Introduction to Statistics

Class Notes

Introduction to Statistics

Introduction

The course will be divided into 3 parts

1) Descriptive statistics
2) Probability
3) Statistical inference.

1) Descriptive Statistics

Idea: We have a set of data and we wish to capture its essence or summerize its main features.

The sigma notation

We want a convenient way to write a sum of numbers $x_1, x_2, \ldots, x_n$

We use the sigma sign, $\sum$ to represent such a sum:

$$\sum_{i=m}^{n} x_i = x_m + x_{m+1} + \cdots + x_n$$

Add the numbers x_i, starting with $i = m$ and ending with $i = n$. $(n \geq m)$

$$\sum_{i=5}^{10} x_i = x_5 + x_6 + \cdots + x_{10}$$

end $\to 10$

start $\to 5$

Properties of summation $\sum$

1) $\sum_{i=1}^{n} c x_i = c \sum_{i=1}^{n} x_i$ if c is a constant,

2) $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

3) $\left(\sum_{i=1} x_i\right)^2 \neq \sum_{i=1}^{n} x_i^2$

$$(x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2 = x_1^2 + x_2^2$$

There are three ways of measuring the center

1) Sample mean $\overline{x}$

2) Sample median m

3) mode.

Sample mean.

Given a set of quantities

$$x_1, x_2 \dots x_n,$$

We call

$$\overline{x} = \sum_{i=1}^{n} x_i / n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

the sample mean of the x_i's

Note:

1) $\bar{x}$ "x bar" represents the "center" of a set of numbers in some sense.

2) Note that $\bar{x}$ may not represent any number in the dataset.

$x_1 = 1$. $x_2 = 3$.

$\bar{x} = (1+3)/2 = 2$ is not equal to 1 or 3!

3) $\bar{x}$ is easily influenced by extreme observations (either very large or small)

Ex $x_1 = 1$, $x_2 = 1$, $x_3 = 1$, $x_4 = 1$, $x_5 = 100$

(Income) $\bar{x} = \dfrac{1+1+1+1+100}{5} = 20.8$

does not provide a picture of typical income. values pull towards 100.

Sample median

Let $x_1, x_2, \ldots, x_n$ be a set of values.

The sample median m is defined as follows.

1) If the number of observations n is odd.

- rank the observations from smallest to largest. $x_1, x_2, \ldots, x_n \Rightarrow x_1^*, x_2^*, \ldots, x_n^*$

- then the median m is the middle number after ranking

$$m = x^*_{\frac{n+1}{2}}$$

Ex

Data: $x_1 = 1$, $x_2 = 5$ $x_3 = 4$ $x_4 = 2$ $x_5 = 3$

After ranking: $x_1^* = 1$ $x_2^* = 2$ $x_3^* = 3$ $x_4^* = 4$ $x_5^* = 5$

$$n = 5 \qquad m = x^*_{\frac{n+1}{2}} = x^*_3 = 3$$

i.e. the sample median is the observation with the property that half of the observations are $\leq$ it and half $\geq$ it.

2) When the number of observations n is even

- rank the observations as before.

- identify two middle observations

the sample median m is the average of these two observations.

$$m = \frac{x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}}{2}$$

Ex

Data:

$x_1 = 5 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 3 \quad x_5 = 1 \quad x_6 = 6$

After ranking:

$x_1^* = 1 \quad x_2^* = 2 \quad x_3^* = 3 \quad x_4^* = 4 \quad x_5^* = 5 \quad x_6^* = 6$

$n = 6.$ $$m = \frac{x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}}{2} = \frac{x_3^* + x_4^*}{2}$$

$$= \frac{3+4}{2} = 3$$

Note :

1) Sample median also represents a measure of centrality – different possibly from $\bar{x}$

2) It is not easily affected by extreme observations as $\bar{x}$. We say that the sample median is more robust than the sample mean.

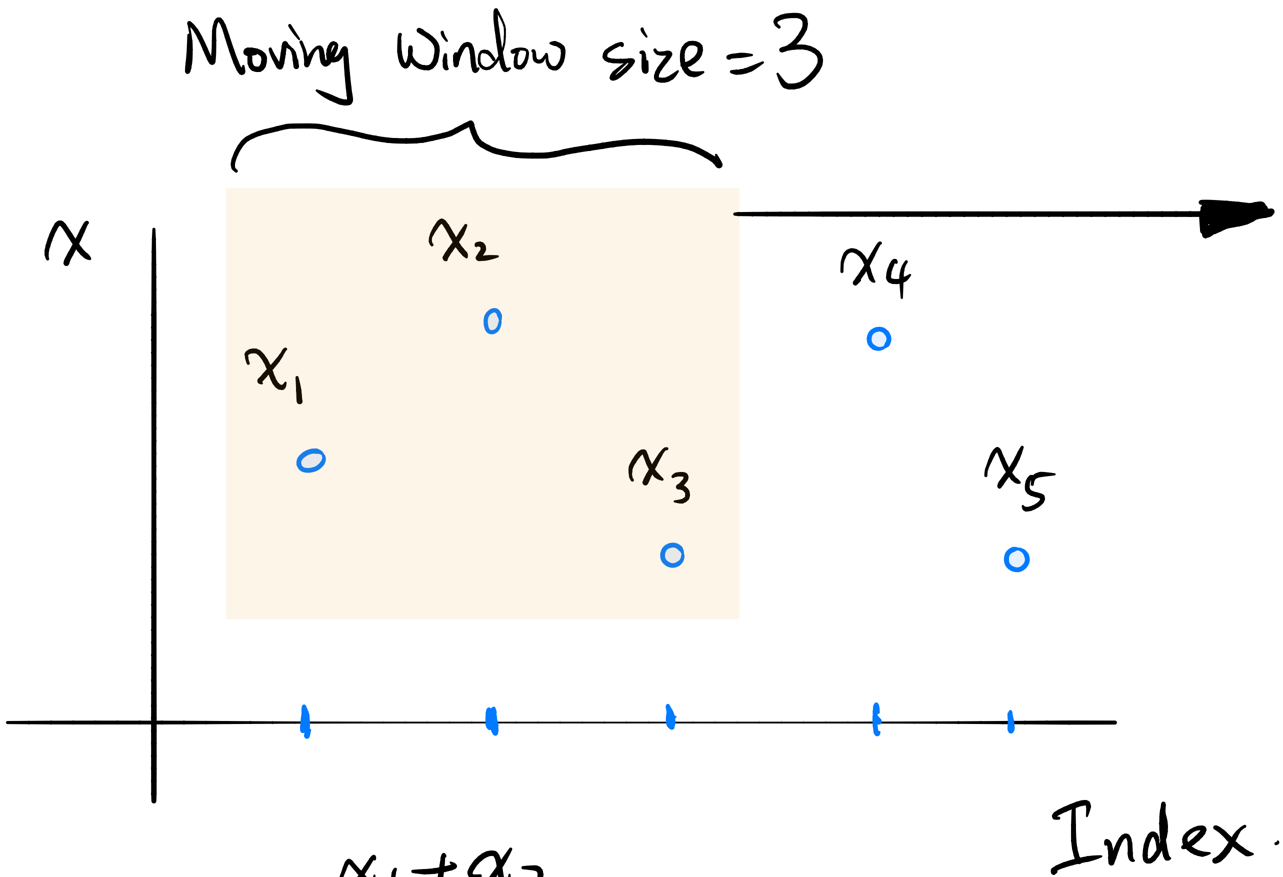
Thus, if $x_1 = 1$, $x_2 = 2$, $x_3 = 1000$

$$\bar{x} = \frac{1+2+1000}{3} \approx 334$$

$m = 2$.

Applications of Sample mean & median

How to compute Moving Average (MA) with a window size 3 ?



$$a_1 = \frac{x_1 + x_2}{2}$$

$$a_2 = \frac{x_1 + x_2 + x_3}{3}$$

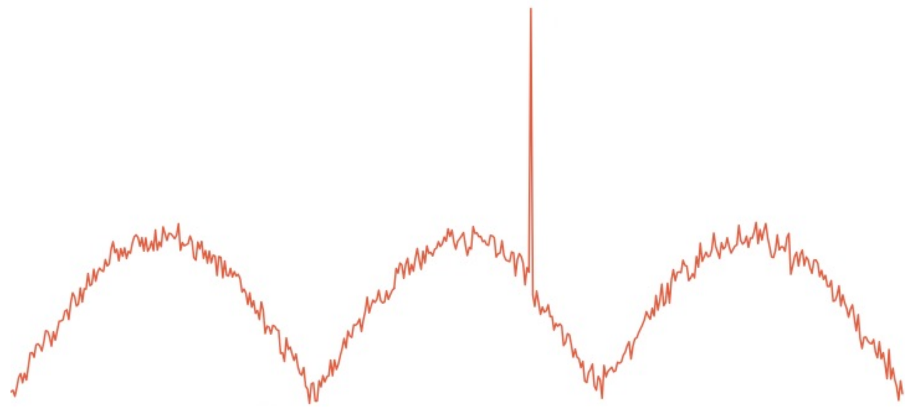
$$a_3 = \frac{x_2 + x_3 + x_4}{3}$$

$$a_4 = \frac{x_3 + x_4 + x_5}{3}$$

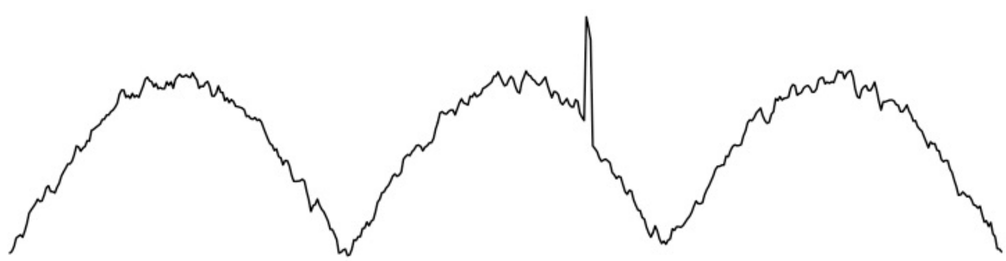
$$a_5 = \frac{x_4 + x_5}{2}$$

Moving Median can be computed using a similar way

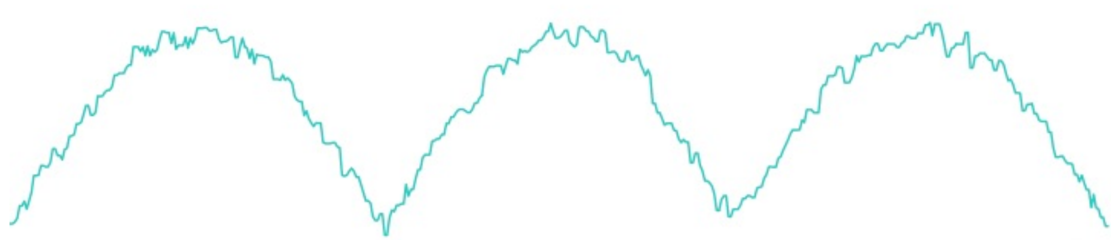
Original Data:



Moving Average:

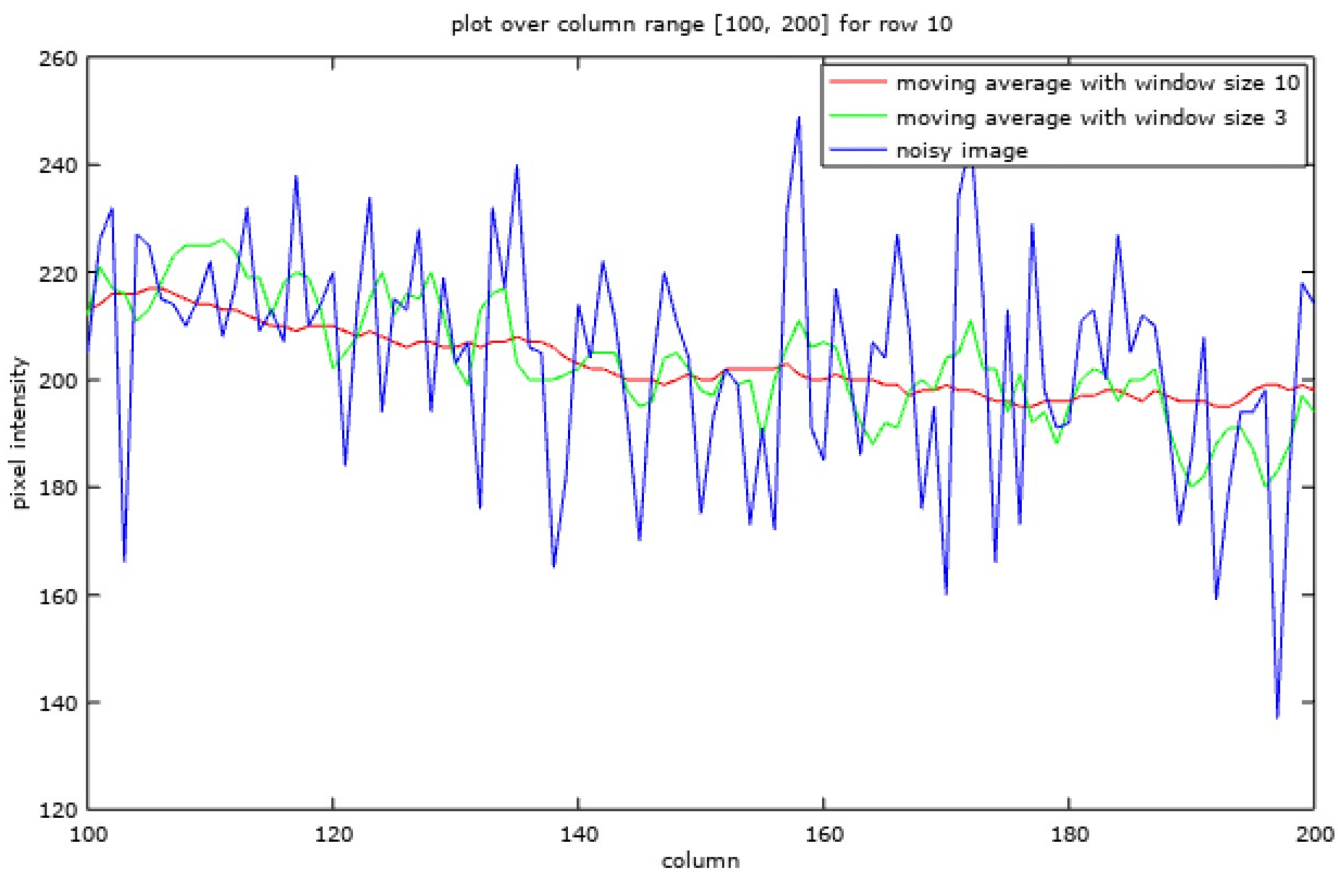


Moving Median:



Moving Median is less affected by extreme values than Moving Average is

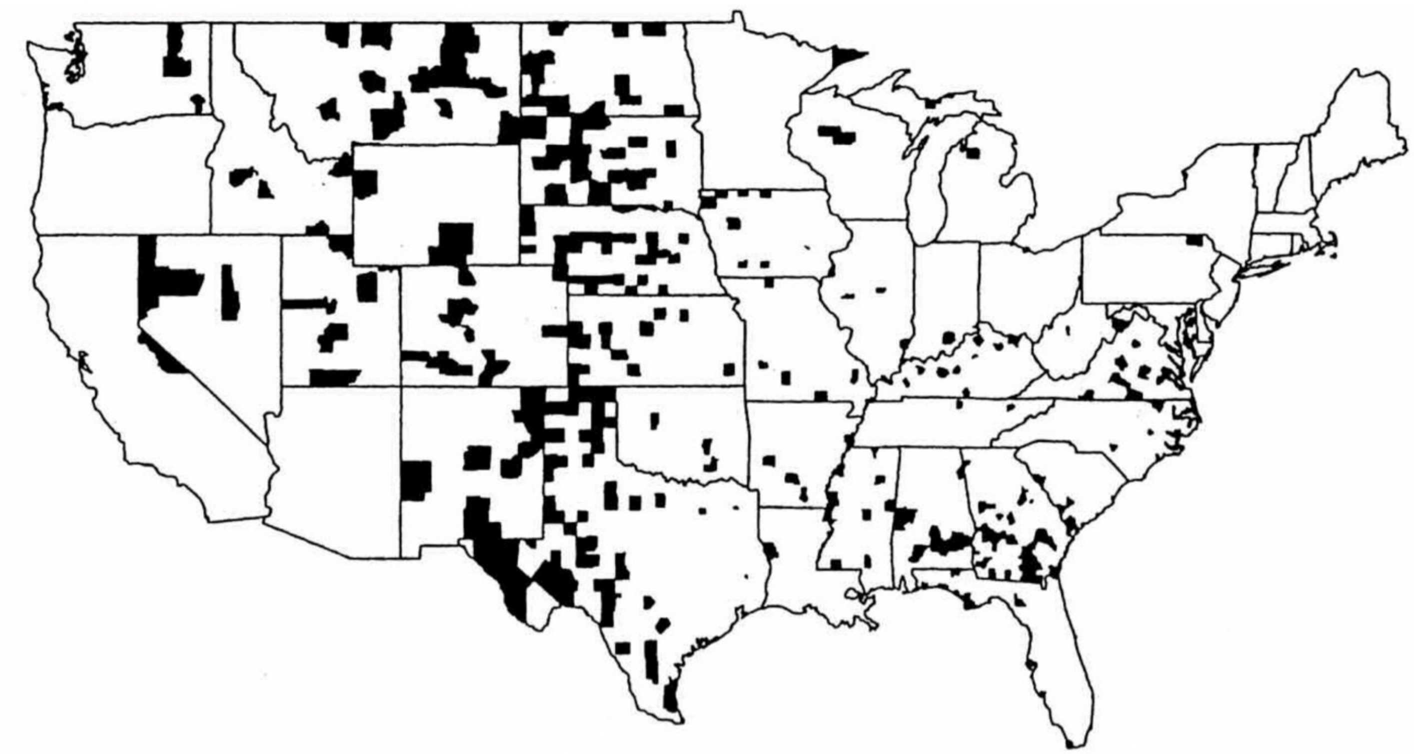
How about the window size ?



The larger the window size, the smoother the curve

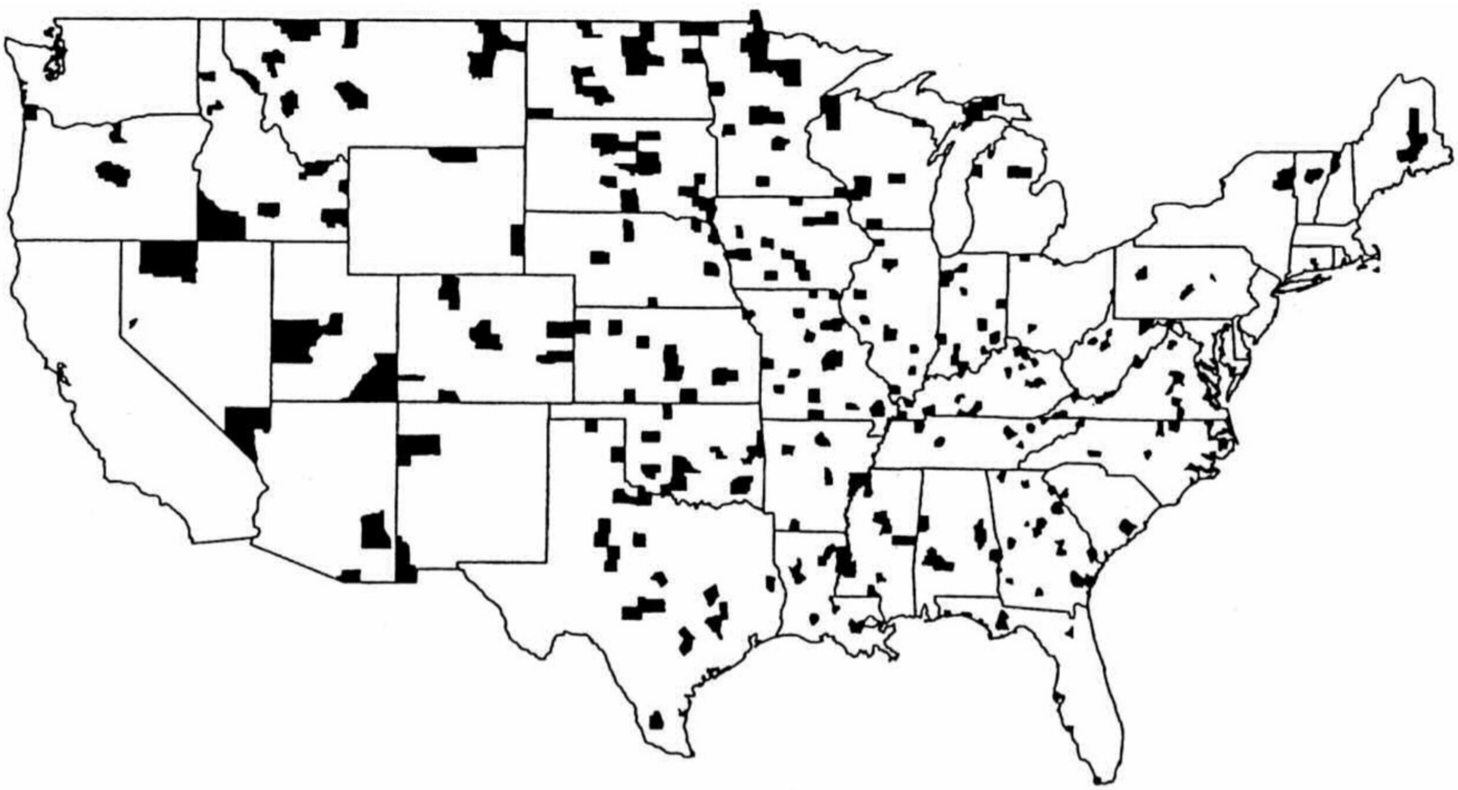
In general, the large the sample size, the more stable the mean (and median)

FIGURE 1.



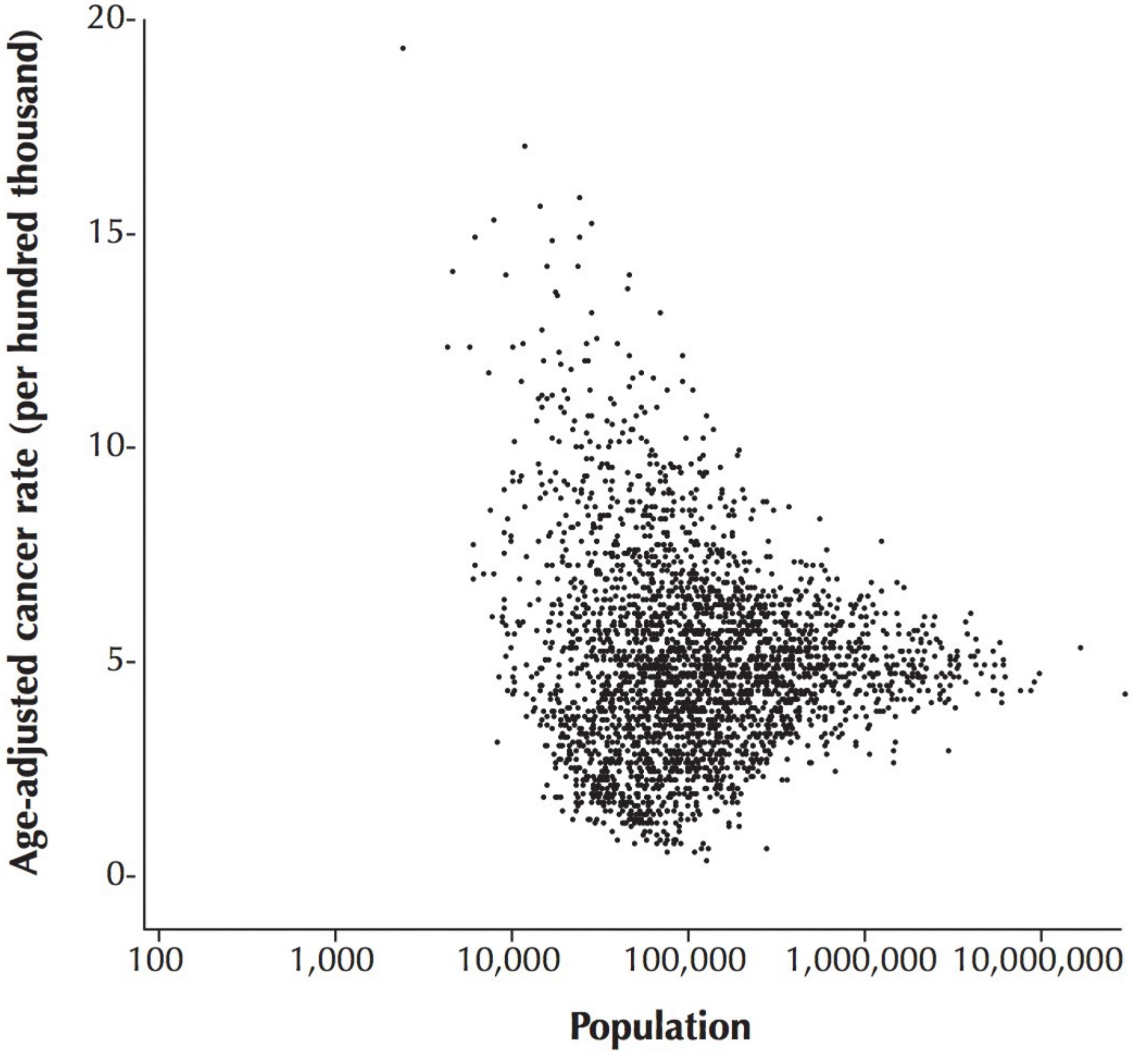
The counties with the lowest 10% age-standardized death rates for cancer of the kidney/ureter for U.S. males, 1980-89. Reprinted, by permission, from Andrew Gelman and Deborah Nolan, *Teaching Statistics: A Bag of Tricks* (New York: Oxford University Press, 2002), p. 15.

