ISQS 5347 Final exam Fall 2012

Closed book and notes. Points (out of 200) in parentheses.

1. Pick an example of your own interest where data can be observed. These data must be *variable*, and describable using a probabilistic model. Do not use an example from the book, or from class discussions. Do not make the example about coins, dice, cards, or any other games of chance either. This must be an example from your own interests. Describe how the phrase “model produces data” applies to your example as follows:

   A. (5) First, describe your data, how they arise, and why they are variable. What is the *real* data generating process? (No statistical model here).

Solution: I like baseball, so my data will be number of runs scored in a game by both teams. This is determined by the flow of the game and the players’ performances – consecutive hits, home runs, errors, etc.

   B. (5) Give a model for how your data are produced, and explain why it must be probabilistic (as opposed to deterministic).

Solution: Let $Y$ = total number of runs scored in a game (pick any one game). I will assume $Y \sim p(y)$ for some generic pdf $p(y)$. The number of runs cannot be predicted deterministically in advance, so I’ll assume a probabilistic model.

   C. (5) Explain why the *average* that is calculated from the data you can collect is *random*. Specifically, why can there be *more than one* average? Relate this answer to your answer for 1.A.

Solution: In any collection of games, pick 10 games for example, the average number of runs scored will be different from the average in any other collection of ten games. Every game is unique and not predictable, so every collection of ten games is similarly unique and unpredictable.

   D. (5) Define the meaning of *parameter* as it relates to your model of 1.B. Explain why the parameter(s) is(are) *unknown*.

Solution: One parameter of my model is the mean $\mu = \Sigma\, y\, p(y)$. This is part of the model that produces my data, and I assume it is a fixed value, because I assume the distribution $p(y)$ is a fixed distribution. My model for runs is that Hans, down the hall is simulating runs data according to some distribution $p(y)$ that only he knows. I don't know what model Hans is using, so I don't know $\mu$. Further, I know that any set of data I collect will have an average number of runs that is random (see 1.C.). Therefore, I know that the parameter can't be from the data, because every set of data gives me a different number.

E. (5) When is your model of 1.B. a *good model*? Explain in terms of the data you describe in 1.A.

Solution: The model is good if it produces data that look like the runs data, for some values of the unknown parameters. Here, the distribution $p(y)$ has parameters $\pi_1$, $\pi_2$, …, the probabilities for 1, 2, … total runs scored (in American baseball, the total number of runs is always more than 0 because there are no ties, only suspensions). Since the parameters $\pi_1$, $\pi_2$, …, can be anything, this model certainly will produce data that look just like the actual runs data, for some settings of the parameters.

