Means and measures of central tendency

Means, or the arithmetic "average", are important statistics in characterizing data. It is a measure that provides an indication of where the center of the distributions lies and is an important reference point. In hypothesis testing it is usually the means that are compared to determine of two samples are potentially drawn from the same population or not. Variances and other parameters can also be tested to compare populations, but the test of the means is more common.

Summation Operations

The symbol Σ is used to represent summation. Given a variable, Y_i, representing a series of observations from Y_1 (the first observation) to Y_n (the last observation out of "n" observations), the notation ΣY_i represents the sum of all of the Y_i values from the first to the last. Since the summation is for values of i from 1 to n the summation sign is often subscripted with "i = 1" and superscripted with an n (e.g. $\sum_{i=1}^{n} Y_i$)

Example of Summation: A variable "length of Bluegill in centimeters" is measured for individuals captured in a seine. This quantitative variable will be called "*Y*", and the number of individuals captured will be represented by "*n*".

For this example let $n = 4$

The variable Y_i is subscripted in order to distinguish between the individual fish (i)

$$Y_1 = 3,\ Y_2 = 4,\ Y_3 = 1,\ Y_4 = 2$$

Summation operation: To indicate that the sum all individuals in the sample (size *n*) write.

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + Y_3 + Y_4 = 3 + 4 + 1 + 2 = 10$$, and where, $n = 4$

the mean is given by $\sum_{i=1}^{n} Y_i \Big/ n = 10/4 = 2.5$

Sum of Squares

Two other values that will have to be calculated are the "sum of the squares" and the "square of the sums". To indicate the sum of squared numbers, simply indicate the square of the variable after the summation notation.

$$\sum_{i=1}^{n} Y_i^2 = Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 = 3^2 + 4^2 + 1^2 + 2^2 = 9 + 16 + 1 + 4 = 30$$

where $n = 4$

This is called the Sum of Squares, and should not be confused with the ...

Square of the Sum: The sum was $\sum_{i=1}^{n} Y_i = 10$

The square of the sum is given by simply squaring the sum, $\left(\sum_{i=1}^{n} Y_i \right)^2 = 10^2 = 100$.

Both of these calculations will be needed in calculating the variance.

Measures of Central Tendency

These measures provide an indication of location on a scale. The most common measure is called the arithmetic mean or the "average". It is the sum of all observations of the variable of interest ($\sum_{i=1}^{n} Y_i$) divided by the number of values summed (n).

Calculation of the Mean

Example: The calculation of the mean for 4 fish lengths.

It was previously determined that

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + Y_3 + Y_4 = 3 + 4 + 1 + 2 = 10$$

where, $n = 4$

The mean is given by $\sum_{i=1}^{n} Y_i \Big/ n = 10/4 = 2.5$

For a larger sample of fish

$Y_i = 7, 9, 9, 3, 6, 5, 0, 7, 0, 7$

$n = 10$

$\Sigma Y_i = (7 + 9 + 9 + 3 + 6 + 5 + 0 + 7 + 0 + 7) = 53$

The mean is then $\sum_{i=1}^{n} Y_i \Big/ n = (7 + 9 + 9 + 3 + 6 + 5 + 0 + 7 + 0 + 7) / 10 = 53/10 = 5.3$.

Other measures of central tendency

MEDIAN – the central-most observation in a ranked (ordered or sorted) set of observations. If the number of observations is even, take the mean of the center most 2 observations

Example: for the fish sample used earlier, rank the observations

$Y_i = 0, 0, 3, 5, 6, 7, 7, 7, 9, 9$

If a single observation was in the center it would be used as the median. In this case the number of observations is even and the center falls between two numbers, 6 and 7, so calculate the mean of those two numbers.

MEDIAN = (6 + 7) / 2 = 6.5

MODE – the value of the most frequently occurring observation

Example: For the fish sample,

$Y = 0, 0, 3, 5, 6, 7, 7, 7, 9, 9$

The most frequently occurring value was "7".

Therefore, the MODE = 7

MIDRANGE – average of the largest and smallest observation.

Example: The smallest observation in the fish sample was 0 and the largest was 9. The midrange is calculated as the midpoint between these values

MIDRANGE = (0 + 9) / 2 = 4.5

Do not make the mistake of subtracting the lower value from the higher value and dividing by 2. That would be half of the RANGE, not the MIDRANGE.

