COLUMBIA UNIVERSITY
Statistical Inference and Modeling
QUIZ 2

Prof. Marco Avella
TA: Ian Kinsella, Shun Xu
February 21, 2019

---

**Name:**

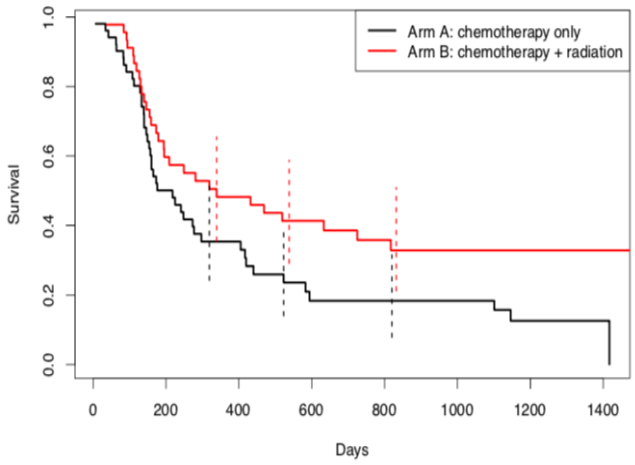
**UNI:**

---

You have 20 minutes to answer the following 10 multiple choice questions. Good luck!

## Question 1

Figure 1: Kaplan-Meier estimates of survival function for study of head/neck cancer patients. Two types of treatments led to the group A (chemotherapy only) and group B (chemotherapy + radiation). Vertical lines indicate approximate 95% confidence intervals.




From Figure 1, which of the following statements is NOT correct?

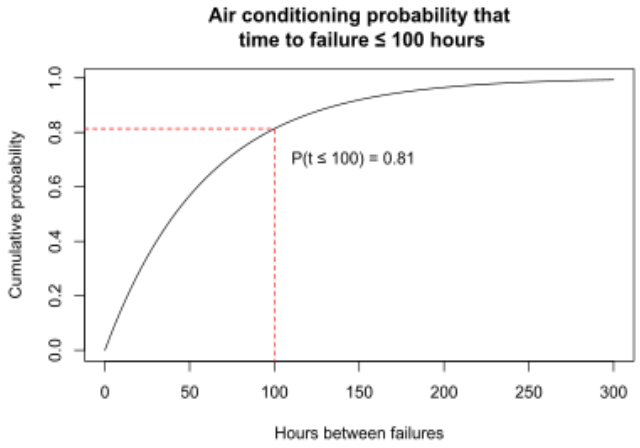
(a) It seems like most of the observed survival times are of less than a year.

(b) Arm A has a better median survival time.

(c) The chemotherapy with radiation treatment seems to be more effective.

(d) The largest survival time recorded in Arm B is right censored.

Figure 2: Estimated exponential disrtibution




**Question 2**

The curve in 2 was obtained by fitting an exponential model to some failure data. Which of the following statements are correct?

(a) The model assumes a non-constant hazard function.

(b) We know that there was censored data.

(c) We can deduce the value of the estimated parameter $\hat{\lambda}$.

(d) The estimated population median is close to $t = 100$

**Question 3**

Let $X_1, \ldots, X_n$ be i.i.d. random variables with probability density function $f(x; \lambda) = \lambda e^{-\lambda x}$ and the survival function is $S(x) = e^{-\lambda x}$ for $x > 0$. Which of the following statement is NOT correct

(a) $\sum_{i=1}^{n} X_i$ is a sufficient statistic for $\lambda$.

(b) $\bar{X}$ is the MVUE for $\lambda^{-1}$.

(c) The hazard function is $h(x) = \lambda$.

(d) The MLE of the population median is $\hat{\lambda}_{ML} \log(1/2)$.

**Question 4**

Let $X = (X_1, \ldots, X_n)$ be and i.i.d. sample of exponential random variables with population mean $\lambda^{-1}$. Which of the following is NOT correct

(a) The log likelihood function is

$$\ell(\lambda; \mathbf{x}) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

(b) $\hat{\lambda}_{ML} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

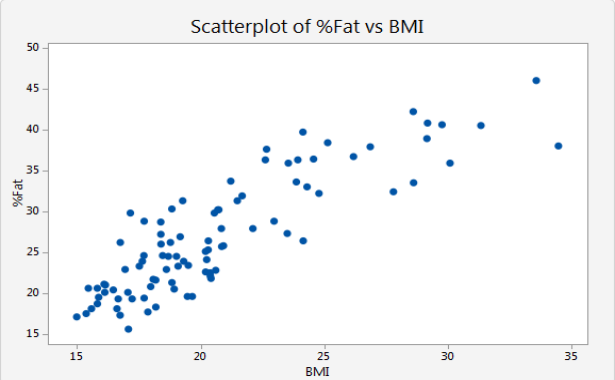
(c) The Fisher information of the whole sample is $I_n(\lambda) = n\lambda^{-2}$.

