IT3030 - Biostatistics

Week #2 (3/10/2020)

Chapter 3
Numerical Summary Measures

Outline

- 3.1 Measures of ***Central Tendency***
  - 3.1.1. Mean
  - 3.1.2. Median
  - 3.1.3 Mode
- 3.2 Measures of ***Dispersion***
  - 3.2.1. Range
  - 3.2.2. Interquartile Range
  - 3.2.3. Variance and Standard Deviation (STD)
  - 3.2.4. Coefficient of Variation
- 3.3 Grouped Data
- 3.4 Chebychev’s Inequality

Introduction

- In Chapter 2 we learned various ways of ***summarizing*** data into tables, charts or graphs. These help people to better "see" or "visualize" what data are collected, including adequate reasoning of these data.
- They, however, do not allow us to make concise, ***quantitative*** statements that characterize the distribution of values as a whole.

Cont'd

- In this chapter we focus on ***numerical summary measures.***
- Together (techniques introduced in Chapters 2 and 3), these various types of ***descriptive statistics*** can provide a great deal of information about **a set of observations**.

3.1 Central Tendency

- A ***center*** within a set of observations (measurements) usually represents the point about which the observations **tend to *cluster***.
- In a way, center can be interpreted more or less as the ***average value*** that can adequately ***represent*** this group of observations.

3.1.1. Mean

- “Mean” is often ***the arithmetic mean***, or called ***average***, of a collection of numerical values.
- It is the ***summation*** of all divided by the ***count***.
- It is apparent that a mean is applicable in discrete and continuous data, but is generally not adequate for either nominal or ordinal data.

TABLE 3.1

Forced expiratory volumes in 1 second for 13 adolescents suffering from asthma

Subject	FEV_1 (liters)
1	2.30
2	2.15
3	3.50
4	2.60
5	2.75
6	2.82
7	4.05
8	2.25
9	2.68
10	3.00
11	4.02
12	2.85
13	3.38

FEV_1 : This is the amount of air that you can forcibly blow out in one second, measured in liters, and is considered one of the primary indicators of lung function.

- One can easily compute the mean value for these 13 observations.

TABLE 3.1
Forced expiratory volumes in 1 second for 13 adolescents suffering from asthma

Subject	FEV_1 (liters)
1	2.30
2	2.15
3	3.50
4	2.60
5	2.75
6	2.82
7	4.05
8	2.25
9	2.68
10	3.00
11	→ **40.2** ~~4.02~~
12	2.85
13	3.38

- Assuming that the value for subject #11 was ***mistakenly*** recorded as 40.2.
- One can also compute the mean value for “these” 13 values.
- How are the two mean values different?

Comments

- What happens if one value is ***very different*** from the others? (This might be true, or might be a mistake.) Should we use it or discard it?
- How “different” is “very” different?
- For any statistical analysis, it should be taken into consideration how the magnitude of every observation in a set of data is ***distributed***.

Cont’d

- ***A mean value is very sensitive to every observation***. (Can you see this?)
- In this case, we may want to use a summary measure ***that is NOT as sensitive to every observation (or more resistant to errors).***

3.1.2. Median

- **A *median* is defined as the *50th* percentile of a set of measurement.**
- If the list of values is ranked from smallest to largest, half of the values are greater than or equal to the median, whereas the other half are less or equal to it.
- In short, it is the "**middle value**" for a set of ***ranked*** values.

Cont’d

- ☑ Can you find the median from both the “true” Table 3.1 and the “contaminated” Table 3.1?
- ☑ Can you see that median is “less sensitive” than the mean from this example?

TABLE 3.1
Forced expiratory volumes in 1 second for 13 adolescents suffering from asthma

Subject	FEV_1 (liters)
1	2.30
2	2.15
3	3.50
4	2.60
5	2.75
6	2.82
7	4.05
8	2.25
9	2.68
10	3.00
11	40.2 ~~4.02~~
12	2.85
13	3.38

How to compute mean and median?

- From hand calculation.
- From MS-Excel or other spreadsheet applications.
- From MATLAB. (Google for something like "MATLAB mean" or "MATLAB median"…)
- Other tools that you know? (For example, some scientific hand calculators)
- Write your own computer program for doing so?

3.1.3. Mode

- The mode is **a set of values** (not necessarily a single value) in the observation that ***occurs most frequently***.
- Alternatively, we may say a mode is a set of values that ***dominate***” the observation.
- Whether a mode exists, or what should be chosen as a mode for the entire set of observations pretty much depends on the distribution itself.

