

Tutorial 5 Solutions

Question 1

(a) (10.46) Consider the large sample α -level test of

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0,$$

which rejects the null hypothesis if

$$\frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} > z_{\alpha}.$$

Show that this is equivalent to rejecting H_0 if θ_0 is less than the $100(1 - \alpha)\%$ lower confidence bound for θ .

We reject H_0 if:

$$\begin{aligned} & \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} > z_{\alpha} \\ \Leftrightarrow & \hat{\theta} - \theta_0 > z_{\alpha} \sigma_{\hat{\theta}} \\ \Leftrightarrow & -\theta_0 > -\hat{\theta} + z_{\alpha} \sigma_{\hat{\theta}} \\ \Leftrightarrow & \theta_0 < \hat{\theta} - z_{\alpha} \sigma_{\hat{\theta}} \end{aligned}$$

The right-hand side is the formula for a $100(1 - \alpha)\%$ lower confidence bound. Therefore, it is equivalent to reject H_0 if θ_0 is less than the $100(1 - \alpha)\%$ lower confidence bound for θ . Equivalently, this also means to reject H_0 if θ_0 is not found in the interval

$$(\hat{\theta} - z_{\alpha} \sigma_{\hat{\theta}}, \infty).$$

(b) (10.48) Consider the large sample α -level test of

$$H_0 : \theta \geq \theta_0 \quad \text{vs} \quad H_1 : \theta < \theta_0,$$

which rejects the null hypothesis if

$$\frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} < -z_{\alpha}.$$

Show that this is equivalent to rejecting H_0 if θ_0 is greater than the $100(1 - \alpha)\%$ upper confidence bound for θ .

We reject H_0 if:

$$\begin{array}{lll}
 \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} & < -z_{\alpha} \\
 \Leftrightarrow \hat{\theta} - \theta_0 & < -z_{\alpha} \sigma_{\hat{\theta}} \\
 \Leftrightarrow -\theta_0 & < -\hat{\theta} - z_{\alpha} \sigma_{\hat{\theta}} \\
 \Leftrightarrow \theta_0 & > \hat{\theta} + z_{\alpha} \sigma_{\hat{\theta}}
 \end{array}$$

The right-hand side is the formula for a $100(1-\alpha)\%$ upper confidence bound. Therefore, it is equivalent to reject H_0 if θ_0 is greater than the $100(1-\alpha)\%$ upper confidence bound for θ . Equivalently, this also means to reject H_0 if θ_0 is not found in the interval

$$(-\infty, \hat{\theta} + z_{\alpha} \sigma_{\hat{\theta}})$$

Question 2

(10.106) A survey of voter sentiment was conducted in four midcity political wards to compare the fraction of voters favouring candidate A. Random samples of 200 voters were polled in each of the four wards, with the results as shown in the table below. The number of voters favouring A in the four samples can be regarded as four independent binomial random variables.

- (a) Construct a likelihood ratio test of the hypothesis that the fractions of voters favour candidate A are the same in the four wards.

The hypotheses of interest are:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p \quad \text{vs} \quad H_1 : \text{At least one } p_i \neq p, \quad i = 1, 2, 3, 4.$$

The general likelihood function is given by the product of the marginal densities:

$$\mathcal{L}(p_1, p_2, p_3, p_4 | \mathbf{y}) = \prod_{i=1}^{n=4} \binom{200}{y_i} p_i^{y_i} (1 - p_i)^{200 - y_i}.$$

Under H_0 , $p_1 = p_2 = p_3 = p_4 = p$, so that

$$\Omega_0 = \{(p_1, p_2, p_3, p_4) : p_1 = p_2 = p_3 = p_4 = p\},$$

and thus, under the conditions of H_0 , the general likelihood function reduces to a function of a single parameter, p , and is given by

$$\mathcal{L}_{\Omega_0}(p | \mathbf{y}) = \prod_{i=1}^{n=4} \binom{200}{y_i} p^{y_i} (1 - p)^{200 - y_i}.$$

We now use the usual procedures for maximum likelihood estimation to find a \hat{p} that will maximize $\mathcal{L}_{\Omega_0}(p|\mathbf{y})$.

$$\begin{aligned}\ell(p|\mathbf{y}) &= \log(\mathcal{L}_{\Omega_0}(p|\mathbf{y})) \\ &= \sum_{i=1}^{n=4} \log\left(\binom{200}{y_i}\right) + \sum_{i=1}^{n=4} y_i \log(p) + \sum_{i=1}^{n=4} (200 - y_i) \log(1 - p)\end{aligned}$$

The first derivative of the log-likelihood (under the conditions of H_0) is:

$$\frac{d\ell}{dp} = \sum_{i=1}^{n=4} \left(\frac{y_i}{p} - \frac{(200 - y_i)}{1 - p} \right)$$

