STAT 135 Spring 2015, Final Exam

May 12, 2015

Name: ________________ **SID:** ________________

Person on left: ________________ Person on right: ________________

- If you are stuck in one question of a problem, you can move to the following questions.
- Partial credit will be given.
- Try to answer short and to the point.
- Good luck!

Please sign the UC Berkeley Honor Code:

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Signature:

Problem 1 [33%]

The following are reasoning questions that require very little derivations.

1. [2 pts] Explain the notion of *regression to the mean.* Plot a population (x_i, y_i) that exhibits regression to the mean, and another that does not.

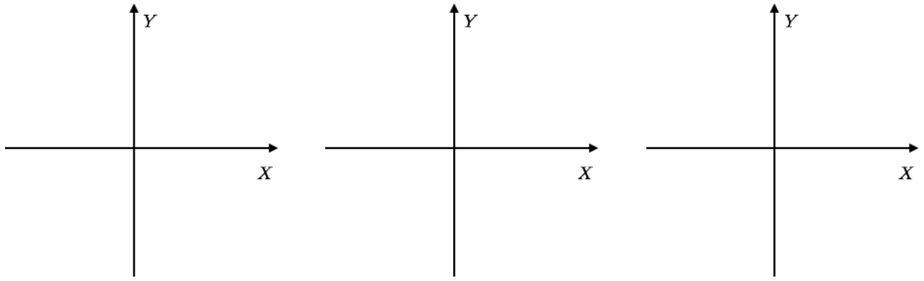
2. [2 pts] Suppose you observe (x_i, y_i), with $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$, for $i = 1 \ldots n$. Denote by $\hat{\beta}$ the estimated linear regression coefficients, by $\hat{y_i} = x_i\hat{\beta}$ the predicted values of the response, and by $e_i = y_i - \hat{y_i}$ the residuals. Can you predict the residuals e_i from x_i using linear regression? (i.e. replacing (x_i, y_i) by (x_i, e_i).) Explain.

3. [2 pts] Find a function f such that $\mathbf{E}(f(X)) = f(\mathbf{E}(X))$ but $\mathrm{var}(f(X)) \neq f(\mathrm{var}(X))$. Find a function g such that $\mathbf{E}(g(X)) \neq g(\mathbf{E}(X))$.

4. [2 pts] When would you use a t-test to compare equality of means as opposed to a rank-sum test? How would you verify that the t-test is “legitimate” ?

5. [2 pts] Scientists A and B obtain independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ respectively, where both $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. They estimate (μ, σ^2) with $(\overline{X}, s_X^2)$ and $(\overline{Y}, s_Y^2)$ respectively. If $|\mu - \overline{X}| < |\mu - \overline{Y}|$ (thus A has a better estimate of μ than B), which scientist is more likely to have a better estimate for σ^2 ? Explain.

6. [2 pts] Draw in the axis below scatter-plots of samples (x_i, y_i) from X and Y, such that (i) X and Y are positively correlated, (ii) X and Y are independent (and thus uncorrelated), and (iii) X and Y are uncorrelated but **not** independent.



(i) *(ii)* *(iii)*

7. [2 pts] We observe $X \sim \mathrm{U}[0, \theta]$ and we want to test whether

$$H_0 \, : \, \theta = 1 \quad \text{or} \quad H_1 \, : \, \theta = 4 \; .$$

Design a test with significance 0. What is the largest power you can get?
Design a test with power 1. What is the smallest significance you can get?

8. [2 pts] If $X_1, \ldots, X_n$ is a sample from a density f_θ with θ unknown, what is the meaning of a 95% confidence interval for θ? In the case where X_i are iid $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known and μ unknown, find how large n needs to be (in terms of σ^2) in order to have a 95% confidence interval for μ of length $\leq \epsilon > 0$. *(Hint: If $Z \sim \mathcal{N}(0, 1)$, then $P(Z > 1.96) = 0.025$.)*

Problem 2 [33%]

The two most popular Macaron flavors at the famous french patisserie *La Durée* are Raspberry and Lemon. The number of Raspberry and Lemon macarons sold in a day can be modeled as a Poisson distributions with parameter λ_R and λ_L respectively. We are interested in studying whether these parameters are the same or not. For that purpose, we gather samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ of the number of raspberry and lemon macarons sold during a period of n days, respectively.

1. [1 pt] Obtain the Maximum-Likelihood Estimate $\hat{\lambda}_{MLE}$ of a Poisson distribution from n iid samples $X_1, \dots, X_n$.

2. [2 pt] Suppose that a prior distribution for λ is given by $\lambda \sim \mathbf{E}(a)$ with fixed a. Show that the *Bayes Estimate*, defined as the mean of the posterior distribution, is given by

$$\hat{\lambda}_B = \int \lambda p(\lambda | x_1, \dots, x_n) d\lambda = \frac{n\overline{X} + 1}{n + a} \ .$$

   Hint: Use the fact that the posterior distribution becomes a Gamma distribution with appropriate shape and rate parameters, with pdf $f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, *and that if* $Z \sim$ *Gamma*(α, β), *then* $\mathbf{E}(Z) = \frac{\alpha}{\beta}$.

