &= P[-\bar{X} \geq t] \\ &= P[\bar{X} \leq -t] \\ &= P\left[\frac{\sqrt{n}(\bar{X} - \mu)}{s} \leq \frac{\sqrt{n}(-t-\mu)}{s}\right] \\ &= P\left[t_{n-1} \leq \frac{\sqrt{n}(-t-\mu)}{s}\right]. \end{aligned}$$

To maximize this we choose μ in $[A, \infty)$ as small as possible which clearly means choosing $\mu = A$. Hence the p-value is

$$P\left[t_{n-1} \leq \frac{\sqrt{n}(-t-A)}{s}\right] = P\left[t_{n-1} \leq \frac{\sqrt{n}(\bar{x}-A)}{s}\right].$$

(c) Suppose $\Theta_0 = \{(A, \sigma) : 0 < \sigma < \infty\}$. Define $T = |\bar{X} - A|$. Let t denote the observed value of T. Then

$$\begin{aligned} p(t;\theta) &= P[|\bar{X}-A| \geq t] = P[\bar{X} \geq A+t] + P[\bar{X} \leq A-t] \\ &= P\left[\frac{\sqrt{n}(\bar{X}-\mu)}{s} \geq \frac{\sqrt{n}(A+t-\mu)}{s}\right] \\ &\quad + P\left[\frac{\sqrt{n}(\bar{X}-\mu)}{s} \leq \frac{\sqrt{n}(A-t-\mu)}{s}\right] \\ &= P\left[t_{n-1} \geq \frac{\sqrt{n}(A+t-\mu)}{s}\right] \\ &\quad + P\left[t_{n-1} \leq \frac{\sqrt{n}(A-t-\mu)}{s}\right]. \end{aligned}$$

The maximization is trivially found by setting $\mu = A$. Hence the p-value is

$$P\left[t_{n-1} \geq \frac{\sqrt{n}t}{s}\right] + P\left[t_{n-1} \leq \frac{-\sqrt{n}t}{s}\right] = 2P\left[t_{n-1} \geq \frac{\sqrt{n}t}{s}\right] = 2P\left[t_{n-1} \geq \frac{\sqrt{n}|\bar{x}-A|}{s}\right].$$

(d) Suppose $\Theta_0 = \{(\mu, \sigma) : -\infty < \mu < \infty, 0 < \sigma \leq A\}$. Define $T = \sum_{i=1}^n (X_i - \bar{X})^2$. Let t denote the observed value of T. Then

$$\begin{aligned} p(t;\sigma) &= P\left[\textstyle\sum_{i=1}^n (X_i-\bar{X})^2 \geq t\right] \\ &= P\left[\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma^2} \geq \frac{t}{\sigma^2}\right] \\ &= P\left[\chi^2_{n-1} \geq \tfrac{t}{\sigma^2}\right]. \end{aligned}$$

To maximize this we choose σ in $(0, A]$ as large as possible which clearly means choosing $\sigma = A$. Hence the p-value is

$$P\left[\chi^2_{n-1} \geq \frac{t}{A^2}\right] = P\left[\chi^2_{n-1} \geq \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{A^2}\right]$$

(e) $\Theta_0 = \{(\mu, \sigma) : -\infty < \mu < \infty, A \leq \sigma < \infty\}$. Define $T = [\sum_{i=1}^n (x_i - \bar{x})^2]^{-1}$, and let t denote the observed value of T. Then

$$\begin{aligned} p(t;\sigma) &= P\left[\frac{1}{\sum_{i=1}^n (X_i-\bar{X})^2} \geq t\right] \\ &= P\left[\sum_{i=1}^n (X_i-\bar{X})^2 \leq \frac{1}{t}\right] \\ &= P\left[\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma^2} \leq \frac{1}{t\sigma^2}\right] \\ &= P\left[\chi^2_{n-1} \leq \frac{1}{t\sigma^2}\right]. \end{aligned}$$

To maximize this we choose σ in $[A, \infty)$ as small as possible which clearly means choosing $\sigma = A$. Hence the p-value is

$$P\left[\chi^2_{n-1} \leq \frac{1}{tA^2}\right] = P\left[\chi^2_{n-1} \leq \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{A^2}\right].$$

(f) Suppose $\Theta_0 = \{(\mu, A) : -\infty < \mu \leq \infty\}$. Define

$$T = \max\left\{\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{A^2}, \frac{A^2}{\sum_{i=1}^n (X_i-\bar{X})^2}\right\}.$$

Let t denote the observed value of T and note that t must be greater than or equal to 1. Then

$$\begin{aligned} p(t;\sigma) &= P\left[\sum_{i=1}^n (X_i-\bar{X})^2 \geq A^2 t\right] + P\left[\sum_{i=1}^n (X_i-\bar{X})^2 \leq \tfrac{A^2}{t}\right] \\ &= P\left[\tfrac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma^2} \geq \tfrac{A^2 t}{\sigma^2}\right] + P\left[\tfrac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma^2} \leq \tfrac{A^2}{t\sigma^2}\right] \\ &= P\left[\chi^2_{n-1} \geq \tfrac{A^2 t}{\sigma^2}\right] + P\left[\chi^2_{n-1} \leq \tfrac{A^2}{t\sigma^2}\right]. \end{aligned}$$

Since A is the only element in Θ_0, the maximization is trivially found by setting $\sigma = A$. Hence the p-value is

$$P[\chi^2_{n-1} \geq t] + P\left[\chi^2_{n-1} \leq \frac{1}{t}\right].$$

where $t = \max\left\{\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{A^2}, \frac{A^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right\}$.

6.5 Two-sample problems

In this section we will consider problems involving sampling from two populations where the hypothesis is a statement of equality of two parameters. The two problems are:

1. Test $H_0 : \mu_1 = \mu_2$ where μ_1 and μ_2 are the means of two normal populations.
2. Test $H_0 : p_1 = p_2$ where p_1 and p_2 are the parameters of two binomial populations.

Example 6.9.
Given independent random samples $X_1, X_2, \ldots, X_{n_1}$ from a normal population with unknown mean μ_1 and known variance σ_1^2 and $Y_1, Y_2, \ldots, Y_{n_2}$ from a normal population with unknown mean μ_2 and known variance σ_2^2, derive a test for the hypothesis $H : \mu_1 = \mu_2$ against one-sided and two-sided alternatives.

Solution of Example 6.5. *Note that the hypothesis can be written as $H : \mu_1 - \mu_2 = 0$. An unbiased estimator of $\mu_1 - \mu_2$ is $\overline{X} - \overline{Y}$ so this will be used as the test statistic. Its distribution is given by*

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

or, in standardized form, ***if*** *H_0* ***is true***

$$\frac{\overline{X} - \overline{Y}}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0,1).$$

For a two-tailed test (corresponding to $H_1 : \mu_1 - \mu_2 \neq 0$) we have a rejection region of the form

$$\frac{|\overline{x} - \overline{y}|}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} > c \tag{6.5.1}$$

where $c = 1.96$ for $\alpha = .05$, $c = 2.58$ for $\alpha = .01$, etc.

For a one-tailed test we have a rejection region

$$\frac{\overline{x} - \overline{y}}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} > c \text{ for } H_1 : \mu_1 - \mu_2 > 0 \tag{6.5.2}$$

$$< -c \text{ for } H_1 : \mu_1 - \mu_2 < 0 \tag{6.5.3}$$

where $c = 1.645$ for $\alpha = .05$, $c = 2.326$ for $\alpha = .01$, etc. Can you see what modification to make to the above rejection regions for testing $H_0 : \mu_1 - \mu_2 = \delta_0$ for some specified constant other than zero?

Example 6.10.
Suppose that n_1 Bernoulli trials where $P(S) = p_1$ resulted in X successes and that n_2 Bernoulli trials where $P(S) = p_2$ resulted in Y successes. How do we test $H : p_1 = p_2$ $(= p$, say)?

Solution of Example 6.6. *Note that H_0 can be written $H_0 : p_1 - p_2 = 0$. Now X is distributed as* $\text{bin}(n_1, p_1)$ *and Y is distributed as* $\text{bin}(n_2, p_2)$ *and we have seen earlier that unbiased estimates of p_1, p_2, are respectively*

$$\overline{p_1} = x/n_1, \overline{p_2} = y/n_2,$$

so an appropriate statistic to use to estimate $p_1 - p_2$ is $\frac{X}{n_1} - \frac{Y}{n_2}$.

For n_1, n_2 large, we can use the Central Limit Theorem to observe that

$$\frac{\frac{X}{n_1} - \frac{Y}{n_2} - E\left[\frac{X}{n_1} - \frac{Y}{n_2}\right]}{\sqrt{\text{Var}\left[\frac{X}{n_1} - \frac{Y}{n_2}\right]}} \sim \textit{approximately } N(0,1) \tag{6.5.4}$$

and

$$E\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) = \frac{n_1 p_1}{n_1} - \frac{n_2 p_2}{n_2} = 0 \textit{ under } H_0, \textit{ and} \tag{6.5.5}$$

$$\text{Var}\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) = \frac{n_1 p_1 q_1}{n_1^2} + \frac{n_2 p_2 q_2}{n_2^2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \textit{ under } H_0 \tag{6.5.6}$$

In (6.5.4) the variance is unknown, but we can replace it by an estimate and it remains to decide what is the best estimate to use. For the binomial distribution, the MLE of p is

$$\overline{p} = \frac{X}{n} = \frac{\textit{number of successes}}{\textit{number of trials}}$$

In our case, we have 2 binomial distributions with the same probability of success under H_0, so intuitively it seems reasonable to "pool" the 2 samples so that we have $X + Y$ successes in $n_1 + n_2$ trials. So we will estimate p by

$$\overline{p} = \frac{x+y}{n_1+n_2}.$$

Using this in (6.5.4) we can say that to test $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ at the $100\alpha\%$ significance level, H_0 is rejected if

$$\frac{|(x/n_1) - (y/n_2)|}{\sqrt{\left(\frac{x+y}{n_1+n_2}\right)\left(1 - \frac{x+y}{n_1+n_2}\right)\left(\frac{n_1+n_2}{n_1 n_2}\right)}} > z_{\alpha/2}. \tag{6.5.7}$$

Of course the appropriate modification can be made for a one-sided alternative.

Example 6.11 (Two Gaussian samples).
Suppose that we have data $X_1, X_2, \ldots, X_n$ which are iid observations from a $\mathcal{N}(\mu_1, \sigma^2)$ density and data $y_1, y_2, \ldots, y_m$ which are iid observations from a $\mathcal{N}(\mu_2, \sigma^2)$ density where μ_1, μ_2, and σ are unknown.

Here

$$\theta = (\mu_1, \mu_2, \sigma)$$

and

$$\Theta = \{(\mu_1, \mu_2, \sigma) : -\infty < \mu_1 < \infty, -\infty < \mu_2 < \infty, 0 < \sigma < \infty\}.$$

Define

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \Sigma_{j=1}^{m}(y_j - \bar{y})^2}{n + m - 2}.$$

(a) Suppose $\Theta_0 = \{(\mu_1, \mu_2, \sigma) : -\infty < \mu_1 < \infty,\ \mu_1 < \mu_2 < \infty,\ 0 < \sigma < \infty\}$. *Define* $T = \bar{x} - \bar{y}$. *Let* t *denote the observed value of* T. *Then*

$$\begin{aligned} p(t;\theta) &= P[\bar{x} - \bar{y} \geq t] \\ &= P\left[\frac{[(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}} \geq \frac{[t - (\mu_1 - \mu_2)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right] \\ &= P\left[t_{n+m-2} \geq \frac{[t - (\mu_1 - \mu_2)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right]. \end{aligned}$$

To maximize this we choose $\mu_2 > \mu_1$ *in such a way as to maximize the probability which clearly implies choosing* $\mu_2 = \mu_1$. *Hence the p-value is*

$$P\left[t_{n+m-2} \geq \frac{t}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right] = P\left[t_{n+m-2} \geq \frac{\bar{x} - \bar{y}}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right].$$

(b) Suppose $\Theta_0 = \{(\mu_1, \mu_2, \sigma) : -\infty < \mu_1 < \infty, -\infty < \mu_2 < \mu_1, 0 < \sigma < \infty\}$. *Define* $T = \bar{y} - \bar{x}$. *Let* t *denote the observed value of* T. *Then*

$$\begin{aligned} p(t;\theta) &= P[\bar{y} - \bar{x} \geq t] \\ &= P\left[\frac{[(\bar{y} - \bar{x}) - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}} \geq \frac{[t - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right] \\ &= P\left[t_{n+m-2} \geq \frac{[t - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right]. \end{aligned}$$

To maximize this we choose $\mu_2 < \mu_1$ *in such a way as to maximize the probability which clearly implies choosing* $\mu_2 = \mu_1$. *Hence the p-value is*

$$P\left[t_{n+m-2} \geq \frac{t}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right] = P\left[t_{n+m-2} \geq \frac{\bar{y} - \bar{x}}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right].$$

