Ch 2 Descriptive Statistics

Ch 2.1 Stemplots and dotplots.

To summarize quantitative data: we look at five areas: CVDOT: center, variation, distribution, outlier, trend.
Summarize quantitative data by graph:
1) Stemplot
2) Dotplot
3) Histogram (main one)
4) Boxplot.

A) Stemplot

Stemplot is a quick way to graph relatively small quantitative data. It can show the overall pattern and outliers. The main problem is data value must be in a relatively small range. But data values can be recovered from a stemplot. Also, stemplot can easily be created without use of technology.

Each data value is separated into stem and leaf (the last digit). Data are arranged in order with the same stem in a row. Back-to-back stem plot can be used to compare two datasets. It can be used to show distribution of data (look side way).
Note: do not skip a stem with no leave to show distribution and outliers correctly.
Online stemplot maker:
http://digitalfirst.bfwpub.com/stats_applet/stats_applet_8_ovc.html

If data consists of decimal, use this online stemplot maker: https://www.geogebra.org/m/zPA7QFe

Ex1. A stemplot is given below: legend 6|3 means 63

```
6|35
7|89
8|12259
9|19
```

a) How many data are there?
b) What is the lowest and highest data?
c) What is the list of data?

Ex2. Graph stemplots for the two different samples of grades. Describe shape of distribution.

Sample 1: 63, 65, 68, 69, 81, 82, 82, 85, 89, 91, 99
Sample 2: 63, 65, 68, 69, 81, 82, 82, 85, 89, 91, 99, 123

Sample 1 Sample 2

B) Dotplot

Doplots are used for graphing small discrete dataset with small range. It is possible to recreate the original list of data values. Also, Dotplots can be easily created without the use of technology.

Online dotplot maker:
https://www.geogebra.org/m/BxqJ4Vag
(enter data to column A only)
Ex1.
Sketch a dotplot for : 3, 4, 5, 7, 8, 9, 5, 6, 7, 7, 7, 7
Describe shape of distribution and outliers.

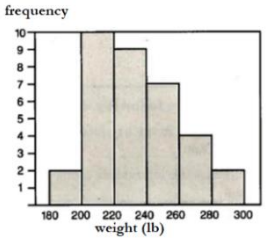
Ch 2.2 Histogram, frequency polygon and Time series graph.

A) Histogram

Histogram can be used to show distribution of medium to large quantitative data. Data are summarized into frequency classes.

- consists of contiguous (adjoining) bars. A Typical histogram should have about 5 to 15 bars or classes. Each bar represents one class of data.
- The horizontal axis is labeled with what the data represents in each class.
- The vertical axis is labelled either frequency or relative frequency.
- Histogram can show shape of distribution of the data, the center, and the spread of the data, but data cannot be recovered.
- Histogram can also show outliers that has a frequency of 1 or 2 but with large gap from the rest of data.

Ex1. The histogram below shows distribution of weights of players in a football team.



a) How many players are on the team?

b) How many players weight 240 lb. or more?

c) Can we tell the actual weights of the players from the histogram?
d) Are there any outliers?

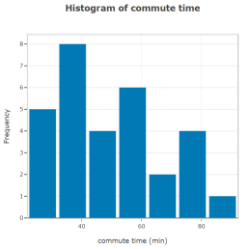
Graph histogram from data

Method 1: Use Statdisk (statdisk.org) Histogram

- Input or copy and paste (ctrl-V) data to Sample Editor in a column.
- Select Data/Histogram.
- Select column where data is located.
- Enter Histogram Title, x and y-axis label.
-Use Statdisk default classes (Auto-fit) or select "user-defined", enter "class-width" and "class start". "Class start" must be lower or equal to the minimum value in the data.
- Select frequency or relative frequency.
- Click plot. screen shot to save the histogram.
- Find the class range by hovering over each bar.

Ex1. Sketch a histogram for the following commute time. Use class width of 10 and lowest value of 20.

22	42	29	61
33	54	38	75
41	73	50	32
53	27	57	40
71	37	74	52
26	46	30	69
35	56	39	83
	73	51	



Method 2: Use Statdisk, Explore data.

- Enter data to one column,
- Select data, select Explore data-descriptive statistics.
- Select column, click evaluate.
- Summary statistics will be on the left and three graphs: Histogram (using auto-fit classes), boxplot and Normal quantile plot will be on the right.

