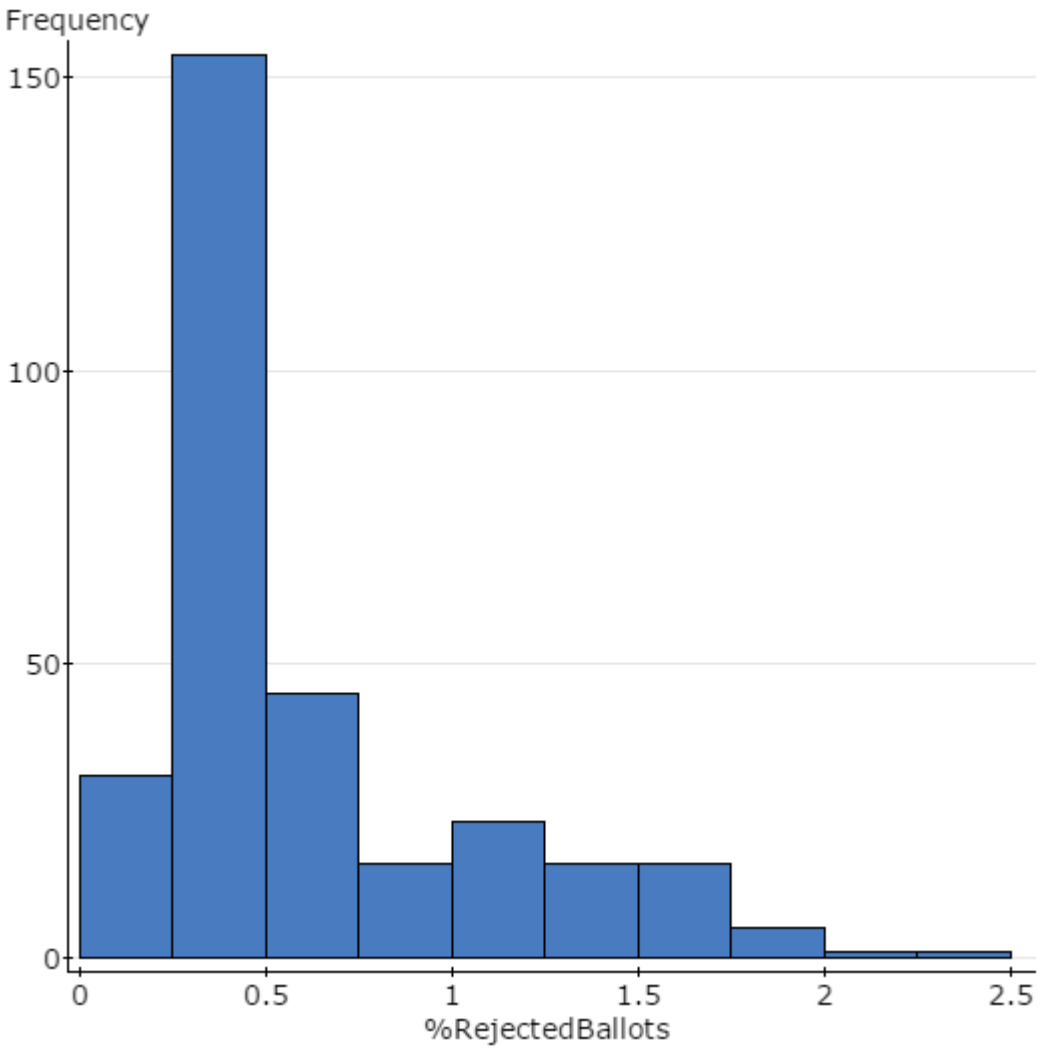
Chapter 3: Displaying and summarizing quantitative data p52

The pattern of variation of a variable is called its **distribution**.