(c) $\Theta_0 = \{(\mu, \mu, \sigma) : -\infty < \mu < \infty,\ 0 < \sigma < \infty\}$. *Define* $T = |\bar{y} - \bar{x}|$. *Let* t *denote the observed value of* T. *Then*

$$\begin{aligned} p(t;\theta) &= P[|\bar{y} - \bar{x}| \geq t] \\ &= P[\bar{y} - \bar{x} \geq t] + P[\bar{y} - \bar{x} \leq -t] \\ &= P\left[\frac{[(\bar{y} - \bar{x}) - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}} \geq \frac{[t - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right] \\ &\quad + P\left[\frac{[(\bar{y} - \bar{x}) - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}} \leq \frac{[-t - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right] \\ &= P\left[t_{n+m-2} \geq \frac{[t - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right] + \\ &\quad P\left[t_{n+m-2} \leq \frac{[-t - (\mu_2 - \mu_1)]}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}\right]. \end{aligned}$$

Since, for all sets of parameter values in Θ_0, *we have* $\mu_1 = \mu_2$ *the maximization is trivial and so the p-value is*

$$\begin{aligned} P\left[t_{n+m-2} \geq \frac{t}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}}\right] + P\left[t_{n+m-2} \leq \frac{-t}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}}\right] \\ = 2P\left[t_{n+m-2} \geq \frac{t}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}}\right] \\ = 2P\left[t_{n+m-2} \geq \frac{|\bar{y} - \bar{x}|}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}}\right]. \end{aligned}$$

(d) Suppose that we have data $X_1, X_2, \ldots, X_n$ *which are iid observations from a* $\mathcal{N}(\mu_1, \sigma_1^2)$ *density and data* $y_1, y_2, \ldots, y_m$ *which are iid observations from a* $\mathcal{N}(\mu_2, \sigma_1^2)$ *density where* $\mu_1, \mu_2, \sigma_1,$ *and* σ_2 *are all unknown. Here* $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ *and* $\Theta = \{(\mu_1, \mu_2, \sigma_1, \sigma_2) : -\infty < \mu_1 < \infty, -\infty < \mu_2 < \infty, 0 < \sigma_1 < \infty, 0 < \sigma_2 < \infty\}$. *Define*

$$s_1^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}, \qquad \text{and} \qquad s_2^2 = \frac{\Sigma_{j=1}^m (y_j - \bar{y})^2}{m-1}.$$

Suppose $\Theta_0 = \{(\mu_1, \mu_2, \sigma, \sigma) : -\infty < \mu_1 < \infty, \mu_1 < \mu_2 < \infty, 0 < \sigma < \infty\}$.

Define

$$T = \max\left\{\frac{s_1^2}{s_2^2}, \frac{s_2^2}{s_1^2}\right\}.$$

Let t *denote the observed value of* T *and observe that* t *must be greater than or equal to 1. Then*

$$\begin{aligned} p(t;\theta) &= P\left[\frac{s_1^2}{s_2^2} \geq t\right] + P\left[\frac{s_2^2}{s_1^2} \geq t\right] \\ &= P\left[\frac{s_1^2}{s_2^2} \geq t\right] + P\left[\frac{s_1^2}{s_2^2} \leq \frac{1}{t}\right] \\ &= P\left[\frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} \geq \frac{\sigma_2^2 t}{\sigma_1^2}\right] + P\left[\frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} \leq \frac{\sigma_2^2}{\sigma_1^2 t}\right] \\ &= P\left[F_{n-1,m-1} \geq \frac{\sigma_2^2 t}{\sigma_1^2}\right] + P\left[F_{n-1,m-1} \leq \frac{\sigma_2^2}{\sigma_1^2 t}\right]. \end{aligned}$$

Since, for all sets of parameter values in Θ_0, we have $\sigma_1 = \sigma_2$ the maximization is trivial and so the p-value is $P\left[F_{n-1,m-1} \geq t\right] + P\left[F_{n-1,m-1} \leq \frac{1}{t}\right]$.

6.6 Connection between hypothesis testing and CI's

Consider the problem where we have a sample of size n from a $N(\mu, \sigma^2)$ distribution where σ^2 is known and μ is unknown. An unbiased estimator of μ is $\overline{x} = \sum_{i=1}^{n} x_i/n$. We can use this information either

1. to test the hypothesis $H_0 : \mu = \mu_0$; or
2. to find a CI for μ and see if the value μ_0 is in it or not.

We will show that testing H_0 at the 5% significance level (that is, with $\alpha = .05$) against a 2-sided alternative is the same as finding out whether or not μ_0 lies in the 95% confidence interval.

1. For $H_1 : \mu \neq \mu_0$ we reject H_0 at the 5% significance level if

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > 1.96 \quad \text{or} \quad \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -1.96. \tag{6.6.1}$$

   That is, if

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} > 1.96.$$

   Or, using the "P-value," if $\overline{x} > \mu_0$ we calculate the probability of a value as extreme or more extreme than this, in either direction. That is, calculate

$$P = 2 \times P(\overline{X} > \overline{x}) = 2 \times P_n\left(Z > \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right).$$

   If $P < .05$ the result is significant at the 5% level. This will happen if $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < -1.96$, as in (6.5.7).

2. A symmetric 95% confidence interval for μ is $\overline{x} \pm 1.96\sigma/\sqrt{n}$ which arose from considering the inequality

$$-1.96 < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < 1.96$$

   which is the event complementary to that in (6.5.7).

   So, to reject H_0 at the 5% significance level is equivalent to saying that "the hypothesized value is not in the 95% CI." Likewise, to reject H_0 at the 1% significance level is equivalent to saying that "the hypothesized value is not in the 99% CI," which is equivalent to saying that "the P-value is less than 1%."

   If $1\% < P < 5\%$ the hypothesized value of μ will not be within the 95% CI but it will lie in the 99% CI.

   This approach is illustrated for the hypothesis-testing situation and the confidence interval approach below.

Computer Exercise 6.3. *Using the data in Computer Exercise 6.2, find a 99% CI for the true mean, μ.*

Solution of Computer Exercise 6.3.

```
#Calculate the upper and lower limits for the 99% confidence interval.
CI <- qnorm(mean=xbar,sd=5/sqrt(25),p=c(0.005,0.995) )

> CI
[1] 18 23
```

So that the 99% CI is $(18, 23)$.

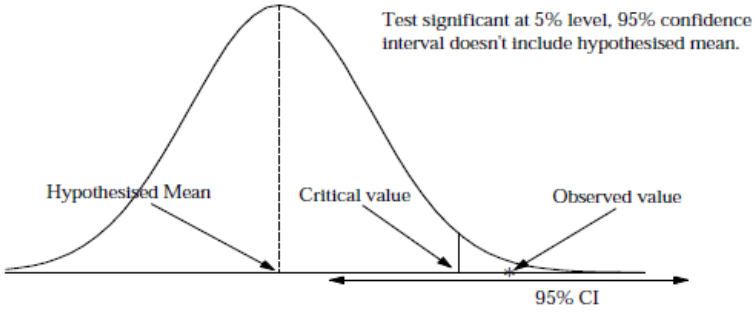


Figure 6.6.1: Relationship between Non-significant Hypothesis Test and Confidence Interval

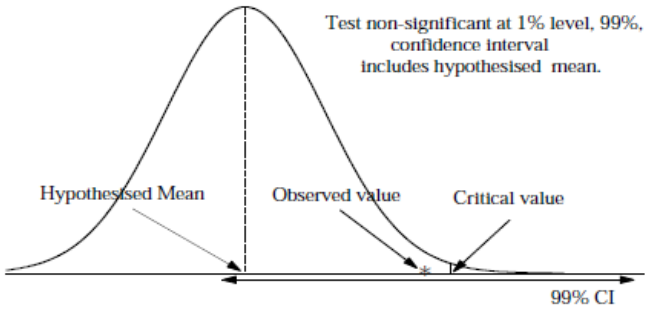


Figure 6.6.2: Relationship between Significant Hypothesis Test and Confidence Interval

6.7 The Neyman-Pearson lemma

Suppose we are testing a simple null hypothesis $H_0 : \theta = \theta'$ against a simple alternative $H_1 : \theta = \theta''$, where θ is the parameter of interest, and θ', θ'' are particular values of θ. Observed values of the i.i.d. random variables $X_1, X_2, \ldots, X_n$, each with p.d.f. $f_X(x|\theta)$, are available. We are going to reject H_0 if $(x_1, x_2, \ldots, x_n) \in \mathcal{C}$, where $\mathcal{C}$ is a region of the n-dimensional space

called the critical region. This specifies a test of hypothesis. We say that the critical region $\mathcal{C}$ has size α if the probability of a Type I error is α:

$$P[\ (X_1, X_2, \ldots, X_n) \in \mathcal{C} | H_0\ ] = \alpha.$$

We call $\mathcal{C}$ a best critical region of size α if it has size α, and

$$P[\ (X_1, X_2, \ldots, X_n) \in \mathcal{C} | H_1\ ] \ \geq \ P[\ (X_1, X_2, \ldots, X_n) \in \mathcal{A} | H_1\ ]$$

for every subset $\mathcal{A}$ of the sample space for which $P[\ (X_1, X_2, \ldots, X_n) \in \mathcal{C} | H_0\ ] = \alpha$. Thus, the power of the test associated with the best critical region $\mathcal{C}$ is at least as great as the power of the test associated with any other critical region $\mathcal{A}$ of size α. The Neyman-Pearson lemma provides us with a way of finding a best critical region.

Lemma 6.1 (The Neyman-Pearson lemma)**.** If $k > 0$ and $\mathcal{C}$ is a subset of the sample space such that

(a) $L(\theta')/L(\theta'') \leq k$ for all $(x_1, x_2, \ldots, x_n) \in \mathcal{C}$

(b) $L(\theta')/L(\theta'') \geq k$ for all $(x_1, x_2, \ldots, x_n) \in \mathcal{C}^*$,

(c) $\alpha = P[(X_1, X_2, \ldots, X_n) \in \mathcal{C} | H_0]$

where $\mathcal{C}^*$ is the complement of $\mathcal{C}$, then $\mathcal{C}$ is a best critical region of size α for testing the simple hypothesis $H_0 : \theta = \theta'$ against the alternative simple hypothesis $H_1 : \theta = \theta''$.

Proof. Suppose for simplicity that the random variables $X_1, X_2, \ldots, X_n$ are continuous. (If they were discrete, the proof would be the same, except that integrals would be replaced by sums.) For any region $\mathcal{R}$ of n-dimensional space, we will denote the probability that $\mathbf{X} \in \mathcal{R}$ by $\int\limits_{\mathcal{R}} L(\theta)$, where θ is the true value of the parameter. The full notation, omitted to save space, would be

$$P(\mathbf{X} \in \mathcal{R} | \theta) = \int\limits_{\mathcal{R}} \cdots \int L(\theta | x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n \ .$$

We need to prove that if $\mathcal{A}$ is another critical region of size α, then the power of the test associated with $\mathcal{C}$ is at least as great as the power of the test associated with $\mathcal{A}$, or in the present notation, that

$$\int\limits_{\mathcal{A}} L(\theta'') \leq \int\limits_{\mathcal{C}} L(\theta''). \tag{6.7.1}$$

Suppose $X \in \mathcal{A}^* \cap \mathcal{C}$. Then $X \in \mathcal{C}$, so by (a),

$$\int\limits_{\mathcal{A}^* \cap \mathcal{C}} L(\theta'') \geq \frac{1}{k} \int\limits_{\mathcal{A}^* \cap \mathcal{C}} L(\theta'). \tag{6.7.2}$$

Next, suppose $X \in \mathcal{A} \cap \mathcal{C}^*$. Then $X \in \mathcal{C}^*$, so by (b),

$$\int\limits_{\mathcal{A} \cap \mathcal{C}^*} L(\theta'') \leq \frac{1}{k} \int\limits_{\mathcal{A} \cap \mathcal{C}^*} L(\theta'). \tag{6.7.3}$$

We now establish (6.7.1), thereby completing the proof.

