Chapter 3

Data summary and transformation

Professor Jung Jin Lee
Soongsil University, Korea
New Uzbekistan University, Uzbekistan

Chapter 3 Data summary and transformation

3.1 Data summary using tables

- 3.1.1 Frequency table for a single variable
- 3.1.2 Two-dimensional frequency table for two variables
- 3.1.3 Multi-dimensional frequency table

3.2 Quantitative data summary using measure

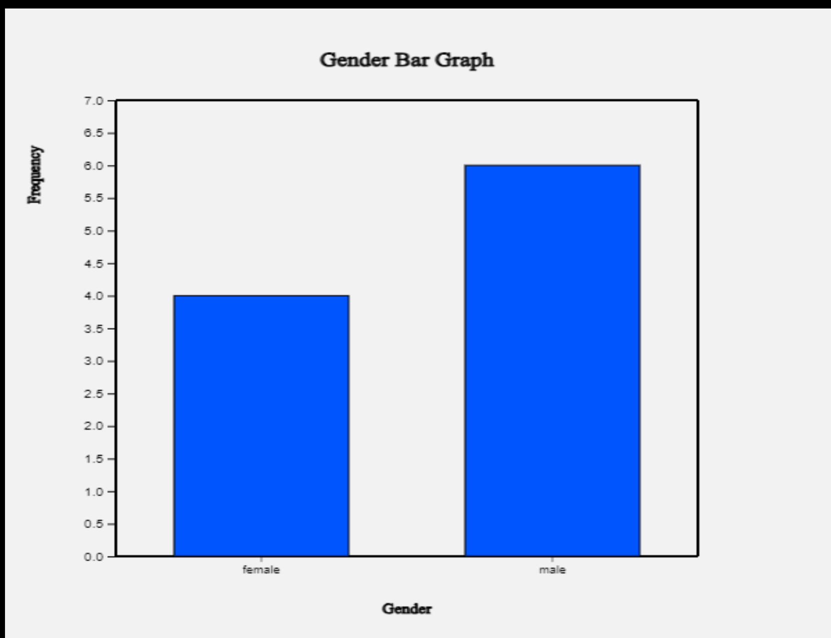
- 3.2.1 Measures of single quantitative variable
- 3.2.2 Measures of several quantitative variables
- 3.2.3 Similarity measures of observations

3.3 Data manipulation and transformation

3.4 Dimension reduction: Principal component analysis

3.1 Data summary using tables

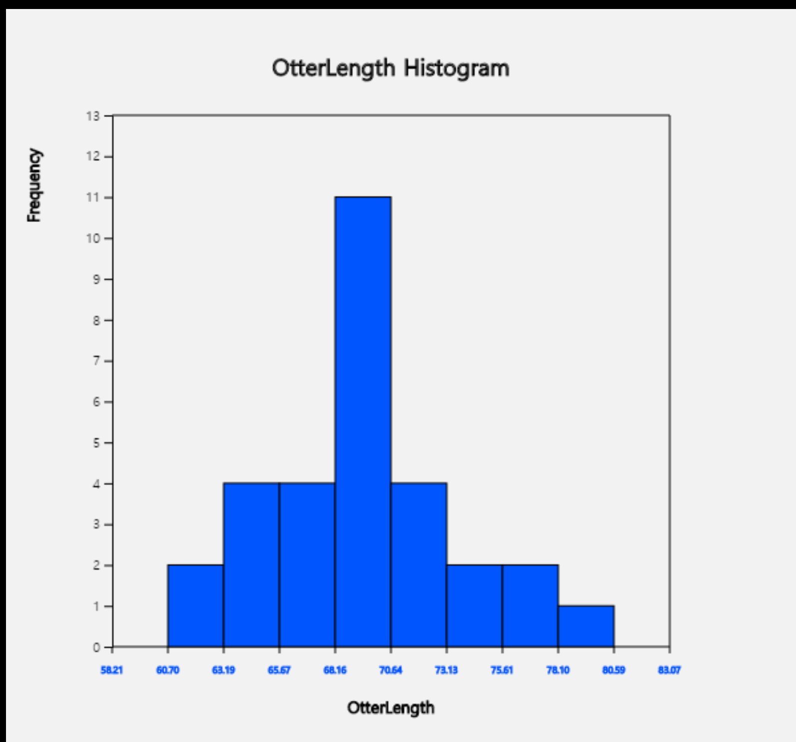
❖ Frequency table for a single variable



Frequency Table	Analysis Var	(Gender)		
Var Value	Value Label	Frequency	Relative Frequency	Cumulated Relative Frequency (%)
female		4	40.0	40.0
male		6	60.0	100.0
Total		10	100.0	
	Missing Observations	0		

