# UNIVERSITY OF OSLO

## Faculty of mathematics and natural sciences

| | |
|---|---|
| Exam in: | STK1000 – Introduction to applied statistics |
| Day of examination: | Tuesday 12. December 2017 |
| Examination hours: | 14.30 – 18.30 |
| This problem set consists of 5 pages. | |
| Permitted aids: | Approved calculator, textbook (any edition and you may write in the book) and dictionary for STK1000. |

Please make sure that your copy of the problem set is complete. All 10 problems count equally in the evaluation of the exam. Good luck!

### Exercise 1

John Arbuthnot is considered to be the first in the word to calculate a p value. His null hypothesis that the numbers of male and female births were equal, was studied in the christening registries in London, 1629 to 1710. He observed that for 82 years in a row, more boys than girls were born (and christened). In this exercise, we will investigate whether Arbuthnot needed 82 years of observations, or if one year would be enough.

| Anno. | Christened. Males. | Females. | Anno. | Christened. Males. | Females. |
|---|---|---|---|---|---|
| 1629 | 5218 | 4683 | 1648 | 3363 | 3181 |
| 30 | 4858 | 4457 | 49 | 3079 | 2746 |
| 31 | 4422 | 4102 | 50 | 2890 | 2722 |
| 32 | 4994 | 4590 | 51 | 3231 | 2840 |
| 33 | 5158 | 4839 | 52 | 3220 | 2908 |
| 34 | 5035 | 4820 | 53 | 3196 | 2959 |
| 35 | 5106 | 4928 | 54 | 3441 | 3179 |
| 36 | 4917 | 4605 | 55 | 3655 | 3349 |
| 37 | 4703 | 4457 | 56 | 3668 | 3382 |
| 38 | 5359 | 4952 | 57 | 3396 | 3289 |
| 39 | 5366 | 4784 | 58 | 3157 | 3013 |
| 40 | 5518 | 5332 | 59 | 3209 | 2781 |
| 41 | 5470 | 5200 | 60 | 3724 | 3247 |
| 42 | 5460 | 4910 | 61 | 4748 | 4107 |
| 43 | 4793 | 4617 | 62 | 5216 | 4803 |
| 44 | 4107 | 3997 | 63 | 5411 | 4881 |
| 45 | 4047 | 3919 | 64 | 6041 | 5681 |
| 46 | 3768 | 3395 | 65 | 5114 | 4858 |
| 47 | 3796 | 3536 | 66 | 4678 | 4319 |

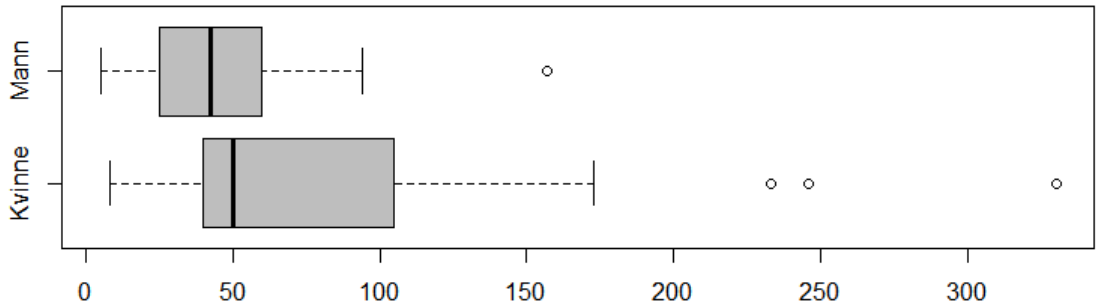
a) Calculate the probability of observing more male births than female births for 82 years in a row, if the probability of giving birth to a boy equals the probability of giving birth to a girl.
Let the stochastic variable X be the number of boys born in 1629. Which probability distribution does X have?
Which three assumptions must be fulfilled for X to have this distribution?

b) Based on X and the parameter p from the distribution in a), formulate the null hypothesis and the alternative hypothesis for the proportion of boys born in 1629.
Give a point estimate for p, and calculate a 95% confidence interval for p.
Give an interpretation of the confidence interval, and explain which conclusion you reach for the hypothesis test if you use a significance level of 0.05.

c) Assume that the true proportion of male births is 0.52. How many births (n) would you have to observe to reject a two-sided $H_0$ at significance level 0.05? (Hint: Use the lower linit in a 95% confidence interval like the one in exercise b).)

## Exercise 2

When 20 male (Mann) and 31 female (Kvinne) STK1000 students in 2016 were asked how many every-day clothes they owned, the distributions of the answers were as follows:




Descriptive statistics for the 20 males:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      sd
    5.0    27.5    42.5    49.0    59.0   157.0    34.3
```

Descriptive statistics for the 31 females:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      sd
    8.0    40.0    50.0    85.0   105.0   330.0    75.2
```

The R outputs below are results from a two-sample t test and a Wilcoxon Rank Sum test:

```
> t.test(hverdag~kjonn)

        Welch Two Sample t-test

data:  hverdag by kjonn
t = 2.3216, df = 45.082, p-value = 0.02483
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
  4.780768 67.383748
sample estimates:
mean in group Kvinne   mean in group Mann
            85.03226             48.95000

> wilcox.test(hverdag~kjonn)

        Wilcoxon rank sum test with continuity correction

data:  hverdag by kjonn
W = 399.5, p-value = 0.08544
alternative hypothesis: true location shift is not equal to 0
```

a) Which numbers from the descriptive statistics will summarize the groups best? Give reasons for your answer.

b) State hypotheses to test whether the number of clothes differ between the genders. Explain which test you would chose, and why.
Chose a significance level of 0.05. Which conclusion do you reach, and why?

