Moodle Questions Chapter 8

Homer White

Thursday, June 19, 2014

# What's the Parameter of Interest?

In this question we describe several studies and state a Research Question. For each study, decide which of the five Basic Parameters is of interest. One or more of the studies may not involve one of the Basic Five Parameters.

Study A: In the tigerstats package we find the beans data frame:

data(beans)  
View(beans)  
help(beans)

This was a repeated measures-experiment performed at UC-Davis; fifteen students participated. Each student was asked to place as many beans into a cup as he/she could, in 15 seconds. Each student performed this task once with the dominant hand, and once with the non-dominant hand, but the order of performance was randomized. Assume that the subjects in the experiment are the fifteen students. (*Reminder about terminology*: your dominant hand is the hand you use the most.)

The Research Question was:

*Which hand is more dextrous: the dominant hand, or the non-dominant hand?*

Parameter: Differences of Means .

Study B: In the tigerstats package we find the data frame napkins:

data(napkins)  
View(napkins)  
help(napkins)

this was an observational study performed at Georgetown College. A small team of student researchers visited the Cafe several times during the lunch hour and recorded the sex of each student in the study as well as how many napkins each student used during the meal.

The Research Question was:

*Who uses more napkins, on average, during lunch: a GC male or a GC female?*

Parameter: Mean of Differences .

Study C: In the tigerstats package we find the data frame hair\_and\_act:

data(hair\_and\_act)  
View(hair\_and\_act)  
help(hair\_and\_act)

This was an observational study performed at Georgetown College. A pair of student researchers attempted to make a random selection of fellow students. Each person in the study was asked his or her sex (male, female), and the student researchers also classified each person's hair-color as "dark" or "light".

The Research Question was:

*Who is more likely to have light-colored hair? A GC male or a GC female?*

Parameter: Difference of Proportions .

Study D: Assume that it is known that the mean ACT score of all student at the University of Kentucky is 24. The hair\_and\_act data frame records the ACT score of a sample of Georgetown College students.

We are interested in the Research Question:

*Is the mean ACT score of all GC students the same as that of all UK students, do the two schools differ in their mean ACT scores?*

Parameter: One Population Mean .

Study E: It is known that 60% of all Ohio State students plan to vote for the Democratic candidate in the next presidential election. In a survey of 200 randomly-selected Georgetown College students, it is found that 101 of them plan to vote for the Democratic candidate.

The Research Question is:

*Is the level of support for the Democratic candidate lower at GC than it is at OSU?*

Parameter: One Population Proportion .

Study F: In the m111survey data frame, the variable **sex** records the sex of each student in the survey, and the variable **seat** tells where the students prefers to sit in a classroom: Front, Middle or back.

We are interested in the knowing whether GC males and GC females differ in their seating preferences.

Parameter: None of the Basic Five parameters are involved.

# Defining Parameters Accurately: Difference of Two Means

In a random sample of 25 males in the U.S., the mean weight is 170 pounds, with a standard deviation of 20 pounds. In an independent random sample of 36 females in the U.S., the mean weight is 130 pounds, with a standard deviation of 15 pounds. We want to know how much taller guys are, on average, than gals are.

Which of the following is the most accurate way to describe the parameter of interest?

Option A:

Define:

the mean weight of all males in the U.S.

the mean weight of all females in the U.S.

We are interested in .

Option B:

Define:

the mean weight of the 25 males in the sample.

the mean weight of the 36 females in the sample.

We are interested in .

Feedback: Parameters are features of the population, not the samples. Everyone knows what's in the samples, because we HAVE the samples on hand!

Option C:

Define:

all males in the U.S.

all females in the U.S.

We are interested in .

Feedback: As defined, and are not parameters: they are not even numbers; they are just sets of people. We need to mention the *weights*!

Option D:

Define:

the mean for males.

the mean for females.

We are interested in .

Feedback: This is too vague! Are we talking about the means of the samples, or the means of populations? Which populations: U.S. males, Japanese males, etc.??

# Defining Parameters Accurately: The Mean of Differences

In package abd, check out the data frame SpiderSpeed:

require(abd)  
data(SpiderSpeed)  
View(SpiderSpeed)  
help(SpiderSpeed)

The researchers took 32 *Tidarren* spiders and recorded how fast each one could run, in centimeters per second. Then they amputated a [pedipalp](http://simple.wikipedia.org/wiki/Pedipalp) from each one of the spiders, and again recorded how fast each spider could run.

The researchers wanted to know if the spiders could run faster, on average, without the hindrance of a pedipalp.

Which of the following is the most accurate way to describe the parameter of interest?

Option A:

The researchers were interested in , the mean difference in running speed (after amputation minus before amputation) for all *Tidarren* spiders.

Feedback: Correct. The population is named specifically, and it is clear which speed is subtracted from which to make the difference.

Option B:

The researchers were interested in , the mean difference in running speed (after amputation minus before amputation) for the 32 *Tidarren* spiders in the sample.

Feedback: Parameters refer to a population, not to sample data.

Option C:

The researchers were interested in , the mean difference in running speed for all *Tidarren* spiders.

