Week 1

Exploring Data

ST 314
Introduction to Statistics for Engineers

Oregon State University

Visualizing Data

Common graphical displays used for quantitative variables:

- Histograms
- Stem plots
- Box plots

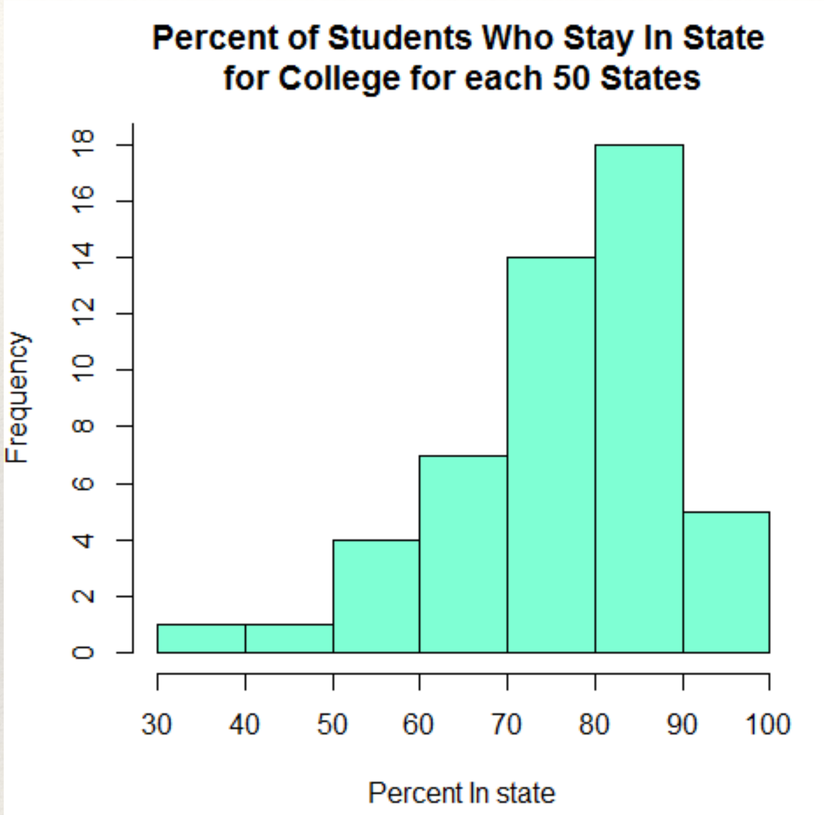
Common graphical displays used for categorical variables:

- Bar charts
- Pie charts

Quantitative Variables: Histograms

state	%	state	%	state	%	state	%
AK	83	IL	83	NC	83	SC	79
AL	75	IN	80	ND	55	SD	68
AR	82	KS	80	NE	81	TN	84
AZ	70	KY	83	NH	55	TX	92
CA	92	LA	86	NJ	92	UT	76
CO	77	MA	63	NM	82	VA	76
CT	69	MD	81	NV	91	VT	36
DE	60	ME	74	NY	81	WA	84
FL	84	MI	91	OH	85	WI	81
GA	83	MN	79	OK	80	WV	59
HI	81	MO	80	OR	76	WY	63
IA	62	MS	85	PA	73		
ID	68	MT	73	RI	44		

Bins	Frequency
31-40	1
41-50	1
51-60	4
61-70	7
71-80	14
81-90	18
91-100	5



- Bins divide data into groups.
- Number of observations in each bin represented by a bar.
- Bar height is on vertical axis and variable range is on horizontal axis.

Quantitative Variables: Stemplots

The stemplot displays the distribution of a quantitative variable with the raw data values.

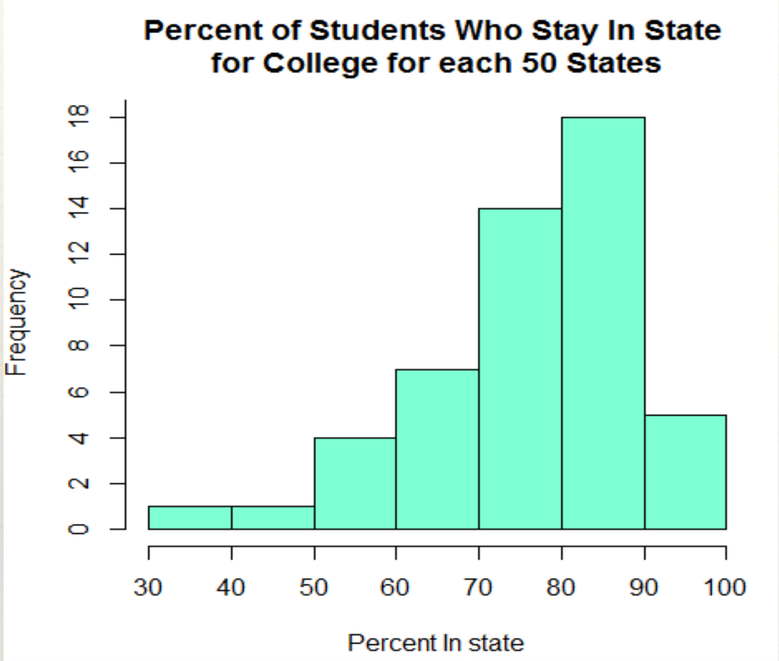
- Each value is ordered and separated into a "stem" and a "leaf" based on place value.
  - Stem = first part of number
  - leaf = remaining part of number
- Stems organized in a vertical column with leaves ascending order to the right.

