LAST NAME (Please Print): KEY

FIRST NAME (Please Print): _____

HONOR PLEDGE (Please Sign): _____

Statistics 111

Midterm 2

- This is a closed book exam.
- You may use your calculator and a single page of notes.
- The room is crowded. Please be careful to look only at your own exam. Try to sit one seat apart; the proctors may ask you to randomize your seating a bit.
- Report all numerical answers to at least two correct decimal places or (when appropriate) write them as a fraction.
- All question parts count for 1 point.

Some possibly useful formulæ:

The Gamma(α, β) distribution has mean α/β, variance α/β^2, and density function

$$f(x) = \frac{\beta^\alpha}{(\alpha-1)!} x^{\alpha-1} \exp(-\beta x) \text{ for } 0 \le x \text{ and } \alpha > 0, \beta > 0.$$

The Beta(α, β) has mean $\alpha/(\alpha+\beta)$, variance $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$, and density function

$$f(x) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0 \le x \le 1 \text{ and } \alpha > 0, \beta > 0.$$

1. Assume the number of phone calls you receive in an hour has a Poisson distribution with parameter λ. Let X be the number of calls you receive between noon and 1 p.m., and let Y be the number of calls you receive between noon and 3:00 p.m.

   $1/\sqrt{3}$ or 0.58 What is the correlation between X and Y?

   Here Y is Pois(3λ). Write $Z = Y - X$. Then Z is Pois(2λ) and independent of X. Since

   $$\text{Cov}(X, Y) = \text{Cov}(X, X + Z) = \mathbb{E}[X(X + Z)] - \mathbb{E}[X] * \mathbb{E}[X + Z]$$

   and using properties of expectations, this simplifies to

   $$\mathbb{E}[X^2] + \mathbb{E}[X] * \mathbb{E}[Z] - \mathbb{E}[X] * (\mathbb{E}[X] + \mathbb{E}[Z]).$$

   Since

   $$\lambda = \text{Var}\,[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - \lambda^2$$

   then

   $$\text{Cov}(X, Y) = \lambda + \lambda^2 + \lambda * 2\lambda - \lambda(\lambda + 2\lambda) = \lambda$$

   and so

   $$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}\,[X] * \text{Var}\,[Y]}} = \frac{\lambda}{\sqrt{\lambda * 3\lambda}} = 1/\sqrt{3}.$$

2. You own 2 shares of Apple, 3 shares of Google, and 4 shares of Alcoa. The annual return on investment for an Apple share is normally distributed with mean 1 and sd 2, or N(1,2). The annual share-wise ROI for Google is N(2,3) and for Alcoa is N(3,4). Since Apple and Google are in the same sector, their ROIs are have correlation r.

   3 If $r = 0.5$, what is the covariance in shares of Apple and Google?

   $0.5 = \frac{\text{Cov}(X,Y)}{\sqrt{2^2 * 3^2}}$ so 3.

   20 Suppose the covariance between Apple and Google is 2. What is your expected total ROI?

   2*1 + 3*2 + 4*3 = 20.

19.41 Suppose the covariance between Apple and Google is 2. What is the sd on your total ROI?

The variance is $2^2 * 2^2 + 3^2 * 3^2 + 4^2 * 4^2 + 2 * 2 * 3 * 2 = 377$. So the sd is 19.416.

0.59 What is the probability that your total ROI exceeds 15.8?

Linear combinations of normal random variables, so the total ROI is N(20, 19.416). The z-transformation gives -0.216, and the table finds this probability as 0.587.

3. Suppose $x_1, \ldots, x_n$ are a random sample from an exponential distribution with parameter λ. You do not know λ but you have a prior distribution for it. Your prior is also exponential, with parameter 2.

What is your posterior distribution for λ? Gamma$(n + 1, 2 + \sum x_i)$

The formula is

$$\pi^*(\lambda|x_1, \ldots, x_n) = \frac{\text{likelihood } * \text{ prior}}{\text{something that does not depend on } \lambda}$$

where the prior is

$$\pi(\lambda) = 2\exp(2\lambda) \text{ for } 0 < \lambda < \infty$$

and the likelihood is

$$\prod_{i=1}^{n} f(x_i; \lambda) = \lambda^n \exp(-\lambda \sum x_i).$$

When one writes the product in the numerator, the terms that involve λ have the same form as in the Gamma distribution. And since the posterior must integrate to 1, this means that the numerator integral gives exactly that are needed to ensure this, which are the values in the Gamma density.