## Exercise 3

Among non-pregnant adults, an association between høy body mass index (BMI), that is, (vekt i kg)/(høyde i m)$^2$, and high insulin resistance. Insulin resistence gives high blood sugar values, in particular after food intake. In a sample of 130 pregnant women, researchers wanted to study the association between the women's pre-pregnancy BMI, and their blood sugar level (measured in mmol/l) two hours after food intake, in the second half of the pregnancy. The following regression analysis was done:

```
> summary(lm(blodsukker ~ BMI ))

Call:
lm(formula = blodsukker  ~ BMI )

Residuals:
    Min      1Q  Median      3Q     Max
-2.8228 -0.7290 -0.2273  0.7749  3.5811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.77942    0.88482   3.141  0.00209 **
BMI          0.08225    0.03897   2.111  0.03676 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.169 on 128 degrees of freedom
Multiple R-squared:  0.03363,   Adjusted R-squared:  0.02608
F-statistic: 4.454 on 1 and 128 DF,  p-value: 0.03676
```

a) What is the effect measure in this analysis?
What is the interpretation of this effect measure?
Give an estimate of the association between pre-pregnancy BMI and blood sugar level two hours after food intake, in second half of pregnancy.
Calculate a 95% confidence interval for this association.

The researchers suspected that weight gain during pregnancy could be an important factor. They based this on research in non-pregnant people, which showed that even a moderate increase in BMI resulted in increased insulin resistance in persons who were followed up over time.
Health advice given to pregnant women concerning weight gain aims at reducing adverse outcomes related to overweight, both for the pregnant woman, and the child she is carrying. Recommended weight gain is therefore a consequence of the woman's pre-pregnancy BMI. Overweight women are recommended to gain less weight than normal weight women.

b) Can you assume that weight gain during pregnancy is a confounder (lurking variable) for the association between maternal pre-pregnancy BMI and blood sugar levels two hours after food intake, in second half of the pregnancy?
Give reasons for your answer.
Comment the following R output based on your answers.

```
> summary(lm(blodsukker  ~ BMI + vektokning))

Call:
lm(formula = blodsukker  ~ BMI  + vektokning)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8334 -0.6902 -0.2317  0.8012  3.5537

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.63692    0.97384   2.708  0.00771 **
BMI          0.08268    0.03912   2.113  0.03653 *
vektokning   0.01789    0.05023   0.356  0.72231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.173 on 127 degrees of freedom
Multiple R-squared:  0.03459,   Adjusted R-squared:  0.01939
F-statistic: 2.275 on 2 and 127 DF,  p-value: 0.1069
```

## Exercise 4

One of the data sets available in R provides measurements of the stem circumference, height and volume of timber in 31 felled black cherry trees. The stem circumference (in cm) is measured 137 cm over ground. In this exercise, the timber volume is converted to litres. One can assume that one litre corresponds to one piece of firewood (in Norwegian: "vedkubbe"). If you plot the timber volume $y_i$ (`vedkubbe` in the output), and the tree circumference $x_i$ (`omkrets` in the output) in a scatter plot, you can see that the association between the two variables can be expressed by the regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma).$$

The means and standard deviations are $\bar{x} = 105.7$, $sd_x = 25.0$, and $\bar{y} = 854.3$, $sd_y = 465.5$.

The output shows the result from the regression analysis of the 31 trees:

```
> summary(lm(vedkubber~omkrets))

Call:
lm(formula = vedkubber ~ omkrets)

Residuals:
     Min       1Q   Median       3Q      Max
-228.386  -87.972    4.303   98.961  271.468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1046.1223    95.2903  -10.98 7.62e-12 ***
omkrets        17.9769     0.8779   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 120.4 on 29 degrees of freedom
Multiple R-squared:  0.9353,    Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

a) How must the relation between x and y in the scatter plot look, to make this regression model a reasonable choice for the analysis?

   Give estimates and interpretations of the estimates for the parameters $\beta_0$ and $\beta_1$.

b) Calculate a 95% prediction interval for the timber volume for a tree with a circumference of 94.2 cm (diameter 30 cm).

c) The R output below shows a similar regression analysis for the tree circumference against the tree height (measured in meters). A corresponding 95% prediction interval for the height of a tree with a circumference of 94.2 cm (diameter 30 cm) is [19.2, 26.2]. Explain what a prediction interval shows.
   Give interpretations of the two prediction intervals in b) and c).
   For the 31 cherry trees: Which will be most precisely predicted by the tree circumference: The timber volume or the height? Give reasons for your answer.

```
> round(c(mean(trehoydeimeter), sd(trehoydeimeter)),1)
[1] 23.2  1.9
> summary(lm(trehoydeimeter~omkrets))

Call:
lm(formula = trehoydeimeter ~ omkrets)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8349 -0.8439  0.0964  0.7537  3.0314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.90714    1.33603  14.152 1.49e-14 ***
omkrets      0.04027    0.01231   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.688 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

**THE END**