Module 2:
Central tendency, shape, and difference in means

MSIR 525

Monday, September 23, 2019

Recap of Module 1 (check list from syllabus; see pages 1-2)

- We learned about the NHST framework

- We developed an understanding of *p*-values and how they can be used to inform evidence-based management decisions

- We compared different types of error that can threaten our inferences and conclusions
  - We also learned how one can attempt to avoid these errors and disclosures that must be given if a study is underpowered

- We contrasted three different research designs (e.g. observational) and two different data collection approaches (e.g., longitudinal)

- We learned about different data sources and data types

- We summarized several types of validity and phenomena that may threaten them

Agenda for Module 2

- 9/23/2019
  - Central tendency and shape; interpretation and communication; issues in datasets

Agenda for Module 2

- 9/23/2019
  - Central tendency and shape; interpretation and communication; issues in datasets

- 9/25/2019
  - Assess whether or not two means are *statistically* different from each other (i.e., a *t*-test)

Agenda for Module 2

- 9/23/2019
  - Central tendency and shape; interpretation and communication; issues in datasets

- 9/25/2019
  - Assess whether or not two means are *statistically* different from each other (i.e., a *t*-test)

- 9/30/2019
  - Assess whether or not multiple means are *statistically different from each other (i.e., ANOVA test)

Agenda for Module 2

- 9/23/2019
  - Central tendency and shape; interpretation and communication; issues in datasets

- 9/25/2019
  - Assess whether or not two means are *statistically* different from each other (i.e., a *t*-test)

- 9/30/2019
  - Assess whether or not multiple means are *statistically different from each other (i.e., ANOVA test)

- 10/2/2019
  - Module 2 recap and software tutorial (R ***must*** be installed by this date!!)

Agenda for Module 2

- 9/23/2019
  - Central tendency and shape; interpretation and communication; issues in datasets
- 9/25/2019
  - Assess whether or not two means are *statistically* different from each other (i.e., a *t*-test)
- 9/30/2019
  - Assess whether or not multiple means are *statistically different from each other (i.e., ANOVA test)
- 10/2/2019
  - Module 2 recap and software tutorial (R ***must*** be installed by this date!!)
- 10/7/2019
  - In-class exercise for credit (i.e., a hackathon)
  - Applying what we learned in M2 to ascertain whether or not a meaningful group difference exists

Agenda for Module 2

- Let’s get started! ☺

Summarizing Data

- Frequency distribution

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Table 2. Frequency Distribution

Rating	Stress	WLB	Job satisfaction	Pay satisfaction
10	0	0	0	3
9	0	0	0	0
8	0	2	1	0
7	3	0	3	2
6	2	0	0	2
5	2	1	0	0
4	0	1	2	0
3	0	0	1	0
2	0	3	0	0
1	0	0	0	0
0	0	0	0	0
Count	7	7	7	7

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Table 2. Frequency Distribution

Rating	Stress	WLB	Job satisfaction	Pay satisfaction
10	0	0	0	3
9	0	0	0	0
8	0	2	1	0
7	3	0	3	2
6	2	0	0	2
5	2	1	0	0
4	0	1	2	0
3	0	0	1	0
2	0	3	0	0
1	0	0	0	0
0	0	0	0	0
Count	7	7	7	7

CAN BE VISUALIZED IN A BARPLOT

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Table 2. Frequency Distribution

Rating	Stress	WLB	Job satisfaction	Pay satisfaction
10	0	0	0	3
9	0	0	0	0
8	0	2	1	0
7	3	0	3	2
6	2	0	0	2
5	2	1	0	0
4	0	1	2	0
3	0	0	1	0
2	0	3	0	0
1	0	0	0	0
0	0	0	0	0
Count	7	7	7	7

CAN BE VISUALIZED IN A BARPLOT

CAN BE USED TO SUMMARIZE ALL TYPES OF DATA (SEE MODULE 1)

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data
  - Represents how often a response is observe relative to the total number of responses
    - “What proportion of the respondents gave a rating of 7 for stress?”