3.1 Data summary using tables

❖ Frequency table for a single quantitative variable



Histogram Frequency Table	Group Name	0
Interval (OtterLength)	Group 1 (null)	Total
1 [60.70, 63.19)	2 (6.7%)	2 (6.7%)
2 [63.19, 65.67)	4 (13.3%)	4 (13.3%)
3 [65.67, 68.16)	4 (13.3%)	4 (13.3%)
4 [68.16, 70.64)	11 (36.7%)	11 (36.7%)
5 [70.64, 73.13)	4 (13.3%)	4 (13.3%)
6 [73.13, 75.61)	2 (6.7%)	2 (6.7%)
7 [75.61, 78.10)	2 (6.7%)	2 (6.7%)
8 [78.10, 80.59)	1 (3.3%)	1 (3.3%)
Total	30 (100%)	30 (100%)

3.1 Data summary using tables

❖ Frequency table for two variables

File: MaritalByGender.csv [EditVar]

Analysis Var: 2: Marital — by Group: 1: Gender

(Selected data: Raw Data) (Summary Data: Multiple Selection)

SelectedVar: V2 by V1, [Cancel]

	Gender	Marital	V3	V4	V5	V
1	1	1				
2	2	2				
3	1	1				
4	2	1				
5	1	2				
6	1	1				
7	1	1				
8	2	2				
9	1	3				
10	2	1				

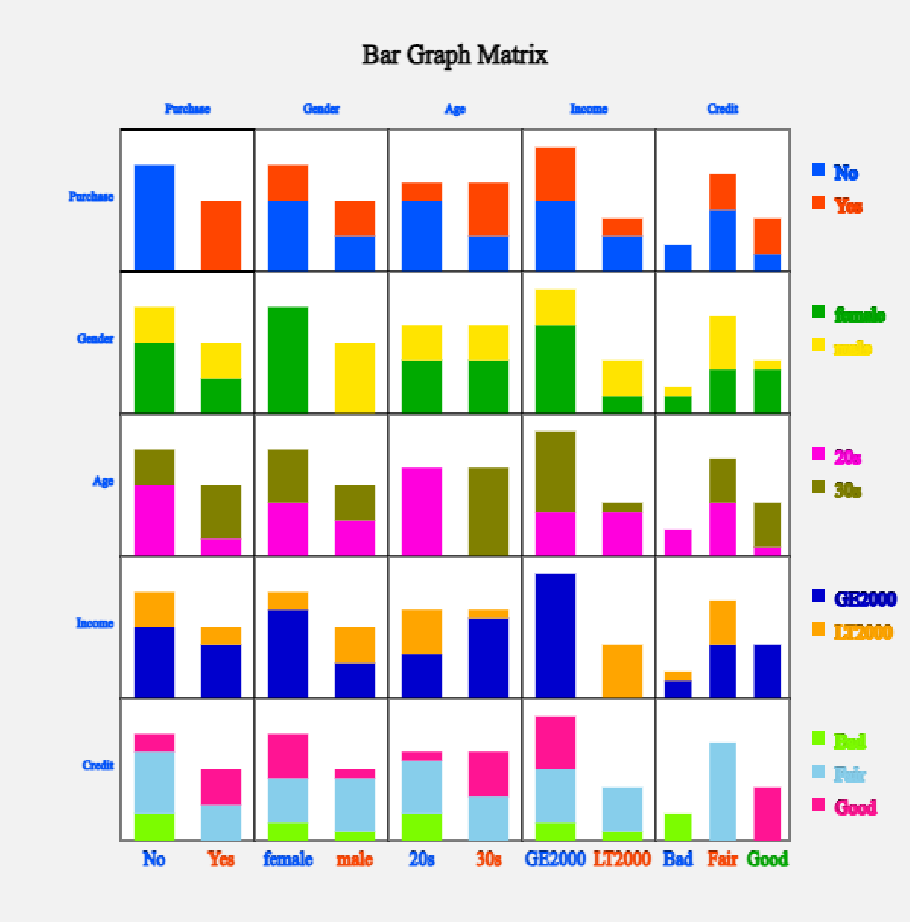
Cross Table	Col Variable	(Marital)			
Row Variable (Gender)	1	2	3	Total	
Group 1 Row % Col % Tot %	4 66.7% 66.7% 40.0%	1 16.7% 33.3% 10.0%	1 16.7% 100.0% 10.0%	6 100.0% 60.0%	
Group 2 Row % Col % Tot %	2 50.0% 33.3% 20.0%	2 50.0% 66.7% 20.0%	0 0.0% 0.0% 0.0%	4 100.0% 40.0%	
Total Row % Col %	6 60.0% 100.0%	3 30.0% 100.0%	1 10.0% 100.0%	10 100.0% 100.0%	
	Missing Observations	0			
Independence Test					
Sum of χ^2 value	1.667	deg of freedom	2	p-value	0.4346