Under what circumstances would the median of your posterior distribution be a good guess?

When the penalty I pay for being wrong about λ is proportional to the absolute value of my error.

4. Let 0.8 and 0.75 be a random sample from the Beta distribution with unknown α. However, you know that $\beta = 1$.

   3.92 What is the maximum likelihood estimate (MLE) of α?

   The density for one observation is

$$f(x) = \frac{\alpha!}{(\alpha-1)!}x^{\alpha-1}(1-x)^0 = \alpha x^{\alpha-1}$$

   so the likelihood is

$$\alpha^2(x_1)^{\alpha-1}(x_2)^{\alpha-1}.$$

   Taking the log gives

$$2\ln\alpha + (\alpha-1)\sum \ln x_i$$

   and taking the derivative wrt α and setting to 0 gives

$$0 = \frac{2}{\alpha} + \sum \ln x_i.$$

   Solving gives $\hat{\alpha} = -n/\sum \ln x_i = 3.9152$.

   0.80 What is the MLE of the mean of this Beta distribution?

   The mean of the Beta is $\alpha/(\alpha+\beta)$, which is a function of α. The MLE of a transformation is the transformation of the MLE, so this is $3.9152/(3.9152 + 1) = 0.7965$.

5. You believe that your probability p of answering True or False questions correctly on a statistics test has a Beta distribution with mean 0.5 and variance 0.1. When your test comes back, you see that you got only 5 of the 15 T/F questions.

   What is your new distribution for p? Beta(5.75, 10.75)

   This is the Beta-Binomial case, and the first step is to find α and β. Since the mean is $0.5 = \alpha/(\alpha+\beta)$ we solve to see $\alpha = \beta$. And since the variance is

$$0.1 = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)] = \alpha^2/[(2\alpha)^2(2\alpha+1)]$$

   Then $\alpha = (1/2)[(1/0.4 - 1)] = 0.75$ and so $\beta = 0.75$. Now, from the Beta-Binomial properties, we know the posterior is Beta(0.75 + 5, 0.75 + 10).

What is your best one-number guess about p under squared error loss? 0.35

This is the mean, or $5.75/(5.75 + 10.75) = 0.3485$.

6. Suppose you take 32 courses during your time at Duke. In each class, the grade you get is a random variable. An A+ counts for 12 points, an A counts for 11, and so forth. Assume your probability of getting a B is 0.25, your probability of a B+ is 0.2, your probability of a B- is 0.2, and all other grades are equally likely.

7.09 What grade point average do you expect?

The expected value of the GPA is just the expected value of one Grade, or

$$\mathbb{E}[\mathrm{G}] = (8)(0.25) + (9)(0.2) + (7)(0.2) + (0.35/10)[0 + 1 + \cdots + 6 + 10 + 11 + 12] = 7.09.$$

0.49 What is the standard error in your average grade?

Find

$$\mathbb{E}[\mathrm{G}^2] = (8^2)(0.25) + (9^2)(0.2) + (7^2)(0.2) + (0.035)[0^2 + \cdots + 6^2 + 10^2 + 11^2 + 12^2] = 57.96$$

and then the variance for a single course grade is $57.96 - (7.09)^2 = 7.6919$. The standard deviation is $\sigma/\sqrt{32} = \sqrt{7.6919/32} = 0.4903$.

0.20 What is the (approximate) probability that your average grade at graduation is greater than 7.5?

This is a CLT question. Your GPA is approximately normal with mean 7.09 and sd 0.4903. The z-transformation is (7.5 - 7.09)/0.4903 = 0.837. From the table, the probability is 0.2033.

7. Durham has 25 fast-food chain restaurants. You visit 10 of them, and find that the average sanitation score is 90.2 with a sample sd of 5.

87.91 Set a one-sided lower 95% confidence interval on the average sanitation score in such restaurants.

This is a FPCF situation with a t-distribution. So your L is

$$90.2 + (5/\sqrt{10})\sqrt{(25-10)/(25-1)}t_{9,0.05}$$

where the t-table has $t_{9,0.05} = -1.833$. You get L = 87.909.

8. Jed Bartlett is running for president. His team polls 100 voters in Durham, and finds that 55 plan to vote for him. Before pulling their ad funding for Durham, his two top strategists set confidence intervals on his support. Josh Lymon sets a two-sided 95% interval; Toby Zeigler sets a one-sided lower 95% interval.