FIGURE 2.



The counties with the highest 10% age-standardized death rates for cancer of the kidney/ureter for U.S. males, 1980-89. Reprinted, by permission, from Andrew Gelman and Deborah Nolan, *Teaching Statistics: A Bag of Tricks* (New York: Oxford University Press, 2002), p. 14.

FIGURE 3.



Age-adjusted death rates for cancer of the kidney/ureter for U.S. males, for all U.S. counties, 1980-89, shown as a function of the log of the county population.

Mode

A much less common measure centrality

Defined as the most commonly occurring observation

(obs. that occurs most often)

e.g. $x_1 = -1.2$. $x_2 = 3.6$ $x_3 = 3.7$

$x_4 = -1.2$

The mode is -1.2

If I add $x_5 = 3.6$, then mode becomes

-1.2, 3.6 (two modes)

In addition to a measure of the center, we need to measure variability of the data.

1) range

2) sample variance

3) sample standard deviation

Range

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Sample Variance

EX (Asset Return)

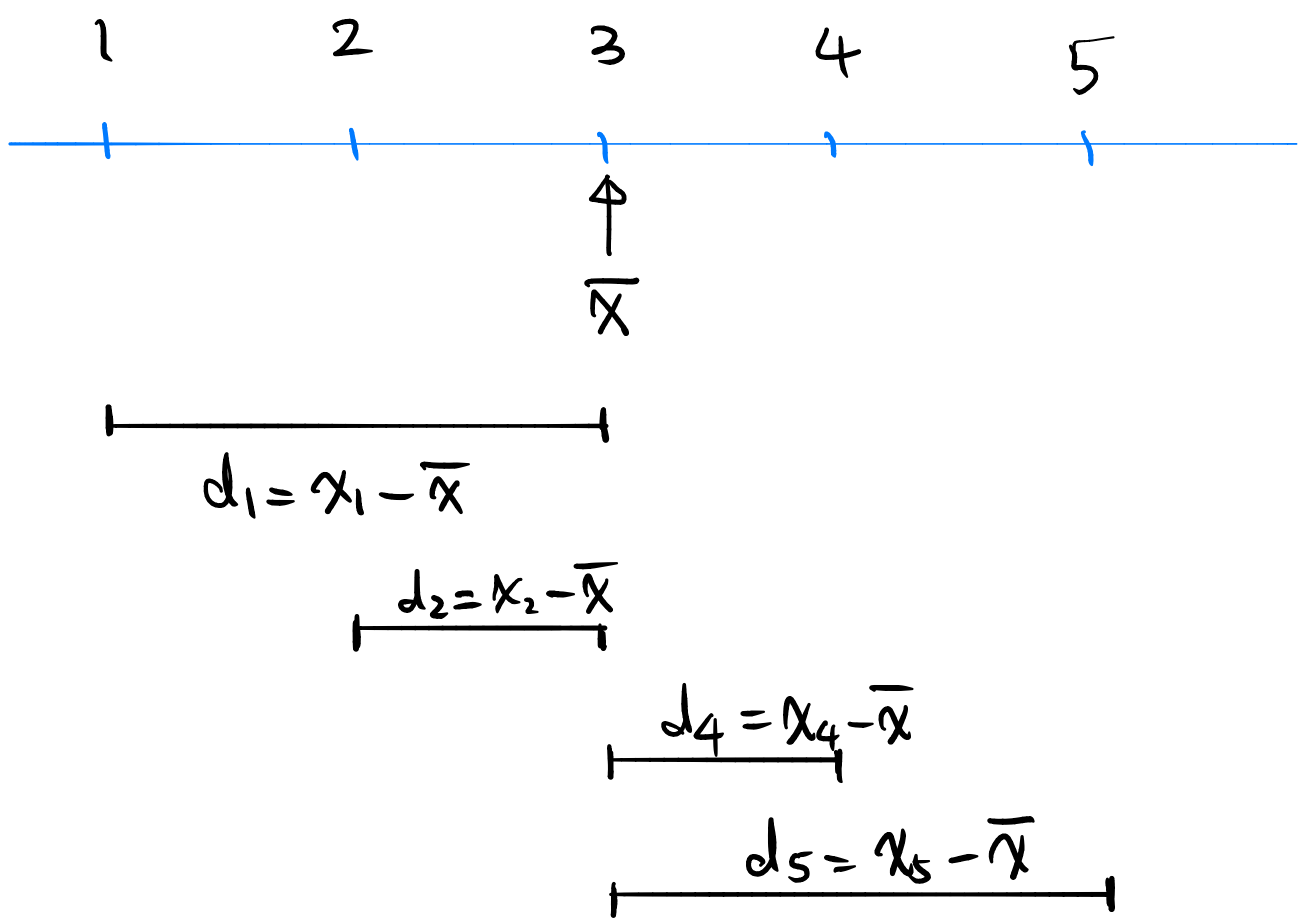
Data 1	Data 2
1, 2. 3. 4. 5	2, 3, 3. 3. 4
$\bar{x}_1 = 3$	$\bar{x}_2 = 3$

Same averaged return. which one has more variation?

Definition:

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

obs. sample mean



Variance of Data 1

$$s_1^2 = \frac{1}{5-1}\left(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2\right)$$

$$= \frac{1}{5-1}\left((1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2\right)$$

$$= \frac{10}{4} = 2.5$$

Variance of Data 2

$$s_2^2 = \frac{1}{5-1}\left(d_1^2 + \cdots + d_5^2\right)$$

$$= \frac{1}{5-1}\left((2-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2 + (4-3)^2\right)$$

$$= \frac{2}{4} = 0.5$$

Note:

1) s^2 is "roughly" ($n-1$ instead of n) the averaged squared divation of x_i's from $\bar{x}$.

2) Reason of using $n-1$ will be explained later

3) Alternative formula for computing s^2.

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\cdot(\bar{x})^2\right]$$

proof: as homework question.

Sketch:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \underbrace{\sum_{i=1}^{n} 2x_i\bar{x}}_{-2\bar{x}\sum_{i=1}^{n} x_i \,=\, -2n\bar{x}^2} + \underbrace{\sum_{i=1}^{n}\bar{x}^2}_{n\bar{x}^2}\right]$$

4) Alternatively, use $|x_i - \bar{x}|$ instead of $(x_i - \bar{x})^2$ to eliminate minus signs with deviations. Hence, it results in

$$L(x) = \frac{1}{n-1} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Sample Standard deviation

Definition :

Standard deviation : $s = \sqrt{s^2}$

variance

Note that the unit of s^2 are the same as the units of x_i^2. Interpretation of s^2 is more difficult. Hence we consider $s = \sqrt{s^2}$ as our measure of spread. S has the same unit as x_i's.

Numerical Example

$x_1 = 4$. $x_2 = 1$ $x_3 = 3$ $x_4 = 1$ $x_5 = 3$

$x_6 = 1$ $x_7 = 2$ $x_8 = 2$.

Solution:

$$\bar{x} = \frac{1}{8} \cdot \sum_{i=1}^{8} x_i = \frac{4+1+\cdots+2}{8}$$

$$= \frac{17}{8} = 2.125$$

x_2	x_4	x_6	x_7	x_8	x_3	x_5	x_1
1	1	1	2	2	3	3	4

$m = 2$. mode $= 1$

$$s^2 = \frac{1}{8-1}\left[\sum_{i=1}^{8} x_i^2 - 8\bar{x}^2\right]$$

$$= \frac{1}{8-1}\left[\sum_{i=1}^{8} x_i^2 - 8 \cdot (2.125)^2\right]$$

$= 1.26$

$$\sum_{i=1}^{8} x_i^2 = 4^2 + 1^2 + \cdots + 2^2 = 45$$

$$s = \sqrt{s^2} = \sqrt{1.26} = 1.126.$$

range $= 4 - 1 = 3$

Types of Data

1) Quantitative.

Definition : measurements recorded on numerical scale.

Ex: Tempreture, unemployment rate, interest rate, exam scores.

2) Qualitative.

Definition : measurements that cannot be measured on a numerical scale. only be classified into categories.

Ex: species, Car type, gender

Note: Qualitative data are often coded in arbitrary numerical values, but can not be meaningfully "+" "−" "×" "÷".

Car Type	code
SUV	1
sedan	2
Truck	3
...	

Describing Qualitative Data

Original Data:

Subject	Education
1	BS
2	Master
3	PhD
4	PhD
5	BS

Education is Qualitative, can be classified into 3 classes, BS. Master. PhD

Such data can be summarized in three ways

1) Class Frequency

the number of obs that falls in a particular class

2) Class relative frequency

$$\text{class relative freq.} = \frac{\text{class freq.}}{n}$$

n is the total number of obs.

