Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 3 (MWF)

Suhasini Subba Rao

Review

- Each individual in a population has several different quantities associated with it. For example, height, subject etc. These are known as variables. Variables come in different forms, numerical/categorical/binary etc.

- A relative frequency histogram can be used to represent a sample. In general the associated plot for the population will be unknown - as we do not know the population. The height of a 'bar' indicates chance of a variable been selected in the corresponding interval.

- To describe the populations of a numerical continuous random variable we use a density plot (rather than a histogram). The area below the density tells us the likelihood of observing the outcome in the interval.

- The shape of the density (distribution) can be used to describe features in the population.

Numerical characteristics

- It is hard to quantatively compare different histograms and other graphical tools.

- In general a numerical characteristic describes some feature in the data (they are often refered to as a parameter).

- They are useful for making statistical inference.

- Types of numerical characteristics:
  - Mean.
  - Median.

- To define these notions we first define a random variable.

Random variables and some notation

- In the previous we introduced the idea of a variable (this is the quantity/feature in the population we are interested in). Now we look at the idea of a $random\ variable$.

- In statistics we call this variable a random variable, because the outcome changes from individual to individual.

- Definition The size of a random sample, are the number of individuals observed. To be general, we usually denote this as n.

- Definition We denote the random sample as $X_1, \ldots, X_n$.

- Here X_i denotes a measurement (height etc) of the ith randomly chosen individual from the population.

- We will often use the notation $X_1, \ldots, X_n$ because we do not want to specify a fixed sample. By using $X_1, \ldots, X_n$ we avoid the need to write down all possible (SRS) subsets of the population that you can draw (too long and sometimes impossible)!

Example: Random variables

- Suppose we have the population $2, 5, 7$ (a rather trivial population!).
- We draw a sample X_1 from the population. It can be any one of

2	5	7

- We draw a simple random sample (SRS) of size two, we denote this as (X_1, X_2).
- (X_1, X_2) can be any one of

X_1	2	5	7	2	2	5
X_2	2	5	7	5	7	7

- Observe the SRS can involve repetitions (as mentioned in Lecture 2).

- We observe that the $random\ variable\ X_1$ can be any one value from the population. It is random - we do not know what it is.

- We observe that the $random\ variables\ (X_1, X_2)$ can be any one of several different subsets of size two. It is also random.

- Hence we see using the notation X_1, X_2 gives us versatility. Rather than stating and an exact sample we can use (X_1, X_2) to denote any sample from the population of size two.

Some formal definitions

- A population parameter is some measure of the population. For example a measure of central tendency, such as the mean.

- The random sample When we don't have numbers or want a more general way of describing an arbitrary random sample is $X_1, X_2, \ldots, X_n$. This is a sample of size n. Hence n draws are made of the population. X_i is called a random variable.

- The sample parameter/statistic is a function of the sample $X_1, \ldots, X_n$ (if you like algebra use $t(X_1, \ldots, X_n)$). Examples include the average (what we formally define as the sample mean).

An example of a measure of central tendency: The mean

- In many problems the goal is to make inference about the population mean.

- The population mean is the average of all outcomes in the population. Usually the population mean is unknown.

- We need to make inference about the population mean based on the observed sample mean.

- The sample mean (formal definition) Suppose $X_1, \ldots, X_n$ are n observations from a population, then the sample mean is

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_{n-1} + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\text{sum of observations}}{\text{sample size}}.$$

Mean: advantages and disadvantages

- It is the average of the data set.
- There is only one mean per data set.
- Properties Mathematically it is very simple.

  It is easy to combine the means from two data sets to obtain the mean of the comined data set.

  – *Example:* Sample 1 is $\{1, 2, 3, 4\}$ and its mean is 2. Sample 2 is $\{10, 11, 12, 13\}$ its mean is 11.5. The mean of the combined samples $1, 2, 3, 4, 10, 11, 12, 13$ is $\frac{2.5\times4+11.5\times4}{8} = 7$ (observe we have only used the individual means and sample size to calculate this average).
- There are other statistical advantages which we come to later.

- Main drawback: It is sensitive to extreme values (outliers).

Mean and its sensitivity to outliers

- Try this example: Calculate the mean of the following samples

Sample 1:	$-1, 0, 0, 0, 0, 0$
Sample 2:	$-1, 0, 0, 0, 0, 2$
Sample 3:	$-1, 0, 0, 0, 0, 20$

- The final measurement in Sample 3 could have got there due to contamination in an experiment. Observe how much it influences the mean!

- We say that the mean is not robust to outliers. Intuitively the 'center' seems to be 0.

- A centrality measure that is less sensitive to outliers is the $median$, which we define below.

The median

- The median of a set of measurements is the middle value when the measurements are arranged from lowest to highest.
- It is the central point of the sample.
- Half the sample have values less than the median.
- Half the sample have values more than the median

Population median The central point of the population. Half the population fall below the median and half the population fall above it.

Sample median The central point of the sample. Half the sample fall below the median and half the sample fall above it.

