Descriptive Statistics

Sunthud Pornprasertmanit Chulalongkorn University

Purpose

1) Describe sets of data, such as mean
2) Examine Irregularities in sets of data, such as outlier

Methods of Descriptive Statistics

1) Graphical
2) Numerical

Graphical Illustration

Categorical Variable

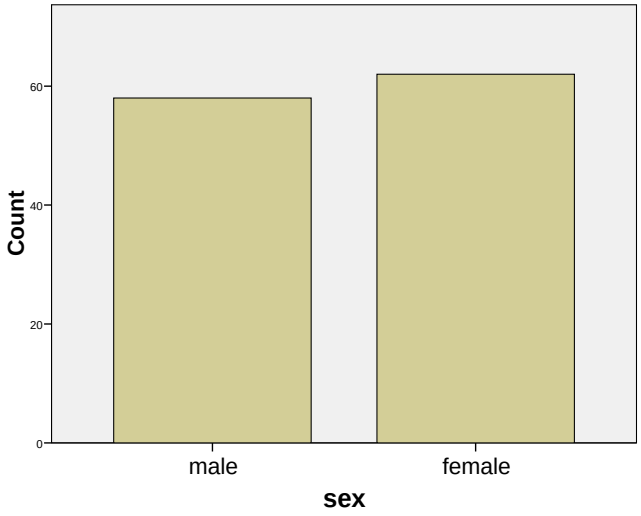
- Frequency Distribution

sex

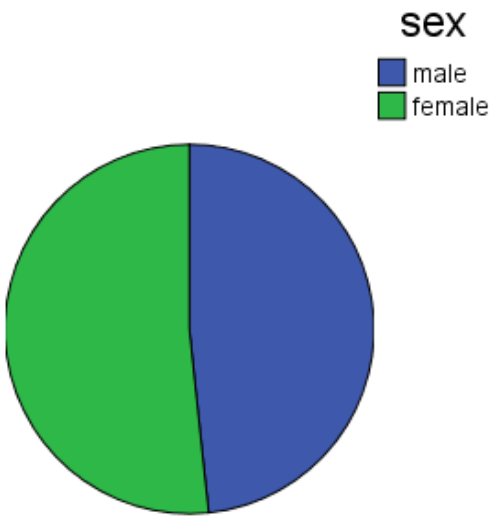
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	57	47.5	47.9	47.9
	female	62	51.7	52.1	100.0
	Total	119	99.2	100.0	
Missing	99	1	.8		
Total		120	100.0		

Author Note
This article was written in June 2007 for teaching in Introduction to Statistics in Psychology Class, Faculty of Psychology, Chulalongkorn University
Correspondence to Sunthud Pornprasertmanit. Email: psunthud@gmail.com

- Bar graph



- Pie chart



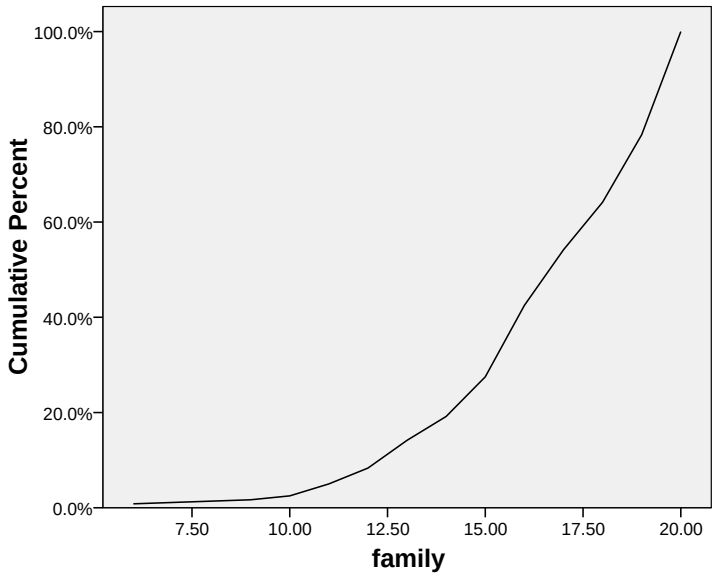
Numerical Variable

- Frequency Distribution

groupss

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	30-32	1	.8	.8	.8
	42-44	1	.8	.8	1.7
	48-50	3	2.5	2.5	4.2
	51-53	6	5.0	5.0	9.2
	54-56	10	8.3	8.3	17.5
	57-59	10	8.3	8.3	25.8
	60-62	20	16.7	16.7	42.5
	63-65	12	10.0	10.0	52.5
	66-68	19	15.8	15.8	68.3
	69-71	17	14.2	14.2	82.5
	72-74	14	11.7	11.7	94.2
	75-77	7	5.8	5.8	100.0
	Total	120	100.0	100.0	

