UPPSALA UNIVERSITY 5LN445 Mathematics for Language Technologists
Fredrik Wahlberg and Mats Dahllöf 30 October 2019

Dugga 2: Statistics and Linear algebra

Pass: at least 60% correct out of max 40p (i.e. 24p). Pass with distinction: at least 80% correct on the two tests added. When in doubt about the interpretation of a question, make reasonable assumptions and motivate those. If you get stuck on a task, try to solve other tasks first, then go back. Please read the whole exam before beginning.
General rules: Mobile phones must be switched off.
Tools: Pen and calculator. Provided: normal cdf table, χ^2 table and a table of equations.

Statistics (24p)

1. Since 1977, the Swedish Scholastic Aptitude Test ("Högskoleprovet") provides a way into higher education without having reached the required grade cut-offs from school. It is taken by approximately 50000 future students each semester. To make the tests comparable from year to year, a normal distribution model is used from normalizing the scores to the range [0, 2]. For the spring test 2019, the future student's scores had the mean 0.88 and standard deviation 0.39 (i.e. $X \sim \mathcal{N}(\mu=0.88, \sigma=0.39)$). For this semester at Uppsala University, the score cut-off for the master's programme in medicine was 1.70, bachelor's programme in economics 1.40, and bachelor's programme in social sciences 1.20. Find:
   a. The expectation of the distribution over scores (i.e. E(X)). (1p)

   For a random person having taken the test, find:
   b. The probability of making the economics cut-off. (1p)
   c. The probability of making the social sciences cut-off. (1p)
   d. The probability of *not* making the social sciences cut-off. (1p)
   e. The probability of making the economics cut-off, but *not* the medicine cut-off. (2p)
   f. What is the probability of having a score within the interval $\mu \pm 1.96\sigma$, and what interval of scores does this represent. (2p)

> a. $E(X) = \mu = 0.88$, by definition
> b. $P(1.40 < X) = 1-\phi((1.40-0.88)/0.39) \approx 1-\phi(1.33) \approx 9.2\%$
> c. $P(1.20 < X) = 1-\phi((1.20-0.88)/0.39) \approx 1-\phi(0.82) \approx 21\%$
> d. $P(X < 1.20) = \phi((1.20-0.88)/0.39) \approx \phi(0.82) \approx 79\%$
> e. $P(1.40 < X < 1.70) = \phi((1.70-0.88)/0.39) - \phi((1.40-0.88)/0.39) \approx \phi(2.1)-\phi(1.33) \approx 7.4\%$
> f. $P(\mu-1.96\sigma \leq X \leq \mu+1.96\sigma)$ is the same for all normal distributions.
> $P(-1.96 \leq Z \leq 1.96) \approx 95\%$
> $\mu-1.96\sigma = 0.88-1.96 \cdot 0.39 \approx 0.12$
> $\mu+1.96\sigma = 0.88+1.96 \cdot 0.39 \approx 1.64$

2. In one of the lecturer's favorite surveys, *"Democrats and Republicans differ on conspiracy theory beliefs"*, 1247 people living in the USA were asked about "conspiracy theories". Voters of their two major political parties answered somewhat differently on the question "Q6: Do you believe there is a link between childhood vaccines and autism, or not?". Out of 474 democrat voters 16% answered yes, and out of 434 republican voters 26% answered yes.
   (For this exam, we assume this is a very high quality study.)
   a. For each party group (i.e. republican and democrat), estimate their respective sampling distribution. (2p)
   b. For each party group, calculate their 95% confidence interval. (2p)
   c. We believe that there is not a significant difference between the party groups. We want to test if there is good reason to hold on to this belief. Formulate a null hypothesis and an alternative hypothesis. (1p)

d. Calculate the distribution of the difference between the party groups, given H_0. (2p)
e. What is the p-value for this test? (2p)
f. What are the α-values for the significance levels 90, 95, and 99? (1p)
g. Can the null hypothesis be refuted at an α of 0.1, 0.05 or 0.01. (1p)

a. From the text: $p_1 = 0.16$, $n_1 = 474$, $p_2 = 0.26$, $n_2 = 434$

$$\hat{p}_1 \sim \mathcal{N}\left(\mu = p_1 = 0.16, \sigma = \sqrt{\frac{p_1(1-p_1)}{n_1}} \approx 0.017\right)$$

