ISQS 5347 Final Exam Fall 2010

Instructions: Closed book, closed notes, no electronic devices. Points (out of 100) in parentheses

1. The assumption of the two sample *t* test is that the Y_{ij} are independently distributed as N(μ_i, σ^2).
    a. (10) State two conclusions, one having to do with Type I error rate and the other having to do with power, that are guaranteed to be true when this assumption is true.

    Solution: If the assumptions are true, then the actual type I error rate is α (usually set at 0.05) and the pooled two sample t test is the most powerful test.

    b. (5) If the assumption is false, are the two conclusions of 1.a. false as well? Explain.

    Solution: A implies B. Truth of assumptions (the model) implies truth of conclusions (correct level and highest power). But if A is false, it is not necessarily true that B is also false. If an animal is not a horse, it still might be a mammal! It is possible that the level is correct and that the two sample t test is still most powerful among two or more competing tests, even when the assumptions are not valid.

2. The data from the “car rating” example was something like this: 3,2,2 etc. , all Likert scale (1,2,3,4,5) responses, from students who answered the question “How likely are you to buy a car in the next two years?”
    a. (5) Explain why the normality assumption is not valid.

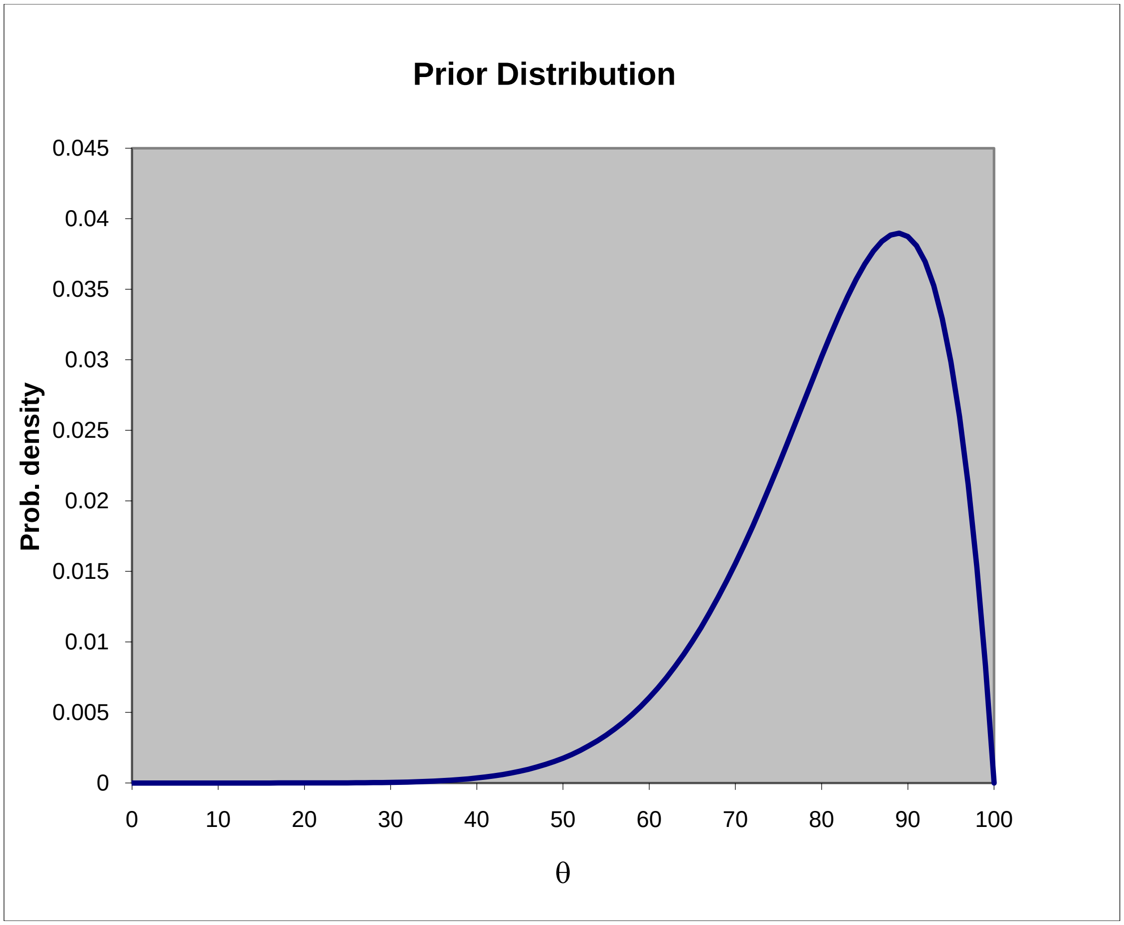
    Solution: The data are discrete, 1,2,3,4 5. Such data clearly could not have been generated from a normal distribution; if they were then there would be decimals.

    b. (5) Explain why the independence assumption might be valid.

