Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

August 19, 2019

Notes

Topics

1. Organizing and Graphing Qualitative Data
2. Organizing and Graphing Quantitative Data

Notes

Objectives

Objectives:

- Make a frequency distribution table qualitative data.
- Make a bar chart of qualitative data.
- Make a frequency distribution table of quantitative data.
- Make a dot plot of quantitative data.
- Make a histogram of quantitative data.
- Identify and interpret the shape, center, and spread of a dot plot or histogram.

Notes

Organizing Qualitative Data (2.2)

Frequency Distribution Tables for Qualitative Data

- One way to organize qualitative data is in a ***frequency distribution table***. To construct one:
  1. ***Group*** the data according to the categories and count how many individuals fall into each category.
  2. Summarize the results in a *table*, showing the categories and their ***frequencies*** (i.e. counts) and/or ***relative frequencies*** (proportions or percentages).

Notes

Exercise

Fill in the **frequency distribution table** (on the next slide) using the data from the in-class responses to the question:

"Whose face paint job do you like the best?"

A. Ace
B. Paul
C. Peter
D. Gene

Notes

Response	Frequency	Relative Frequency
Ace		
Paul		
Peter		
Gene		
Total		

Notes

Graphing Qualitative Data (2.2)

Bar Charts

- The most useful way to display *qualitative* data is with a ***bar chart***. To construct one:
  1. Make a *frequency distribution table* of the data.
  2. Write the names of the categories on the horizontal axis.

Notes

- (cont'd):
  3. On the vertical axis, label the scale ("Frequency" or "Relative Frequency") and draw tick marks and their values.
  4. Place a bar over each category name, with bar height equal to the frequency or relative frequency of the category.
  5. Add a title.

Notes

Notes

Exercise

Make a a **bar chart** of the data from the in-class responses to the question:

"Whose face paint job do you like the best?"

using the frequency distribution table from the last example.

Then draw any **conclusions** that seem appropriate.

Notes

Organizing Discrete Quantitative Data (2.3)

- We can organize *quantitative* data in a *frequency distribution table* too.

  The table is created differently depending on whether the data are *discrete* or *continuous*.

Notes

Frequency Distribution Tables for *Discrete* Quantitative Data

- To construct the ***frequency distribution table*** for *discrete* data:

  1. Determine which distinct values occur in the data set.
  2. ***Group*** the data by these distinct values ("***single-value grouping***"), then count the number times each value occurs.
  3. Summarize the results in a *table* showing the distinct values and their ***frequencies*** (counts) and/or ***relative frequencies*** (proportions or percentages).

Notes

Example

Here are data on sizes of litters for $n = 36$ sows. Litter size (number of piglets) is an integer (discrete).

10	12	10	7	14	11
14	11	10	13	10	10
8	11	7	13	12	13
10	8	5	11	11	12
11	11	9	8	12	10
9	11	10	12	10	9

Here's the **frequency distribution table**:

Distinct Litter Sizes	Frequency	Relative Frequency
5	1	0.028
6	0	0.000
7	2	0.056
8	3	0.083
9	3	0.083
10	9	0.250
11	8	0.222
12	5	0.139
13	3	0.083
14	2	0.056

Notes

Frequency Distribution Tables for *Continuous* Quantitative Data

- To construct the ***frequency distribution table*** for *continuous* data:
  1. Choose the number of ***class intervals***. Usually 5-20 works well. The number of class intervals will depend on the size of the data set - for smaller data sets, use around 5 and for larger data sets, around 20.
  2. Decide on appropriate ***cutpoints*** (values at the borders of the class intervals). The class intervals should all be the same width.

Notes

- Constructing a frequency distribution table (cont'd):
  3. Use class intervals of the form:

     $$\textit{Lower Cutpoint} \leq \text{Data Value} < \textit{Upper Cutpoint}$$

     so that borderline data values go into the *upper* class interval.
  4. ***Group*** the data according to the class intervals ("***cutpoint grouping***") and count the number of observations that fall into each class interval.
  5. Summarize the results in a *table* showing the class intervals and their ***frequencies*** (counts) and/or ***relative frequencies*** (proportions or percentages).