3.1 Data summary using tables

❖ Multidimensional frequency table

Table 2.1.3 Survey on twenty customers of a computer store

id	Gender	Age	Income	Credit	Purchase
1	male	20s	LT2000	Fair	Yes
2	female	30s	GE2000	Good	No
3	female	20s	GE2000	Fair	No
4	female	20s	GE2000	Fair	Yes
5	female	20s	LT2000	Bad	No
6	female	30s	GE2000	Fair	No
7	female	30s	GE2000	Good	Yes
8	male	20s	LT2000	Fair	No
9	female	20s	GE2000	Good	No
10	male	30s	GE2000	Fair	Yes
11	female	30s	GE2000	Good	Yes
12	female	20s	LT2000	Fair	No
13	male	30s	GE2000	Fair	No
14	male	30s	LT2000	Fair	Yes
15	female	30s	GE2000	Good	Yes
16	female	30s	GE2000	Fair	No
17	female	20s	GE2000	Bad	No
18	male	20s	GE2000	Bad	No
19	male	30s	GE2000	Good	Yes
20	male	20s	LT2000	Fair	No



3.1 Data summary using tables

❖ Multidimensional frequency table

Cross Table	Purchase		Gender		Age		Income		Credit		
	No	Yes	female	male	20s	30s	GE2000	LT2000	Bad	Fair	Good
Purchase: No	12	0	8	4	8	4	8	4	3	7	2
Purchase: Yes	0	8	4	4	2	6	6	2	0	4	4
Gender: female	8	4	12	0	6	6	10	2	2	5	5
Gender: male	4	4	0	8	4	4	4	4	1	6	1
Age: 20s	8	2	6	4	10	0	5	5	3	6	1
Age: 30s	4	6	6	4	0	10	9	1	0	5	5
Income: GE2000	8	6	10	4	5	9	14	0	2	6	6
Income: LT2000	4	2	2	4	5	1	0	6	1	5	0
Credit: Bad	3	0	2	1	3	0	2	1	3	0	0
Credit: Fair	7	4	5	6	6	5	6	5	0	11	0
Credit: Good	2	4	5	1	1	5	6	0	0	0	6

Multidimension Frequency Table	Purchase	Gender	Age	Income	Credit	Frequency	%
1	No	female	20s	GE2000	Bad	1	5.00
2	No	female	20s	GE2000	Fair	1	5.00
3	No	female	20s	GE2000	Good	1	5.00
4	No	female	20s	LT2000	Bad	1	5.00
5	No	female	20s	LT2000	Fair	1	5.00
6	No	female	20s	LT2000	Good	0	0.00
7	No	female	30s	GE2000	Bad	0	0.00
8	No	female	30s	GE2000	Fair	2	10.00
9	No	female	30s	GE2000	Good	1	5.00
10	No	female	30s	LT2000	Bad	0	0.00
11	No	female	30s	LT2000	Fair	0	0.00
12	No	female	30s	LT2000	Good	0	0.00
13	No	male	20s	GE2000	Bad	1	5.00
14	No	male	20s	GE2000	Fair	0	0.00
15	No	male	20s	GE2000	Good	0	0.00
16	No	male	20s	LT2000	Bad	0	0.00
17	No	male	20s	LT2000	Fair	2	10.00
18	No	male	20s	LT2000	Good	0	0.00
19	No	male	30s	GE2000	Bad	0	0.00
20	No	male	30s	GE2000	Fair	1	5.00

3.2 Quantitative data summary using measures

❖ Measures for central tendency

- Average(mean), median, mode, weighted average

$$\text{Average} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- population mean: μ, sample mean: $\overline{x}$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Mean is influenced by extreme points:
  very large or small value.

- Sample mean has a good characteristic to estimate population mean.

3.2 Quantitative data summary using measures

❖ Measures for central tendency

- **Median** is the value placed centrally when data is listed in order of size
  - Sample median m, population median M

$$Median = \begin{cases} \frac{(n+1)}{2}th\ data & if\ n\ is\ odd \\ Mean\ of\ (\frac{n}{2})th,\ (\frac{n+2}{2})th & if\ n\ is\ even \end{cases}$$

- The median is not sensitive for an extreme point.

- **Mode** is the most frequently occurred value.

3.2 Quantitative data summary using measures

❖ Measures for central tendency

- **Trimmed mean** compensates for the disadvantage of the simple mean.
  => list data in order
  => remove certain portions of large and small values
  => take an average of the remaining data
- It is often used to prevent biased judging by referees in sports such as gymnastics and figure skating

- **Weighted Mean** $= \dfrac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n} = \dfrac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

3.2 Quantitative data summary using measures

❖ Measures for central tendency

[Example] Quiz scores of seven students in a class:

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

Find the mean and median.