    Solution: We can assume that people make up their minds as to how to answer the survey independently of other people. On the other hand, if people discussed the questions and decided to answer the same based on their discussion, then the independence assumption would be violated.

3. (15) The percentage of TTU graduate students receiving some type of financial assistance is θ. Putting numbers and labels on the vertical and horizontal axes, draw your prior distribution for θ, and explain why you think that way.

    Solution: Percent can range from 0 to 100. I think most students receive some kind of aid, so I will specify a distribution like this:



4. Model produces data, model has uncertain parameters, and data reduce uncertainty about unknown parameters.

   The recipe for Bayesian statistics is Posterior = C * Prior * Likelihood.

   a. (5) How does "Model produces data" relate to the likelihood function?

      Solution: The likelihood function is equal to the probability distribution function. And the data are assumed to be produced by (or sampled from) the probability distribution function.

   b. (5) How does "Model has uncertain parameters" relate to the prior distribution?

      Solution: Before the study, you do not know the values of the parameters, but you have some ideas. Your uncertainty about the parameters before you see the data is expressed via your prior distribution.

   c. (5) How does "Data reduce uncertainty about unknown parameters" relate to the posterior distribution?

Solution: After the study, you have a better idea about where the parameters lie. The posterior distribution shows your understanding of where the parameters lie after collecting data. With a large sample size, the posterior becomes narrower and narrower, reflecting greater reduction in uncertainty with greater sample size.

5. (10) You wish to select sample sizes for a study where the data will be analyzed using the pooled two-sample *t* test. Explain how you will select those sample sizes.

   Solution: You need to specify (i) the true standard deviation (assumed common for the two groups), (ii) a minimal difference between true means that you would like to call "statistically significant," (iii) a power figure (80% is default) and (iv) a significance level (5% is default). Given these inputs you can find the n's using Russ Lenth's JAVA page, or by using SAS with simulations as we used in class.

6. Consider the following distribution.

x	p(x)
0	.80
1	.20

Let X~p(x) denoted a randomly sampled response from this distribution.

A. (5) Calculate E(X).

Solution: E(X) = 0*.80 + 1*.20 = .20.

