סטטיסטיקה תיאורית

מהי סטטיסטיקה תיאורית

- לאחר שאספנו נתונים, אנו רוצים לגלות את המידע הבסיסי שנמצא בהם.
- מידע זה הוא מידע על התפלגות המשתנים שמרכיבים את הנתונים שלנו, ולעיתים גם על הקשרים ביניהם
- האופן בו ניתן לתאר כל משתנה תלוי בסולם המדידה בו הוא נמדד
- ניתן לתאר את המשתנים באופן כמותי ובאופן ויזואלי

© Yossi Levy 2021

דוגמא: נתוני הטיטניק

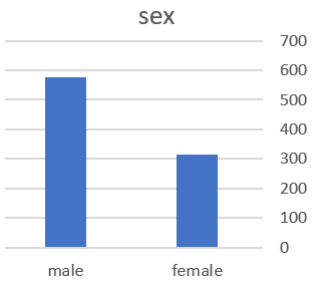
titanic.xlsx - Excel

File Home Insert Page Layout Formulas Data Review View Add-ins Help Tell me what you want to do

O51

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen H	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Lai	female	26	0	0	STON/O2. 3	7.925		S
5	4	1	1	Futrelle, Mrs. Jacque	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William He	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timo	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gc	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Osca	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nichol	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, Miss. M	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Miss. Elizat	female	58	0	0	113783	26.55	C103	S

תיאור של משתנה שמי



sex	count	percent
female	314	35.2%
male	577	64.8%
total	891	100.0%

דיאגרמת עוגה

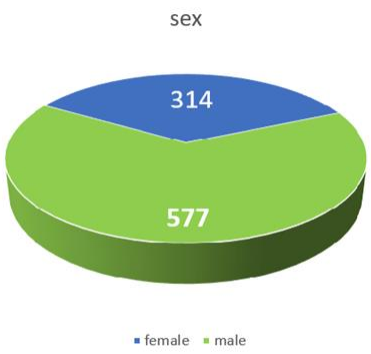
דיאגרמת עמודות

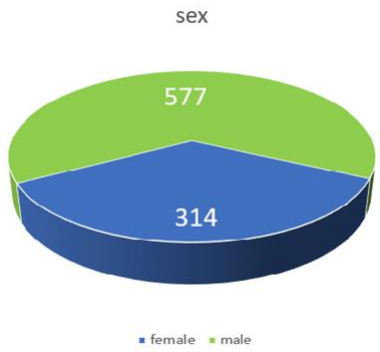
טבלת שכיחות

הערך השכיח: male

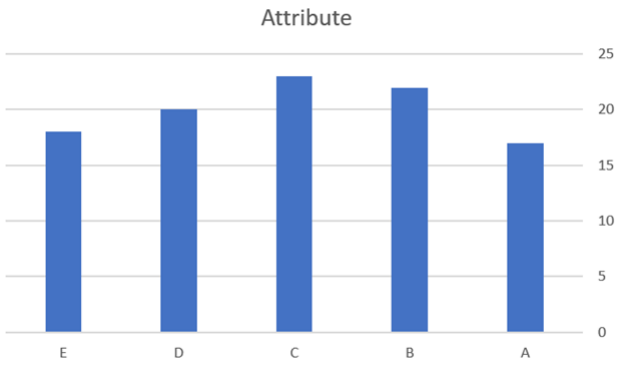
© Yossi Levy 2021

הזהרו מדיאגרמות עוגה!





הזהרו מדיאגרמת עוגה!



© Yossi Levy 2021

כמה נוסעים היו בכל מחלקה?

Class	Count	Percent
1	216	24.2%
2	184	20.7%
3	491	55.1%
Total	891	100.0%

השכיח הוא **מדד מיקום מרכזי**

רוב הנוסעים נסעו במחלקה השלישית.
הערך השכיח הוא 3

אבל: שימו לב כי המחלקה היא משתנה סדר!
זה מאפשר לנו לומר דברים נוספים על הנתונים.

תיאור משתנים בסולם סדר

age group	Freq	Pct	Cum Freq	Cum Pct
0-9	62	8.7%	62	8.7%
10-19	102	14.3%	164	23.0%
20-29	220	30.8%	384	53.8%
30-39	167	23.4%	551	77.2%
40-49	89	12.5%	640	89.6%
50-59	48	6.7%	688	96.4%
60-69	19	2.7%	707	99.0%
70-79	6	0.8%	713	99.9%
80-89	1	0.1%	714	100.0%
Total	714	100.0%		

- תחילה נחלק את נתוני הגיל לקבוצות גיל – וכך למעשה נייצר משתנה סדר לצורך הדוגמא, ו-"נשכח" את הנתונים המקוריים
- יותר מ-50% מהנוסעים היו בקבוצת הגיל 20-29 או בקבוצת גיל יותר נמוכה
- כמו כן, יותר מ-50% מהנוסעים היו בקבוצת הגיל 20-29 או בקבוצת גיל יותר גבוה
- במקרה כזה נאמר כי קבוצת הגיל 20-29 היא קבוצת הגיל החציונית

