Tekniska högskolan, LiU
Matematiska institutionen
Matematisk statistik

# TAMS11: Probability and Statistics — computer lab

- The lab is done by **Matlab**. All students at LiU can download and install Matlab for free on your own computer through LiU IT service. In order to develop a good habit of programming in Matlab, it is better to write commands/codes in m-files, so that you can easily go back and make changes.
- After you finish each problem, write a brief report addressing the questions/suggestions within each problem. You might need to paste outputs of Matlab to support your report.
- When you are done with all problems, combine the report of each problem as **one pdf** file (on the first page of the pdf file please write **your name** and **personal number**!) and send it to
  Xiangfeng Yang (xiangfeng.yang@liu.se)

## Problem 1. Simulation of observations from a discrete distribution

In our lectures, we have often used the experiment of throwing a die to illustrate various calculations of probabilities. In this problem you should now use Matlab to simulate such throws. Start **Matlab** and in the menu **Home** click **New Script**, then an m-file is open. In the menu **Editor**, you can choose to save this m-file, and please name it as problem1.m. Now type the following commands in the m-file:

```
n=600; % n throws of the die

% Simulate one throw of the die
throw = randi(6,n,1);
throw(101:115)
```

In order to compile/run the above codes, in the menu **Editor** click **Run**, and click **Change Folder**.

Write `help randi` in the command window to see how `randi` works.

What is the meaning of the code `throw(101:115)`?

We have 6 different outcomes: 1, 2, 3, 4, 5, 6. Now we want to know the frequency $f_i$ of each outcome $i = 1, 2, \ldots, 6$. To do this, we write in m-file `sum(throw==i)` for each $i = 1, 2, \ldots, 6$. What frequencies do you get?

We can even use Histogram to plot the frequencies. To this end, we should first "comment out" the throws using % in front, that is, `% throw = randi(6,n,1)` (if we don't comment out, then Matlab will simulate new throws). Write in the m-file the following

```
% Histogram
figure
hist(throw,[1:6])
```

Now we want to find the sample mean and sample standard deviation of these 600 throws. Write in the m-file:

```
% Sample mean
sample_mean = mean(throw)
% Sample standard deviation
sample_standard_deviation = std(throw)
```

Compare these two (simulated) values with the theoretical values (namely $\mu$ and $\sigma$ which you need to compute by hand).

Now please write a brief report of Problem 1 addressing all the above questions/suggestions with possible supporting outputs from Matlab. When you are done, you can close the m-file problem1.m and the histogram graph, and continue to Problem 2.

## Problem 2. CI and HT for the difference of two population means

Use the same steps as in Problem 1 to create an m-file named problem2.m. In the first line of the m-file, write `clear;` in order to clear all previous memories stored in Matlab.

We now generate 10 observations from $N(22, 2^2)$ and 12 observations from $N(16, 2^2)$, using `normrnd`. Write in the command window `help normrnd` to see how this works. Write in the m-file:

```
n = 10;
m = 12;
mu1 = 22;
mu2 = 16;
sigma = 2;

% Generate observations
x = normrnd(mu1,sigma,n,1);
y = normrnd(mu2,sigma,m,1);
```

We may think that these two sets of observations $x$ and $y$ are from two independent populations $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively (note that the two population variances are the same!). Our goal of this problem is to use Matlab to (i) construct a 95% confidence interval of $\mu_1 - \mu_2$, and (ii) perform hypotheses test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$. To this end, write in the m-file:

```
[H,P,CI,STATS] = ttest2(x,y,0.05,'both','equal')
```

where 'equal' means two equal population variances. You can write in the command window `help ttest2` to get to know the meaning of it.

(i) The output of `CI` gives the 95% confidence interval of $\mu_1 - \mu_2$. Can you also construct the 95% confidence interval of $\mu_1 - \mu_2$ by hand? (which should coincide with the output). You should need the two sample standard deviations $s_x$ and $s_y$ which can be found by using `std(x)` and `std(y)` in the command window.

(ii) The output of `H` or `P` gives us the answer whether or not $H_0$ should be rejected. Can you reach the same conclusion by using $TS$ and $C$ by hand?

## Problem 3. CI using normal approximations

Now create an m-file named problem3.m. Again in the first line of the m-file, write `clear;` in order to clear all previous memories stored in Matlab.

