Lecture 3: Sample Estimates Versus Population Measures

STAT 515 — Statistical Methods I

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Outline

- Population Versus Sample
- Random Sampling
- Sample Distribution Versus Underlying “Population Distribution"

Summarizing Data — Measures of Central Tendency

- **mean (average)**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **median** = middle observation after ordering data
- **mode** = most frequent observation

Summarizing Data — Measures of Dispersion (or Spread)

- **Range** = difference between largest and smallest values
- **Variance** = "average" of the squared differences of observations from the sample mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard deviation**

$$s = \sqrt{s^2}$$

Notes on s

- The bigger s is, the more variability there is
- s measures the spread about the mean
- s can equal 0 only if there is no spread
  - All n observations have the same value
- The units of s are the same as the units of the data (for example, mm Hg)
- Often abbreviated *SD* or *sd*
- s^2 is the best estimate from the sample of the population variance σ^2; s is the best estimate of the population standard deviation σ

Population Versus Sample

- *Sample*: a subset (part) of a larger group (population) from which information is collected to learn about the larger group
  - For example, sample of blood pressures $n =$ five 18-year-old male college students in the United States

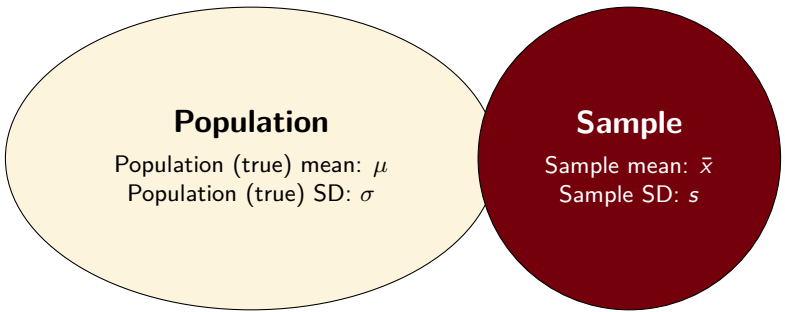
- *Population*: the entire group for which information is wanted
  - For example, the blood pressure of all 18-year-old male college students in the United States

Random Sampling

- For studies it is optimal if the sample which provides the data is representative of the population under study
  - Certainly not always possible!
- For this term, we will make this assumption unless otherwise specified
- One way of getting a representative sample: simple random sampling
  - A sampling scheme in which every possible sub-sample of size n from a population is equally likely to be selected
  - How to do it? More detail later, but think of the "names in a hat" idea

Population Versus Sample

- The sample summary measures (mean, median, sd) are called statistics, and are just estimates of their population (process) counterparts
- Assuming the sample is representative of the population from which it is taken (for example, a randomly drawn sample) these sample estimates should be "good" estimates of true quantities



Population Versus Sample

- For example, we will never know the population mean μ but would like to know it
- We draw a sample from the population
- We calculate the sample mean $\bar{x}$
- How close is $\bar{x}$ to μ?
- Statistical theory allow us to estimate how close $\bar{x}$ is to μ using information computed from the same single sample we use to estimate $\bar{x}$

SD: Why Do We Divide by n-1 Instead of n?

- We really want to replace $\bar{x}$ with μ in the formula for s^2

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- Since we don't know μ, we use $\bar{x}$

- But generally, $\sum_{i=1}^{n}(x_i - \bar{x})^2$ tends to be smaller than $\sum_{i=1}^{n}(x_i - \mu)^2$
  - To compensate, we divide by a smaller number: $n-1$ instead of n

$n - 1$

- $n - 1$ is called the *degrees of freedom of the variance* or *SD*
- Why?
  - The sum of the deviations is zero
  - The last deviation can be found once we know the other $n - 1$
  - Only $n - 1$ of the squared deviations can vary freely
- The term *degrees of freedom* arises in other areas of statistics
- It is not always $n - 1$, but it is in this case

Why SD as Measure of Variation

- Why note use the range of the data for example?
  - Range = maximum - minimum
- What happens to the sample maximum and minimum as sample size increases?
  - As it turns out, as sample size increases, the maximum tends to increase, and the minimum tends to decrease: Extreme values are more likely with larger samples!
  - This will tend to increase the range systematically with increased sample size

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

- Target population:

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

- Target population: all weekday lunch trays served during the semester.
- Sample:

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

- Target population: all weekday lunch trays served during the semester.
- Sample: the 60 selected trays.
- Observation unit:

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

- Target population: all weekday lunch trays served during the semester.
- Sample: the 60 selected trays.
- Observation unit: one lunch tray.
- Response:

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

- Target population: all weekday lunch trays served during the semester.
- Sample: the 60 selected trays.
- Observation unit: one lunch tray.
- Response: grams discarded.
- Parameter:

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

- Target population: all weekday lunch trays served during the semester.
- Sample: the 60 selected trays.
- Observation unit: one lunch tray.
- Response: grams discarded.
- Parameter: the population mean grams of waste per tray.
- Sample statistic:

Exercise

A university sustainability office weighs food discarded from 60 lunch trays selected from three dining halls. The mean waste is 112 grams per tray. The office wants to estimate food waste for all weekday lunch trays served during the semester.

(a) Identify the target population, sample, observation unit, response variable, population parameter, and sample statistic.

- Target population: all weekday lunch trays served during the semester.
- Sample: the 60 selected trays.
- Observation unit: one lunch tray.
- Response: grams discarded.
- Parameter: the population mean grams of waste per tray.
- Sample statistic: sample mean is 112 grams.

Exercise (Cont.)

Name one way the sampling plan could fail to represent the target population.

Exercise (Cont.)

Name one way the sampling plan could fail to represent the target population.

