

2019-2

Introduction to Biostatistics - Lecture 1: Introduction and Descriptive Statistics

Jonggyu Baek
University of Massachusetts Medical School

Let us know how access to this document benefits you.

Follow this and additional works at: https://escholarship.umassmed.edu/liberia_peer



Part of the [Biostatistics Commons](#), [Family Medicine Commons](#), [Infectious Disease Commons](#), [Medical Education Commons](#), and the [Public Health Commons](#)

Repository Citation

Baek J. (2019). Introduction to Biostatistics - Lecture 1: Introduction and Descriptive Statistics. PEER Liberia Project. <https://doi.org/10.13028/rrh1-1470>. Retrieved from https://escholarship.umassmed.edu/liberia_peer/10

This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in PEER Liberia Project by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.

Introduction to Biostatistics

2/26/2019

Jonggyu Baek, PhD

Outline

- Purpose
- Introduction to biostatistics
- Descriptive Statistics

Purpose of the course

- Basic principles and applications of statistics to problems in clinical and public health settings.
- Will cover tools for statistical inference: t-test, chi-square tests, ANOVA, Linear regression.
- Interpretation and presentation of the results
- Introductory foundation for the implementation of those techniques using R or R studio software.

References

- Multiple authors, Openstax College

Introductory Statistics

Publisher: OpenStx, Pubdate: 2013

<https://open.umn.edu/opentextbooks/textbooks/introductory-statistics-2013>

- Quick-R: <https://www.statmethods.net/>
- UCLA statistical computing: <https://stats.idre.ucla.edu/>

What is Statistics?

- **Statistics** is the **science** of learning from **data**, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances (*Davidian, M. and Louis, T. A., 10.1126/science.1218685*).
- **Statistics** is also an **ART** ...
of conducting a study, analyzing the data,
and derive useful conclusions from numerical
outcomes about real life problems...



What is Biostatistics?

- **Biostatistics** is the application of statistics in medical research, e.g.:
 - Clinical trials
 - Epidemiology
 - Pharmacology
 - Medical decision making
 - Comparative Effectiveness Research
 - etc.

Statistical Analysis

Key steps for a complete and accurate statistical analysis:

- State a valid research question
- Collect information (DATA) for answering this question
- Validate/clean/organize the collected information
- Exploratory Data Analysis (EDA)
- Analyze this information
- Translate numerical results into answers
- Interpret results and derive conclusions
- Present the results and communicate with people

Terms in Biostatistics

- **Data :**
 - all the information we collect to answer the research question
- **Variables :**
 - Outcome, treatment, study population characteristics
- **Subjects :**
 - units on which characteristics are measured
- **Observations :**
 - data elements
- **Population :**
 - all the subjects of interest
- **Sample :**
 - a subset of the population for which data are collected

Sample from Population

	Population	Sample	
Descriptive Measure	Parameter	statistic	Summary of a characteristic
Size	N	n	Total # of subjects
Mean	μ	\bar{x}	Average
Variance	σ^2	s^2	Variance

Impossible/impractical to analyze the entire population →

→ thus we only analyze a sample

Statistical Inference

Collect and analyze data from samples →

→ Calculate summary statistics →

→ Make Inference about unknown population parameters (e.g., population average from sample mean)

The Framingham Heart Study

<https://www.framinghamheartstudy.org/fhs-about/history/epidemiological-background/>

- ... *“a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts.”* ...
- Identifying risk factors for cardiovascular disease (CVD)
- N=4,434 participants (subset of the original sample)
- Follow-up period: 1956 – 1968
- Longitudinal data: measurements approximately every 6 years
- 1 to 3 observations for each participant (total 11,627 obs)

The Framingham Heart Study

- Information:
 - ID
 - Age
 - Sex
 - Period (1st, 2nd, or 3rd exam)
 - Systolic Blood Pressure (mmHg)
 - Diastolic Blood Pressure (mmHg)
 - Use of Anti-hypertensive medication at exam (yes/no)
 - Current smoking status (yes/no)
 - Average number of cigarettes smoked/day
 - Prevalent coronary Heart disease (yes/no)
 - ... etc