In this problem we will construct confidence intervals when the population is a Binomial random variable $Bin(n, p)$. In order to use normal approximation, in Lecture it is required that $n$ is large enough such that $np \geq 10$ and $n(1-p) \geq 10$. It is interesting to see what happens if $n$ is not that large. We will construct a 95% confidence interval for $p$ using normal approximations for large $n$, and not so large $n$.

(i) We first generate 1000 samples (each sample size is 1) from $Bin(16, 0.3)$. Write in the m-file:

```
n = 16;
p = 0.3;
x = binornd(n,p,1000,1); % one can think of x as 1000 samples (each sample size is 1)
```

For each sample one can estimate $p$ by using $\hat{p}$. So 1000 samples give us 1000 estimated $\hat{p}$. Write in m-file:

```
phat = x/n; % these are 1000 estimated values of p
```

The corresponding 1000 (95%) confidence intervals are: write in the m-file

```
lower_lim = phat - 1.96*sqrt(phat.*(1-phat)/n); % this is a vector with 1000 values
upper_lim = phat + 1.96*sqrt(phat.*(1-phat)/n); % this is a vector with 1000 values
```

Now we have 1000 such 95% confidence intervals $I_p$=(`lower_lim`, `upper_lim`). Since it is 95%, there should be around 950 such intervals containing the real value $p = 0.3$, and around 50 not containing the real value. Now we count how many such intervals not containing $p = 0.3$ : write in m-file:

```
missing = sum(lower_lim > p) + sum(upper_lim < p)
```

where '`missing`' gives you the number of intervals not containing the real value $p = 0.3$. Compare this number with the expected number 50 (is it far away from 50 or very close to 50? why?)

(ii) Repeat these procedures in (i) for a new population $Bin(80, 0.3)$, and compare again.

## Problem 4. Simple linear regression

Now create an m-file named problem4.m. Again in the first line of the m-file, write `clear;` in order to clear all previous memories stored in Matlab.

A geyser is a hot spring, which more or less regularly erupts. During an eruption, the water can spray high into the air. Old Faithful Geyser in Wyoming is one such source which has become a tourist attraction. Time between two consecutive eruptions is usually long, it is therefore interested in being able to predict the time until the next eruption. It is believed that this time depends on the length of the previous eruption. To construct such a model we put

$$\begin{aligned} x = &\ \text{the length of the last eruption (unit:min) and} \\ y = &\ \text{time till the next eruption (unit:min).} \end{aligned}$$

We will use the model

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

where $\varepsilon_1, ..., \varepsilon_n$ are independent $N(0, \sigma^2)$. The sample is given as follows (copy $x$ and $y$ in the m-file)

```
x = [4.4 3.9 4.0 4.0 3.5 4.1 2.3 4.7 1.7 4.9 1.7 4.6 3.4 4.3 1.7 3.9 3.7 3.1 4.0 1.8 ...
     4.1 1.8 3.2 1.9 4.6 2.0 4.5 3.9 4.3 2.3 3.8 1.9 4.6 1.8 4.7 1.8 4.6 1.9 3.5 4.0 ...
     3.7 3.7 4.3 3.6 3.8 3.8 3.8 2.5 4.5 4.1 3.7 3.8 3.4 4.0 2.3 4.4 4.1 4.3 3.3 2.0 ...
     4.3 2.9 4.6 1.9 3.6 3.7 3.7 1.8 4.6 3.5 4.0 3.7 1.7 4.6 1.7 4.0 1.8 4.4 1.9 4.6 ...
     2.9 3.5 2.0 4.3 1.8 4.1 1.8 4.7 4.2 3.9 4.3 1.8 4.5 2.0 4.2 4.4 4.1 4.1 4.0 4.1 ...
     2.7 4.6 1.9 4.5 2.0 4.8 4.1]';

y = [78 74 68 76 80 84 50 93 55 76 58 74 75 80 56 80 69 57 90 42 91 51 79 53 82 51 76 ...
     82 84 53 86 51 85 45 88 51 80 49 82 75 73 67 68 86 72 75 75 66 84 70 79 60 86 71 ...
     67 81 76 83 76 55 73 56 83 57 71 72 77 55 75 73 70 83 50 95 51 82 54 83 51 80 78 ...
     81 53 89 44 78 61 73 75 73 76 55 86 48 77 73 70 88 75 83 61 78 61 81 51 80 79]';
```

(i) We first plot $y$ against $x$ in order to see if there is a linear relation between them, write in the m-file:

```
plot(x,y,'b*')
```

We can even find the correlation coefficient $\rho_{X,Y}$ using the code in the m-file

```
corr(x,y)
```

Note that if $|\rho_{X,Y}| \approx 1$ then it means that $x$ and $y$ have linear relation.