© Yossi Levy 2021

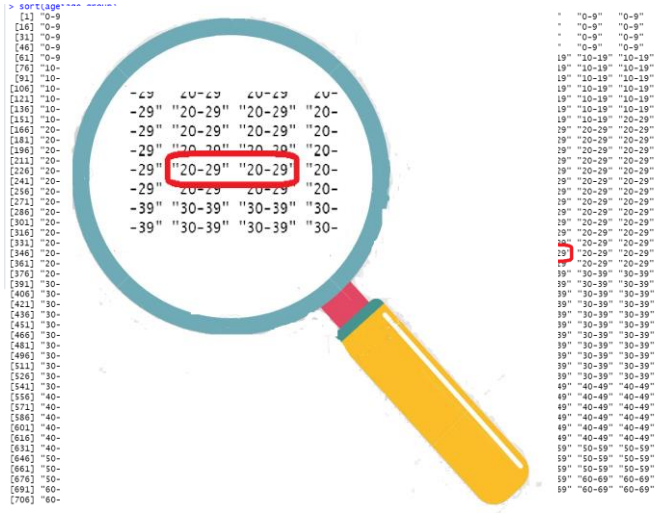
אינטואיציה

```
> sort(age$age_group)
  [1] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
  [8] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
 [15] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
 [22] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
 [29] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
 [36] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
 [43] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
 [50] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"
 [57] "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "0-9"   "10-19"
 [64] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
 [71] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
 [78] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
 [85] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
 [92] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
 [99] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
[106] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
[113] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
[120] "10-19" "10-19" "10-19" "10-19" "10-19" "10-19" "10-19"
```

נסדר את כל הנוסעים בשורה על פי הגיל שלהם:

- ראשונים יהיו הנוסעים בגילאי 0 עד 9
- אחריהם יעמדו כל הנוסעים בגילאי 10 עד 19
- וכן הלאה....
- מי עומד בדיוק באמצע השורה?

אינטואיציה



- יש לנו בסך הכל נתונים על 714 נוסעים
- אמצע השורה נמצא בין המקום ה-357 והמקום ה-358
- מי עומד שם?

© Yossi Levy 2021

חציון, רבעונים ואחוזונים

הגדרות:

- החציון הוא הערך שלפחות 50% מהנתונים נמוכים ממנו ולפחות 50% גבוהים ממנו
- הרבעון התחתון הוא הערך שלפחות 25% מהנתונים נמוכים ממנו ולפחות 75% גבוהים ממנו
- הרבעון העליון הוא הערך שלפחות 75% מהנתונים נמוכים ממנו ולפחות 25% גבוהים ממנו

באופן כללי

- האחוזון ה-p הוא הערך שלפחות p% מהנתונים נמוכים ממנו ולפחות (100-p)% גבוהים ממנו

החציון הוא **מדד מיקום מרכזי**

האחוזונים השונים הם **מדדי מיקום יחסי**

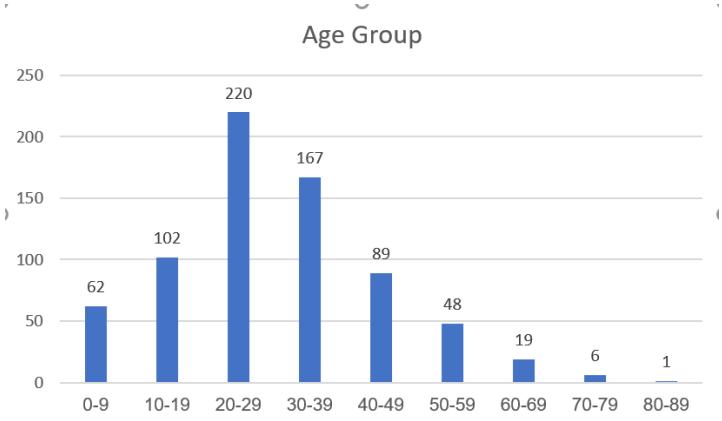
חישוב אחוזונים: דוגמאות

age group	Freq	Pct	Cum Freq	Cum Pct
0-9	62	8.7%	62	8.7%
10-19	102	14.3%	164	23.0%
20-29	220	30.8%	384	53.8%
30-39	167	23.4%	551	77.2%
40-49	89	12.5%	640	89.6%
50-59	48	6.7%	688	96.4%
60-69	19	2.7%	707	99.0%
70-79	6	0.8%	713	99.9%
80-89	1	0.1%	714	100.0%
Total	714	100.0%		

- רבעון תחתון: הערך בו השכיחות המצטברת עולה לראשונה על 25%
- חציון: הערך בו השכיחות המצטברת עולה לראשונה על 50%
- רבעון עליון: הערך בו השכיחות המצטברת עולה לראשונה על 75%
- עשירון עליון: הערך בו השכיחות המצטברת עולה לראשונה על 90%4

© Yossi Levy 2021

הצגה ויזואלית: דיאגרמת עמודות



שימו לב כי באן יש לשמור על סדר העמודות שעל הציר האופקי

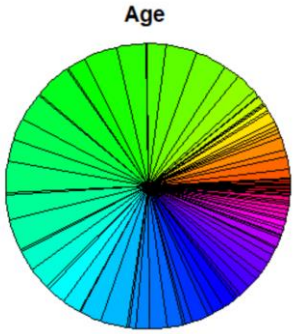
משתנים כמותיים

- מדדי מיקום מרכזי
- מדדי פיזור
- מדדי מיקום יחסי

© Yossi Levy 2021

תיאור משתנים כמותיים

- בזכור – משתנים כמותיים הם משתנים הנמדדים בסולם רווח או סולם מנה
- השיטות לתיאור משתנים כמותיים הן אותן שיטות למשתני רווח ולמשתני מנה
- על פי ההיררכיה של סולמות המדידה, כל השיטות לתיאור משתנים הנמדדים בסולם שמי או סולם סדר מתאימות גם לתיאור משתנים כמותיים
- בפועל – רוב השיטות האלה אינן יעילות



