UNIVERSITETET I OSLO

Matematisk Institutt

Exam in:	**STK 4900/9900 – Statistical Methods and Applications**
With:	**Nils Lid Hjort**
Auxilia:	**All printed material, also the candidate's own notes Calculator**
Time for exam:	**Friday 7/vi/2024, 15:00–19:00**

This exam set contains four exercises and comprises five pages. Write your solutions in bokmål, nynorsk, riksmål, Danish, Swedish, English, or Latin.

Exercise 1: happy birthday

How to prolong your life: well, just wait till after your birthday. There's indeed a theory that many elderly people, in societies where birthdays are seen as important, somehow manage to postpone their deaths to avoid missing one's birthday. Some years ago, a Salt Lake City newspaper reported having studied obituaries for 747 decedents, where 60 of them had died during the three-months-before-birthday time window.

(a) Discuss briefly that it might be reasonable to take the number $X = 60$ to be the outcome of a binomial $(747, p)$. Give a verbal definition of the probability p in question.

(b) Forgetting Salt Lake City for a moment, suppose X is binomial $(747, 0.25)$. Give an interval inside which X will fall with probability approximately 95 percent.

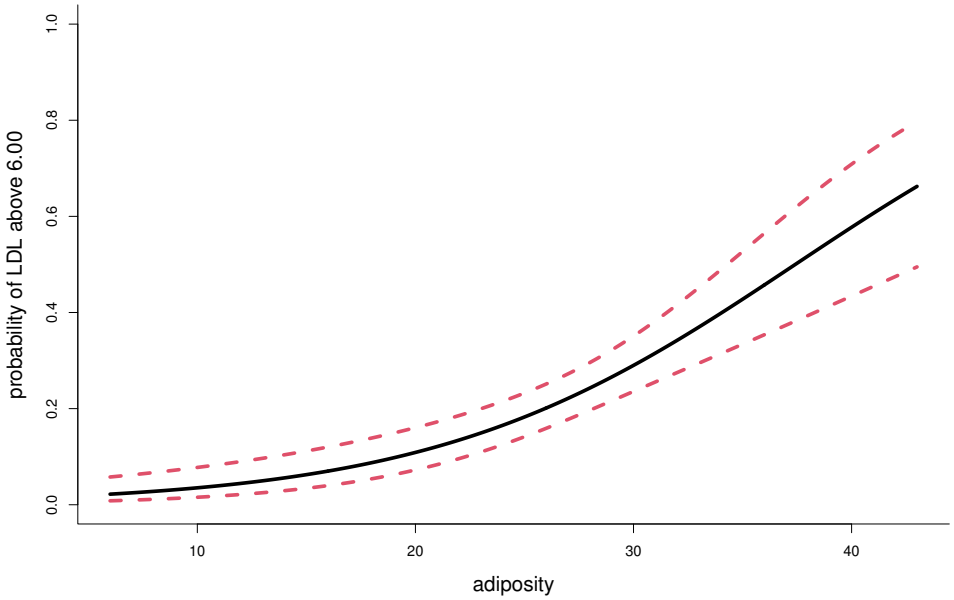
(c) Test the null hypothesis that there is no connection between people's birth dates and death dates, and formulate a conclusion, based on the Salt Lake City numbers.

Exercise 2: adiposity, LDL, and heart disease

Men menneskenes hjerter forandres aldeles intet i alle dager, says Sigrid Undset, though we sometimes try, particularly if our hearts are at risk for entering cardiovascular difficulties. I have analysed a certain dataset, pertaining to $n = 462$ South African men, with information about certain cardiovascular risk factors. The focus is on the *LDL level*, for low-density lipoprotein (also associated with so-called 'bad cholesterol'), which is recognised as a strong predictor for coronary heart problems. The LDL level for this dataset varies from 0.98 to 15.33, but we shall care here about the outcome y, which is 1 if LDL is 6.00 or more, and 0 if it is less than 6.00; the idea is that such high levels are judged as a serious threat to coronary health, perhaps with important recommendations for changes in lifestyle or medication.

The risk factors we care about here are

. x_1, adiposity (a measure of fatness, but different from e.g. the bmi);
. x_2, age (in years) at the onset of the study;
. x_3, tobacco use (average no. of cigarettes per day);
. x_4, alcohol use.



The estimated probability of having LDL level above threshold 6.00 (full curve), as a function of adiposity level, for a man of median age (45), median cigarette consumption (2 per day), and median alcohol use (7.1). The lower and upper curves correspond to pointwise 95 percent confidence.

