PROBABILITY AND STATISTICS

Sections 1-4

Elmer Poliquit

Introduction

- The raw material of statistics is **data**.
- Statistics is a field of study concerned with
  1. The collection, organization, summarization, and analysis of data; and
  2. The drawing of inference about a body of data when only a part of the data is observed.

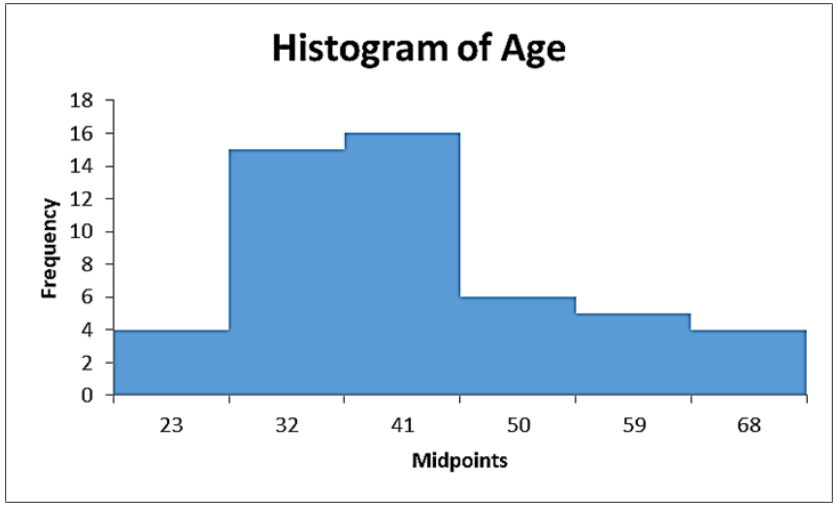
Gender	Age	Systolic BP (mm Hg)	Gender	Age	Systolic BP (mm Hg)
F	38	125	M	40	110
F	27	130	F	67	152
F	32	120	F	35	122
M	55	126	F	45	137
M	42	131	M	55	125
F	40	125	F	33	138
F	35	131	F	61	156
M	34	115	M	50	129
F	29	125	F	58	160
F	50	163	M	38	118
F	30	125	F	36	121
F	34	114	F	43	125
M	41	132	F	30	125
M	33	105	M	41	130
F	39	110	M	65	147
M	35	133	F	19	125
M	43	150	F	48	142
F	20	109	M	46	132
F	25	115	F	60	148
F	39	139	F	70	148
M	72	142	F	36	149
M	41	146	M	40	117
F	46	154	M	33	127
F	32	116	F	50	143
M	37	105	M	41	134

Load Per Week in Pesos

Stem-and-Leaf Display

Stem unit 10

Stem	Leaf
5	045689
6	58
7	12489
8	013467899
9	0123



Introduction

Primary Data Source Facts

- Data collected by the evaluator using methods such as observations, surveys, or interviews
- Can be more expensive and time-consuming, but it allows for more targeted data collection
- Offers an opportunity to review any and all secondary data available before collecting primary data (saving time)

Secondary Data Source Facts

- Provides information if existing data on a topic or project is not current or directly applicable to the chosen evaluation questions
- Information that has already been collected, processed, and reported by another researcher or entity
- Will reveal which questions still need to be addressed and what data has yet to be collected

Introduction

- If, as we observe a characteristic, we find that it takes on different values in different persons, places, or things, we label the characteristic a **variable**.
- A **quantitative variable** is one that can be measured in the usual sense.
  - ❑Counted or expressed numerically
  - ❑Variables can be identified and relationships measured
  - ❑Requires use of statistical analysis
  - ❑Often perceived as a more objective method of data analysis
  - ❑Typically collected with surveys or questionnaires
  - ❑Often represented visually using graphs or charts

Introduction

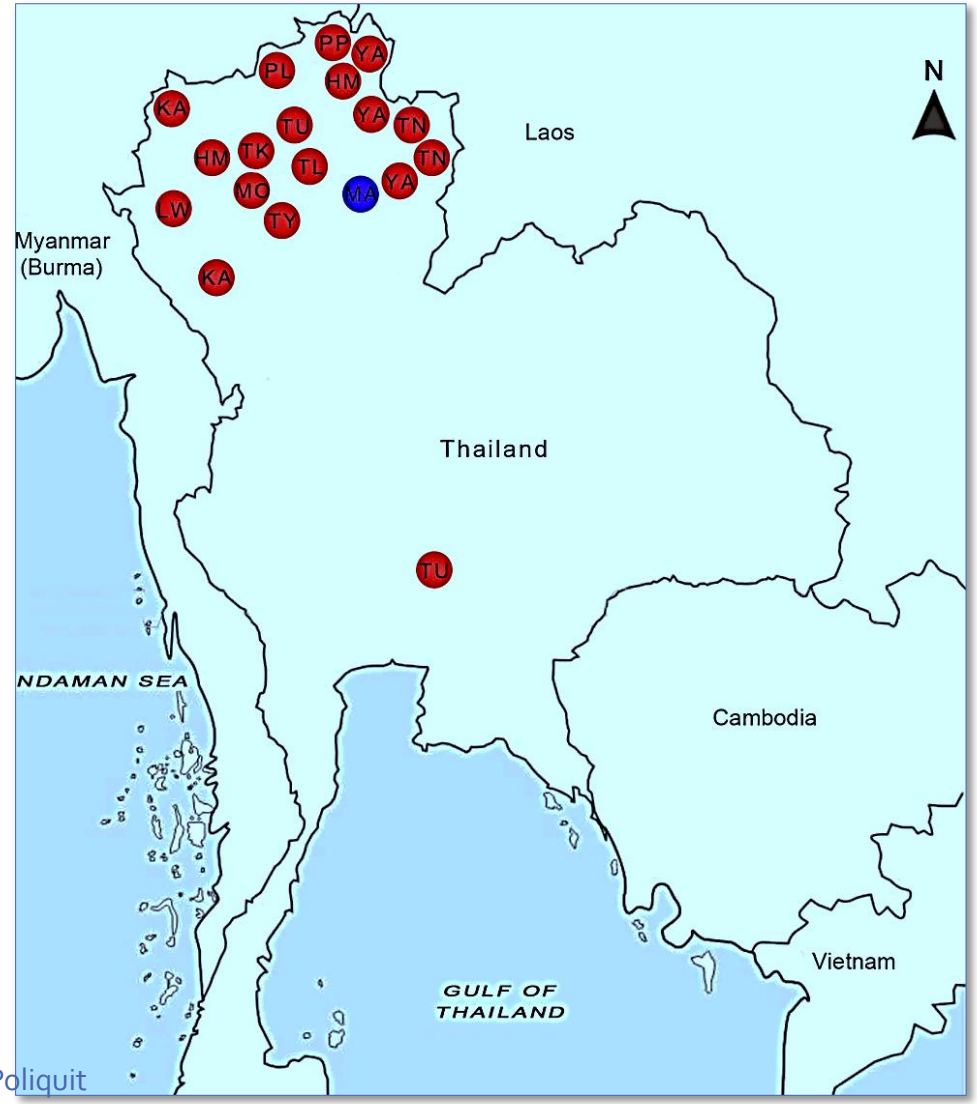
- If, as we observe a characteristic, we find that it takes on different values in different persons, places, or things, we label the characteristic a **variable**.
- A **qualitative variable** is where measuring consists of categorizing.
  - Nonnumerical data examined for patterns and meanings
  - Often described as being more "rich" than quantitative data
  - Is gathered and analyzed by an individual; can be more subjective
  - Can be collected through observation techniques, focus groups, interviews, case studies, etc.

Introduction

- When the values obtained arise as a result of chance factors, so that they cannot be exactly predicted in advance, the variable is called a **random variable**.
- A **discrete random variable** is characterized by gaps or interruptions in the values that it can assume.
  - Number of admissions, number of missing teeth per child
- A **continuous random variable** does not possess the gap or interruptions characteristics of a discrete random variable.
  - Skull circumference, waistline

Introduction

- A **population** of values as the largest collection of values of a random variable for which we have interest at a particular time.
  - Covid-19 Infected in Cebu Province
- A **sample** is a part of a population.
  - Covid-19 Infected in Talisay City



Introduction

- **Measurement** may be defined as the assignment of numbers to objects or events according to a set of rules.

Measurement Scales

1. **Nominal Scale** – lowest measurement scale; it can be grouped
   - Class of animal: bird, mammal, reptile, etc.;
   - Automobile registration plates;
   - Taxpayer registration numbers.

2. **Ordinal Scale** – it can be ranked according to some criterion
   - Consumer preference ranks: "like", "accept", "dislike", "reject", etc.;
   - Military ranks: private, corporal, sergeant, lieutenant, captain, etc.;
   - Certainty degrees: "unsure", "possible", "probable", "sure", etc.

Introduction

- ***Measurement*** may be defined as the assignment of numbers to objects or events according to a set of rules.

Measurement Scales

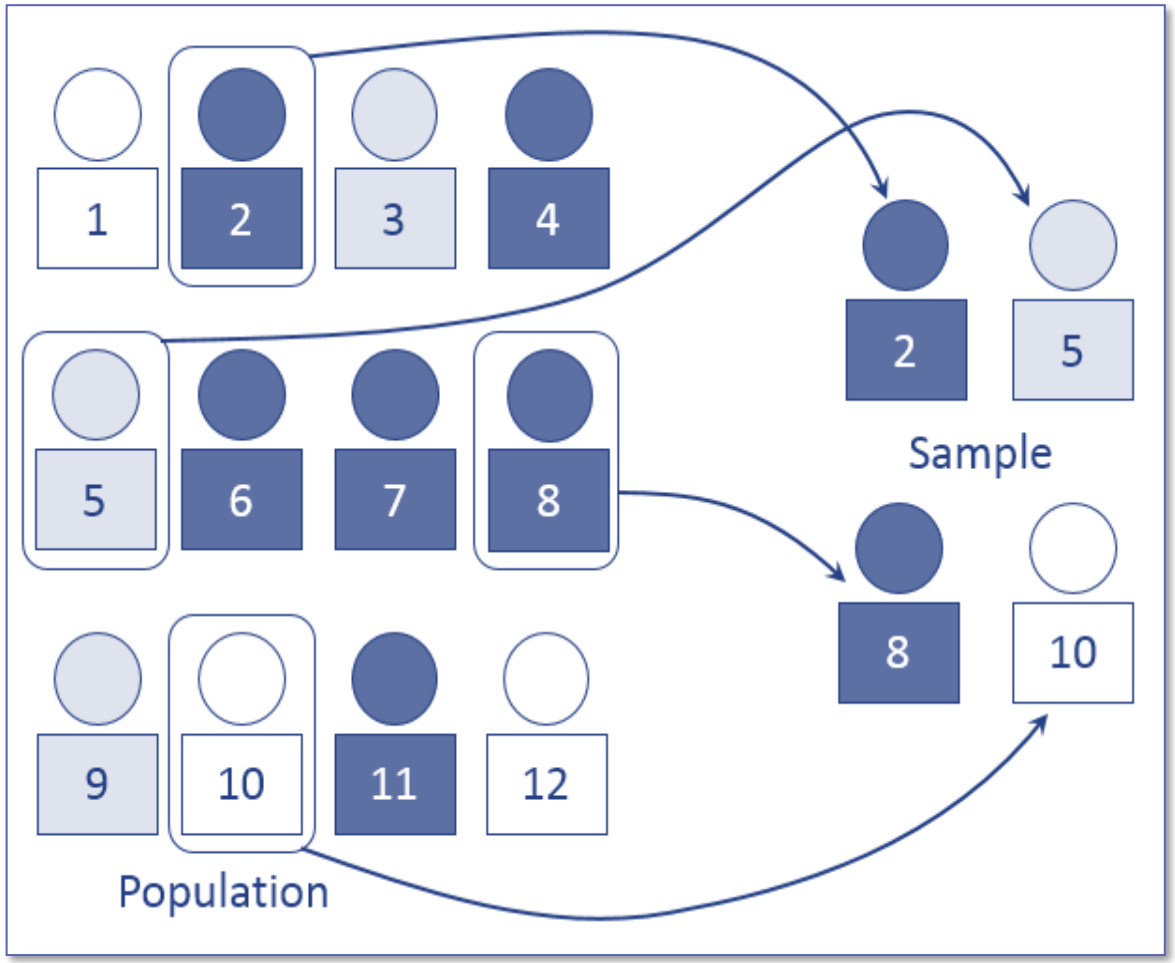
3. **Interval Scale** – the distance between any two measurements is known; zero is arbitrary
   - Temperature
4. **Ratio Scale** – highest level of measurement; fundamental to this scale is a true zero point
   - Height, weight, length

Introduction

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Introduction

- **Statistical inference** is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample that has been drawn from that population.
- If a sample of size **n** is drawn from a population of size **N** in such a way that every possible of size **n** has the same chance of being selected, the sample is called a **simple random sample**.



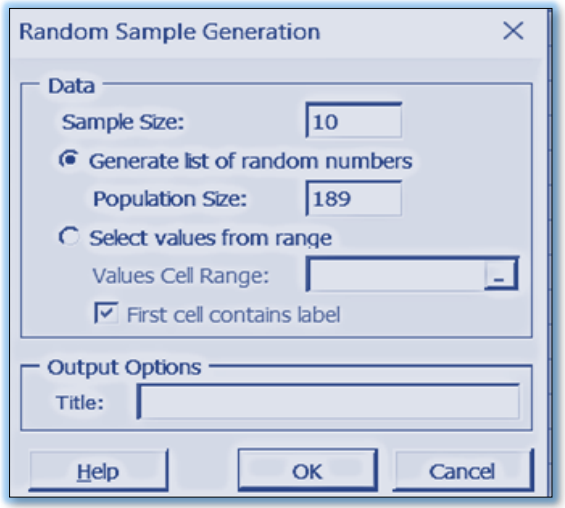
Ages of 189 Subjects Who Participated in a Study on Smoking Cessation

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
1	48	21	53	41	47	61	58	81	62	101	63	121	78	141	46	161	61	181	57
2	35	22	66	42	44	62	54	82	52	102	50	122	66	142	48	162	60	182	50
3	46	23	71	43	48	63	59	83	62	103	59	123	68	143	47	163	51	183	64
4	44	24	75	44	43	64	56	84	57	104	54	124	71	144	43	164	50	184	63
5	43	25	72	45	45	65	62	85	59	105	60	125	69	145	52	165	53	185	65
6	42	26	65	46	40	66	50	86	59	106	50	126	77	146	53	166	64	186	71
7	39	27	67	47	48	67	64	87	56	107	56	127	76	147	61	167	64	187	71
8	44	28	38	48	49	68	53	88	57	108	68	128	71	148	60	168	53	188	73
9	49	29	37	49	38	69	61	89	53	109	66	129	43	149	53	169	60	189	66
10	49	30	46	50	44	70	53	90	59	110	71	130	47	150	53	170	54		
11	44	31	44	51	43	71	62	91	61	111	82	131	48	151	50	171	55		
12	39	32	44	52	47	72	57	92	55	112	68	132	37	152	53	172	58		
13	38	33	48	53	46	73	52	93	61	113	78	133	40	153	54	173	62		
14	49	34	49	54	57	74	54	94	56	114	66	134	42	154	61	174	62		
15	49	35	30	55	52	75	61	95	52	115	70	135	38	155	61	175	54		
16	53	36	45	56	54	76	59	96	54	116	66	136	49	156	61	176	53		
17	56	37	47	57	56	77	57	97	51	117	78	137	43	157	64	177	61		
18	57	38	45	58	53	78	52	98	50	118	69	138	46	158	53	178	54		
19	51	39	48	59	64	79	54	99	50	119	71	139	34	159	53	179	51		
20	61	40	47	60	53	80	53	100	55	120	69	140	46	160	54	180	62		

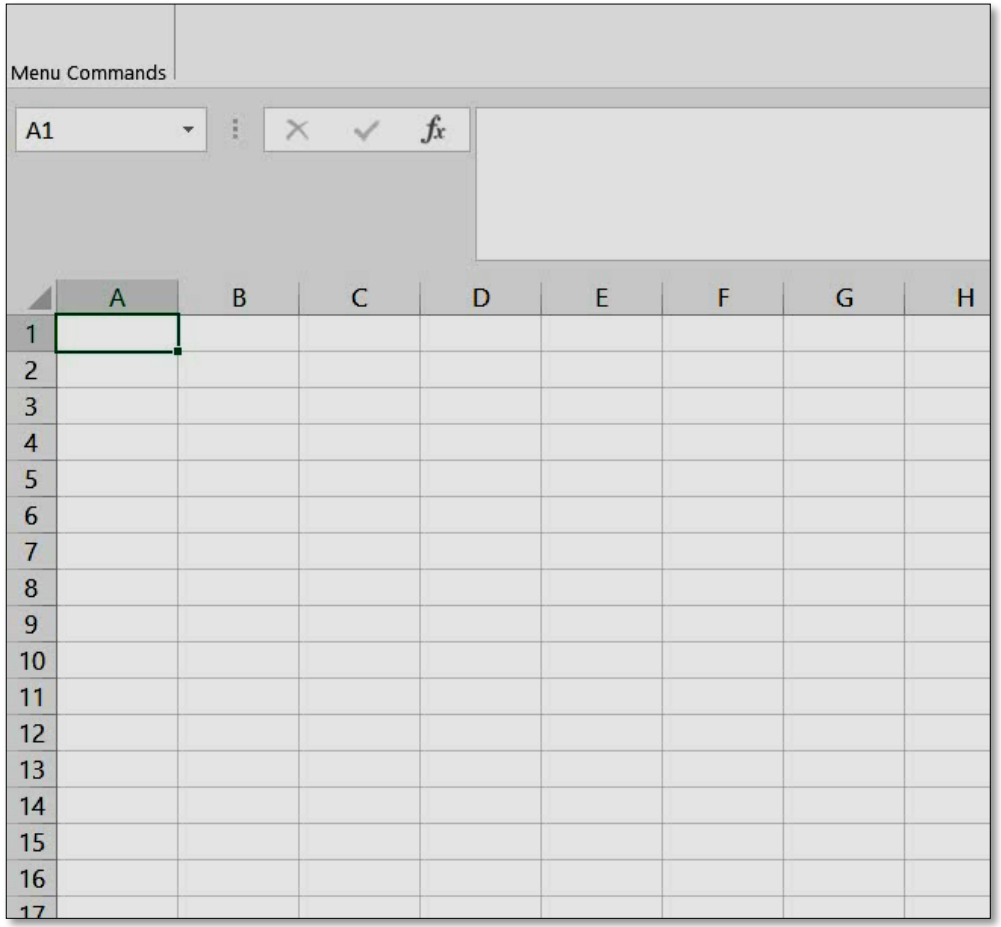
Introduction

SAMPLING AND STATISTICAL INFERENCE

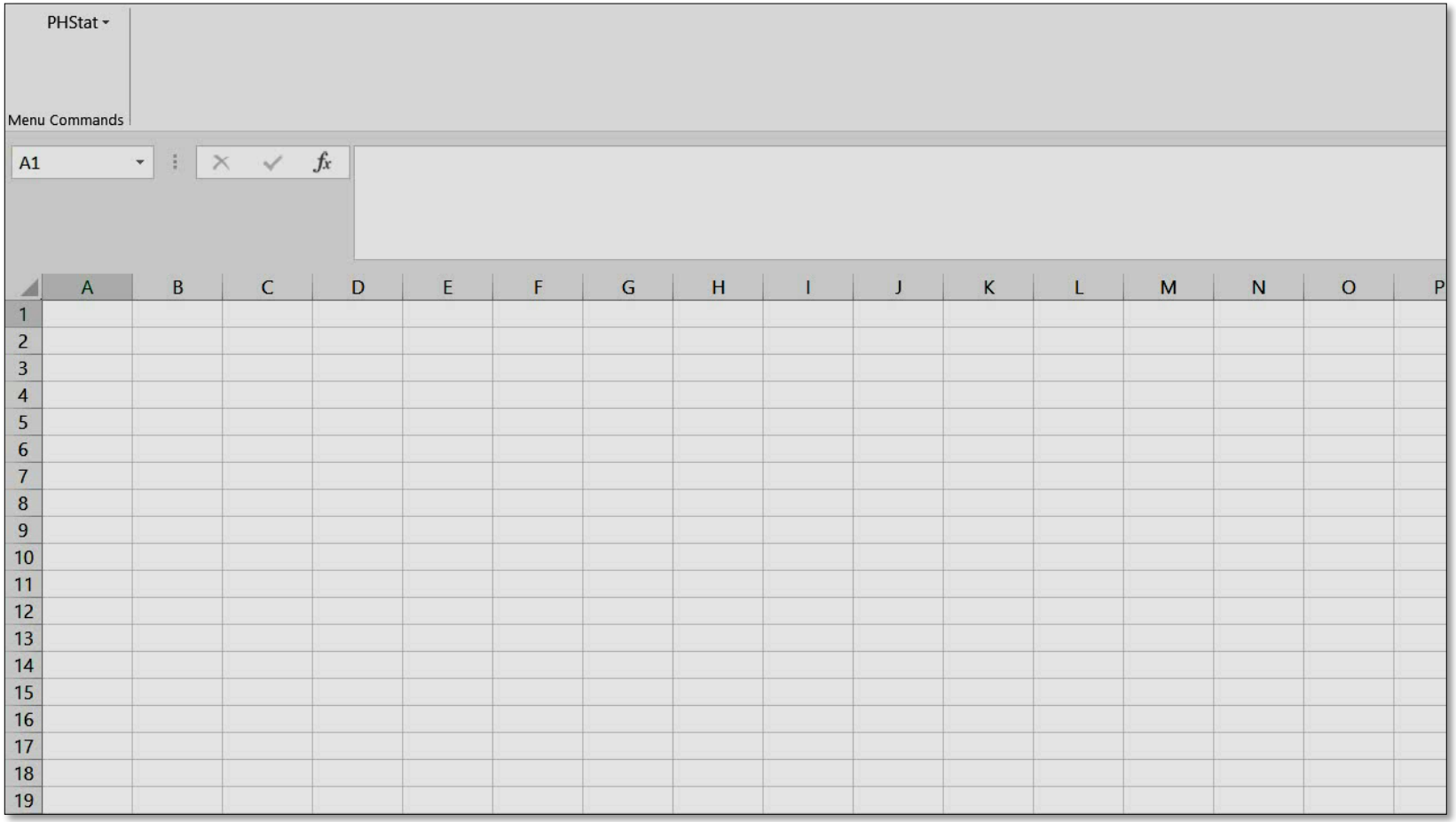
- One way of selecting a **simple random sample** is to use a table of random numbers like that shown in the Appendix, Table A.
- Another is using RANBETWEEN of excel or
- PHStat



Excel RANBETWEEN Function



PHStat Random Sample Generator



Random Sample using R

r sample - simple random sampling in r

```
sample (c(1:189))
```

r sample with replacement from vector

```
sample (c(1:189), replace =T)
```

r sample multiple times without replacement

```
sample (c(1:189), size=5, replace =F)
```

r sample with replacement from vector

```
sample (c(1:189), size=5, replace=T)
```

r sample - generate random sample in r

```
sample (c('Joe','Karl','Jack','Larry','Curly',
        'Moe','Kim','Kathy','Sam','Jim'), size=3)
```

```
> sample (c(1:189))
  [1]  73 121 161  44 169   1  32  75  40 109  70 174  29 102  25  23 123  54
 [19] 104  46  88  71  87 186 149  78  86  33 124  31  90  13  50 101 100 115
 [37] 138 125 181  17 111  35  26 120  36 133  85 136  98  18 135  12  48 113
 [55]  14 167 141 163  10 156  49  52  93 175  30  68 107  37   6   8 184 173
 [73] 142  24 185  20 180 148  74  64  55 105  69  21  27 188  95  47 183 137
 [91] 122  62 114  53 144 118 131 157  39 165 119 172  96 134  60  97  72  89
[109]  19 187   5  65 159 176   4 170  94 130  11 177 155  83 145 150 152  51
[127] 158 108  16  58  15 171  42 103  38 106  81  59 160 112 162  67 140 139
[145]  61   9 151  79  82 129 178 110 189  84 127 164 146 143 116  22 126  99
[163]   7  76  34 117 182  63  80  41 128  56 179   2 168  91   3 147  92  66
[181] 154 166 153  28  45  77 132  57  43
> # r sample with replacement from vector
> sample (c(1:189), replace =T)
  [1] 110 186 110  36  79  48  87 106 157  64 135  56  65 176   4  35  27 182
 [19] 158 100  49  90 159  22 153 102 118  40 136 188 152  55   8 119 100 102
 [37]  32 132  37  26 151  11   7 165 143 140  60 182 132 137 170 156  90 180
 [55] 160 152 111 131 158  34 166 106 151 139 181 128  55  49 164  25 171 144
 [73]  62  91 136 105  52 141  36  75  35 164 187  52 120  82 172 180  45 115
 [91] 154  96 131  44 168  21  13 113 176 111  17 177 163  66 135 114  32  87
[109]   2 173 158  57 130 124  38 174 163 148 154 148 155  67 167 130  75  27
[127] 111 131  48 124  22 138  93 156  12 154  82 169 157  54  10  61  99 137
[145]  94 153  42  25 142  86 131  20 159  91  72 163 178  93  95  26  62  10
[163]  81 185 181   5  16  47  51  50 177  76  65  97   7  62  97 147 113 173
[181]  25  45  94 145   2  89 131  21 156
> # r sample multiple times without replacement
> sample (c(1:189), size=5, replace =F)
[1] 13 60 30 41 34
> # r sample with replacement from vector
> sample (c(1:189), size=5, replace=T)
[1] 139 183 161  39  38
> # r sample - generate random sample in r
> sample (c('Joe','Karl','Jack','Larry','Curly',
+           'Moe','Kim','Kathy','Sam','Jim'), size=3)
[1] "Jack"  "Kim"   "Kathy"
```

Introduction

SAMPLING AND STATISTICAL INFERENCE

- A **research study** is a scientific study of a phenomenon of interest. Research studies involve designing sampling protocols, collecting and analyzing data, and providing valid conclusions based on the results of the analyses.
- **Experiments** are a special type of research study in which observations are made after specific manipulations of conditions have been carried out; they provide the foundation for scientific research.

