

A Brief Review of Statistics Concepts

0

Statistics?

A set of principles and procedures for collecting, compiling, analyzing and interpreting data in order to assist in making decisions in the presence of uncertainty.

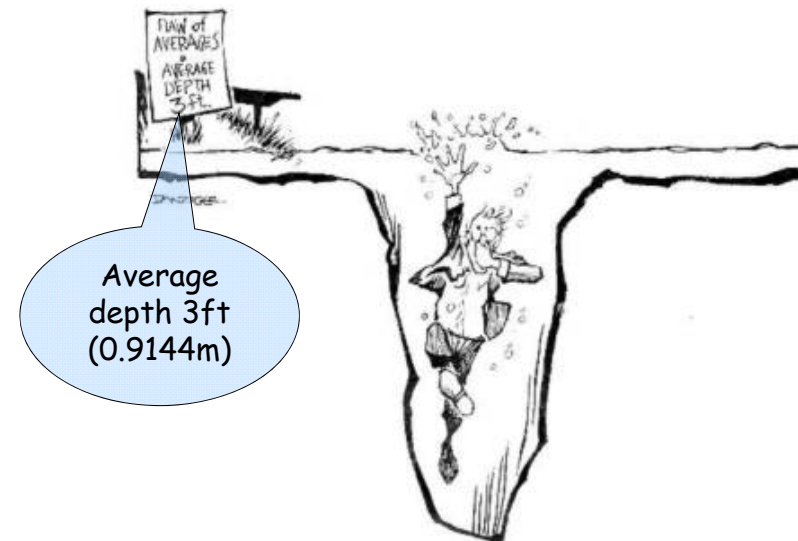
1

Herbert George Wells,
English author, said (circa 1940),

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write”



2



3



Why do we need to understand statistics?

Reasoning with Uncertainty

- *from*
- Peter Donnelly: How juries are fooled by statistics
- <http://www.ted.com/index.php/talks/view/id/67>

4

Ex 1 - Coin Tossing

- Imagine tossing a coin successively, and waiting till the first time a particular pattern appears, say **HTT**
- For example, if the sequence of tosses was

HHTHH**THHTT**HHTTTHTH

- The pattern **HTT** would first appear after the 10th toss

5

- Imagine that half of you toss a coin several times, each time till the sequence **HTT** occurs.
 - Record the average number of tosses till **HTT** occurs
- The other half of you prefer to count **HTH**
 - Record the average number of tosses till **HTH** occurs

6

- Which of the following is true:
 - A. The average number of tosses until **HTH** is larger than the average number of tosses until **HTT**
 - B. The average number of tosses until **HTH** is the same as the average number of tosses until **HTT**
 - C. The average number of tosses until **HTH** is smaller than the average number of tosses until **HTT**

Most people think that B is true but **A is true**. The average number of tosses till **HTH** is 10 and the average number of tosses till **HTT** is 8.

7

- Intuitive explanation:
- Imagine that you win if HTH occurs
 - If the first toss gives a H you are excited and you get even more excited if the second is a T. If the third is H you win but if it is a T you have to start again and wait for the next H.
- If you win when HTT occurs
 - For the first two tosses the experience is the same. However, if the third toss is a H you loose but you already have the first H and are 1/3 of the way to your pattern.

8

It was an example of a simple question on probabilities that most people get wrong.

9

Conclusions from the example

- Randomness, uncertainty and chance are part of our life.
- People make errors of logic when reasoning with uncertainty.
- Errors in statistics may have serious consequences.

It is very important to understand statistics!

10



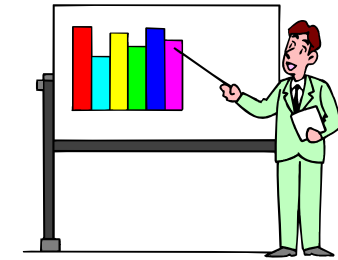
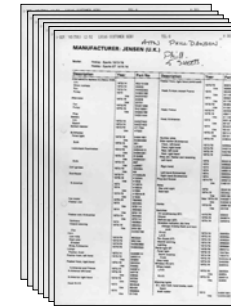
What is the problem here?

On average the temperature is very nice...

11

Descriptive statistics

Descriptive statistics seeks to synthesize and represent in a comprehensible manner the information contained in a data set (through the construction of tables, graphs, calculation of measures)



The purpose of descriptive statistics is to summarize the information contained in data