(a) Logistic regression for the dataset, organised into a matrix of (x_1, x_2, x_3, x_4, y) for the $n = 462$ men, uses the model

$$p_i = P(\text{LDL} \geq 6.00 \,|\, x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}) = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_4 x_{4,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_4 x_{4,i})}$$

for $i = 1, \ldots, n$. I have run the `R` routine

`cordial = glm(yy ~ x1 + x2 + x3 + x4, family=binomial)`

with `summary(cordial)` giving a certain output, of which the following is an edited part (with some parts left out):

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	−3.9384	0.4761	.	.
x1	0.1266	0.0190	.	.
x2	0.0051	0.0099	.	.
x3	0.0168	0.0253	.	.
x4	−0.0047	0.0043	.	.

Which of the regression coefficients are significantly nonzero, at the 0.05 level? Write a few sentences interpreting what this means, in the present context.

(b) Meet Mr. Jones, from Cape Town, whose adiposity level is 30, of age 50, he smokes 20 cigarettes a day, with an alcohol score 25. Estimate p_{jones}, the probability that he will have an LDL level above 6.00.

(c) Explain, with sentences and formulae, how you could go about finding a 95 percent confidence interval for p_{jones} – you would however need more output from the logistic regression fitting than what is given above, to actually find the numbers. Such calculations underlie the figure above, where we see the probability of a man having an LDL level above 6.00, as a function of his adiposity level, given that he is of median age, a median smoker, and a median drinker.

(d) Of the 462 men in this study, 192 were classified as having had a family history with coronary heart disease, whereas 270 had none such prior family history. To see whether these two groups might have different types of LDL above threshold mechanisms, I examined the two groups separately, with logistic regressions, as above. This led to the following (edited) output, with estimates for regression coefficients (having approximately normal distributions) and their standard errors for the two groups. Are the regression coefficients for the two groups essentially similar, or are there significant differences for any of the risk factors?

	family history		no family history	
	Estimate	Std. Error	Estimate	Std. Error
Intercept	−3.598	0.821	−3.832	0.594
x1	0.141	0.029	0.117	0.026
x2	−0.001	0.016	0.001	0.014
x3	−0.025	0.041	0.044	0.032
x4	−0.004	0.006	−0.006	0.007

Exercise 3: getting children (and children (and children))

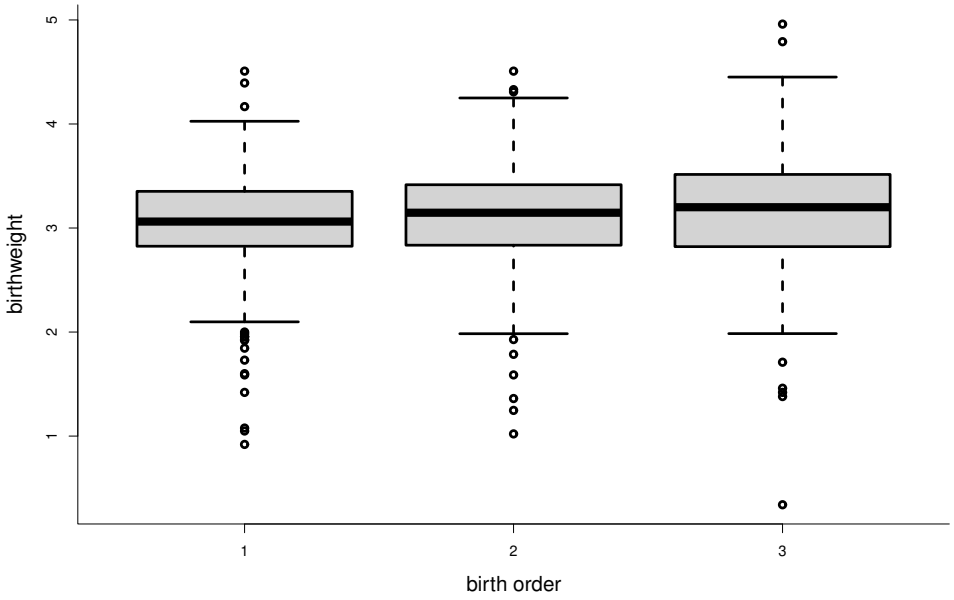
Make more children, proclaimed the Norwegian Prime Minister in her New Year's nation-wide televised speech, a few years back. Here we consider the birthweights, of children 1, 2, 3, from 199 mothers. Are they getting slightly bigger, with birth order? The figure below gives boxplots for these three connected datasets.

(a) Computing empirical means and standard deviations, for the three datasets corresponding to children 1, 2, 3, one finds (with weight being in kg)

	means	stdevs
child 1	3.027	0.557
child 2	3.103	0.551
child 3	3.144	0.602

Compute 95 percent confidence intervals for the population means μ_1, μ_2, μ_3, for the birthweights of sibling chidren 1, 2, 3. Here you may assume normal distributions for birthweights, and the 0.975 quantile of the t-distribution with degrees of freedom 198 is 1.972 (i.e. quite close to the well-known 1.960 for the normal distribution).

