## Assignment 1 (20%)

Let $X$ denote a continuous stochastic variable with the following cumulative probability function

$$F(x) = \begin{cases} 0 & for\ x \leq -1 \\ \frac{1}{2}(x+1) & for - 1 < x < 1 \\ 1 & for\ x \geq 1 \end{cases}$$

a) Compute $P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right)$ and $P(X > 0{,}75)$

$$P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right) = F\left(\frac{1}{2}\right) - F\left(-\frac{1}{2}\right) = 0{,}75 - 0{,}25$$

$= 0{,}5$

$$(X > 0{,}75) = 1 - P(X \leq 0{,}75) = 1 - \frac{1}{2}(0{,}75 + 1)$$

$= {}^{1}/_{8}$

b) Show that the density function $f(x)$ for $X$ is

$$f(x) = \begin{cases} \frac{1}{2} & for - 1 < x < 1 \\ 0 & Otherwise \end{cases}$$

Differentiate

c) Find the expected value and variance of $X$

$$E(X) = \int_{-1}^{1} \frac{1}{2} x\, dx = \frac{1}{4} x^2 \Big|_{-1}^{1} = \frac{1}{4} - \frac{1}{4} = 0$$

$$Var(X) = \int_{-1}^{1} \frac{1}{2} x^2\, dx = \frac{1}{6} x^3 \Big|_{-1}^{1} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

## Assignment 2 (20%)

A batch of 1000 hard drives from three suppliers were tested. 2% of the hard drives from Toshiba and 2% of the hard drives from Seagate were defective, and in the entire batch there were 3% defectives in total. In the batch, 50% were Western Digital hard drives and 30% were Toshibas.

a) Based on this information, create a 3 x 2 contingency table

| | Non-defective | Defective | Sum R |
|---|---|---|---|
| Toshiba | 294 | 6 | 300 |
| Seagate | 196 | 4 | 200 |

| | | | |
|---|---|---|---|
| WD | 480 | 20 | 500 |
| Sum C | 970 | 30 | 1000 |

b) What is the probability that a defective product came from Seagate?

$$\mathrm{P(Seagate|Defective)} = \frac{0{,}004}{0{,}03}$$

$= 0{,}13$

c) What is the probability of randomly selecting a Western Digital hard drive from the entire batch?

$$P(WD) = \frac{500}{100}$$

$= 0{,}5$

**Assignment 3 (10%)**

Different screens and their hue bias were tested and the result is displayed in the following table:

| | Blueish | Reddish | Greenish |
|---|---|---|---|
| Display 1 | 46 | 82 | 72 |
| Display 2 | 42 | 38 | 20 |
| Display 3 | 52 | 40 | 8 |

Is there sufficient evidence to conclude that screens and hue bias depend significantly? Design an appropriate test to answer this question.

H0: Screens and hue bias are independent

H1: Screens and hue bias are dependent

From template we obtain p-value = 0,0000. From this we reject the null hypothesis and conclude that screens and hue bias are dependent.

**Assignment 4 (20%)**

Two different machines, A and B, which are used to measure blood pressure, are tested on 12 different patients such that each patient has his/her blood pressure measured by both machines. The results for the systolic blood pressure are displayed in the table below:

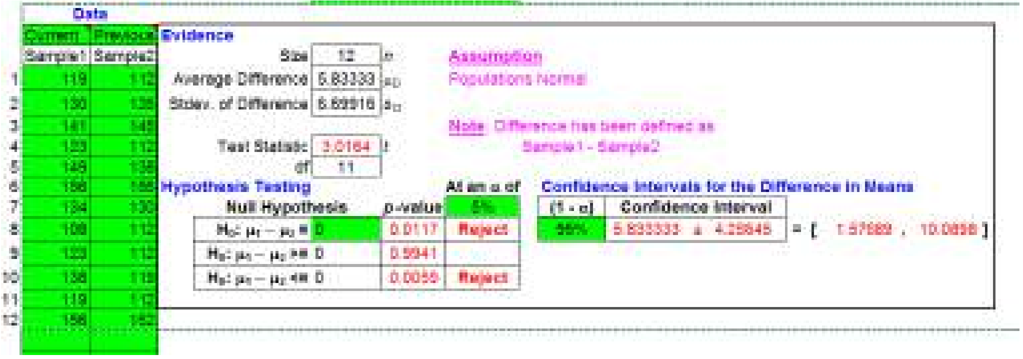
| Patient | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Machine A | 119 | 130 | 141 | 123 | 149 | 156 | 134 | 108 | 123 | 138 | 119 | 156 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Machine B | 112 | 126 | 145 | 112 | 138 | 156 | 130 | 112 | 112 | 119 | 112 | 152 |

