COLUMBIA UNIVERSITY
Statistical Inference and Modeling
First Exam

Prof. Marco Avella
TA: Ian Kinsella
October 22, 2019

Name:

UNI:

You have 1 hour and 10 minutes to answer the following questions. Buena suerte!

Exercise 1 (10 points)

Let $X_1, \dots, X_n$ be random sample of i.i.d. with density function $f(x;\beta) = \beta x^{-(\beta+1)}$ for $x > 1$ and $\beta > 2$.

1. Find a minimal sufficient statistic for β.
2. Show that the method of moments estimator is not a function of a sufficient statistic.
3. Compute the asymptotic distribution of the MLE of β.

1) 4 pts

$$f(x_1,\dots,x_n;\beta) = \prod_{i=1}^n \beta x_i^{-(\beta+1)} = \beta^n e^{-\beta \sum_{i=1}^n \log x_i} \cdot \left(\prod_{i=1}^n x_i\right)^{-1}$$

$$\Rightarrow T(\mathbf{X}) = \sum_{i=1}^n \log X_i \text{ is minimal sufficient}$$

2) 3 pts

$$\mu_1 = \mathbb{E}\{X\} = \int_1^\infty x \cdot \beta e^{-(\beta+1)} dx = \beta \int_1^\infty x^{-\beta} dx = \frac{\beta}{1-\beta} x^{-\beta+1} \Big|_{x=1}^{x=\infty}$$

$$= \frac{\beta}{\beta-1} \Leftrightarrow \beta = \frac{\mu_1}{\mu_1 - 1} \Rightarrow \hat{\beta}_{MM} = \frac{\bar{X}}{\bar{X}-1}$$

not a fct of $\sum_i \log X_i$

3) 3 pts

$$\log f(x;\beta) = \log \beta - (\beta+1) x$$

$$I(\beta) = -\mathbb{E}\left\{\frac{\partial^2}{\partial \beta^2} \log f(X_1;\beta)\right\} = \frac{1}{\beta^2}$$

$$\Rightarrow \sqrt{n}(\hat{\beta}_{ML} - \beta) \xrightarrow[n\to\infty]{D} \mathcal{N}(0, \beta^2)$$

Exercise 2 (10 points)

The life time X (in hours) of a device is modeled using an exponential distribution with density function $f(x;\theta) = \theta e^{-\theta x}$ for $x > 0$. In order to estimate the parameter population life time, we collect data a random i.i.d. sample $X_1, \dots, X_n$ of these devices. However, the life times of less than x_0 hours are not observed.

1. Show that the complete case mean estimator is not consistent.

2. Propose a an unbiased and consistent mean estimator.

1) Let $R_i = \mathbb{1}\{X_i \geq x_0\}$

5 pts

$$\hat{\mu}^c = \frac{\sum_{i=1}^n R_i X_i}{\sum_{i=1}^n R_i} \xrightarrow[n\to]{} \frac{\mathbb{E}[X \mathbb{1}\{X \geq x_0\}]}{\mathbb{P}(X \geq x_0)} = 1 + 1/\theta = 1 + \mu$$

Since $\mathbb{E}[X] = 1/\theta$, $\mathbb{P}(X \geq x_0) = \int_{x_0}^{\infty} \theta e^{-\theta x} dx = e^{-\theta x_0}$

and

$$\mathbb{E}[X \mathbb{1}\{X \geq x_0\}] = \int_{x_0}^{\infty} x \theta e^{-\theta x} dx = \theta \left[-\frac{1}{\theta} e^{-\theta x} x \Big|_{x_0}^{\infty} + \int_{x_0} \frac{1}{\theta} e^{-\theta x} dx \right]$$

$$= e^{-\theta x_0} - \frac{1}{\theta} e^{-\theta x} \Big|_{x_0}^{\infty} = \left(1 + \frac{1}{\theta}\right) e^{-\theta x_0}$$

2) Unbiased estimator:

5 pts

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \frac{R_i X_i}{\mathbb{P}(X_i \geq x_0)} - 1 = \frac{1}{n e^{-\theta x_0}} \sum_{i=1}^n R_i X_i - 1$$

Indeed, by the above calculations $\mathbb{E}[\hat{\mu}] = 1/\theta$

and by the LLN $\hat{\mu} \xrightarrow[n\to\infty]{P} 1/\theta$

Exercise 3 (10 points)

The table below reports daily rainfall over three years at Alofi in the Niue Island group in the Pacific Ocean. The states are 1 (no rain), 2 (up to 5mm rain), 3 (over 5mm).