2. (10) Suppose $Y_1, Y_2, \ldots, Y_{16} \sim_{\text{iid}} p(y)$, where

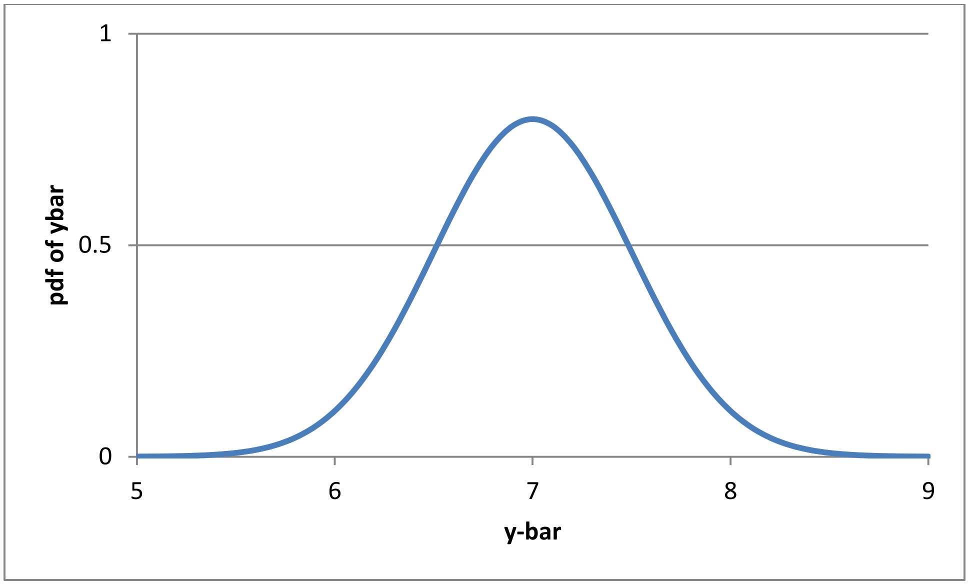
$$\mu = \int y\, p(y)\, dy = 7.0$$

and

$$\sigma^2 = \int (y - 7.0)^2\, p(y)\, dy = 4.0.$$

Carefully draw the probability distribution function of $\bar{Y} = (1/16)(\, Y_1 + Y_2 + \ldots + Y_{16})$ as best you can. Label both the vertical and horizontal axes, and put numbers on them that make sense. Explain your reasoning in words and symbols.

Solution: Note that $E(\bar{Y}) = \mu = 7.0$ and $\mathrm{Var}(\bar{Y}) = \sigma^2/n = 4.0/16 = 0.25$. So $\mathrm{StdDev}(\bar{Y}) = (0.25)^{1/2} = 0.5$. Further, by the CLT, the distribution of $\bar{Y}$ is approximately a normal distribution. Putting it all together, $\bar{Y} \sim N(7.0, 0.5^2)$. Here is a graph:



3. In the example in class where you rated the color "green," you were each given a colored piece of paper – some yellow, some green, some white. It was stated that the *F* statistic calculated from this experiment will have an *F* distribution under the chance-only model. The *F* statistic was computed to be *f* = 0.19 for our class.

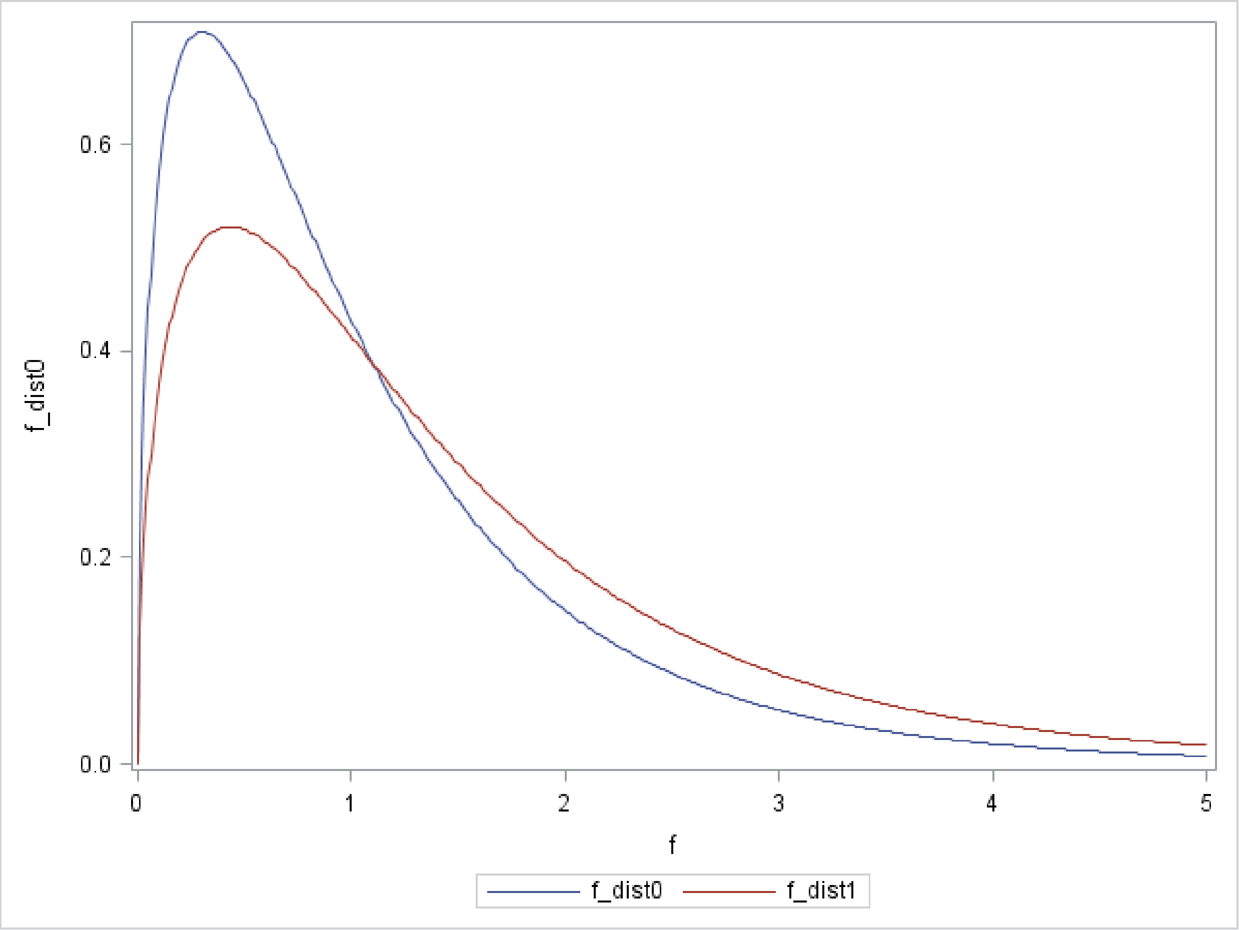
   But a "*distribution*" refers to a large possible set of data values, not just one observation.