$$\hat{p}_2 \sim \mathcal{N}\left(\mu = p_2 = 0.26, \sigma = \sqrt{\frac{p_2(1-p_2)}{n_2}} \approx 0.021\right)$$

b. From the text: $p_1 = 0.16$, $n_1 = 474$, $p_2 = 0.26$, $n_2 = 434$, CI 95% $z^*=1.96$

$$p_1 \pm z^* \cdot \sqrt{\frac{p_1(1-p_1)}{n_1}} = 0.16 \pm 1.96 \cdot \sqrt{\frac{0.16 \cdot 0.84}{474}} \Rightarrow (0.12, 0.20)$$

$$p_2 \pm z^* \cdot \sqrt{\frac{p_2(1-p_2)}{n_2}} = 0.26 \pm 1.96 \cdot \sqrt{\frac{0.26 \cdot 0.74}{434}} \Rightarrow (0.21, 0.31)$$

c. H_0: There is no difference between the party groups
H_a: There is a difference between the party groups

d. Under H_0 the probabilities are pooled for the distribution of the difference.

$$p_{pooled} = \frac{p_1 \cdot n_1 + p_2 \cdot n_2}{n_1 + n_2} \approx 0.21$$

$\Rightarrow (\hat{p}_2 - \hat{p}_1) \sim \mathcal{N}(\mu, \sigma)$, where:

$$\mu = p_2 - p_1 = 0.10$$

$$\sigma = \sqrt{\frac{p_{pooled}(1-p_{pooled})}{474} + \frac{p_{pooled}(1-p_{pooled})}{434}} \approx 0.027$$

e. We are interested in the probability of a 10 percentage point difference under the null hypothesis (i.e. no difference) given the spread in the distribution of the difference from (d).

$$P(0.10 < \hat{p}_2 - \hat{p}_1 | H_0) = 1 - \Phi\left(\frac{0.10 - 0}{0.027}\right) \approx 1 - \Phi(3.7) \approx 0.0001$$

f. Significance levels 90, 95, and 99 corresponds to α-values .1, .05, and .01 .

g. When α=0.1: H_0 can be refuted
When α=0.05: H_0 can be refuted
When α=0.01: H_0 can be refuted

3. In the paper *"Athletics: momentous sprint at the 2156 Olympics?"* by Tatem at al. (Nature, 2004), the authors propose that female and male Olympic 100-meter sprinters will be equally fast in the year 2156. In the papers appendix, two data sets are given. One data set where x were years and y were male running times, the other where x were years and y were female running times. Below are the means and covariance matrices for the male and female data sets.
(This is likely a hoax paper. However, we will treat it as truth for this task.)

$$\bar{x}_{male} = 1954, \bar{x}_{female} = 1969, \bar{y}_{male} = 10.32, \bar{y}_{female} = 11.23$$

$$C_{male} = \begin{pmatrix} 1024 & -11.27 \\ -11.27 & 0.1406 \end{pmatrix} \text{ and } C_{female} = \begin{pmatrix} 520 & -8.740 \\ -8.740 & 0.1864 \end{pmatrix}$$

a. From the given data, find the least square linear model $y=\beta_0+x\beta_1$ for *male* sprinters. (1p)
b. From the given data, find the least square linear model $y=\beta_0+x\beta_1$ for *female* sprinters. (1p)

c. At what future year do the models predict an equal outcome? (1p)
d. How much of the data variance is explained by the respective models? (2p)

Four significant digits are used in these solutions.

a. The model parameter β_1 can be found from the covariance matrix of the data.

$$C = \begin{pmatrix} s_x^2 & Rs_xs_y \\ Rs_ys_x & s_y^2 \end{pmatrix}$$

$$\beta_1 = R \cdot \frac{s_y}{s_x} = \frac{Rs_xs_y}{s_x^2} = \frac{-11.27}{1024} \approx -0.01101$$

All values needed for β_0 are now found above and in the given data.

$$\beta_0 = \bar{y}_{male} - \beta_1 \cdot \bar{x}_{male} = 10.32 - (-0.01101) \cdot 1954 \approx 31.83$$

Note that there is no need for calculating any covariance here. All information is already in the C for each data set.