3) Class percentage.

$$\text{class percentage} = (\text{class relative freq.}) \times 100$$

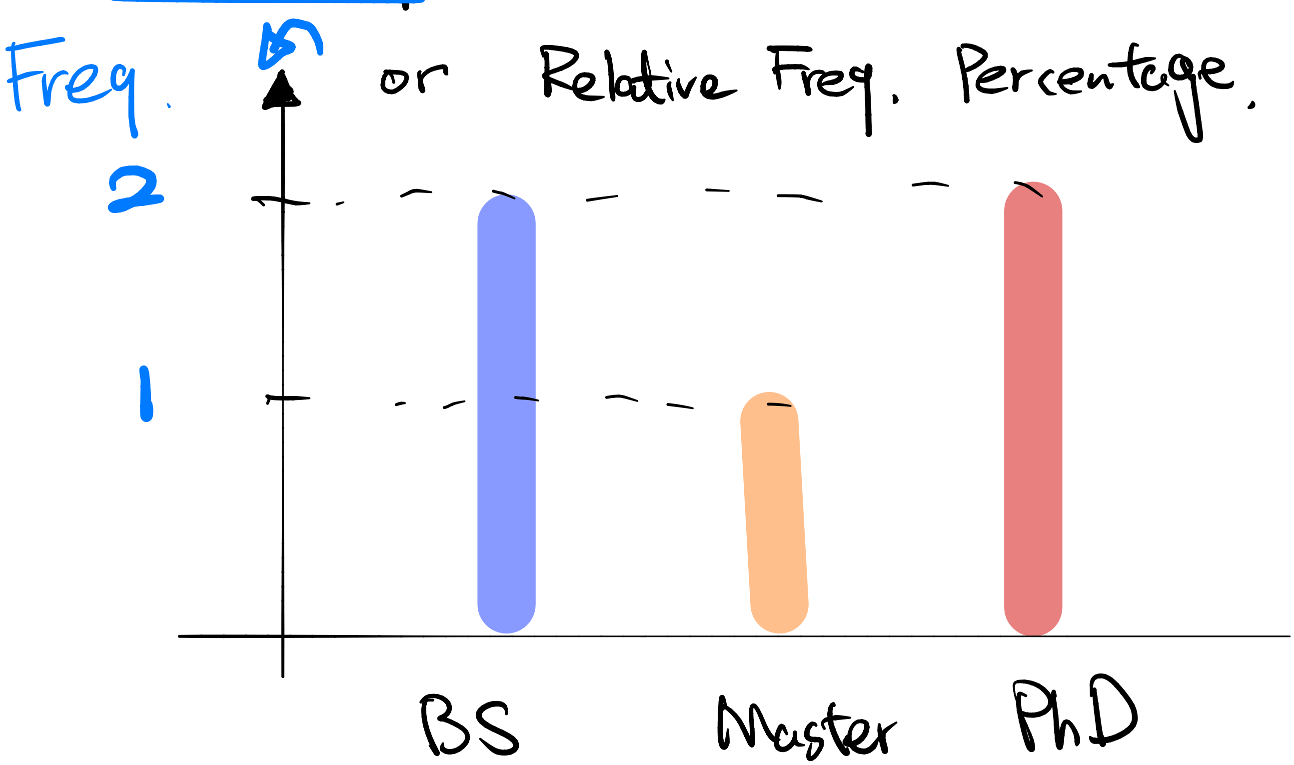
Table:

$n = 5$

	Freq	Relative Freq.	Percentage
BS	2	2/5	40%
Master	1	1/5	20%
PhD	2	2/5	40%
Total	5	1	100%

Graphs

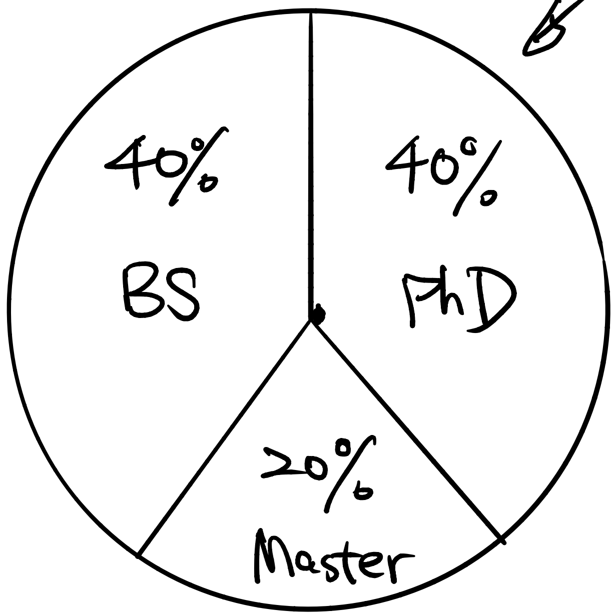
Bar Graph



Pie Chart

relative Freq.

percentage.



In January 1971 the Gallup poll asked: "A proposal has been made in Congress to require the U.S. government to bring home all U.S. troops before the end of this year. Would you like to have your congressman vote for or against this proposal?"
Guess the results, for respondents in each education category, and fill out this table (the two numbers in each column should add up to 100%):

	Adults with:			
	Grade school education	High school education	College education	Total adults
% for withdrawal of U.S. troops (doves)				73%
% against withdrawal of U.S. troops (hawks)				27%
Total	100%	100%	100%	100%

p_1	p_2	p_3
$1-p_1$	$1-p_2$	$1-p_3$

the proportions of adults with grade school, high school, and college education are λ_1, λ_2, λ_3 respectively. They must satisfy

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

Total proportion of adults against war is 0.73

Therefore.

$$\lambda_1 \times p_1 + \lambda_2 \times p_2 + \lambda_3 \times p_3 = 0.73$$

Assume

$$\lambda_1 = \lambda_2 = \lambda_3 = 1/3$$

Then p_1, p_2 and p_3 must satisfy.

$$\boxed{(p_1 + p_2 + p_3)/3 = 0.73}$$

In January 1971 the Gallup poll asked: "A proposal has been made in Congress to require the U.S. government to bring home all U.S. troops before the end of this year. Would you like to have your congressman vote for or against this proposal?"
Guess the results, for respondents in each education category, and fill out this table (the two numbers in each column should add up to 100%):

	Adults with: Grade school education	High school education	College education	Total adults
% for withdrawal of U.S. troops (doves)	80%	75%	60%	73%
% against withdrawal of U.S. troops (hawks)	20%	25%	40%	27%
Total	100%	100%	100%	100%

- "Back in Vietnam days, the anti-war movement spread from the intelligentsia into the rest of the population, eventually paralyzing the country's will to fight." -*The Economist* (2000)

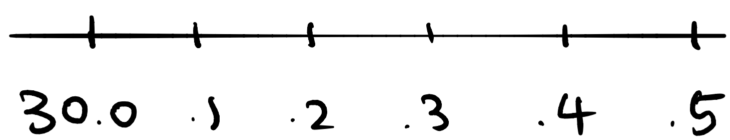
Describing Quantitative Data

- Dot plot
- Histogram.

Dot Plot

30.3 rounding 30.5 "to nearest half"

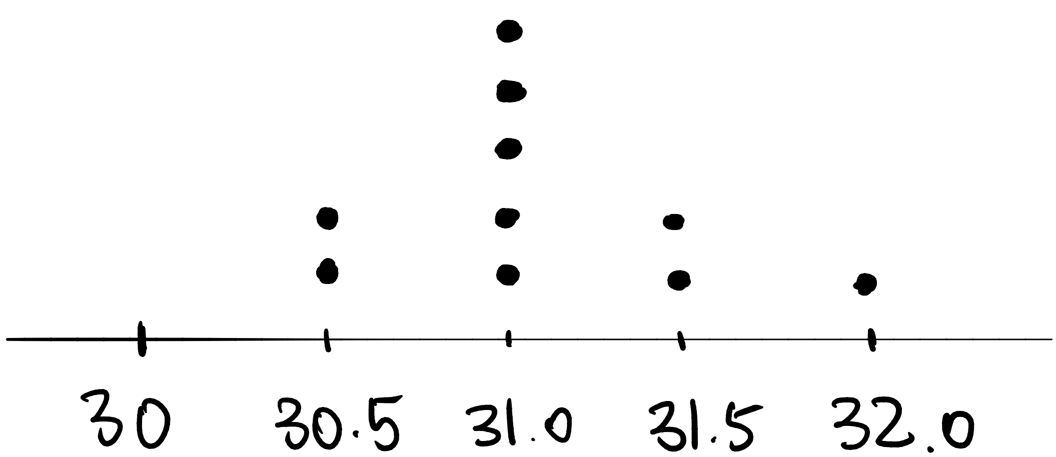
Ex



Data:

30.5 ← 30.3 30.6 30.9 30.8 31.1

31.2 31.0 31.7 31.4 32.1



Histogram (Discretization of quantitative data)

- Divide the range of the data into K equal intervals or bins.
- Allocate data into bins
- Count the frequency of observations in bins.

Ex

Data: 30.1 . 30.2 . 30.5 . 30.6

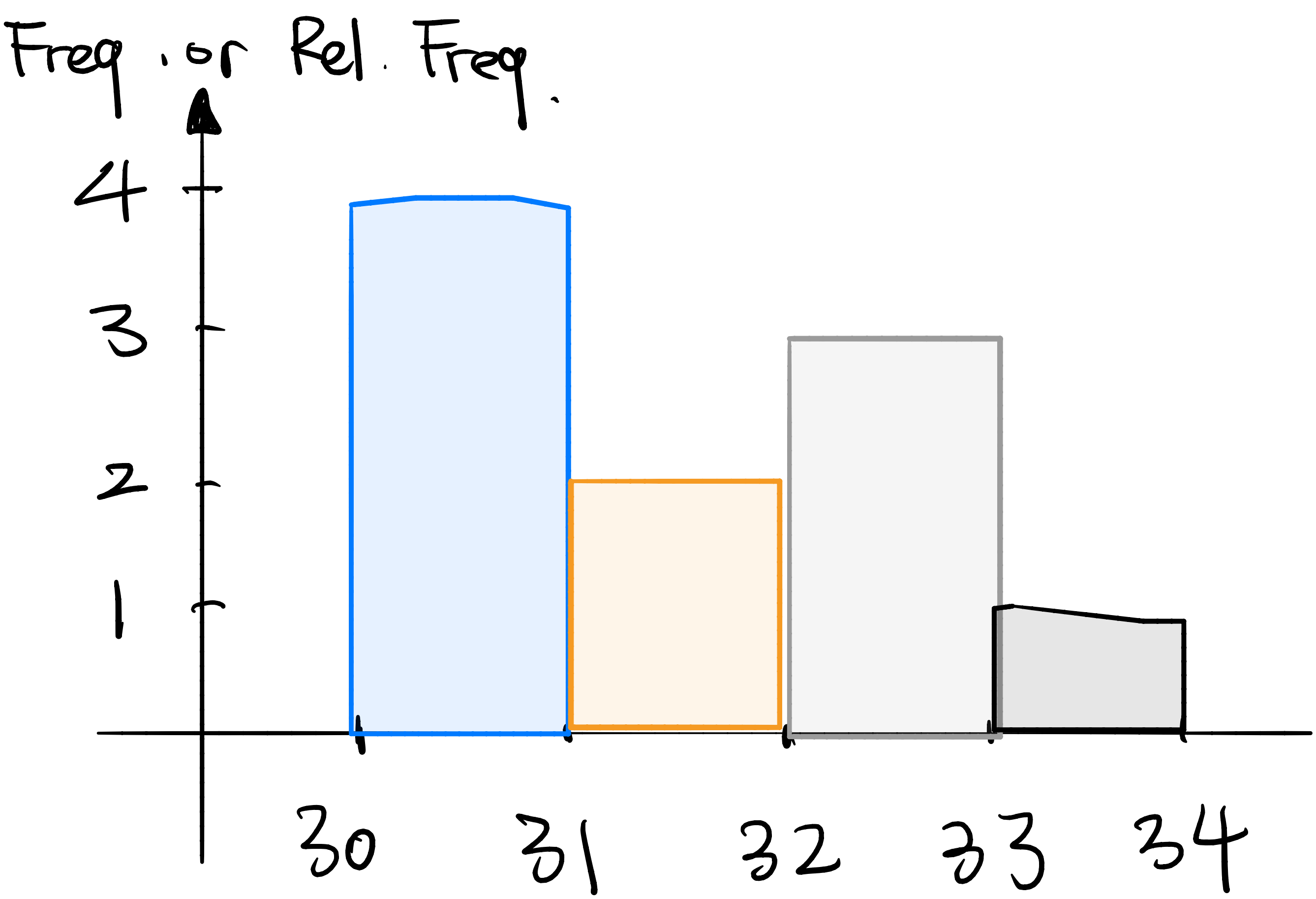
31.5 31.3

32.3 32.7 32.9

33.1

$n = 10$ obs.