   A. (10) Explain how the other *f* 's (other than the *f* = 0.19 that we computed) from this distribution can arise. Use the context of our in-class experiment to explain where the other *f* 's come from.

Solution: If I had re-arranged the pieces of paper so that many of you were in different color groups, then the *F* statistic would differ. It would also differ if the students in the class were a different collection of students.

   B. (10) Draw two curves on the same graph: (i) The distribution of these *F* statistics under the chance-only model. (ii) The distribution of these *F* statistics when the people who have green paper *tend* rate the green color *slightly* more highly than people who have either white or yellow paper. Label and number the axes.

Solution: Something like this:



4. The *p*-value is the probability of observing a difference as extreme as the observed difference, when the data are produced by a model where there is in fact no difference. (In other words, when the data are produced by the null, or chance-only model.)

   A. (10) Explain the *logic* for why *pv* = 0.03 allows you to "reject" the chance-only model. Include a graph. *DO NOT REFER TO* $\alpha = 0.05$.

Solution: The graph of 3.B. is good. In the null (blue) graph showing the values of *f* that are explained by chance only, a p-value of 0.03 corresponds to an f-statistic around 3.0. So it is unusual to see an f value that big by chance alone. The fact that it is unusual to see such a big *f* value, under the null, is the logic for rejecting the null.

B. (10) Explain why $pv = 0.03$ *does not prove* that the chance-only model is *wrong*. Again, do not refer to $\alpha = 0.05$. Give the *logic* instead, and refer to your graph in 4.A.

Solution: While a 0.03 probability is small, it is not zero. You do in fact see $f$ statistics as large as the observed value in 3 out of 100 studies where the null is true; this is the area to the right of the observed $f$ statistic in the graph of the null distribution in 4A. So, this result is explainable by chance alone, although it seems unusual.

5. A student's time to get to class is obviously related to their commute distance, although not deterministically.

   A. (10) Give a model for how $Y$ (time to get to class) *is produced*, as a function of $X$ (commute distance), and some *unknown* parameters. Describe in words how your model produces data, and what kind of data it produces.

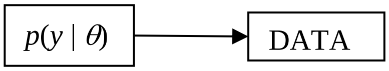
Solution: One model is $Y \mid X = x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$. This states that, for a particular commute distance $x$, the commute time is produced at random from a normal distribution with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2$. The unknown parameters are $\beta_0$, $\beta_1$, and $\sigma^2$.

   B. (5) Explain how the mantra "nature favors continuity over discontinuity" is demonstrated by your model.

Solution: When the commute distance $x$ increases a tiny bit, like by 0.01 for example, the distribution shifts a tiny bit, like to a mean that is $0.01\beta_1$ higher, for example.

