ISQS 5347 Spring ’11 Final Exam

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

1. Suppose Y has the uniform distribution between 0 and 100.

1.A. (5) This model produces data. Explain, in a sentence, how the data produced by this model will look, and then write down your own "random sample" (from your mind) of 4 observations produced by this distribution.

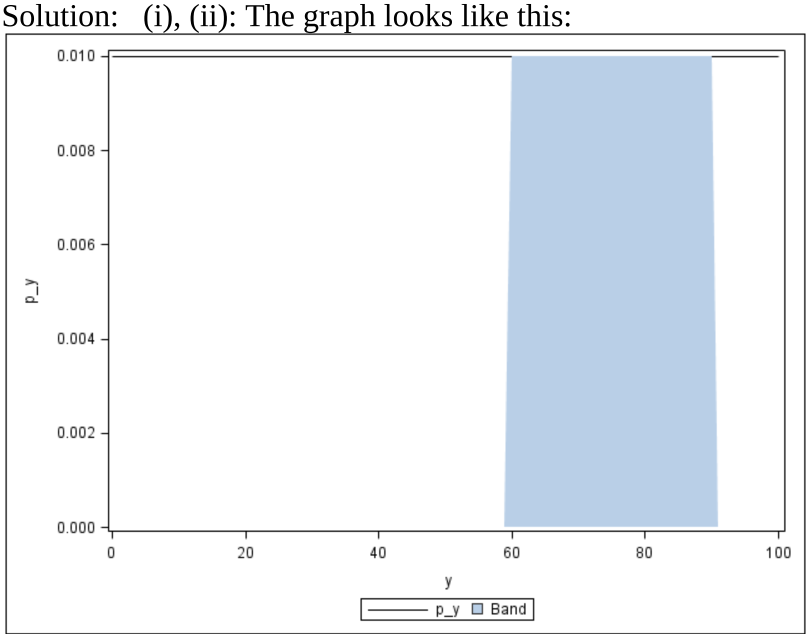
Solution: The data will be continuous, therefore with decimals that theoretically extend forever, and the frequency of observed values in any 10-point range will be the same – approximately 10%. Here is a random sample with decimals truncated to the tenths place: 53.1, 23.5, 78.7, 12.4.

1.B. (8) Suppose this model is used for test scores on a 100-question multiple choice exam. Give two reasons that the model is not a good model.

Solution: (i) The test scores are discrete, having no decimals, not continuous. (ii) test scores ordinarily are in the higher range, with many more observation above 50 than below 50. The model produces half of the data values above 50 and the other half below 50.

1.C. (12) If the model is valid, what percentage of scores will be between 60 and 90? (i) Draw the appropriate graph with numbers and labels on the axes, (ii) mark in pencil the area, (iii) express the area as an integral, and (iv) find the integral using calculus.

Solution: (i), (ii): The graph looks like this:



(iii), (iv) Area = $\int_{60}^{90} p(y)dy = \int_{60}^{90}(0.01)dy = 0.01y\Big|_{60}^{90} = 0.01(90) - 0.01(60) = 0.9 - 0.6 = 0.3.$

2. (10) What does the following SAS code do? Explain every line.

Solution: See the comments.

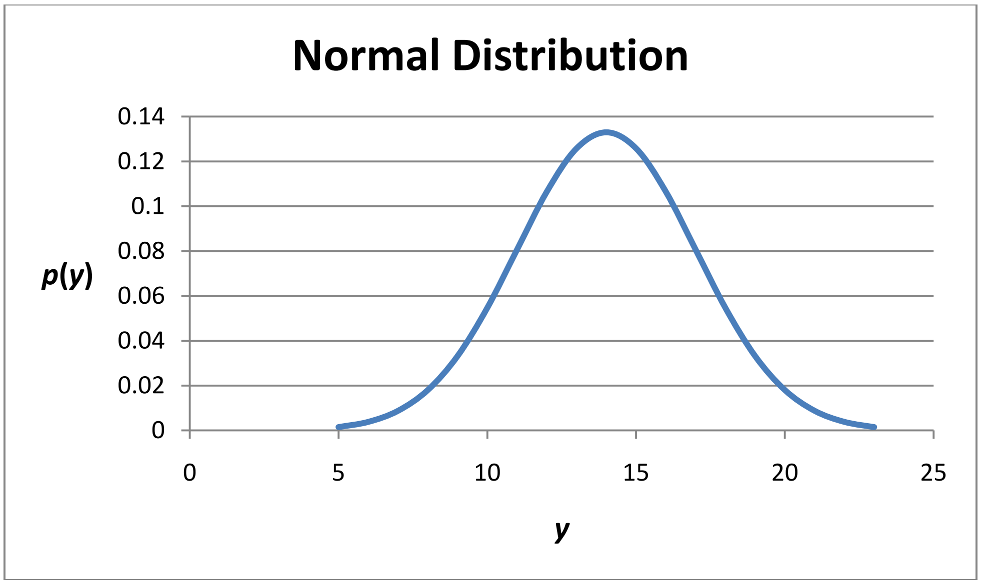
```
data poisson; /* tells SAS to create a data set
work.poisson */
   do day = 1 to 100; /* increment day from 1 to 100 by 1
*/
      cars_sold = rand("poisson", .5);  /* Generate
cars_sold at random from the Poisson distribution with mean
.5 */
      output; /* Write the current values to work.poisson
*/
   end;  /* End of the ‘do loop’ code */
run;   /* Ends the code used to define the work.poisson
data set */

 proc means data=poisson;   /* Analyze the  data in
work.poisson using PROC MEANS, which computed averages and
other summary statistics */
   var cars_sold;  /* Performs analysis of the cars_sold
variable only */
run;  /* Ends the code used to define what PROC MEANS is
suppose to do */
```