Notes

Exercise

Here are surgery times (in hours) for emergency surgeries of $n = 50$ animals at a local animal hospital (ordered from shortest to longest).

0.33	0.33	0.33	0.33	0.50	0.50	0.50	0.67	0.75
0.75	0.75	0.83	0.92	0.92	1.00	1.00	1.00	1.00
1.08	1.08	1.17	1.25	1.27	1.33	1.42	1.42	1.50
1.50	1.50	1.50	1.58	1.67	1.67	1.67	1.67	1.67
1.70	1.75	1.75	1.83	1.83	2.00	2.00	2.00	2.33
2.42	2.50	2.67	3.08	4.50				

Fill in the **frequency distribution table** (on the next slide) using **cutpoints 0.0**, **0.5**, **1.0**, **1.5**, **...**, **5.0**.

(Check your answer two slides ahead.)

Notes

Class Interval (Surgery Times)	Frequency (Number of Surgeries)	Relative Frequency
0.0 - under 0.5		
0.5 - under 1.0		
1.0 - under 1.5		
1.5 - under 2.0		
2.0 - under 2.5		
2.5 - under 3.0		
3.0 - under 3.5		
3.5 - under 4.0		
4.0 - under 4.5		
4.5 - under 5.0		
	$n = 36$	1.00

Notes

Class Interval (Surgery Times)	Frequency (Number of Surgeries)	Relative Frequency
0.0 - under 0.5	4	0.08
0.5 - under 1.0	10	0.20
1.0 - under 1.5	12	0.24
1.5 - under 2.0	15	0.30
2.0 - under 2.5	5	0.10
2.5 - under 3.0	2	0.04
3.0 - under 3.5	1	0.02
3.5 - under 4.0	0	0.00
4.0 - under 4.5	0	0.00
4.5 - under 5.0	1	0.02
	$n = 36$	1.00

Notes

Graphing Quantitative Data (2.3)

Introduction

- There are many ways to display quantitative data graphically. We'll look at a few:
  1. Dot plots
  2. Histograms
  3. Box plots (later)
  4. Scatter plots (later)

Notes

Dot Plots

- Dot plots are useful for *quantitative* data when only a small number of distinct values appear in the data set. To make a ***dot plot***:
  1. Create a *frequency distribution table* of the data using *single-value grouping*.
  2. Draw a horizontal axis, label the horizontal axis scale (same units as the data), and mark the distinct values on this axis.

Notes

- (cont'd):
  3. For each distinct value in the data set, stack as many dots above horizontal axis as there are occurrences of that value in the data set.
  4. Add a title.

Notes

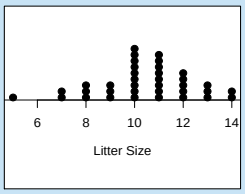
Example

For the litter size data, the **frequency distribution table** and **dot plot** are below.

Frequency Distribution Table

Distinct Litter Sizes	Frequency
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2

Dot Plot of of Litter Size



Notes

Exercise

In an experiment, one female and one male restaurant server drew happy faces on the checks of randomly chosen dining parties.

The data are given as tip percentages for the two servers ($n = 22$ for the female and $n = 23$ for the male) in the tables on the next slide.

Notes

female	17	21	21	22	23	24	24	25	27	27
	28	29	30	33	33	34	40	41	44	48
	65	72								

male	9	9	12	13	14	14	15	15	15	16
	17	17	17	18	18	19	20	20	22	22
	26	27	31							

Make separate **dot plots** of the female and male tip percentages. Use the same horizontal axis scale for the two plots so that a comparison can be made, and draw any **conclusions** that seem appropriate.