Histograms p53

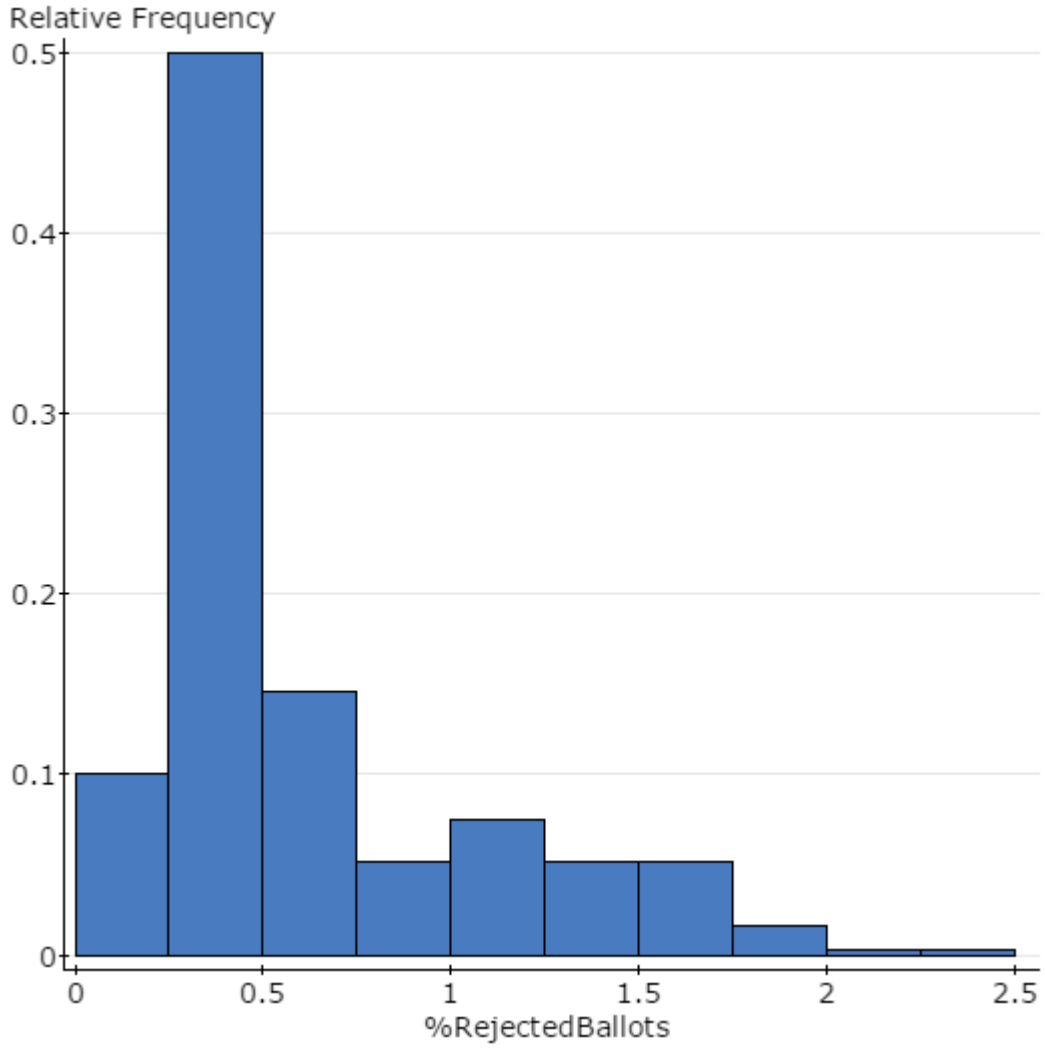
Spoiled ballots are a real threat to democracy. Below are displays of data from Elections Canada showing percentage of spoiled ballots (number of spoiled ballots divided by total number of ballots cast) for all 308 electoral ridings in the 2006 Canadian federal election



- % rejected ballots varies from 0 to 2.5%

- Most of them in the range 0.25%-0.5%..
- A very few electoral districts have more than 2% rejected ballots.
- Shape *?*

Relative frequency histogram (Relative frequency distribution)



Describing a distribution (*p57*)

- •In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern (e.g.. gaps. P54)

-

- –Overall pattern of a distribution can be described by its **shape**, **centre**, and **spread**.

-

- –An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

-

- –For now we can describe the centre of a distribution by the median (the midpoint)

-

- –Some other things to look for in describing shape are:

-

- •Does the distribution have one or several major peaks, called **modes**? A distribution with one major peak is called **unimodal**.

-

- •Is it approx. symmetric or skewed in one direction.