The Framingham Heart Study

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Lec1.R x dat1 x

Filter

	sex	randid	totchol	age	sysbp	diabp	cursmoke	cigpday	bmi	diabetes	bpmeds	hearttrte
1	1	2448	195	39	106.0	70.0	0	0	26.97	0	0	
2	1	2448	209	52	121.0	66.0	0	0	NA	0	0	
3	2	6238	250	46	121.0	81.0	0	0	28.73	0	0	
4	2	6238	260	52	105.0	69.5	0	0	29.43	0	0	
5	2	6238	237	58	108.0	66.0	0	0	28.50	0	0	
6	1	9428	245	48	127.5	80.0	1	20	25.34	0	0	
7	1	9428	283	54	141.0	89.0	1	30	25.34	0	0	
8	2	10552	225	61	150.0	95.0	1	30	28.58	0	0	
9	2	10552	232	67	183.0	109.0	1	20	30.18	0	0	
10	2	11252	285	46	130.0	84.0	1	23	23.10	0	0	
11	2	11252	343	51	109.0	77.0	1	30	23.48	0	0	
12	2	11252	NA	58	155.0	90.0	1	30	24.61	0	0	
13	2	11263	228	43	180.0	110.0	0	0	30.30	0	0	
14	2	11263	230	49	177.0	102.0	0	0	31.36	0	1	
15	2	11263	220	55	180.0	106.0	0	0	31.17	1	1	
16	2	12629	205	63	138.0	71.0	0	0	33.11	0	0	
17	2	12629	220	70	149.0	81.0	0	0	36.76	0	0	
18	2	12806	313	45	100.0	71.0	1	20	21.68	0	0	
19	2	12806	NA	51	109.5	72.5	1	30	22.19	0	0	
20	2	12806	320	57	110.0	46.0	1	30	22.02	0	0	
21	1	14367	260	52	141.5	89.0	0	0	26.36	0	0	
22	1	14367	292	58	132.0	90.0	0	0	25.39	0	0	
23	1	14367	280	64	168.0	100.0	0	0	25.72	0	0	
24	1	16365	225	43	162.0	107.0	1	30	23.61	0	0	
25	1	16365	258	48	147.0	102.0	0	0	27.50	0	0	

Showing 1 to 26 of 11,627 entries



Statistical Concepts: Example 1

The Framingham Heart Study

- **Data :**
 -
- **Variables :**
 -
- **Subjects :**
 -
- **Observations :**
 -
- **Population :**
 -
- **Sample :**
 -



Statistical Concepts: Example 1

The Framingham Heart Study

- **Data :**
 - all the collected information for the purposes of this study
- **Variables :**
 -
- **Subjects :**
 -
- **Observations :**
- **Population :**
- **Sample :**

Statistical Concepts: Example 1

The Framingham Heart Study

- **Data :**
 - all the collected information for the purposes of this study
- **Variables :**
 - “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc
- **Subjects :**
 -
- **Observations :**
- **Population :**
- **Sample :**

Statistical Concepts: Example 1

The Framingham Heart Study

- **Data :**
 - all the collected information for the purposes of this study
- **Variables :**
 - “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, ..., etc
- **Subjects :**
 - participants (each one with unique ID number “randid”)
- **Observations :**
- **Population :**
- **Sample :**

Statistical Concepts: Example 1

The Framingham Heart Study

- **Data :**
 - all the collected information for the purposes of this study
- **Variables :**
 - “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, ..., etc
- **Subjects :**
 - participants (each one with unique ID number “randid”)
- **Observations :**
 - Each element of the dataset, e.g. for participant with “randin”=9428 :
 - “period”=2, “totchol”=283, “age”=54, ... etc.
- **Population :**
- **Sample :**