<Answer>

- The sample mean is as follows:

$$\bar{x} = \frac{5+6+3+7+9+4+8}{7} = 6$$

- To find the sample median, arrange data in ascending order

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

- Since the sample size is an odd number, median is $(\frac{n+1}{2})^{th}$ data which is $(\frac{7+1}{2})^{th}$ that is m = 6,

3.2 Quantitative data summary using measures

[Example] An Olympic Gymnastics competition was judged by eight referees, and their scores were as follows.

9.0 9.5 9.3 7.2 10.0 9.1 9.4 9.0

Find mean, median, trimmed mean excluding maximum and minimum.

<Answer>

- This data mean is is not a sample but a population.

  $\mu = (9.0 + 9.5 + 9.3 + 7.2 + 10.0 + 9.1 + 9.4 + 9.0) / 8 = 9.063$
- To find the median, arrange the data in ascending order.

  7.2 9.0 9.0 9.1 9.3 9.4 9.5 10.0
- Since n=8 is an even number, median is the average of $(\frac{n}{2})^{th} = (\frac{8}{2})^{th} =$ (=9.1) and $(\frac{n+2}{2})^{th} = (\frac{8+2}{2})^{th}$(=9.3). M = (9.1 + 9.3)/2 = 9.2.
- Trimmed mean is the average of the remaining numbers except the minimum of 7.2 and the maximum of 10.0.

  Trimmed mean = (9.0 + 9.0 + 9.1 + 9.3 + 9.4 + 9.5) / 6 = = 9.217
- Median or trimmed better representative of the data than mean.

3.2 Quantitative data summary using measures

❖ Measures for dispersion

- Variance is the average of the squared distances from data to the mean,
  - If data are spread widely around mean, variance increase
  - If data is concentrated around the mean, variance is small

Population variance $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \quad (N:\ \textit{number of population data})$$

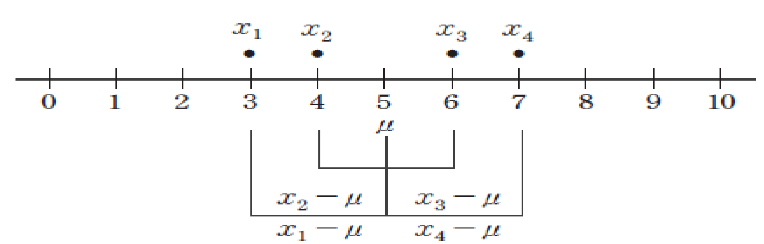
Sample variance $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (n:\ \textit{number of sample data})$$

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

✓ There are important reasons for using n-1 instead n when calculating the sample variance.

✓ => Correct estimation for population mean



$$\sigma^2 = \frac{(-2)^2 + (-1)^2 + 1^2 + 2^2}{4} = 2.5$$

3.2 Quantitative data summary using measures

❖ Measures for dispersion

[Example } Calculate mean and standard deviation from sample data
5, 6, 3, 7, 9, 4, 8.

<Answer>

- $\bar{x} = \frac{5+6+3+7+9+4+8}{7} = 6$
- $s^2 = \frac{(5-6)^2+(6-6)^2+(3-6)^2+(7-6)^2+(9-6)^2+(4-6)^2+(8-6)^2}{7-1} = \frac{28}{6} = 4.6$
- $s = \sqrt{s^2} = \sqrt{4.667} = 2.16$

3.2 Quantitative data summary using measures

❖ Measures for dispersion

- Coefficient of variation is the division of the standard deviation by its mean to compare data in different units

Population coefficient of variation $C = \frac{\sigma}{\mu} \times 100 \quad \text{(unit \%)}$

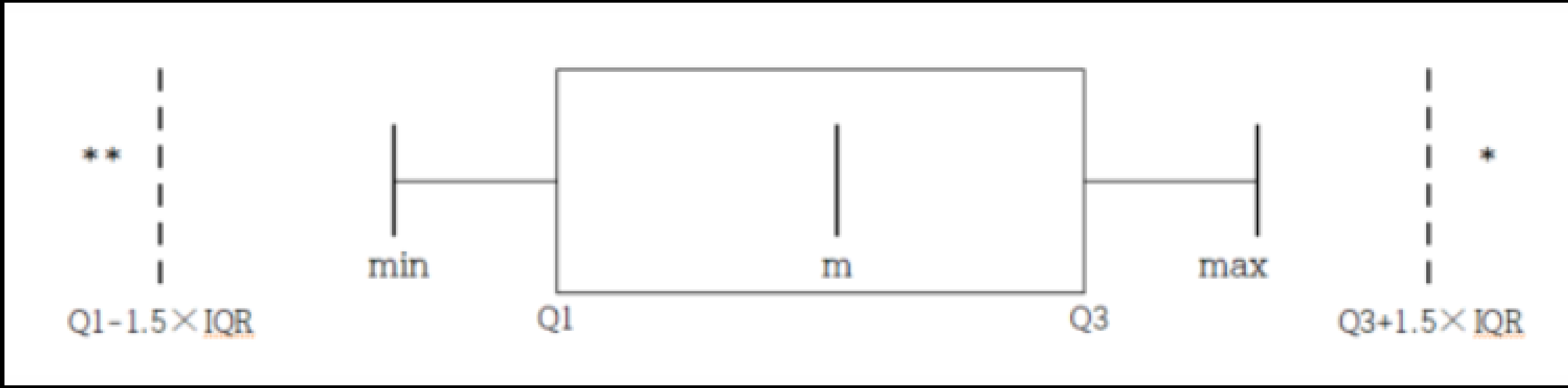
Sample coefficient of variation $c = \frac{s}{\bar{x}} \times 100 \quad \text{(unit \%)}$

