Introduction
Definitions, Sampling, Measures, Plots

Statistics

Mehmet Güray Güler, PhD

Last updated: 23.02.2021

Definitions

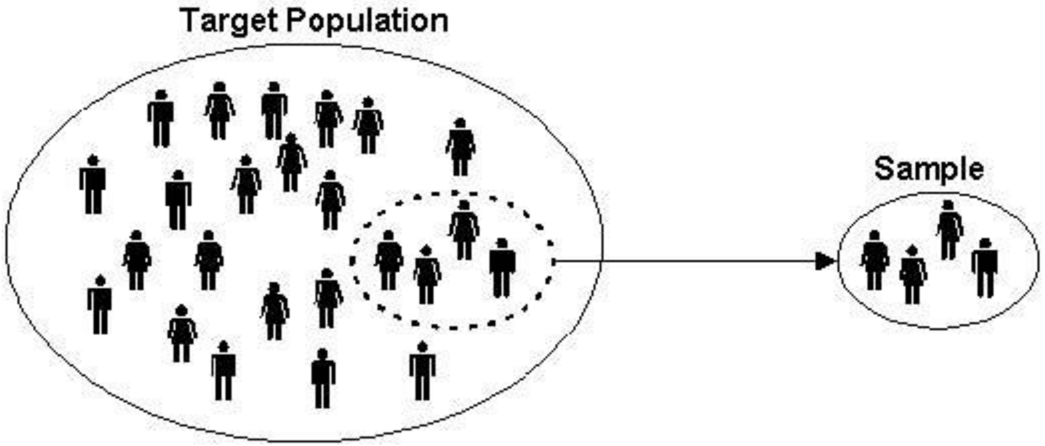
- **Population**: The collection of all individuals or observations of a particular type.
  - All students taking freshman courses at a university.
  - All chips manufactured on a certain day, in a given production process of a factory.
- **Population Size**: The number of all elements in a population.
  - Typically denoted by N.

Definitions

- In practice we will be studying large populations.
- That is, we want to infer or extract information regarding the two important properties:
  - The mean of the population
  - The variance of the population
- It is either not possible or not economical to observe all elements of such a population, recall:
  - The heights of students in the university
  - The diameter of all produced steel valves
- **So how to do it?**

Definitions

- Take a sample!!!
- The sample will (hopefully) have the characteristics of the population
- Their properties will reflect the properties of the population



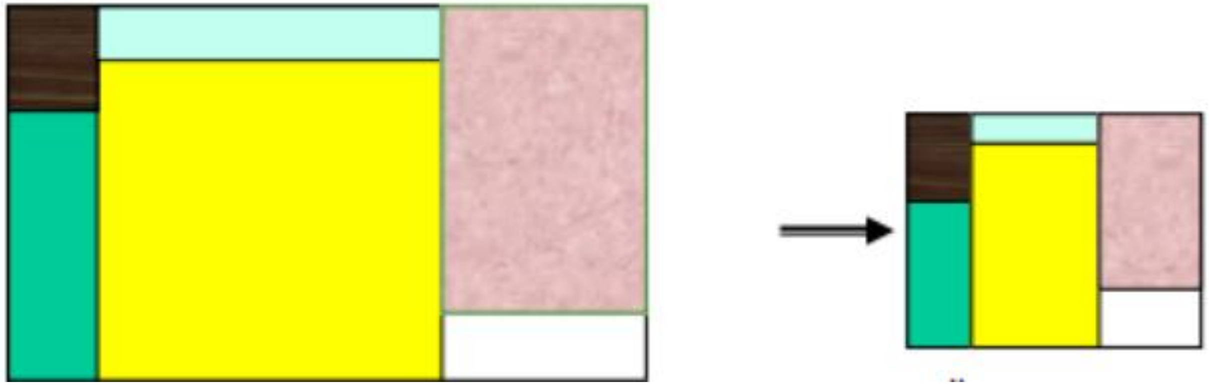
Definitions

- **Sample**: A subset of the population of size ***n***
  - ***Should reflect the population characteristics properly.***
  - Students taking this course can be a sample for this university.

- **1. Biased Sample**:
  - The elements in the sample are not selected at random,
  - therefore the sample may not represent the population accurately.
  - Example?
- **2. Random Sampling.**

Random Sampling

- The important thing in sampling is whether the sample represents the population in terms of the variables subject to research if the population is correctly and clearly defined.
- A good, representative sampling is only a small sample in terms of the number of units of the population, similar in terms of features and model.



Definitions

- A **statistic**
  - any function of the sample data that does not contain unknown parameters
- For the sample we can calculate the following statistics:
  - The sample mean

$$\bar{X} = \frac{\sum_i^n X_i}{n}$$

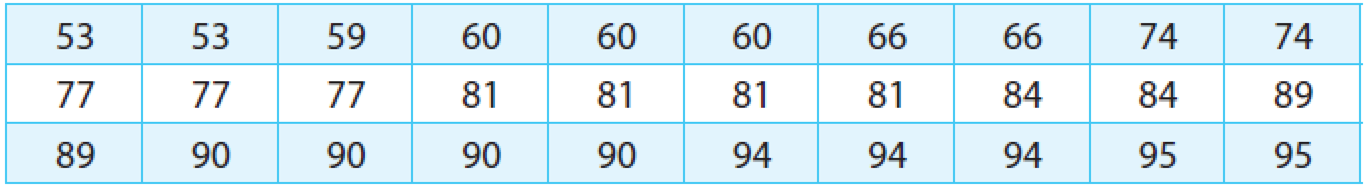
  - The sample variance

$$S^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n-1}$$

Sample Data

Data Series

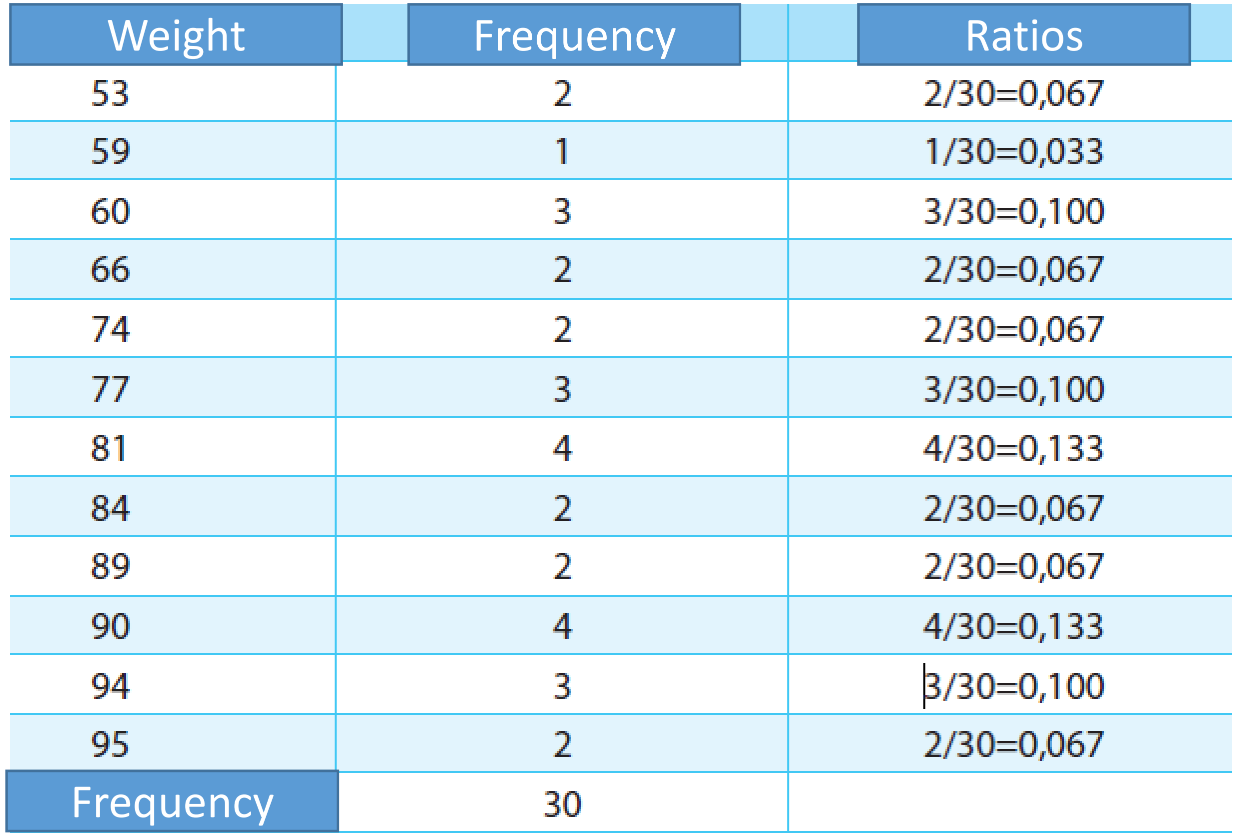
- Dispersion Series
  - **Array**
  - Quantitative
    - Frequency
    - Grouped



53	53	59	60	60	60	66	66	74	74
77	77	77	81	81	81	81	84	84	89
89	90	90	90	90	94	94	94	95	95

Data Series

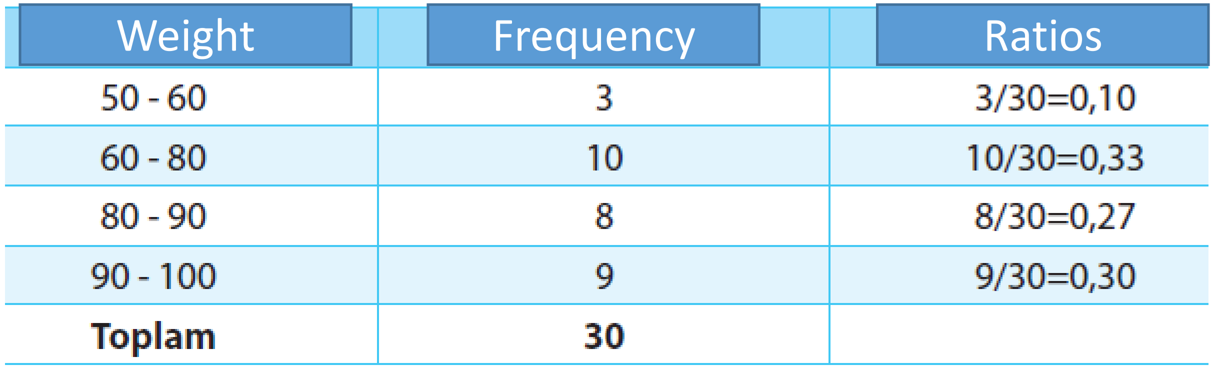
- Dispersion Series
  - Array
  - Quantitative
    - **Frequency**
    - Grouped



Weight	Frequency	Ratios
53	2	2/30=0,067
59	1	1/30=0,033
60	3	3/30=0,100
66	2	2/30=0,067
74	2	2/30=0,067
77	3	3/30=0,100
81	4	4/30=0,133
84	2	2/30=0,067
89	2	2/30=0,067
90	4	4/30=0,133
94	3	3/30=0,100
95	2	2/30=0,067
Frequency	30	

Data Series

- Dispersion Series
  - Array
  - Quantitative
    - Frequency
    - **Grouped**



Weight	Frequency	Ratios
50 - 60	3	3/30=0,10
60 - 80	10	10/30=0,33
80 - 90	8	8/30=0,27
90 - 100	9	9/30=0,30
Toplam	**30**	

Measures of Centrality and Dispersion

Measures for Centrality and Dispersion

- Two very very important measures for data:
  - Where is the center?
    - Averages
  - How does the data scatter around the center?
    - Deviations

Averages

- **Definition:**
    - Categories or scores that describe what is average or characteristics of the distribution.
- Types
    - Arithmetic average
    - Geometric average
    - Harmonic average
    - Median
    - Mode

Arithmetic Mean

- Sample

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Weighted Average

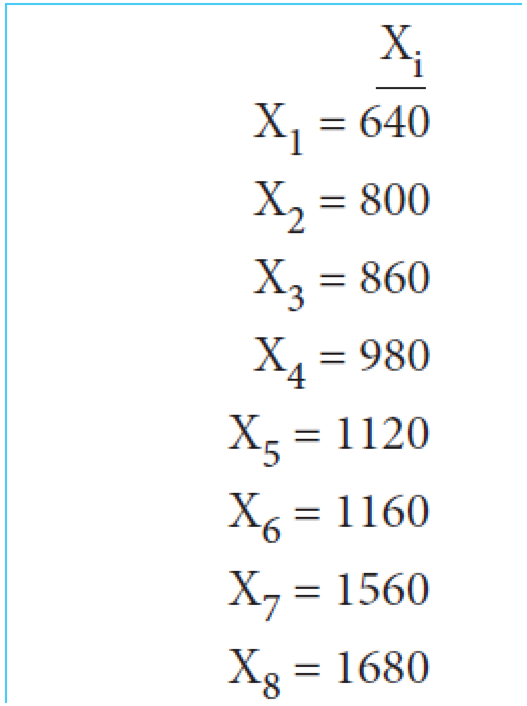
$$\overline{X}_t = \frac{\sum X_i t_i}{\sum t_i}$$

- Frequency series

$$\bar{X} = \frac{x_1 f_1 + x_2 f_2 + \; \dots + x_k f_k}{f_1 + f_2 + \; \dots + f_k} = \frac{\sum_i^k x_i f_i}{\sum_i^k f_i}$$

Arithmetic Mean

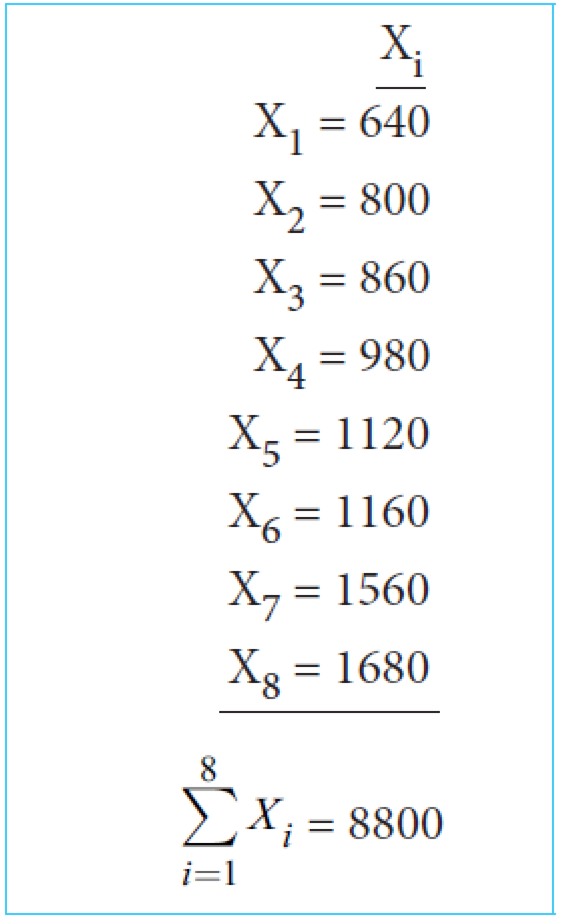
- The average income of 8 workers are given in the following.
- Find the average income.