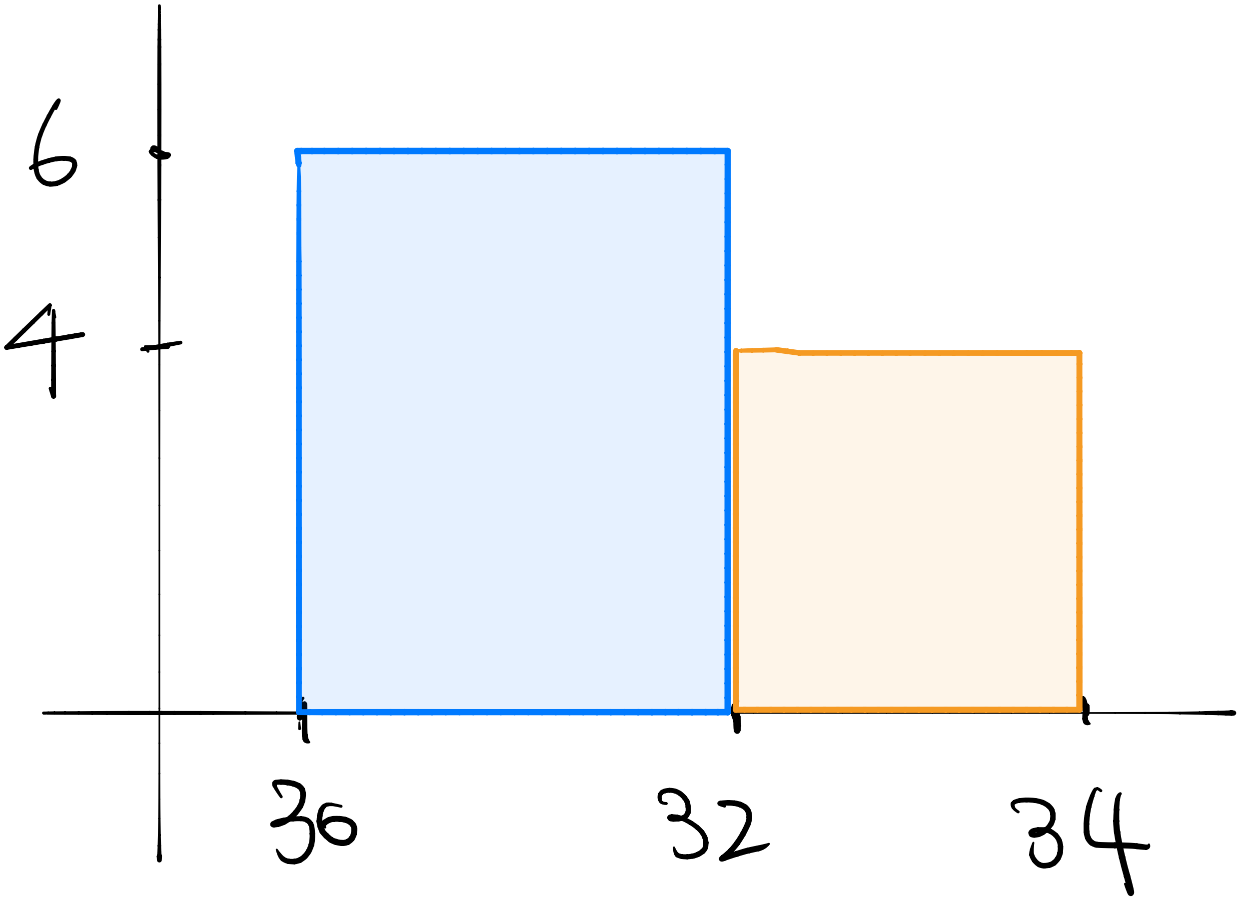
Class Interval	Freq	Rel. Freq
[30, 31]	4	4/10
[31, 32]	2	2/10
[32, 33]	3	3/10
[33, 34]	1	1/10
	10	1.



If you use the relative frequency then the total area represented by the histogram is 1.

If the size of the interval is 2 instead of 1.

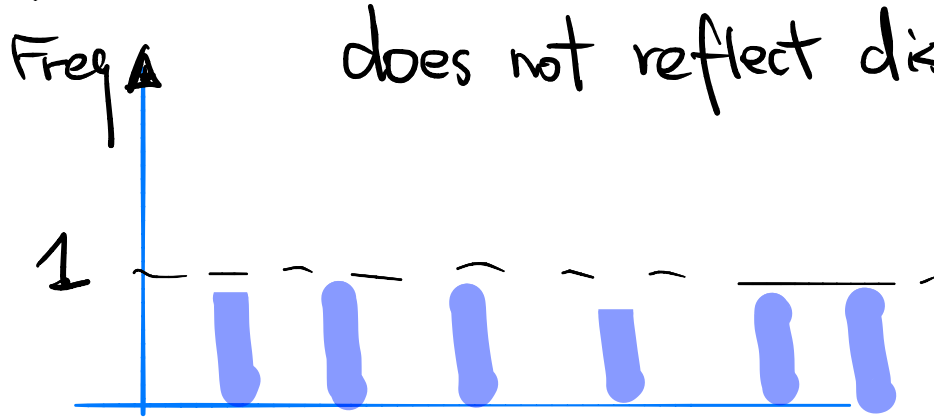
Interval	Freq	Rel. Freq.
[30, 32]	6	6/10
[32, 34]	4	4/10
	10	1



Note:

1) The interval size is reasonably important.

If size is too small, too detailed, does not reflect distribution



If size is too large, the shape will be lost and it conveys little information.

2) Rule for selecting the interval size.

# obs	# of intervals
<25	$5-6$
$25-50$	$7-14$
>50	$15-20$

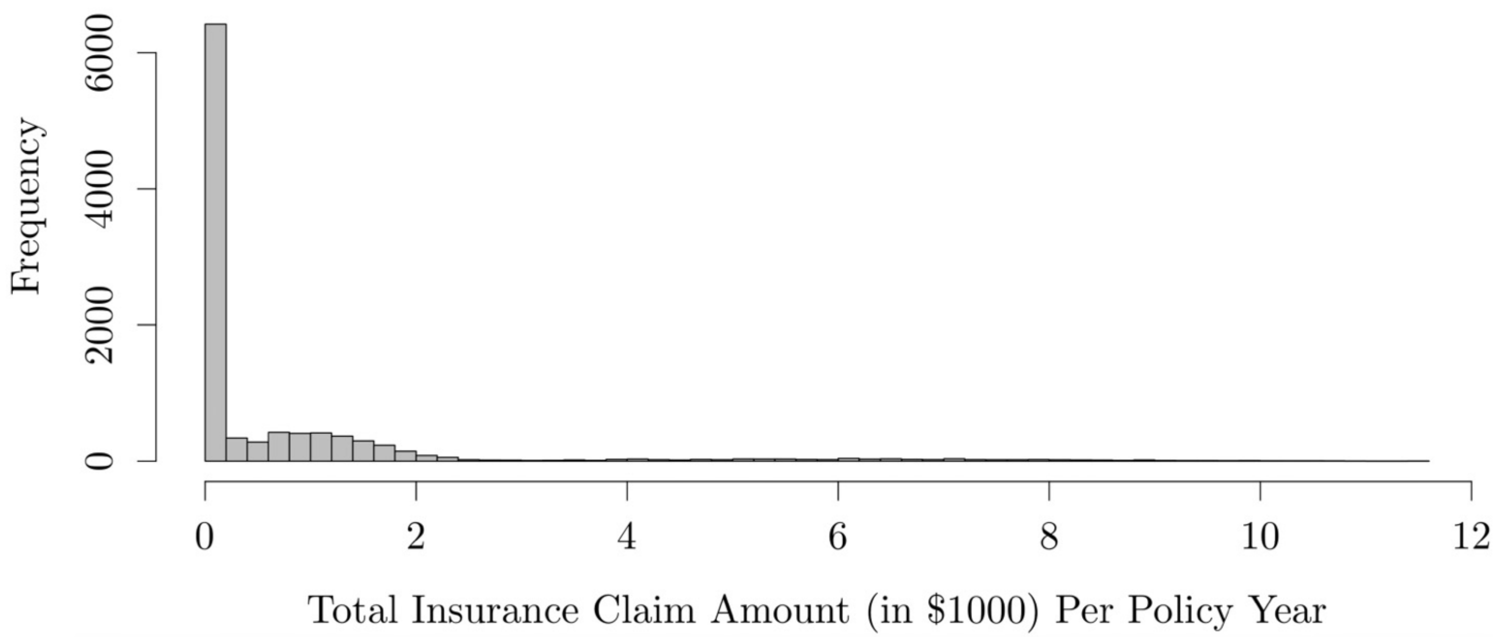
3) The scale is important.

When comparing hists. make sure the scales are the same.

Need a separate interval for the value "zero"

Zero - inflated

right - skewed



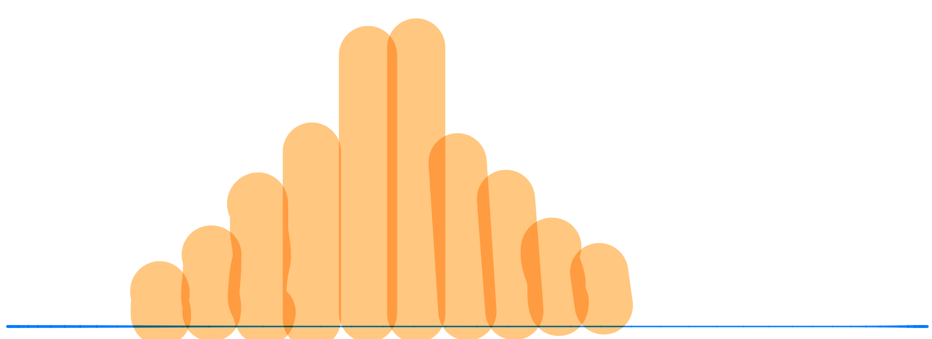
If the insurance company charge everyone at the price of $\overline{X}$, can they cover the basic cost and won't lose money in a long run?

4) You can get an idea of the shape of the dist. of the data.

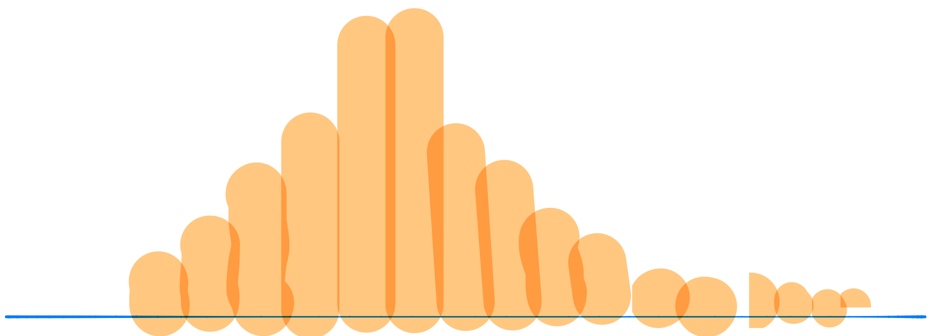
Skewness

A dataset is said to be skewed if one tail of the distribution has more extreme observations than the other tail.

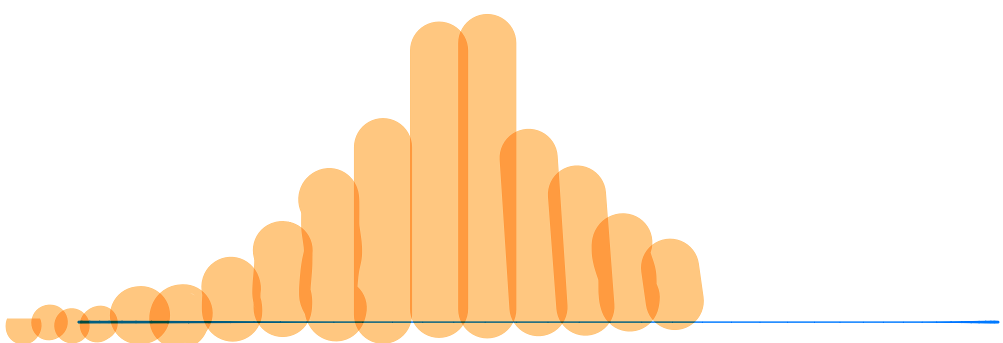
Symmetric



right skewed



left skewed



Detecting Skewness

- Check the shape of the histogram.
- or, compare the $\bar{x}$ and M.

  mean median

Rule :