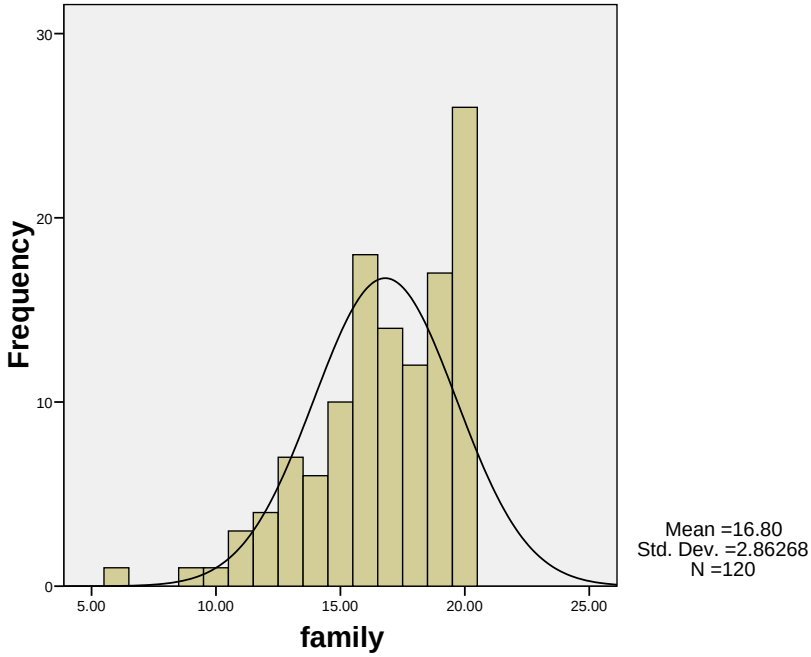
- Cumulative Frequency Distribution



Bell Shaped Distribution

Nonnormal Distribution

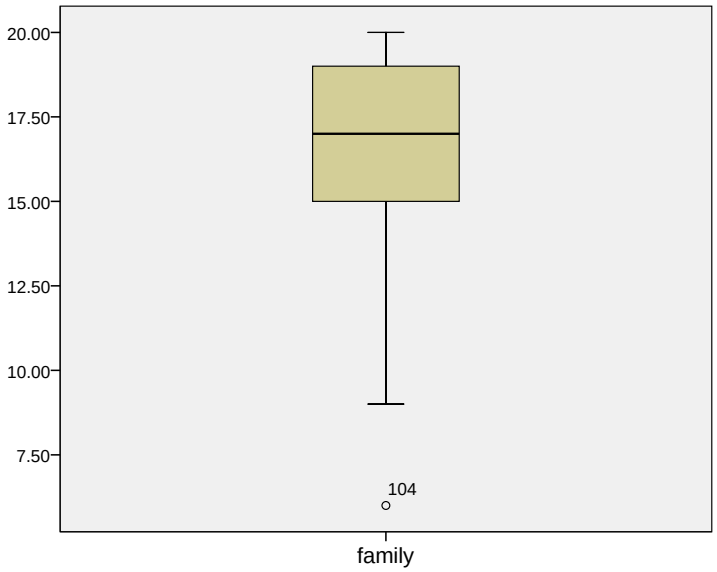
- Histogram and Polygon



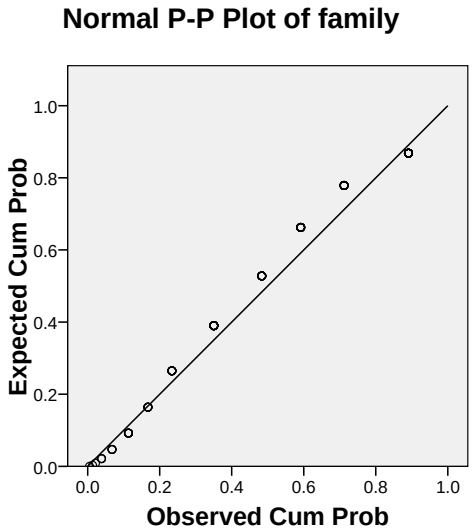
- Stem and Leaf Displays

```
family Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1.00 Extremes    (=<6.0)
     1.00         9 .  0
     1.00        10 .  0
     3.00        11 .  000
     4.00        12 .  0000
     7.00        13 .  0000000
     6.00        14 .  000000
    10.00        15 .  0000000000
    18.00        16 .  000000000000000000
    14.00        17 .  00000000000000
    12.00        18 .  000000000000
    17.00        19 .  00000000000000000
    26.00        20 .  00000000000000000000000000

