Statistics

Contents

Chapter 1

1.1 – Key Words and Definitions

1.1 Key words

Data	Any observations that hvae been collected.
Statistics	Collect, analyze, summarize, interpret and draw conclustions from there.
Population	The complete set of elements being studied.
Samples	Some subset of the population.
Census	Collection from every member of a population.

Table 1.1: Statistics Vocabulary

→ If you take a sample, it must be collected **randomly**.

1.2 Types pf Data

P-P	**Parameter**	A characteristic of a population.
S-S	**Statistic**	A characteristic of a sample.

Table 1.2: Statistics Vocabulary

1.3 Two Types of Data

Qualitative (Categorical)	Data that is non-numerical e.g. color, gender, race, zip-codes... Mathematical operations are **meaningless**.
Quantitative	Numerical e.g. height/weight, wages, temperature, time. Mathematical operations are **meaningful**.

Table 1.3: table

1.3.1 Two types of Quantitative Data

Discrete data	Countabe or finite Numbers of eggs, dice...
Continious Data:	Infinite number of possible values (not countable) Usually a **measurement**, e.g. temperature.

Table 1.4: Quantitative data

1.4 4 Levels of Measurement

Nominal	Categories **not** ordered. e.g. religion
Ordinal	Can be ordered, differences are meaningless Rank, color (spectrum)...
Interval	Ordered, differences are meaningful, no ”Natural Zero” e.g. temperature
Ratio	Just like interval, but with a natural zero. e.g. amount of money

Table 1.5: Measurements

1.5 Design of Experiments/Observations

1.5.1 Observation vs. Experiment

An **observation** measures specific traits, but noes **not** modify subjects.

An **experiment** applies a treatment and then measuers the effect on the subjects.

1.5.2 Random

Each member of a population, has an equal chance of being selected in a sample.

Simple random sample

Each group of size ’n’ has an equal chance of being selected.

1.5.3 Common techniques to get a sample

Table 1.6: 4 Common techniques to get a sample

Convenience sample	You use the results, which you easily get (not random)
Systematic sampling	Put a population in some order and select every ”k^{th}” member.
Stratafied Sample	Breaking population into sub-groups based on some characteristic, and then take a simple random sample out of each sub-groups.
Cluster sample	Divide population into ”clusters” (regardless of characteristic), randomly select a certain number of clusters, and then collect data from the entire cluster.

Chapter 2

Frequency Distribution

A frequency distribution is a list of values with corresponding frequencies.

Class width	Difference between two ”lower class limits”
Lower class limit	Smallest value belonging to a class
Upper class limit	Highest value belonging to a class

Table 2.1: Frequency Distribution Terms

Steps:

1. Determine number of classes: 8
2. class width:

$$\frac{\text{Max Value - Min value}}{\text{number of classes}} \rightsquigarrow \frac{44-18}{8} \rightsquigarrow \frac{26}{8} \rightsquigarrow 3.25$$

   Round **up**. $\rightsquigarrow 4$
3. Start with the minimum value: 18
4. Creat classes with class width (4)
5. Find the class midpoint:

$$\frac{\text{upper class limit} + \text{lower class limit}}{2} \rightsquigarrow .$$

6. Class boundaries: used to seperate classes without gaps.

class width: 4

Lower class limit: 18, 22, 26, . . . 46

upper class limit: 21,25 . . . 49

class midpoint:

$$\frac{\text{upper class limit} + \text{lower class limit}}{2}$$

$\rightsquigarrow 19.5, 23.5, 27.5, 31.5, 35.5, 39.5, 43.5, 47.5$

class-width inbetween

class boundaries: Used to seperate classes without gaps. 17.5, 21.5, 25.5, 29.5, 33.5, 37.5, 41.5, 49.5

Relative frequency distribution: Percentage

$$\frac{\text{class } f.}{\sum f.(n)}$$

Cumulative Frequency Distribution Adds sequential classes together.

Age	Freq.	Rel. Freq.	Cum. Freq.
18-21	25	58.1%	25
22-25	10	23.3%	35
26-29	4	9.3%	39
30-33	2	4.7%	41
34-37	1	2.3%	42
38-41	0	0%	42
42-43	1	2.3%	43
46-49	0	0%	43
	n=43 $\sum f \uparrow$	100%	

Table 2.2: Frequency Distribution

2.1 Touching Bar Chart

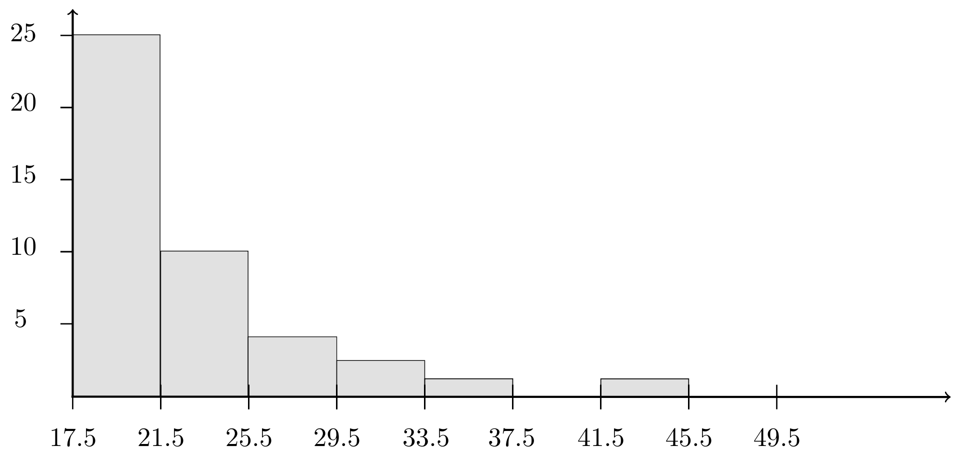

Figure 2.1: figures/stats-1

A cumulative chart would look exactly the same, but instead of having boundaries numered it'd be in the middle of the bars with teh cumulative frequency from class 1-8. And also the y-axis would be the percentage.

