Descriptive Statistics

Mahbub Latif, PhD

October 2024

Describing data sets

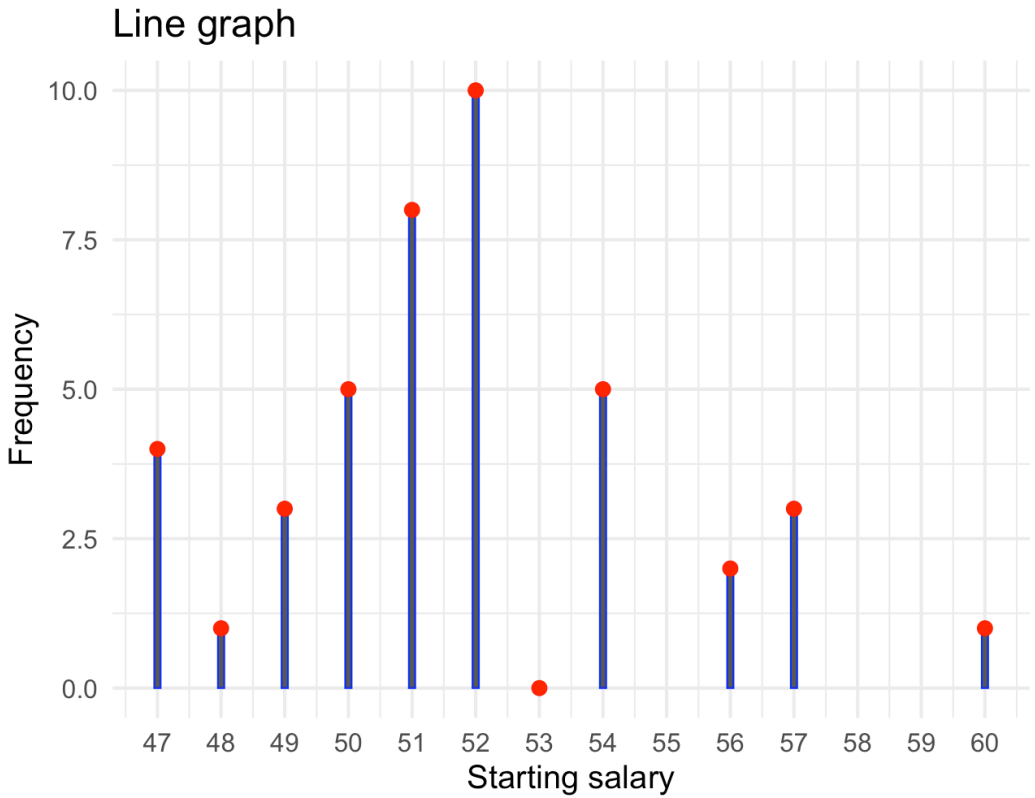
- The numerical findings of a study should be presented clearly, concisely and in such a manner that others can quickly obtain a feel for the essential characteristics of the data
- Tables and graphs are useful ways of presenting data
- The choice of graphs depends on the type of data
  - Bar graph is used for qualitative data
  - For quantitative data, histogram, line graph, and frequency plygons are used

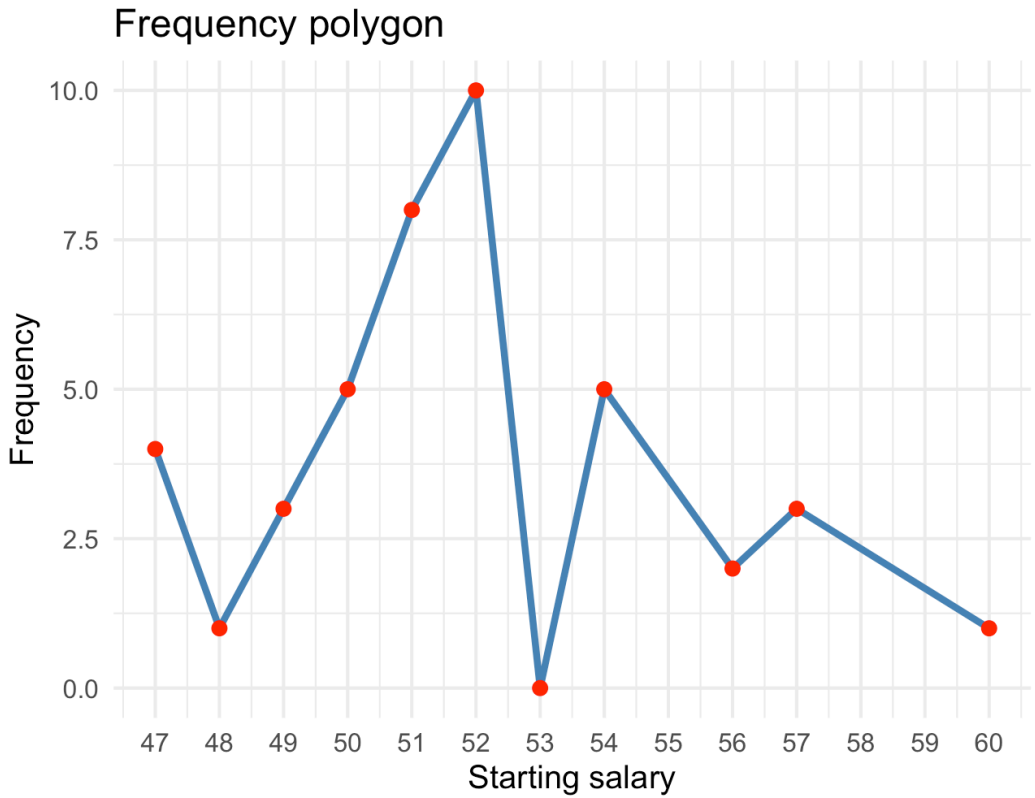
Frequency tables

- A data set can be conveniently presented in a frequency table if the data set has a relatively small number of distinct values
- Consider a sample of starting salary of 42 recent graduates with BS degrees in electrical engineering
- Salary is a quantitative variable

Starting salary	*Frequency*
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

