Dr Isabella Gollini

Probability and Statistics

Solutions 5

1. (i) We assume that the data are a random sample from a $N(\mu, \sigma^2)$ distribution, where μ and σ^2 are unknown. We test the null hypothesis $H_0 : \mu = 58$ against the one-sided alternative $H_1 : \mu < 58$.

(ii) We first calculate the sample mean, $\bar{x} = 55.995$, and the sample variance, $s^2 = 83.148$. To test H_0 we use the statistic

$$t = \frac{(\bar{x} - 24.9)}{\frac{s}{\sqrt{n}}} = \frac{(55.995 - 58)}{\sqrt{\frac{83.148}{40}}} = -1.391,$$

which under H_0 has the t_{39} distribution. Using a one-tail test, we reject H_0 if t is small enough. If F denotes the distribution function of the t_{39} distribution then, using Table 9 of *Lindley and Scott*, with some interpolation, the p-value is given by

$$p = \Pr(t \leq -1.391) = 1 - F(1.391) = 1 - 0.914 = 0.086 \quad \text{(to 3 decimal places)}.$$

We do not reject the null hypothesis at the 5% significance level, i.e., there is no strong evidence that the mean weight of dogs of the given breed is less than the 58 ounces stated by the Kennel Club.

Alternatively we could compare our calculated t-value with the percentage points of the t-distribution with 39 degrees of freedom. From Table 10, $t_{39}(5) = 1.685$. Since the modulus of our calculated value 1.39 is smaller than this percentage point, we do not reject the null hypothesis at the 5% significance level.

(iii) An estimate of the mean weight μ for this breed is $\bar{x} = 55.995$ ounces.

A 95% confidence interval for the mean weight is given by

$$\left(55.995 - t_{39}(2.5)\sqrt{\frac{83.148}{40}}\ ,\ 55.995 + t_{39}(2.5)\sqrt{\frac{83.148}{40}}\right)$$

i.e.

$$(55.995 - (2.023)(1.442)\ ,\ 55.995 + (2.023)(1.442)) \;=\; (53.08, 58.91)\ .$$

A 99% confidence interval is given by

$$\left(55.995 - t_{39}(0.5)\sqrt{\frac{83.148}{40}}\ ,\ 55.995 + t_{39}(0.5)\sqrt{\frac{83.148}{40}}\right)$$

i.e.

$$(55.995 - (2.708)(1.442)\ ,\ 55.995 + (2.708)(1.442)) \;=\; (52.09, 59.90)\ .$$

(iv) An estimate of the variance of weights within the breed is $s^2 = 83.148$. A 95% confidence interval for the variance is given by

$$\left(\frac{39s^2}{\chi^2_{39}(2.5)} \, , \, \frac{39s^2}{\chi^2_{39}(97.5)} \right)$$

i.e., using Table 8 of *Lindley and Scott*,

$$\left(\frac{(39)(83.148)}{58.12} \, , \, \frac{(39)(83.148)}{23.65} \right) = (55.8, 137.1) \, .$$

Using R:

(i)
```
weight <- c(66.2, 59.2, 70.8, 58.0, 64.3, 50.7, 62.5,
            58.4, 48.7, 52.4, 51.0, 35.7, 62.6, 52.3,
            41.2, 61.1, 52.9, 58.8, 64.1, 48.9, 74.3,
            50.3, 55.7, 55.5, 51.8, 55.8, 48.9, 51.8,
            63.1, 44.6, 47.0, 49.0, 62.5, 45.0, 78.6,
            54.2, 72.2, 52.4, 60.5, 46.8)
```

```
n <- length(weight)
n

## [1] 40

xbar <- mean(weight)
xbar

## [1] 55.995

s2 <- var(weight)
s2

## [1] 83.14818

se <- sqrt(s2 / n)
se

## [1] 1.441771
```

(ii)
```
t.test(weight,
       mu = 58,
       alternative = "less")

##
## 	One Sample t-test
##
## data:  weight
## t = -1.3907, df = 39, p-value = 0.08611
```

```
## alternative hypothesis: true mean is less than 58
## 95 percent confidence interval:
##      -Inf 58.4242
## sample estimates:
## mean of x
##    55.995
```

(iii) The point estimate of the mean is:

```
xbar

## [1] 55.995
```

A 95% confidence interval for the mean is

```
alpha <- 0.05
tval95 <- qt(1 - alpha / 2, df = n - 1)
tval95

## [1] 2.022691

c(xbar - tval95 * se,
xbar + tval95 * se)

## [1] 53.07874 58.91126
```

A 99% confidence interval for the mean is:

```
alpha <- 0.01
tval99 <- qt(1 - alpha / 2, df = n - 1)
tval99

## [1] 2.707913

c(xbar - tval99 * se,
xbar + tval99 * se)

## [1] 52.09081 59.89919
```