Notes

Notes

Histograms

- Histograms are useful for any set of **quantitative** data (**discrete** or **continuous**). They're especially useful for large data sets.

Notes

Histograms for Discrete Data

- To construct a ***histogram*** for *discrete* data:
  1. Create a *frequency distribution table* of the data using *single-value grouping*.
  2. Mark the distinct values on a horizontal axis and label the horizontal axis scale (same units as the data).
  3. Label the vertical axis ("Frequency" or "Relative Frequency") and draw tick marks and their values.

Notes

- (con'td):
  4. Place a bar over each distinct value, with bar height equal to the frequency or relative frequency of the value.
  5. Bars over adjacent distinct values should touch each other (i.e. no gaps between adjacent bars).
  6. Add a title.

Notes

Example

For the litter size data, the **frequency distribution table** and **histogram** are below.

Frequency Distribution Table

Distinct Litter Sizes	Frequency
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2

Notes

Histograms for Continuous Data

- To construct a ***histogram*** for *continuous* data:
  1. Create a *frequency distribution table* of the data using *cutpoint grouping*.
  2. Mark the cutpoints on a horizontal axis and label the horizontal axis scale (same units as the data).
  3. Label the vertical axis ("Frequency" or "Relative Frequency") and draw tick marks and their values.

Notes

- (cont'd):
  4. Place a bar over each class interval, with bar height equal to the frequency or relative frequency of the class interval.
  5. Bars over adjacent class intervals should touch each other (i.e. no gaps between adjacent bars).
  6. Add a title.

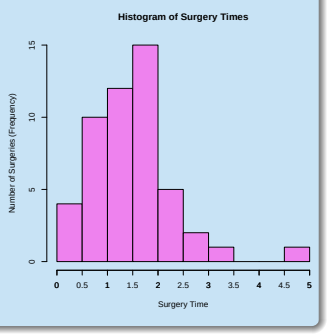
Notes

Example

For the surgery times data, the **frequency distribution table** and **histogram** are below.

Frequency Distribution Table

Class Interval (Surgery Times)	Frequency
0.0 - under 0.5	4
0.5 - under 1.0	10
1.0 - under 1.5	12
1.5 - under 2.0	15
2.0 - under 2.5	5
2.5 - under 3.0	2
3.0 - under 3.5	1
3.5 - under 4.0	0
4.0 - under 4.5	0
4.5 - under 5.0	1



Notes

- **Interpretation of histograms**: The **area** under the histogram above *any* interval, relative to the total area of the histogram, represents the **proportion** of observations that lie in that interval.

Exercise

Using the histogram from the last example, about what **proportion** of the surgery times are shorter than **1** hour? About what **proportion** are longer than **2** hours? About what **proportion** are between **1** and **2** hours?

What to Look For in Histograms and Dot Plots (2.4)

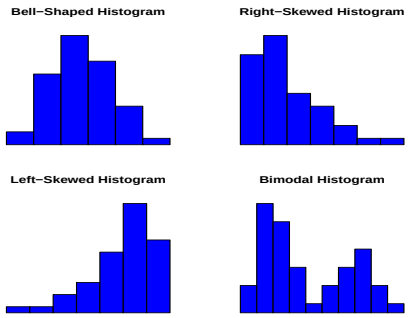
- Some things to look for in histograms and dot plots are:
  - ***Shape*** of the distribution of the data:
    - ***Symmetric*** (left and right halves are mirror images).
    - ***Bell-shaped*** .
    - ***Right skewed*** (long tail to the right).
    - ***Left skewed*** (long tail to the left).
    - ***Unimodal*** , , (one main peak), ***bimodal*** (two peaks), or ***multimodal*** (multiple peaks).
  - ***Outliers*** (observations lying apart from the overall pattern in the plot).