X_i
$X_1 = 640$
$X_2 = 800$
$X_3 = 860$
$X_4 = 980$
$X_5 = 1120$
$X_6 = 1160$
$X_7 = 1560$
$X_8 = 1680$

Arithmetic Mean

- The average income of 8 workers are given in the following.
- Find the average income.
- $\bar{X} = \frac{8800}{8} = 1100$



$$\underline{X_i}$$

$$X_1 = 640$$

$$X_2 = 800$$

$$X_3 = 860$$

$$X_4 = 980$$

$$X_5 = 1120$$

$$X_6 = 1160$$

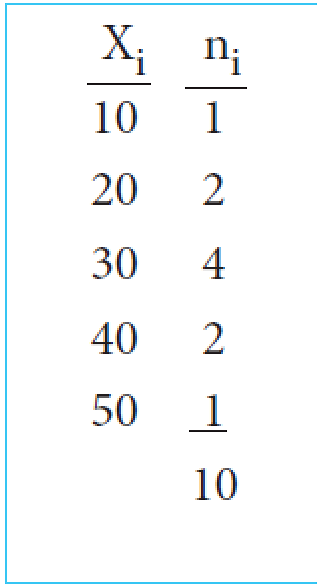
$$X_7 = 1560$$

$$X_8 = 1680$$

$$\sum_{i=1}^{8} X_i = 8800$$

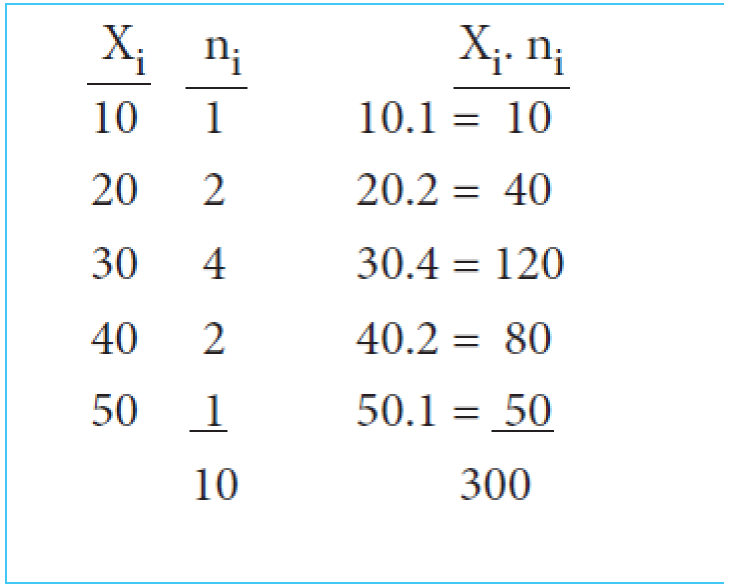
Arithmetic Mean

- ***Frequency Series:*** Parcels of a company have the following frequency table. Find the mean parcel weight.



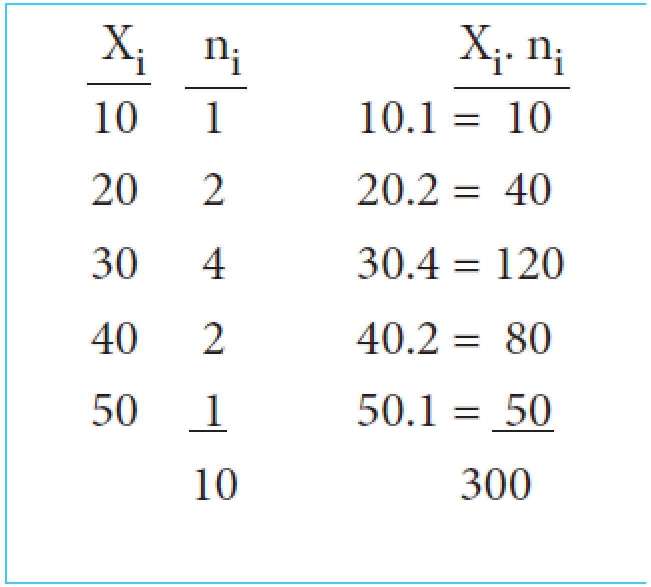
X_i	n_i
10	1
20	2
30	4
40	2
50	1
	10

Arithmetic Mean



X_i	n_i	$X_i . n_i$
10	1	10.1 = 10
20	2	20.2 = 40
30	4	30.4 = 120
40	2	40.2 = 80
50	1	50.1 = 50
	10	300

Arithmetic Mean

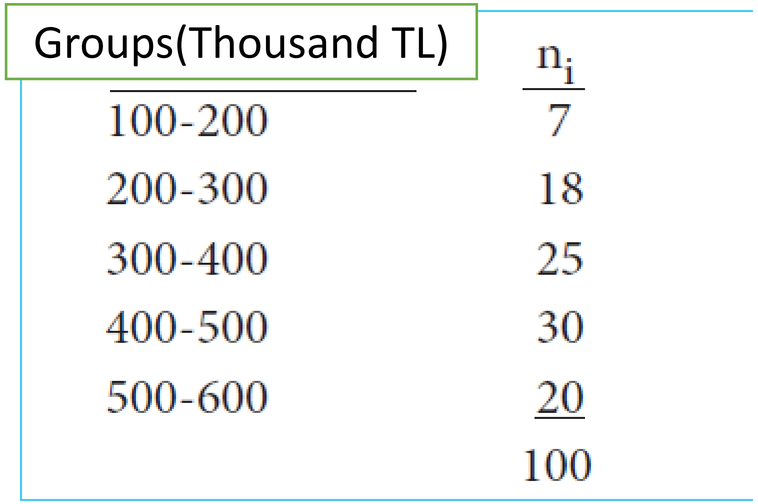


X_i	n_i	$X_i . n_i$
10	1	10.1 = 10
20	2	20.2 = 40
30	4	30.4 = 120
40	2	40.2 = 80
50	1	50.1 = 50
	10	300

$$\overline{X} = \frac{\sum_{i=1}^{10} X_i n_i}{\sum_{i=1}^{10} n_i} = \frac{300}{10} = 30 \text{ Kg}$$

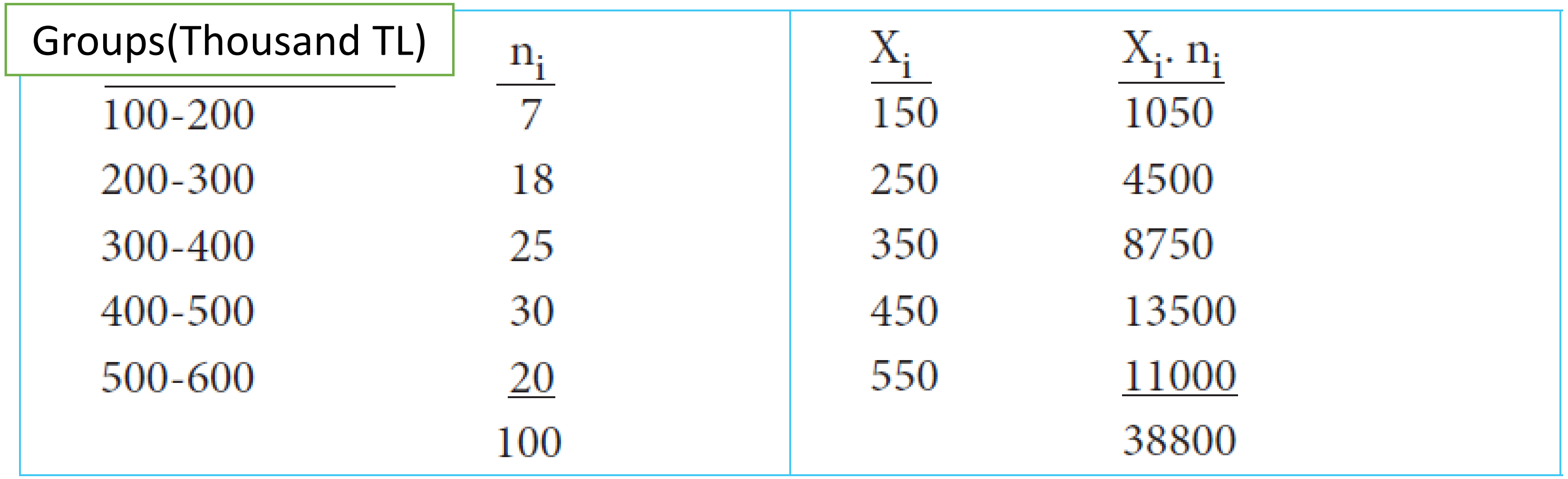
Arithmetic Mean

- ***Grouped Series:*** What is the average tax paid by 100 companies?



Groups(Thousand TL)	n_i
100-200	7
200-300	18
300-400	25
400-500	30
500-600	20
	100

Arithmetic Mean



Groups(Thousand TL)	n_i	X_i	$X_i . n_i$
100-200	7	150	1050
200-300	18	250	4500
300-400	25	350	8750
400-500	30	450	13500
500-600	20	550	11000
	100		38800

$$\overline{X} = \frac{38800}{100} = ₺\ 388$$

Median

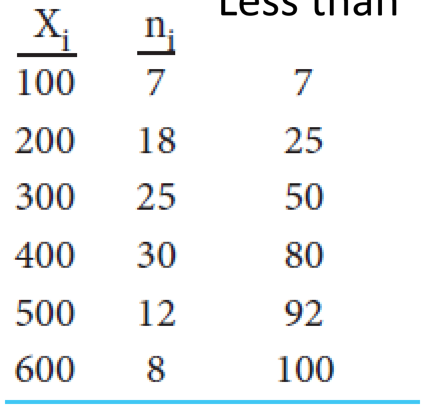
- **(Sample) Median:**
- the ordered statistics (from smallest to largest) by: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$
- Then the sample median is also a statistic, defined by:

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)} \right) & \text{if } \boldsymbol{n} \text{ is even.} \end{cases}$$

Median

- **Definition**:
  - Order the values.
    - İf n is odd, then the middle value is the median
    - If n is even, then the average of two middle values is the median.
  - : 1,6,3,5,2,8,5
  - : 1,35,22,54,3
  - : 1,35,22,5400,3

Median



X_i	n_i	Less than
100	7	7
200	18	25
300	25	50
400	30	80
500	12	92
600	8	100

Median

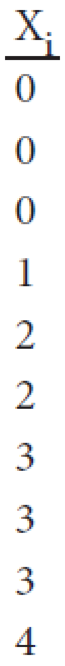
X_i	n_i	Less than
100	7	7
200	18	25
300	25	50
400	30	80
500	12	92
600	8	100

The values are sorted from smallest to largest. The number of observations is 100. Then the median is:

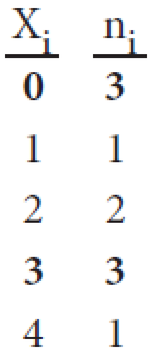
$$Med = \frac{X_{50} + X_{51}}{2} = \frac{300 \quad + \quad 400}{2} = 350 \text{ Ton}$$

(Sample) Mode

- Definition:
  - The most frequent observation in a series
  - Can be found by converting to a frequency series
- **Example**: 3, 2, 0, 0, 2, 3, 3, 1, 0, 4
- Two modes here: bi modal



X_i
0
0
0
1
2
2
3
3
3
4



X_i	n_i
0	**3**
1	1
2	2
3	**3**
4	1

Averages – Summary

- The most used average is
  - Arithmetic mean
  - Median
- Median and mode are **insensitive** averages, i.e., they are insensitive to observations.
- Arithmetic mean is **sensitive** to the observations.
- https://www.mathsisfun.com/data/frequency-grouped-mean-median-mode.html