Graphical summary of salary data



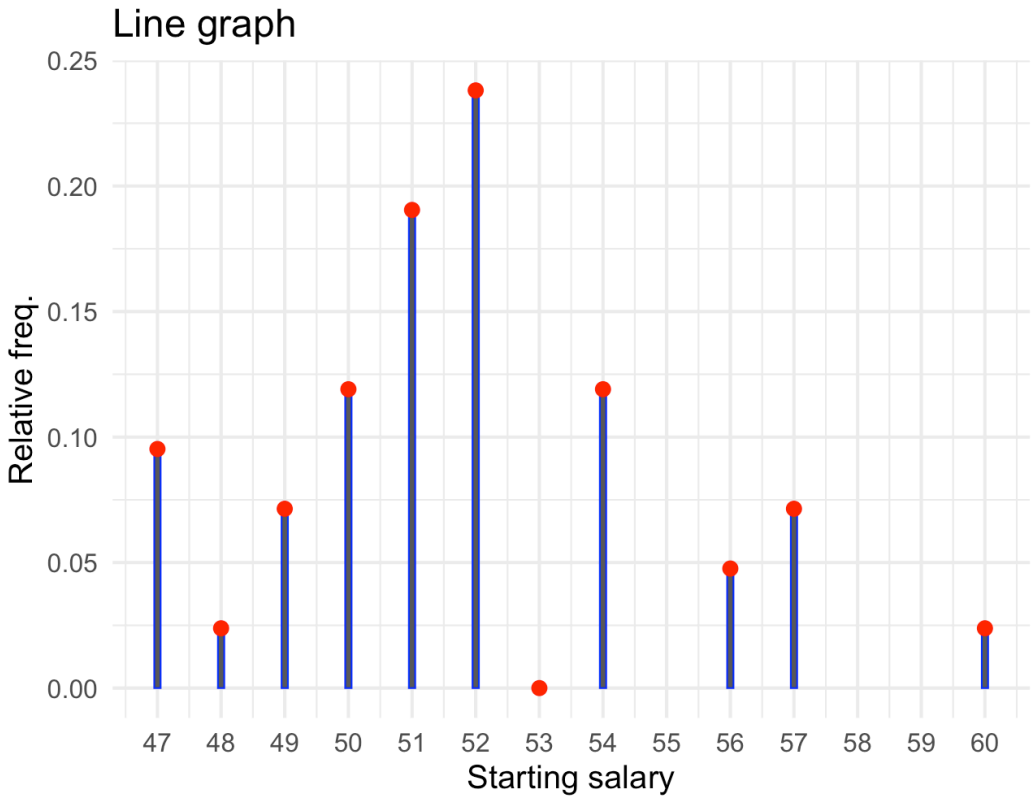


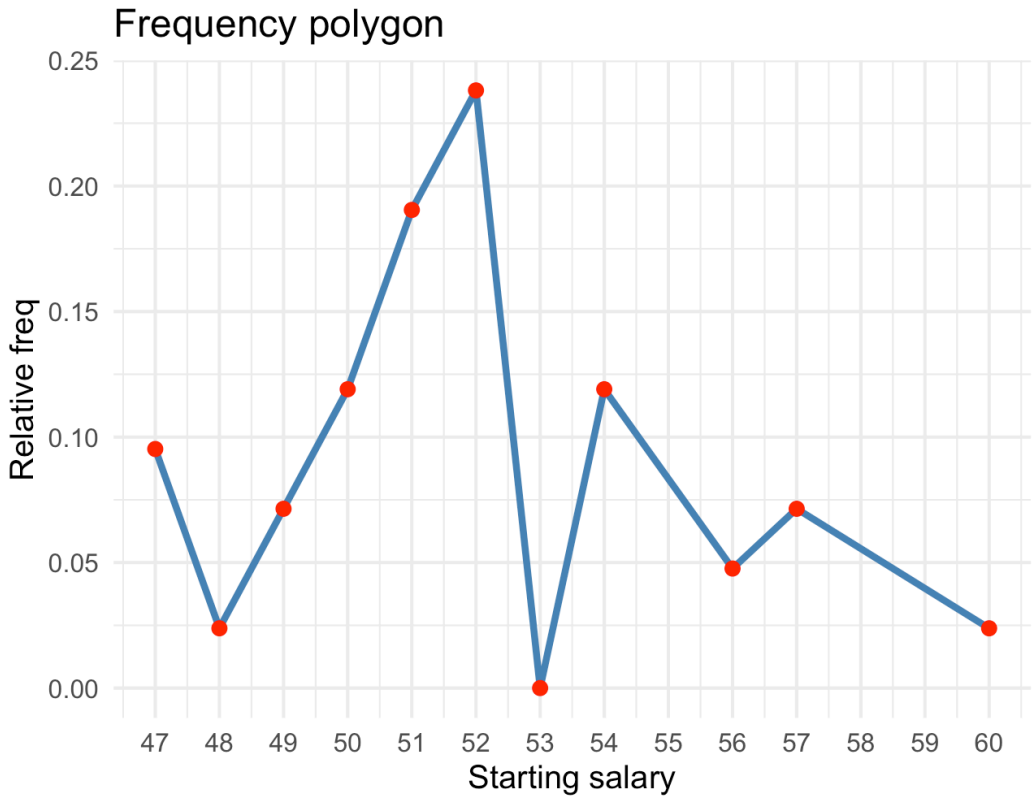
Relative frequency

- Relative frequency of a data value is the proportion of the data that have that value
- If there are n values in the data set and f is the frequency of a particular value
  - $(f/n) \rightarrow$ relative frequency

Salary, x	*Freq., f*	*relative freq. f/n*
47	4	0.095 (=4/42)
48	1	0.024 (=1/42)
49	3	0.071 (=3/42)
50	5	0.119 (=5/42)
51	8	0.190 (=8/42)
52	10	0.238 (=10/42)
53	0	0.000 (=0/42)
54	5	0.119 (=5/42)
56	2	0.048 (=2/42)
57	3	0.071 (=3/42)
60	1	0.024 (=1/42)

Graphical summary of salary data

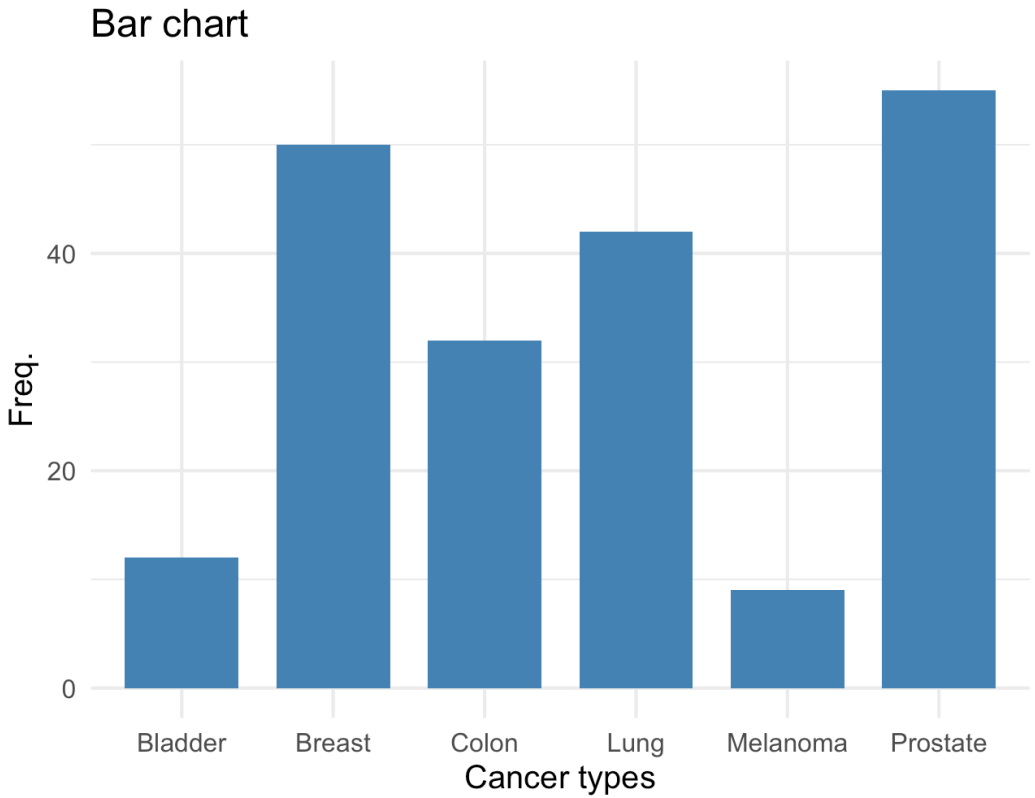


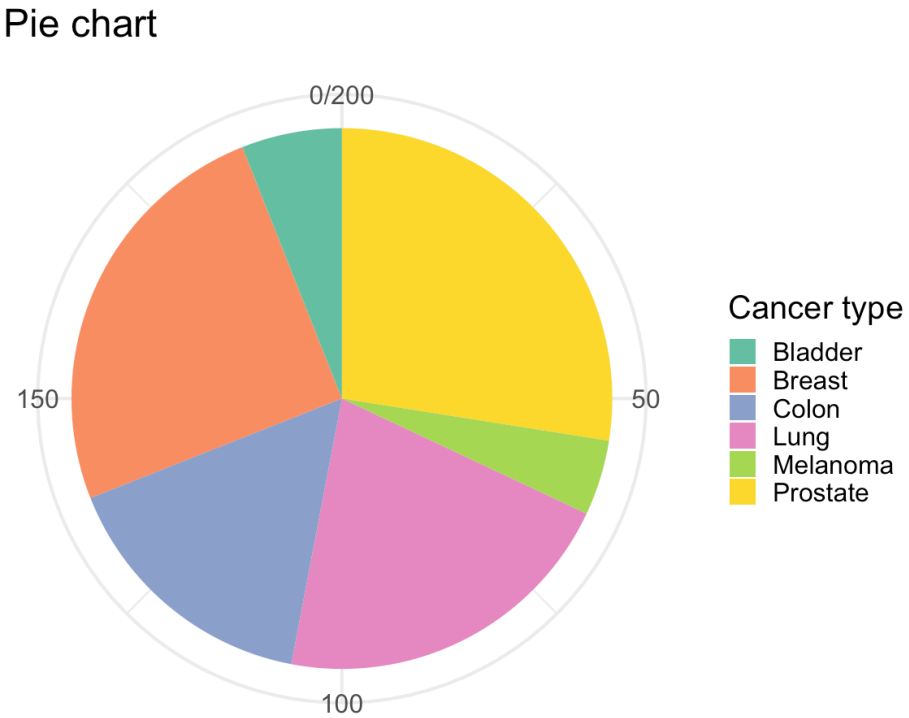


Qualitative data

- Data available on different types of cancers affecting the 200 most recent patients to enroll at a clinic specializing in cancer
- "Cancer type" is a qualitative variable and Bar chart or pie chart can be used as graphical tools

Cancer type	*Frequency*	*Relative freq.*
Lung	42	0.210
Breast	50	0.250
Colon	32	0.160
Prostate	55	0.275
Melanoma	9	0.045
Bladder	12	0.060





- Width and position of bars are not important

- Bar chart is more useful than a pie chart as a graphical tool

Group data and related graphical tools

Lifetime of lamps

- Lifetime (in hour) is a quantitative variable and frequency distribution of raw data (200 observations) would not be very useful as number of distinct values is too large

- Divide the data into groups (known as class interval) and frequency distribution of the grouped data would be useful

Lifetime of lamps

TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

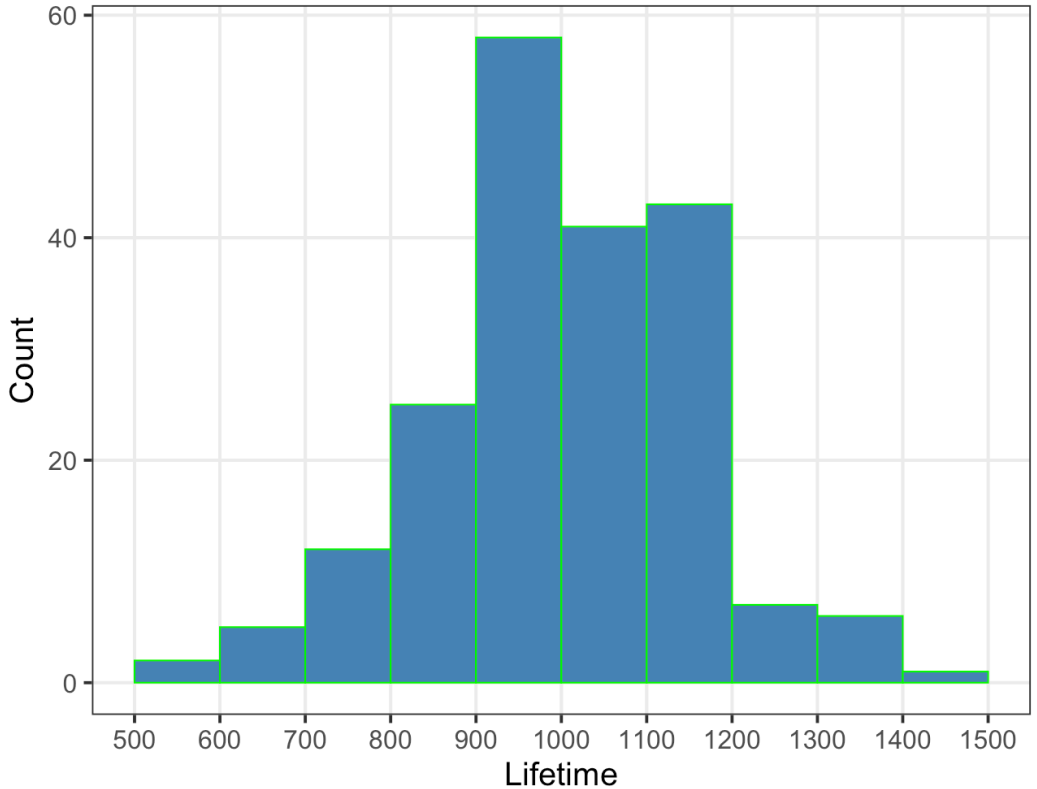
Grouped data

- The number of class intervals should not be too large or too small
  - 5-10 class intervals are typical
- Endpoints of a class interval is known as class boundaries and we will adopt left-end inclusion convention, i.e. class interval contains its left-end but not the right-end boundary point
  - E.g. the interval 20–30 contains all the values greater than or equal to 20 and less than 30, i.e. $(20 \leq \text{Lifetime} < 30)$
- *Histogram* is commonly used as a graphical tool for grouped quantitative data

