UPPSALA UNIVERSITY
Fredrik Wahlberg and Mats Dahllöf

Exam

This exam is split up into two parts, each representing one half of the course. If you passed dugga 1, you should skip part 1 of this exam. If you passed dugga 2, you should skip part 2 of this exam. The cut-off for a passing grade is 60% per part. For a pass with distinction, the cut-off is 80% on the two parts (or duggas) combined. 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.
Tools: Scientific calculator.
Provided tables: normal cdf table, χ^2 table and a table of equations.

Part 1: Sets, functions and probability

Set theory (10p)

1. For sets $A=\{x\in\mathbb{N} \mid 4 \leq x \leq 6\}$ and $B=\{x\in\mathbb{N} \mid 6 \leq x \leq 9\}$, show the resulting sets for the following statements. (6p)

   a. A
   b. $A \cup B$
   c. $A \cap B$
   d. $|A| + |B| - |A \cap B|$
   e. $A \times B$
   f. $A \times \varnothing$

   a. $\{4, 5, 6\}$
   b. $\{4, 5, 6, 7, 8, 9\}$
   c. $\{6\}$
   d. $3 + 4 - 1 = 6$
   e. $\{(4, 6), (4, 7), (4, 8), (4, 9), (5, 6) \ldots (6, 9)\}$
   f. $\varnothing$

2. For *any finite sets* A and B, argue for each of the following statements that it is true for all A and B, false for all A and B, or true for some A and B. (4p)

   a. $A \subseteq (A \cup B)$
   b. $(A - B) \subseteq A$
   c. $|A \cup B| = |A| + |B| - |A \cap B|$
   d. $|A - B| = |B - A|$

   a. Necessarily true for all A and B. A is always a subset of a union between itself and any B. The union operation never removes anything.
   b. Necessarily true for all A and B. (A-B) must include A or less by definition. One can not remove zero or more elements from A and get something that is larger than the original A.
   c. Necessarily true for all A and B. Can be shown with a Venn diagram or derived. Compare to the general addition rule.
   d. Can be true for disjunct A and B where $|A|=|B|$, including empty sets.

Functions (4p)

3. For the following recursive function $f:\mathbb{N}\rightarrow\mathbb{N}$, enumerate the results of f(n) where $n\in\{x\in\mathbb{N} \mid x \leq 10\}$. (4p)

$$f(n) = \begin{cases} 1 & , n = 1 \\ 1 & , n = 2 \\ f(n-1) + f(n-2) & , n > 2 \end{cases}$$

{x∈ℕ| x ≤ 10} = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
f(n) is the Fibonacci series
f(1) = 1, f(2) = 1, f(3) = 2, f(4) = 3, f(5) = 5, f(6) = 8, f(7) = 13, f(8) = 21, f(9) = 34, f(10) = 55

Probability (26p)

4. The Monty Hall problem comes from a game show that is played as follows: You are shown three closed doors by the game show host and told that behind two randomly selected doors are goats and behind the third door is a car (winner's prize). Your task is to choose a door and at the end of the game you will win the car, assuming you chose the door with the car behind it. After you have chosen a door, the game show host opens a door you have not chosen, revealing a goat. You are asked if you want to switch door, you do the switch (or not) and then the game finishes. Analyse this problem using probability theory.
   a. Define a suitable space of outcomes Ω. (2p)
   b. Define relevant events. (2p)
   c. Draw an event tree showing all possibles game progressions (with event symbols where applicable). (2p)

   Calculate the following probabilities (assume uniform probabilities for all choices):
   d. P(win | change door). (1p)
   e. P(win | not change door). (1p)

   a. Ω = {car behind door 1, car behind door 2, car behind door 3}×{choose door 1, choose door 2, choose door 3}×{change door, not change door}
   b. Several reasonable alternatives.
      Ex: "Win": A = {x∈Ω | (car behind door y **and** choose door y **and** not change door) **or** (car behind door y **and not** choose door y **and** change door))}
      B = {x∈Ω | switch door}
   c. Several alternatives (full tree or a collapsed tree).
      Ex: A tree with 3 splits followed by 3 splits and, lastly, 2 splits.
   d. P(win | change door) = $|A \cap B|/|B| = ⅔$
   e. P(win | not change door) = $|A \cap B^C|/|B^C| = ⅓$

