

Descriptive Statistics

Gökmen ZARARSIZ, Phd.

Erciyes University, Faculty of Medicine, Department of Biostatistics

gokmen.zararsiz@gmail.com

March 10, 2022

Copyright 2019 ©. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part.

Table of Contents

1 Introduction

2 Measures of Location

- The Arithmetic Mean
- The Geometric Mean
- The Harmonic Mean
- The Median
- The Mode

3 Measures of Spread

- The Range
- The Interquartile Range
- The Variance and Standard Deviation

4 Relationship Between The Mean and Standard Deviation

- Normal Distribution
- Coefficient of Variation and Coefficient of Dispersion

5 Measures of Shape

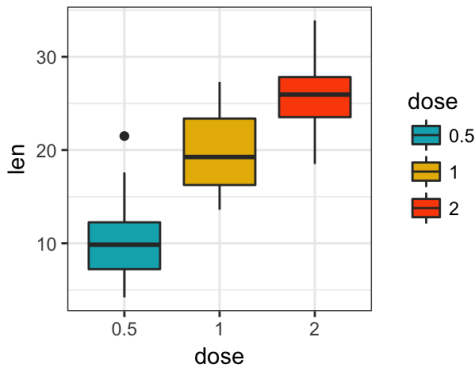
- Skewness
- Kurtosis

6 References

Descriptive Statistics

Definition

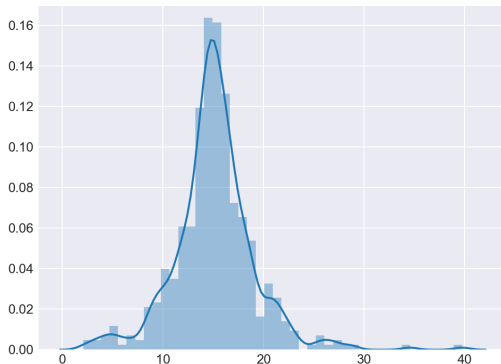
Descriptive statistics: Organizing, summarizing, and displaying data



Descriptive Statistics

Distribution of data

Distribution: The spread of the data about some central tendency value. A distribution is characterized by location, spread and shape.



Descriptive Statistics

Distribution of data

Measures of Location

Includes measures to describe the central tendency of the data.

Measures of Spread

Includes measures to describe the spread of the data.

Measures of Shape

Includes measures to describe how the variation is distributed about the location.

Descriptive Statistics

Example

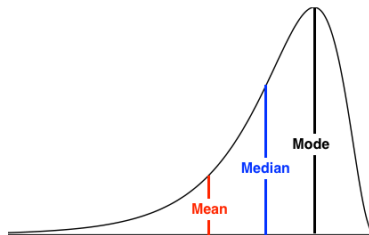
The estimated glomerular filtration rate (eGFR, $ml/min/1.73m^2$) levels of 20 patients with autosomal dominant polycystic kidney disease (ADPKD) at the time of diagnosis were measured and given below. Descriptive statistics, which will be mentioned throughout this chapter, will be calculated on this data.

85	85	70	60	35	55	75	60	65	90
75	65	75	95	70	75	55	80	65	85

Table: eGFR levels of ADPKD patients

Measures of Location

- The arithmetic mean
- The geometric mean
- The harmonic mean
- The median
- The mode



The Arithmetic Mean

Definition

The arithmetic mean (commonly used as mean): The sum of all individual observations divided by the number of observations. In statistics, arithmetic mean is the most widely used measure of location.

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

The Arithmetic Mean

Properties

- If same constant c is added to each data point, then the mean of the translated data will be $\bar{x} + c$

$$\frac{1}{n} \sum_{i=1}^n (x_i + c) = \bar{x} + c$$

- If same constant c is multiplied with each data point, then the mean of the rescaled data will be $c\bar{x}$

$$\frac{1}{n} \sum_{i=1}^n cx_i = c\bar{x}$$

- If same constant c_1 is multiplied with each data point, and the same constant c_2 is added to each data point, then the mean of the rescaled data will be $c_1\bar{x} + c_2$

$$\frac{1}{n} \sum_{i=1}^n (c_1x_i + c_2) = c_1\bar{x} + c_2$$

The Arithmetic Mean

Solution of the Example

$$\bar{x} = \frac{35 + 55 + \dots + 95}{20} = 71$$

The mean eGFR of the ADPKD patients is $71 \text{ ml/min/1.73m}^2$.