Dispersion Measures

Variability (Dispersion) Measures

- DEFINITION. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population.
- Let $X_{\max} = X_{(n)}$ be the largest of these sample values and $X_{\min} = X_{(1)}$ be the smallest of them.
- The sample range is a simple measure of the spread (variability) of the data, defined by:
- $R = X_{\max} - X_{\min}$ or $R = X_{(n)} - X_{(1)}$
- 180, 192, 175, 167, 188 → 167, 175, 180, 188, 192 → 192-167 = 25

Variability (Dispersion) Measures

- The sample variance is a statistic, defined by:

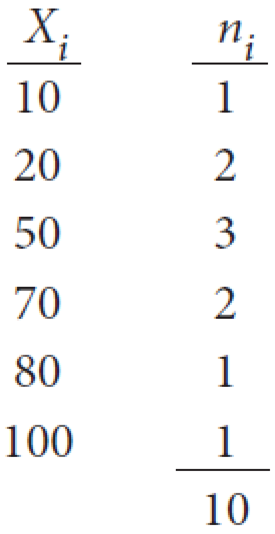
$$S^2 = \frac{1}{n - 1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$= \frac{1}{n(n - 1)}\left[n\sum_{i=1}^{n}X_i^2 - (\sum_{i=1}^{n}X_i)^2\right]$$

- The first form is the common definition, the second form is easier to use for hand calculations.
- Sample standard deviation

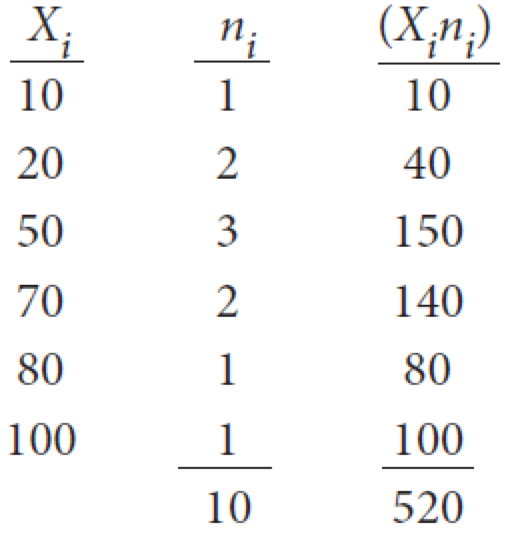
Standard Deviation

- **Frequency Series**
- **Ex:** Daily consumption of flour in a breakery is given in the following as a frequency series. Find its standard deviation.



X_i	n_i
10	1
20	2
50	3
70	2
80	1
100	1
	10

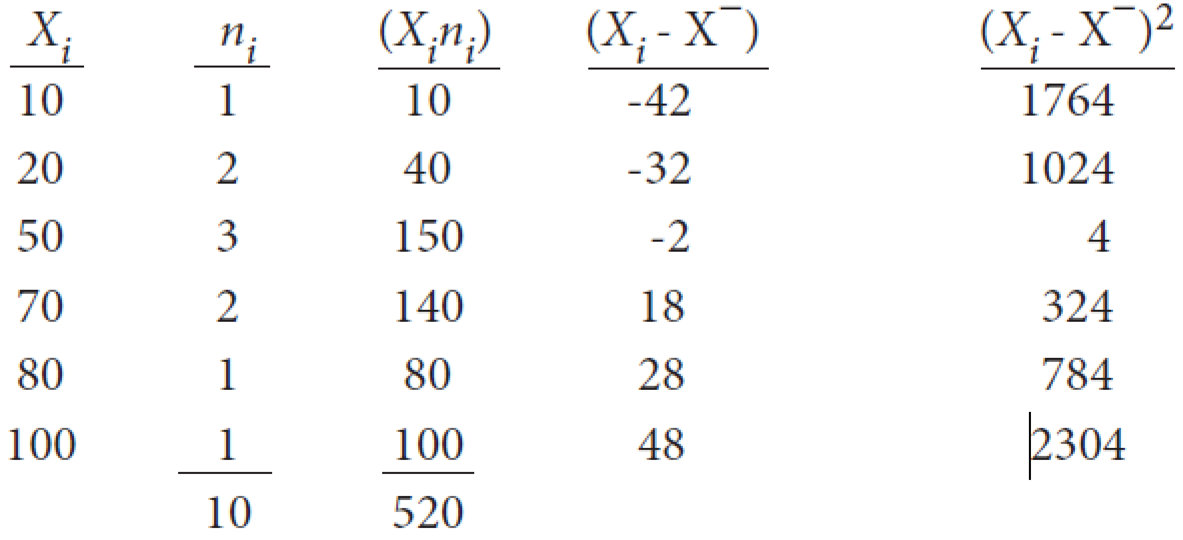
Standard Deviation



X_i	n_i	(X_in_i)
10	1	10
20	2	40
50	3	150
70	2	140
80	1	80
100	1	100
	10	520

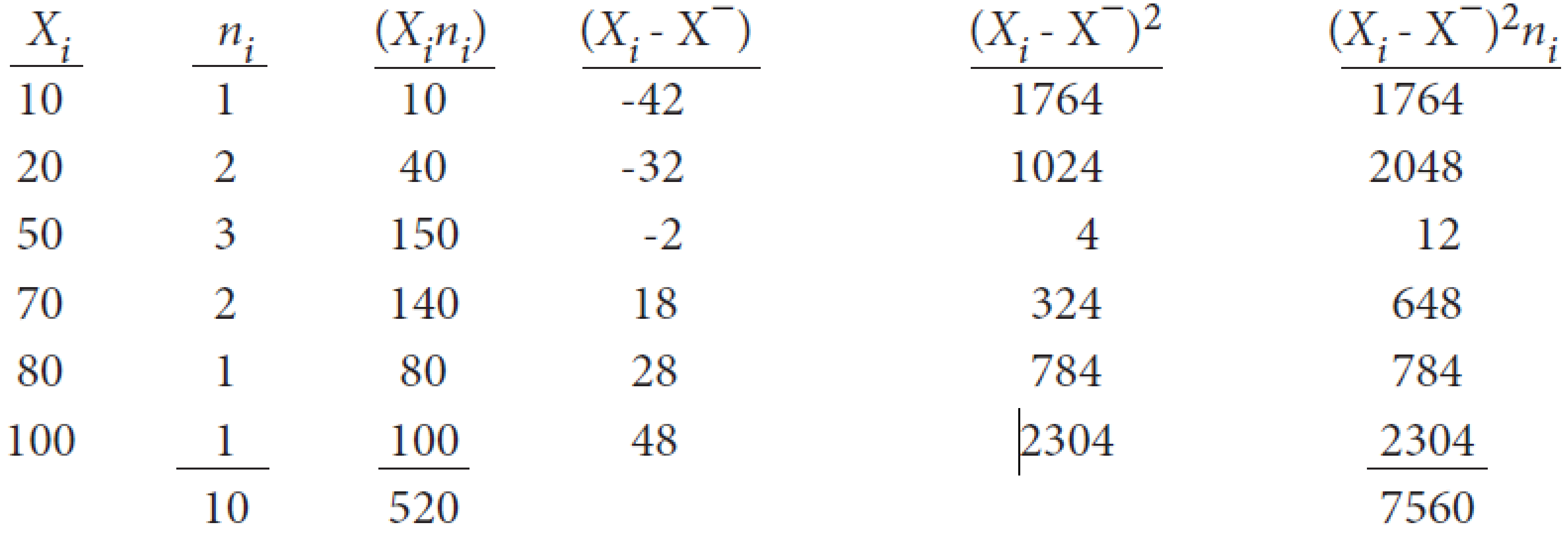
$$\bar{X} = \frac{520}{10} = 52$$

Standard Deviation



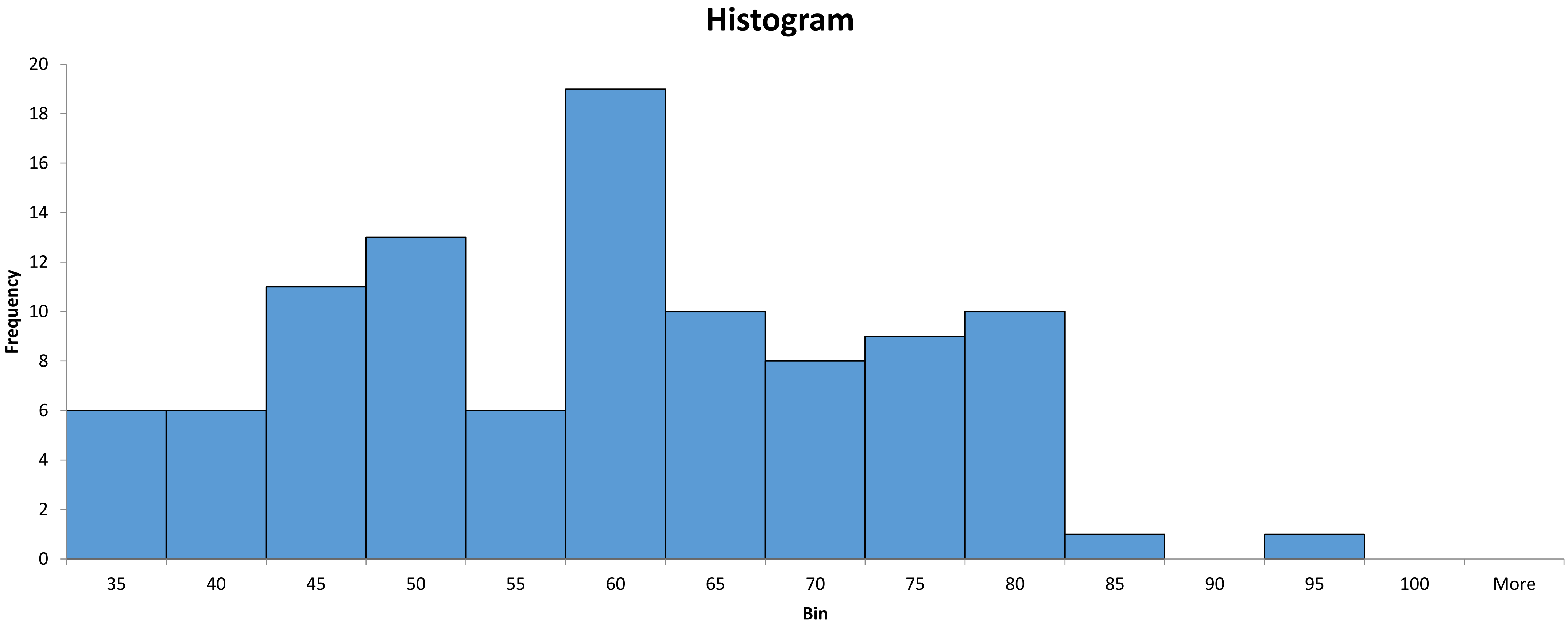
X_i	n_i	$(X_i n_i)$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
10	1	10	-42	1764
20	2	40	-32	1024
50	3	150	-2	4
70	2	140	18	324
80	1	80	28	784
100	1	100	48	2304
	10	520		

Standard Deviation

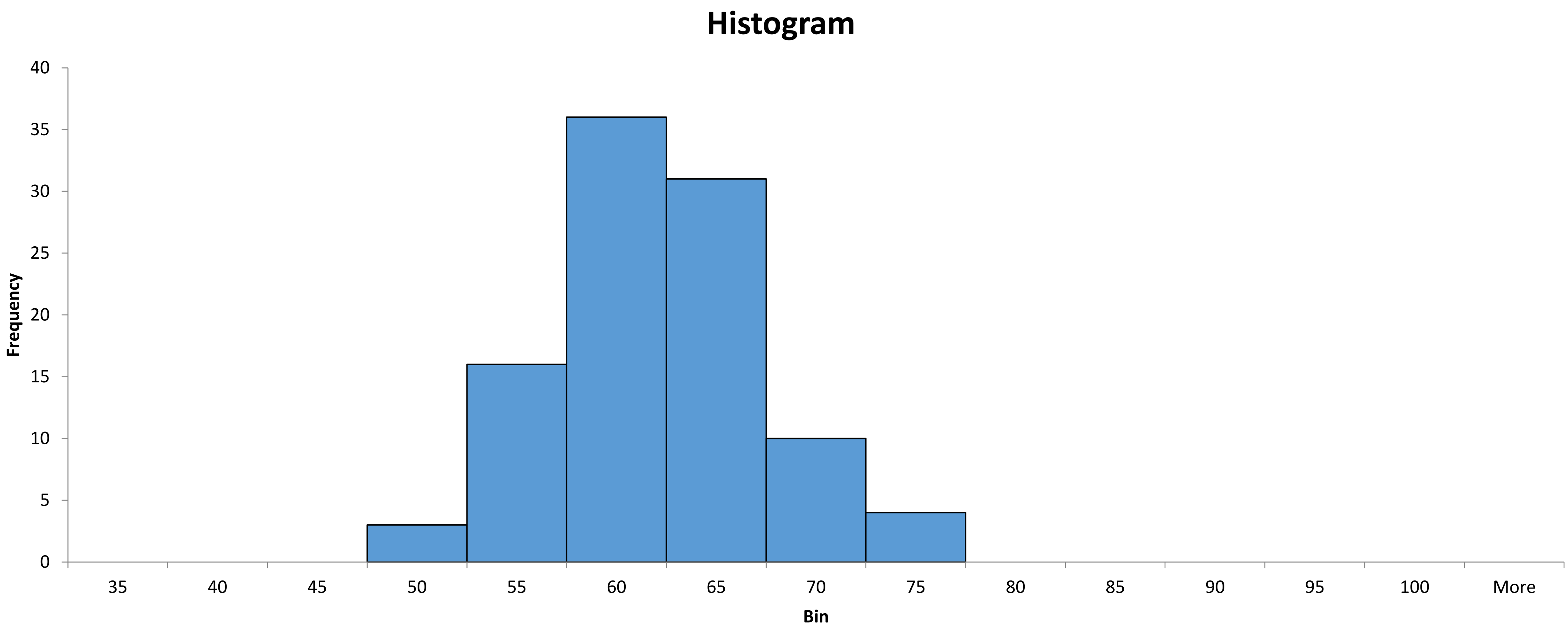


X_i	n_i	$(X_i n_i)$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 n_i$
10	1	10	-42	1764	1764
20	2	40	-32	1024	2048
50	3	150	-2	4	12
70	2	140	18	324	648
80	1	80	28	784	784
100	1	100	48	2304	2304
	10	520			7560

High Variance



Low Variance



EXAMPLE 1

Find the sample mean, the sample median, the quartiles, and the sample standard deviation for the data set:

3, 5, 3, 7, 6, 6, 8, 5, 7, 4, 4, 6, 9, 5, 7.