Stem-and-leaf plots

Variable: %RejectedBallots

Decimal point is 1 digit(s) to the left of the colon.
Leaf unit = 0.01

```
 1 : 79999
 2 : 000011112222223333333444445556667777777777888888899999999
 3 : 00000011122222222333333333333334444455555555566666666677777777788888999999
 4 : 00111112222223334444444444555555666666677788888899
 5 : 00001112222336666888999
 6 : 001113345578899999
 7 : 013366779
 8 : 022237
 9 : 33488
10 : 0113568
11 : 12233444669
12 : 002245666889
13 : 229
14 : 022445
15 : 02256
16 : 223444556
17 : 14
18 : 2468
19 : 6
20 : 8

High: 2.32
```

- same shape , i.e. skewed to the right (turned on side)
- unusual values listed at bottom (will also be at the top if there are unusually small values)

Percentage Voter turnout:

Variable: %VoterTurnout

Decimal point is at the colon.
Leaf unit = 0.1

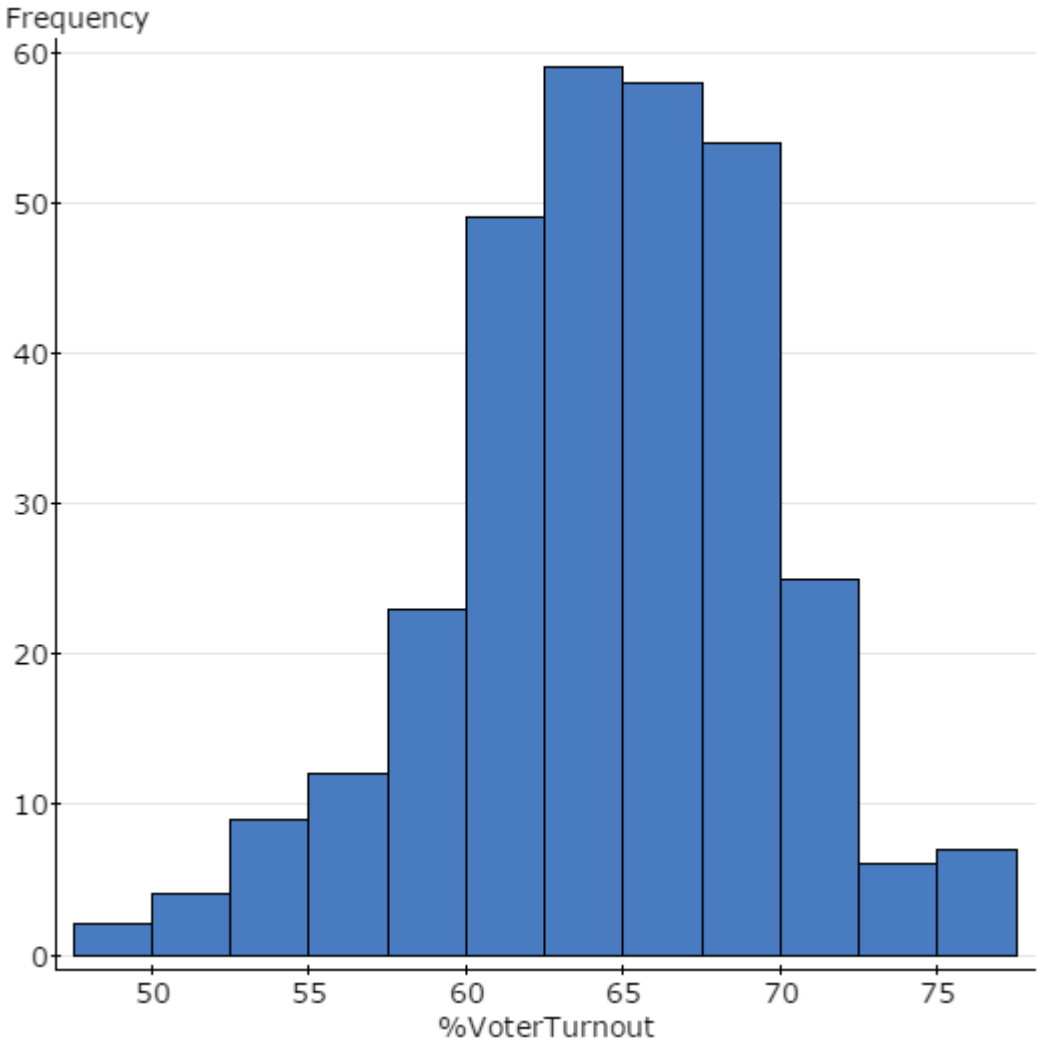
```
48 : 3
49 : 0
50 : 46
51 :
52 : 118
53 : 3678
54 : 1278
55 : 2334
56 : 2334
57 : 000399
58 : 012334779
59 : 012245667779
60 : 0001223334456689
61 : 111112344577778999
62 : 00111123334444556778889
63 : 000011223455567788888889
64 : 000001111222344567778999
65 : 02233333345678999
66 : 000111222333444555666667778899999
67 : 11112333455556666666777899999
68 : 00122333445556789
69 : 0122334445578899
70 : 2334455667789
71 : 00347789
72 : 12248
73 : 0279
74 : 8
75 : 1123568
```