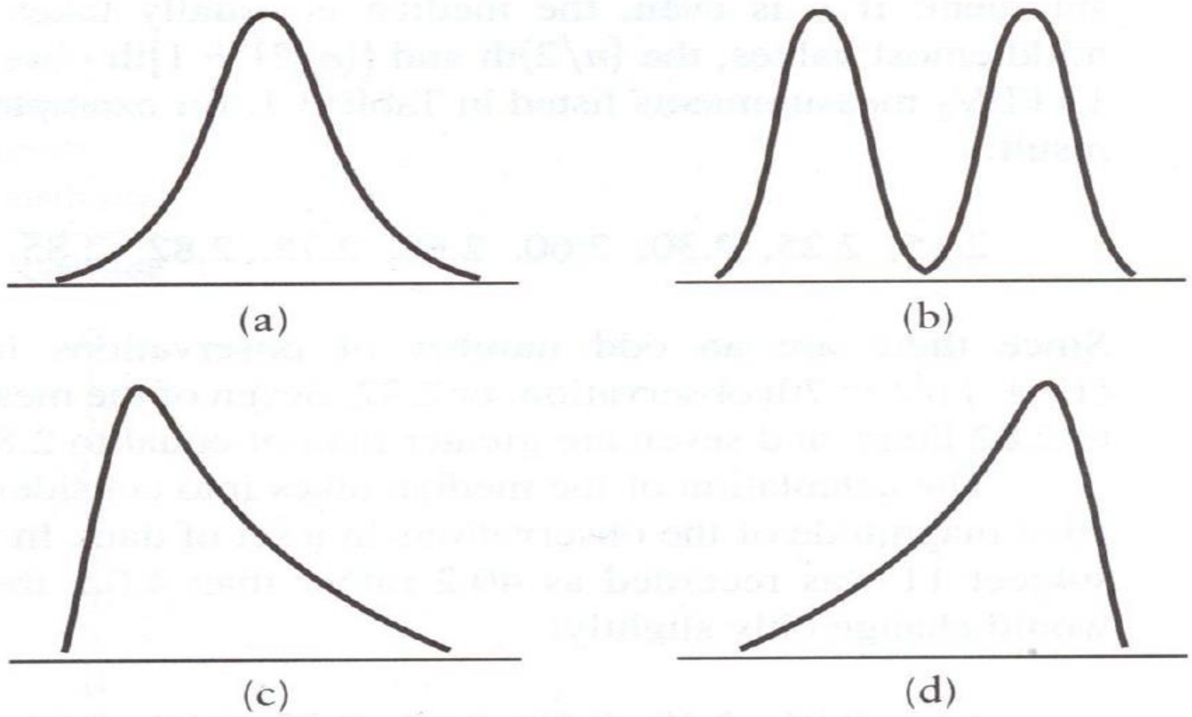
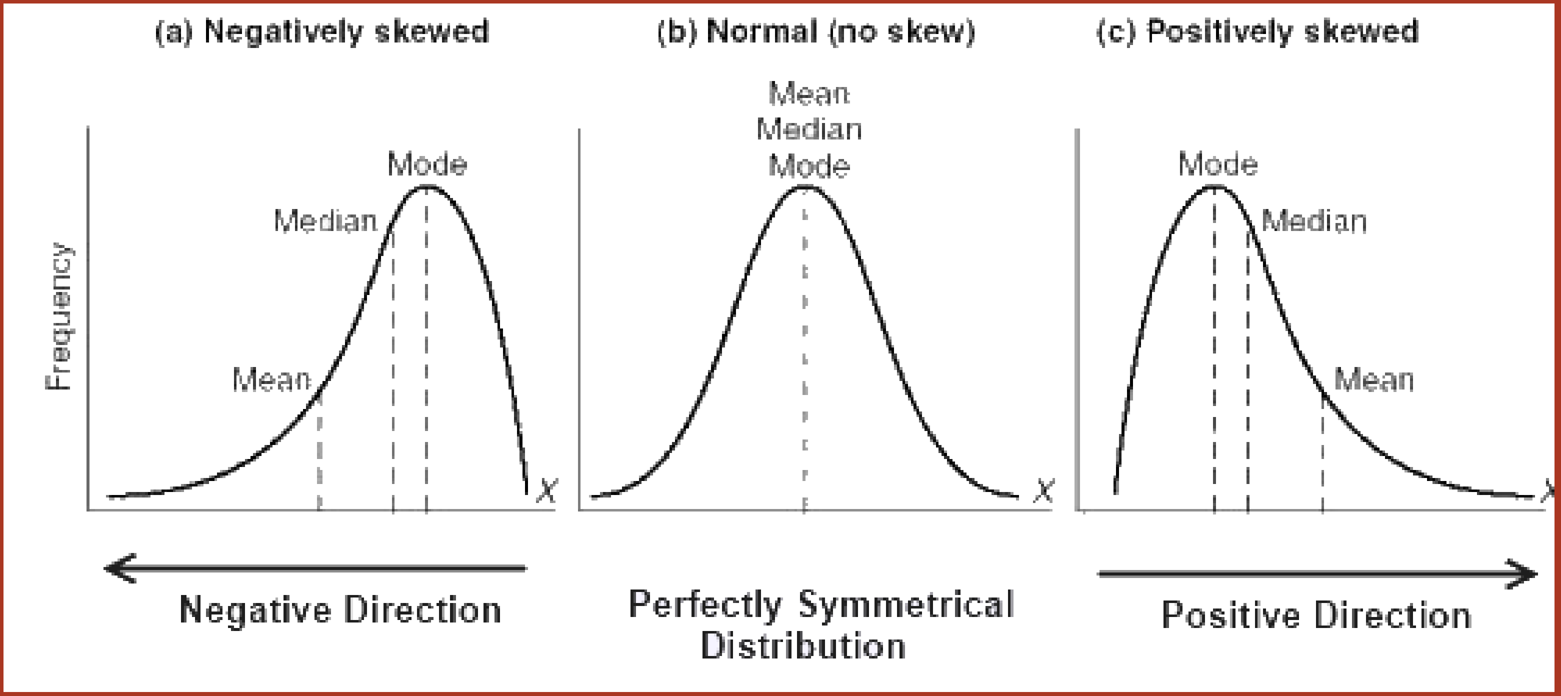