The Arithmetic Mean

Advantages and Disadvantages

Advantages

- Every data point is used during the calculations. Thus, this statistic is a good representative of the data.
- Samples repeatedly drawn from the same population tend to have similar arithmetic means.
- The arithmetic mean has a close relation with the most common dispersion statistic, i.e. standard deviation.

Disadvantages

- The arithmetic mean is sensitive to extreme values and outliers. It is not an appropriate measure of location for skewed distribution.

The Arithmetic Mean

Weighted Mean

Weighted mean: In some cases, certain data points may have more importance than the others. By calculating the weighted mean, these data points contribute more “weight” to the final mean.

$$\mu_w = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} \qquad \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

e.g. The mean Biostatistics score of each student may be calculated as follows:

$$\bar{x}_w = 0.15Mid_1 + 0.15Mid_2 + 0.20Assignments + 0.20Quiz + 0.30Final$$

The Geometric Mean

Definition

The geometric mean: The geometric mean statistic is the n^{th} root of the product of n data points (for population parameters, use N).

$$\mu_g = \sqrt[N]{\prod_{i=1}^N x_i} \qquad \bar{x}_g = \sqrt[n]{\prod_{i=1}^n x_i}$$

This statistic is usually used for data that are products or exponential in nature.

E.g. growth rates, concentrations of one substance in another, investment return, etc. $X = \{2, 6, 18, 54, 162, 486, 1458, 4374\}$

It is assumed that the logarithmic transformation of the data points have a symmetrical distribution, while the distribution of the untransformed data is skewed.

The Geometric Mean

Solution of the Example

$$\bar{x}_g = \sqrt[20]{35 \times 55 \times \dots \times 95} \approx 69.45$$

The geometric mean eGFR of the ADPKD patients is 69.45 *ml/min/1.73m²*.

The Harmonic Mean

Definition

The harmonic mean: Harmonic mean is defined as the value obtained when the number of observations in the data set is divided by the sum of the reciprocals of each data point.

$$\mu_h = \frac{N}{\sum_{i=1}^N \frac{1}{x_i}}$$

$$\bar{x}_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Harmonic mean is mostly appropriate when the average of rates is desired.

E.g. Speed, price-to-earning ratio, other rates and ratios.

The Harmonic Mean

Solution of the Example

$$\bar{x}_h = \frac{20}{\frac{1}{35} + \frac{1}{55} + \dots + \frac{1}{95}} \approx 67.62$$

The harmonic mean eGFR of the ADPKD patients is 67.62 *ml/min/1.73m²*.

The Median

Definition

The median: After ordering the observations from smallest to largest, the middle value of the data points is the median. The 50% of the data points take values less (or greater) than the median. The population and sample median are denoted by η and M .

- The median is the $(\frac{n+1}{2})$ th largest observation, if n is odd
- The median is the average of $(\frac{n}{2})$ th and $(\frac{n+2}{2})$ th largest observation, if n is even

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = $(4 + 5) \div 2$
= **4.5**

The Median

Solution of the Example

35	55	55	60	60	65	65	65	70	70		75	75	75	75	80	85	85	85	90	95
----	----	----	----	----	----	----	----	----	----	--	----	----	----	----	----	----	----	----	----	----

$$M = \frac{70 + 75}{2} = 72.5$$

The 50% of the ADPKD patients have eGFR values less than 72.5 *ml/min/1.73m²*.

The Median

Advantages and Disadvantages

Advantages

- Easy calculation, simple interpretation.
- Less affected by extreme values and outliers. Median is the preferred measure of location when the distribution is not symmetrical.