משתנים כמותיים – מדדי מיקום מרכזי

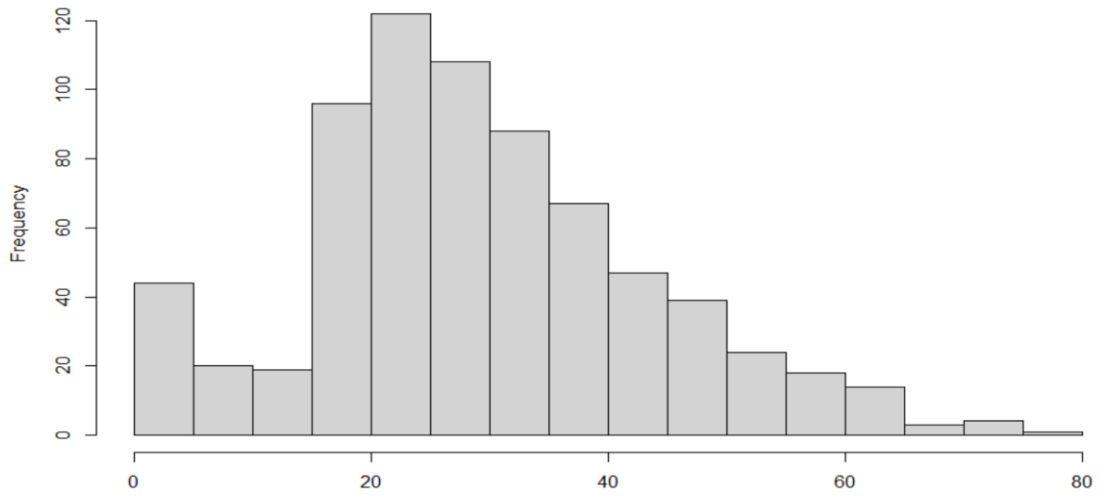
- השכיח – בדרך כלל חסר משמעות
- החציון - הערך שלפחות 50% מהנתונים נמוכים ממנו ולפחות 50% גבוהים ממנו
- הממוצע – סכום כל התצפיות חלקי מספר התצפיות

בנתוני הטיטניק

- הגיל השכיח – 24 (30 מתוך 714, 4.2%)
- הגיל החציוני - 28
- הגיל הממוצע 29.7

www.sci-princes.info © Yossi Levy 2021

גילאי הנוסעים בטיטניק - היסטוגרמה

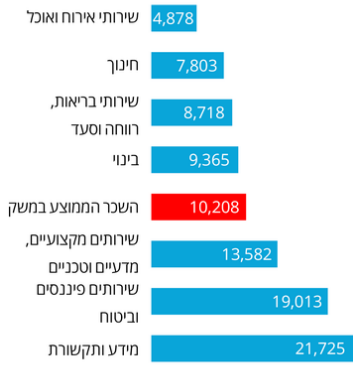


רגישות לערכים קיצוניים: הממוצע מול החציון

- במקום עבודה מסויים, שבו יש 50 עובדים, השכר הממוצע הוא כ-7200 ₪ בחודש
- אבל: 49 מהעובדים משתכרים שכר מינימום – כ-5300 ₪ לחודש
- איך זה קורה? שכר המנהל הוא 100000 ₪ לחודש
- אם המנהל ייתן לעצמו תוספת שכר של 50000 ₪ לחודש – השכר הממוצע יעלה ל-8200 ₪ בחודש
- לעומת זאת: השכר החציוני הוא 5300 ₪ לפני ההעלאה ואחרי ההעלאה

הפערים נותרו רחבים

השכר הממוצע בינואר 2018, בשקלים



מקור: הלמ"ס

מקור: https://www.themarker.com/allnews/1.5979205

להרחבה בנושא חפשו בגוגל: נסיכת המדעים המנהל והפועלים

© Yossi Levy 2021

ממוצע משוקלל

אדם יצא לדרך של 240 ק"מ במכוניתו. את המחצית הראשונה של הדרך, 120 ק"מ, נסע במהירות של 80 קמ"ש. את המחצית השניה נסע במהירות 120 קמ"ש. מה הייתה מהירותו הממוצעת לאורך הדרך?

פתרון

- את 120 הק"מ הראשונים עבר במהירות 80 קמ"ש ולכן משך הזמן בו עבר מרחק זה הוא 1.5 שעות.
- את 120 הק"מ האחרונים עבר במהירות 120 קמ"ש ולכן משך הזמן בו עבר מרחק זה הוא שעה אחת.
- סך כל הזמן הנסיעה היה לכן 2.5 שעות
- המהירות הממוצעת=אורך הדרך חלקי משך הזמן = 240/3 = 96 קמ"ש

ממוצע משוקלל

התרגיל שעשינו הוא

$$V = \frac{240}{\frac{120}{80} + \frac{120}{120}} = \frac{240}{1.5 + 1} = \frac{1.5 \cdot 80 + 1 \cdot 120}{2.5} = 96$$

שקללנו את המהירויות במשכי הזמן בו נסע בכל מהירות

www.sci-princes.info © Yossi Levy 2021

ממוצע משוקלל

- דרך אחרת להסתכל על זה:
- בדקה הראשונה של הנסיעה הוא נסע במהירות 80 קמ"ש
- בדקה הראשונה של הנסיעה הוא נסע במהירות 60 קמ"ש

...

- בדקה ה-90 של הנסיעה הוא נסע במהירות 60 קמ"ש

- בדקה ה-91 של הנסיעה הוא נסע במהירות 120 קמ"ש
- בדקה ה-92 של הנסיעה הוא נסע במהירות 120 קמ"ש

...