Frequency distribution

Class interval	Frequency
500-600	2
600-700	5
700-800	12
800-900	25
900-1000	58
1000-1100	41
1100-1200	43
1200-1300	7
1300-1400	6
1400-1500	1

Histogram



Relative frequency and density

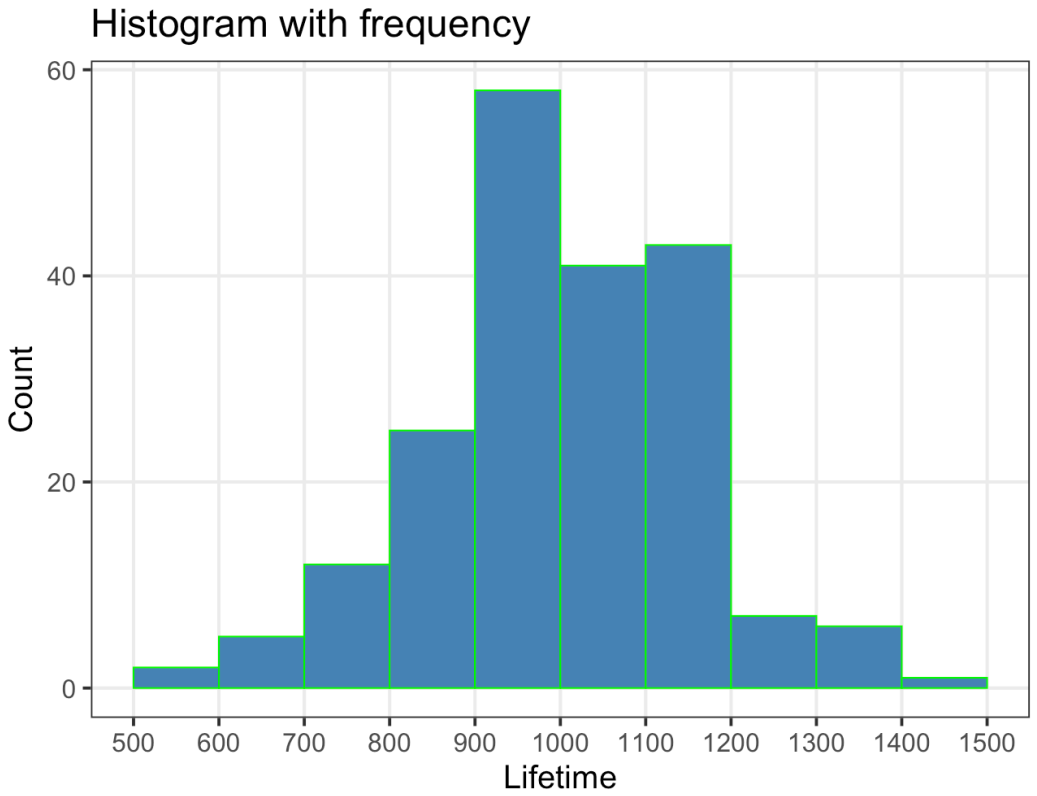
$$\text{Relative freq} = \frac{\text{Class frequency}}{\text{Total frequency}}$$

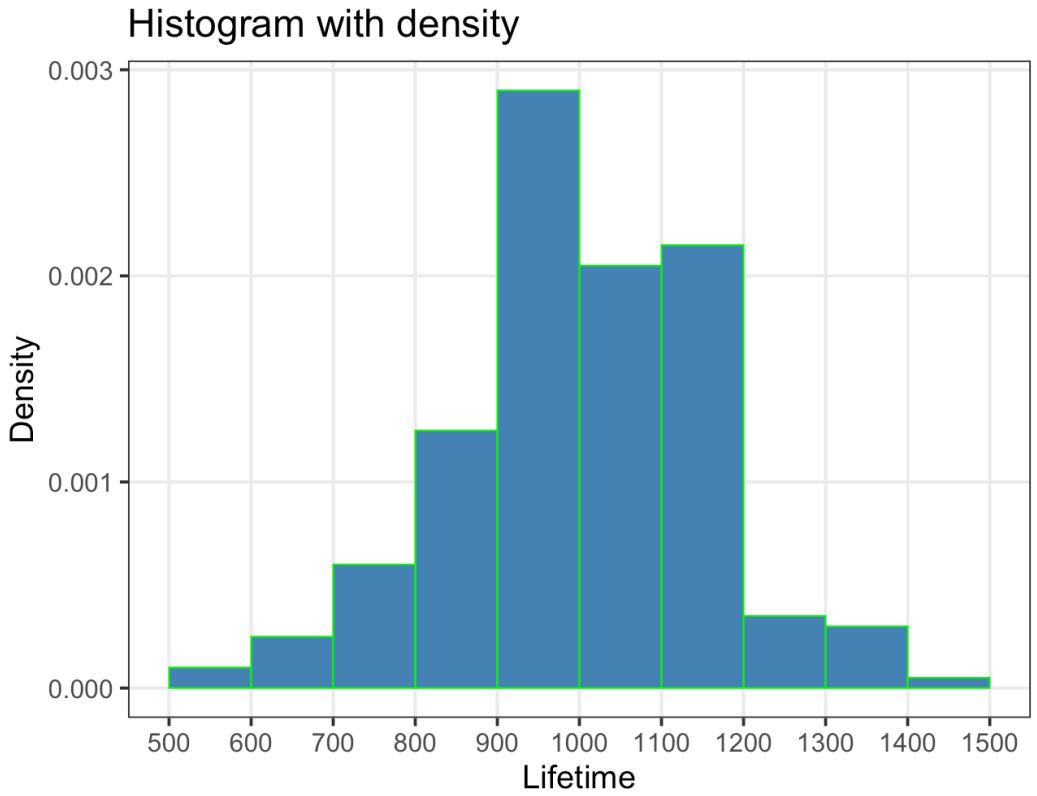
$$\text{Density} = \frac{\text{Relative freq}}{\text{Class interval}}$$

- Histograms can be drawn using relative frequency and density

Class interval	Frequency	Relative freq.	Density
500-600	2	0.010	0.00010
600-700	5	0.025	0.00025
700-800	12	0.060	0.00060
800-900	25	0.125	0.00125
900-1000	58	0.290	0.00290
1000-1100	41	0.205	0.00205
1100-1200	43	0.215	0.00215
1200-1300	7	0.035	0.00035
1300-1400	6	0.030	0.00030
1400-1500	1	0.005	0.00005

