basic concepts and definitions

contents

- introduction
- variable and variable types
- frequency distributions
- sample and population
- descriptive sample statistics and parameters
- data encoding for statistical calculations
- methods for calculating mean and std. dev.
- coefficient of variance

introduction

- statistics studies on **methods** for gathering, analysis and interpretation of data

introduction

- statistics studies on **methods** for gathering, analysis and interpretation of data
- statistics is required for **objective** decision making

introduction

- statistics studies on **methods** for gathering, analysis and interpretation of data
- statistics is required for **objective** decision making

Popul Bull. 1982 Mar;37(1):1-55.

The World Fertility Survey: charting global childbearing.

Lightbourne R Jr, Singh S, Green CP.

Abstract

Interviewing some 350,000 women in 42 developing countries and 20 developed countries representing nearly 40% of the world's population, the World Fertility Survey (WFS) is in a unique position to document the historic 1970s slowdown in global population growth. This Bulletin describes efforts begun in 1972 to ensure high quality, internationally comparable, accessible data, the data's importance for policymakers, planners and researchers, and major findings available by early 1982 from directly assisted WFS surveys in 29 developing countries and contraceptive use data from WFS-type surveys in 16 developed countries. Marital fertility has declined in all developing regions except Africa but still averages from 4.6 children/woman in Latin America to 6.7 in Africa, while preferred family size ranges from 3.0 children in Turkey to 8.9 in Senegal--far above the average 2.2-2.5 children/woman needed to end developing countries' population growth in the long run. However, women ages 15-19 prefer nearly 2 children fewer than the oldest women ages 45-49; 3.8 vs. 5.7 on the average. Nearly 1/2 (48%) of married women surveyed in 27 countreis said they wanted no more children. Preventing all unwanted births would reduce birth rates up to 15 births/1000 population in these countries. Overall, 32% of married, fecund women in developing countries are using contraception compared to an average 72% in 16 developed countries. Education, literacy, and more available family planning services increase contraceptive use. Age at marriage is rising in Asia, but this factor alone has little effect on fertility. Infant mortality is higher in many developing countries than previously thought. Breastfeeding is an important restraint on fertility in most developing countries but is declining among more educated, employed, and urban women which could raise fertility if not compensated for by gains in contraceptive use.

variable *(değişken)*

observation of a **measurement *(ölçüm)*** (weight), a **count *(sayım)*** (number of customers) or a **categorization *(sınıflandırma)*** (sex)

variable *(değişken)*

observation of a **measurement *(ölçüm)*** (weight), a **count *(sayım)*** (number of customers) or a **categorization *(sınıflandırma)*** (sex)

if the observation results could be depicted by probabilistic functions, those are named as **random variables *(şans değişkenleri)***

variable *(değişken)*

observation of a **measurement *(ölçüm)*** (weight), a **count *(sayım)*** (number of customers) or a **categorization *(sınıflandırma)*** (sex)

if the observation results could be depicted by probabilistic functions, those are named as **random variables *(şans değişkenleri)***

the events that you can expect different outcomes even on identical conditions (e.g. blood sample)

variable *(değişken)*

observation of a **measurement *(ölçüm)*** (weight), a **count *(sayım)*** (number of customers) or a **categorization *(sınıflandırma)*** (sex)

if the observation results could be depicted by probabilistic functions, those are named as **random variables *(şans değişkenleri)***

the events that you can expect different outcomes even on identical conditions (e.g. blood sample)

if there had been no difference on observations, there would be no need for statistics

variable *(değişken)*

observation of a **measurement *(ölçüm)*** (weight), a **count *(sayım)*** (number of customers) or a **categorization *(sınıflandırma)*** (sex)

if the observation results could be depicted by probabilistic functions, those are named as **random variables *(şans değişkenleri)***

the events that you can expect different outcomes even on identical conditions (e.g. blood sample)

if there had been no difference on observations, there would be no need for statistics

example: (**Y**: weight of each block - variable)

$\mathbf{y_5}$ = 200 kg (result of 5th observation)

variable types (nature based)

- **continuous variables** (e.g. weight, sensitivity of measurement transforms it into a discrete variable ~ kg, g, mg)

variable types (nature based)

- **continuous variables** (e.g. weight, sensitivity of measurement transforms it into a discrete variable ~ kg, g, mg)
- **discrete variables** (cannot take any value between two integers ~ #of rotten apples)

variable types (nature based)

