EXERCISE SET 6,
MS-A0503, FIRST COURSE IN PROBABILITY AND STATISTICS

Explorative exercises

I will expect that you study the explorative problems before the first lecture of the week. It is very strongly recommended that you work on them in groups.

Problem 1. Eight patients have their blood pressure measured before and after trying a new medicine. The results are listed in the table below:

	1	2	3	4	5	6	7	8
Before	134	174	118	152	187	136	125	168
After	128	176	110	149	183	136	118	158

We want to test whether the medicine has any effect on the blood pressure.

(1) Suggest a test statistic and a null hypothesis. (Hint: We are interested in whether the difference $Y_i = X_{i,\text{after}} - X_{i,\text{before}}$ of the blood pressure of the i^{th} patient cen be explained by measuring errors or random variations from one day to the other. What distribution would such random fluctuations have?)
(2) Is the change statistically significant with confidence level 5%? With confidence level 1%?

Problem 2. Recall that the *covariance* of two random variables X and Y is

$$\mathrm{Cov}(X,Y) = E\left((X-\mu)(Y-\nu)\right) = E(XY) - \mu\nu.$$

Suggest a way to estimate $\mathrm{Cov}(X,Y)$ from n observations

$$(x_1,y_1),\ldots,(x_n,y_n).$$

(Hint: Compute the expected value of $\sum_i (x_i-\bar{x})(y_i-\bar{y})$.)

Problem 3. The following is the height (in inches) of ten pairs of a father and his son.

Fathers' height	60	62	64	65	66	67	68	70	72	74
Sons' height	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

(1) Estimate (using your answer in Problem 2) the covariance of the height of a father and his son.
(2) (Challenging:) Discuss how (and whether) this could be used to test the hypothesis that the height of a father is independent from that of his son.

Homework problems

The homework problems are reported during the second exercise session of the week. You are allowed and encouraged to work in groups, but every student should be prepared to present the solutions individually. During the last exercise session of the week, the teacher will ask you to mark what problems you have solved, and you get points according to how many problems you marked as solved. If you mark a problem as solved, however, you should also be prepared to present your solution in front of the class.

Homework 1. A factory manufactures nails with target length 10 cm. However, the length of manufactured nails varies randomly according to normal distribution. The quality of nails is controlled such that on each full hour, 30 nails are selected randomly and measured.

In a sample the average length of nails is 10.05 and the sample variance is 0.16cm^2.Test the null hypothesis that the length of these nails is on average 10 cm using 5% significance level, under the alternative hypothesis that the average length differs from the target length.

Homework 2. There are two machines, M_1 and M_2, which manufacture screws in a screw factory. The thickness of screws manufactured by these machines vary randomly and independently according to a normal distribution. We pick two independent random samples of screws manufactured by each machine and compute the sample variances of their thickness. Data from the samples is shown on the table below. Test the null hypothesis that the machines manufacture equally thick screws on average, when the alternative hypothesis is that there is a difference between the machines. Use 5% significance level.

Machine	Average (mm)	Sample variance (mm^2)	Sample size
M_1	9.9	0.25	31
M_2	10.3	0.16	21

Homework 3. In an opinion poll, 3433 random Finns are asked which party they intend to vote for in the next general election. In the September poll, 17.6% claimed that they would vote for Keskusta, whereas only 16.5% claimed they would in the November poll. On confidence level 95%, is it true that the support for Keskusta has fallen?

Week 6, Homework 1

$X_1 \dots X_{30}$ length of nails.

H_0 : $X_1 .. X_{30}$ i.i.d. $\mathcal{N}(10, \sigma)$

Test statistic: $T(X) = \dfrac{\bar{X} - 10}{S/\sqrt{30}} \sim t_{29}$

assuming H_0.

Plugging in $\bar{X} = 10.05$ $S^2 = 0.16$, we get

$T(X) = \dfrac{0.05}{0.4/\sqrt{30}} \approx 0.68$.

If $T \sim t_{29}$, then

$\mathbb{P}[|T| \geq 0.68] > \mathbb{P}[|T| \geq 2.045] = 0.05$, so

we accept the null hypothesis with significance level 5%.

Week 6, Homework 2

$X_1 \dots X_{31}$ screws from M_1

$Y_1 \dots Y_{21}$ — " — M_2

~~Hypothesis~~ $X_1 \dots X_{31}$ iid, $\cancel{N(\mu_1, \sigma_1)}$ mean μ_1, var σ_1^2

$Y_1 \dots Y_{21}$ iid, mean μ_2, var σ_2^2.

H_0 : $\mu_1 = \mu_2$.

$\bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. Unfortunately, we do not know the <u>exact</u> distribution of $\bar{X} - \bar{Y}$, even after normalizing with S_1 and S_2.

But $Var(\bar{X} - \bar{Y}) = Var\,\bar{X} + Var\,\bar{Y} = \frac{\sigma_1^2}{31} + \frac{\sigma_2^2}{21}$,

so by CLT, $\bar{X} - \bar{Y}$ can be <u>approximated</u> by $N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{31} + \frac{\sigma_2^2}{21}\right)$.

Assuming H_0, we would have

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{31} + \frac{\sigma_2^2}{21}}} \underset{\text{approx}}{\sim} N(0,1).$$

Week 6, Homework 2, Continued

Assuming large sample size (31 & 21 are large enough), we have

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{31}+\frac{\sigma_2^2}{21}}} \approx \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_1^2}{31}+\frac{S_2^2}{21}}} = \frac{9.9-10.3}{\sqrt{\frac{0.25}{31}+\frac{0.16}{21}}} \approx 3.19$$

But if $Z \sim N(0,1)$, then

$$\mathbb{P}[|Z|>3.19] = 2\,\Phi(-3.19)$$

$$\approx 0.0014,$$ so we reject H_0.

Week 6, Homework 3

Denote: $X_i = \mathbb{I}_{\{i^{th} \text{ voter in September poll votes Kesk}\}}$

$Y_i = \mathbb{I}_{\{i^{th} \text{ voter in November poll votes Kesk}\}}$

$\mathbb{E}[X_i] = \mu_{Sept} \qquad \mathbb{E}[Y_i] = \mu_{Nov}$

$H_0: \ \mu_{Sept} = \mu_{Nov}, \quad X_i, Y_j \ \text{i.i.d.}$

Assuming H_0, an $\overset{\text{approximately}}{\chi^2}$-distributed unbiased estimator for $Var(X_i - Y_j) = Var(X_i + Y_j)$ is

$$S^2 = \frac{\sum X_i^2 + \sum Y_i^2 - \sum \left(\frac{\bar{X}+\bar{Y}}{2}\right)^2}{2 \cdot 3433 - 1} = \frac{3433 \cdot 0.176 + 3433 \cdot 0.165 - 0.1705^2 \cdot 3433}{2 \cdot 3433 - 1}$$

$$\approx \frac{1.705 - 0.175^2}{2} \approx 0.1414.$$

So we observed $\dfrac{\bar{X} - \bar{Y}}{S/\sqrt{n}} = \dfrac{0.176 - 0.165}{\sqrt{0.1414}/\sqrt{3433}} \approx 1.713.$

Assuming H_0, $\dfrac{\bar{X}-\bar{Y}}{S/\sqrt{n}} \underset{\text{approx}}{\sim} N(0,1)$, and if $Z \sim N(0,1)$,

then $\mathbb{P}\left[|Z| \geq 1.713\right] \approx 0.086$, so we cannot reject H_0.