Notes

- Things to look for (cont'd):
  - ***Central location*** - what's a typical value in the data set?
  - ***Spread*** - do the values in the data set vary greatly , or are they all fairly consistent ?
  - ***Other interesting features*** - are there gaps, clumps, etc.?

Notes

The figure below illustrates some common **histogram shapes**.



Notes

Exercise

Using the histogram of the surgery times from the earlier example,

a) Describe the **shape** of the distribution of the data.

b) Determine a **typical** surgery time (**central location**).

c) Describe the **variation** (**spread**) in the surgery times.

Notes

Exercise

Using the dotplots of tipping percentages from the earlier exercise,

a) Compare the **typical** tip amounts (**central locations**) for the female and male servers. Does it appear that one receives better tips than the other?

b) Decide for which server, female or male, do the tips exhibit more variation (spread).

c) Decide if there are any **outliers** in either data set.

Notes

- The following example illustrates an **outlier** in a data set.

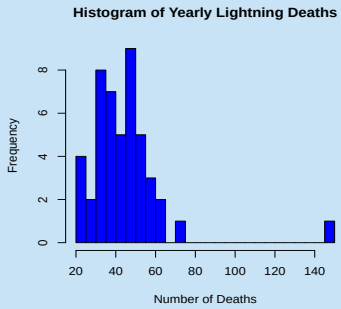
Example

The table below shows (part of) a data set on the numbers deaths by lightning strikes in the U.S. for each of the years 1959 - 2005, as reported by the National Weather Service.

Year	Deaths
1959	75
1960	48
1961	61
1962	48
1963	150
1964	49
⋮	⋮
2005	38

Notes

A histogram of the deaths data, shown below, reveals a clear outlier.



Notes

Referring to the outlier, the National Weather Service report states:

> *On December 8, 1963 the crash of a jetliner killing 81 people near Elkin, Maryland, was attributed to lightning by the Civil Aeronautics Board investigators.*

Notes

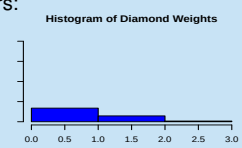
- This next example shows an **"interesting pattern"** that only reveals itself when we increase the number of bars in the histogram.

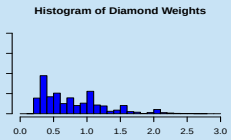
Example

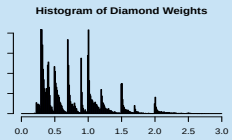
Shown below is (part of) a data set containing the weights of $n = 50,000$ cut diamonds from reputable dealers around the world. The weights are in carats (cts). A carat is 0.2 of a gram.

0.36	1.20	0.72	0.54	0.33	0.71	1.01	1.20	2.22	0.90
0.55	0.77	1.21	1.22	2.01	0.31	0.32	0.73	0.32	1.47
2.00	0.33	1.01	1.50	0.72	0.33	0.27	0.53	1.00	2.10
1.32	0.41	0.71	0.91	0.51	0.38	0.31	0.52	0.54	1.03
				⋮					
0.44	0.50	0.31	0.34	0.90	0.70	1.52	1.10	0.76	0.72

Here are histograms of the data using different class interval widths (1, 0.1, and 0.01 carat), and hence different numbers of bars:





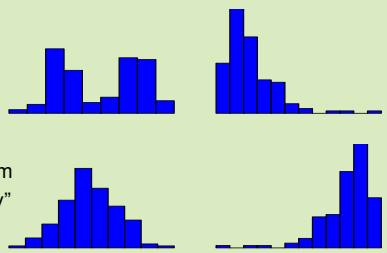


The spikes that appear (only) in the third histogram are due to the dealers rounding the weights *up* to the nearest tenth of a carat. (They're trying to *sell* the diamonds for the highest price, after all!)

Exercise

Match the data set on the left with its histogram:

A. People's heights
B. People's incomes
C. Weights of pets (cats and dogs)
D. Scores on an exam that was "too easy"



Notes

Notes

Notes

Notes