Example: Final marks on a given course

Aluno	Freq	Final	Aluno	Freq	Final	Aluno	Freq	Final	Aluno	Freq	Final	Aluno	Freq	Final	Aluno	Freq	Final	
1			31	7.3	5.7425	61	5.1		91	8.7	9.50511	121	10.2	11.625	151	5.5	1.65	
2			32	10.3	10.19511	62	7.4	8.17021	92	9.4	10.5201	122	11.5	13.7501	152	11.4	14.13011	
3			33	11.9	10.15025	63	11.0	13.11761	93			123	8.1	10.1476	153	5.9		
4			34	9.8	12.92564	64	8.5	10.985	94	10.4	7.47764	124	9.6	9.60007	154			
5			35	10.0	11.87254	65			95	6.6	10	125	8.3	10.855	155	9.6	9.81	
6	9.6	10.50661	36	16.0	16.5145	66	13.0	13.1225	96	8.5	13.275	126			156	9.3	11.50511	
7	7.2	7.480175	37			67	12.3	9.67507	97	5.5		127	14.6	16.5075	157	6.3	4.890035	
8	5.7	9.620105	38			68	7.3	7.7375	98	6.4		128			158	11.4	12.52011	
9			39	13.3	14.875	69	7.0		99			129	7.7	3.64021	159	11.8	11.3275	
10	7.7	9.52	40			70	9.1		100	10.0	10.665	130	9.5	8.06504	160	8.3		
11	10.2	11.075	41	9.9	11.0725	71			101	10.9	11.9851	131	7.0	6.61514	161			
12	5.8		42	12.3	14.61011	72			102			132	10.2	13.5075	162	7.3	10	
13	15.6	17.685	43	8.7	9.5225	73	10.4	9.560105	103	17.3	16.84	133	12.8	13.0101	163	13.2	13.69	
14	12.0	10.79254	44	7.4	15.55	74	8.8	9.51064	104	7.6		134			164	8.2	11.10025	
15	13.3	15.75	45			75			105			135			165	6.1	7.01	
16	7.8	11.51011	46	3.6	4.580175	76	11.3	13.52981	106	7.7	8.435	136	7.1	6.43507	166			
17	11.6	9.83271	47	7.8	7.065	77	8.7	10.66004	107	10.1	9.785	137	10.9	11.18	167			
18	7.5	7.447675	48	10.4	12.58761	78	10.2	10.5675	108	9.5	13.91	138	8.6	12.625	168			
19			49	12.3	12.28261	79	13.6	17.17	109	10.0	9.72011	139		6.02014	169	10.5	10.95511	
20	12.3	9.5105	50	12.2	9.60005	80	6.7	9.517605	110			140	7.5	9.635	170	9.2	6.97764	
21			51	14.9	15.915	81	7.5	10.23	111			141	9.1	10.2026	171	9.4		
22	11.2	11.37514	52	9.8	13.2002	82	12.9	16.19	112	7.1	10	142	9.9	12.5251	172	9.0	16.9102	
23	4.4	6.325	53	6.1	8.025	83	5.6		113	10.2	12.545	143	6.1	8.13	173	10.1	11.5175	
24	10.5	12.74	54	13.7	11.51261	84	9.8	10.78	114	8.0	9.61	144	6.3	8.34771	174	10.3	12.96011	
25	7.5	11.175	55	15.0	14.93	85	11.3	9.508	115	14.4	12.6501	145	5.6		175	9.4	9.54	
26			56	8.3	9.509605	86			116	12.5	15.8775	146	7.9	8.77507	176	5.4		
27	13.4	14.065	57	7.4		87	12.2	9.52271	117			147	13.9	13.5675	177	13.4	12.01757	
28	9.4	8.105035	58	7.4	10	88	4.6		118	8.3	7.18	148			178	9.5	9.780105	
29	8.8	11.56511	59	12.5	12.535	89	10.7	10.595	119	10.3	8.62021	149	10.5	9.78292	179	10.8	13.705	
30	9.4	12.95261	60	7.3	3.74757	90	7.2	10	120			150	14.5	13.5376	180			
																181	8.8	9.535
																182	5.5	2.94514
																183	8.1	

- Average** 10.52
- Median** 10.51
- Range** 16.29
- Maximum** 17.67
- Minimum** 1.38
- First quartile** 9.068
- Third quartile** 12.68
- Standard deviation** 3.208
- Variance** 10.291
- Asymmetry** -0.25

Average and Median

Example:

\underline{x}
10
12
14
11
7
14
10
12

Average:

$$\bar{x} = (10 + 12 + 14 + 11 + 7 + 14 + 10 + 2) / 8 = 11.25$$

Median:

$$7 \quad 10 \quad 10 \quad 11 \quad | \quad 12 \quad 12 \quad 14 \quad 14$$

11.5

Average and Median

Example

\underline{x}
10
12
14
11
7
200
10
12

Average:

$$\bar{x} = (10 + 12 + 14 + 11 + 7 + 200 + 10 + 2) / 8 = 34.5$$

Median:

$$7 \quad 10 \quad 10 \quad 11 \quad | \quad 12 \quad 12 \quad 14 \quad 200$$

11.5

**The average is more sensitive to extreme values.
ex. average salary vs median salary**

Variance and Standard Deviation

In order to infer the variability of a population from a sample it should be used the sample variance (s^2)

$$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2$$

The sample standard deviation (s), square root of sample variance, has the advantage of being expressed in the same unit as the original data