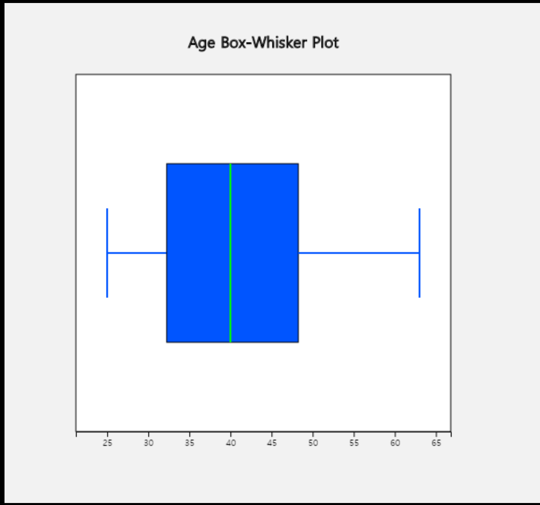
- Range = maximum - minimum
  - easy to calculate, but not a good measure if extreme points.

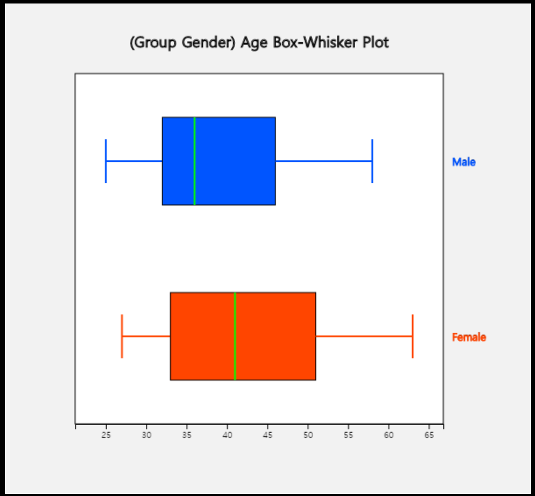
- p percentile : there are p% of observations less than(≤) this value, (100-p)% of observations above(≥) this value
  - 25 percentile: 1st quartile (Q1), 75 percentile: 3rd quartile (Q3).
- Inter-quartile range (IQR) = Q3 - Q1

3.2 Quantitative data summary using measures

❖ Box-whiskers plot







3.2 Quantitative data summary using measures

❖ Measures for several variables

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$S = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1m} \\ s_{21} & s_2^2 & \dots & s_{2m} \\ \dots & \dots & \dots & \dots \\ s_{m1} & s_{m2} & \dots & s_m^2 \end{bmatrix}$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$R = \begin{bmatrix} 1 & r_{12} & \dots & r_{1m} \\ r_{21} & 1 & \dots & r_{2m} \\ \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & \dots & 1 \end{bmatrix}$$

