POLI 30 D: Political Inquiry

Professor Umberto Mignozzetti
(Based on DSS Materials)

Lecture 06 | Measuring Population Characteristics I

Before we start

Announcements:

- ▶ Quizzes and Participation: On Canvas.

- ▶ Github page:
  https://github.com/umbertomig/POLI30Dpublic

- ▶ Piazza forum: https://piazza.com/ucsd/winter2023/17221

Before we start

Recap:

- ▶ We learned the definitions of Theory, Scientific Theory, and Hypotheses.
- ▶ Data, datasets, variables, and how to compute means.
- ▶ Causal effect, treatments, outcomes, randomization, and ATE.

Great job!

- ▶ Do you have any questions about these contents?

Plan for Today

- Sample vs. Population
- Representative samples
- Random Sampling
- Random Treatment Assignment vs.
  Random Sampling

- Exploring One Variable At a Time
  - Table of frequencies
  - Table of proportions
  - Histogram
  - Descriptive Statistics: mean, median,
    standard deviation, and variance

Why Do We Analyze Data?

1. MEASURE: **To infer population characteristics via survey research**
   – what proportion of constituents support a particular policy?

2. PREDICT: **To make predictions**
   – who is the most likely candidate to win an upcoming election?

3. EXPLAIN: **To estimate the causal effect of a treatment on an outcome**
   – what is the effect of small classrooms on student performance?

Why Do We Analyze Data?

1. MEASURE: **To infer population characteristics via survey research**
   – what proportion of constituents support a particular policy?

2. PREDICT: **To make predictions**
   – who is the most likely candidate to win an upcoming election?

3. EXPLAIN: **To estimate the causal effect of a treatment on an outcome**
   – what is the effect of small classrooms on student performance?

Sample vs. Population

▶ We often want to know the characteristics of a large **population** such as the residents of a country

▶ Yet collecting data from every individual in the population is either prohibitively expensive or infeasible.

▶ In the US, we try to collect data from each individual every ten years

  ▶ The 2020 census cost $14.2 billion, approximately (the population at that time was around 331 mi)
  ▶ This is not feasible for research purposes!

▶ We use surveys to collect data from a small subset of observations in order to understand the population

Sample vs. Population

▶ The subset of individuals chosen for study is called a **sample**
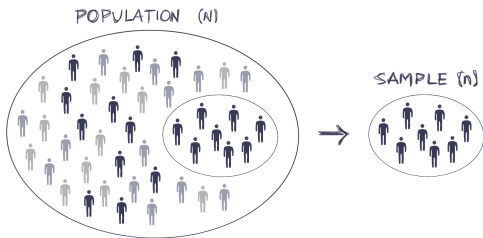
▶ Researchers typically survey about 1,000 people to infer the characteristics of more than 200 million US citizens

▶ n=1,000, N=200 million

Representative Samples

▶ In survey research, the sample needs to be representative of the population of interest

▶ **Representative sample**: Accurately reflects the characteristics of the population from which it is drawn

▶ If the sample is not representative, our inferences regarding the population characteristics will be wrong

Representative Samples

▶ Are you a representative sample of US residents?

▶ Are you a representative sample of UCSD students?

▶ Are you a representative sample of UCSD Poli majors?

▶ Are you a representative sample of POLI 30 D students?

▶ What would be the best way to draw a representative sample of UCSD students?

Representative Samples

▶ You should be careful about how representative your sample is:

Random Sampling

▶ The best way to draw a representative sample is to select individuals at *random* from the population.

Random sampling makes the **sample** and the **target population** *on average* identical.

▶ Random sampling: enables us to infer *valid* population characteristics from the sample,

Random Treatment Assignment vs. Random Sampling

- ▶ They both use a random process but are different concepts.

- ▶ **Random treatment assignment** means that treatment is assigned at random:
  - ▶ makes treatment and control groups comparable.

- ▶ **Random sampling** means that individuals are selected at random from the population into the sample:
  - ▶ makes the sample representative of the population.

- ▶ For this class, we assume we are always studying a representative sample.

Exploring One Variable At a Time

▶ Suppose we have collected data from a sample. Now what?

▶ To understand the content and distribution of each variable, we can:

   ▶ Create a table of frequencies
   ▶ Create a table of proportions
   ▶ Create a histogram
   ▶ Compute descriptive statistics

▶ Let us return to the voting experiment

   ▶ data collected from a sample of registered voters in the state of Michigan

The *voting* dataset

Unit of observation: registered voters

Description of variables:

variable	description
birth	year of birth of registered voter
message	whether registered voter received message: "yes", "no"
voted	whether registered voter voted: 1=voted, 0=didn't vote

Table of Frequencies

▶ The **frequency table** shows the values the variable takes and the number of times each value appears in the variable

▶ R function: `table()`

```
table(voting$voted)
##
##      0      1
## 158276  71168
```

▶ Interpretation?

Table of Proportions

▶ The **table of proportions** shows the proportion of observations that take each value in the variable

▶ The proportions in the table should add up to 1

▶ R function: `prop.table(table( ))`

```
prop.table(table(voting$voted))
##
##         0         1
## 0.6898241 0.3101759
```
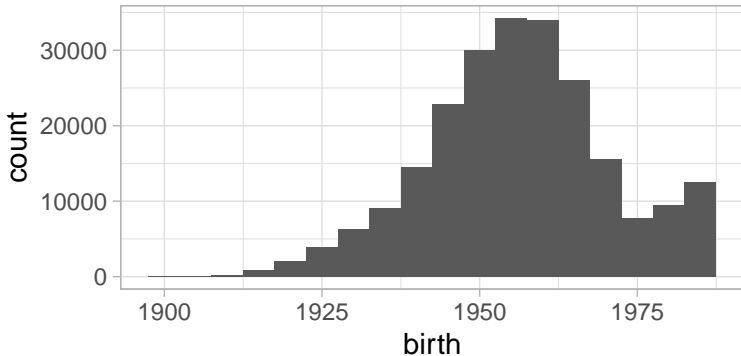
▶ Interpretation?