3. (10) Give the formula for the "standard error of $\bar{y}$," and explain what the "standard error of $\bar{y}$" is used for.

Solution: The formula is s.e.($\bar{y}$) = $\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ and n is the sample size used to compute $\bar{y}$. This quantity is an estimate of the standard deviation of the random variable $\bar{Y}$, and tells you how far $\bar{y}$ can be from the mean μ. A 95% confidence interval for μ is approximately $\bar{y} \pm 2\{\text{s.e.}(\bar{y})\}$, although the precise critical value is not 2.0, but rather the .975 quantile of the t_{n-1} distribution.

4. (5) The model is assumed to be Y|X=x ~ N(2+ 3x , 3^2). Draw a graph of the distribution of Y when X is known to be 4. Put labels and numbers on the horizontal and vertical axes.

Solution: The mean will be 14 and the variance will be 9, so the standard deviation is 3.0. Here is a graph:



5. (10) Suppose the data 3.2, 6.1, 1.0 are assumed to come from an iid sample from the N(μ,1) distribution. Write down the likelihood function for μ. There is no need to simplify the function.

Solution: The normal pdf is given by $p(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$. Here, σ=1, so the formula becomes $p(y) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(y-\mu)^2}{2}\right)$. Since data are assumed to be from an independent sampling process, the likelihood of the sample is the product of the likelihood of the individual observations, so we get

$$L(\mu) = \left[\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(3.2-\mu)^2}{2}\right)\right] \times \left[\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(6.1-\mu)^2}{2}\right)\right] \times \left[\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(1.0-\mu)^2}{2}\right)\right].$$

This can be simplified somewhat, but as per instructions, “there is no need to simplify the function.”

6. (10) State the Central Limit Theorem.

Solution: If Y_1, Y_2,…,Y_n are an iid sample from some distribution $p(y)$ with finite variance σ^2, then the distribution of the sum $Y_1+Y_2+\ldots+Y_n$ is approximately a normal distribution. The approximation becomes better as the sample size n gets larger.

The same statement can also be made about the average $(Y_1+Y_2+\ldots+Y_n)/n$.

7. (10) State the Law of Large Numbers.

Solution: If Y_1, Y_2,…,Y_n are an iid sample from some distribution $p(y)$ whose mean is μ, where $\mu = \int yp(y)\, dy$ (or $\mu = \Sigma yp(y)$ in the discrete case), then the average $(Y_1+Y_2+\ldots+Y_n)/n$ converges to μ as n increases.

Multiple Choice. (3 points each)

8. What is true about the "beta" distribution?
A. It produces data that are either 0 or 1. (No, that's Bernoulli)
B. It produces data that are between 0 and 1. (Yes.)
C. It is produced by data that are either 0 or 1. (No, model produces data).
D. It is produced by data that are between 0 and 1. (No, model produces data).

9. The histogram is used to estimate _______.
A. μ (No, $(Y_1+Y_2+\ldots+Y_n)/n$ is used to estimate μ).
B. σ
C. β
D. $p(y)$ (Yes)

10. What is the approximate outcome of the following code:

```
data sim;
    do i = 1 to 1000000;
      Y = rand('normal', 0, 1);
      P = (Y > 1.96);
      output;
   end;
run;
proc means mean;
    var P;
run;
```

A. 0.950
B. 1.960
C. 0.050
D. 0.025 (recall that the 1.96 number puts 95% probability between $\mu \pm 1.96\sigma$. So there is 0.025 above $\mu + 1.96\sigma = 0 + 1.96(1) = 1.96$).