Disadvantages

- The median is less representative of the data, since it only uses the middle value of the data points.

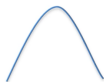
The Mode

Definition

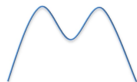
The mode: The mode is the most frequently occurred value of the data points.

- A distribution with only one mode is called as **unimodal**.
- A distribution with two modes is called as **bimodal**.
- A distribution with more than two modes is called as **multimodal**.

Unimodal



Bimodal



Multimodal



The Mode

Solution of the Example

35	55	55	60	60	65	65	65	70	70	75	75	75	75	80	85	85	85	85	90	95
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\text{Mode} = 75$$

The most frequently occurred eGFR value of the ADPKD patients is 75 $\text{ml}/\text{min}/1.73\text{m}^2$.

The Mode

Advantages and Disadvantages

Advantages

- Easy calculation, simple interpretation.
- Less affected by extreme values and outliers.

Disadvantages

- The mode may not always represent the center of the distribution.

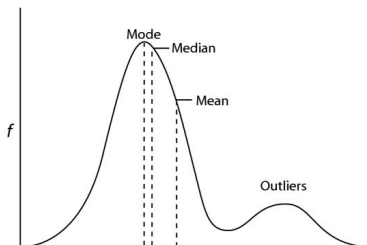
Outliers and Extreme Values

Relation with the measures of location

Outliers: Data point(s) which are substantially different or far from the rest of the data.

Extreme values: Outliers which show extreme deviation from the rest of the data.

In the presence of outliers and extreme values, the data distribution becomes skewed.



Outliers and Extreme Values

Relation with the measures of location

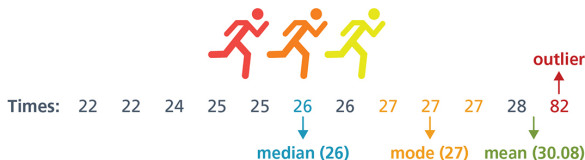
Symmetrical data

- The mean statistic should be used as a measure of location. Since, the mean statistic uses every data point during the calculations.

Skewed data

- The median statistic should be used as an alternative to the mean statistic. Since, the median statistic is less affected by the outliers.

E.g. 12 students who participated in the regional 200m under 15's running finals had their times recorded. There is an outlier amongst the figures. One student (Trevor) completed the race in 82 seconds, all other recorded times were less than 30 seconds.



Outliers and Extreme Values

Relation with the measures of location



Mean Salary

4.48*million*\$

Median Salary

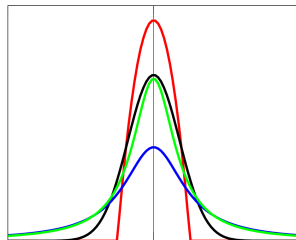
1.3*million*\$

Player	Salary (\$)
Jordan	30,140,000
Rodman	9,000,000
Kukoc	3,960,000
Harper	3,840,000
Longley	2,790,000
Pippen	2,250,000
Brown	1,300,000
Simpkins	1,040,000
Parish	1,000,000
Wennington	1,000,000
Kerr	750,000
Caffey	700,000
Buechler	500,000

Table: Salary distribution of Chicago-Bulls players in 1996-1997 season

Measures of Spread

- The range
- The interquartile range
- The variance and standard deviation



The Range

Definition

The range: The range is the distance between the maximum and the minimum of the data points.

$$\text{Range} = X_{\max} - X_{\min}$$

- Range represents the 100% of a data distribution.
- The range is sensitive to extreme values and outliers.



The Range

Solution of the Example

35	55	55	60	60	65	65	65	70	70	75	75	75	75	80	85	85	85	90	95
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\text{Range} = 95 - 35 = 60$$

The eGFR value of the ADPKD patients ranges between 35 and 95 $\text{ml}/\text{min}/1.73\text{m}^2$.

Quantiles

Definition

Quantiles: Cut points which divides the range of a probability distribution into q continuous intervals with equal probabilities.

