3. MEAN, SD and the NORMAL CURVE: FPP Ch 4,5

3.1 SAMPLES AND POPULATIONS

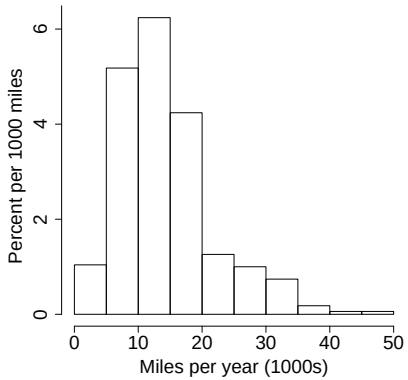
• **Remember, we take a sample to find out about a population.**

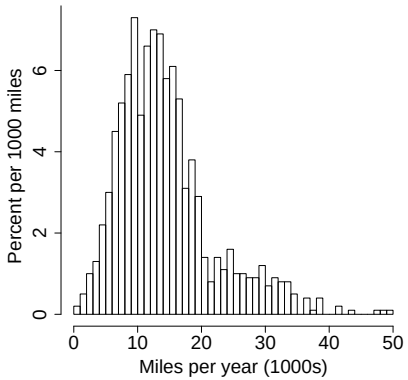
• **As we sample more and more subjects, the histogram gets more and more like the (unknown) histogram of the whole population, so long as there are no biases in the way we are sampling.**

• **For a (continuous) quantitative variable, as we have more and more subjects, we can draw our histogram boxes on finer and finer scale.**

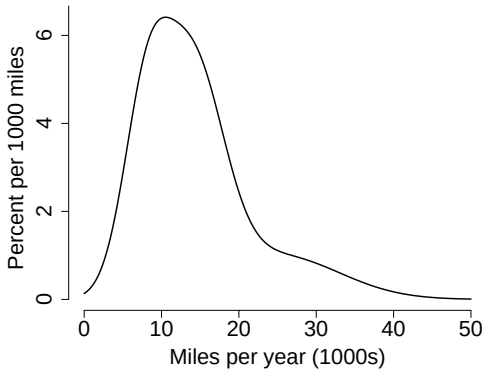
• **Our histogram gets more like a continuous curve: the population distribution.**

• **Example of the 1000 cars, in units of 1000 miles per year:**





• **In fact, this sample came from the following population distribution:**



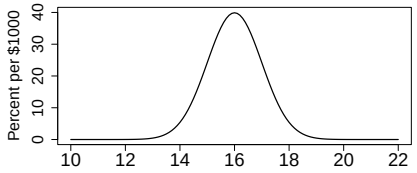
Rest of this page for your notes

3.2 THE (SAMPLE) AVERAGE OR (POPULATION) MEAN

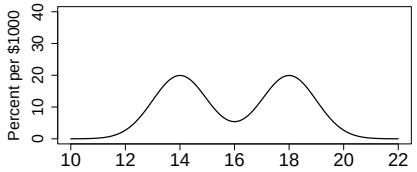
- **Add up the list. Divide by the number in the list.**
- **Examples: average (mean) height of men age 40**
  average (mean) weight of women age 25-29
  your GPA – a weighted average?
- **A very incomplete picture.**

Want to know the "shape of the histogram"

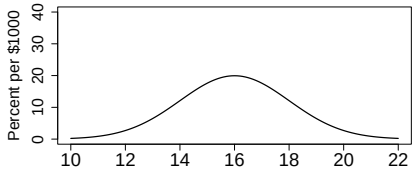
- **Three population distributions with the same mean:**
  average (mean) income of TAs (in $1000) at 3
  U.California campuses:



mean=median=16
Percent below 16: 50%
Percent 15 to 17: 67%



mean=median=16
Percent below 16: 50%
Percent 15 to 17: 16%

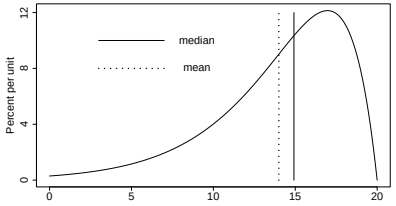


mean=median=16
Percent below 16: 50%
Percent 15 to 17: 38%

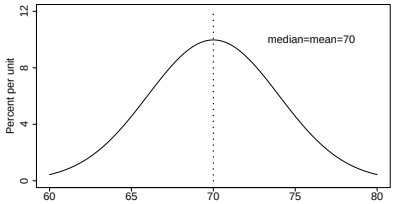
Blank page for your notes

3.3 THE MEDIAN

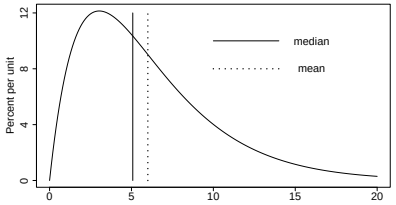
- **"All the children are above average" (??)**
  Almost all the children can be above their average.
- **Half the values are above the median, and half below.**
- **Exam scores: median bigger than average**



- **Heights of men age 40: (range 60" to 80")**
  median and average about the same



- **Incomes: units of $ 10,000.**
  The median is less than the average.