11. "Markov Chain Monte Carlo" used to
A. calculate the true mean, μ. (No, use calculus for that).
B. calculate the true standard deviation, σ. (Again, calculus)
C. find the maximum likelihood estimates of the parameters ($\hat{\theta}$). (Use the maximum likelihood method – take the log, take the derivative of the likelihood function. No simulation.)
D. simulate parameters (θ) from the posterior distribution. (Yes).

12. To test a hypothesis using Bayesian methods you must calculate the
A. the posterior probability that the hypothesis is true. (Right).
B. the frequentist p-value. (No – that's frequentist, which is the classical approach).
C. the likelihood that the p-value is true. (Doesn't make sense).
D. the probability that you reject the null hypothesis. (That's called "power" of a test).

13. A Bayesian range for the values of an unknown parameter is typically called a
A. confidence interval. (That's frequentist)
B. credible interval. (yes)
C. posterior mean. (that's not a range)
D. posterior median. (that's not a range)

14. The model $Y|X=x \sim N(\beta_0 + \beta_1 x, \sigma^2)$ is best called what?
A. A Bayesian model. (no, because you'd need a prior)
B. A normal regression model. (yes)
C. A logistic regression model. (no, that's when Y is Bernoulli).
D. A heteroscedastic variance model. (No, the variance of Y is constant here).

15. The least squares line is the one that
A. minimizes the sum of squared deviations. (yes)
B. maximizes the sum of squared deviations.
C. minimizes the sum of likelihoods.
D. maximizes the sum of likelihoods. (close, but the likelihood of the sample is the product, not the sum, of the likelihoods).

16. Which problem can Bayesian analysis solve for easily than frequentist analysis?
A. Interval estimation of nonlinear functions of parameters. (yes, recall the estimation of probability of a stock return being positive when the S&P return is zero)
B. Interval estimation of linear functions of parameters.
C. Maximum likelihood estimation of regression parameters.
D. Least squares estimation of regression parameters.

17. If a prior distribution is uniform from negative infinity to positive infinity, then it is usually called
A. a proper prior.
B. an improper prior. (yes)
C. a normally distributed prior.
D. a non-normally distributed prior. (could be, but this is not usual.)

18. What are the two approximations used when calling the interval $\bar{y} \pm 1.96\hat{\sigma}/\sqrt{n}$ an *approximate* 95% confidence interval for μ?
A. (i) $\bar{y} \approx \mu$ and (ii) $\hat{\sigma} \approx \sigma$
B. (i) $\bar{y} \approx \mu$ and (ii) $\bar{Y}$ has an approximately normal distribution
C. (i) $\hat{\sigma} \approx \sigma$ and (ii) $\bar{Y}$ has an approximately normal distribution (yes)
D. (i) $\bar{y}$ has an approximately normal distribution, and (ii) $\bar{Y}$ has an approximately normal distribution

19. If data 4.1, 3.9, 5.2, 5.8, 1.0 are the result of an iid sampling from a continuous distribution $p(y)$. Select the true statement.
A. The mean of the distribution that produced the data is $\mu = 4.0$. (no, $\bar{y}$ =4)
B. The distribution that produced the data is normal with unknown σ^2. (Can't assume normal).
C. The skewness of the distribution that produced the data is a positive number. (Can't know this from such a small sample)
D. The variance of the distribution that produced the data is $\int (y-\mu)^2 p(y)dy$. (By definition of variance, yes).

20. If the data 45 76 87 65 87 45 56 67 89 67 90 91 are test scores sampled from some distribution $p(y)$, then
A. P(Y = 80) = 0, because the distribution is continuous. (no, the distribution is actually discrete)
B. P(Y = 80) = 0, because there is no '80' in the data set. (No, probability refers to process, not data).
C. P(Y = 80) is closer to 0.06 than it is to 0.60. (yes, because there are 100 potential values of Y.