- **continuous variables** (e.g. weight, sensitivity of measurement transforms it into a discrete variable ~ kg, g, mg)
- **discrete variables** (cannot take any value between two integers ~ #of rotten apples)
- **categorical variables** (

  you cannot measure or count them ~ degree of pain;

  transforming the continuous variable ~ thin, obese;

  transforming the discrete variable sets ~ top quality olive, fair quality olive

)

variable types (scale based)

- **nominal scale** (qualitative, separate categories, corresponds to categorical variables, e.g. condition of a product)

variable types (scale based)

- **nominal scale** (qualitative, separate categories, corresponds to categorical variables, e.g. condition of a product)
- **ordinal scale** (+ degree of quality is important on categories, logical ordering, e.g. degree of pain)

variable types (scale based)

- **nominal scale** (qualitative, separate categories, corresponds to categorical variables, e.g. condition of a product)
- **ordinal scale** (+ degree of quality is important on categories, logical ordering, e.g. degree of pain)
- **interval scale** (+ information about how far data pairs are from each other, quantitative, no real zero reference point, can transform between interval scales however no idea about ratio between data pairs – only distance between them e.g. IQ, temperature)

variable types (scale based)

- **nominal scale** (qualitative, separate categories, corresponds to categorical variables, e.g. condition of a product)
- **ordinal scale** (+ degree of quality is important on categories, logical ordering, e.g. degree of pain)
- **interval scale** (+ information about how far data pairs are from each other, quantitative, no real zero reference point, can transform between interval scales however no idea about ratio between data pairs – only distance between them e.g. IQ, temperature)
- **ratio scale** (+ Stevensen's scale, absolute zero – starting point, e.g. weight, # of women)

variable types (scale based)

- **nominal scale** (qualitative, separate categories, corresponds to categorical variables, e.g. condition of a product)
- **ordinal scale** (+ degree of quality is important on categories, logical ordering, e.g. degree of pain)
- **interval scale** (+ information about how far data pairs are from each other, quantitative, no real zero reference point, can transform between interval scales however no idea about ratio between data pairs – only distance between them e.g. IQ, temperature)
- **ratio scale** (+ Stevensen's scale, absolute zero – starting point, e.g. weight, # of women)

transformation between scales (loss of information)

frequency distributions

- organizing and summarizing gathered data

frequency distributions

- organizing and summarizing gathered data
- works well for discrete and categorical data

frequency distributions

- organizing and summarizing gathered data
- works well for discrete and categorical data
- relatively harder for continuous data

continuous dataset

Tablo 1.3. 156 hastada kolestrol düşürücü ilaç alımından sonra gözlenen serum kolesterol değişiklikleri

17	-12	25	-37	-29	-39
-22	0	-22	-63	34	-31
-64	-2	-49	5	-8	33
-50	-7	16	-11	-38	-17
0	-9	-21	1	2	-30
-32	-34	-14	-18	5	6
24	-6	-49	-8	-49	-37
-25	-12	14	10	-41	-66
-31	35	21	-19	-27	17
-6	-17	-6	1	-28	40
-31	17	-54	-27	-16	16
-44	10	-3	-3	5	6
-19	9	-10	-20	-9	-8
-10	-11	11	-39	19	-32
4	-15	-18	35	6	20
46	24	-27	-19	5	-60
27	23	-22	-1	12	-27
-13	-39	39	-34	-97	-26
38	14	-47	8	16	-15
-62	12	-53	11	21	-47
-54	-11	-5	0	55	34
-69	-11	-44	20	-50	19
0	-25	-24	-4	14	2
-34	16	-23	-71	-58	9
9	2	-2	-58	13	14
17	-13	-22	-3	-17	1

frequency distributions

- organizing and summarizing gathered data
- works well for discrete and categorical data
- relatively harder for continuous data
  - comprehension (7,8 vs 10,20)
  - cover whole range of data
  - avoid sparse categories *(Bolton method, %50 of categories > %10 of whole data)*
  - equal range on each category (for easing future calculations)

frequency table

Tablo 1.4. Tablo 1.3'de yer alan verilerin 16 sınıflı bir Frekans Dağılışı halinde gösterir