- בדקה ה-150 של הנסיעה הוא נסע במהירות 120 קמ"ש

ממוצע משוקלל

ולכן כאן החישוב הוא

$$V = \frac{80 + 80 + .. + 80 + 120 + 120 + \cdots 120}{150} =$$

$$= \frac{90 \cdot 80 + 60 \cdot 120}{150} = 96$$

שימו לב כי בשני המקרים שיקללנו את המהירויות על פי המכנה – כלומר על פי הזמנים

www.sci-princes.info © Yossi Levy 2021

דוגמה: נתוני תחלואה בקורונה

מה מספר מקרי המוות למיליון תושבים ב-15 מדינות אלה?

location	last updated date	total cases	total deaths	total cases per million	total deaths per million	population	human development index
Norway	07/03/2021	74942	632	13823.8	116.578	5421242	0.957
Ireland	07/03/2021	223219	4422	45206.2	895.541	4937796	0.955
Switzerlar	07/03/2021	562290	10048	64969.9	1161	8654618	0.955
Iceland	07/03/2021	6059	29	17755.3	84.982	341250	0.949
Germany	07/03/2021	2508655	71984	29942	859.162	83783945	0.947
Sweden	07/03/2021	684961	13003	67822.8	1287.52	10099270	0.945
Australia	07/03/2021	29046	909	1139.06	35.647	25499881	0.944
Netherlar	07/03/2021	1135258	15962	66254.2	931.551	17134873	0.944
Denmark	07/03/2021	215528	2380	37210	410.897	5792203	0.94
Finland	07/03/2021	62063	767	11201.3	138.43	5540718	0.938
Singapor	07/03/2021	60033	29	10261.5	4.957	5850343	0.938
United Ki	07/03/2021	4231166	124736	62327.5	1837.43	67886004	0.932
Belgium	07/03/2021	787891	22261	67982.5	1920.77	11589616	0.931
New Zeal	07/03/2021	2405	26	498.732	5.392	4822233	0.931

25 המדינות בעלות מדד הפיתוח האנושי הגבוה ביותר (להוציא הונג-קונג)

מקור: https://github.com/owid/covid-19-data/blob/master/public/data/latest/owid-covid-latest.csv

דוגמה: נתוני תחלואה בקורונה

פתרון לא נכון: לחשב את הממוצע של מספר מקרי המוות למיליון. התוצאה: 832.8

location	last updated date	total cases	total deaths	total cases per million	total deaths per million	population	human development index
Norway	07/03/2021	74942	632	13823.8	116.578	5421242	0.957
Ireland	07/03/2021	223219	4422	45206.2	895.541	4937796	0.955
Switzerlar	07/03/2021	562290	10048	64969.9	1161	8654618	0.955
Iceland	07/03/2021	6059	29	17755.3	84.982	341250	0.949
Germany	07/03/2021	2508655	71984	29942	859.162	83783945	0.947
Sweden	07/03/2021	684961	13003	67822.8	1287.52	10099270	0.945
Australia	07/03/2021	29046	909	1139.06	35.647	25499881	0.944
Netherlar	07/03/2021	1135258	15962	66254.2	931.551	17134873	0.944
Denmark	07/03/2021	215528	2380	37210	410.897	5792203	0.94
Finland	07/03/2021	62063	767	11201.3	138.43	5540718	0.938
Singapor	07/03/2021	60033	29	10261.5	4.957	5850343	0.938
United Ki	07/03/2021	4231166	124736	62327.5	1837.43	67886004	0.932
Belgium	07/03/2021	787891	22261	67982.5	1920.77	11589616	0.931
New Zeal	07/03/2021	2405	26	498.732	5.392	4822233	0.931

© Yossi Levy 2021

דוגמה: נתוני תחלואה בקורונה

location	last updated date	total cases	total deaths	total cases per million	total deaths per million	population	human development index
Norway	07/03/2021	74942	632	13823.8	116.57	5421242	0.957
Ireland	07/03/2021	223219	4422	45206.2	895.54	4937796	0.955
Switzerla	07/03/2021	562290	10048	64969.9	116	8654618	0.955
Iceland	07/03/2021	6059	29	17755.3	84.98	341250	0.949
Germany	07/03/2021	2508655	71984	29942	859.16	83783945	0.947
Sweden	07/03/2021	684961	13003	67822.8	1287.5	10099270	0.945
Australia	07/03/2021	29046	909	1139.06	35.64	25499881	0.944
Netherla	07/03/2021	1135258	15962	66254.2	931.55	17134873	0.944
Denmark	07/03/2021	215528	2380	37210	410.89	5792203	0.94
Finland	07/03/2021	62063	767	11201.3	138.4	5540718	0.938
Singapor	07/03/2021	60033	29	10261.5	4.95	5850343	0.938
United Ki	07/03/2021	4231166	124736	62327.5	1837.4	67886004	0.932
Belgium	07/03/2021	787891	22261	67982.5	1920.7	11589616	0.931
New Zeal	07/03/2021	2405	26	498.732	5.39	4822233	0.931

פתרון נבון:

- לסכם את סך מקרי המוות
- לסכם את סך האוכלוסייה
- לחלק את סך מקרי המוות בסך האוכלוסייה

1003416 מיליון 936.3

1003416/936.3= 1071.7

דוגמה: נתוני תחלואה בקורונה

פתרון שקול: לחשב ממוצע של מספר מקרי המוות למיליון משוקלל על פי גודל האוכלוסייה (במיליונים)

- נורבגיה: 116.6 מקרי מוות למיליון, אוכלוסייה – 5.4 מיליון
- אירלנד: 895.5 מקרי מוות למיליון, אוכלוסייה – 4.9 מיליון

...