- There are $q - 1$ of the q quantiles, one for each integer k satisfying $0 < k < p$.
- The data points should be sorted before calculating the desired quantiles.
- $k^{th} \text{ quantile} = \frac{k(n+1)}{q} \text{th data point}$

q	Quantile
2	Median
3	Tertiles
4	Quartiles
5	Quintiles
10	Deciles
100	Percentiles

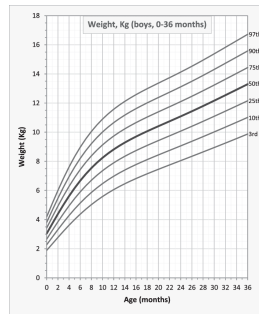
Percentile and Quartile

Definition

Percentile: A type of quantile indicating the value below which a given percentage of observations in a group of observations falls. E.g. the 95th percentile is the value below which 95% of the observations may be found.

Quartile: A type of quantile which divides the range of a probability distribution into 4 continuous intervals with equal probabilities.

- $k^{th} \text{ percentile} = \frac{k(n+1)}{100} \text{th data point}$
- The middle value of the sorted data is $\text{median} = 2^{nd} \text{ quartile} = 50^{th} \text{ percentile}$
- $1^{st} \text{ quartile}(Q_1) = 25^{th} \text{ percentile}$
- $3^{rd} \text{ quartile}(Q_3) = 75^{th} \text{ percentile}$



The Interquartile Range

Definition

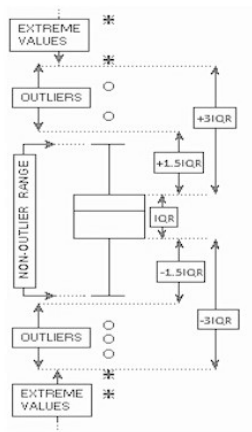
The Interquartile range: Interquartile range is the distance between Q_3 and Q_1 .

$$IQR = Q_3 - Q_1$$

- IQR represents the 50% central portion of a data distribution.
- IQR is less affected by extreme values and outliers.

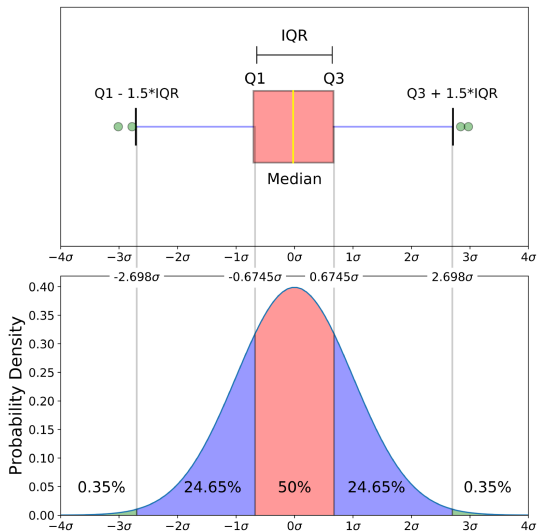
Tukey's fences: $[Q_1 - c \times IQR, Q_3 + c \times IQR]$

- Any observation outside these ranges can be defined as;
outliers ($c = 1.5$) or **extreme values** ($c = 3$).



The Interquartile Range

Definition



The Interquartile Range

Solution of the Example

35 55 55 60 60 65 65 70 70 75 75 75 75 80 85 85 85 90 95

$$k^{th} \text{ percentile} = \frac{k(n+1)}{100} \text{th data point}$$

$$Q_1 = 25^{th} \text{ percentile} = \frac{25(20+1)}{100} = 5.25 \text{th data point} = 61.25$$

$$Q_3 = 75^{th} \text{ percentile} = \frac{75(20+1)}{100} = 15.75 \text{th data point} = 83.75$$

The Interquartile Range

Solution of the Example

35	55	55	60	60	65	65	65	70	70	75	75	75	75	80	85	85	85	90	95
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$IQR = 83.75 - 61.25 = 22.5$$

The eGFR value of the middle 50% of the ADPKD patients ranges between 61.25 and 83.75 $ml/min/1.73m^2$.