(b) Write $Y_{i,1}, Y_{i,2}, Y_{i,3}$ for the birthweights of children 1, 2, 3 for mother i. For this point we wish to compare birthweights for child 3 vs. child 1. Explain why the traditional t-test for comparing two normal datasets might *not* be appropriate here. We may however address the 199 differences $D_i = Y_{i,3} - Y_{i,1}$ directly. Find $\bar{D}$, their average. Their standard deviation is computed to be 0.673. Test the hypothesis that $\mu_1 = \mu_3$.



Boxplots for the birthweigths of children 1, 2, 3, in kg, born from the same 199 mothers.

(c) A natural model for the three mean parameters, for investigating whether birthweights are increasing with birth order, takes $\mu_1 = \mu$, $\mu_2 = \mu + \beta$, $\mu_3 = \mu + 2\beta$. Explain how β may be interpreted in such a model. Explain also why traditional regression models might not be well-working for making inference about β.

(d) A model taking potential sibling dependency into account is the following, for the $n = 199$ mothers and their children,

$$Y_{i,1} = \mu + M_i + \varepsilon_{i,1},$$
$$Y_{i,2} = \mu + \beta + M_i + \varepsilon_{i,2},$$
$$Y_{i,3} = \mu + 2\beta + M_i + \varepsilon_{i,3},$$

where $M_1, \ldots, M_n$ are independent with distribution $\mathrm{N}(0, \tau^2)$, and all the $\varepsilon_{i,1}, \varepsilon_{i,2}, \varepsilon_{i,3}$ are independent, among themselves and of the M_i, with $\mathrm{N}(0, \sigma^2)$ distribution. You would need more time, and `R` coding wth the actual data, to analyse the consequences of this model, but explain how the τ and σ can be interpreted. Also, show that the variance of $Y_{i,j}$ is $\tau^2 + \sigma^2$, and that the covariance between birthweights of two children from the same mother is τ^2.

(e) I have fitted the four-parameter model above, via maximum likelihood theory, yielding the following parameter estimates, with estimated standard deviations:

3.033	0.038	`mu`
0.058	0.023	`beta`
0.452	0.016	`sigma`
0.346	0.028	`tau`

(i) Is β positive? (ii) Estimate the correlation between birthweights of siblings.

Exercise 4: bad-tempered and good-tempered husbands and wives

ARE BAD-TEMPERED MEN BETTER AT FINDING GOOD-TEMPERED WOMEN than the good-tempered men are? Or, to rephrase such a delicate and intricate question, do good-tempered women in their good-temperedness have a certain tendency to penetrate the shields of even bad-tempered men? Sir Francis Galton did not merely invent fingerprinting and correlation and regression and the two-dimensional normal distribution while working on anthropology and genetics and meteorology or exploring the tropics, but had a formidable appetite for even arcane psychometrics and for actually attempting to answer half-imprecise but good questions like the above in meaningful ways – by going out in the world to observe, to note, to think, to analyse (just as his perhaps even more famous cousin did).

On an inspired day in 1887 he therefore sat down and examined interview results pertaining to 111 married couples, and classified the wives and husbands into 'bad-tempered' and 'good-tempered', reaching the following table:

		wife :	
		good-tempered	bad-tempered
husband :	good-tempered	24	27
	bad-tempered	34	26

We see this as the outcome of a four-nomial experiment, with $n = 111$ randomly sampled pairs in the relevant population of married English couples, giving counts $N_{0,0} = 24$ for $(X = 0, Y = 0)$, $N_{0,1} = 27$ for $(X = 0, Y = 1)$, $N_{1,0} = 34$ for $(X = 1, Y = 0)$, $N_{1,1} = 26$ for $(X = 1, Y = 1)$, for the four categories in question, in which X is 0 or 1 for good- or bad-tempered for the husband and Y similarly is 0 or 1 for good- og bad-tempered for the wife.

(a) So there are four probabilities $p_{i,j} = P(X = i, Y = j)$ to examine here. Define

$$a_i = p_{i,0} + p_{i,1} \quad \text{for } i = 0, 1,$$
$$b_j = p_{0,j} + p_{1,j} \quad \text{for } j = 0, 1.$$

Give interpretations for these a_i and b_j.

(b) The natural hypothesis to test is that of independence between factors X and Y. Explain that this corresponds to $p_{i,j} = a_i b_j$ for $i = 0, 1$, $j = 0, 1$.

(c) Estimate a_i and b_j from the data, and compute from these the expected numbers $E_{i,j} = n\widehat{a}_i\widehat{b}_j$, under the hypothesis of independence.

(d) Test the independence hypothesis. Should you need it, the 0.95 quantiles of the chi-squared distribution, with degrees of freedom 1, 2, 3, 4, are 3.841, 5.991, 7.815, 9.488.