Introduction

SAMPLING AND STATISTICAL INFERENCE

- A sampling method that is widely used in healthcare research is the **systematic sample**.
- A random numbers table is then employed to select a starting point in the file system. The record located at this starting point is called record **x**. A second number, determined by the number of records desired, is selected to define the sampling interval (call this interval **k**). Consequently, the data set would consist of records **x**, **x + k**, **x + 2k**, **x + 3k**, and so on, until the necessary number of records are obtained.

Introduction

SAMPLING AND STATISTICAL INFERENCE

- Suppose that the first random digit is a 4 and will serve as our starting point, **x**. Since we are starting at subject 4, this leaves 185 remaining subjects (i.e., 189–4) from which to choose. Since we wish to select 10 subjects, one method to define the sample interval, **k**, would be to take 185/10 = 18.5. To ensure that there will be enough subjects, it is customary to round this quotient down, and hence we will round the result to 18. The resulting sample is shown.

Systematically Selected Subject Number	Age
4	44
22	66
40	47
58	53
76	59
94	56
112	68
130	47
148	60
166	64

Introduction

Using R

```
#data
age=read.csv('Ages189.csv')
age
#define function to obtain systematic sample
obtain_sys = function(N,n){
 k = ceiling(N/n)
 r = sample(1:k, 1)
 seq(r, r + k*(n-1), k)
}
#obtain systematic sample
sys_sample_age = age[obtain_sys(nrow(age), 10), ]
sys_sample_age
```

```
> sys_sample_age
    Subject.No. Age
15           15  49
34           34  49
53           53  46
72           72  57
91           91  61
110         110  71
129         129  43
148         148  60
167         167  64
186         186  71
```

Introduction

SAMPLING AND STATISTICAL INFERENCE

- A common situation that may be encountered in a population under study is one in which the sample units occur together in a grouped fashion. On occasion, when the sample units are not inherently grouped, it may be possible and desirable to group them for sampling purposes. In other words, it may be desirable to partition a population of interest into **groups**, or **strata**, in which the sample units within a particular stratum are more similar to each other than they are to the sample units that compose the other strata. After the population is stratified, it is customary to take a random sample independently from each stratum. This technique is called **stratified random sampling**. The resulting sample is called a **stratified random sample**.

Example

- Suppose we want 30 samples of COVID-19 patients in Cebu hospitals then we have 10 random sample patients from Chong Hua, 10 random sample patients from Vicente Sotto and 10 random sample patients from Cebu Doctors.

Introduction

THE SCIENTIFIC METHOD AND THE DESIGN OF EXPERIMENTS

- The **scientific method** is a process by which scientific information is collected, analyzed, and reported in order to produce unbiased and replicable results in an effort to provide an accurate representation of observable phenomena.

Key Elements

1. **Making an Observation**

   This observation leads to the formulation of questions or uncertainties that can be answered in a scientifically rigorous way.

2. **Formulating a Hypothesis**

   It is formulated to explain the observation and to make quantitative predictions of new observations. Often hypotheses are generated as a result of extensive background research and literature reviews.

Introduction

THE SCIENTIFIC METHOD AND THE DESIGN OF EXPERIMENTS

3. Designing an Experiment

It will yield the data necessary to validly test an appropriate statistical hypothesis. This step of the scientific method, like that of data analysis, requires the expertise of a statistician.

Those who properly design research experiments make every effort to ensure that the measurement of the phenomenon of interest is both **accurate** and **precise**.

Accuracy refers to *the correctness of a measurement*. **Precision**, on the other hand, refers to the *consistency of a measurement*. It should be noted that in the social sciences, the term *validity* is sometimes used to mean **accuracy** and that *reliability* is sometimes used to mean **precision**.

A *true experimental design* is one in which study subjects are randomly assigned to an *experimental group* (or *treatment group*) and a *control group* that is not directly exposed to a treatment.

Introduction

THE SCIENTIFIC METHOD AND THE DESIGN OF EXPERIMENTS

4. Conclusion

In the execution of a research study or experiment, one would hope to have collected the data necessary to draw conclusions, with some degree of confidence, about the hypotheses that were posed as part of the design.

It is often the case that hypotheses need to be modified and retested with new data and a different design.

Whatever the conclusions of the scientific process, however, results are rarely considered to be conclusive. That is, results need to be replicated, often a large number of times, before scientific credence is granted them.

Descriptive Statistics

- An **ordered array** is a listing of the values of a collection (either population or sample) in order of magnitude from the smallest value to the largest value.

  - The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint.

2.5	2.8	2.8	2.9	3.0	3.3	3.4	3.6
3.7	4.0	4.4	4.8	4.8	5.2	5.6	5.7

How do we describe the walk-in adult patients at a medical mission? The data set below presents the patient's gender, age and systolic BP.

Gender	Age	Systolic BP (mm Hg)	Gender	Age	Systolic BP (mm Hg)
F	38	125	M	40	110
F	27	130	F	67	152
F	32	120	F	35	122
M	55	126	F	45	137
M	42	131	M	55	125
F	40	125	F	33	138
F	35	131	F	61	156
M	34	115	M	50	129
F	29	125	F	58	160
F	50	163	M	38	118
F	30	125	F	36	121
F	34	114	F	43	125
M	41	132	F	30	125
M	33	105	M	41	130
F	39	110	M	65	147
M	35	133	F	19	125
M	43	150	F	48	142
F	20	109	M	46	132
F	25	115	F	60	148
F	39	139	F	70	148
M	72	142	F	36	149
M	41	146	M	40	117
F	46	154	M	33	127
F	32	116	F	50	143
M	37	105	M	41	134

Organizing and Presenting Data

I. **Tabular presentation**
- Frequency tables
  - Stem-and-leaf display or stem-and-leaf plot
  - Double stem-and-leaf display
- Grouped frequency tables
- Contingency tables

The Oxford English dictionary defines a **table** as "**a set of facts or figures systematically displayed, especially in columns**".

II. **Graphical presentation of data**
- Dot plots
- Scatter plots
- Bar graph
- Line graph
- Pie graph
- Histogram
- Frequency polygon

I. Tabular Presentation

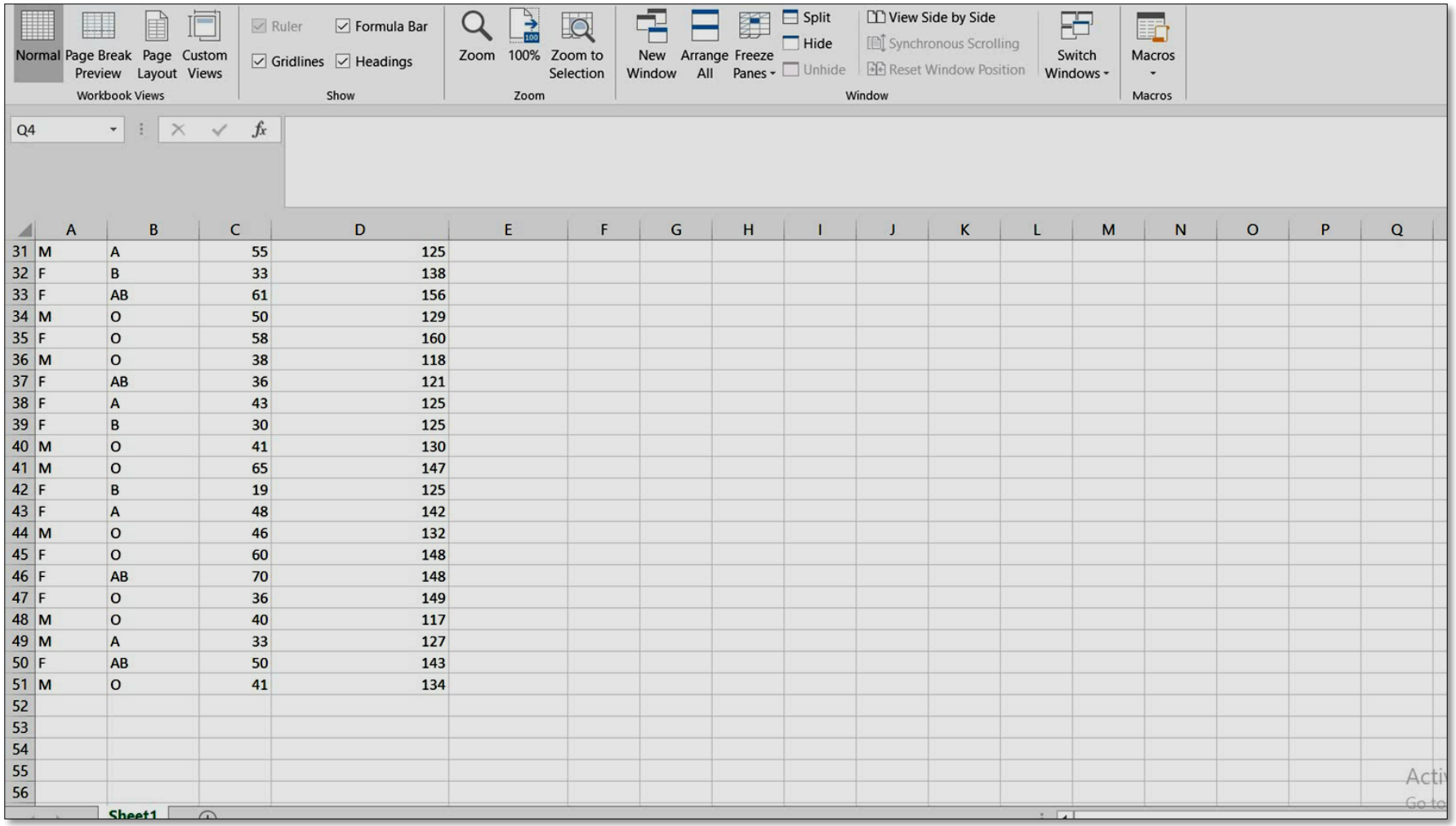
Frequency Table

The frequency of a particular data value is the number of times the data value occurs.

Gender	No. of Walk-in Adult Patients
Male	20
Female	30
Total	50

In using PHStat, choose only the best presentation of your data.

PHStat One-Way Table and Charts



	A	B	C	D
31	M	A	55	125
32	F	B	33	138
33	F	AB	61	156
34	M	O	50	129
35	F	O	58	160
36	M	O	38	118
37	F	AB	36	121
38	F	A	43	125
39	F	B	30	125
40	M	O	41	130
41	M	O	65	147
42	F	B	19	125
43	F	A	48	142
44	M	O	46	132
45	F	O	60	148
46	F	AB	70	148
47	F	O	36	149
48	M	O	40	117
49	M	A	33	127
50	F	AB	50	143
51	M	O	41	134
52				
53				
54				
55				
56				

One-Way Table and Charts in R

```
patients=read.csv('Patients.csv')
#One way table
mytable <- with(patients, table(Gender))
mytable  # frequencies
prop.table(mytable) # proportions
prop.table(mytable)*100 # percentages
```

```
> #One way table
> mytable <- with(patients, table(Gender))
> mytable  # frequencies
Gender
 F  M
30 20
> prop.table(mytable) # proportions
Gender
   F   M
0.6 0.4
> prop.table(mytable)*100 # percentages
Gender
 F  M
60 40
```

I. Tabular Presentation

A **stem-and-leaf display** or **stem-and-leaf plot** is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution.

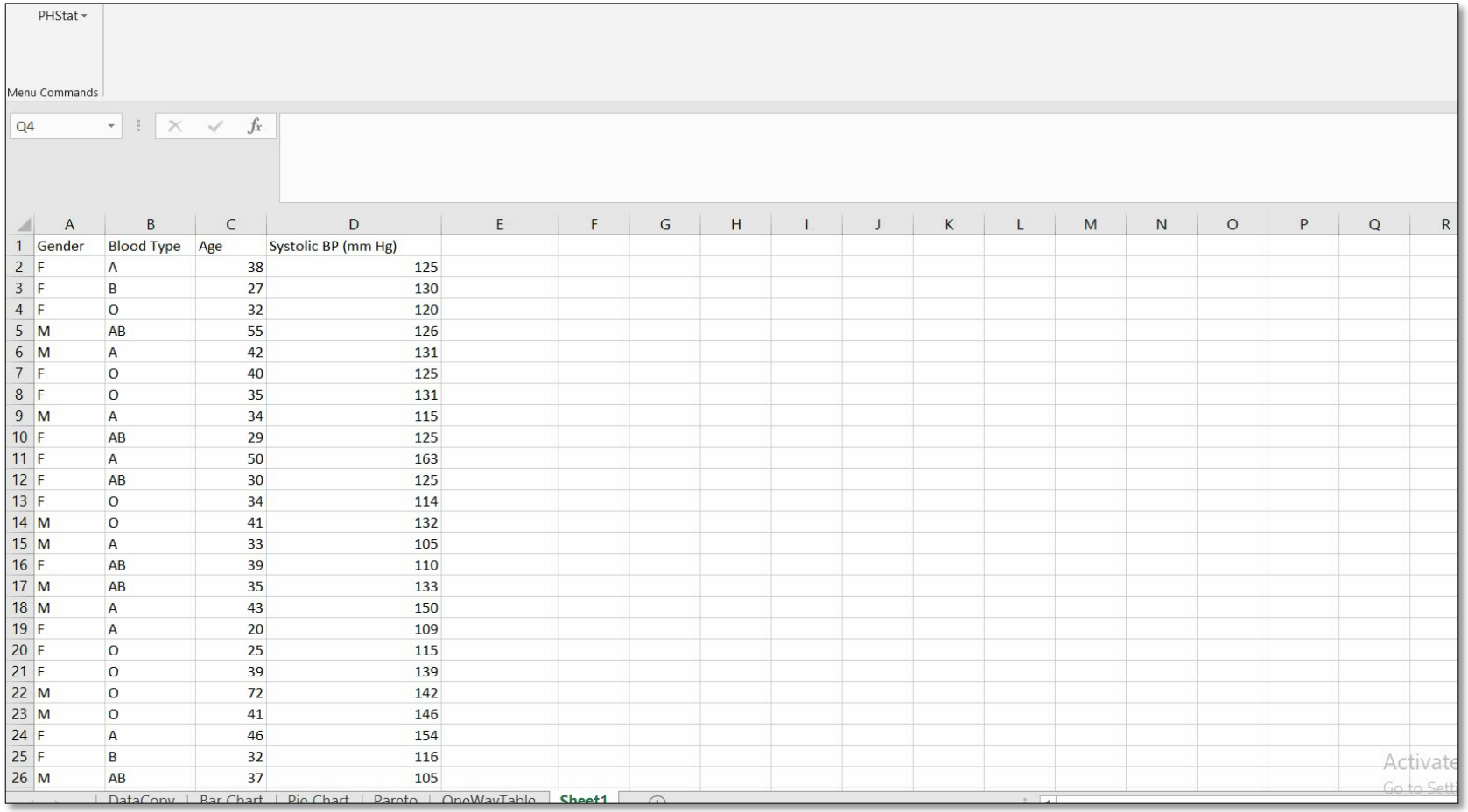
Statistics	
Sample Size	50
Mean	130.9200
Median	129.5000
Std. Deviation	14.6383
Minimum	105.0000
Maximum	163.0000

Stem-and-Leaf Display

Stem unit: 10

10	5 5 9
11	0 0 4 5 5 6 7 8
	0 1 2 5 5 5 5 5 5 5 6
12	7 9
13	0 0 1 1 2 2 3 4 7 8 9
14	2 2 3 6 7 8 8 9
15	0 2 4 6
16	0 3

Stem-and-Leaf Display using PHStat



	A	B	C	D
1	Gender	Blood Type	Age	Systolic BP (mm Hg)
2	F	A	38	125
3	F	B	27	130
4	F	O	32	120
5	M	AB	55	126
6	M	A	42	131
7	F	O	40	125
8	F	O	35	131
9	M	A	34	115
10	F	AB	29	125
11	F	A	50	163
12	F	AB	30	125
13	F	O	34	114
14	M	O	41	132
15	M	A	33	105
16	F	AB	39	110
17	M	AB	35	133
18	M	A	43	150
19	F	A	20	109
20	F	O	25	115
21	F	O	39	139
22	M	O	72	142
23	M	O	41	146
24	F	A	46	154
25	F	B	32	116
26	M	AB	37	105

Stem-and-Leaf Display using R

#Stem and Leaf Display

stem(patients$Age)

stem(patients$SystolicBPmmHg)

```
> #Stem and Leaf Display
> stem(patients$Age)

  The decimal point is 1 digit(s) to the right of the |

  1 | 9
  2 | 0579
  3 | 00223334455566678899
  4 | 00011112335668
  5 | 000558
  6 | 0157
  7 | 02

> stem(patients$Systolic.BP..mm.Hg.)

  The decimal point is 1 digit(s) to the right of the |

  10 | 559
  11 | 00455678
  12 | 01255555555679
  13 | 00112234789
  14 | 22367889
  15 | 0246
  16 | 03
```

I. Tabular Presentation

The **frequency of a group** (or class interval) is the number of data values that fall in the range specified by that group (or class interval).

To construct a frequency table, we need the following:

- Range (**R**) = Maximum Value – Minimum Value
- Desired number of intervals (**d**) (Suggested: 5 – 15 intervals)
- Class interval width (**i**) = $\frac{\mathbf{R}}{\mathbf{d}}$
  - ✓ Round up to the next integer

Then construct a table with three columns (class interval, tally, frequency), and then write the data groups or class intervals in the first column…Tally…Count the tally marks. (Conventional)

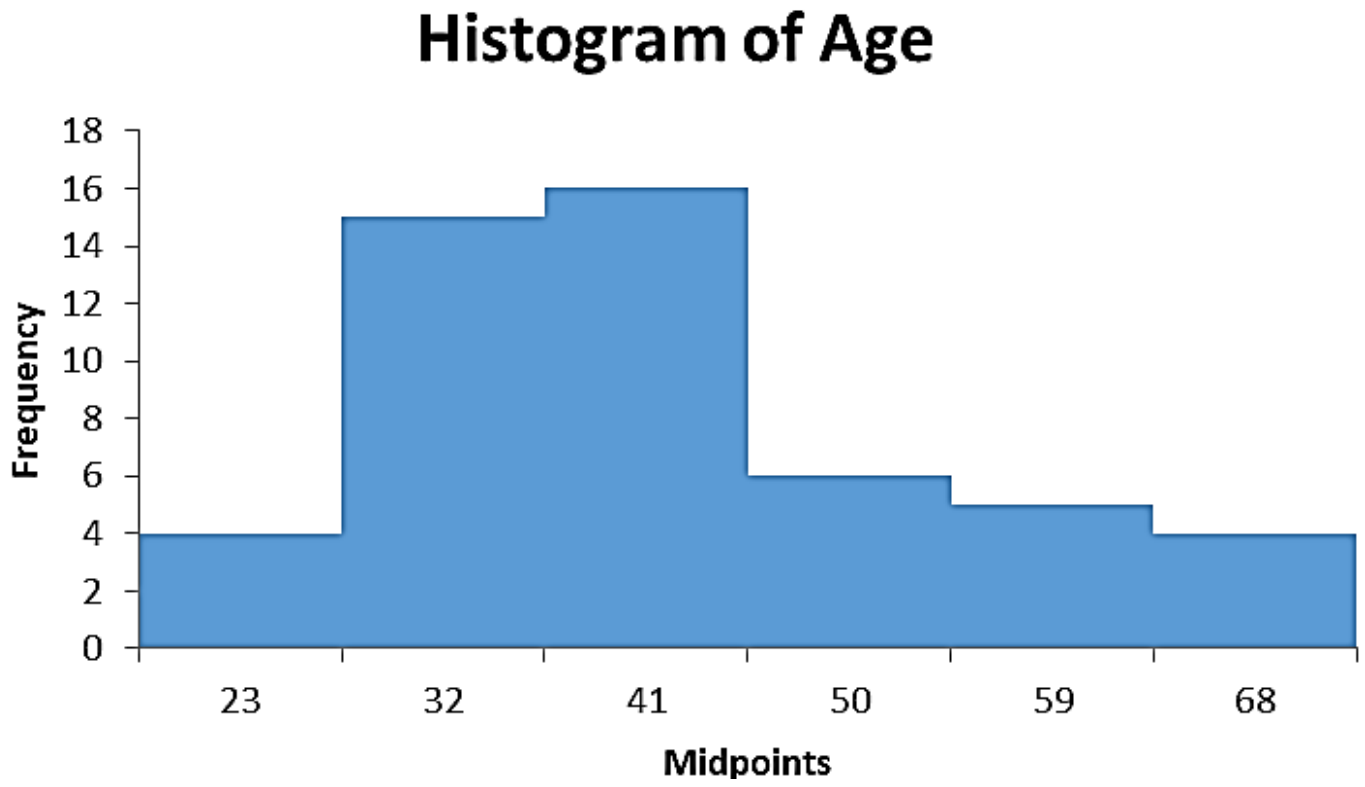
Descriptive Statistics

- Grouped Data
  - The **Frequency Distribution** (**Cumulative Frequency**, **Relative Frequency**, **Cumulative Relative Frequency**)

Age Interval	No. of Walk-in Female Patients
19-29	5
30-40	14
41-51	6
52-62	3
63-73	2
Total	30

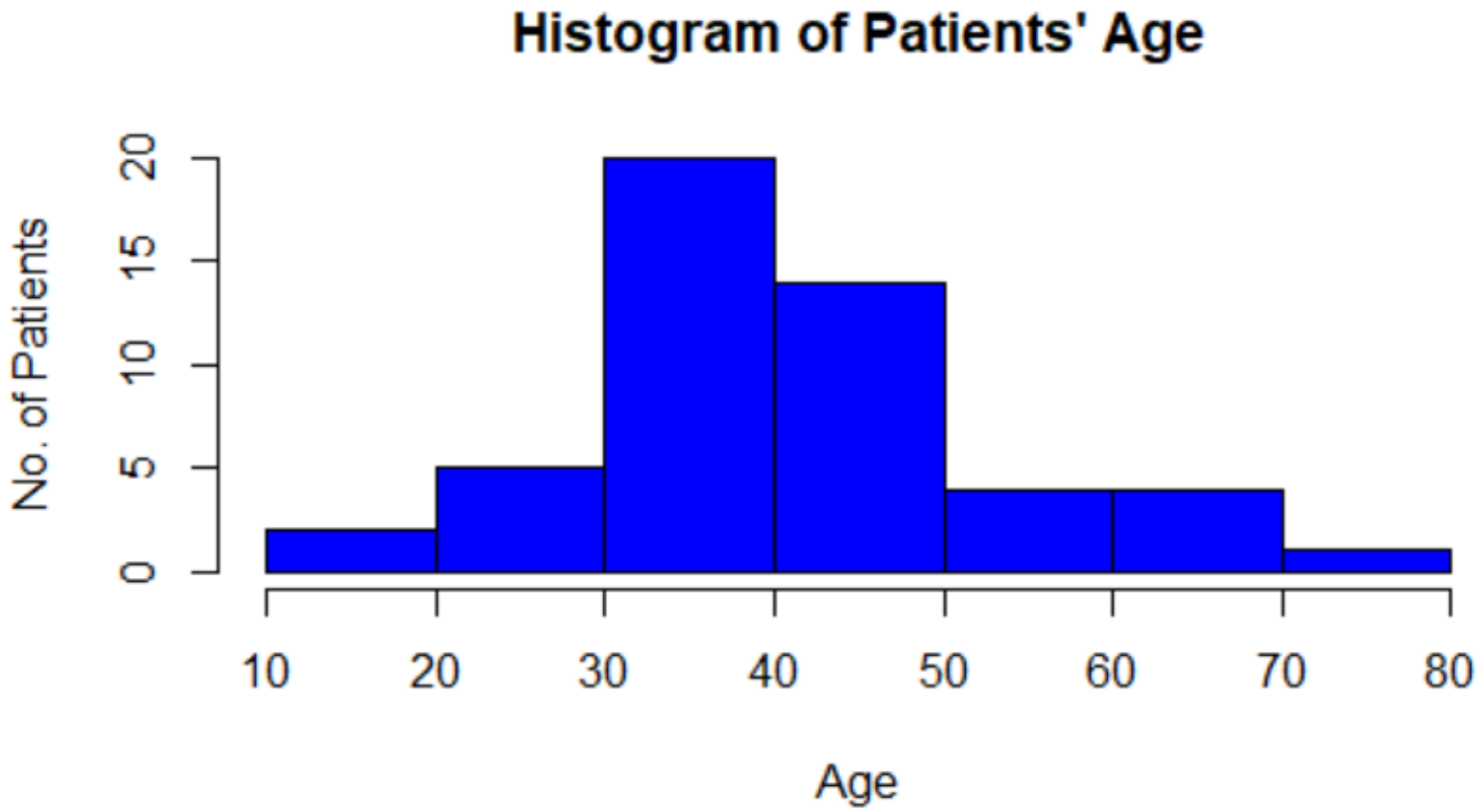
Descriptive Statistics

- **Graphs**
  - The **Histogram**



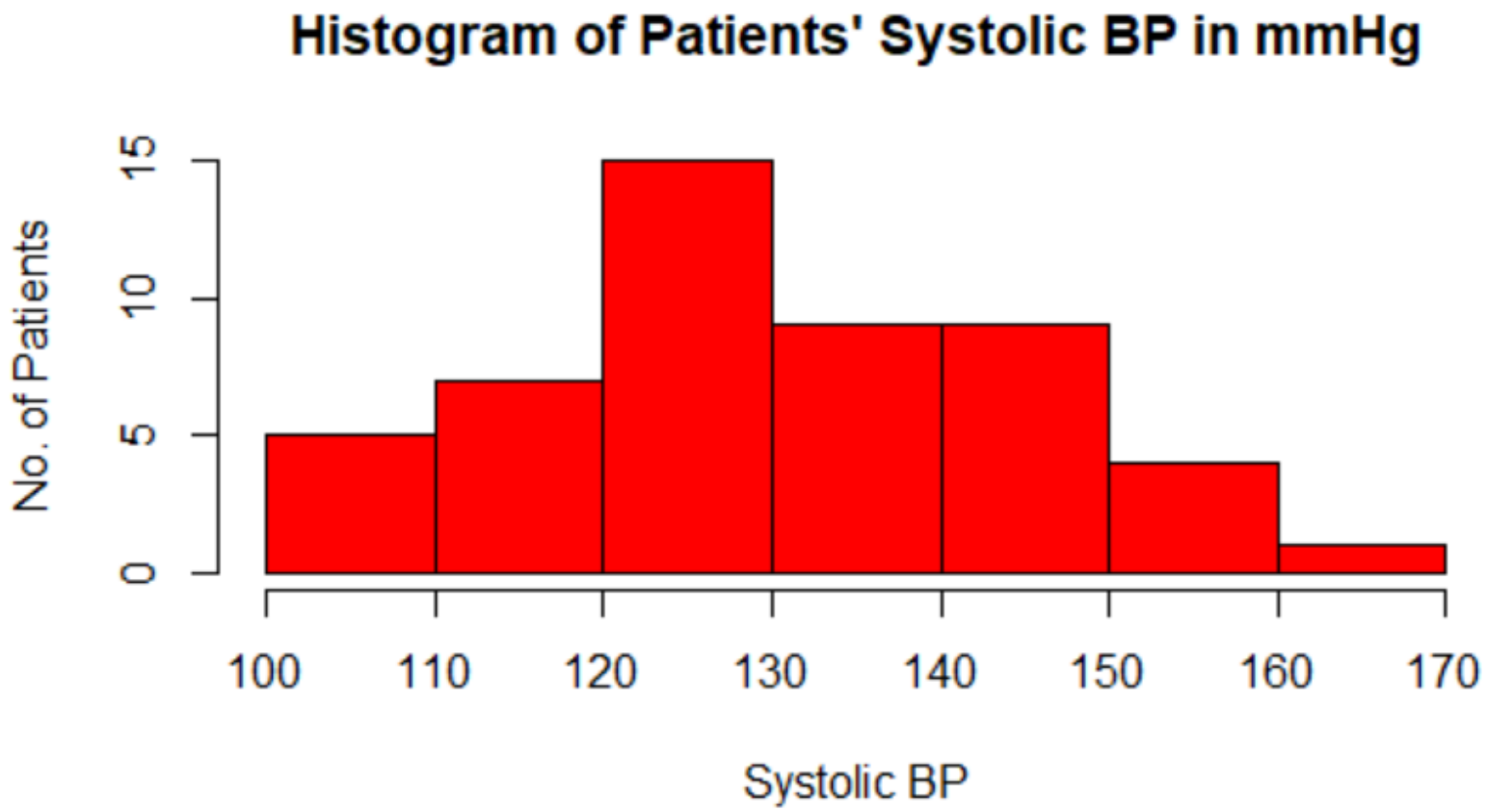
#Histogram

hist(patients$Age, xlab="Age", ylab = "No. of Patients", main = "Histogram of Patients' Age", col='blue')