6. (10) How, *specifically*, is the concept "model produces data," as pictured below, used when you analyze your data using Bayesian methods?

Model produces data picture:



Solution: Bayesian methods are based on the posterior distribution, which is proportional to the likelihood function times your prior distribution. Your model $p(y \mid \theta)$ gives you the likelihood function; for an iid sample it is specifically:

$$L(\theta \mid \text{data}) = p(y_1 \mid \theta) \times p(y_2 \mid \theta) \times \ldots \times p(y_n \mid \theta)$$

7. Often, we make the assumption that our DATA are produced by normal distributions.

   A. (5) Give an example of a particular type of confidence interval that assumes normality.

Solution: The ordinary confidence interval for the mean, which uses the *t* distribution critical value *c*, assumes normality of the data generating process. The interval is $\bar{y} \pm c \times \hat{\sigma} / \sqrt{n}$ .

   B. (5) Explain what goes wrong in 7.A. when the DATA are produced by non-normal distributions.

Solution: The confidence level is different from advertised. For example, if your critical value is the .975 quantile, you are claiming 95% confidence. But if the data are produced by non-normal distributions, then the true confidence level is not 95%.

8. (10) Define what it means for an estimator to be unbiased. *Define all terms*.

Solution: An estimator $\hat{\theta}$ is a function of observable data. It is used to estimate a parameter $\theta$ (the estimand), which is a characteristic of the data-generating process. The estimator is unbiased if $E(\hat{\theta}) = \theta$; in words, this equation states that the mean of the probability distribution of possible values of the random estimator is equal to the fixed estimand.

9. (5) How is the concept "model produces data" related to the Law of Large Numbers?

Solution: The Law of Large Numbers refers to data $Y_1, Y_2, \ldots, Y_n$ that are produced as iid from a model $p(y)$, and states that the average of these data values gets closer to the mean ($\mu$) of the pdf

$p(y)$ as $n$ increases. (Note that $\mu$ is either $\mu = \Sigma\, y\, p(y)$ or $\mu = \int y\, p(y)\, dy$, depending on whether $p(y)$ is discrete or continuous.)

10. (5) We often suppose that

$$Y_1, Y_2, \ldots, Y_n \sim_{\text{iid}} \text{N}(\mu, \sigma^2).$$

Explain what

$$Y_1, Y_2, \ldots, Y_n \sim_{\text{iid}} \text{N}(\mu, \sigma^2)$$

means, but *only* in terms of "model produces data."

Don't discuss anything else except how "model produces data" explains the meaning of the symbolic expression $Y_1, Y_2, \ldots, Y_n \sim_{\text{iid}} \text{N}(\mu, \sigma^2)$.

Solution: The statement $Y_1, Y_2, \ldots, Y_n \sim_{\text{iid}} \text{N}(\mu, \sigma^2)$ describes how the data are produced. They are produced independently, and all from the same distribution, which is a normal distribution with mean $\mu$ and variance $\sigma^2$. You can imagine Hans sitting down the hall using the Excel random number generator's normal distribution to produce these values, using some values for the mean and standard deviation input parameters that only he knows.

11. Suppose $Y_1, Y_2, \ldots, Y_n \sim_{\text{iid}} p(y)$, where $\mu = \text{E}(Y_1)$ and $\sigma^2 = \text{Var}(Y_1)$. Let $Z = \dfrac{(\bar{Y} - \mu)}{\sigma / \sqrt{n}}$,

where $\bar{Y} = (Y_1 + Y_2 + \ldots + Y_n) / n$ .

A. (10) Show that E(Z) = 0. Use the linearity and additivity properties as appropriate.

Solution: First, $E(\bar{Y}) = E\{(Y_1 + Y_2 + \ldots + Y_n)/n\} = (1/n)E\{(Y_1 + Y_2 + \ldots + Y_n)\}$ by the linearity property of expected value. By the additivity property of expected value, we have further that $(1/n)E\{(Y_1 + Y_2 + \ldots + Y_n)\} = (1/n)\{E(Y_1) + E(Y_2) + \ldots + E(Y_n)\}$ . Now, since the data are identically distributed, the mean of the pdf that produced them all is the same number, and identical to the mean of the distribution that produced $Y_1$; namely, $\mu$. So

$(1/n)\{E(Y_1)+E(Y_2)+\ldots+E(Y_n)\} = (1/n)\{\mu+\mu+\ldots+\mu\} = (1/n)(n\mu) = \mu$ , with the last steps explained by algebra.