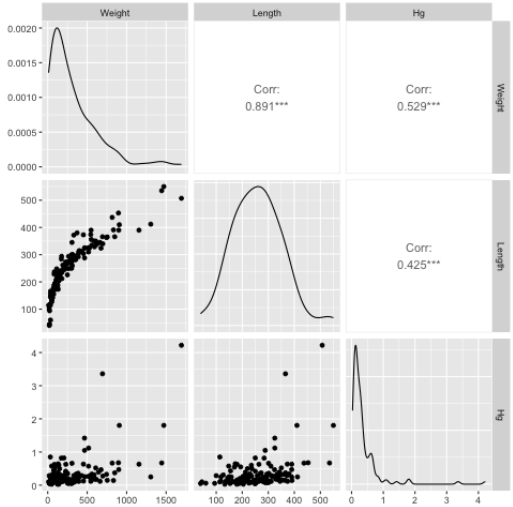
Examples include sampling only one day, sampling only early diners, excluding one dining hall, or selecting trays that are easiest to reach.

Outline

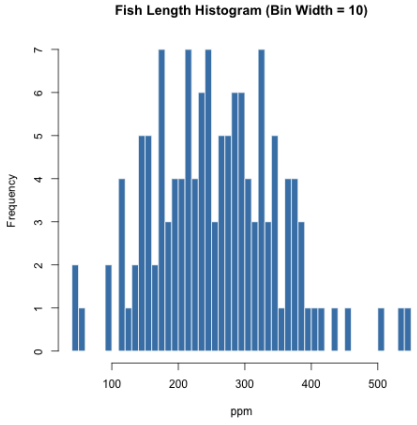
- Population Versus Sample
- Random Sampling
- Sample Distribution Versus Underlying “Population Distribution"

Mercury in Fish Data

Mercury concentrations in 133 fish collected across the United States.

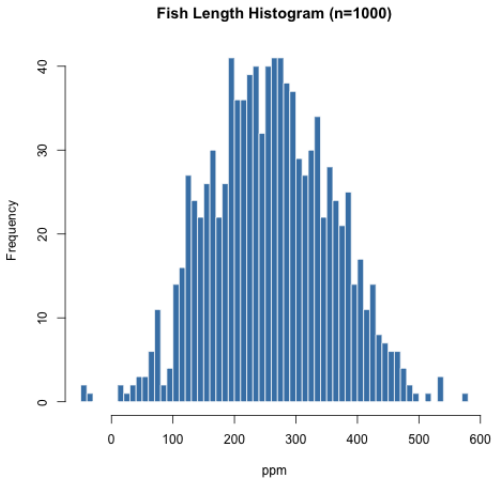


Mercury in Fish Data (n=133)

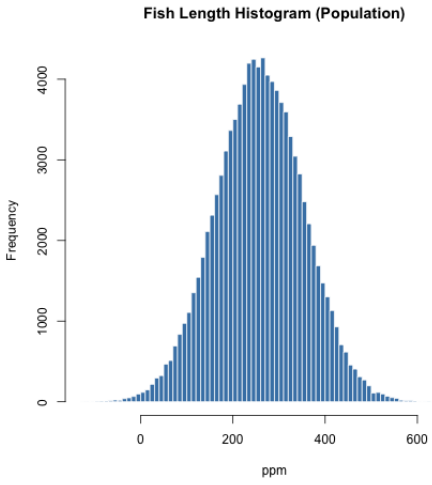


$$\overline{x} = 258.14(mm)$$
$$sd = 95.09(mm)$$

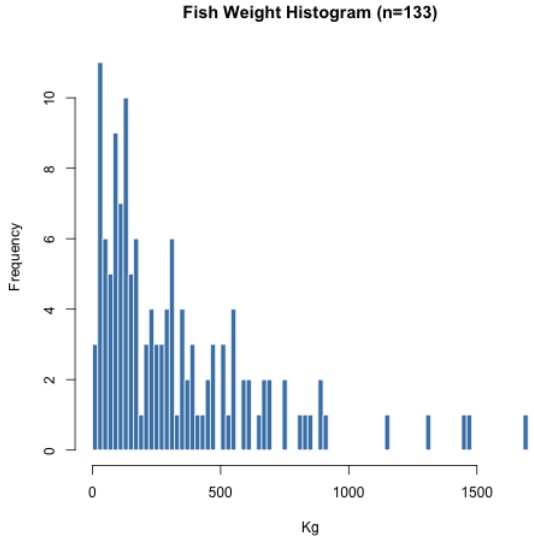
Mercury in Fish Data (n=1000)



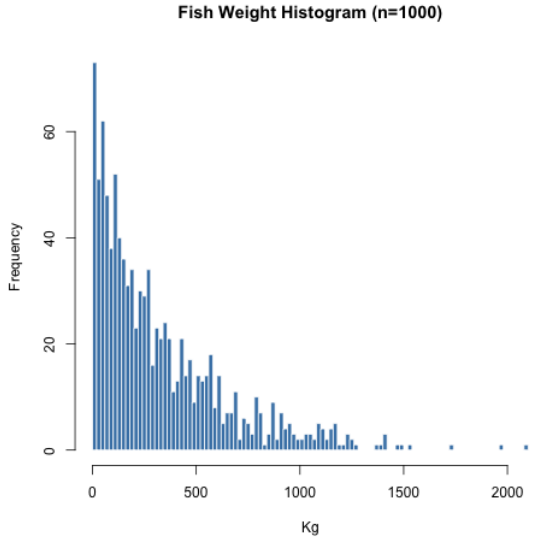
Mercury in Fish Data (Population)



Mercury in Fish Data (n=133)



Mercury in Fish Data (n=1000)



Mercury in Fish Data (Population)