MAT 2779, Introduction à la biostatistique

Devoir 4
SOLUTIONNAIRE

Professeur: M'hammed Mountassir

SVP CORRIGER JUSTE LES 3 PREMIÈRES QUESTIONS POUR UN TOTAL DE 15 POINTS. MERCI

Question 1:9.6 (5 points)
Soit p le taux de succès de PN.
(a) Un estimateur ponctuel de p est $\hat{p} = 289/350 = 0.8257$ et l'estimation de son erreur standard est $s\{\hat{p}\} = \sqrt{\hat{p}(1-\hat{p})/n} = 0.02028$.

(b) On veut tester $H_0 : p = 0.78$ contre $H_1 : p \neq 0.78$. La valeurobservée de la statistique est

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

La $valeur - P$ est $2\,P(Z > 2.06) = 2\,(1 - 0.9803) = 0.0394$. À un seuil de significationde 1%, on ne rejette pas l'hypothèse nulle et on en conclut que le taux de succès de PN n'est pas différent de 0,78 . i

(c) Un intervalle de confiance de niveau 95% pour r p est

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

On est confiant à 95% que le taux de succès de PN est compris entre 78.6% et 86.5%.
REMARQUE: Si on utilise le logiciel R, on va trouver un intervalle de confiance légéremment différent.

Question 2:10.14 (5 points)

Soit μ_i la densité moyenne de l'organisme (en nombre d'organismes par mètre carré) pour la position $i = 1, 2$. On veut tester

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

La valeur observée de t est:

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

Puisqu'il s'agit d'un test bilatéral , alors la $Valeur - P$ est $2P(T > 6.40)$, avec T admet une distribution $t(\nu)$

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

De la Table 17.4, $P(T(12) > 6.4) < 0.005$. Alors, que $Valeur - P = 2\,P(T(12) > 6.4) < 2\,(0.005) = 0.01$. Avec un $\alpha = 5\%$, on rejette l'hypothèse nulle et on conclut que les niveaux moyens des densités aux deux endroits sont significativement différents.

Question 3: 14.8 (5 points)

Ce sont des données appariées. Soit D le rendement de l'acre avec des engrais traditionnels moins le rendement de l'acre avec les engrais organiques. On veut tester $H_0 : \mu_D = 0$ contre $H_1 : \mu_D > 0$.

Pour les $n = 10$ différences, on obtient $\bar{d} = 0.059$ et $s_d = 0.0722$. Donc:

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

La valeur-p est $P(T > 2.58)$, où $T \sim T(9)$. Alors la valeur-p est comprise entre 0.01 et 0.025. Puisque la valeur-p dépasse $\alpha = 0.01$, on ne rejette pas H_0. On ne peut pas dire que les engrais traditionnels donnent des rendements supérieurs à un niveau de signification de $\alpha = 0.01$.

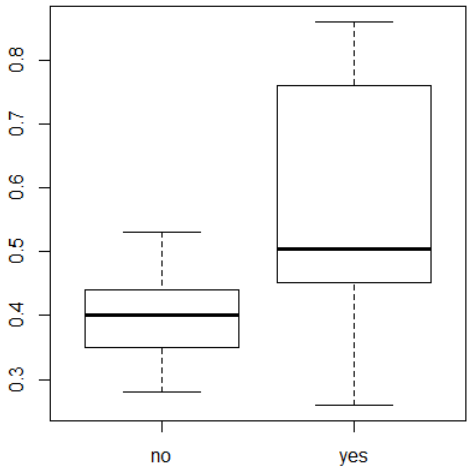
Part (II)

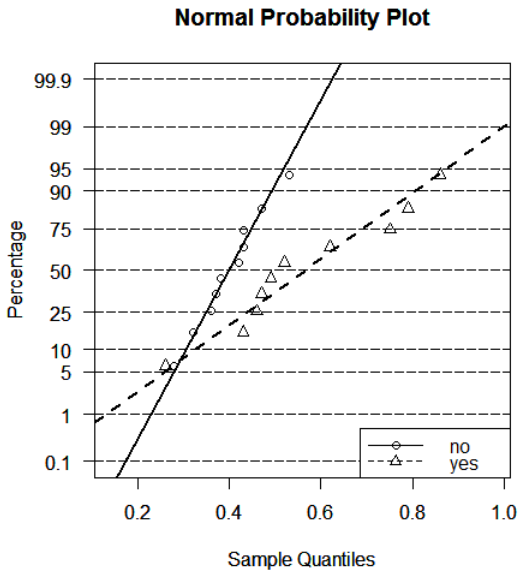
Question 4:

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.

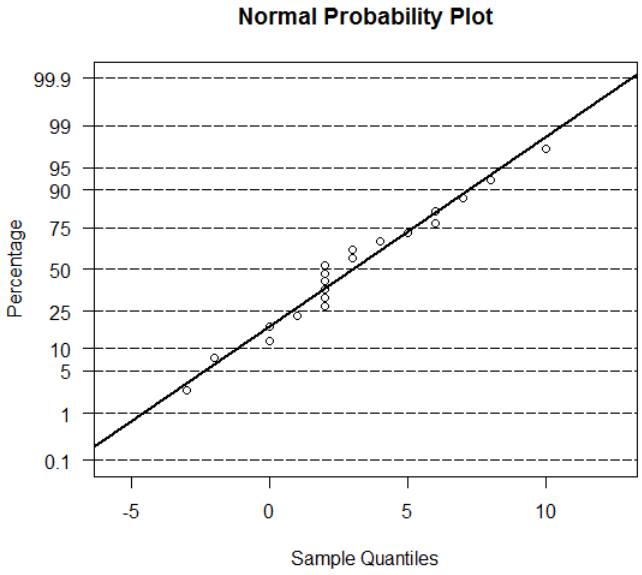
(c) A 95% confidence interval for the difference between the mean weight of the stems with nitrogen and the mean weight of the stems without nitrogen is $[0.027, 0.304]$. Refer to the R output in (b) from the `t.test` function.

Question 5: 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.

Question 6: (a) The correlation is $r = s_{xy}/(s_x\, s_y) = -1.8003/[(5.6851)(0.4036)] = -0.78$. The association between the melanoma mortality rate and the latitude is approximately linear, negative and moderately strong with a correlation of -0.78.

(b) The estimated slope is $\widehat{\beta} = s_{xy}/s_x^2 = -1.8003/(5.6851)^2 = -0.0557$ and the estimated intercept is $\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 3.5996$. Thus, the least squares line is

$$\widehat{y} = 3.5996 - 0.0557\, x.$$

We interpret the slope as follows. For every increase of one unit in latitude, there is a reduction of 0.0557 in the melanoma mortality rate.

(c) We imported the data in `R` and assigned the mortality rates for the female population to y and the latitude to x. We used the following commands to build the scatter plot and to display the estimation of the coefficient of the least squares line.

```
> plot(x,y,ylab="Mortality Rates",xlab="Latitude")
> lm(y~x)
```

Here is the output .

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     2.4157      -0.0345
```