$$\begin{aligned}
\int_{\mathcal{A}} L(\theta'') &= \int_{\mathcal{A}\cap\mathcal{C}} L(\theta'') + \int_{\mathcal{A}\cap\mathcal{C}^*} L(\theta'') \\
&= \int_{\mathcal{C}} L(\theta'') - \int_{\mathcal{A}^*\cap\mathcal{C}} L(\theta'') + \int_{\mathcal{A}\cap\mathcal{C}^*} L(\theta'') \\
&\le \int_{\mathcal{C}} L(\theta'') - \frac{1}{k}\int_{\mathcal{A}^*\cap\mathcal{C}} L(\theta') + \frac{1}{k}\int_{\mathcal{A}\cap\mathcal{C}^*} L(\theta') \qquad (\textit{ see (6.7.2), (6.7.3) }) \\
&\qquad + \left[\frac{1}{k}\int_{\mathcal{A}\cap\mathcal{C}} L(\theta') - \frac{1}{k}\int_{\mathcal{A}\cap\mathcal{C}} L(\theta')\right] \qquad (\textit{add zero}) \\
&= \int_{\mathcal{C}} L(\theta'') - \frac{1}{k}\int_{\mathcal{C}} L(\theta') + \frac{1}{k}\int_{\mathcal{A}} L(\theta') \qquad (\textit{collect terms}) \\
&= \int_{\mathcal{C}} L(\theta'') - \frac{\alpha}{k} + \frac{\alpha}{k} \\
&= \int_{\mathcal{C}} L(\theta'')
\end{aligned}$$

since both $\mathcal{C}$ and $\mathcal{A}$ have size α. □

Example 6.12.
Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(0,1)$, and and we want to test $H_0 : \theta = \theta'$ versus $H_1 : \theta = \theta''$, where $\theta'' > \theta'$. According to the Z-test, we should reject H_0 if $Z = \sqrt{n}(\bar{X} - \theta')$ is large, or equivalently if $\bar{X}$ is large. We can now use the Neyman-Pearson lemma to show that the Z-test is "best". The likelihood function is

$$L(\theta) = (2\pi)^{-n/2} \exp\{-\sum_{i=1}^{n}(x_i - \theta)^2/2\}.$$

According to the Neyman-Pearson lemma, a best critical region is given by the set of $(x_1, \ldots, x_n)$ such that $L(\theta')/L(\theta'') \le k_1$, or equivalently, such that

$$\frac{1}{n}\ln[L(\theta'')/L(\theta')] \ge k_2.$$

But

$$\begin{aligned}
\frac{1}{n}\ln[L(\theta'')/L(\theta')] &= \frac{1}{n}\sum_{i=1}^{n}[(x_i - \theta')^2/2 - (x_1 - \theta'')^2/2] \\
&= \frac{1}{2n}\sum_{i=1}^{n}[(x_i^2 - 2\theta' x_i + \theta'^2) - (x_i^2 - 2\theta'' x_i + \theta''^2)] \\
&= \frac{1}{2n}\sum_{i=1}^{n}[2(\theta'' - \theta')x_i + \theta'^2 - \theta''^2] \\
&= (\theta'' - \theta')\bar{x} + \frac{1}{2}[\theta'^2 - \theta''^2].
\end{aligned}$$

So the best test rejects H_0 when $\bar{x} \geq k$, where k is a constant. But this is exactly the form of the rejection region for the Z-test. Therefore, the Z-test is "best".

6.8 Summary

We have only considered 4 hypothesis testing problems at this stage. Further problems will be dealt with in later chapters after more sampling distributions are introduced. The following might be helpful as a pattern to follow in doing examples in hypothesis testing.

1. State the hypothesis and the alternative. This must always be a statement about the unknown parameter in a distribution.

2. Select the appropriate *statistic* (function of the data). In the problems considered so far this is an unbiased estimate of the parameter or a function of it. State the distribution of the statistic and its particular form when H_0 is true.

Alternative Procedures

1. Find the critical region using the appropriate value of α (.05 or .01 usually).

2. Find the observed value of the statistic (using the data).

3. Draw conclusions. If the calculated value falls in the CR, this provides evidence against H_0. You could say that the result is significant at the 5% (or 1% or .1% level).

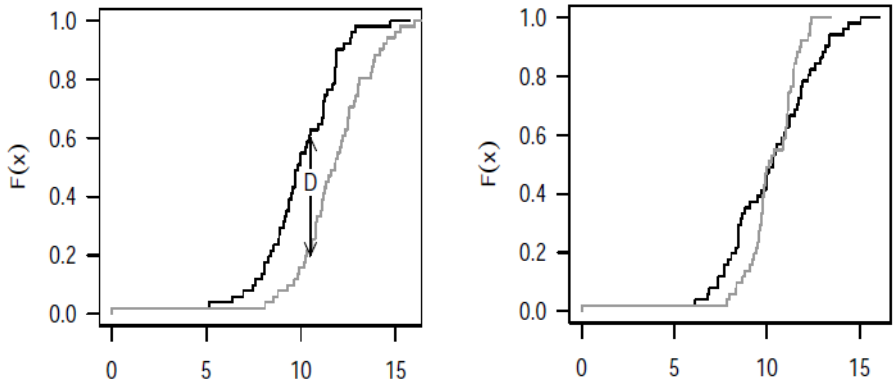
Other way:

2. Calculate P, the probability associated with values as extreme or more extreme than that observed. For a 2-sided H_1, you'll need to double a probability such as $P(X \geq k)$.

3. Draw conclusions. For example, if $P < .1\%$ we say that there is *very strong* evidence against H_0. If $.1\% < P < 1\%$ we say there is *strong* evidence. If $1\% < P < 5\%$ we say there is *some* evidence. For larger values of P we conclude that the event is not an unusual one **if H_0 is true**, and say that this set of data is consistent with H_0.

6.9 Non-parametric hypothesis testing

Figure 6.9.1 shows two ways in which distributions differ. The difference depicted in Figure 6.1(a) is a shift in location (mean) and in Figure 6.1(b) there is a shift in the scale (variance).



(a) Due to shifts in location. (b) Due to shifts in scale.

Figure 6.9.1: Distributions that differ due to shifts in (a) location and (b) scale.

6.9.1 Kolmogorov-Smirnov (KS)

The KS test is a test of whether 2 independent samples have been drawn from the same population or from populations with the same distribution. It is concerned with the agreement between 2 cumulative distribution functions. If the 2 samples have been drawn from the same population, then the cdf's can be expected to be close to each other and only differ by random deviations. If they are too far apart at any point, this suggests that the samples come from different populations.

The KS test statistics is

$$D = \max\left(\left|\widehat{F_1(x)} - \widehat{F_2(y)}\right|\right) \tag{6.9.1}$$

Exact sampling distribution

The exact sampling distribution of D under $H_0 : F_1 = F_2$ can be enumerated.

If H_0 is true, then $[(X_1, X_2, \ldots, X_m), (Y_1, Y_2, \ldots, Y_n)]$ can be regarded as a random sample from the same population with actual realised samples

$$[(x_1, x_2, \ldots, x_m), (y_1, y_2, \ldots, y_n)]$$

Thus (under H_0) an equally likely sample would be

$$[(y_1, x_2, \ldots, x_m), (x_1, y_2, \ldots, y_n)]$$

where x_1 and y_1 were swapped.

There are $\binom{m+n}{m}$ possible realisations of allocation the combined sample to 2 groups of sizes m and n and under H_0 the probability of each realisation is $\frac{1}{\binom{m+n}{m}}$. For each sample generated this way, a $D^\star$ is observed.

Now $F_1(x)$ is steps of $\frac{1}{m+1}$ and $F_2(y)$ is steps of $\frac{1}{n+1}$ so for given m and n, it would be possible to enumerate all $D^\star_{m,n}$ if H_0 is true. From this enumeration the upper $100\alpha\%$ point of $\{D^\star_{m,n}\}$, $\{D_{m,n}; \alpha\}$, gives the critical value for the α sized test. If the observed $\{D_{m,n}\}$ is greater than $\{D_{m,n}; \alpha\}$, reject H_0.

6.9.2 Asymptotic distribution

If m and n become even moderately large, the enumeration is huge. In that case, we can utilize the large sample approximation that

$$\chi^2 = \frac{4D^2(nm)}{n+m}$$

Example 6.13.
These data are the energies of sway signals from 2 groups of subjects, Normal group and Whiplash group. Whiplash injuries can lead to unsteadiness and the subject may not be able to maintain balance. each subject had their sway pattern measured by standing on a plate blindfolded. Does the distribution of energies differ between groups?

Normal	33 211 284 545 570 591 602 786 945 951
	1161 1420 1529 1642 1994 2329 2682 2766 3025 13537
Whipl	269 352 386 1048 1247 1276 1305 1538 2037 2241
	2462 2780 2890 4081 5358 6498 7542 13791 23862 34734

Table 6.1: Wavelet energies of the sway signals from normal subjects and subjects with whiplash injury.

Solution of Example 6.7. *The plots of the ecdf suggest a difference. We apply the Kolmogorov-Smirnov test to these data.*

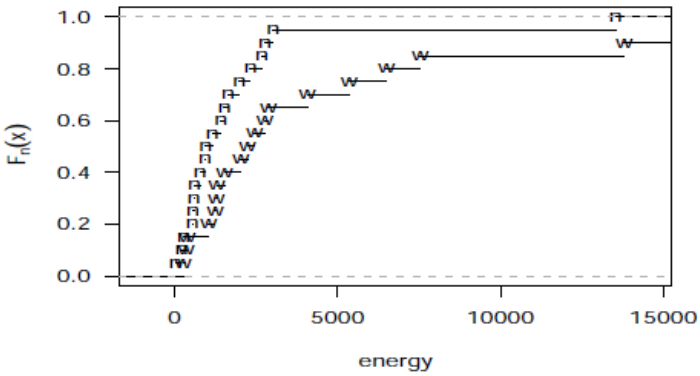


Figure 6.9.2: The ecdf's of sway signal energies for N & W groups

```
N.energy <- c(33,211,284,545,570,591,602,786,945,951,1161,1420,
  1529,1642,1994,2329,2682,2766,3025,13537)
W.energy <- c(269,352,386,1048,1247,1276,1305,1538,2037,2241,2462,2780,
  2890,4081,5358,6498,754,1379,23862,34734)
KS <- ks.test(N.energy,W.energy,alternative="greater")
> KS

Two-sample Kolmogorov-Smirnov test
```

```
data: N.energy and W.energy
D^+ = 0.35, p-value = 0.0863
alternative hypothesis: the CDF of x lies above

 that of y
# ______ the Asymptotic distribution __________
D <- KS$statistic
Chi <- 4*(KS$statistic^2)*m*n/(m+n)
P <- pchisq(q=Chi,df=2,lower.tail=F)

> cat("X2 = ",round(Chi,2),"P( > X2) = ",P,"\n")
X2 = 4.9 P( > X2) = 0.08629
```

1. *The Kolmogorov-Smirnov test of whether the null hypothesis can be rejected is a permutation test.*

2. *The equality $F_1 = F_2$ means that F_1 and F_2 assign equal probabilities to all sets; $P_{F_1}(A) = P_{F_2}(A)$ for and A subset of the common sample space of x and y. If H_0 is true, there is no difference between the randomness of x or y.*

3. *The null hypothesis is set up to be rejected. If however, the data are such that the null hypothesis cannot be decisively rejected, then the experiment has not demonstrated a difference.*

4. *A hypothesis test requires a statistic, $\widehat{\theta}$, for comparing the distributions. In the Kolmogorov-Smirnov test $\widehat{\theta} = D$.*

5. *Having observed $\widehat{\theta}$, the achieved significance level of the test is the probability of observing at least as large a value when H_0 is true, $P_{H_0}(\widehat{\theta^\star} \geq \widehat{\theta})$. The observed statistic, $\widehat{\theta}$ is fixed and the random variable $\widehat{\theta^\star}$ is distributed according to H_0.*

6. *The KS test enumerated all permutations of elements in the samples. This is also termed sampling without replacement. Not all permutations are necessary but an accurate test does require a large number of permutations.*

7. *The permutation test applies to any test statistic. For the example in Figure 6.1(b), we might use $\widehat{\theta} = \frac{\widehat{\sigma_x^2}}{\widehat{\sigma_y^2}}$.*

6.9.3 Bootstrap Hypothesis tests

The link between confidence intervals and hypothesis tests also holds in a bootstrap setting. The bootstrap is an approximation to a permutation test and a strategic difference is that bootstrap uses *sampling with replacement*.

A permutation test of whether $H_0 : F_1(x) = F_2(y)$ is true relies upon the ranking of the combined data set $(\mathbf{x}, \mathbf{y})$. The data were ordered smallest to largest and each permutation was an allocation of the group labels to each ordered datum. In 1 permutation, the label x was ascribed to the first number and in another, the label y is given to that number and so on.

The test statistic can be a function of the data (it need not be an estimate of a parameter) and so denote this a $t(\mathbf{z})$.

The principle of bootstrap hypothesis testing is that if H_0 is true, a probability atom of $\frac{1}{m+n}$ can be attributed to each member of the combined data $\mathbf{z} = (\mathbf{x}, \mathbf{y})$.

The empirical distribution function of $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, call it $\widehat{F_0}(z)$, is a non-parametric estimate of the common population that gave rise to $\mathbf{x}$ and $\mathbf{y}$, assuming that H_0 is true. Bootstrap hypothesis testing of H_0 takes these steps,

1. Get the observed value of t, e.g. $t_{\text{obs}} = \overline{x} - \overline{y}$.
2. Nominate how many bootstrap samples (replications) will be done, e.g. $B = 499$.
3. For b in $1 : B$, draw samples of size $m + n$ *with replacement* from $\mathbf{z}$. Label the first m of these $x_b^\star$ and the remaining n be labelled $y_b^\star$.
4. Calculate $t(z_b^\star)$ for each sample. For example, $t(z_b^\star) = x_b^\star - y_b^\star$.
5. Approximate the probability of t_{obs} or greater by $\frac{\text{number of } t(\mathbf{z}_b^\star) \geq t_{\text{obs}}}{B}$.