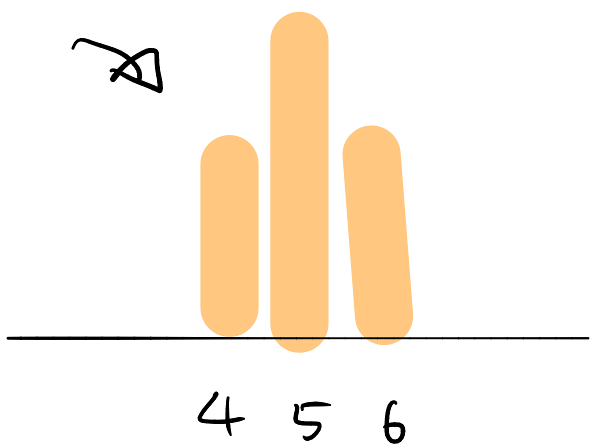
$$\bar{x} > M \Rightarrow \text{right skewed}$$

$$\bar{x} < M \Rightarrow \text{left skewed.}$$

Ex : symmetric

4, 4, 4, 5, 5, 5, 5, 6, 6, 6

symmetric



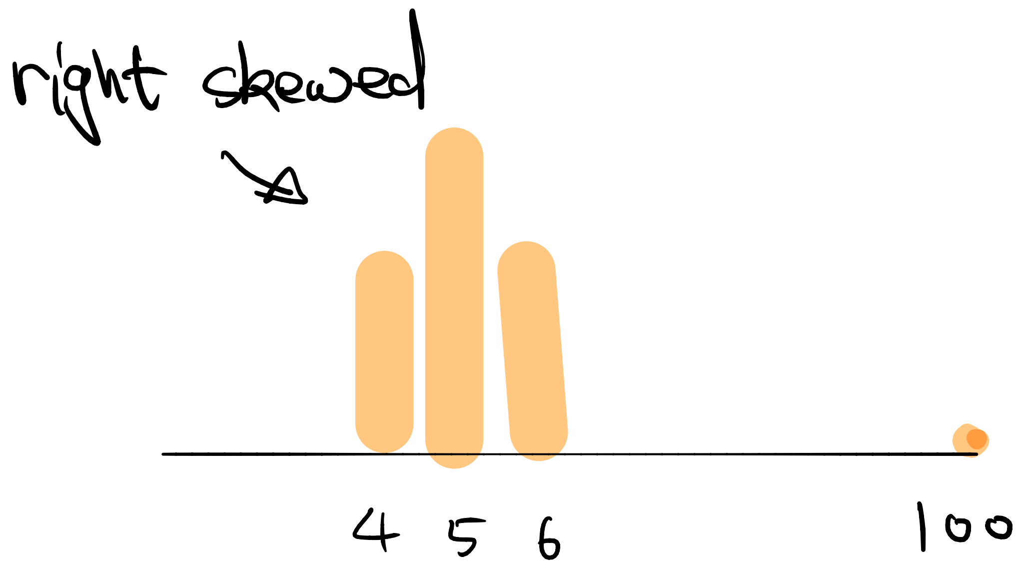
$$\bar{x} = 5$$

$$M = 5$$

$$\bar{x} = M$$

Ex: right skewed.

4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 100



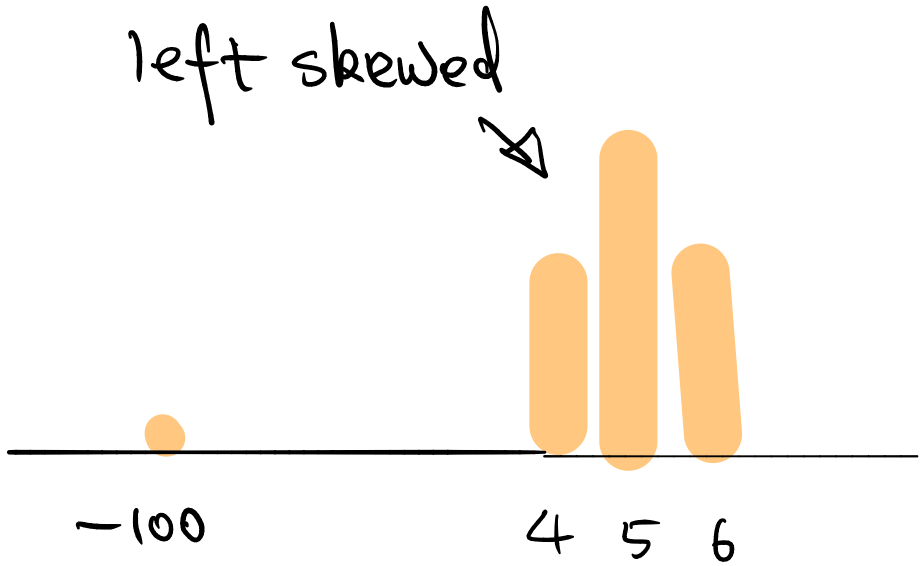
$$\bar{x} = 13.6$$

$$M = 5$$

$$\bar{x} > M$$

Ex: left skewed.

−100, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6

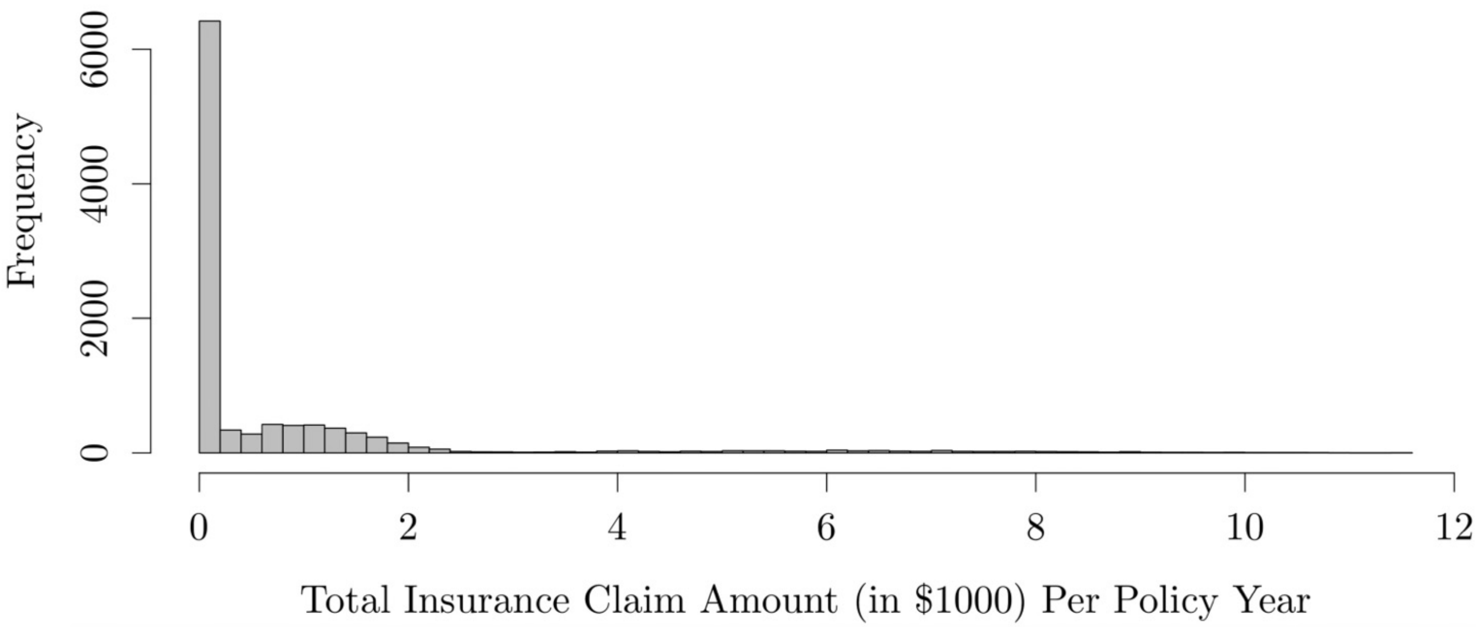


$$\bar{x} = -4.5$$

$$M = 5$$

$$\bar{x} < M.$$

Is this data skewed?



$M = 0$

$\bar{x} > 0$

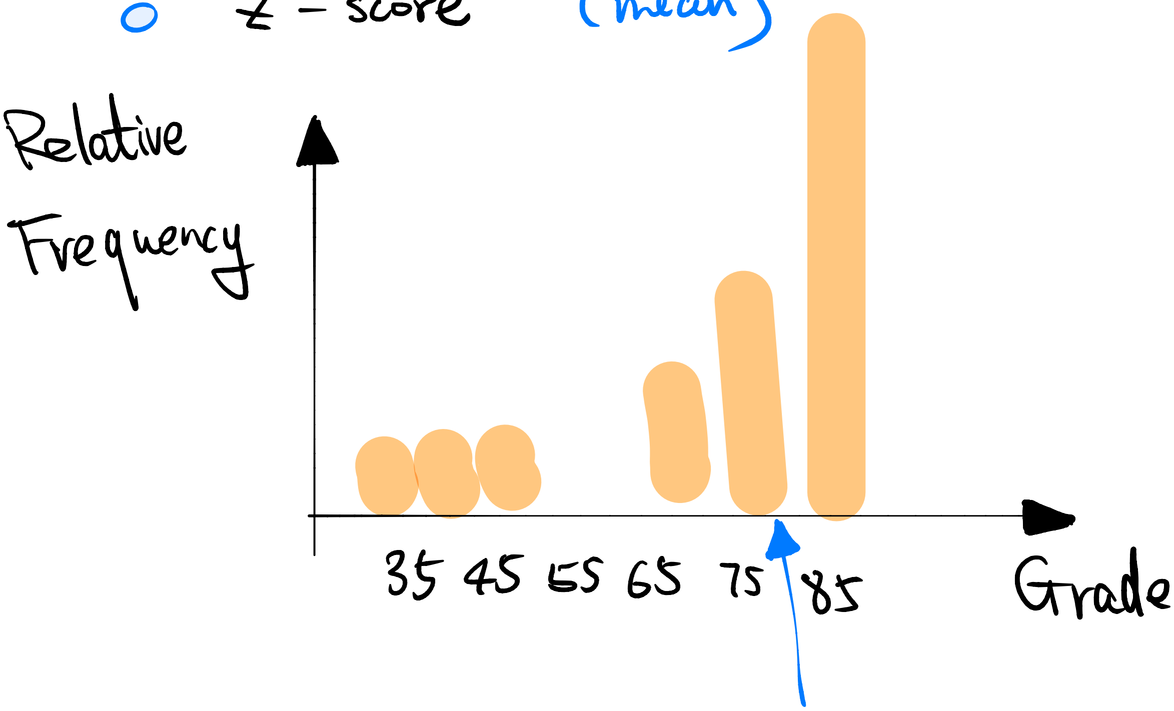
$\bar{x} > M$

The data is right skewed

Measures of Relative Standing

Describing the quantitative location of a particular measurement within a dataset.

- Percentile ranking (median)
- z - score (mean)



80 is 25th pencentile

Percentile

25% of students their grades ≤ 80

Definition:

p-th percentile ($p \in [0, 100]$) is the number such that $p\%$ of the observations fall below the number and $(100-p)\%$ fall above it.

Quartile (special percentile)

- lower quartile

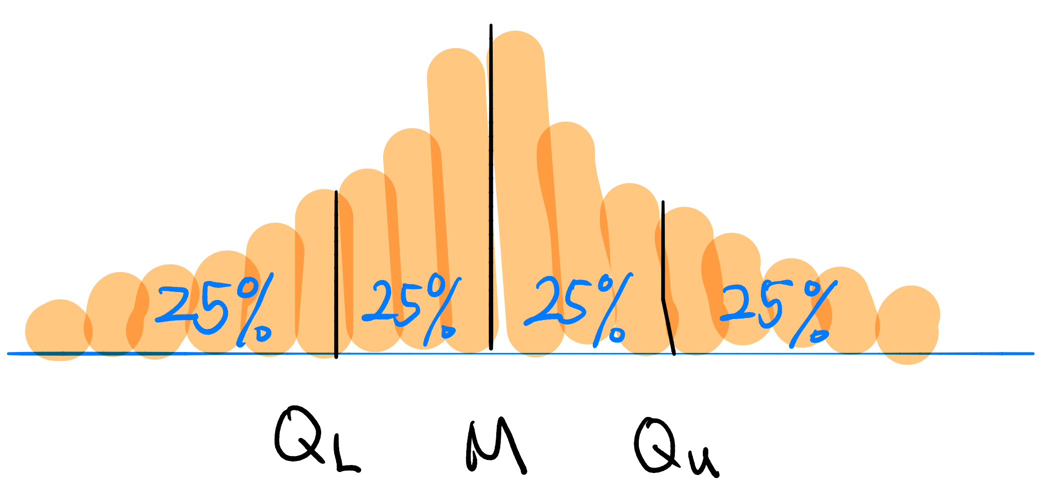
  Q_L = 25-th percentile

- median

  M = 50-th percentile.