For Josh, the formula is

$$(55/100) \pm \sqrt{\frac{0.55 * (1-0.55)}{100}} z_{0.025} = 0.55 \pm 0.44975 * (1.96).$$

For Toby, the formual is

$$(55/100) + \sqrt{\frac{0.55 * (1-0.55)100}{z}}_{0.05} = 0.55 - 0.44975 * (1.65 \text{ or } 1.64$$

Josh's interval: L = 0.45 U = 0.65

Toby's interval: L = 0.47

Which advisor is correct and why?

Toby is right. Bartlett doesn't need an upper bound—he only wants to know whether he has sewn up at least 50% of the vote.

9. Assume $X_1, \ldots, X_n$ are a random sample from the Gamma$(\alpha, 3)$ distribution.

$3\bar{X}$ Find an unbiased estimate of α.

For a Gamma, the expected value of $\bar{X}$ is $\alpha/\beta = \alpha/3$. So an unbiased estimator is $3\bar{X}$.

$\sqrt{\alpha/n}$ What is the standard error of your estimate?

$\text{Var}\,[3\bar{X}] = (3/n)^2 \sum \text{Var}\,[X] = (3/n)^2(n\alpha/9) = \alpha/n.$

10. You estimate the probability of Heads on a coin by tossing it n times, counting the number of Heads, and dividing by $n-1$. What is the bias, variance, and mean squared error in your estimator?

bias = $\frac{p}{n-1}$ variance = $\frac{np(1-p)}{(n-1)^2}$

mean squared error = $\frac{np(1-p)+p^2}{(n-1)^2}$

$\mathbb{E}[\frac{1}{n-1}\sum X_i] = \frac{1}{n-1}\mathbb{E}[\sum X_i] = \frac{np}{n-1}$.
$\text{Var}\,[[\frac{1}{n-1}\sum X_i] = \frac{1}{(n-1)^2}\text{Var}\,[\sum X_i] = \frac{1}{(n-1)^2} n * np(1-p)$
MSE = variance + bias^2.

11. Consider the linear congruential generator $X_n \equiv 7X_{n-1} \pmod 5$. Set the seed for the random number generator to be 11.

4 What is X_2?

$X_1 = 7 * (11) \pmod 5 = 2$, so $X_2 = 7 * (2) \pmod 5 = 4$.

12. You think that women may be more likely to vote Democrat than men. You sample 100 women in NC and find that 60 support the Democrat Kay Hagan. You sample 200 men in NC and find that 150 support Thom Tillis, the Republican. Set an approximate two-sided confidence 90% interval on the difference between the proportion of female Democrats and male Democrats in NC (subtract men from women).

L = 0.25 U = 0.45 Grader: accept $\pm$ 0.01.

$\hat{p}_w = 60/100 = 0.6$ $\hat{p}_m = 50/200 = 0.25$. So the point estimate is 0.6 - 0.25 = 0.35. The variance of the difference is the sum of the variances, or $0.6(1-0.6)/100+0.25(1-$

0.25)/200 = 0.00334. The critical value comes from a z-table and is 1.645. So pe + se * cv gives 0.25 and 0.45.

13. The Medicare program maintains a sample of 5% of its participants. When a person is selected into the sample, they are followed until their death, and Medicare tracks the services they use and the associated costs. When someone dies, or when the number of enrollees increases (as has happened each year since the program began, due to the Baby Boomers), additional people are enrolled, and these are chosen at random from all current Medicare patients not currently in the sample. In what way is the sample not representative of the population of Medicare patients?

    It will be healthier and younger. The sickest people die and are replaced. Eventually, they are replaced by someone who is unusually healthy, and probably relatively young. Once that happens, those people stay in the pool a long time, biasing the sample.

14. True or False. List only the true statments. A, B, D, E, H, I, K

    A. There was a stomach cancer hotspot centered on Pittsburgh.

    B. Response bias occurs if the wording of a survey question affects the answer.

    C. Respondent bias occurs if people who refuse to answer are unlike those who do.

    D. Holding all else constant, as your confidence level increases, the width of your confidence interval increases.

    E. Holding all else constant, as your variance increases, the width of your confidence interval increases.

    F. Holding all else constant, as your sample size increases, the width of your confidence interval increases.

    G. Mean squared error is the sum of the variance and the bias.

    H. Linear combinations of normal random variables are normally distributed.

    I. As the sample size increases, maximum likelihood estimates have minimum MSE.

    J. John Snow stopped a smallpox outbreak by removing the handle from the Broad Street pump.

    K. The population pyramid in a developing nation tends to be short and wide.