What is the max value?

What is the min value?

What is the shape of the distribution?

Here is the histogram



Centre of the Distribution p 57

The mean and median (measures of "centre")

- **Measuring center: mean**
  - Two common measures of center are the *mean* and the *median*.
  - The two measures behave differently.
- Example
  Find the mean of the following observations.

  4, 5, 9, 3, 6

  Solution:

$$mean = \frac{4+6+9+3+6}{5} = \frac{28}{5} = 5.6$$

- If there are *n* observations $y_1, y_2, \ldots, y_n$ in a sample, the sample mean (denoted by $\bar{y}$) is given by

$$\bar{y} = \frac{sum\ of\ \ y's}{n} = \frac{\sum y}{n}.$$

- Example
  The annual salaries (in thousands) of a random sample of five employees of a company are:

  40, 30, 25, 200, 28

$$mean = \frac{40+30+25+200+28}{5} = \frac{323}{5} = 64.6$$

If we exclude 200 as an outlier,

$$mean = \frac{40+30+25+28}{4} = \frac{123}{4} = 30.75$$

- Mean is sensitive to the influence of extreme observations. It cannot resist influence of the extreme values. Mean is NOT a **resistant measure** of center. (*p62*)

Measuring center: the median (p62)

The median is the midpoint of the distribution, the number such that half the observations are smaller and other the half are larger.

- To find the median of a distribution:
    1. Arrange all observations in order of size, from smallest to largest.
    2. If the number of observations is odd the median is the center observation in the ordered list.
    3. If the number of observations is even the median is the average of the two center observations in the ordered list.
- Examples
    1. The annual salaries (in thousands) of a random sample of five employees of a company are:
    40, 30, 25, 200, 28

Arranging the values in increasing order:

25 28 30 40 200
median = 30

Excluding 200 median = (28+30)/2.

- Note that the mean for this data set was 64.6 and the influence of the extreme value 200 is much less.

- StatCrunch commands *Stat > Summary Statistics*
- StatCrunch output for the data in Example above is as follows:

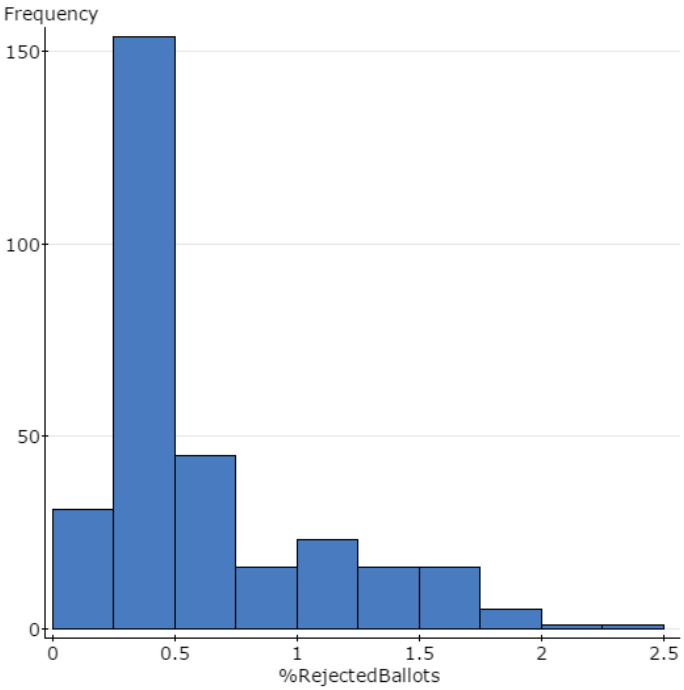
Summary statistics:

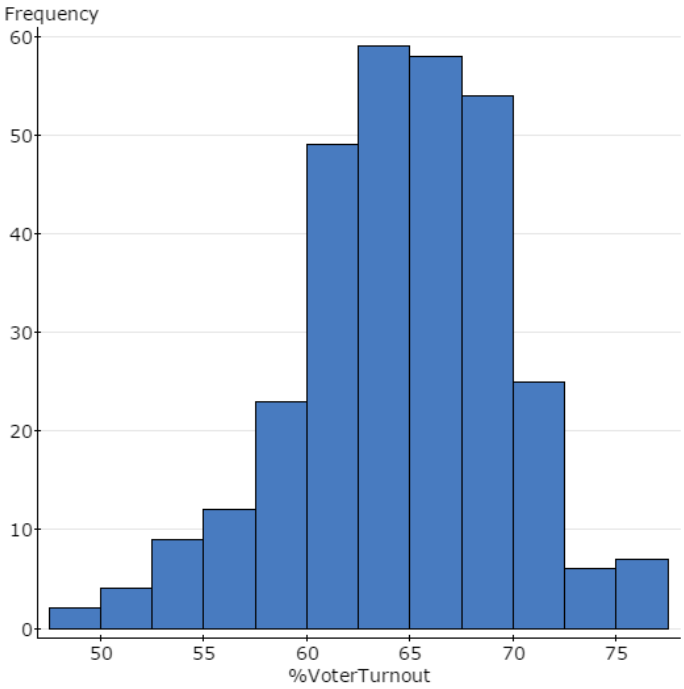
Column	n	Mean	Median
salary	5	64.6	30

Mean versus median

- The median and mean are the most common measures of the center of a distribution.
- If the distribution is exactly symmetric, the mean and median are exactly the same.
- Median is less influenced by extreme values.
- If the distribution is skewed to the right, median < mean
- If the distribution is skewed to the left, mean < median

Example:





Summary statistics:

Column	n	Mean	Median
%RejectedBallots	308	0.6174026	0.44
%VoterTurnout	308	64.512662	64.75

Questions

1. You are asked to recommend a measure of center to characterize the following data:
   0.6, 0.2, 0.1, 0.2, 0.2, 0.3, 0.7, 0.1, 0.0, 22.5, 0.4.

   What is your recommendation and why?

2. The mean is ____ sensitive to extreme values than the median
   a) more
   b) less
   c) equally
   d) can't say without data

3. Changing the value of a single score in a data set will necessarily cause the mean to change. (T/F)

4. Changing the value of a single score in a data set will necessarily cause the median to change. (T/F)

Spread: interquartile range p68

- 1st quartile Q_1 has ¼ of data values below it and ¾ above
- 3rd quartile Q_3 has ¾ of data values below it and ¼ above
- Find a quartile by taking lower (upper) half of data, and finding median of that half.
- Interquartile range is IQR=Q_3-Q_1. Larger = more spread out.

Example: 3, 5, 7, 7, 8

- lower half 3, 5, 7 (include middle), so Q_1=5
- upper half 7, 7, 8 so Q_3= 7
- IQR=7-5=2
- *IQR not affected by extremely high or low values, like median.*

- **Measuring spread Standard deviation** (p64)

  The variance (s^2) of a set of n observations $y_1, y_2, \ldots, y_n$ is

$$s^2 = \frac{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}{n-1} = \frac{\sum(y-\bar{y})^2}{n-1}.$$

  The standard deviation(s) is the square root of the variance (s^2).

i.e. $$s = \sqrt{\frac{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}{n-1}} = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$$

Example

Find the standard deviation of the following data set: 5, 8, 2

$$n=3,\ \text{Mean}\ (\bar{x}) = \frac{5+8+2}{3} = \frac{15}{3} = 5$$

$$s^2 = \frac{(5-5)^2+(8-5)^2+(2-5)^2}{3-1} = \frac{18}{2} = 9$$

$$s = \sqrt{9} = 3.$$

StatCrunch: Percentage rejected ballots

Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
%RejectedBallots	308	0.6176	0.1965	0.4432	0.02526	0.44	2.15	0.17	2.32	0.32	0.76

Notes

s measures the spread about the mean $\bar{x}$.

$s = 0$ only when there is no spread. This happens only when all observations have the same value.

s, like the mean $\bar{x}$, is not resistant.

