BIOL 220 Problem Set 01: Biostats Basics

Answer Key

Populations, estimates, sources of variation

Populations, estimates, sources of variation

1. In one or two sentences, what is the difference between a sample and a population? [1 point]

Possible answer

A population is the collection of all possible sampling units that we could ever collect data about. We will never be able to measure/collect data about the entire population. Instead, we are only able to collect data about a sample. A sample is a finite subset of all the units from the population

2. If you're trying to study differences in species richness across locations, taxonomic misidentification can occur, what category of error or source of variation is this? [1 point]

   - Bad luck
   - Random variation
   - Scientific
   - Measurement error

Answer

Measurement error

3. What/which is an estimate (choose all that apply)? [1 point]

   - A property of a population
   - A statistic calculated from a sample

- A best guess of the value of a parameter
- The average body mass of 20 Kōlea birds

Answer

- A statistic calculated from a sample
- A best guess of the value of a parameter
- The average body mass of 20 Kōlea birds

Random sampling for a clinical trial

Suppose you are designing a clinical trial for a new drug and you have enough budget to afford a trial with 100 participants randomly selected from the total (finite) population of adults living in the fictional country of El Dorado. Assume no one is added or subtracted (dies) from the population during sampling. El Dorado is divided into 10 regions that vary widely in their population size. To randomly select participants, you are given a list of all living adults in El Dorado arranged by Region and alphabetically by surname (last name) within Region like this:

Region	Name	ID Number
A	Aaronson, A	1
⋮	⋮	⋮
A	Zykowski, Z	65184
B	Aaronson, A	65185
⋮	⋮	⋮
B	Zykowski, Z	187706
C	Aaronson, A	187707
⋮	⋮	⋮
C	Zykowski, Z	245906

Each adult also has unique ID Number determined by their Region and Name.

Consider the following four sampling procedures to select participants, then answer questions 4 & 5.

A. Choose the first 100 individuals in the list B. Choose the first 10 individuals from each Region C. Use a random number generator to select 10 ID Numbers from each region D. Use a random number generator to select 100 ID Numbers from the country population

4. Which of the sampling procedures (A, B, C, or D) is most likely to produce a *biased* estimate of the true population response to the drug? [1 point]

Answer

A. Choose the first 100 individuals in the list

5. After starting the trial on participants, you learn that a technician accidentally used a nonrandom sampling procedure that is likely to bias your estimate. The technician suggests that you ask for more funding to increase the sample size to 200 while continuing to use the same biased sampling procedure. If you adopt their suggestion, what is most likely true about the resulting estimate from the sample of 200? [1 point]

    - More precise and less biased
    - More precise and equally biased
    - Less precise and less biased
    - Less precise and equally biased

Answer

More precise and equally biased

Randomly assign individuals to "treatments"

Randomly assign students to project groups.

6. [2 points] As discussed in the Syllabus, the Group Project is a major part of this class. All students in your section will be divided into groups of 4 to 5 students each. For this activity, you will **randomly assign students to groups**. Each student must be assigned to one and only one group. Assume there are 23 students in your section, and you can simply "name" each student "Student 01", "Student 02", ..., "Student 23". You can use any method you want to assign groups, as long as it is truly random. You need to upload your work to show how your randomization procedure works. This could be any of the following:

    - a picture of a piece of paper showing randomization
    - a Google Sheet/Doc with randomization procedure

- a short video describing and showing your randomization procedure
- an *R* script
- something else, as long as it's clear what you did.

You cannot just send the "names" of students in each group; you must show your randomization procedure.

A possible R-based answer

```
# create a vector of numbers 1 to 23 representing students
students <- 1:23

# create a vector of group names
groups <- c(rep("group A", 5), rep("group B", 5), rep("group C", 5),
            rep("group D", 4), rep("group E", 4))

# if we line up the student "names" (aka numbers) and the group names in a
# data.frame, that will tell us which student belongs to which group:

group_memebership <- data.frame(student = students, group = groups)
head(group_memebership)
```

```
  student   group
1       1 group A
2       2 group A
3       3 group A
4       4 group A
5       5 group A
6       6 group B
```

```
# but that is *not* random, to make random we need to shuffle the order of
# either the students or the groups; let s shuffle the groups

# the sample function used like this will randomly shuffle the order of the
# groups vector
rand_groups <- sample(groups)

group_memebership <- data.frame(student = students, group = rand_groups)
head(group_memebership)
```

```
  student   group
1       1 group A
2       2 group B
3       3 group B
4       4 group E
5       5 group C
6       6 group D
```

Types of variables

Imagine you have a dataset from a different clinical trial that looks like this:

patient_ID	disease_status	treatment	age_in_years
4oncnyz0	tested positive	control	44
e01i28hz	tested positive	low_dose	38
gis9nqlo	tested positive	medium_dose	53
5gijw0g0	tested negative	high_dose	47
5y72ehlm	tested positive	control	55
r60yuoz4	tested negative	low_dose	58
p3fhvv4q	tested negative	medium_dose	59
7dmeyvy5	tested negative	high_dose	60

(Assume the above shows only a small subset of the full data)

Types of variables

7a. Which of the following is a *categorical* variable?
7b. Which of the following is a *numerical* variable?
7c. Which of the following is a *ordinal* variable?

- `patient_ID`
- `disease_status`
- `treatment`
- `age_in_years`

Questions 7a-c are worth [3] points in total.

Answers

7a. Which of the following is a *categorical* variable?

- `disease_status`
- `patient_ID` is not correct: it is true that `patient_ID` takes discrete values like a categorical variable, but in this situation `patient_ID` is not data, it is just a "book keeping" column to help us organize the data

7b. Which of the following is a *numerical* variable?

- `age_in_years`

7c. Which of the following is a *ordinal* variable?

- `treatment`