Percentiles and Quartiles

Percentiles – the value of an observation that has a given percent of the observations below that value and the remaining observations above that value.

The 50th percentile is the value where 50% of the sample observations would have values below it and 50% would be above it. This is also known as the median.

It is often useful to know what value has 5% of the observations below it and 95% above it. This is the 5th percentile. Conversely the 95th percentile is the observation whose value exceeds 95% of the observations in the data set and is exceeded by 5% of the values.

Likewise, the value of the 75th percentile would have 75% of the observations below the value and 25% above.

Quartiles – observations that have one, two or three quarters of the observations above and below their value.

The first quartile is the value of the observation that has one quarter of the observations below it and three quarters above the value. It is the 25th percentile.

The second quartile is the value of the observation that has half (two quarters) of the observations below and above. This value is the same as the MEDIAN or 50th percentile.

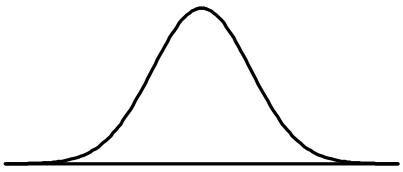
The third quartile is the value of the observation that has three quarters of the observations below and one quarter of the observations above the value. It is the 75th percentile.

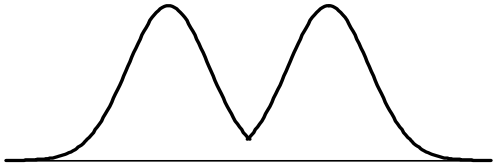
Which measure of Central Tendency is best?

This depends on the distribution. If the distribution is monomodal and symmetric then the

MEAN = MEDIAN = MODE = MIDRANGE

This is true for the NORMAL bell-shaped curve.



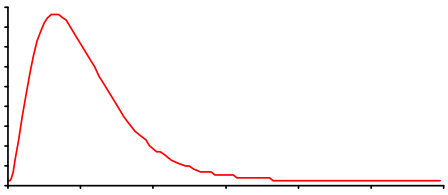


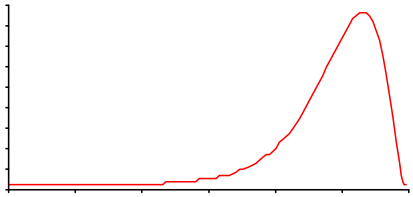
Bimodal distributions are not well described by any measure of central tendency, particularly a single MODE.

Asymmetrical distributions may be best described by the MEDIAN or MODE, depending on the objectives.

POSITIVE SKEW (the long tail of the distribution to the right)

Mode < MEDIAN < Mean





NEGATIVE SKEW

Mean < MEDIAN < Mode

Relative positions of the MEAN, MEDIAN and MODE for asymmetric distributions

- The MEAN is closest to the drawn out tail of the distribution.
- The MODE is farthest from the tail.
- The MEDIAN is intermediate.

Statistic	Negatively Skewed	Symmetric	Positively Skewed
MODE	Largest	Middle	Smallest
MEDIAN	Middle	Middle	Middle
MEAN	Smallest	Middle	Largest

Selecting a measure of central tendency

The MEAN is generally preferred because:

- it utilizes all information in the data set
- it is widely recognized and is easy to work with
- the distribution of the means tends to be normally distributed even if the original observations are not.
- it is generally more sensitive to changes in the form of the distribution (e.g. asymmetry) though this is not always an advantage.

The MEDIAN or MODE may be desirable for asymmetric data sets and may give a more representative measure of location of the center of the data.

Example: Find a "typical" salary for a business employing 5 individuals.

The salaries are: $100,000 $30,000 $20,000 $15,000 $10,000

MEAN = $175000 / 5 = $35,000

MEDIAN = $20,000

MODE – there is no mode unless the data are arbitrarily grouped, and if grouped, two different groupings may not give the same mode

Salary Interval	Frequency	
1 – 10000	1	
10001 – 20000	2	<=== MODE here
20001 – 30000	1	
30001 and over	1	

Since the MEDIAN and the MODE do not use all of the information in the data set for calculation, so they are less sensitive to change.

For example if the top person above gets a raise to $200,000 the MEDIAN and MODE do not change. This can be either an advantage or a disadvantage, depending on the objectives. However, the MEAN would increase from $35,000 to $75, 000.

