MAT 2379, Introduction to biostatistics

Assignment 4 - solutions

Part I: Solve the following problems using a calculator permitted by the Faculty of Science (TI30, TI34, Casio fx-260 and Casio fx-300), and the table for the cumulative distribution of the standard normal.

Question 1: Problem 8.12 from the textbook.

(a) The estimated standard error of the mean is $s/\sqrt{n} = 0.275$. Therefore, the sample standard deviation is $s = 0.275\sqrt{25} = 1.375$ noanolitres per second.

(b) A 95% confidence interval for the population mean maximal nitric oxide diffusion rate is

$$\overline{x} \pm t\,\frac{s}{\sqrt{n}} = [3.18; 4.32],$$

where $\overline{x} = 3.75$, $s/\sqrt{n} = 0.275$ and $t = t_{0.025;24} = 2.064$.

(c) In the context of repeated sampling, a smaller proportion (i.e. 92%) of the intervals are going to contain the population mean, so the intervals should be shorter.

(d) In the context of repeated sampling, a larger proportion (i.e. 97%) of the intervals are going to contain the population mean, so the intervals should be longer.

[8] **Question 2:** Problem 9.4 (a to c) and (e, f) from the textbook.

(a) Let μ be the mean carbon monoxide concentration. We want to test $H_0 : \mu = 100$ against $H_1 : \mu > 100$.

(b) We committed an error of type I, if we conclude that the mean carbon monoxide concentration is larger than 100, but in reality it is not the case.

(c) We committed an error of type II, if we fail to conclude that the mean carbon monoxide concentration is larger than 100, but in reality it is.

(e) The observed value of the test statistic is

$$t_0 = \frac{\overline{x} - 100}{s/\sqrt{n}} = \frac{101.012 - 100}{5.7644/\sqrt{25}} = 0.878.$$

The p-value is $P(T > 0.878)$, where T has a $T(24)$ distribution. Therefore, the p-value is between 10% and 25%. At a level of significance of 10%, the evidence against the null hypothesis is not significant. We cannot conclude that the mean monoxide concentration is larger than 100 ppm.

(f) We failed to reject the null hypothesis, this means that we committed an error of type II, if we did indeed commit an error.

[7] **Question 3:** Problem 9.6.

Let p be the true success rate of PN in treating kidney stones.
(a) A point estimate for p is $\hat{p} = 289/350 = 0.8257$ and its estimated standard error is $s\{\hat{p}\} = \sqrt{\hat{p}(1-\hat{p})/n} = 0.02028$.

(b) We want to $H_0 : p = 0.78$ against $H_1 : p \neq 0.78$. The observed value of the test statistic is

$$z_0 = \frac{\hat{p} - 0.78}{\sqrt{.78(1-.78)/350}} = 2.06.$$

The p-value is $2\,P(Z > 2.06) = 2\,(1 - 0.9803) = 0.0394$. At a level of significance of 1%, the evidence that the success rate of PN in treating kidney stones is different than the success rate of open surgery is not significant.

(c) A 95% confidence interval for p is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = [0.786; 0.865].$$

We are 95% confident that the success rate of PN in treating kidney stones is between 78.6% and 86.5%. We are 95% confident that PN is more successful for treating kidney stones than open surgery.

[4] **Question 4:** Problem 14.8.

There are paired measurements. Let D be the yield for the acre with traditional fertilizer minus the yield for the acre with organic fertilizer. We want to test $H_0 : \mu_D = 0$ against $H_1 : \mu_D > 0$.

From the $n = 10$ differences, we computed $\bar{d} = 0.059$ and $s_d = 0.0722$. The observed value of the test statistic is:

$$t_0 = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{0.059}{0.0722/\sqrt{10}} = 2.58.$$

The p-value is $P(T > 2.58)$, where $T \sim T(9)$. The p-value is between 0.01 and 0.025. Since the p-value is larger that $\alpha = 0.01$, we cannot reject H_0. There is not enough evidence that the traditional fertilizer produces a larger yield, at level $\alpha = 0.01$.