FIGURE 3.1
Possible distributions of data values

(a) **Unimodal** - single peak and symmetric. The mean, median and mode should all be roughly the same.

(b) (b) **Bimodal** – two symmetric peaks. Mean and median will be roughly the same, which is possibly located in the. **This is mathematically correct, but definitely not a "dominant" value.**

(c) (c)&(d) Data is ***skewed***. **A median is better to measure the central tendency rather than a mean or mode.**

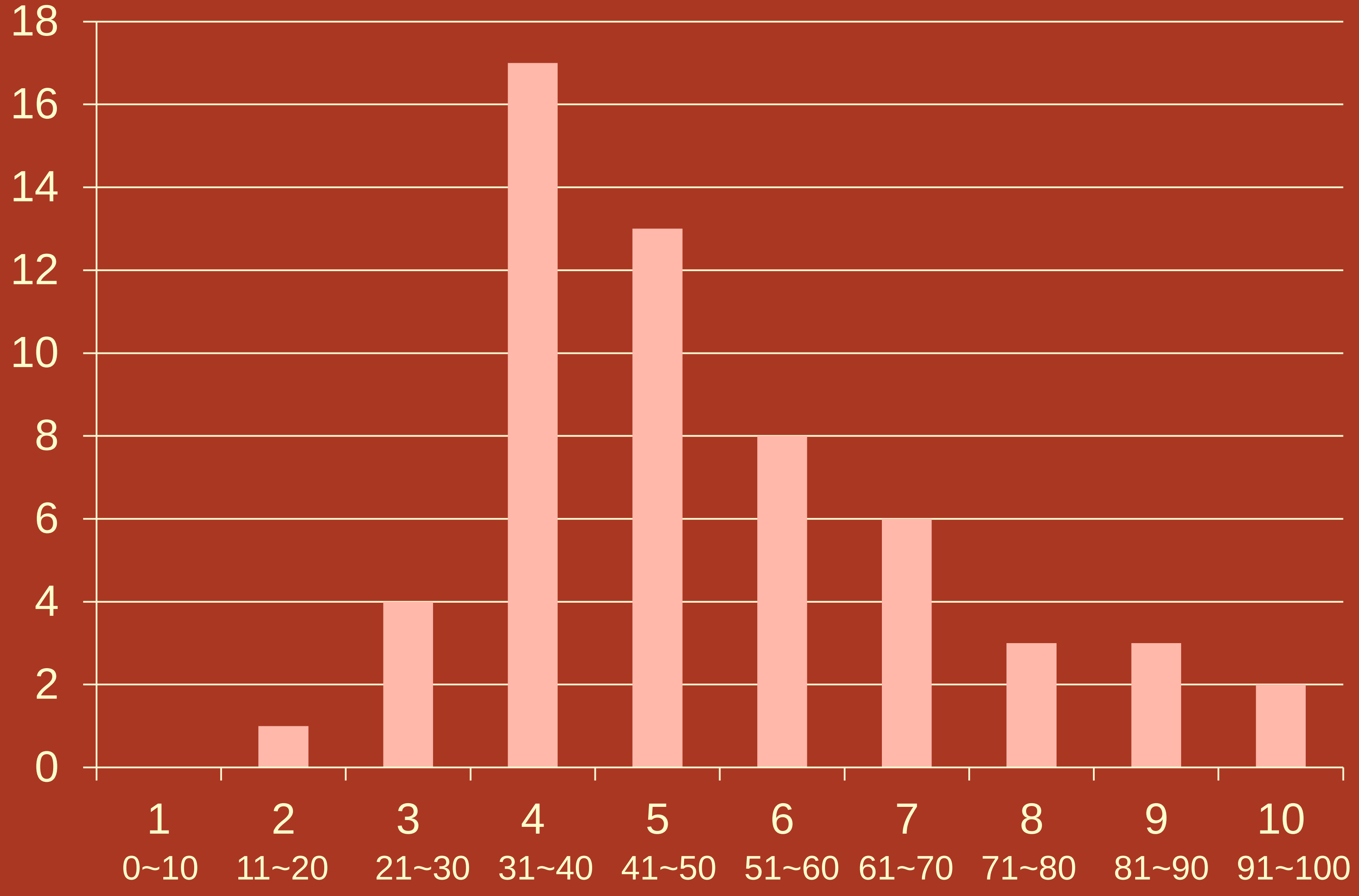


- What is x-axis? What is y-axis?
- What are the heights?