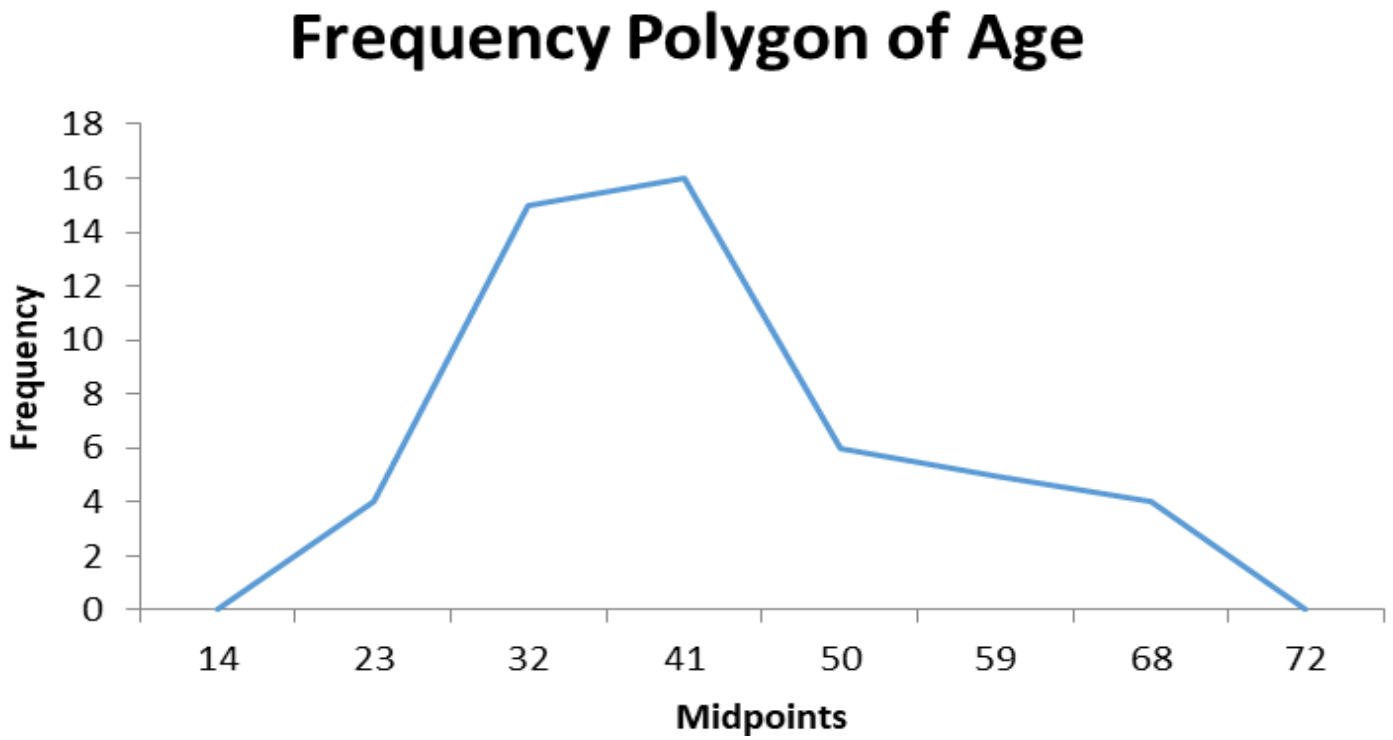
#Histogram

hist(patients$SystolicBPmmHg., xlab="Systolic BP", ylab = "No. of Patients", main = "Histogram of Patients' Systolic BP in mmHg", col='red')



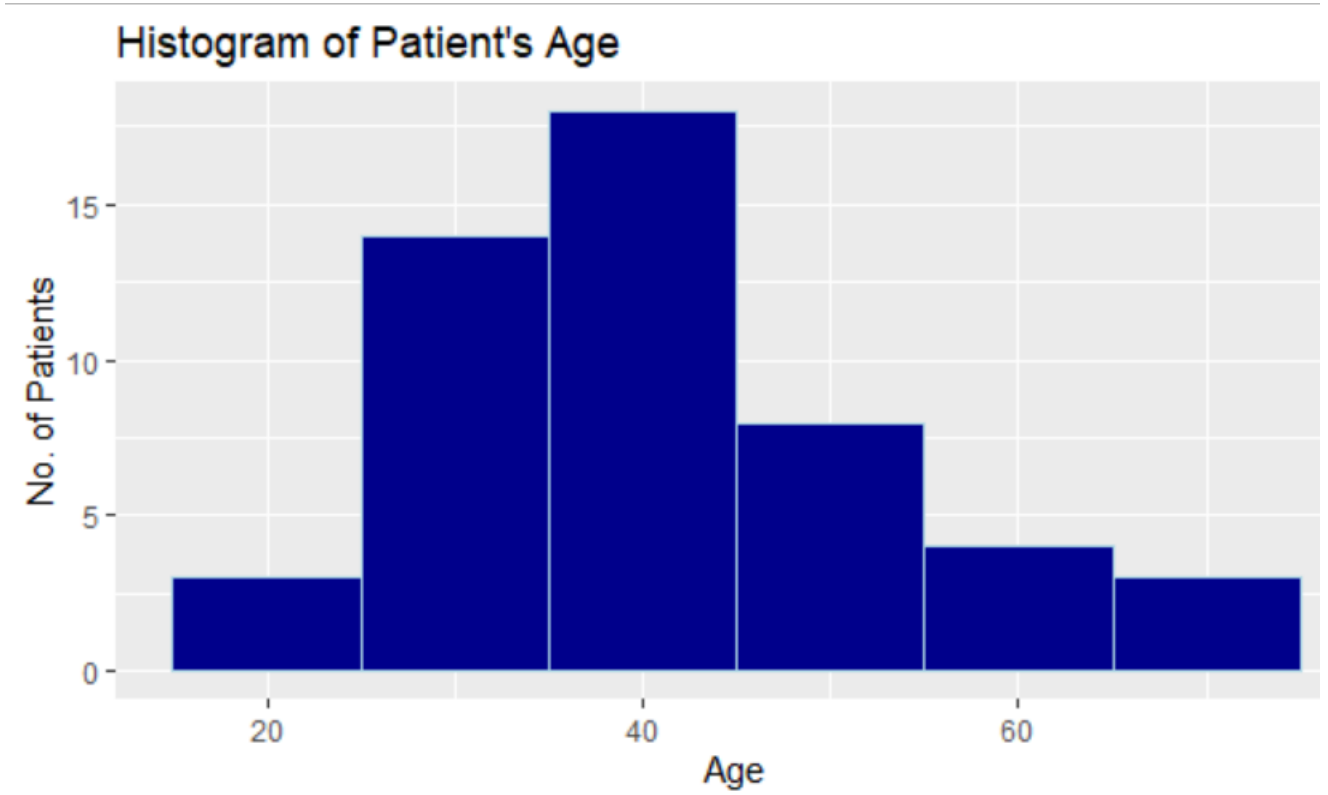
Descriptive Statistics

- **Graphs**
  - The **Frequency Polygon**



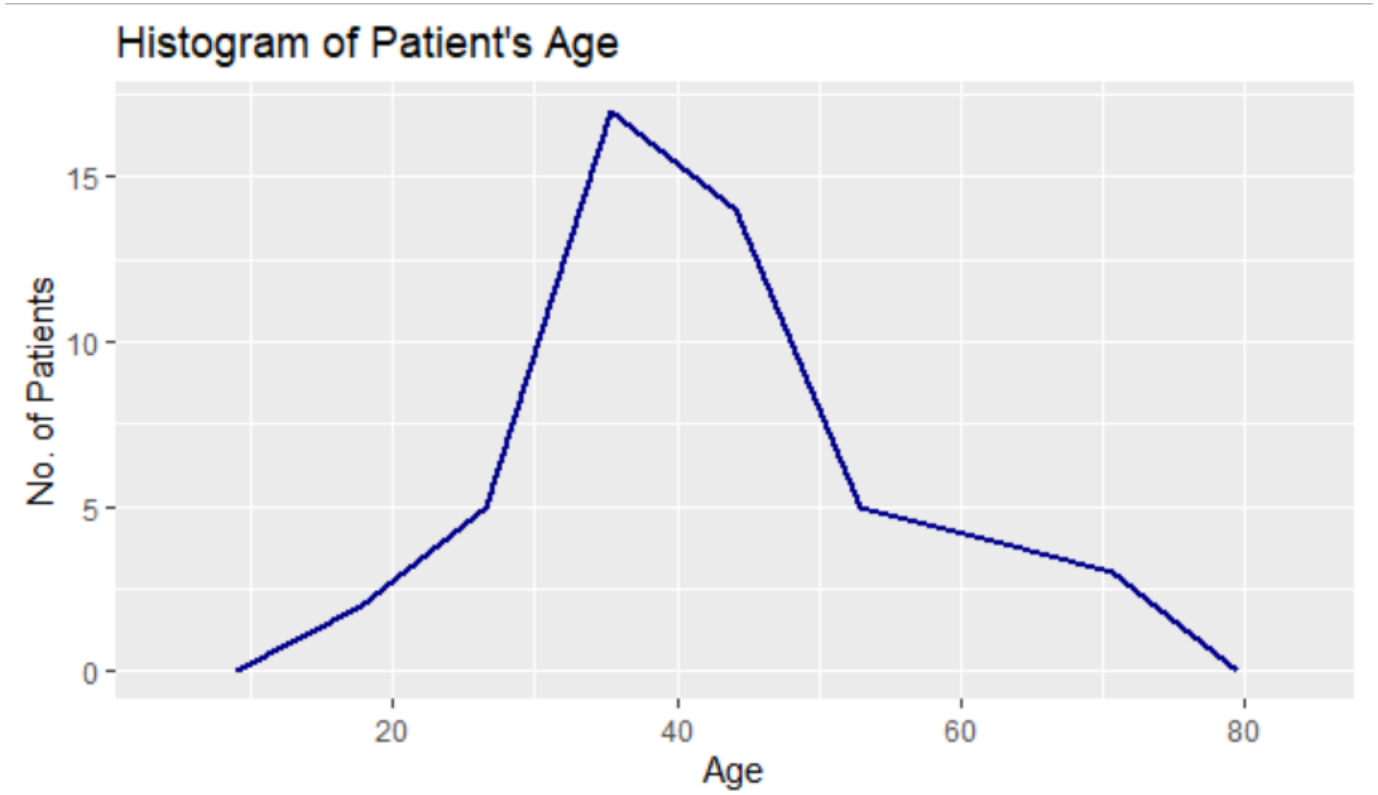
#Histogram & Frequency Polygon using ggplot2

```
ggplot(patients, aes(Age)) +          # histogram

  geom_histogram(bins = 7,binwidth=10, color="lightblue", fill="darkblue")+

  labs(title = "Histogram of Patient's Age", y="No. of Patients")
```

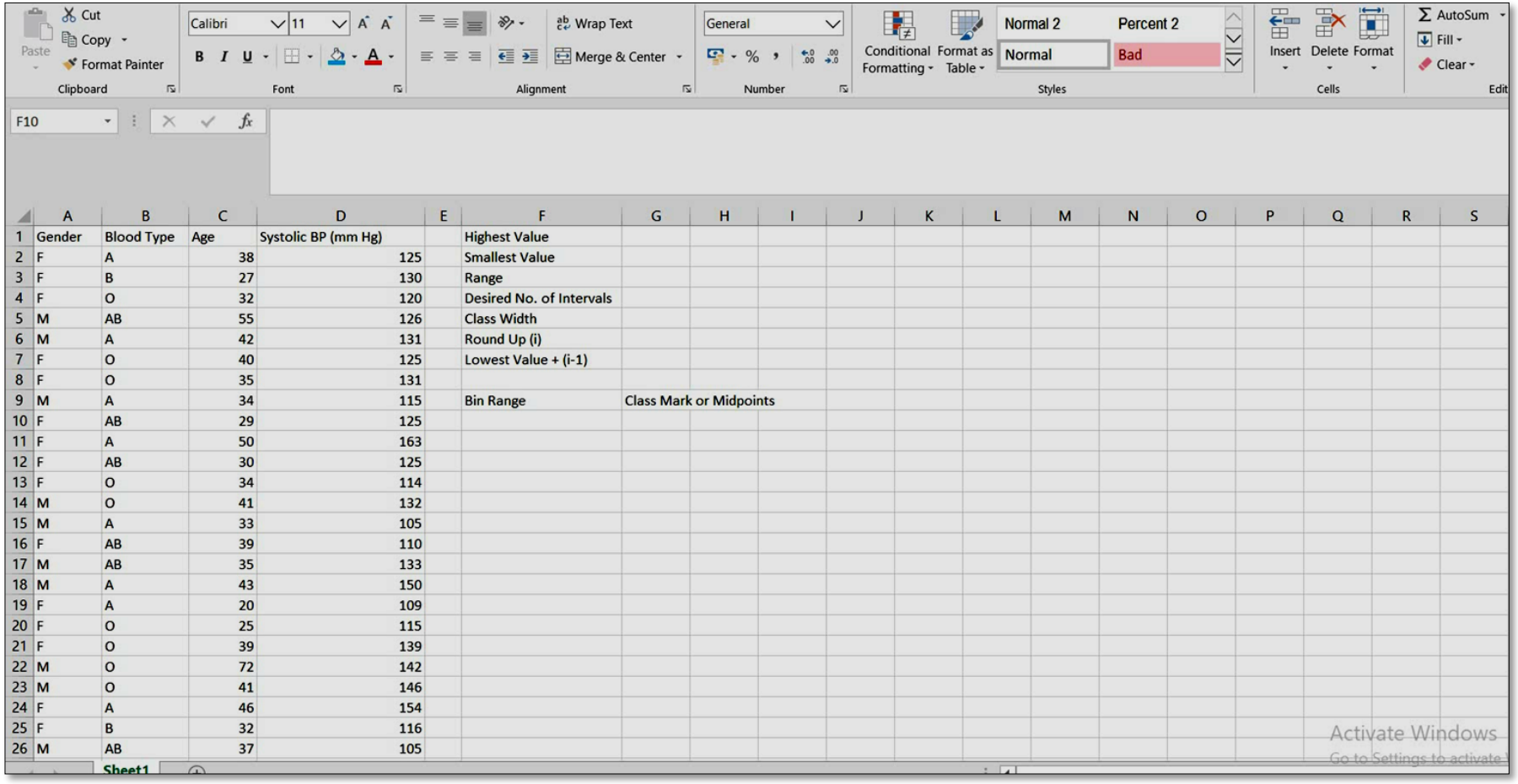


#Histogram & Frequency Polygon using ggplot2

```
ggplot(patients, aes(Age)) +          # frequency polygon
  geom_freqpoly(bins = 7, size = 1, color = "darkblue")+
  labs(title = "Histogram of Patient's Age", y="No. of Patients")
```



Using PHStat



	A	B	C	D	E	F	G
1	Gender	Blood Type	Age	Systolic BP (mm Hg)		Highest Value	
2	F	A	38	125		Smallest Value	
3	F	B	27	130		Range	
4	F	O	32	120		Desired No. of Intervals	
5	M	AB	55	126		Class Width	
6	M	A	42	131		Round Up (i)	
7	F	O	40	125		Lowest Value + (i-1)	
8	F	O	35	131			
9	M	A	34	115		Bin Range	Class Mark or Midpoints
10	F	AB	29	125			
11	F	A	50	163			
12	F	AB	30	125			
13	F	O	34	114			
14	M	O	41	132			
15	M	A	33	105			
16	F	AB	39	110			
17	M	AB	35	133			
18	M	A	43	150			
19	F	A	20	109			
20	F	O	25	115			
21	F	O	39	139			
22	M	O	72	142			
23	M	O	41	146			
24	F	A	46	154			
25	F	B	32	116			
26	M	AB	37	105			

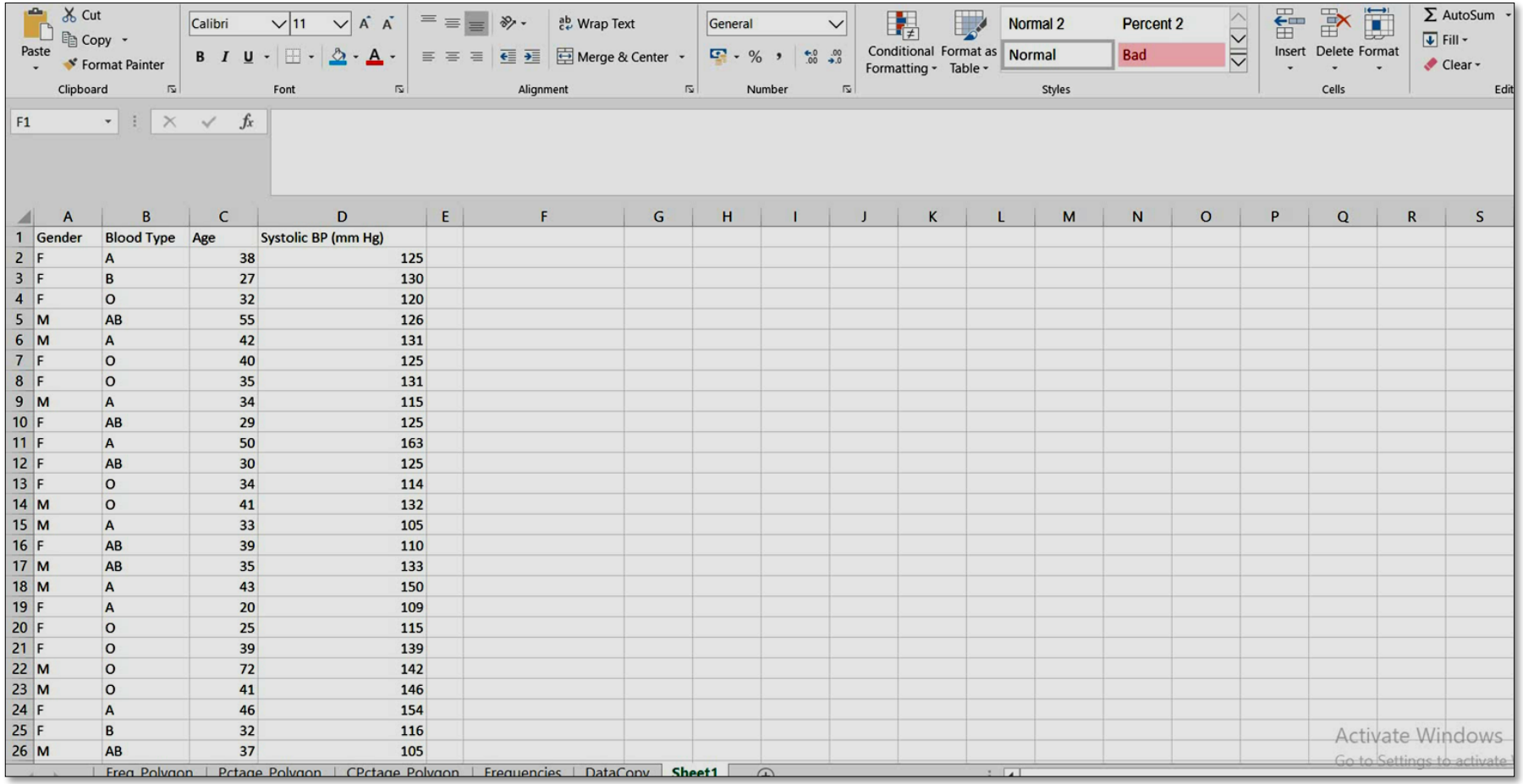
I. Tabular Presentation Contingency Table

A **contingency table** (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables.

The table allows users to see at a glance the proportion of men and women who have different blood type.