- ספרד: 1521.5 מקרי מוות למיליון, אוכלוסייה – 46.6 מיליון
- צרפת: 1359.3 מקרי מוות למיליון, אוכלוסייה – 65.3 מיליון

location	last updated date	total cases	total deaths	total cases per million	total deaths per million	population	human development index
Norway	07/03/2021	74942	632	13823.8	116.57	5421242	0.957
Ireland	07/03/2021	223219	4422	45206.2	895.54	4937796	0.955
Switzerla	07/03/2021	562290	10048	64969.9	116	8654618	0.955
Iceland	07/03/2021	6059	29	17755.3	84.98	341250	0.949
Germany	07/03/2021	2508655	71984	29942	859.16	83783945	0.947
Sweden	07/03/2021	684961	13003	67822.8	1287.5	10099270	0.945
Australia	07/03/2021	29046	909	1139.06	35.64	25499881	0.944
Netherla	07/03/2021	1135258	15962	66254.2	931.55	17134873	0.944
Denmark	07/03/2021	215528	2380	37210	410.89	5792203	0.94
Finland	07/03/2021	62063	767	11201.3	138.4	5540718	0.938
Singapor	07/03/2021	60033	29	10261.5	4.95	5850343	0.938
United Ki	07/03/2021	4231166	124736	62327.5	1837.4	67886004	0.932
Belgium	07/03/2021	787891	22261	67982.5	1920.7	11589616	0.931
New Zeal	07/03/2021	2405	26	498.732	5.39	4822233	0.931

$$\frac{116.6 \cdot 5.4 + 895.5 \cdot 4.9 + \cdots + 1359.3 \cdot 65.3}{5.4 + 4.9 + \cdots + 65.3} = 1071.8$$

© Yossi Levy 2021

חישוב הממוצע מתוך טבלת שכיחות

דוגמה: הבליץ על לונדון

מספר הפגיעות	מספר האיזורים
0	229
1	211
2	93
3	35
4	7
5*	1
סך הכל	576

$$\bar{X}_n = \frac{0+0+\cdots+0+1+1+\cdots+1+\cdots 4+4+\cdots 4+5}{576}$$

$$= \frac{229\cdot 0 + 211\cdot 1 + 93\cdot 2 + 35\cdot 3 + 7\cdot 4 + 1\cdot 5}{576}$$

$$= \frac{535}{576} = 0.93$$

חישוב הממוצע מתוך טבלת שכיחות

דוגמה: נתוני הטיטניק

age group	Freq	Midpoint
0-9	62	5
10-19	102	15
20-29	220	25
30-39	167	35
40-49	89	45
50-59	48	55
60-69	19	65
70-79	6	75
80-89	1	85
Total	714	

$$\bar{X}_n = \frac{62\cdot 5 + 102\cdot 15 + \cdots + 6\cdot 75 + 1\cdot 85}{714}$$

$$= \frac{21600}{714} = 30.25$$

בזכור: בחישוב ישיר קיבלנו כי הגיל הממוצע הוא 29.7

© Yossi Levy 2021

תזכורת: חציון, רבעונים ואחוזונים

הגדרות:

- החציון הוא הערך שלפחות 50% מהנתונים נמוכים ממנו ולפחות 50% גבוהים ממנו
- הרבעון התחתון הוא הערך שלפחות 25% מהנתונים נמוכים ממנו ולפחות 75% גבוהים ממנו
- הרבעון העליון הוא הערך שלפחות 75% מהנתונים נמוכים ממנו ולפחות 25% גבוהים ממנו

באופן כללי

- האחוזון ה-p הוא הערך שלפחות p% מהנתונים נמוכים ממנו ולפחות %(100-p) גבוהים ממנו

החציון הוא **מדד מיקום מרכזי**

האחוזונים השונים הם **מדדי מיקום יחסי**

חישובי חציון, רבעונים ואחוזונים – נתוני הקורונה

- נמיין את הנתונים מהקטן לגדול

```
> sort(covid$total_deaths_per_million)
 [1]    4.957    5.392   31.110   35.647   61.814   84.982  114.734
 [8]  133.918  406.408  580.650  658.191  834.754  870.834  911.941
[15]  945.439 1017.611 1150.946 1269.993 1313.519 1415.948 1478.822
[22] 1542.157 1806.676 1838.925 1901.185
>
```


© Yossi Levy 2021

חישוב החציון

```
> sort(covid$total_deaths_per_million)
 [1]    4.957    5.392   31.110   35.647   61.814   84.982  114.734
 [8]  133.918  406.408  580.650  658.191  834.754  870.834  911.941
[15]  945.439 1017.611 1150.946 1269.993 1313.519 1415.948 1478.822
[22] 1542.157 1806.676 1838.925 1901.185
>
```

- תזכורת: החציון הוא הערך שלפחות 50% מהנתונים קטנים ממנו או שווים לו ולפחות 50% גדולים ממנו או שווים לו
- הגדרה לא מדוייקת: "החציון נמצא באמצע"
- יש לנו 25 נתונים, ולכן 50% מהנתונים הם 12.5
- לכן החציון הוא הנתון הנמצא במקום ה-13: 870.8
- יש 13 נתונים הקטנים מ- או שווים ל-870.834 ו-13 נתונים הגדולים מ- או שווים ל-870.834

חישובי הרבעונים

```
> sort(covid$total_deaths_per_million)
 [1]    4.957    5.392   31.110   35.647   61.814   84.982  114.734
 [8]  133.918  406.408  580.650  658.191  834.754  870.834  911.941
[15]  945.439 1017.611 1150.946 1269.993 1313.519 1415.948 1478.822
[22] 1542.157 1806.676 1838.925 1901.185
>
```

- תזכורת: הרבעון התחתון הוא הערך שלפחות 25% מהנתונים קטנים ממנו ולפחות 75% גדולים ממנו
- יש לנו 25 נתונים, ולכן 25% מהנתונים הם 6.25
- לכן הרבעון התחתון הוא הנתון הנמצא במקום ה-7: 114.7
- באופן דומה: הרבעון העליון הוא הנתון הנמצא במקום ה-18: 1313.5