Sınıf Aralığı	Tarama	Frekans
-100 ile -91	/	1
-90 ile -81		0
-80 ile -71	/	1
-70 ile -61	/////	5
-60 ile -51	///// /	6
-50 ile -41	///// /////	10
-40 ile -31	///// ///// ///	14
-30 ile -21	///// ///// ///// //	17
-20 ile -11	///// ///// ///// ///// //	22
-10 ile -1	///// ///// ///// ///	18
0 ile +9	///// ///// ///// ///// //	22
+10 ile +19	///// ///// ///// ///// /	21
+20 lie +29	///// ////	9
+30 ile +39	///// //	7
+40 ile +49	//	2
+50 ile +59	/	1
Toplam		156

stem and leaf table

- Tukey (1974), reason: loss of original value

Tablo 1.5. Tablo 1.3'de yer alan verilerin gövde-ve-yaprak Frekans dağılışı

de	Yapraklar
-9	7
-8	
-7	1
-6	4 2 9 3 6 0
-5	0 4 4 3 8 0 8
-4	4 9 9 7 4 1 9 7
-3	2 1 1 4 4 9 7 9 4 8 9 1 0 7 2
-2	2 5 5 2 1 7 2 4 3 2 7 0 9 7 8 6 7
-1	9 0 3 2 2 2 7 1 5 1 1 3 4 0 8 1 8 9 9 6 7 7 5
-0	6 7 9 6 6 3 5 2 8 3 1 4 3 8 9 8
+0	0 4 0 9 0 9 2 5 1 1 8 0 2 5 5 6 5 6 6 2 9 1
+1	7 7 0 7 4 2 6 6 4 1 0 1 9 2 6 4 3 7 6 9 4
+2	4 7 4 3 5 1 0 1 0
+3	8 9 9 5 4 3 4
+4	6 0
+5	5

stem and leaf table

- Tukey (1974), reason: loss of original value

Tablo 1.5. Tablo 1.3'de yer alan verilerin gövde-ve-yaprak Frekans dağılışı

de	Yapraklar
-9	7
-8	
-7	1
-6	4 2 9 3 6 0
-5	0 4 4 3 8 0 8
-4	4 9 9 7 4 1 9 7
-3	2 1 1 4 4 9 7 9 4 8 9 1 0 7 2
-2	2 5 5 2 1 7 2 4 3 2 7 0 9 7 8 6 7
-1	9 0 3 2 2 2 7 1 5 1 1 3 4 0 8 1 8 9 9 6 7 7 5
-0	6 7 9 6 6 3 5 2 8 3 1 4 3 8 9 8
+0	0 4 0 9 0 9 2 5 1 1 8 0 2 5 5 6 5 6 6 2 9 1
+1	7 7 0 7 4 2 6 6 4 1 0 1 9 2 6 4 3 7 6 9 4
+2	4 7 4 3 5 1 0 1 0
+3	8 9 9 5 4 3 4
+4	6 0
+5	5

sensitivity of leaves

cumulative frequency distribution

Tablo 1.6. Kolesterol verileri için hazırlanan Tablo 1.4'deki frekans dağılışından, eklemeli ve oransal frekans dağılışlarının bulunması

Sınıf Aralığı	Sınıf Orta Değeri	Frekans	Oransal Frekans	Eklemeli Frekans	Eklemeli Oransal Frekans
-100 ile -91	-95.5	1	0.64	1	0.64
-90 ile -81	-85.5	0	0.00	1	0.64
-80 ile -71	-75.5	1	0.64	2	1.28
-70 ile -61	-65.5	5	3.21	7	4.49
-60 ile -51	-55.5	6	3.85	13	8.34
-50 ile -41	-45.5	10	6.41	23	14.75
-40 ile -31	-35.5	14	8.97	37	23.72
-30 ile -21	-25.5	17	10.89	54	34.61
-20 ile -11	-15.5	22	14.10	76	48.71
-10 ile -1	-5.5	18	11.54	94	60.25
0 ile +9	4.5	22	14.10	116	74.35
+10 ile +19	14.5	21	13.46	137	87.81
+20 ile +29	24.5	9	5.77	146	93.58
+30 ile +39	34.5	7	4.50	153	98.08
+40 ile +49	44.5	2	1.28	155	99.36
+50 ile +59	54.5	1	0.64	156	100.00
Toplam		156	100.00		