Gender	Blood Type				
	A	AB	B	O	Grand Total
Female	6	7	6	11	30
Male	7	3	0	10	20
Grand Total	13	10	6	21	50

Two-Way Chart



	A	B	C	D
1	Gender	Blood Type	Age	Systolic BP (mm Hg)
2	F	A	38	125
3	F	B	27	130
4	F	O	32	120
5	M	AB	55	126
6	M	A	42	131
7	F	O	40	125
8	F	O	35	131
9	M	A	34	115
10	F	AB	29	125
11	F	A	50	163
12	F	AB	30	125
13	F	O	34	114
14	M	O	41	132
15	M	A	33	105
16	F	AB	39	110
17	M	AB	35	133
18	M	A	43	150
19	F	A	20	109
20	F	O	25	115
21	F	O	39	139
22	M	O	72	142
23	M	O	41	146
24	F	A	46	154
25	F	B	32	116
26	M	AB	37	105

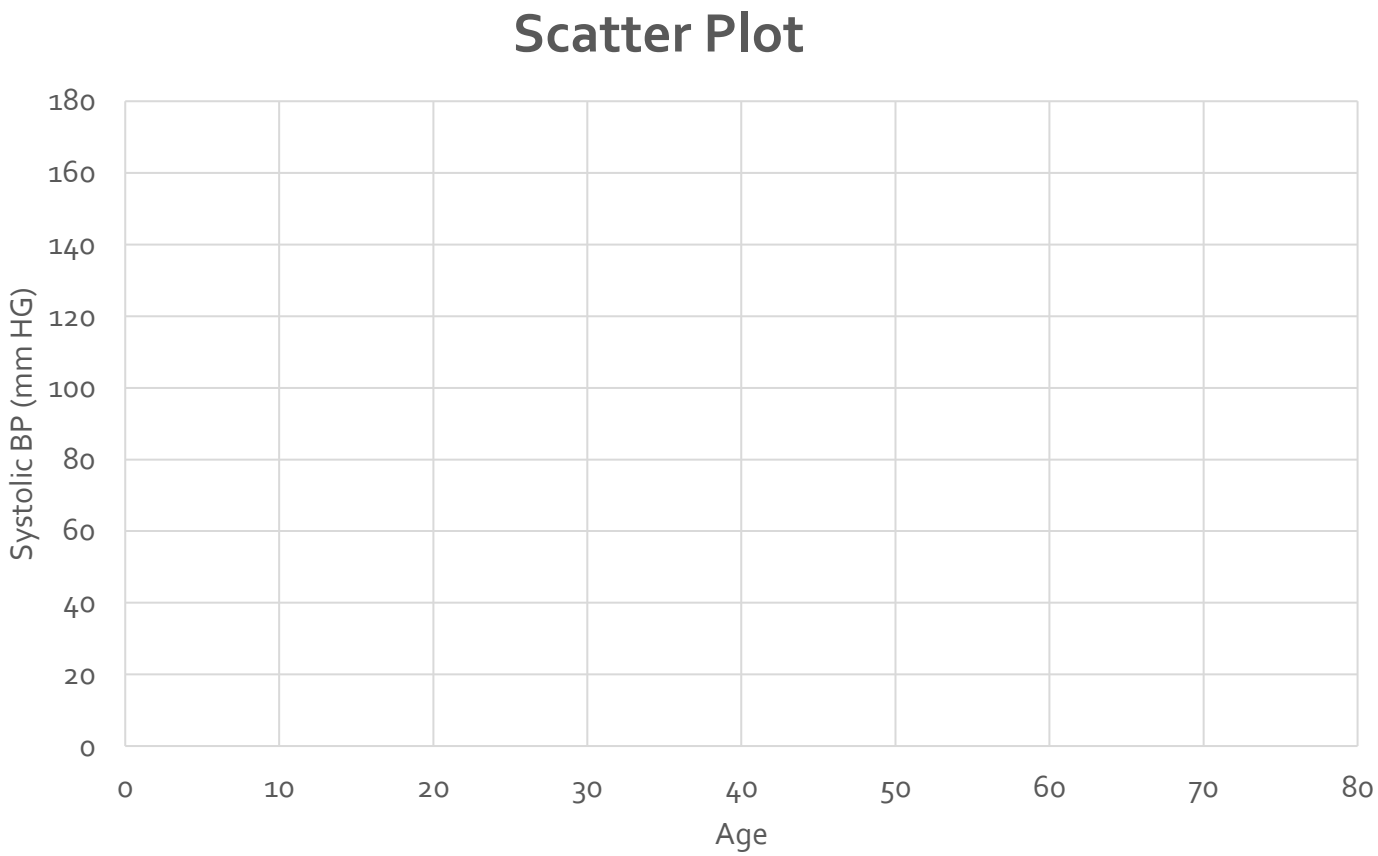
Two-Way Chart in R

library(gmodels)
CrossTable(patients$Gender, patients$BloodType)

```
> CrossTable(patients$Gender, patients$Blood.Type)


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  50


                | patients$Blood.Type
patients$Gender |         A |         AB |         B |         O | Row Total |
----------------|-----------|------------|-----------|-----------|-----------|
              F |         6 |          7 |         6 |        11 |        30 |
                |     0.415 |      0.167 |     1.600 |     0.203 |           |
                |     0.200 |      0.233 |     0.200 |     0.367 |     0.600 |
                |     0.462 |      0.700 |     1.000 |     0.524 |           |
                |     0.120 |      0.140 |     0.120 |     0.220 |           |
----------------|-----------|------------|-----------|-----------|-----------|
              M |         7 |          3 |         0 |        10 |        20 |
                |     0.623 |      0.250 |     2.400 |     0.305 |           |
                |     0.350 |      0.150 |     0.000 |     0.500 |     0.400 |
                |     0.538 |      0.300 |     0.000 |     0.476 |           |
                |     0.140 |      0.060 |     0.000 |     0.200 |           |
----------------|-----------|------------|-----------|-----------|-----------|
   Column Total |        13 |         10 |         6 |        21 |        50 |
                |     0.260 |      0.200 |     0.120 |     0.420 |           |
----------------|-----------|------------|-----------|-----------|-----------|
```

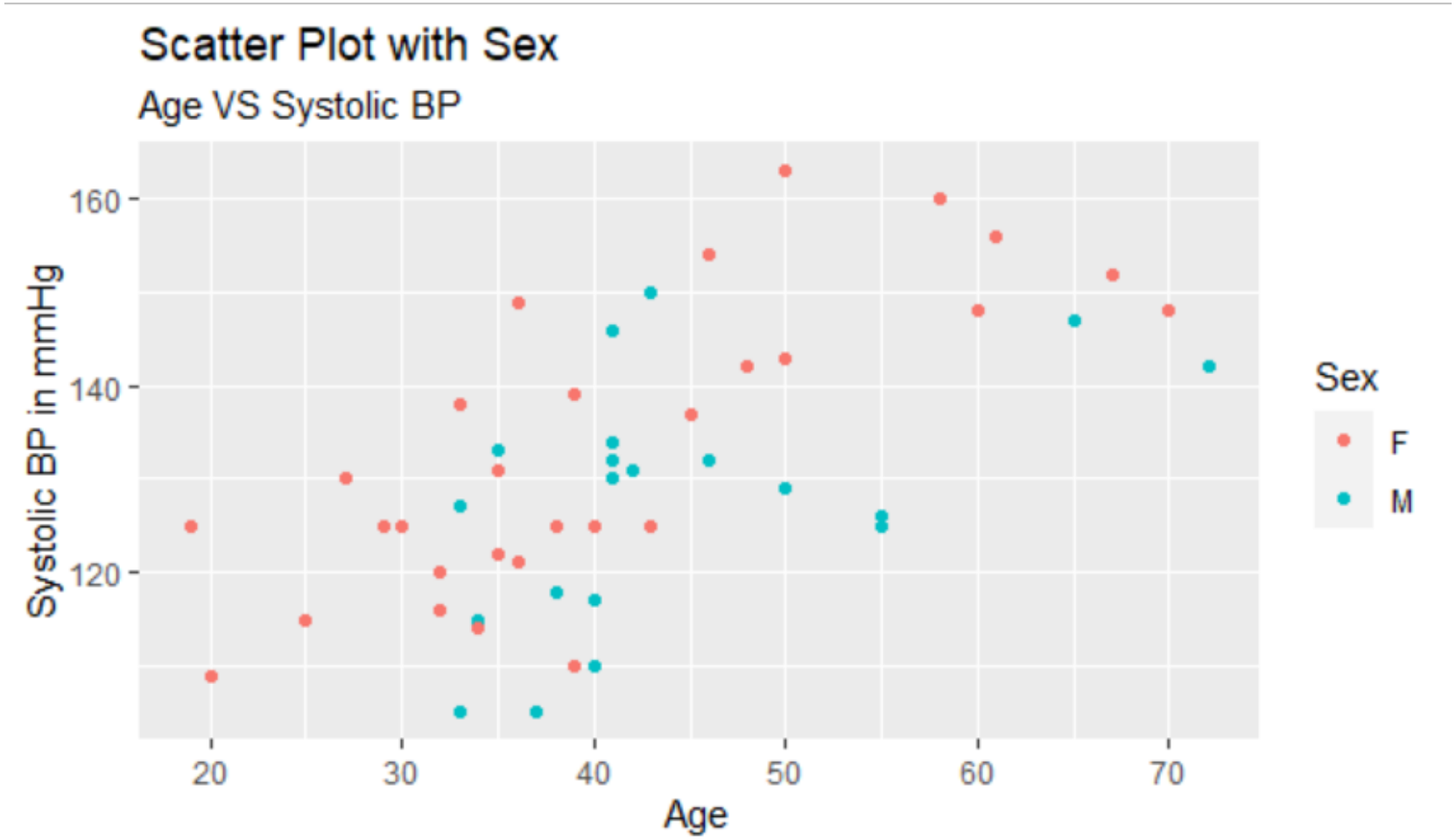
Descriptive Statistics

- **Graphs**
  - **Scatter Plot**

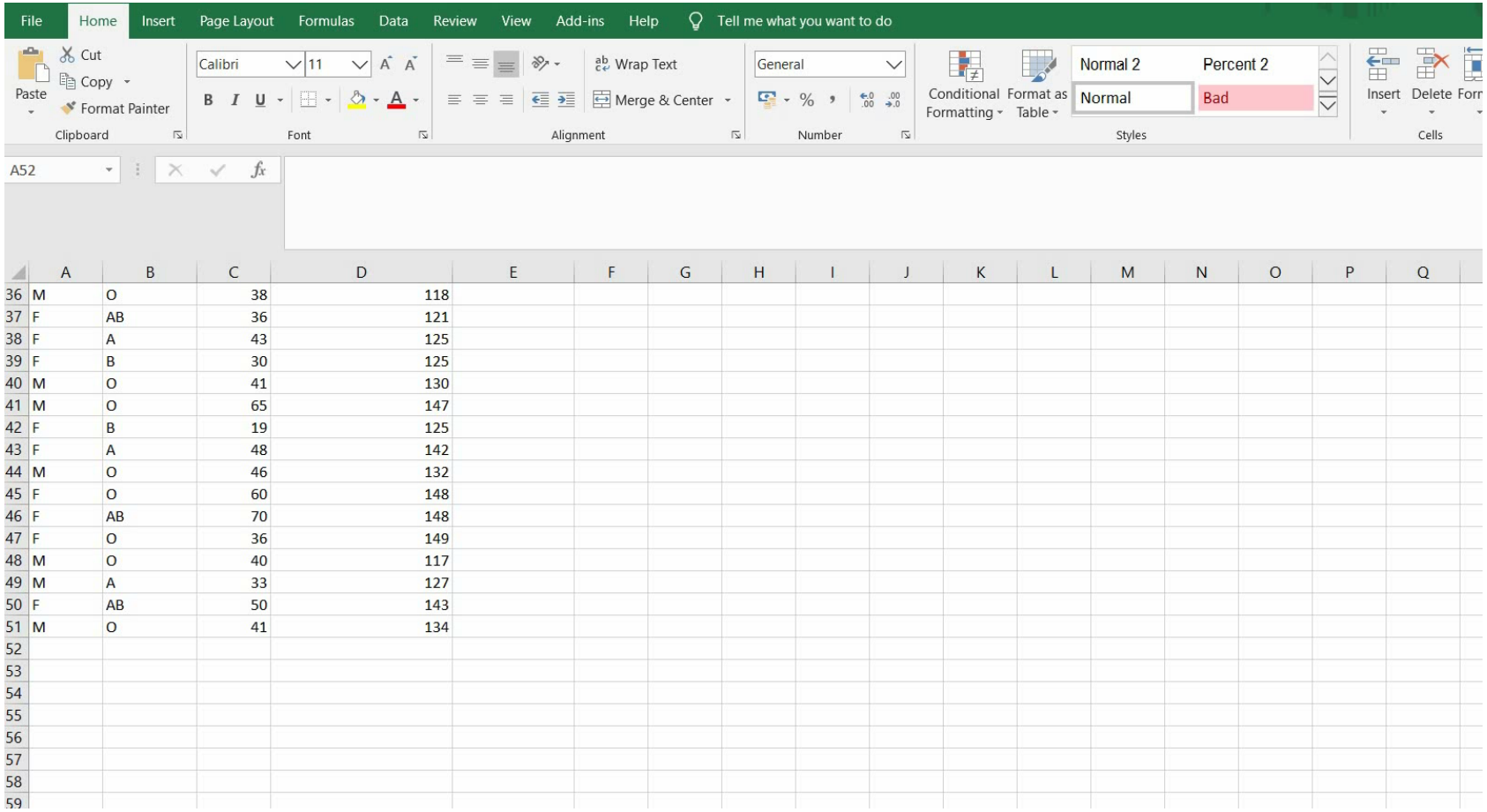


#Scatter Plot in R

```
ggplot(patients, aes(Age, SystolicBPmmHg, color=Gender))+
 geom_point()+
 labs(title = "Scatter Plot with Gender", subtitle ="Age VS Systolic BP", y="Systolic BP in mmHg", color='sex')
```



Scatter Plot using PHStat



	A	B	C	D
36	M	O	38	118
37	F	AB	36	121
38	F	A	43	125
39	F	B	30	125
40	M	O	41	130
41	M	O	65	147
42	F	B	19	125
43	F	A	48	142
44	M	O	46	132
45	F	O	60	148
46	F	AB	70	148
47	F	O	36	149
48	M	O	40	117
49	M	A	33	127
50	F	AB	50	143
51	M	O	41	134
52				
53				
54				
55				
56				
57				
58				
59				

Descriptive Statistics

- A descriptive measure computed from the data of a sample is called a **statistic**.
  - Average systolic BP of walk-in patients
- A descriptive measure computed from the data of a population is called a **parameter**.
  - Average age of COVID-19 infected in Cebu city

Descriptive Statistics

Measures of Central Tendency

1. **Mean** – average value
   - Formula:

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Properties

a. Uniqueness
b. Simplicity
c. It is affected by extreme values.

Descriptive Statistics

Measures of Central Tendency

1. **Mean**

Example: Find the mean of these numbers:

3, −7, 5, 13, −2

Solution:

$$\bar{x} = \frac{3 + (-7) + 5 + 13 + (-2)}{5} = \frac{12}{5} = \mathbf{2.4}$$

Descriptive Statistics

Measures of Central Tendency

2. **Median** – middle value of an ordered array

- First, arrange the observation in an ascending or descending order.
- The location of the median = $\frac{\mathrm{n}+1}{2}\,th\ place$, where **n** is the total number of data values in the sample.

OR

- If **n** is odd, the median is the middle value.
- If **n** is even, the median is the average of the two middle values.

Properties

a. Uniqueness
b. Simplicity
c. It is not affected by extreme values.

Descriptive Statistics

Measures of Central Tendency

2. Median

Example: Find the median of these numbers:

3, −7, 5, 13, −2

Solution:

- First, arrange the observation in an ascending or descending order.

-7, -2, 3, 5, 13

- The location of the median = $\frac{5+1}{2} th\ place = 3^{rd}\ place$, where **5** is the total number of data values in the sample.

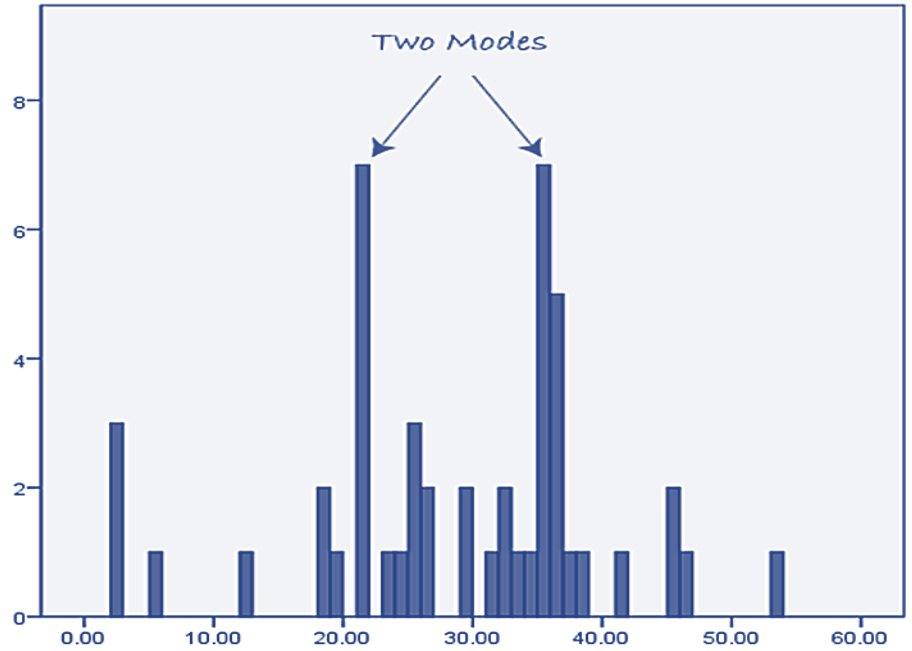
Median = 3

- Or simply, the middle value of the arranged dataset which is 3.

Descriptive Statistics

Measures of Central Tendency

3. **Mode** – occurs most frequently value
   - The most frequent score in the dataset.
   - On a histogram it represents the highest bar in a bar chart or histogram.
   - You can, therefore, sometimes consider the mode as being the most popular option.



Descriptive Statistics

Measures of Dispersion

1. **Range** – difference between the largest and smallest values

   R = LV - SV

2. **Variance** – subtracting the mean from each of the values, square the resulting differences, and get the average of the resulting difference (divisor is N if solving for population variance and n-1 for a sample variance)

Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$

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

Descriptive Statistics

Measures of Dispersion

3. **Standard Deviation** – the square root of variance

Population Standard Deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Sample Standard Deviation:

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

Descriptive Statistics

Measures of Dispersion

4. **Coefficient of Variation** – quotient of standard deviation and mean
   - Use to compare variability of datasets
   - The formula is given as

$$CV = \frac{s}{\bar{x}} X\ 100\ \%$$

Descriptive Statistics

Example

Find the range, variance, standard deviation, and coefficient of variation of the given sample dataset:

9, 2, 5, 4, 12, 7, 8, 11

Solution

- Range $\rightarrow R = 12 - 2 = 10$
- Variance $\rightarrow$ First find the mean. $\bar{x} = \frac{9+2+5+4+12+7+8+11}{8} = \frac{58}{8} = 7.25$

$$\boldsymbol{s^2 = \frac{(9-7.25)^2+(2-7.25)^2+\cdots+(8-7.25)^2+(11-7.25)^2}{8-1} = 11.93}$$

- Standard Deviation $\rightarrow s = \sqrt{11.93} = 3.45$
- Coefficient of Variation $\rightarrow cv = \frac{3.45}{7.25} \times 100\% = 47.6\%$

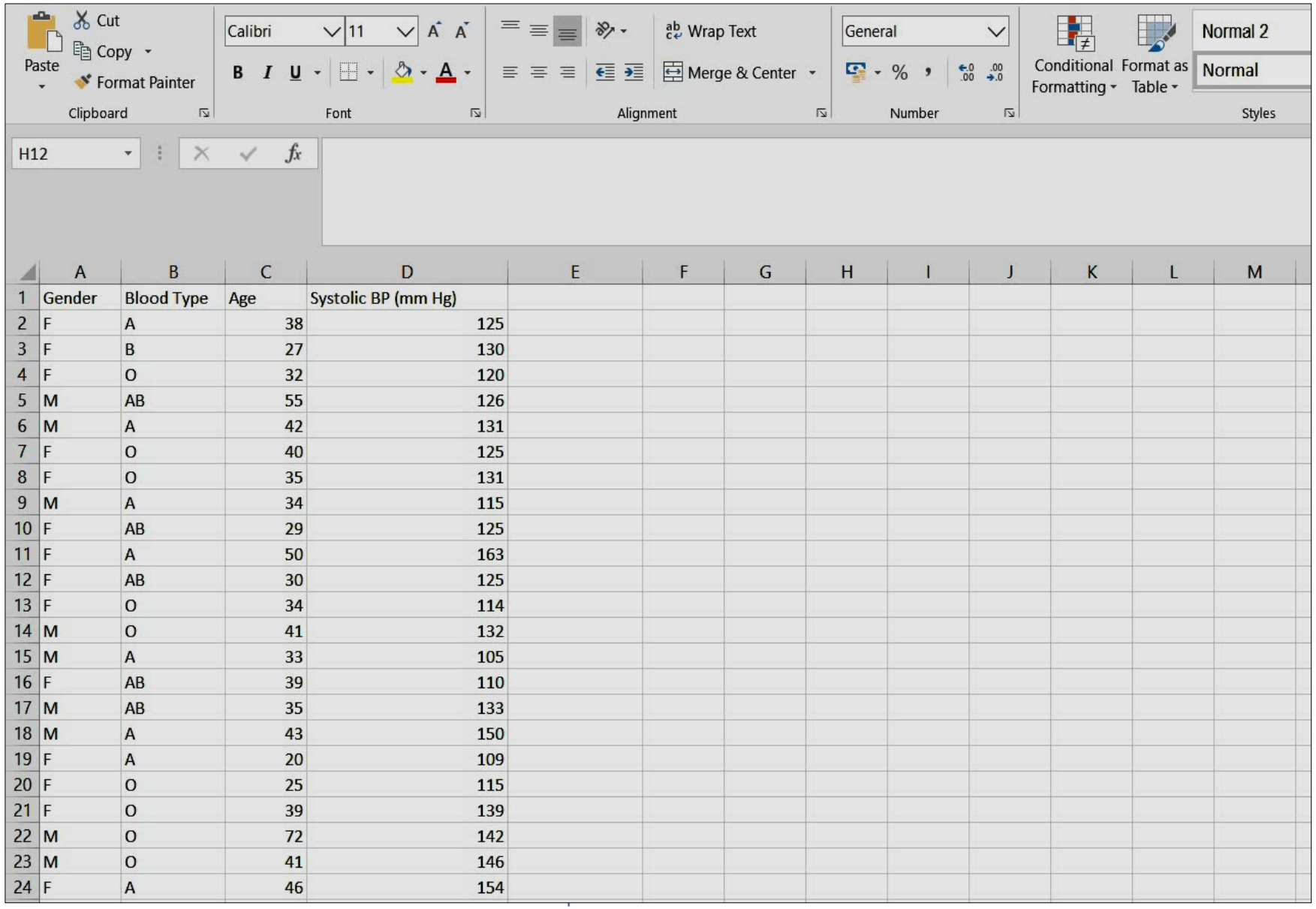
Descriptive Statistics

PHStat Output for Multiple Descriptive Summary of Example-Patients Data

Descriptive Summary	Age	Systolic BP (mm Hg)
Mean	41.78	130.92
Median	40	129.5
Mode	41	125
Minimum	19	105
Maximum	72	163
Range	53	58
Variance	147.9710	214.2792
Standard Deviation	12.1643	14.6383
Coeff. of Variation	29.12%	11.18%
Skewness	0.7187	0.2812
Kurtosis	0.2797	-0.6196
Count	50	50
Standard Error	1.7203	2.0702

Descriptive Statistics

Using PHStat



	A	B	C	D
1	Gender	Blood Type	Age	Systolic BP (mm Hg)
2	F	A	38	125
3	F	B	27	130
4	F	O	32	120
5	M	AB	55	126
6	M	A	42	131
7	F	O	40	125
8	F	O	35	131
9	M	A	34	115
10	F	AB	29	125
11	F	A	50	163
12	F	AB	30	125
13	F	O	34	114
14	M	O	41	132
15	M	A	33	105
16	F	AB	39	110
17	M	AB	35	133
18	M	A	43	150
19	F	A	20	109
20	F	O	25	115
21	F	O	39	139
22	M	O	72	142
23	M	O	41	146
24	F	A	46	154

#Descriptive Summary

```
install.packages('pastecs')
library(pastecs)
res <- stat.desc(patients[,3:4])
round(res, 2)
```

```
> res <- stat.desc(patients[,3:4])
> round(res, 2)
                    Age SystolicBPmmHg
nbr.val           50.00           50.00
nbr.null           0.00            0.00
nbr.na             0.00            0.00
min               19.00          105.00
max               72.00          163.00
range             53.00           58.00
sum             2089.00         6546.00
median            40.00          129.50
mean              41.78          130.92
SE.mean            1.72            2.07
CI.mean.0.95       3.46            4.16
var              147.97          214.28
std.dev           12.16           14.64
coef.var           0.29            0.11
```

Descriptive Statistics

Measures of Location

- **Quartile** – ordered array is divided into 4 groups
- **Percentile** – ordered array is divided into 100 groups
  - Given a set of n observations $x_1, x_2, \cdots, x_n$, the p^{th} percentile P is the value of X such that p percent or less of the observations are less than P and (100-p) percent or less of the observations are greater than P.

Location: $L_i = \frac{i}{r}(n+1)^{th}$ position

✓ i = 1, 2, 3, 4 for quartiles Q_r, r = 4

✓ i = 1, 2, 3, …, 100 for percentiles P_r, r = 100

Descriptive Statistics

Measures of Location

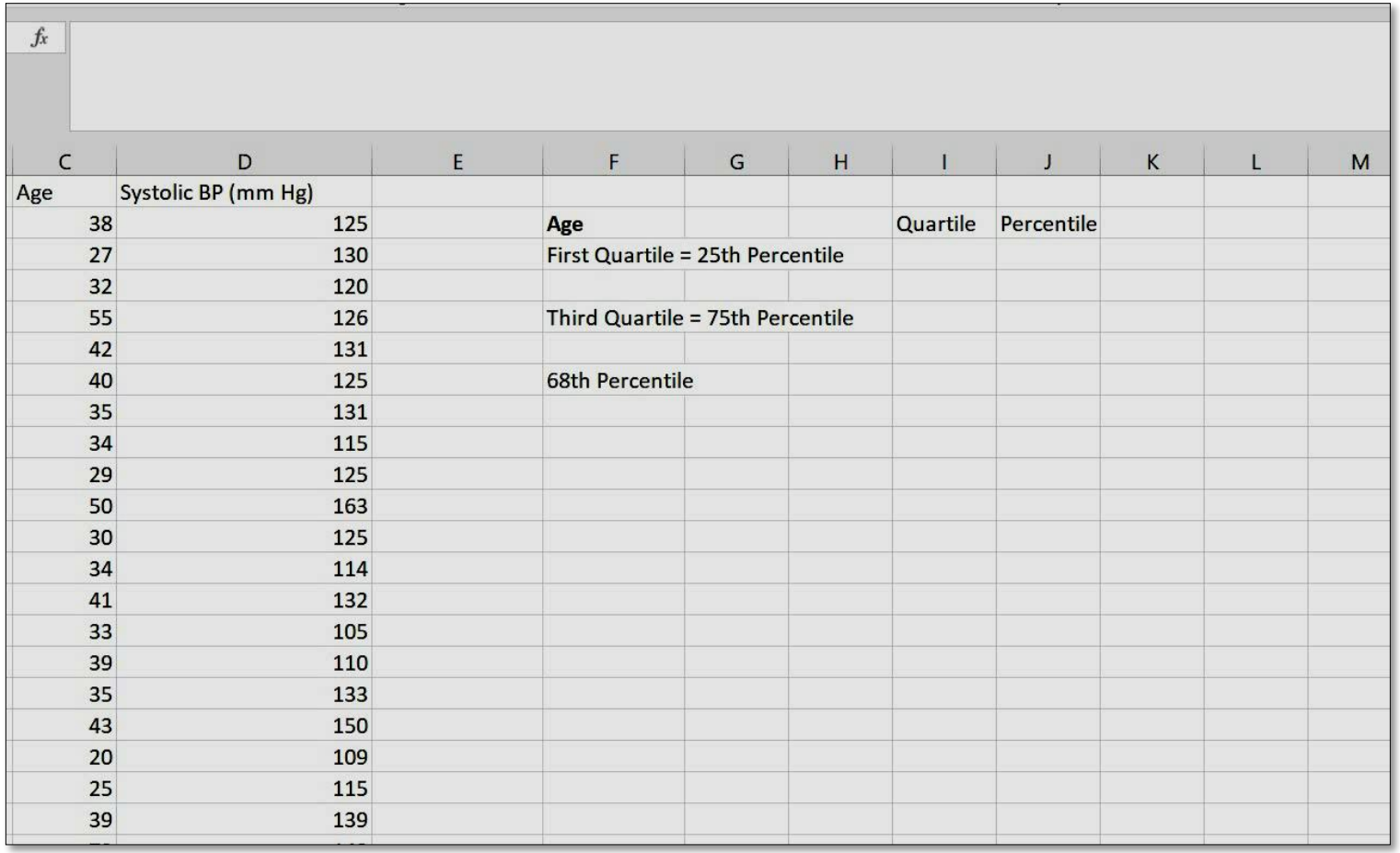
Example

Find the first and third quartiles of the following:

1, 1, 3, 4, 6, 7, 8, 8

- $Q_1 \rightarrow L_1 = \frac{1(8+1)^{th}position}{4} = 2.25^{th}position$ → Any number between 1 and 3.
  $Q_1 = 2$, 25% of the dataset is below 2 and 75% is above 2.
- $Q_3 \rightarrow L_3 = \frac{3(8+1)^{th}position}{4} = 6.75^{th}position$ → Any number between 7 and 8.
  $Q_3 = 7.5$, 75% of the dataset is below 7.5 and 25% is above 7.5.