(iv) A point estimate of the variance is

```
s2

## [1] 83.14818
```

A 95% confidence interval for the variance is

```r
alpha <- 0.05
chisq25 <- qchisq(alpha / 2, df = n - 1, lower.tail = FALSE)
chisq975 <- qchisq(1 - alpha / 2, df = n - 1, lower.tail = FALSE)
```

```r
c((n - 1) * s2 / chisq25,
  (n - 1) * s2 / chisq975)

## [1]  55.79449 137.09032
```

2. (i) We assume that the first sample comes from a $N(\mu_1, \sigma^2)$ distribution and that the second sample comes from a $N(\mu_2, \sigma^2)$ distribution, with common variance σ^2, where μ_1, μ_2 and σ^2 are unknown. We test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the two-sided alternative $H_1 : \mu_1 \neq \mu_2$,where μ_1, μ_2 and σ^2 are unknown.

(ii) We carry out a two-sample t-test. First calculate the basic sample statistics:

	Variety 1	Variety 2
Sample size	10	10
Sample mean	156.8	151.5
Sample variance	166.4	169.4

The pooled estimate s^2 of the population variance σ^2 is given by

$$s^2 = \frac{9\,(166.4) + 9\,(169.4)}{18} = 167.9 \ .$$

The t-statistic for testing H_0 is given by

$$t = \frac{156.8 - 151.5}{\sqrt{167.9\left(\frac{1}{10} + \frac{1}{10}\right)}} = 0.915$$

with 18 degrees of freedom.

Because we are using a two-sided alternative hypothesis, we use a two-tail test. If F denotes the distribution function of the t_{18} distribution then in this case, using Table 9 of *Lindley and Scott*, with some interpolation, the p-value is given by

$$p = \Pr(|t| \geq 0.915) = 2(1 - F(0.915)) \approx 2(1 - 0.8137) = 0.373.$$

We do not reject the null hypothesis at the 5% significance level. There is no strong evidence that the two varieties have different mean wingspans.

Alternatively we could compare our calculated t-value with the percentage points of the t-distribution with 18 degrees of freedom. From Table 10, $t_{18}(2.5) = 2.101$. Since our calculated value, $t = 0.915$, is smaller than this percentage point, we do not reject the null hypothesis at the 5% significance level.

(iii) An estimate of the difference of the mean wingspans between the two varieties, $\mu_1 - \mu_2$, is given by 156.8 - 151.5 = 5.3.

A 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\left(5.3 - t_{18}(2.5)s\sqrt{\frac{1}{10} + \frac{1}{10}}\ ,\ 5.3 + t_{18}(2.5)s\sqrt{\frac{1}{10} + \frac{1}{10}}\right)$$

$$= \left(5.3 - 2.101\sqrt{167.9/5}\ ,\ 5.3 + 2.101\sqrt{167.9/5}\ \right) \ = \ (-6.87, 17.47) \ .$$

(iv) For the modified data, for Variety 1, the sample mean is modified to 159.8 and the sample variance to 64.4.

$$s^2 = \frac{9\,(64.4) + 9\,(169.4)}{18} = 116.9 \ .$$

The t-statistic for testing H_0 is given by

$$t = \frac{159.8 - 151.5}{\sqrt{116.9\left(\frac{1}{10} + \frac{1}{10}\right)}} = 1.717$$

with 18 degrees of freedom. From Table 9,

$$p = \Pr(|t| \geq 1.717) = 2(1 - F(1.717)) \approx 2(1 - 0.9483) = 0.103.$$

Although the t-value is now larger and the p-value smaller, we still do not reject the null hypothesis at the 5% significance level. There is no strong evidence that the two varieties have different mean wingspans.
An estimate of $\mu_1 - \mu_2$, is given by 159.8 - 151.5 = 8.3.
A 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned} &\left(8.3 - t_{18}(2.5)s\sqrt{\frac{1}{10} + \frac{1}{10}}\ ,\ 8.3 + t_{18}(2.5)s\sqrt{\frac{1}{10} + \frac{1}{10}}\right) \\ = &\left(8.3 - 2.101\sqrt{116.9/5}\ ,\ 8.3 + 2.101\sqrt{116.9/5}\right) = (-1.86, 18.46)\ . \end{aligned}$$