Outlier limits: $[61.25 - 1.5 \times 22.5, 83.75 + 1.5 \times 22.5] = [27.5, 117.5]$

Extreme v. limits: $[61.25 - 3 \times 22.5, 83.75 + 3 \times 22.5] = [-6.25, 151.25]$

No outliers or extreme values are present in the data.

The Variance and Standard Deviation

Definition

The variance: A measure of how far each data point are spread out from the mean. The population and sample variances are denoted by σ^2 and s^2 .

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

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

The standard deviation: A measure of the average distance to the mean. Standard deviation is square root of variance. The population and sample standard deviations are denoted by σ and s .

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

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

The Variance and Standard Deviation

Properties

- If same constant c is added to each data point, then the variance of the translated data (x_c) will remain the same.

$$\frac{\sum_{i=1}^n (x_i + c - \bar{x}_c)^2}{n-1} = s^2, \quad \bar{x}_c = E[x_i + c] = \bar{x} + c$$

- If same constant c is multiplied with each data point, then the variance of the rescaled data (x_c) will be

$$\frac{\sum_{i=1}^n (cx_i - \bar{x}_c)^2}{n-1} = c^2 s^2, \quad \bar{x}_c = E[cx_i] = c\bar{x}$$

The Variance and Standard Deviation

Solution of the Example

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
35	-36	1,296
55	-16	256
55	-16	256
60	-11	121
60	-11	121
65	-6	36
65	-6	36
65	-6	36
70	-1	1
70	-1	1
75	4	16
75	4	16
75	4	16
75	4	16
80	9	81
85	14	196
85	14	196
85	14	196
90	19	361
95	24	576
1,420	0	3,830

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{3,830}{19} \approx 201.58$$

$$s = \sqrt{s^2} = \sqrt{201.58} \approx 14.2$$

The average distance to mean eGFR value of the ADPKD patients is $14.2 \text{ ml/min/1.73m}^2$.

The Variance and Standard Deviation

Advantages and Disadvantages

Advantages

- Every data point is used during the calculations. Thus, this statistic is a good representative of the data.
- The standard deviation has the same unit and has a close relation with the most common measure of location, i.e. arithmetic mean.

Disadvantages

- The standard deviation is sensitive to extreme values and outliers. It is not an appropriate measure of location for skewed distribution.

Outliers and Extreme Values

Relation with the measures of spread

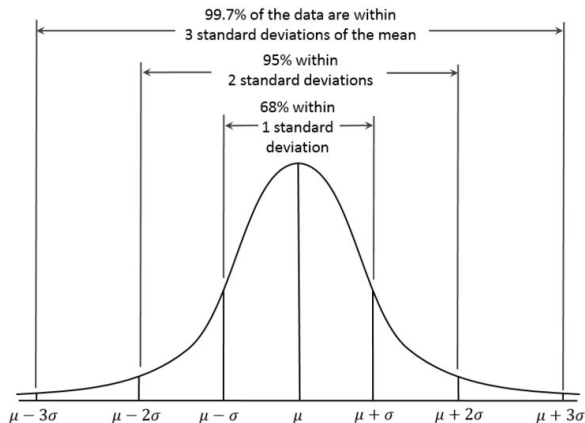
Symmetrical data

- The standard deviation statistic should be used with the mean. Since, both statistics use every data point during the calculations.

Skewed data

- The interquartile range should be used with the median statistic. Since, both statistics are less affected by the outliers. Alternatively, the range may be used with the median statistic; when the sample size of the data is low.