Histogram of lifetime



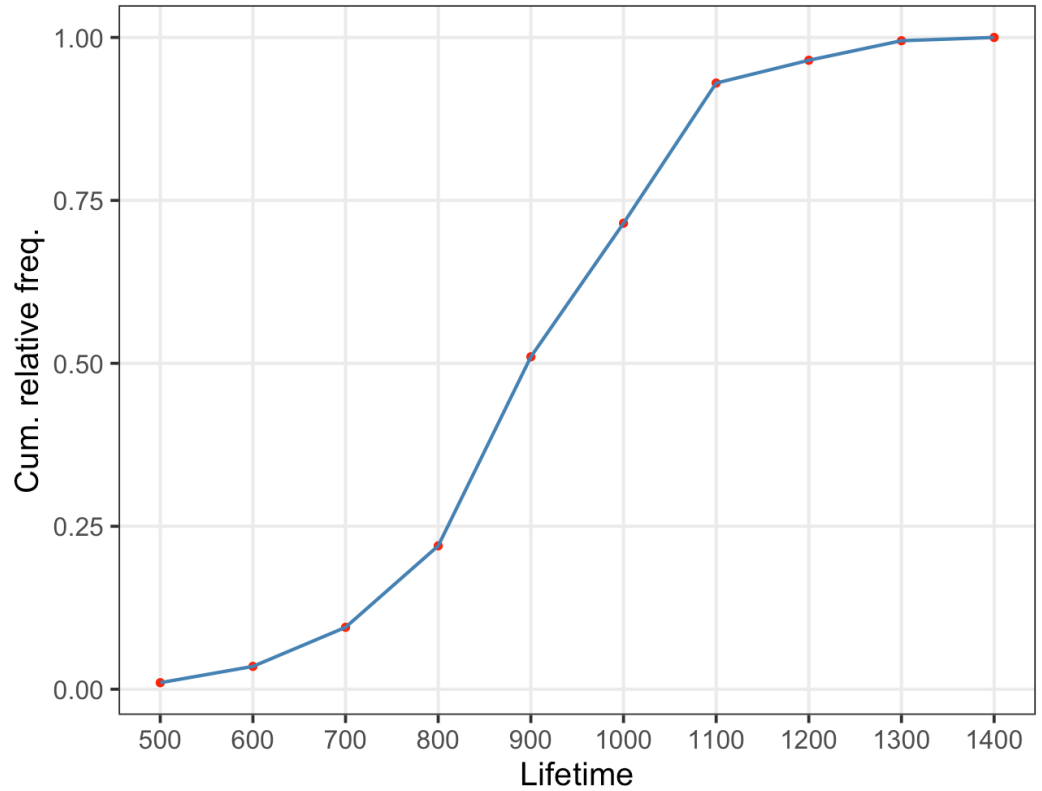


Cumulative frequency and cumulative relative frequency

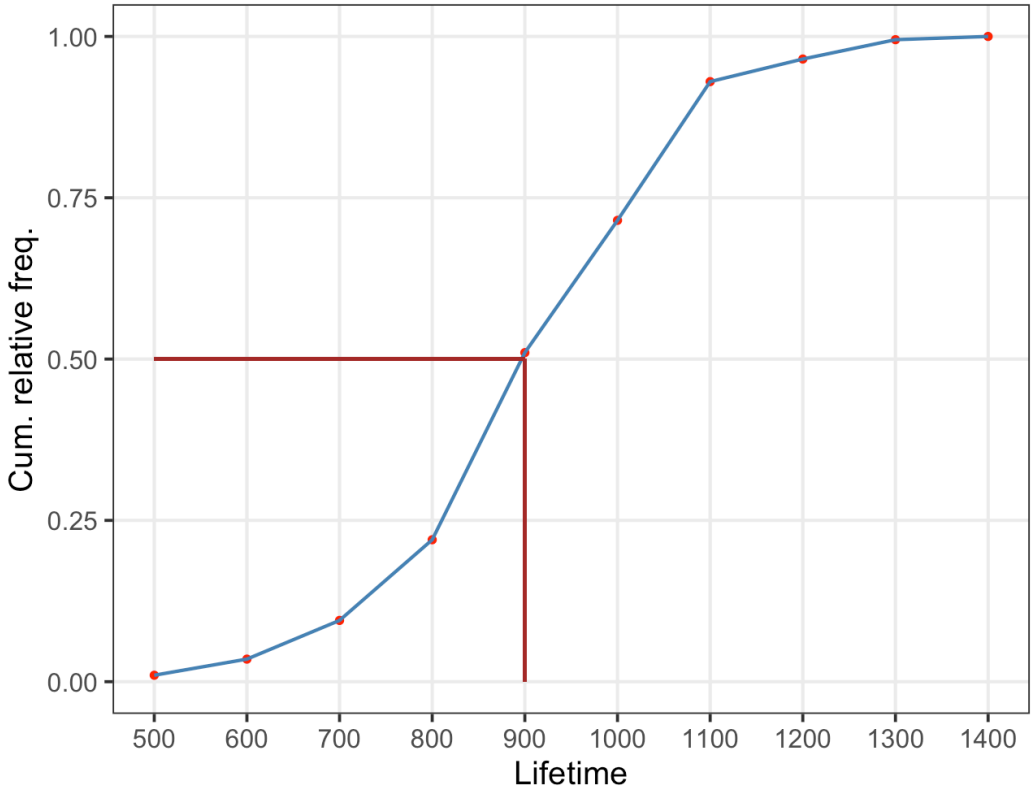
Class interval	Frequency	Cumulative freq.	Cumulative relative freq
500-600	2	2	0.010
600-700	5	7	0.035
700-800	12	19	0.095
800-900	25	44	0.220
900-1000	58	102	0.510
1000-1100	41	143	0.715
1100-1200	43	186	0.930
1200-1300	7	193	0.965
1300-1400	6	199	0.995
1400-1500	1	200	1.000

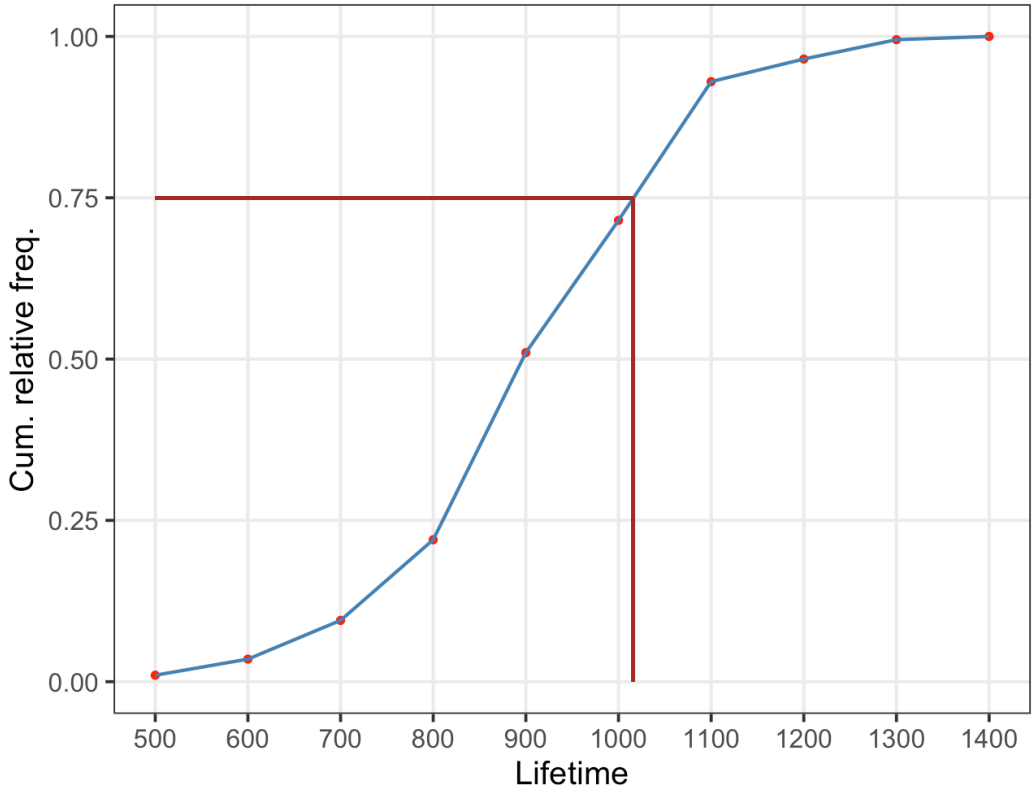
Cumulative relative frequency plot

- Cumulative relative frequency plot is useful to obtain some useful characteristics of the distribution
- E.g. about 50% lamps have lifetime less than or equal to 900 hours
- Cumulative relative frequency plot is also known as *ogive*



Cumulative relative frequency plot





Stem and Leaf plot

- Stem and leaf plot is an efficient way of organizing a small- to moderate-sized data set
- It is obtained by dividing each data value into two parts – stem and leaf
  - E.g. the data value 62 is divided into *Stem* $(= 6)$ and *Leaf* $(= 2)$
  - Similarly, for data values 62 and 67, Stem will be 6 and Leaf will be 2 and 7

Stem and Leaf plot

- **The annual average daily minimum temperatures in 35 U.S. cities**

7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2	9.0, 9.8

Summarizing Data Sets

Summarizing Data Sets

- There are two important numerical measures of summarizing data
  - Measures of central tendency or location
  - Measures of spread or dispersion
- A measure of central tendency is a representative value of the data set
- A measure of dispersion quantifies the average distances of the data points from its location value

Measures of central tendency or location

- Most common measures of location
  - Sample mean, median, and mode

Sample mean

- The sample mean is a measure of location, which can be obtained only for quantitative data

- Let $x_1, \ldots, x_n$ be n numerical values and the sample mean of these values is denoted by $\bar{x}$ (read "x-bar") and is defined as

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

Sample mean

- The sample mean of 5 numerical values $\{10, 15, 20, 18, 12\}$ is

$$\bar{x} = \frac{1}{5} \sum_{i=1}^{5} x_i = \frac{75}{5} = 15$$

Sample mean

- Transformed data $y_i = ax_i + b$, where a and b are constants
- Sample mean of transformed data

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\ &= a\bar{x} + b\end{aligned}$$

Sample mean

- For x values $\{10, 15, 20, 18, 12\}$, sample mean of $y_i = 2x_i + 5$

$$\bar{y} = \frac{1}{5}\sum_{i=1}^{n} y_i$$
$$= \frac{175}{5} = 35 = 2\bar{x} + 5$$

  - $\bar{x} = 15$

Sample mean

- Calculate the sample mean of the following winning scores of US masters golf tournament

$$280, 278, 272, 276, 281, 279, 276, 281, 289, 280$$

- It is easier to calculate sample mean $\bar{x}$ using the sample mean of the transformed scores $(y_i = x_i - 280)$ than by direct calculation using the x values

Sample mean as an weighted average

- Suppose the data consist of k values $v_1, \ldots, v_k$, each of which repeated $f_1, \ldots, f_k$ times, respectively, such that $n = \sum_{i=1}^{n} f_i$
- The sample mean is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} f_i v_i = \sum_{i=1}^{k} w_i v_i, \quad where\ w_i = f_i / n$$

- Expressing a sample mean in terms of weights w's and frequencies f's is known as *weighted mean*

Example 2.3b

- The frequency table shows the ages of members of a symphony orchestra for young adults.
- Find the sample mean of the ages of the 54 members of the symphony.