a) Determine the mean, standard deviation and interquartile range for both sets of data

| | Machine A | Machine B |
|---|---|---|
| | 119 | 112 |
| | 130 | 126 |
| | 141 | 145 |
| | 123 | 112 |
| | 149 | 138 |
| | 156 | 156 |
| | 134 | 130 |
| | 108 | 112 |
| | 123 | 112 |
| | 138 | 119 |
| | 119 | 112 |
| | 156 | 152 |
| | | |
| Mean | 133 | 127,16667 |
| St. Dev. | 15,462565 | 16,813595 |
| IQR | 21 | 27,75 |

b) Is it possible to conclude with statistical significance that the two machines give different measurement? Design an appropriate test to answer this question.
We will test this by testing difference in means
H0: Mean machine A is equal to mean of machine B
H1: Mean machine A is not equal to mean of machine B



Data

| | Current | Previous |
|---|---|---|
| | Sample1 | Sample2 |
| 1 | 119 | 112 |
| 2 | 130 | 126 |
| 3 | 141 | 145 |
| 4 | 123 | 112 |
| 5 | 149 | 138 |
| 6 | 156 | 156 |
| 7 | 134 | 130 |
| 8 | 108 | 112 |
| 9 | 123 | 112 |
| 10 | 138 | 119 |
| 11 | 119 | 112 |
| 12 | 156 | 152 |

Evidence

| | | |
|---|---|---|
| Size | 12 | n |
| Average Difference | 5.83333 | $\bar{x}_D$ |
| Stdev. of Difference | 6.69916 | $s_D$ |
| Test Statistic | 3.0164 | t |
| df | 11 | |

Assumption
Populations Normal

Note: Difference has been defined as
Sample1 - Sample2

Hypothesis Testing

| Null Hypothesis | p-value | At an α of 5% |
|---|---|---|
| $H_0$: $\mu_1 - \mu_2$ = 0 | 0.0117 | Reject |
| $H_0$: $\mu_1 - \mu_2$ >= 0 | 0.9941 | |
| $H_0$: $\mu_1 - \mu_2$ <= 0 | 0.0059 | Reject |

Confidence Intervals for the Difference in Means

| (1 - α) | Confidence Interval |
|---|---|
| 95% | 5.833333 ± 4.25645 = [ 1.57689 , 10.0898 ] |

We use a t-test since the samples are small. Also, the F-test shows that we are unable to reject different variances and thus assume equal variance. We obtain a p-value = 0,0117. From this we reject the null hypothesis and conclude that the machines are significantly different.

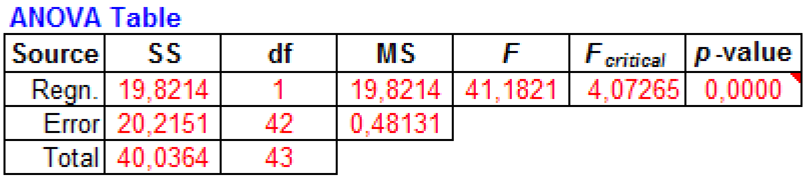
c) Explain what the P-value obtained in b) actually means.

The p-value indicates the probability of obtaining the samples given that the null hypothesis is true, i.e. under the assumption that the two machines yield similar measurements, the probability of obtaining the results from assignment a) is 0,0117.