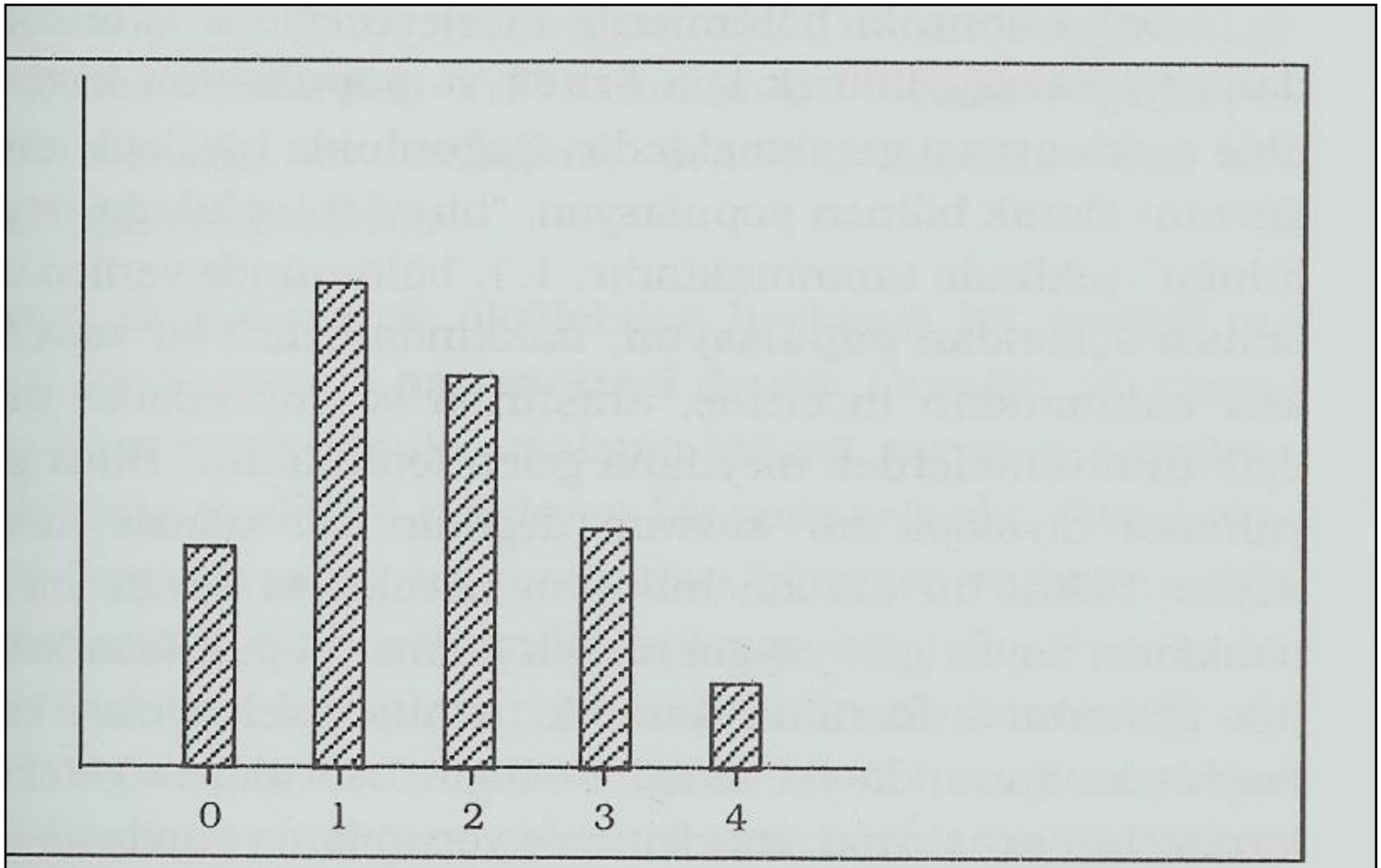
ile frekans tablosu haline getirilir ve oransal frekanslar dikkate alınarak