- **The five-number summary p69**
  - The **five-number** summary of a set of observations consists of the minimum, the first quartile, median, the third quartile and the maximum.
  - These five numbers give a quick summary of the both center and the spread of the distribution.
  - StatCrunch commands:*Stat > Summary Statistics*

Example

Data 10, 11, 14, 15, 17, 19, 21, 28, 35:

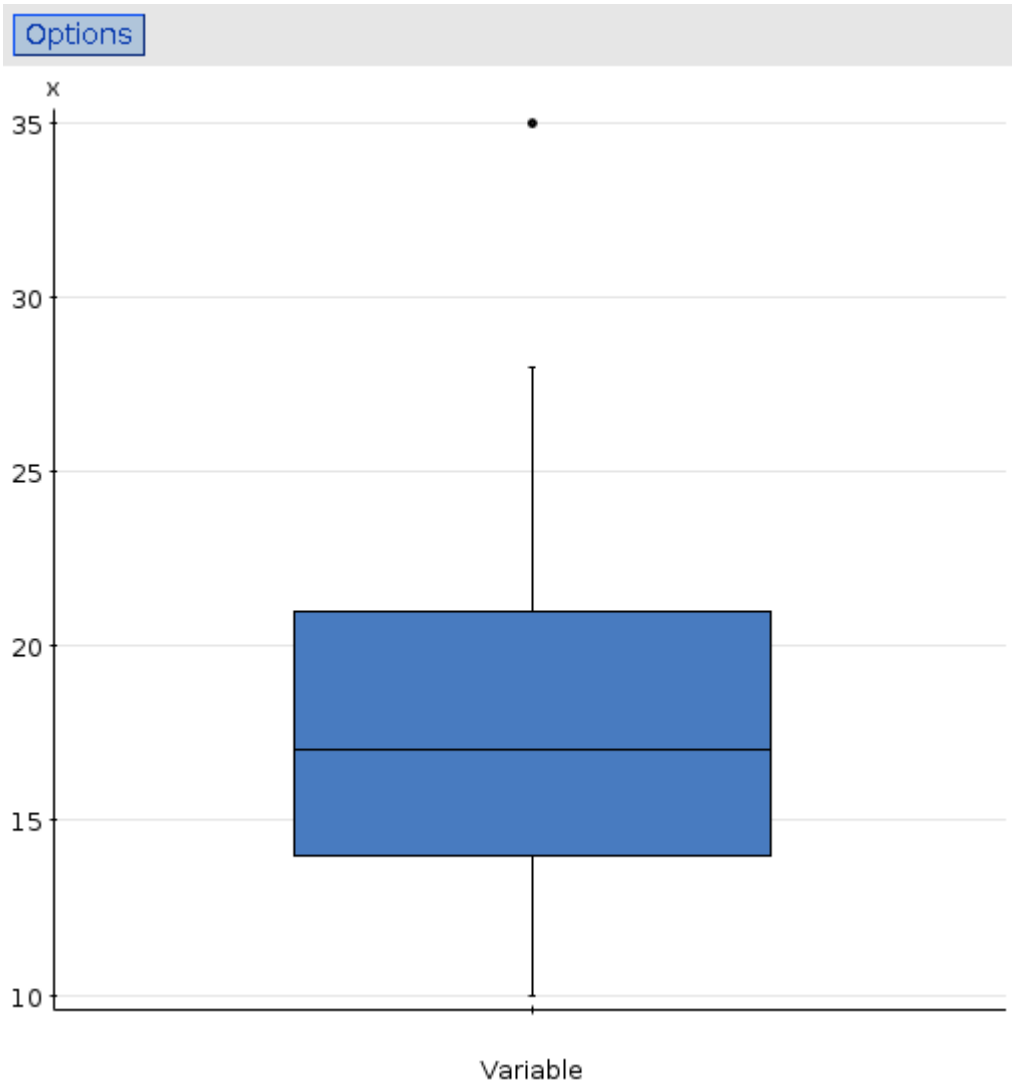
- find median
  - 17, Q1= median of 10,11,14,15, 17 ie. 14
  - Q3=median of 17, 19, 21, 28, 35 ie. 21
- find Q_1 and Q_3
- find interquartile range
  - 21-14=7
- find 5-number summary min, Q_1, median, Q_3, max.
  - 10, 14, 17, 21, 35

Boxplot (numbers from example above) p70

- box goes down the page, with scale on left.
- centre of box at **median (17)**
- top of box at (21) $\mathbf{Q_3}$
- bottom of box at (14) $\mathbf{Q_1}$
- calculate R=1.5 x IQR= 1.5(21-14)=10.5
  - *upper fence* at Q_3+R=21+10.5=31.5
  - *lower fence* at Q_1-R= 14-10.5=3.5
- draw lines connecting box to most extreme value within fences

- plot values outside fences individually. These are suspected outliers and deserve to be investigated.

