MAT 2379, Introduction to biostatistics

Solution to Assignment 5

Due date: Tuesday December 2, 2014

$Total = 100$ marks

Problem 11.1 Let X be the depth of the ice sheet and $\mu = E(X)$. We want to test $H_0 : \mu = 3140$ against $H_1 : \mu < 3140$. We have $\bar{x} = 3126.1$ and $s = 8.63$. The observed value of the test statistic is:

$$\frac{3126.1 - 3140}{8.63/\sqrt{10}} = -5.09$$

The p-value is $P(T < -5.09) = P(T > 5.09) < 0.005$ where T has a $T(9)$ distribution. Since the p-value is smaller than $\alpha = 0.05$, we reject H_0. There is evidence that $\mu < 3140$.

Problem 12.8 (a) The pooled standard deviation is

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}} = \sqrt{\frac{(15-1)(1.2)^2 + (15-1)(1.75)^2}{(15+15-2)}} = 1.500.$$

(b) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, the observed value of the test statistic is

$$\frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{118 - 122}{1.500\sqrt{1/15 + 1/15}} = -7.30.$$

The p-value is $2\,P(T > |-7.30|) = 2\,P(T > 7.30)$, where T has a $T(15+15-2) = T(28)$ distribution. Since $P(T > 7.30) < 0.005$, then the p-value is smaller than 0.01. We can reject H_0 and conclude that the means are different.
(c) A 95% confidence interval for $\mu_1 - \mu_2$ is

$$\overline{x}_1 - \overline{x}_2 \pm 2.048\, s_p\sqrt{1/n_1 + 1/n_2} = 118 - 122 \pm (2.048)(1.500)\sqrt{1/15 + 1/15} = [-5.122, -2.878].$$

Since the interval contains only negative values, we can conclude that $\mu_1 < \mu_2$.

Problem 13.4 These are paired observations. We define the difference D between the two measurements:

$$D = \text{measure with method 1} - \text{measure with method 2}.$$

We want to test $H_0 : \mu_d = 0$ against $H_1 : \mu_d \neq 0$. The summary statistics of the differences are

$$\overline{d} = -0.0610, s_d = 0.0415, n = 10.$$

The observed value of the test statistic is

$$\frac{\overline{d} - 0}{s_d/\sqrt{n}} = \frac{-0.0610 - 0}{0.0415/\sqrt{10}} = -4.648.$$

The p-value is $2P(T > |-4.648|) = 2\,P(T > 4.648)$, where T has a $T(9)$ distribution. Since $P(T > 4.648) < 0.005$, then the p-value is less than 0.01. At a level of significance of $\alpha = 0.05$, we have sufficient evident to conclude that the method give different measurements on average.

Problem 16.13 (a) A point estimator for p is $\hat{p} = 148/495 = 0.299$. The interval is:

$$0.299 \pm 1.96\sqrt{\frac{(0.299)(0.701)}{495}} = 0.299 \pm 0.04 = [0.259; 0.339]$$

(b) We test $H_0 : p = 0.25$ against $H_1 : p > 0.25$. The observed value of the test statistic is:

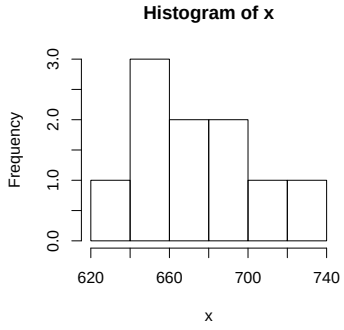
$$\frac{0.299 - 0.25}{\sqrt{(0.25)(0.75)/495}} = 2.52$$

The p-value is $P(Z > 2.52) = 1 - 0.9941 = 0.0059$. Since the p-value is smaller than 0.01, we reject H_0 in favor of H_1. We conclude that there is evidence that p is larger than 0.25.

Problem 11.5 (a) We create a variable x in R which contains the values in the column "Yield" of the file barley.txt. To verify that the yield is normally distributed we produce the histogram of x. For this, we type in the R console:

```
hist(x)
```

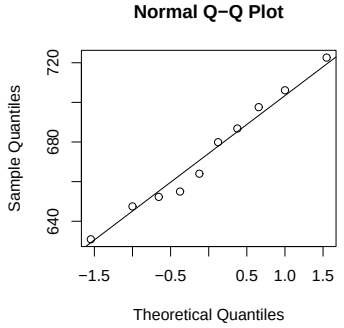
Below is the histogram:



To produce the QQ-plot together with the best-fitted line, we type in the R console:

```
> qqnorm(x)
> abline(mean(x),sd(x))
```

Below is the QQplot:



Since the plot appears to have a linear tendency and the histogram is approximately bell-shaped, we infer the yields are normally distributed.

