

STATISTICAL STUDY PROPOSAL FORMAT

Goal. State what population parameter you are trying to estimate, and stipulate the desired confidence level α .

Population. Describe the intended population.

Sampling frame. Describe the sampling frame you will be using for taking your sample. Describe in detail how exactly you will select individuals for measurement: will you be picking random items from a large list? Or measuring convenient individuals at randomly selected times and locations?

Sampling method. Imagine yourself performing the data collection, and describe in minute detail what you will do and how. State the intended sample size. The main idea behind this section is to provide a description of a procedure so detailed, that anyone can repeat your experiment just the way you conducted it. Another reason to be thorough here is to lay bare all the limitations of your chosen sampling technique.

Discussion. Provide constructive criticism of your proposal. In other words, enumerate the flaws which are present by design. There is no shame in finding flaws: almost every time you can blame them on either the lack of resources or an ethical dilemma or whatnot. There is, however, a great amount of shame in being willfully blind to the consequences of your own choices, and biases inherent in your sampling procedure. Think of it this way: if you don't point out the built-in flaws now, someone will surely bring them up later in an attempt to discredit your study.

STATISTICAL STUDY PUBLICATION FORMAT

Data analysis. Provide full raw data, either as a separate file or inline. State the relevant sample statistics. State the confidence interval and, if appropriate, the margin of error.

Discussion. This is your time to be both creative and reflective. Some of the points you may want to bring up:

- Mention all the things that went wrong, but also pat yourself on the back for all the things that went right.
- If you encountered unexpected difficulties and had to deviate from your proposal in any way, describe all of that in detail, and explain how that might have affected the data and the conclusion.
- Re-examine the assumptions and discuss how well they were met in your sample. For proportion studies, you would want a large sample and a much larger population, especially if the point estimate is extreme. For mean and standard deviation studies, you would like to see an approximately normal population or else a very large sample size.

EXAMPLE: CI FOR THE MEAN PROPOSAL

Goal. The purpose of this study is to produce a 99% confidence estimate for the mean **girth of a tree** in Alhambra Triangle neighborhood of Sacramento CA.

Population. Every tree within the shown area will be considered for the study, given that

- the height of a tree is 6 feet or more
- the tree grows on a patch of land open to general public

Sampling frame. Instead of selecting individuals, we will be selecting geographical locations, and then measuring nearby individuals. Each geographical location will be represented by a point on the map (x and y coordinates) where we will look for an individual to measure.

Sampling method: choosing location. We will take a cluster sample by selecting 5 random locations on the map of the Alhambra Triangle, walking over to these spots, and measuring a number of trees at each location. R will be used to generate random numbers.

To pick a location, we will print the map of the Alhambra Triangle, measure its dimensions, and then roll two random numbers: an x coordinate and a y coordinate. We will then mark this spot on the map with a circle.

For example, we will use a 12 by 6 cm map image. We need to roll 2 random numbers which are uniformly distributed in their respective ranges:

```
runif(1, 0, 12) # roll 1 number between 0 and 12
```

```
[1] 3.454436
```

```
runif(1, 0, 6) # roll 1 number between 0 and 6
```

```
[1] 2.633399
```



Map generated by Google

If the random location is within the Alhambra Triangle area, then we mark the location on the map, as seen above. If the random location is outside of the highlighted area, or overlaps with a formerly chosen location, then we discard both coordinates and roll again.

Sampling method: individual measurement. Once we have all 5 random locations marked on the map, we will travel to each one and attempt to make 8 individual measurements at each location. If the location is on an intersection, then we will take two measurements from each of the four corners. If the location is in the middle of a street, we will take 4 measurements from each side of the street. In all other circumstances we will measure 8 individual trees which are closest to the chosen coordinates.

To measure a single tree, we will use a metric measuring tape. The girth will be measured at 3 feet above the ground, and recorded as a single number measured in millimeters. If the tree is too thick to be measured with the tape, NA will be entered.

Discussion.

- Chosen locations are large and imprecise, so picking the 8 closest trees may be difficult without resorting to a sample of convenience.
- Some locations, like the areas under the freeway, may be inaccessible or lacking in trees, resulting in a reduced sample size.

EXAMPLE: CI FOR THE MEAN PUBLICATION

Data analysis. Raw data is attached in a separate file: **sample-data-mu-estimate.csv**

Adjusted sample size: $n = 34$

Sample mean (point estimate): $\bar{x} = 613.0588$ mm

Sample standard deviation: $s = 296.2632$ mm

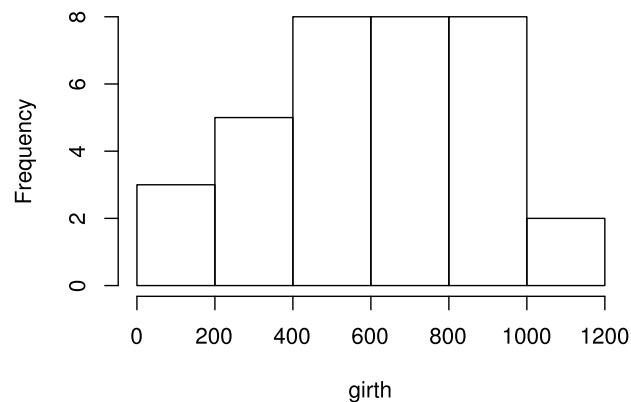
99% CI for population mean using t_{33} as the sampling distribution, in mm:

(474.1845, 751.9331)

Margin of error: 138.8743 mm

Discussion. The sample size of 34 is rather small, but the sample appears to be approximately normal: the histogram is somewhat mound-shaped and symmetric, and the Q-Q plot shows that the bulk of the sample data follows normal quantiles.

Histogram of girth



Normal Q-Q Plot