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data
  - Represents how often a response is observe relative to the total number of responses
    - "What proportion of the respondents gave a rating of 7 for stress?"

$$\text{Relative frequency} = \frac{frequency\ of\ response}{total\ number\ of\ responses}$$

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data
  - Represents how often a response is observe relative to the total number of responses
    - "What proportion of the respondents gave a rating of 7 for stress?"

$$\text{Relative frequency} = \frac{frequency\ of\ response}{total\ number\ of\ responses}$$

$$= \frac{3}{7} = 43\%$$

Summarizing Data

- Cumulative frequency and cumulative percentage
  - An assessment of the total frequency (percentage) of all categories up to and including the category of interest

Summarizing Data

- Cumulative frequency and cumulative percentage
  - An assessment of the total frequency (percentage) of all categories up to and including the category of interest

Cumulative frequency$_n$ = frequency$_n$ + *cumulative frequency*$_{n-1}$

Table 3. Frequency Distributions for Stress

		Relative	Cumulative	Cumulative
Rating	Frequency	frequency	frequency	percentage
10	0	0 (0%)	7	1.0 (100%)
9	0	0 (0%)	7	1.0 (100%)
8	0	0 (0%)	7	1.0 (100%)
7	3	.43 (43%)	7	1.0 (100%)
6	2	29 (29%)	4	.58 (58%)
5	2	.29 (29%)	2	.29 (29%)
4	0	0 (0%)	0	0 (0%)
3	0	0 (0%)	0	0 (0%)
2	0	0 (0%)	0	0 (0%)
1	0	0 (0%)	0	0 (0%)
0	0	0 (0%)	0	0 (0%)

Summarizing Data

- Cumulative frequency and cumulative percentage
  - An assessment of the total frequency (percentage) of all categories up to and including the category of interest

Cumulative percentage$_n$ = percentage$_n$ + cumulative percentage$_{n-1}$

Table 3. Frequency Distributions for Stress

Rating	Frequency	Relative frequency	Cumulative frequency	Cumulative percentage
10	0	0 (0%)	7	1.0 (100%)
9	0	0 (0%)	7	1.0 (100%)
8	0	0 (0%)	7	1.0 (100%)
7	3	.43 (43%)	7	1.0 (100%)
6	2	29 (29%)	4	.58 (58%)
5	2	.29 (29%)	2	.29 (29%)
4	0	0 (0%)	0	0 (0%)
3	0	0 (0%)	0	0 (0%)
2	0	0 (0%)	0	0 (0%)
1	0	0 (0%)	0	0 (0%)
0	0	0 (0%)	0	0 (0%)

Summarizing Data

- Mean (or median) splits
    - A method used to estimate the number of “high” vs. “low” responses observed in a dataset

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of "high" vs. "low" responses observed in a dataset
  - Example: How many people have "high" and "low" levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Step 3: Identify “high” (i.e., > 5.71) vs. “low” (i.e., < 5.71) scores

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Step 3: Identify “high” (i.e., > 5.71) vs. “low” (i.e., < 5.71) scores

Step 4: Calculate “high” vs. “low” frequencies and percentages

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

4 out 7 =“high” scores

4/7 = .57 (57%)

3 out 7 =“low” scores

3/7 = .43 (43%)

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Step 3: Identify “high” (i.e., > 5.71) vs. “low” (i.e., < 5.71) scores

Step 4: Calculate “high” vs. “low” frequencies and percentages

Central tendency

- Mean, median, mode

Central tendency

- Mean, median, mode

Central tendency

- Mean, median, mode

- Honestly, we are mostly just interested in the **mean**

Variance

- Skewness

- Kurtosis

Shape

- Skewness

- Kurtosis

Threats to descriptive statistics

- Missing data

- Outliers

- Range restriction

Interpreting descriptive statistics