- Is a mode (dominant value) occurring on the x- or y-axis?

Example : 57 scores listed below (ranked from top to bottom and left to right) with ***mean = 48.98 and median = 43***.

19	35	38	45	59	74
25	36	39	45	59	74
26	36	41	46	59	85
28	36	41	48	62	87
28	36	41	49	63	89
33	37	43	50	63	91
34	37	43	51	63	92
34	37	43	51	65	
34	37	**43**	53	65	
35	37	44	58	70	

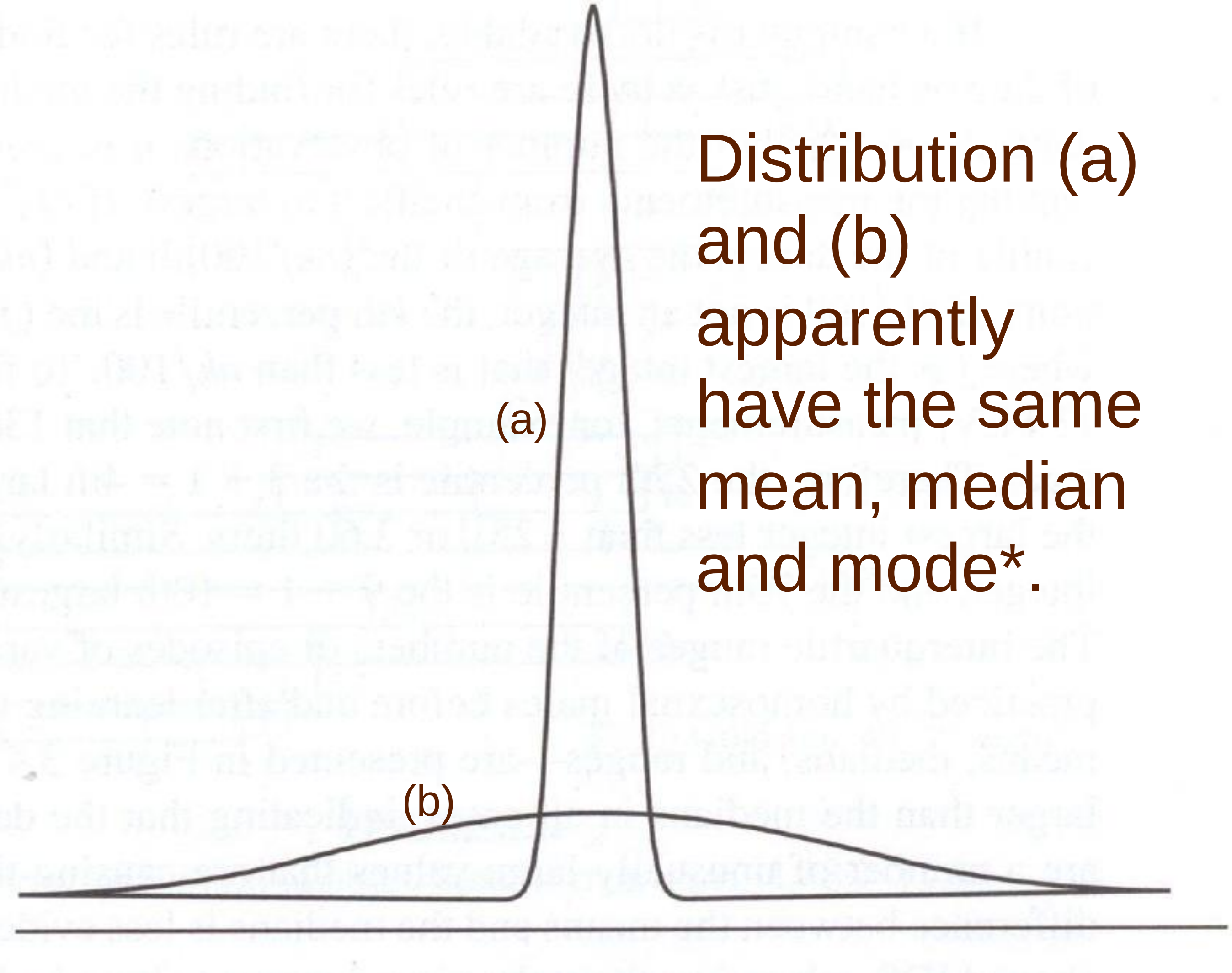
Category	frequency (heads)
0~10 (1)	0
11~20(2)	1
21~30(3)	4
31~40(4)	**17**
41~50(5)	13
51~60(6)	8
61~70(7)	6
71~80(8)	3
81~90(9)	3
91~100(10)	2