The sample mean is simply average of these 15 numbers:

TOTAL = 3+5+3+…+7 = 85 ➔ MEAN = 85/15 ≈ 5.67

For **the sample median and the quartiles** we first order the sample values from smallest to largest:

3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7, 8, 9.

EXAMPLE 1

The sample median and the quartiles:

Ordered data: 3, 3, 4, 4, 5, 5, 5, **6**, 6, 6, 7, 7, 7, 8, 9

Since $\boldsymbol{n}=15$ is odd, the median is the middle observation, this is the 8th observation in the ordered set

Hence **MEDIAN** = 6.

Lower quartile Q_1 is the median of the lower half of data:

LOWER HALF: 3, 3, 4, **4**, 5, 5, 5 ➔ Q_1 = 4.

Upper quartile Q_3 is the median of the upper half of data:

UPPER HALF: 6, 6, 7, **7**, 7, 8, 9 ➔ Q_3 = 7.

EXAMPLE 1

To find the sample standard deviation for this data set, we first calculate the two important statistics, namely

$$\sum_{i=1}^{n} X_i \quad \text{and} \quad \sum_{i=1}^{n} X_i^2$$

For 3, 5, 3, 7, 6, 6, 8, 5, 7, 4, 4, 6, 9, 5, 7 we have

$$\sum x_i = 85 \quad \text{and} \quad \sum x_i^2 = 525$$

EXAMPLE 1

Hence the **sample variance** is

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2\right]$$

$$= \frac{1}{(15)(14)}\left[15\cdot 525 - (85)^2\right] => s^2 = 3.095.$$

and the sample standard deviation = $\boldsymbol{s = 1.76.}$

Note that the sample range is $\boldsymbol{R = 9 - 3 = 6.}$

We can also find the **interquartile range** as $\mathbf{Q_3 - Q_1 = 7 - 4 = 3.}$

Numerical Summary of Data

	Sample1	Sample2	Sample3
	3	1	103
	5	5	105
	7	9	107
$\bar{X}$	**5**	**5**	**105**
S			

Numerical Summary of Data

	Sample1	Sample2	Sample3
	3	1	103
	5	5	105
	7	9	107
$\bar{X}$	**5**	**5**	**105**
S	**2**	**4**	

Numerical Summary of Data

	Sample1	Sample2	Sample3
	3	1	103
	5	5	105
	7	9	107
$\bar{X}$	**5**	**5**	**105**
S	**2**	**4**	**2**

The standard deviation does not reflect the magnitude of the sample data, only the scatter about the average

Coefficient of variation (CV)

- deviation per unit average
- $DK = \frac{s}{\bar{X}}$
- $DK = \frac{\sigma}{\mu}$
- Sometimes multiplied by 100.

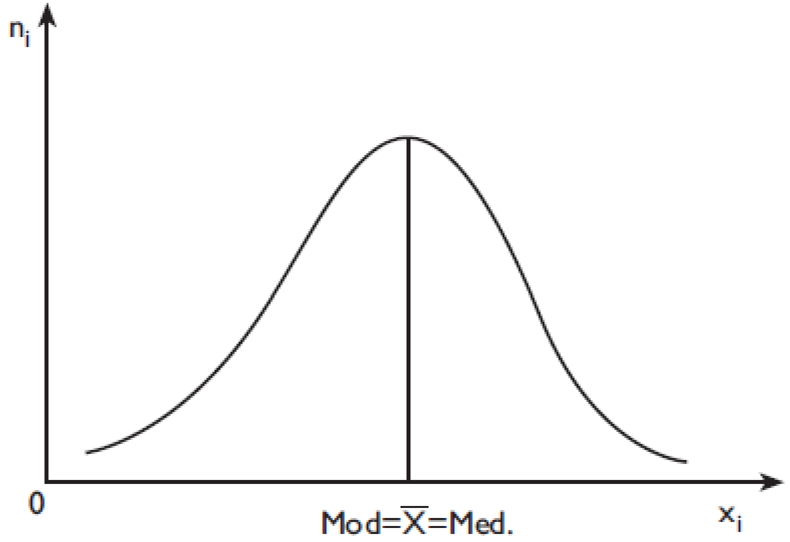
Coefficient of variation (CV)

- **Ex:**
- Stock A and Stock B's yearly return has the following average and standard deviation values. What can you say about their deviation?
- A: 20% average, 5% SD
- B: 40% average, 20% SD

- CV(A) = 0.25 = 25%
- CV(B) = 0.50 = 50%
- Stock B has a higher relative risk.

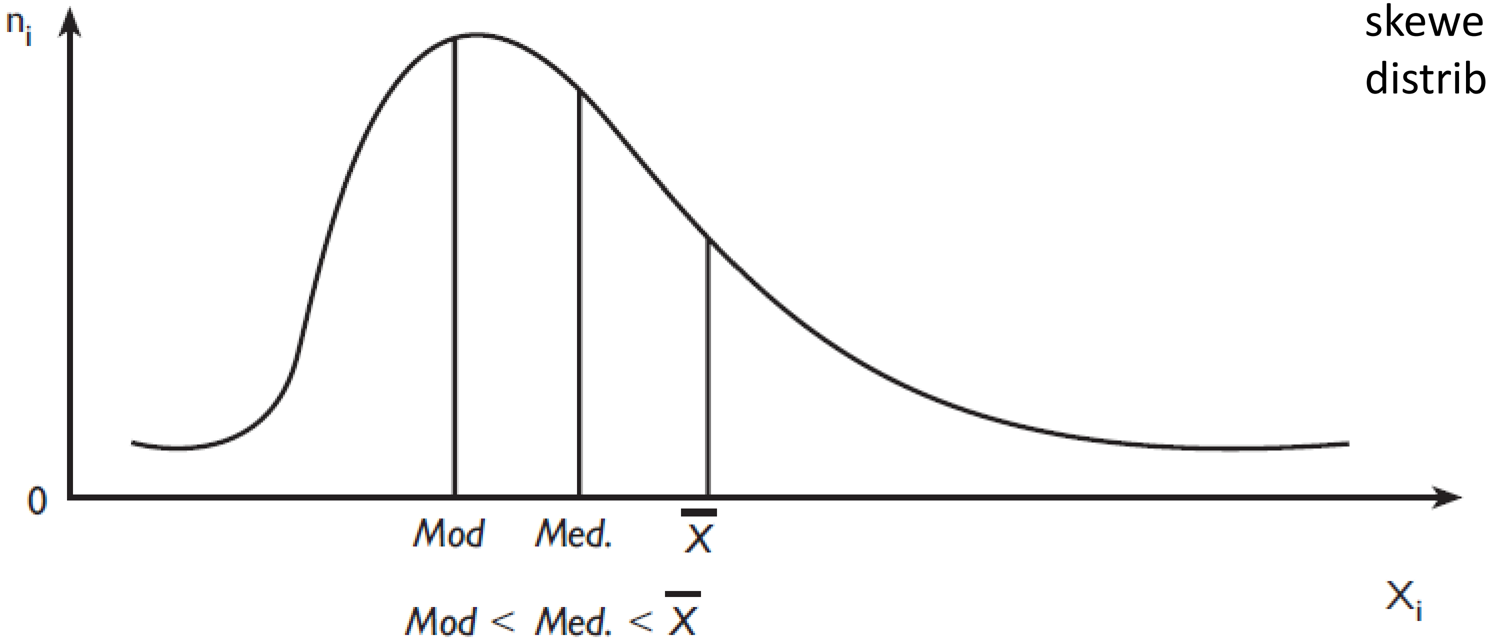
Shape of a distribution

- We have covered two important measures of a data set:
  - Average
  - Dispersion
- Shape is another important measure.



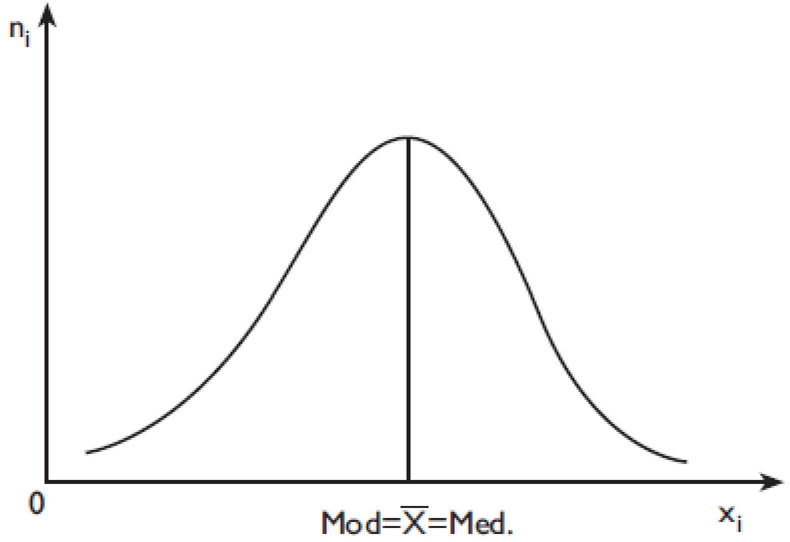
$\overline{X}$=172cm = Med=172cm= Mod=172cm

Positively skewed distribution



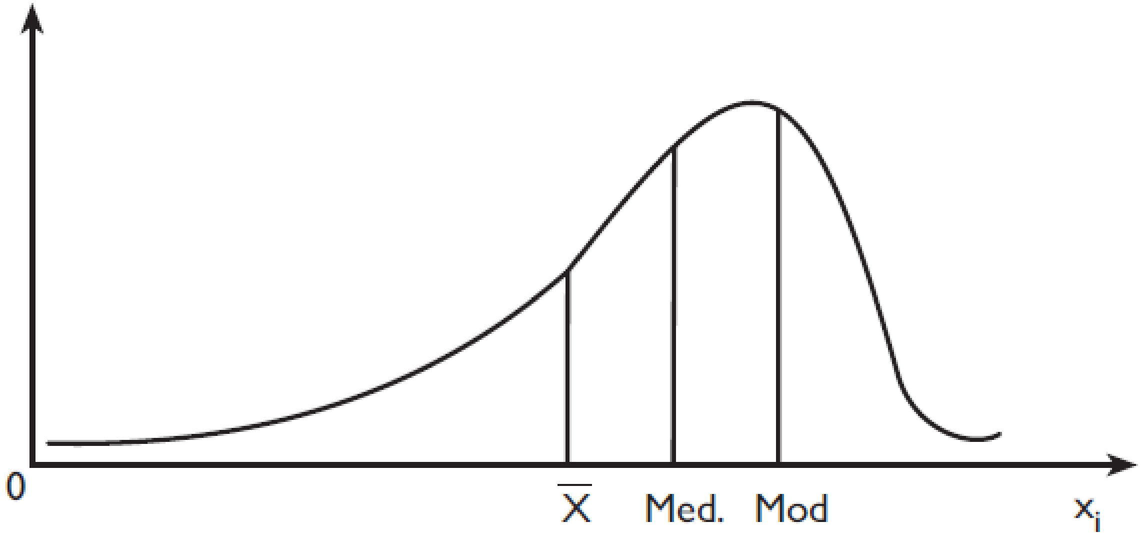
Mod < *Med.* < $\overline{X}$

Mod=165<Med=169<$\overline{X}$ =172



$\overline{X}$=172cm = Med=172cm= Mod=172cm

Negatively (left) skewed



Mod<$\overline{X}$<Med.

Data Plots

Example 2

- The interior temperature of a drying oven has been measured every 15 minutes for the duration of one production cycle and the following data are obtained:
- 56 46 48 50 42 43 49 48 56 50 52 47 48 56 41 37 47 49 45 44 40 55 45 44 50 45 44 64 48 48 32 40 52 43 51 59 63 59 47 38 50 49 40 54 46 51 48 54 49 45 50 56 44 52 37 61
- As such, the data set is a confusing list of numbers.

