ISQS 5347 Final exam.

Closed book, notes, and no electronic devices. Points (out of 200) are in parentheses.

1. (15) Suppose X and Y are independent random variables. Suppose also that the marginal pdfs of X and Y are as follows:

x	$p(x)$
A	0.2
B	0.8
Total:	1.00

y	$p(y)$
1	0.4
2	0.3
3	0.3
Total:	1.00

If n = 100 pairs (X_i, Y_i) are sampled, and arranged in the following table, what are the expected values in each of the six cells? No discussion is needed here; just numbers, but I'll give part credit for good logic and wrong answers. Note: All six entries must add to 100.

	Y = 1	Y = 2	Y = 3
X = A			
X = B			

Solution: Since the variables are independent, the joint probabilities are products of the marginals. Thus the table looks like this:

	Y = 1	Y = 2	Y = 3
X = A	8	6	6
X = B	32	24	24

2. Here is a data set for a variable called Y: 2, 4, 3, 4, 1, 1, 7, 3, 5, 3.
   A. (5) Give the bootstrap distribution of Y using all n = 10 observations.

Solution: Here it is:

y	$\hat{p}(y)$
1	0.2
2	0.1
3	0.3
4	0.2
5	0.1
7	0.1
Total:	1.00

B. (5) Suppose the first five observations are data from ENG students and the second five observations are data from BIOL students. Calculate the difference between the averages of the two groups, first five minus second five.

Solution: $14/5 - 19/5 = -1.0$.

C. (10) Give an iid null model under which the difference in B is explained by chance alone. Why is the word “explained” used here rather than “explainable”?

Solution: Such a model is that the data are produced according to $Y_1, \ldots, Y_{10} \sim_{\text{iid}} p(y)$. If this model is true, then the differences between the averages of the first five and second five are explained by chance alone, since the model is exactly the same (the same $p(y)$) in either case.

D. (10) Explain in detail how you would use simulation with the bootstrap distribution of A to find the two sided p-value to assess whether the difference is explainable by chance alone.

Solution: I’d simulate 10,000 data sets, each having 10 observations, using the distribution in A. Then for each of those 10,000 data sets, I’d calculate the difference between the average of the first five and the last five. Then I’d estimate the two-sided p-value as the proportion of those 10,000 data sets for which the difference between averages was either ≤ -1.0 or $\geq +1.0$.

E. (5) Why is the word “explainable” rather than “explained” used in D? Use “ENG” and “BIOL” somewhere in your answer.

Solution: If the difference -1.0 between BIOL and ENG (say, biology and engineering) is within the realm of chance variation, that does not prove that BIOL and ENG have the precise same distribution. They really might be different, in which case the difference is explained not only by chance, but also by systematic differences between BIOL and ENG students.

3. An equation relating product complexity (x) to consumer preference (y) is as follows:

$$f(x) = 2 + 0.5x - 0.1x^2$$

A. (5) Show that $f(x)$ is a concave function by calculating its second derivative.

Solution: Here, $f'(x) = 0.5 - 0.2x$ and thus $f''(x) = -0.2$. Since the second derivative is negative, the function is concave.

B. (5) Using the first derivative of $f(x)$, find the x that makes $f(x)$ a maximum. Explain the logic of your methods; also explain how you know from A that the result is a maximum rather than a minimum.

Solution: The extreme of the function (either max or min) is located where the derivative is zero; ie, where the slope is flat. Here the derivative is $f'(x) = 0.5 – 0.2x$; setting this to zero yields $x = 2.5$. This value locates the maximum, rather than a minimum, because the function is concave.

4. The logistic regression model specifies the following conditional distribution for success probability.

y	$p(y \mid x)$
Failure	$1/\{1 + \exp(\beta_0 + \beta_1 x)\}$
Success	$\exp(\beta_0 + \beta_1 x)/\{1 + \exp(\beta_0 + \beta_1 x)\}$
	Total: 1.0

Suppose Y_i = purchase of a particular item (Yes or No) for customer i, $i = 1, 2, \ldots, n$, who enters a store, and X_i = age of the customer. The first two of the n data pairs are (Yes, 32), (No, 20), …

A. (5) Give a model for how the n Y_i observations are produced. (Independence should be mentioned somewhere, as well as the pdf above.

Solution: The model is $Y_i \mid X_i = x_i \sim_{\text{independent}} p(y \mid x_i)$, where $p(y \mid x)$ is given above.

B. (5) Give a model for how the n Y_i observations are produced according to an appropriate null model.

Solution: The model is $Y_i \mid X_i = x_i \sim_{\text{iid}} p(y)$, where $p(y)$ is Bernoulli(π), independent of x. Equivalently, $Y_i \mid X_i = x_i \sim_{\text{independent}} p(y \mid x_i)$, where $p(y \mid x)$ is given above but with $\beta_1 = 0$.

C. (20) Write down the expressions for the likelihood functions and the log likelihood functions under both models A and B. Four separate expressions are needed here; use "…" as needed.

Solution:
Unrestricted model:
$L = \exp(\beta_0 + \beta_1(32))/\{1 + \exp(\beta_0 + \beta_1(32))\} \times 1/\{1 + \exp(\beta_0 + \beta_1(20))\} \times \ldots$
$LL = \ln(L)$, where L is given in the line above.

Restricted model:
$L = \pi \times (1 - \pi) \times \ldots$
$LL = \ln(L)$, where L is given in the line above.

D. (10) Explain how the likelihood ratio chi-square statistic for comparing the two models is related to your answer of C.

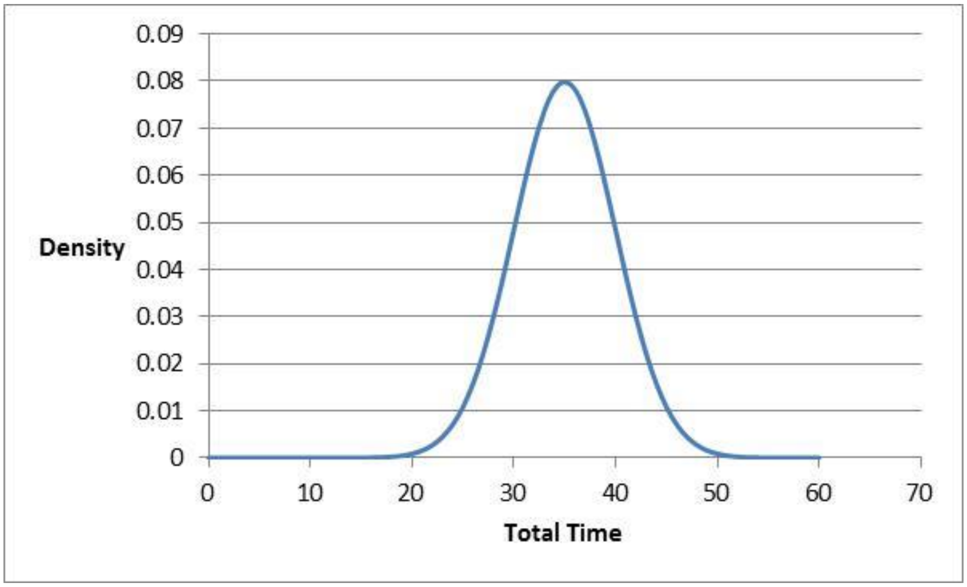
Solution: Find the MLEs for β_0 and β_1 in the unrestricted model, then plug them into the expression for LL to get LL_1. Find the MLE for π in the restricted model, then plug it into the expression for LL to get LL_0. Then the LR statistic is $\chi^2 = 2(LL_1 - LL_0)$.

E. (5) How many degrees of freedom are there for the chi-squared statistic in D? Explain.

Solution: Since there are two parameters in the unrestricted model (β_0 and β_1) and there is one parameter in the restricted model (π), there are 2 – 1 = 1 degrees of freedom for the chi-square test.