Now, E(*Z*) = $E\left\{\frac{(\bar{Y}-\mu)}{\sigma/\sqrt{n}}\right\} = \frac{1}{\sigma/\sqrt{n}}E\{\bar{Y}-\mu\}$ by the linearity property of expectation. But $E\{\bar{Y}-\mu\} = E\{\bar{Y}\}-\mu$ , also by the linearity property of expectation. As shown above, $E\{\bar{Y}\}=\mu$ , hence E(*Z*) = 0.

B. (10) Show that Var(*Z*) = 1. Use the linearity and additivity properties as appropriate.

Solution: First, $Var(\bar{Y}) = Var\{(Y_1+Y_2+\ldots+Y_n)/n\} = (1/n)^2 Var\{(Y_1+Y_2+\ldots+Y_n)\}$ by the linearity property of variance. By the additivity property of variance when the data are independent, we have further that $(1/n)^2 Var\{(Y_1+Y_2+\ldots+Y_n)\} = (1/n)^2\{Var(Y_1)+Var(Y_2)+\ldots+Var(Y_n)\}$ . Now, since the data are identically distributed, the variance of the pdf that produced them all is the same number, and identical to the variance of the distribution that produced $Y_1$; namely, $\sigma^2$. So $(1/n)^2\{Var(Y_1)+Var(Y_2)+\ldots+Var(Y_n)\} = (1/n)^2\{\sigma^2+\sigma^2+\ldots+\sigma^2\} = (1/n)^2(n\sigma^2) = \sigma^2/n$ , with the last steps explained by algebra.

Now, Var(*Z*) = $Var\left\{\frac{(\bar{Y}-\mu)}{\sigma/\sqrt{n}}\right\} = \left(\frac{1}{\sigma/\sqrt{n}}\right)^2 Var\{\bar{Y}-\mu\}$ by the linearity property of variance. But $Var\{\bar{Y}-\mu\} = Var\{\bar{Y}\}$ , also by the linearity property of variance. As shown above, $Var\{\bar{Y}\} = \sigma^2/n$ , hence Var(*Z*) = $\left(\frac{1}{\sigma/\sqrt{n}}\right)^2\sigma^2/n$ =1.0.

12. (5) Suppose $Y_1, Y_2, \ldots, Y_n \sim_{\text{iid}} \text{N}(\mu, \sigma^2)$. Let $Z = \frac{(\bar{Y}-\mu)}{\sigma/\sqrt{n}}$. Let $T = \frac{(\bar{Y}-\mu)}{\hat{\sigma}/\sqrt{n}}$. Explain, using these formulas, why the variance of *T* is larger than the variance of *Z*.

Solution: The formulas differ only in that there is $\hat{\sigma}$ in the *T* statistic and a $\sigma$ is the *Z* statistic. Since $\hat{\sigma}$ is a function of random data, it is also random. This randomness increases the variability of the *T* statistic relative to the *Z* statistic.

For a more technical explanation, consider the case where the variance of *T* is finite. Then its mean is zero, hence Var(*T*) = E{(*T* – 0)$^2$} = E($T^2$).