Example 6.14.
The data in Table 6.1 are used to demonstrate bootstrap hypothesis testing with the test statistic,

$$t(\mathbf{z}) = \frac{\overline{y} - \overline{x}}{\widehat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

The R code is written to show the required calculations more explicitly but a good program minimises the variables which are saved in the iterations loop.

Solution of Example 6.8.

```
#_____________ Bootstrap Hypothesis Test ____________________
N.energy <- c(33,211,284,545,570,591,602,786,945,951,1161,1420,
              1529,1642,1994,2329,2682,2766,3025,13537)
W.energy <- c(269,352,386,1048,1247,1276,1305,1538,2037,2241,2462,2780,
              2890,4081,5358,6498,754,1379,23862,34734)
Z <- c(N.energy,W.energy)
m <- length(N.energy)
n <- length(W.energy)
T.obs <- (mean(W.energy) - mean(N.energy))/(sd(Z)*sqrt(1/m + 1/n))

nBS <- 999

T.star <- numeric(nBS)
for (j in 1:nBS){
z.star <- sample(Z,size=(m+n))
w.star <- z.star[(m+1):(m+n)]
n.star <- z.star[1:m]
T.star[j] <- ( mean(w.star) - mean(n.star) )/( sd(z.star) * sqrt(1/m + 1/n) )
}
```

```
p1 <- sum(T.star >= T.obs)/nBS

cat( "P(T > ",round(T.obs,1),"|H0) = ",round(p1,2),"\n")
```

The results are:

$$T = 1.4$$
$$P(t > 1.4|H_0) = 0.09$$

Thus this statistic does not provide evidence that the 2 distributions are different.

6.10 Likelihood Ratio and Score Tests

Suppose that we observe the value of a random vector $\mathbf{X}$ whose probability density function is $g(\mathbf{X}|\boldsymbol{\theta})$ for $\mathbf{x} \in \mathcal{X}$ where the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ is some unknown element of the set $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$. Let $\boldsymbol{\Theta}_0$ be a specified subset of $\boldsymbol{\Theta}$. Consider the hypothesis $H_0 : \theta \in \Theta_0$. In this section we consider three ways in which good test statistics may be found for this general problem.

The Likelihood Ratio Test: This test statistic is based on the idea that the maximum of the log likelihood over the subset $\boldsymbol{\Theta}_0$ should not be too much less than the maximum over the whole set $\boldsymbol{\Theta}$ if, in fact, the parameter $\boldsymbol{\theta}$ actually does lie in the subset $\boldsymbol{\Theta}_0$. Let $\ell(\boldsymbol{\theta})$ denote the log likelihood function. The test statistic is

$$T_1(\mathbf{x}) = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)]$$

where $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ in the set $\boldsymbol{\Theta}$ for which $\ell(\boldsymbol{\theta})$ is a maximum and $\hat{\boldsymbol{\theta}}_0$ is the value of $\boldsymbol{\theta}$ in the set $\boldsymbol{\Theta}_0$ for which $\ell(\boldsymbol{\theta})$ is a maximum.

The Maximum Likelihood Test Statistic: This test statistic is based on the idea that $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_0$ should be close to one another. Let $\mathbf{I}(\boldsymbol{\theta})$ be the $p \times p$ information matrix. Let $\mathbf{B} = \mathbf{I}(\hat{\boldsymbol{\theta}})$. The test statistic is

$$T_2(\mathbf{x}) = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_0)^T \mathbf{B} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_0)$$

Other forms of this test statistic follow by choosing $\mathbf{B}$ to be $\mathbf{I}(\hat{\boldsymbol{\theta}}_0)$ or $\mathrm{E}\mathbf{I}(\hat{\boldsymbol{\theta}})$ or $\mathrm{E}\mathbf{I}(\hat{\boldsymbol{\theta}}_0)$.

The Score Test Statistic: This test statistic is based on the idea that $\hat{\boldsymbol{\theta}_0}$ should almost solve the likelihood equations. Let $\mathbf{S}(\theta)$ be the $p \times 1$ vector whose rth element is given by $\partial\ell/\partial\theta_r$. Let $\mathbf{C}$ be the inverse of $\mathbf{I}(\hat{\boldsymbol{\theta}}_0)$ i.e. $\mathbf{C} = \mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1}$. The test statistic is

$$T_3(\mathbf{x}) = \mathbf{S}(\hat{\boldsymbol{\theta}}_0)^T \mathbf{C} \mathbf{S}(\hat{\boldsymbol{\theta}}_0)$$

In order to calculate p-values we need to know the probability distribution of the test statistic under the null hypothesis. Deriving the exact probability distribution may be difficult but approximations suitable for situations in which the sample size is large are available in the special case where $\boldsymbol{\Theta}$ is a p dimensional set and $\boldsymbol{\Theta}_0$ is a q dimensional subset of Θ for $q < p$, whence it can be shown that, when H_0 is true, the probability distributions of $T_1(\mathbf{x}), T_2(\mathbf{x})$ and $T_3(\mathbf{x})$ are all approximated by a χ^2_{p-q} density.

Example 6.15.
Let $X_1, X_2, \ldots, X_n$ be iid each having a Poisson distribution with parameter θ. Consider testing $H_0 : \theta = \theta_0$ where θ_0 is some specified constant. Recall that

$$\ell(\theta) = \left[\sum_{i=1}^{n} x_i\right] \log[\theta] - n\theta - \log\left[\prod_{i=1}^{n} x_i!\right].$$

Here $\Theta = [0, \infty)$ and the value of $\theta \in \Theta$ for which $\ell(\theta)$ is a maximum is $\hat{\theta} = \bar{x}$. Also $\Theta_0 = \{\theta_0\}$ and so trivially $\hat{\theta}_0 = \theta_0$. We saw also that

$$S(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta} - n$$

and that

$$I(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta^2}.$$

Suppose that $\theta_0 = 2, n = 40$ and that when we observe the data we get $\bar{x} = 2.50$. Hence $\sum_{i=1}^{n} x_i = 100$. Then

$$\begin{aligned} T_1 &= 2[\ell(2.5) - \ell(2.0)] \\ &= 200 \log(2.5) - 200 - 200 \log(2.0) + 160 = 4.62. \end{aligned}$$

The information is $B = I(\hat{\theta}) = 100/2.5^2 = 16$. Hence

$$T_2 = (\hat{\theta} - \hat{\theta}_0)^2 B = 0.25 \times 16 = 4.$$

We have $S(\theta_0) = S(2.0) = 10$ and $I(\theta_0) = 25$ and so

$$T_3 = 10^2/25 = 4.$$

Here $p = 1,\ q = 0$ implying $p - q = 1$. Since $P[\chi_1^2 \geq 3.84] = 0.05$ all three test statistics produce a p-value less than 0.05 and lead to the rejection of $H_0 : \theta = 2$. □

Example 6.16.
Let $X_1, X_2, \ldots, X_n$ be iid with density $f(x|\alpha,\beta) = \alpha\beta x^{\beta-1}exp(-\alpha x^\beta)$ for $x \geq 0$. Consider testing $H_0 : \beta = 1$. Here $\Theta = \{(\alpha,\beta) : 0 < \alpha < \infty, 0 < \beta < \infty\}$ and $\Theta_0 = \{(\alpha, 1) : 0 < \alpha < \infty\}$ is a one-dimensional subset of the two-dimensional set Θ. Recall that $\ell(\alpha,\beta) = n\log[\alpha] + n\log[\beta] + (\beta-1)\sum_{i=1}^n \log[x_i] - \alpha\sum_{i=1}^n x_i^\beta$. Hence the vector $\mathbf{S}(\alpha,\beta)$ is given by

$$\begin{pmatrix} n/\alpha - \sum_{i=1}^n x_i^\beta \\ n/\beta + \sum_{i=1}^n \log[x_i] - \alpha\sum_{i=1}^n x_i^\beta \log[x_i] \end{pmatrix}$$

and the matrix $\mathbf{I}(\alpha,\beta)$ is given by

$$\begin{pmatrix} n/\alpha^2 & \sum_{i=1}^n x_i^\beta \log[x_i] \\ \sum_{i=1}^n x_i^\beta \log[x_i] & n/\beta^2 + \alpha\sum_{i=1}^n x_i^\beta \log[x_i]^2 \end{pmatrix}$$

We have that $\hat\theta = (\hat\alpha, \hat\beta)$ which require Newton's method for their calculation. Also $\hat\theta_0 = (\hat\alpha_0, 1)$ where $\hat\alpha_0 = 1/\bar x$. Suppose that the observed value of $T_1(\mathbf{x})$ is 3.20. Then the p-value is $P[T_1(\mathbf{x}) \geq 3.20] \approx P[\chi_1^2 \geq 3.20] = 0.0736$. In order to get the maximum likelihood test statistic plug in the values $\hat\alpha, \hat\beta$ for α, β in the formula for $\mathbf{I}(\alpha,\beta)$ to get the matrix B. Then calculate $T_1(\mathbf{X}) = (\hat\theta - \hat\theta_0)^T B(\hat\theta - \hat\theta_0)$ and use the χ_1^2 tables to calculate the p-value. Finally, to calculate the score test statistic note that the vector $\mathbf{S}(\hat\theta_0)$ is given by

$$\begin{pmatrix} 0 \\ n + \sum_{i=1}^n \log[x_i] - \sum_{i=1}^n x_i \log[x_i]/\bar x \end{pmatrix}$$

and the matrix $\mathbf{I}(\hat\theta_0)$ is given by

$$\begin{pmatrix} n\bar x^2 & \sum_{i=1}^n x_i \log[x_i] \\ \sum_{i=1}^n x_i \log[x_i] & n + \sum_{i=1}^n x_i \log[x_i]^2/\bar x \end{pmatrix}$$

Since $T_2(\mathbf{x}) = \mathbf{S}(\hat\theta_0)^T C \mathbf{S}(\hat\theta_0)$ where $\mathbf{C} = \mathbf{I}(\hat\theta_0)^{-1}$ we have that $T_2(\mathbf{x})$ is

$$[n + \sum_{i=1}^n \log[x_i] - \sum_{i=1}^n x_i \log[x_i]/\bar x]^2$$

multiplied by the lower diagonal element of $\mathbf{C}$ which is given by

$$\frac{n\bar x^2}{[n\bar x^2][n + \sum_{i=1}^n x_i \log[x_i]^2/\bar x] - [\sum_{i=1}^n x_i \log[x_i]]^2}$$

Hence we get that

$$T_2(\mathbf{x}) = \frac{[n + \sum_{i=1}^n \log[x_i] - \sum_{i=1}^n x_i \log[x_i]/\bar x]^2 n\bar x^2}{[n\bar x^2][n + \sum_{i=1}^n x_i \log[x_i]^2/\bar x] - [\sum_{i=1}^n x_i \log[x_i]]^2}$$

No iterative techniques are need to calculate the value of $T_2(\mathbf{X})$ and for this reason the score test is often preferred to the other two. However there is some evidence that the likelihood ratio test is more powerful in the sense that it has a better chance of detecting departures from the null hypothesis.

6.11 Goodness of fit tests

Suppose that we have a random experiment with a random variable Y of interest. Assume additionally that Y is discrete with density function f on a finite set $\mathcal{S}$. We repeat the experiment n times to generate a random sample $Y_1, Y_2, \ldots, Y_n$ from the distribution of Y. These are independent variables, each with the distribution of Y.

In this section, we assume that the distribution of Y is unknown. For a given density function f_0, we will test the hypotheses $H_0 : f = f_0$ versus $H_1 : f \neq f_0$. The test that we will construct is known as the goodness of fit test for the conjectured density f_0. As usual, our challenge

in developing the test is to find an appropriate test statistic – one that gives us information about the hypotheses and whose distribution, under the null hypothesis, is known, at least approximately.

Suppose that $\mathcal{S} = y_1, y_2, \ldots, y_k$. To simplify the notation, let $p_j = f_0(y_j)$ for $j = 1, 2, \ldots, k$. Now let $N_j = \#\{i \in 1, 2, ..., n : y_i = y_j\}$ for $j = 1, 2, \ldots, k$. Under the null hypothesis, $(N_1, N_2, \ldots, N_k)$ has the multinomial distribution with parameters n and $p_1, p_2, \ldots, p_k$ with $\mathrm{E}(N_j) = np_j$ and $\mathrm{Var}(N_j) = np_j(1 - p_j)$. This results indicates how we might begin to construct our test: for each j we can compare the observed frequency of y_j (namely N_j) with the expected frequency of value y_j (namely np_j), under the null hypothesis. Specifically, our test statistic will be

$$X^2 = \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} + \cdots + \frac{(N_k - np_k)^2}{np_k}.$$

Note that the test statistic is based on the squared errors (the differences between the expected frequencies and the observed frequencies). The reason that the squared errors are scaled as they are is the following crucial fact, which we will accept without proof: under the null hypothesis, as n increases to infinity, the distribution of X^2 converges to the chi-square distribution with $k - 1$ degrees of freedom.