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2}$$

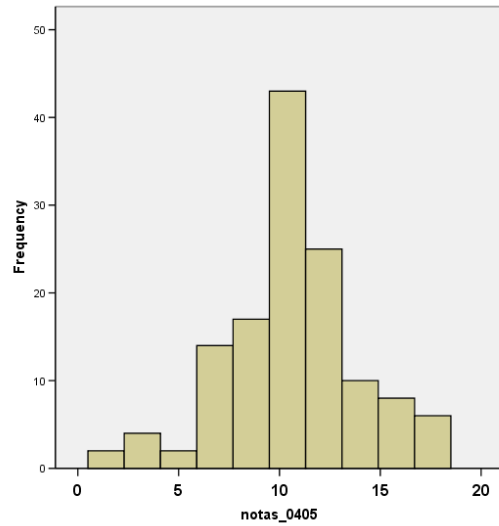
Example: Compute the standard deviation for the following values:
- 4 , -3 , -2 , 3 , 5

X_i	\bar{X}	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
-4	-0,2	-3,8	14,44
-3		-2,8	7,84
-2		-1,8	3,24
3		3,2	10,24
5		5,2	27,04
		Soma=	62,8

Given that $n = 5$ e $62,8 / (5-1) = 15,7$

The square root of 15,7 gives the **standard deviation = 3,96**

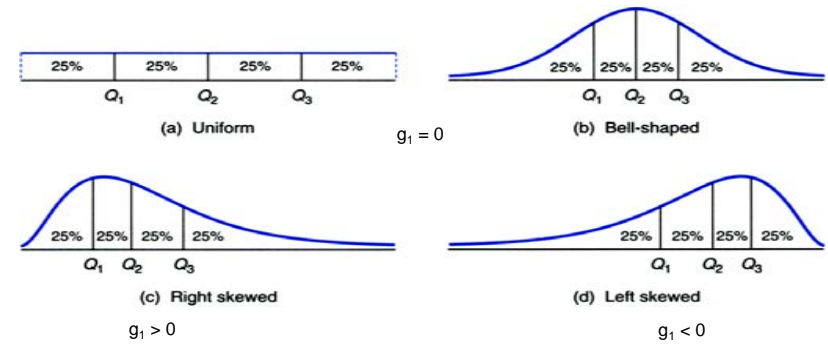
Histogram for the marks



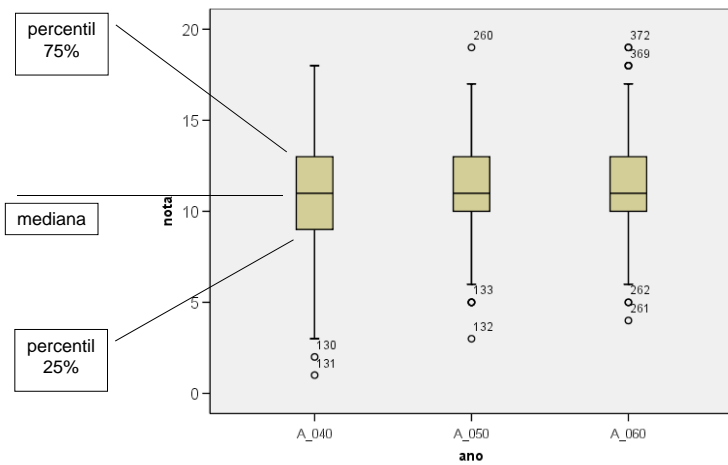
<http://www.stat.tamu.edu/~west/javahtml/Histogram.html>

Assimetry coefficient (g_1)

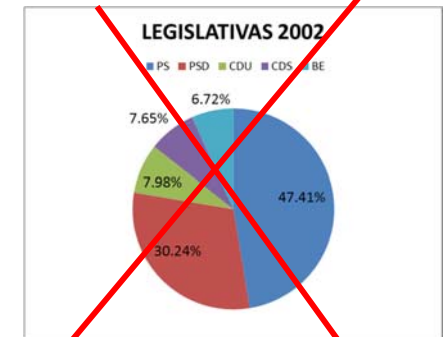
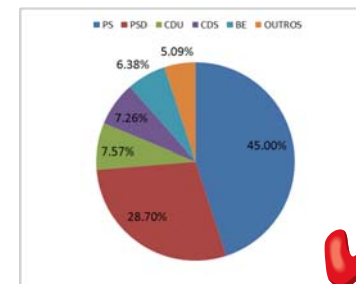
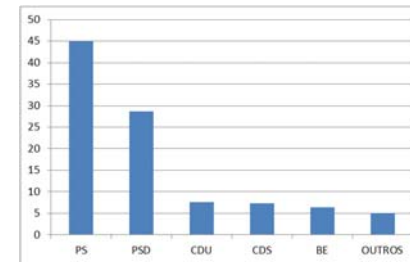
$$g_1 = \frac{k_3}{s^3}, \quad \text{COM} \quad k_3 = \frac{N^2}{(N-1) \cdot (N-2)} \cdot \left(\frac{1}{N} \cdot \sum_{i=1}^N (x_i - \bar{x})^3 \right)$$