Feedback: What difference are we talking about? Speed before amputation minus speed after? Or speed after minus speed before? Or some other pair of speeds?

Option D:

The researchers were interested in , the mean running speed for all *Tidarren* spiders.

Feedback: That looks like one mean, not a mean of differences at all.

Option E:

The researchers were interested in , the mean of all *Tidarren* spiders.

Feedback: The mean WHAT of spiders? Spiders are animals, not numbers, so they don't have a "mean".

# Expected Value and Standard Deviation for a Sample Mean

Remember the imagpop data frame in the tigerstats package. Here is some information on the heights for this population:

## min Q1 median Q3 max mean sd n missing  
## 53.8 64.7 67.5 70.3 80.2 67.53 3.907 10000 0

As you can see, the mean height is 67.53 inches, and the standard deviation is 3.907 inches.

George is about to take a simple random sample of 25 people from this population. He will then compute the sample mean of the 25 heights. Compute the expected value and the standard deviation of , and choose all correct options below.

The expected value of is 67.53 inches, and the standard deviation of is about 0.78 inches.

The expected value of is 67.53 inches, and the standard deviation of is about 3.907 inches.

George's sample mean should turn out to be about 67.53 inches, give or take 0.78 inches or so.

George's sample mean should turn out to be about 67.53 inches, give or take 3.907 inches or so.

3.907014/sqrt(25)

## [1] 0.7814

# Defining Parameters in an Experiment:Pseudoscorpions

The data for this study is in data frame Pseudoscorpions, from package abd:

require(abd)  
data(Pseudoscorpions)  
View(Pseudoscorpions)  
help(Pseudoscorpions)

From the text *Analysis of Biological Data*, by Michael C. Whitlock and Dolph Schluter:

Pseudoscorpions of the species *Chordylochernes scorpioides* live in tropical forests where they ride on the backs of harlequin beetles to reach rotting figs where they feed. Females of the species are promiscuous and mate with multiple males over short lifetimes. It is unclear what advantages there are for a female to mate multiple times, because the males don't help to care for her young, and mating just once provides all the sperm she needs to fertilize her eggs.

One possible advantage is that the sperm of some males is genetically incompatible with a given female and, by mating multiple times, a female increases the chances of mating with at least one male whose sperm is compatible with her. To investigate this idea, Newcomer et al. (1999) recorded the number of successful broods by female Pseudoscorpions randomly assigned to one of two treatments. One group of females was each mated to two different males (DM), whereas females in the other group were each mated twice to the same male (SM). By mating each female twice, the same amount of sperm was provided in both treatments, but the DM females received genetically more diverse sperm than the SM females.

The Pseudoscorpions data frame gives the results of this study.

From the above description, it appears that the investigators were interested in the following Research Question:

*On average, do female pseudoscorpions who mate with two males produce a larger number of successful broods than do female pseudoscorpions who mate with only one male?*

From the description of how the data was collected, you can see that this was a randomized experiment with two treatment groups. The subjects were the female Pseudoscorpions, and the treatments were:

* mate twice, with one male (SM);
* mate with two males, once each (DM).

We also see that this was a completely randomized design: no blocking or repeated measures.

Assume that the subjects in the study may be thought of as a random sample from the population of all female Pseudoscorpions. What's the best was to define the parameters of interest in this experiment?

Option A:

Define

mu-1 = the mean number of successful broods produced by all female Pseudoscorpions in the population, if all of them were to mate with two males, once each.

mu-2 = the mean number of successful broods produced by all female Pseudoscorpions in the population, if all of them were to mate twice with one male.

We are interested in mu-1 - mu-2.

Option B:

Define:

mu-1 = the mean number of successful broods produced by all female Pseudoscorpions in the study, if all of them were to mate with two males, once each.

mu-2 = the mean number of successful broods produced by all female Pseudoscorpions in the study, if all of them were to mate twice with one male.

We are interested in mu-1 - mu-2.

Feedback: This would be correct if the Pseudoscorpions in the study were thought not to be representative of the general population in some way relevant to the variables being measured. However, the female Pseudoscorpions in the study were said to be like a random sample from the population, so probably they ARE fairly representative of the population. The results of this experiment can be applied to the general population, so we should indicate that in the way we define our parameters!

Option C:

Define:

p-1 = the proportion of all female Pseudoscorpions in the population who would produce a successful brood, if all of them were to mate with two males, once each.

p-2 = the proportion of all female Pseudoscorpions in the population who would produce a successful brood, if all of them were to mate twice with a single male.

We are interested in p-1 - p-2.

Feedback: The response variable in this experiment is the numerical variable **successful.broods**, which measures the *number of times* each female produced a successful brood during her lifetime. In order for proportions to be involved, we would have had to have a response variable that indicated only *whether or not* each female produced at least one successful brood.

Option D:

Define

mu-1 = the mean for females who mate with two males once each.

mu-2 = the mean for females who mate with a single male twice.

We are interested in mu-1 - mu-2.

Feedback: The mean WHAT? And what does "all" refer to: all females in the study, or all females in the population? This sort of talk is acceptable only as "shorthand", AFTER you have defined the parameters precisely.