Since the distribution is slightly skewed to the left, median=43 is apparently closer to the mode (between Category ‘4’, or 31~40, to Category ‘5’, or 41~50) than to the mean=48.98.

Comments

- It must be careful not to wrongfully interpret that the measure of central tendency (previously mentioned mean, median or mode) is the representative of ALL observations in the group.
- A rule of thumb is to seriously consider what this “center” means when there are considerable variance (not normal distribution) involved.
- **Remember that when we summarize the center of a set of data, information is always lost.**



Distribution (a) and (b) apparently have the same mean, median and mode*.

FIGURE 3.2
Two distributions with identical means, medians, and modes

*Recall that a mode is the peak **location** (on the horizontal axis), not the height of this peak on the vertical axis.

Comments

- To know how good our measurement of central tendency actually is, we need to have some idea about the **variation** among the measurements.
- Do all the observations tend to be quite similar and therefore lie close to the center (graph (a) from previous slide), or are they spread out across a broad range of values (graph (b) from previous slide)?

Cont'd

- We have discussed about how to determine the measures of central tendency, with the purpose of knowing what our observations are **centered to**.
- That is, what value(s) are most appropriate to represent the whole collection of observations.

Cont'd

- Following these, we now proceed to investigate **the overall distribution** of these observations, to better let people understand the quantitative features in addition to the center values (such as mean, median, etc.)
- ***Range, percentile, interquartile range, variance, etc.***

3.2.1 - Range

- A **range** is a number that describes how “wide” our collections is.
- This is basically the difference between the largest observation and the smallest (or the other way around).
- This “range”, however, considers ALL data including the **extreme values** rather than the ***majority*** of the observations.

3.2.2 – Interquartile Range

- Recall that Q1 is the 25th percentile, and Q3 the 75th percentile. All data are ***sorted from smallest to largest***.
- ***Interquartile range*** is defined as Q3–Q1, which encompasses the middle 50% (Q2, or the median).
- This is more or less like an "effective" or "major" range, since it excludes both 25% of extremes.
- In fact, this is the height (or width) of the box we have seen in a box plot.