Statistical Concepts: Example 1

The Framingham Heart Study

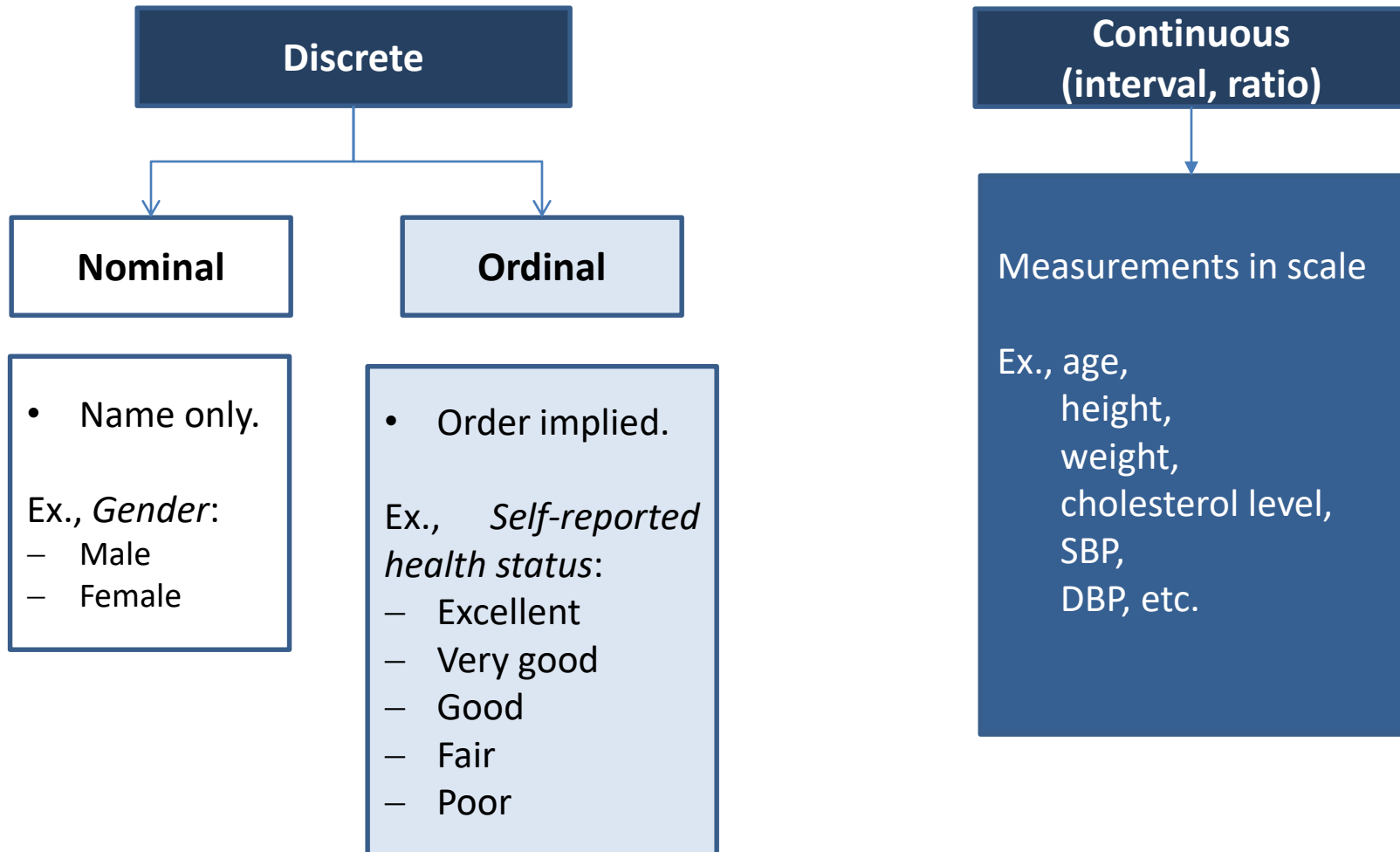
- **Data :**
 - all the collected information for the purposes of this study
- **Variables :**
 - “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc
- **Subjects :**
 - participants (each one with unique ID number “randid”)
- **Observations :**
 - Each element of the dataset, e.g. for participant with “randin”=9428 :
 - “period”=2, “totchol”=283, “age”=54, ... etc.
- **Population :**
 - ... *“a population of free living subjects in the community of Framingham, Massachusetts.”* ...
- **Sample :**
 -

Statistical Concepts: Example 1

The Framingham Heart Study

- **Data :**
 - all the collected information for the purposes of this study
- **Variables :**
 - “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc
- **Subjects :**
 - participants (each one with unique ID number “randid”)
- **Observations :**
 - Each element of the dataset, e.g. for participant with “randin”=9428 :
 - “period”=2, “totchol”=283, “age”=54, ... etc.
- **Population :**
 - ... *“a population of free living subjects in the community of Framingham, Massachusetts.”* ...
- **Sample :**
 - Subset of the population of size $n=4,434$

Classification of Variables



Classification of Variables: Example

The Framingham Heart Study

- Discrete Variables:
 - Nominal:
 - Ordinal:
- Continuous Variables:

The Framingham Heart Study

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Lec1.R x dat1 x

Filter

	sex	randid	totchol	age	sysbp	diabp	cursmoke	cigpday	bmi	diabetes	bpmeds	hearttrte
1	1	2448	195	39	106.0	70.0	0	0	26.97	0	0	
2	1	2448	209	52	121.0	66.0	0	0	NA	0	0	
3	2	6238	250	46	121.0	81.0	0	0	28.73	0	0	
4	2	6238	260	52	105.0	69.5	0	0	29.43	0	0	
5	2	6238	237	58	108.0	66.0	0	0	28.50	0	0	
6	1	9428	245	48	127.5	80.0	1	20	25.34	0	0	
7	1	9428	283	54	141.0	89.0	1	30	25.34	0	0	
8	2	10552	225	61	150.0	95.0	1	30	28.58	0	0	
9	2	10552	232	67	183.0	109.0	1	20	30.18	0	0	
10	2	11252	285	46	130.0	84.0	1	23	23.10	0	0	
11	2	11252	343	51	109.0	77.0	1	30	23.48	0	0	

Classification of Variables: Example

The Framingham Heart Study

- Discrete Variables:
 - Nominal: “sex”, “cursmoke”, etc.
 - Ordinal: “period”
- Continuous Variables:
 - “sysbp”, “bmi”, etc

Descriptive statistics for Discrete variables

- **Frequency (f):** Number (#) of subjects in each category.
- **Relative frequency ($\frac{f}{n} \cdot 100$):** Proportion (%) of subjects in each category.

Example: calculate number/proportion of subjects in each period

Period	Frequency (f)	Relative Frequency (%)	Cumulative Relative Frequency (%)
1	4434	$\frac{4434}{11627} \cdot 100 = 38.1$	38.1
2	3930	33.8	71.9
3	3260	28.1	100
Total	11627	100	

Descriptive statistics for Discrete variables

- **Frequency (f):** Number (#) of subjects in each category.
- **Relative frequency ($\frac{f}{n} \cdot 100$):** Proportion (%) of subjects in each category.

Example: calculate number/proportion of subjects in each period in R

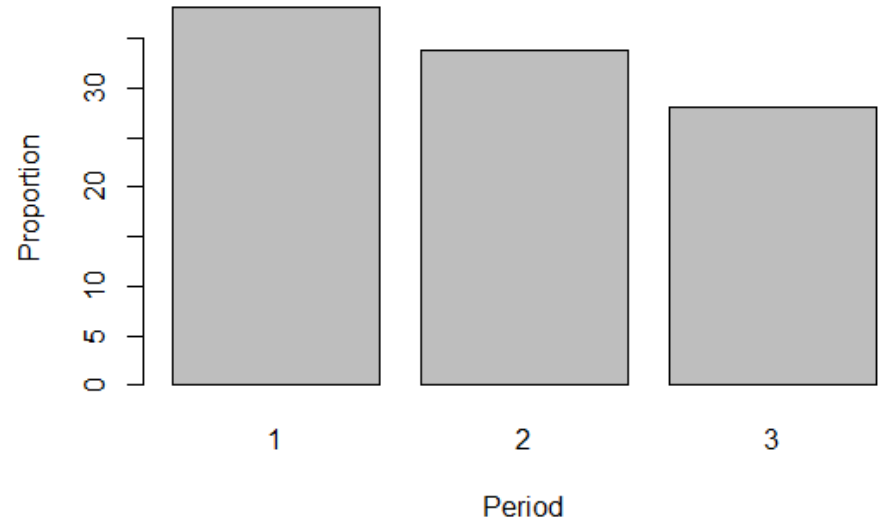
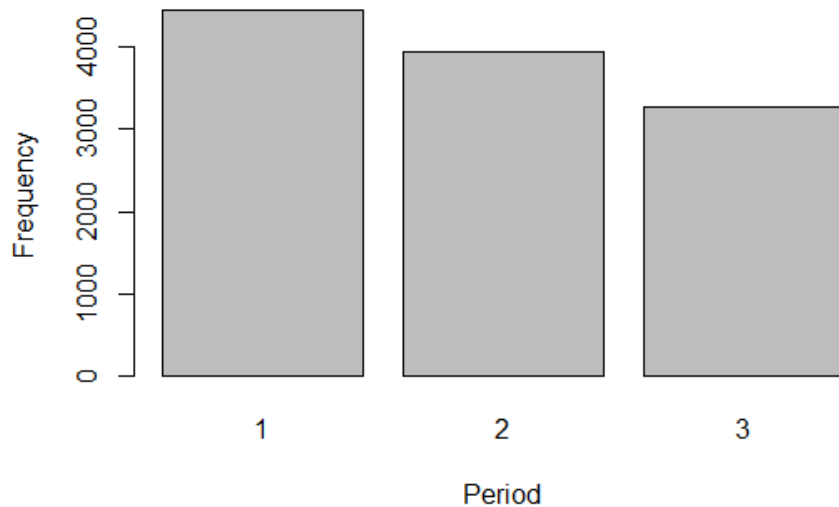
```
## frequency and relative frequency of period ##
tab1 = table(dat1$period)
n = sum(tab1)
rel_tab1 = tab1/n*100 ## alternative way: prop.table(tab1)*100
cum_tab1 = cumsum(rel_tab1)
cbind(tab1, rel_tab1, cum_tab1)
```

```
> cbind(tab1, rel_tab1, cum_tab1)
  tab1 rel_tab1 cum_tab1
1 4434 38.13537 38.13537
2 3930 33.80064 71.93601
3 3263 28.06399 100.00000
```