 Stem width:      1.00
 Each leaf:       1 case(s)
```

- Boxplot



- Normal Probability Plot



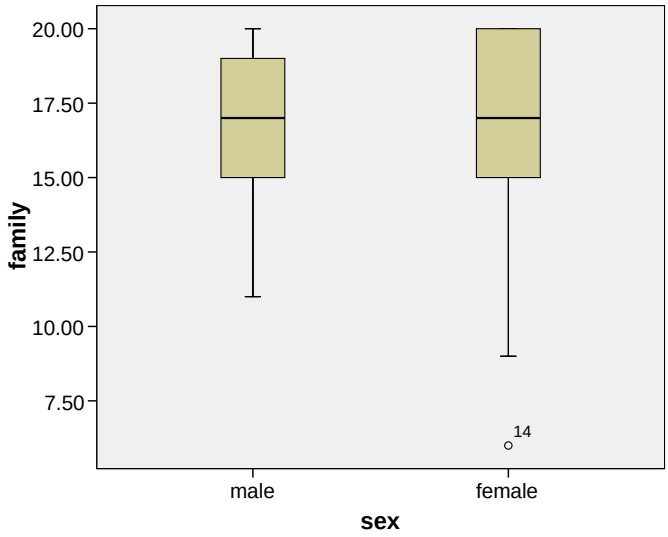
Two or more variables

- Crosstabs

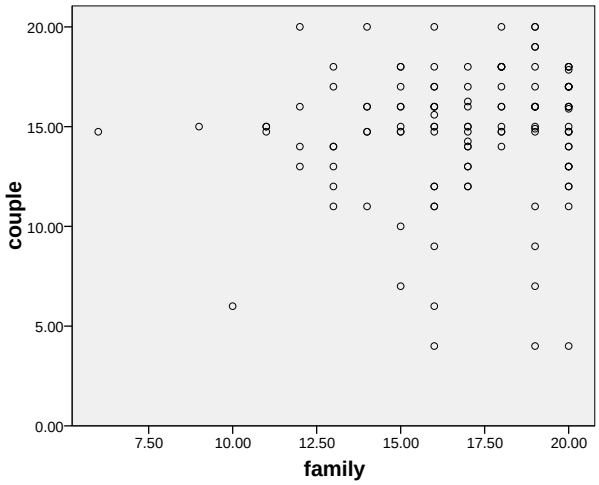
sex * education level Crosstabulation

			education level		Total
			high school	undergraduate	high school
sex	male	Count	29	29	58
		% within education level	47.5%	49.2%	48.3%
	female	Count	32	30	62
		% within education level	52.5%	50.8%	51.7%
Total		Count	61	59	120
		% within education level	100.0%	100.0%	100.0%

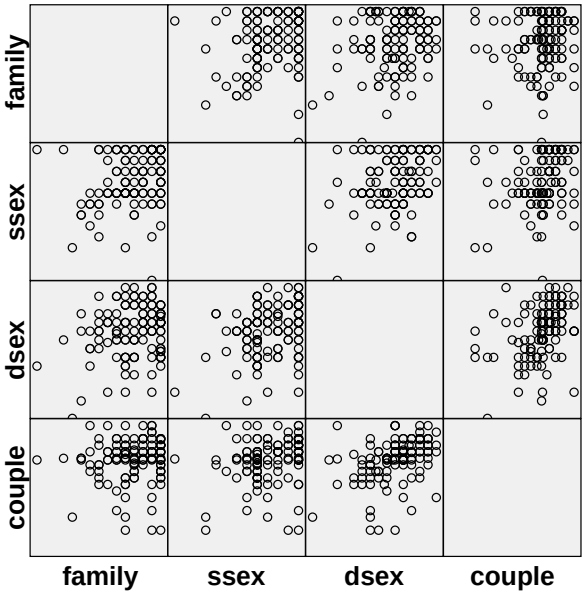
- Boxplot



- Scatterplot



- Scatterplot Matrix



Numerical Illustration

Measures of Central Tendency

- Mode
- Median
- Mean

Summation Notation Extreme Score (Outlier)

