Name: Section:

STAT 113 – Midterm 3

1) Otis[1] (1979) interviewed people waiting to see the *space aliens* film "Close Encounters of the Third Kind." Each person was asked to state his or her degree of agreement with the statement "Life on Earth is being observed by intelligent aliens," on a scale from 1 (strongly disagree) to 7 (strongly agree). Denote with x_i the i-th response. Assume that $x_i \sim N(\mu, \sigma^2)$, with known standard deviation is $\sigma = 1$. The purpose of the study was to test Otis' assertion that individuals selected movies that they were predisposed to believe in intelligent aliens. Thus we want to test

$$H_0 : \mu = 4.0 \text{ vs. } H_1 : \mu > 4.0$$

1a) [4pts] The test adopted can be described as follows:

If the sample mean of $n = 25$ responses is larger than 4.4, reject H_0.

Find α for the test.

Solution: Note that $\bar{x} \sim N(4.0, \sigma^2/5)$ with $\sigma^2/5 = 0.2$, and hence

$$P(\bar{x} > 4.4) = P(z > \frac{4.4 - 4.0}{0.2}) = P(z > 2.0) = 0.0228.$$

Note, since $\sigma^2 = 1$ is know you can use the z statistic instead of the t statistic.

1b) [3pts] Find the power $(1 - \beta)$ against $H_1 : \mu = 5$.

Solution: Power = Pr(reject false H_0). If $\mu = 5$ then $\bar{x} \sim N(5, 1/5)$ and:

$$P(\bar{x} > 4.4) = P(z > \frac{4.4 - 5}{1/5}) = P(z > -3.0) = 0.9987.$$

1c) [3pts] If the observed $\bar{X}$ for $n = 25$ was in fact 4.5, what is p-value (observed significance level)?

Solution: The alternative H_1 is one-sided to the right. Thus

$$p = P(\bar{x} > 4.5) = P(z > \frac{4.5 - 4}{1/5}) = P(z > 2.5) = 0.0062.$$

2 [10pts] Let $y_1, y_2, \ldots, y_n$ be a random sample of n observations from a normal distribution with *known* mean μ and *unknown* variance σ^2. Find the maximum likelihood estimator of σ^2.

Solution: Remember $p(y_i) = 1/(\sqrt{2\pi}\sigma) e^{-\frac{1}{2}(y_i - \mu)^2/\sigma^2}$. Hence

$$\begin{aligned}
L &= f(y_1) \cdot f(y_2) \cdot \ldots f(y_n) = \\
&= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2} \\
\log(L) &= -n/2 \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \\
\frac{\partial logL}{\partial \sigma} &= -n/\sigma + \frac{1}{\sigma^3} \sum (y_i - \mu)^2 = 0
\end{aligned}$$

and hence $\hat{\sigma}^2 = 1/n \cdot \sum_{i=1}^n (y_i - \mu)^2$.

[1] Otis, L. (1979). Selective exposure to the film "Close Encounters". *Journal of Psychology, 101, 293-295.*

3. Here are annual US data on beer production in mio gallons (X) and the number of married people (Y):

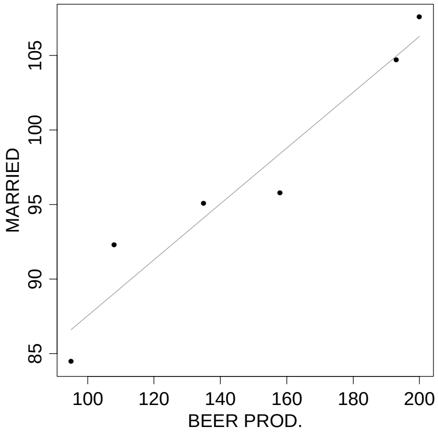
	1960	1965	1970	1975	1980	1985
X	95	108	135	158	193	200
Y	84.4	92.2	95.0	95.7	104.6	107.5

Sample standard deviations and covariances are computed to be $s_x = 43.4$, $s_y = 8.4$, and $r = 0.96$. You might also need $S_{xx} = 9407$, $S_{yy} = 354$, $S_{xy} = 1761$.

3a [2pts] Does the high correlation of $r = 0.96$ mean that drinking leads to marriage, or marriage leads to drinking, or what? (explain in no more than 10 words).

Solution: no, correllation is not the same as causation

3b [3pts] The following figure shows a scatter plot of marriages against beer consumption. In the figure, indicate the distances (residuals, deviations) whose squares are minimized to obtain the least squares regression line.



3c [3pts] If beer consumption is measured in mio liters, i.e. $z = 4 \cdot x$, how does the correlation coefficient change?