For $m > 0$ and r in $(0, 1)$, we will let $\chi^2_{m,r}$ denote the quantile of order r for the chi-square distribution with m degrees of freedom. Then, the following test has approximate significance level α: reject $H_0 : f = f_0$ versus $H_1 : f \neq f_0$, if and only if $X^2 > \chi^2_{k-1,1-\alpha}$. The test is an approximate one and works best when n is large. Just how large n needs to be depends on the p_j. One popular rule of thumb proposes that the test will work well if all the expected frequencies satisfy $np_j \geq 1$ and at least 80% of the expected frequencies satisfy $np_j \geq 5$.

Example 6.17 (Genetical inheritance).
In crosses between two types of maize four distinct types of plants were found in the second generation. In a sample of 1301 plants there were 773 green, 231 golden, 238 green-striped, 59 golden-green-striped. According to a simple theory of genetical inheritance the probabilities of obtaining these four plants are $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}$ and $\frac{1}{16}$ respectively. Is the theory acceptable as a model for this experiment?

Formally we will consider the hypotheses:

$H_0 :$ *$p_1 = \frac{9}{16}$, and $p_2 = \frac{3}{16}$, and $p_3 = \frac{3}{16}$ and $p_4 = \frac{1}{16}$;*

H_1 : *not all the above probabilities are correct.*

The expected frequencies for any plant under H_0 is $np_i = 1301p_i$. We therefore calculate the following table:

Observed Counts O_i	*Expected Counts* E_i	*Contributions to* X^2 $(O_i - E_i)^2/E_i$
773	*731.8125*	*2.318*
231	*243.9375*	*0.686*
238	*243.9375*	*0.145*
59	*81.3125*	*6.123*
		$X^2 = 9.272$

Since X^2 embodies the differences between the observed and expected values we can say that if X^2 is large that there is a big difference between what we observe and what we expect so the theory does not seem to be supported by the observations. If X^2 is small the observations apparently conform to the theory and act as support for the theory. The test statistic X^2 is distributed $X^2 \sim \chi^2_{3\mathrm{df}}$. In order to define what we would consider to be an unusually large value of X^2 we will choose a significance level of $\alpha = 0.05$. The `R` *command* `qchisq(p=0.05,df=3,lower.tail=FALSE)` *calculates the 5% critical value for the test as 7.815. Since our value of X^2 is greater than the critical value 7.815 we reject H_0 and conclude that the theory is not a good model*

for these data. The `R` *command* `pchisq(q=9.272,df=3,lower.tail=FALSE)` *calculates the p-value for the test equal to 0.026. (These data are examined further in chapter 9 of Snedecor and Cochoran.)*

Very often we do not have a list of probabilities to specify our hypothesis as we had in the above example. Rather our hypothesis relates to the probability distribution of the counts without necessarily specifying the parameters of the distribution. For instance, we might want to test that the number of male babies born on successive days in a maternity hospital followed a binomial distribution, without specifying the probability that any given baby will be male. Or, we might want to test that the number of defective items in large consignments of spare parts for cars, follows a Poisson distribution, again without specifying the parameter of the distribution.

The X^2 test is applicable when all the probabilities depend on unknown parameters, provided that the unknown parameters are replaced by their maximum likelihood estimates and provided that one degree of freedom is deducted for each parameter estimated.

Example 6.18.
Feller reports an analysis of flying-bomb hits in the south of London during World War II. Investigators partitioned the area into 576 sectors each beng $\frac{1}{4}km^2$. The following table gives the resulting data:

No. of hits (x)	*0*	*1*	*2*	*3*	*4*	5
No. of sectors with x hits	*229*	*221*	*93*	*35*	*7*	*1*

If the hit pattern is random in the sense that the probability that a bomb will land in any particular sector in constant, irrespective of the landing place of previous bombs, a Poisson distribution might be expected to model the data.

x	$P(x) = \frac{\hat{\theta}^x e^{-\hat{\theta}}}{x!}$	*Expected* $576 \times P(X)$	*Observed*	*Contributions to* X^2 $(O_i - E_i)^2/E_i$
0	*0.395*	*227.53*	*229*	*0.0095*
1	*0.367*	*211.34*	*211*	*0.0005*
2	*0.170*	*98.15*	*93*	*0.2702*
3	*0.053*	*30.39*	*35*	*0.6993*
4	*0.012*	*7.06*	*7*	*0.0005*
5	*0.002*	*1.31*	*1*	*0.0734*
				$X^2 = 1.0534$

The MLE of θ was calculated as $\hat{\theta} = 535/576 = 0.9288$, that is, the total number of observed hits divided by the number of sectors. We carry out the chi-squared test as before except that we now subtract one additional degree of freedom because we had to estimate θ. The test statistic X^2 is distributed $X^2 \sim \chi^2_{4\text{df}}$. The `R` *command* `qchisq(p=0.05,df=4,lower.tail=FALSE)` *calculates the 5% critical value for the test as 9.488. Alternatively, the* `R` *command* `pchisq(q=1.0534,df=4,lower.tail=FALSE)` *calculates the p-value for the test equal to 0.90. The result of the chi-squared test is not statistically significant indicating that the divergence between the observed and expected counts can be regarded as random fluctuations about the expected values. Feller comments, "It is interesting to note that most people believed in a tendency of the points of impact to cluster. It this were true, there would be a higher frequency of sectors with either many hits or no hits and a deficiency in the intermediate classes. the above table indicates perfect randomness and homogeneity of the area; we have here an instructive illustration of the established fact that to the untrained eye randomness appears a regularity or tendency to cluster."* □

6.12 The χ^2 test for contingency tables

Let X and Y be a pair of categorical variables and suppose there are r possible values for X and c possible values for Y. Examples of categorical variables are Religion, Race, Social Class,

Blood Group, Wind Direction, Fertiliser Type etc. The random variables X and Y are said to be independent if $P[X = a, Y = b] = P[X = a]P[Y = b]$ for all possible values a of X and b of Y. In this section we consider how to test the null hypothesis of independence using data consisting of a random sample of N observations from the joint distribution of X and Y.

Example 6.19.
A study was carried out to investigate whether hair colour (columns) and eye colour (rows) were genetically linked. A genetic link would be supported if the proportions of people having various eye colourings varied from one hair colour grouping to another. 955 people were chosen at random and their hair colour and eye colour recorded. The data are summarised in the following table :

O_{ij}	*Black*	*Brown*	*Fair*	*Red*	*Total*
Brown	*60*	*110*	*42*	*30*	*242*
Green	*67*	*142*	*28*	*35*	*272*
Blue	*123*	*248*	*90*	*25*	*486*
Total	*250*	*500*	*160*	*90*	*1000*

The proportion of people with red hair is $90/1000 = 0.09$ *and the proportion having blue eyes is* $486/1000 = 0.486$. *So if eye colour and hair colour were truly independent we would expect the proportion of people having both black hair and brown eyes to be approximately equal to* $(0.090)(0.486) = 0.04374$ *or equivalently we would expect the number of people having both black hair and brown eyes to be close to* $(1000)(0.04374) = 43.74$. *The observed number of people having both black hair and brown eyes is* 60.5. *We can do similar calculations for all other combinations of hair colour and eye colour to derive the following table of expected counts :*

E_{ij}	*Black*	*Brown*	*Fair*	*Red*	*Total*
Brown	*60.5*	*121*	*38.72*	*21.78*	*242*
Green	*68.0*	*136*	*43.52*	*24.48*	*272*
Blue	*121.5*	*243*	*77.76*	*43.74*	*486*
Total	*250.0*	*500*	*160.00*	*90.00*	*1000*

In order to test the null hypothesis of independence we need a test statistic which measures the magnitude of the discrepancy between the observed table and the table that would be expected if independence were in fact true. In the early part of this century, long before the invention of maximum likelihood or the formal theory of hypothesis testing, Karl Pearson (one of the founding fathers of Statistics) proposed the following method of constructing such a measure of discrepancy:

$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	*Black*	*Brown*	*Fair*	*Red*
Brown	*0.004*	*1.000*	*0.278*	*3.102*
Green	*0.015*	*0.265*	*5.535*	*4.521*
Blue	*0.019*	*0.103*	*1.927*	*8.029*

For each cell in the table calculate $(O_{ij} - E_{ij})^2/E_{ij}$ *where* O_{ij} *is the observed count and* E_{ij} *is the expected count and add the resulting values across all cells of the table. The resulting total is called the* χ^2 *test statistic which we will denote by* W. *The null hypothesis of independence is rejected if the observed value of* W *is surprisingly large. In the hair and eye colour example the discrepancies are as follows :*

$$W = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 24.796$$

What we would now like to calculate is the p-value which is the probability of getting a value for W *as large as* 24.796 *if the hypothesis of independence were in fact true. Fisher showed that, when the hypothesis of independence is true,* W *behaves somewhat like a* χ^2 *random variable with degrees of freedom given by* $(r-1)(c-1)$ *where* r *is the number of rows in the table and* c *is the number of columns. In our example* $r = 3, c = 4$ *and so* $(r-1)(c-1) = 6$ *and so the p-value is* $P[W \geq 24.796] \approx P[\chi^2_6 \geq 24.796] = 0.0004$. *Hence we reject the independence hypothesis.*

CHAPTER

7

CHI-SQUARE DISTRIBUTION

7.1 Distribution of S^2

Recall that if $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution then

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$$

is an unbiased estimator of σ^2. We will find the probability distribution of this random variable. Firstly note that the numerator of S^2 is a sum of n squares but they are not independent as each involves $\overline{X}$. This sum of squares can be rewritten as the sum of squares of $n-1$ *independent* variables by the method which is illustrated below for the cases $n = 2, 3, 4$.

For $n = 2$,

$$\sum_{i=1}^{2} (X_i - \widehat{X})^2 = Y_1^2 \quad \text{where} \quad Y_1 = (X_1 - X_2)/\sqrt{2};$$

for $n = 3$,

$$\sum_{i=1}^{3} (X_i - \overline{X})^2 = \sum_{j=1}^{2} Y_j^2 \quad \text{where} \quad Y_1 = (X_1 - X_2)/\sqrt{2}, Y_2 = (X_1 + X_2 - 2X_3)/\sqrt{6};$$

for $n = 4$,

$$\sum_{i=1}^{4} (X_i - \widehat{X})^2 = \sum_{j=1}^{3} Y_j^2 \quad \text{where } Y_1, Y_2 \text{ are as defined above and}$$

$$Y_3 = (X_1 + X_2 + X_3 - 3X_4)/\sqrt{12}.$$

Note that Y_1, Y_2, Y_3 are linear functions of X_1, X_2, X_3, X_4 which are mutually orthogonal with the sum of the squares of their coefficients equal to 1.

Consider now the properties of the X_i and the Y_j as random variables. Since Y_1, Y_2, Y_3 are mutually orthogonal linear functions of X_1, X_2, X_3, X_4 they are uncorrelated, and since they are normally distributed (being sums of normal random variables), they are independent. Also,

$$E(Y_1) = 0 = E(Y_2) = E(Y_3)$$

and,

$$Var(Y_1) = \frac{1}{2}(Var(X_1) + Var(X_2)) = \sigma^2$$

$$Var(Y_2) = \frac{1}{6}Var(X_1) + \frac{1}{6}Var(X_2) + \frac{4}{6}Var(X_3) = \sigma^2.$$

Similarly, $Var(Y_3) = \sigma^2$.

In general the sum of n squares involving the X's can be expressed as the sum of $n-1$ squares involving the Y's. Thus $\sum_{i=1}^{n}(X_i - \overline{X})^2$ can be expressed as

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{j=1}^{n-1} Y_j^2 = \sum_{j=1}^{\nu} Y_j^2$$

where $\nu = n - 1$ is called the number of **degrees of freedom** and

$$Y_j = \frac{X_1 + X_2 + \cdots + X_j - jX_{j+1}}{\sqrt{j(j+1)}}, \quad j = 1, 2, \cdots, n-1.$$

The random variables $Y_1, Y_2, \ldots, Y_\nu$ each have mean zero and variance σ^2. So each $Y_j \sim N(0, \sigma^2)$ and the Y_j's are independent.

Now write $S^2 = \frac{\sum_{j=1}^{\nu} Y_j^2}{\nu}$ and recall that

1. If $X \sim N(\mu, \sigma^2)$ then $\frac{(X-\mu)^2}{2\sigma^2} \sim \Gamma\left(\frac{1}{2}\right)$, [Statistics 260, (8.16)]
2. If $X_1, X_2, \ldots, X_\nu$ are independent $N(\mu, \sigma^2)$ variates, then $\frac{\sum_{j=1}^{\nu}(X_j-\mu)^2}{2\sigma^2}$ is distributed as $\Gamma\left(\frac{\nu}{2}\right)$ [Statistics 260, section **7.4**].