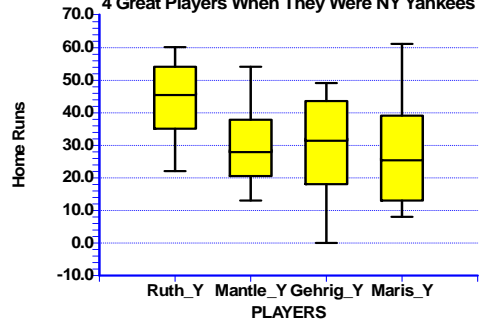
Box-Plot: useful to compare distributions



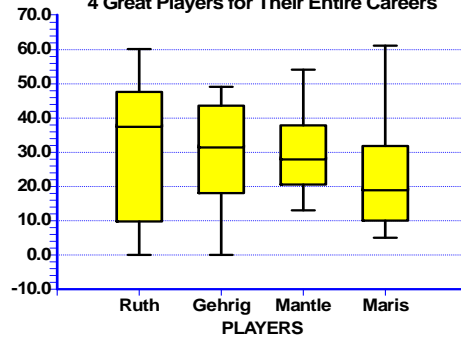
2002 elections



Box Plot of Home Runs per Season for 4 Great Players When They Were NY Yankees



Box Plot of Home Runs per Season for 4 Great Players for Their Entire Careers

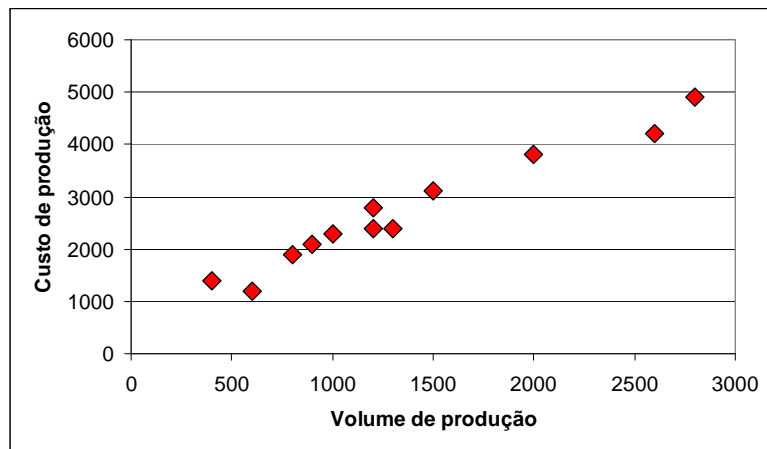


Bivariate samples - quantitative data

The relationship between the two attributes of a bivariate sample with quantitative data can be evidenced by a diagram (X, Y) or, more synthetic, by calculating the degree of fit of a particular relationship

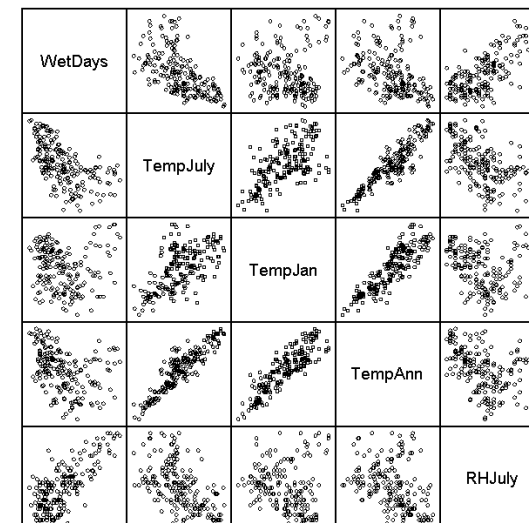
LOTE	VOLUME DE PRODUÇÃO (unidades)	CUSTO DE PRODUÇÃO (contos)
1	1500	3100
2	800	1900
3	2600	4200
4	1000	2300
5	600	1200
6	2800	4900
7	1200	2800
8	900	2100
9	400	1400
10	1300	2400
11	1200	2400
12	2000	3800

SCATTERPLOT

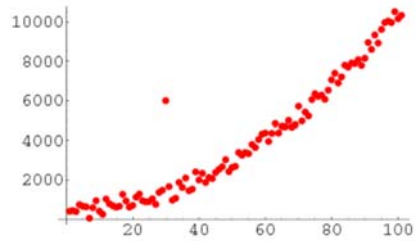
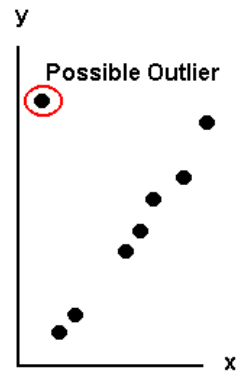
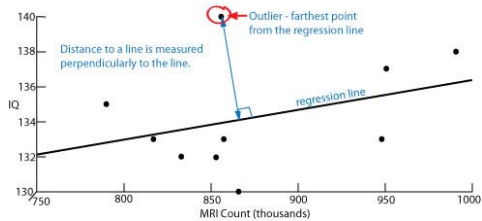


matriz de scatter plots

Climatic predictors



Scatterplots are useful to detect outliers



Measures the degree of adjustment of a linear relationship:

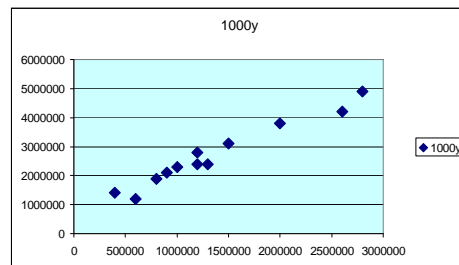
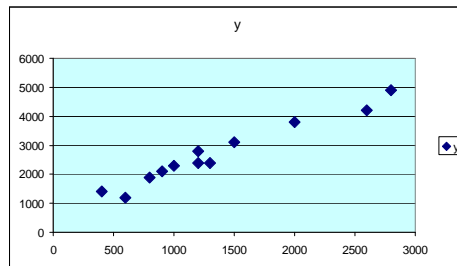
Sample covariance (infer about the population)

$$c_{XY} = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x}) \cdot (y_n - \bar{y})$$

Sample correlation coefficient (adimensional measure)

$$r_{XY} = \frac{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x}) \cdot (y_n - \bar{y})}{\sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \cdot \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (y_n - \bar{y})^2}} = \frac{c_{XY}}{s_X \cdot s_Y} \quad (-1 \leq r_{XY} \leq 1)$$

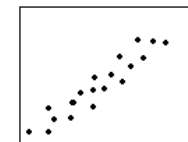
x	y	1000x	1000y
1500	3100	1500000	3100000
800	1900	800000	1900000
2600	4200	2600000	4200000
1000	2300	1000000	2300000
600	1200	600000	1200000
2800	4900	2800000	4900000
1200	2800	1200000	2800000
900	2100	900000	2100000
400	1400	400000	1400000
1300	2400	1300000	2400000
1200	2400	1200000	2400000
2000	3800	2000000	3800000



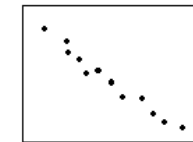
correl: 0.9811009 0.9811009
 cov: 757847.22 7.578E+11

Covariance is affected by the unit in which the variable is expressed.

Degree of Correlation



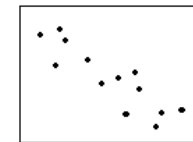
Strong Positive



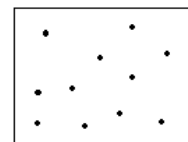
Strong Negative



Weak Positive



Moderate Negative

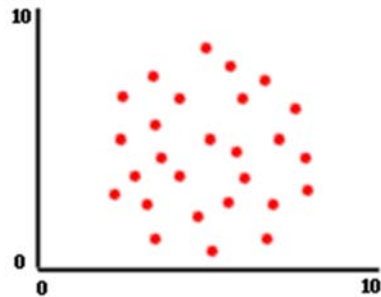


None



Weak Negative

Correlation= 0



http://bcs.whfreeman.com/ips4e/cat_010/applets/CorrelationRegression.html

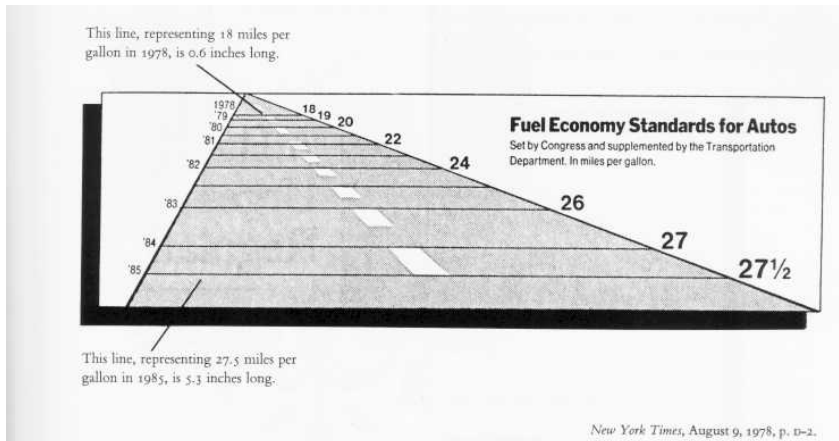


Expresso – 18 Jan. 2003

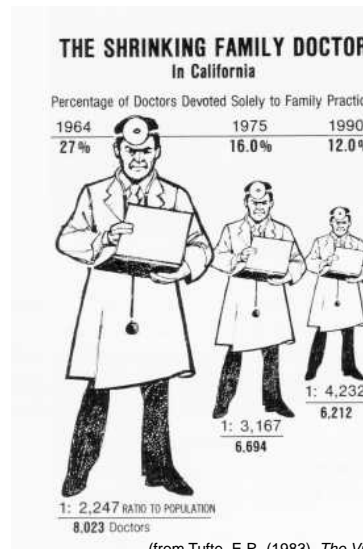
$$\frac{9657}{3449} = 2.8$$

Classical example of how to lie with statistics.