state	%	state	%	state	%	state	%
AK	83	IL	83	NC	83	SC	79
AL	75	IN	80	ND	55	SD	68
AR	82	KS	80	NE	81	TN	84
AZ	70	KY	83	NH	55	TX	92
CA	92	LA	86	NJ	92	UT	76
CO	77	MA	63	NM	82	VA	76
CT	69	MD	81	NV	91	VT	36
DE	60	ME	74	NY	81	WA	84
FL	84	MI	91	OH	85	WI	81
GA	83	MN	79	OK	80	WV	59
HI	81	MO	80	OR	76	WY	63
IA	62	MS	85	PA	73		
ID	68	MT	73	RI	44		

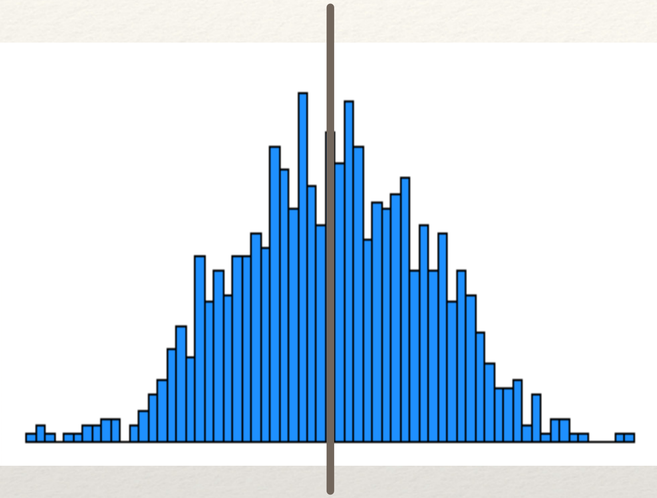
Stem	Leaves
3	6
4	4
5	5 5 9
6	0 2 3 3 8 8 9
7	0 3 3 4 5 6 6 6 7 9 9
8	0 0 0 0 1 1 1 1 1 2 2 3 3 3 3 3 4 4 4 5 5 6
9	1 1 2 2 2

Describing Distributions

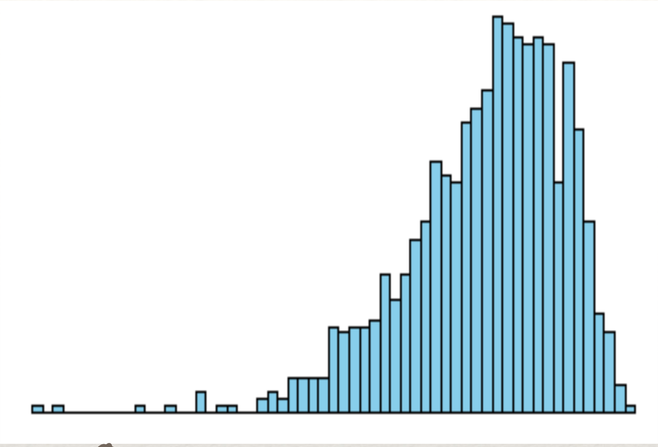
shape	Describe the overall pattern of the data. Include whether the data is **symmetric**, **skewed**, **unimodal** or **multimodal**.
center	Use the **mean** or the **median** to describe the center of the data.
spread	The **range**, **interquartile range**, or **standard deviation** can be used to describe the spread of the data.
outliers	Remark on observations that fall outside the pattern of the data or **extreme values** that fall far from the center.