Histogram

▶ The **histogram** is the visual representation of a variables distribution through bins of different heights

▶ The position of the bins along the x-axis indicates the interval of values

▶ The height of the bins indicates the frequency (or count) of the interval of values

▶ R functions: `hist()` or `ggplot() + geom_histogram()`

▶ Great for `quantitative` variables (the numeric R data types)

Histogram

```
ggplot(voting, aes(x = birth)) +
  geom_histogram(binwidth = 5) + theme_light()
```
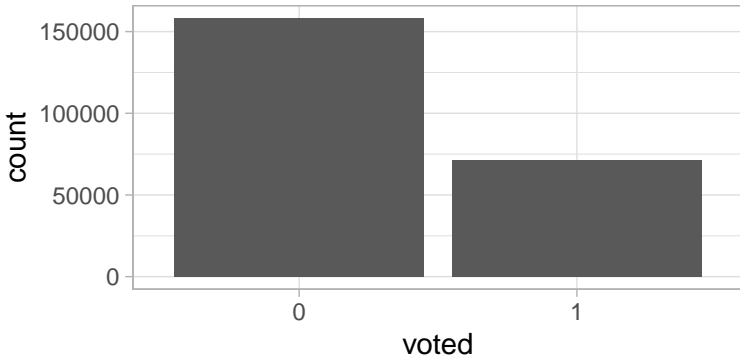

▶ Interpretation?

Barplots

▶ The **barplot** is similar to a histogram, but discretizes the variation

▶ The position of the bins along the x–axis indicates a value

▶ The height of the bins indicates the frequency (or count) of the values

▶ R functions: `barplot(table())` or `ggplot() + geom_bar()`

▶ Great for `qualitative` variables (numeric binary or character)

Barplots

```
ggplot(voting, aes(x = voted)) +
  geom_bar(aes(x = as.character(voted))) + theme_light()
```


- ► Interpretation?

Descriptive Statistics

▶ The **descriptive statistics** of a variable numerically summarizes the main characteristics of its distribution

▶ Measures of centrality
(center of the distribution):
  ▶ mean
  ▶ median

▶ Measures of spread
(amount of variation from the center):
  ▶ standard deviation
  ▶ variance

Mean

▶ The **mean** of a variable equals the sum of the values across all observations divided by the total number of observations

▶ What is the function in R?

▶ Example:

```
mean(voting$birth)
## [1] 1956.18
mean(voting$voted)
## [1] 0.3101759
```

▶ Interpretations?

Median

▶ The **median** of a variable is the value at the midpoint of the distribution that divides the data into two equal-size groups

▶ When the variable contains an odd number of observations, the median is the middle value of the distribution

▶ When the variable contains an even number of observations, the median is the average of the two middle values

Median

▶ Example, if $X=\{10, 4, 6, 8, 22\}$, what is the median of X?
  ▶ First, we need to sort the values of X in ascending order (as they would be in the distribution): $\{4, 6, 8, 10, 22\}$
  ▶ The value in the middle of the distribution is 8 so the median is 8.

▶ R function: **median()**

```
median(voting$birth)
## [1] 1956
```

▶ Interpretations?

Standard Deviation

▶ The **standard deviation** of a variable is a measure of the spread of its distribution

$$\text{sd}(X) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}$$

- $sd(X)$ stands for the standard deviation of X
- X_i is a particular observation of X
- $\overline{X}$ stands for the mean of X
- n is the total number of observations in the variable
- $\sum_{i=1}^{n}(X_i - \overline{X})^2$ means the sum of all $(X_i - \overline{X})^2$
  from $i = 1$ to $i = n$

Standard Deviation

The **standard deviation** of variable measures the average distance of the observations to the mean.

▶ The larger the standard deviation, the `flatter` the distribution

▶ It gives us a sense of the range of the data, especially when dealing with bell–shaped distributions

▶ In bell–shaped (normal) distributions, 95% of the observations fall within two standard deviations from the mean

Standard Deviation

- ▶ R function: **sd()**

```
sd(voting$birth)
## [1] 14.46019
```

- ▶ If *birth* were normally distributed, about 95% of the registered voters would have been born between 1927 and 1985:
  - ▶ $\overline{X} - 2 \times \text{sd}(X) = 1956 - 2 \times 14.5 = 1927$
  - ▶ $\overline{X} + 2 \times \text{sd}(X) = 1956 + 2 \times 14.5 = 1985$

Variance

▶ Another measure of the spread of the distribution

▶ The **variance** of a variable is simply the square of the standard deviation

$$\text{var(X)} = [\text{sd(X)}]^2$$

    – *var(X)* stands for the variance of *X*
    – *sd(X)* stands for the standard deviation of *X*

Variance

- ▶ R function: **`var()`**

```
var(voting$birth)
## [1] 209.0971
```

- ▶ Alternatively: **`sd()^2`**

```
sd(voting$birth)^2
## [1] 209.0971
```

- ▶ We are usually better off using standard deviations as our measure of spread:
- ▶ Same unit of measurement as the variable

Summary

- **Today's Class:**
  - Sample vs. Population
  - Representative Samples and Random Sampling
  - Exploring a single variable

- **Next class:**
  - Correlations
  - Scatter-plots

Questions?

See you in the next class!