Calculating the median in practice

Calculating the median:

- Let n be the size of the sample.
- Order the data
- If n is odd, then the median is the $(n+1)/2$ point in the ordering.
- If n is even, then the median is the average of the $n/2$ and $(n/2+1)$ values.

Example (odd number of observations):

- Data $97, 99, 93, 96, 91, 90, 95$. We see that $n = 7$.
- Ordered data: $90, 91, 93, 95, 96, 97, 99$
- $(n + 1)/2 = 4$
- $90, 91, 93, \underbrace{95}_{\text{middle value}}, 96, 97, 99$. We see that 95 is the 4th value in the ordered row.

Example (even number of observations):

- Data $97, 99, 93, 96, 91, 90, 95, 100$. We see that $n = 8$.
- Ordered data: $90, 91, 93, 95, 96, 97, 99, 100$
- $n/2 = 4$
- The 4th value is 95, the $5th$ value is 96.
- The median is $(95 + 96)/2 = 95.5$.

The median and extreme values

Let us consider the example above:

- Now calculate the median for the samples below:

Sample 1:	$-1, 0, 0, 0, 0, 0$
Sample 2:	$-1, 0, 0, 0, 0, 2$
Sample 3:	$-1, 0, 0, 0, 0, 20$

- We recall that the sample means were $-1/6, 1/6, 19/6$ respectively.

- Compare this with the medians of the samples.

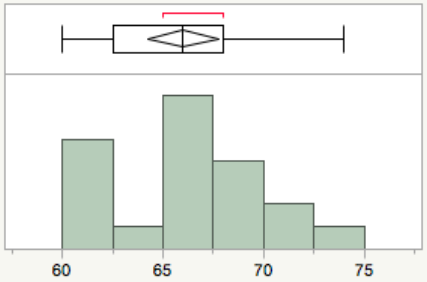
Motivating spread: The heights of students

- Here are four samples each of size 5. For each sample the average is taken.

	Sample 1	Sample 2	Sample 3	Sample 4
	68	65	67	69
	74	62	65	62
	68	60	64	71
	61	66	68	72
	61	66	65	66
Average	66.4	63.8	65.8	68

Distribution of heights and averages

height



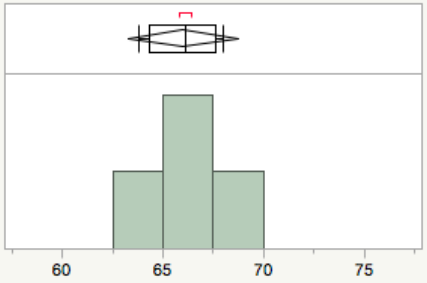
Quantiles

100.0%	maximum	74
99.5%		74
97.5%		74
90.0%		71.9
75.0%	quartile	68
50.0%	median	66
25.0%	quartile	62.5
10.0%		61
2.5%		60
0.5%		60
0.0%	minimum	60

Summary Statistics

Mean	66
Std Dev	3.7836212
Std Err Mean	0.8460434
Upper 95% Mean	67.770789
Lower 95% Mean	64.229211
N	20

average



Quantiles

100.0%	maximum	68
99.5%		68
97.5%		68
90.0%		68
75.0%	quartile	67.6
50.0%	median	66.1
25.0%	quartile	64.3
10.0%		63.8
2.5%		63.8
0.5%		63.8
0.0%	minimum	63.8

Summary Statistics

Mean	66
Std Dev	1.7358955
Std Err Mean	0.8679478
Upper 95% Mean	68.762197
Lower 95% Mean	63.237803
N	4

- What are the main similarities and differences between the histograms?

Measures of variability/spread

- The simplest measure of spread is to use the range. The smallest and largest observations eg. the range of $-1, 0, 0, 0, 0, 20$ is $[-1, 20]$.

