Tutorial 5

Research Methods for Political Science - PO3110

Andrea Salvi

30 October 2018

Trinity College Dublin,
https://andrsalvi.github.io/research-methods/

Table of contents

Brief Recap of Previous Concepts

Estimate Mean, Standard Deviation and Standard Error

1. Estimate Mean: $\bar{x} = \frac{\sum x}{n}$

Estimate Mean, Standard Deviation and Standard Error

1. Estimate Mean: $\bar{x} = \frac{\sum x}{n}$
2. Sum of Squared Errors (SS): $\sum(x - \bar{x})^2$

Estimate Mean, Standard Deviation and Standard Error

1. Estimate Mean: $\bar{x} = \frac{\sum x}{n}$
2. Sum of Squared Errors (SS): $\sum(x - \bar{x})^2$
3. Estimate Variance: $\sigma^2 = \frac{SS}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$

Estimate Mean, Standard Deviation and Standard Error

1. Estimate Mean: $\bar{x} = \frac{\sum x}{n}$
2. Sum of Squared Errors (SS): $\sum(x - \bar{x})^2$
3. Estimate Variance: $\sigma^2 = \frac{SS}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$
4. Estimate Standard Deviation: $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\sigma^2}$

Estimate Mean, Standard Deviation and Standard Error

1. Estimate Mean: $\bar{x} = \frac{\sum x}{n}$
2. Sum of Squared Errors (SS): $\sum(x - \bar{x})^2$
3. Estimate Variance: $\sigma^2 = \frac{SS}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$
4. Estimate Standard Deviation: $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\sigma^2}$
5. Estimate standard error of the mean: $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Estimate Mean, Standard Deviation and Standard Error

1. Estimate Mean: $\bar{x} = \frac{\sum x}{n}$
2. Sum of Squared Errors (SS): $\sum(x - \bar{x})^2$
3. Estimate Variance: $\sigma^2 = \frac{SS}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$
4. Estimate Standard Deviation: $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\sigma^2}$
5. Estimate standard error of the mean: $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Estimate Mean, Standard Deviation and Standard Error

1. Estimate Mean: $\bar{x} = \frac{\sum x}{n}$
2. Sum of Squared Errors (SS): $\sum(x - \bar{x})^2$
3. Estimate Variance: $\sigma^2 = \frac{SS}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$
4. Estimate Standard Deviation: $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\sigma^2}$
5. Estimate standard error of the mean: $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

How do you calculate those in SPSS?

Confidence Intervals

For a given statistic calculated from a sample, **the confidence interval is a range of values around that statistic that are believed to contain, with a certain probability, the true value of that statistic (population value)**. A 95% confidence interval will contain the population mean 19 out of 20 times.

In short:

$$CI95 = \bar{x} \pm 1.96 * sd(\bar{X})$$

Confidence Interval: Common Mistakes

WRONG: There is a 95% probability that the population mean lies between CI LOW and CI HIGH.

Confidence Interval: Common Mistakes

WRONG: There is a 95% probability that the population mean lies between CI LOW and CI HIGH.
Why: The population mean is assumed to be fixed, thus there is no randomness. The probability that the confidence intervals cover the true mean is 0 or 1.

Confidence Interval: Common Mistakes

WRONG: There is a 95% probability that the population mean lies between CI LOW and CI HIGH.
Why: The population mean is assumed to be fixed, thus there is no randomness. The probability that the confidence intervals cover the true mean is 0 or 1.
Correct interpretation: 95% of the time, when we calculate a confidence interval in this way, the true mean will be between the two values. 5% of the time, it will not.

Confidence Interval: Common Mistakes

WRONG: There is a 95% probability that the population mean lies between CI LOW and CI HIGH.
Why: The population mean is assumed to be fixed, thus there is no randomness. The probability that the confidence intervals cover the true mean is 0 or 1.
Correct interpretation: 95% of the time, when we calculate a confidence interval in this way, the true mean will be between the two values. 5% of the time, it will not.
Since the true mean (population mean μ) is an unknown value, we don't know if we are in the 5% or the 95%.

One sample T-test

IQ in general population is 100. We take a random sample of 30 high school students and find $\bar{x} = 110$, $sd = 10$.

One sample T-test

IQ in general population is 100. We take a random sample of 30 high school students and find $\bar{x} = 110$, $sd = 10$.

State the null hypothesis and alternative hypothesis:

- $H0$: The observed average IQ equals the general population's IQ.

One sample T-test

IQ in general population is 100. We take a random sample of 30 high school students and find $\bar{x} = 110$, $sd = 10$.

State the null hypothesis and alternative hypothesis:

- $H0$: The observed average IQ equals the general population's IQ.
- $H1$: The observed average IQ differs from the general population IQ.

One sample T-test

IQ in general population is 100. We take a random sample of 30 high school students and find $\bar{x} = 110$, $sd = 10$.

State the null hypothesis and alternative hypothesis:

- $H0$: The observed average IQ equals the general population's IQ.
- $H1$: The observed average IQ differs from the general population IQ.

We ask: if the null hypothesis were true, how likely would we be to collect the data we have?

One sample T-test

IQ in general population is 100. We take a random sample of 30 high school students and find $\bar{x} = 110$, $sd = 10$.

State the null hypothesis and alternative hypothesis:

- $H0$: The observed average IQ equals the general population's IQ.
- $H1$: The observed average IQ differs from the general population IQ.

We ask: if the null hypothesis were true, how likely would we be to collect the data we have?

Choose a level of significance

- IC95

One sample T-test

IQ in general population is 100. We take a random sample of 30 high school students and find $\bar{x} = 110$, $sd = 10$.