- upper quartile

  Q_U = 75-th percentile

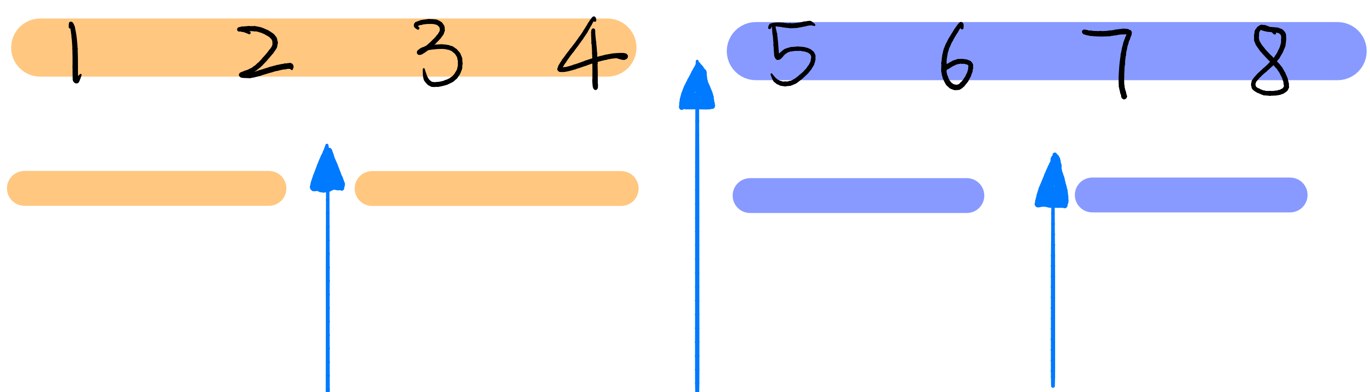


How to find quartiles

- Rank the observations in order (from small to large)
- Cut the order sequence into 4 equal parts
- Find the quartiles at these "cuts"

Ex

Data:



$$Q_L = \frac{2+3}{2} \qquad M = \frac{4+5}{2} \qquad Q_U = \frac{6+7}{2}$$

Data:

1 2 3 4 5 6 7 8 9

$Q_L = 3$ $M = 5$ $Q_u = 7$

Z - score

Definition:

obs.

$$Z = \frac{x - \bar{x}}{s}$$

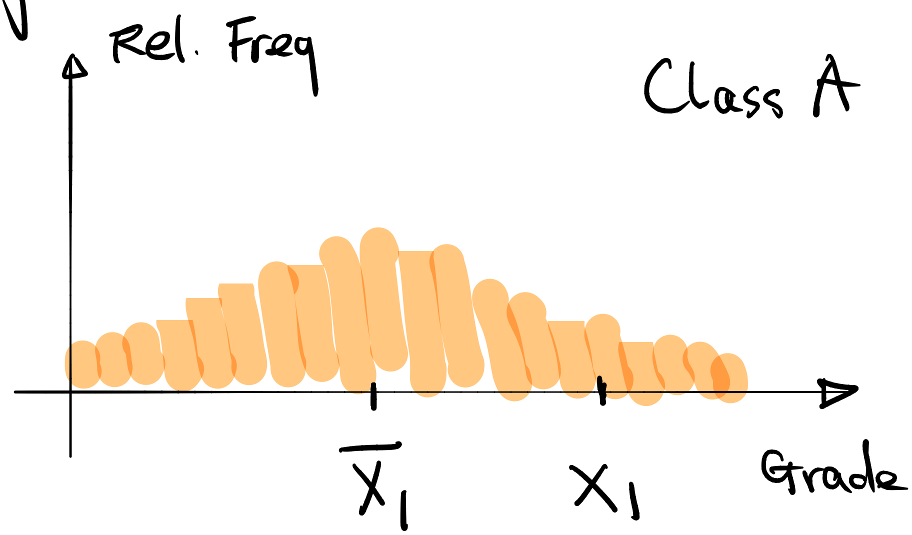
Idea: $x - \bar{x}$ measure the deviation from the mean

$\frac{1}{s}$ is the weight of that deviation.

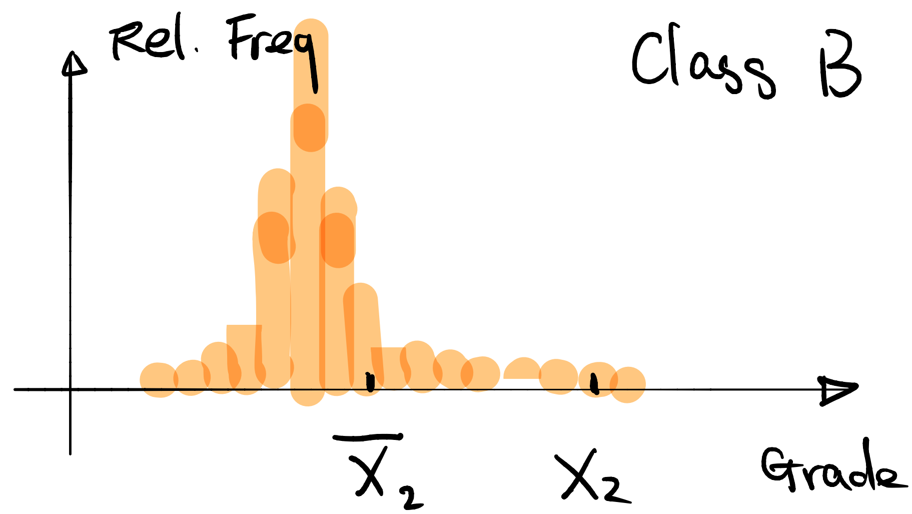
Why there is a weight $\frac{1}{s}$?

S is standard deviation, representing the variability of the data

Histogram



S_1



S_2

$$S_1 > S_2 \quad \bar{X}_1 = \bar{X}_2$$

If $x_1 - \bar{x} = x_2 - \bar{x}$

but $s_1 > s_2$, $\frac{1}{s_1} < \frac{1}{s_2}$

then
$$\frac{x_1 - \bar{x}_1}{s_1} < \frac{x_2 - \bar{x}_2}{s_2}$$

i.e. $z_1 < z_2$

Interpretation:

z represents the distance between a given observation x and mean $\bar{x}$, expressed by in standard deviation s. (adjusted by variability)

Given the same deviation $x-\bar{x}$ the higher the variability, the less extreme the observation x is, relatively to the other observations.

Ex

A sample of 2000 students scores $\bar{x}=550$ and $s=75$

$x_1=475$. $x_2=625$ $x_3=700$

Solution:

$$z_1 = \frac{475-\bar{x}}{s} = \frac{475-550}{75} = -1$$

$$z_2 = \frac{625-550}{75} = 1$$

$$z_3 = \frac{700-550}{75} = 2.$$

Student 3 has the best relative performance.

What if we change $S = 25$?

$$z_1' = \frac{475-550}{25} = -3$$

$$z_2' = \frac{625-550}{25} = 3$$

$$z_3' = \frac{700-550}{25} = 6.$$

Empirical rule

If data dist. is bell-shape and symmetric

then

1) 68% of obs will have z-score between -1 and 1

   68% of x's $\in (\bar{x}-s, \bar{x}+s)$

2) 95% of obs. will have z-score between -2 and 2.

   95% of x's $\in (\bar{x}-2s, \bar{x}+2s)$

$$z = \frac{x-\bar{x}}{s} \in (-2, 2) \Rightarrow x-\bar{x} \in (-2s, s)$$

$$\Rightarrow x \in (\bar{x}-2s, \bar{x}+2s)$$

3) 99.7% $\cdots$ ----

$z \in (-3, 3)$, $x \in (\bar{x} - 3s, \bar{x} + 3s)$

Detecting outlier

To identify inconsistent or unusual observations in a dataset.

Definition :

An outlier is an obs that is actually large or small observations relative to the other observations.

Causes :

1) incorrectly recorded , wrong entry.
2) come from a different population,
3) correct entry , but rare event.

Methods to detect outlier

- Boxplot

- z - score

Boxplot

Range A. Interquartile Range (IQR)

$$IQR = Q_u - Q_L$$

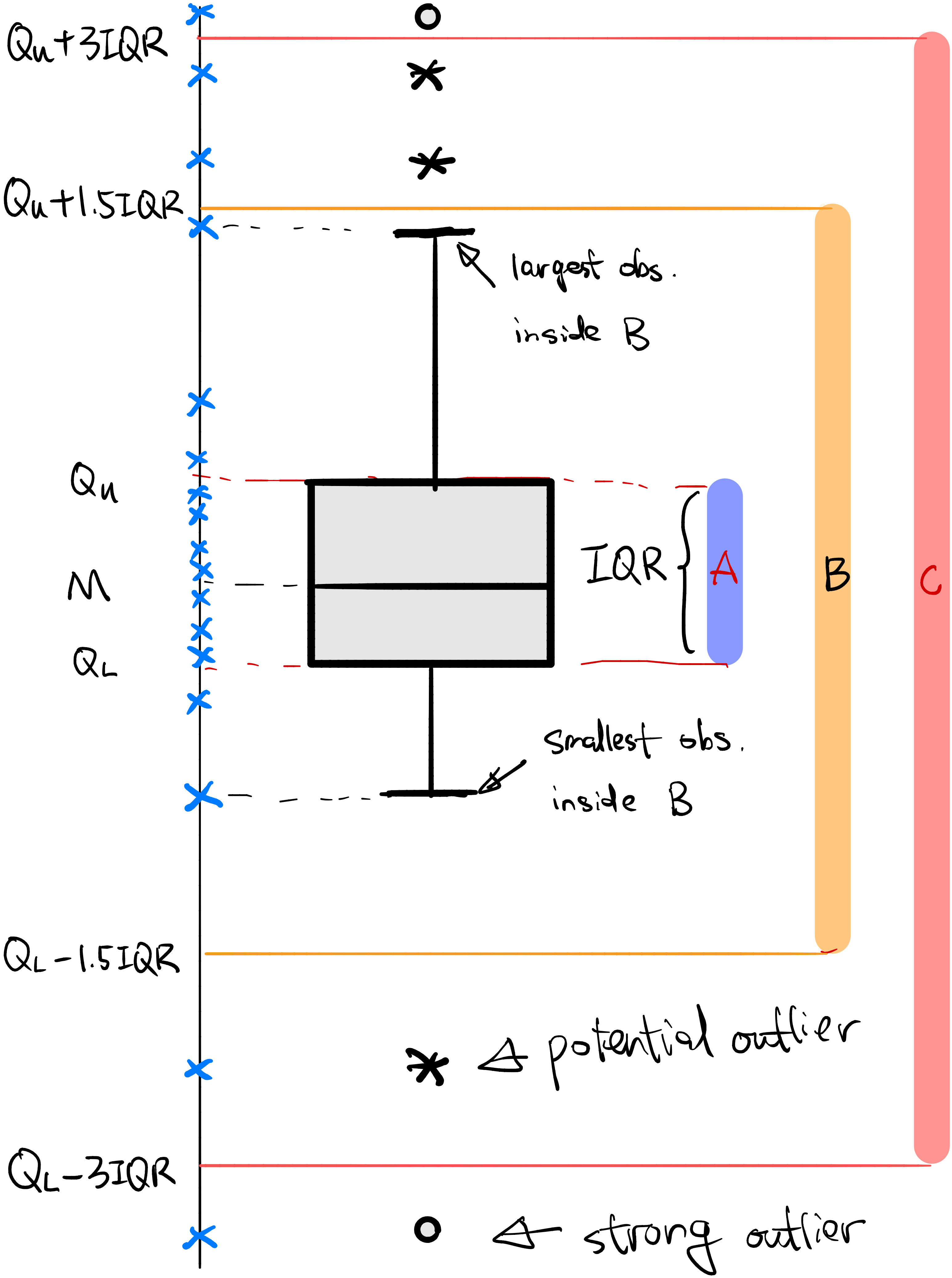
This size of IQR indicates the range of the middle 50% of the observations.