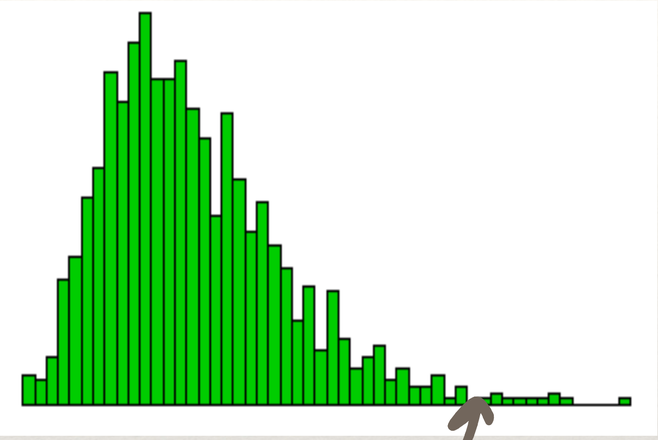
Symmetric vs. Skewed



symmetric



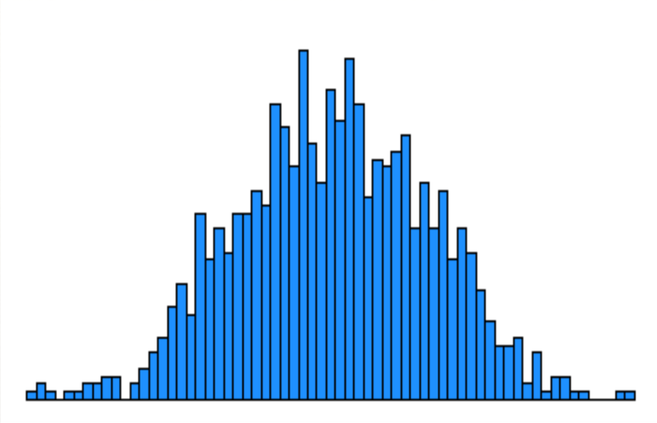
Left-skewed



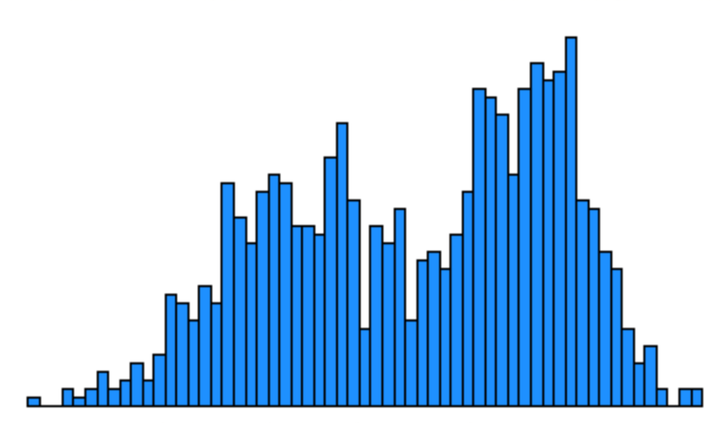
Right-skewed

Unimodal, Bimodal, Multimodal

How many peaks?



Unimodal

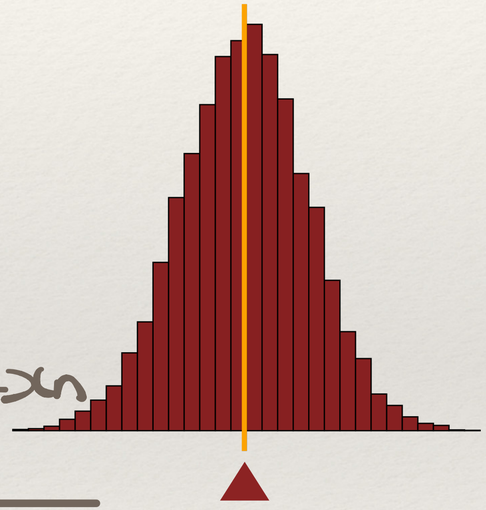


Bimodal

Mean

The **mean** is the average or expected value of a quantitative characteristic of interest. The mean equally weights each observation value in the set.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$



Median

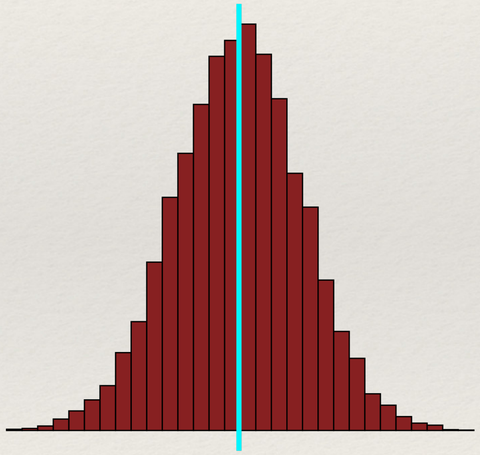
The **median** is the value in which approximately 50% of observations fall below.

