Tufts University
Department of Mathematics
Spring 2022

MA 166: Statistics

Practice final exam (v1.0) [1]

16 March 2022

PLEASE READ THESE INSTRUCTIONS BEFORE ANYTHING ELSE.

This exam is closed book and closed notes. While you are taking this exam, you may not communicate or otherwise exchange information regarding this exam or related content, with any human, either in person or by electronic means, and either to give or to receive help. The work you present must be your own, and yours only. Violations of the letter or even the spirit of this rule would be considered an extremely serious breach of ethics, honor, and conduct, and I would be obliged by Tufts University to report even so much as any suspicion I might have of such a violation to the Office of the Dean of Students.

Please sign and print your name below to indicate that you are aware of the above instructions, and that you will comply with them while you are taking this exam. You must turn in this entire exam booklet, signed where indicated, as the first four pages of your completed final examination.

Name: ______________________ Signature: ______________________

YOU MAY NOW TURN THE PAGE.

THE EXAM QUESTIONS

Point values for each problem are given. You must show all your work and justify all your reasoning in order to receive credit, and also to receive partial credit. You do not have to provide formal two-column proofs for your arguments, but you do have to present them in such a way that a mathematically literate reader will understand and be convinced by your reasoning.

1. (25 points) Explain, in your own words, without using equations, what the following terms mean
   (a) Unbiasedness
   (b) Efficiency
   (c) Sufficiency
   (d) Consistency

2. (25 points) Let $x_1, \ldots, x_n$ be n independent samples of the random variable $X \geq 0$, distributed according to the pdf

$$f_X(x;\theta) = \begin{cases} \frac{2x}{\theta^2} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise,} \end{cases}$$

   where the parameter θ is unknown. Consider the null hypothesis $H_0 : \ \theta \geq \theta_0$, where θ_0 is given.
   (a) Derive the Generalized Likelihood Ratio λ for the given H_0.
   (b) Is it true or false that rejecting H_0 if $\lambda \leq \lambda^*$ is equivalent to rejecting it if $\max_j x_j$ is greater than a certain cutoff θ_c that depends on λ^*, n, and θ_0. Justify your answer.

3. (25 points) Can I do a least-squares fit to the four points $(+1, +1)$, $(-1, +1)$, $(-1, -1)$, $(+1, -1)$ to find a line $y = ax + b$ such that the sum of the squares of the vertical distances from the points to the line is minimized? If so, find the values of a and b that result and sketch the points and the line. If not, explain what goes wrong.

4. (25 points) Ten observations are drawn at random from the pdf $f_X(x) = 2(1 - x)$ for $0 \leq x \leq 1$. What is the probability that three of the observations lie in $[0, 1/4)$, three lie in $[1/4, 1/2)$, two lie in $[1/2, 3/4)$, and two lie in $[3/4, 1]$? Your answer should be exact, but you may leave it in terms of factorials and powers of integers if you wish.

POTENTIALLY USEFUL INFORMATION

You may use the following information without proof or justification.

- You can use any of the quantities mentioned below without explaining what they are, whenever you need them.
  - * $f_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ (Normal distribution)
  - * $\int_{-\infty}^{z_{1-\alpha}} dx\ f_Z(x) = \int_{z_\alpha}^{+\infty} dx\ f_Z(x) = \alpha$
  - * $f_{\chi^2_n}(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{(n/2)-1} e^{-x/2}$ (χ^2 (chi squared) distribution)
  - * $\int_0^{\chi^2_{\alpha,n}} dx\ f_{\chi^2_n}(x) = \int_{\chi^2_{1-\alpha,n}}^{+\infty} dx\ f_{\chi^2_n}(x) = \alpha$
  - * $f_{F_{m,n}}(x) = \frac{\Gamma\left(\frac{m+n}{2}\right) m^{m/2} n^{n/2} x^{(m/2)-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)(n+mx)^{(m+n)/2}}$ (Fisher distribution)
  - * $\int_0^{f_{\alpha,m,n}} dx\ f_{F_{m,n}}(x) = \int_{f_{1-\alpha,m,n}}^{+\infty} dx\ f_{F_{m,n}}(x) = \alpha$
  - * $f_{T_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$ (Student T distribution)
  - * $\int_{-\infty}^{t_{1-\alpha,n}} dx\ f_{T_n}(x) = \int_{t_{\alpha,n}}^{+\infty} dx\ f_{T_n}(x) = \alpha$
- The gamma function $\Gamma(z)$ that appears in the above is defined by $\Gamma(z) = \int_0^\infty dt\ e^{-t} t^{z-1}$. Its recurrence relation is $\Gamma(z+1) = z\Gamma(z)$. For positive integer arguments, it is related to the factorial function by $\Gamma(n+1) = n!$.
- For data $X_1 = x_1, \ldots, X_n = x_n$ sampled from a pdf $f_X(x;\theta)$, the *likelihood function* is