Range B Range of inner fences.

$$(Q_L - 1.5 \times IQR \ , \ Q_u + 1.5 \times IQR)$$

Range C Range of outer fences

$$(Q_L - 3 \times IQR \ , \ Q_u + 3 \times IQR)$$



Qu+3IQR
Qu+1.5IQR
largest obs. inside B
Qu
M
QL
IQR
A
B
C
smallest obs. inside B
QL-1.5IQR
potential outlier
QL-3IQR
strong outlier

Rule of Thumb for detecting outlier

(1) $x \in$ Range C

$x \notin$ Range B

$x \in (Q_L - 3IQR, Q_L - 1.5IQR)$

or $x \in (Q_u + 1.5IQR, Q_u + 3IQR)$

Then x is a potential outlier

(2) x is outside range C.

$x \in (-\infty, Q_L - 3IQR)$

$x \in (Q_u + 3IQR, +\infty)$

Then x is a strong outlier.

The largest and smallest observations in the dataset are not necessarily outliers.

Ex.

Data: $n = 13$

$0, 1, 2, \boxed{4}, 5, 5, \boxed{7}, 10, 10, \boxed{12}, 13, 17, 39$

$$M = 7$$

$$Q_L = 4$$

$$Q_u = 12$$

$$IQR = Q_u - Q_L = 8$$

B: inner fences.

$$(Q_L - 1.5\,IQR,\ Q_u + 1.5\,IQR)$$

$$(4 - 1.5 \times 8,\ 12 + 1.5 \times 8) = (-8, 24)$$

C. outer fences

$$(Q_L - 3\,IQR,\ Q_u + 3\,IQR)$$

$$(4 - 3 \times 8,\ 12 + 3 \times 8) = (-20, 36)$$

The 39 is a strong outlier.

Z - score

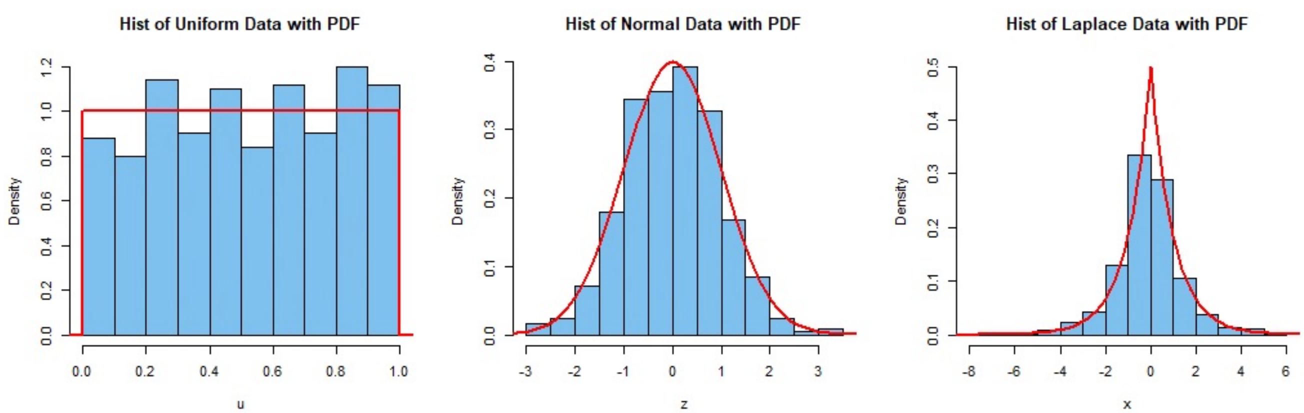
$|Z| > 3 \Rightarrow$ outlier

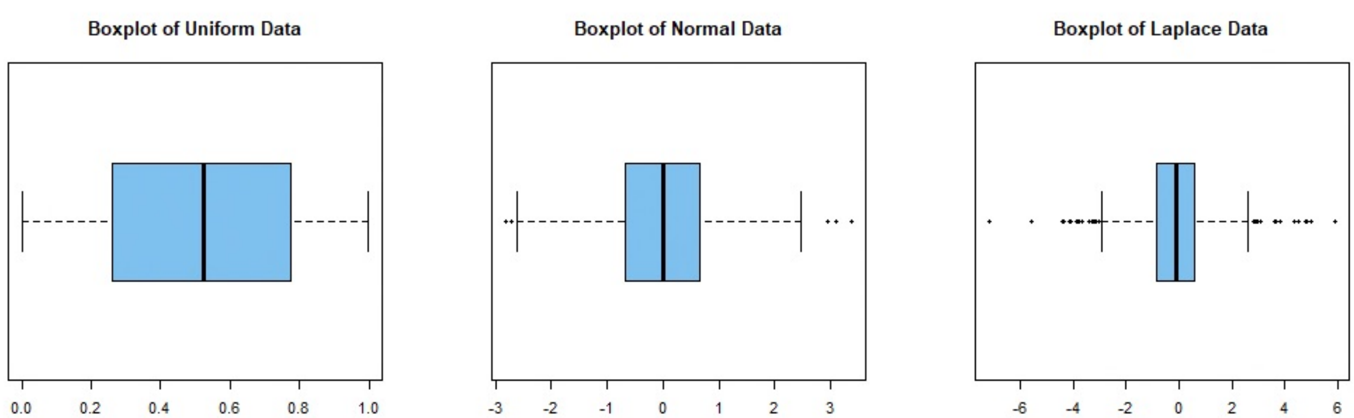
$$\left|\frac{39-\bar{X}}{s}\right| > 3.$$

Interpretation of boxplots

1. The line inside the boxplot is the center (median) of the distribution of the data.

2. Examine the length of the box. ($IQR = Q_U - Q_L$) measures data variation

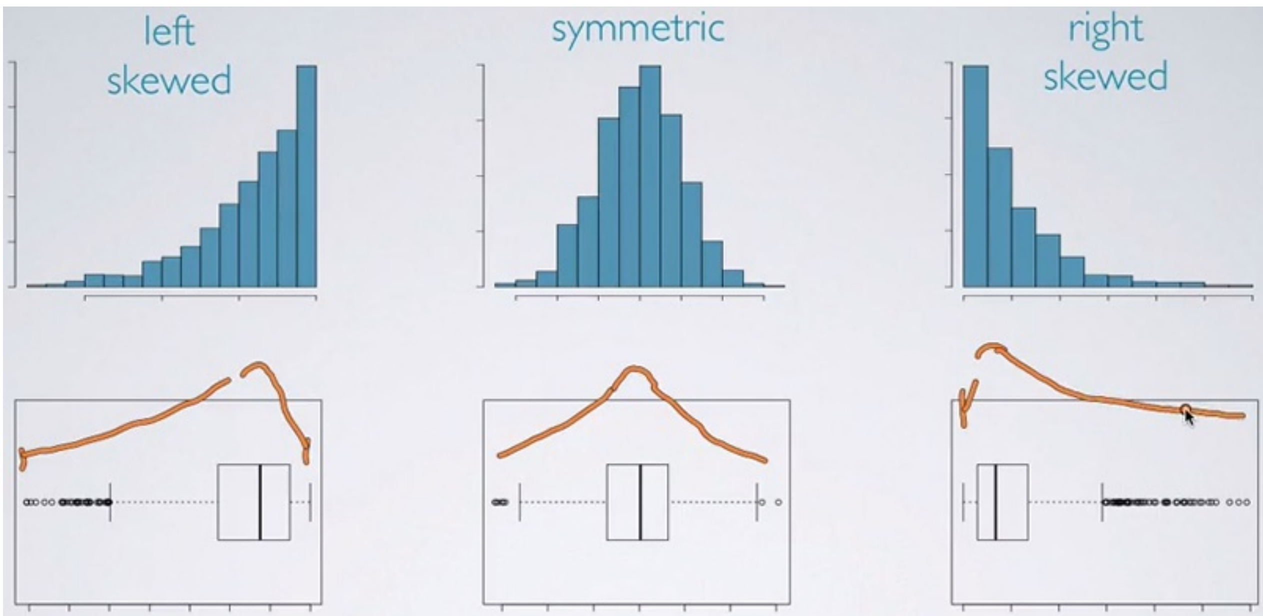




3. Visually compare the lengths of the whiskers

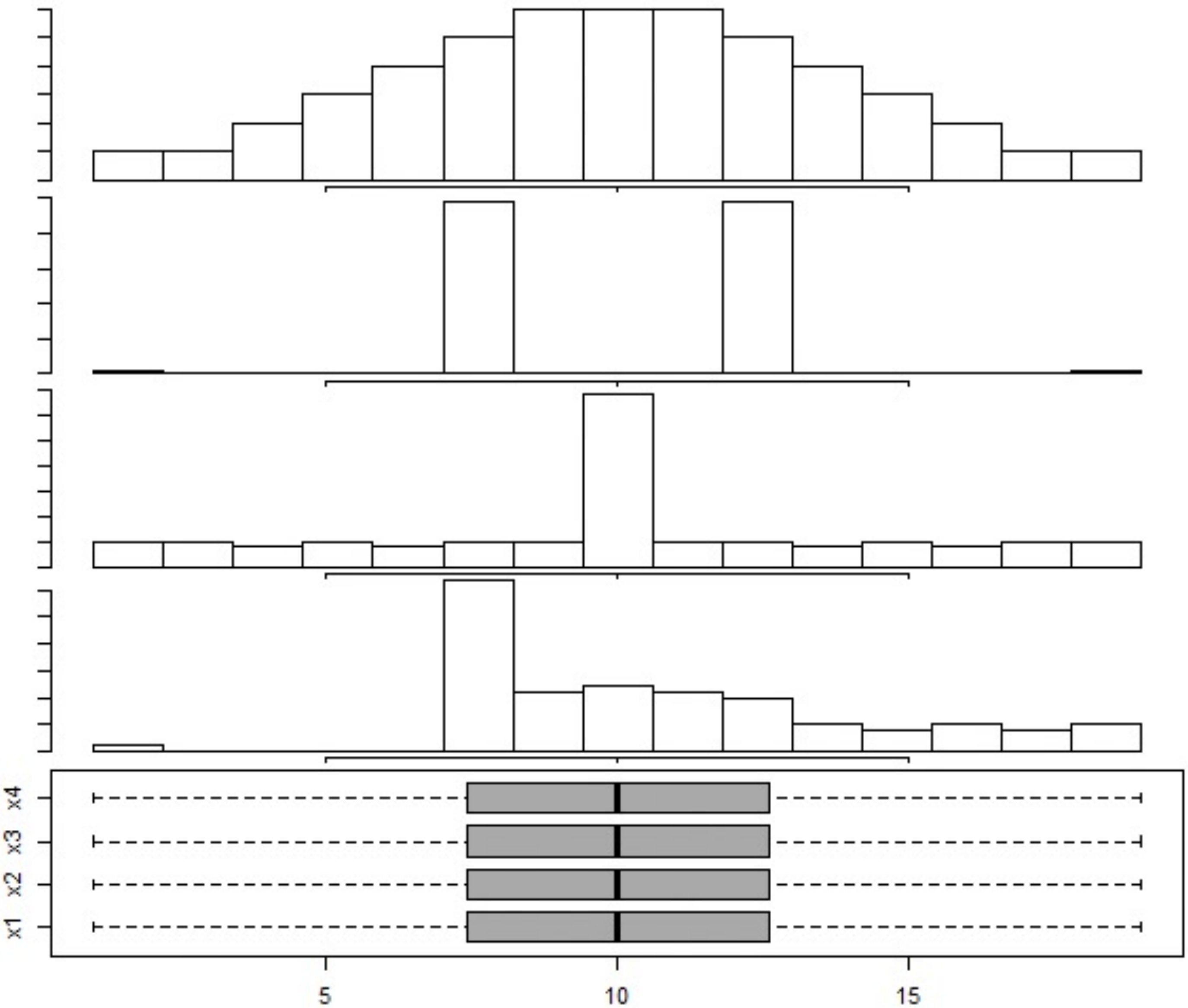
Rule of thumb: the distribution of data is skewed in the direction of longer whisker.

(the data distribution must be unimodal)



4. The boxplot is less informative than histogram

It only gives a few pieces of information about the whole data.



x4
x3
x2
x1
5
10
15