

Problem Set (Week 4)

Econ 103

Lecture 12

Note: In the following five questions $X_1, X_2 \sim \text{iid } N(\mu, \sigma^2)$, $Y = (X_1 - \mu)/\sigma$, $Z = (X_2 - \mu)/\sigma$.

1. (a) What is the distribution of $X_1 + X_2$?

Solution: $X_1 + X_2 \sim N(2\mu, 2\sigma^2)$

- (b) Use R to calculate $P(X_1 + X_2 > 5)$ if $\mu = 5$ and $\sigma^2 = 50$.

Solution: In this case, $X_1 + X_2 \sim N(10, 100)$, hence

$$\begin{aligned} P(X_1 + X_2 > 5) &= 1 - P(X_1 + X_2 \leq 5) \\ &= 1 - P\left(\frac{X_1 + X_2 - 10}{10} \leq \frac{5 - 10}{10}\right) \\ &= 1 - \text{pnorm}(-0.5) \\ &\approx 0.6914625 \end{aligned}$$

Alternatively, we could use `1 - pnorm(5, mean = 10, sd = 10)`, which gives the same result.

- (c) Use R to calculate the 10th percentile of the distribution of $X_1 + X_2$.

Solution: `qnorm(p = 0.1, mean = 10, sd = 10)` gives -2.815516.

2. (a) What is the distribution of $Z/\sqrt{Y^2}$?

Solution: Since it is the ratio of a standard normal to the square root of an independent χ^2 random variable divided by its degrees of freedom (in this case one), $Z/\sqrt{Y^2} \sim t(1)$.

- (b) What value of c satisfies $P(-c \leq Z/\sqrt{Y^2} \leq c) = 0.95$?

Solution: By the symmetry of the t -distribution, it suffices to find the 97.5th percentile (this allocates 2.5% probability to the upper and lower tails). The command `qt(p = 0.975, df = 1)` gives 12.7062, so $c \approx 12.7$. Alternatively, we could have calculated the 2.5th percentile: `qt(p = 0.025, df = 1)` gives -12.7062.

- (c) How does the interval in part (b) compare to the corresponding interval for Z ?

Solution: Since Z is a standard normal RV, $P(-2 \leq Z \leq 2) \approx 0.95$. We see that the interval for a $t(1)$ RV is *much wider* than the corresponding interval for a standard normal. In other words, extreme outcomes are much more likely under the $t(1)$ distribution.

3. (a) What is the distribution of Y^2/Z^2 ?

Solution: This is the ratio of two independent χ^2 random variables, each divided by its degrees of freedom (in this case, one). Hence $Y^2/Z^2 \sim F(1, 1)$.

- (b) Use R to calculate the 95th percentile of the distribution of Y^2/Z^2 .

Solution: `qf(p = 0.95, df1 = 1, df2 = 1)` gives 161.4476

Lecture 13

4. In this question, you will use R to study the sampling distribution of the sample mean where we take as our population the heights of all students in Econ 103.
- (a) Load the class survey data used in R Tutorial # 2, extract the height column and assign it to a variable called `height`. Use `!is.na` to remove all missing values from `height`.

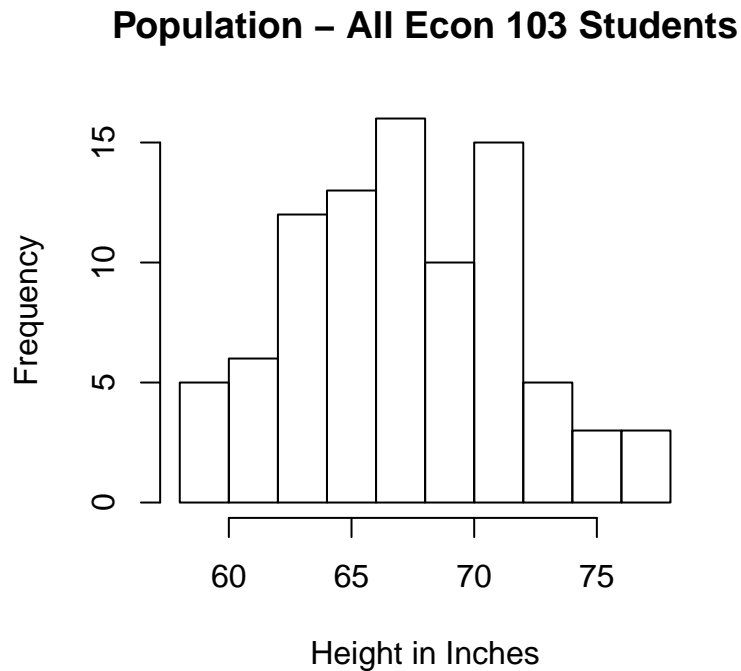
Solution:

```
data.url <- "http://www.ditraglia.com/econ103/old_survey.csv"
survey <- read.csv(data.url)
height <- survey$height
height <- height[!is.na(height)]
```

- (b) Make a histogram of height and calculate the mean height for students in the class. For the purposes of this exercise, these correspond to the *population*.

Solution:

```
hist(height, main = 'Population - All Econ 103 Students',  
      xlab = 'Height in Inches')
```



```
mean(height)  
## [1] 67.54545
```

- (c) Write a function that takes n as its only input and returns the sample mean of an iid random sample of size n drawn from the vector `height`. Call this function `x.bar.draw`. [Hint: use `sample` with `replace = TRUE`.]