3.2 Quantitative data summary using measures

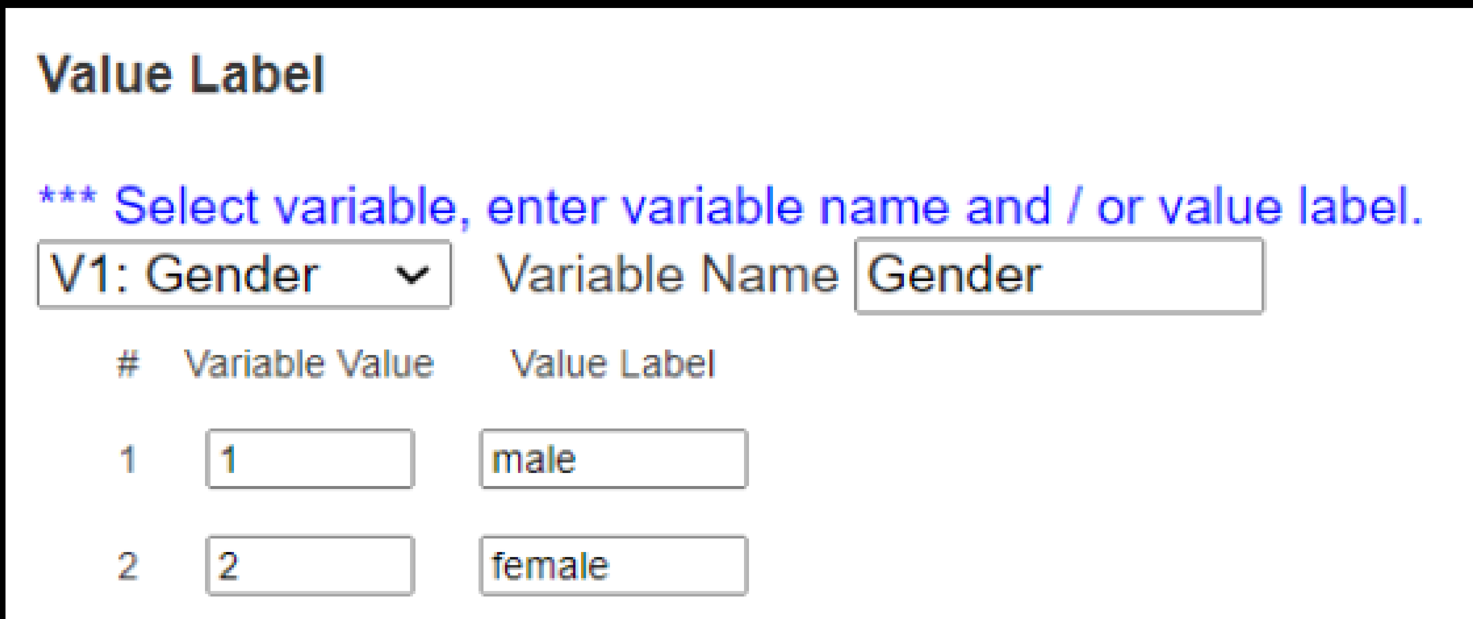
❖ Similarity measures between observations

Table 3.2.4 Distance measures between data of observations

Data type	Distance	Note
Qualitative	$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{00}+f_{11}}{f_{00}+f_{01}+f_{10}+f_{11}}$	Simple match coefficient f_{00} : number of variables such as $x_j = 0$ and $y_j = 0$ f_{01} : number of variables such as $x_j = 0$ and $y_j = 1$ f_{10} : number of variables such as $x_j = 1$ and $y_j = 0$ f_{11} : number of variables such as $x_j = 1$ and $y_j = 1$
Quantitative	$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{m} \lvert x_j - y_j \rvert^r \right)^{1/r}$	Minkowski distance
	if r = 1, it is called L_1 distance. $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{m} \lvert x_j - y_j \rvert$	Manhattan distance or city block distance
	if r = 2, it is called L_2 distance. $d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{m} \lvert x_j - y_j \rvert^2 \right)^{1/2}$	Euclid distance
	if r = ∞, it is called L_∞ distance. $d(\boldsymbol{x}, \boldsymbol{y}) = max_{j=1}^{m} \lvert x_j - y_j \rvert$	Maximum distance