© Yossi Levy 2021

חישובי עשירונים

```
> sort(covid$total_deaths_per_million)
 [1]    4.957    5.392   31.110   35.647   61.814   84.982  114.734
 [8]  133.918  406.408  580.650  658.191  834.754  870.834  911.941
[15]  945.439 1017.611 1150.946 1269.993 1313.519 1415.948 1478.822
[22] 1542.157 1806.676 1838.925 1901.185
>
```

- תזכורת: העשירון העליון/העשירי הוא הערך שלפחות 90% מהנתונים קטנים ממנו ולפחות 10% גדולים ממנו
- יש לנו 25 נתונים, ולכן 90% מהנתונים הם 22.5
- לכן העשירון העליון הוא הנתון הנמצא במקום ה-23: 1806.7
- באופן דומה: העשירון ה-4 הוא 580.7

מדדי פיזור מוחלט

- הטווח: ערך התצפית הגדולה ביותר פחות ערך התצפית הנמוכה ביותר
- התחום הבין רבעוני: הרבעון העליון פחות הרבעון התחתון
- הטווח מושפע מערכי קיצון, אך התחום הבין רבעוני אינו מושפע מהם

- בנתוני הטיטניק, הטווח הוא 79.58, התחום הבין רבעוני הוא 17.88

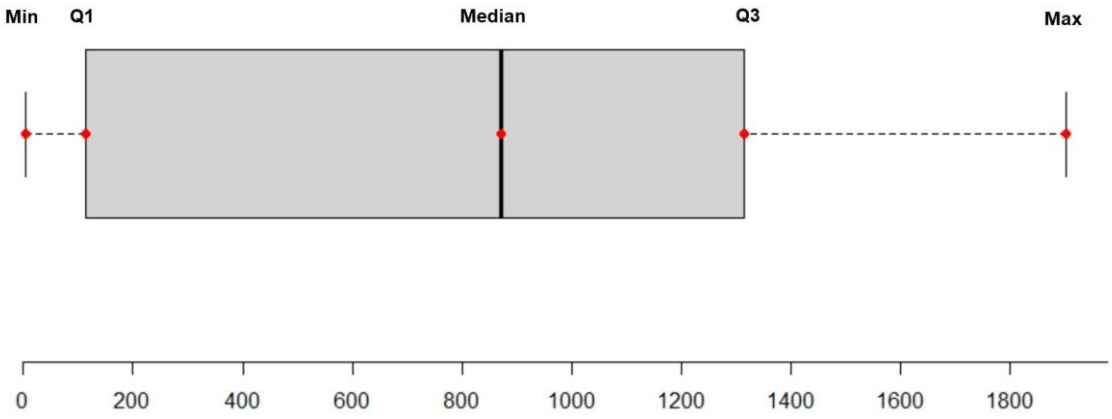
```
> summary(titanic$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.42   20.12   28.00   29.70   38.00   80.00     177
>
```

- בנתוני הקורונה, הטווח הוא 1896.2, התחום הבין רבעוני הוא 1198.785

```
> summary(covid$total_deaths_per_million)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   4.957  114.734  870.834  816.662 1313.519 1901.185
```


© Yossi Levy 2021

דיאגרמת קופסה (boxplot) נתוני קורונה



זיהוי תצפיות חריגות

- בנתוני הטיטניק, ששת האנשים המבוגרים ביותר הם בגילאים: 70, 70.5, 71, 71, 74, 80.

- האם האנשים האלה חריגים בגילם יחסית לשאר הנוסעים?
- פתרון אפשרי: נקבע סף ונגדיר נתונים חריגים כבאלה שערכם גדול מהסף. באופן דומה נגדיר סף עבור ערכים נמוכים במיוחד.

- הבעיה: איך לקבוע את הסף?
- הצעה (John Tukey): נקבע אותו על פי מדדי הפיזור, למשל על פי ערכי התחום הבין רבעוני

© Yossi Levy 2021

זיהוי גילאים חריגים על פי ערכי הרבעונים (טיטניק)

- מחשבים את התחום הבין רבעוני

$$IQR = Q_3 - Q_1 = 38 - 20.12 = 17.88$$

- מכפילים אותו ב-1.5

$$1.5 \cdot IQR = 26.82$$

- מחסירים ערך זה מהרבעון התחתון כדי לקבל את הסף התחתון

$$L = Q_1 - 1.5 \cdot IQR = 20.12 - 26.82 = -6.7$$

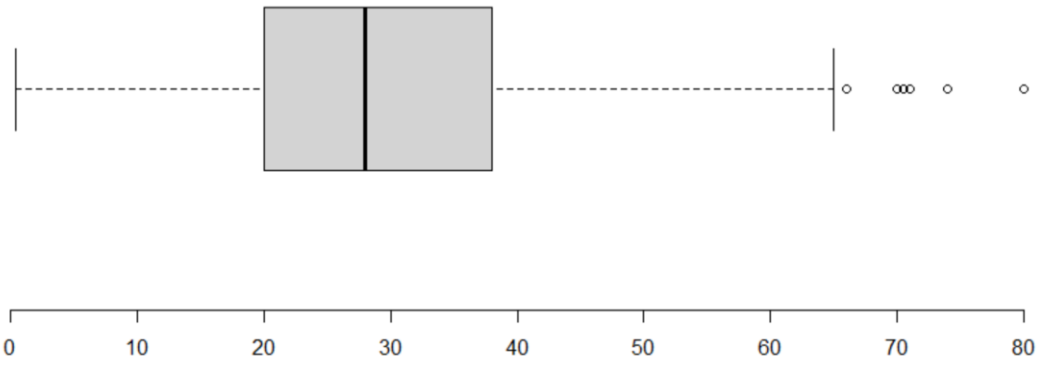
- מוסיפים ערך זה לרבעון העליון כדי לקבל את הסף העליון

$$U = Q_3 + 1.5 \cdot IQR = 64.82$$

מסקנה: כל מי שגילו גבוה מ-64.82 נחשב כבעל גיל חריג. יש 11 נוסעים כאלה.

