How to practice for Exam updated 11/16/2023

Description of this document:

This document is a review for the test. The form of the review is different for the probability questions and the statistics questions.

In the probability section of this review, there are actual problems, and, in the first line of each problem there are key words. Key words like this **will not** appear in the actual exam. ~~Also, on the actual exam, all answers will be multiple choice, as in homework and quiz problems.~~

Expect that most problems on the exam will require numerical answers rather than the multiple choice "closest value" answers as some homework and quiz problems have.

Remember that, if a problem has one numerical answer that depends on a previous numerical answer, it is crucial to use the actual numerical answer for the first question, correct to many decimal places, to compute the answer to the second question, correct to however many decimal places are required.

The statistics portion here is a discussion and does not include problems.

Instructions for the exam itself:

The exam has between 7 and 9 statistics questions and 4-5 probability questions, each of which counts 1 point.

The exam is open-book (you can use any book, hardcopy or electronic) and open-notes (your course notes, lecture notes from edX). A calculator or spreadsheet is ok (including R or any other statistics program on a computer.) You are expected to use StatKey and the datasets accompanying StatKey on the web, but please do not use any other on-line resources and not help from anyone else (in person or online.)

Review for probability portion:

1. *MGF*
Let $X \sim N(\mu, \sigma^2)$ be normal distributed r.v. Find the moment generating function $M_X(t)$.

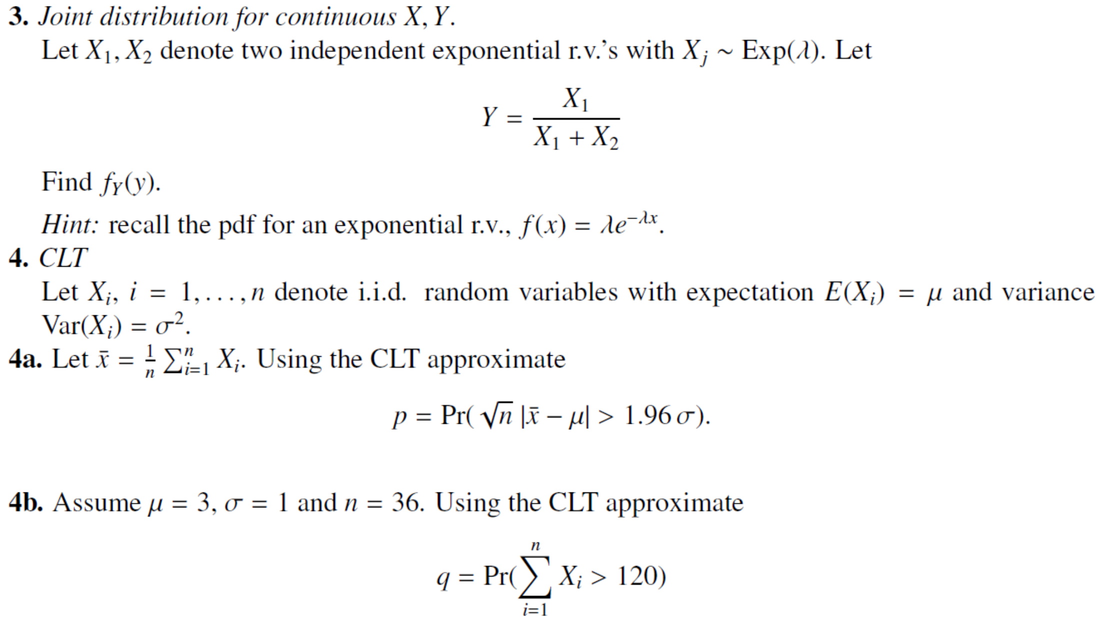
Hint: recall from lecture (proof of the CLT) the mgf for a standard normal r.v. $Z \sim N(0, 1)$ as $M_Z(t) = e^{t^2/2}$.

2. *Chernoff bound*
Let $Z \sim N(0, 1)$ be a standard normal r.v. For $a > 0$, find a Chernoff bound for

$$p = \Pr(Z \geq a).$$

2a. Find the sharpest (i.e., smallest) upper bound.
2b. Let $X \sim N(\mu, \sigma^2)$. For $a > 0$, find a Chernoff bound for

$$p_2 = \Pr(|X - \mu| \geq a)$$



3. *Joint distribution for continuous X, Y.*

Let X_1, X_2 denote two independent exponential r.v.'s with $X_j \sim \text{Exp}(\lambda)$. Let

$$Y = \frac{X_1}{X_1 + X_2}$$

Find $f_Y(y)$.

Hint: recall the pdf for an exponential r.v., $f(x) = \lambda e^{-\lambda x}$.

4. *CLT*

Let X_i, $i = 1, \ldots, n$ denote i.i.d. random variables with expectation $E(X_i) = \mu$ and variance $\text{Var}(X_i) = \sigma^2$.

4a. Let $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Using the CLT approximate

$$p = \Pr(\sqrt{n}\,|\bar{x} - \mu| > 1.96\,\sigma).$$

4b. Assume $\mu = 3$, $\sigma = 1$ and $n = 36$. Using the CLT approximate

$$q = \Pr(\sum_{i=1}^{n} X_i > 120)$$

For the statistics portion:

In general, in the statistics portion of the class, the problems on the homework, quizzes, and exams are all about the same level. So practice for the exam using the homework and similar questions.

Expect a problem on the MLE material in Week 13. The questions will be set up in the same manner as the homework questions on this material.

Expect **between 6 and 8 problems** on the material from Weeks 5, 10, 11, and 12. These include forming confidence intervals, doing hypothesis tests, and computing the required sample size for a confidence interval.

There will not be questions explicitly asking you to check conditions, determine what type of conclusion can be drawn from a particular study design, interpreting p-values, or interpreting confidence intervals. All of those topics are quite important – more important than any particular computations. But I don't find them ideal for multiple-choice questions on a relatively high-stakes exam. Thus, they aren't on this exam. Of course, when you are not told whether to use a theoretical dist'n method or a simulation method, you should always check the conditions before relying on the method from a theoretical dist'n (Before relying on any method that has conditions!)

The solution key for involving statistics or probability questions for the theoretical dist'ns given in StatKey was prepared using StatKey, as has been explained with the previous homework questions.

In this course, you learned that it is as easy to find CI and do HT for more types of problems than our theoretical dist'n methods cover. You may be tested on any of those that appear types which appear in StatKey, whether we have formulas for them or not, just as you had homework on several of them. This includes regression slope, correlation coefficient, standard deviation, and median. Look over the StatKey menu carefully to see all that you could be tested on. (This does NOT include ANOVA for Regression. That is the only thing on the StatKey menu not covered in our course.)

Within StatKey, when there is a choice of the method of resampling (testing a difference of means or a difference of proportions) the solution key is made by using the “default” method. So just don’t ever change those during the exam.

The method I use of evaluating your MLE work requires you to be skilled in computation. I strongly recommend a spreadsheet for this instead of a hand-held calculator. For this type of calculation, as well as various other calculations you will do in statistics classes, make sure you are skilled in the following:

- Find the sum of a column of numbers efficiently.
- Distinguish between the functions for $\log_e$ (natural log) and $\log_{10}$ (common log.) The only log function you should use in statistics formulas is the natural log. (That’s true of most classes that have calculus as a prerequisite, because, if we might differentiate or integrate formulas, the natural log is much more convenient.) So we just move completely away from using common log.
- Calculate the gamma function of a number. (You may not have needed to do that yet, but you are working with distributions whose pdf and/or cdf formulas include gamma functions, so it’s time to learn.)