3. [3 pt] Write down the generalized likelihood ratio Λ for testing

$$H_0 : \lambda_R = \lambda_L \ , \text{ against } \ H_1 : \lambda_R \neq \lambda_L \ ,$$

as a function of $\nu = \frac{\overline{X}}{\overline{X}+\overline{Y}}$ and $T = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$, to show that

$$\Lambda = \left(\nu^\nu + (1-\nu)^{(1-\nu)}\right)^{-T} \ .$$

4. [2 pt] The function $H(\alpha) = -\alpha \log(\alpha) - (1-\alpha)\log(1-\alpha)$, $\alpha \in (0,1)$ is called the entropy and is plotted in Figure 1. Use the plot to show that an acceptance region for H_0 is given by

$$\left\{(\overline{X}+\overline{Y})(\frac{1}{2}-\epsilon) \leq \overline{X} \leq (\overline{X}+\overline{Y})(\frac{1}{2}+\epsilon)\right\}$$

for a certain value of $\epsilon > 0$.

5. [3 pt] Use the fact that $\overline{X}$ and $\overline{Y}$ are approximately Gaussian when n is large to derive a t-test with 5% significance level. Show that an alternative test when n is large is given by

$$|\overline{X} - \overline{Y}| > \frac{\sqrt{\overline{X} - \overline{Y}}}{\sqrt{n}} 1.96 \ .$$

6. [2 pt] Finally, suppose we only record for each day whether we have sold between 0 and 100 units, between 101 and 200 units or more than 200 units for each flavor. How would you test the previous hypothesis? (no need to derive the expressions). At the same significance level, do you think you would get smaller, equal or larger power? Explain.

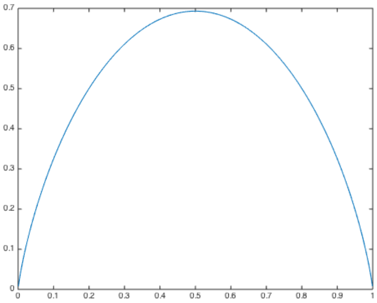


Figure 1: Entropy Function

Problem 3 [33%]

It is year 2199. A giant telescope has detected a potentially fatal meteorite approaching Earth. However, not all scientists agree on the path the meteorite will follow: one group believes in Theory A, that predicts the meteorite will avoid Earth, whereas Theory B believes the meteorite will hit the Earth, unless we employ a latest generation laser to destroy the meteorite. However, triggering the laser would create a huge hole in the ozone layer, so it should be used only as a weapon of last resort.

You, the top scientific adviser of the President, have to make a decision. You would like to know as soon as possible which trajectory the meteorite is actually following, and therefore if the laser needs to be triggered. According to Theory A, the distance of the meteorite to Earth at day n should be d_n^A , whereas according to Theory B, it should be d_n^B. The telescope produces noisy measurements each day x_n, which can be modeled as **independent** Normal $\sim \mathcal{N}(\mu_n, \sigma^2)$, where μ_n is the position of the meteorite, which is d_n^A according to Theory A or d_n^B according to Theory B.

1. [2 pt] Formulate the problem in terms of Null and Alternative Hypothesis. Think about what sort of error (Type-I or Type-II) would be more catastrophic.

2. [2 pt] At day n, the President asks you to give him a test plan that will guarantee the survival of the planet with probability $1-\alpha$ (α here would be much smaller than 0.05!), without having to activate the laser. Show that a good option is to test according to the quantity

$$K_n = \sum_{i=1}^{n} x_i (d_i^A - d_i^B) \ .$$

Can you reassure him that this is the best possible test, given the amount of data available?

3. [2 pt] Show that, under Theory B, $K_n \sim \mathcal{N}(\gamma_n, \beta_n)$, with $\gamma_n = \sum_{i \leq n} d_i^B (d_i^A - d_i^B)$ and $\beta_n = \sigma^2 \sum_{i \leq n} (d_i^A - d_i^B)^2$, and that under Theory A, $K_n \sim \mathcal{N}(\tilde{\gamma_n}, \beta_n)$, with $\tilde{\gamma_n} = \sum_{i \leq n} d_i^A (d_i^A - d_i^B)$. Specify the Rejection Region for K_n at a given significance level α. What is the power of the test and how does it (roughly) behave as n increases? *(Hint: Observe that* $|\gamma_n - \tilde{\gamma_n}| = \beta_n / \sigma^2$. *Perhaps drawing a picture will help.)*

4. [2 pt] You try to tell the President that a premature decision might be too conservative, and that, despite the panic, it is better to wait a few days to gather more evidence in favor of using the laser. Can you explain this intuition in statistical terms?

Useful Formulas

- If $X \sim \text{Poisson}(\lambda)$ then

  $$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \ , \ k = 0, 1, 2 \ldots \ .$$

  We have

  $$\mathbf{E}\,(X) = \lambda \ , \ \text{var}\,(X) = \lambda \ .$$

- The Fisher Information of a parameter θ with respect to a random variable X with density f_θ and parameter $\theta = \theta_0$ is

  $$\mathrm{I}(\theta) = \mathbf{E}\left(\left(\left. \frac{\partial \log f_\theta(X)}{\partial \theta} \right|_{\theta=\theta_0} \right)^2 \right) = -\mathbf{E}\left(\left. \frac{\partial^2 \log f_\theta(X)}{\partial \theta^2} \right|_{\theta=\theta_0} \right) \ .$$

  If $\beta = g(\theta)$ is another parameter and g is invertible and smooth, then $\mathrm{I}(\beta) = \frac{\mathrm{I}(\theta)}{|g'(\theta)|^2}$.

- The Central Limit Theorem says that if $X_1, \ldots, X_n$ are iid random variables with $\mathbf{E}\,(X) = \mu$ and $\text{var}\,(X) = \sigma^2 < \infty$, then

  $$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, 1) \ .$$

- The pdf of a normal distribution with mean μ and variance σ^2 is

  $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \ .$$