Solution:

```
x.bar.draw <- function(n){

  sim <- sample(height, size = n, replace = TRUE)
  return(mean(sim))

}#END x.bar.draw
```

- (d) Test the function you wrote for part (d) by running it with $n = 10000$. What value do you get? Your answer should be approximately equal the population mean you calculated above. If it isn't, something is wrong with your code.

Solution:

```
x.bar.draw(10000)
## [1] 67.4819
```

Lecture 14

5. Textbook question: Chapter 7-18.

Solution: 7-18

The point of this question is non-response bias: the people who respond are not representative of the population as a whole. Note that P and P^* as defined in this question *do not correspond* to question 7-13. Our goal is to estimate the population proportion who will buy a computer. Using the table, we calculate the total number of people who will buy a computer as:

$$0.02 \times 40 + 0.04 \times 5 + 0.1 \times 3 + 0.2 \times 2 = 1.7 \text{ million}$$

which corresponds to a fraction $\pi^* = 1.7/50 = 0.034$. Again using the table, the number of people who will buy a computer *among the sub-population who would respond* can be calculated as:

$$0.02 \times 7 + 0.04 \times 1 + 0.1 \times 1 + 0.2 \times 1 = 0.48 \text{ million}$$

which corresponds to a fraction $\pi = 0.48/10 = 0.048$. The point is that $\pi \neq \pi^*$. In other words, the proportion of people who would buy a computer *differs* across

people who would and would not respond to the phone survey. The estimator P is based on calling 1000 people chosen at random and recording responses for *only those who reply*. The proportion of people who will reply is $10/50 = 1/5$. Thus, P will end up with a sample size of approximately $n = 200$ individuals. These individuals correspond to the sub-population for which the proportion who would buy a computer is $\pi^* = 0.048$. In contrast, the estimator P^* is based on calling $n^* = 100$ people chosen at random and then following up with these people repeatedly until *all of them respond*. Thus, P^* draws from the *full population*, in which a proportion $\pi^* = 0.034$ of people will buy a computer. The *true parameter* is π^* since we want to estimate the *overall* fraction of people who will buy a computer, *not* the fraction of people who would buy a computer among those who are likely to respond to a telephone survey. Hence bias is calculated *relative to* π^* . Variance is calculated *relative to the mean of each sampling distribution*. For P this mean is π while for P^* it is π^* . That is:

$$\begin{aligned} MSE(P) &= \text{Bias}(P)^2 + \text{Var}(P) = (E[P] - \pi^*)^2 + E[(P - \pi)^2] \\ &= (\pi - \pi^*)^2 + E[(P - \pi)^2] \\ &= (\pi - \pi^*)^2 + \pi(1 - \pi)/n \\ MSE(P^*) &= \text{Bias}(P^*)^2 + \text{Var}(P^*) = (\pi^* - \pi^*)^2 + E[(P^* - \pi^*)^2] \\ &= E[(P^* - \pi^*)^2] = \pi^*(1 - \pi^*)/n^* \end{aligned}$$

The estimator P^* does not have any bias because of the follow-ups to ensure that everyone in the original random sample responds. However, since it is based on a smaller sample, we would expect it to have a higher variance. The question is how this trade-off comes out in the expressions for MSE. To find out, we simply plug in the values $\pi^* = 0.034$, $\pi = 0.048$, $n = 200$ and $n^* = 100$. We find $MSE(P^*) \approx 0.000328$ and $MSE(P) \approx 0.000424$.

Lecture 15-16

6. For this question assume that we have a random sample from a normal distribution with unknown mean but *known* variance.
 - (a) Suppose that we have 36 observations, the sample mean is 5, and the population variance is 9. Construct a 95% confidence interval for the population mean.

Solution: Since the population variance is 9, the population standard deviation is 3. Hence, the desired confidence interval is $5 \pm 2 \times 3/\sqrt{36} = 5 \pm 1 = (4, 6)$

- (b) Repeat the preceding with a population variance of 25 rather than 9.

Solution: The population standard deviation becomes 5 so the confidence interval becomes $5 \pm 2 \times 5/\sqrt{36} = 5 \pm 10/6 \approx (3.3, 6.7)$

- (c) Repeat the preceding with a sample size of 25 rather than 36.

Solution: $5 \pm 2 \times 5/\sqrt{25} = 5 \pm 2 = (3, 7)$

- (d) Repeat the preceding but construct a 99% rather than a 95% confidence interval.

Solution: Again we use R to get the appropriate quantile:

```
SE <- 5/sqrt(25)
ME <- qnorm(1 - 0.01/2) * SE
Lower <- 5 - ME
Upper <- 5 + ME
c(Lower, Upper)
## [1] 2.424 7.576
```

- (e) How would you construct the confidence interval if the variance is *unknown*? Explain what estimator you will use to approximate the variance and what distribution you will use to construct the interval.

Solution: Since we do not know the population variance (standard deviation), we will use the sample variance (standard deviation) to construct the relevant random variable as in the lecture note. Then we will use t distribution with degrees of freedom $n - 1 = 36 - 1$ to get the associated quantiles.