5. (15) John and Mary are working on a homework assignment with two problems. Mary will do Problem 1, and when she is finished, she will give it to John and he will do Problem 2. Mary's time comes from a normal distribution with mean 15 minutes and standard deviation 3 minutes, while John's time comes from normal distribution with mean 20 minutes and standard deviation 4 minutes, and is independent of Mary's time. How long will it take them to complete the assignment? There is not one answer here, there is a discussion. **Draw the relevant graph *first***, then give your discussion. ***Give the relevant logic*** behind *all* of your calculations.

Solution: The total time is $T = X + Y$, where X and Y are Mary's and John's times, respectively. Then the pdf of T is $N(15 + 20, 3^2 + 4^2)$, by the additivity property of the normal distribution assuming independence. Simplifying, we have $T \sim N(35, 5^2)$. Here is a graph:



Their total time will likely be between 25 and 45 minutes, according to the 68-95-99.7 rule.

6. (15) In the two-sample t-test, the standard error of the difference between means is given by $s.e.(\overline{Y}_1 - \overline{Y}_2) = \hat{\sigma}(1/n_1 + 1/n_2)^{1/2}$. This result is based on the fact that $Var\,(\overline{Y}_1 - \overline{Y}_2) = \sigma^2(1/n_1 + 1/n_2)$. Starting with the assumed model

$$Y_{ij} \sim_{\text{independent}} \mathrm{N}(\mu_i, \sigma^2), \text{ for } i = 1,2 \text{ and for } j = 1,\ldots,n_i,$$

give the step-by-step logic that explains why $Var\,(\overline{Y}_1 - \overline{Y}_2) = \sigma^2(1/n_1 + 1/n_2)$.

Solution: Start with the first group:

$Var\,(\overline{Y}_1) = Var\,\{(1/n_1)(Y_{11} + Y_{12} + \ldots + Y_{1n_1})\}$ (by substitution)

$= (1/n_1)^2 Var\,(Y_{11} + Y_{12} + \ldots + Y_{1n_1})$ (by the linearity property of variance)

$= (1/n_1)^2\{Var\,(Y_{11}) + Var\,(Y_{12}) + \ldots + Var\,(Y_{1n_1})\}$ (by the additivity property of variance for independent RVs)

$= (1/n_1)^2\{\sigma^2 + \sigma^2 + \ldots + \sigma^2\}$ (since, by assumption, the variance of every observation in group 1 is σ^2)

$= (1/n_1)^2 n_1 \sigma^2$ (since there are n_1 terms in the summation)

$= (1/n_1)\sigma^2$ (by algebra)

Following the same logic for group 2's data, you get $Var\,(\overline{Y}_2) = (1/n_2)\sigma^2$.

Now,

$Var\,(\overline{Y}_1 - \overline{Y}_2) = Var\,\{\overline{Y}_1 + (-1)\overline{Y}_2\}$ (by arithmetic)

$= Var\,(\overline{Y}_1) + Var\,\{(-1)\overline{Y}_2\}$ (since group 1's data are independent of group 2's data we have that $\overline{Y}_1, \overline{Y}_2$ are independent and can therefore apply the additivity property of variance for independent random variables)

$= Var\,(\overline{Y}_1) + (-1)^2 Var\,(\overline{Y}_2)$ (by the linearity property of variance)

$= Var\,(\overline{Y}_1) + Var\,(\overline{Y}_2)$ (by algebra)

$= (1/n_1)\sigma^2 + (1/n_2)\sigma^2$ (by substitution, using results shown above)

$= \sigma^2(1/n_1 + 1/n_2)$ (by algebra)

7. (10) Why are critical values for Student's *T* distribution larger than the corresponding critical values from the standard normal distribution? Explain.