Using Excel



C	D	E	F	G	H	I	J	K	L	M
Age	Systolic BP (mm Hg)									
38	125		**Age**			Quartile	Percentile			
27	130		First Quartile = 25th Percentile							
32	120									
55	126		Third Quartile = 75th Percentile							
42	131									
40	125		68th Percentile							
35	131									
34	115									
29	125									
50	163									
30	125									
34	114									
41	132									
33	105									
39	110									
35	133									
43	150									
20	109									
25	115									
39	139									

#Measures of Location in R

summary(patients) #quartiles

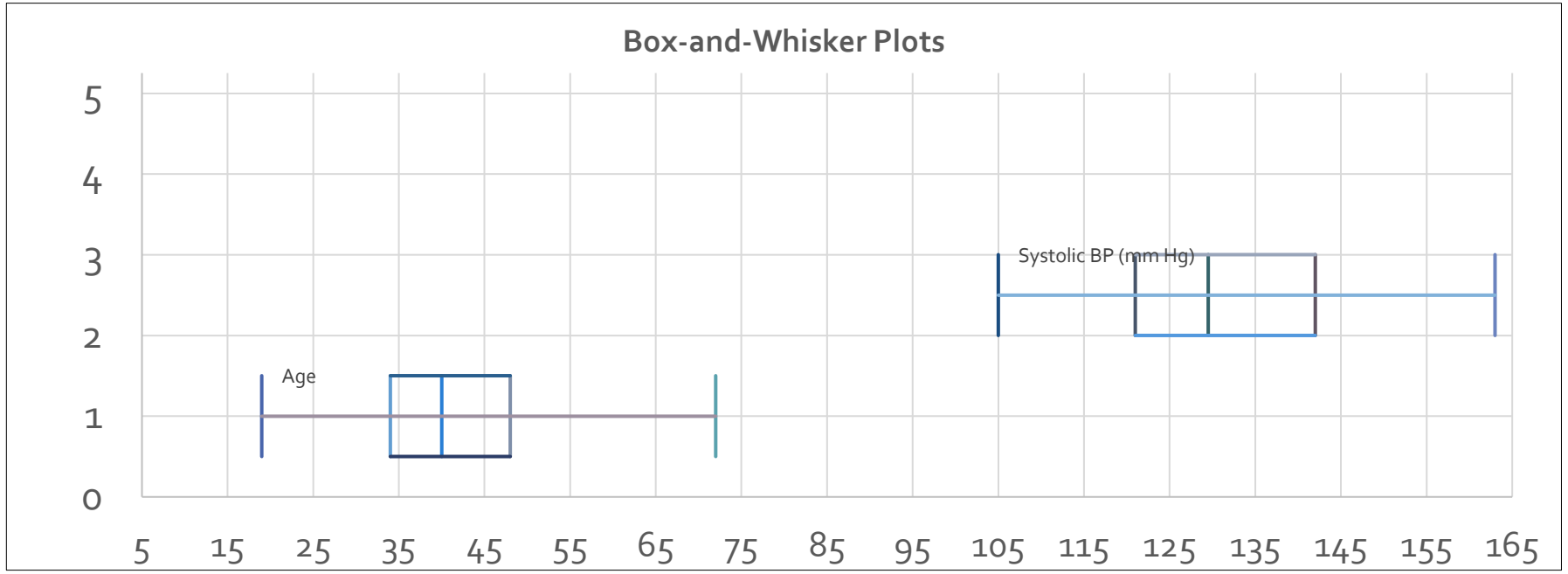
sapply(patients[, 3:4], quantile)#0,25,50,75,100th Percentile

sapply(patients[, 3:4], quantile, c(.32, .57, .98))#Specific Percentile

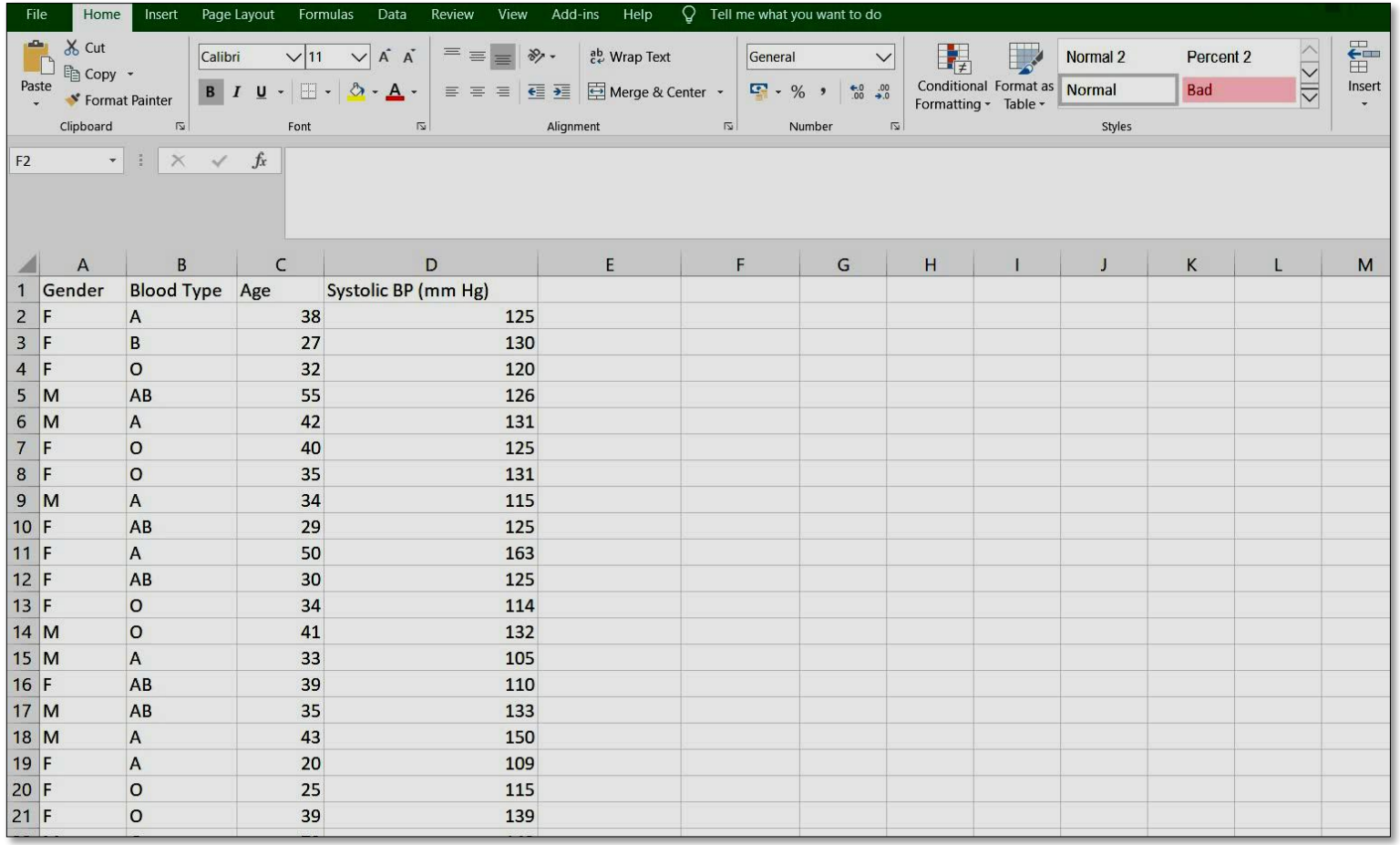
```
> summary(patients) #quartiles
    Gender            BloodType              Age         SystolicBPmmHg
 Length:50          Length:50          Min.   :19.00   Min.   :105.0
 Class :character   Class :character   1st Qu.:34.00   1st Qu.:121.2
 Mode  :character   Mode  :character   Median :40.00   Median :129.5
                                       Mean   :41.78   Mean   :130.9
                                       3rd Qu.:47.50   3rd Qu.:142.0
                                       Max.   :72.00   Max.   :163.0
> sapply(patients[, 3:4], quantile)#0,25,50,75,100th Percentile
      Age SystolicBPmmHg
0%   19.0         105.00
25%  34.0         121.25
50%  40.0         129.50
75%  47.5         142.00
100% 72.0         163.00
> sapply(patients[, 3:4], quantile, c(.32, .57, .98))#Specific Percentile
      Age SystolicBPmmHg
32% 35.00         125.00
57% 41.00         131.00
98% 70.04         160.06
```

Descriptive Statistics

- The **interquartile range (IQR)** is the difference between the third and first quartiles: that is, **IQR = $\mathbf{Q_3 - Q_1}$**. It is a measure of variability, based on dividing a data set into quartiles.
- **Box-and-Whisker Plots** – a plot that contains the 5-number summary (SV, Q_1, Q_2 = Median, Q_3, LV)



Box-and-Whiskers Plot Using PHStat

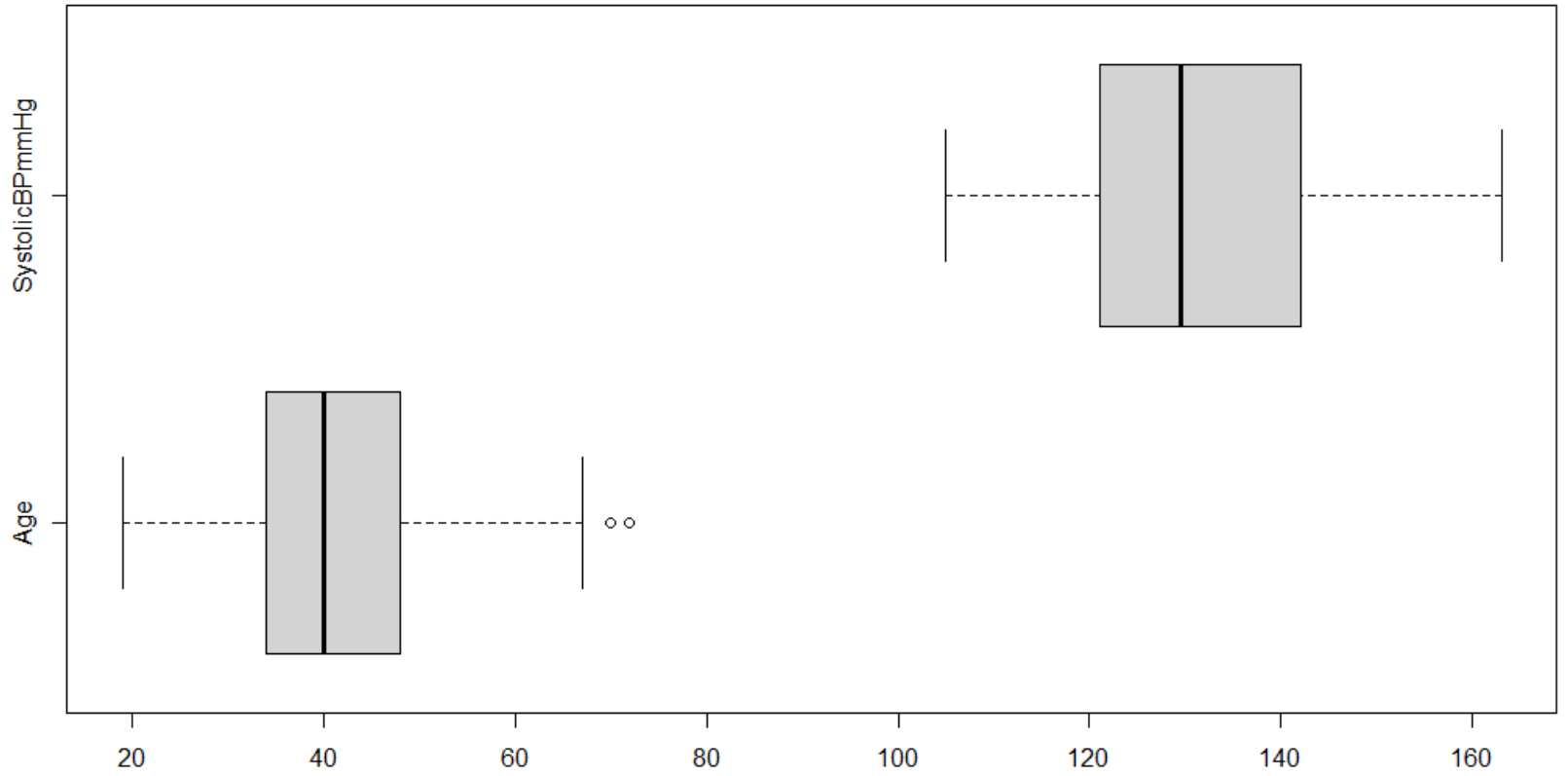


	A	B	C	D
1	Gender	Blood Type	Age	Systolic BP (mm Hg)
2	F	A	38	125
3	F	B	27	130
4	F	O	32	120
5	M	AB	55	126
6	M	A	42	131
7	F	O	40	125
8	F	O	35	131
9	M	A	34	115
10	F	AB	29	125
11	F	A	50	163
12	F	AB	30	125
13	F	O	34	114
14	M	O	41	132
15	M	A	33	105
16	F	AB	39	110
17	M	AB	35	133
18	M	A	43	150
19	F	A	20	109
20	F	O	25	115
21	F	O	39	139

Box-and-Whiskers Plot UsingR

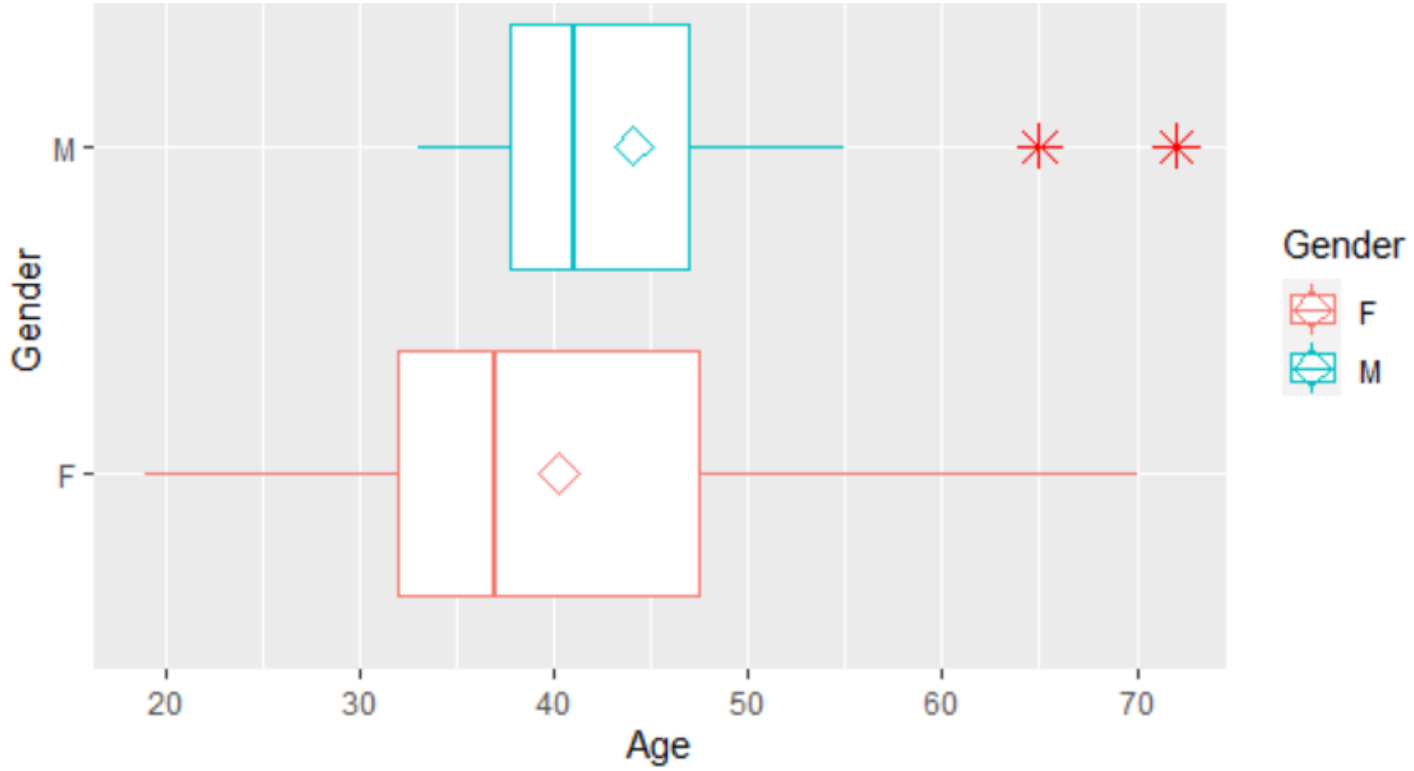
#Box Plots

boxplot(patients[,3:4], horizontal=T)



#Boxplots with bold outliers+mean

```
ggplot(patients, aes(x=Age, y=Gender, color=Gender)) +
 geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4) +
 stat_summary(fun=mean, geom="point", shape=23, size=4)
```



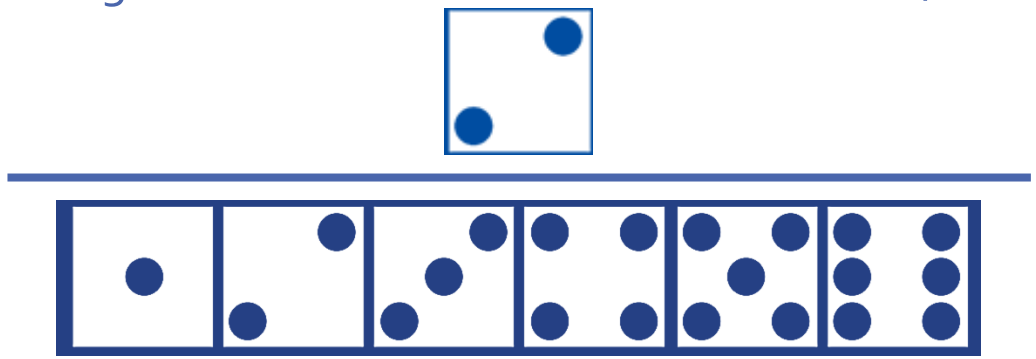
Some Basic Probability Concepts

Classical Probability

- If an event can occur in **N** mutually exclusive and equally likely ways, and if **m** of these possess a trait, **E**, the probability of the occurrence of **E** is equal to **m/N**.
- It is written as **P(E) =** $\frac{\mathbf{m}}{\mathbf{N}}$.

Example

- The probability of getting a 2 when a fair 6-sided die is rolled is 1/6.



Subjective Probability

- "personalistic"

Example

- I am 90% certain that today is going to rain.

Some Basic Probability Concepts

Bayesian Methods

- **Bayesian methods** are named in honor of the Reverend Thomas Bayes (1702–1761), an English clergyman who had an interest in mathematics.
- **Bayesian methods** are an example of subjective probability, since it takes into consideration the degree of belief that one has in the chance that an event will occur.
- While probabilities based on classical or relative frequency concepts are designed to allow for decisions to be made solely on the basis of collected data, Bayesian methods make use of what are known as *prior probabilities* and *posterior probabilities*.

Some Basic Probability Concepts

Bayesian Methods

- The **prior probability** of an event is a probability based on prior knowledge, prior experience, or results derived from prior data collection activity.
- The **posterior probability** of an event is a probability obtained by using new information to update or revise a prior probability.

Some Basic Probability Concepts

Elementary Properties of Probability

1. Given some process (or experiment) with n mutually exclusive outcomes (called events), $\mathbf{E_1}, \mathbf{E_2}, \cdots, \mathbf{E_n}$, the probability of any event $\mathbf{E_i}$ is assigned a nonnegative number. That is, $\mathbf{P(E_i) \geq 0}$.
2. The sum of the probabilities of the mutually exclusive outcomes is equal to 1. $\mathbf{P(E_1) + P(E_2) + \cdots + P(E_n) = 1}$.
3. Consider any two mutually exclusive events $\mathbf{E_i}$ and $\mathbf{E_j}$. The probability of the occurrence of either $\mathbf{E_i}$ or $\mathbf{E_j}$ is equal to the sum of their individual probabilities.

$$\mathbf{P(E_i \text{ or } E_j) = P(E_i) + P(E_j)}$$

Some Basic Probability Concepts

Conditional Probability

If E_i and E_j are two events in a sample space S, then the **conditional probability** of E_i given E_j is defined as

$$P(E_i|E_j) = \frac{P(E_i \text{ and } E_j)}{P(E_i)}, \text{ when } P(E_i) > 0.$$

- The conditional probability of A given B is equal to the probability of A ∩ B (A and B) divided by the probability of B, provided the probability of B is not zero.

Joint Probability

The probability that a subject picked at random from a group of subjects possesses two characteristics at the same time. Such a probability is referred to as a **joint probability**.

Some Basic Probability Concepts

Cold Length	Medicine Taken		Total
	Yes	No	
1 – 3 days	85	20	105
4 – 7 days	17	78	95
Total	102	98	200

Contingency Table

1. What is the probability that a person has a cold and not taking a medicine for 4 - 7 days?
2. What is the probability that a person has a cold for 1 – 3 days provided he is not taking medicine?
3. What is the probability that a person who is not taking medicine given he has a cold for 1 – 3 days?
4. Which one has a greater probability no. 2 or no. 3?

Some Basic Probability Concepts

The Multiplication Rule

For any two events A and B,

$$P(A \cap B) = P(A)P(A|B), \text{ if } P(B) \neq 0 \text{ or}$$
$$P(A \cap B) = P(A)P(B|A), \text{ if } P(A) \neq 0$$

- *Refer to the contingency table, what is the probability that two persons have colds in 4 – 7 days and not taking medicine?*

Some Basic Probability Concepts

The Addition Rule

Given two events A and B, the probability that event A or event B, or both occur is equal to the probability that event A occurs, plus the probability that event B occurs, minus the probability that the events occur simultaneously.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- *Refer to the contingency table, what is the probability that a person has a cold in 4 – 7 days or not taking medicine?*

Some Basic Probability Concepts

Independent Events

Suppose that the probability of event A is the same regardless of whether or not B occurs, i.e. P(A|B) = P(A). We say that A and B are independent events.

In multiplication rule for two independent events,

$$P(A \cap B) = P(A) \cdot P(B), \quad P(A) \neq 0 \text{ and } P(B) \neq 0.$$

- *Refer to the contingency table, if two persons are chosen for free medical consultation, what is the probability that the first person is taking medicine and the other has not?*

Some Basic Probability Concepts

Complementary Events

The probability of an event A is equal to 1 minus the probability of its complement, which is written $\bar{A}$, and P($\bar{A}$) = 1-P(A).