AGE	FREQUENCY
15	2
16	5
17	11
18	9
19	14
20	13

Example 2.3b

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{6} f_i v_i$$

$$= \frac{1}{54}(985) = 18.24$$

- $n = \sum_{i=1}^{6} f_i = 54$

Age, *v*	Freq, *f*	*fv*
15	2	30
16	5	80
17	11	187
18	9	162
19	14	266
20	13	260

Sample median

- Sample median is another measure of location and it indicates the middle value when the data set is arranged in increasing or decreasing order
- Let $x_1, \ldots, x_n$ be the sample observations and $x_{(1)}, \ldots, x_{(n)}$ is the corresponding ordered sample
- The sample median is defined as

$$median = \begin{cases} x_{((n+1)/2))} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Sample median

- Obtain sample median of the $x_1, \ldots, x_n$

$$23, 14, 54, 22, 19, 32, 27$$

- Ordered sample $x_{(1)}, \ldots, x_{(n)}$

$$14, 19, 22, 23, 27, 32, 54$$

- Since $n = 7$, the sample median is the 4^{th} $[= (7+1)/2]$ observation of the ordered sample, which is 23

- Calculate the sample median of the sample

$$14, 13, 15, 14, 11, 8, 21, 9$$

Sample mean and sample median

- Sample mean is applicable only for quantitative data and sample median is applicable for both quantitative and ordinal data
- Sample median is preferable over sample mean for a data set that contains extreme values (also known as outliers)

Measures of spread

- Commonly used measures of spread or dispersion
  - sample range, variance, stanard deviation

The sample variance

- The sample variance and standard deviation are measures of spread
- For a sample of n observations, $x_1, \ldots, x_n$, the sample variance is

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

- Sample variance is the average squared distances, where distances are defined as the difference between observations and its sample mean
- The reason for using $(n-1)$ instead of n in calculating the average will be discussed later in the course

The sample variance

- Calculate sample of the following two samples and which sample has more variability
  - $A: \ 3, 4, 6, 7, 10$
  - $B: \ -20, 5, 15, 24$

The sample variance

An algebraic identity

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

Sample variance of $y_i = ax_i + b$

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= a^2 s_x^2 \end{aligned}$$

- $s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$

The sample variance

- The data give the worldwide number of fatal airline accidents of commercially scheduled air transports in the years from 1997 to 2005.
- Find sample variance using transformed observations $y_i = x_i - 18$

Year	accidents
1997	25
1998	20
1999	21
2000	18
2001	13
2002	13
2003	7
2004	9
2005	18

The sample standard deviation

- The positive square root of sample variance is known as the sample standard deviation, which is denoted by s

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

- The main advantage of standard deviation over variance as a measure of spread is that the standard deviation is measured in the same unit as the data

Sample percentiles

Sample percentiles

- The sample $100p^{th}$ *percentile* is the data value such that $100p$ percent of the data are less than or equal to it and $100(1-p)$ percent of the data are greater than or equal to it $(0 \leq p \leq 1)$
- To determine the sample $100p$ percentile of a data set of size n, we need to determine the data values such that
  - At least np of the values are less than or equal to it
  - At least $n(1-p)$ of the values are greater than or equal to it

Sample percentiles

- If np is a fraction then there will be only one data value which satisfies the above two conditions
  - E.g. $n = 22$ and $p = 0.8$, then $np = 17.6$ and $n(1-p) = 4.4$ and the 18^{th} observation is the 80^{th} percentile
- If np is an integer then two values in positions np and $(np + 1)$ satisfy the conditions, the arithmetic average of these values is the $100pth$ percentile
  - E.g. $n = 32$ and $p = 0.25$, then $np = 8$ and then average of the 8^{th} and 9^{th} observations is the 25^{th} percentile

TABLE 2.6 *Population of 25 Largest U.S. Cities, July 2006*

Rank	City	Population
1	New York, NY	8,250,567
2	Los Angeles, CA	3,849,378
3	Chicago, IL	2,833,321
4	Houston, TX	2,144,491
5	Phoenix, AR	1,512,986
6	Philadelphia, PA	1,448,394
7	San Antonio, TX	1,296,682
8	San Diego, CA	1,256,951
9	Dallas, TX	1,232,940
10	San Jose, CA	929,936
11	Detroit, MI	918,849
12	Jacksonville, FL	794,555
13	Indianapolis, IN	785,597
14	San Francisco, CA	744,041
15	Columbus, OH	733,203
16	Austin, TX	709,893
17	Memphis, TN	670,902
18	Fort Worth, TX	653,320
19	Baltimore, MD	640,961
20	Charlotte, NC	630,478
21	El Paso, TX	609,415
22	Milwaukee, WI	602,782
23	Boston, MA	590,763
24	Seattle, WA	582,454
25	Washington, DC	581,530

- Find the sample 10^{th} and 80^{th} percentiles

Quartiles

- Sample quartiles equally divide the data into four parts so that each part contains roughly 25% of observations
  - The first quartile (Q_1) $\rightarrow 25^{th}$ percentile
  - The second quartile (Q_2) $\rightarrow 50^{th}$ percentile, median
  - The third quartile (Q_3) $\rightarrow 75^{th}$ percentile

Quartiles

- The difference between the third and first quartile is known as the *Inter-Quartile Range (IQR)*, which is a measure of spread

$$IQR = Q_3 - Q_1$$

Quartiles

- Obtain the quartiles from the following data on noise levels measured at 36 different times outside Grand Central Station in Manhattan , New York City

```
## 82 89 94 110 74 122 112 95 100 78 65 60 90 83 87 75 114 85

## 69 94 124 115 107 88 97 74 72 68 83 91 90 102 77 125 108 65
```

Trimmed and Winsorized means

- In the presence of outliers in the data, trimmed and Winsorized means are used because these are more robust compared to the sample mean
- Trimming and Winsorization are methods for reducing the effects of extreme values in the sample

Trimmed and Winsorized means

- The k-times trimmed mean is calculated as

$$\bar{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

- For a sample $x_1, \ldots, x_{20}$, the 3-times trimmed mean is defined as

$$\bar{x}_{t3} = \frac{1}{14} \sum_{i=4}^{17} x_{(i)}$$

  - It excluded $\{x_{(1)}, x_{(2)}, x_{(3)}\}$ and $\{x_{(18)}, x_{(19)}, x_{(20)}\}$

Trimmed and Winsorized means

- The k-times Winsorized mean is calculated as

$$\bar{x}_{wk} = \frac{1}{n}\left\{(k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right\}$$

- For a sample $x_1, \ldots, x_{20}$, the 3-times Winsorized mean is defined as

$$\bar{x}_{w3} = \frac{1}{20}\left\{4x_{(4)} + \sum_{i=5}^{16} x_{(i)} + 4x_{(17)}\right\}$$

  - It replaced $\{x_{(1)}, x_{(2)}, x_{(3)}\}$ by $x_{(4)}$ and $\{x_{(18)}, x_{(19)}, x_{(20)}\}$ by $x_{(17)}$

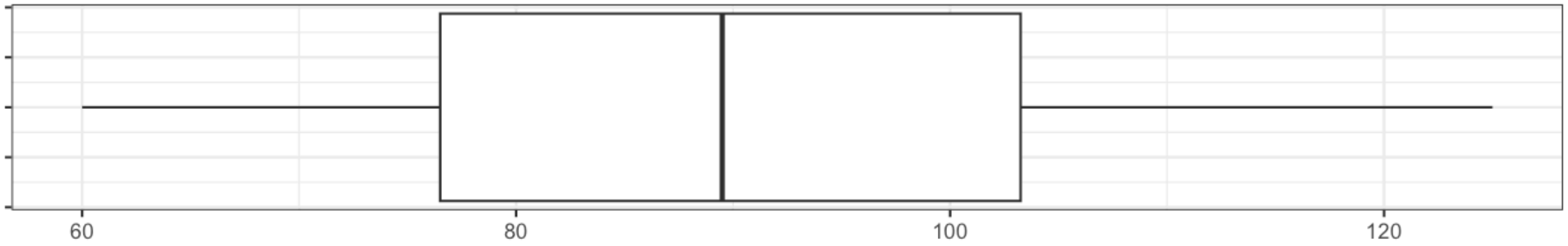
Trimmed and Winsorized means

- The Winsorized sum of squared deviations is defined as

$$ss_{wk} = \Big\{(k+1)(x_{(k+1)} - \bar{x}_{wk})^2 + \sum_{i=k+2}^{n-k-1} (x_{(i)} - \bar{x}_{wk})^2 + (k+1)(x_{(n-k)} - \bar{x}_{wk})^2\Big\}$$

Boxplot

- A boxplot is used to plot five summary statistics, which are
  - Lower-limit, first quartile, second quartile, third quartile, and upper-limit

