Descriptive Analysis

Project ENABLE

May 28, 2019

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

What is Descriptive Analysis (vs. Prescriptive Analysis)?

- Descriptive analysis is used to describe one or more characteristics of data.
  - It can provide meaningful summaries of data .
  - E.g., Where do the values tend to center? How far are those values from one another or how much are they spread? What is the shape of the set of values? Is it symmetric or not?
- Prescriptive analysis is used to make reasonable inferences about the greater population or selected population.
- The human brain is an amaz
- ing processor. However it comes with its limits.

A Sample Case

- Let's imagine that you are a senior physician at the UNC-health care system or a risk manager at the Blue Cross Blue Shield (BCBS). Let's imagine that you are a senior physician at the UNC-health care system or a risk manager at the Blue Cross Blue Shield (BCBS). The number of asthma patients suddenly increased over last winter, and the number of deaths and related costs were also increased. You will want to identify the characteristics of people who are at high risk for asthma, to reduce misdiagnosis, and to reduce corresponding medical expenditures through applying proper treatments.
- How can you make a sense by using data? Can you read all data instances? Absolutely, you don't want to do that. How can you summarize characteristics of asthma patients?

A Sample Case: Aspirin Exacerbated Respiratory Disease (AERD)

- AERD
  - A medical condition consisting of three key features: **asthma**, respiratory symptoms **exacerbated by aspirin** and other nonsteroidal anti-inflammatory drugs (NSAIDs), and nasal polyps.
  - Symptoms are hypersensitivity reactions to NSAIDs rather than the typical allergic reactions that trigger other common allergen-induced asthma, rhinitis, or hives.
  - The disorder is thought to be caused by an anomaly in the arachidonic acid metabolizing cascade that leads to increased production of pro-inflammatory cysteinyl leukotrienes, a series of chemicals involved in the body's inflammatory response. When medications such as NSAIDs or aspirin block the COX-1 enzyme, production of thromboxane and some anti-inflammatory prostaglandins is decreased. This results in the overproduction of pro-inflammatory leukotrienes, which can cause severe exacerbations of asthma and allergy-like symptoms.
  - Reference: Samter M, Beers RF (1968). Intolerance to aspirin. Clinical studies and consideration of its pathogenesis. *Annals of Internal Medicine*. 68 (5): 975–83. doi:10.7326/0003-4819-68-5-975. PMID 5646829.

Describing Quantitative Data with Frequency Distribution

- According to the case study, you are interested in looking at total medical expenditures of asthma patients.
- In particular, you are interested in the possibility of differences between a group taking aspirin regularly and another group not taking aspirin.
- We draw a random sample of 100 from the 338 records for asthma patients taking aspirin regularly and a random sample of 100 from the 1,148 records for asthma patients not taking aspirin.
- Now you want to examine different characteristics of two groups.

Medical Expenditure for 100 Patients Taking Aspirin

38408	66304	4773	4950	7801	5329	1073	4603	22642	742
38701	39103	4737	9328	29079	660	1360	64414	6859	88717
3293	58649	25443	13408	28402	6152	5554	3834	23988	5751
0	70535	1972	6021	12418	3281	10260	103067	2367	3913
1178	1845	6700	11962	713	71096	978	23447	4018	10120
16195	5903	15349	5118	9053	1865	6386	6343	3723	4763
4796	27423	25604	36657	25315	11510	13561	1785	5429	14706
23382	9949	35933	17673	9793	24967	2240	1327	8048	0
7161	169	2208	1387	3943	4558	8504	6821	95898	7579
38	8095	7265	6269	3958	11690	0	17129	9930	7077

Medical Expenditure for 100 Patients Not Taking Aspirin

12388	10710	99	11078	407	105405	3258	0	4926	3796
1867	3264	1283	6434	2947	8374	1694	394	1877	335
291	1690	276	766	540	1355	8069	2319	11077	187
43957	1963	10040	21484	3313	14865	0	8198	599	135662
269	66238	41	1519	2860	1133	2759	27829	1415	22484
689	10136	3727	427	1174	5340	3395	49555	1661	5789
0	234	4162	4486	74861	1210	95019	127	3295	10867
1935	17489	186	952	386	10101	339	965	9008	1174
0	2724	2461	8348	125	0	544	4920	5598	5220
2755	18856	6725	3007	7282	491	501	4456	1184	15860

Types of Descriptive Analysis

- Univariate statistics
  - Used to describe one variable in data.
- Bivariate statistics:
  - Used to describe the relationships between two variables in data.
- Multivariate statistics:
  - Used to describe the relationships among two or more variables in data.