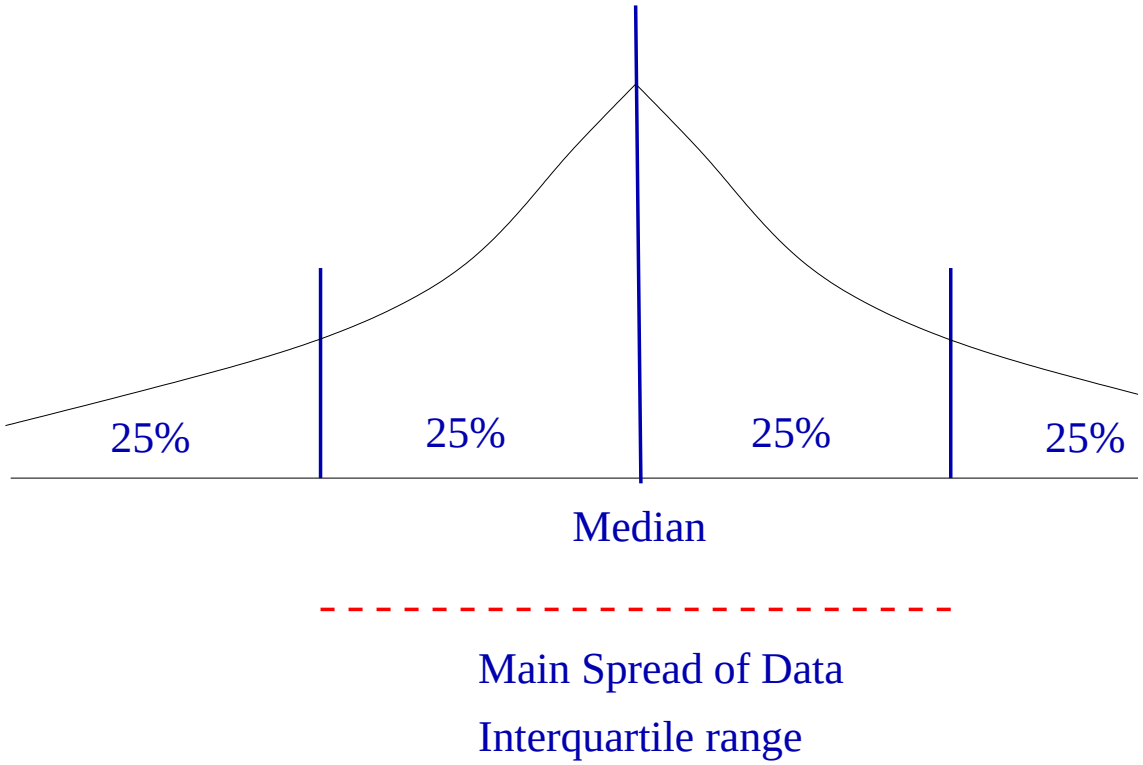
- We see that the range is extremely sensitive to outliers.

- We want a measure of variability that discriminates between different degrees of concentrations of data.

- We recall that the median is the 'half way' or 50% mark in data.

- We can use other percentile marks.
  - The 25% percentile is the value where 25% of the observations lie below the value and 75% above the value.

– The 75% percentile is the value where 75% of the observations lie below the value and 25% above the value.



This is a density plot together with their corresponding percentiles.

Calculating the 25th and 75th percentile

- The 25th and 75th percentiles are known as the first and third quartiles respectively. The 50th percentile is the median.

- Calculation of percentiles is best done with a computer.

- A simple (approximation) of the first and third quartile is to is to find the integer closest to $n/4$ and $3n/4$, where n is the sample size. The $n/4$ and $3n/4$ values in the ordered data are approximations of the first and third quartile respectively.

Measures of spread - The interquartile ranges (IQR)

- We know how to evaluate the 25% th, 50% th and 75% percentile (1st, 2nd and third quartile).

- The 1st quartile is a centrality measure, the median.

- The larger the difference between the 3rd and the 1st quartile the more spread out the population.

- Definition The Interquartile Range (IQR) = 3rd quartile - 1st quartile.

Example

The number of people volunteering to give blood at a center was recordered for each of 20 successive Fridays. The data is summarized below.

320	370	386	334	325	315	334	301	270	310
274	308	315	368	332	260	295	356	333	250

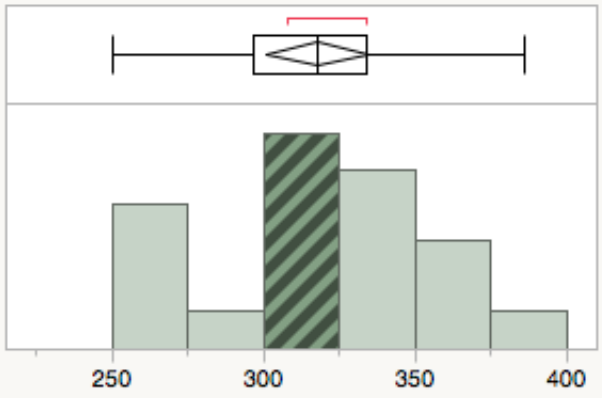
The data can be found here: `https://www.stat.tamu.edu/~suhasini/teaching651/lecture4_data_blood_donation.dat`

Question Summarize the data set.

In JMP

Distributions

BloodDonation



Quantiles

100.0%	maximum	386
99.5%		386
97.5%		386
90.0%		369.8
75.0%	quartile	334
50.0%	median	317.5
25.0%	quartile	296.5
10.0%		261
2.5%		250
0.5%		250
0.0%	minimum	250

Summary Statistics

Mean	317.8
Std Dev	36.624122
Std Err Mean	8.1894027
Upper 95% Mean	334.94062
Lower 95% Mean	300.65938
N	20

Eye-balling the quartiles

- Order the observations from smallest to largest

250	260	270	274	295	301	308	310	315	315
320	325	332	333	334	334	356	368	370	386

- Then, since there are an even number of observations, find the average of the 10th and 11th values, which is $(315 + 320)/2 = 317.5$. 317.5 is the median of the observations.

- To evaluate the 1st and 3rd quartile find the 'median' of the first and second half of the data.

- First quartile: take the average of the 5th and 6th ordered observation The 5th observations is 295, the 6th observation is 301. The 1st quartile is $(295 + 301)/2 = 298$.

- Third Quartile: the average of the 15th and 16th value, which is $(334 + 334)/2 = 334$.

- These almost match what JMP gives.