- comparing A and B experiments with different samples

graphical representations

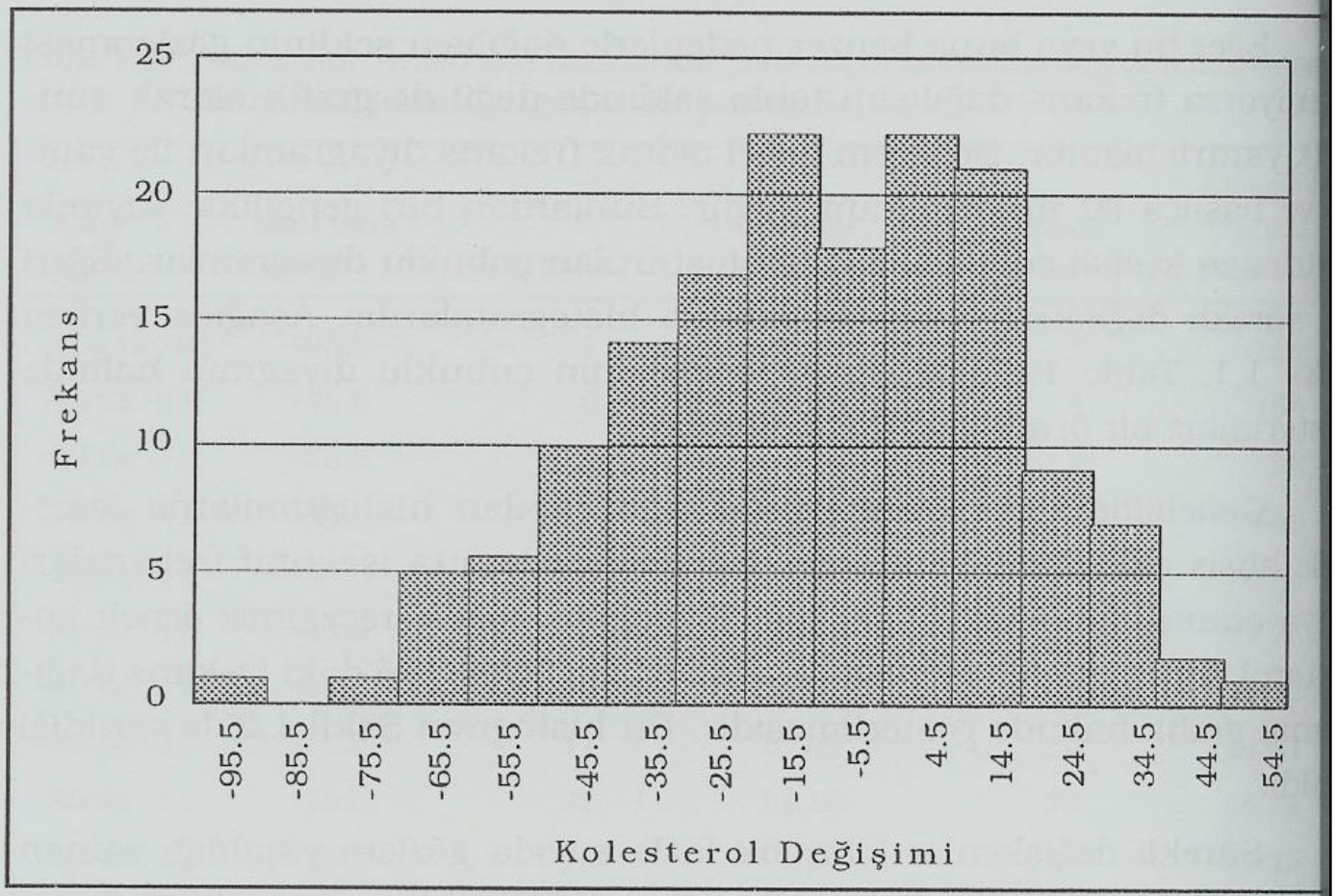
- bar graph for discrete variables
- histogram for continuous variables (you can combine upper edges for achieving envelope of frequency distribution)

bar graph



.1'deki verilen çürük diş sayılarına ilişkin verilerin çubuklu diyagram
gösterimi

histogram



Şekil 1.2. Tablo 1.6'daki verilerin kolesterol değişimine ilişkin verilerin histogram halinc gösterilmesi

1.6

sample and population

- population: a group of units requiring inspection, research and interpretation based on one or more variables (broad cover with respect to biological definition)

sample and population

- population: a group of units requiring inspection, research and interpretation based on one or more variables (broad cover with respect to biological definition)

- **two** main aspects of population are important: what is the **population**? and what are the **observation variable(s)** (e.g. 20 y.o. male athletes + leukocyte, screws manufactured in a particular factory + length)

sample and population

- population: a group of units requiring inspection, research and interpretation based on one or more variables (broad cover with respect to biological definition)