Using `R`,

(i)
```
Variety1 <- c(162, 159, 154, 176, 165, 164, 145, 157, 128, 158)
Variety2 <- c(147, 180, 153, 135, 157, 153, 141, 138, 161, 150)
```

```
n <- length(Variety1)
n

## [1] 10

xbar <- mean(Variety1)
xbar

## [1] 156.8

s2_1 <- var(Variety1)
s2_1

## [1] 166.4
```

```
m <- length(Variety2)
m

## [1] 10

ybar <- mean(Variety2)
ybar
```

```
## [1] 151.5
```

```
s2_2 <- var(Variety2)
s2_2
```

```
## [1] 169.3889
```

(ii)

```
t.test(Variety1, Variety2,
  var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Variety1 and Variety2
## t = 0.91462, df = 18, p-value = 0.3725
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.874275 17.474275
## sample estimates:
## mean of x mean of y
##     156.8     151.5
```

(iii) Point estimate of the difference between the two means:

```
xbar - ybar
```

```
## [1] 5.3
```

a 95% confidence interval is in the output of (ii)

(iv)

```
Variety1 <- c(162, 159, 154, 176, 165, 164, 145, 157, 158, 158)
```

```
xbar <- mean(Variety1)
xbar
```

```
## [1] 159.8
```

```
s2_1 <- var(Variety1)
s2_1
```

```
## [1] 64.4
```

```
t.test(Variety1, Variety2,
  var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
```

```
## data:  Variety1 and Variety2
## t = 1.7166, df = 18, p-value = 0.1032
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.85832 18.45832
## sample estimates:
## mean of x mean of y
##     159.8     151.5
```

3. (i) We have a "paired comparison" situation. Let x_i be the time at sea level and y_i the time at high altitude for Runner i. Let $d_i = x_i - y_i$ $(1 \le i \le 8)$. We assume that $d_1, d_2, \ldots, d_8$ is a random sample from a $N(\mu_D, \sigma_D^2)$ distribution, where μ_D and σ_D^2 are unknown. We test the null hypothesis $H_0 : \mu_D = 0$ against the one-sided alternative $H_1 : \mu_D < 0$.

(ii) The time differences are

-2.1 0.3 -1.6 -2.0 1.1 -3.4 0.1 -1.8

Their sample mean is -1.175 and their sample variance is 2.291. Hence, the t-statistic for testing H_0 is

$$t = -\frac{\sqrt{8} \times 1.175}{\sqrt{2.291}} = -2.196$$

with 7 degrees of freedom. Because we are using a one-sided alternative hypothesis, we use a one-tail test. From Table 9 of *Lindley and Scott*,

$$p = \Pr(t \le -2.196) = 1 - F(2.196) = 1 - 0.968 = 0.032 \quad \text{(to 3 decimal places).}$$

It follows that we reject H_0 at the 5% level although not at the 1% level. There is strong evidence that, on average, performance times are greater at high altitude than at sea level, i.e., athletes perform better at sea level. Alternatively, from Table 10, $t_7(5) = 1.895$ and $t_7(1) = 2.998$, which yields the same conclusion.

(iii) The estimated mean difference between performance times at the sea level location and the high altitude location is -1.175.
A 95% confidence interval for the underlying mean difference is given by

$$\left(-1.175 - t_7(2.5)\sqrt{\frac{2.291}{8}} \;,\; -1.175 + t_7(2.5)\sqrt{\frac{2.291}{8}} \right)$$
$$= \; (-1.175 - (2.365)(0.535) \;,\; -1.175 + (2.365)(0.535)\,) = (-2.44, 0.09) \;.$$

Using `R`:

(i)
```
SeaLevel <- c(48.3, 47.6, 49.2, 50.3, 48.8, 51.1, 49.0, 48.1)
HighAlt <- c(50.4, 47.3, 50.8, 52.3, 47.7, 54.5, 48.9, 49.9)
```

```
cbind(SeaLevel, HighAlt)

##      SeaLevel HighAlt
## [1,]     48.3    50.4
## [2,]     47.6    47.3
## [3,]     49.2    50.8
## [4,]     50.3    52.3
## [5,]     48.8    47.7
## [6,]     51.1    54.5
## [7,]     49.0    48.9
## [8,]     48.1    49.9
```

(ii)
```r
t.test(SeaLevel, HighAlt,
  alternative = "less",
  paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  SeaLevel and HighAlt
## t = -2.1958, df = 7, p-value = 0.03206
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1611981
## sample estimates:
## mean of the differences
##                  -1.175
```

(iii)
```r
diff <- SeaLevel - HighAlt

dbar <- mean(diff)
dbar
```

```
## [1] -1.175
```

```r
sD <- sd(diff)
n <- length(diff)

se <- sD / sqrt(n)
alpha <- 0.05

tval95 <- qt(1 - alpha / 2, df = n - 1)

c(dbar - tval95 * se,
  dbar + tval95 * se)
```

```
## [1] -2.44032651  0.09032651
```