דיאגרמת קופסה משופרת (נתוני הטיטניק)

- הזרועות יגיעו רק עד ערך התחום התחתון ו/או ערך התחום העליון.
- תצפיות חריגות יסומנו בנקודות



© Yossi Levy 2021

מדדי פיזור יחסי

- הרעיון: למדוד את הפיזור ביחס למרכז ההתפלגות
- מרכז ההתפלגות יכול להיות, למשל מדד מיקום מרכזי כלשהו
- אנו נתמקד במדד פיזור יחסי אחד בלבד, המודד את הפיזור סביב הממוצע של הנתונים

- דוגמה מלאכותית:
  - מספר ההודעות בווטסאפ המשפחתי בימים א-ה: 14, 19, 28, 24, 18
  - הממוצע הוא 20.6
  - ההפרשים מהממוצע הם: 14-20.6, 19-20.6, 28-20.6, 24-20.6, 18-20.6
  - כלומר: -6.6 , -1.6, 7.4, 3.4, -2.6
  - ממוצע ההפרשים הוא 0. אופס...

השונות

במקום להסתכל על המרחק מהממוצע, נסתכל על ריבוע המרחק

בחזרה לדוגמה:

- מספר ההודעות בווטסאפ המשפחתי בימים א-ה: 14, 19, 28, 24, 18
- הממוצע הוא 20.6
- ריבועי המרחקים מהממוצע הם: $(14-20.6)^2$, $(19-20.6)^2$, $(28-20.6)^2$, $(24-20.6)^2$, $(18-20.6)^2$
- כלומר: 6.76, 11.56, 54.76, 2.56, 43.56
- ממוצע ריבועי המרחקים הוא 23.84. ערך זה הוא **השונות של הנתונים**
- דרך קלה לחישוב: השונות היא ממוצע הריבועים פחות הממוצע בריבוע

$$\frac{18^2 + 24^2 + 28^2 + 19^2 + 14^2}{5} - 20.6^2 = 23.84$$

© Yossi Levy 2021

השונות

הגדרה: **השונות** היא ממוצע ריבועי המרחקים מהממוצע

- סוג של "נוסחה": השונות היא ממוצע הריבועים פחות הממוצע בריבוע

תכונות:

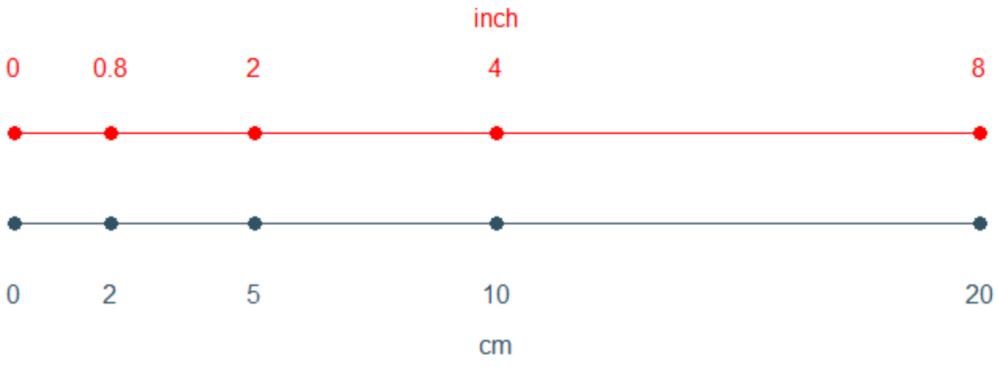
- השונות לעולם אינה קטנה מאפס
- השונות שווה לאפס אם ורק אם כל הנתונים שווים זה לזה
- השונות מושפעת מערכי קיצון, וזאת בניגוד לתחום הבין רבעוני
- יחידות המדידה של השונות הן ריבוע יחידות המדידה של הנתונים