- **two** main aspects of population are important: what is the **population**? and what are the **observation variable(s)** (e.g. 20 y.o. male athletes + leukocyte, screws manufactured in a particular factory + length)

- two types of populations with respect to size of group: finite (real) and infinite (hypothetic)

sample and population

- sample -> troubles in considering all units -- sampling theory (must be *GOOD*)

sample and population

- sample -> troubles in considering all units -- sampling theory (must be *GOOD*)

- parameter: defines a quantitative property of a population (can be measured numerically) (e.g. 20 y.o. male athletes + leukocyte mean defines that population's leukocyte mean)

sample and population

- sample -> troubles in considering all units -- sampling theory (must be *GOOD*)

- parameter: defines a quantitative property of a population (can be measured numerically) (e.g. 20 y.o. male athletes + leukocyte mean defines that population's leukocyte mean)

- parameter's real value is **unknown** and is predicted by researcher with respect to samples (sample statistics) (must be *GOOD*)

sample and population

- sample -> troubles in considering all units -- sampling theory (must be *GOOD*)

- parameter: defines a quantitative property of a population (can be measured numerically) (e.g. 20 y.o. male athletes + leukocyte mean defines that population's leukocyte mean)

- parameter's real value is **unknown** and is predicted by researcher with respect to samples (sample statistics) (must be *GOOD*)

- parameters are depicted by Greek letters while sample statistics are depicted by Latin letters

sample and population

- parameter is a **constant** value (REAL – from all units)

sample and population

- parameter is a **constant** value (REAL – from all units)
- sample mean can **vary** from one from sample set to another

sample and population

- parameter is a **constant** value (REAL – from all units)
- sample mean can **vary** from one from sample set to another
- statatistics is crucial for defining population properties by predicting population parameter by means of calculating sample statistics

descriptive sample statistics

- summarizing observation data by means of scientific methods (important)

descriptive sample statistics

- summarizing observation data by means of scientific methods (important)
  - table-based summarization
  - graph-based summarization
  - numerical summarization (namely descriptive sample statistics) (both on x-axis a.k.a. *abscissa*)
    - location-based measures
    - difference-based measures

location measures


1.8 **1.12**

- mean
  - *arithmetic*
  - *other means (converting data by logarithm and reciprocal -> and then converting back)*
    - *geometric, harmonic*
  - *weighted*
- median
  - also quartile, decile and percentile
- mod

difference measures

- range (rough idea on difference in data)
- standard deviation (range based on only two data, trying to get all data in consideration)

difference measures

Tablo 1.7. 25 birey üzerindeki kol uzunluğu gözlemleri (cm)

34	36	43	35	40
33	43	39	46	38
39	44	38	47	36
41	44	45	36	38
44	41	36	42	39

difference measures

Tablo 1.8. Tablo 1.7'deki kol uzunluğu verilerinin 5 sınıflı bir frekans tablosunda özetlenmesi

Sınıflar	Sınıf Orta Değeri (x_i)	Frekans (f_i)
32.5-35.5	34	2
35.5-38.5	37	8
38.5-41.5	40	6
41.5-44.5	43	6
44.5-47.5	46	3

$\Sigma f_i = 25$

Verilerin dağılış merkezinden uzaklıklarının nasıl hesaplandığı ise

difference measures

Tablo 1.9. Kol uzunluğu verilerinde ortalamadan ayrılışlar

(1) x_i	(2) f_i	(3) $x_i f_i$	(4) $x_i - \bar{x}$	(5) $f_i (x_i - \bar{x})$
34	2	68	-6	-12
37	8	296	-3	-24
40	6	240	0	0
43	6	258	3	18
46	3	138	6	18
Toplam	25	1000		0

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1000}{25} = 40 \text{ cm}$$

variance and standard deviation

Tablo 1.10. Kol uzunluğu verileri için standart sapmanın bulunması

(1) x_i	(2) f_i	(3) $x_i - \bar{x}$	(4) $(x_i - \bar{x})^2$	(5) $f_i (x_i - \bar{x})^2$
34	2	-6	36	72
37	8	-3	9	72
40	6	0	0	0
43	6	3	9	54
46	3	6	36	108
Toplam	25			306