The *Lie Factor* is simply the ratio of the difference in the proportion of the graphic elements versus the difference in the quantities they represent. The most informative graphics are those with a Lie Factor of 1. Here is an example of a badly scaled graphic, with a lie factor of 14.8:



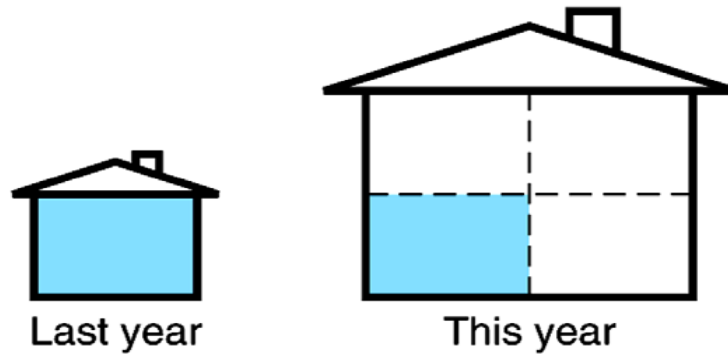
(from Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press)



(from Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press)

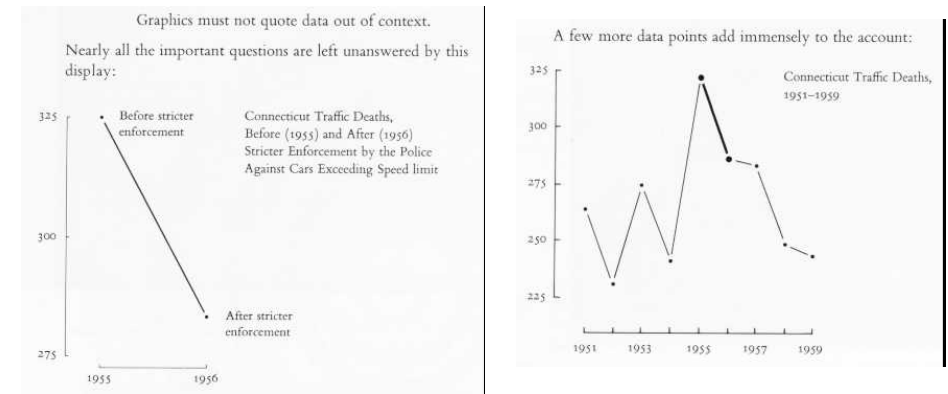
An example of a graph where two-dimensional figures are used to represent one-dimensional values. What often happens is that the size of the graphic is scaled both horizontally and vertically according to the value being graphed. However, this results in the area of the graphic varying with the *square* of the underlying data, causing the eye to read an exaggerated effect in the graph. This graph has a *lie factor* of about 2.8, based on the variation between the area of each doctor graphic and the number it represents.

Los Angeles Times, August 5, 1979, p. 3.



36

One more point about graphs: be sure to include enough context to make the graph meaningful. For instance, one may be tempted to draw unwarranted conclusions based on this graph:



(from Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press)

37

Population and Sample

A population (or universe) is the set of facts that express the characteristic in question for all objects on which the analysis is focused.

A sample corresponds to a subset of the data belonging to the population.

38

Selection of samples

When all elements of the population have equal probability of being sampled we avoid any bias in selection,

such processes are called random sampling

39

The 1936 election: the literary digest poll

- Candidates: Democrat FD Roosevelt and Republican Alfred Landon
- Prediction: Landon to win with 57% of the vote
- Outcome: Landon lost with only 38% of the vote
- Sample Size: **2.3 million** people!
- Literary Digest went bankrupt soon after

40

Why the Digest went wrong:

- **Bias in selection of sample**
 - 10,000,000 questionnaires sent out to
 - Magazine subscribers, car owners, telephone owners
- **Bias from non-response**
 - 20% bothered to reply
 - Presumably, those with strong views about the forthcoming election

Large sample size cannot compensate for poor sample design!!!

41

Data analysis

Resorted to the techniques of descriptive statistics to summarize the information contained in the data

Establishment of inferences about the population

Based on the information contained in the sample the aim is to withdraw conclusions on the population and assign them a degree of credibility

42

In statistical inference, based on analysis of a limited set of data (sample) the goal is to characterize the set from which such data were obtained (population)

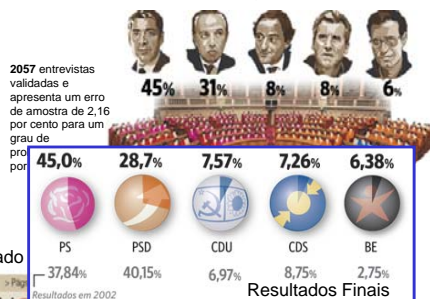
Exemple

From a sample of 100 available balance sheets, drawing conclusions about the behavior of all lawyers customers

43



EXPRESSO-SIC-Renascença -> Eurosondagem



819 entrevistas e apresenta um erro de amostragem para um intervalo de confiança de 95 por cento, de mais ou menos 3,42 por cento.