Graphical Methods for Discrete variables

- **Bar plots** : indicate frequency or relative frequency distribution

```
barplot(tab1, xlab="Period", ylab = "Frequency")
barplot(rel_tab1, xlab="Period", ylab="Proportion")
```



Descriptive statistics for Discrete variables

- Frequency and relative frequency ($\frac{f}{n} \cdot 100$) by groups

Example: calculate number/proportion of subjects in each period in R by sex (female if sex=2)

```
## period by sex ##
tab2 = table(dat1$period, dat1$sex)
tab2
rel_tab2 = prop.table(tab2, margin=2)*100 ## the option margin = 2 for column sum to be 100%
rel_tab2

cbind(tab2, rel_tab2)

> ## period by sex ##
> tab2 = table(dat1$period, dat1$sex)
> tab2

      1    2
1 1944 2490
2 1691 2239
3 1387 1876
> rel_tab2 = prop.table(tab2, margin=2)*100 ## the option margin = 2 for column sum to be 100%
> rel_tab2

      1      2
1 38.70968 37.69871
2 33.67184 33.89856
3 27.61848 28.40273
>
> cbind(tab2, rel_tab2)
      1    2      1      2
1 1944 2490 38.70968 37.69871
2 1691 2239 33.67184 33.89856
3 1387 1876 27.61848 28.40273
```

Descriptive statistics for Continuous variables

Measures of location	Measures of dispersion
<p>Indicate where the collected values of a variable are “located” compared to the range of possible values it can take.</p>	<p>Indicate how dispersed the collected values of a variable are.</p>

Descriptive statistics for Continuous variables

Measures of location	Measures of dispersion
<p>Indicate where the collected values of a variable are “located” compared to the range of possible values it can take.</p>	<p>Indicate how dispersed the collected values of a variable are.</p>
<ul style="list-style-type: none"> • Mean • Median • Quartiles • Mode • Min • Max 	<ul style="list-style-type: none"> • Range • Variance • Standard Deviation • Interquartile range (IQR) • Mean Absolute Deviation (MAD) • Coefficient of variation

Measures of Location :

Mean (\bar{x})

Definition	Formula
<ul style="list-style-type: none"> Average value. A typical value for the variable of interest. 	$\bar{x} = \frac{\sum_{i=1}^n X_i}{n}$

- Sample of $n=7$
- X = Systolic Blood Pressure in mmHg:

X_1	X_2	X_3	X_4	X_5	X_6	X_7
121	110	114	100	160	130	130

Measures of Location :

Mean (\bar{x})

Definition	Formula
<ul style="list-style-type: none"> Average value. A typical value for the variable of interest. 	$\bar{x} = \frac{\sum_{i=1}^n X_i}{n}$

- Sample of $n=7$
- X = Systolic Blood Pressure in mmHg:

X_1	X_2	X_3	X_4	X_5	X_6	X_7
121	110	114	100	160	130	130

$$\bar{x} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + X_3 + \dots + X_7}{n} = \frac{121 + 110 + 114 + \dots + 130}{7} = \frac{865}{7} = 123.57 \approx 123.6$$

Measures of Location :

Median

Definition	Formula
<ul style="list-style-type: none"> The middle value of the variable of interest. 50% of the collected values are less and 50% are greater than the median. 	<ul style="list-style-type: none"> If n odd: the $\frac{(n+1)^{th}}{2}$ observation If n even: mean of the $\frac{n}{2}^{th}$ and the $(\frac{n}{2} + 1)^{th}$ observations <p>in the ordered sample</p>

Unordered	X_1	X_2	X_3	X_4	X_5	X_6	X_7
	121	110	114	100	160	130	130
Ordered	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$
	100	110	114	121	130	130	160

Measures of Location :

