lecture2: descriptive statistics

20 Şubat 2025 Perşembe 10:39

data types:

1- qualitative data: nonnumerical values·
2- quantitative data: obtained by measurement or counting, numerical values·
   2·1- measure of location
   2·2- measures of dispersion

MEASURES OF LOCATION:

arithmetic mean: sum of the data values divided by number of observation· if the data set is from a sample, then the sample mean is as observed values over sample size·

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

the disadvantage of arithmetic mean is that it is affected by the extreme values (aka outliers)

sum of deviations of each value from the mean is zero 0· This means the mean is the balancing point of the data·

MEDIAN:

in an ordered list, median is the middle number, half the data is below the median, half is above·

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

Example Facebook is a popular social networking website. Users can add friends and send them messages, and update their personal profiles to notify friends about themselves and their activities. A sample of 10 adults revealed they spent the following number of hours last month using Facebook.

3	5	7	5	9	1	3	9	17	10

Find the median number of hours.

Solution Note that the number of adults sampled is even (10). The first step, as before, is to order the hours using Facebook from low to high. Then identify the two middle times. The arithmetic mean of the two middle observations gives us the median hours. Arranging the values from low to high:

1	3	3	5	5	7	9	9	10	17

The median is found by averaging the two middle values. The middle values are 5 hours and 7 hours, and the mean of these two values is 6. We conclude that the typical Facebook user spends 6 hours per month at the website. Notice that the median is not one of the values. Also, half of the times are below the median and half are above it.

QUARTILES:

split the ranked data into 4 segments with an equal number of values per segment·

positions of the quartiles;

$$\boxed{\begin{aligned} Q1 &= 0{\cdot}25(n+1) \\ Q2 &= 0{\cdot}50(n+1) \\ Q3 &= 0{\cdot}75(n+1) \end{aligned}}$$

(n is the number of observed values)

Example: Find the first quartile

Sample Ranked Data: 11 12 13 16 16 17 18 21 22

x_2 x_3 x_7 x_8 7.5th position

(n = 9)
Q1 = is in the 0.25(9+1) = 2.5 position of the ranked data
so use the value half way between the 2nd and 3rd values,
so Q1 = 12.5

Q_1 = the value at 2.5th position $= \frac{12+13}{2} = 12.5$

Q_3 = the value at 7.5th position $= \frac{18+21}{2} = 19.5$

including 25;

$0.25(10+1) = 2.75$

Q_1 = the value at 2.75th position

$x_2 \to 12$ } 0.75
$x_3 \to 13$ } 0.25

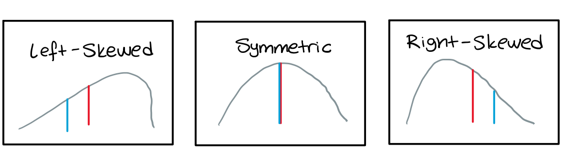
$Q_1 = 12 + 0.75(13-12) = \underline{\underline{12.75}}$

$Q_3 = 8.25^{th} \rightsquigarrow 21 - 22 : 21.25$

MODE:

measure of central tendency·
value that occurs most often·
not affected by extreme values·
there may be no mode or several modes·

DISTRIBUTION SHAPES:



Mean < Median (negative skewed) | Median = Mean | Median > Mean (positive skewed)

VARIABILITY:

indicates how spread out the scores are·
large differences among scores=lot of variability
high variability=low predictability

range: greater the spread of data from the center of distribution, larger the range will be·
range could be disadvantageous if there are outliers and ignores the way in which data are distributed·

INTERQUARTILE RANGE: (IQR)

IQR= Q3-Q1

difference between third and firdt quartile· eliminates high and low valued observations·

VARIANCE:

The **sample variance**, denoted by s^2, is given by

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

The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2}.$$

Population variance:

$$\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$$

- The population variance is calculated with N, the population size. Why isn't the sample variance calculated with n, the sample size?
- The true variance is based on data deviations from the true mean, μ.
- The sample calculation is based on the data deviations from $\bar{x}$, not μ.
- $\bar{x}$ is an estimator of μ; close but not the same.
- So the $n-1$ divisor is used to compensate for the error in the mean estimation.