[5] **Question 5:** Problem 10.4

We denote by μ_1 the mean bunching intensity for families in the culled population and μ_2 the mean bunching intensity for families in a non-culled population. We want to test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 < 0$. The pooled standard deviation is

$$s_p = \sqrt{\frac{18(0.81) + 11(0.49)}{19 + 12 - 2}} = 0.83.$$

The observed value of the test statistic is

$$t_0 = \frac{1.2 - 2.5}{s_p\sqrt{1/19 + 1/12}} = -4.25$$

The p-value is $P(T < -4.25) = P(T > 4.25)$, where $T \sim T(29)$. We have p-value$= P(T_{29} > 4.25) < 0.005$. We reject H_0 at level $\alpha = 0.005$. We conclude that families from culled populations have a lower bunching intensity than families from non-culled populations.

[4] **Question 6:** Problem 10.14

Let μ_i be the mean density of organism (in number of organisms per square meter) from location $i = 1, 2$. We want to test

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

The observed value of the t-test statistic is

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{9,168.75 - 2,168.33}{\sqrt{3,700.57^2/12 + 815.26^2/12}} = 6.40.$$

Since it is a two-sided alternative, then the p-value is $2P(T > 6.40)$, where T has an approximate $t(\nu)$ distribution

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 2)} = 12.06 \approx 12.$$

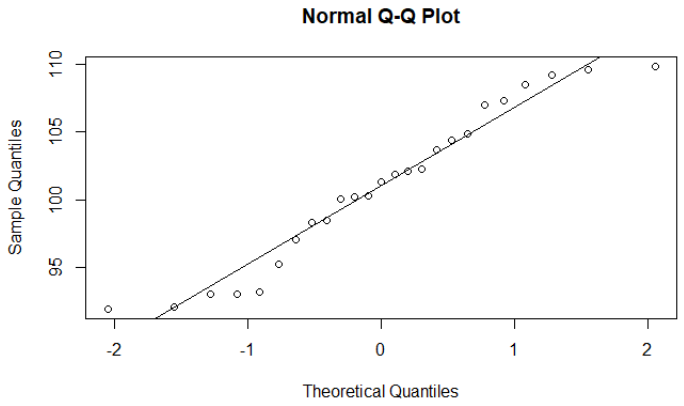
Since, $P(T(12) > 6.4) < 0.005$. Thus, p-value$= 2\,P(T(12) > 6.4) < 2\,(0.005) = 0.01$. At $\alpha = 5\%$, we have significant evidence that the mean density of organisms at the two locations are different.

Part (II) Use `R` for your computations to the following problems. Please attach the `R` commands, output and graphs that you used to answer the question. The `R` output **alone** is not an answer to the question. Please write a sentence or two to properly answer each question.

Question 7:

(a) With `R`, we assigned the values to the numeric vector `x` and used to following commands to produced the QQ-plot found below.

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



There is a linear tendency in the QQ-plot with slight deviations in the tails. It is reasonable to assume that the carbon monoxide concentration is normally distributed.

(b) The p-value as computed with `R` is 0.1944.

```
> t.test(x,mu=100,alternative="greater")

        One Sample t-test

data:  x
t = 0.8778, df = 24, p-value = 0.1944
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 99.03955      Inf
sample estimates:
mean of x
  101.012
```

[8] **Question 8:** In a study conducted at Virginia Tech on the development of ectomycorrhizal, a symbiotic relationship between between the roots of trees and a fungus, in which minerals are transferred from the fungus to the trees and sugars from the trees to the fungus, 20 northern red oak seedlings exposed to the fungus *Pisolithus tinctorus* were grown in a greenhouse. All seedlings were planted in the same type of soil and received the same amount of sunshine and water. Half received no nitrogen at planting time, to served as a control, and the other half received 368 ppm of nitrogen in the form $NaNO_3$. The stem weights, in grams, at the end of 140 days were recorded and are found in the file *nitrogen.txt*.

(a) Produce overlayed quantile-quantile plots (or normal probability plots) for the two groups of stem weights and produce comparative boxplots for the two groups of stem

weights. Is it reasonable to assume that the stem weight is normally distributed? Is it reasonable to assume that the variance of the stem weight is the same for both groups?