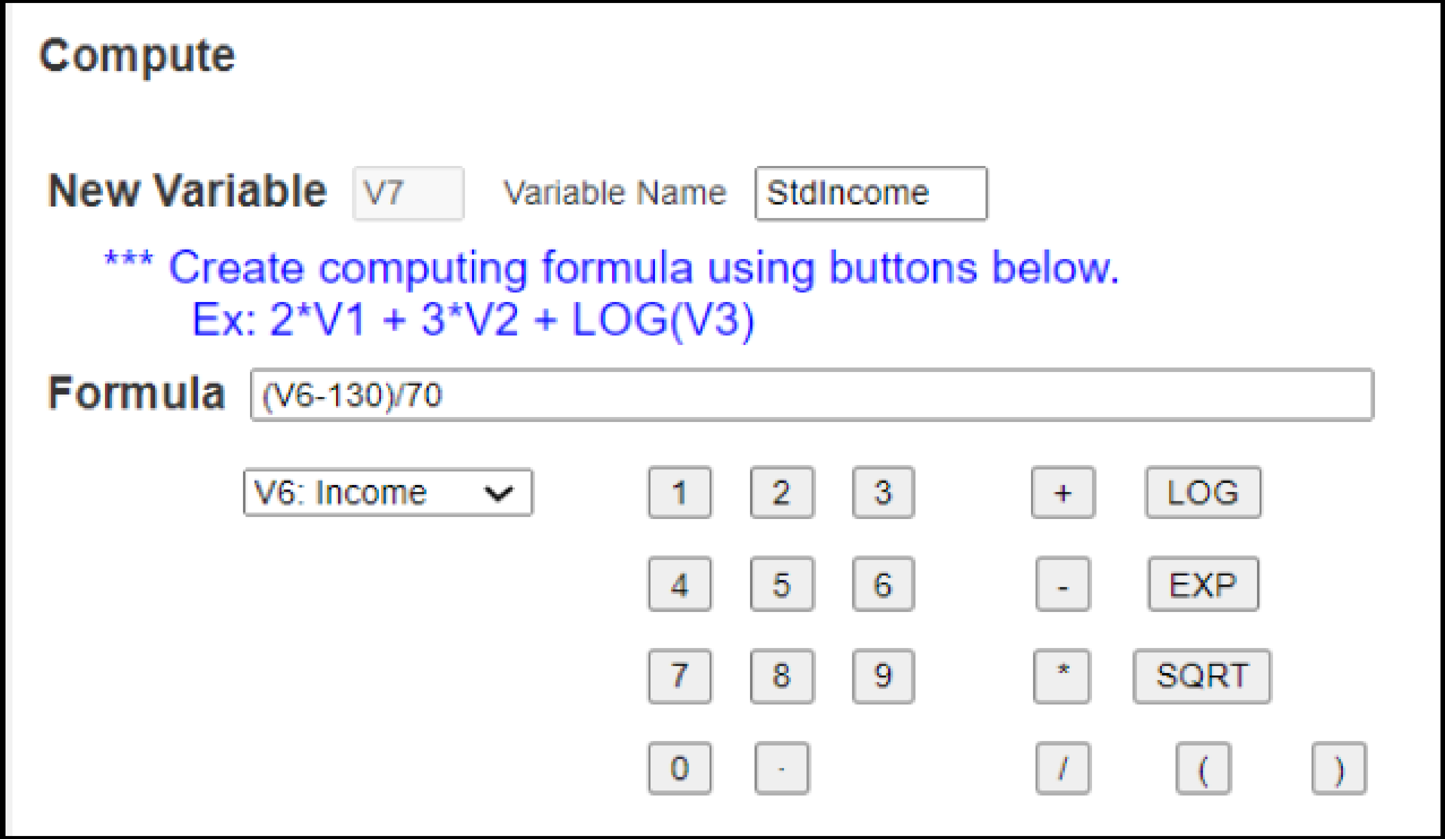
3.3 Data manipulation and transformation

❖ Value label



3.3 Data manipulation and transformation

❖ Compute



3.3 Data manipulation and transformation

❖ Recode: Categorize

Recode: Category

*** Select variable for Category, enter 'Interval Start' and 'Interval Width'.

New Variable — **Categorize Variable**

V8 Variable Name: AgeCategory — V3: Age — min = 20, max = 59

Interval Start: 20 (≤ min)

Interval Width: 10 (≤ 9 Category)

[Category List Check]

#	Category Interval					Category Label
1	20	≤	V3	<	30	[20, 30)
2	30	≤	V3	<	40	[30, 40)
3	40	≤	V3	<	50	[40, 50)
4	50	≤	V3	<	60	[50, 60)

