UPPSALA UNIVERSITY
Fredrik Wahlberg and Mats Dahllöf

5LN445 Mathematics for Language Technologists
7 November 2018

Dugga 2: Statistics, Linear algebra and Graphs

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, paper (provided), calculator, probit table (provided) and table of equations (provided).

Statistics (24p)

1. The function *f(x)*, below, can be interpreted as a probability density function (continuous, $x \in \mathbb{R}$) or a probability mass function (discrete, $x \in \mathbb{Z}$).

$$f(x) = \begin{cases} 0, & x < -2 \\ C(2-|x|), & -2 \leq x \leq 2 \\ 0, & 2 < x \end{cases}$$

For *both* the discrete and continuous case, find:
    a. The normalization constants C. (2p)
    b. P(X ≤ 10). (1p)
    c. P(-1 ≤ X ≤ 1). (2p)
    d. P(X = 0). (2p)

> a. Continuous: By symmetry around x=0 and the triangular shape of the areas under the curve, $2 \cdot \frac{1}{2}(2 \cdot 2C) = 1$ gives C=¼
> Discrete: 0+1+2+1+0=4 gives C=¼
> b. All probability mass is below 10, hence P(X ≤ 10)=1 in both cases.
> c. Continuous: Symmetry around 0 leaves a rectangle and a triangle under the curve.
> $2 \cdot \left(1 \cdot f(1) + \frac{1}{2}(1 \cdot (f(2) - f(1)))\right) = 2C + C = \frac{3}{4}$
> Discrete: f(-1)+f(0)+f(1)= 1C+2C+1C = 1
> d. Continuous: C = ¼ ⇒ P(X=0) = 2C = 2/4 = .5. P(X=0) can also be interpreted as always 0.
> Discrete: C = ¼ ⇒ P(X=0) = 2C = 2/4 = .5

2. Since 1977, the Swedish Scholastic Aptitude Test ("Högskoleprovet") provides a way into higher education without having reached the required grade cut-offs. The distribution over scores is approximately normal and limited to the range [0, 2]. For the spring test 2018, the score had the mean 0.88 and standard deviation 0.39. For this semester at Uppsala University, the score cut-off for the engineering physics master was 1.35, Bachelors in Economics 1.45, and bachelors in history 0.70. Here, the distribution of scores is modelled as $X \sim \mathcal{N}(0.88, (0.39)^2)$. Find:
    a. E(X). (1p)

For a random person taking the test, find:
    b. The probability of making the economics cut-off. (1p)
    c. The probability of making the history cut-off. (1p)
    d. The probability of *not* making the history cut-off. (1p)
    e. The probability of making the engineering physic cut-off, but *not* the economics cut-off. (1p)
    f. What is the probability of getting a score within the interval μ±2σ, and what range of scores does this represent in real terms. (1p)

a. $E(X) = \mu$, by definition
b. $P(1.45 \le X) = 1-\phi((1.45-0.88)/0.39) \approx 1-\phi(1.46) \approx 7.2\%$
c. $P(0.70 \le X) = 1-\phi((0.70-0.88)/0.39) \approx 1-\phi(-0.46) = 1-(1-\phi(0.46)) \approx 67.8\%$
d. $P(X < 0.70) = \phi((0.70-0.88)/0.39) \approx 1-\phi(0.46) \approx 32.2\%$
e. $P(1.35 \le X < 1.45) = \phi((1.45-0.88)/0.39) - \phi((1.35-0.88)/0.39) \approx \phi(1.46)-\phi(1.21) \approx 4.2\%$
f. $P(\mu-2\sigma \le X \le \mu+2\sigma)$ is the same for all normal distributions. $P(-2 \le Z \le 2) \approx 95.5\%$
$\mu-2\sigma = 0.88-2*0.39 \approx 0.10$, $\mu+2\sigma = 0.88+2*0.39 = 1.66$