D. P(Y = 80) is a random variable.   (No, probabilities are fixed numbers, like 0.06).

21. If the confidence level of a confidence interval is stated as "approximately 95%", then
A. you should assume the true confidence level is between 90% and 99%.   (Can't assume that. A cow is approximately a horse).
B. you should assume that the sample size is more than 30.   (Can't tell. It depends on how non-normal the process is that produced the data).
C. you should assume the standard deviation is known.   (Usually the standard deviation is not known, so you shouldn't assume it.)
D. you should question the assumptions of the analysis. (yes, always)

22. The "Student t" distribution was discovered by
A. R.A. Fisher.
B. Isaac Newton.
C. William Gosset.   (this guy, the Guinness brewer)
D. Albert Einstein.

23.   The standard normal distribution has
A. mean=0 and variance=0.
B. mean=0 and variance=1.   (this one)
C. mean=1 and variance=0.
D. mean=1 and variance=1.

24. In the quality control example from class, "degrees of freedom" was
A. the mean of the t distribution.
B. the variance of the t distribution.
C. the sample size.
D. the sample size minus one.   (yes)

25. The reason for the extra variability in the "Student t" distribution compared to the standard normal distribution is that
A. use of $\bar{Y}$ instead of μ increases variability.
B. use of μ instead of $\bar{Y}$ increases variability.
C. use of $\hat{\sigma}$ instead of σ increases variability. (yes)
D. use of σ instead of $\hat{\sigma}$ increases variability.

26. Which statement is true?
A. Real processes produce real data. (yes)
B. Simulation models produce real data. (models aren't real)

C. Bayesian models produce real data. (models aren't real)
D. Frequentist models produce real data. (models aren't real)

27. Data that are collected in time sequence are best modeled as
A. randomly sampled from a population. (gag)
B. randomly sampled from a process. (yes, always)
C. independent but not identically distributed. (could be, but it would depend on the process)
D. identically distributed but not independent.(could be, but it would depend on the process).

28. Suppose that the confidence interval for μ is $248.1 < \mu < 250.3$. Then
A. we can conclude that $\mu = 250$.
B. we can conclude that $\mu \neq 250$.
C. we can conclude that $\mu > 249$.
D. we can conclude that $\mu > 248$. (this is the only logical one)

29. The critical value for the usual 95% confidence interval for μ is the _______ quantile of the appropriate t distribution.
A. 1.960
B. 1.645
C. 0.950
D. 0.975 (yes)

30. The term "α" usually refers to
A. the power of the test.
B. the probability of a Type I error. (yes, usually assumed to .05)
C. the probability of a Type II error.
D. 95%.

31. Which of the following questions best exemplifies the controversial nature of classical (frequentist) hypotheses testing?
A. The null hypothesis $\mu = \mu_0$ is likely false, *a priori*, so why test it at all? (yes)
B. How can you decide which prior distribution for μ to use? (this is not "classical")
C. If the null hypothesis is rejected, how can you state that $\mu \neq \mu_0$? (easy, this is not controversial at all. If the interval for μ excludes μ_0, then we can easily believe that $\mu \neq \mu_0$).
D. If the confidence interval for μ includes μ_0, why do you conclude that $\mu \neq \mu_0$? (We don't. We think that μ could be either more than, less than, or equal to μ_0 in this case.)

32. Which of the following is a valid null hypothesis?
A. H_0: $\bar{Y}$ =250 (hypotheses are about process that produces the data, not about the data themselves)
B. H_0: $\bar{y}$ =250 (hypotheses are about process that produces the data, not about the data themselves)
C. H_0: μ =250 (yes)
D. H_0: $\hat{\mu}$ =250 (hypotheses are about process that produces the data, not about the data themselves)

33. Why can't you state that the null hypothesis is true after finding an insignificant result?
A. Because the normality assumption is false. (not the reason, although nonnormality does indeed have effects on error probabilities)
B. Because the independence assumption is false. (see answer to A.)
C. Because the identical distribution assumption is false. (see answer to A.)
D. Because the Type II error rate is uncontrolled. (this is the reason. There is a large probability of a type II error when the true mean differs only slightly from the hypothesized mean).

34. When is the "model produces data" paradigm useful?
A. When the sample size, *n*, is large enough so that we can apply the law of large numbers.   (not bad, but we used the paradigm for much more than LLN)
B.   When the standard deviation is known, so that we can use the normal distribution instead of the t distribution. (kind of a narrow application, again)
C. When the data look as if they could have been produced in this way. (this is what was emphasized throughout)
D. When the parameters look as if they could have been produced in this way. (we understood that "model produces data" was about how real data appear, not how parameters might look. The latter is an application of Bayesian analysis)