- *Refer to the contingency table, what is the probability that the first person is not taking medicine?*

Some Basic Probability Concepts

- **Marginal Probability**
  Given some variable that can be broken down into **m** categories designated by $\mathbf{A_1, A_2, \cdots, A_i, \cdots, A_m}$ and another jointly occurring variable that is broken down into **n** categories designated by $\mathbf{B_1, B_2, \cdots, B_j, \cdots, B_n}$, the marginal probability of $\mathbf{A_i}$, $P(A_i)$, is equal to the sum of the joint probabilities of **A**, with all categories of **B**. That is,

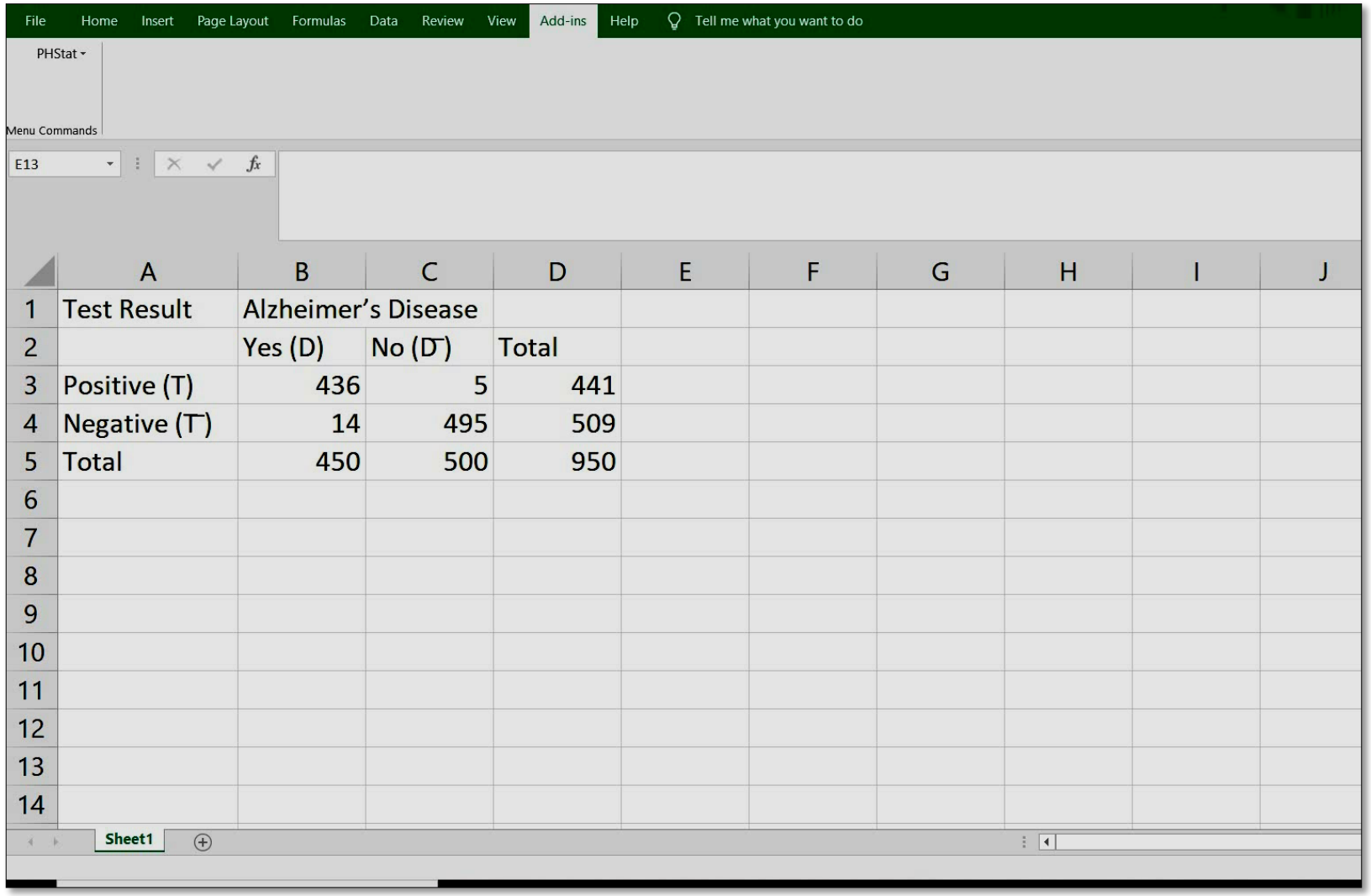
  $$\mathbf{P(A_i) = \sum P(A_i \cap B_j)}, \text{ for all values of } \mathbf{j}.$$

  - *Refer to the contingency table, what is the probability that the first person is taking medicine?*

Using PHStat

Probabilities Calculations				
Sample Space		Event B		
		B1	B2	Totals
Event A	A1	85	20	105
	A2	17	78	95
	Totals	102	98	200
Simple Probabilities				
P(A1)	0.53			
P(A2)	0.48			
P(B1)	0.51			
P(B2)	0.49			
Joint Probabilities				
P(A1 and B1)	0.43			
P(A1 and B2)	0.10			
P(A2 and B1)	0.09			
P(A2 and B2)	0.39			
Addition Rule				
P(A1 or B1)	0.61			
P(A1 or B2)	0.92			
P(A2 or B1)	0.90			
P(A2 or B2)	0.58			

Using PHStat



	A	B	C	D	E	F	G	H	I	J
1	Test Result	Alzheimer's Disease								
2		Yes (D)	No ($\bar{D}$)	Total						
3	Positive (T)	436	5	441						
4	Negative ($\bar{T}$)	14	495	509						
5	Total	450	500	950						
6										
7										
8										
9										
10										
11										
12										
13										
14										

Bayes' Rule

- Bayesian statistics is a collection of tools that is used in a special form of statistical inference which applies in the analysis of experimental data in many practical situations in science and engineering.
- Bayes' rule is one of the most important rules in probability theory.

Bayes' Rule

(Bayes' Rule) If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S such that $P(A) \neq 0$,

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \quad \text{for } r = 1, 2, \ldots, k.$$

Bayes' Rule

Example

A manufacturing firm employs three analytical plans for the design and development of a particular product. For cost reasons, all three are used at varying times. In fact, plans 1, 2, and 3 are used for 30%, 20%, and 50% of the products, respectively. The defect rate is different for the three procedures as follows:

$$P(D|P1) = 0.01,\ P(D|P2) = 0.03,\ P(D|P3) = 0.02,$$

where $P(D|P_j)$ is the probability of a defective product, given plan j. If a random product was observed and found to be defective, which plan was most likely used and thus responsible?

Bayes' Rule

Solution

From the statement of the problem

$$P(P1) = 0.30,\ P(P2) = 0.20, \text{ and } P(P3) = 0.50,$$

$$P(D|P1) = 0.01,\ P(D|P2) = 0.03,\ P(D|P3) = 0.02,$$

Find $P(P_j|D)$ for j = 1, 2, 3.

By the rule,

- $$P(P1|D) = \frac{P(P1)P(D|P1)}{P(P1)P(D|P1) + P(P2)P(D|P2) + P(P3)P(D|P3)}$$

$$= \frac{(0.30)(0.01)}{(0.3)(0.01) + (0.20)(0.03) + (0.50)(0.02)} = \frac{0.003}{0.019} = 0.158.$$

Similarly,

- $P(P2|D) = \frac{(0.03)(0.20)}{0.019} = 0.316$ and $P(P3|D) = \frac{(0.02)(0.50)}{0.019} = 0.526$

The conditional probability of a defect given plan 3 is the largest of the three; thus a defective for a random product is most likely the result of the use of plan 3.

Probability Distributions

Discrete Variables

The set of ordered pairs $(x, f(x))$ is a **probability function, probability mass function**, or **probability distribution** of the discrete random variable X if, for each possible outcome x,

1. $f(x) \geq 0$,
2. $\sum_x f(x) = 1$,
3. $P(X = x) = f(x)$.

Probability Distributions

Discrete Variables

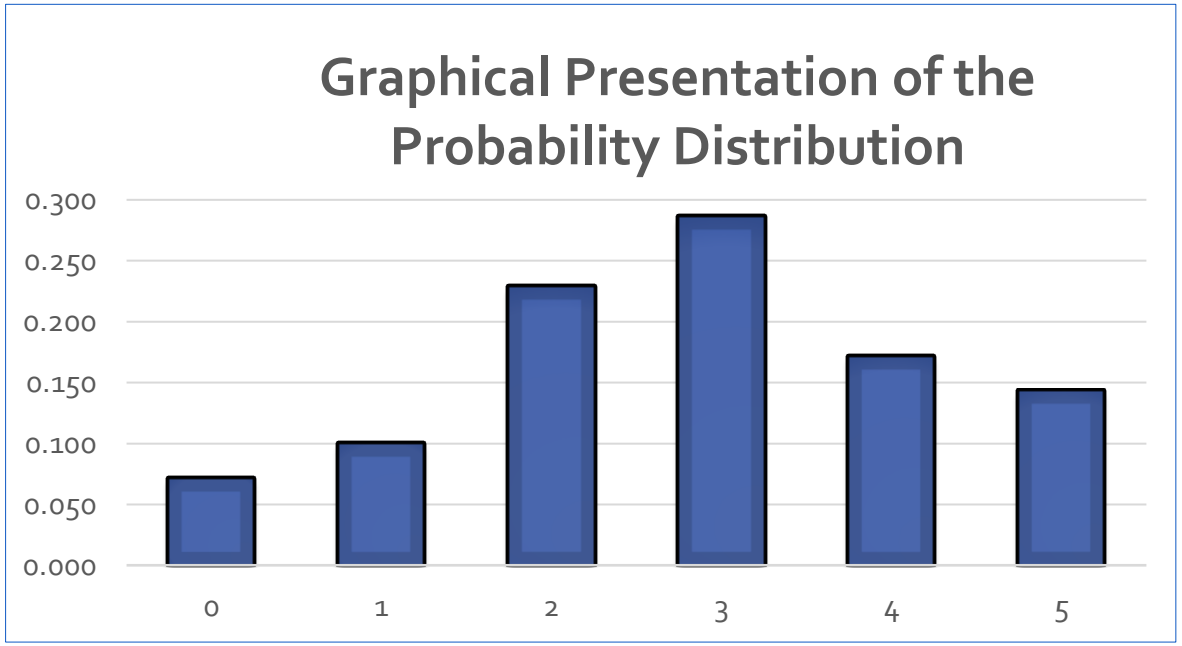
- The **probability distribution of a discrete random variable** is a table, graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.
  - Probability distribution of the number of siblings of 70 families:

Number of Siblings (x)	P(X = x)
0	0.0714
1	0.1000
2	0.2286
3	0.2857
4	0.1714
5	0.1429
Total	1.0000

Probability Distributions

Discrete Variables

- The **probability distribution of a discrete random variable** is a table, graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.
  - Probability distribution of the number of siblings of 70 families:



Probability Distributions

Discrete Variables - Cumulative Distribution

The **cumulative distribution function** $F(x)$ of a discrete random variable X with probability distribution $f(x)$ is

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for } -\infty < x < \infty.$$

Probability Distributions

Discrete Variables

Cumulative Distribution

- It is obtained by successively adding the probabilities.

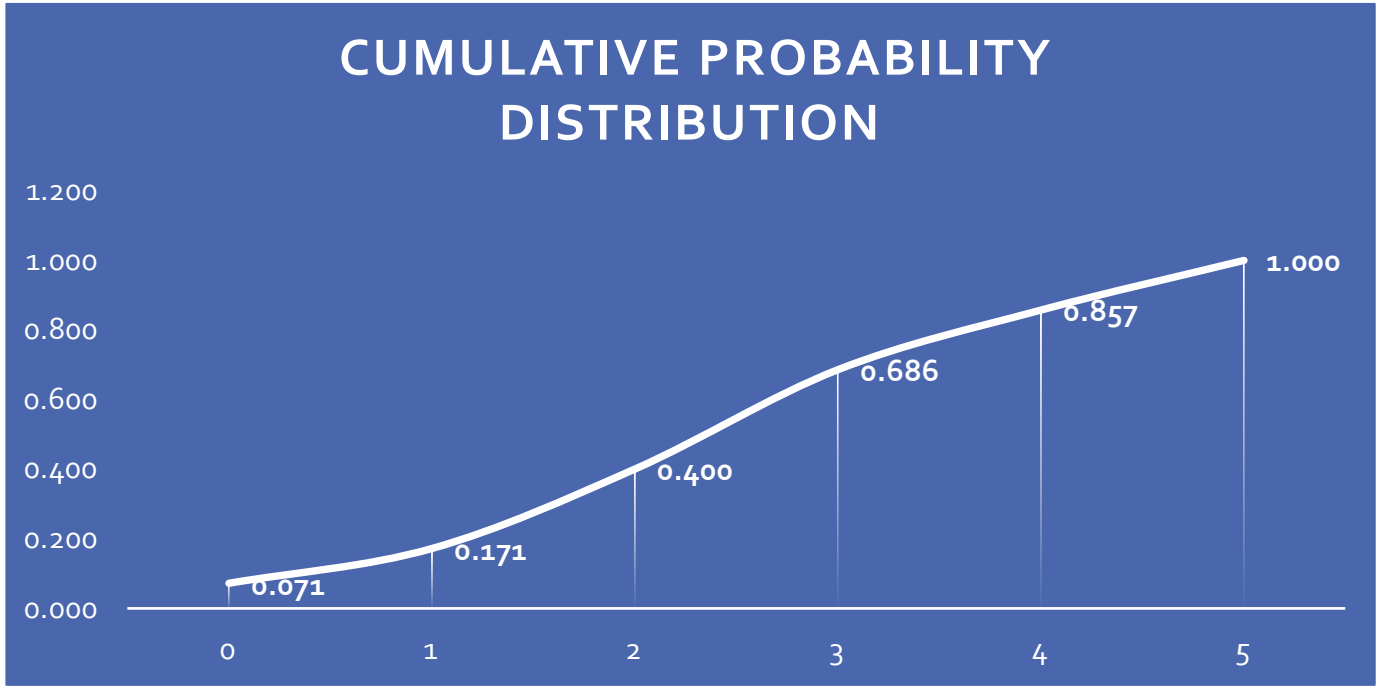
Number of Siblings (x)	Cumulative Probability $P(X \leq x)$
0	0.071
1	0.171
2	0.400
3	0.686
4	0.857
5	1.000

Probability Distributions

Discrete Variables

Cumulative Distribution

- It is obtained by successively adding the probabilities.



Probability Distributions

Example

If a car agency sells 50% of its inventory of a certain foreign car equipped with side airbags, find a formula for the probability distribution of the number of cars with side airbags among the next 4 cars sold by the agency.

Solution

Since the probability of selling an automobile with side airbags is 0.5, the $2^4 = 16$ points in the sample space are equally likely to occur.

$$f(x) = \frac{1}{16}\binom{4}{x} \text{ where } x = 0, 1, 2, 3, 4$$

Probability Distributions

Example

Find the cumulative distribution function of the random variable *X* in the example.

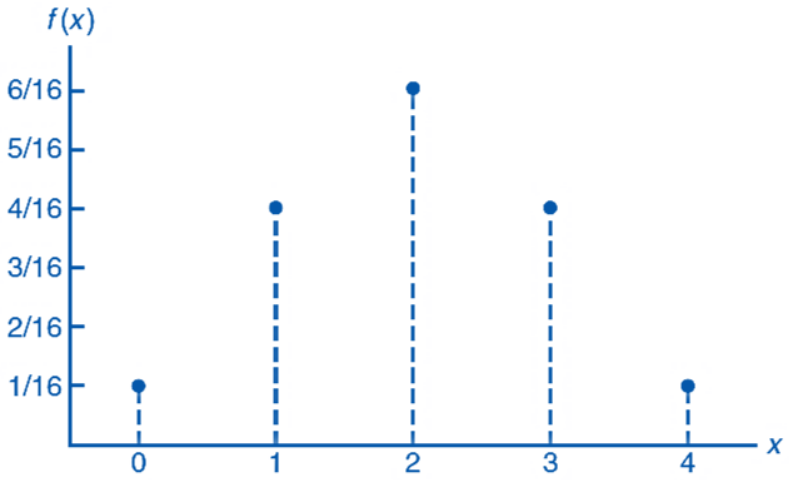
Solution $f(x) = \frac{1}{16}\binom{4}{x}$ where $x = 0, 1, 2, 3, 4$.

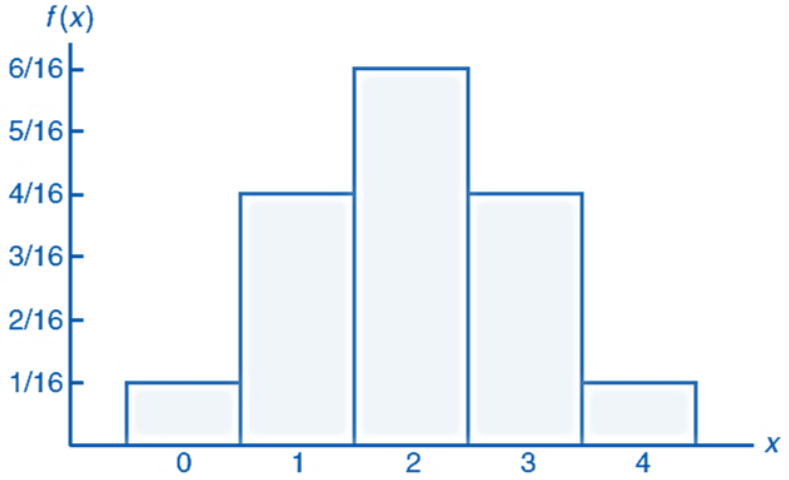
$f(0) = \frac{1}{16} \times \frac{4!}{0!(4-0)!} = \frac{1}{16}$ when $x = 0$.

Similarly, $f(1) = \frac{1}{14}, f(2) = \frac{3}{8}, f(3) = \frac{1}{4}$ and $f(4) = \frac{1}{16}$.

Probability Distributions

$$F(x) = \begin{cases} 0 & for \quad x < 0 \\ \frac{1}{16} & for \quad 0 \le x < 1 \\ \frac{5}{16} & for \quad 1 \le x < 2 \\ \frac{11}{16} & for \quad 2 \le x < 3 \\ \frac{15}{16} & for \quad 3 \le x < 4 \\ 1 & for \quad x \le 4 \end{cases}$$





Probability Distributions

The Binomial Distribution

- It is one of the most widely encountered probability distributions in applied statistics.
- The distribution is derived from a process known as a Bernoulli trial, named in honor of the Swiss mathematician James Bernoulli (1654-1705), who made significant contributions in the field of probability, including , in particular, the binomial distribution.

Probability Distributions

The Bernoulli Process

- A sequence of Bernoulli trials forms a **Bernoulli process** under the following conditions.

1. Each trial results in one of two possible, mutually exclusive, outcomes. One of the possible outcomes is denoted (arbitrarily) as a **success**, and the other is denoted a **failure**.
2. The *probability of a success*, denoted by **p**, remains constant from trial to trial. The *probability of a failure*, **1-p**, is denoted by **q**.
3. The trials are *independent*, that is, the outcome of any particular trial is not affected by the outcome of any other trial.

Probability Distributions

Large Sample Procedure: Use of Combinations

- A **combination** of **n** objects taken **x** at a time is an unordered subset of **x** of the **n** objects.
- The number of combinations of **n** objects that can be formed by taking **x** of them at a time is given by

$$_{n}C_{x} = \frac{n!}{x!\,(n-x)!}$$

where **x!**, read **x** *factorial*, is the product of all the whole numbers from **x** down to **1**.

Probability Distributions

Binomial Probability Formula for *P(X = x)*

$$P\left(X = x\right) = \frac{n!}{x!(n-x)!}\, p^x \left(1 - p\right)^{(n-x)}$$

where *x* may take any value ***0***, ***1***, **…** , ***n***.

- **Example:** Overall, the proportion of people with degree is 0.4. In other words, roughly 40% of the population has degree. Suppose we sample 4 people at random, what is the probability that 3 people will have degree?

Solution: p(A) = 0.4, x = 3 and n = 4 so we have

$$\mathrm{P}(\mathrm{X} = 3) = \frac{4!}{3!(4-3)!}\, 0.4^3 (1 - 0.4)^{(4-3)} = 0.1536 \text{ or } 15.36\%$$

Probability Distributions

The Binomial Parameters

- The mean and variance of the binomial distribution are $\boldsymbol{\mu = np}$ and $\boldsymbol{\sigma^2 = np(1-p)}$.
  - The probability distribution in our example is shown below:

x	Px)	x.P(x)
0	0.1296	0
1	0.3456	0.3456
2	0.3456	0.6912
3	0.1536	0.4608
4	0.0256	0.1024
Total	1.0000	1.6000

  - The mean of the binomial distribution is $\boldsymbol{\sum_{i=1}^{5} x_i \cdot P(x_i) = 1.6}$ which is equal to 4(0.4) (i.e. **np**).
  - The variance of the binomial distribution is $\boldsymbol{\sigma^2 = 4 \cdot 0.4(1-0.4) = 0.96}$

Binomial Probability Distributions Using PHStat

Binomial Probabilities

Data	
Sample size	4
Probability of an event of interest	0.4

Statistics	
Mean	1.6
Variance	0.9600
Standard deviation	0.9798

Binomial Probabilities Table

X	P(X)	P(<=X)	P(<X)	P(>X)	P(>=X)
0	0.1296	0.1296	0.0000	0.8704	1.0000
1	0.3456	0.4752	0.1296	0.5248	0.8704
2	0.3456	0.8208	0.4752	0.1792	0.5248
3	0.1536	0.9744	0.8208	0.0256	0.1792
4	0.0256	1.0000	0.9744	0.0000	0.0256

Binomial Probability Distributions Using R

#PBinomial Probability Distribution

#dbinom(N,size,prob)

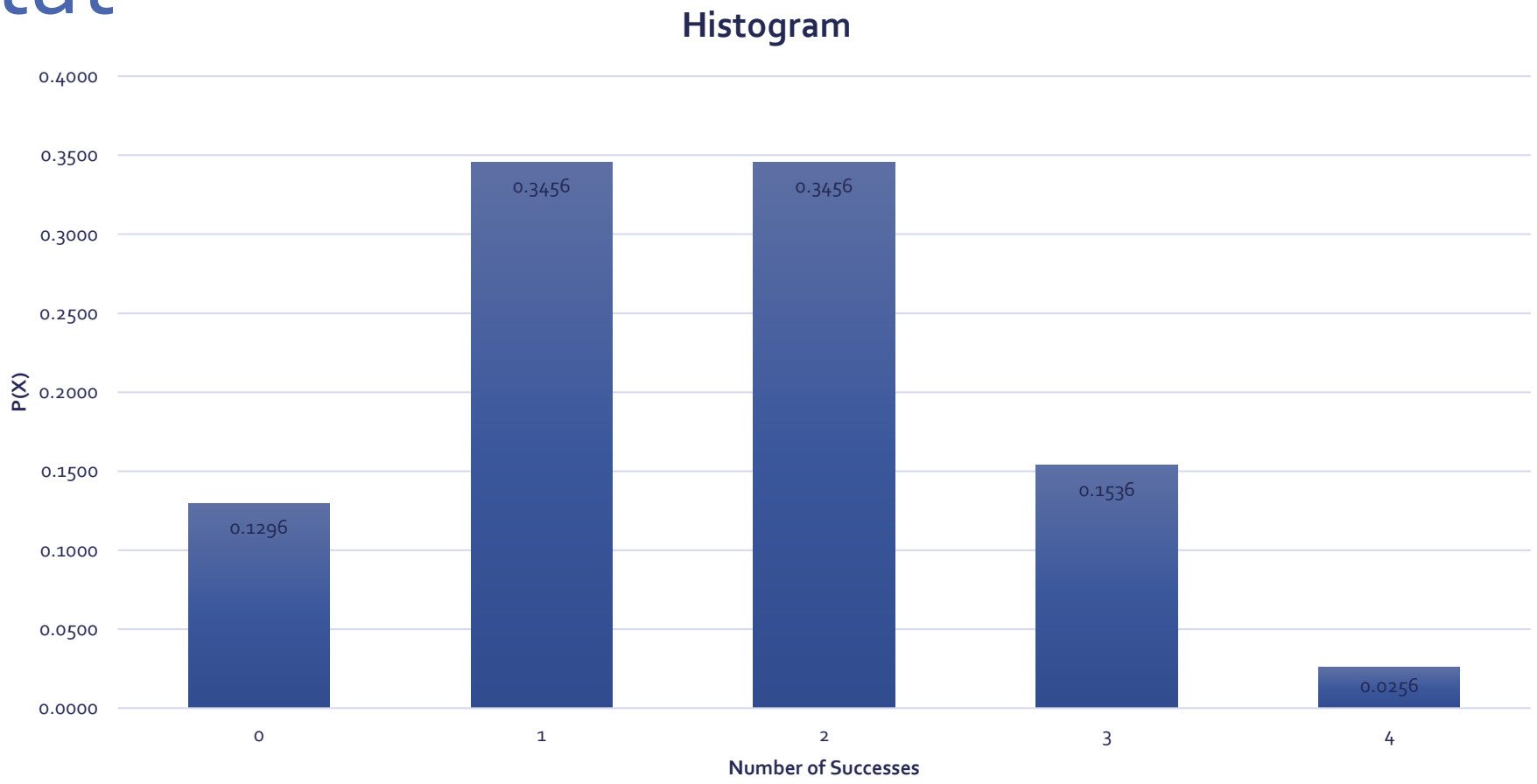
x=c(0,1,2,3,4)

dbinom(x,size=4,prob = 0.4) #size is 4 taken x

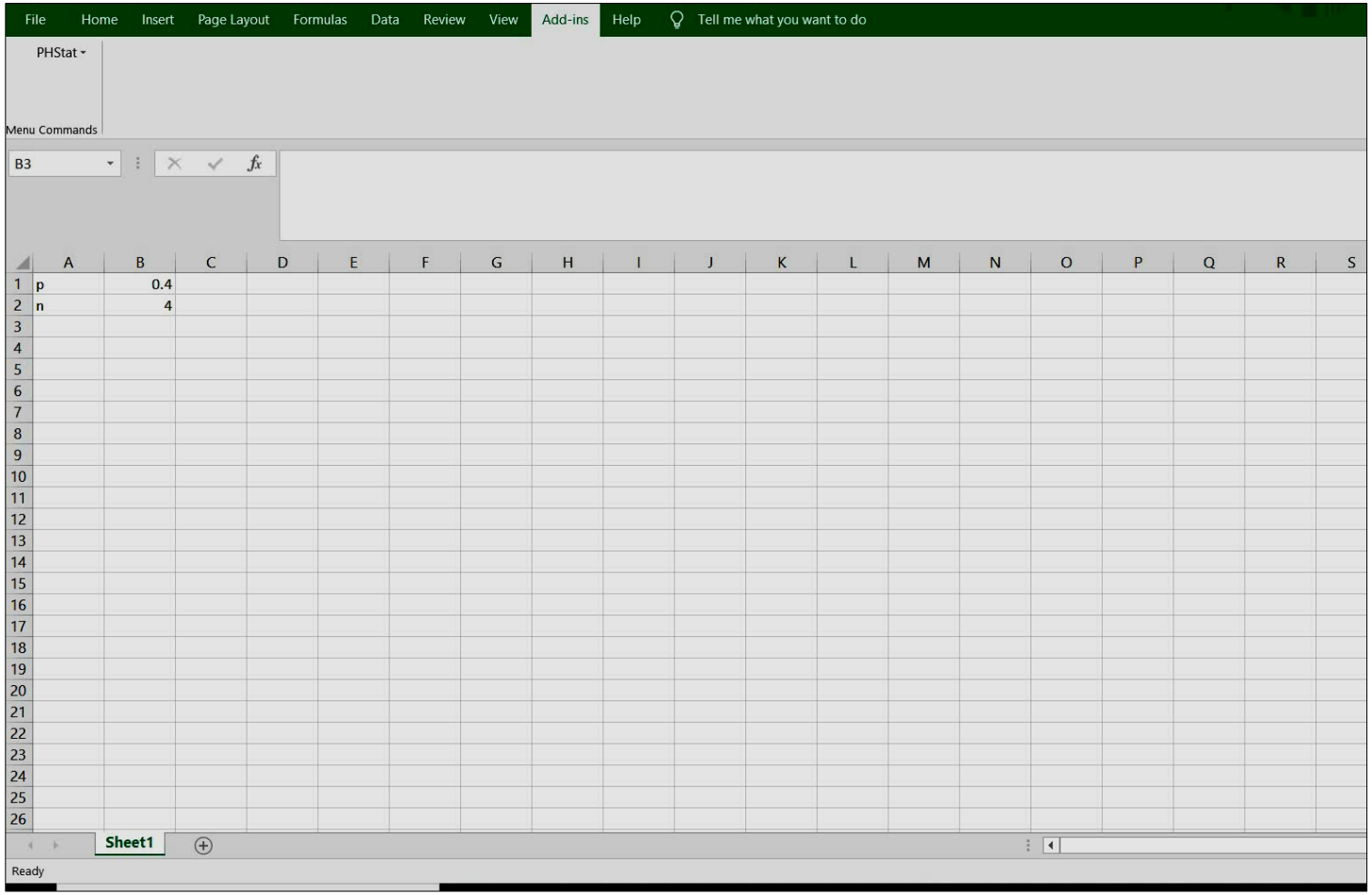
pbinom(q=1, size=4, prob =0.4)# cumulative P(0)+P(1)

```
> x=c(0,1,2,3,4)
> dbinom(x,size=4,prob = 0.4) #size is 4 taken x
[1] 0.1296 0.3456 0.3456 0.1536 0.0256
> pbinom(q=1, size=4, prob =0.4)# cumulative P(0)+P(1)
[1] 0.4752
```