Table 1. Frequency Distribution of Oven Temperature Data

Frequency distribution of temperature data

Interval	***Frequency***	***Cumulative Count***	***Percent***
30 up to 35	1	1	1.8
35 up to 40	3	4	5.5
40 up to 45	11	15	20.0
45 up to 50	18	33	32.7
50 up to 55	12	45	21.8
55 up to 60	7	52	12.7
60 up to 65	3	55	5.5

NOTE: Lower interval limit is included and the upper limit is excluded.

Figure 2. Histogram of Temperature Data

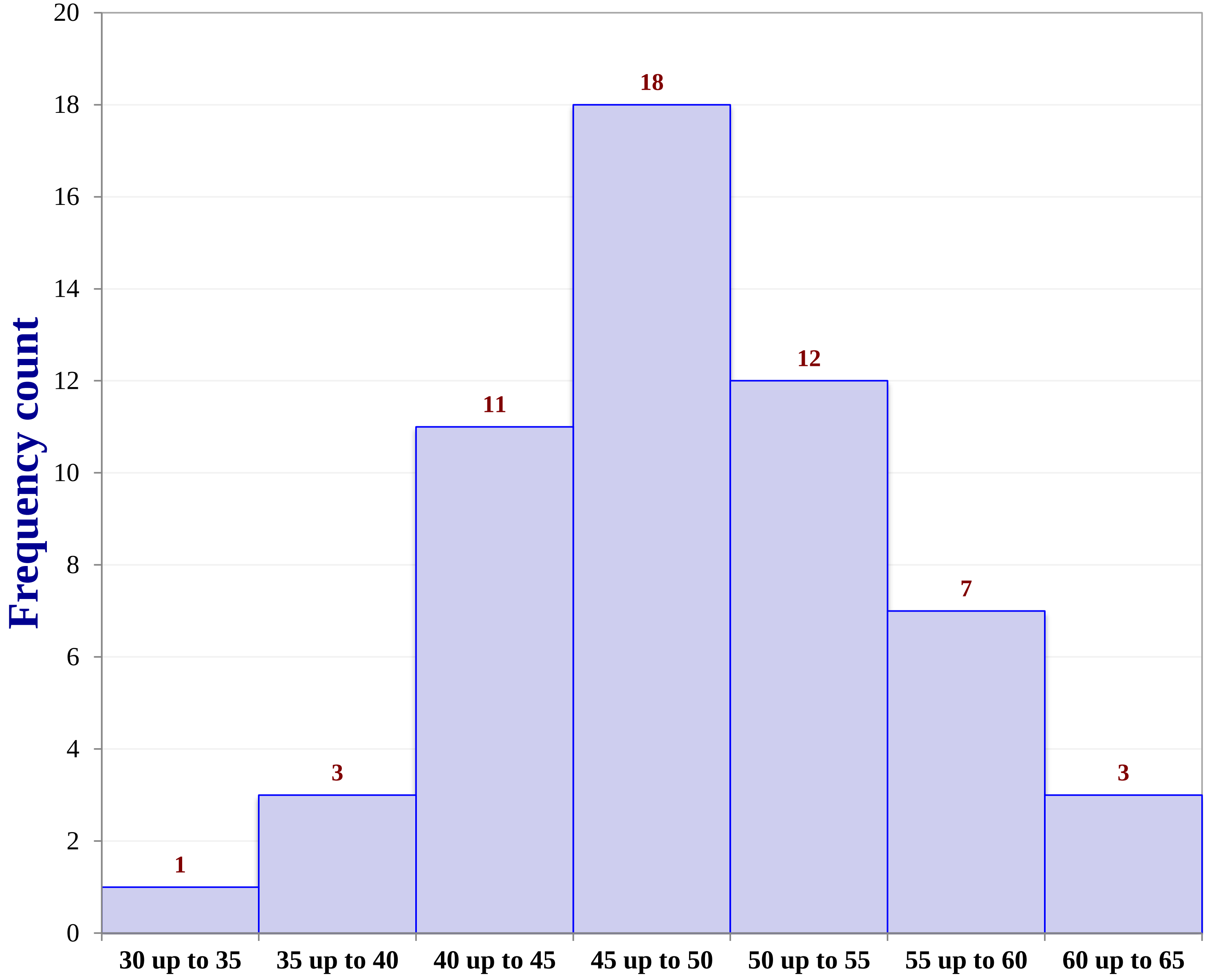


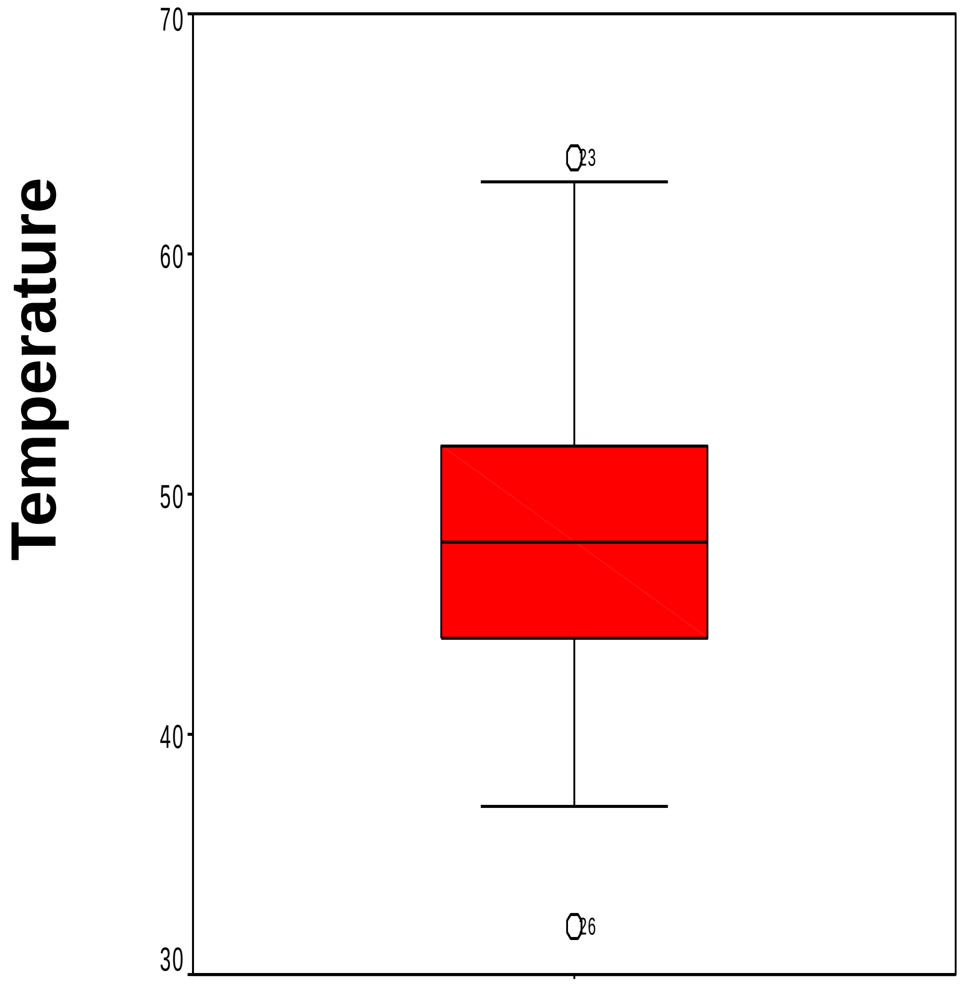
Figure 3. Stem-and-Leaf Diagram for the Oven Temperature Data

3	2
3+	7 7 8
4	0 0 0 1 2 3 3 4 4 4 4
4+	5 5 5 6 6 6 7 7 7 8 8 8 8 8 8 9 9 9 9
5	0 0 0 0 0 1 1 2 2 2 4 4
5+	5 6 6 6 6 9 9
6	1 3 4

This looks like a horizontal histogram, but it shows more details than the histogram.

Stem-and-leaf diagram is suitable for data sets that are not very large.

Figure 4. Box-Plot for the Oven Temperature Data



Box-plot uses five important summary statistics:

Maximum

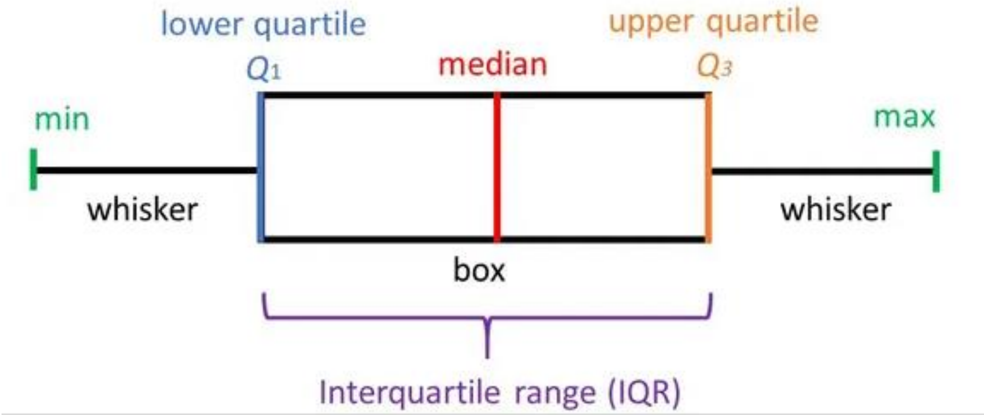
Upper Quartile,

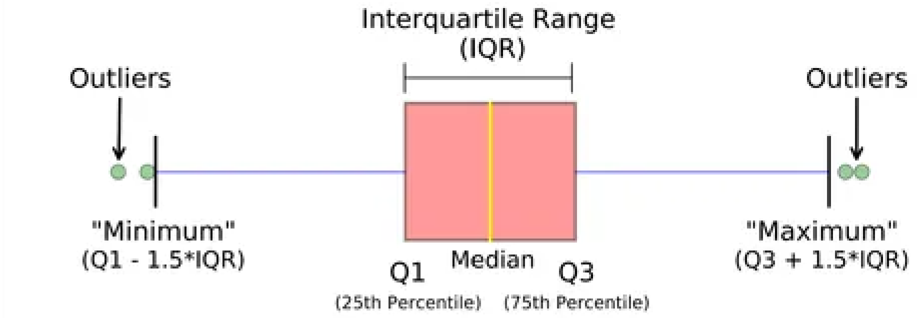
Median,

Lower Quartile,

Minimum,

$\bar{x} = 48.3$ and $s = 6.6.$


lower quartile
Q_1
median
upper quartile
Q_3
min
max
whisker
whisker
box
Interquartile range (IQR)


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)

Definitions

- Recall examples for a RV:
  - Number of customers
  - The income
- Again, two important properties for a random variable.

- ***Where is the center?*** Generally we use the following measure:
  - $\mu = E[X]$ =Mean or expected value

- ***How the data is scattered around center?*** Generally we use the following measure
  - $\sigma^2 = E[(X-\mu)^2]$ = Variance
- We will cover both of them in details…

Oops! What about the sample mean and variance and those formulas?

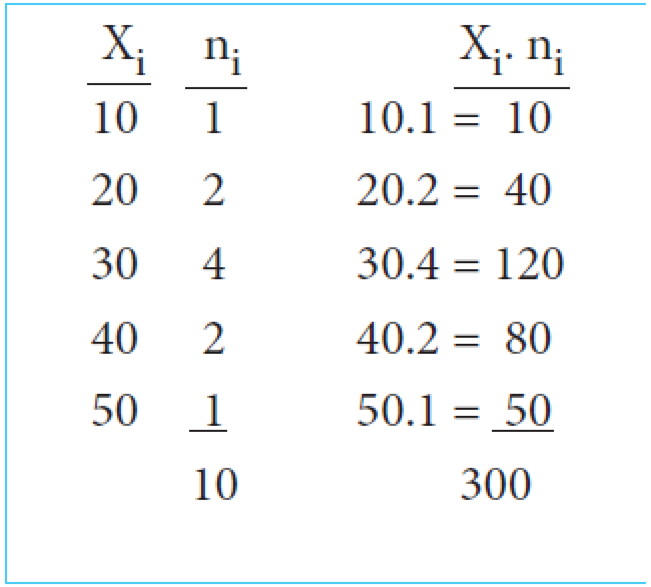
- $E[X] = \int xf(x)dx$ $\qquad$ $E[X] = \sum xf(x)$