Table 1: Observed counts of Alofi dataset

	1	2	3	Total
1	362	126	60	548
2	136	89	68	293
3	50	78	124	252

We will assume a first order Markov chain model for the data reported in the table.

1. Give the maximum likelihood estimate of the probability of having a three consecutive days without rain after one day with heavy rains.
2. Does a zeroth order model seem appropriate for this data? Give the likelihood ratio test statistic for testing this hypothesis and give its critical threshold rejecting the null with level 0.05.
3. Imagine a friend tells you that on average there is a 50% chance of having a rainy day in Alofi, and that the rain is mild half of the time. What is the underlying null hypothesis of your friends assertion? how would you test it?

2 pts

1) $\hat{\mathbb{P}}(X_1 = X_2 = X_3 = 1 \mid X_0 = 3) = \left(\frac{50}{252}\right)\left(\frac{362}{548}\right)^2$

4 pts

2) A quick look at Table 1 suggests that a ~~z~~zeroth order model might not be enoug since for example the prob. of no rain seems to clearly depend on the previous state.

$$\Lambda_n = 2 \sum_{r=1}^{3} \sum_{s=1}^{3} n_{rs} \log\left(\hat{p}_{rs} / \hat{p}_s\right) \xrightarrow[n \to \infty]{D} \chi^2_4$$

under H_0: $p_{rs} = p_s$ for all $r, s = \{1,2,3\}$

Threshold for rejecting the null is the 0,95 quantile of a χ^2_4 i.e the k such that $\mathbb{P}(\chi^2_4 > k) = 0,95$

4 pts

3) One could relate this to a claim about the stationary distribution. Namely, H_0: $\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = \begin{pmatrix} 0,5 \\ 0,25 \\ 0,25 \end{pmatrix}$. Could use the asymptotic normality of $\hat{P}$ to test this

Exercise 4 (10 points)

Consider an ARMA(1,1) process defined

$$Y_t = \alpha Y_{t-1} + \beta \epsilon_{t-1} + \epsilon_t, \quad t = ..., -2, -1, 0, 1, 2, ...$$

where $\{\epsilon_t\}$ is white noise with variance σ^2.

1. Suggest a condition for $\{Y_t\}$ to be stationary and give a brief intuition for it.
2. Assuming the process is stationary, compute the mean and variance of Y_t. (*Hint: Y_{t-1} and ϵ_{t-1} are correlated*).
3. Derive the coefficients c_j such that Y_t can be expressed as a linear process:

$$Y_t = \sum_{j=0}^{\infty} c_j \epsilon_{t-j}, \quad \sum_{j=0}^{\infty} |c_j| < \infty.$$

3pts

1) $|\alpha| < 1$ should guarantee stationarity since it guarantees that $\sum_{k \geq 0} \alpha^{2k} = \frac{1}{1-\alpha^2}$. Intuitively, this ensures that distant observations from the past have a small impact in the present.

4pts

2) $\underbrace{\mathbb{E}\{Y_t\}}_{\mu_t} = \alpha \underbrace{\mathbb{E}[Y_{t-1}]}_{\mu_t} + \beta \underbrace{\mathbb{E}\{\varepsilon_{t-1}\}}_{=0} + \underbrace{\mathbb{E}\{\varepsilon_t\}}_{=0} \iff \mu_t = 0$

$\underbrace{Var\{Y_t\}}_{=\gamma_t} = \alpha^2 \underbrace{Var\{Y_{t-1}\}}_{\gamma_t} + \beta^2\sigma^2 + \sigma^2 + 2\alpha\beta \underbrace{Cov(Y_{t-1}, \varepsilon_{t-1})}_{=\sigma^2}$

$$\iff \gamma_t = \frac{\sigma^2(1 + 2\alpha\beta + \beta^2)}{1-\alpha^2}$$

3pts

3) $Y_t = \alpha(\alpha Y_{t-2} + \beta\varepsilon_{t-2} + \varepsilon_{t-1}) + \beta\varepsilon_{t-1} + \varepsilon_t$