The MLE occurs when the first derivative equals zero:

$$\begin{aligned}\left. \frac{d\ell}{dp} \right|_{p=\hat{p}} &= 0 \\ \sum_{i=1}^{n=4} \left(\frac{y_i}{\hat{p}} - \frac{(200 - y_i)}{1 - \hat{p}} \right) &= 0 \\ \sum_{i=1}^{n=4} (y_i(1 - \hat{p}) - (200 - y_i)\hat{p}) &= 0 \\ \sum_{i=1}^{n=4} (y_i - y_i\hat{p} - 200\hat{p} + y_i\hat{p}) &= 0 \\ \sum_{i=1}^{n=4} (y_i - 200\hat{p}) &= 0 \\ \sum_{i=1}^{n=4} y_i - 200\hat{p} \sum_{i=1}^{n=4} 1 &= 0 \\ \sum_{i=1}^{n=4} y_i - 200(4)\hat{p} &= 0 \\ \frac{\sum_{i=1}^{n=4} y_i}{800} &= \hat{p}\end{aligned}$$

We should verify that \hat{p} obtained above does indeed maximize $\ell(p)$ by checking that

$$\left. \frac{d^2\ell}{dp^2} \right|_{p=\hat{p}} < 0.$$

I leave this as an exercise for the reader.

Using the above information, we find the numerator of our likelihood ratio as:

$$\mathcal{L}_{\Omega_0}(\hat{p}|\mathbf{y}) = \prod_{i=1}^{n=4} \binom{200}{y_i} \left(\frac{\sum_{i=1}^{n=4} y_i}{800} \right)^{y_i} \left(1 - \frac{\sum_{i=1}^{n=4} y_i}{800} \right)^{200-y_i}$$

Under the alternative hypothesis, we have

$$\Omega_1 = \{(p_1, p_2, p_3, p_4) : p_i \neq p \text{ for at least one of } i = 1, 2, 3, 4\}.$$

Then

$$\begin{aligned}\Omega &= \Omega_0 \cup \Omega_1 \\ &= \{(p_1, p_2, p_3, p_4) : p_1 > 0, p_2 > 0, p_3 > 0, p_4 > 0\}.\end{aligned}$$

Since the general likelihood was a product of the component likelihoods (marginal densities), intuitively, to maximize the general likelihood under Ω would be to maximize the individual component likelihoods. We know that for a single binomial sample with a known n , the likelihood function is maximized at $\hat{p} = y/n$. It then follows that the general likelihood is maximized over Ω using $\hat{p}_i = y_i/200$, $i = 1, 2, 3, 4$. The denominator of our likelihood ratio is:

$$\mathcal{L}_\Omega(\hat{\mathbf{p}} | \mathbf{y}) = \prod_{i=1}^{n=4} \binom{200}{y_i} \left(\frac{y_i}{200}\right)^{y_i} \left(1 - \frac{y_i}{200}\right)^{200-y_i}$$

Our likelihood ratio is:

$$\begin{aligned}\lambda &= \frac{\mathcal{L}_{\Omega_0}(\hat{p} | \mathbf{y})}{\mathcal{L}_\Omega(\hat{\mathbf{p}} | \mathbf{y})} \\ &= \frac{\prod_{i=1}^{n=4} \binom{200}{y_i} \left(\frac{\sum_{i=1}^{n=4} y_i}{800}\right)^{y_i} \left(1 - \frac{\sum_{i=1}^{n=4} y_i}{800}\right)^{200-y_i}}{\prod_{i=1}^{n=4} \binom{200}{y_i} \left(\frac{y_i}{200}\right)^{y_i} \left(1 - \frac{y_i}{200}\right)^{200-y_i}} \\ &= \frac{\left(\frac{\sum_{i=1}^{n=4} y_i}{800}\right)^{\sum_{i=1}^{n=4} y_i} \left(1 - \frac{\sum_{i=1}^{n=4} y_i}{800}\right)^{800 - \sum_{i=1}^{n=4} y_i}}{\prod_{i=1}^{n=4} \left(\frac{y_i}{200}\right)^{y_i} \left(1 - \frac{y_i}{200}\right)^{200-y_i}}\end{aligned}$$

For large n , $-2 \log(\lambda)$ has an approximate chi-square distribution with degrees of freedom given by

$$k = \dim(\Omega) - \dim(\Omega_0),$$

where $\dim(\cdot)$ is the number of free parameters. In the context of this question,

- $\dim(\Omega)$ has four free parameters since there are no restrictions on the values of the p_i s.
- $\dim(\Omega_0)$ has one free parameter, for if we fix one p_i , the condition under H_0 requires that the other p_i s are identical.