(ii) Now we do a full analysis of the linear regression using `regstats`, write in the m-file the following codes (try to understand the meaning of each line),

```
stats = regstats(y,x,'linear','all');
betahat = stats.tstat.beta
se = stats.tstat.se
t = stats.tstat.t
s2 = stats.mse
```

What is the estimated regression line?

(iii) To see how the estimated regression line fits the points, write in the m-file:

```
figure
scatter(x,y,'*')
xlabel('x'), ylabel('y')
hold on
lsline % ls = least square, this is how we obtain the regression line
```

(iii) In the output of Matlab, find the standard errors of the coefficients, namely, $s_{\hat{\beta}_0}$ and $s_{\hat{\beta}_1}$. (Note that in Lecture, we also use the notations $s\sqrt{h_{00}} = d(\hat{\beta}_0)$ and $s\sqrt{h_{11}} = d(\hat{\beta}_1)$ for $s_{\hat{\beta}_0}$ and $s_{\hat{\beta}_1}$).

(iv) With a significance level $\alpha = 0.01$, test the hypotheses

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0$$

Is $H_0$ rejected?

(v) It is assumed that the error terms $\varepsilon_j \sim N(0, \sigma^2)$, but is this really true ? We will study this by looking at the residuals. Write in the m-file:

```
residualer = stats.r;

figure
scatter(x,residualer,'filled')
title('Residualer')
```

The idea is: if there is no obvious pattern in the plot of residuals, then it is reasonable to say $\varepsilon_j \sim N(0, \sigma^2)$. But if there is an obvious pattern, then the error terms are not normal. Do you think $\varepsilon_j \sim N(0, \sigma^2)$ ?

## Problem 5. Logistic regression

Now create an m-file named problem5.m. Again in the first line of the m-file, write `clear;` in order to clear all previous memories stored in Matlab.

Copy the following $x$ and $y$ in the m-file:

```
x = [41 41 42 43 54 53 57 58 63 66 67 67 67 68 69 70 70 70 70 72 73 ...
     75 75 76 76 78 79 81 85 86 86 88]';

y = [1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0]';
```

The variable $x$ represents the launch temperature, and $y$ denotes incidence of failure of O-rings in 32 space shuttle launches prior to the Challenger disaster of 1986. It is noted that $y$ can only take two values: $y = 1$ (failure) and $y = 0$ (success). From the above data it is suspected that space shuttle launch will be likely to fail ($y = 1$) if launch temperature is low (say $x \leq c$ for some threshold $c$). We now use **logistic regression to model such relation**. Namely let $Y$ be a Bernoulli random variable with

$$P(Y = 1) = p(x), \quad P(Y = 0) = 1 - p(x),$$

where the failure probability $p = p(x)$ depends on the launch temperature $x$. For logistic regression, we assume that $p(x)$ depends on $x$ via the logit function

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

(i) Use Matlab to find the estimated logit function

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}.$$

To do this, write in the m-file:

```
fitglm(x,y,'distr','binomial')
```

**Run** the m-file and it is from the output that one can get the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ : namely in the output **Estimated Coefficients**, the column **Estimate** gives the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.

(ii) Now we want to test, with a significance level $\alpha = 0.05$,

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_a : \beta_1 \neq 0.$$

If we reject $H_0$, then it means that $\beta_1 \neq 0$ which suggests that the launch temperature indeed affects the launch failure. Based on the data, do we reject $H_0$ ? Use $TS$ and $C$ to answer the question: since $n = 32$ is large, we have the random variable $\frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \approx N(0,1)$, therefore $TS = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}}$ (the column **SE** in the output **Estimated Coefficients** gives us both $s_{\hat{\beta}_0}$ and $s_{\hat{\beta}_1}$), and $C = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty)$.

(iii) Now if a new space shuttle is going to be launched at a temperature $x = 65$ (which is not a temperature in the data), then what is the estimated probability $\hat{p}(65)$ that the launch will fail? If $\hat{p}(65) \geq 0.5$, then we classify $y(65) = 1$ (failure), otherwise we classify it as success.