1.1 Overview: Statistical Inference, Samples, Populations, and Experimental Design

- -Use of Scientific Data
- -Variability in Scientific Data
- -The Role of Probability

43.0
43.9
43.1

قراءة قيم مختلفة
لكن جميعها متقاربة
لأخذ قرار معين صحيح

إضافة نيتروجين لشجر البلوط

Example 1.2

Often the nature of the scientific study will dictate the role that probability and deductive reasoning play in statistical inference. Exercise 9.40 on page 297 provides data associated with a study conducted at the Virginia Polytechnic Institute and State University on the development, of a relationship between the roots of trees and the action of a fungus. Minerals are transferred from the fungus to the trees and sugars from the trees to the fungus. Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen. All other environmental conditions are held constant. All seedlings contain the fungus *Pisolithus tinctorus*. More details are supplied in Chapter 9. The stem weights in grams were recorded after the end of 140 days. The data are given in Table 1.1.

جذور الشجرة

بيوت زجاجية

urpose of this experiment : داخل بيوت زجاجية وهنا ودرسنا تأثير الـ N

f the use of nitrogen has an influence on the growth of the roots.

Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26 — أقل
أعلى — 0.53	0.43
أقل — 0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86 — أعلى
0.38	0.62
0.43	0.46
X	O

(0.28 - 0.53) (0.26 - 0.86)

وزن النباتات بعد إضافة الـ N

بندرس تغير الـ N وإضافة مع تثبيت باقي العوامل

one factor → نغيره وباقي العوامل نثبته

في المثال وضع في بيوت زجاجية

إضافة الـ N يحسن جودة نبات البلوط

* الهدف تقليل التكلفة وتحسين شجر البلوط

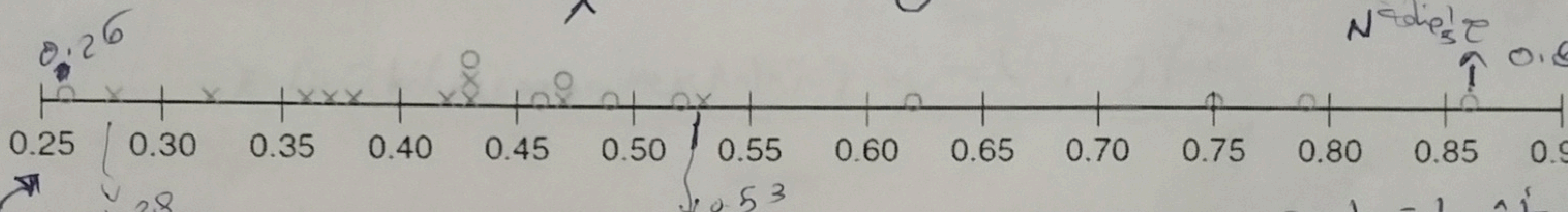


Figure 1.1: A dot plot of stem weight data.

* ندرس تأثير مادة واحدة بتثبيت باقي العوامل

(Chart إحصاء وصفي (على شكل
Descriptive statistics

إحصاء تلخيصي
summer statistics → tabel
أو ملخص بشكل جدول
سلايد 9

* الإضافة يجب أن تكون مدروسة وليست عشوائية

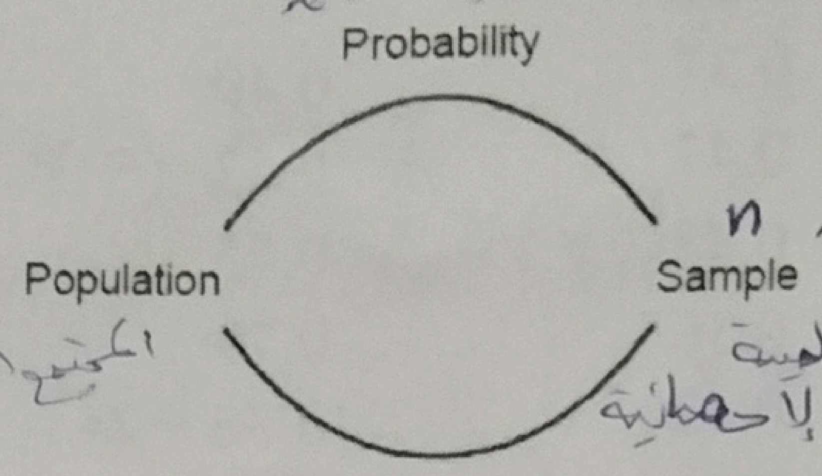
- On average, the use of nitrogen increases the stem weight. إضافة النيتروجين يحسن وزن نباتات البلوط
- Nitrogen is effective but ,how can this be quantified. ?

← أخذ عينة بشكل عشوائي، ما بناخذ أول 3 مثلاً

01010 10,010

احتمالية حدوث شي من خلال عامل معين، مثل دراسة سبب قلة الرواتب لعائلة معينة أو مكان السكن ودوره في تحصيل الرواتب

الاحتمالية



① عشوائية random
② ممثلة للمجتمع represented
تمثل الكل من المجتمع

← تأخذ عينة من مجتمع مثلاً لدراسة وتأخذ avg النتيجة كمتوسط لمعرفة الدخل مثلاً

مثلاً 1.5 مليون عائلة نأخذ 2000 عائلة من هذه العائلات مثلاً ونأخذ avg

المجتمع الإحصائي ← عدد كبير مثل دراسة في المجتمع تشمل كل عدد سكان المجتمع أو عن مصنع ندرس المصابيح بكاملها بنفس المنتج

العينة الإحصائية

الإحصاء الاستدلالي

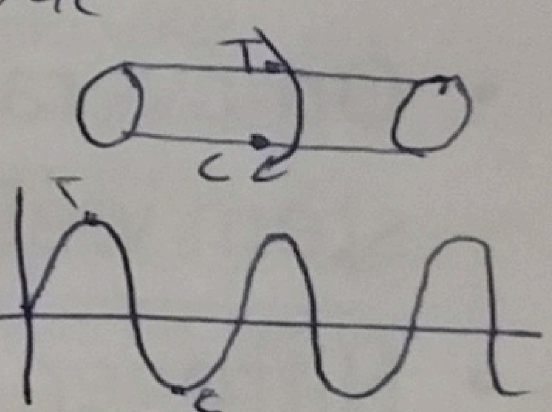
Figure 1.2: Fundamental relationship between probability and inferential statistics.

← استدلينا عن معلومة في population من خلال avg sample

1.3 Sampling Procedures; Collection of Data

failure يحصل دينكسر ← 10000 cycle
تآكل أكثر ← 4000 cycle
صدأ أكثر

Simple Random Sampling
Experimental Design



← نحسب عدد الدوران

دهن في مادة معينة، للتحقق من الصدأ

Example 1.3:1 A corrosion study was made in order to determine whether corrosion of an aluminum metal coated with a corrosion retardation substance reduced the amount of corrosion. The coating is a protectant that is advertised to minimize fatigue damage in this type of material. Also of interest is the influence of humidity on the amount of corrosion. A corrosion measurement can be expressed in thousands of cycles to failure. Two levels of coating, no coating and chemical corrosion coating, were used. In addition, the two relative humidity levels are 20% relative humidity and 80% relative humidity.

← نسبة الرطوبة

نقيس الصدأ من خلال عدد الدوران كل ما زاد عدد الدوران قبل التآكل، بتأخر ليصدي

رطوبة

مرة دهنا في الرطوبة 20 / 80
لا ,, ,, ,, 20 / 80

رطوبة

ندرس عاملين: رطوبة، دهان المادة
نثبت واحد ونغير واحد

Table 1.2: Data for Example 1.3

Coating	Humidity	Average Corrosion in Thousands of Cycles to Failure
Uncoated	20%	975
	80%	350
Chemical Corrosion	20%	1750
	80%	1550

إذا بدنا نقارنهم مع بعض بنثبت الرطوبة

الرطوبة العالية تقلل عدد الـ cycles بالتالي انكسار أسرع

* دهن مادة يزيد number of cycles
* والرطوبة العالية تقلل الـ n of cycles

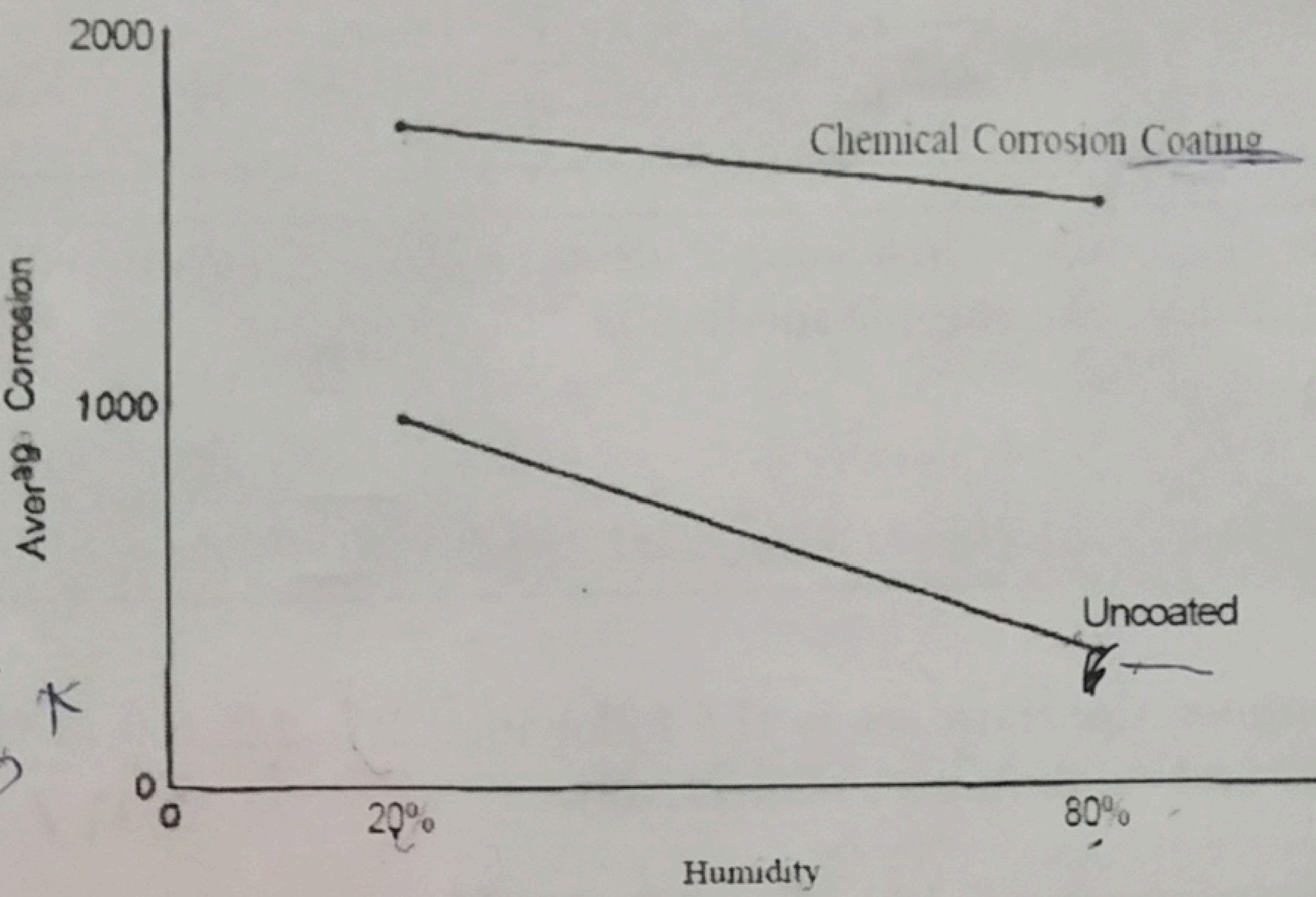


Figure 1.3: Corrosion results for Example 1.3.

مهم نتعلم قراءة الـ chart

بالامتحان بيجي chart or table

* اقل ما راد بعد نقطة عن المركز قل ما راد التشتت لـ S.D

تحليل البيانات

Data Analysis
Measures

Measures of center of Location (Central Tendency) (where the data center is in a sample)

- Mean
- Median
- Mode

Measures of variability (Dispersion) measures of spread

- Variance σ^2 / S^2
- Standard Deviation
- Range

Other methods of quantifying the center of location:

Trimmed mean

"trimming away" a certain percent of both the largest and smallest set of values)

1 2 $\boxed{3\,4}$ 5 6 even

$\frac{3+4}{2}$

3.5

1 2 $\boxed{3}$ 4 5 odd

الوسيط

avg

1.4 Measures of Location: The Sample Mean and Median

Definition 1.1: Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The sample mean, denoted by $\bar{x}$, is

avgال mean

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

Definition 1.2: 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

الوسيط median

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

As an example, suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

3.11 / 3.9

$$\bar{x} = 5.12, \quad \tilde{x} = \cancel{3.9}. \; 3.11$$

1.7 / 2.2 / 3.11 / 3.9 / 14.7

بسبب تباعد $\tilde{x}$ و $\bar{x}$ هو التباين في القيم

variance

Clearly, the mean is influenced considerably by the presence of the extreme observation, 14.7, whereas the median places emphasis on the true "center" of the data set. In the case of the two-sample data set of Example 1.2, the two measures of central tendency for the individual samples are

$$\text{avg } \bar{x} \text{ (no nitrogen)} = 0.399 \text{ gram},$$

$$\text{mean } \tilde{x} \text{ (no nitrogen)} = \frac{0.38 + 0.42}{2} = 0.400 \text{ gram},$$

$$\text{avg } \bar{x} \text{ (nitrogen)} = 0.565 \text{ gram},$$

$$\text{mean } \tilde{x} \text{ (nitrogen)} = \frac{0.49 + 0.52}{2} = 0.505 \text{ gram}.$$

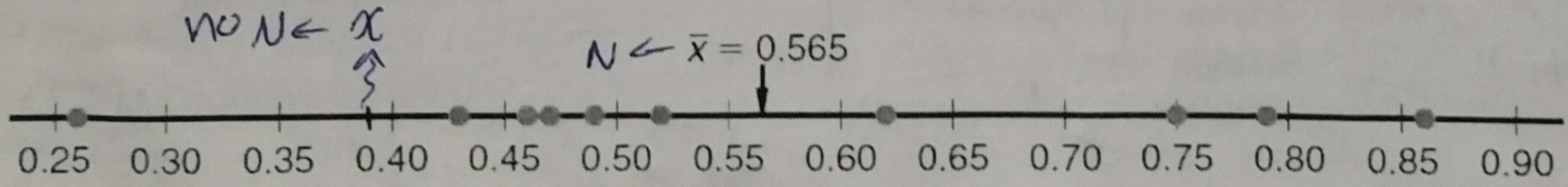


Figure 1.4: Sample mean as a centroid of the with-nitrogen stem weight.

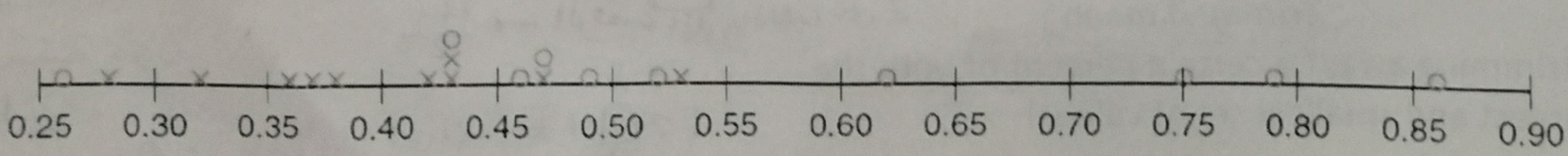


Figure 1.1: A dot plot of stem weight data.

sample 1:

1 2 5 8 9

$\bar{x}_1 \rightarrow$ mean :- 25/5 = 5

median : 5

mode : no mode

العنصر الأكثر تكرار ممكن يكون أكثر من رقم إذا تكرروا نفس عدد المرات

3, 4, 5, 6, 7

$\bar{x}_2$ mean :- 5

median : 5

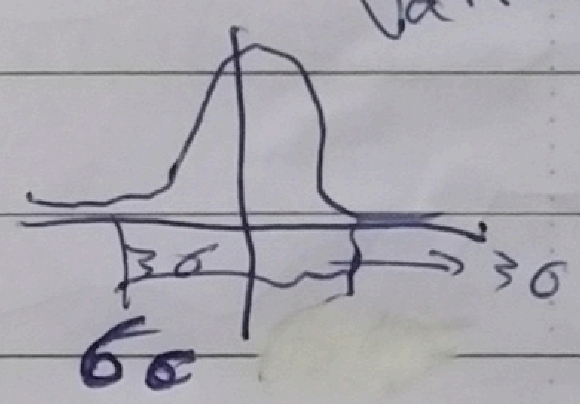
mode : no mode

population symbols

mean	μ	$\bar{x}$
median	Med	$\hat{x}$
mode	mode	mode
variance	σ^2	s^2
S.D	σ	s

→ على مركز الـ data

66 → variation تقلل



4, 5, 10, 16, 28, 5, 11, 5, 22, 19

mean $\bar{x} = \frac{\quad}{10} = 12.5$

median $\hat{x}$ = 4/5/5/5/10/11/16/19/22/28

$\hat{x} = \frac{10+11}{2} = 10.5$

mode = 5

* سبب تباعد مقاييس النزعة المركزية، أنّ عدد العينة قليل

$\bar{x} = 12.5$

$\hat{x} = 10.5$

mode = 5

symbols

$s^2 = \sum (x - \bar{x})^2 / n-1$

$s = \sqrt{s^2}$

* بحسب كم بعد نقطة عن المركز لنعرف مقدار التشتت ؟

$\sigma^2 \rightarrow$ population

$\sigma \rightarrow$

Subject Date No.

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

$$s^2 = \frac{(4-12.5)^2 + (8-12.5)^2 + (5-12.5)^2 \cdots}{10-1} = 70.5$$

المركز

variance

10.5 { 70.5

تباين

10.5 < Variat

ذكا

$s = \sqrt{70.5} = 8.39$

S.D

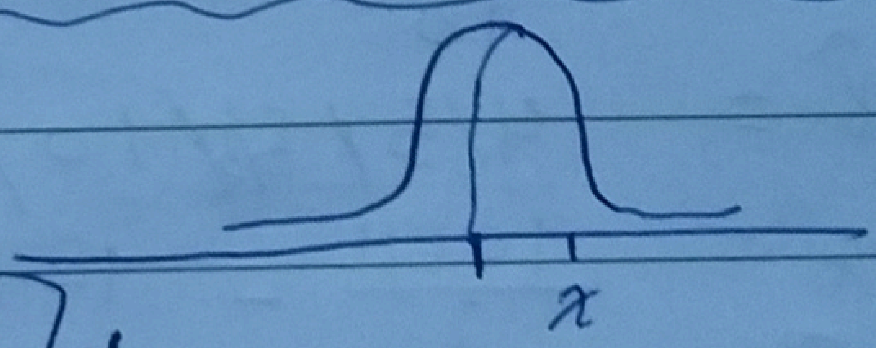
$$\boxed{\text{range} = \text{max point} - \text{min point} = 28 - 4 = 24}$$

$$\sigma^2 = \frac{(4-12.5)^2 + (8-12.5)^2 \cdots}{N = 10} = 36.45 \rightarrow \text{population}$$

عادة العدد يكون كبير

وليس 10 (طالبي بالامتحان !!)

$\sigma = 7.97$



$[(x-\bar{x})^2] / n-1$

← نطرح كل نقطة من المركز

ونربعها (Variation)

D.O.F ← (n-1)

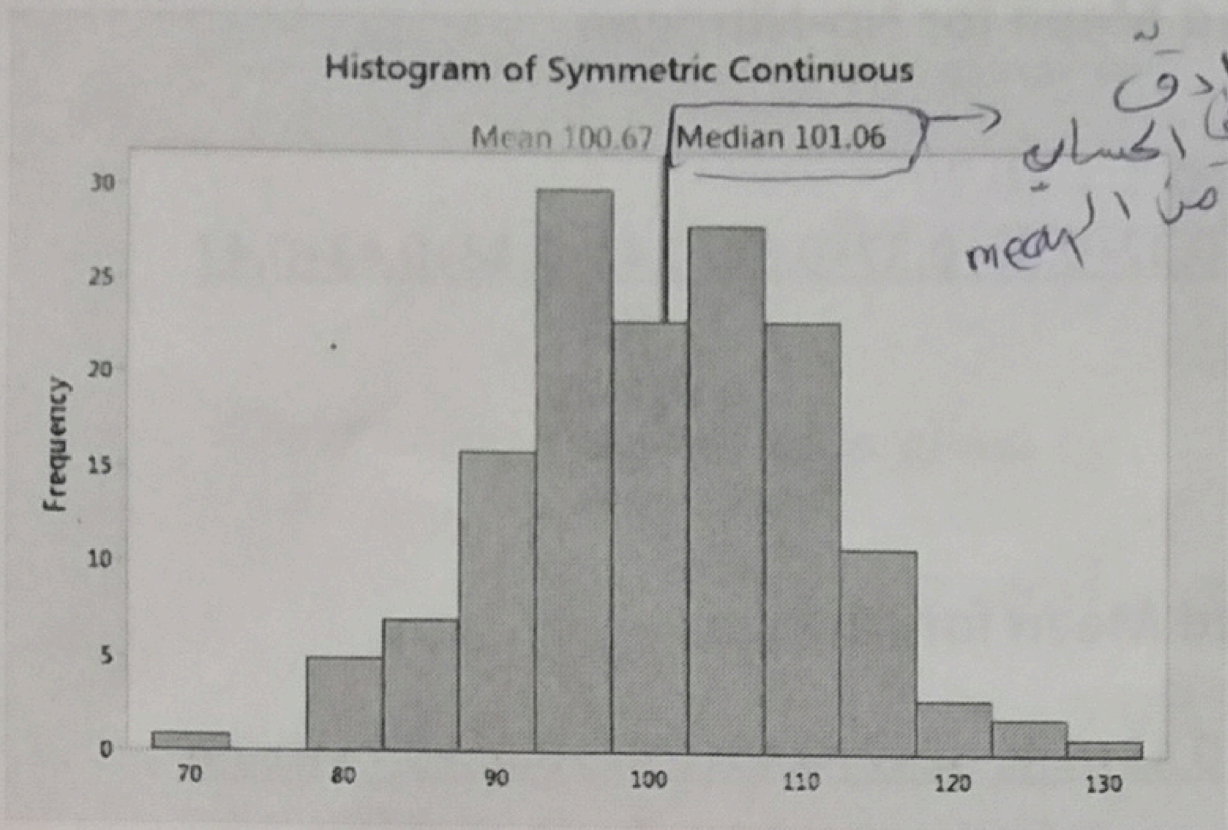
N O T E B O O K

* الظواهر حول العالم يجب أن تتوزع بشكل طبيعي

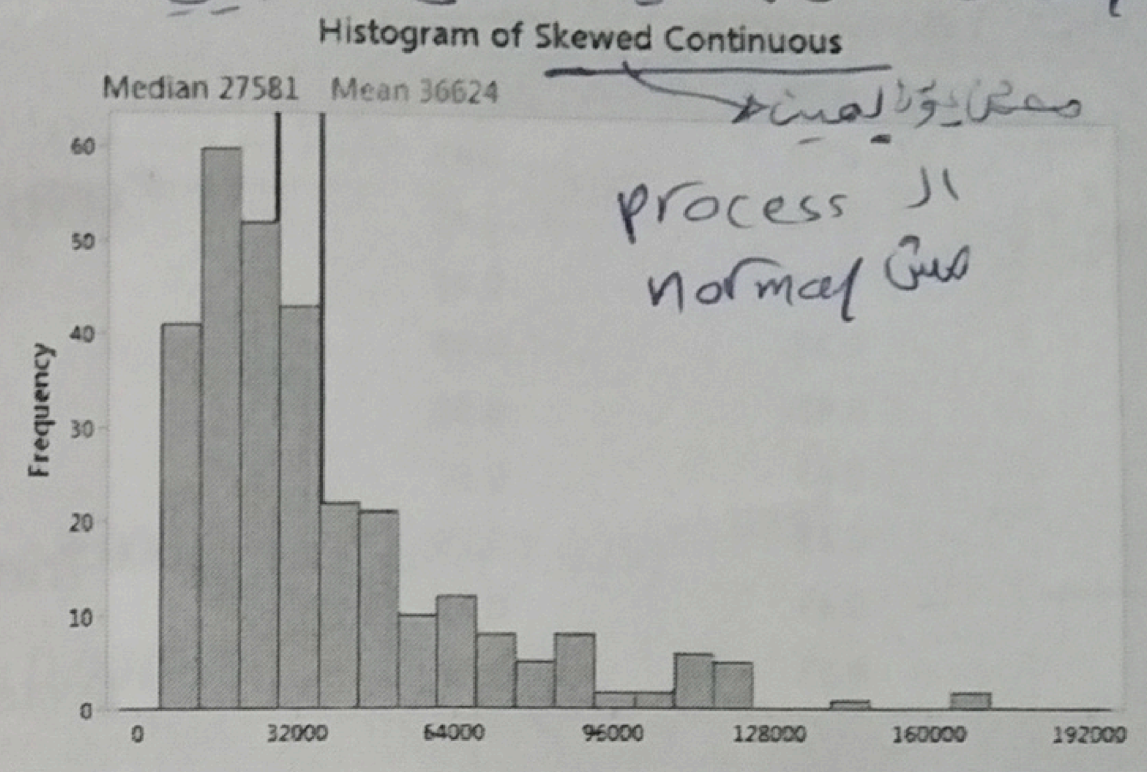
Mean and Median

* مثلاً شركة لها مواصفات معينة لمنتج معين، يجب أن يكون المنتج مطابق للمواصفات

المواصفات = ولادة في منحنى الطبيعي



أدق في الحساب من الـ mean



معظم قيم البيانات

الـ process مش normal

skewed left

البيانات تتحسن و منع التصنيع مثلاً

مسموح للتصنيع منتج normal

0.5 ± 1

[0.5+1
0.5-1]

عبارات لطرف منتج لحتون normal

median أكثر دقة

* السبب أنه لا يعتمد على الأطراف (بيكون ملغية !!)

① مرتبة data

② النسبة تعتمد على السؤال

Other Measures of Locations

-Trimmed Mean → قص الـ mean

e.g., in computing 10% trimmed mean, we cancel the highest 10% and the lowers 10% of our data

-Benefit:

TR 10%

1)Having a mean close to median

* بفعل الـ data cleaning من خلال حساب الـ outliers ثم حذفها

2) Reduce the effect of very high and very low value

size is 10 for each sample. So for the without-nitrogen group the 10% trimmed mean is given by

$$\bar{x}_{\mathrm{tr}(10)} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

and for the 10% trimmed mean for the with-nitrogen group we have

14

$$\bar{x}_{\mathrm{tr}(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Subject: Trimmed Mean (avg) Date: الوسيط (med) No. slide 14

~~7~~ / 8 / 9 / 9 / 11 / 13 / 17 / 18 / 19 / ~~21~~

Mean = 13.2

out layer بنحسبهم ونحذفهم

10% Tr

sample size ← $10 * \frac{10}{100} = 1$

من كل طرف بدنا نشيل واحد

meadian = $\frac{13+11}{2} = \frac{24}{2} = 12$

tr % 10 mean = $\frac{8+9+\dots+19}{8}$

ال mean اللي بنحسبه بعد ال tr يسمى trimmed mean

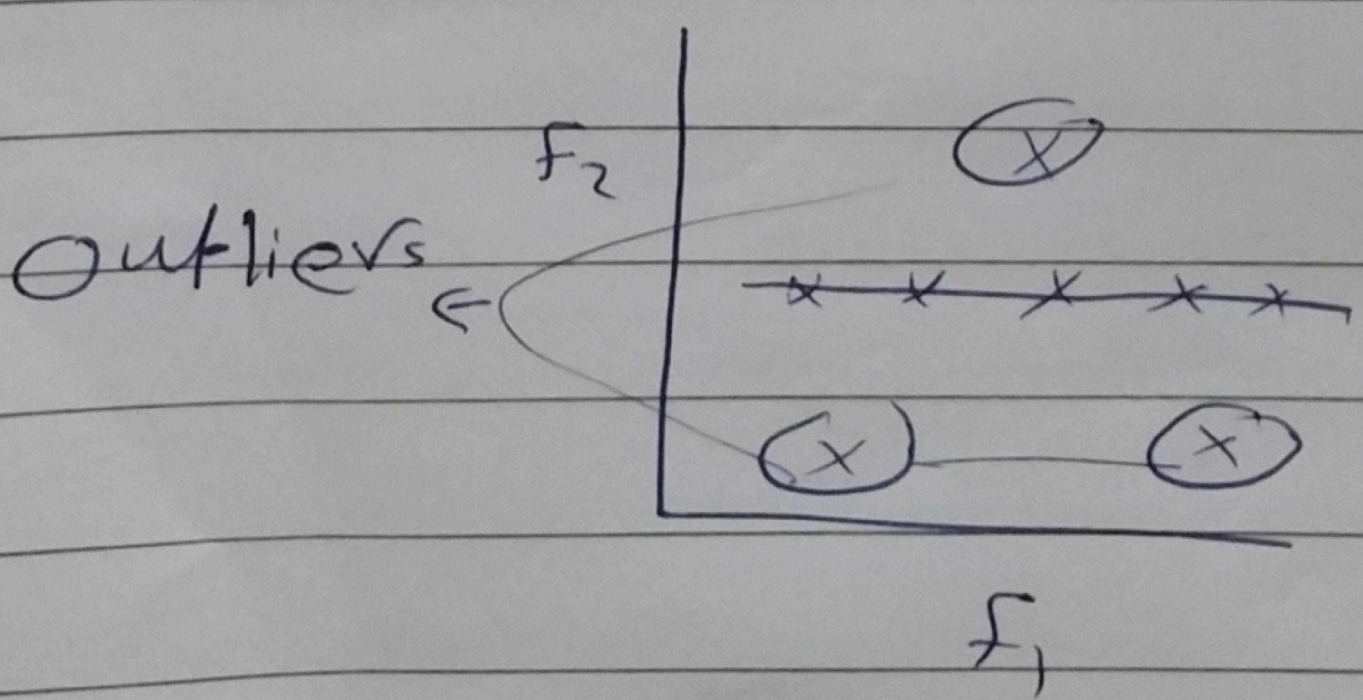
20% Tr = $\frac{20}{100} * 10 = 2$

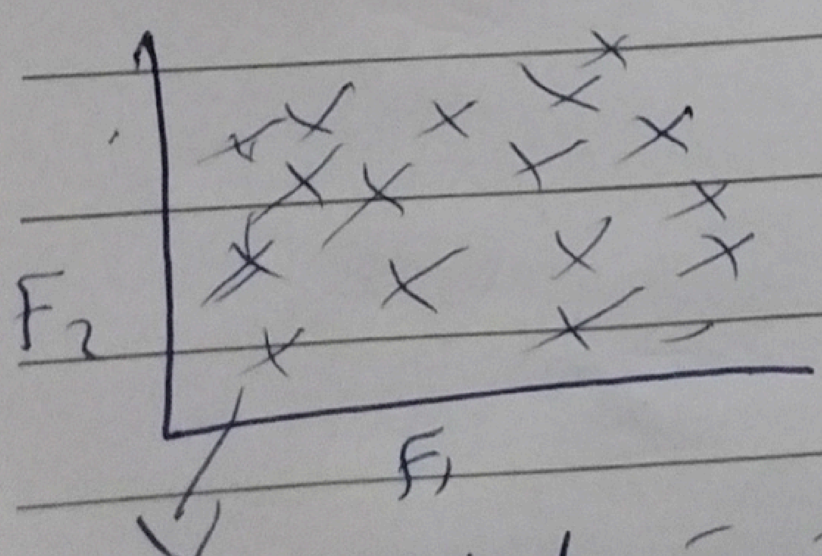
* من كل طرف 2 / 100%

* إذا استمرينا بعمل trim بنوصل لل median لأنه تخلصنا من جميع الأطراف التي يمكن أن تحسن ال data

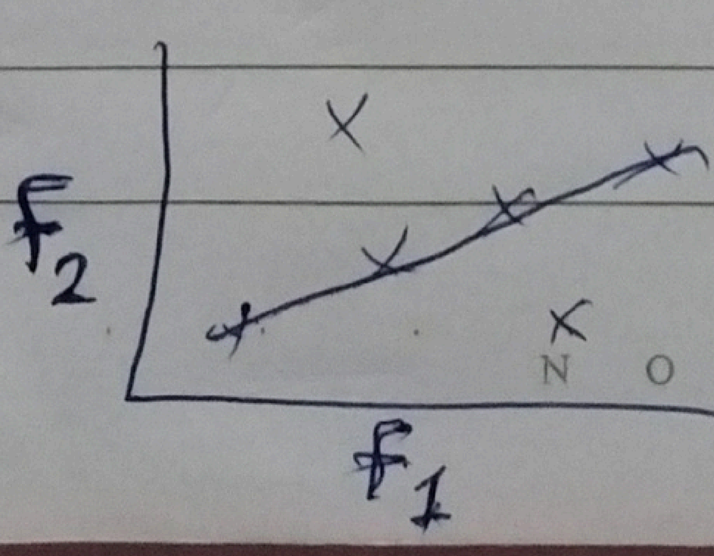
1) median أدق
2) trimmean → لأنه يلغي تأثير الأطراف — أكثر دقة من > mean
3) avg ↓ insensitive

sensitive → دقيق





Independent factor علاقة عشوائية
- no relation



* ملاحظة

Date No.

itive

vg(mean) تزداد
rimmed mean الحساسية
median للأطراف

Accurace

تزداد الدقة
1) median
2) trimmed m
3) avg(mea

filtration نسوي outliers لل قبل إزالة

No-Nitrogen	Nitrogen
0.28	0.26
0.32	0.43
0.36	0.46
0.37	0.47
0.38	0.49
0.42	0.52
0.43	0.62
0.43	0.75
0.47	0.79
0.53	0.86

** Trimmed Mean for No-Nitrogen

$$x_{tr(10)} = \frac{0.32+0.36+0.37+0.38+0.42+0.43+0.43+0.47}{8}$$

$$= 0.3975$$

** Trimmed Mean for Nitrogen

$$x_{tr(10)} = \frac{0.43+0.46+0.47+0.49+0.52+0.62+0.75+0.79}{8}$$

$$= 0.56625$$

* نتيجة ال trimmed mean لازم تكون قريبة من ال mean وال median معاً

قريب

- The trimmed means are close to both the mean and median for the individual samples
- Trimmed mean is more insensitive to outliers than the sample mean but not as insensitive as the median.

دقة أقل point

- The sample median is a special case of the trimmed mean in which all of the sample data are eliminated apart from the middle one or two observations.

15

* ال median حالة خاصة من ال trimmed mean

Measures of Variability
-Range
- Variance and Standard deviation

Range $= X_{\max} - X_{\min}$

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

symbol *

population

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

$$\sigma^2 = \frac{\sum (x-\bar{x})^2}{N} \rightarrow \text{size of sample}$$

عدد العينة * sample size ←

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

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

ample 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

* ال mean لازم يطلع close من القيم الموجودة في السؤال.

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

The sample variance s^2 is given by

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

10-1

As a result, the sample standard deviation is given by

S.D

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

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

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

إحصاء وصفي

Descriptive statstics

1) Scatter Plot

سكاتر

النقاط المبعثرة

At times the model postulated may take on a somewhat complicated form. Consider, for example, a textile manufacturer who designs an experiment where cloth specimen that contain various percentages of cotton are produced. Consider the data in Table 1.3.

مصنع يصنع ملابس بها نسبة قطن

Table 1.3: Tensile Strength

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

* بدنا نرسم Chart يوضح العلاقة بين نسبة القطن و tensile strength

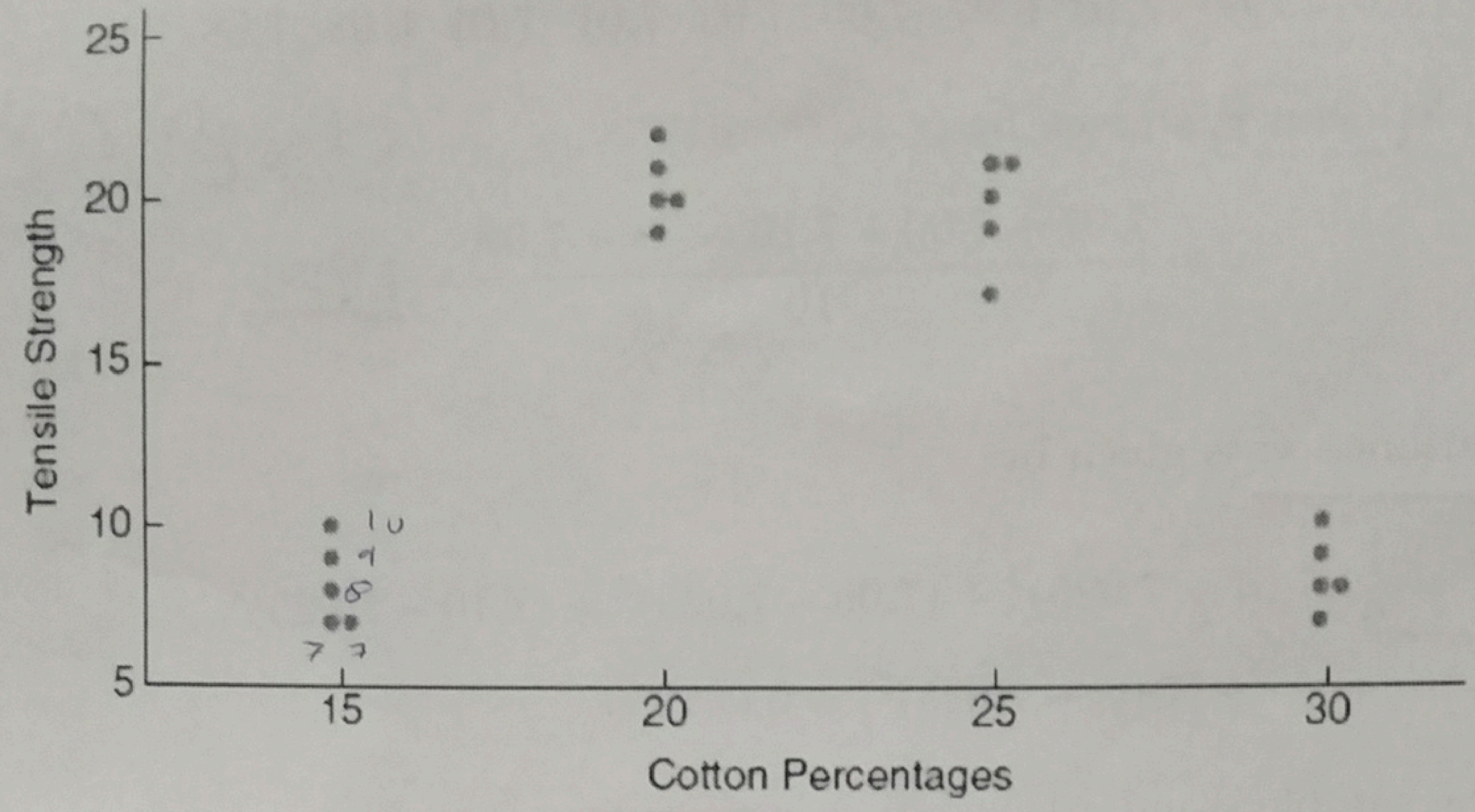


Figure 1.5: Scatter plot of tensile strength and cotton percentages.

* العلاقة بالرسم مش واضحة لأن النسب متباعدة !!

* العلاقة بين نسبة القطن في الملابس و tensile strength
كل ما زاد القطن تقل tensile strength (عند نسبة معينة)

19

2) Stem-and-Leaf Plot الغصن والورق

e.g., To illustrate the construction of a stem-and-leaf plot, consider the data of Table 1.4, which specifies the "life" of 40 similar car batteries recorded to the nearest tenth of a year.

دراسة لبطارية سيارة

Table 1.4: Car Battery Life

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

2.2 stem leaf

2 | 25669

* شكل ال data يحكم في اختيار مين ال stem ومين ال leaf

Table 1.5: Stem-and-Leaf Plot of Battery Life

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

mean تقريبا

$1^{\circ} \rightarrow (1-4)$ 1.1 1.3 1.2 1.4

$1^{*} \rightarrow (5-9)$