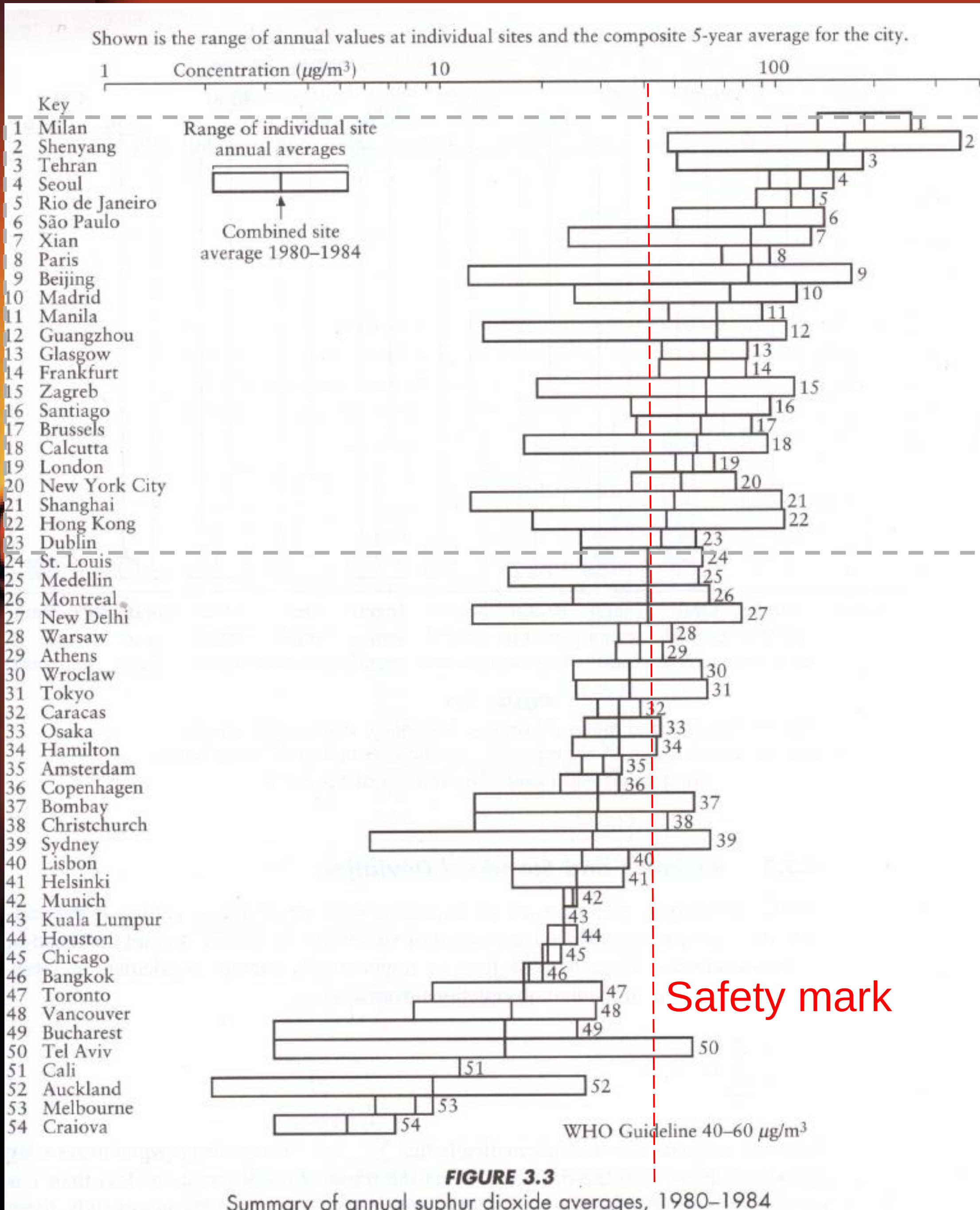
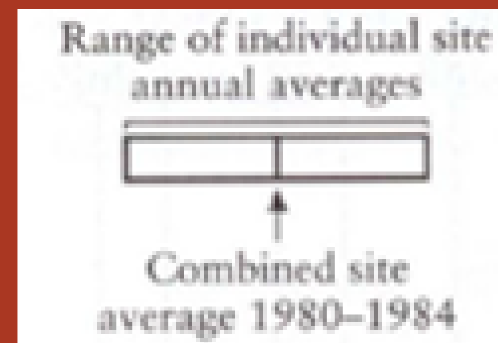


FIGURE 3.3
Summary of annual suphur dioxide averages, 1980–1984

☑ What information can you get from organizing these measurements like this?

☑ What cities are considered safe?

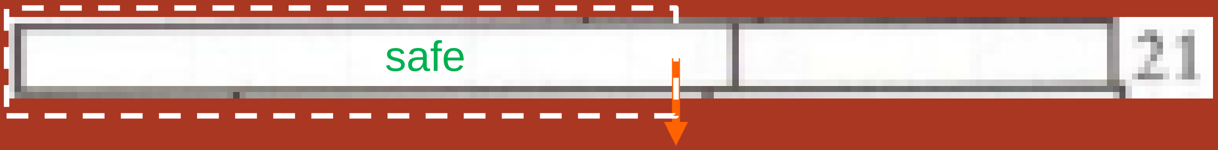


Range (not boxplot) of annual sulphur dioxide (二氧化硫) average.