Normal Distribution



Coefficient of Variation and Coefficient of Dispersion

Definition

Coefficient of variation: The ratio of the standard deviation to the absolute value of the mean. Coefficient of variation is a suitable measure to compare the variations of variables with different units.

$$\text{Population CV} = \frac{\sigma}{\mu} \times 100 \qquad \text{Sample CV} = \frac{s}{\bar{x}} \times 100$$

Coefficient of dispersion: The ratio of the variance to the absolute value of the mean. This measure is related to the Poisson distribution. A Poisson random variable, which includes discrete non-negative counts, has a coefficient of dispersion equal to 1.

$$\phi = \frac{\sigma^2}{\mu}$$

$$f = \frac{s^2}{\bar{x}}$$

Coefficient of Variation and Coefficient of Dispersion

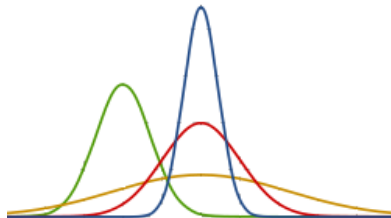
Solution of the Example

$$CV = \frac{s}{\bar{x}} \times 100 = \frac{14.2}{71} \times 100 = 20\%$$

$$f = \frac{s^2}{\bar{x}} = \frac{201.58}{71} \approx 2.84$$

Measures of Shape

- Skewness
- Kurtosis



Skewness

Definition

Skewness: A measure of asymmetry in the data distribution. Using skewness statistic, we can interpret whether the data is positively or negatively skewed relative to the normal distribution.

$$\nu = \left[\frac{N}{(N-1)(N-2)} \right] \frac{\sum_{i=1}^N (x_i - \mu)^3}{\sigma^3}$$

$$g_1 = \left[\frac{n}{(n-1)(n-2)} \right] \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

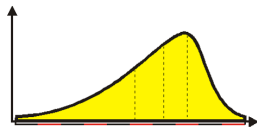
- Skewness of the normal distribution is zero.
- If skewness statistic is greater than zero, then the data distribution is positively skewed (right-skewed).
- If skewness statistic is less than zero, then the data distribution is negatively skewed (left-skewed).

Skewness

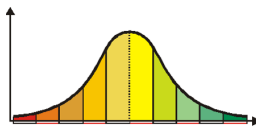
General Forms of Skewness

The mean, median and mode can be used to figure out the skewness of the data distribution.

Negatively skewed
($\nu < 0$)



Symmetrical
($\nu = 0$)



Positively skewed
($\nu > 0$)



Mean < Median < Mode Mean = Median = Mode Mode < Median < Mean

Skewness

Solution of the Example

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^3$
35	-36	-46,656
55	-16	-4,096
55	-16	-4,096
60	-11	-1,331
60	-11	-1,331
65	-6	-216
65	-6	-216
65	-6	-216
70	-1	-1
70	-1	-1
75	4	64
75	4	64
75	4	64
75	4	64
80	9	729
85	14	2,744
85	14	2,744
85	14	2,744
90	19	6,859
95	24	13,824
1,420	0	-28,260

$$g_1 = \left[\frac{n}{(n-1)(n-2)} \right] \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

$$g_1 = \left[\frac{20}{(19)(18)} \right] \frac{-28260}{14.20^3}$$

$$g_1 = -0.577$$

The eGFR variable of the ADPKD patients has a slightly skewed distribution to the left.

Kurtosis

Definition

Kurtosis: A measure of tailedness in the data distribution. Using kurtosis statistic, we can interpret whether the data is heavy-tailed or light-tailed relative to the normal distribution.

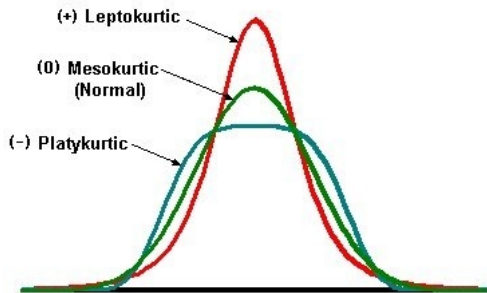
$$\tau = \left[\frac{N(N+1)}{(N-1)(N-2)(N-3)} \right] \frac{\sum_{i=1}^N (x_i - \mu)^4}{\sigma^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

$$g_2 = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} \right] \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Kurtosis

General Forms of Kurtosis

- Kurtosis of the standard normal distribution is zero (mesokurtic distribution).
- If kurtosis statistic is greater than zero, then the data distribution is leptokurtic (heavy-tailed).
- If kurtosis statistic is less than zero, then the data distribution is platykurtic (light-tailed).