Therefore,

$$k = \dim(\Omega) - \dim(\Omega_0) = 4 - 1 = 3.$$

This hypothesis test is an upper-tailed test. As such, we will reject H_0 if

$$-2\log(\lambda) > \chi_{k,\alpha}^2.$$

Note that:

$$\begin{aligned} & -2\log(\lambda) \\ &= -2 \left(\sum_{i=1}^{n=4} y_i \log \left(\frac{\sum_{i=1}^{n=4} y_i}{800} \right) + \left(800 - \sum_{i=1}^{n=4} y_i \right) \log \left(1 - \frac{\sum_{i=1}^{n=4} y_i}{800} \right) \right. \\ & \quad \left. - \sum_{i=1}^{n=4} y_i \log \left(\frac{y_i}{200} \right) - \sum_{i=1}^{n=4} (200 - y_i) \log \left(1 - \frac{y_i}{200} \right) \right) \end{aligned}$$

- (b)  Using the data provided in the table below, carry out the hypothesis test using $\alpha = 0.05$.

Opinion	Ward				Total
	1	2	3	4	
Favor A	76	53	59	48	236
Do not favor A	124	147	141	152	564
Total	200	200	200	200	800

We will store the data into a variable named `y`. I will also store the sum of the `y` values into a variable named `sum_y` since we will require the sum of the `y` values many times.

```
y <- c(76, 53, 59, 48)
sum_y <- sum(y)
```

We then compute $-2\log(\lambda)$ as given above. This is another instance where we take advantage of the vectorization of mathematical operations in **R**.

```
statistic <- -2 * (sum_y * log(sum_y / 800) + (800 - sum_y) * log(1 - sum_y / 800)
- sum(y * log(y / 200)) - sum((200 - y) * log(1 - y / 200)))
```

As stated in (a), the test statistic has an approximate chi-square distribution on 3 degrees of freedom and we reject the null hypothesis if the value of the observed test statistic exceeds the critical value of $\chi_{3,0.05}^2$.

```
qchisq(0.05, df=3, lower.tail=FALSE)
```

```
## [1] 7.814728
```

Since $10.535 > 7.815$, we reject the null hypothesis in favour of the alternative and conclude that there is sufficient evidence to support the claim that the proportion of voters supporting candidate A is not the same among the four wards.


As this is an upper-tailed test, we also could have computed an (approximate) p -value, which is the area to the right of the observed value of our test statistic, under the chi-square distribution on 3 degrees of freedom.

```
pchisq(statistic, df=3, lower.tail=FALSE)
```

```
## [1] 0.01452352
```

As $0.0145 < 0.05$, we reject the null hypothesis in favour of the alternative once again, and make the same conclusion.

Question 3

(10.124)  The data in the table below gives readings in foot-pounds of the impact strength of two kinds of packaging material, type A and type B. Determine whether the data suggests a difference in mean strength between the two kinds of material. Test at the $\alpha = 0.10$ level of significance.

A	B
1.25	.89
1.16	1.01
1.33	.97
1.15	.95
1.23	.94
1.20	1.02
1.32	.98
1.28	1.06
1.21	.98
$\sum y_i = 11.13$	$\sum y_i = 8.80$
$\bar{y} = 1.237$	$\bar{y} = .978$
$\sum y_i^2 = 13.7973$	$\sum y_i^2 = 8.6240$

The hypotheses that we wish to test are:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

We will perform a pooled t -test. In order to perform a t -test we make the assumption that the data come from a normal distribution. In order to use the pooled variance, we make the assumption that the two samples come from populations with identical variances, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

We begin by entering the data into **R**.

```
sample_a <- c(1.25, 1.16, 1.33, 1.15, 1.23, 1.20, 1.32, 1.28, 1.21)
sample_b <- c(0.89, 1.01, 0.97, 0.95, 0.94, 1.02, 0.98, 1.06, 0.98)
```

You can read the documentation on how to perform t -tests in **R** by calling `?t.test` in the console. Make note of the default values of the arguments.

A two-sided pooled t -test for a difference in means using $\alpha = 0.10$ is performed as follows:

```
t.test(sample_a, sample_b, var.equal=TRUE, conf.level=0.90)
```

```
##
## Two Sample t-test
##
## data: sample_a and sample_b
## t = 9.5641, df = 16, p-value = 5.088e-08
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
## 0.2116300 0.3061477
```

```
## sample estimates:
## mean of x mean of y
## 1.2366667 0.9777778
```

The p -value of this test is obtained as 5.088×10^{-8} . Since the p -value is less than 0.10, we reject the null hypothesis. We conclude that there is sufficient evidence to support the claim that there is a difference in mean strength between the two types of materials.

Exercise: Use the summary values in the table given above to perform this hypothesis test by hand. It may be useful to recall that

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \left(\left(\sum_{i=1}^n X_i^2 \right) - n\bar{X}^2 \right) \\ &= \frac{1}{n-1} \left(\left(\sum_{i=1}^n X_i^2 \right) - \frac{(\sum_{i=1}^n X_i)^2}{n} \right) \end{aligned}$$