The **sample median** denoted by **M**, represents the middle observation when data is ordered from smallest to largest.

If n is odd: $\left(\frac{n+1}{2}\right)^{th}$ obs.

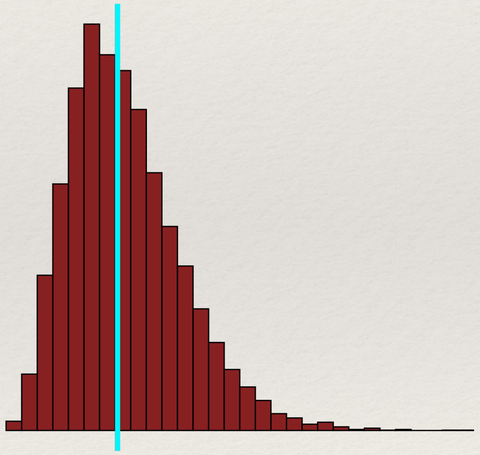
If n is even: average of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ obs.

Symmetric Distribution



Median

Skewed Distribution



Median

Calculating Mean & Median Example

These data are from applicants to a rigorous engineering course that requires students to take an entrance exam prior to admission into the course.

Individual	ExamScore	GPA	Gender	Year
1	79	3.3	Man	Junior
2	88	2.9	Man	Senior
3	92	3.8	Non-binary/non-conforming	Soph.
4	87	3.7	Man	Junior
5	68	3.4	Woman	Junior

Calculate the mean exam score for this sample:

$$\bar{x} = \frac{79+88+92+87+68}{5} = 82.8$$

Calculate the median exam score for this sample:

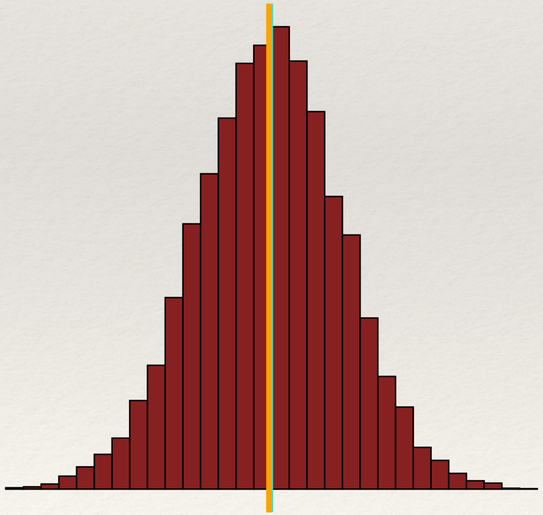
68 79 87 88 92

$\mu = 87$

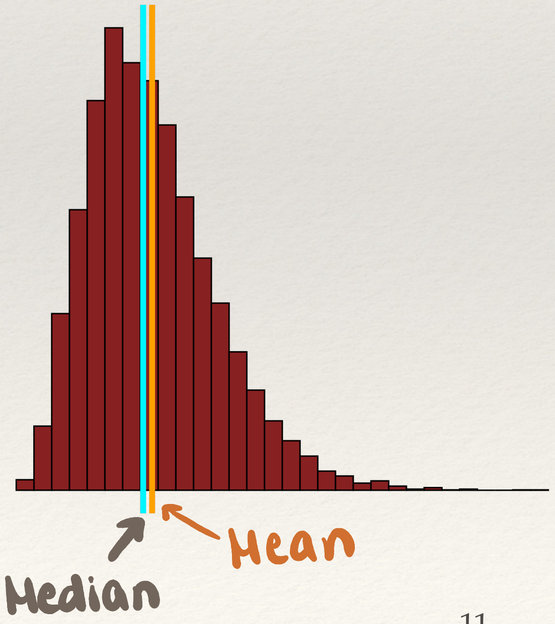
Comparing Mean & Median

When data is skewed or has outliers the mean will be influenced by the outliers or extreme observations.

Symmetric Distribution



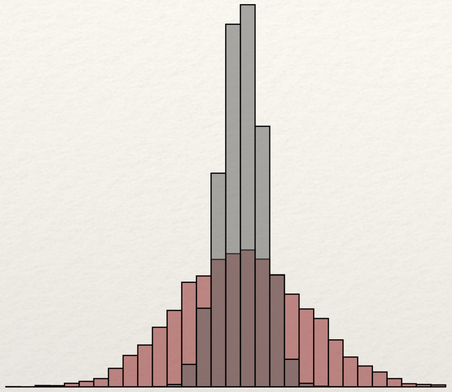
Skewed Distribution



"Robust"

The median is a better representation of a typical value for skewed data.