$= \alpha^2 Y_{t-2} + \alpha\beta\varepsilon_{t-2} + (\alpha+\beta)\varepsilon_{t-1} + \varepsilon_t$

$= \alpha^3 Y_{t-3} + \alpha^2\beta\varepsilon_{t-3} + \alpha(\alpha+\beta)\varepsilon_{t-2} + (\alpha+\beta)\varepsilon_{t-1} + \varepsilon_t$

$= \ldots = \alpha^k Y_{t-k} + \alpha^{k-1}\beta\varepsilon_{t-k} + (\alpha+\beta)\sum_{j=1}^{k-1}\alpha^{j-1}\varepsilon_{t-j} + \varepsilon_t$

$\xrightarrow{k\to\infty} = (\alpha+\beta)\sum_{j=1}^{\infty}\alpha^{j-1}\varepsilon_{t-j} + \varepsilon_t$

Exercise 5 (10 points)

A data analysis in R gave the following summary and plot. [1]

```
Call:
survreg(formula = Surv(t, d) ~ type, data = ncog, dist = "exp")
              Value Std. Error     z        p
(Intercept)  3.742      0.213 17.55 6.01e-69
type2       -0.325      0.285 -1.14 2.55e-01

Scale fixed at 1

Exponential distribution
Loglik(model)= -228   Loglik(intercept only)= -228.6
Chisq= 1.31 on 1 degrees of freedom, p= 0.25
Number of Newton-Raphson Iterations: 5
n= 101
```

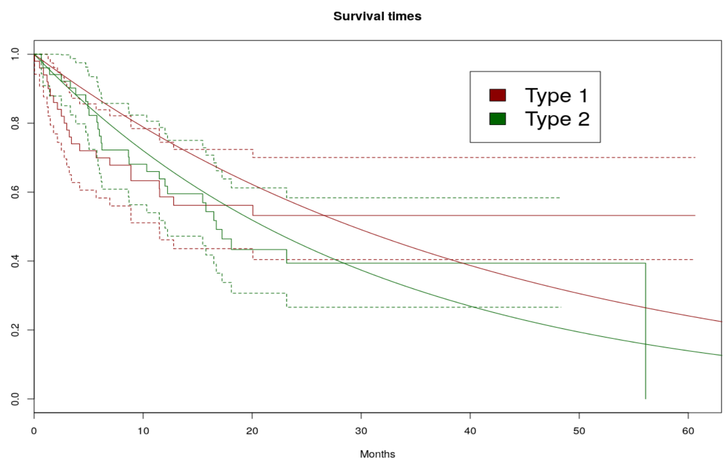

Figure 1: The exponential fits and Kaplan-Meier survival curves for each arm with 95% pointwise confidence bands.

For the following, assume a parametric model.

3 pts
1. Does the sign of the fit coefficients match what is displayed in the corresponding figure? Justify your answer.

3 pts
2. Based on the R output, give the estimated survival function for each arm.

4 pts
3. Based on the R output, does the data provide sufficient evidence to conclude that the two treatments lead to significantly different survival times? (You should reference the null hypothesis, test, and statistic used to arrive at a conclusion.)

[1] We consider data regarding survival rate of patients undergoing allogenic and autologous transplant operations. For each patient, the dataset gives time to death or relapse (in months) and the types of transplant received (1=allogeneic, 2=autologous).

Solution:

1. Yes, for the fitted models $\lambda_1 = \exp(-1 * 3.742)$ and $\lambda_2 = \exp(-1 * (3.742 - 0.325))$ we will have that $\lambda_2 > \lambda_1$. This agrees with the figure since we see that $S_1(x) > S_2(x)$ where $S_i(x) = \exp(-\lambda_i x)$ so with $\lambda_2 > \lambda_1$ this is what we would expect.

2. We are looking for the students to correctly give the coefficients $\hat{\lambda}_1, \hat{\lambda}_2$ based on the coefficients in the R output (see above) and then plug into the correct form of the survival function (see above again) to yield:

$$\hat{S}_1(x) = \exp\big[-e^{-3.742}x\big]$$
$$\hat{S}_2(x) = \exp\big[-e^{-3.417}x\big]$$

3. No, according to the likelihood ratio test with H_0: *the transplant types have the same survival function* (ie, single parameter model) we obtain a p-value of $p = .25$. This is not very significant so we cannot reject the null.