There is also a last one, where one takes the cumulative stuff, so that the graph colums are getting bigger and bigger...

Horizontal: Class midpoints or boundaries.

Vertical: Frequency.

Chapter 3

Describing Data

3.1 5 Caracteristics

1. Center
2. Variation
3. Distribution
4. Outliers
5. Changes over time.

Center The ”middle: of the data set. 3 ways:

1. **mean:** Arthimetric Average, add all the values and divide by the numbers you added.

$$\begin{aligned}
\text{Mean} &= \frac{\sum x}{\text{Number of values}}\\
\sum &= \text{sum}\\
x &= \text{data value}\\
n &= \text{number of items in a sample}\\
N &= \text{Number of items in a population}\\
\overline{x} &= \text{sample mean}\\
\mu &= \text{population mean}
\end{aligned}$$

We can write the sameple mean then as:

$$\overline{x} = \frac{\sum x}{n}.$$

And the population mean as

$$\mu = \frac{\sum x}{N}.$$

Example

Sample data: 5.40, 1.10, 0.42, 0.73, 0.48, 1.10
$\overline{x} = \frac{\sum x}{n}$ is the formula we have to use, because it's a sample, then we get:

$$\overline{x} = \frac{5.40 + 1.10 + 0.42 + 0.73 + 0.48 + 1.10}{6} = \frac{9.23}{6} = 1.54.$$

2. **Median:** The middle value of the dataset.
   - Must be in order.
   - Find middle value.
     - If odd number of values, the median is the middle number.
     - If even number of values, the median is the **mean** of the two middle values.

   Example

   8, 3, 5, 11, 13, 4, 6
   To find the median we first need to order em, so:
   3, 4, 5, 6, 8, 11, 13.
   We have seven values so we can just take the middle one which is 6.
   If we'd then add 412, so our numbers are:
   3, 4, 5, 6, 8, 11, 13, 412.
   Then our median is: $M = \frac{6+8}{2} = 7$

   And it's obviously the same with decimals.

   The Median is **not** affected by outliers, the mean is.

3. **Mode:** The most commonly ocurring data value.

Example

(a) 5.40, 1.10, 0.42, 0.73, 0.48, 1.10
Here the mode is 1.10 becuase it's ocuring most often.
(b) 27, 27, 27, 55, 55, 55, 88, 88, 89
Modes: 27, 55
(c) 1, 2, 4, 7, 9, 10, 12
Mode: $\emptyset$

One rounds always to one more value than the beginning values, so one more decibel, and rounded is not before the most final step.

3.2 Mean of a frequency distribution

Age	freq.	x (midpoint)	$freq. \cdot x$
21-30	28	25.5	714
31-40	30	35.5	1065
41-50	12	45.5	546
51-60	2	55.5	111
61-70	2	65.5	131
71-80	2	75.5	151
	n=76		$\sum f \cdot x = 2718$

Table 3.1: Another age distribution

So now we can get the sample mean:

$$\overline{x} = \frac{\sum f \cdot x}{n} = \frac{2718}{76} = 35.76.$$

And here another table, this time about a grade's distribution.

	w	points	$x \cdot w$
Hw	15%	70	10.5
T_1	20%	90	18.0
T_2	20%	68	13.6
T_3	20%	85	17.0
F	25%	95	23.75
			$\sum x \cdot w = 82.85$

Table 3.2: Grade example

$$\overline{x} = \frac{\sum x \cdot w}{\sum \cdot w} \rightarrow \frac{82.85}{100} = .8285 \rightarrow 82.85\%$$

We can also do the same just half way in the class with the following table:

	w	points	$x \cdot w$
Hw	15%	70	10.5
T_1	20%	90	18.0
T_2	20%	68	13.6
	55%		$\sum x \cdot w = 42.10$

Table 3.3: Grade example

$$\overline{x} = \frac{\sum x \cdot w}{\sum \cdot w} \rightarrow \frac{42.10}{55} = .765 \rightarrow 76.50\%$$

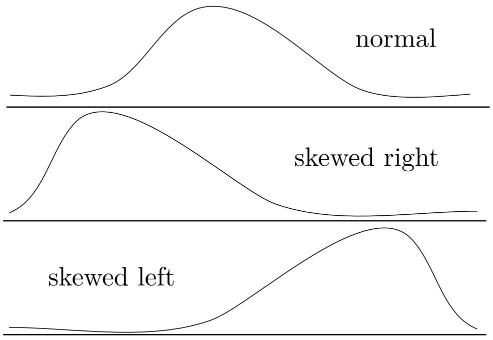

Figure 3.1: figures/stats-2

3.3 Finding the standard Deviation

Notitz:

next lesson is 3.3