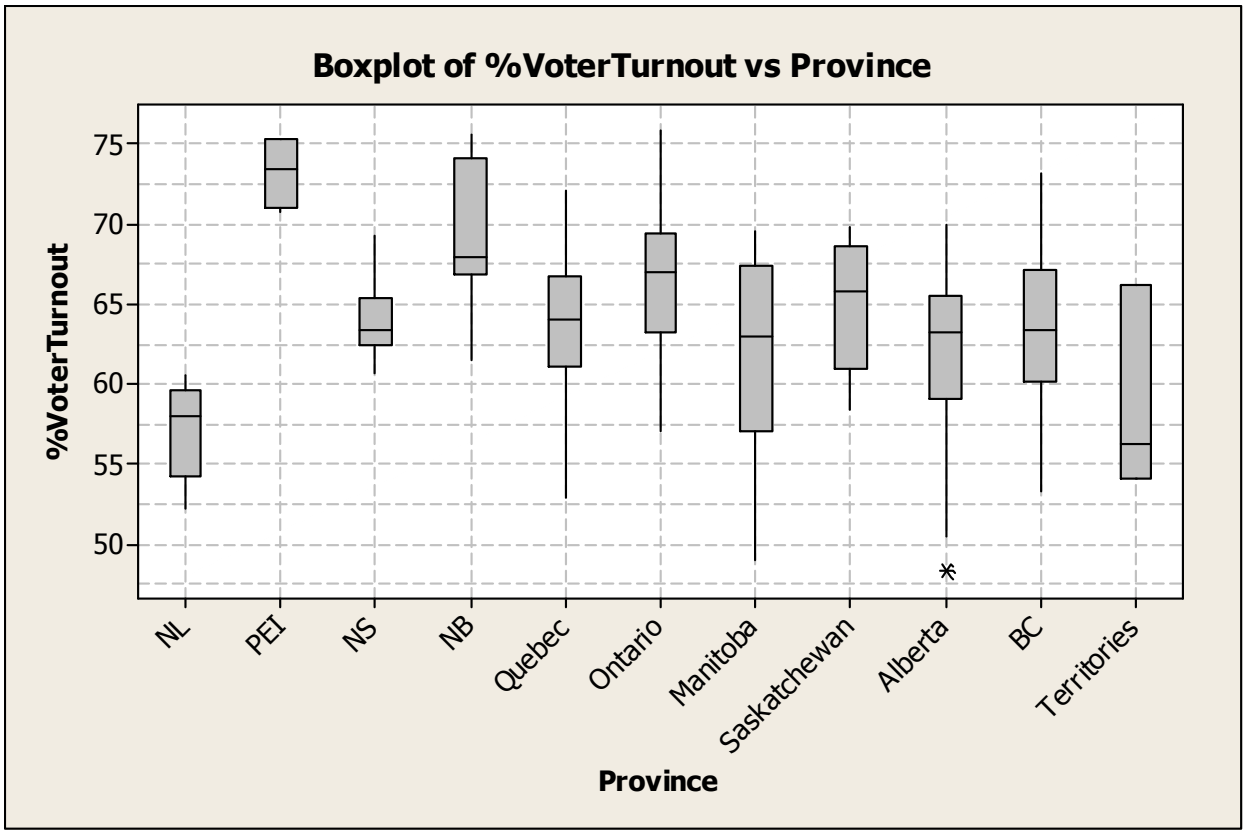
StatCrunch boxplot (select “use fences to identify outliers”):