Median

Definition	Formula
<ul style="list-style-type: none"> The middle value of the variable of interest. 50% of the collected values are less and 50% are greater than the median. 	<ul style="list-style-type: none"> If n odd: the $\frac{(n+1)^{th}}{2}$ observation If n even: mean of the $(\frac{n}{2})^{th}$ and the $(\frac{n}{2} + 1)^{th}$ observations <p>in the ordered sample</p>

Unordered	X_1	X_2	X_3	X_4	X_5	X_6	X_7
	121	110	114	100	160	130	130
Ordered	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$
	100	110	114	121	130	130	160

Median

$n=7 \rightarrow$ **odd** # \rightarrow median: $\frac{(7+1)}{2} = 4^{th}$ observation in the ordered sample
 \rightarrow median = $X_{(4)} = 121$

Measures of Location :

Median

Unordered	X_1	X_2	X_3	X_4	X_5	X_6
	121	110	114	100	160	130
Ordered	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$
	100	110	114	121	130	130

Measures of Location :

Median

Unordered	X_1	X_2	X_3	X_4	X_5	X_6
	121	110	114	100	160	130
Ordered	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$
	100	110	114	121	130	130

3th

4th

$n=6 \rightarrow$ even # \rightarrow median: mean of the $(\frac{6}{2})=3^{\text{th}}$ and the $(\frac{6}{2} + 1)=4^{\text{th}}$ observations in the ordered sample

$$\rightarrow \text{median} = \frac{X_{(3)} + X_{(4)}}{2} = \frac{114 + 121}{2} = 117.5$$

Measures of Location :

Quartiles

Definition

- First (Q_1): 25% of the collected values are less than Q_1 .
- Second (Q_2): 50% of the collected values are less than Q_2 (**median**).
- Third (Q_3): 75% of the collected values are less than Q_3 .

Measures of Location :

Percentiles

Definition

- q_p : $p\%$ of the collected values are less than q_p .
- E.g., q_1 is that value of the population (or sample) with 1% of the observed values being less and 99% being greater than it.

Measures of Location :

Mode / Min / Max

Definition

- **Min:** the minimum of the collected values ($X_{(1)}$).
- **Max:** the maximum of the collected values ($X_{(n)}$).
- **Mode:** the most frequent of the collected values.

Unordered	X_1	X_2	X_3	X_4	X_5	X_6	X_7
	121	110	114	100	160	130	130
Ordered	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$
	100	110	114	121	130	130	160

Measures of Location :

Mode / Min / Max

Definition

- **Min:** the minimum of the collected values ($X_{(1)}$).
- **Max:** the maximum of the collected values ($X_{(n)}$).
- **Mode:** the most frequent of the collected values.

Unordered	X_1	X_2	X_3	X_4	X_5	X_6	X_7
	121	110	114	100	160	130	130
Ordered	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$
	100	110	114	121	130	130	160

Min

Max

Mode = 130

Measures of Dispersion: Variance (s^2)

Definition	Formula
<ul style="list-style-type: none"> Average squared deviation from the mean. 	$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$

- $\bar{X} = 123.6$

X_1	X_2	X_3	X_4	X_5	X_6	X_7
121	110	114	100	160	130	130

Measures of Dispersion: Variance (s^2)

Definition	Formula
<ul style="list-style-type: none"> Average squared deviation from the mean. 	$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$

- $\bar{X} = 123.6$

X_1	X_2	X_3	X_4	X_5	X_6	X_7
121	110	114	100	160	130	130

$$\begin{aligned}
 S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{(X_1 - \bar{X})^2 + \dots + (X_7 - \bar{X})^2}{n-1} = \frac{(121 - 123.6)^2 + \dots + (130 - 123.6)^2}{7-1} = \\
 &= \frac{2247.72}{6} = 374.62 \approx 374.6
 \end{aligned}$$

Other Measures of Dispersion:

Definition	Formula
<ul style="list-style-type: none"> Standard deviation 	$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$
<ul style="list-style-type: none"> Mean Absolute Deviation (MAD) 	$\text{MAD} = \frac{\sum_{i=1}^n X_i - \bar{X} }{n}$
<ul style="list-style-type: none"> Range 	$\text{Max} - \text{Min}$
<ul style="list-style-type: none"> Interquartile Range (IQR) 	$Q_3 - Q_1$
<ul style="list-style-type: none"> Coefficient of variation 	$\frac{s}{\bar{X}}$