- $\mathrm{Var(X)} = \int (x-\mu)^2 f(x)dx$ $\qquad$ $\mathrm{Var(X)} = \sum (x-\mu)^2 f(x)$

$$\bar{X} = \frac{\sum_i^n X_i}{n}$$

$$S^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n-1}$$

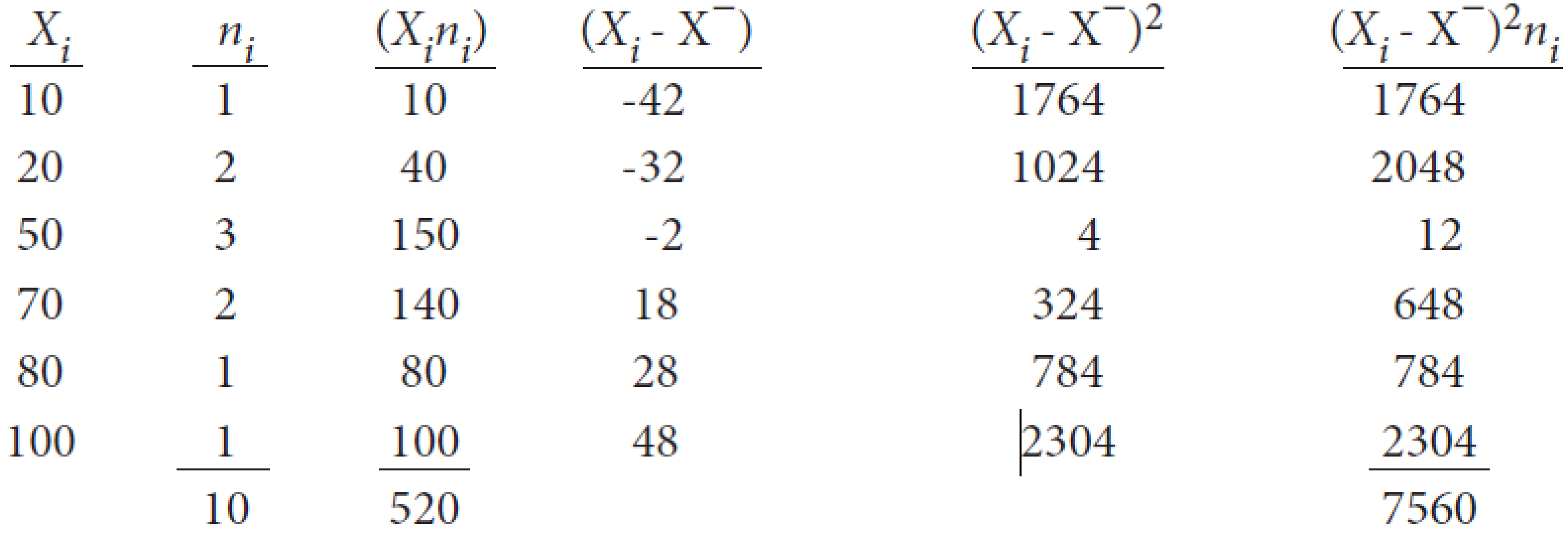
Arithmetic Mean - Remember



X_i	n_i	$X_i . n_i$
10	1	10.1 = 10
20	2	20.2 = 40
30	4	30.4 = 120
40	2	40.2 = 80
50	1	50.1 = 50
	10	300

$$\overline{X} = \frac{\sum_{i=1}^{10} X_i n_i}{\sum_{i=1}^{10} n_i} = \frac{300}{10} = 30 \text{ Kg}$$

Standard Deviation – Remember

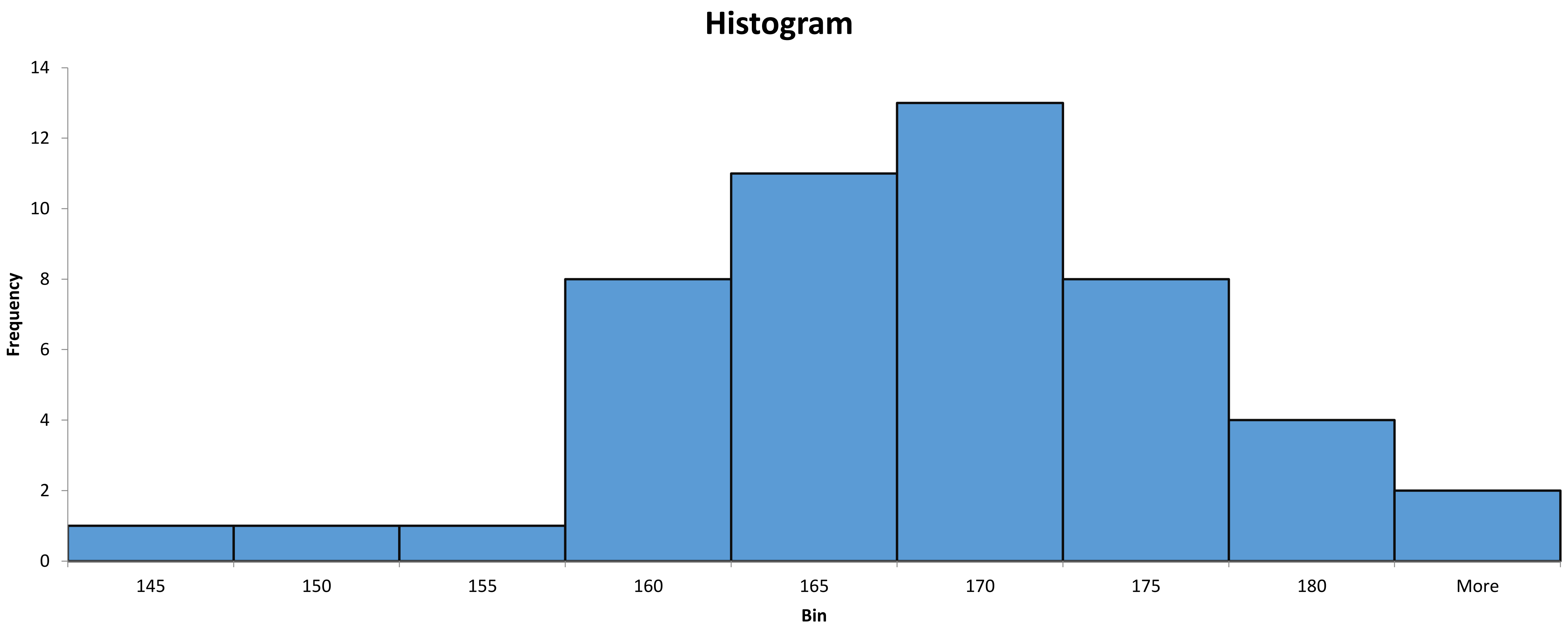


X_i	n_i	$(X_i n_i)$	$(X_i - X^{-})$	$(X_i - X^{-})^2$	$(X_i - X^{-})^2 n_i$
10	1	10	-42	1764	1764
20	2	40	-32	1024	2048
50	3	150	-2	4	12
70	2	140	18	324	648
80	1	80	28	784	784
100	1	100	48	2304	2304
	10	520			7560