Applying this to the Y_j where $\mu = 0$, $\frac{Y_j^2}{2\sigma^2} \sim \Gamma\left(\frac{1}{2}\right)$ and

$$V = \frac{1}{2}\sum_{j=1}^{\nu}\frac{Y_j^2}{\sigma^2} \quad \text{is distributed as } \Gamma\left(\frac{\nu}{2}\right). \tag{7.1.1}$$

Thus the pdf of V is given by

$$f(v) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} v^{\left(\frac{\nu}{2}-1\right)} e^{-v}, \quad v \in (0, \infty)$$

with V and S^2 being related by

$$S^2 = \frac{\sum_{j=1}^{\nu} Y_j^2}{\nu} = \frac{2\sigma^2 V}{\nu}$$

or

$$V = \frac{\nu}{2\sigma^2} S^2 \tag{7.1.2}$$

Now V is a strictly monotone function of S^2 so, by the change-of-variable technique, the pdf of S^2 is

$$\begin{aligned} g(s^2) &= f(v) \left| dv/ds^2 \right| \\ &= \frac{e^{-\nu s^2/2\sigma^2}}{\Gamma(\nu/2)} \left(\frac{\nu s^2}{2\sigma^2} \right)^{(\nu/2)-1} \cdot \left(\frac{\nu}{2\sigma^2} \right), s^2 \in (0, \infty) \end{aligned}$$

which is

$$\frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \times (s^2)^{\left(\frac{\nu}{2}-1\right)} \left(\frac{\nu}{2\sigma^2} \right)^{\nu/2} \exp\left\{ -\frac{\nu}{2\sigma^2} s^2 \right\} \tag{7.1.3}$$

This is the pdf of S^2 derived from a $N(\mu, \sigma^2)$ distribution.

7.2 Chi-Square Distribution

Define the random variable W as

$$W = \nu S^2/\sigma^2 = 2V,$$

where V is defined in (7.1.2). Note that W is a "sum of squares" divided by σ^2, and can be thought of as a standardized sum of squares. Then the p.d.f. of W is

$$h(w) = g(s^2) \left| \frac{ds^2}{dw} \right|, \text{where } \frac{ds^2}{dw} = \frac{\sigma^2}{\nu}$$

which can be written as

$$\frac{e^{-w/2} w^{(\nu/2)-1}}{2^{\nu/2}\Gamma(\nu/2)}, w \in [0, \infty]. \tag{7.2.1}$$

A random variable W with this pdf is said to have a **chi-square distribution on ν degrees of freedom** (or with parameter ν) and we write $W \sim \chi^2_\nu$.

Notes:

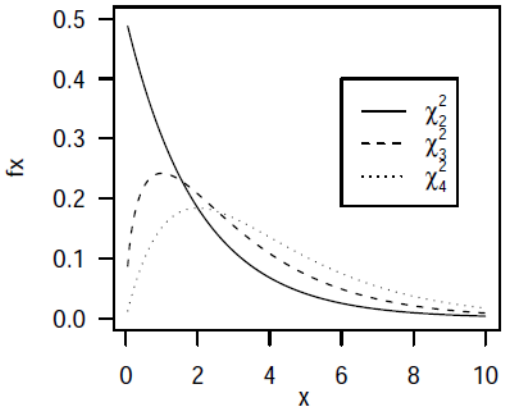
1. $W/2 \sim \gamma(\nu/2)$.
2. This distribution can be thought of as a special case of the generalized gamma distribution.
3. When $\nu = 2$, (7.2.1) becomes $h(w) = \frac{1}{2} e^{-w/2}$, $w \in [0, \infty]$, which is the exponential distribution.

Computer Exercise 7.1. *Graph the chi-square distributions with* $2, 3$ *and* 4 *degrees of freedom for* $x = 0.05, 0.1, 0.15, \ldots, 10$*, using one set of axes.*

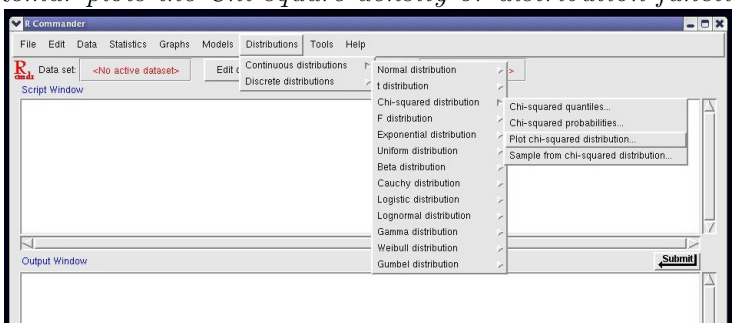
Solution of Computer Exercise 7.1.

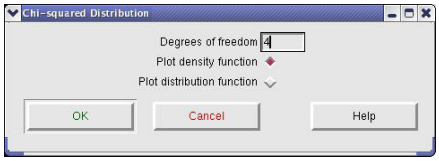
```
x <- seq(from=0.05,to=10,by=0.05)

for (d in 2:4){
 fx <- dchisq(x,df=d)
 if(d==2) plot(fx ~ x,type='l',
            ylim=c(0,0.5),las=1)
 else lines(x,fx,lty=(d-1))
} # end of d loop
legend(6,0.4,
 expression(chi[2]^2,chi[3]^2,chi[4]^2),
             lty=1:3)
```



Rcmdr plots the Chi-square density or distribution function readily,





Cumulative Distribution Function

If $W \sim \chi^2_\nu$, percentiles (i.e. 100P) of the chi-square distribution are determined by the inverse of the function

$$\frac{P}{100} = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^{w_{1-.01P}} w^{\frac{1}{2}\nu-1} e^{-w/2} dw = P(W \leq w_{1-.01P}).$$

Fig 7.2.1 depicts the tail areas corresponding to P (lower tail) and $1-P$ (upper tail) for the density function and superimposed is the distribution function. The scales for the Y-axes of the density function (left side) and the distribution function (right side) are different.

The R function for calculating tail area probabilities for given quantiles is

```
pchisq(q= , df = ,lower.tail= T (or F) )
```

and for calculating quantiles corresponding to a probability, `qchisq(p = , df = )`

These functions are included in the Rcmdr menus. The following example requires us to find a probability.

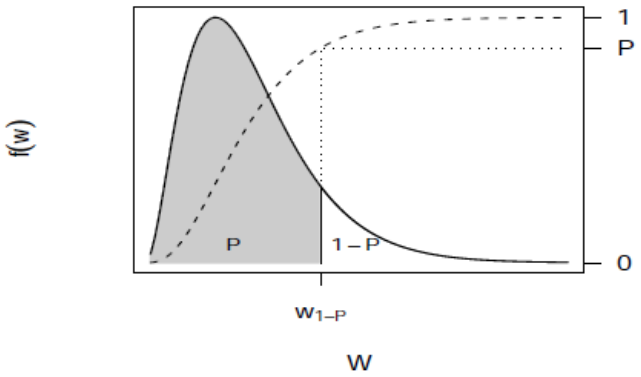


Figure 7.2.1: Area corresponding to the 100P percentile of the χ^2 random variable w.

Example 7.1.
A random sample of size 6 is drawn from a $N(\mu, 12)$ distribution. Find $P(2.76 < S^2 < 22.2)$.

Solution of Example 7.1. *We wish to express this as a probability statement about the random variable W. That is,*

$$\begin{aligned} P(2.76 < S^2 < 22.2) &= P\left(\frac{5}{12} \times 2.76 < \frac{\nu S^2}{\sigma^2} < \frac{5}{12} \times 22.2\right) \\ &= P(1.15 < W < 9.25) \quad \textit{where } W \sim \chi^2_5 \\ &= P(W < 9.25) - P(W < 1.15) \end{aligned}$$

```
#___ Pint.R _______
Q <- c(2.76,22.2)*5/12
Pint <- diff( pchisq(q=Q,df=5))
cat("P(2.76 < S2 < 22.2) = ",Pint,"\n")

> source("Pint.R")
P(2.76 < S2 < 22.2) = 0.85
```

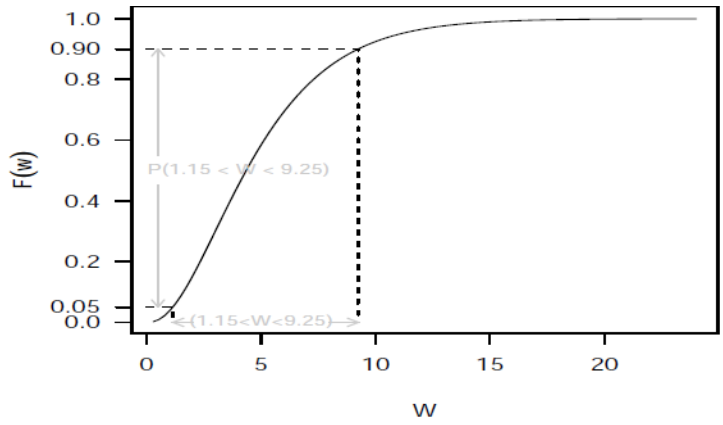


Figure 7.2.2: $P(2.76 < S^2 < 22.2)$

Moments

As V (defined in (7.1.2)) has a gamma distribution its mean and variance can be written down. That is, $V \sim \gamma(\nu/2)$, so that

$$E(V) = \nu/2 \quad \text{and} \quad Var(V) = \nu/2$$

Then since W is related to V by $W = 2V$

$$E(W) = 2(\nu/2) = \nu$$

and

$$Var(W) = 4(\nu/2) = 2\nu. \tag{7.2.2}$$

Thus, a random variable $W \sim \chi^2_\nu$ has mean ν and variance 2ν.

Exercise: find $E(W)$ and $Var(W)$ directly from $h(w)$.

Moment Generating Function

The MGF of a chi-square variate can be deduced from that of a gamma variate. Let $V \sim \gamma(\nu/2)$ and let $W = 2V$. We know $M_V(t) = (1-t)^{-\nu/2}$ from Statistics 260, Theorem 4.4. Hence

$$M_W(t) = M_{2V}() = M_V(2t) = (1-2t)^{-\nu/2}.$$

So if $W \sim \chi^2_\nu$ then

$$M_W(t) = (1-2t)^{-\nu/2}. \tag{7.2.3}$$

Exercise: Find the MGF of W directly from the pdf of W. (Hint: Use the substitution $u = w(1-2t)/2$ when integrating.)

To find moments, we will use the power series expansion of $M_W(t)$.

$$\begin{aligned} M_W(t) &= 1 + \frac{\nu}{2}\cdot 2t + \frac{\nu}{2}\left(\frac{\nu}{2}+1\right)\frac{(2t)^2}{2!} + \frac{\nu}{2}\left(\frac{\nu}{2}+1\right)\left(\frac{\nu}{2}+2\right)\frac{(2t)^3}{3!} + \cdots \\ &= 1 + \nu t + \nu(\nu+2)\frac{t^2}{2!} + \nu(\nu+2)(\nu+4)\frac{t^3}{3!} + \cdots \end{aligned}$$

Moments can be read off as appropriate coefficients here. Note that $\mu'_1 = \nu$ and $\mu'_2 = \nu(\nu+2)$. The cumulant generating function is

$$\begin{aligned} K_W(t) = \log M_W(t) &= -\frac{\nu}{2}\log(1-2t) \\ &= -\frac{\nu}{2}\left[-2t - \frac{2^2t^2}{2} - \frac{2^3t^3}{3} - \frac{2^4t^4}{4} - \cdots\right] \\ &= \nu t + \frac{2\nu t^2}{2!} + \frac{8\nu t^3}{3!} + \frac{48\nu t^4}{4!} + \cdots \end{aligned}$$

so the cumulants are

$$\kappa_1 = \nu \kappa_2 = 2\nu, \kappa_3 = 8\nu, \kappa_4 = 48\nu.$$

We will now use these cumulants to find measures of skewness and kurtosis for the chi-square distribution.

Comparison with Normal

1. Coefficient of skewness,

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{\frac{3}{2}}} = \frac{8\nu}{2\nu\sqrt{2\nu}} \text{for the } \chi^2_\nu \text{ distribution}$$
$$\to 0 \quad \text{as} \quad \nu \to \infty$$

That is, the χ^2 distribution becomes symmetric for $\nu \to \infty$.

2. Coefficient of kurtosis,

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} \text{for any distribution}$$
$$= \frac{48\nu}{4\nu^2} \text{for the } \chi^2 \text{ distribution}$$
$$\to 0 \quad \text{as} \quad \nu \to \infty.$$

This is the value γ_2 has for the normal distribution.

Additive Property

Let $W_1 \sim \chi^2_{\nu_1}$ and W_2 (independent of W_1) $\sim \chi^2_{\nu_2}$. Then from (7.2.3) $W_1 + W_2$ has moment generating function

$$M_{W_1+W_2}(t) = M_{W_1}(t)M_{W_2}(t) = (1-2t)^{-\nu_1/2}(1-2t)^{-\nu_2/2}$$
$$= (1-2t)^{-(\nu_1+\nu_2)/2}$$

This is also of the form (7.2.3); that is, we recognize it as the MGF of a χ^2 random variable on $(\nu_1 + \nu_2)$ degrees of freedom.