Descriptive Statistics for Continuous Variables

Example: The Framingham Heart Study

```
> ## the overall summary stat for sysbp ##
> describe(dat1$sysbp)
  vars      n  mean   sd median trimmed   mad  min max range skew kurtosis  se
x1     1 11627 136.32 22.8   132  134.34 20.76 83.5 295 211.5 0.94    1.37 0.21
>
> ## the summary stat for sysbp by sex ##
> describeBy(dat1$sysbp, dat1$sex)

Descriptive statistics by group
group: 1
  vars      n  mean   sd median trimmed   mad  min max range skew kurtosis  se
x1     1  5022 135.07 20.3   132  133.37 19.27 83.5 235 151.5 0.86    0.93 0.29
-----
group: 2
  vars      n  mean   sd median trimmed   mad  min max range skew kurtosis  se
x1     1  6605 137.28 24.49   133  135.15 22.24 83.5 295 211.5 0.93    1.28 0.3
```

SEX = 1 for male, 2 for female

Std.dev = $Var(X_i)$ to explain variation of sysbp

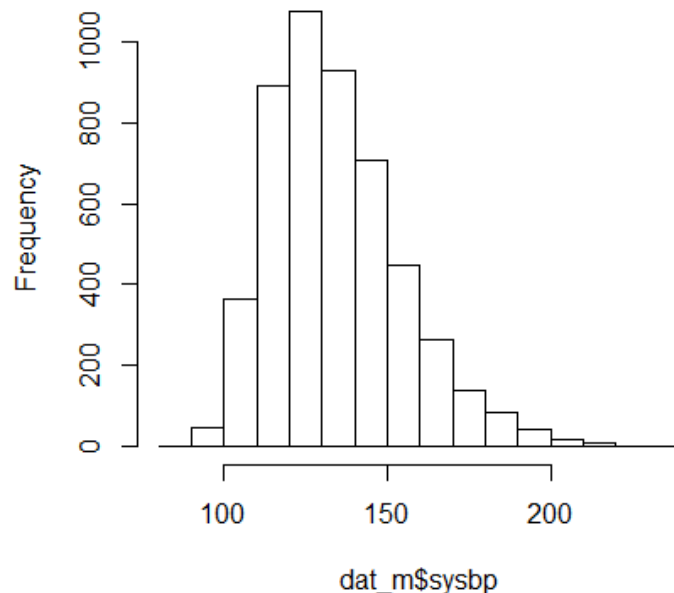
SE.mean = $\sqrt{Var(\bar{X})}$ to explain variation of MEAN sysbp

Graphical Methods for Continuous variables

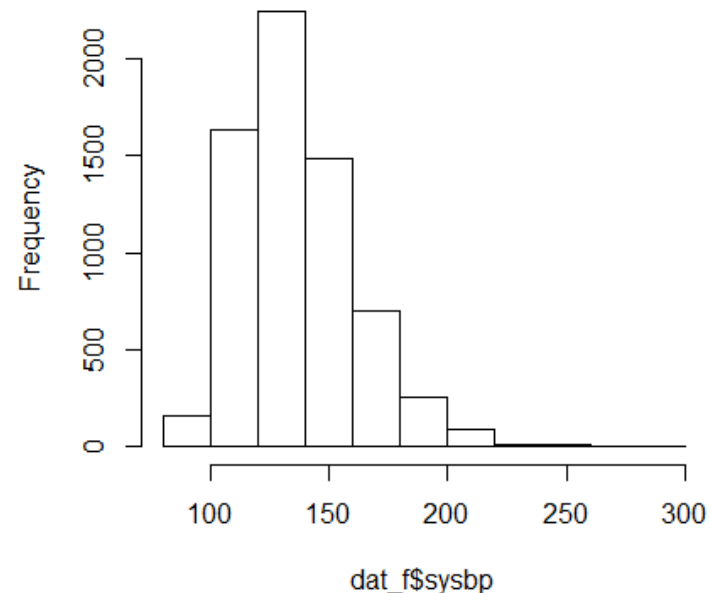
- **Histogram** : indicate the distribution of the values of a continuous variable.

```
## Histogram of sysbp by sex ##
dat_m = subset(dat1, sex==1) ## get a subset for male
dat_f = subset(dat1, sex==2) ## get a subset for female
par(mfrow = c(1,2)) ## to draw two plots side by side
hist(dat_m$sysbp, main="Histogram of sysbp for male")
hist(dat_f$sysbp, main="Histogram of sysbp for female")
```