$\bar{x} = 40$

$\sum f_i (x_i - \bar{x})^2 = 306$

Varyans $= \frac{306}{25} = 12.24$, Standart Sapma = 3.499

Bu tablonun (1), (2) ve (3) nolu sütunları daha önceki tabloda yer

descriptive sample statistics and parameters

- GOOD sample -> GOOD prediction

descriptive sample statistics and parameters

- GOOD sample -> GOOD prediction
- prediction of population parameters by sample statistics

descriptive sample statistics and parameters

- GOOD sample -> GOOD prediction
- prediction of population parameters by sample statistics
- biased and unbiased prediction

descriptive sample statistics and parameters

Tablo 1.10. Kol uzunluğu verileri için standart sapmanın bulunması

(1) x_i	(2) f_i	(3) $x_i - \bar{x}$	(4) $(x_i - \bar{x})^2$	(5) $f_i (x_i - \bar{x})^2$
34	2	-6	36	72
37	8	-3	9	72
40	6	0	0	0
43	6	3	9	54
46	3	6	36	108
Toplam	25			306

$\bar{x} = 40$

$\sum f_i (x_i - \bar{x})^2 = 306$

Varyans $= \frac{306}{25} = 12.24$, Standart Sapma = 3.499

Bu tablonun (1), (2) ve (3) nolu sütunları daha önceki tabloda yer

data encoding for statistical calculations

- by addition, subtraction, multiplication, division

data encoding for statistical calculations

- by addition, subtraction, multiplication, division

effect on arithmetic mean

addition or subtraction

multiplication or division

mixed

data encoding for statistical calculations

- by addition, subtraction, multiplication, division

effect on arithmetic mean

addition or subtraction

multiplication or division

mixed

effect on variance and standard deviation

addition or subtraction

multiplication or division

mixed

practical methods for the calculation of mean and standard deviation

- correction term

Tablo 1.11. 9465 kişi üzerinde gözlenen kalp atım hızı değerlerinin frekans dağılışı

Sınıf Orta Değeri (x)	Frekans (f)	Kodlanmış Sınıf Orta Değeri (x_c)
59.5	2	0
67.5	6	1
75.5	39	2
83.5	385	3
91.5	888	4
99.5	1729	5
107.5	2240	6
115.5	2007	7
123.5	1233	8
131.5	641	9
139.5	201	10
147.5	74	11
155.5	14	12
163.5	5	13
171.5	1	14
	n = 9465	

X değişkeni $x_c = \frac{x - 59.5}{8}$ şeklinde kodlanmıştır.

Kodlanmış sınıf orta değerlerinin frekans sütunu ile karşılıklı

descriptive sample statistics and parameters

- correction term
- finger calculation

  midrange mean

  standard dev approx.

Örnek Büyüklüğü	Değişim Aralığının Bölüneceği Katsayı
10	3
30	4
100	5
500	6
1000	6.5

Tablo 1.7. 25 birey üzerindeki kol uzunluğu gözlemleri (cm)

34	36	43	35	40
33	43	39	46	38
39	44	38	47	36
41	44	45	36	38
44	41	36	42	39

coefficient of variance

- comparison of parameters for populations A and B (e.g. Tails of elephants vs. mouses, sugar level on blood and urine samples etc.
- be careful about irrelevant comparisons!

coefficient of variance

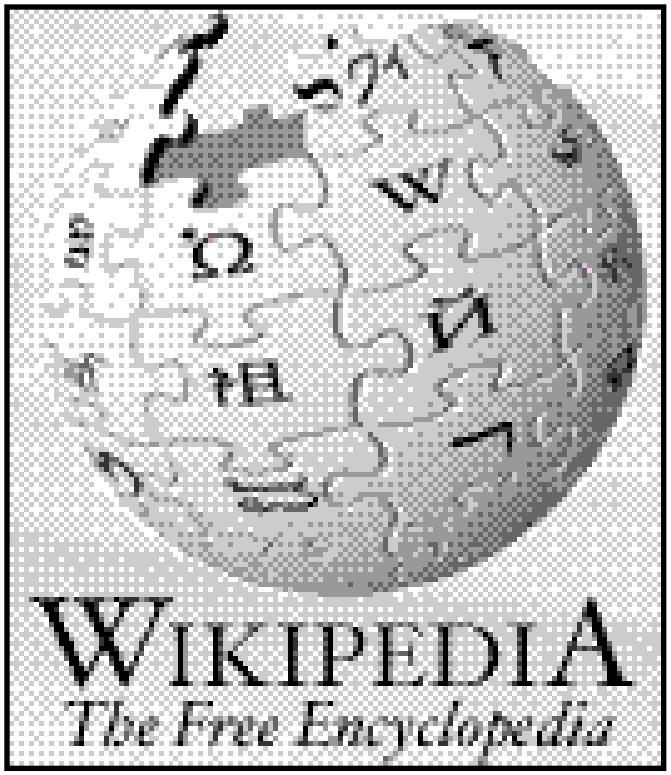
Tablo 1.11. 9465 kişi üzerinde gözlenen kalp atım hızı değerlerinin frekans dağılışı