Standard Deviation



The **standard deviation** is a measurement for the amount of spread or variability of a quantitative random variable.

The **sample standard deviation,** s, describes the typical deviation from $\bar{x}$ in a *sample*.

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

The standard deviation is highly influenced by skewed data and extreme observations.

Standard Deviation Example

These data are from applicants to a rigorous engineering course that requires students to take an entrance exam prior to admission into the course.

Individual	ExamScore	GPA	Gender	Year
1	79	3.3	Man	Junior
2	88	2.9	Man	Senior
3	92	3.8	Non-binary/non-conforming	Soph.
4	87	3.7	Man	Junior
5	68	3.4	Woman	Junior

Calculate the standard deviation exam score for this sample:

① Start by calculating the deviations for each observation $(x_i - \bar{x})$:

$79 - 82.8 = -3.8$
$88 - 82.8 = 5.2$
$92 - 82.8 = 9.2$
$87 - 82.8 = 4.2$
$68 - 82.8 = -14.8$

② Square each deviation:

$(-3.8)^2 = 14.44$
$(5.2)^2 = 27.04$
$(9.2)^2 = 84.64$
$(4.2)^2 = 17.64$
$(-14.8)^2 = 219.04$

③ Find the sum of the squared deviations:

$14.44 + 27.04 + 84.64 + 17.64 + 219.04 = 362.8$

④ Divide the sum of the sum of the squared deviations by $n-1$:

$n = 5 \Rightarrow n-1 = 4$

$\frac{362.8}{4} = 90.7$

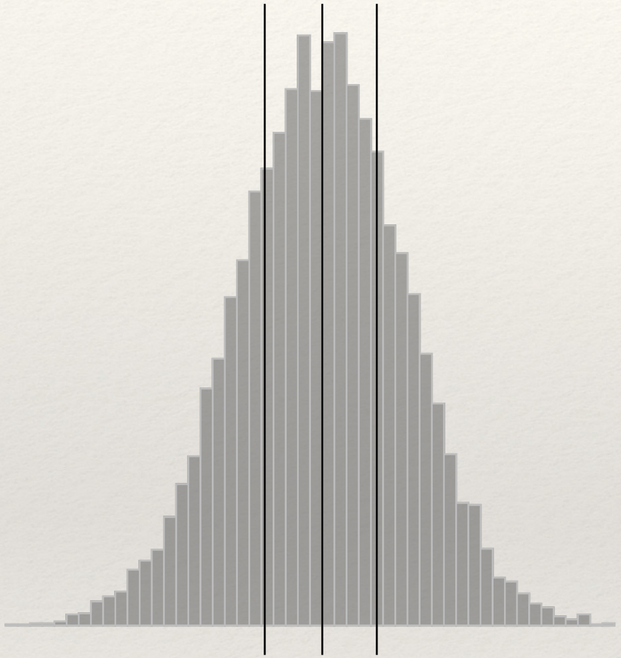
if you stop here, you've calculated the variance. Go to step ⑤ for the standard deviation.

⑤ Take the square root of the variance (the value calculated in the previous problem).

$s = \sqrt{90.7} = 9.52$

The typical deviation from the sample mean for the exam score data set is 9.52 points.

Quartiles & IQR



25th percentile:
25% of observations fall below this value

75th percentile:
75% of observations fall below this value

The values that cut the data into four quarters are the **quartiles** which are the 25th, 50th and 75th percentiles.

The ***1st quartile*, Q1**, and ***3rd quartile*, Q3,** are the respective medians of the first and second half of the data.