B. (5) Calculate Var(X).

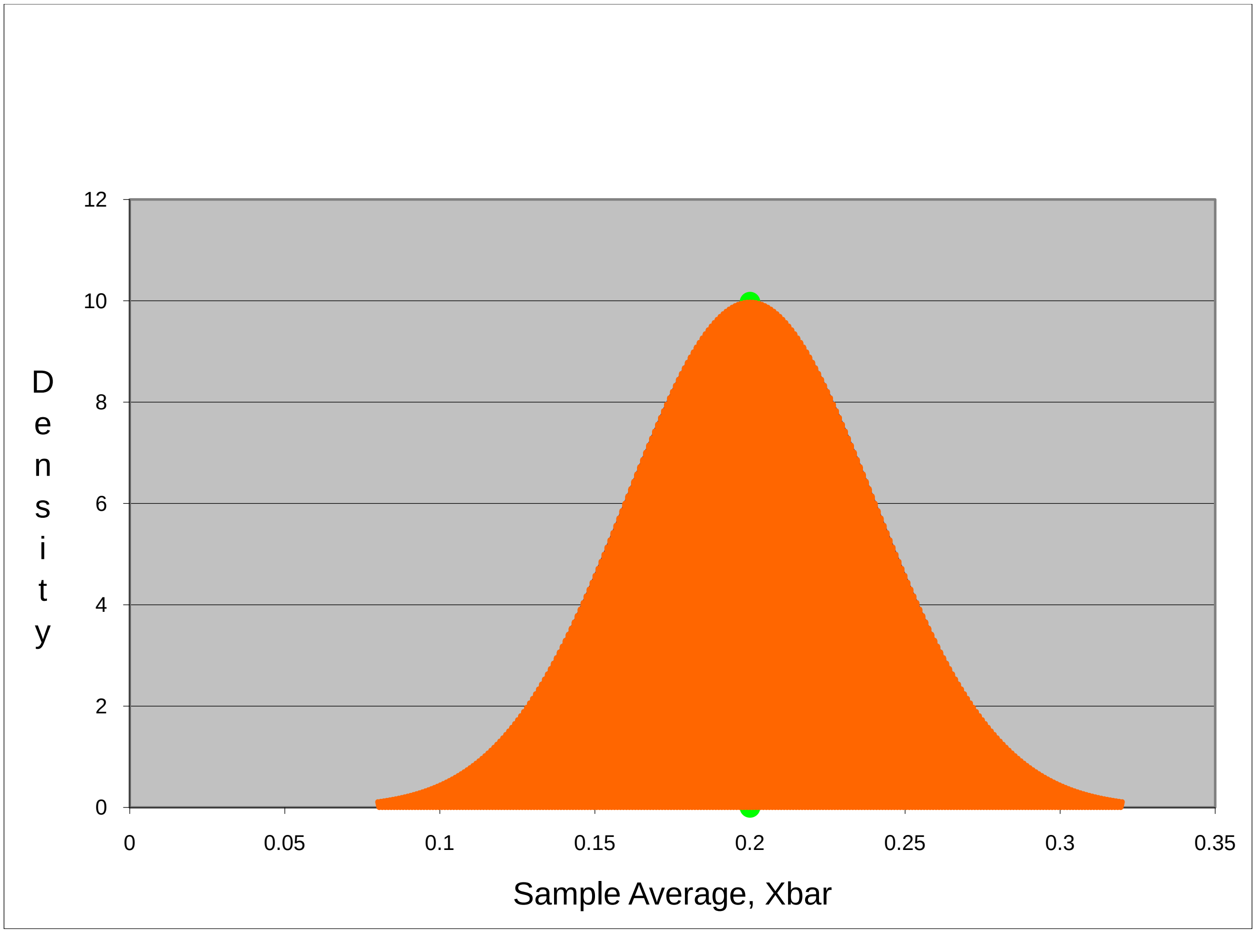
Solution: Var(X) = $(0-.2)^2$*.80 + $(1-.2)^2$*.20 = .032 + .128 = .16.

C. (5) Calculate the standard deviation of X.

Solution: StdDev(X) = Sqrt(0.16) = 0.4.

D. (10) Let $\overline{X}$ = (X_1 + X_2 + …+ X_{100})/100 be the average value of an iid sample from p(x). Draw a graph of the approximate distribution of $\overline{X}$. Put numbers and labels on the vertical and horizontal axes.

Solution: By the linearity and additivity properties of mean and variance, the expected value and variance of $\overline{X}$ are .20 and .16/100, respectively. Hence the standard deviation of $\overline{X}$ is sqrt(.16/100) = .4/10 = .04. By the central limit theorem, the distribution of $\overline{X}$ is approximately normal, so the graph should be the normal distribution with mean .2 and standard deviation .04, like this:



7. See the following code. Explain the meaning of each of the lines labeled ‘A’, ‘B’, ‘C’, ‘D’, and ‘E’ (2 each).

   Solution: See comments inside the code below.

```
data house;
   do subject_id = 1 to 1000;
      income = rand('normal',70,10);       /*  A: This line generates a
random income value from the normal distribution with mean 70 and standard
deviation 10 */
       house_expense = .15*income*exp(.3*rand('normal',0,1));
       output;                       /* B:  This line writes the current
data to the file work.house  */
   end;
run;

proc gplot data=house;
   plot house_expense*income;              /* C:  This line creates a
scatterplot of the house_expense (y axis) and income (x axis) data */
run;

proc univariate data=house;
   var house_expense income;
   histogram;
run;

proc univariate data=house(where=(income<60));    /* D: This line says to
analyze the house data using the univariate procedure, but only analyzing
cases where the income is <60 */
   var house_expense;
   histogram / endpoints = 0 to 40 by 2.5;
run;

proc univariate data=house(where=(income>80));
   var house_expense;
   histogram/ endpoints = 0 to 40 by 2.5;         /* E:  This line says to
create a histogram of the house_expense variable, using endpoints 0, 2.5,
5, …, 40 for the bars */
run;
```