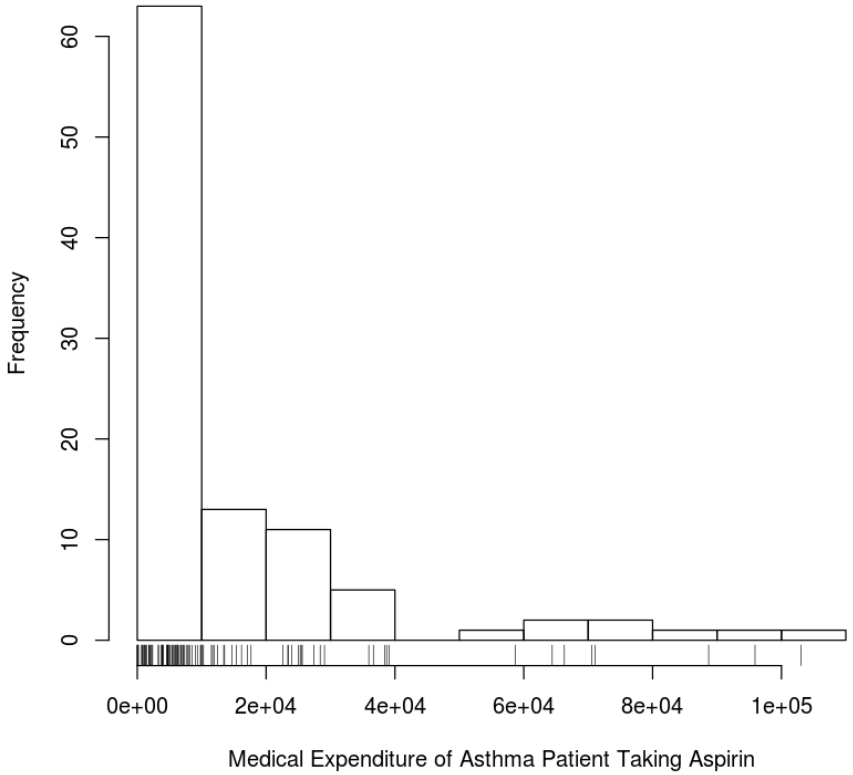
Frequency Distributions (Medical Expenditure for 100 Patients Taking Aspirin)

E	0	38	169	660	713	742	978	1073	1178	1327	1360
f(E)	3	1	1	1	1	1	1	1	1	1	1
E	1387	1785	1845	1865	1972	2208	2240	2367	3281	3293	3723
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	3834	3913	3943	3958	4018	4558	4603	4737	4763	4773	4796
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	4950	5118	5329	5429	5554	5751	5903	6021	6152	6269	6343
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	6386	6700	6821	6859	7077	7161	7265	7579	7801	8048	8095
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	8504	9053	9328	9793	9930	9949	10120	10260	11510	11690	11962
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	12418	13408	13561	14706	15349	16195	17129	17673	22642	23382	23447
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	23988	24967	25315	25443	25604	27423	28402	29079	35933	36657	38408
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	38701	39103	58649	64414	66304	70535	71096	88717	95898	103067	
f(E)	1	1	1	1	1	1	1	1	1	1	