Thus if $W_1 \sim \chi^2_{\nu_1}$ and $W_2 \sim \chi^2_{\nu_2}$ and W_1 and W_2 are independent then

$$W_1 + W_2 \sim \chi^2_{\nu_1+\nu_2}$$

The result can be extended to the sum of k independent χ^2 random variables.

$$\text{If } W_1, \ldots, W_k \text{ are independent } \chi^2_{\nu_1}, \ldots, \chi^2_{\nu_k} \text{ then } \sum_{i=1}^{k} W_i \sim \chi^2_\nu \tag{7.2.4}$$

where $\nu = \sum \nu_i$. Note also that a χ^2_ν variate can be decomposed into a sum of ν independent chi-squares each on 1 d.f.

Chi-square on 1 degree of freedom

For the special case $\nu = 1$, note that from (7.1.1) if $Y \sim N(0, \sigma^2)$ then $V = \frac{Y^2}{2\sigma^2} \sim \gamma(1/2)$ and $W = 2V = Y^2/\sigma^2 \sim \chi^2_1$.

Thus if $Z = Y/\sigma$, it follows $Z \sim N(0,1)$ and

$$Z^2 \sim \chi^2_1. \tag{7.2.5}$$

(The square of a $N(0,1)$ random variable has a chi-square distribution on 1 df.)

Summary

You may find the following summary of relationships between χ^2, gamma, S^2 and normal distributions useful.

Define $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$, the X_i being independent $N(\mu, \sigma^2)$ variates, then

1. $W = \nu S^2/\sigma^2 \sim \chi^2_\nu$ where $\nu = n - 1$,
2. $\frac{1}{2}W = \nu S^2/2\sigma^2 \sim \gamma(\nu/2)$,
3. If $Z_i = \frac{X_i - \mu}{\sigma}$, (that is, $Z_i \sim N(0,1)$) then

$$Z_i^2 \sim \chi^2_1 \quad \text{and} \quad Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi^2_k$$

7.3 Independence of $\overline{X}$ and S^2

When $\overline{X}$ and S^2 are defined for a sample from a normal distribution, $\overline{X}$ and S^2 are statistically independent. This may seem surprising as the expression for S^2 involves $\overline{X}$.

Consider again the transformation from X's to Y's given in 7.1. We've seen that $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$ can be expressed as $\sum_{j=1}^{\nu} Y_j^2$ where the Y_j defined by

$$Y_j = \frac{X_1 + X_2 + \cdots + X_j - jX_{j+1}}{\sqrt{j(j+1)}}, \quad j = 1, 2, \cdots, n-1,$$

have zero means and variances σ^2. note also that the sample mean,

$$\overline{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

is a linear function of $X_1, \ldots, X_n$ which is orthogonal to each of the Y_j, and hence uncorrelated with each Y_j. Since the X_i are normally distributed, $\overline{X}$ is thus independent of each of the Y_j and therefore independent of any function of them.

Thus when $X_1, \ldots, X_n$ are normally and independently distributed random variables $\overline{X}$ and S^2 are statistically independent.

7.4 Confidence intervals for σ^2

We will use the method indicated in 4.11 to find a confidence interval for σ^2 in a normal distribution, based on a sample of size n. The two cases (i) μ unknown; (ii) μ known must be considered separately.

Case (i)

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. It has been shown that S^2 is an unbiased estimate of σ^2 (Theorem 4.14) and we can find a confidence interval for σ^2 using the χ^2 distribution. Recall that $W = \nu S^2/\sigma^2 \sim \chi^2_\nu$. By way of notation, let $w_{\nu,\alpha}$ be defined by $P(W > w_{\nu,\alpha}) = \alpha$, where $W \sim \chi^2_\nu$.

The quantile for the upper 5% region is obtained by

```
qchisq(p=0.05,df=5,lower.tail=F)
```

or

```
qchisq(p=0.95,df=5)
```

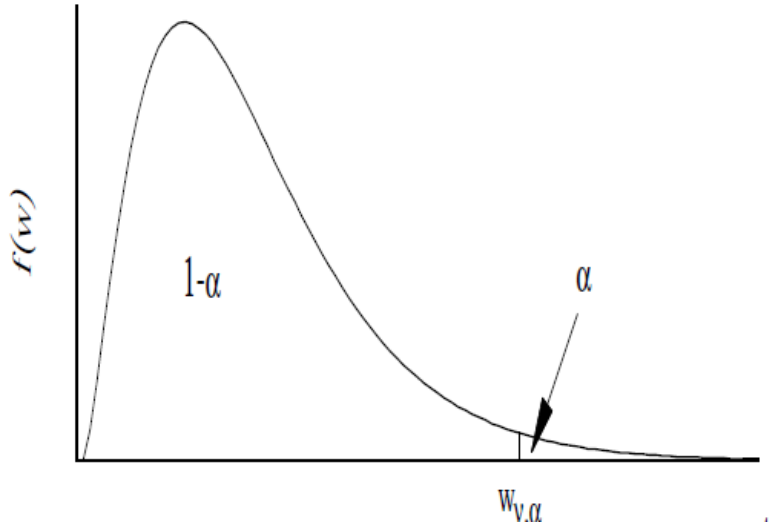


Figure 7.4.1: Area above $w_{\nu,\alpha}$

We find two values of W, $w_{\nu,\alpha/2}$ and $w_{\nu,1-(\alpha/2)}$, such that

$$P\left(w_{\nu,1-(\alpha/2)} < W < w_{v,\alpha/2}\right) = 1 - \alpha.$$

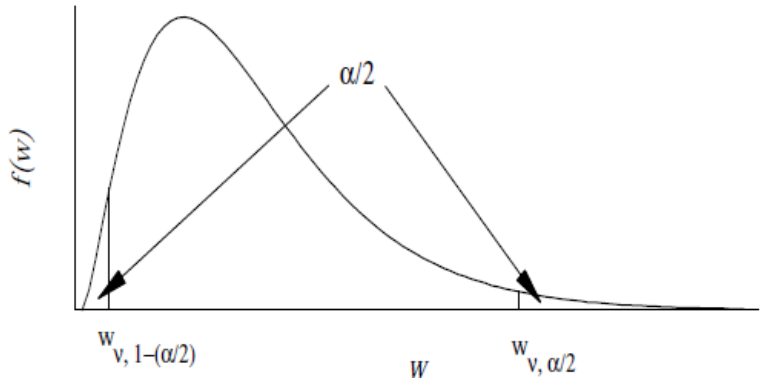


Figure 7.4.2: Upper and lower values for w

The event $w_{\nu,1-(\alpha/2)} < W < w_{\nu,\alpha/2}$ occurs iff the events

$$\sigma^2 < \nu S^2/w_{\nu,1-(\alpha/2)}, \sigma^2 > \nu S^2/w_{\nu,\alpha/2}$$

occur. So

$$P\left(w_{\nu,1-(\alpha/2)} < W < w_{\nu,\alpha/2}\right) = P\left(\nu S^2/w_{\nu,\alpha/2} < \sigma^2 < \nu S^2/w_{\nu,1-(\alpha/2)}\right)$$

and thus

$$\text{A } 100(1-\alpha)\% \text{ CI for } \sigma^2 \text{ is } \left(\nu s^2/w_{\nu,\alpha/2}, \nu s^2/w_{\nu,1-(\alpha/2)}\right) \tag{7.4.1}$$

Example 7.2.
For a sample of size $n = 10$ *from a normal distribution* s^2 *was calculated and found to be* 6.4. *Find a* 95% *CI for* σ^2.

Solution of Example 7.2. *Now* $\nu = 9$, *and*

```
qchisq(p=c(0.025,0.975),df=9,lower.tail=F)
[1] 19.0 2.7
```

$w_{9,.025} = 19$ *and* $w_{9,.975} = 2.7$.
Hence, $\nu s^2/w_{9,.025} = 3.02$, *and* $\nu s^2/w_{9,.975} = 21.33$.
That is, the 95% *CI for* σ^2 *is* $(3.02, 21.33)$.

Case (ii)

Suppose now that $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ where μ is known and we wish to find a CI for the unknown σ^2. Recall that the maximum likelihood estimator of σ^2 (which we'll denote by S^{*2}) is

$$S^{*2} = \sum_{i=1}^{n} (X_i - \mu)^2/n.$$

We can easily show that this is unbiased.

$$E(S^{*2}) = \sum_{i=1}^{n} \frac{E(X_i - \mu)^2}{n} = n\frac{1}{n}\sigma^2 = \sigma^2$$

The distribution of S^{*2} is found by noting that $nS^{*2}/\sigma^2 = \sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2$ is the sum of squares of n independent $N(0,1)$ variates and is therefore distributed as χ^2_n (using (7.2.4) and (7.2.5)). Proceeding in the same way as in Case (i) we find

$$\text{A } 100(1-\alpha)\% \text{ CI for } \sigma^2 \text{ when } \mu \text{ is known is } \left(\frac{ns^{*2}}{w_{n,\alpha/2}}, \frac{ns^{*2}}{w_{n,1-(\alpha/2)}} \right) \tag{7.4.2}$$

7.5 Testing hypotheses about σ^2

Again the cases (i) μ unknown; and (ii) μ known are considered separately.

Case (i)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution where μ **is unknown**, and suppose we wish to test the hypothesis

$$H : \sigma^2 = \sigma_0^2 \quad \text{against} \quad A : \sigma^2 \neq \sigma_0^2.$$

Under H, $\nu S^2/\sigma_0^2 \sim \chi^2_\nu$ and values of $\nu s^2/\sigma_0^2$ too large or too small would support A. For $\alpha = .05$, say, and equal-tail probabilities we have as critical region

$$R = \left\{ s^2 : \frac{\nu s^2}{\sigma_0^2} > w_{\nu,.025} \text{ or } \frac{\nu s^2}{\sigma_0^2} < w_{\nu,.975} \right\}.$$

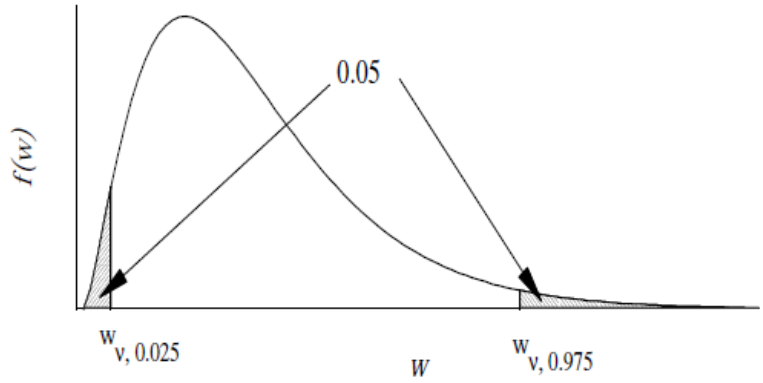


Figure 7.5.1: Critical Region

Consider now a one-sided alternative. Suppose we wish to test

$$H : \sigma^2 = \sigma_0^2 \text{ against } A : \sigma^2 > \sigma_0^2.$$

Large values of s^2 would support this alternative. That is, for $\alpha = .05$, use as critical region

$$\{s^2 : \nu s^2/\sigma_0^2 > w_{\nu,.05}\}.$$

Similarly, for the alternative $A : \sigma^2 < \sigma_0^2$, a critical region is

$$\{s^2 : \nu s^2/\sigma_0^2 < w_{\nu,.95}\}.$$

Example 7.3.
A normal random variable has been assumed to have standard deviation $\sigma = 7.5$. If a sample of size 25 has $s^2 = 95.0$, is there reason to believe that σ is greater than 7.5?

Solution of Example 7.3. *We wish to test $H : \sigma^2 = 7.5^2$ $(= \sigma_0^2)$ against $A : \sigma^2 > 7.5^2$.*
Using $\alpha = .05$, the rejection region is $\{s^2 : \nu s^2/\sigma_0^2 > 36.4\}$.
The calculated value of $\nu s^2/\sigma^2$ is $\frac{24\times 95}{56.25} = 40.53$.

```
> pchisq(q=40.53,df=24,lower.tail=F)
[1] 0.019
```

When testing at the 5% level, there is evidence that the standard deviation is greater than 7.5.

Case (ii)

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where μ **is known**, and suppose we wish to test $H : \sigma^2 = \sigma_0^2$. Again we use the fact that **if** H **is true**, $nS^{*2}/\sigma_0^2 \sim \chi_n^2$ where $S^{*2} = \sum_{i=1}^n (X_i - \mu)^2/n$, and the rejection region for a size-α 2-tailed test, for example, would be

$$\left\{ s^{*2} : \frac{ns^{*2}}{\sigma_0^2} > w_{n,\alpha/2} \quad \text{or} \quad \frac{ns^{*2}}{\sigma_0^2} < w_{n,1-(\alpha/2)} \right\}$$

7.6 χ^2 and Inv-χ^2 distributions in Bayesian inference

7.6.1 Non-informative priors

A prior which does not change very much over the region in which the likelihood is appreciable and does not take very large values outside that region is said to be locally uniform.