Binomial Probability Distributions Using PHStat



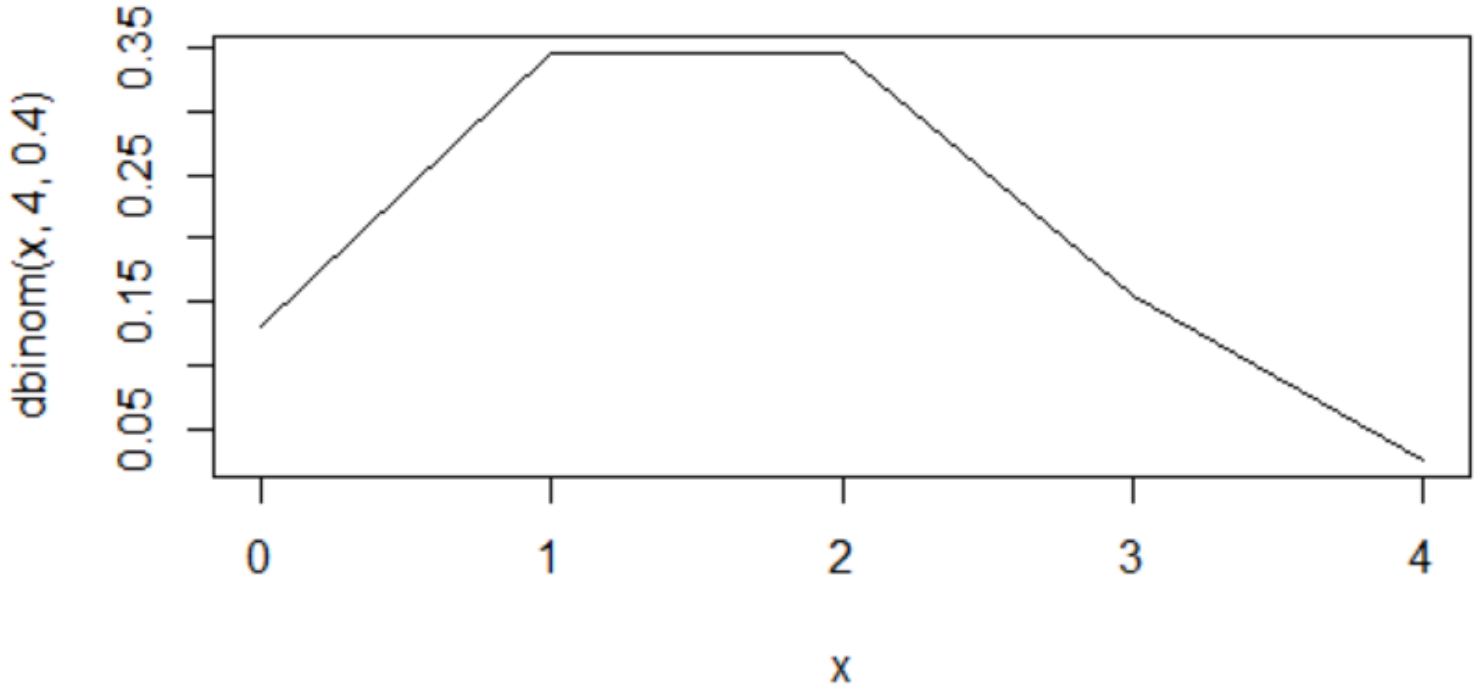
Binomial Probability Distributions Using PHStat



Binomial Probability Distributions Using R

#Binomial Probability Plot

```
x <- c(0,1,2,3,4)

plot(x, dbinom(x, 4, 0.4), type = "l")
```



Probability Distributions

The Poisson Distribution

- It is named for the French mathematician *Simeon Denis Poisson* (1781-1840), who is generally credited for publishing its derivation in 1837.
- If x is the number of occurrences of some random event in an interval of time or space (or some volume of matter), the probability that x will occur is given by

$$\mathbf{f(x)} = \frac{\mathbf{e^{-\lambda}\lambda^{x}}}{\mathbf{x!}}, \text{ x = 0, 1, 2, ...}$$

- The Greek $\boldsymbol{\lambda}$ (lambda) is called the parameter of the distribution and in the average number of occurrences of the random event in the interval (or volume).

Probability Distributions

The Poisson Process

- The following statements describe what is known as the **Poisson process**.

1. The occurrences of the events are independent. The occurrence of an event in an interval of space or time has no effect on the probability of a second occurrence of the event in the same, or any other, interval.
2. Theoretically, an infinite number of occurrences of the event must be possible in the interval.
3. The probability of the single occurrence of the event in a given interval is proportional to the length of the interval.
4. In any infinitesimally small portion of the interval, the probability of more than one occurrence of the event is negligible.

Probability Distributions

Example

How would you decide whether four accidents in a night are reasonably likely? The first thing is to look at past data, and so learn about the distribution of accidents. Since the bypass was opened nearly two years ago, the figures are as follows:

Number of accidents per day, x	0	1	2	3	>3
Frequency, f	395	235	73	17	0

In this case, the given time interval is one day, or 24 hours. An event is an accident.

- The total number of accidents has been $0 \times 395 + 1 \times 235 + 2 \times 73 + 3 \times 17 = 432$
- The number of days has been $395 + 235 + 73 + 17 = 720$
- So the mean number of accidents per day has been $\frac{432}{720} = 0.6$

Probability Distributions

Example

Number of accidents per day, x	0	1	2	3	>3
Frequency, f	395	235	73	17	0

So, the probability of

0 occurrences is $e^{-\lambda}$

1 occurrence is $\lambda e^{-\lambda}$

2 occurrences is $\frac{\lambda^2}{2!}e^{-\lambda}$

3 occurrences is $\frac{\lambda^3}{3!}e^{-\lambda}$

and so on.

Probability Distributions

In this example, $\lambda = 0.6$ and so the probabilities and expected frequencies in 720 days are as follows:

Number of accidents per day	0	1	2	3	4	5	>5
Probability	0.5488	0.3293	0.0988	0.0198	0.0030	0.0004	0.0000
Expected frequency	395.1444	237.0866	71.1260	14.2252	2.1338	0.2561	0.0000

- The table shows that with this model you would expect 2.1 days in 720 (i.e. just over 1 a year) where there would be 4 accidents. There is no need to jump to the conclusion that there was another factor, such as full moon, that influenced the data.

Poisson Probability Distributions Using PHStat

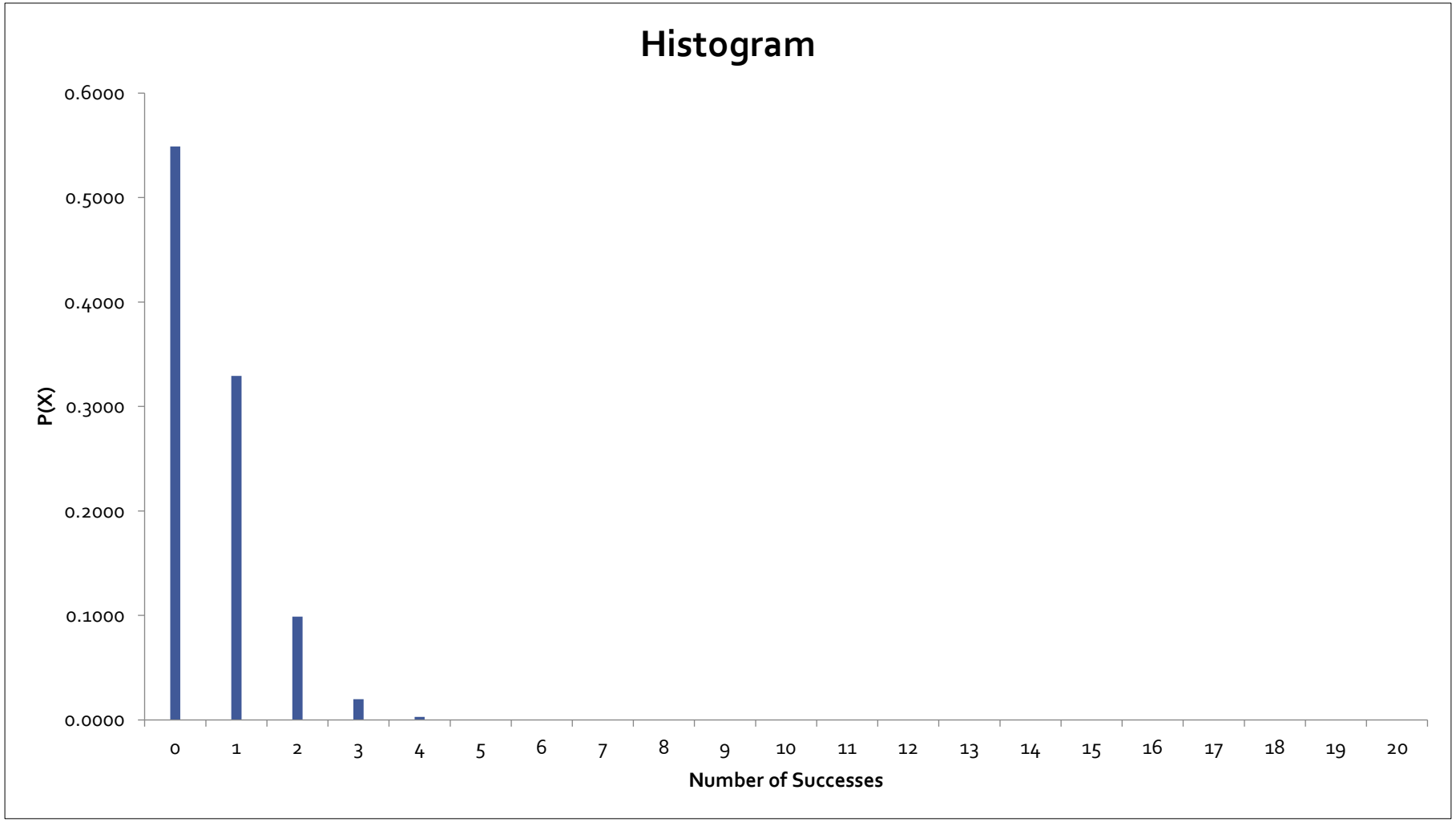
POISSON.DIST Probabilities

Data					
Mean/Expected number of events of interest:				0.6	

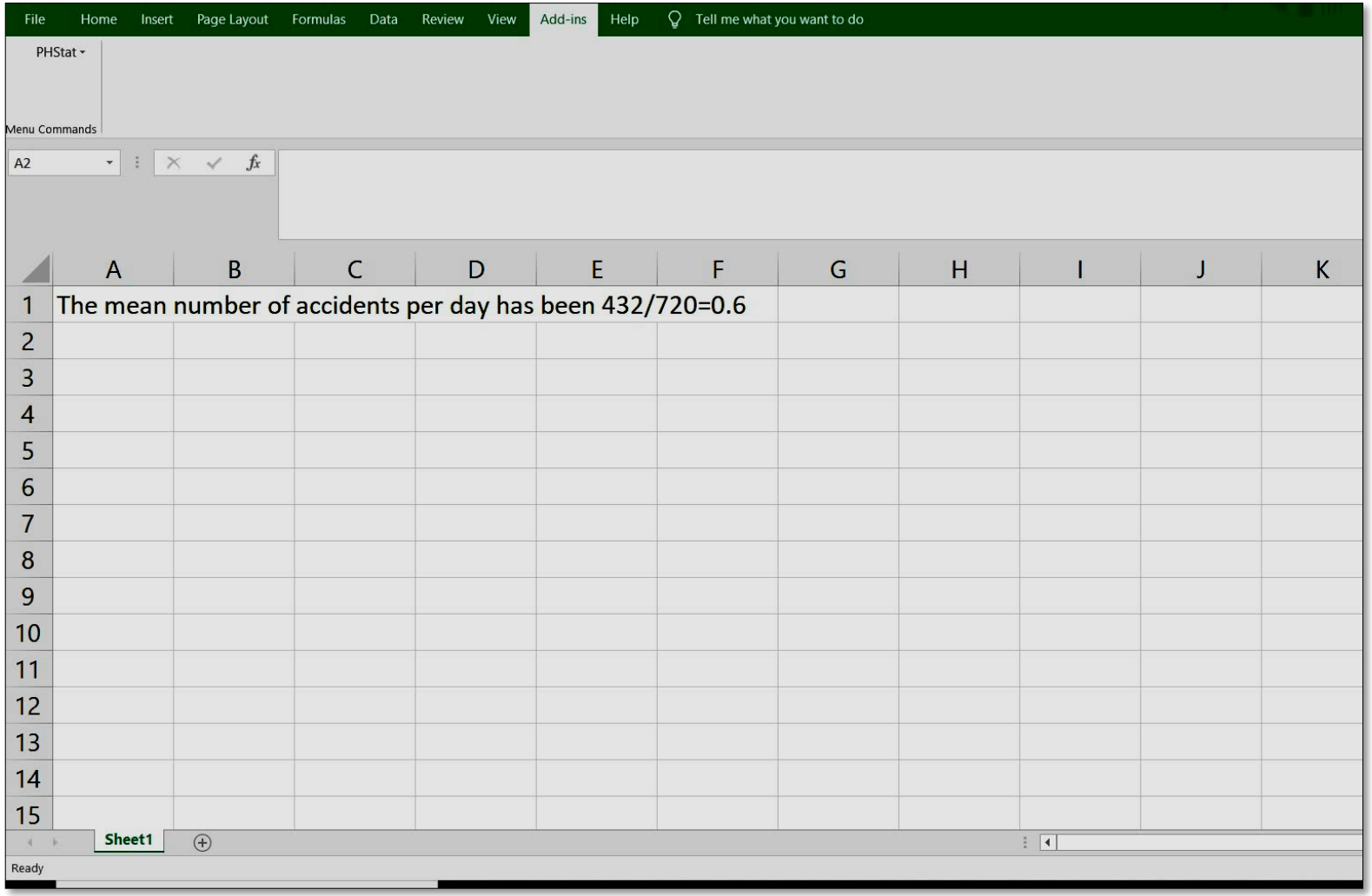
POISSON.DIST Probabilities Table

	X	P(X)	P(<=X)	P(<X)	P(>X)	P(>=X)
	0	0.5488	0.5488	0.0000	0.4512	1.0000
	1	0.3293	0.8781	0.5488	0.1219	0.4512
	2	0.0988	0.9769	0.8781	0.0231	0.1219
	3	0.0198	0.9966	0.9769	0.0034	0.0231
	4	0.0030	0.9996	0.9966	0.0004	0.0034
	5	0.0004	1.0000	0.9996	0.0000	0.0004
	6	0.0000	1.0000	1.0000	0.0000	0.0000
	7	0.0000	1.0000	1.0000	0.0000	0.0000
	8	0.0000	1.0000	1.0000	0.0000	0.0000
	9	0.0000	1.0000	1.0000	0.0000	0.0000
	10	0.0000	1.0000	1.0000	0.0000	0.0000
	11	0.0000	1.0000	1.0000	0.0000	0.0000
	12	0.0000	1.0000	1.0000	0.0000	0.0000
	13	0.0000	1.0000	1.0000	0.0000	0.0000
	14	0.0000	1.0000	1.0000	0.0000	0.0000
	15	0.0000	1.0000	1.0000	0.0000	0.0000
	16	0.0000	1.0000	1.0000	0.0000	0.0000
	17	0.0000	1.0000	1.0000	0.0000	0.0000
	18	0.0000	1.0000	1.0000	0.0000	0.0000
	19	0.0000	1.0000	1.0000	0.0000	0.0000
	20	0.0000	1.0000	1.0000	0.0000	0.0000

Poisson Probability Distributions Using PHStat



Poisson Probability Distributions Using PHStat



Poisson Probability Distributions Using R

#Poisson Probability Distribution

x=0:20

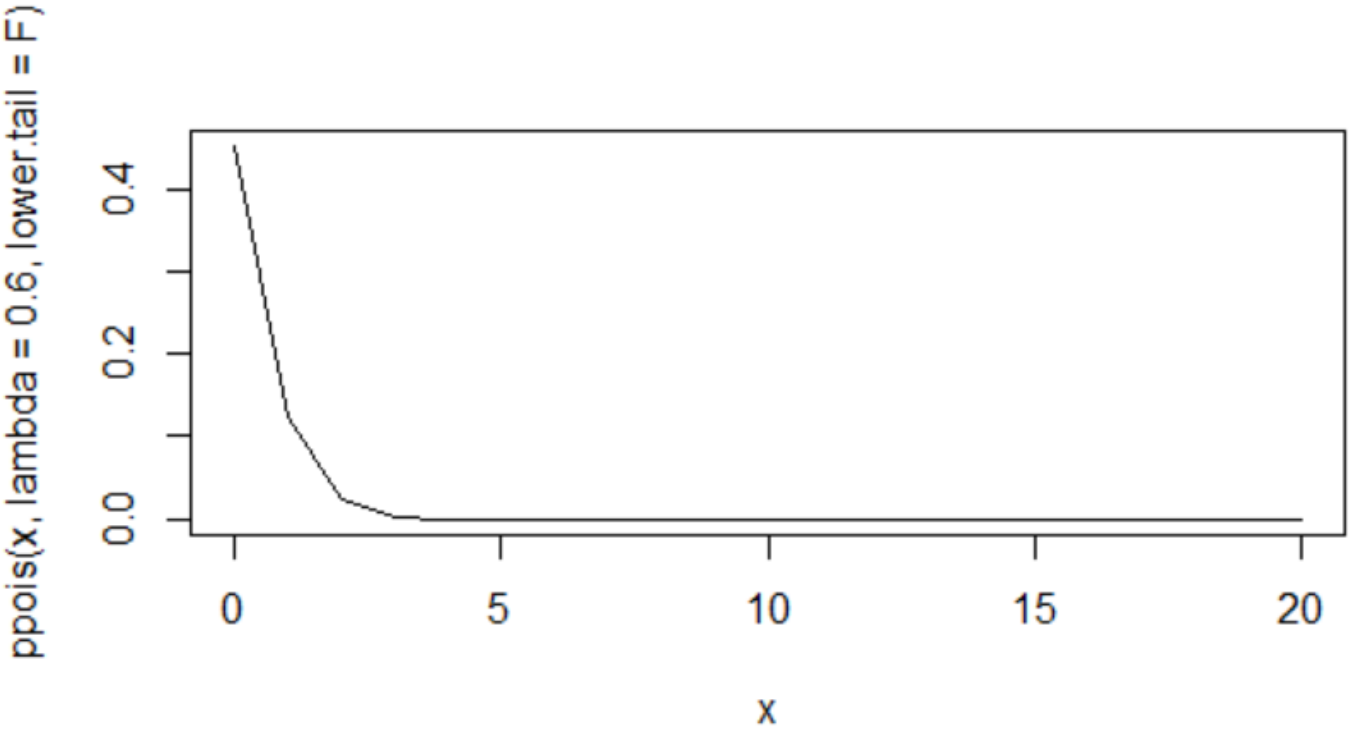
ppois(x, lambda = 0.6, lower.tail = T) #Cumulative < or =

```
> x=0:20
> ppois(x, lambda = 0.6, lower.tail = T) #Cumulative < or =
 [1] 0.5488116 0.8780986 0.9768847 0.9966419 0.9996055 0.9999611 0.9999967
 [8] 0.9999998 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[15] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

Poisson Probability Distributions Using R

#Poisson Probability Distribution

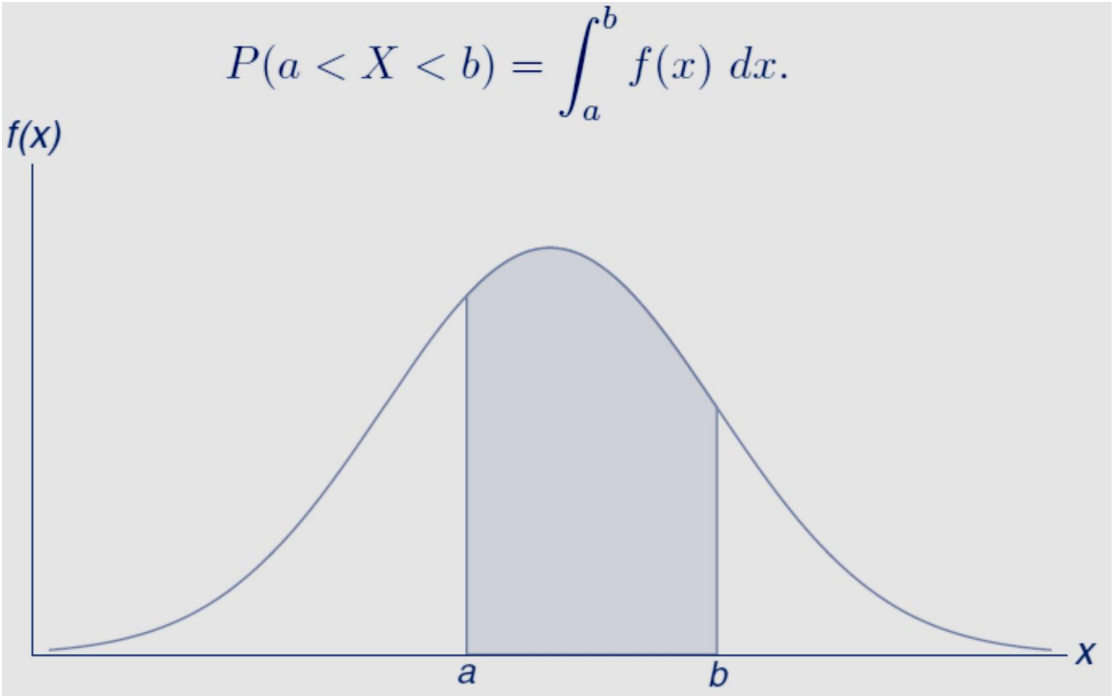
```
x=0:20