Final model: y = -0.01101x+31.83

b. Analogous with (a) but with numbers from the other covariance matrix:

$$\beta_1 = R \cdot \frac{s_y}{s_x} = \frac{Rs_xs_y}{s_x^2} = \frac{-8.740}{520} \approx -0.01681$$

$$\beta_0 = \bar{y}_{female} - \beta_1 \cdot \bar{x}_{female} = 11.23 - (-0.01681) \cdot 1969 \approx 44.32$$

Final model: y = -0.01681x+44.32

c. Both models will at some point in the future predict the same y:

$$-0.01681x + 44.32 = -0.01101x + 31.83$$

$$\Rightarrow 0.0058x = 12.49 \Rightarrow x = \frac{12.49}{0.0058} \approx 2153$$

d. The explained variance is the R^2 score. The R^2 score can be found in terms of the given data without much algebra.

Male model:

$$R^2 = \frac{Rs_xs_y \cdot Rs_xs_y}{s_x^2 \cdot s_y^2} = \frac{-11.27 \cdot -11.27}{1024 \cdot 0.1406} \approx 0.8822$$

Female model:

$$R^2 = \frac{Rs_xs_y \cdot Rs_xs_y}{s_x^2 \cdot s_y^2} = \frac{-8.740 \cdot -8.740}{520 \cdot 0.1864} \approx 0.7881$$

Linear algebra (16p)

4. Given the vectors ($\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{u}_1$, $\mathbf{u}_2$), give the resulting vector or scalar for the expressions a-f below. (6p)

$\mathbf{v}_1 = (1, 5, 2)^T$ $\quad$ $\mathbf{u}_1 = (4, 1, 8)^T$

$\mathbf{v}_2 = (0, 3/2, -1)^T$ $\quad$ $\mathbf{u}_2 = (3, 4, -2)^T$

a. $-\mathbf{v}_1$
b. $\mathbf{v}_1 + \mathbf{v}_2$
c. $2(4\mathbf{v}_1 - 3\mathbf{v}_2)$
d. $||\mathbf{u}_1||$
e. $\mathbf{u}_1/||\mathbf{u}_1||$
f. $\mathbf{u}_1 \cdot \mathbf{u}_2$

a. $-\mathbf{v}_1 = (-1, -5, -2)^T$
b. $\mathbf{v}_1 + \mathbf{v}_2 = (1, 6.5, 1)^T$
c. $2(4\mathbf{v}_1 - 3\mathbf{v}_2) = (8, 31, 22)^T$
d. $||\mathbf{u}_1|| = 9$
e. $\mathbf{u}_1/||\mathbf{u}_1|| = (4/9, 1/9, 8/9)^T$
f. $\mathbf{u}_1 \cdot \mathbf{u}_2 = 4 \cdot 3+1 \cdot 4+8 \cdot -2=0$

5. Given the following geometric shapes (P,R,Q,S), give solutions to the tasks below.
(If a number has many decimals, it can be given as a quotient in the final answer.)

P: The line $(1, 2, 3)^T + t \cdot (1, -1, 1)^T$, where $t \in \mathbb{R}$

R: The line $(8, -1, 2)^T + s \cdot (2, 0, -2)^T$, where $s \in \mathbb{R}$

Q: The plane 2x+y+2z+5=0

S: A sphere with its centre at $(4, -2, 4)^T$ and radius 4

a. Find a point where the lines P and R intersect. (3p)

b. Find a point where the line P intersects the plane Q. (3p)

c. Find a point on the line P that is inside the sphere S, if such a point exists. (4p)

a. Setting P equal to R: $(1, 2, 3)^T + t \cdot (1, -1, 1)^T = (8, -1, 2)^T + s \cdot (2, 0, -2)^T$

$$\begin{cases} 1+t &= 8+2s \\ 2-t &= -1 \\ 3+t &= 2-2s \end{cases}$$

Eq 2: $2 - t = -1 \Rightarrow t = 3$

Eq 1 and t=3: $1 + 3 = 8 + 2s \Rightarrow s = -2$

P and t=3: $(1, 2, 3)^T + 3 \cdot (1, -1, 1)^T = (4, -1, 6)^T$

R and s=-2: $(8, -1, 2)^T + -2 \cdot (2, 0, -2)^T = (4, -1, 6)^T$

b. Substituting P into Q, where $(1, 2, 3)^T + t \cdot (1, -1, 1)^T = (x, y, z)^T$ and 2x+y+2z+5=0 :

$$2(1+t) + (2-t) + 2(3+t) + 5 = 0 \Rightarrow 2 + 2t + 2 - t + 6 + 2t + 5 = 0$$

$$\Rightarrow 3t = -15 \Rightarrow t = -5$$

P and t=-5: $(1, 2, 3)^T + -5 \cdot (1, -1, 1)^T = (-4, 7, -2)^T$

c. All points $\bar{p}$ inside the sphere S must satisfy:

$||\bar{p} - (4, -2, 4)^T|| \leq 4$, where $\bar{p} = (1, 2, 3)^T + t \cdot (1, -1, 1)^T$

Taking the square of the inequality and substituting the line equation:

$$||(1, 2, 3)^T + t \cdot (1, -1, 1)^T - (4, -2, 4)^T||^2 \leq 4^2$$

Expanding the euclidean distance:

$$(t-3)^2 + (4-t)^2 + (t-1)^2 \leq 16$$

Only one value of t that fits the equation is needed. A good start is to try to minimize one of the squares on the left hand side.

Try t=1: $(1-3)^2+(4-1)^2+(1-1)^2 = 4+9+0 \leq 16$ (inequality holds)

Try t=3: $(3-3)^2+(4-3)^2+(3-1)^2 = 0+1+4 \leq 16$ (inequality holds)

Try t=4: $(4-3)^2+(4-4)^2+(4-1)^2 = 1+0+9 \leq 16$ (inequality holds)

At least one point inside S, with t=1: $(1, 2, 3)^T + 1 \cdot (1, -1, 1)^T = (2, 1, 4)^T$

Table of equations

Statistics

Binomial distribution

$B \sim Binom(n,p)$

$P(B=k) = \binom{n}{k} p^k (1-p)^{n-k}$

$P(B \leq x) = \sum_{k=1}^{x} P(B=k)$

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Normal approximation

$\mu = np$

$\sigma^2 = np(1-p)$

$P(a \leq B \leq b) = \Phi\left(\frac{b+\frac{1}{2}-\mu}{\sigma}\right) - \Phi\left(\frac{a-\frac{1}{2}-\mu}{\sigma}\right)$

Normal distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$

$P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$

Linear combinations

$aX_1 + bX_2 + c$, where: $a,b,c \in \mathbb{R}$

$\mu_{new} = a\mu_1 + b\mu_2 + c$

$\sigma^2_{new} = (a\sigma_1)^2 + (b\sigma_2)^2$

Estimators

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Significance and confidence

$\hat{p} \sim N\left(\mu = p, SE = \sqrt{\frac{p\cdot(1-p)}{n}}\right)$

$n_{sample} \geq \left(\frac{z^*}{\mathrm{ME}}\right)^2 p(1-p)$

CI ⇒p±z* · SE

For 95% CI, z*=1.96. For 99% CI, z*=2.58.

p-value = P(observations with condition | H_0)

χ^2 test

$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df=k-1}$

Regression

$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$

$\beta_1 = R \cdot \frac{s_y}{s_x}$

$C = \begin{pmatrix} s_x^2 & Rs_x s_y \\ Rs_y s_x & s_y^2 \end{pmatrix} = \begin{pmatrix} Var(x,x) & Var(x,y) \\ Var(y,x) & Var(y,y) \end{pmatrix}$

$Var(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$residual = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

Linear Algebra

$\bar{p}, \bar{q} \in \mathbb{R}^n$

$\bar{p} \cdot \bar{q} = \sum_{i=1}^{n} p_i q_i = ||\bar{p}||\, ||\bar{q}||\, cos\theta$

$||\bar{p}|| = \sqrt{\sum_{i=1}^{n} p_i^2}$

$d(\bar{p}, \bar{q}) = ||\bar{p} - \bar{q}|| = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$

Geometry

Line: $\bar{p} = \bar{p_0} + t\bar{v}$

$\frac{x-x_0}{v_1} = \frac{y-y_0}{v_2} = \frac{z-z_0}{v_3}$

Plane: $\bar{n}(\bar{p} - \bar{p_0}) = 0$

$Ax + By + Cz + D = 0, \bar{n} = (A,B,C)^T$

Sphere: $||\bar{p} - \bar{p_0}|| = r$

$(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2 = r^2$