  $$L(\theta) = \prod_{j=1}^{n} f_X(x_j;\theta).$$

  This expression also works for discrete r.v.s if we replace $f_X(x_j;\theta)$ by $p_X(k_j;\theta)$.
- An *estimator* $\hat{\theta}$, applied to data $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, results in an *estimate* $\theta_e = \hat{\theta}(x_1, \ldots, x_n)$.
- An *estimator* $\hat{\theta}$ is defined to be *sufficient* if the likelihood function obeys the *first factorization criterion for sufficiency*,

  $$L(\theta) = f_{\hat{\theta}}(\theta_e)\ b(x_1, \ldots, x_n).$$

  for some function b. In other words, the likelihood function factors into the pdf for the estimator, times a function of the data alone.
- We proved that the *second factorization criterion for sufficiency*,

  $$L(\theta) = g\left[\hat{\theta}(x_1, \ldots, x_n)\,;\,\theta\right]\ b(x_1, \ldots, x_n),$$

is completely equivalent to the first factorization criterion for sufficiency. It states that the likelihood function factors into some function g of the estimator and the parameter, times another function b of the data alone. This is clearly a weaker requirement for sufficiency than that of the first factorization criterion, and hence often more useful.

- An estimator $\hat{\theta}_n = h(W_1, \ldots, W_n)$ is said to be *consistent* for θ if it converges in probability to θ, that is, if

$$\forall \epsilon > 0 : \lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| < \epsilon\right) = 1.$$

- The Poisson distribution with parameter λ is

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

where $k = 0, 1, 2, \ldots$.

- Bayesian estimation (formulated for a continuous r.v. W with a continuous parameter θ): Let W be a statistic dependent on a parameter θ. Call its pdf $f_W(w \,|\, \theta)$. Assume that the parameter θ is the value of a continuous random variable Θ, whose *prior distribution* is denoted $f_\Theta(\theta)$. The *posterior distribution* of Θ, given the observation $W = w$, is the quotient

$$g_\Theta(\theta \,|\, W = w) = \frac{f_W(w \,|\, \theta) f_\Theta(\theta)}{\int d\xi \; f_W(w \,|\, \xi) f_\Theta(\xi)},$$

where the region of integration for the integral in the denominator is the set of all possible Θ.

- Let $y_1, \ldots, y_n$ be a random sample of size n from a normal distribution where σ is known. Let $z = \frac{\overline{y}-\mu_0}{\sigma/\sqrt{n}}$.
  - To test $H_0: \mu = \mu_0$ versus $H_1: \mu > \mu_0$ at the α level of significance, reject H_0 if $z \geq +z_\alpha$.
  - To test $H_0: \mu = \mu_0$ versus $H_1: \mu < \mu_0$ at the α level of significance, reject H_0 if $z \leq -z_\alpha$.
  - To test $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ at the α level of significance, reject H_0 if either $z \geq +z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$.

  We often invoke the Central Limit Theorem to apply the above test for random samples that are not normally distributed, but for which n is large.