plot(x,ppois(x, lambda = 0.6, lower.tail = F), type = "l")
```



Probability Distributions

The Continuous Distributions

- A nonnegative function f(x) is called a **probability distribution** (sometimes called a **probability density function**) **of the continuous random variable X** if the total area bounded by its curve and the **x**-axis is equal to **1** and if the subarea under the curve bounded by the curve, the **x**-axis, and the perpendiculars erected at any two points *a* and *b* give the probability that **X** is between the points **a** and **b**.



$P(a < X < b)$

Probability Distributions

The Continuous Distributions

The function $f(x)$ is a **probability density function** (pdf) for the continuous random variable X, defined over the set of real numbers, if

1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_{a}^{b} f(x)\, dx$.

Probability Distributions

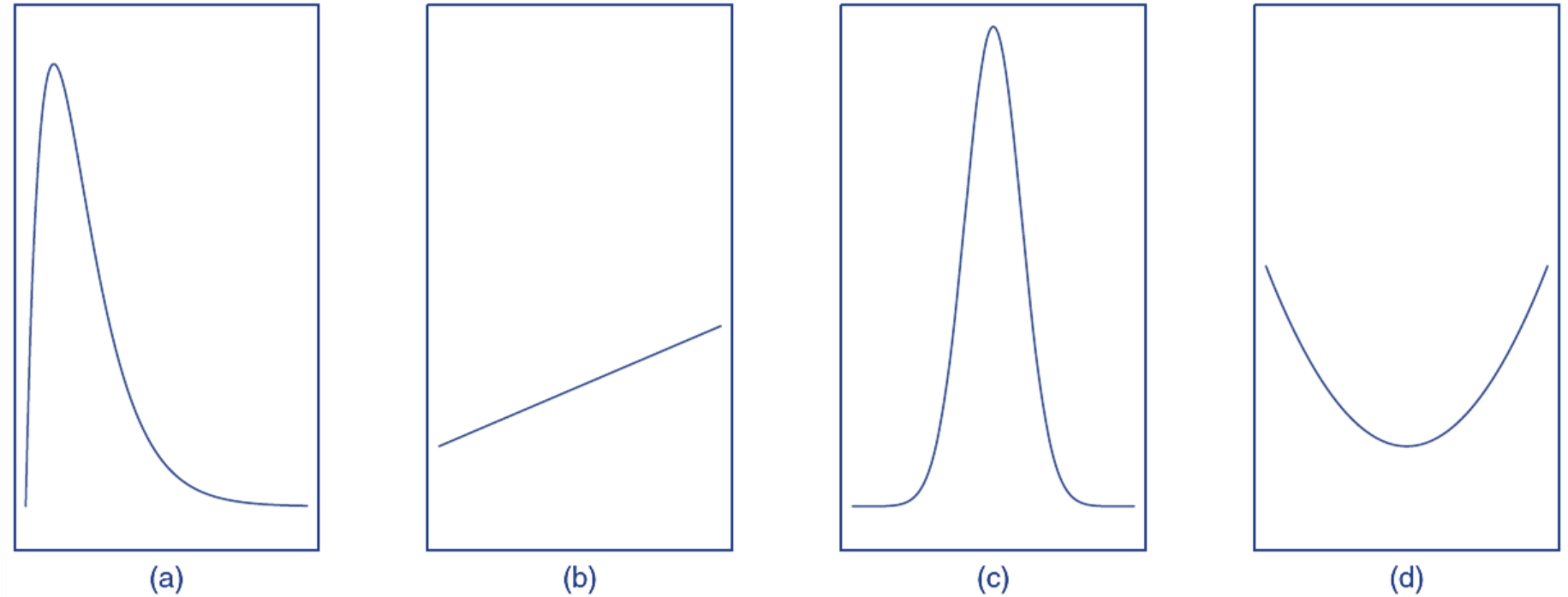
The Continuous Distributions

The **cumulative distribution function** $F(x)$ of a continuous random variable X with density function $f(x)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt, \quad \text{for } -\infty < x < \infty.$$

Probability Distributions

Typical Density Functions



Probability Distributions

Example

Suppose that the error in the reaction temperature, in ◦C, for a controlled laboratory experiment is a continuous random variable X having the probability density function

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Verify that $f(x)$ is a density function.

b. Find $P(0 < X \leq 1)$.

c. Find the cumulative distribution function $F(x)$ and use it to evaluate.

Probability Distributions

Solution

Suppose that the error in the reaction temperature, in ◦C, for a controlled laboratory experiment is a continuous random variable *X* having the probability density function

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

a. To verify

1. Obviously, $f(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{2} \frac{x^2}{3} dx = \frac{x^3}{9}\Big|_{-1}^{2} = \frac{2^3}{9} - \frac{(-1)^3}{9} = 1$

b. $P(0 < x \leq 1) = \int_{0}^{1} \frac{x^2}{3} dx = \frac{1}{9}$

c. For $-1 < x < 2$, $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-1}^{x} \frac{t^2}{3} dt = \frac{t^3}{9}\Big|_{-1}^{x} = \frac{x^3+1}{9}$.

Probability Distributions

Solution

a. To verify

1. Obviously, $f(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{2} \frac{x^2}{3} dx = \frac{x^3}{9}\Big|_{-1}^{2} = \frac{2^3}{9} - \frac{(-1)^3}{9} = 1$

b. $P(0 < x \leq 1) = \int_{0}^{1} \frac{x^2}{3} dx = \frac{1}{9}$

c. For $-1 < x < 2$, $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-1}^{x} \frac{t^2}{3} dt = \frac{t^3}{9}\Big|_{-1}^{x} = \frac{x^3+1}{9}$.

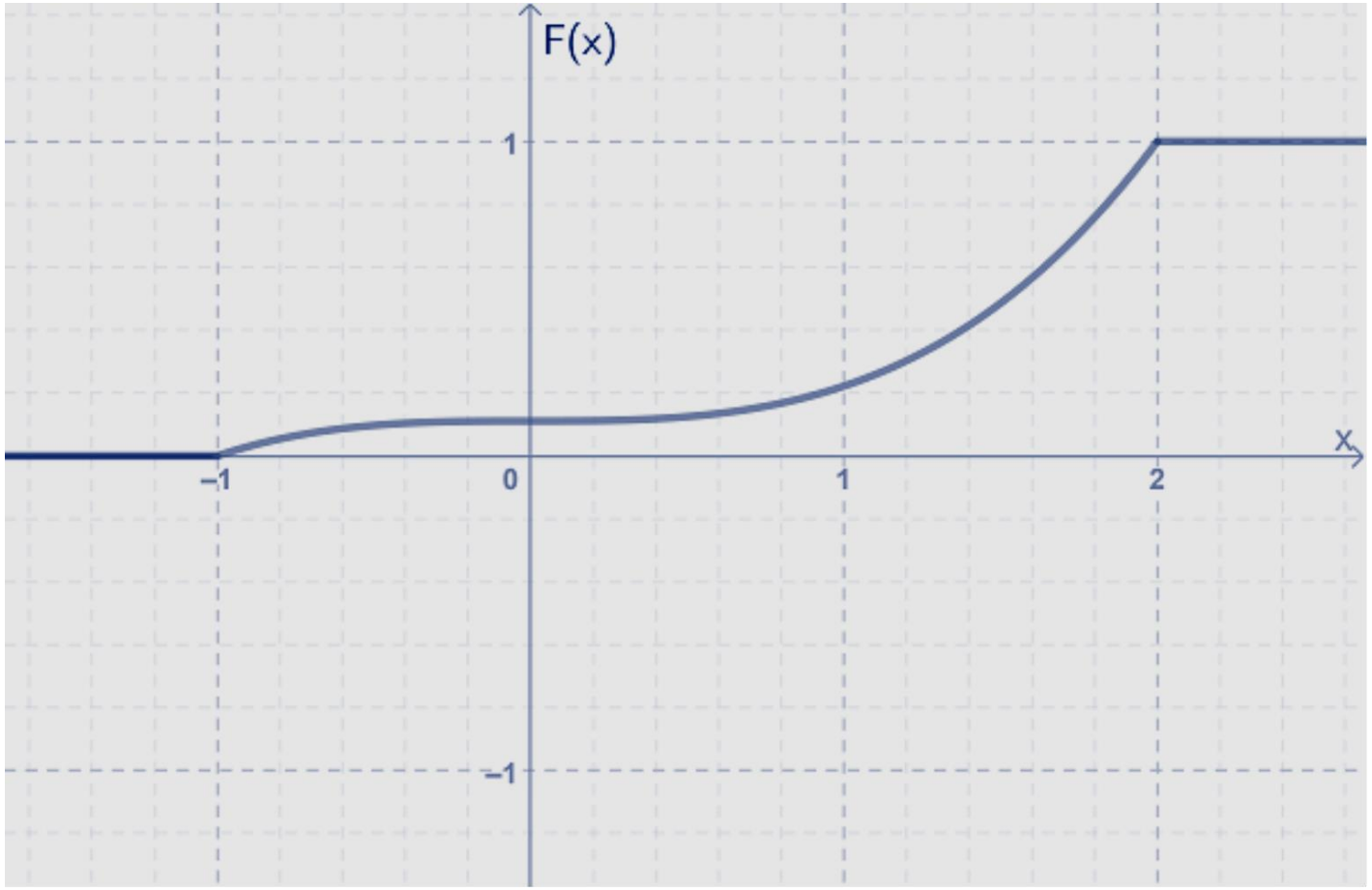
Therefore, $$F(x) = \begin{cases} 0, & x < -1, \\ \frac{x^3+1}{9}, & -1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

$P(0 < x \leq 1) = F(1) - F(0) = \frac{2}{9} - \frac{1}{9} = \frac{1}{2}$ which is equal to b.

Probability Distributions

Solution

$$F(x) = \begin{cases} 0, & x < -1, \\ \frac{x^3+1}{9}, & -1 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$



Probability Distributions

The Normal Distribution

- It is the most important distribution in all of statistics.
- The formula for this distribution was first published by *Abraham De Moivre* (1667-1754) on November 12, 1733.
- Many other mathematicians figure prominently in the history of the normal distribution, including *Carl Friedrich Gauss* (1777-1855).
- The distribution is frequently called the **Gaussian distribution** in recognition of his contributions.
- The normal density is given by

$$\mathbf{f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty.}$$

Probability Distributions

The Normal Distribution

- It is the most important distribution in all of statistics.
- The formula for this distribution was first published by *Abraham De Moivre* (1667-1754) on November 12, 1733.
- Many other mathematicians figure prominently in the history of the normal distribution, including *Carl Friedrich Gauss* (1777-1855).
- The distribution is frequently called the **Gaussian distribution** in recognition of his contributions.
- The normal density is given by

$$\mathbf{f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty.}$$

Probability Distributions

The Normal Distribution

- Characteristics of the Normal Distribution

1. It is symmetrical about its mean, μ,
2. The mean, the median, and the mode are all equal.
3. The total area under the curve above the **x**-axis is one square unit.
4. If we erect perpendiculars a distance of 1 standard deviation from the mean in both directions, the area enclosed by these perpendiculars, the x-axis, and the curve will be approximately 68% of the total area. For 2 standard deviations, approximately 95%. For 3 standard deviations, approximately 99.7%.
5. The normal distribution is completely determined by the parameters μ and σ.

Probability Distributions

The Standard Normal Distribution

- The **standard normal distribution** has a *mean* of **0** and a *standard deviation* of **1**.
- To transform all values of **X** (normal) to corresponding values of **z** (standard normal known as *z-score*), we use the formula

$$\mathbf{z} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

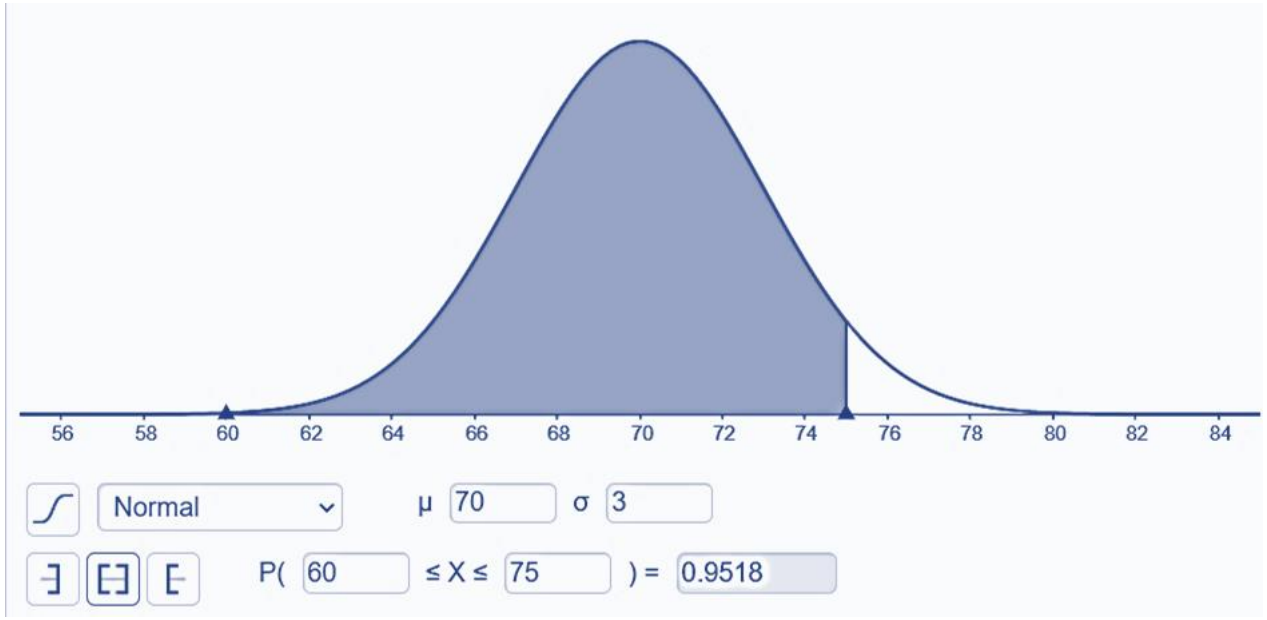
Probability Distributions

Normal Distribution Application

The mean weight of 500 college students is 70 kg and the standard deviation is 3 kg. Assuming that the weight is normally distributed, determine how many students weigh:

1. Between 60 kg and 75 kg.

   0.9518 x 500 ≈ **476 students**



Graphs are provided by GeoGebra.

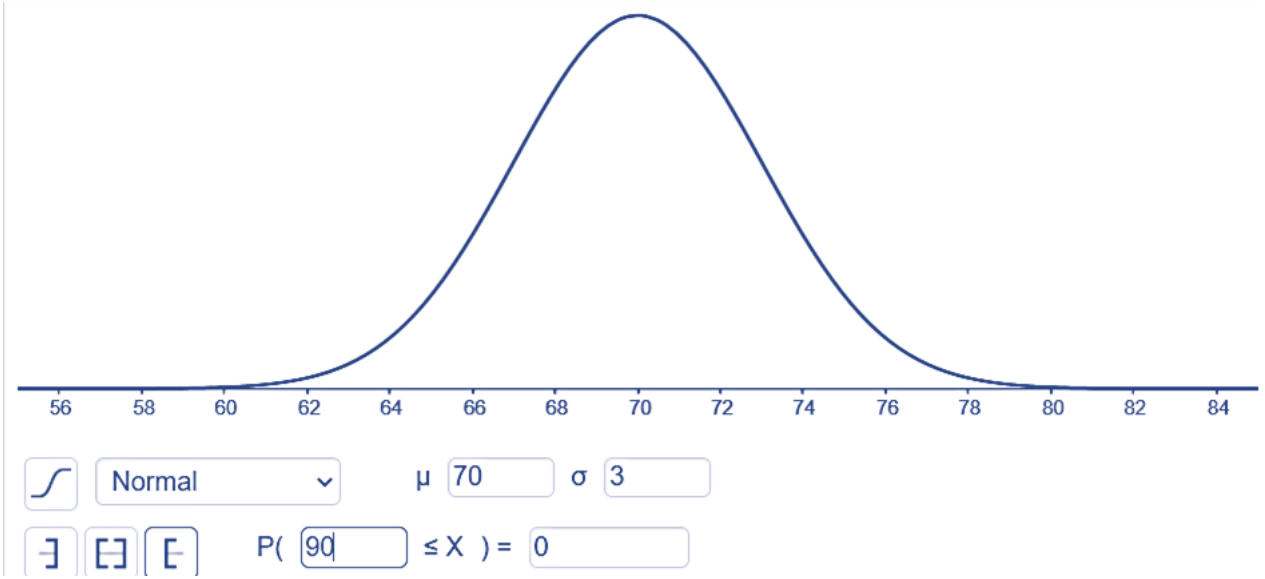
Probability Distributions

Normal Distribution Application

The mean weight of 500 college students is 70 kg and the standard deviation is 3 kg. Assuming that the weight is normally distributed, determine how many students weigh:

2. More than 90 kg.

   0 x 500 = **0 students**



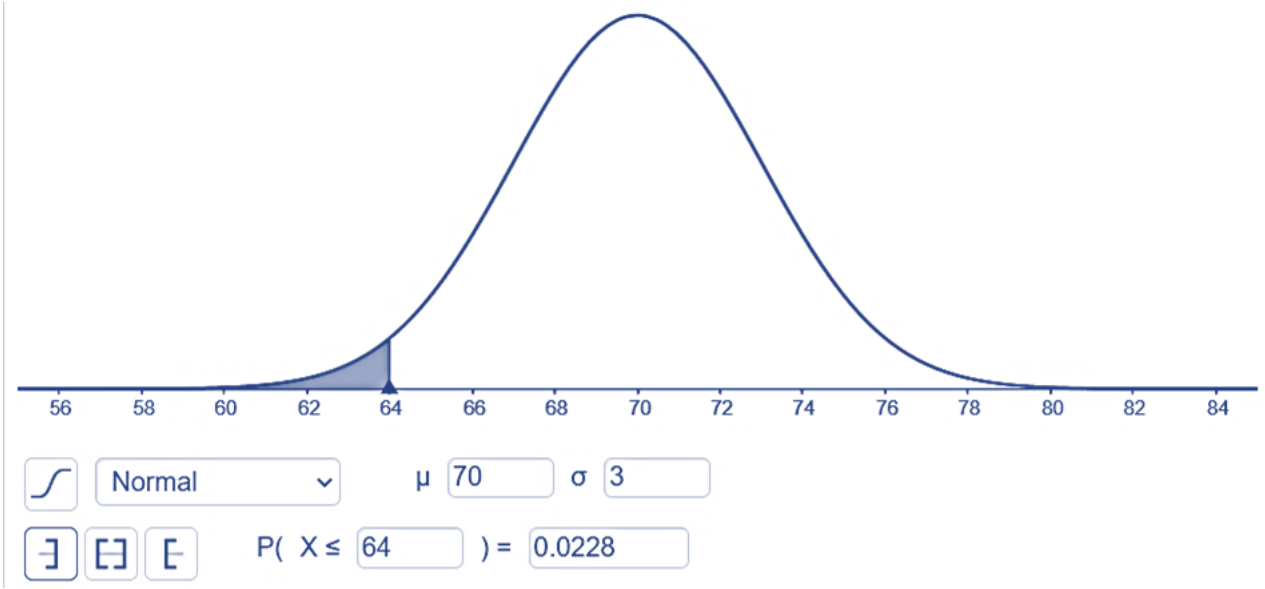
Probability Distributions

Normal Distribution Application

The mean weight of 500 college students is 70 kg and the standard deviation is 3 kg. Assuming that the weight is normally distributed, determine how many students weigh:

3. 64 kg or less.

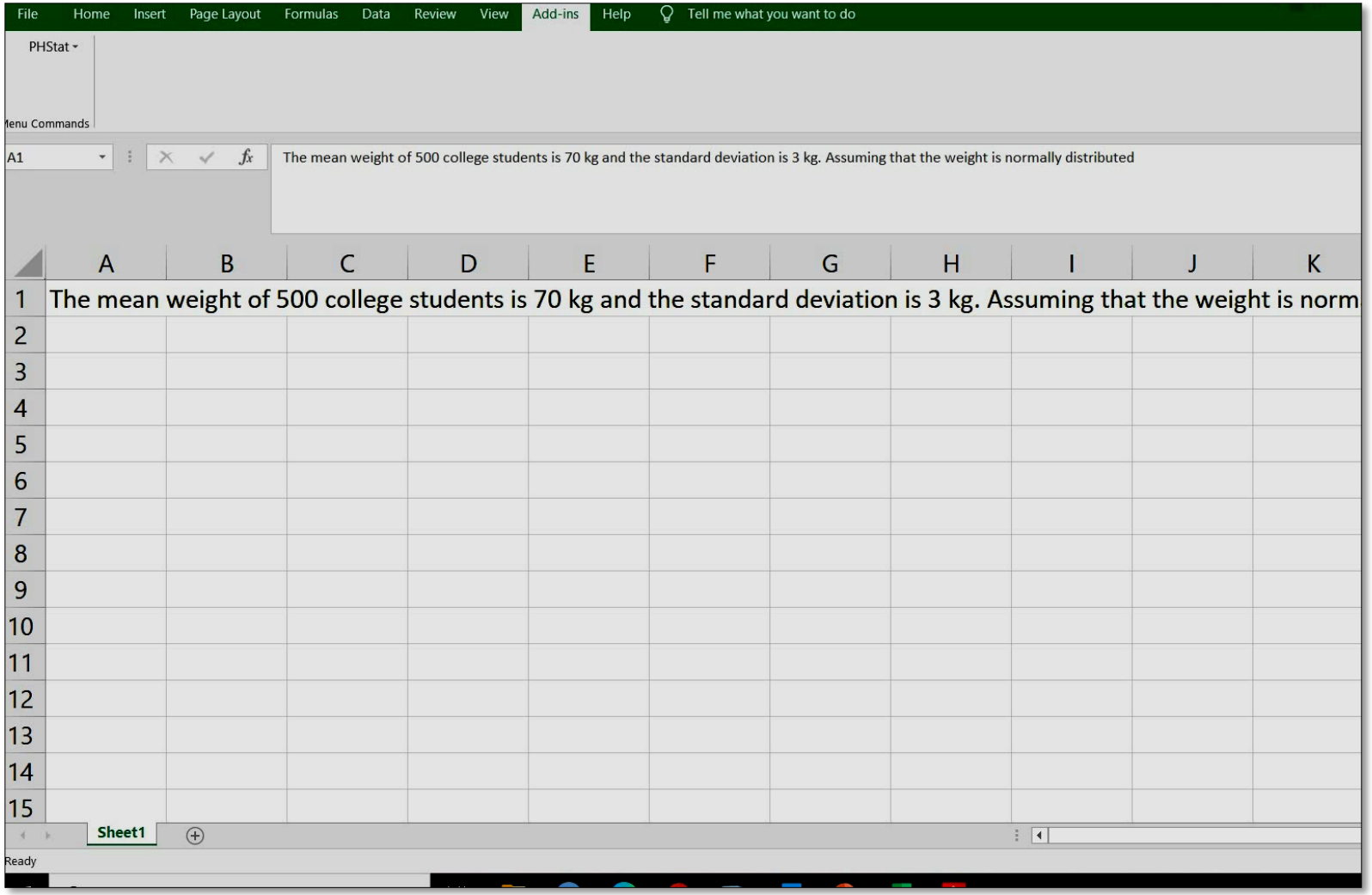
   0.0228 x 500 ≈ **11 students**



Normal Probability Distributions Using PHStat

Normal Probabilities			
Common Data			
Mean	70		
Standard Deviation	3		
		Probability for a Range	
		From X Value	60
		To X Value	75
		Z Value for 60	-3.333333
		Z Value for 75	1.666667
		P(X<=60)	0.0004
		P(X<=75)	0.9522
		P(60<=X<=75)	0.9518

Normal Probability Distributions Using PHStat



Normal Probability Distributions Using R

y=pnorm(75, mean = 70, sd = 3) - pnorm(60, mean = 70, sd = 3)#1

y*500

(1-pnorm(90, mean = 70, sd = 3))*500#2

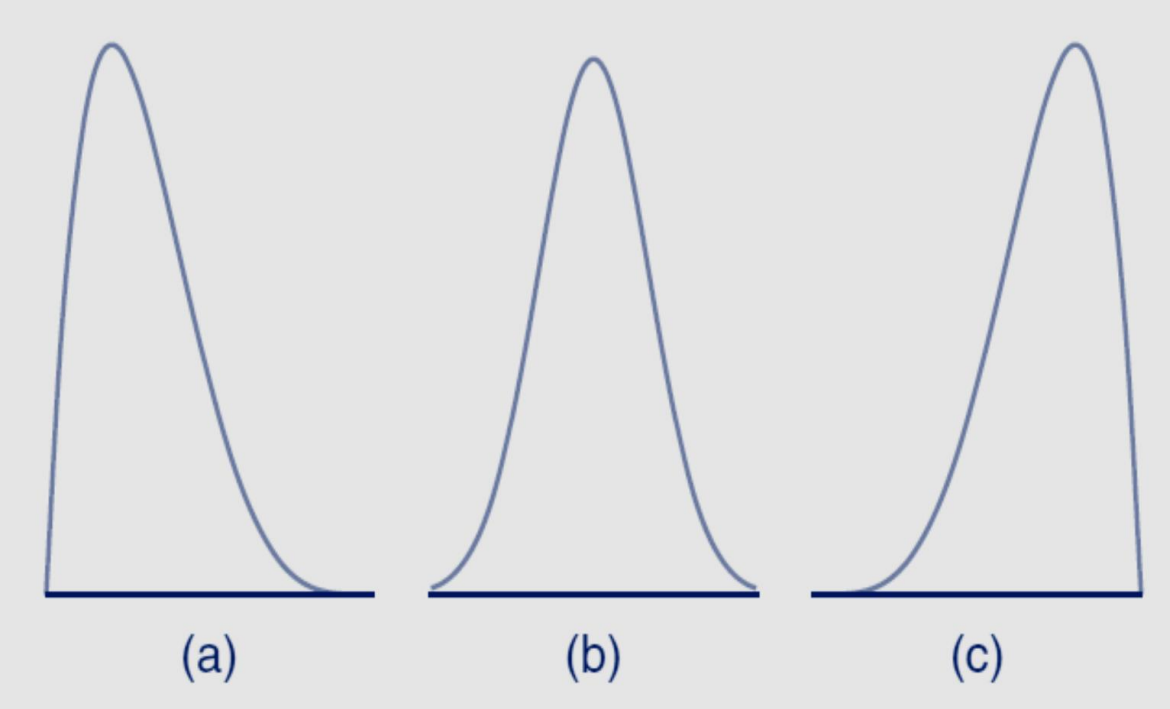
pnorm(64, mean = 70, sd = 3)*500 #3

```
> y=pnorm(75, mean = 70, sd = 3) - pnorm(60, mean = 70, sd = 3)#1
> y*500
[1] 475.8903
> (1-pnorm(90, mean = 70, sd = 3))*500#2
[1] 6.541989e-09
> pnorm(64, mean = 70, sd = 3)*500 #3
[1] 11.37507
```

Probability Distribution

Skewness

A distribution is said to be **symmetric** if it can be folded along a vertical axis so that the two sides coincide. A distribution that lacks symmetry with respect to a vertical axis is said to be **skewed**. The distribution illustrated below, figure (a) is said to be *skewed to the right* since it has a long right tail and a much shorter left tail. Figure (b) we see that the distribution is symmetric, while in figure (c) it is skewed to the left.



Skewness of Data

REFERENCES

- Walpole, Ronald E. Probability and Statistics for Engineers and Scientists. New York, USA: MacMillan Publishing Co., Inc., 1985
- Daniel, Wayne W. & Cross, Chad L., BIOSTATISTICS A Foundation for Analysis in the Health Sciences. 10th ed., John Wiley & Sons, Inc, 2013(text)
- Sprinthall, Richard C., Basic Statistical Analysis. 8th ed., USA: Progressive Publishing Alternatives, 2007
- Downie, N. M. and R. W. Heath. Basic Statistical Methods. New York, USA: McGraw-Hill Publishing Co.Inc., 1977
- Walpole, Ronald E. Introduction to Statistics. 3rd ed., New York, USA: MacMillan Publishing Co., 1981
- Snedecor and Cochran. Statistical Methods. Lexington, Mass.: Addison-Wesley Publishing Com Inc., 1984.
- Joaquim P. Marques de Sá, Applied Statistics Using SPSS, STATISTICA, MATLAB and R

Identify the following variable as quantitative or qualitative variable and identify the scale of each variable.

- Example: Name: qualitative and nominal.

1. Smartphone brand
2. COVID patients temperature
3. Number of students in a class
4. Weight of vegetables
5. Likert scale like: strongly agree, agree, disagree and strongly disagree.