Comments to Fig. 3.3

- The ***averaged*** SO_2 are still above the lower safety mark for 23rd Dublin and above.
- Although the average for 31st Tokyo is below the mark and considered safe, still a good portion of these readings are polluted.

- While the average of 21st Shanghai is hazardous, most of the readings are below the safety mark (less polluted).



- A few cities, such as Kuala Lumpur, has very limited range.

Excel 說明

搜尋

QUARTILE 函數

傳回一個資料組的四分位數。四分位數通常用於把銷售和市場調查資料從母群體內加以分類。例如，可用 QUARTILE 來找出一個母群體中前 25% 的收入。

＋全部顯示

語法

quartile(a1..a10,**3**)

QUARTILE(array,quart)

Array 是要求得四分位數的一個數值陣列或儲存格範圍。

Quart 指出要傳回的數值。

如果 QUART 等於	QUARTILE 傳回值
0	最小值
1	第一四分位數 (第百分之二十五)
2	中間值 (第百分之五十)
3	第三四分位數 (第百分之七十五)
4	最大值

所有 Excel

已連線至 Office Online

percentile(a1..a10,**0.3**)

3.2.3 – Variance and Standard Deviation (STD)

- Variance is a common way of quantifying the amount of variability, or spread, around the mean of the measurements.
- It can be intuitively represented by the following formula, where $\bar{x}$ is the mean value, and x_i the individual measurement for *n* observations.

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

Cont'd

- In short, variance represented this way is the average distance of the individual observations from its mean value.
- **Can you see, however, this formula would always give zero value?**

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

Cont'd

- Although we can conveniently use the absolute value for these distances, a more widely used procedure is to square the deviations from the mean, and then find the average of the squared distances.
- This summary measure is **the *variance* of the observations s^2**.(See next slide.)

Cont'd

- The commonly known ***standard deviation* (STD)**, now, is the value s itself (or the positive *square root* of the variance).

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

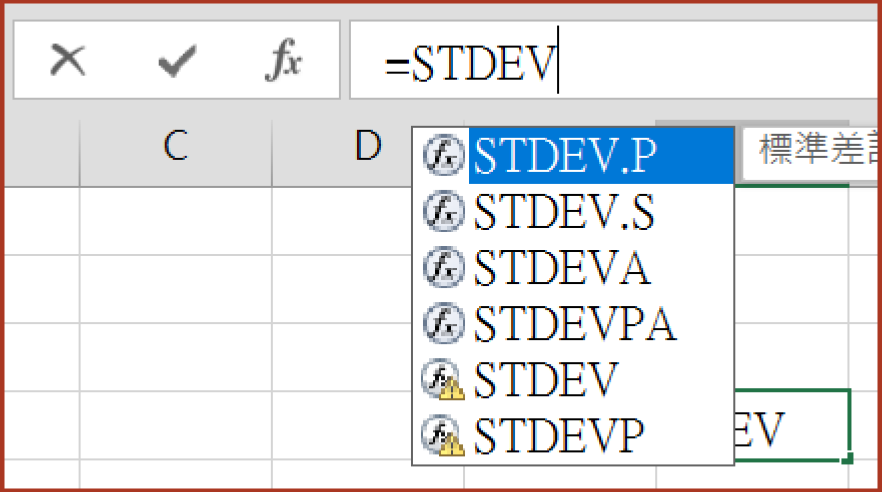
This is not the number of observations *n*. (Losing one degree of freedom by estimating the sample mean $\bar{x}$)

Cont'd

- It can be seen that *STD has the same units of measurements as the mean*.
- As a result, it is more used than the variance itself (squared units).

Using Excel

- The variance described here can be computed using Excel's function ***VAR***.
- The standard deviation can be computed using Excel's function ***STDEV***.



Excel 2016 說明

搜尋

- STDEV.P 的計算公式是：

$$\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

其中，x 為樣本平均數 AVERAGE (number1,number2,...)，而 n 為樣本大小。

Excel 2016 說明

STDEV.S

- STDEV.S 的計算公式是：

$$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

其中，x 為樣本平均數 AVERAGE (number1,number2,...)，而 n 為樣本大小。

P: Population. S: Sample.