Write $T^2 = \left\{ \frac{(\bar{Y} - \mu)}{\hat{\sigma} / \sqrt{n}} \right\}^2 = \left\{ \frac{(\bar{Y} - \mu)}{\sigma / \sqrt{n}} \right\}^2 (\sigma / \hat{\sigma})^2$. By independence of $\bar{Y}$ and $\hat{\sigma}$, the product rule for expected value applies; hence Var(*T*) = E($T^2$) = $\mathrm{E}\left[ \left\{ \frac{(\bar{Y} - \mu)}{\sigma / \sqrt{n}} \right\}^2 \right] \mathrm{E}\left\{ (\sigma / \hat{\sigma})^2 \right\}$. By the previous problem, $\mathrm{E}\left[ \left\{ \frac{(\bar{Y} - \mu)}{\sigma / \sqrt{n}} \right\}^2 \right] = \mathrm{E}(Z^2) = \mathrm{Var}(Z) = 1$. That leaves $\mathrm{E}\left\{ (\sigma / \hat{\sigma})^2 \right\} = \sigma^2 \mathrm{E}(1 / \hat{\sigma}^2) > \sigma^2 \{1 / \mathrm{E}(\hat{\sigma}^2)\}$, by Jensen's inequality. But $\hat{\sigma}^2$ is an unbiased estimator; hence $\{1 / \mathrm{E}(\hat{\sigma}^2)\} = 1 / \sigma^2$. Putting it all together, Var(*T*) > Var(*Z*) = 1.

13. Here is a data set. 3, 4, 5, 6, 3.

A. (5) Give the bootstrap distribution.

Solution:

| *y* | $\hat{p}(y)$ |
|---|---|
| 3 | 0.4 |
| 4 | 0.2 |
| 5 | 0.2 |
| 6 | 0.2 |

B. (5) Show how to calculate the plug-in estimate of the expected value, using the bootstrap distribution. Be sure to use the expected value formula that involves the probability distribution.

Solution: The mean of a discrete distribution is $\mu = \Sigma\, y\, p(y)$. Using the bootstrap estimate of *p*(*y*), we get 3(0.4) + 4(0.2) + 5(0.2) + 6(0.2) = 1.2 + 0.8 + 1.0 + 1.2 = 4.2.

14. (10) Mary takes, on average, 20 minutes to get ready for her day, with a standard deviation of 3 minutes. Jane takes, on average, 30 minutes with a standard deviation of 4 minutes. Over 100 days, here are Mary's and Jane's data on time to get ready:

| Day | Mary | Jane | Difference |
|---|---|---|---|
| 1 | 26 | 33 | -7 |
| 2 | 23 | 31 | -8 |
| … | … | … | … |
| 100 | 19 | 29 | -10 |

What's your best guess of the standard deviation of the numbers in the "Difference" column? DO NOT USE the numbers you see in the "Difference" column AT ALL. Explain your logic.

Solution: We can assume they are independent. So the variance of the difference is

Var(Mary – Jane) = Var(Mary + (-1)Jane) = Var(Mary) + $(-1)^2$Var(Jane) = $3^2 + 4^2 = 25$. Hence the standard deviation of the differences is $(25)^{1/2} = 5.0$.

15. (10) Your data set has a column of values that you call *Y*. Suppose you re-code your data where $Y > 10$ as 1, and the data where $Y \leq 10$. Why is the average of 0/1 data an estimate of Pr($Y > 10$)?

Solution: The average of the 0s and 1s is the total number of 1's divided by the total number of observations. This translates to the total number of observations where $Y > 10$ divided by the total number of observations, which is an estimate of Pr($Y > 10$).

16. (5) Explain how you would use simulation to demonstrate that the normal quantile-quantile plot data do not fall precisely on a straight line as expected, even when the data are in fact produced by a normal distribution.

Solution: I would simulate data using Excel or some other software from a normal distribution. I'd have to pick the mean and variance of that distribution; say 0 and 1 for starters. Then I'd draw a normal q-q plot of those simulated data, and I'd see that they do not fall precisely on a straight line. These differences are purely explained by chance.

17. (10) Describe the differences between a generic model *p*(*y*) that produces discrete DATA and a generic model *p*(*y*) that produces continuous DATA.

Solution: In the discrete case, the $p(y)$ is positive (non-zero) only for a collection of $y$ values that you can list. It is not positive over any continuous range. It has values that *add* to 1.0.

In the continuous case, the $p(y)$ is positive over continuous range(s). It has values that *integrate* to 1.0.

The interpretations differ: In the discrete case $p(y)$ is the probability that $Y = y$. In the continuous case, $p(y)$ is not a probability. Instead, $p(y)\Delta$ is the approximate probability that $Y$ is in a $\pm\Delta/2$ range of $y$.