UPPSALA UNIVERSITY
Fredrik Wahlberg

Dugga 2: Statistics and Linear Algebra

Pass: at least 60% correct (i.e. 24p). Pass with distinction: at least 80% correct on the two tests added.
When in doubt about the interpretation of a task, 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. While taking the exam remotely, you must be connected to the video conference room until you have handed in your exam.
Tools: Pen, handwritten notes and a calculator. You will be given probit and χ^2 tables as a part of the exam.

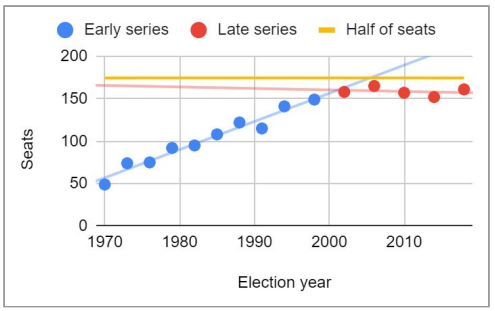
Statistics (20p)

1. In total, the swedish parliament ('riksdagen') has 349 seats. At the bottom of this question, a data set of the gender distribution in the Swedish parliament is presented, spanning from the 1970 election until today. 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 a binary definition of gender (which is what there is official statistics on), solve the following:
   (Reasonable approximations are encouraged.)
   (a) Find the distribution parameters for modelling the expected distribution of seats for women (or men) as a binomial distribution in the ideal case of 50/50. (1p)
   (b) What is the approximate range of seats, as natural numbers, for a 95% confidence interval of the distribution in (a)? *(Using a normal approximation without continuity correction is recommended.)* (3p)
   (c) Find the parameters for a normal approximation of the *proportion* of seats for the distribution in (a) (i.e. the gender distribution in parliamentary seats modelled as Bernoulli trials). *(hint: think sampling distribution)* (2p)

   For the following, the significance level is 5% (α=0.05).
   (d) For the hypotheses "H_0: p=50% (equality)" and "H_a: p≠50% (not equality)", find the first election where H_0 could *not* be refuted. (3p)
   (e) For the hypotheses "H_0: p=50% (equality)" and "H_a: p<50% ("patriarchy")", find the first election where H_0 could *not* be refuted. (3p)

Year	1970	1973	1976	1979	1982	1985	1988	1991	1994	1998	2002	2006	2010	2014	2018
Women	49	74	75	92	95	108	122	115	141	149	158	165	157	152	161

2. The trend over time is a tendency towards a 50/50 gender distribution of seats in the swedish parliament. The data from question one can be seen in the plot to the right (red and blue, with trendlines). It is clear that the data does not lie on a line and is not fit for our linear modelling. We solve this by splitting up the data into two ranges and create two linear models, one for each of these ranges, called a piecewise linear model. We can define the range spanning from 1970 to circa 2000 as the *early series* (blue), and the range circa 2000 until today as the *late series* (red). For the full piecewise linear model, we need to find the parameters of the two linear models and a transition point where the linear models cross. Below are the means and covariance matrices for the early series and the late series.



$$\bar{x}_{early} = 1984 \qquad \bar{y}_{early} = 102 \qquad \bar{x}_{late} = 2010 \qquad \bar{y}_{late} = 159$$

$$C_{early} = \begin{pmatrix} 77 & 256 \\ 256 & 881 \end{pmatrix} \qquad C_{late} = \begin{pmatrix} 32 & -6 \\ -6 & 19 \end{pmatrix}$$

(Some quotients will give you many decimals. Choose a number of significant digits and try to be consistent.)

(a) From the given data, find the least square linear model $y=\beta_0+x\beta_1$ for the *early series*. (2p)
(b) From the given data, find the least square linear model $y=\beta_0+x\beta_1$ for the *late series*. (2p)
(c) What year do the linear models from (a) and (b) cross? (2p)
(d) How much of the data variance is explained by the respective models *(the R^2 number)*? (2p)

Linear algebra (20p)

3. Given the following geometric shapes (S, P, Q, R) in $\mathbb{R}^3$, give solutions to the tasks below.
(If a number has many decimals, it can be given as a quotient in the final answer.)

S: The line $(10, -13, 8)^T+t\cdot(1, -3, 2)^T$, where $t \in \mathbb{R}$
P: The line $(-4, -1, 1)^T+s\cdot(3, 1, -1)^T$, where $s \in \mathbb{R}$
Q: The plane $6x-4y+2z-42=0$
R: A sphere with its centre at $(5, 3, -2)^T$ and radius 7

(a) Find a point where the lines S and P intersect. (4p)
(b) Find a point where the line P intersects the plane Q. (4p)
(c) Find a point on the line S that is inside the sphere R. (4p)

4. The data matrix **D,** below, contains vectors in some euclidean space $\mathbb{R}^2$, belonging to two clusters. The first three rows belong to cluster A (blue dots in the figure) and the last three rows are the vectors belonging to cluster B (red dots in the figure). There are also some useful expressions here.

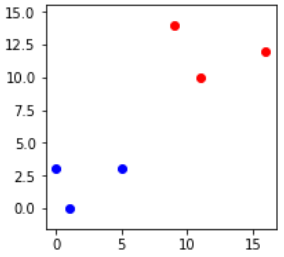
$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \\ 5 & 3 \\ 16 & 12 \\ 9 & 14 \\ 11 & 10 \end{pmatrix}$$

$$(\bar{p} - \bar{q}) \cdot \bar{n} = 0$$

$$\bar{c} = \frac{1}{n}\sum_{i=1}^{n} \bar{v}^{(i)}$$

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

$$\hat{n} = \frac{\bar{n}}{||\hat{n}||}$$



Given this data find the following:
(If a number has many decimals, it can be given as a quotient or rounded in the final answer.)

(a) Find the centroids of clusters A and B, respectively. (2p)
(b) Find a decision boundary as a plane between the two clusters (i.e. a plane between the two clusters, separating them). (2p)
(c) Find the margin of clusters A and B, given your decision boundary from (b) (i.e. the distance between the plane from (b) and the closest point for each cluster). (4p)