- Let $k_1, \ldots, k_n$ be a random sample of n Bernoulli random variables for which

$$0 < np_0 - 3\sqrt{np_0(1-p_0)} < np_0 + 3\sqrt{np_0(1-p_0)} < n.$$

Let $k = k_1 + \cdots + k_n$ be the number of "successes" in the n trials. Define $z = \frac{k-np_0}{\sqrt{np_0(1-p_0)}}$.

- To test $H_0: \ p = p_0$ versus $H_1: \ p > p_0$ at the α level of significance, reject H_0 if $z \geq +z_\alpha$.
- To test $H_0: \ p = p_0$ versus $H_1: \ p < p_0$ at the α level of significance, reject H_0 if $z \leq -z_\alpha$.
- To test $H_0: \ p = p_0$ versus $H_1: \ p \neq p_0$ at the α level of significance, reject H_0 if either $z \geq +z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$.

- A Type I error is that of rejecting H_0 when it is in fact true. The probability of making a Type I error is denoted by α.

- A Type II error is that of accepting H_0 when H_1 is in fact true. The probability of making a Type II error is denoted by β, and $1-\beta$ is called the *power of the test*.

- For example, if the statistic being tested is μ and if $H_0: \ \mu = \mu_0$, then a Type II error occurs if H_0 is accepted when in fact $\mu = \mu' \neq \mu_0$. A plot of $1-\beta$ versus μ' is called a *power curve*.

- If Ω is the set of all possible values of parameter(s) θ, and ω is the set of all possible values of parameter(s) θ consistent with H_0, then the *Generalized Likelihood Ratio (GLR)* is
$$\lambda = \frac{\max_\omega L(\theta)}{\max_\Omega L(\theta)}.$$
A *Generalized Likelihood Ratio Test (GLRT)* is one that rejects H_0 if $\lambda \leq \lambda^*$ for some threshold value λ^*.

- The pdf of $U = \sum_{j=1}^m Z_j^2$ where $Z_1, \ldots, Z_m$ are standard normal r.v.s is the χ^2 distribution with m degrees of freedom.

- Let $Y_1, \ldots, Y_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be its *sample standard deviation*. Then
  - S^2 and $\overline{Y}$ are independent.
  - $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^n \left(Y_j - \overline{Y}\right)^2$ has a χ^2 distribution with $n-1$ degrees of freedom.

- Let $y_1, \ldots, y_n$ be a random sample of size n from a normal distribution where σ is unknown. Let s^2 be its sample variance. Let $t = \frac{\overline{y}-\mu_0}{s/\sqrt{n}}$.
  - To test $H_0: \ \mu = \mu_0$ versus $H_1: \ \mu > \mu_0$ at the α level of significance, reject H_0 if $t \geq +t_{\alpha,n-1}$.
  - To test $H_0: \ \mu = \mu_0$ versus $H_1: \ \mu < \mu_0$ at the α level of significance, reject H_0 if $t \leq -t_{\alpha,n-1}$.
  - To test $H_0: \ \mu = \mu_0$ versus $H_1: \ \mu \neq \mu_0$ at the α level of significance, reject H_0 if either $t \geq +t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$.

- Let $y_1, \ldots, y_n$ be a random sample of size n from a normal distribution where σ is unknown. Let s^2 be its sample variance. Let $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$.

- To test $H_0: \sigma^2 = \sigma_0^2$ versus $H_1: \sigma^2 > \sigma_0^2$ at the α level of significance, reject H_0 if $\chi^2 \geq \chi^2_{1-\alpha,n-1}$.
- To test $H_0: \sigma^2 = \sigma_0^2$ versus $H_1: \sigma^2 < \sigma_0^2$ at the α level of significance, reject H_0 if $\chi^2 \leq \chi^2_{\alpha,n-1}$.
- To test $H_0: \sigma^2 = \sigma_0^2$ versus $H_1: \sigma^2 \neq \sigma_0^2$ at the α level of significance, reject H_0 if either $\chi^2 \geq \chi^2_{1-\alpha/2,n-1}$ or $\chi^2 \leq \chi^2_{\alpha/2,n-1}$.

• Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution with mean μ_X and standard deviation σ, and let $Y_1, \ldots, Y_m$ be an independent random sample of size m from a normal distribution with mean μ_Y and standard deviation σ. Let S_X^2 and S_Y^2 be the two corresponding sample variances, and define the *pooled variance*,

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Then the quantity

$$T_{n+m-2} = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

has a Student T distribution with $n+m-2$ degrees of freedom.

• If the standard deviations are not known to be the same in the situation described in the previous bullet point, then it has been shown that the statistic

$$W = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

is approximately distributed as a Student T_ν r.v., where ν is the closest integer to

$$\frac{\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)^2}{\frac{\sigma_X^4}{n^2(n-1)} + \frac{\sigma_Y^4}{m^2(m-1)}}$$

• Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be independent random samples from normal distributions with means μ_X and μ_Y, and standard deviations σ_X and σ_Y, respectively.

  - To test $H_0: \sigma_X^2 = \sigma_Y^2$ versus $H_1: \sigma_X^2 > \sigma_Y^2$ at the α level of significance, reject H_0 if $s_Y^2/s_X^2 \leq F_{\alpha,m-1n-1}$.
  - To test $H_0: \sigma_X^2 = \sigma_Y^2$ versus $H_1: \sigma_X^2 < \sigma_Y^2$ at the α level of significance, reject H_0 if $s_Y^2/s_X^2 \geq F_{1-\alpha,m-1,n-1}$.
  - To test $H_0: \sigma_X^2 = \sigma_Y^2$ versus $H_1: \sigma_X^2 \neq \sigma_Y^2$ at the α level of significance, reject H_0 if either $s_Y^2/s_X^2 \leq F_{\alpha/2,n-1}$ or $s_Y^2/s_X^2 \geq F_{1-\alpha/2,m-1,n-1}$.

- Let X_j, where $j = 1, \ldots, t$, denote the number of times that the outcome r_j occurs in a series of n independent trials, where $p_j = P(r_j)$. Then the vector $(X_1, \ldots, X_t)$ has a multinomial distribution and

$$p_{X_1,\ldots,X_t}(k_1, \ldots, k_t) = P\left(X_1 = k_1, \ldots, X_t = k_t\right) = \frac{n!}{k_1! \cdots k_t!} p_1^{k_1} \cdots p_t^{k_t},$$

where $k_j = 0, \ldots, n$, where $j = 1, \ldots, t$, and where $\sum_{j=1}^{t} k_j = n$. Moreover, the marginal distribution of X_ℓ is the binomial pdf with parameters n and p_ℓ.

- Let $r_1, \ldots, r_t$ be the set of possible outcomes (or ranges of outcomes) associated with each of the n independent trials, where $p_j = P(r_j)$, and $j = 1, \ldots, t$. Let X_j be the number of times r_j occurs. Then

  - The random variable
  $$D = \sum_{j=1}^{t} \frac{(X_j - np_j)^2}{np_j}$$
  has approximately a χ^2 distribution with $t - 1$ degrees of freedom. For the approximation to be adequate, the t classes should be defined so that $np_j \geq 5$ for all $j = 1, \ldots, t$.

  - Let $k_1, \ldots, k_t$ be the observed frequencies for the outcomes $r_1, \ldots, r_t$, respectively, and let $np_{1_0}, \ldots, np_{t_0}$ be the corresponding expected frequencies, based on H_0. At the α level of significance, $H_0: \ f_Y(y) = f_0(y)$ is rejected if
  $$d = \sum_{j=1}^{t} \frac{(k_j - np_{j_0})^2}{np_{j_0}} \geq \chi^2_{1-\alpha, t-1},$$
  where $np_{j_0} \geq 5$ for all j.