For such a prior,

$$p(\theta|y) \propto p(y|\theta) = \ell(\theta|y)$$

The term pivotal quantity was introduced p. 107 and now is defined for (i) location parameter and (ii) scale parameter.

1. If the density of y, $p(y|\theta)$, is such that $p(y-\theta|\theta)$ is a function that is free of y and θ, say $f(u)$ where $u = y - \theta$, then $y - \theta$ is a pivotal quantity and θ is a *location parameter*.

   Example: if $(y|\mu, \sigma^2) \sim N(\mu, \sigma^2)$, then $(y - \mu|\mu, \sigma^2) \sim N(0, \sigma^2)$ and $y - \mu$ is a pivotal quantitiy.

2. If $p\left(\frac{y}{\phi}|\phi\right)$ is a function free of ϕ and y, say $g(u)$ where $u = \frac{y}{\phi}$, then u is a pivotal quantitity and ϕ is a *scale parameter*.

   Example: if $(y|\mu, \sigma^2) \sim N(\mu, \sigma^2)$, then $\frac{y-\mu}{\sigma} \sim N(0, 1)$.

A non-informative prior for a **location** parameter, θ, would give $f(y - \theta)$ for the posterior distribution $p(y - \theta|y)$. That is under the posterior distribution, $(y - \theta)$ should still be a pivotal quantity.

Using Bayes' rule,

$$p(y - \theta|y) \propto p(\theta)p(y - \theta|\theta)$$

Thus $p(\theta) \propto C$, where C is a constant.

For the case of a **scale** parameter, ϕ, Bayes' rule is

$$p\left(\frac{y}{\phi}|y\right) \propto p(\phi)p\left(\frac{y}{\phi}|\phi\right) \tag{7.6.1}$$

$$p(u|y) \propto p(\phi)p(u|\phi) \tag{7.6.2}$$

(The LHS of the first of those equations is the posterior of a parameter say $\phi^* = \frac{y}{\phi}$ and the RHS is the density of a scaled variable $y^* = \frac{y}{\phi}$. Both sides are free of y and ϕ.)

$$p(y|\phi) = p(u|\phi)\left|\frac{du}{dy}\right| = \frac{1}{\phi}p(u|\phi)$$

$$p(\phi|y) = p(u|y)\left|\frac{du}{d\phi}\right| = \frac{y}{\phi^2}p(u|y)$$

Thus, from the last numbered equation, equate $p(u|y)$ to $p(u|\phi)$,

$$p(\phi|y) = \frac{y}{\phi}p(y|\phi)$$

so that the uninformative prior is

$$p(\phi) \propto \frac{1}{\phi} \tag{7.6.3}$$

7.7 The posterior distribution of the Normal variance

Consider normally distributed data,

$$y|\mu,\sigma^2 \sim N(\mu,\sigma^2)$$

The joint posterior density of parameters μ,σ^2 is given by

$$p(\mu,\sigma^2) \propto p(y|\mu,\sigma^2) \times p(\mu,\sigma^2) \tag{7.7.1}$$

To get the marginal posterior distribution of the variance, integrate with respect to μ,

$$p(\sigma^2|y) = \int p(\mu,\sigma^2|y)d\mu \tag{7.7.2}$$

$$= \int p(\sigma^2|\mu,y)p(\mu|y)d\mu \tag{7.7.3}$$

Choose the prior

$$p(\mu,\sigma^2) \propto p(\mu)p(\sigma^2) \quad (\mu \perp \sigma^2) \tag{7.7.4}$$

$$p(\mu,\sigma^2) \propto (\sigma^2)^{-1} \quad (p(\mu) \propto C, \text{constant}) \tag{7.7.5}$$

Write the posterior density as

$$\begin{aligned} p(\mu,\sigma^2) &\propto \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y-\mu)^2\right\} \\ &= \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\overline{y})^2 + n(\overline{y}-\mu)^2\right]\right\} \\ &= \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\overline{y}-\mu)^2\right]\right\} \end{aligned}$$

where $S^2 = \frac{\sum(y_i-\overline{y})^2}{(n-1)}$.

Now integrate the joint density with respect to μ,

$$\begin{aligned} p(\sigma^2|y) &\propto \int \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\overline{y}-\mu)^2\right]\right\} \\ &= \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\}\int \exp\left\{-\frac{1}{2\sigma^2/n}(\overline{y}-\mu)^2\right\}d\mu \\ &= \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\}\sqrt{2\pi\sigma^2/n} \end{aligned}$$

which equals

$$(\sigma^2)^{-\frac{n+1}{2}}\exp\left\{-\frac{(n-1)S^2}{2\sigma^2}\right\} \tag{7.7.6}$$

The pdf of S^2 was derived at (7.1.3),

$$\begin{aligned} g(s^2) &= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \times (s^2)^{\left(\frac{\nu}{2}-1\right)}\left(\frac{\nu}{2\sigma^2}\right)^{\frac{\nu}{2}}\exp\left\{-\frac{\nu s^2}{2\sigma^2}\right\} \\ &\propto (s^2)^{\left(\frac{\nu}{2}-1\right)}\exp\left\{-\frac{\nu s^2}{2\sigma^2}\right\} \end{aligned}$$

with $\nu = (n-1)$ and this is a $\Gamma\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right)$ distribution.

7.7.1 Inverse Chi-squared distribution

Its Bayesian counterpart at (7.7.6) is a *Scaled Inverse Chi-squared distribution*. Since the prior was uninformative, simular outcomes are expected.

The inverse χ^2 distribution has density function

$$p(\sigma^2|\nu) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{1}{2}\right)^{\frac{\nu}{2}}\left(\frac{1}{\sigma^2}\right)^{\frac{\nu}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}\right\} \times I_{(0,\infty)}(\sigma^2).$$

The scaled inverse chi-squared distribution has density

$$p(\sigma^2|\nu, s^2) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}(\sigma^2)^{-\left(\frac{\nu}{2}+1\right)}\exp\left\{-\frac{\nu s^2}{2\sigma^2}\right\}$$

The prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$ can be said to be an inverse chi-squared distribution on $\nu = 0$ degrees of freedom or sample size $n = 1$. Is there any value in it? Although uninformative, it ensures a mathematical "smoothness" and numerical problems are reduced.

The posterior density is Scaled Inverse Chi-squared with degrees of freedom $\nu = (n-1)$ and scale parameter s.

7.8 Relationship between χ^2_ν and Inv-χ^2_ν

Recall that χ^2_ν is $\Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right)$. The Inverse-Gamma distribution is also prominent in Bayesian statistics so we examine it first.

7.8.1 Gamma and Inverse Gamma

The densities of the Gamma and Inverse Gamma are:

$$\text{Gamma } p(\theta|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\theta^{(\alpha-1)}\beta^\alpha \exp\{-\beta\theta\} \times I_{0,\infty}(\theta)\ \alpha,\beta > 0 \tag{7.8.1}$$

$$\text{Inverse Gamma } p(\theta|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\beta^\alpha \exp\{-\beta/\theta\} \times I_{0,\infty}(\theta)\ \alpha,\beta > 0 \tag{7.8.2}$$

If $\theta^{-1} \sim \Gamma(\alpha,\beta)$, then $\theta \sim \Gamma^{-1}(\alpha,\beta)$.

Put $\phi = \theta^{-1}$. Then

$$\begin{aligned} f(\theta;\alpha,\beta) &= f(\phi^{-1};\alpha,\beta)\left|\frac{d\phi}{d\theta}\right| \\ &= \frac{1}{\Gamma(\alpha)}\theta^{-(\alpha-1)}\beta^\alpha \exp\left\{-\frac{\beta}{\theta}\right\}\theta^{-2} \\ &= \frac{1}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\beta^\alpha \exp\left\{-\frac{\beta}{\theta}\right\} \end{aligned}$$

7.8.2 Chi-squared and Inverse Chi-squared

If $Y = SX$ such that $Y^{-1} \sim S^{-1}\chi^2_\nu$, then Y is S times an inverse χ^2 distribution. The Inverse-$\chi^2(\nu, s^2)$ distribution is a special case of the Inverse Gamma distribution with $\alpha = \frac{\nu}{2}$ and $\beta = \frac{\nu s^2}{2}$.

7.8.3 Simulating Inverse Gamma and Inverse-χ^2 random variables

- **InvGa.** Draw X from $\Gamma(\alpha, \beta)$ and invert it.
- **ScaledInv** – χ^2_{ν,s^2}. Draw X from χ^2_ν and let $Y = \frac{\nu s^2}{X}$.

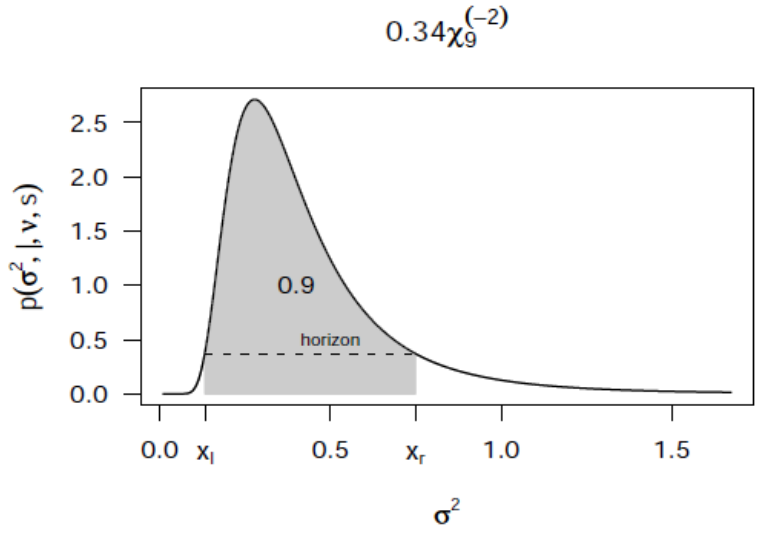
Example 7.4.
Give a 90% HDR for the variance of the population from which the following sample is drawn.
4.17 5.58 5.18 6.11 4.50 4.61 5.17 4.53 5.33 5.14

$$S^2 = 0.34$$
$$p(\sigma^2|\nu, S^2) = 0.34\chi_9^{-2}$$

The 90% CI for σ^2 *is* $(0.18, 0.92)$. *The mode of the posterior density of* σ^2 *is 0.28 and the 90% HDR for* σ^2 *is* $(0.13, 0.75)$.

The HDR was calculated numerically in this fashion,

1. *Calculate the posterior density,* (7.8.2)
2. *Set an initial value for the "horizon," estimate the abcissas (left and right of the mode) whose density is at the horizon. Call these* x_l *and* x_r
3. *Integrate the density function over* (x_i, x_r).
4. *Adjust the horizon until this is 0.9. The HDR is then* (x_l, x_r) *at the current values.*




```
#_________________ to calculate HDR of \sigma^2 _____________________
options(digits=2)
#_________________ functions to use later in the job ___________________
closest <- function(s,v){
delta <- abs(s-v)
p <- delta==min(delta)
return(p) }
# _________________
IGamma <- function(v,a=df/2,b=0.5*df*S2){
p <- (1/gamma(a))* (v**(-(a+1)) ) * (b**a) * exp(-b/v)
```

```
return(p) }
#_____________________________________________________________________________
wts <- c(4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14) # the data
n <- length(wts); S2 <- Var(wts); df <- n - 1 # statistics
cat("S-sq = ",S2,"\n")
# ___________ 90% CI _____________
Q <- qchisq(p=c(0.95,0.05),df=df)
CI <- df*S2/Q
cat("CI.sigma = ",CI,"\n")
# _____________ Posterior ___________________
Ew <- df*S2/(df-2)
Vw <- (2*df^2*S2^2)/((df-2)^2*(df-4)^2)
w <- seq(0.01,(Ew+10*sqrt(Vw)),length=501)
ifw <- IGamma(v=w)
mode <- w[closest(max(ifw),ifw)]
# ________ deriving the HDR by numerical integration ___________
PHDR <- 0.9 # this is the level of HDR we want
step <- 0.5; convergence.test <- 1e3; prop <- 0.9 # scalar variables for the numerical steps
while (convergence.test > 1e-3 ){ # iterate until the area is very close to 0.9
horizon <- max(ifw)*prop
left.ifw <- subset(ifw,subset=w < mode);lw <- w[w < mode]
right.ifw <- subset(ifw,subset=w > mode);rw <- w[w > mode]
xl <- lw[closest(horizon,left.ifw)]
xr <- rw[closest(horizon,right.ifw)]
Pint <- integrate(f=IGamma,lower=xl,upper=xr)
convergence.test <- abs(Pint$value - PHDR)
adjust.direction <- 2*(0.5 - as.numeric(Pint$value < PHDR)) # -1 if < +1 if >
prop <- prop+ adjust.direction*step*convergence.test
} # end of while loop
HDR <- c(xl,xr)
cat("HDR = ",HDR,"\n")
```