Solution: When $Y_1,\ldots,Y_n \sim_{\text{iid}} \mathrm{N}(\mu, \sigma^2)$, we know from the linearity and additivity properties of the normal distribution that

$$Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}}$$

is distributed as N(0,1); i.e., *Z* has the standard normal distribution. But if we replace σ by $\hat{\sigma}$, there is an additional source of variability in the statistic.

The distribution of

$$T = \frac{\bar{Y} - \mu}{\hat{\sigma} / \sqrt{n}}$$

is the T_{n-1} distribution, which is wider than the standard normal distribution because it accounts for the variability inherent in the estimator $\hat{\sigma}$. Since the T distribution has more variability, you have to go farther out in the tail to find critical values that trap 95% of the area, hence the T distribution critical values are larger than the standard normal critical values.

8. (10) Suppose you have 1,000,000 values of θ^*, which you have obtained by simulating from the posterior distribution $p(\theta \,|\, \text{data})$. Describe how to obtain the 90% equal-tail credible interval for θ using these 1,000,000 values.

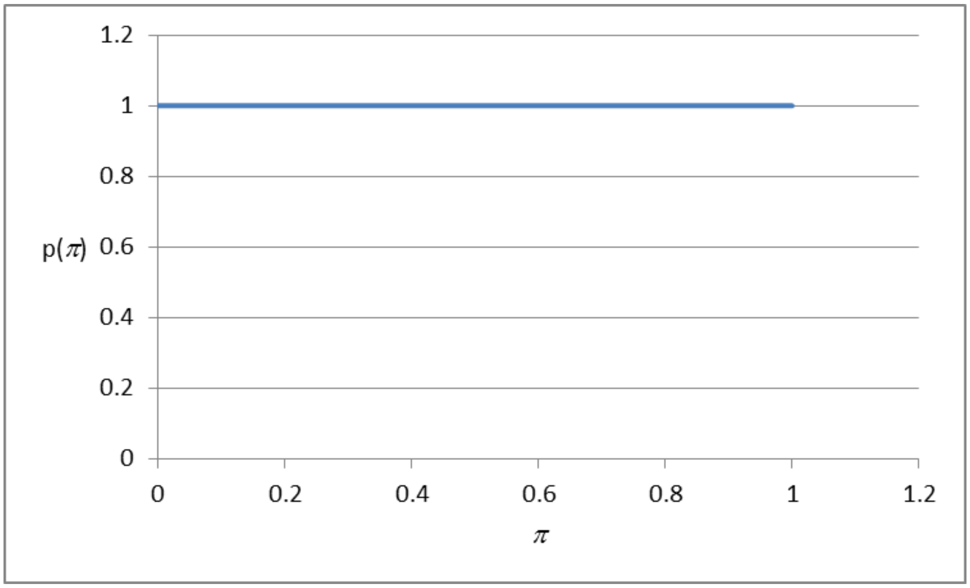
Solution: Find the 5th percentile of these data and the 95th percentile. You can do this by sorting the data set and taking the 50,000th value and the 950,000th values from the bottom. Call these values L and U --- They are reasonably good approximations to the true quantiles because the number of simulated data values is large. The 90% equal-tail credible interval for θ is then L < θ < U.

9. (10) Give an example where you use the beta distribution as a model. Also, explain why you would *not* use the normal distribution as a model for that example.