With a few multi-millionaires, almost everyone could be below average

Blank page for your notes

3.4 THE STANDARD DEVIATION (SD)

• Our 3 distributions in 3.2 all had the same mean and the same median, and the mean equal to the median.

• We need a measure of spread about the mean.

• The SD is the root-mean-square deviation from the average.

• Computing the SD:

The data 5 obsv	The deviation from avg	The square dev of col. 2
3.0	-1.3	1.69
4.2	-0.1	0.01
5.8	1.5	2.25
6.0	1.7	2.89
2.5	-1.8	3.24
sum = 21.5	sum =0.0	sum = 10.08
average = 4.3	(check)	average = 2.016
		sqrt = 1.42

• Your calculator will probably say 1.59 not 1.42 ??
For SD we take sum (10.08) divided by sample size (5).
For $(SD)^+$ we take sum divided by one less (4).
Most books/calculators use $(SD)^+$.
For large samples, it will not matter: $10,000 - 1 = 9,999$

• For the three curves in 3.2, the SDs are 1, 3 and 2.

• For a "bell-shaped distribution":
67% (2/3) is within ± 1 SD of the mean
95% (19/20) is within ± 2 SD of the mean
99.7% (almost all) is within ± 3 SD of the mean.

Blank page for your notes

3.5 HOMEOPATHY STUDY 4
QUANTITATIVE LAB1 RESULTS

• Here are the 40 quantitative observations from Lab1, arranged in increasing order

2.0	2.8	2.9	2.9	3.8	4.1	4.2	4.2
4.8	5.0	5.2	5.7	5.7	5.8	6.1	6.2
6.4	6.6	6.7	6.8	6.9	7.0	7.0	7.2
8.0	8.2	8.5	9.5	9.5	10.0	10.0	10.2
10.8	11.0	11.5	12.0	12.5	12.8	14.5	16.5

•The average or mean is 7.54.

•The standard deviation (SD^+ ?) is 3.40.

•The median is 6.85.
Strictly, it is any number between 6.8 and 6.9.

•The median is a little bit less than the mean.
There are 2 somewhat large observations.

• 27 of 40 observations are within 1 SD of the average (67.5%).

• 38 of 40 observations are within 2 SD of the average (95%).

• The mean is sensitive to a few extreme observations: remember, the billionaire on page 3.3.

• The median would be the same, even if those large observations were 34.5 and 36.5 (for example).

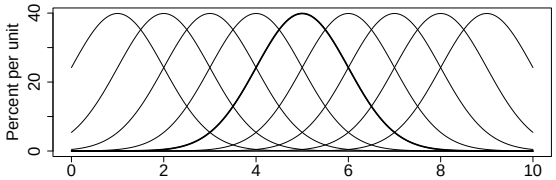
3.6 THE NORMAL CURVE

• Many quantitative measurements have a typical "bell-shaped" distribution in the population.

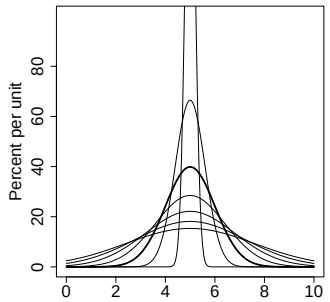
• This is known as the normal distribution.

• The standard normal distribution has mean 0 and SD 1.

• If we slide it sideways, we change the mean, but not the spread or shape (SD).



• If we stretch it sideways we change the SD but not the mean.



• When talking about a given value, relative to the distribution, it is how many SDs different from the mean that matters. That is, we talk in SD units. Remember, about 2/3 (67%) of values are within 1 SD of the mean.

3.7 HOMEOPATHY STUDY 5
THE LAB1 RESULTS UNBLINDED

• When the Lab1 results were unblinded (on air), here they are. They are again arranged in increasing order within each group.