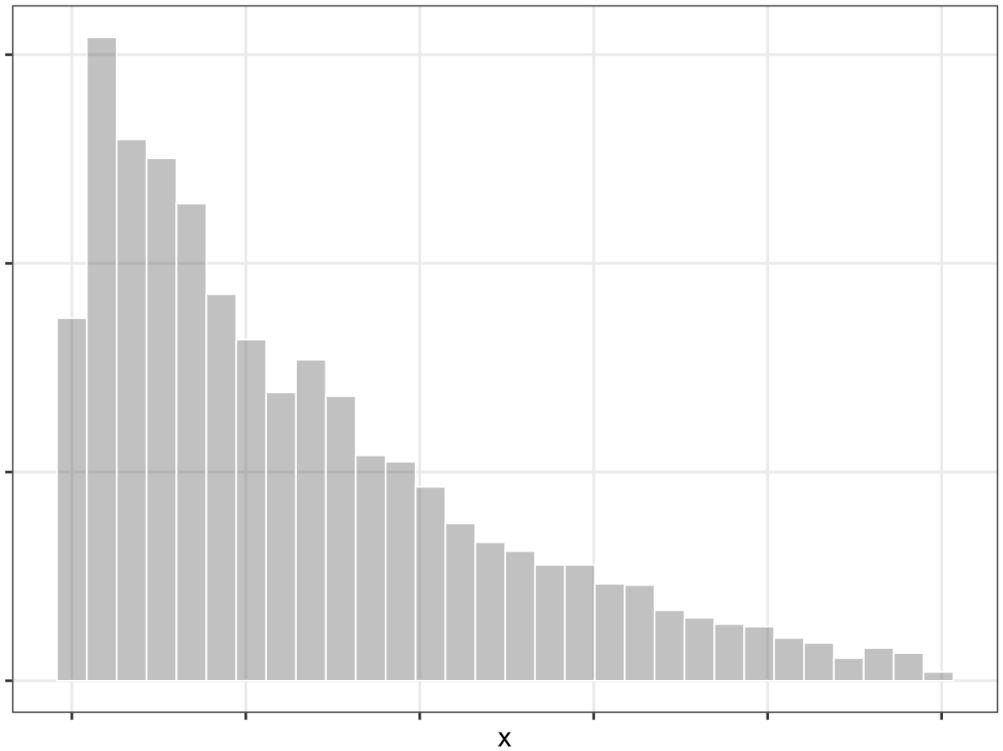


Boxplot

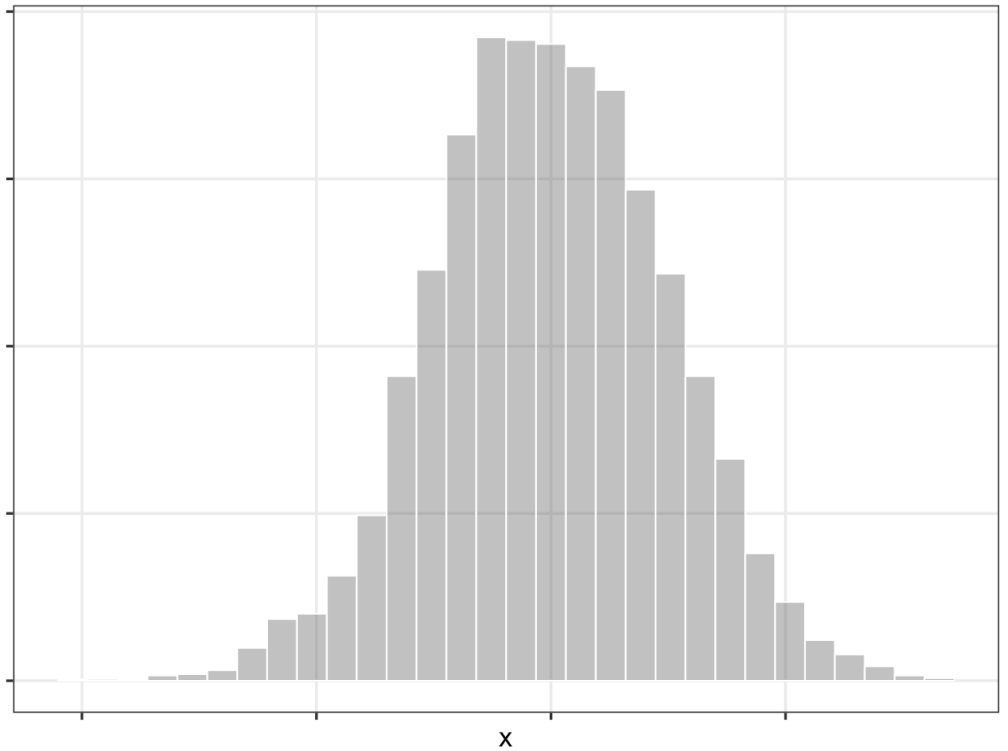
- The upper- and lower-limits are defined as
  - upper-limit = $Q_3 + (1.5 \times IQR)$
  - lower-limit = $Q_1 - (1.5 \times IQR)$
- Outliers are the observations that are greater than upper-limit or smaller than lower-limit