Ex2. Sketch a Histogram of number of customers in a sample of stores.

12, 34, 45, 67, 43, 55, 57, 89, 77, 72, 56, 37, 45, 49, 51

Use default class of statdisk. Copy Histogram below. Describe shape of distribution.

B) Other graphs for quantitative data:

1) **Frequency polygon/relative frequency polygon** – Use for large datasets from frequency distribution. Use a line to show frequency distribution instead of bars. Line starts and ends at the horizontal axis.

Multiple frequency polygons can be graphed on the same graph for comparing two or more datasets.

C) Shape of distribution and skewness

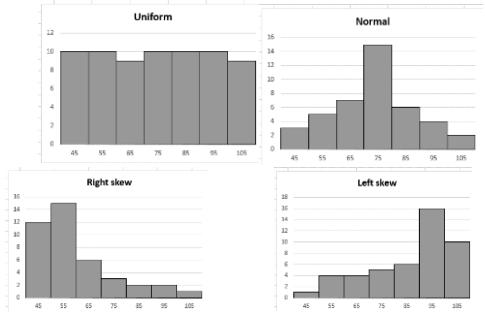
Histogram and frequency polygon can be used to show **shape** of distribution and skewness of data values.

Normal distribution means data are in a symmetrical bell shape.

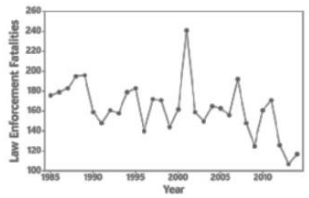
Skewed to the right – most data are in the low values.

Skewed to the left – most data are in the high values.

Uniform – data are evenly distributed.



D) Time Series Graph: Show trend of data collected over times. Data are not summarized. Time series graph shows increasing and decreasing data values over time.



Data are represented by points connected with lines.

Graph time-series graph by Excel

- Input data with time in one column and data in another column.
- Select the data values, insert chart, line graph.
- Select the x-axis label, right click, select data, select
- Edit horizontal axis label by selecting the year column. Select Ok and Ok.
- Input Chart Title, x and y-axis label and marker.

Ch 2.3 and 2.4 Percentile, Boxplot, outliers

A) Percentile

Percentile: are measures of location. Denoted by P_1, P_2, … P_{99} which divide a set of data into 100 groups with about 1% of the values in each group.

If x is at 90th percentile, means 90% of all data are less than x. Note, percentile is not the same as percentage.

Quartiles: (Q_1, Q_2, Q_3)

Quartiles are measures of location, which divide a set of data into four groups with about 25% of the values in each group.

Q_1 – First quartile or P_{25}. It separates the bottom 25% of value from the top 75%.

Q_2 - Second quartile or P_{50} or median. It separates the bottom 50% of values from the top 50%.

Q_3 – Third quartile or P_{75}. It separates the bottom 75% of values from the top 25%.

B) Five-number-summary, IQR and Boxplot:

Five-number-summary are:

Mininum, Q_1, Median, Q_3 and Maximum divides the data into four groups of 25% each.

$IQR = Q_3 - Q_1$ (Inter-quantile Range)

The **interquartile range (IQR)** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

A **boxplot** shows graphical image of concentration of data. A boxplot is constructed using 5-number summary with Q1, median and Q3 in a box containing 50% of all data. It gives good distribution of data in 25%, 50% and 75%.

- Maximum and Minimum values are extended as whiskers at the two ends of the box.

<u>Find 5-number-summary and boxplot by Statdisk</u>

- Enter data in a column in Statdisk.
- Select Data, Explore data, descriptive statistic, select column and Click evaluate.

Ex1. The time(in min.) a sample of 15 student spent on exercising daily is given:

0, 40, 60, 30, 60, 10, 46, 30, 300, 90, 30, 120, 60, 0, 20

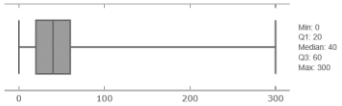
a) Find the 5-number summary and sketch a boxplot.

b) What percent of student exercise from 0 to 60 min?

c) What percent of student exercise between 20 to 60 min?

Use Statdisk:

Min = 0, Q1=20, Med=40, Q3=60, Max = 300



b) Since Q3 = 60, hence 75% of students exercise from 0 to 60 min.

c) Since Q1 = 20 and Q3=60, Hence 50% of students exercise from 20 to 60 min.