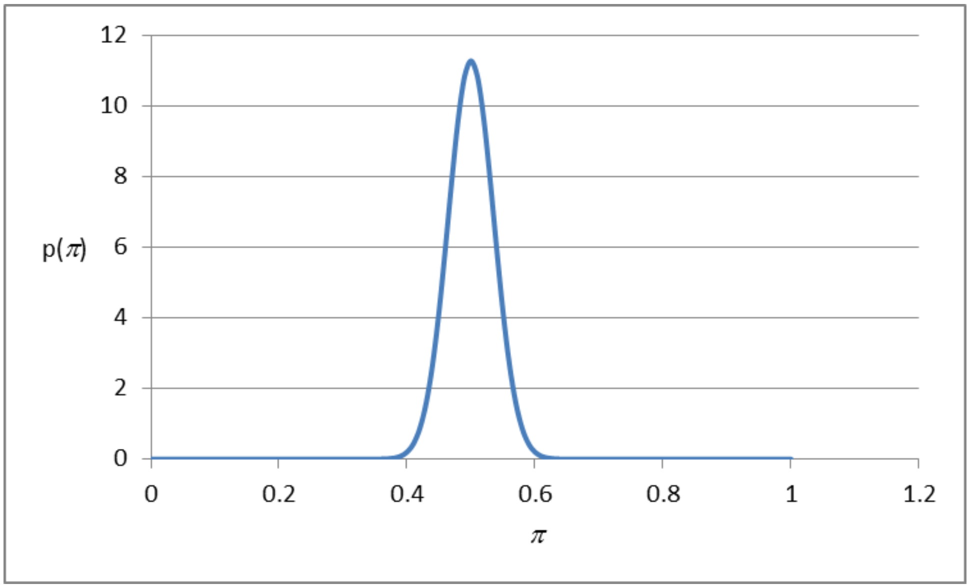
Solution: I'd use it to model my prior opinion about a Bernoulli parameter π, which I know has to be between 0 and 1. I wouldn't use a normal distribution here because a normal distribution states that values less than 0 and more than 1 are possible. (Specifically, anything between negative infinity and positive infinity is possible with a normal distribution, although most of that infinite range has such infinitesimally small probability that it can be ignored.)

10. (15) What is the difference between an "ignorant prior" and an "informative prior"? Draw graphs of each for a particular example, either one discussed in class or one of your own choosing. Be sure to explain the meaning of the parameter θ for that example, and give an informative prior that makes sense.

Solution: A typical "ignorance prior" is the uniform distribution. We used this for the thumbtack toss. Here it is:



In the case of a coin toss, we can be more informative, since we think a coin has probability near 0.50. Here is an example:



The difference is that with the ignorance prior, you state that you are ignorant about the value of the parameter (e.g., the probability of an object landed on one particular side). With an informative prior, you are expressing an opinion about which values of the parameter are more likely. With an ignorance prior, you let the data talk for themselves in the analysis. With an informative prior, you modify your interpretation of the data based on your prior opinion.

11. (20) An approximate 95% confidence interval for the mean μ of the Poisson distribution is given by $\bar{Y} \pm 1.96\sqrt{\bar{Y}}/\sqrt{n}$. Describe, step-by-step, how to estimate the true confidence level by using a simulation study. Assume n= 10 and $\mu = 2$ for your study. Don't bother with SAS code or EXCEL statements, just explain the methods clearly enough so that someone could carry the instructions out using any software that can simulate data from the Poisson distribution.

Solution:

(i) Simulate 10 values from the Poisson distribution with mean = 2.

(ii) Construct the interval $\bar{Y} \pm 1.96\sqrt{\bar{Y}}/\sqrt{n}$ using the 10 values.

(iii) Check to see whether 2.0 is inside the interval you constructed in (ii).

(iv) Repeat (i), (ii), and (iii) 1,000,000 times. The proportion of the 1,000,000 intervals having 2.0 inside the interval is an estimate of the true confidence level.

It wasn't requested, but here is SAS code. The procedure works fine, with true confidence level near 95%, although the interval misses more often on the low side. A Bayesian interval would be better because it would be asymmetric.

```
data sim;
    do sim = 1 to 1000000;
      sum = 0;
        do i = 1 to 10;
      sum + rand('poisson',2);
      end; ave = sum/10;
        Low_int = 2 > ave + 1.96*sqrt(ave/10);
        Hgh_int = 2 < ave - 1.96*sqrt(ave/10);
        Wrng_int = Low_int + Hgh_int;
          output;
    end;
run;

proc freq;
  tables Low_int Hgh_int Wrng_int;
run;
```