Boxplot and skewness

Positively skewed distribution

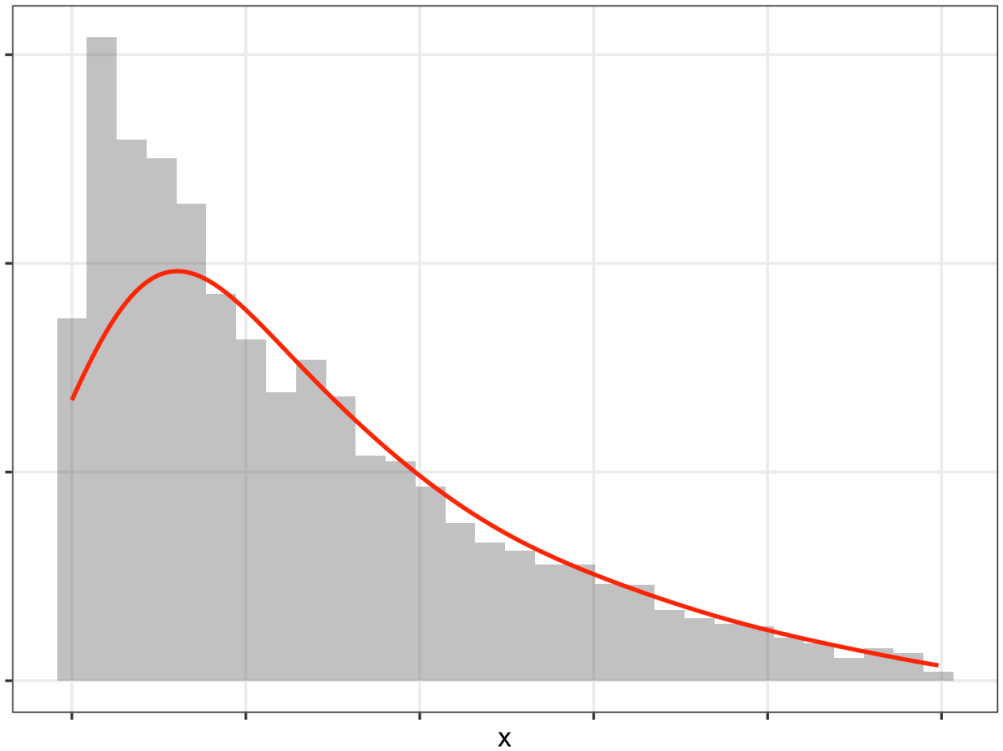


Symmetric distribution

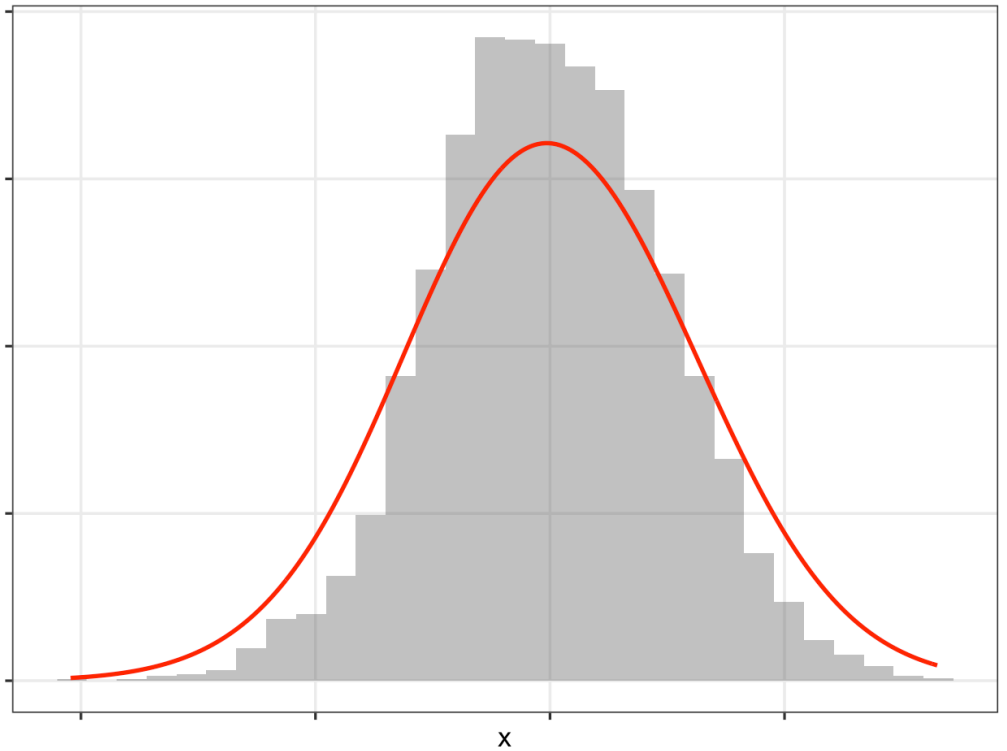


Boxplot and skewness

Positively skewed distribution

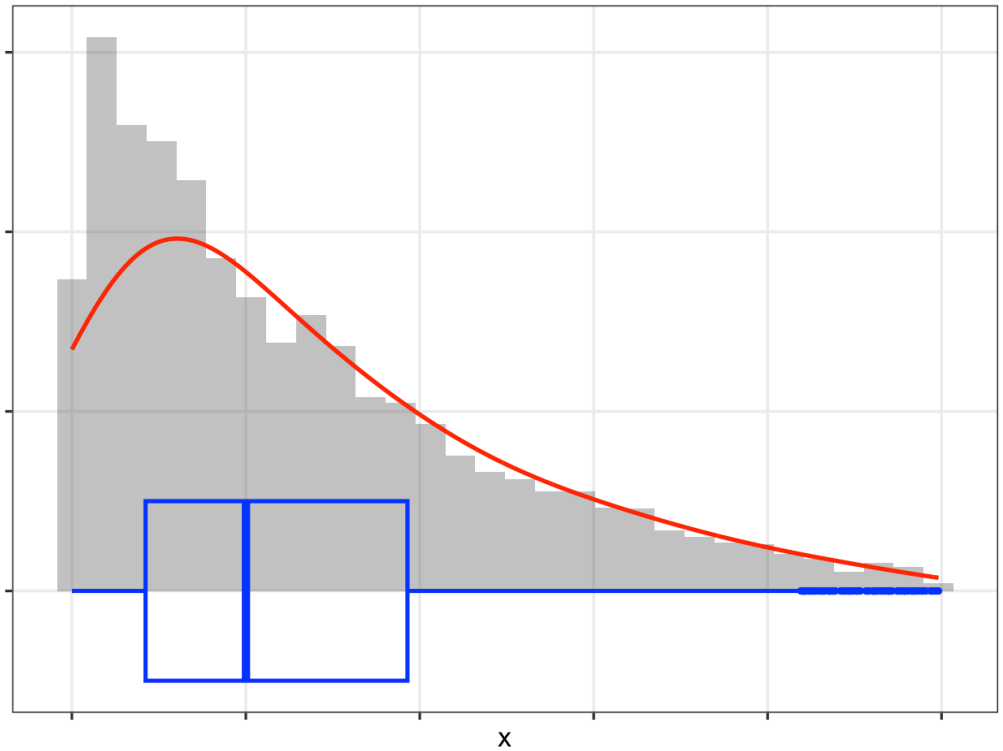


Symmetric distribution

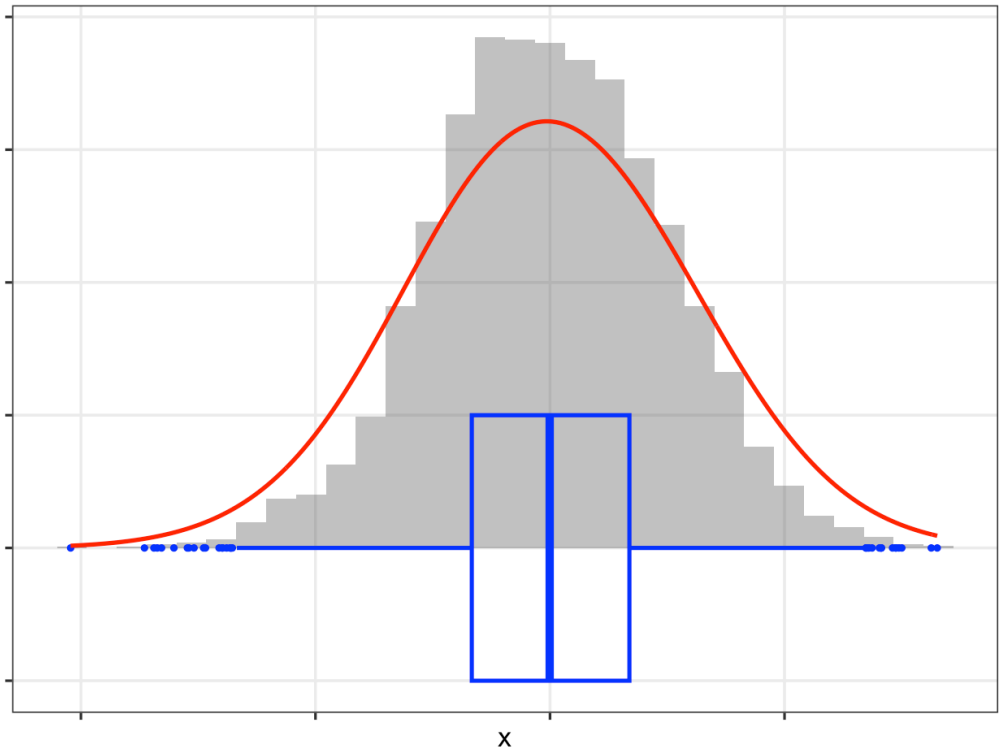


Boxplot and skewness

Positively skewed distribution



Symmetric distribution



Chebyshev's Inequality

- Let $\bar{x}$ and s be sample mean and standard deviation of a data set
- Chebyshev's inequality states that

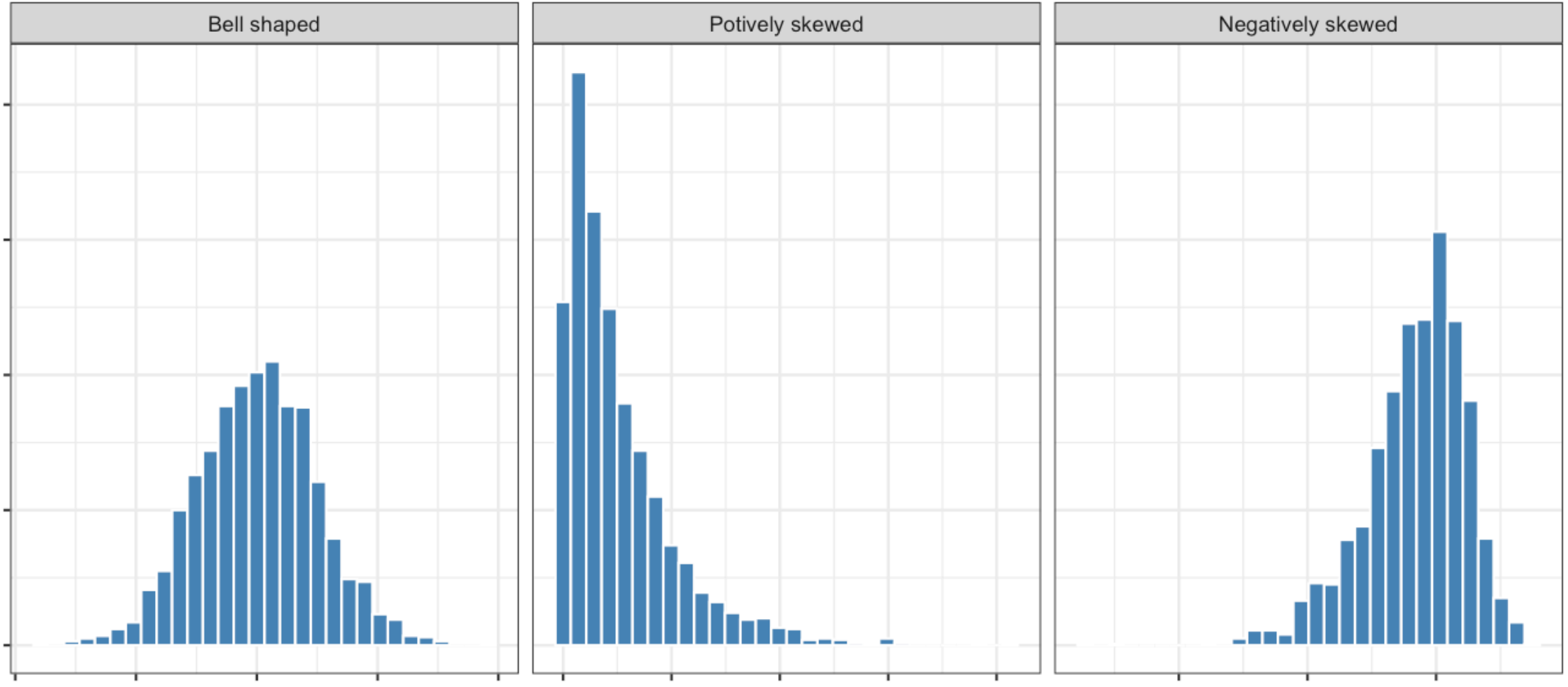
> for any value $k \geq 1$, at least $100(1 - 1/k^2)$ percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$

- For $k = 2$, at least 75% of the data lie within $\bar{x} - 2s$ and $\bar{x} + 2s$