- In reference to the last bullet point, if the s of the t parameters are unknown you may replace the corresponding p_j with their maximum likelihood estimators $\hat{p}_j$ in the expression for d, but then you should expect D to be distributed with $t - 1 - s$ degrees of freedom.

- Given points $(x_1, y_1), \ldots, (x_n, y_n)$, the straight line $y = a + bx$ minimizing the sum of the squares of the vertical distances between the points and the line is determined by

$$\begin{aligned} b =& \frac{\frac{1}{n}\sum_j^n x_j y_j - \left(\frac{1}{n}\sum_j^n x_j\right)\left(\frac{1}{n}\sum_j^n y_j\right)}{\frac{1}{n}\sum_j^n x_j^2 - \left(\frac{1}{n}\sum_j^n x_j\right)^2} \\ a =& \overline{y} - b\overline{x}. \end{aligned}$$

- The *Simple Linear Model* is a statistical model for the points $(x_1, Y_1), \ldots, (x_n, Y_n)$ (where the Y_j are now r.v.s), with the assumptions

  - $f_{Y|x}(y)$ is a normal pdf for all x.

- The standard deviation σ associated with $f_{Y|x}(y)$ is independent of x.
- The means of all the conditional Y are collinear, so

$$y = E(Y|x) = \beta_0 + \beta_1 x.$$

- All of the conditional distributions represent independent random variables.

- Maximum likelihood estimation for the Simple Linear Model yields expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ that are identical in form to those given above for a and b, respectively, except with all of the y_j replaced by Y_j. It also yields

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j\right)^2.$$

- For the Simple Linear Model described above,
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed.
  - $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$.
  - $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_j^n (x_j - \overline{x})^2}$
  - $\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_j^n (x_j - \overline{x})^2}\right]$
  - $\hat{\beta}_1$, $\overline{Y}$ and $\hat{\sigma}^2$ are mutually independent.
  - $\frac{n\hat{\sigma}^2}{\sigma^2}$ has a χ^2 distribution with $n-2$ degrees of freedom.
  - $S^2 = \frac{n}{n-2}\hat{\sigma}^2$ is an unbiased estimator for σ^2.

- The covariance of X and Y is

$$\mathrm{Cov}(X, Y) = E((X - E(X))(Y - E(Y))).$$

- The correlation of X and Y is

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_Y \sigma_Y}.$$

- The bivariate normal distribution is

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{1-\rho^2}\right)\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}.$$

- Let $y_1, \ldots, y_n$ be a random sample of size n from any continuous distribution having median $\tilde{\mu}$, where $n \geq 10$. Let k denote the number of y_j's greater than $\tilde{\mu}_0$, and let $z = \frac{k - n/2}{\sqrt{n/4}}$.

- To test $H_0: \tilde{\mu} = \tilde{\mu}_0$ versus $H_1: \tilde{\mu} > \tilde{\mu}_0$ at the α level of significance, reject H_0 if $z \geq +z_\alpha$.
- To test $H_0: \tilde{\mu} = \tilde{\mu}_0$ versus $H_1: \tilde{\mu} < \tilde{\mu}_0$ at the α level of significance, reject H_0 if $z \leq -z_\alpha$.
- To test $H_0: \tilde{\mu} = \tilde{\mu}_0$ versus $H_1: \tilde{\mu} \neq \tilde{\mu}_0$ at the α level of significance, reject H_0 if either $z \geq +z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$.

- Let $y_1, \ldots, y_n$ be a set of independent observations drawn, respectively, from the continuous and symmetric (but not necessarily identical) pdfs $f_{Y_j}(y)$, where $j = 1, \ldots, n$. Suppose that each of the $f_{Y_j}(y)$'s has the same mean μ. If $H_0: \mu = \mu_0$ is true, the pdf of the data's signed rank statistic, $p_W(w)$ is given by $p_W(w) = P(W = w) = 2^{-n}c(w)$ where $c(w)$ is the coefficient of e^{wt} in the expansion of

$$\prod_{j=1}^{n} \left(1 + e^{jt}\right)$$