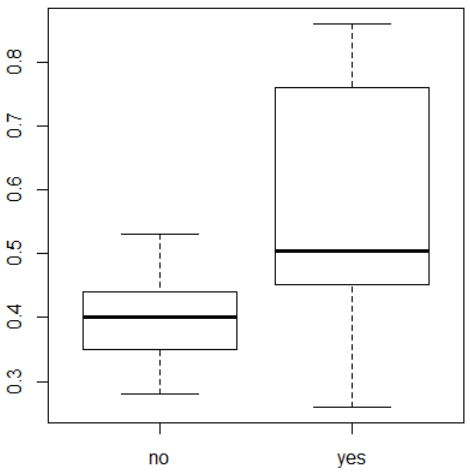
(b) Does the use of nitrogen have a significant effect on the stem weights on average? (Use $\alpha = 5\%$.)

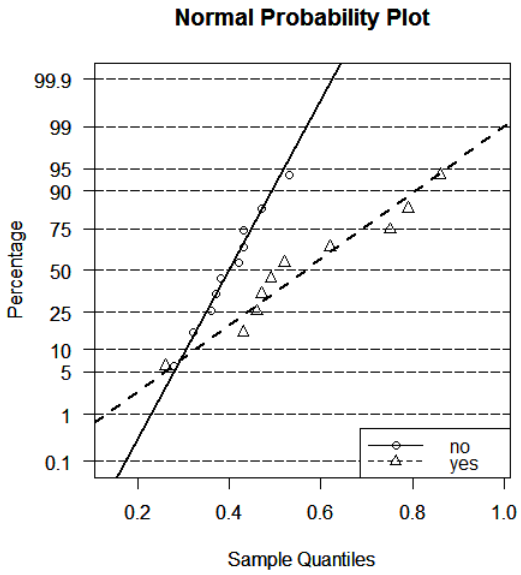
(c) Give a 95% confidence interval for the difference between the mean weight of the stems with nitrogen and the mean weight of the stems without nitrogen.

With the following commands we imported the data from the file `nitrogen.txt` and we displayed the names of the columns.

```
> nitrogen=read.table(file.choose(),header=TRUE,sep="\t")
> names(nitrogen)
[1] "Stem.Weight" "Nitrogen"
```

(a) You will find below the comparative boxplots for the two groups of stem weights and also the overlayed normal plots. There are linear tendencies in the normal probability plots. So it is reasonable to assume that the populations are normally distributed. However, we observe in the comparative boxplots that the stem weights for the units that received nitrogen are much more dispersed compared to those that did not received nitrogen. Furthermore, the slopes in the normal probability plots are very different. It is not reasonable to assume that the population variances are equal.





With R:

```
> ## source plots.R
> source(file.choose())
```

```
> ## a 1 by 2 graphics window
> par(mfrow=c(1,2))
> BoxPlot(Stem.Weight~Nitrogen,data=nitrogen)
> ppnorm(Stem.Weight~Nitrogen,data=nitrogen)
```

(b) We use the `t.test` function with `R` to test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$, where μ_1 is the mean weight stem weight without nitrogen and μ_2 is the mean weight stem weight with nitrogen.

```
> t.test(Stem.Weight~Nitrogen,data=nitrogen)

        Welch Two Sample t-test

data:  Stem.Weight by Nitrogen
t = -2.6191, df = 11.673, p-value = 0.02286
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.30452438 -0.02747562
sample estimates:
 mean in group no mean in group yes
            0.399             0.565
```

Since the p-value is 0.02286 (which is smaller than $\alpha = 5\%$), then we have significant evidence against H_0 in favour of H_1. We have significant evidence that nitrogen has an effect on the stem weight.