Kurtosis

Solution of the Example

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^4$
35	-36	1,679,616
55	-16	65,536
55	-16	65,536
60	-11	14,641
60	-11	14,641
65	-6	1,296
65	-6	1,296
65	-6	1,296
70	-1	1
70	-1	1
75	4	256
75	4	256
75	4	256
75	4	256
80	9	6,561
85	14	38,416
85	14	38,416
85	14	38,416
90	19	130,321
95	24	331,776
1,420	0	2,428,790



$$g_2 = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} \right] \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$g_2 = \left[\frac{20(21)}{(19)(18)(17)} \right] \frac{2428790}{14.20^4} - \frac{3(19)^2}{(18)(17)}$$

$$g_2 = 0.779$$

The eGFR distribution of the ADPKD patients is slightly heavy-tailed.

References I

-  Alpar, R. [2016]. *Uygulamalı İstatistik ve Geçerlik Güvenirlik*, 4th ed. Detay Yayıncılık, Ankara
-  Alpar, R. [2017]. *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*, 5th ed. Detay Yayıncılık, Ankara
-  Chernick, M. R., and Friis, R. H. [2003]. *Introductory Biostatistics for the Health Sciences: Modern Applications Including Bootstrap*, 1st ed. Wiley Interscience, New Jersey
-  Crawley, M.J. [2004]. *The R Book*, 1st ed. Wiley, England
-  Elston, R. C., and Johnson, W. D. [2008]. *Basic Biostatistics for Geneticists and Epidemiologists: A Practical Approach*, 1st ed. Wiley, UK
-  Fisher, R. A. [1950]. *Statistical Methods for Research Workers*, 11th ed. Hafner, New York.
-  Fisher, L. D., van Belle, G., Heagerty, P. J., Lumley, T. [2004]. *Biostatistics: A Methodology for the Health Sciences*, 2nd ed. Wiley Interscience, New Jersey

References II



Forthofer, R. N., Lee, E. S., Hernandez, M. [2007]. *Biostatistics: A Guide to Design, Analysis, and Discovery*, 2nd ed. Elsevier, London



Kendall, M. G., and Stuart, A. [1963]. *The Advanced Theory of Statistics*, Vol. 1, 2nd ed. Charles Griffin, London.



Kruskal, W. [1968]. *In International Encyclopedia of the Social Sciences*, D. L. Sills (ed). Macmillan, New York.



Logan, M. [2010]. *Biostatistical Design and Analysis Using R: A Practical Guide*, 1st ed. Wiley Blackwell, UK



Mainland, D. [1963]. *Elementary Medical Statistics*, 2nd ed. Saunders, Philadelphia.



Mood, A. M. [1950]. *Introduction to the Theory of Statistics*. McGraw-Hill, New York.



Rosner, B. [2011]. *Fundamentals of Biostatistics*, 7th ed. Brooks/Cole, Cengage Learning, Boston

References III



Savage, I. R. [1968]. *Statistics: Uncertainty and Behavior*. Houghton Mifflin, Boston.



Sokal, R. R., and Rohlf, F. J. [1969]. *Introduction to Biostatistics*. W. H. Freeman and Co., Ltd., New York.



von Mises, R. [1957]. *Probability, Statistics and Truth*, 2nd ed. Macmillan, New York.



Wassertheil Smoller S. [2004]. *Biostatistics and Epidemiology: A Primer for Health and Biomedical Professionals*, 3rd ed. Springer-Verlag, New York



Wayne W. D. [2005]. *Biostatistics: A Foundation for Analysis in the Health Sciences*, Ninth Edition. Wiley, New York.