5. Suppose that in some population group, 5% have a certain disease. A random subgroup is given a screening test with an *accuracy* of 90% (i.e. the probability that the test tells the truth). When a person gets a positive screening result, what is the probability that they actually have the disease?
   a. Derive Bayes' rule from the equation for conditional probability *(i.e. P(A|B) = P(AB)/P(B))* (3p)
   b. Define a suitable space of outcomes Ω. (1p)
   c. Define the events. (2p)
   d. Find numbers for the probabilities needed to use Bayes' rule (i.e. having the disease, getting a positive result, getting a positive result given that one has the disease etc) in terms of your chosen events (2p).
   e. What is the probability of having the disease, given a positive screening result? (2p)

   a. P(A|B) = P(AB)/P(B) ⇔ P(A|B)P(B) = P(AB) = P(B|A)P(A)
      ⇒ P(A|B) = P(B|A)P(A)/P(B)
   b. Ω = {sick, not sick}×{positive test, negative test}
   c. A (or any arbitrary character) is defined as having the disease

B (or any arbitrary character) is defined as getting a positive result

d. P(A)=0.05 *(prevalence)*
   P(B|A) = 0.9 *(accuracy)*
   $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$ *(total probability theorem)*
   $P(B|A^c) = 0.1$ *(accuracy)*
   $P(A^c) = 1 - P(A) = 0.95$
   $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0.9 \cdot 0.05 + 0.1 \cdot 0.95 = 0.14$

e. P(A|B) = P(B|A)P(A)/P(B) *(Bayes' rule)*
   $P(A|B) = P(B|A)P(A)/P(B) = 0.9*0.05/0.14 \approx 32\%$

6. In a game of dice, two players take turns in throwing two six-sided dice. The player throwing the dice wins the turn (and gets a point) if the difference between the dice is less than 3, otherwise the non-throwing player wins the turn. The game is played until one of the two players has 10 points. *Give probabilities as percentages and answer with relevant calculations/reasoning.*
   a. Define a suitable space of outcomes, Ω, for one turn in this game. (1p)
   b. Define the relevant events for one turn. (1p)
   c. Show (using a table or event tree) the number of outcomes in Ω that favour the player throwing the dice (for one turn). (2p)
   d. Calculate the probability of the throwing player or the non-throwing player winning a turn, respectively. (2p)
   e. After 10 turns (i.e. 5 turns throwing the dice for each player), what is the probability that the game has ended. (2p)

a. $T = \{1, 2, 3, 4, 5, 6\}$, $\Omega = T \times T = \{(1, 1), (1, 2), (1,3), (1, 4) \ldots (6,5), (6,6)\}$

b. The set of possible differences is {0, 1, 2, 3, 4, 5}
   $A = \{(a, b) \in \Omega \mid |a-b| \leq 2\}$ (caster wins)
   Ex: $(1, 1) \in A$, $(2, 6) \notin A$

c. Event A is marked in bold

|a - b|	1	2	3	4	5	6
1	**0**	**1**	**2**	*3*	*4*	*5*
2	**1**	**0**	**1**	**2**	*3*	*4*
3	**2**	**1**	**0**	**1**	**2**	*3*
4	*3*	**2**	**1**	**0**	**1**	**2**
5	*4*	*3*	**2**	**1**	**0**	**1**
6	*5*	*4*	*3*	**2**	**1**	**0**

d. Since all outcomes are equally likely: $P(A) = |A|/|\Omega|$
   $\Rightarrow P(A) = |A|/|\Omega| = 24/36 = \frac{2}{3} \approx 67\%$
   From axioms: $P(A)+P(A^C)=1$
   $\Rightarrow P(A^C) = 1-P(A) = \frac{1}{3} \approx 33\%$

e. One can think of this as a tree diagram with two outcomes per level. Since each turn is independent, 10 wins require five throwing wins and 5 non-throwing wins (due to the players switching roles between turns). The end criterion can be reached by either player winning. $2 \cdot P(A)^5 \cdot P(A^C)^5 = 2 \cdot (\frac{2}{3})^5 (\frac{1}{3})^5 \approx 0.11\%$

Part 2: Statistics and Linear algebra

Statistics (24p)