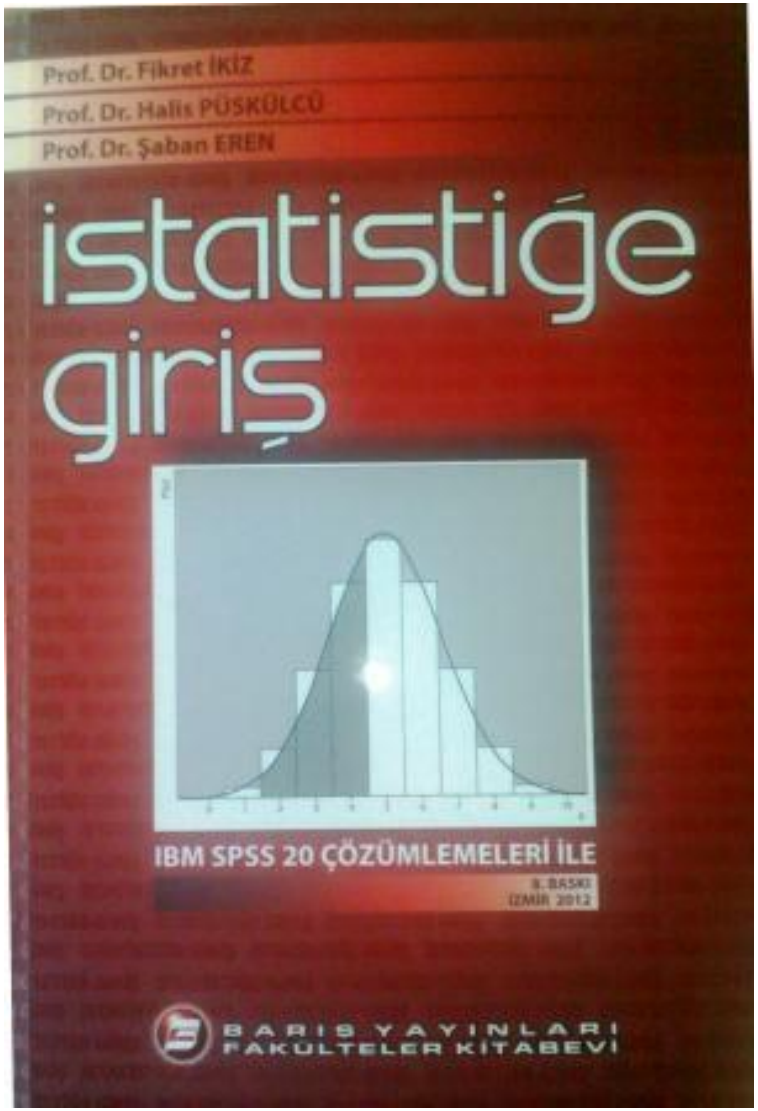
Sınıf Orta Değeri (x)	Frekans (f)	Kodlanmış Sınıf Orta Değeri (x_c)
59.5	2	0
67.5	6	1
75.5	39	2
83.5	385	3
91.5	888	4
99.5	1729	5
107.5	2240	6
115.5	2007	7
123.5	1233	8
131.5	641	9
139.5	201	10
147.5	74	11
155.5	14	12
163.5	5	13
171.5	1	14
n = 9465		

X değişkeni $x_c = \frac{x - 59.5}{8}$ şeklinde kodlanmıştır.

Kodlanmış sınıf orta değerlerinin frekans sütunu ile karşılıklı

references


WIKIPEDIA
The Free Encyclopedia


Prof. Dr. Fikret İKİZ
Prof. Dr. Halis PÜSKÜLCÜ
Prof. Dr. Şaban EREN
istatistiğe giriş
IBM SPSS 20 ÇÖZÜMLEMELERİ İLE
BARIŞ YAYINLARI
FAKÜLTELER KİTABEVİ