The *IQR* is used to determine potential **outliers**.

1. Find the quartiles Q_1, Q_2, and Q_3.
2. Find the interquartile range (IQR), where IQR $= Q_3 - Q_1$.
3. Evaluate $1.5 \times$ IQR.
4. **In a modified boxplot, a data value is an *outlier* if it is above Q_3, by an amount greater than $1.5 \times$ IQR or below Q_1, by an amount greater than $1.5 \times$ IQR**

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

C) Modified boxplot and outliers:

A modified boxplot can be graphed to show outliers without calculating IQR and applying the Q1-1.5IQR, Q3+1.5IQR. Outliers are shown as markers in the boxplot.

- use Statdisk, click data , Boxplot,
- Select the column of data, click **modified boxplot**. The outlier will be shown as marker at the lowest or highest end of the boxplot.
- If there are no markers, there is no outliers in the dataset.
- To find the values of the outlier, sort the data. The outliers will be at the top and end of the sorted data.

Ex2. Determine if outliers exist in the exercise time from 15 students.

0, 40, 60, 30, 60, 10, 46, 30, 300, 90, 30, 120, 60, 0, 20

By calculation:

Since Q1 = 20, Q3 = 60, So IQR = 60 – 20 = 40

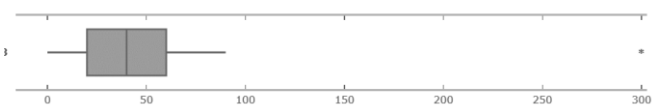
Lower fence = Q1 – 1.5(IQR) = 20 – 1.5(40) = -40

upper fence = Q3 + 1.5(IQR) = 60 + 1.5(40) = 120

Values lower than -40 and higher than 120 is an outlier. So the value 300 is an outlier.

Graph a modified boxplot to identify outliers.

Use Statdisk, Data/Boxplot/Modified boxplot.



There is one outlier in the high end of the data.

Use Statdisk, Sort, one column, select the column containing the data. The last data (300) is the outlier.

Ch 2.5 and 2.6 Measure of Center and skewness

Center: A measure of center is a value at the center or middle of a data set. It is used to provide a representative value that "summarize" the data.

Measure of center: (mean, median, mode, midrange)

1**) Mean**: the average of the data.

$sample\ mean\ \bar{x} = \frac{\sum x}{n} = \frac{sum\ of\ all\ data\ values}{number\ of\ data\ values}$;

n = sample size

$population\ mean\ \mu = \frac{\sum x}{N} = \frac{sum\ of\ all\ data\ values}{number\ of\ data\ values}$;

N =population. size.

$\bar{x}$ is read as x-bar, μ is read as mu (Greek letter)

*the term "average" is not used by statistician.

Important properties of mean:

a) Use every data value (in a sample or population)

b) Extreme value can change the value of mean substantially. Mean is not resistance.

c) $\bar{x}$ can be used to estimate μ if sample is not bias. If the sample is voluntary response or biased $\bar{x}$ will not be a good estimate of μ.

Ex1: Mean of 78, 89, 75, 92, 66, 82 is 80.3

Mean of 78, 89, 75, 92, 66, 82, 5 is 69.6 (not middle)

Ex2: Mean of top 5 scores: 90, 95, 92, 98, 89; $\bar{x}$ is 92.8 but is not a good estimate for the whole class μ because the sample is bias, not from random selection.

2) Median: the middle value when the data are arranged in order of increasing or decreasing magnitude.

How to calculate the median:

a) Sort the values

b) Odd number of values: Median is the middle one

Even number of values: Median is the mean of the middle two values.

Properties of Median:

a) The median does not change when we add a few extreme values. It is resistance.

b) Median does not use every value.

Ex1: Median of 78, 89, 75, 92, 66, 82 is 80

Median of 78, 89, 75, 92, 66, 82, 5 is 78 (middle)

3) Mode: the value (s) that occur(s) with the greatest frequency. Can be calculated for qualitative data. Not common for quantitative continuous data.

Properties of mode:

a) Data can have one, two, multiple or no modes.

b) Bimodal – two data values occur with same greatest frequency.

c) Multimodal – more than two data values occur with the same greatest frequency.

d) No mode – no data value is repeated.

4) **Midrange**: the value that is midway between the maximum and minimum values in the dataset.