Using Excel 2016 (left) and MATLAB (right)

1	AVERAGE	5.5
2		
3	VAR	9.166667
4	**STDEV**	**3.02765**
5		
6	VAR.P	8.25
7	STDEV.P	2.872281
8		
9	VAR.S	9.166667
10	**STDEV.S**	**3.02765**

It is clear that both Excel's STDEV and MATLAB's std use the sample (S) rather than the population (P).

```
>> a=1:10
a =
     1     2     3     4     5     6     7
8     9    10
>> mean(a)
ans = 5.5000
>> var(a)
ans = 9.1667
>> std(a)
ans =    3.0277
>>
```

3.3.1 – Grouped Mean

- We can compute a mean value by summarizing all values and divided by the count.
- Or we can ***group them into adequate partitions*** and compute the sum in each group individually, followed by the division of the total sum by the total count.

TABLE 3.3

Duration of transfusion therapy for ten patients with sickle cell disease

Subject	Duration (years)
1	12
2	11
3	12
4	6
5	11
6	11
7	8
8	5
9	5
10	5

3(5)+1(6)+1(8)+3(11)+2(12)=86

Compute the mean value by the following two approaches:

☑ Sum all 10 measurements and divided by the count (which is apparently 10).

☑ Group the measurements into partitions according to the value. For example, subjects 8, 9 and 10 are grouped together because they all have a value of 5 years. Compute the sum of values and sum of counts for each group, then use them to compute the mean value of the complete data set.

Cont'd

- Certainly there is no reason that the two answers you have obtained from the previous exercise in computing the averaged duration would be different.
- The technique of grouping measurements that have equal values before calculating the mean **has one distinct advantage** over the standard method: this procedure can be applied to **data that have been summarized in the form of a frequency distribution**.

TABLE 3.4
Absolute frequencies of serum cholesterol levels for U.S. males, aged 25 to 34 years, 1976–1980

Cholesterol Level (mg/100 ml)	Number of Men
80–119	13
120–159	150
160–199	442
200–239	299
240–279	115
280–319	34
320–359	9
360–399	5
Total	1067

Grouped mean is the mean obtained from a grouped data set as seen here.

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i},$$

Here *k* is the count for group, that is 8 in this case, m_i is the midpoint of the i^{th} interval, and f_i is the frequency associated with the i^{th} interval. For example, when i=1, we have m_1 =(80+119)/2=99.5 and f_1 =13.

According to this formula, we may have the grouped mean as 198.8 mg/100 ml in this case. (Verify this by yourself.)

Comments

- The grouped mean computed earlier is actually **a weighted average of the interval midpoints**; each mid-point is weighted by the **frequency** of observations within the interval.
- This, of course, would likely to be less accurate than computing the average by dividing the grand total by the count of 1,067 US males.

3.3.2 – Grouped Variance

- Likewise we may compute a grouped variance based on the same concept in computing for grouped mean, as well as the definition for the standard variation s^2 we seen earlier. The formula is given below.

$$s^2 = \frac{\sum_{i=1}^{k} (m_i - \bar{x})^2 f_i}{\left[\sum_{i=1}^{k} f_i\right] - 1},$$

- Accordingly, the grouped variance s^2 for Table 3.4 can be computed as 1930.9, or s = 43.9 mg/100 ml.
- Together with the grouped mean computer earlier, we may interpret Table 3.4 as **"The averaged serum cholesterol level for US males, aged 25 to 34 years, 1976-1980, is 198.8±43.9 mg/100 ml."**.

3.4 Chebychev's Inequality

- Observing that the previous result we had regarding the serum cholesterol level, 198.8±43.9, does not imply that the lowest cholesterol level is 198.8–43.9=154.9 and highest is 198.8+43.9=242.7. (From Table 3.4 you can surely see that there are extreme values beyond that.)

Cont'd

- We *usually* estimate 2*STD from the mean is roughly OK for most of the data, assuming that the data distribution is "OK".
- Of course, if we knew something about the actual distribution, we can do better estimation than that.

Cont'd

- If the data are symmetric and unimodal, we can say that approximately **67%** of the observations are within **±1*STD**, **95%** are within **±2*STD**, and **99%** are within **±3*STD**.

Cont'd

- If, however, we know nothing about the distribution in advance, we may use ***Chebychev's inequality*** to have a ***conservative*** estimate of that.
- **For any number *k* that is greater than 1, at least [1-(1/*k*)^2] measurements lie within *k* STDs of their mean.**

$$1-\left(\frac{1}{2}\right)^2=\frac{3}{4}$$

For *k* = 2. This value of ¾ means that the interval of 2*STD from the mean value encompasses at least 75% of the observation in the group. (Recall for a normal distribution we have 95%.)

$$1-\left(\frac{1}{3}\right)^2=\frac{8}{9}$$

For *k* = 3. This value of 8/9 means that the interval of 3*STD from the mean value encompasses at least 88.9% of the observation in the group. (Recall for a normal distribution we have 99%.)