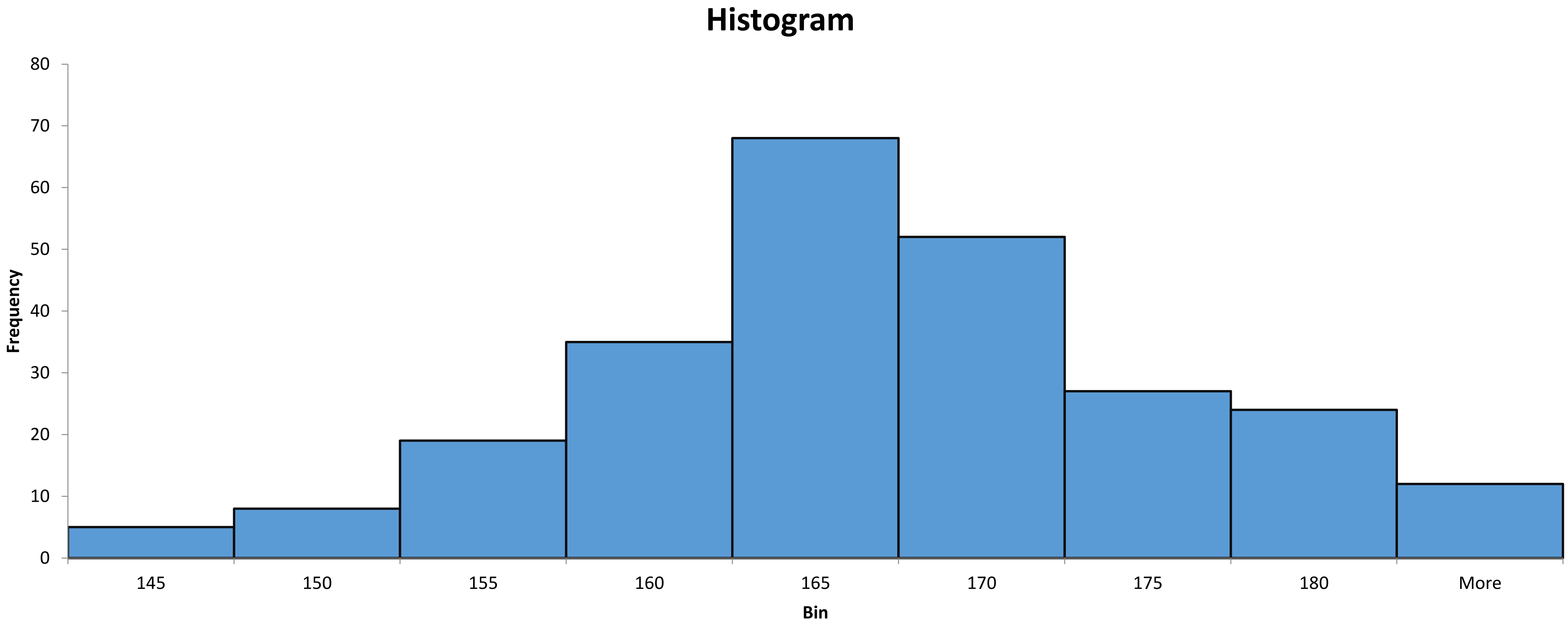
About a die

Distribution Function and Histogram

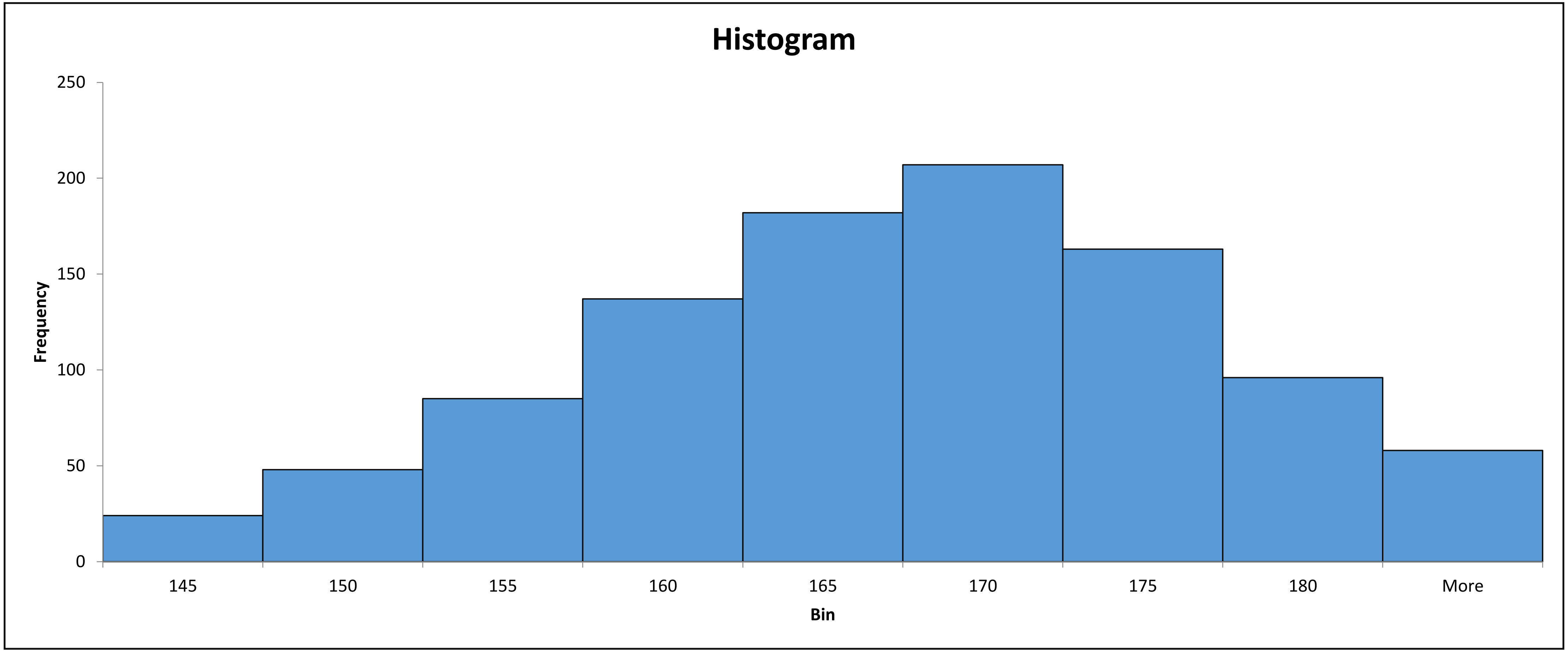
Histogram – 50 data



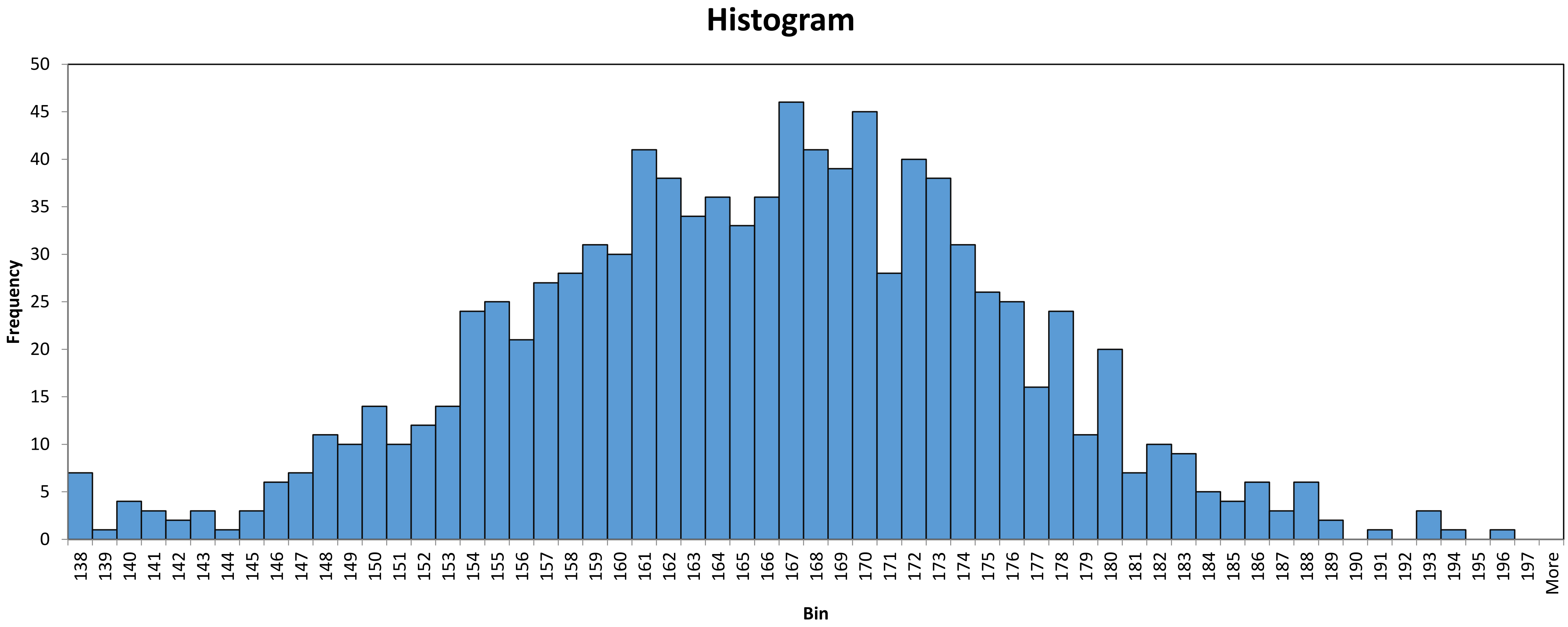
Histogram – 250 Data



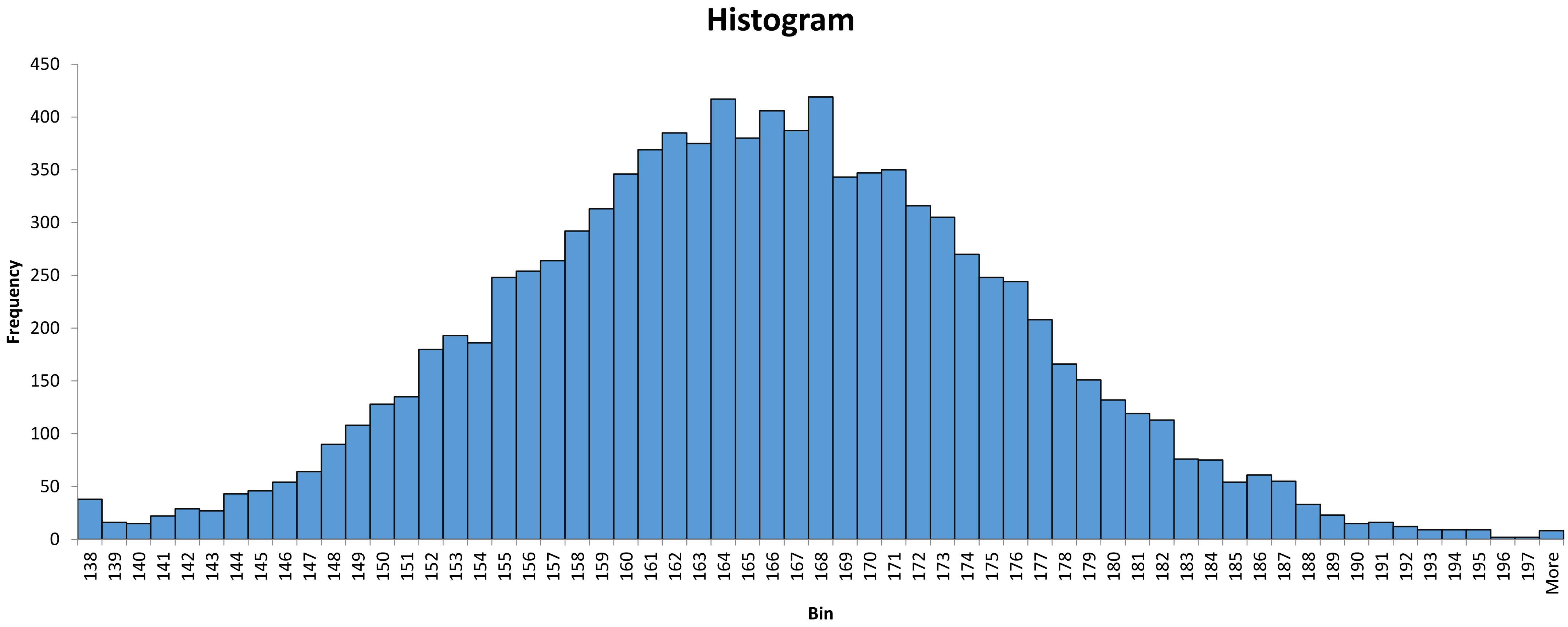
Histogram – 1000 Data



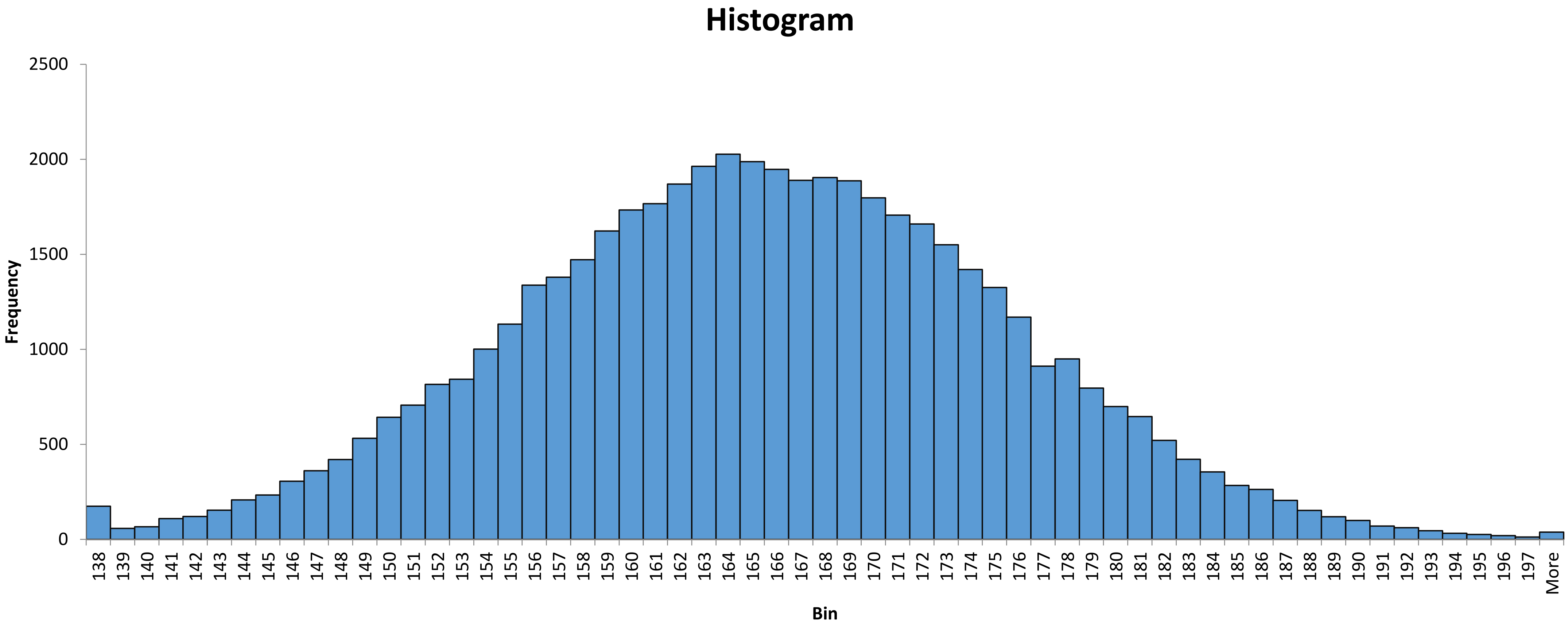
Histogram – 1000 data, thinner bins



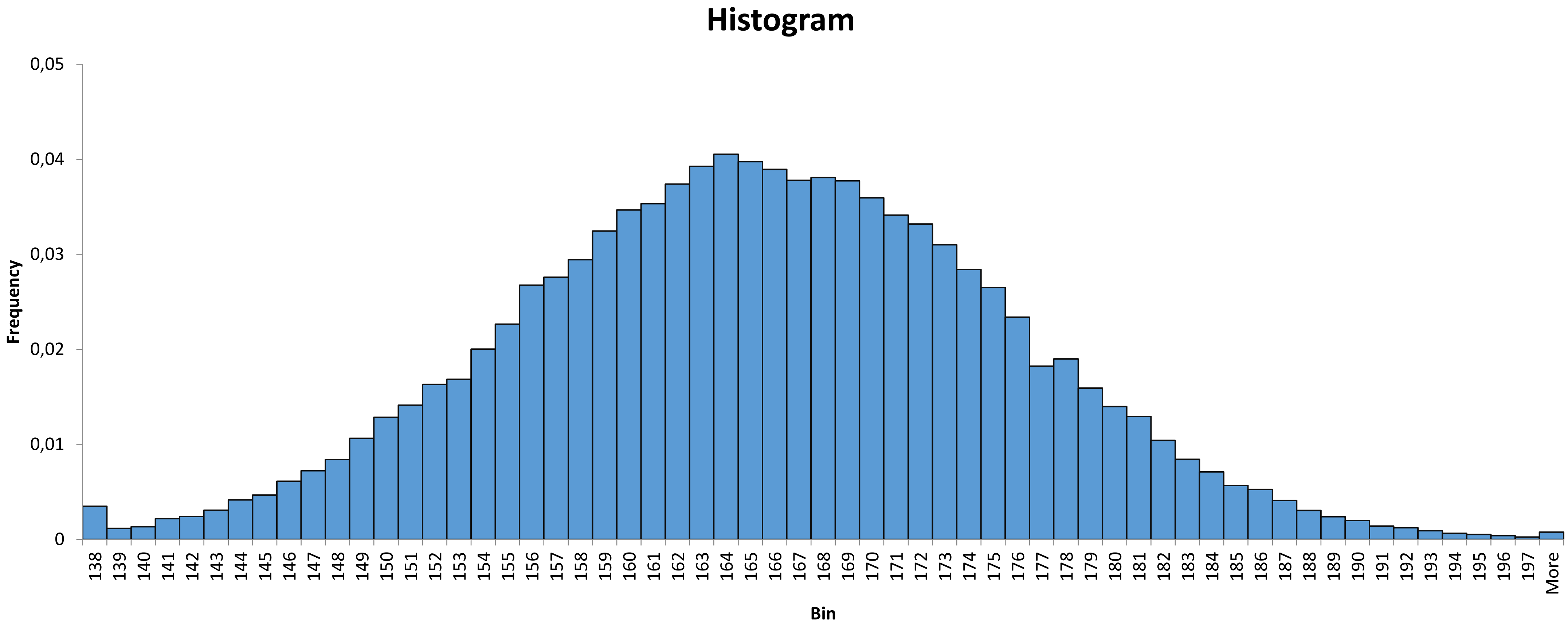
Histogram – 10.000 data



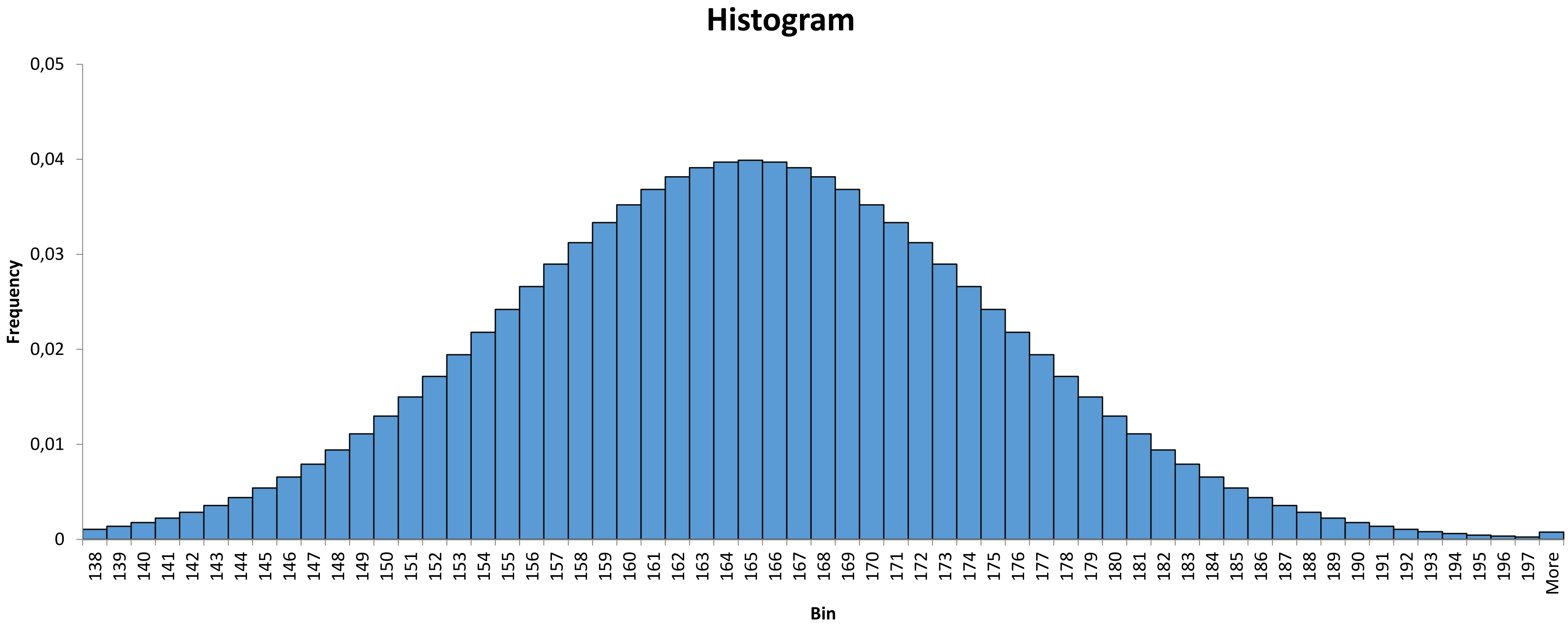
Histogram – 100.000 data



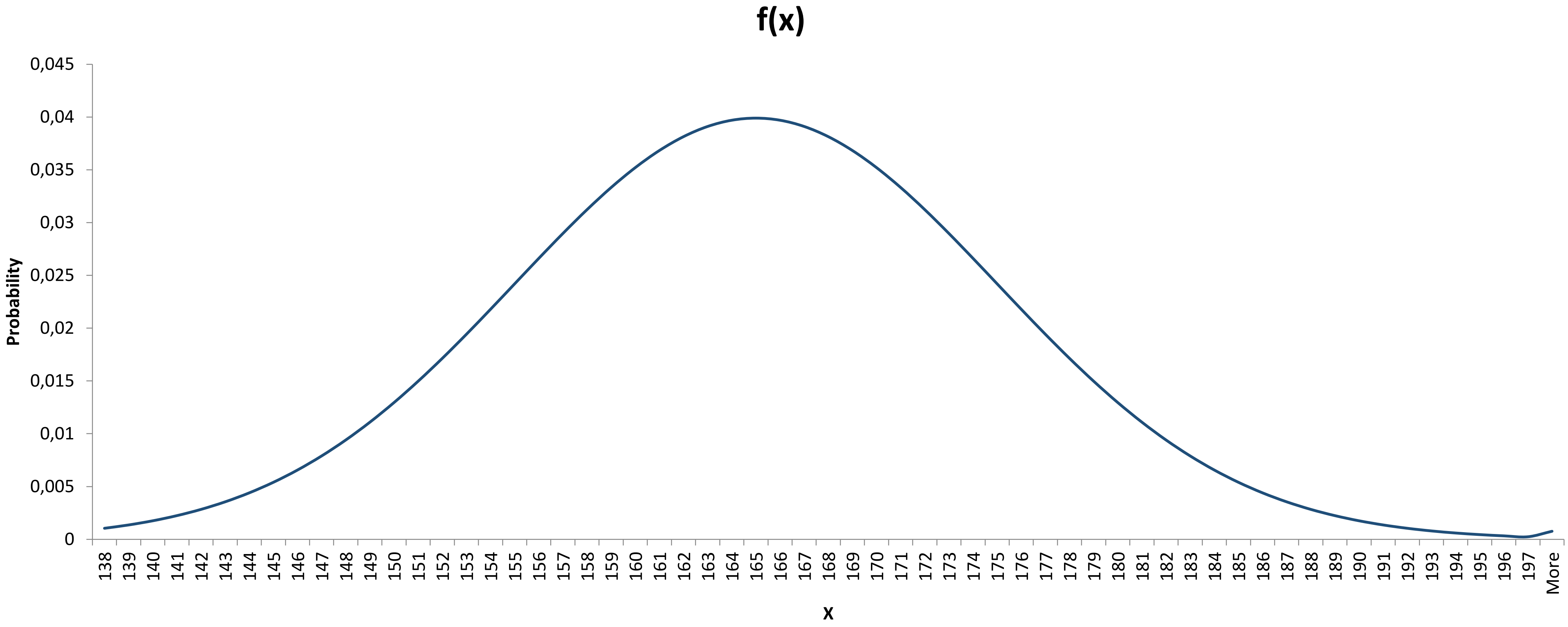
Histogram – 100.000 data - %



Histogram –∞ data - %



Distribution function



Definitions

- The whole story is this…

	Mean	**Variance**
Population	μ	σ^2
Sample	$\bar{X}$	S^2

Central Limit Theorem

Sample mean $\bar{X}$

- Note that $\bar{X}$ is a RV:

$$\bar{X} = \frac{\sum_i^n X_i}{n}$$

- Hence it has
  - a mean
  - a variance
  - a distribution
- In order to infer (extract) information for the unknown μ, we will need all of them.

Sample mean $\bar{X}$

- First find $E[\bar{X}] = \mu_{\bar{X}}$
- Recall:
- If $X_1, X_2, \ldots and\ X_n$ are independent RVs with μ_i and σ_i^2 as their expected values and variances, then

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

- Hence

$$E[\bar{X}] = \mu$$

- **What does this mean? Think of a student just coming into the class..**

Sample mean $\bar{X}$

- Then find $var(\bar{X})$
- Recall:
- If $X_1, X_2, \ldots and\ X_n$ are independent RVs with μ_i and σ_i^2 as their expected values and variances, then

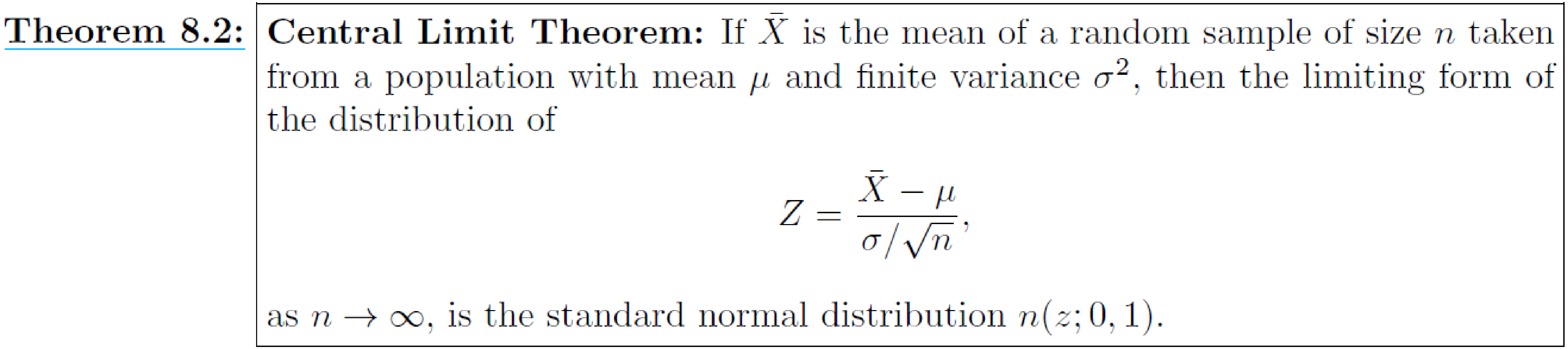
$$\sigma^2_{a_1X_1+a_2X_2+\ \ldots+a_nX_n} = a_1^2\sigma_1^2 + a_1^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

- Hence

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Sample mean $\bar{X}$

- Finally, what is the distribution of $\bar{X}$?
- We have a great result



Theorem 8.2: **Central Limit Theorem:** If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

Probability Reminder

- Random variables:
  - Die toss
  - Coin Toss