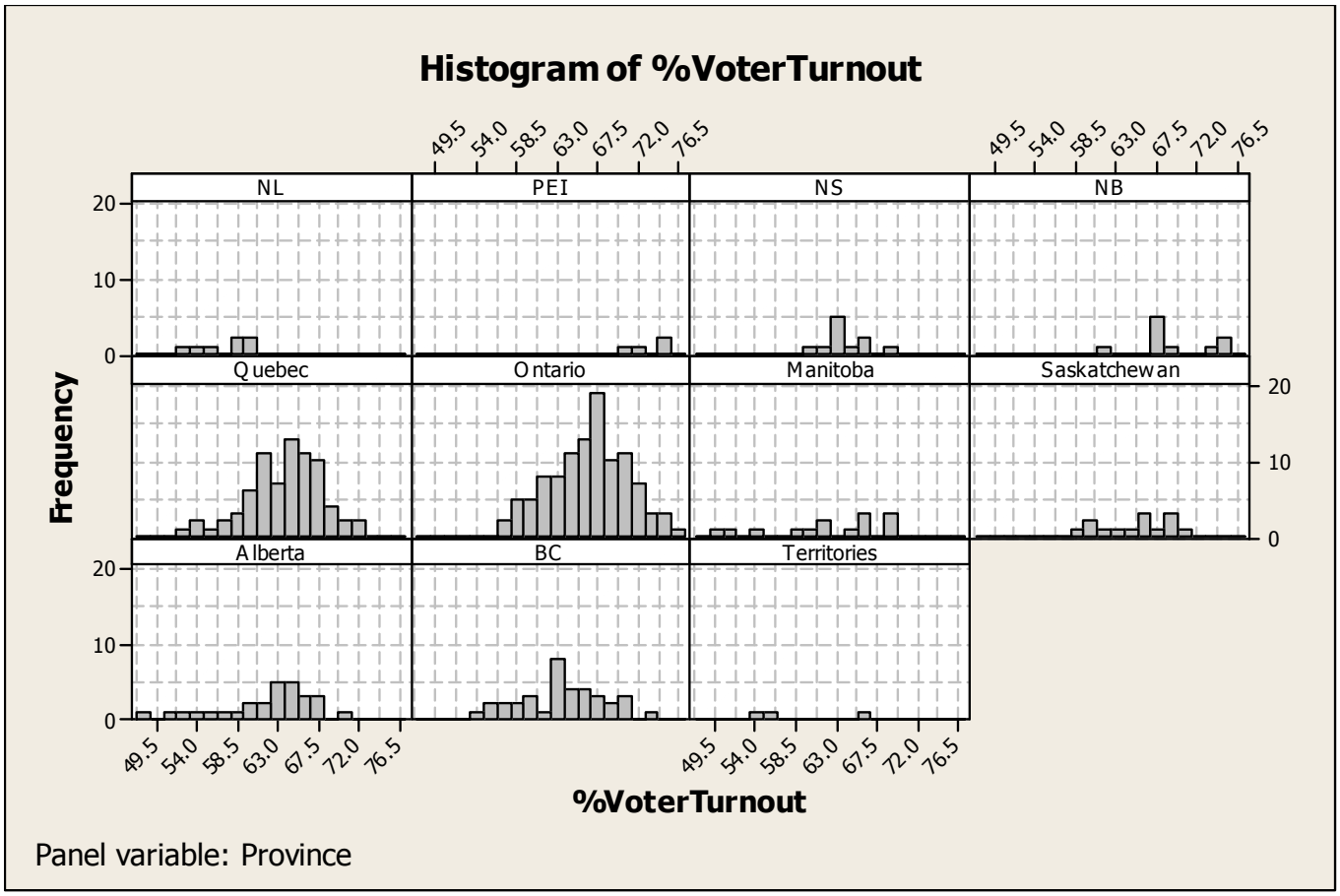


Chap 4: Understanding and Comparing Distributions p99

Election 2006

Side-by-side boxplots for voter ternout in different provinces




Histogram of %VoterTurnout
NL
PEI
NS
NB
Quebec
Ontario
Manitoba
Saskatchewan
Alberta
BC
Territories
Frequency
%VoterTurnout
49.5
54.0
58.5
63.0
67.5
72.0
76.5
0
10
20
Panel variable: Province

Choosing a summary p76 (in ch 3)

The five-number summary is usually better than the mean and the standard deviation for describing skewed distributions or distributions with strong outliers.

Use mean and std. deviation for reasonably symmetric distributions that are free of outliers.