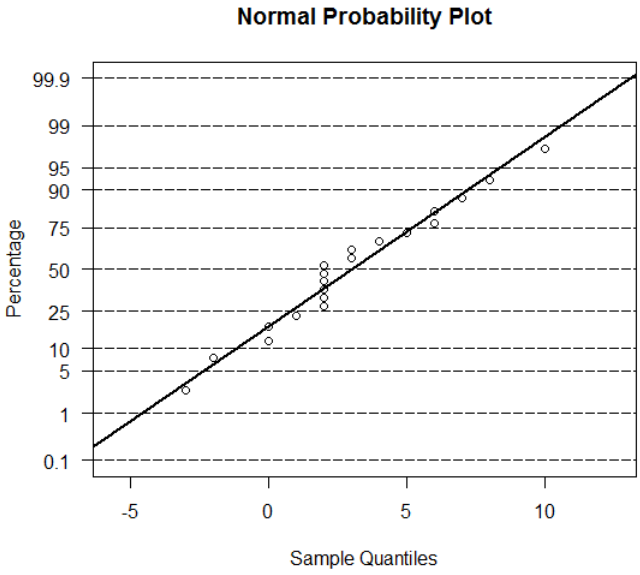
(c) A 95% confidence interval for the difference between the mean weight of the stems with nitrogen and the mean weight of the stems without nitrogen is $[0.027, 0.304]$. Refer to the `R` output in (b) from the `t.test` function.

Question 9: We start by importing the data from the file `TOBACCO.txt` and display the names of the columns.

```
> tobacco=read.table(file.choose(),header=TRUE,sep="\t")
> names(tobacco)
[1] "preparation.1" "preparation.2"
```

(a) These are paired measurements. We will compute the difference and construct a normal probability plot of the differences.

```
> d=tobacco$preparation.1-tobacco$preparation.2
> ## source plots.r
> ppnorm(d)
```



There is a linear tendency in the normal probability plot. It is reasonable to assume that the difference of the paired measurements is normally distributed.

Remark: You can have provided a qq-plot instead of the normal probability. But, the conclusion should be the same.

(b) The p-value for the paired t-test is 0.0005896, which is much smaller than the level of significance of $\alpha = 1\%$. We have significant evidence that the preparation have different effects on the tobacco plants.

```
> t.test(tobacco$preparation.1,tobacco$preparation.2,paired=TRUE)

        Paired t-test

data:  tobacco$preparation.1 and tobacco$preparation.2
t = 4.1147, df = 19, p-value = 0.0005896
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.473987 4.526013
sample estimates:
mean of the differences
                      3
```

(c) A 95% confidence interval for μ_D is $1.47, 4.53$ (see part (b)). We are 95% confident that on average there will be between 1.5 to 4.5 more lesions, if we use preparation 1 compared to preparation 2.

Marking Scheme:

Question 2:

- (a) 1/2pt for H_0; 1/2pt for H_1
- 1pt for (b); 1pt for (c)
- (e) 1pt for value of t_0; 1pt to write the p-value as a right-sided prob; 1pt for $0.1 < p\text{-value} < 0.25$; 1pt for conclusion to not reject H_0.
- (f) 1pt

Question 3:

- (a) 1/2pt for point estimate; 1/2pt for standard error
- (b) 1/2pt H_0; 1/2pt for H_1
- (b) 1pt for value of z_0; 1pt to write the p-value as a two-sided prob; 1pt value of the p-value; 1pt for conclusion to not reject H_0.
- (c) 1pt to compute the correct confidence interval

Question 4:

- 1/2pt for H_0; 1/2pt for H_1
- 1/2pt for $\bar{d}$, 1/2pt for s_d
- 1pt for t_0 and 1pt for correct conclusion to not reject H_0.

Question 5:

- 1/2pt for H_0; 1/2pt for H_1
- 1 pt for pooled standard deviation s_p
- 1 pt for computing 29 degrees of freedom
- 1pt for t_0 and 1pt for correct conclusion to reject H_0.

Question 6:

- 1/2pt for H_0; 1/2pt for H_1
- 1 pt for computing 12 degrees of freedom
- 1pt for t_0 and 1pt for correct conclusion to reject H_0.

Question 8:

- (a) 1 pt for providing the plots; 1 pt for concluding that it is reasonable to assume that the populations are normal; 1 pt for concluding that it is reasonable to assume that the variances are unequal.
- 1 pt stating H_0 and H_1; 1 pt for using `t.test` correctly; 1pt for identifying the p-value; 1pt to concluding that we reject H_0.
- 1 pt for giving the confidence interval

[**/36**]