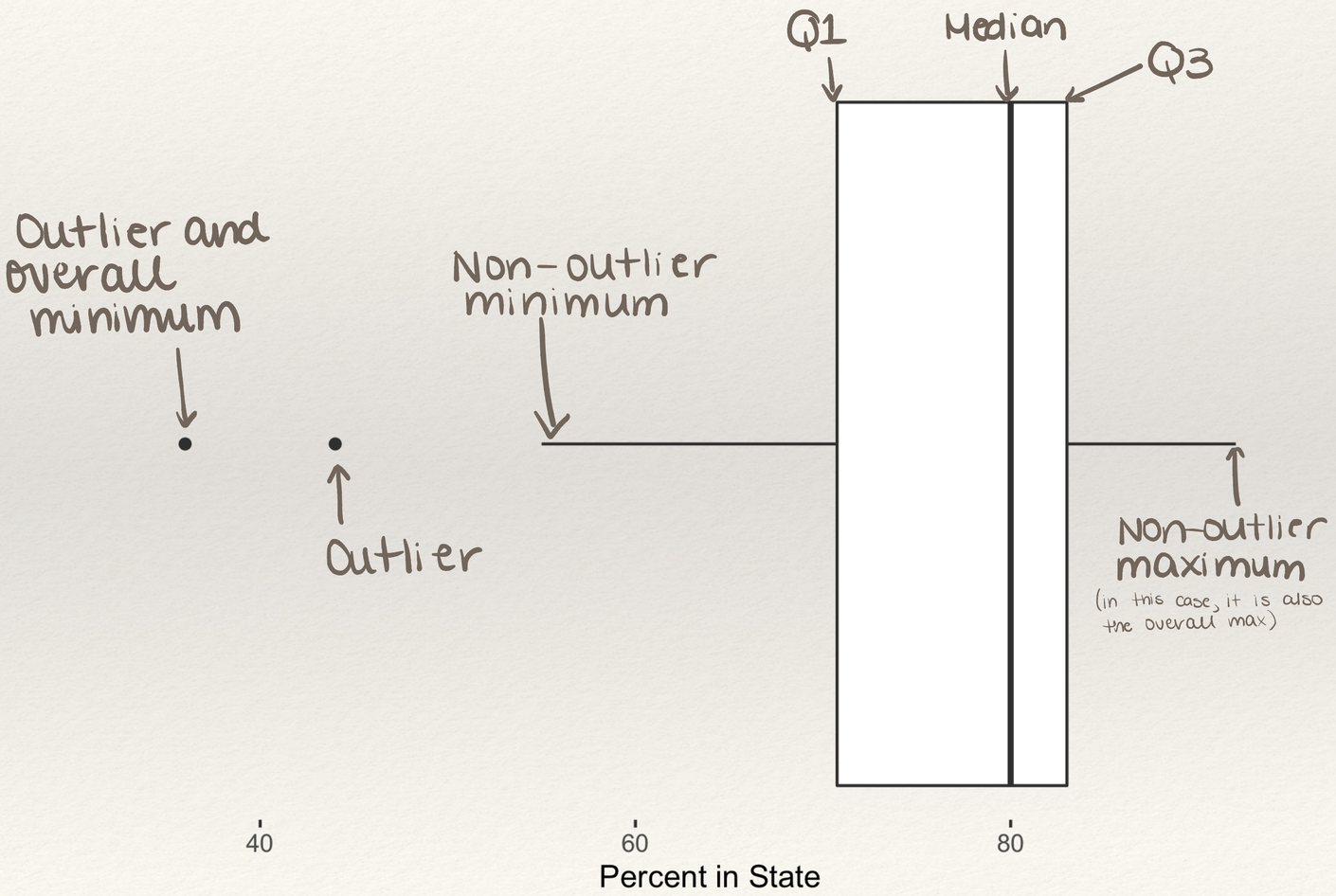
The **Interquartile Range (IQR)** is a measurement for variability that is resistant to extreme observations or outliers.

The **IQR** is the range for the middle 50% of the data.

$$IQR = Q3 - Q1$$

Five Number Summary & Boxplot

The five number summary includes five characteristic values of a quantitative variable:



Determining Outliers

An **outlier** is an extreme observation that falls far from the center of the data.

The 1.5 × IQR rule for outliers determines that any point outside the cutoff is an outlier.

The lower cutoff for outliers is: $Q_1 - 1.5(IQR)$
any values less than this cutoff are considered outliers

The upper cutoff for outliers is: $Q_3 + 1.5(IQR)$
any values more than this cutoff are considered outliers

Outliers Example

state	%	state	%	state	%	state	%
AK	83	IL	83	NC	83	SC	79
AL	75	IN	80	ND	55	SD	68
AR	82	KS	80	NE	81	TN	84
AZ	70	KY	83	NH	55	TX	92
CA	92	LA	86	NJ	92	UT	76
CO	77	MA	63	NM	82	VA	76
CT	69	MD	81	NV	91	VT	36
DE	60	ME	74	NY	81	WA	84
FL	84	MI	91	OH	85	WI	81
GA	83	MN	79	OK	80	WV	59
HI	81	MO	80	OR	76	WY	63
IA	62	MS	85	PA	73		
ID	68	MT	73	RI	44		

Five Number Summary:

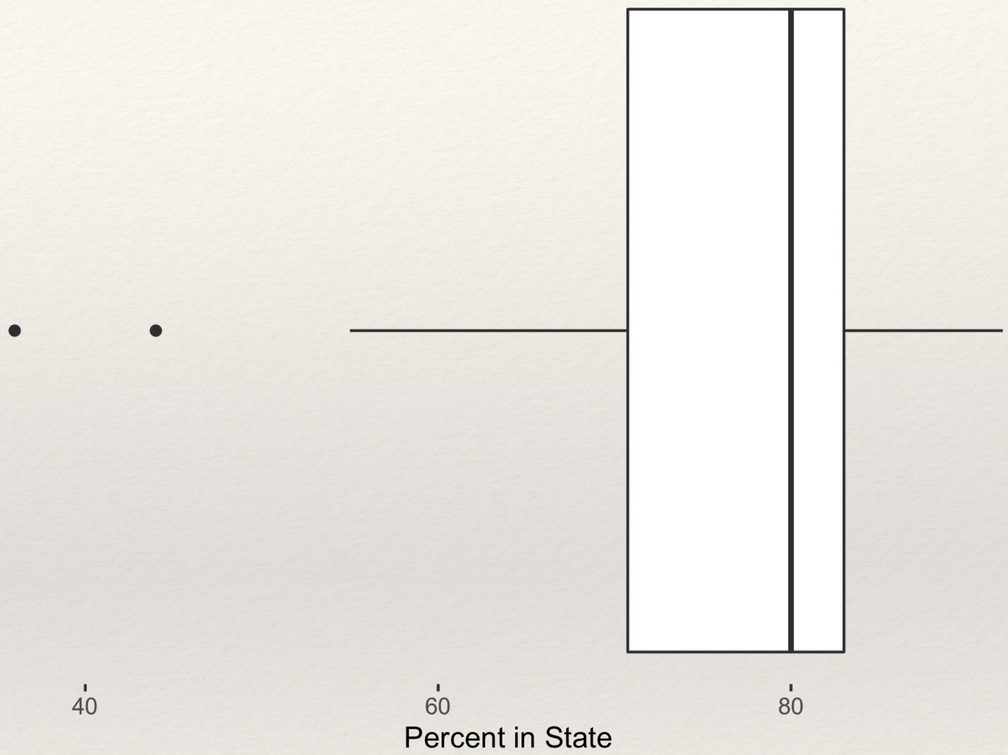
Min = 36

Q1 = 70.75

M = 80

Q3 = 83

Max = 92



What are the lower and upper cutoffs for outliers in this dataset?

IQR = 83 − 70.75 = 12.25

lower = 70.75 − 1.5(12.25) = 52.375

upper = 83 + 1.5(12.25) = 101.375

Categorical Variables: Proportions & Frequency Tables

- Categorical variables are numerically summarized using groups or categories.

- A sample proportion, $\hat{p}$, represents the proportion of individuals in a sample within a category and is a special case of a sample mean.

Example

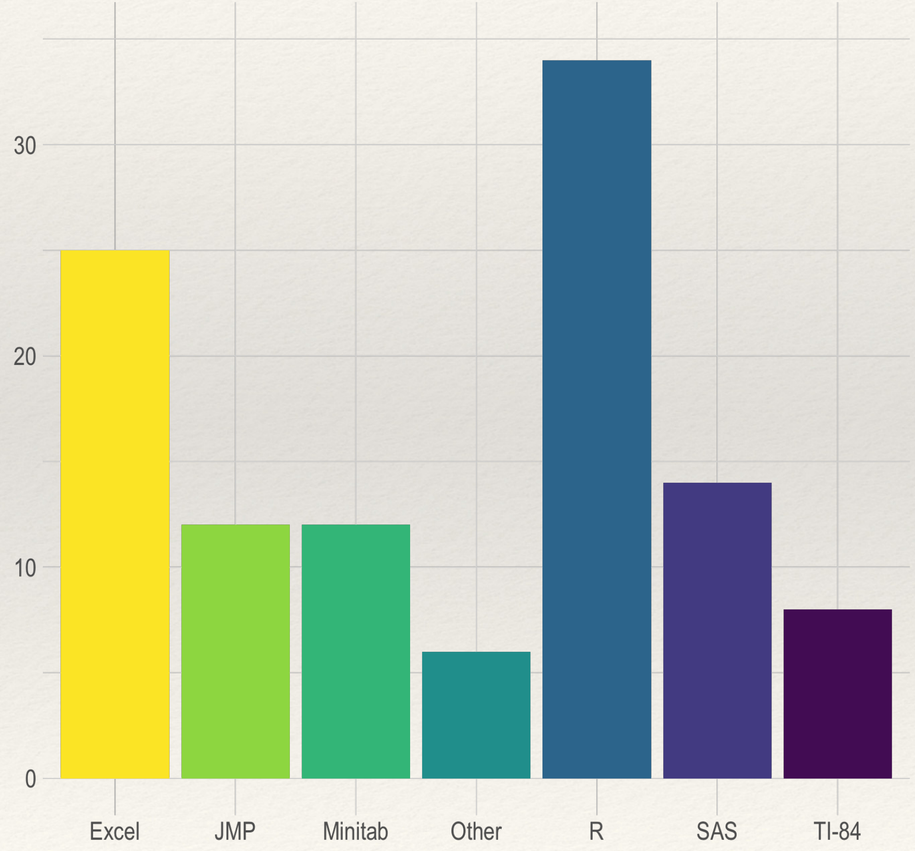
A random sample of 111 college level introductory engineering statistics courses were sampled to see what type of software was used in the course.