$Midrange = \frac{maximum+minimum}{2}$

Properties of Midrange:

a) Not resistance to extreme values.

b) Easy to compute but rarely used.

c) Midrange does not use all data. It is not the median and it is not half of range.

Find Mean, median, and midrange:

Use Statdisk - Explore Data, Descriptive statistics to find mean, median, midrange .

Find mode:

Use online calculator: (for multiple mode)

https://www.calculatorsoup.com/calculators/statistics/mean-median-mode.php

Input data to the window, click Calculate.

Round off rules:

a) Mean, median and Midrange: carry one more decimal place than original data.

b) Mode: leave the value as is without rounding.

Ex1. Find the mean, median, midrange, mode of the length of boats (in ft) parked in a marina.

16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

a) Find the mean, median, mode and midrange. Which of the above is the best measure of center?

b) Can this mean be used to estimate mean length of all boats in all marina? Explain.

a) Use Statdisk: Mean = 27.3, Median = 27, Midrange = 28, Mode = 25 and 27.

Mean is a better choice because data is symmetrical and there are no outliers in the data.

b) This mean is not representative because the sample is from one marina only.

Ex. 2 Which is the greatest, mean, median or mode?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

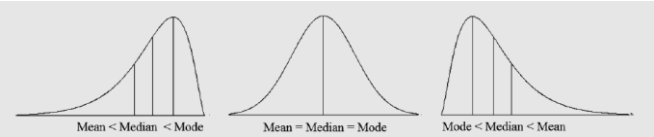
Mean =

Median =

Mode =

Midrange =

Mean, median and skewness:



Choose the best measure of center:

a) Use mode for nominal data as center.
b) Use mean if there is no extreme data.
c) Use median if extreme data exist.
d) Use median if data is skewed.

Find mean from a frequency distribution:

$\bar{x} = \frac{\sum(f \cdot x)}{\sum f}$ is the mean from GFDT where

x is the class midpoint of each class.
f is the frequency of each class

To **find mean for GFDT**, use Statdisk

- Click data, frequency Table Generator.
- Enter lower limit to start, upper limit to end, frequency to freq for each row in the GFDT.
- scroll to the bottom and select the Editor column for generated data. (make sure the column is available.)
- Click generate, a sample of data will be generated.
- Use Data, Explore Data- Descriptive statistic and select the column containing the generated data to find the grouped mean.

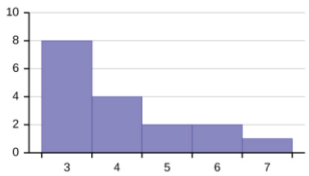
Ex1. Find grouped mean from GFDT:

Grades	Freq
50 - 59	1
60 - 69	0
70 - 79	2
80 - 89	4
90 - 99	5

i) Select Frequency Table Generator in Statdisk and generated data in an empty column.

ii) Select Explore data – Descriptive Statistic to find the mean.

Ex2. Find the mean, median from the data in the histogram.



Write the data out:

The mean and median are about the same but the data are very skewed to the right.
Note: when data is extremely skewed in a particular way, the mean and median can be the same.

Try Book 2.31

Ch 2.7 Measure of Spread and Variation

A measure of variation is used to describe the degree of spread of data.: **Range, standard deviation, Interquartile Range.**

1) Range: Difference between max and min.
= (maximum value – minimum value)
Properties:
a) Not resistance. Affected by extreme value.
b) Does not take every value into account.

2) **Standard Deviation of a <u>sample</u>**: Measure of how much data values deviate away from the mean.
Notation: s = sample standard deviation

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

n = sample size. (n -1) = degree of freedom.

Properties:

a) s is never negative. It is zero when all data values are the same. Large value of s indicates greater amount of variation.
b) s increases dramatically with one or more outliers. Not resistance to extreme values.
c) s has the same unit as the original data values.
d) s is a biased estimator of population standard deviation. It does not center around the value of σ.
e) Compare standard deviation for data with similar mean only.
f) s increases when data are more spread out.

Ex1:

Find standard deviation of 15, 15,17, 21 by manual calculation : $\bar{x} = 17$;

x	$x-\bar{x}$	$(x-\bar{x})^2$
15	-2	4
15	-2	4
17	0	0
21	4	16

$s = \sqrt{\frac{4+4+0+16}{3}} = \sqrt{8} = 2.8$

Ex2. Find range, standard deviation of the following data: 5, 6, 3, 2, 6, 8, 10, 12, 17

Use Statdisk:

- Enter data in column 1, Data-Explore Data-Descriptive Statistics, Select column 1,

Sample standard deviation = 4.72 (rounded)
Range =

Ex 3. Find standard deviation of Grouped data.