  - Number of customers in bank
- Discrete
  - Binomial
  - Poission

- Random variables
  - The height of a person
  - The income of a worker
  - The lifetime of a lamp
- Continuous
  - Exponential
  - Normal
    - **extremely important**
    - **Statistics is based on normality.**

Probability Reminder

- RVs are denoted by capital letters, X
- Small letters show a value of RV. For example x=7.
- $P(X \geq 7)$
    - Probability that the RV X is greater than or equal to the scalar 7
- How to calculate the probabilities?
    - We are lucky, scientists have developed the probability functions for different types of random variables.

Probability Review (Normal Distribution)*

- If X is normal with $\mu = 10 \ and \ \sigma = 2$?

$$P(X \leq 7) = \int_0^7 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\frac{1}{2}(x-\mu)^\wedge 2}{\sigma^\wedge 2}}$$

- Hard to find the integral!!!
- How to deal with it?
  - Standardize
  - Use tables

*an extended version is in another slide set.

Probability Review

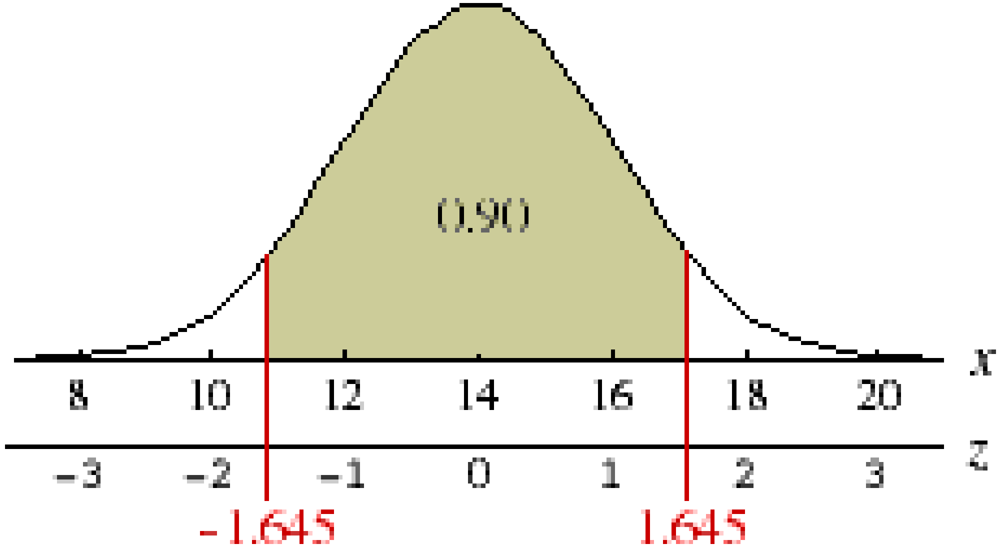
- Z:
  - Standard normal RV
  - Normal RV with $\mu = 0$ and $\sigma = 1$

$$P(X \leq 7) = P\left(\frac{X-\mu}{\sigma} \leq \frac{7-\mu}{\sigma}\right) = P\left(Z \leq \frac{7-10}{2}\right) = P(Z \leq -1.5)$$

- You can plot this!
- We can check the table to find the answer.

Probability Review

- A different notation
- $z_{0.05}$:
  - The z value that leaves an area of 0.05 (5% the right.
  - 1.645



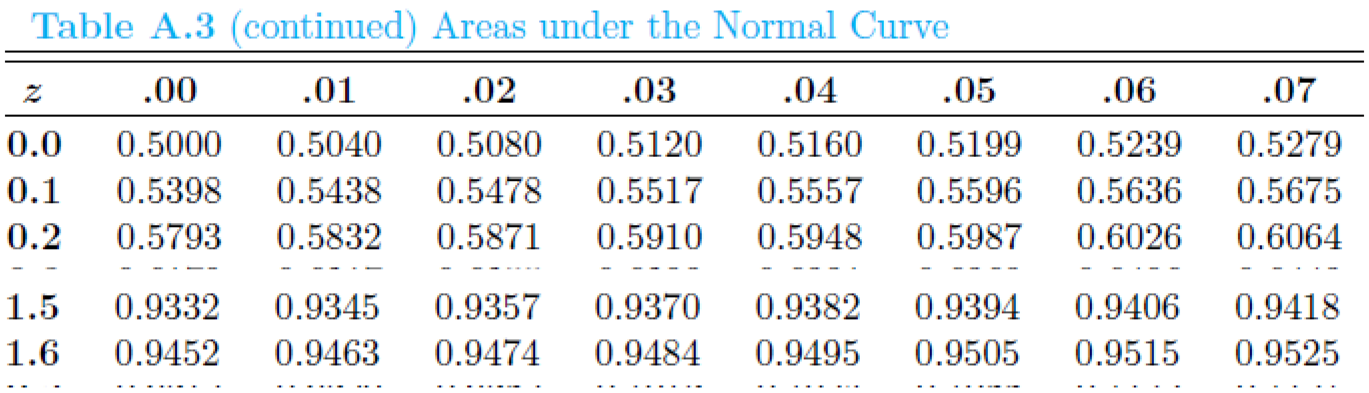


Table A.3 (continued) Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525

- How did we find the probability functions?

A Wrap Up Question

- Sampling Distribution of Means and the CLT
- **Example:** Assume that height of a student in this university is normally distributed with $\mu = 170$ and $\sigma = 10$
- (a) What is the probability that a student coming from the door is shorter that 165?
- (b) What is the probability that the average height of a class with 25 students is shorter than 165?