3.3 Data manipulation and transformation

❖ Recode: Value

Recode: Value

*** Select variable for Recode, enter 'New Value'.

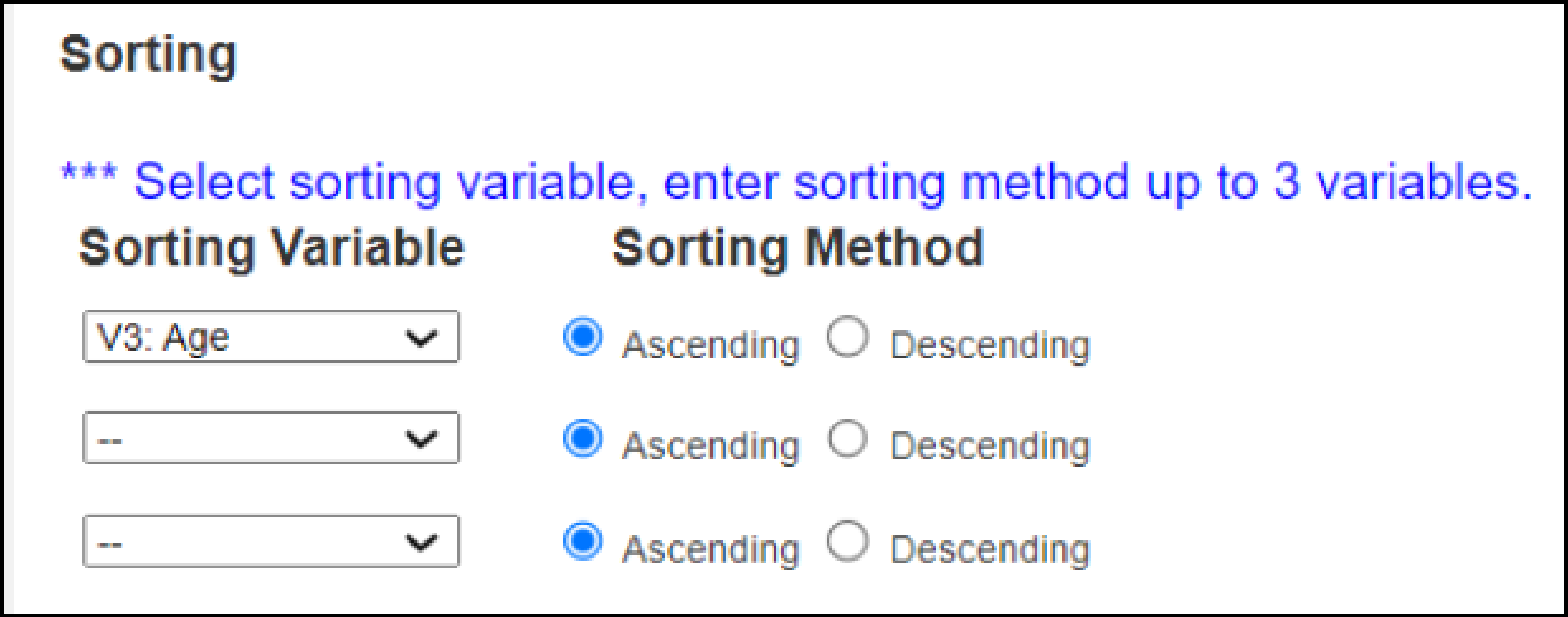
New Variable | **Recode Variable**

V9 Variable Name JobNew | V4: Job | * Allow recoding up to 9 values.

#	Current Value	New Value (Missing value: "MISSING")
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	8
7	7	7
8	8	8

3.3 Data manipulation and transformation

❖ Sorting



3.3 Data manipulation and transformation

❖ Conditional selection: Select if

Select If

*** Select up to 3 variables, enter their conditions.

Variable for Select	Relation Operator	Value
V3: Age	=	1
V3: Age	≥	30
--		

3.4 Dimension reduction

- Reducing data size using sampling
  - Simple random sampling
  - Stratified sampling

- Reducing variable size using principle component analysis

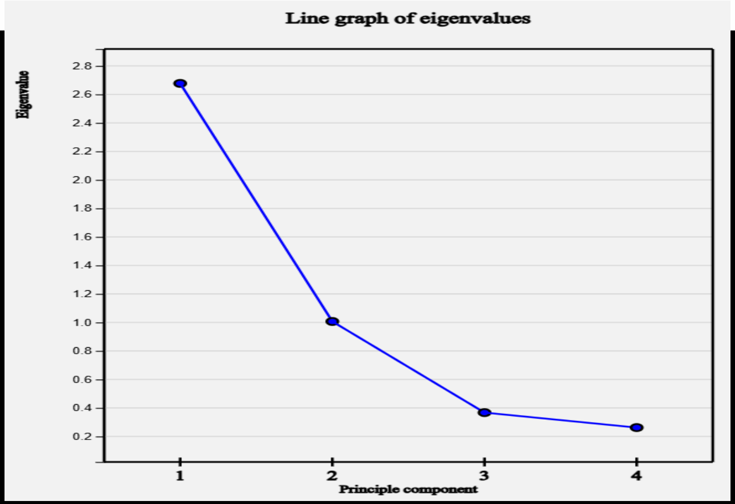
3.4 Dimension reduction

❖ Principle component analysis

Assume that a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ has a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The diagonal elements of $\boldsymbol{\Sigma}$ are the variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$ of each random variable. Let the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ be $\lambda_1, \lambda_2, \ldots, \lambda_m$, which are arranged in descending order of magnitude, and let the eigenvectors corresponding to each eigenvalue be $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m$. If $\boldsymbol{E}$ is a $m \times m$ matrix with these eigenvectors as columns such as $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m]$, the linear transformation $\boldsymbol{Y} = \boldsymbol{E}\boldsymbol{X}$ creates new variables $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$, which are called **principal components**. The principal component Y_j is a linear combination of $X_1, X_2, \ldots, X_m$ with coefficients of the eigenvectors.

$$\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_m^2 = \lambda_1 + \lambda_2 + \ldots + \lambda_m$$

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{E}'\boldsymbol{\Sigma}\boldsymbol{E} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_m \end{bmatrix}$$



Summary

- Categorical data summary using tables:
  - one-dimension, two-dimension, multi-dimension frequency table
- Quantitative data summary using measures:
  - central tendency: average, median, mode, weighted average
  - dispersion: variance, standard deviation, range, inter-quartile range
  - distance matrix
- Data manipulation and transformation:
  - value label, compute, recode-categorization, recode-value, sorting, select if
- Dimension reduction: sampling, principle component analysis

Thank you !!!