# Defining Parameters in an Experiment: The Attitudes Study

Review the attitudes data frame (as discussed in Chapter 6 Course Notes and slides):

data(attitudes)  
View(attitudes)  
help(attitudes)

Suppose that the subjects in the study can be thought of as like a random sample from the population of all Georgetown College students, and that we are interested in the Research Question:

*who tends to recommend longer sentences, on average: a GC student who thinks the defendant is white, or a GC student who thinks the defendant is black?*

Which of the following options is the best way to define the parameters of interest in this study?

Option A:

Define

mu-1 = the mean sentence recommended by all GC students, if all of them could have been given a form in which the defendant's suggested race was Black.

mu-2 = the mean sentence recommended by all GC students, if all of them could have been given a form in which the defendant's suggested race was White.

We are interested in mu-1 - mu-2.

Option B:

Define

mu-1 = the mean sentence recommended by all GC students in the study, if all 267 of them could have been given a form in which the defendant's suggested race was Black.

mu-2 = the mean sentence recommended by all GC students in the study , if all of them could have been given a form in which the defendant's suggested race was White.

We are interested in mu-1 - mu-2.

Feedback: This would be correct if we had reason to believe that the students in the study were not representative of some larger population, but in the lead-up info we learned that the subjects were like a random sample from the population of all GC students.

Option C:

Define

mu-1 = the mean for a Black defendant.

mu-2 = the mean for a White defendant.

We are interested in mu-1 - mu-2.

Feedback: This is much too vague, useful only as "shorthand" after you have already defined the parameters precisely.

Option d:

Define:

p-1 = the proportion of all GC students who would recommend a sentence for a White defendant.

p-2 = the proportion of all GC students who would recommend a sentence for a black defendant.

We are interested in p-1 - p-2.

Feedback: The response variable is **sentence**, which is numerical, so proportions are not at issue.

# What Changes from Sample to Sample?

Today Jill plans to take a simple random sample of size 50 form a population, and to compute the sample mean. She will then replace the 50 items that she sampled.

Tomorrow she will take another simple random sample of of size 50, and compute the sample mean from this new sample.

Some of the following quantities will stay the same from one day to the next, and others are liable to change. Select all of those that are liable to change.

The mean of the population.

The expected value of x-bar.

The standard deviation of the population.

The standard deviation of x-bar.

The size of the sample.

x-bar

The standard deviation of the sample

The standard error of x-bar

# Computing a standard error

You take a sample of size 64 from a population, and you find that sample mean x-bar is 25. The SD of the sample is 4. Unknown to you, the SD of the population is 5.

Which of the following bits of R-code gives you the value of the standard error of x-bar?

Option A:

4/sqrt(16)

## [1] 1

Option B:

5/sqrt(16)

## [1] 1.25

4/16

## [1] 0.25

5/16

## [1] 0.3125

25/sqrt(16)

## [1] 6.25

4/sqrt(25)

## [1] 0.8

Feedback for wrong answers: You need (SD of the sample)/sqrt(sample size), which is 4/sqrt(16), which works out to 1.

# Computing a Standard Error: II

In the m111survey data frame, we have the fastest speed ever driven for a sample of GC students.

Here are some results from favstats():

favstats(~fastest,data=m111survey)

## min Q1 median Q3 max mean sd n missing  
## 60 90.5 102 119.5 190 105.9 20.88 71 0

One of the following statements is true (up to some round-off error). Which one?

The sample mean is 105.9 and the standard deviation of x-bar is 2.478.

The sample mean is 105.9 and the standard error of x-bar is 2.478.

The sample mean is 105.9 and the standard error of x-bar is 0.294.

The population mean is 105.9 and the standard error of x-bar is 2.478.

20.8773/sqrt(71)

## [1] 2.478

# Interpreting Standard Errors

You would like to estimate the mean height of all GC students, so you consult the m111survey data frame:

favstats(~height,data=m111survey)

## min Q1 median Q3 max mean sd n missing  
## 51 65 68 71.75 79 67.99 5.296 71 0

As you can see, the sample mean x-bar is 67.99 inches. But you know that, due to chance error in the sampling process, this is unlikely to be exactly mu, the population mean.

So you also compute the standard error of x-bar:

5.296414/sqrt(71)

## [1] 0.6286

What's a good way to summarize your results to other people?

I figure that the mean height of all GC students is somewhere around 67.99 inches, give or take 0.63 inches or so for chance error in the sampling process.

The mean height of all GC students has to be within 0.63 inches of my sample mean of 67.99 inches.

The sample mean is somewhere around 67.99 inches, give or take 0.63 inches or so for chance error in the sampling process.

# The 68-95 Rule for Estimation

Back to those heights at Georgetown College:

favstats(~height,data=m111survey)

## min Q1 median Q3 max mean sd n missing  
## 51 65 68 71.75 79 67.99 5.296 71 0

Using the 68-95 Rule for Estimation, construct a rough 95%-confidence interval for mu, the mean height of all GC students. Report the lower bound of this interval below. Make sure your answer is correct to at least two decimal places.

67.98662-2\*5.296414/sqrt(71)

## [1] 66.73