Independente -> Instituto de Pesquisa de Opinião e Mercado



997 entrevistas validadas e apresenta um erro de amostragem, para um nível de confiança de 95,5 por cento, de mais ou menos 3,1 pontos percentuais.

JN -> Intercampus



1015 entrevistas, e apresenta um erro de amostragem, para um intervalo de confiança de 95 por cento, de mais ou menos 3,1 por cento.

PS: 46% (118-124 deputados)
 PSD: 31% (80-84)
 CDU: 7% (8-12)
 BE: 7% (8-12)
 CDS-PP: 6% (6-10)
 Outros: 1% (0)
 Brancos/nulos: 2%

5051 inquiridos, de 1,4 por cento com um nível de confiança de 95 por cento.

PÚBLICO, RTP e Antena 1 -> Universidade Católica

Testing Hypothesis about proportions (N > 20 e N·p > 7)

Defining the hypothesis

$$H_0: p_A - p_B = p_0$$

$$H_1: p_A - p_B = p_0 \neq p_0, \quad p_A - p_B = p_0 > p_0 \quad \text{ou} \quad p_A - p_B = p_0 < p_0$$

The test is performed using the following statistics

$$ET = \frac{(Y_A/N_A - Y_B/N_B) - p_0}{\sqrt{Y_A \cdot (N_A - Y_A)/N_A^3 + Y_B \cdot (N_B - Y_B)/N_B^3}} \rightarrow N(0, 1)$$

Z 0.10	Z 0.05	Z 0.025	Z 0.01	Z 0.005
1.28	1.645	1.96	2.33	2.575

Confidence Intervals for Comparing Two Proportions

(N > 20 e N·p > 7)

$$\left(\frac{Y_A}{N_A} - \frac{Y_B}{N_B} \right) \pm z(\alpha/2) \cdot \sqrt{\frac{Y_A \cdot (N_A - Y_A)}{N_A^3} + \frac{Y_B \cdot (N_B - Y_B)}{N_B^3}}$$

Example

In the evaluation of a classification problem we used two algorithms. The first algorithm (A) correctly classified 27 of 45 examples while the second algorithm (B) correctly classified 32 of 65 examples. Do you think we can say that algorithm A is significantly more accurate than algorithm B?

$$p_a = \frac{27}{45} = 0.60 \quad e \quad p_b = \frac{32}{65} = 0.49$$

$$ET = \frac{\frac{27}{45} - \frac{32}{65}}{\sqrt{\frac{27(45-27)}{45^3} + \frac{32(65-32)}{65^3}}} = 1.12$$

1.12 < 1.645 thus, the difference is not statistically significant

Basic Statistics

<http://www.liaad.up.pt/~ltorgo/Regression/DataSets.html>
http://www.liaad.up.pt/~ltorgo/Regression/cal_housing.tgz

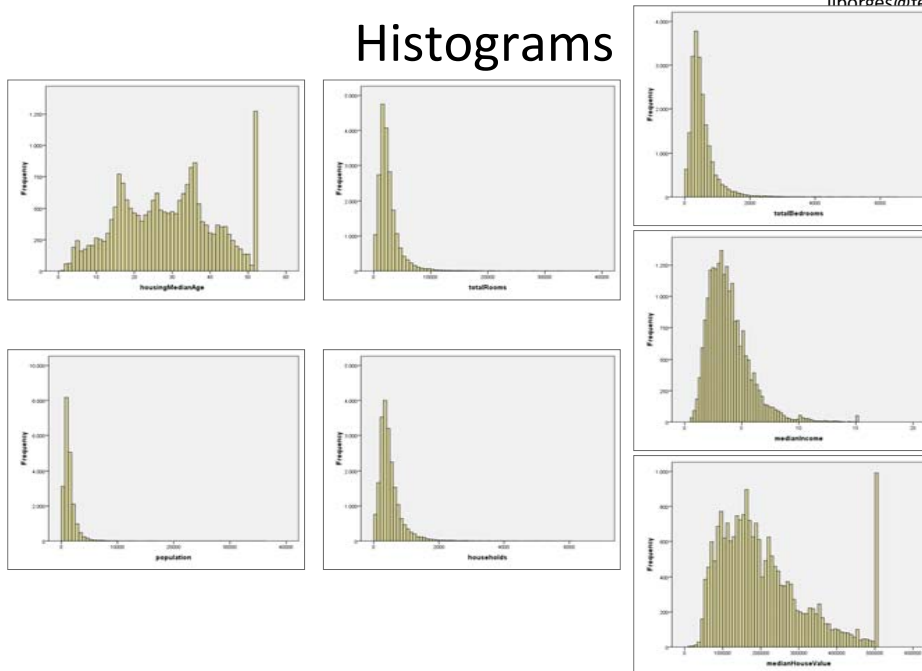
	longitude	latitude	housing Median Age	total Rooms	total Bedrooms	population	households	median Income	median House Value
6	-122.23	37.88	41	880	129	322	126	8.3252	452600
7	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500
8	-122.24	37.85	52	1467	190	496	177	7.2574	352100
9	-122.25	37.85	52	1274	235	558	219	5.6431	341300
10	-122.25	37.85	52	1627	280	565	259	3.8462	342200
11	-122.25	37.85	52	919	213	413	193	4.0368	269700
12	-122.25	37.84	52	2535	489	1094	514	3.6591	299200
20638	-121.45	39.26	15	2319	416	1047	385	3.125	115600
20639	-121.53	39.19	27	2080	412	1082	382	2.5495	98300
20640	-121.56	39.27	28	2332	395	1041	344	3.7125	116800
20641	-121.09	39.48	25	1665	374	845	330	1.5603	78100
20642	-121.21	39.49	18	697	150	356	114	2.5568	77100
20643	-121.22	39.43	17	2254	485	1007	433	1.7	92300
20644	-121.32	39.43	18	1860	409	741	349	1.8672	84700
20645	-121.24	39.37	16	2785	616	1387	530	2.3886	89400