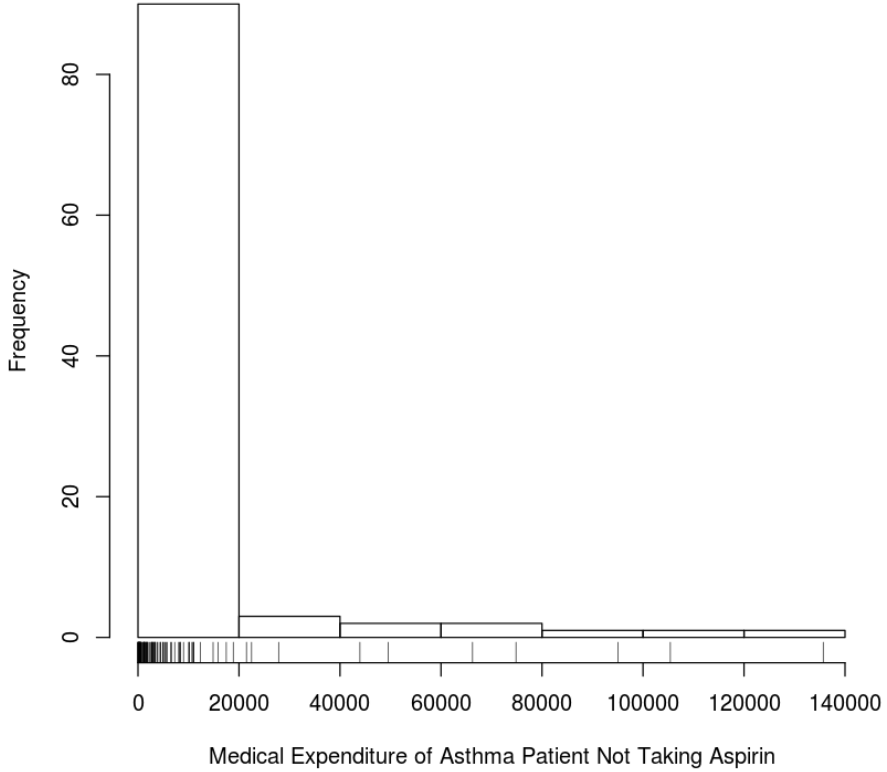
Frequency Distributions (Medical Expenditure for 100 Patients Not Taking Aspirin)

E	0	41	99	125	127	186	187	234	269	276	291
f(E)	5	1	1	1	1	1	1	1	1	1	1
E	335	339	386	394	407	427	491	501	540	544	599
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	689	766	952	965	1133	1174	1184	1210	1283	1355	1415
f(E)	1	1	1	1	1	2	1	1	1	1	1
E	1519	1661	1690	1694	1867	1877	1935	1963	2319	2461	2724
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	2755	2759	2860	2947	3007	3258	3264	3295	3313	3395	3727
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	3796	4162	4456	4486	4920	4926	5220	5340	5598	5789	6434
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	6725	7282	8069	8198	8348	8374	9008	10040	10101	10136	10710
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	10867	11077	11078	12388	14865	15860	17489	18856	21484	22484	27829
f(E)	1	1	1	1	1	1	1	1	1	1	1
E	43957	49555	66238	74861	95019	105405	135662				
f(E)	1	1	1	1	1	1	1				

Histogram (Medical Expenditure for 100 Patients Taking Aspirin)



Histogram (Medical Expenditure for 100 Patients Not Taking Aspirin)