7. In the INSARK dataset, collected data of 20th century Swedish conscripts are reported. The histogram over heights is, approximately, shaped like a normal distribution. Assuming the distribution of heights is modelled, in centimeters, as $X \sim \mathcal{N}(179, (6.2)^2)$, find:
   a. E(X). (1p)
   b. Var(X). (1p)
   c. The three standard deviation span of heights, i.e. [μ-3σ, μ+3σ] (1p)

   For a random person in the dataset, find the probability of:
   d. Being able to reach the highest kitchen shelf, i.e. being longer than 190 cm. (2p)
   e. Being short enough to operate a vehicle with a small space for the driver, i.e. shorter than 165 cm. (2p)
   f. Being very average, i.e. a height between 175 cm and 185 cm. (2p)

   a. $E(X) = \mu = 179$
   b. $Var(X) = \sigma^2 = (6.2)^2 \approx 38.4$
   c. [μ-3σ, μ+3σ] → [179-3*6.2, 179+3*6.2] → [160.4, 197.4]
   d. $P(190 < X) = 1-\phi((190-179)/6.2) \approx 1-\phi(1.77) \approx 3.8\%$
   e. $P(X < 165) = \phi((165-179)/6.2) \approx \phi(-2.26) = 1-\phi(2.26) \approx 1.2\%$
   f. $P(175 \leq X \leq 185) = \phi((185-179)/6.2) - \phi((175-179)/6.2) \approx \phi(0.97)-\phi(-0.65) \approx 57\%$

8. Knowing if a die is loaded is tricky due to the fact that the effect of the loading hardly shows for any single throw. You suspect that a six sided die is loaded, and have an afternoon free. After throwing the die 100 times, a six has come up 23 times. You would expect a six to come up 100/6≈16.7 times. Is this result *significantly* off from a fair die? To determine this, find:
   a. The distribution parameters for modelling the trows of the die as bernoulli trials, i.e. for a binomial distributions. (1p)
   b. Not having the internet at hand, find the normal approximation for this binomial distribution, in order to simplify later calculations. *(Note that 100 throws should be considered a small number of samples.)* (1p)

   Assuming: H_0 (fair die) and H_a (loaded die giving more sixes)
   c. To get an idea of what to expect for random variation of outcomes, find a 95% confidence interval under H_0. (2p)
   d. What is the one sided p-value for refuting H_0? (2p)
   e. Are the results from part d significant at a 90%, 95% or 99% significance level, respectively? (1p)

   a. $B \sim Binom(n=100, p=1/6)$
   b. Normal approximation of the binomial distribution:
      $\mu = np \approx 16.7$, $\sigma = \sqrt{np(1-p)} \approx 3.73$
      $X \sim \mathcal{N}(16.7, (3.73)^2)$
   c. [μ-1.96σ, μ+1.96σ] → [9.4, 24]
   d. Since n is so small, we need to use continuity correction.
      $$P(B \leq x|H_0) \approx 1 - \Phi\left(\frac{x-\frac{1}{2}-\mu}{\sigma}\right)$$
      $$= 1 - \Phi\left(\frac{22.5-16.7}{3.73}\right) \approx 1 - \Phi(1.55) \approx 0.0606$$
   e. 90%: Yes
      95%: No

99%: No

9. 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 paper's appendix, the years and respective best times used in the calculations are given. Below in the table, four data points from the male data set are given.
(This is likely a hoax paper. However, we will treat it as truth for this task. Choose a number of significant digits and try to be consistent.)

x [year]	1928	1960	1984	2004
y [seconds]	10.8	10.2	9.99	9.85

   a. From the given data, find the least square linear model $y=\beta_0+x\beta_1$ for the male sprinters. (6p)
   b. Assuming that the regression model for the female sprinters is $y=-0.01681x+44.32$, at what future year do the models predict an equal outcome? (2p)

   a. The model parameter β_1 can be found from the covariance matrix of the data.
   $$C = \begin{pmatrix} s_x^2 & Rs_xs_y \\ Rs_ys_x & s_y^2 \end{pmatrix}$$
   For this, the variance of x and the covariance needs to be calculated:

Label	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
	1928	10.8	-41	0.59	1681	-24.19
	1960	10.2	-9	-0.01	81	0.09
	1984	9.99	15	-0.22	225	-3.3
	2004	9.85	35	-0.36	1225	-12.6
Averages	1969	10.21			803	-10

   $$\beta_1 = R \cdot \frac{s_y}{s_x} = \frac{Rs_xs_y}{s_x^2} = \frac{-10}{803} \approx -0.0125$$
   $$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x} = 10.21 - (-0.0125) \cdot 1969 \approx 34.7$$
   Final model: y = -0.0125x+34.7
   b. Both models will at some point in the future predict the same y:
   Given model for female sprinters: y = −0.01681x+44.32
   $$-0.01681x + 44.32 = -0.0125x + 34.7$$
   $$\Rightarrow 0.00431x = 9.62 \Rightarrow x = \frac{9.62}{0.00431} \approx 2232$$

Linear algebra (16p)

10. 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)

$\mathbf{v}_1 = (-1, -5, -2)^T$ $\quad$ $\mathbf{u}_1 = (4, 1, 8)^T$
$\mathbf{v}_2 = (0, 3, -2)^T$ $\quad$ $\mathbf{u}_2 = (3, 4, -2)^T$

   a. $-\mathbf{v}_1$
   b. $\mathbf{v}_1 + \mathbf{v}_2$
   c. $4(3\mathbf{v}_1 - 2\mathbf{v}_2)$
   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. $\mathbf{v}_1 + \mathbf{v}_2 = (-1, -2, -4)^T$
c. $4(3\mathbf{v}_1 - 2\mathbf{v}_2) = (-12, -84, -8)^T$
d. $\|\mathbf{u}_1\| = 9$
e. $\mathbf{u}_1/\|\mathbf{u}_1\| = (4/9, 1/9, 8/9)^T$
f. $\mathbf{u}_1 \cdot \mathbf{u}_2 = 4 \cdot 3+1 \cdot 4+8 \cdot -2=0$

11. 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 $(2, 2, 3)^T+t \cdot (1, -1, 2)^T$, where $t \in \mathbb{R}$
R: The line $(9, 1, 5)^T+s \cdot (1, 1, -2)^T$, where $s \in \mathbb{R}$
Q: The plane $3x+2y+4z+5=0$
S: A sphere with its centre at $(3, -1, 5)^T$ 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: $(2, 2, 3)^T+t \cdot (1, -1, 2)^T = (9, 1, 5)^T+s \cdot (1, 1, -2)^T$

$$\begin{cases} 2+t & = 9+s \\ 2-t & = 1+s \\ 3+2t & = 5-2s \end{cases}$$

Eq 1: $2+t = 9+s \Rightarrow t = 7+s$

Eq 2 and substitute t from eq. 1: $2-(7+s) = 1+s \Rightarrow s = -3$

$t = 7+s$ and $s = -3 \Rightarrow t = 4$

P and t=4: $(2, 2, 3)^T+4 \cdot (1, -1, 2)^T = (6, -2, 11)^T$

R and s=-3: $(9, 1, 5)^T+-3 \cdot (1, 1, -2)^T = (6, -2, 11)^T$

b. Substituting P into Q, where $(2, 2, 3)^T+t \cdot (1, -1, 2)^T = (x, y, z)^T$ and $3x+2y+4z+5=0$:

$$3(2+t)+2(2-t)+4(3+2t)+5 = 0 \Rightarrow 6+3t+4-2t+12+8t+5 = 0$$

$$\Rightarrow 9t = -27 \Rightarrow t = -3$$

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

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

$||\bar{p}-(3,-1,5)^T|| \leq 4$, where $\bar{p} = (2,2,3)^T + t \cdot (1,-1,2)^T$

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

$||(2,2,3)^T + t \cdot (1,-1,2)^T - (3,-1,5)^T||^2 \leq 4^2$

Expanding the euclidean distance:

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

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

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

Try t=3: $(3-1)^2+(3-3)^2+(2 \cdot 3-2)^2 = 4+0+16 \leq 16$ (inequality does *not* hold)

At least one point inside S, with t=1: $(2, 2, 3)^T+1 \cdot (1, -1, 2)^T = (3, 1, 5)^T$

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

$\mu = np$

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

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

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

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

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

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

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

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

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