**Assignment 5 (30%)**
Data collected in 1960 from the National Cancer Institute provides the per capita numbers of cigarettes sold along with death rates for various forms of cancer (see the Excel file Smoking and Cancer. Note: The column about "state" is irrelevant).

a) Use the coefficient of correlation to determine if a significant relationship exists between the number of cigarettes sold and each form of cancer

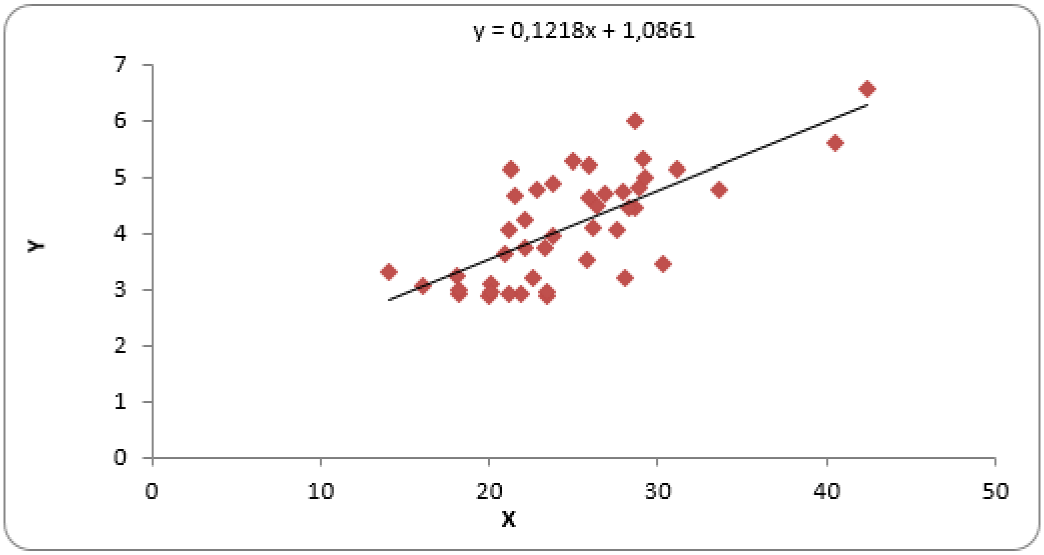
**Summary of data from template: Cigs sold vs. bladder cancer**



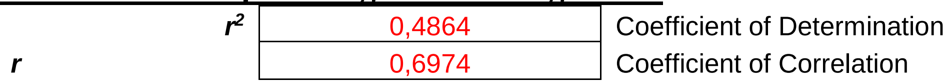
ANOVA Table

| Source | SS | df | MS | $F$ | $F_{critical}$ | $p$-value |
|---|---|---|---|---|---|---|
| Regn. | 19,8214 | 1 | 19,8214 | 41,1821 | 4,07265 | 0,0000 |
| Error | 20,2151 | 42 | 0,48131 | | | |
| Total | 40,0364 | 43 | | | | |

Scatter Plot, Regression Line and Regression Equation



**Summary of data from template: Cigs sold vs. lung cancer**

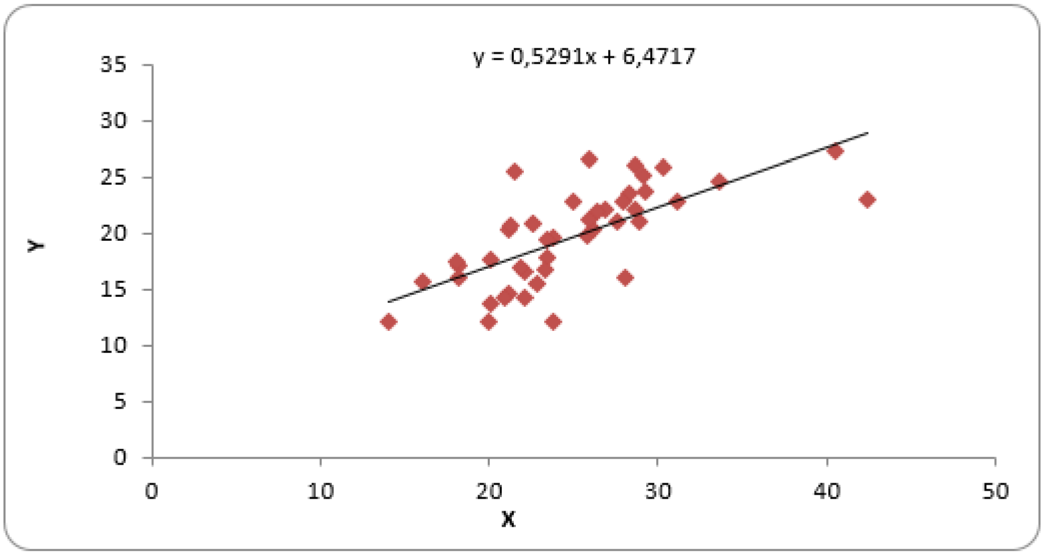


| | | | |
|---|---|---|---|
| | $r^2$ | 0,4864 | Coefficient of Determination |
| $r$ | | 0,6974 | Coefficient of Correlation |