3. After the Swedish election of 2018, 161 seats of the 349 seats in parliaments were occupied by women. The same figure for the 2014 election was 153 seats. With the assumptions that the proportion of men and women in parliament should match the proportion in the general population (i.e. 50/50) and using a binary definition of gender, the null hypothesis is defined as equality. For *both elections*, find: *(Reasonable approximations are encouraged but should be motivated.)*
    a. The distribution parameters for modelling the gender balance, for both elections, as bernoulli trials i.e. binomial distributions. (1p)
    b. What is the approximate range of seats, in real numbers, for a two-tailed 95% confidence level (i.e. $\mu \pm 1.96\sigma$)? (1p)
    c. What are the p-values for refuting H_0 in favour of H_a: discrimination *against women*? (3p)

a. The distributions for both elections have the same n and p. B ~ Binom(n=349, p=.5)
b. Normal approximation of the binomial distribution:
$\mu = np = 174.5$, $\sigma = \sqrt{np(1-p)} = \sqrt{n}/2 \approx 9.34 \Rightarrow \mu \pm 1.96\sigma \approx [156, 193]$
c. "discrimination against women", i.e. a one-sided test
$P(X \le 161 \mid H_0) = \phi((161.5-\mu)/\sigma) \approx 0.082$
$P(X \le 153 \mid H_0) = \phi((153.5-\mu)/\sigma) \approx 0.012$

4. In a study, exhaustion disorder is treated with acceptance and commitment therapy (ACT). You are, in your capacity as having taken a statistics courses, asked to help with determining significance. The severity of the disorder, in the relevant population, is quantified using the Karolinska Exhaustion Disorder Scale (KEDS). It turns out that the afflicted population is normally distributed with a mean of 25 and a standard deviation of 4.5. After the treatment of a group of 20 patients, their mean KEDS score is 23. Use the large sample Z-test to solve the following:
    a. Formulate a null hypothesis H_0 and an alternative hypothesis H_a. We are only interested in if the treatment reduced the KEDS score. (1p)
    b. What are the α-values for the significance levels 90, 95, and 99? (1p)
    c. Find the rejection region, given a significance level of 95%. (2p)
    d. What is the p-value of this test? (2p)

a. H_0: The treatment has no effect.
H_a: The treatment was successful.
b. Significance levels 90, 95, and 99 correspond to α-values .1, .05, and .01
c. We are interested in a reduction, hence, the rejection region is on the form $[-\infty, c]$

$$P(\bar{X} \le c|H_0) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{c-\mu}{\sigma/\sqrt{n}}\right) \approx \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right) = \alpha$$

$$\frac{c-25}{4.5/\sqrt{20}} = \Phi^{-1}(.05) \approx -1.645 \Rightarrow c \approx 23.3$$

d. P-value: $P(X \le 23 \mid H_0)$

$$P(\bar{X} \le 23|H_0) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{23-\mu}{\sigma/\sqrt{n}}\right)$$

$$= \Phi\left(\frac{23-25}{4.5/\sqrt{20}}\right) \approx \Phi(-1.99) \approx 0.023$$

Linear algebra (10p)

5. Given the following vectors ($\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{u}_1$, $\mathbf{u}_2$), give the resulting vector from expressions a-f below. (6p)

   $\mathbf{v}_1 = (1, 5, 2)^T$ $\quad\mathbf{u}_1 = (2, 6, 3)^T$
   $\mathbf{v}_2 = (0, 3/2, -1)^T$ $\quad\mathbf{u}_2 = (1, -2, 3)^T$

a. $-\mathbf{v}_1$ b. $\mathbf{v}_1 + \mathbf{v}_2$ c. $4(3\mathbf{v}_1 - 2\mathbf{v}_2)$	d. $\lVert\mathbf{u}_1\rVert$ e. $\mathbf{u}_1/\lVert\mathbf{u}_1\rVert$ 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. $4(3\mathbf{v}_1 - 2\mathbf{v}_2) = (12, 48, 32)^T$	d. $\lVert\mathbf{u}_1\rVert = 7$ e. $\mathbf{u}_1/\lVert\mathbf{u}_1\rVert = (2/7, 6/7, 3/7)^T$ f. $\mathbf{u}_1 \cdot \mathbf{u}_2 = -1$