Summary Statistics

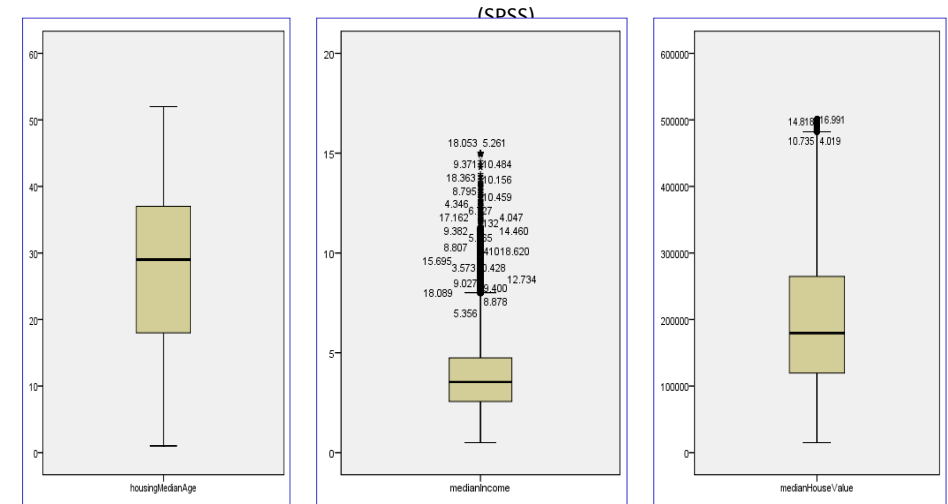
(Excel)

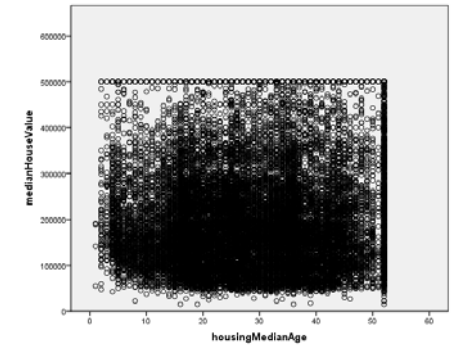
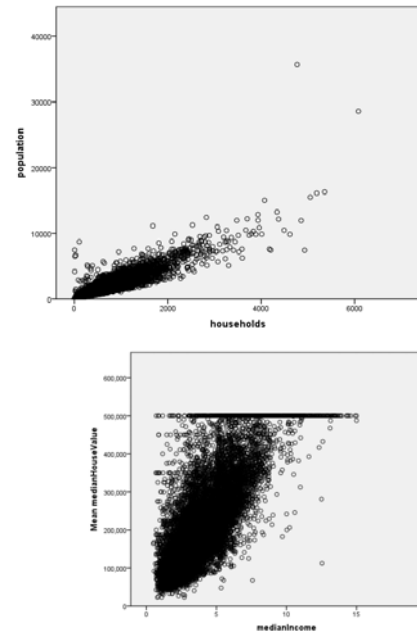
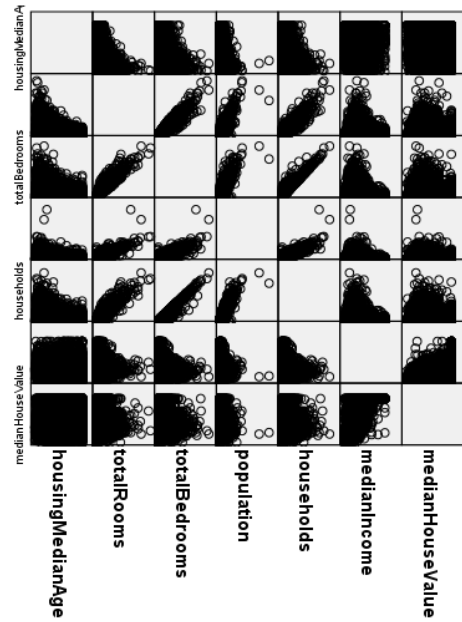
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									

Histograms



Box Plots





A Brief Review of Statistics Concepts