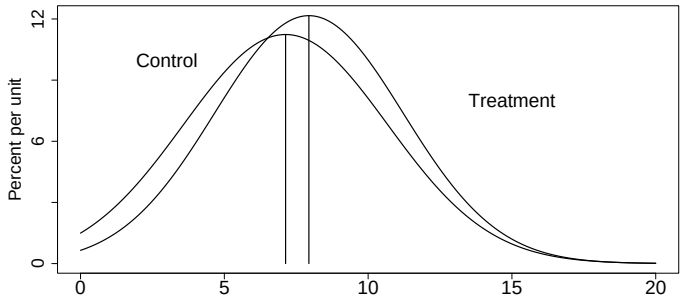
Treatment (Dilution)					Control (Pure Water)			
2.0	3.8	4.2	5.0		2.8	2.9	2.9	4.1
5.7	5.7	5.8	6.4		4.2	4.8	5.2	6.1
6.6	7.0	8.0	8.5		6.2	6.7	6.8	6.9
9.5	10.0	10.0	10.2		7.0	7.2	8.2	9.5
11.5	12.0	12.5	14.5		10.8	11.0	12.8	16.5

• For the Treatment: mean = 7.94, SD^+ = 3.28

• For the Control: mean = 7.13, SD^+ = 3.55

Relative to the treatment distribution, the control mean (7.13) is (7.94-7.13)/3.28 or 0.25 SD units below mean.

Relative to the control distribution, the treatment mean (7.94) is 0.23 SD units above the mean



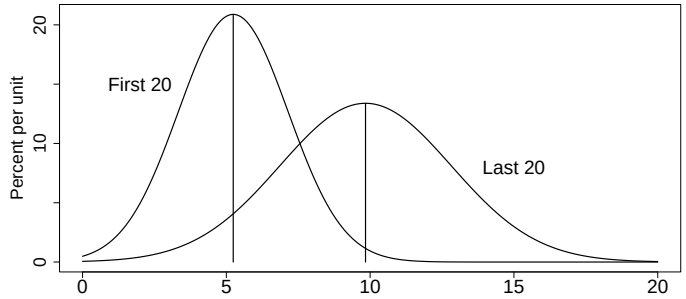
Blank page for your notes

3.8 HOMEOPATHY STUDY 6
THE LAB1 RESULTS BY TIME

• Martin Bland (the statistician) did not want to struggle with "incompatible Excel files" on air in front of a million viewers! So he insisted on having the data ahead of time (but still blinded).

• Lab 1 gave him their 40 observations, in the order those observations were made. So, just as an example, he pretended the first 20 were treatment and the last 20 were control.

• For the first 20, mean = 5.24, SD = 1.91
For the last 20, mean=9.84, SD = 2.98



• Relative to the second 20, the mean of the first 20 is 1.54 SD below the mean.

• Relative to the first 20, the mean of last 20 is 2.41 SD above the mean.

• Recall the tubes were tested on "live blood". For the later samples, the blood from the donors was not quite so fresh!!

Blank page for your notes

3.9 FINDING AREAS UNDER THE CURVE

• A value which is a deviation (difference) in SD units from the mean is often known as a z-score.

• For a bell-shaped population distribution, z-scores have a standard normal distribution: mean 0, SD 1.

• The total area under a distribution curve is 1, or 100%

• The area between two values of the (horizontal axis) variable is the proportion or percentage in that range.

• If 58% of the area is between values 6.93 and 12.89, this represents than 58% of the population distribution between values 6.93 and 12.89.

• The 58th percentile of a population distribution is the value such that 58% of the population values fall below it.

• For any distribution: the median is the 50th percentile.

• For any distribution: the interquartile range is the difference between the 75th and the 25th percentiles. It is another measure of spread – different from the SD.

• For the standard normal distribution, FPP P.A-105 tabulates the area between any value, and the negative of that value.

• If we have a bell-shaped curve and convert to SD units, then we can find the area in any interval

For example, 68.27% is within 1 SD of mean.

For example, 95.45% is within 2 SD of mean.

Blank page for your notes

3.10 More on A(z): Cholesterol example

• **Suppose cholesterol levels in some population have mean 200 mg/dl, SD 25 mg/dl and follow a bell-shaped curve. (Not very realistic.)**

• **Cholesterol levels are "high" if over 230, "very high" if over 250, and "too low" if under 160. What are these limits in SD units? (Answers: 1.2, 2.0 and -1.6)**

• **What percentage of the population has "very high" levels? (Answer: middle area = 95.54%: so 2.23% (or approx 2.5%))**

• **What percentage of the population does NOT have too low levels? (Answer: middle area = 89%: so 94.5%)**

• **What percentage of the population has "normal" levels (160 to 230)? (Answer: below 160 is 5.5%, above 230 is 11.5%, so 83%)**

• **Fred has level 230? What percentile is he at? (Answer: 88.5 percentile)**

• **Anna is at as much below the mean as Fred is above? What percentile is she at? (Answer: 11.5 percentile).**

• **Jane is at the 60th percentile. What is her cholesterol level? (Answer: need between-area = 20%: z-score=0.25, level = 206.25)**

• **Joe is at the 30th percentile. What is his cholesterol level? (Answer: need between-area = 40%: z-score = -0.52: level = 187)**

• **Exact vs approximate A(z). (Theory vs practice.)**

Blank page for your notes