הגדרה: **סטיית התקן** היא השורש (החיובי) של השונות

- יחידות המדידה של סטיית התקן הן יחידות המדידה של הנתונים

מדדים למיקום יחסי: ציון התקן

אינטואיציה: מדידת אורך

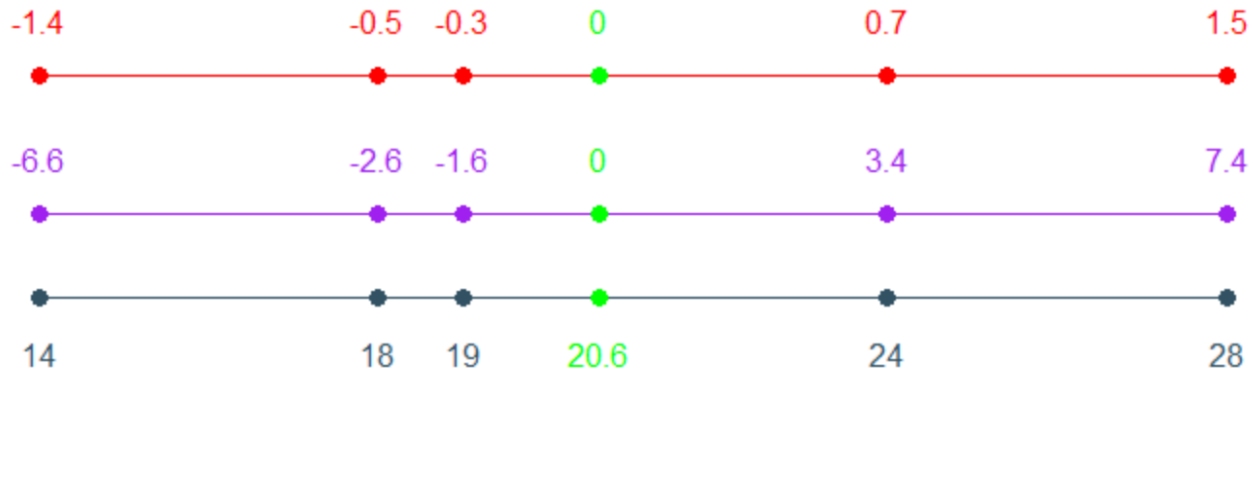


© Yossi Levy 2021

מדדים למיקום יחסי: ציון התקן

- אינטואיציה: לכל נתון מודדים את המרחק מהממוצע על פי יחידת מידה סטנדרטית.
- יחידה זו היא סטיית התקן
- הגדרה: (הנתון פחות הממוצע) חלקי סטיית התקן
- בחזרה לדוגמה שלנו:
  - מספר ההודעות בווטסאפ המשפחתי בימים א-ה: 14, 19, 28, 24 , 18
  - הממוצע הוא 20.6 וסטיית התקן היא 4.88
  - ההפרשים מהממוצע הם: 6.6- , 1.6- , 7.4, 3.4, 2.6-
  - נחלק כל אחד מההפרשים בסטיית התקן ונקבל: 1.4- , 0.3- , 1.5, 0.7, 0.5-
  - אלה הם **ציוני התקן** של כל אחד מהנתונים

מדדים למיקום יחסי: ציוני התקן



© Yossi Levy 2021

תכונות ציון התקן

- ציון התקן הוא מספר "טהור" – ללא יחידות מדידה
- הוא מאפשר השוואה יחסית בין קבוצות נתונים
  - דוגמה: למי היו באופן יחסי יותר הודעות בקבוצת הווטסאפ המשפחתית שלו?
  - לי 28 הודעות ביום ג היו עם ציון תקן 1.5
  - לך היו 17 הודעות ביום ג עם ציון תקן 1.9
  - יום ג שלך היה יותר עמוס מיום ג שלי, באופן יחסי

זיהוי תצפיות חריגות על ידי ציוני תקן

- ערכים גבוהים במיוחד או נמוכים במיוחד של ציוני התקן מתריעים על תצפיות חריגות
- ערכי סף מקובלים הם ±1.96 שמעוגלים ל- ±2 , וכן ±3

- בנתוני הקורונה ציון התקן הנמוך ביותר הוא 1.26- והגבוה היותר הוא 1.58, כלומר אין תצפיות חריגות

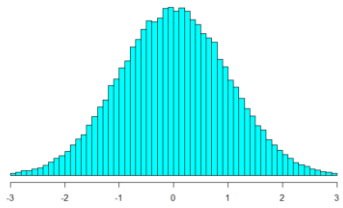
- בנתוני הטיטניק:
  - יש נתון אחד עם ציון תקן נמוך מ-2-: תינוק בן 5 חודשים
  - יש 28 נתונים עם ציון תקן גבוה מ-2: הצעיר שבהם בן 59
  - יש 2 נתונים עם ציון תקן גבוה מ-3: נוסעים בגילאי 74 ו-80

© Yossi Levy 2021

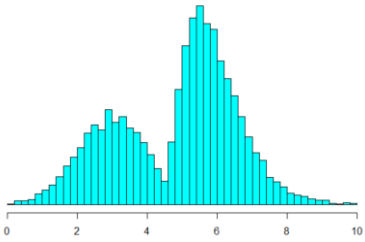
מה עושים עם תצפיות חריגות?

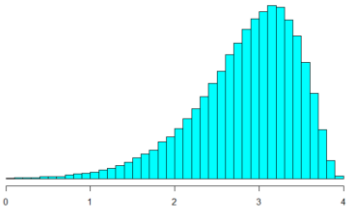
- בדרך כלל בודקים תצפיות חריגות כדי לזהות טעויות אפשריות בנתונים ואז לתקן אותן.
- כבלל, אין להשמיט תצפיות חריגות מהנתונים
- אם בכל זאת מחליטים להשמיט תצפיות חריגות, יש לנמק בנפרד עבור כל תצפית מדוע הוחלט להשמיט אותה

מיון התפלגויות

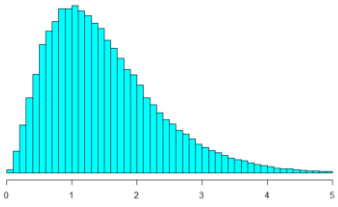


התפלגות סימטרית





זנב שמאלי



זנב ימני

© Yossi Levy 2021

סיכום

ראינו מספר דרכים לתיאור נתונים של משתנה יחיד

- טבלת שכיחות
- הצגות גרפיות: דיאגרמת עמודות, היסטוגרמה, דיאגרמת קופסה
- מדדי מיקום מרכזי: שכיח, חציון, ממוצע
- מדדי פיזור: טווח, תחום בין רבעוני, שונות, סטיית תקן
- מדדי מיקום יחסי: רבעונים, עשירונים, אחוזונים, ציוני תקן
- זיהוי תצפיות חריגות
  - זיהוי חריגה מעל הרבעון העליון או מתחת לרבעון התחתון
  - ציוני תקן חריגים
- מיון התפלגויות: סימטרית/זנב ימני/זנב שמאלי

© Yossi Levy 2021