(d) $\sum_{i=1}^{n} X_i$ is a minimal sufficient statistic for $\lambda$.

## Question 5

Which of the following statements is most accurate given Figure 3?

Figure 3: Scatter plot




(a) The data definitely follows a linear pattern given the values in the axes.

(b) The slope parameter of the LS estimator will be positive.

(c) We need assumptions on error distributions in order to use LS or LAD fits.

(b) If we fit a line using LAD, we must assume a linear model with Laplace errors.

## Question 6

Assume that we fitted a straight line to the points $(x_1, y_1), \ldots, (x_n, y_n)$ using least squares and obtained the coefficients $(\hat{\alpha}, \hat{\beta}) = (0.3, 0.9)$. Which of the following sentences is NOT correct

(a) The empirical covariance between $x$ and $y$ is positive.

(b) We don't have enough information to assess the significance of the estimated parameters.

(c) The slope parameter ($\hat{\beta}$) of the LAD fit will also be positive.

(d) The least squares prediction of a response variable with $x = 2$ is $\hat{y} = 2.1$

### Question 7

A least squares fit in R gave the following output:

```
Call:
lm(formula = y ~ year)

Residuals:
    Min      1Q  Median      3Q     Max
-38.948  -9.557  -0.658   7.253  70.414

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.62105    1.74907  62.674  < 2e-16 ***
year          0.43641    0.06378   6.842 8.16e-10 ***
---
Signif. codes:
0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 17.05 on 93 degrees of freedom
Multiple R-squared:  0.3348,Adjusted R-squared:  0.3277
F-statistic: 46.82 on 1 and 93 DF,  p-value: 8.158e-10
```

Which of the following sentences is NOT correct.

(a) When year= 1 the predicted value of $y$ is about 110.

(b) The reported p-values for each coefficient (denoted $\beta$) correspond to the null-hypothesis $H_0 : \beta = 0$ and alternative-hypothesis $H_a : \beta > 0$.

(c) The response variable and the covariate year are positively correlated.

(d) We can deduce the sample size from this output.

### Question 8

Which of the following sentences is NOT correct

(a) Imputation is a bad way to handle missing data.

(b) The EM-algorithm works for normal data.

(c) When data is missing at random, complete case analysis can be inconsistent.

(d) Estimating consistently a population mean is an easy task with missing data when that data is assumed to be missing completely at random.

**Question 9**

Assume that we have a sample of size $n$ where we either observe i.i.d. univariate random variables $Y_i$ or missing values `N.A`. Let $R_i = 1$ if $Y_i$ is observed and 0 otherwise. If the assume that the unobserved data is missing completely at random, which of the following is NOT correct.

(a) $\frac{\sum_{i=1} R_i Y_i^2}{\sum_{i=1}^n R_i}$ is an unbiased estimator of $\mathbb{E}[Y^2]$.

(b) We need auxiliary information to estimate $\text{var}(Y)$.

(c) We cannot compute $\frac{1}{n}\sum_{i=1}^n Y_i$.

(d) $\frac{\sum_{i=1}^n R_i Y_i^2}{\sum_{i=1}^n R_i}$ is a consistent estimator of $\mathbb{E}[Y^2]$.

**Question 10**

Let $X_1, \ldots, X_n$ be an i.i.d. sample with common cumulative distribution function $F$ i.e. $\mathbb{P}(X_1 \leq x) = F(x)$. Let $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \leq x}$ be the empirical CDF. Which of the following sentences is NOT correct.

(a) $\hat{S}_n(x) = 1 - \hat{F}_n(x)$ is an unbiased estimator of $P(X_1 > x)$.

(b) Applying the central limit theorem we see that $\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, \sigma_x^2)$, where $\sigma_x^2 = F(x)\{1 - F(x)\}$.

(c) $\lim_{t\to\infty} \hat{S}_n(t) = 1$.

(d) $n\hat{F}_n(t) \sim Bin(n, F(t))$.