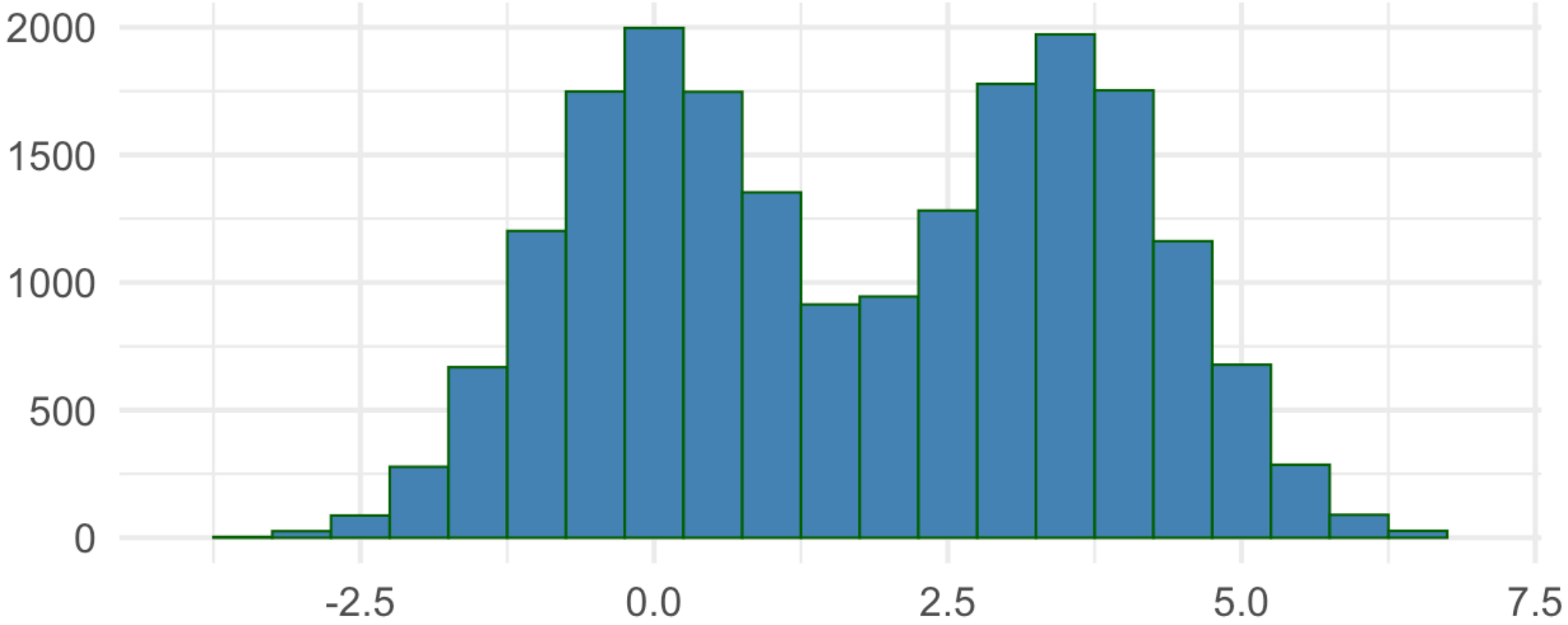
Normal data sets

Normal data sets

- Histograms of normal data sets are bell-shaped, and mean and median of normal data sets are approximately equal
- The statistical theory for analyzing normal distributed data are well established and distributions of many variables (e.g. adult male's height, blood pressure, etc.) are approximately normal
- A data set is said to be skewed if the histogram is not bell-shaped, for skewed data mean and median will not be equal


Bell shaped
Potively skewed
Negatively skewed

Histogram of bimodal data set



Empirical rule

- If a data set is approximately normal with mean $\bar{x}$ and standard deviation s, then the following statements are true:
  - Approximately 68% of the observations lie within
    $$\bar{x} \pm s$$
  - Approximately 95% of the observations lie within
    $$\bar{x} \pm 2s$$
  - Approximately 99.5% of the observations lie within
    $$\bar{x} \pm 3s$$

Paired data sets

Paired data sets

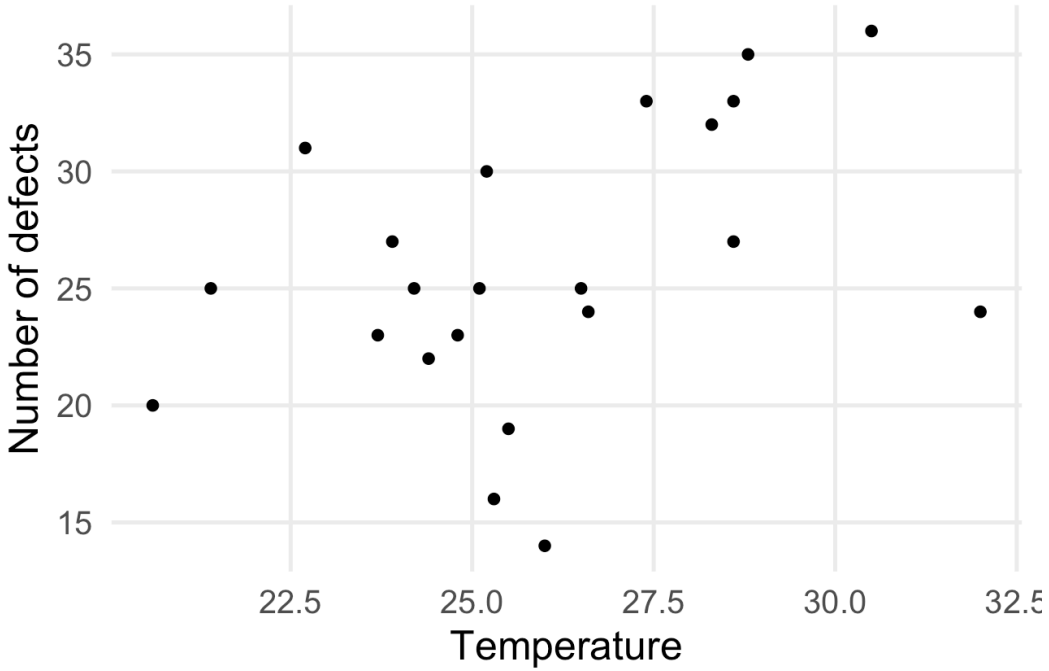
- Sometime pair of values are observed from each element and goal is to examine the relationship between the two
  - E.g. one may be interested in examining whether the daily temperature has any relation with the number of defective items produced per day
  - Paired data $(x_i, y_i),\ i = 1, 2, \ldots, n$

TABLE 2.8 *Temperature and Defect Data*

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

Scaltter plot

- Scatter plot is used to examine the association between two quantitative variables



Sample correlation coefficient

- For the paired data $\{(y_i, x_i), i = 1, \ldots, n\}$, the sample correlation coefficient is defined

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}} = \frac{s_{xy}}{\left[s_{xx}\, s_{yy}\right]^{1/2}}$$

  - $r > 0 \longrightarrow$ the sample data pairs are positively correlated
  - $r < 0 \longrightarrow$ the sample data pairs are negatively correlated

Properties of sample correlation coefficient

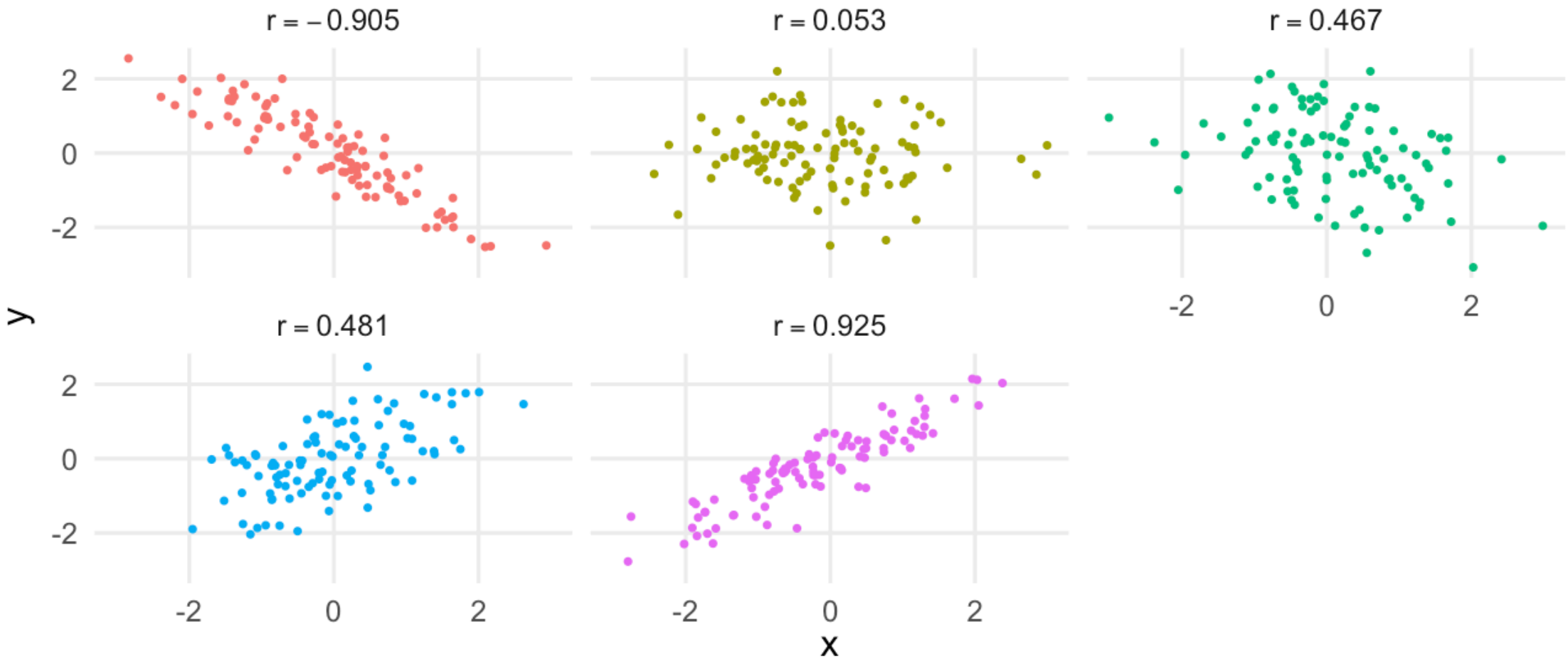
- $-1 \leq r(x, y) \leq 1$
- Let $y_i = a + bx_i$

$$r(x, y) = \begin{cases} +1 & \text{if } b > 0 \\ -1 & \text{if } b < 0 \end{cases}$$

- If r is the sample correlation coefficient between values x_i and y_i, then it is also the sample correlation coefficient between the values

$$a + bx_i \quad \text{and} \quad c + dy_i$$

  - provided the constants b and d have the same sign


r = −0.905
r = 0.053
r = 0.467
r = 0.481
r = 0.925
y
x
2
0
-2

- The following data show the resting pulse rate (in bears per minute) and the years of schooling.
- Obtain the sample correlation coefficient of these two variables.

Person	Years of schooling	Pulse rate
1	12	73
2	16	67
3	13	74
4	18	63
5	19	73
6	12	84
7	18	60
8	19	62
9	12	76
10	14	71

Homeworks

- 1, 6, 7, 8, 14, 16, 26, 30, 36