### ANOVA Table

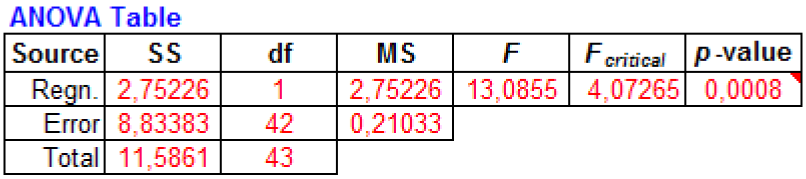
| Source | SS | df | MS | $F$ | $F_{critical}$ | $p$-value |
|---|---|---|---|---|---|---|
| Regn. | 373,878 | 1 | 373,878 | 39,771 | 4,07265 | 0,0000 |
| Error | 394,833 | 42 | 9,40079 | | | |
| Total | 768,712 | 43 | | | | |

### Scatter Plot, Regression Line and Regression Equation



**Summary of data from template: Cigs sold vs. kidney cancer**

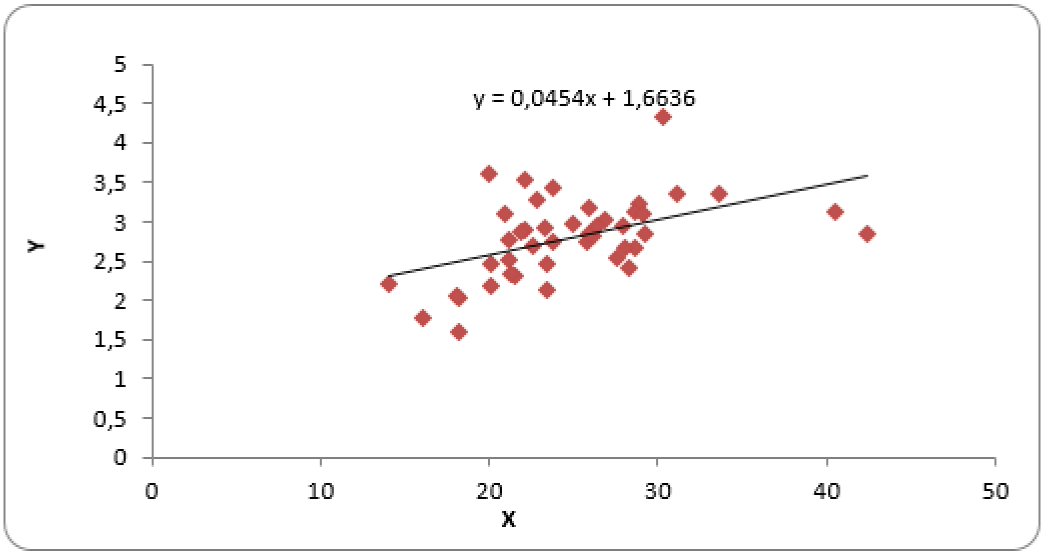
| | | |
|---|---|---|
| $r^2$ | 0,2375 | Coefficient of Determination |
| $r$ | 0,4874 | Coefficient of Correlation |



### ANOVA Table

| Source | SS | df | MS | F | $F_{critical}$ | p-value |
|---|---|---|---|---|---|---|
| Regn. | 2,75226 | 1 | 2,75226 | 13,0855 | 4,07265 | 0,0008 |
| Error | 8,83383 | 42 | 0,21033 | | | |
| Total | 11,5861 | 43 | | | | |

