

BST 140.651

Problem Set 2

Due in class on September 29

- Problem 1. Suppose that we have a data set consisting of 100 ages (recorded in years). The mean, median, mode, standard deviation, variance, range, and geometric mean are computed. Suppose now that the ages are reexpressed in the terms of months (i.e. all ages are multiplied by 12). What would happen to all of these statistics if they were recalculated?
- Problem 2. Investigator A takes a random sample of 100 men ages 18 - 24 in a community. Investigator B takes a random sample of 1,000 such men. Answer the following. (A formal mathematical proof is not required. Give an intuitive explanation for your answers.)
- Which investigator will tend to get a bigger standard deviation for the heights of men in his sample? Or can it not be determined?
 - Which investigator will likely get a bigger standard error of the mean height? Or can it not be determined?
 - Which investigator is likely to get the tallest man? Or, are the chances about the same for both investigators.
 - Which investigator is likely to get the shortest man? Or are the chances about the same for both investigators.
- Problem 3. Recall the exponential problem from the last homework. Do the following:
- Plot the exponential density for $\beta = 1$.
 - Draw a random sample of 100 draws from the exponential distribution. Plot a histogram of these draws. Does it look like your answer to Part a? (It should.)
 - Draw 100 sample means of 20 observations from the exponential distribution. (Hence you will have to draw 2,000 exponentials.) For each of the 100 simulated sample means, subtract off the population mean and divide by its standard error. Plot a histogram of the 100 normalized means. What does it look like?
- Problem 4. A special study is conducted to test the hypothesis that persons with glaucoma have higher blood pressure than average. Two hundred subjects with glaucoma are recruited with a sample mean systolic blood pressure of 140mm and a sample standard deviation of 25mm. (Do not use a computer for this problem.)
- What is the estimated standard error of the mean? What is the difference in interpretation between the standard error of the mean and the standard deviation (25mm)? Explain your answer in words.
 - Construct a 95% confidence interval for the mean systolic blood pressure among persons with glaucoma. Do you need to assume normality. Explain.

- c. If the average systolic blood pressure for persons without glaucoma of comparable age is 130mm, is there statistical evidence that the blood pressure is elevated?
- d. Briefly discuss the limitations of this study design (if any) for drawing conclusions about the association between blood pressure and glaucoma.

Problem 5. A special study is conducted to test the hypothesis that persons with glaucoma have higher blood pressure than average, and 200 persons with glaucoma are recruited with mean systolic blood pressure of 140mm and standard deviation of 25mm. (Do not use a computer for this problem.)

- a. a. What is the standard error of the mean? What is the difference in interpretation between the standard error of the mean and the standard deviation of 25mm (Explain in words.)?
- b. b. Construct a 95% confidence interval for the mean systolic blood pressure among persons with glaucoma. Do you need to assume normality? Explain.
- c. c. If the average systolic blood pressure for persons without glaucoma of comparable age is 130mm, is there statistical evidence that the blood pressure is elevated for persons with glaucoma?
- d. d. Briefly discuss the limitations of this study design (if any) for drawing definitive conclusions about the association between blood pressure and glaucoma.

Problem 6. Suppose we wish to estimate the concentration $\mu\text{g}/\text{m}\ell$ of a specific dose of ampicillin in the urine. We recruit 25 volunteers and find that they have sample mean concentration of $7.0\ \mu\text{g}/\text{m}\ell$ with sample standard deviation $3.0\ \mu\text{g}/\text{m}\ell$. Let us assume that the underlying population distribution of concentrations is normally distributed.

- a. a. Find a 90% confidence interval for the population mean distribution.
- b. b. How large a sample would be needed to insure that the length of the confidence interval is $0.5\ \mu\text{g}/\text{m}\ell$ if it is assumed that the sample standard deviation remains at $3.0\ \mu\text{g}/\text{m}\ell$?

Problem 7. Perform a computer simulation. Generate 100 random variables from a standard normal distribution ($\mu = 0, \sigma = 1$).