- When the sample variance is calculated with the quantity $n-1$ in the denominator, the quantity $n-1$ is called the **degrees of freedom**
- Origin of term:
  - There are n deviations from the $\bar{x}$ in the sample
  - The sum of the deviations is zero
  - $n-1$ of the observations can be freely determined but the n^{th} observation is fixed to maintain the zero sum

each value in the data set is used in the call· Values far from the mean are given extra weight (because deviations from the mean are squared·

Example 1.4: In an example discussed extensively in Chapter 10, an engineer is interested in testing the "bias" in a pH meter. Data are collected on the meter by measuring the pH of a neutral substance (pH = 7.0). A sample of size 10 is taken, with results given by

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08.

The sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \cdots + 7.08}{10} = 7.0250.$$

The sample variance s^2 is given by

$$s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 + \cdots + (7.08 - 7.025)^2] = 0.001939.$$

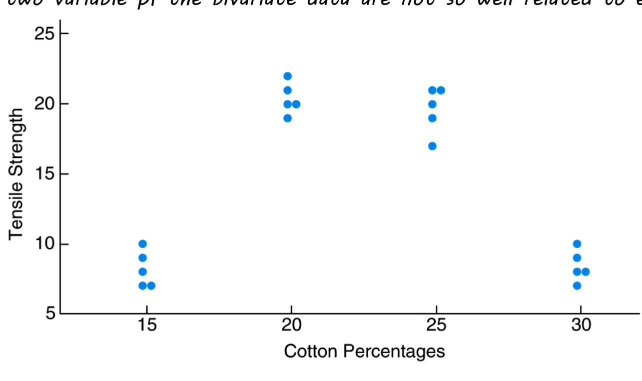
As a result, the sample standard deviation is given by

$$s = \sqrt{0.001939} = 0.044.$$

So the sample standard deviation is 0.0440 with $n - 1 = 9$ degrees of freedom.

SCATTERPLOT:

bivariate: data for items consisting of a pair of values·
graphical summary for bivariate data is scatterplot·
if dots on the scatterplot are spread out in random scatter, this means the two variable pf the bivariate data are not so well related to each other·



Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

STEM AND LEAF PLOT:

stem: leftmost one or two digits·
leaf: the next digit·

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Stem	Leaf	Frequency
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

Frequency → Number of observations

Question: how can you create stem and leaf for the data set?

data set: 423 , 312 , 322 , 125 , 100, 125

Stem	Leaf	Frequency
1·	69	2
2⋆	2	1
2·	5669	4
3⋆	001111222333444	15
3·	5567778899	10
4⋆	11234	5
4·	577	3

double stem and leaf

the more the number of stems, higher the accuracy is

FREQUENCY DISTRIBUTION

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

data:

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Sort raw data in ascending order:
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Find range: 58 - 12 = 46

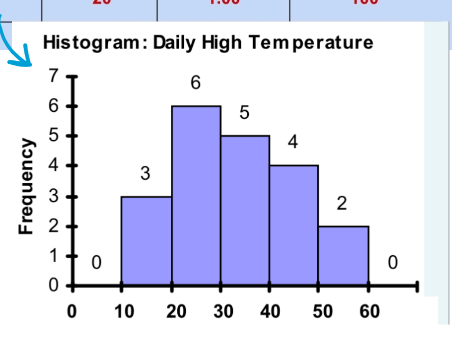
Select number of classes: 5 (usually between 5 and 15)

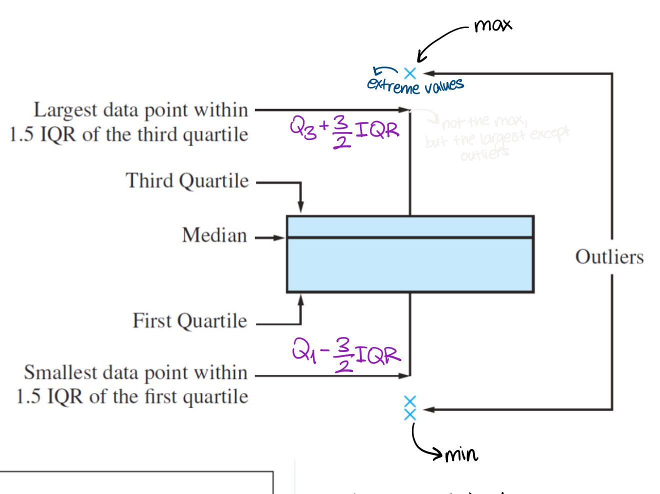
Compute interval width: 10 (46/5 then round up)

Determine interval boundaries: 10 but less than 20, 20 but less than 30, . . . , 60 but less than 70

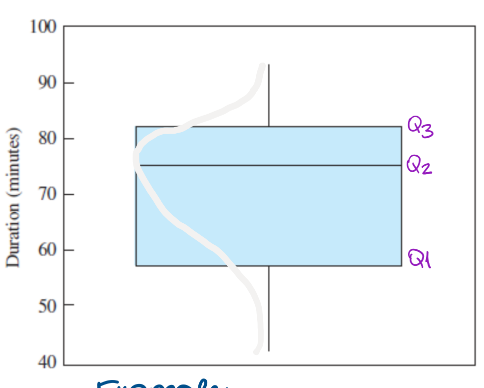
Count observations & assign to classes

Interval	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

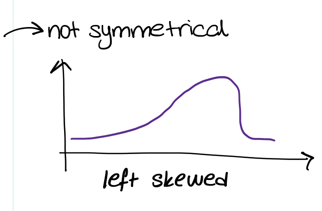




$IQR = Q_3 - Q_1$

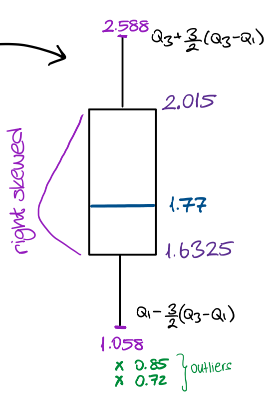


→ not symmetrical



Example:

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69



min = 0.72

max = 2.55

Position of $Q_1 = 0.25(n+1) = 0.25(41) = \underline{10.25}$

$x_{10} = 1.63$
$x_{11} = 1.64$

$x_{10.25} = x_{10} + 0.25(x_{11} - x_{10})$
$= 1.63 + 0.25(1.64 - 1.63) = \underline{1.6325}$

Position of $Q_2 = 0.5(n+1) = 0.5(41) = 20.5$

$x_{20} = 1.75$
$x_{21} = 1.79$

$x_{20.5} = x_{20} + 0.5(x_{21} - x_{20})$
$= 1.77$

↳ same procedure $Q_3 = 2.015$

$x_{30} + 0.75(x_{31} - x_{30})$

$1.97 + 0.75(2.03 - 1.97)$

Probability Plots

- A **probability plot** is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data
- To construct a probability plot:
  - Rank the data observations in the sample from smallest to largest: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
  - The observed value $x_{(j)}$ is plotted against the observed cumulative frequency $(j - 0.5)/n$.
  - The paired numbers are plotted on the probability paper of the proposed distribution.
- If the plotted points deviate a straight line, then the hypothesized distribution adequately describes the data.

The effective service life (X_j in minutes) of batteries used in a laptop are given in the table.

We hypothesize that battery life is adequately modeled by a normal distribution.

To test this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies and plot them.

TABLE 6.6 Calculation for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j-0.5)/10$	z_j
1	176	0.05	−1.64
2	183	0.15	−1.04
3	185	0.25	−0.67
4	190	0.35	−0.39
5	191	0.45	−0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

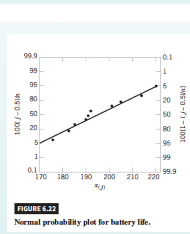


FIGURE 6.22 Normal probability plot for battery life.

- Can be constructed on ordinary axes by plotting the standardized normal scores z_j against $x(j)$, where the standardized normal scores satisfy

$$\frac{j - 0.5}{n} = P(Z \le z_j) = \Phi(z_j)$$

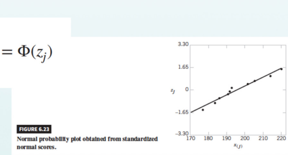


FIGURE 6.23 Normal probability plot obtained from standardized normal scores.

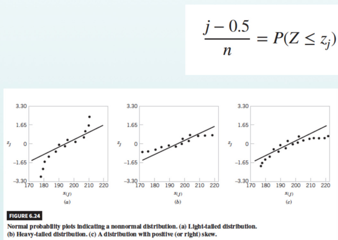


FIGURE 6.24 Normal probability plots indicating a nonnormal distribution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.