BIO2310 Mathematics and statistics for bioinformatics, exam 26.2.2016

Juha Kesseli, juha.kesseli@uta.fi

Only standard writing equipment allowed, no calculators or tables. Notice that there are two pages in this question sheet. You may take the question sheet with you when you leave the exam.

1. For the following statements a-f, indicate whether the statement is true or false. You don't have to write down your justification. Each correct response is worth +1p, each incorrect response is $-\frac{1}{2}$p, and each statement with no response is 0p.
   a) Maximum is an example of a robust statistic.
   b) Vectors in a basis are always linearly independent.
   c) The higher the rank of a matrix $A \in \mathbb{R}^{n \times n}$, the higher the dimensionality of its kernel.
   d) `qpois` function in R can be used to calculate the quantiles of a Poisson distribution.
   e) A maximum likelihood estimator is always unbiased.
   f) The null hypothesis of a two-sample Kolmogorov-Smirnov test is that two sets of observations come from the same distribution.

2. a) Is the statement $\exists x \in \mathbb{R}^2 : \forall y \in \mathbb{R}^2 : x^T y = 0$ true or false? Justify your answer. (1p)
   b) Simplify $\log_g(3^h) + 2\log_g(3^{-h})$ (1p)
   c) If $P \in \mathbb{R}^{n \times n}$ is an orthogonal projection matrix and $x \in \mathbb{R}^n$ is a vector, what is $((I - P)x)^T Px$? Justify your answer. (2p)
   d) Assuming we have a set of 5 red balls and 2 green balls, and we randomly draw 3 balls without replacement, what's the probability of getting 2 red balls and 1 green ball (in any order)? (2p)

3. a) Given that we are doing a linear transformation of a random variable $X \in \mathbb{R}^n$ with mean $\mu \in \mathbb{R}^n$ and variance $\Sigma \in \mathbb{R}^{n \times n}$ into a new variable $Y \in \mathbb{R}^m$: $Y = AX + b$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, what can we know in general about the random variable Y? (2p)
   b) If
   $$A = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix},$$
   what's the rank of A? Justify your answer. (2p)
   c) If A is the same matrix as in b) and $b = [1, 1, 1]^T$, what are the solutions x to the equation $Ax = b$? (2p)

4. a) Have a look at the following R script:

```
N=50;x=rnorm(N);z=x^2;w=x^3
y=-2*x+z+.1*rnorm(50)
m=lm(y~x)
m2=lm(y~x+z)
m3=lm(y~x+z+w)
print(anova(m,m2)$P[2])
print(anova(m2,m3)$P[2])
```

What kinds of outputs would you expect from the `print` commands? Why? What is the statistical meaning of the operations being performed? How can the operations be useful in practice? (3p)

b) Have a look at the following R script:

```
N=1e5; N2=1e3
res=c()
for(i in 1:N2)
{
  res[i]=sum(c(runif(N,min=-1,max=1),rt(N,df=5)))
}
testres=shapiro.test(res)
print(testres$p.value)
```

What kind of an output would you expect from the `print` command? What is the statistical meaning of the operations being performed, i.e. what does a script like this demonstrate? How can this be useful? (3p)

5. Let's assume we are observing a population of cells growing in controlled conditions during a period of 5 hours. The population consists of two types of cells A and B. It is known that in these experimental conditions, cells of type A can be considered to grow their number exponentially, $n_A(t) = c_1 e^{c_2 t}$ with a known and fixed constant c_2. Cells of type B are known to increase their number at a constant rate, so that $n_B(t) = c_3 + c_4 t$. We are measuring the total number of cells $n = n_A + n_B$ in the population using a device which gives us an approximative output on a ratio scale. A total number of 6 measurements $y_i = y(t_i)$ are obtained at time points $t_1 = 0$, $t_2 = 1$, ..., $t_6 = 5$ (in hours). Assuming i.i.d. Gaussian error ϵ_i, $i = 1, 2, \ldots, 6$ in the measurements, write down the model matrix X such that the measurements y can be written as a linear model $y = X\beta + \epsilon$, where vector β contains the coefficients c_1, c_3 and c_4 to be estimated from measurement data y. (3p) Given this linear model formulation, how can you estimate the coefficients of interest from the data? (1p) How would the situation change if c_2 was unknown as well? How would you estimate the parameters in this case? (2p)