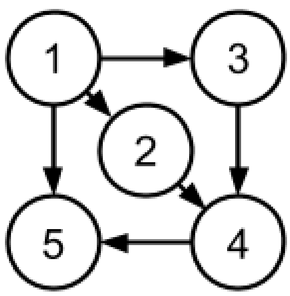
6. Given the lines (on parametric form) $\mathbf{p}=\mathbf{u}_1+t\mathbf{v}_1$ and $\mathbf{q}=\mathbf{u}_2+s\mathbf{v}_2$, where $t,s \in \mathbb{R}$, the vectors ($\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{u}_1$, $\mathbf{u}_2$) from above, and the plane $2x+y+2z+3=0$:
   a. Find a pair (t, s) for where the lines intersect. (2p)
   b. Find the points where the respective lines intersect with the plane. (2p)

a. $\mathbf{p} = \mathbf{u}_1 + t\mathbf{v}_1 = (2+t, 6+5t, 3+2t)^T$
$\mathbf{q} = \mathbf{u}_2 + s\mathbf{v}_2 = (1, -2+3s/2, 3-s)^T \Rightarrow 2+t=1,\ 6+5t=-2+3s/2,\ 3+2t=3-s$
$2+t=1 \Rightarrow t=-1$
$3+2t=3-s$ and $t=-1 \Rightarrow s=2$

b. $2x+y+2z+3=0$
$\mathbf{p} = \mathbf{u}_1 + t\mathbf{v}_1 = (2+t, 6+5t, 3+2t)^T \Rightarrow 2(2+t)+(6+5t)+2(3+2t)+3=4+2t+6+5t+6+4t+3=0$
$\Rightarrow 19+11t=0 \Rightarrow t=-19/11$
$\mathbf{u}_1 + -(19/11)\mathbf{v}_1 \approx (0.28, -2.64, -0.45)^T$
$\mathbf{q} = \mathbf{u}_2 + s\mathbf{v}_2 = (1, -2+3s/2, 3-s)^T \Rightarrow 2(1)+(-2+3s/2)+2(3-s)+3=2-2+3s/2+6-2s+3=0$
$\Rightarrow 9-½s=0 \Rightarrow s=18$
$\mathbf{u}_2 + 18\mathbf{v}_2 = (1, 25, -15)^T$

Graph theory (6p)

7. For the directed graph to the right:
   a. Make an adjacency matrix. (3p)
   b. Is this graph acyclic (i.e. are there no possible cycles)? (1p)
   c. Find all paths from vertex 1 to vertex 5. (2p)



1. A = (0 1 1 0 1)
(0 0 0 1 0)
(0 0 0 1 0)

(0 0 0 0 1)
(0 0 0 0 0)

2. Yes, there are no possible cycles.
3. 1 (1 to 2) 2 (2 to 4) 4 (4 to 5) 5
   1 (1 to 3) 3 (3 to 4) 4 (4 to 5) 5
   1 (1 to 5) 5

Table of equations for dugga 2

Statistics

Mass/density/distribution functions

$$\sum_K f(k) = 1$$

$$\int f(x)dx = 1$$

PDF: f(x) = P(X = x)
CDF: F(x) = P(X ≤ x)

Binomial distribution

$$B \sim Binom(n,p)$$

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

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

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

Normal approximation

$$\mu = np$$

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

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

Normal distribution

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

$$f_{PDF}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Linear combinations

$aX_1 + bX_2 + c$, where: $a, b \in \mathbb{R} \ \wedge \ a, b \neq 0$

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

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

$$\text{sample mean} \sim \mathcal{N}(\bar{X}, \sigma^2/\sqrt{n})$$

$$\bar{X}^* = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

Maximum likelihood estimators (MLE)

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

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

Standard normal

$$X \sim \mathcal{N}(0, 1^2)$$

$$Z = \frac{X-\mu}{\sigma}$$

$$P(X \le x) = \Phi(x)$$

$$\Phi(-x) = 1 - \Phi(x)$$

Linear Algebra

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

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

Geometry

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

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

Graph theory

$$A_{ij} = \begin{cases} 1, & [v_i \to v_j] \in E(G) \\ 0, & otherwise \end{cases}$$