Solution: Denote with r_{zy} and r_{xy} the correllation coefficient for (z, y) and (x, y), respectively. Then

$$r_{zy} = \frac{SS_{zy}}{\sqrt{SS_{zz}SS_{yy}}} == \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sqrt{\sum(z_i - \bar{z})^2 \cdot \sum(y_i - \bar{y})^2}} = \frac{4\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{4^2\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} = r_{xy}$$

3d [2pts] Using the method of least squares, estimate the slope β_1 of the regression line $y = \beta_0 + \beta_1 x$.

Solution:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 0.19.$$

Name: Section:

STAT 113 – Midterm 2
Problem 4

Take home problem. Work on this problem at home and hand in your solution together with the in-class part of the exam on Friday.
No group work on this problem.

4) Prostate cancer is one of the most virulent forms of cancer. Generally, it has spread before being detected and is usually fatal. One dietary factor that has been studied for its relationship with prostate cancer is fat consumption. The following table lists fat consumption (in g/day, first column) and prostate cancer death rates (per 100,000, second column) for several countries.

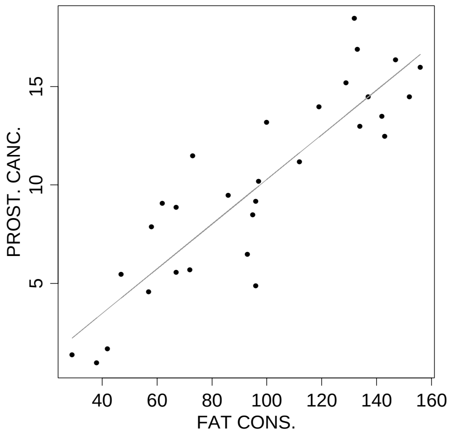
El Salvador	38	0.9	Spain	97	10.1
Philippines	29	1.3	Portugal	73	11.4
Japan	42	1.6	Finland	112	11.1
Mexico	57	4.5	Hungary	100	13.1
Greece	96	4.8	UK	143	12.4
Colombia	47	5.4	Germany	134	12.9
Bulgaria	67	5.5	Canada	142	13.4
Yugoslavia	72	5.6	Austira	119	13.9
Poland	93	6.4	France	137	14.4
Panama	58	7.8	Netherlands	152	14.4
Isreal	95	8.4	Australia	129	15.1
Romania	67	8.8	Denmark	156	15.9
Venezuela	62	9.0	US	147	16.3
Czechoslovakia	96	9.1	Norway	133	16.8
Italy	86	9.4	Sweden	132	18.4

The data are available on the STA 113 homepage (click “exams”). Or copy the data set by typing from your acpub account:

```
cp /afs/acpub/project/sta215/fat.data fat.data
```

4a [2pts] Construct a scattergram for the data (attach your plot on a seperate sheet).

Solution:



4b [2pts] Assuming the relationship between the variabels is best described by a straight line, use the method of least squares to estimate the intercept and the slope of the line.

Solution:

```
The regression equation is
C2 = - 1.06 + 0.113 C1

Predictor       Coef       Stdev    t-ratio        p
Constant      -1.063       1.170      -0.91    0.371
C1           0.11336     0.01126      10.06    0.000

s = 2.295       R-sq = 78.3%     R-sq(adj) = 77.6%

Analysis of Variance

SOURCE       DF          SS          MS        F        p
Regression    1      533.31      533.31   101.30    0.000
Error        28      147.42        5.26
Total        29      680.73
```

$\hat{\beta}_0 = -1.06$ *and* $\hat{\beta}_1 = 0.11$.

4c [2pts] Plot the least squares line on your scattergram.

4d [3pts] Do the data provide sufficient evidence to indicate that fat consumption x contributes information for the prediction of prostate cancer death rate? Test using $\alpha = 0.05$.

Solution:

$$H_0 : \beta_1 = 0 \ \textit{vs.}\ H_1 : \beta_1 > 0$$

Note: can argue from substantive knowledge of the problem that if fat consumption has any influence on cancer death rate, it's a positive slope, i.e., can use one-sided alternative (but you didn't have to).

$$t = \frac{\hat{\beta}_1}{\sigma_{\beta_1}} = \frac{0.11}{0.011} = 10.06.$$

The p-value $p = 0.000$ *(see printout) is below* $\alpha = 0.05 \Rightarrow$ *reject* H_0. *There is sufficient evidence to conclude that fat consumption does contribute information for the prediction of prostate cancer test rate.*

4e [3pts] Find a 90% confidence interval for the prostate cancer death rate in Ecuador, a country with a fat consumption of 52 g/day.

Solution: Note: The question is about predicting one *specific future observation* y_p*(not an average). The point estimate for a response* y_p *at* $x_p = 52$ *is* $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p$. *The 90% c.i. for* y_p *is given by*

$$\hat{y} \pm t_{\alpha/2,n-2}\, s\sqrt{\mathbf{1} + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}},$$

where $\alpha = 10\%$.