35. Why is $\bar{Y}$ different from μ?
A. Because the model that is assumed to produce the data is different from the true model. (No, even under the true model $\bar{Y}$ is different from μ.)
B. Because $\bar{Y}$ is a random variable.   (Yes)
C. Because σ is unknown. (Irrelevant.)
D. Because different samples of $\{Y_i\}$ data give different values of μ. (no, μ is fixed)

36. In the in-class data collection, the process that our class data comes from is best described as
A. the population of TTU graduate students. (close, but population rankles)

B. the potential population of TTU graduate students.   (better, more like process)
C. the population of ISQS 5347 class attendees.   (population again…)
D. the potential population of ISQS 5347 class attendees.   (the most sensible answer)

37.   If the p-value is 0.04 then
A. the probability that H_0 is true is more than 0.04.   (p-value is not the probability that H_0 is true)
B. the probability that H_0 is true is less than 0.04. (“)
C. the probability that H_0 is true is equal to 0.04. (“)
D. the probability that the test statistic is more extreme than the observed test statistic is 0.04, assuming that the null hypothesis is true.   (this is the definition)

38. The p-value is calculated using
A. the cumulative distribution of the test statistic.   (yes)
B. the quantile of the distribution of the test statistic.
C. the likelihood function of the data. (that’s for MLEs)
D. the posterior distribution of the parameters, given the data.   (no, that’s Bayesian analysis)

39. The true power of the a test depends on
A. the true mean of the process. (yes)
B. the sample mean of the data.   (power is not from data, it’s from process)
C. the sample standard deviation of the data. (“)
D. the probability that the null hypothesis is true. (power assumes the alternative is true, so there is nothing about the null being true here)

40. If the distribution $p(y)$ that produces the data $\{Y_i\}$ has positive skewness, then the distribution of the t-statistic $T = \dfrac{\bar{Y} - \mu}{\hat{\sigma} / \sqrt{n}}$ has ______ skewness.
A. negative (yes)
B. positive
C. zero
D. leptokurtic

41.   If the data are sampled as iid from $N(\mu, \sigma^2)$, then the distribution of the test statistic $T = \dfrac{\bar{Y} - \mu_0}{\hat{\sigma} / \sqrt{n}}$ when $\mu \neq \mu_0$ is called
A. the standard normal distribution.
B. the noncentral normal distribution.
C. the standard t distribution.
D. the noncentral t distribution.   (yes)

42. All else fixed, what is the effect of sample size *n* on power?
A. If the null hypothesis is false, then larger *n* means larger power. (yes, look at the shift)
B. If the null hypothesis is true, then larger *n* means larger power. (power assumes a false null)
C. If the null hypothesis is false, then smaller *n* means larger power. (backwards)
D. If the null hypothesis is true, then smaller *n* means larger power. (see B)

43. All else fixed, what is the effect of process standard deviation (σ) on power?
A. If the null hypothesis is false, then larger σ means larger power. (backwards)
B. If the null hypothesis is true, then larger σ means larger power. (see 42B)
C. If the null hypothesis is false, then smaller σ means larger power. (yes, look at the shift)
D. If the null hypothesis is true, then smaller σ means larger power. (see 42B)

44. All else fixed, what is the effect of α on power?
A. If the null hypothesis is false, then larger α means larger power. (yes, the critical value is smaller so there is more area under the alternative distribution in the rejection region)
B. If the null hypothesis is true, then larger α means larger power. (see 42B)
C. If the null hypothesis is false, then smaller α means larger power. (backwards)
D. If the null hypothesis is true, then smaller α means larger power. (see 42B)

45. What does "significance level" refer to in a simulation study?
A. The proportion of simulated samples for which a true null hypothesis is rejected. (right)
B. The proportion of simulated samples for which a false null hypothesis is rejected. (power)
C. The proportion of simulated samples for which the null hypothesis is true. (no, the null is either true or false in all of our simulation studies, anyway, if that weeren't the the case, it still wouldn't be called sig level)
D. The proportion of simulated samples for which the null hypothesis is false. (see 45C)

46. A "statistically significant difference" means that
A. the test for the null hypothesis of no difference has been rejected. (yes)
B. you have failed to reject the null hypothesis of no difference. (that's insig)
C. the difference is practically important. (no – see distinction between practical and statistical sig)
D. the difference is practically unimportant. (see C)

47. If Y* is a single observation drawn using bootstrap sampling, then

A. E(Y*) = μ . (Not for bootstrap sampling)

B. E(Y*) = $\bar{Y}$. (No because $\bar{Y}$ is a random variable)

C. E(Y*) = $\bar{y}$. (yes)

D. E(Y*) = $(\bar{y} + \bar{Y})/2$. (See B.)