The following is a frequency table of the results.

Softwar	Count	Percent	$\hat{p}$
R	34	30.6%	0.306
Excel	25	22.5%	0.225
SAS	14	12.6%	0.126
JMP	12	10.8%	0.108
Minitab	12	10.8%	0.108
TI-84	8	7.2%	0.072
Other	6	5.4%	0.054

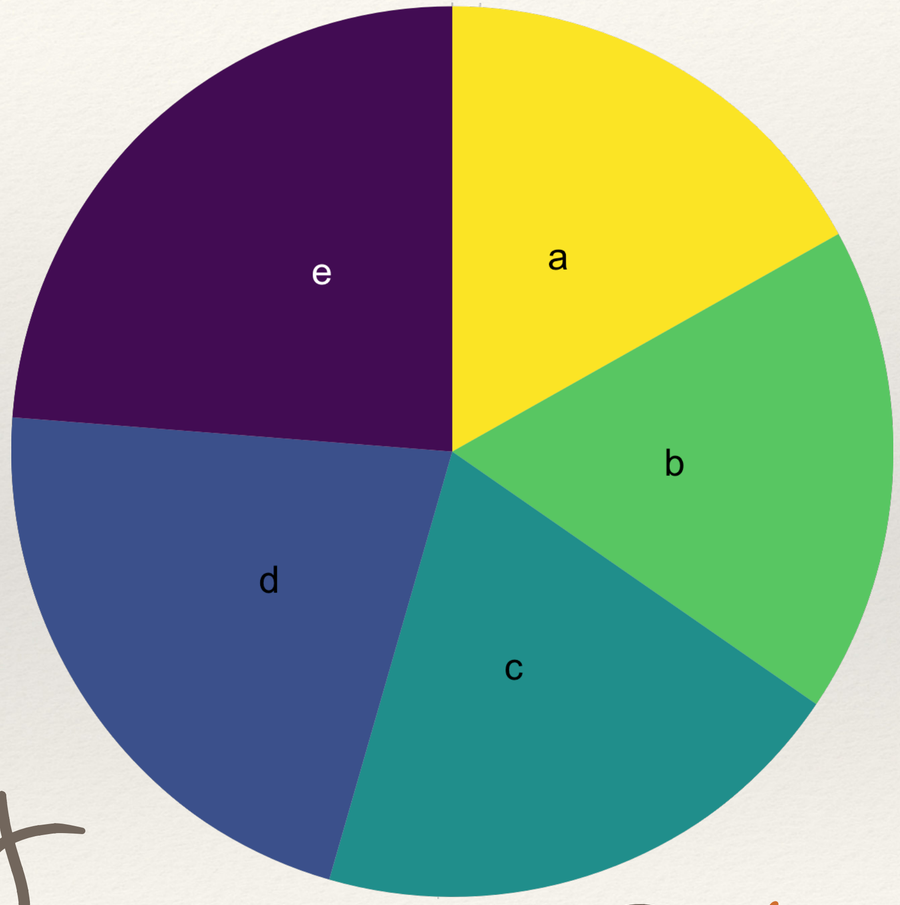
Categorical Variables: Bar Charts

- Height of bars represent count or percent of a category with respect to the entire sample.



Categorical Variables: Pie Charts

Order the groups (a, b, c, d, e) from largest to smallest.



Wedge size represents the proportion of observations per category. Can you easily tell which category is the largest? (No!)

Contingency Tables

A contingency table is a tool used to display data collected for two categorical variables. The example below (ex. 2.2.25 from OpenIntro) contains information on 3921 emails that were classified as *spam* or *not spam*. The format of each email was also recorded: *text* or *HTML*.

What proportion of text emails are classified as spam?

$$\frac{209}{1195}$$

	text	*HTML*	total
spam	209	158	367
not spam	986	2568	3554
total	1195	2726	3921