BIO2310 Mathematics and statistics for bioinformatics, exam 29.1.2016

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) All consistent estimators are unbiased.
   b) The null hypothesis of a two-sample Kolmogorov-Smirnov test is that the means of two distributions are equal.
   c) 95% confidence interval contains the true parameter value with 95% probability.
   d) Vectors in a basis of vector space V always span the whole vector space.
   e) Matrix $A \in \mathbb{R}^{n \times n}$ such that $A = A^T$ always has n eigenvalues.
   f) `rt` function in R can be used to generate random data from the t distribution.

2. a) Is the statement $\exists x \in \mathbb{R}^2 : \forall y \in \mathbb{R}^2 : x^T y = 1$ true or false? Justify your answer. (1p)
   b) Find $\int 2^{(2x+1)} dx$ (1p)
   c) Find the limit (if it exists) of $f : \mathbb{R} \setminus \{1\} \mapsto \mathbb{R} : f(x) = \frac{x^2+x-2}{x-1}$ when $x \to 1$. Can you define a value $f(1)$ such that f is continuous? Justify your answer. (2p)
   d) Assuming we have a set of 5 red balls and 2 green balls, and we randomly draw 3 balls with replacement, what's the probability of getting 2 red balls and 1 green ball (in any order)? (2p)

3. Have a look at the following R script:

```
W=matrix(runif(4),nrow=2)
A=matrix(0,nrow=2,ncol=100)
for(i in 1:100)
{
  A[,i]=W%*%rnorm(2)
}
print(shapiro.test(A[1,])$p.value)
print(shapiro.test(A[2,])$p.value)
```

   Describe and explain each variable defined and calculated in the script. (2p) What kind of an output would you expect from the `print` commands? (1p) What is the statistical meaning of the operations being performed, i.e. what does a script like this demonstrate? (2p) Would it be possible to replace the `for` loop with a single line of R code without the loop? If yes, what would that line be? (1p)

4. a) Given that we have a pair of random variables (X_1, X_2) with a joint density function

$$f(x_1, x_2) = \begin{cases} \lambda^2 e^{-\lambda(x_1+x_2)}, & x_1 \geq 0 \wedge x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

with $\lambda > 0$, what's the conditional density $f(x_1|x_2)$? Justify your answer.(2p)

b) If

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix},$$

what's the orthogonal complement of $col(A)$? Justify your answer. (2p)

c) If

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix},$$

what are the eigenvalues and eigenvectors of A? (2p)

5. Let's assume we are observing a population of cells growing in controlled conditions during a period of 50 hours. The population consists of type A cells of interest as well as other types of cells. It is known that in these experimental conditions, cells of type A can be considered to grow their number exponentially, $n_A(t) = c_1 e^{c_2 t}$. We are quantitating the number of cells of type A, n_A in the population using a device which gives us a sampled count of cells as the output. A total number of 51 measurements $y_i = y(t_i)$ are obtained at time points $t_1 = 0$, $t_2 = 1$, ..., $t_{51} = 50$ (in hours). Assuming that the observed counts can be modeled with a Poisson model[1] such that the intensity parameter at time t follows the exponential model and different measurements can be considered independent, write down the likelihood function $L(c_1, c_2; y)$. (4p) How could you use R to get the ML estimate of the parameters? (1p) Can we use one of the simpler estimation methods presented on the course to find comparable estimates? Why/why not? (1p)

[1] Point probability of the Poisson distribution can be obtained from the formula $f(n; \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$.