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

- Weighted Mean

$$\bar{X} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

- Merits of each measures of central tendency

	Mean	Median	Mode
Nominal scale	?		✓
Ordinal scale	?	?	✓
Quantitative scale	✓	✓	✓
Sampling stability	1	2	3
Mathematical tractability	✓		
Very skewed distribution	?	✓	
Extreme score	?	✓	
Truncated distribution	?	✓	
Open-ended distribution		✓	
Discrete variable	?	?	?
Use all values to measure CT	✓		
One value for one operation	✓	✓	

Measures of Location

- Median
- Quartile
- Decide
- Percentile
  - Percentile Point → Value
  - Percentile Rank → Rank, Percentage below

$$P_{\%} = X_{ll} + i\left(\frac{n(P_R/100) - \sum f_b}{f_i}\right)$$

$$P_R = \frac{100}{n} + \left[\sum f_b + \frac{f_i(P_{\%} - X_{ll})}{i}\right]$$

Measures of Variability

- Range
- Semi-interquartile Range

$$Q = \frac{Q_3 - Q_1}{2}$$

Deviation Score Sum of squared deviation (SS)

- Standard Deviation and Variance

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

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

- Index of Dispersion

$$D = \frac{DP}{DP_{max}}$$

	Range	IQR	SD	ID
Nominal scale				✓
Ordinal scale				✓
Quantitative scale	✓	✓	✓	?
Sampling stability	3	2	1	?
Mathematical tractability			✓	
Very skewed distribution		✓	?	?
Extreme score		✓	?	?
Truncated distribution		✓	?	?
Open-ended distribution		✓		?
Use all values to measure dispersion			✓	✓

- Detecting Outlier
  - First Criterion: $Mdn \pm 2(Q_3 - Q_1)$ → This criterion represents in boxplot.
  - In SPSS: Outlier = $Mdn \pm 2(Q_3 - Q_1)$; Extreme score = $Mdn \pm 3.5(Q_3 - Q_1)$
  - Second Criterion: $\bar{X} \pm 2.5S$

- Standardized Measures of Variability
  - Standardized term of range

$$\frac{Max - Min}{Max + Min}$$

  - Standardized term of interquartile range

$$\frac{Q_3 - Q_1}{Q_3 + Q_1}$$

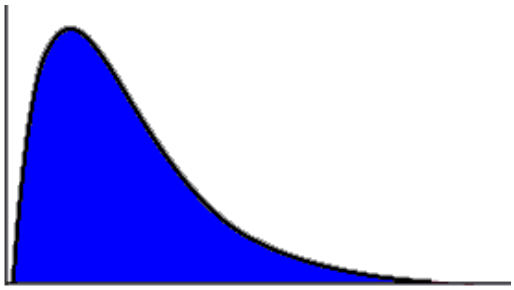
- Coefficient of Variation

$$CV = \frac{S_X}{\bar{X}}$$

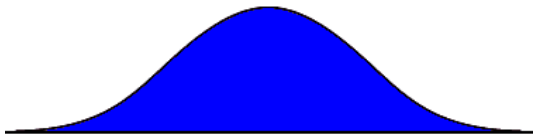
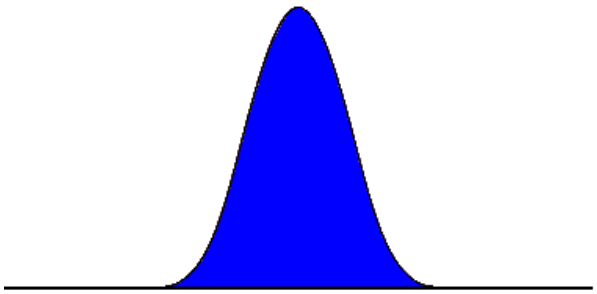
Measures of Nonnormality

- Skewness
  - Negative Skewness

  - Positive Skewness



- Kurtosis
  - Negative Kurtosis (Platykurtic)



  - Positive Kurtosis (Leptokurtic)