- a. Graphically display the data. Does it look normal?
- b. Calculate a 95% confidence interval for the mean based on the 100 data points (don't assume σ is known). Does your confidence interval include 0? Must it always?

Problem 8. Use the computer to do this computer simulation problem. An experiment consists of choosing $n = 9$ men at random and measuring their heights. Assume that the heights of men are normal, with $\mu = 69$ inches and $\sigma = 3$ inches. Simulate the results of this experiment. Then find a 90% confidence interval for μ , assuming σ is unknown (using your simulated values of \bar{X} and S). Repeat this for a total of 20 samples, thus obtaining 20 confidence intervals.

- a. How many of the 20 intervals contain μ ?
- b. Would you expect all 20 of the intervals to contain μ ? Explain.
- c. Do all the intervals have the same width? Why or why not?
- d. Suppose you took 95% intervals instead of 90%. Would they be narrower or wider?
- e. How many of your intervals contain 72? 70? 69?
- f. Suppose you took samples of size $n = 100$ instead of $n = 9$. Would you expect more or fewer intervals to cover 72? 70? 69? What about the width of the intervals for $n = 100$? Would they be longer or shorter than $n = 9$?
- g. Suppose you calculated twenty 90% confidence intervals for real data. About how many would you expect to contain the true μ ? Could you tell which?
- h. Graphically display the 20 sample means. What is the sample variance of these 20 numbers? What is the theoretical population variance of these numbers?

SOME HELPFUL HINTS ABOUT SIMULATING DATA IN STATA

1. In Problem 7, I asked you to generate 100 random variables from a standard normal distribution. You can do this in STATA by

```
set seed 991
set obs 100
generate nran = invnorm(uniform())
```

Now, I know the last command seems weird and awkward. Here is what STATA is doing. STATA starts with a random number between 0 and 1 that is uniformly distributed which is what the command UNIFORM() does. Then, STATA transforms this random variable to make it normally distributed, which is what the command INVNORM does. Actually a course in mathematical statistics would show that if F is the cumulative distribution function of a standard normal distribution and u is a uniformly distributed random variable on the interval (0,1) then $X = F^{-1}(u)$ has a standard normal distribution. So, in summary, you can simulate a random normal deviate by taking the normal quantile of a random uniform deviate.

2. Here is a STATA hint on problem 8.

I have asked you to generate simulated values from a normal distribution with mean 69 and variance 9. But in the comment above I have only told you how to generate a standard normal random variable Z with mean 0 and variance 1. How can we get a normal variable with the specified mean and variance?

Well, remember that $Z = (Y - E(y))/SD$. So, $Y = E(Y) + SD * Z$ where SD is the population standard deviation and $E(Y)$ is the population mean.

So what this means is this: if we take standard normal random variables, multiply by the population standard deviation and add the population mean, we will get what we want.

So, all you have to do is

- (a) Generate standard normal random variables using `generate nran= invnorm(uniform())`.
- (b) Then say

Generate $Y = 69 + 3 * nran$

If you then type "list" you will see the data that you generated. Graph the data. Compare Y to $nran$. Is Y centered at "around" 69, $nran$ at around 0. They should be.

3. Note that for Problem 3 you can get a random exponential, by taking $-\log$ (the negative natural logarithm) of a random uniform deviate!

SOME HELPFUL HINTS ABOUT SIMULATING DATA IN SPLUS/R

1. In Problem 7 and 8, you can generate random variables from a normal distribution using Splus/R by using the *rnorm* command. This command requires three arguments (in this order):

n: the number of random variables you want to generate
mean: the mean of the normal distribution to sample from (default = 0)
sd: the standard deviation of the normal distribution to sample from
 (default = 1)

For Problem 7, to generate 100 random variables from the standard normal distribution, simply type "y <- rnorm(100)". For Problem 3, to get 100 random exponential deviates, use the function "rexp(100)".