Histogram of sysbp for male

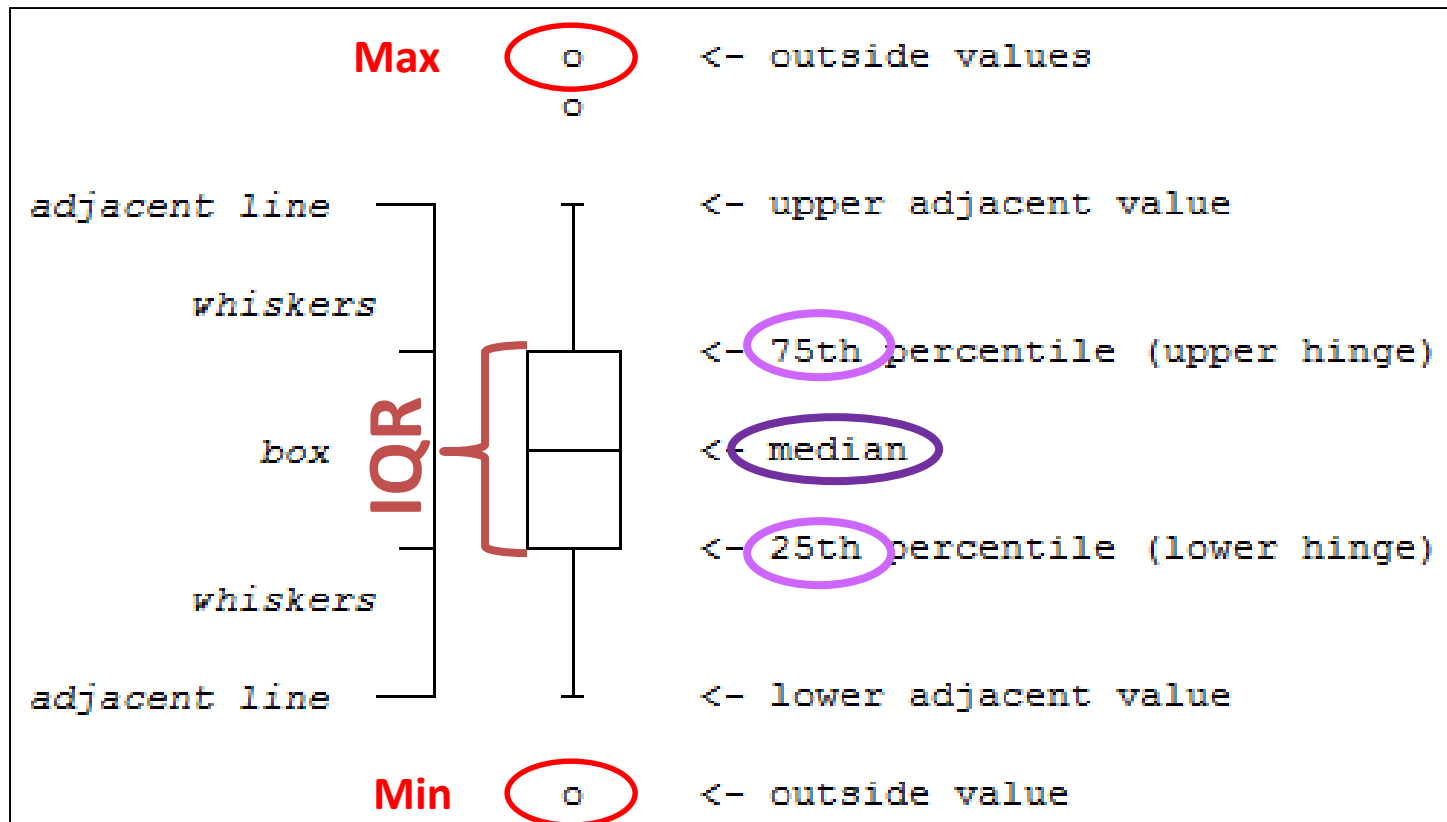


Histogram of sysbp for female



Graphical Methods for Continuous variables

Box - Plot : indicate the distribution of the values of a continuous variable, pointing out the following quantities:



Outliers

- Observations above $Q3 + 1.5IQR$ or below $Q1 - 1.5IQR$ are called, “outliers”, in the box plot.
- Outliers are not caused by typo or errors.
- Outliers are simply part of data, which can not be ignored.
- Outliers explain how many extreme values are located at tails of a distribution.

Graphical Methods for Continuous variables

- **Box-Plot** : the distribution of the values of a continuous variable.

```
## A box plot of sysbp by sex ##
par(mfrow = c(1,1))
boxplot(sysbp ~ sex, data=dat1, main="Box plot of sysbp by sex")
```