State the null hypothesis and alternative hypothesis:

- $H0$: The observed average IQ equals the general population's IQ.
- $H1$: The observed average IQ differs from the general population IQ.

We ask: if the null hypothesis were true, how likely would we be to collect the data we have?

Choose a level of significance

- IC95
- $\alpha = 0.05$

One Sample T-test

Selecting sampling distribution and critical region:

One Sample T-test

Selecting sampling distribution and critical region:

- Population standard deviation (σ) is unknown.

One Sample T-test

Selecting sampling distribution and critical region:

- Population standard deviation (σ) is unknown.
- t-distribution with $n - 1$ degrees of freedom (n = 30; df = 29). $= 0.05$

One Sample T-test

Selecting sampling distribution and critical region:

- Population standard deviation (σ) is unknown.
- t-distribution with $n - 1$ degrees of freedom (n = 30; df = 29). $= 0.05$
- Two-tailed test!

One Sample T-test

Selecting sampling distribution and critical region:

- Population standard deviation (σ) is unknown.
- t-distribution with $n - 1$ degrees of freedom (n = 30; df = 29). $= 0.05$
- Two-tailed test!
- Let's find our critical value:

Critical Values

one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042

One Sample T-test

- Calculate test statistic

One Sample T-test

- Calculate test statistic
- $t = \frac{\text{observed value - expected value under } H_0}{\text{standard error}}$

One Sample T-test

- Calculate test statistic
- $t = \frac{\text{observed value - expected value under } H_0}{\text{standard error}}$
- $t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

One Sample T-test

- Calculate test statistic
- $t = \frac{\text{observed value - expected value under } H_0}{\text{standard error}}$
- $t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- $t = \frac{110 - 100}{\frac{10}{\sqrt{30}}}$

One Sample T-test

- Calculate test statistic
- $t = \frac{\text{observed value - expected value under } H_0}{\text{standard error}}$
- $t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- $t = \frac{110 - 100}{\frac{10}{\sqrt{30}}}$
- $t = 5.47$

Problems in the assignments

Problems in the assignments

- Low quality graphs

Problems in the assignments

- Low quality graphs
  - Please use the *EXPORT* feature. (See Guides on Tutorial's website)

Problems in the assignments

- Low quality graphs
    - Please use the *EXPORT* feature. (See Guides on Tutorial's website)
- Screen-shots of SPSS output (*syntax* may help!)

Problems in the assignments

- Low quality graphs
    - Please use the *EXPORT* feature. (See Guides on Tutorial's website)
- Screen-shots of SPSS output (*syntax* may help!)
- Equation mode not used:

Problems in the assignments

- Low quality graphs
  - Please use the *EXPORT* feature. (See Guides on Tutorial's website)
- Screen-shots of SPSS output (*syntax* may help!)
- Equation mode not used:
  - sqrt(30/(9-1))= 1.936492

Problems in the assignments

- Low quality graphs
  - Please use the *EXPORT* feature. (See Guides on Tutorial's website)
- Screen-shots of SPSS output (*syntax* may help!)
- Equation mode not used:
  - sqrt(30/(9-1))= 1.936492
  - $\sqrt{\frac{30}{9-1} = 1.94}$

Problems in the assignments

- Low quality graphs
  - Please use the *EXPORT* feature. (See Guides on Tutorial's website)
- Screen-shots of SPSS output (*syntax* may help!)
- Equation mode not used:
  - sqrt(30/(9-1))= 1.936492
  - $\sqrt{\frac{30}{9-1} = 1.94}$
- Table formatting
  - Avoid reporting unnecessary info.
  - Tables should be readable!

Problems in the assignments

Research Design

- Don't: Is there a correlation between X and Y.
- Don't: What are the factors that resulted in Y.
- Don't: The impact of X on A, B, C, D

Conceptualization of variables

- Be precise.
- Define what you want to measure.
- Don't mix up independent and dependent variable
- Hypothesis should be clear and testable!

In-class Exercise

T-test in practice

- Download the following dataset on simulated rent prices in Dublin [1]: `https://tinyurl.com/MT4dublinrent`
- Area: North Dublin; South Dublin; price: simulated price
- Create a box-plot that shows the distribution of north and south rents.
- Observed value of 400; conduct one-sample t-test (for entire sample)
- Conduct independent samples t-test (compare means of South and North Dublin)

[1] Credits to Stefan Mueller

In-class Exercise 2

In-class Exercise 2

- Download ”UCDP.sav” `https://tinyurl.com/ucdp-mt5`
- You can work in pairs

In-class Exercise 2

- Create a new syntax file and store your output there!
- Paste the following into the syntax file: what does that do?

```
DATASET ACTIVATE DataSet1.
COMPUTE duration=end_year - start_year.
EXECUTE.
```

In-class Exercise 2

- Subset the data in order to select just the following conflicts (hint: ”type_of_conflict”): ”Internal armed conflict occurs between the government of a state and one or more groups (no int)”.
- Subset the data in order to select just conflict occurring in Africa and Middle-East
- Plot a histogram of your choice that conveys meaningful information.
- Have a look at the visualization options in SPSS. Any hints on which ones suits our data?

In-class Exercise 2

- ”Get rid of cases with missing values in the duration variable.
- ”Split” the data-set based on the ”region” variable (Hint: Data -> Split File). Now try to calculate the mean, standard deviation and Standard Error of the Mean for the ”duration” variable. What happened?
- Conduct a independent sample t-test to compare the duration in Africa and in the Middle East. Are they significantly different?