Grades	Freq
50 - 59	1
60 - 69	0
70 - 79	2
80 - 89	4
90 - 99	5

Use the same procedure as finding Grouped mean. Use Statdisk – Frequency Table Generator and Descriptive Statistics.

s = 12.06

3) IQR: Q3 – Q1 = middle 50% range of data. IQR is resistance to extreme values.

Ex1. Find Range, Standard Deviation and IQR for

dataset A: 5, 6, 3, 2, 6, 8, 10, 12, 17

dataset B: 5, 6, 3, 2, 6, 8, 10, 12, 17, 60

	Range	SD	IQR
Dataset A:	15	4.7	5
Dataset B:	58	17.1	7

IQR is not influenced by one extreme data.

Other measures related to measure of variation.

1) Population Standard deviation (σ)

$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$ where x are data from population

N = population size.

Find Population standard deviation online:

https://www.socscistatistics.com/descriptive/variance/default.aspx

2) Variances:

Population variance: σ^2:

Sample variance: s^2

Variances are used in statistical analysis.

Relative standing and Z-score.

Standard deviation is used to describe how far away a value is from the mean.

Sample data = mean + (# of stdev) * stdev.

When comparing values from different dataset, it is best to compare how each value is from their respective mean. The number of standard deviation is a measure of relative standing.

z-score is the number of standard deviations a data value is from the mean.

$z = \frac{x-\mu}{\sigma}$ for population data.

$z = \frac{x-\bar{x}}{s}$ for sample data.

Round-off rule: round to 2 decimal places.

Properties of z-score:

a) z-score tells how many standard deviations a value is from the mean. Negative means below the mean. Positive means above the mean.0

b) z-score has no units.

Ex1:

A student's English score is 83 when mean = 80 with sd = 10. The student's History score is 75 when class mean is 72 with sd= 7.

Is the score better in English or History?

English z-score = $\frac{83-80}{10}$ =0.30

History z-score = $\frac{75-72}{7} = 0.43$

The student score less than 1 standard deviation from the mean but since 0.43 > 0.30, hence the student scores better in History.

Ex2.

Ages of 20 fifth graders are given below:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

a) Find the mean and standard deviation of ages.

b) What age is one standard deviation above the mean?

c) What age is two standard deviation below the mean?

d) How many data are two standard deviation below the mean or two standard deviation above the mean?

Ex3.

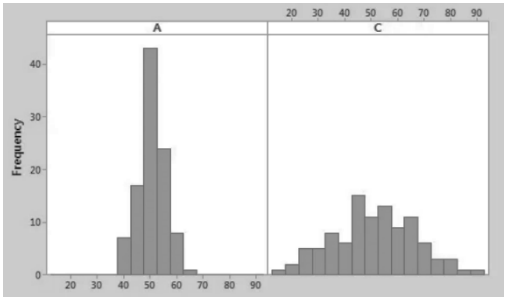
Measurements of diameter of a bottle cap manufactured in a factory are collected, what situation is best? Explain.

a) High standard deviation.

b) Low standard deviation.

Ex4.

The following histogram show distribution of measurement of diameters of bottle cap from two production line.



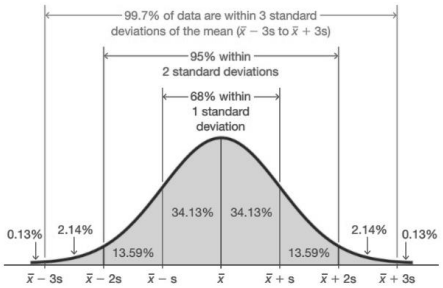
Which histogram,A or C, represents sample with low standard deviation? Explain.

Empirical Rule:

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.

More than 99% of the data is within three standard deviations of the mean



Ex1. The distribution of daily number of cash withdrawal from an ATM machine has a bell-shape distribution with a mean of 42 and a standard deviation of 6.

a) Use the empirical rule to calculate the percentage of cash withdrawal falls between 30 and 54.

c) What percent of cash withdrawal falls between 24 and 60?