### Scatter Plot, Regression Line and Regression Equation



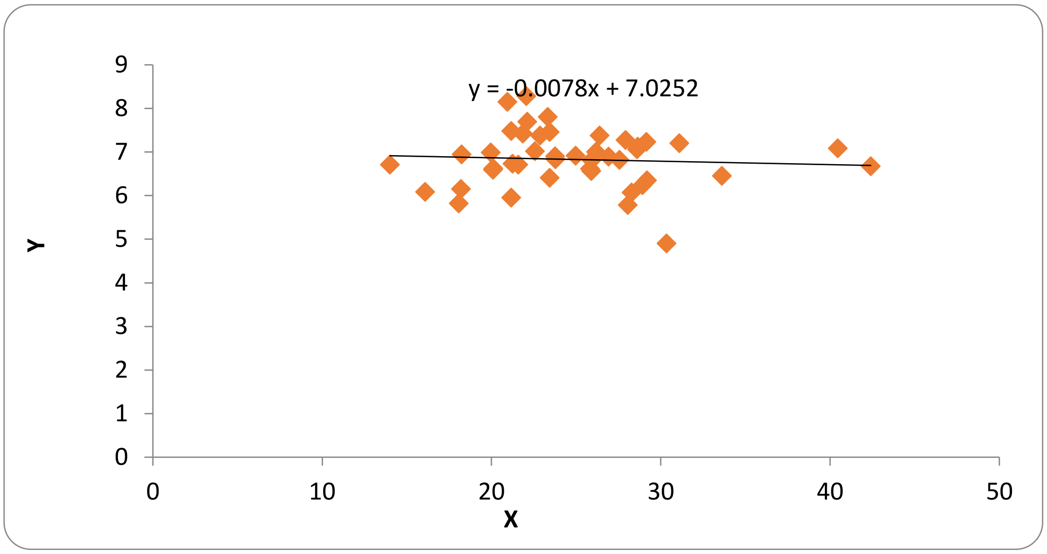
## Summary of data from template: Cigs sold vs. leukemia

| | | |
|---|---|---|
| $r^2$ | 0,0047 | Coefficient of Determination |
| $r$ | -0,0685 | Coefficient of Correlation |

### ANOVA Table

| Source | SS | df | MS | F | $F_{critical}$ | p-value |
|---|---|---|---|---|---|---|
| Regn. | 0,08215 | 1 | 0,08215 | 0,19789 | 4,07265 | 0,6587 |
| Error | 17,4349 | 42 | 0,41512 | | | |
| Total | 17,5171 | 43 | | | | |

### Scatter Plot, Regression Line and Regression Equation



I will define “significant” as a correlation greater than 0,6. That means that the highest correlation exists between bladder cancer and cigs sold followed by lung cancer. There seems only to be a small correlation between kidney cancer and cigs sold and almost no correlation between cigs sold and leukemia. Thus, only bladder cancer and lung cancer can be said to correlate significantly with cigs sold

b) In each of the cases in a) determine the correlation of determ99ination and comment on its meaning.´
r-squared bladder : 0,4951
r-squared lung: 0,4864
r-squared kidney: 0,2375
r-squared leukemia: 0,0047

The correlation of determination states the amount of variability that the model is able to explain. Thus, the model for bladder cancer is the “best” model and the one for leukemia is not a good model. We might also state that cigs sold may be used as a predictor for bladder (and lung) cancer and cannot at all be used as a predictor for leukemia.

c) Which types of cancer seems to have the highest and lowest, respectively, statistical relationship with number of cigarettes sold? (Hint: Look at the correlation of coefficients)

Bladder cancer has the highest and leukemia has the lowest.

d) For the type of cancer that has the highest relationship with number of cigarettes sold, determine what the maximum number of cigarettes sold per capita must be if we want to keep death rates below 2, 3, 4 and 5 per 100K respectively.

y = -0,0078x + 7,0252

$2 \ > \ 0{,}1218x + 1{,}0861 \Rightarrow 2 - 1{,}0861 > 0{,}1218x \Rightarrow x < 7{,}50$
$3 \ > \ 0{,}1218x + 1{,}0861 \Rightarrow 3 - 1{,}0861 > 0{,}1218x \Rightarrow x < 15{,}71$
$4 \ > \ 0{,}1218x + 1{,}0861 \Rightarrow 4 - 1{,}0861 > 0{,}1218x \Rightarrow x < 23{,}92$
$5 \ > \ 0{,}1218x + 1{,}0861 \Rightarrow 5 - 1{,}0861 > 0{,}1218x \Rightarrow x < 32{,}13$