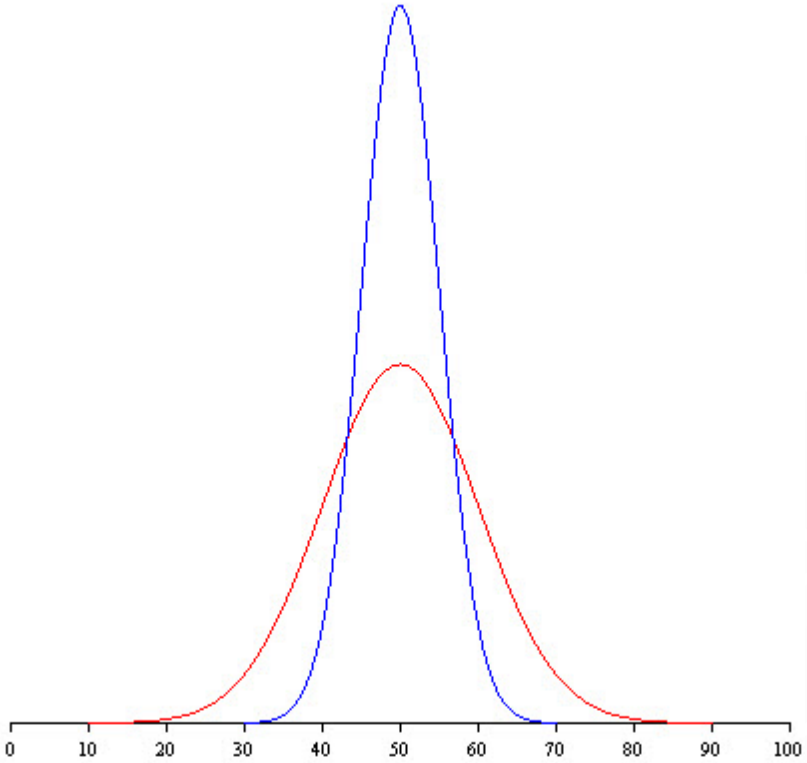
Central Tendency

- Mode: the value that occurs most often.
- Median: the point on a scale that separated a distribution of scores into two groups; one half of the data are above the median and the other half are below the median
- Mean: the arithmetic mean; or the sum of the scores divided by the number of scores

Central Tendency (cont …)

- {1, 2, 2, 3, 4, 7, 9}
- Mode
  - 1, **2**, **2**, 3, 4, 7, 9
- Median
  - 1, 2, 2, **3**, 4, 7, 9
  - {1, 2, 3, 4, 5, 6, 7, 8} ?
  - (4+5) / 2 = 4.5
- Mean
  - (1+2+2+3+4+7+9)/7 = 4

Variability

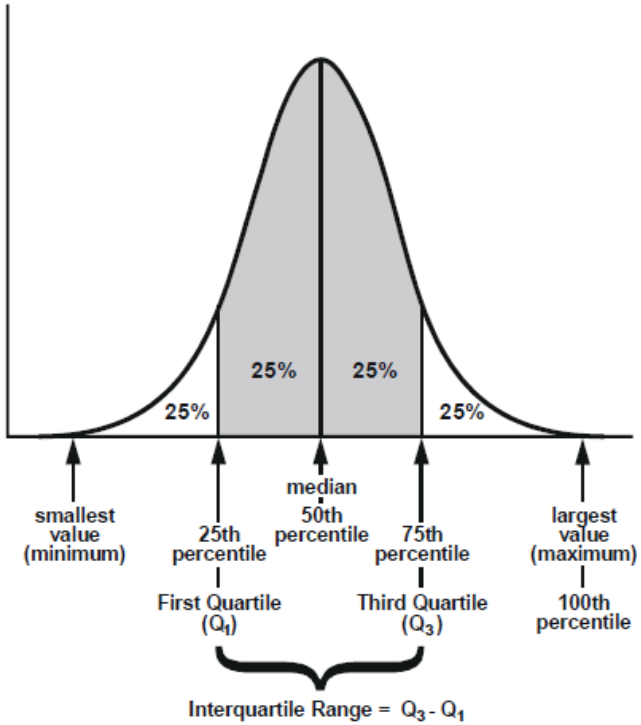


Variability

- Measures of variability represent the degree of the spread of scores.
  - Range: the difference between the largest score and the smallest score
  - Interquartile range: the distance between 75th and 25th percentiles (3rd or 1st quartiles).
  - Variance: the average squared deviation from the mean

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

Interquartile Range



Source: make me analyst

Variability (cont …)

1 2 3 4 5 6 7 8 9 10 11 12 13

- {7, 7, 31, 31, 47, 75, 87, 115, 116, 119, 119, 155, 177}

Q_1 Median Q_3

- Range: 177- 7 = 170
- Interquartile range: 119 – 31 = 88
- Variance:
  - μ = (7+7+31+31+47+75+87+115+116+119+119+155+177)/13 = 83.54
  - σ^2 = 2850.56

Any Questions?

Data Treatment

Next Class

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL