Central concepts in statistics

RHS 481

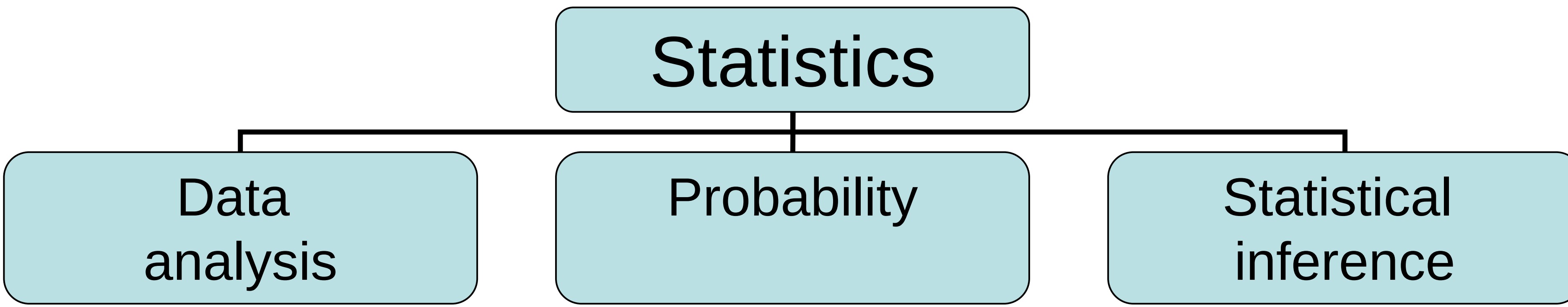
Lecture **5**

Dr. Einas Al-Eisa

- We muddle through life making choices based on incomplete information
- *Statistics* help us quantify *uncertainty*

Statistics

- = a discipline in which mathematics and probability are applied in ways that allow researchers to make sense of their data

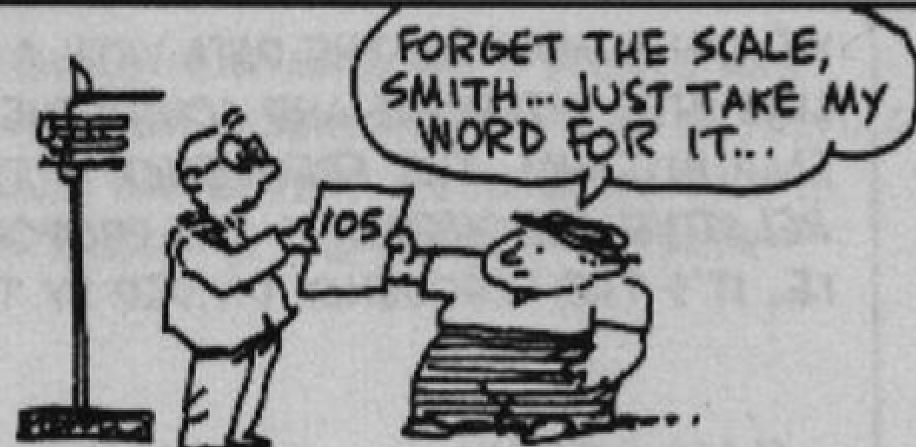

Statistics
Data analysis
Probability
Statistical inference

Definitions

- **Data analysis** = the gathering, display, and summary of data
- **Probability** = the laws of chance
- **Statistical inference** = the science of drawing statistical conclusions from data, using a knowledge of probability

Example: 92 students reported their weight (in pounds)

HERE IS SOME REAL DATA:
AS PART OF A CLASSROOM
EXPERIMENT, 92 PENN STATE
STUDENTS REPORTED THEIR
WEIGHT, WITH THESE
RESULTS:



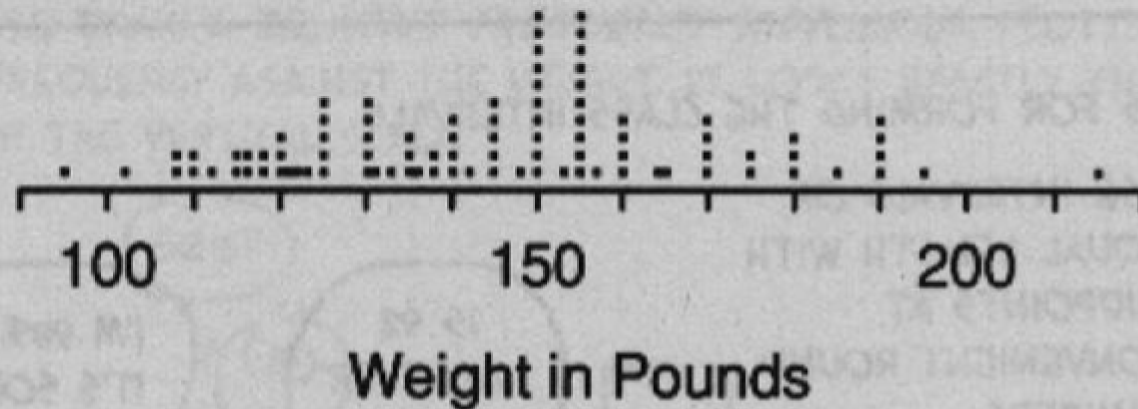
MALES

140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180 135
170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160 135 160
130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155

FEMALES

140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135 125
118 122 115 102 115 150 110 116 108 95 125 133 110 150 108

GETTING RIGHT DOWN TO BUSINESS, WE DRAW A *DOT PLOT*: ONE DOT PER STUDENT GOES OVER EACH STUDENT'S REPORTED WEIGHT.



YOU MAY SEE A *PROBLEM* HERE: THE CLUMPS AT *150* AND *155* POUNDS. THE STUDENTS TENDED TO REPORT THEIR WEIGHT IN *FIVE-POUND INCREMENTS.* IN REAL-LIFE SITUATIONS LIKE THIS ONE, SUCH ROUNDING OFF CAN OBSCURE GENERAL PATTERNS IN DATA... BUT FOR NOW, WE'LL JUST WORK AROUND IT.

WE CAN SUMMARIZE THE DATA WITH A ***FREQUENCY TABLE***. DIVIDE THE NUMBER LINE INTO INTERVALS AND COUNT THE NUMBER OF STUDENT WEIGHTS WITHIN EACH INTERVAL. THE ***FREQUENCY*** IS THE COUNT IN ANY GIVEN INTERVAL. THE ***RELATIVE FREQUENCY*** IS THE PROPORTION OF WEIGHTS IN EACH INTERVAL, I.E., IT'S THE FREQUENCY DIVIDED BY THE TOTAL NUMBER OF STUDENTS.

CLASS INTERVAL	MIDPOINT	FREQUENCY	RELATIVE FREQUENCY
87.5-102.4	95	2	.022
102.5-117.5	110	9	.098
117.5-132.4	125	19	.206
132.5-147.4	140	17	.185
147.5-162.4	155	27	.293
162.5-177.4	170	8	.087
177.5-192.4	185	8	.087
192.5-207.5	200	1	.011
207.5-222.4	215	1	.011
TOTAL		92	1.000

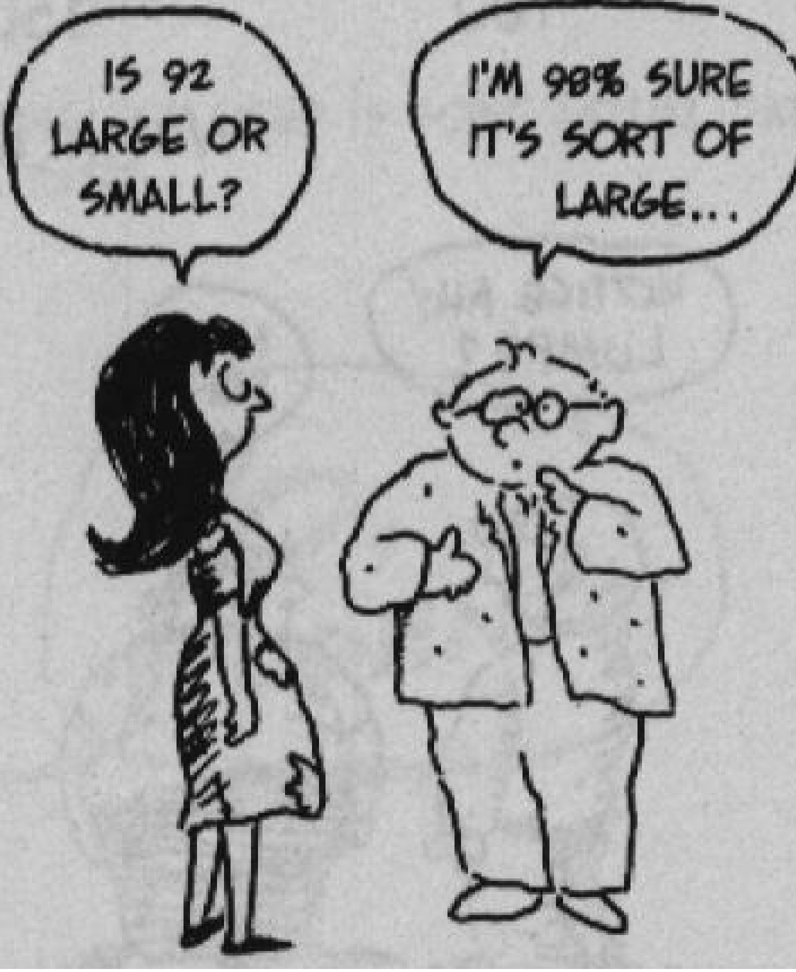
NOTE: WE KEPT THE INTERVAL BOUNDARIES AWAY FROM THOSE TROUBLESOME 5-POUND MULTIPLES. THIS GETS AROUND THE STUDENTS' REPORTING BIAS.

GUIDELINES FOR FORMING THE CLASS INTERVALS:

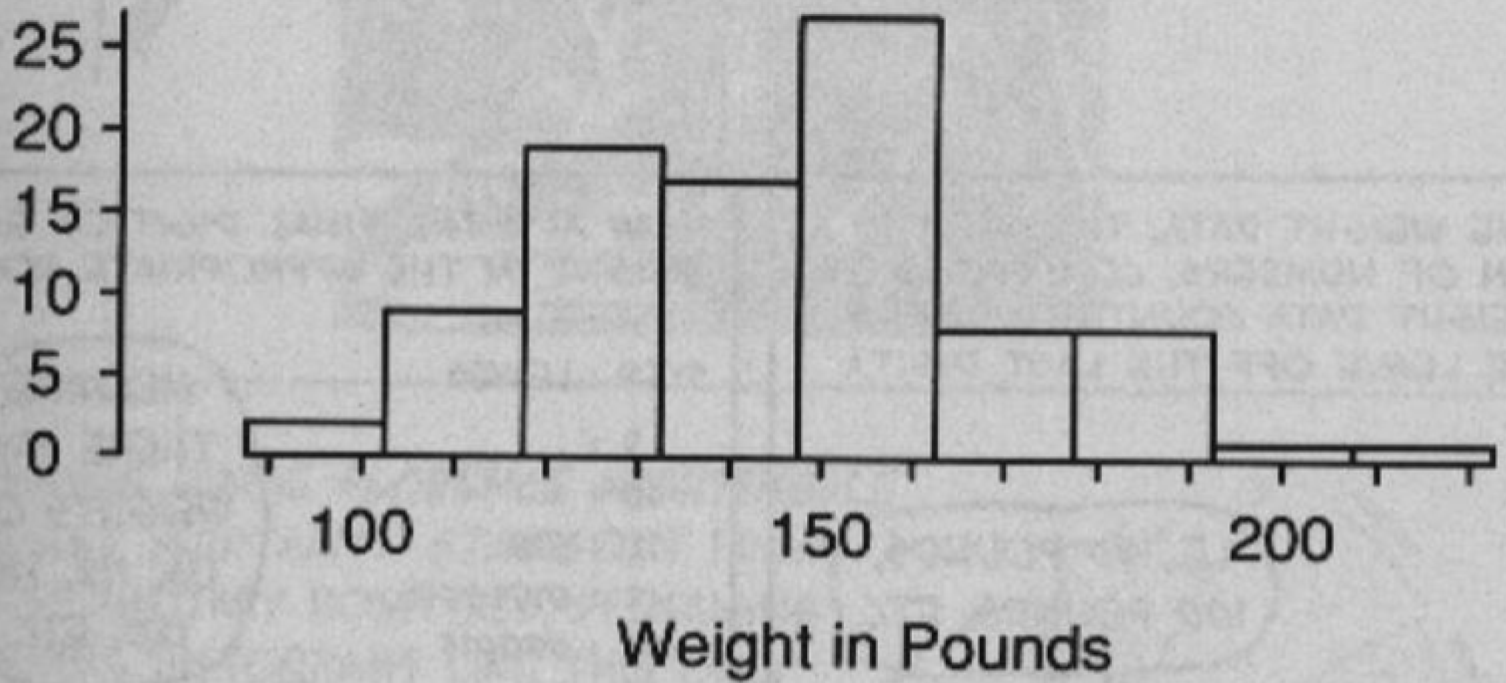
1) USE INTERVALS OF EQUAL LENGTH WITH MIDPOINTS AT CONVENIENT ROUND NUMBERS.

2) FOR A SMALL DATA SET, USE A SMALL NUMBER OF INTERVALS.

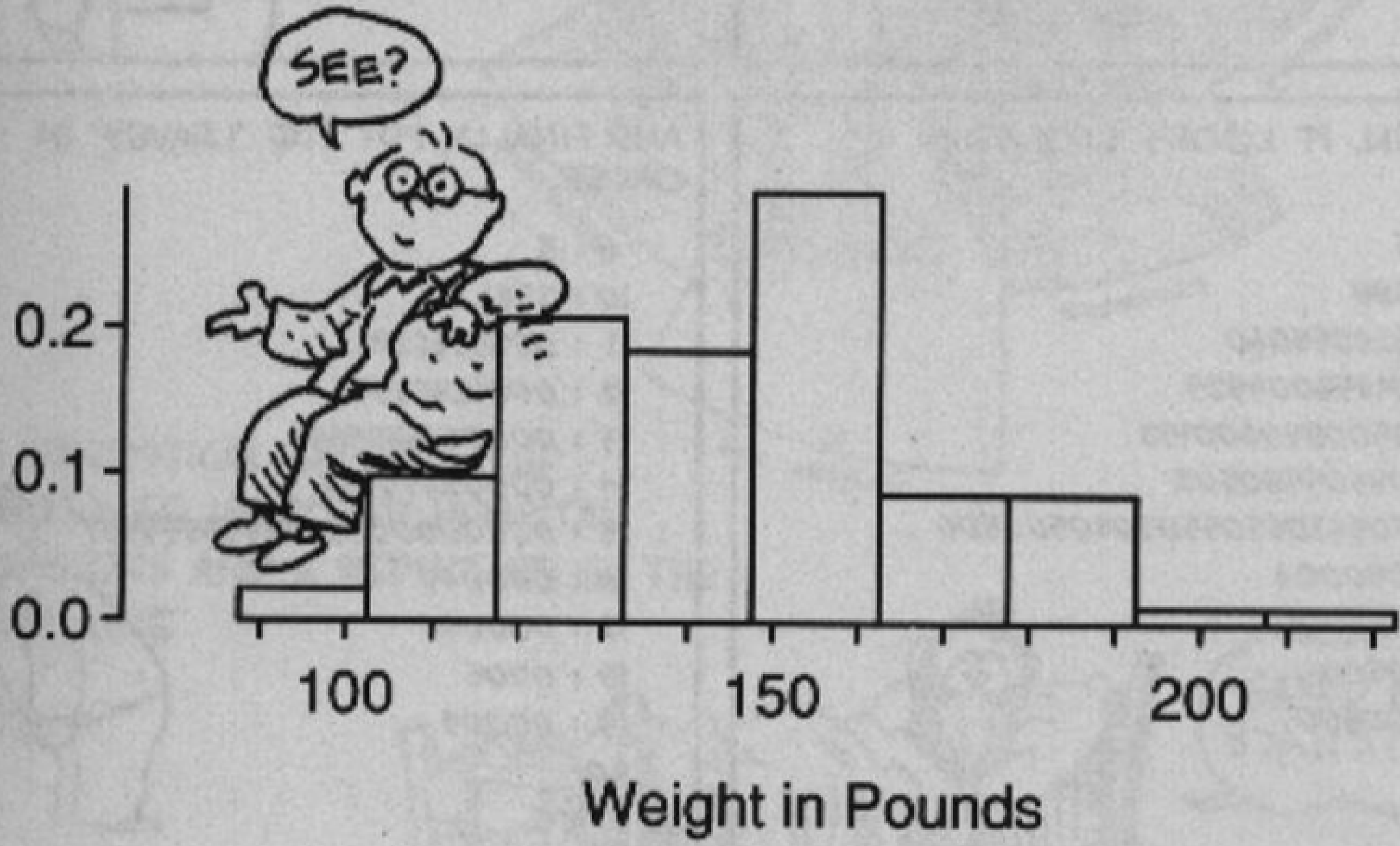
3) FOR A LARGE DATA SET, USE MORE INTERVALS!

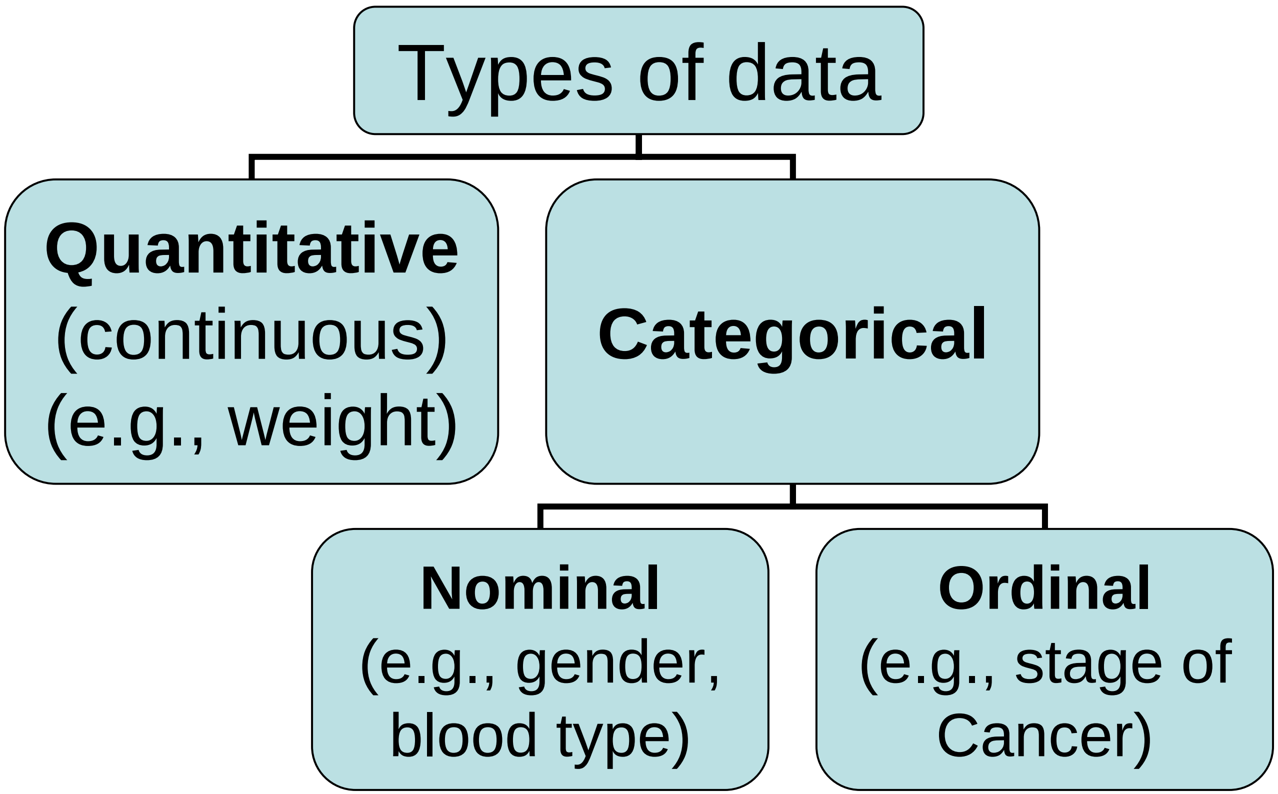


IN THE FREQUENCY TABLE, WE ARE SHOWING HOW MANY DATA POINTS ARE "AROUND" EACH VALUE. WE CAN GRAPH THIS INFORMATION, TOO. THE RESULTING BAR GRAPH IS CALLED A *HISTOGRAM*. EACH BAR COVERS AN INTERVAL AND IS CENTERED AT THE MIDPOINT. THE BAR'S HEIGHT IS THE NUMBER OF DATA POINTS IN THE INTERVAL.



WE CAN ALSO DRAW A ***RELATIVE FREQUENCY HISTOGRAM***, PLOTTING THE RELATIVE FREQUENCY AGAINST THE WEIGHT. IT LOOKS EXACTLY THE SAME, EXCEPT FOR THE VERTICAL SCALE.




Types of data
Quantitative
(continuous)
(e.g., weight)
Categorical
Nominal
(e.g., gender, blood type)
Ordinal
(e.g., stage of Cancer)

Types of statistics

Descriptive statistics:
Describe the data
and its distribution

Inferential statistics:
To *generalize* the data
or to
infer *cause and effect*
or *differences*
between groups

The normal distribution

- A symmetric frequency distribution (bell-shaped curve) that can be defined by the mean and standard deviation

- The distribution is symmetric around the mean

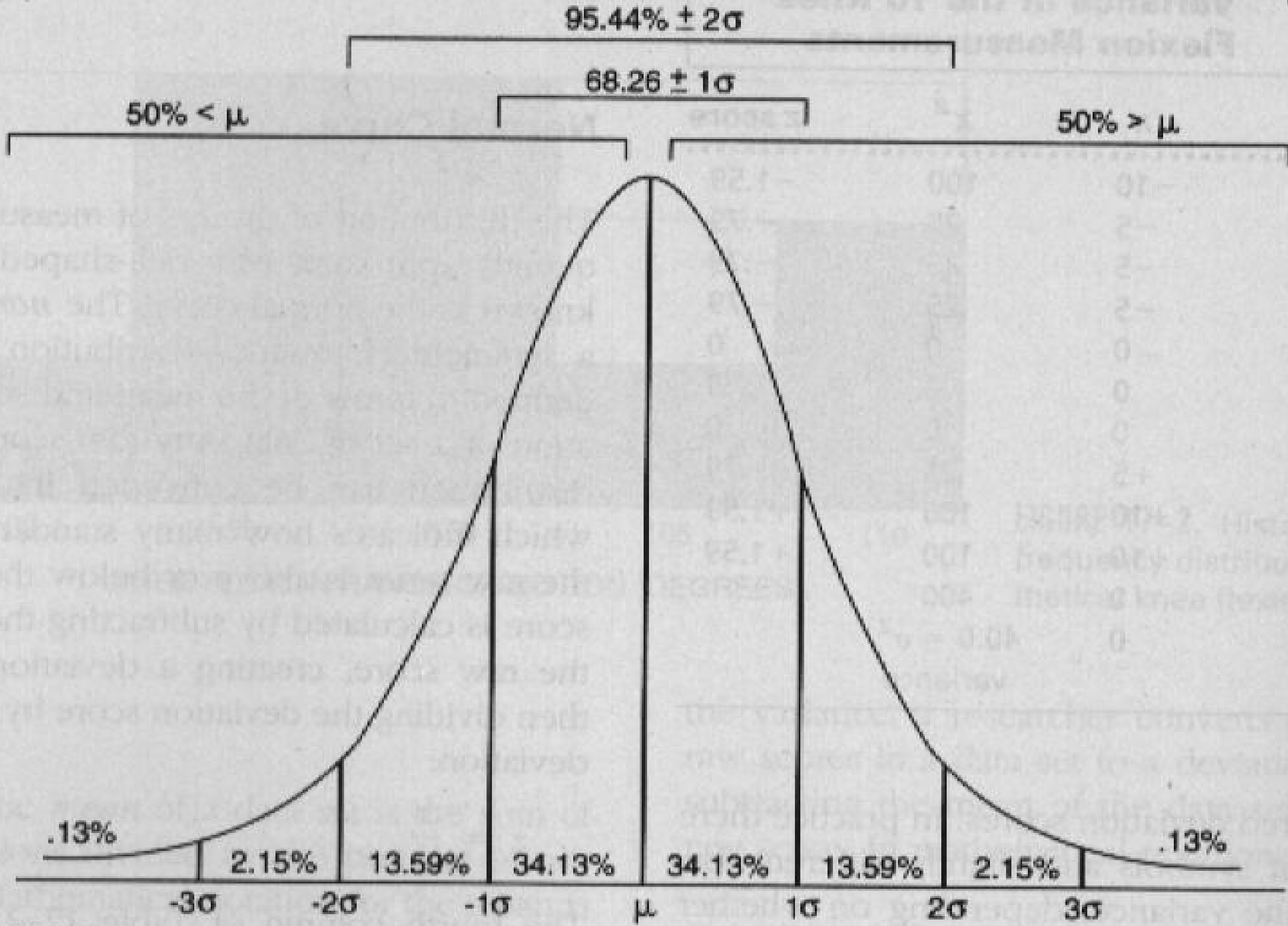


FIGURE 17–3. Probabilities of the normal curve. μ = mean; σ = standard deviation.

The normal distribution

- The mean $\pm$ 1 standard deviation covers approximately 68% of the data
- The mean $\pm$ 2 standard deviations covers approximately 95% of the data

Central limit theorem

- The basis of most statistical tests
- The mean value of repeated samples of a population has approximately normal distribution
- For purposes of experimentation, we assume that our samples are representative of the total population

Types of statistical tests

- **Parametric statistics:**
  - To describe normally distributed data
  - For continuous variable

- **Non-parametric statistics:**
  - When the distribution is not-symmetrical or unknown
  - For nominal or ordinal data

Parametric statistics

- **Z-test:**
  - Used to determine how many standard deviations your sample falls from the known population mean

- **T-test:**
  - Used to compare two means
  - As the sample size increase, the T-test becomes equivalent to Z-test

Parametric statistics

- **Unpaired T-test:**
  - To compare two independent groups

- **Paired T-test:**
  - Uses before and after data
  - Less variability → easier to achieve significance

Parametric statistics

- **Analysis of Variance (ANOVA):**
  - Tells you if more than 2 groups are different
  - H_o: all the means are equal

    H_1: not all the means are equal
  - Compares variances within groups to variances between groups (F-value)
  - It does not tell you which group is different!

Parametric statistics

- **Multiple Analysis of Variance (MANOVA):**
  - Used to determine not only that there are differences between the means, but what differences are significant

Non-parametric statistics

- **Ordinal data:**
  - Wilcoxon signed rank test

- **Proportions:**
  - Chi-square test
  - Fisher’s exact test

Correlation

- How closely do two factors follow each other? (e.g., height and weight)

- Does not assume cause-and-effect relationship

Linear Regression

- Can height predict weight?

$$\text{weight} = a + \mathbf{b}\,(\text{height})$$

- We can calculate the significance of **b** (is **b** significantly different from zero)

Multiple Linear Regression

- Weight = **a** + **b** (height) + **c** (calories)

- Can calculate the significance of any of **a**, **b**, **c**, ….etc.

Logistic Regression

- Used to determine the effect of a variable on a binominal outcome

  (e.g., dead or alive)

Types of Error

- **Type I (alpha error) = p-value:**
  - Probability that your results occurred by chance alone
  - Probability of rejecting H_o when it is correct
- **Type II (beta error):**
  - Probability of missing a true difference
  - Probability of accepting H_o when is not correct

Power

- = probability of finding a true difference (1-Beta)