To gain evidence for the fact that the mean yield has increased due to the kiln-drying procedure, we test the hypothesis $H_0 : \mu = 672$ against $H_1 : \mu > 672$, using R. We type in the R console:

```
t.test(x,mu=672,alternative="greater")
```

Below is the R output:

```
        One Sample t-test

data:  x
t = 0.2464, df = 9, p-value = 0.4055
alternative hypothesis: true mean is greater than 672
95 percent confidence interval:
 657.3817      Inf
sample estimates:
mean of x
   674.27
```

Since the p-value is larger than $\alpha = 0.05$, we fail to reject H_0. There is not enough evidence that the mean yield has increased due to the kiln-drying procedure.

(b) To produce a 95% confidence interval for the mean yield μ of kiln-dried barley, we type in the R console:

```
t.test(x)$conf.int
```

Below is the R output:

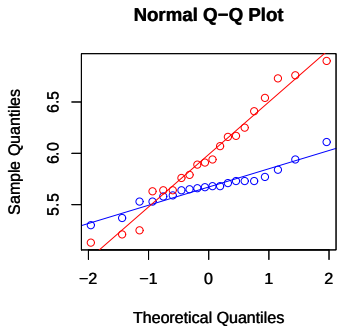
```
[1] 653.4289 695.1111
attr(,"conf.level")
[1] 0.95
```

Problem 12.5 (a,b) We create two variables $x1$ and $x2$ in R which contain the values in the column "Ph (lot 1)" and "Ph (lot 2)" of the file phcomparison.txt.

To produce the overlayed QQ-plots together with the lines of best fit, we type in the R console:

```
lmts=range(x1,x2)
qqnorm(x1,ylim=lmts,col="blue")        # produces the QQ-plot for x1 in blue
abline(mean(x1),sd(x1),col="blue")     # produces the line of best-fit for x1
par(new=T)                             # begin overlay
qqnorm(x2,ylim=lmts,col="red")         # produces the QQ-plot for x2 in red
abline(mean(x2),sd(x2),col="red")      # produces the line of best fit for x2
par(new=F)                             # end overlay
```

Below are the QQ-plots for the two data sets and the lines:



Since there is a linear tendency in both QQ plots, it is reasonable to assume that both populations are normal. But the slopes in the QQ plots are different; therefore, the assumption of equal variances is not verified.

(c) Since we have independent normal populations with unequal variances, we will use Welch's approximate two-sample t-test to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu 1 \neq \mu_2$. For this, we type in the R console:

```
t.test(x1,x2)
```

Below is the R output:

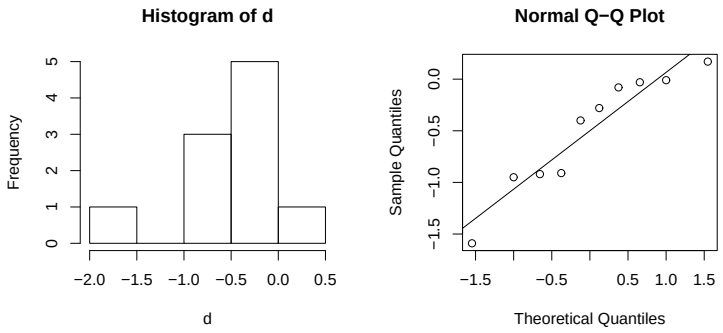
```
Welch Two Sample t-test

data:  x1 and x2
t = -2.6196, df = 23.526, p-value = 0.01516
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.56702362 -0.06697638
sample estimates:
mean of x mean of y
    5.672     5.989
```

Since the p-value is smaller than 0.05, we reject H_0. We have evidence that the mean pH levels are different.

Problem 13.5 We create in R two variables x and y which contain the values in the columns "Control" and "GH" of the file cow-hormone.txt: x=control, y=GH. We then create the variable $d = x - y$, for which we produce the histogram and the normal QQ plot (with the line of best fit). For this, we type in the R console:

```
d=x-y
hist(d)
qqnorm(d)
abline(mean(d),sd(d))
```

The histogram and QQ plot are given below:



The histogram does not show strong evidence against normality. Since the QQ plot appears to be linear, we conclude that the differences are normally distributed.

We would like to test the hypotheses $H_0 : \mu_D = 0$ against $H_1 : \mu_D < 0$, where $\mu_D = \mu_X - \mu_Y$. For this, we type in the R console:

```
t.test(x,y,paired=TRUE,alternative="less")
```

Below is the R output:

```
    Paired t-test

data:  x and y
t = -2.7952, df = 9, p-value = 0.01044
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1720942
sample estimates:
mean of the differences
                  -0.5
```

Since p-value is smaller than $\alpha = 0.025$, we reject H_0. Based on this test, we can say that there is enough evidence that the growth hormone increases the milk production.

To produce the 98% confidence interval for μ_D, we type in the R console:

```
t.test(x,y,conf.lev=0.98,paired=TRUE)
```

Below is the R output:

```
     Paired t-test
data:  x and y
t = -2.7952, df = 9, p-value = 0.02088
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval:
 -1.004696634  0.004696634
sample estimates:
mean of the differences
                  -0.5
```

Since 0 falls in the interval, we should not conclude that there is a difference in the milk production. At a level of confidence of 98%, there is no evidence that the growth hormone has increased the milk production.