Parametric statistics

In this course we will examine primarily "parametric" statistical techniques. These techniques generally assume that the data conforms to a normal, bell-shaped curve. This distribution is referred to as the normal distribution or Gaussian distribution.

We are able to assume that these techniques are adequate under the following conditions.

The distribution of the data is normal or "approximately" normal (e.g. symmetric, bell shaped) since the parametric techniques are robust to violations of the assumption of normality.

The sample size is large and hypotheses of the means are to be tested, since the means will tend to be normally distributed even if the original observations are not.

The distribution is known, or can be determined from a large sample, and can therefore be transformed to approximate normality.

For example, the number of individuals in many biological situations is commonly distributed as a negative binomial. The original observations can be transformed by taking logarithms to approximate a normal distribution.

Other types of means

ARITHMETIC means: no transformation

GEOMETRIC means: result from a logarithmic transformation

$GM(Y_i) = n^{th}$ root of $(Y_1 * Y_2 * Y_3 * Y_4 * ... * Y_n) = \sqrt[n]{Y_1 \times Y_2 \times Y_3 \times Y_4 \times ... \times Y_n}$ or

$$\exp\left(\left(\log(Y_1) + \log(Y_2) + ... + \log(Y_n)\right)/n\right) = e^{\frac{\log(Y_1)+\log(Y_2)+\log(Y_3)+...+\log(Y_n)}{n}}$$

HARMONIC means: result from an inverse transformation

$$HM(Y_i) = INV(\{1/Y_1+1/Y_2+1/Y_3+1/Y_4...+1/Y_n\}/n) =$$

$$1\Big/\left(\left(1/Y_1 + 1/Y_2 + 1/Y_3 + 1/Y_4 + ... + 1/Y_n\right)/n\right)$$

Harmonic mean – used for particular cases where the probability of being sampled is an inverse function of the variable of interest.

Suppose we want to calculate the mean time a fisherman spends fishing on a lake. The 9 fishermen are as follows. Eight fishermen each fish for one hour, one starting at each hour from 8AM to 3PM. The ninth fisherman fishes for 8 hours from 8AM to 3PM. In order to determine the amount of fishing effort a biologist from wildlife and fisheries goes to the lake once and interviews all available fishermen, asking "How long will you fish today?" He then calculates the "average effort".

Time	7 am	8 am	9 am	10 am	11 am	noon	1 pm	2 pm
Fisherman #	1	2	3	4	5	6	7	8
Fisherman #	9	9	9	9	9	9	9	9

e.g. What is the average effort by fishermen in a day?

Since 8 fishermen go for 1 hour and one fisherman goes fishing for 8 hours the total effort is actually 16 hours and the true mean is 16/9=1.778 hours. However, a

biologist visiting once, at any given hour, will meet only 2 fishermen, one fishing for one hour and one fishing for 8 hours.

The arithmetic mean of the sample = (8+1)/2 = 4.5 hours

The harmonic mean of the sample = 1/((0.125+1)/2) =1.778

This type of mean is appropriate for samples where the probability of being included in the sample is a function of the value being measured.

Summary

The true population mean is denoted μ, the greek letter mu.

The sample estimate of the mean is denoted $\bar{Y}$, and is called "y-bar".

Remember always that our sample estimate of the mean is just one of many possible samples. Each sample is an attempt to estimate the true population mean. Our sample may be one of the good ones, pretty close to the true population mean, or it may be one of the not so good ones. We won't really know.

How good is our estimate from the sample? This depends on how good our sample is and on how much variability there is in the population.

Our best guarantee of getting a good sample is to sample at random. This should at least give us a representative sample of the population.

Variability is the other big problem in sampling, so we need to estimate how variable the population is, our next topic.

Applications for other types of means

Arithmetic mean – the usual case

Geometric mean – Used as a transformation for some non-normal distributions, particularly the negative binomial, a strongly skewed distribution.

Harmonic mean – Used for particular cases where the probability of being sampled is an inverse function of the variable of interest.

SAS example (#1a) from Freund & Wilson (1997) Table 1.1

```
PROC UNIVARIATE DATA=HouseSales PLOT; VAR SP;
    TITLE4 'Proc Univariate of house sales price'; RUN;
```

See SAS output for results

Probability distributions

PROBABILITY – a measure of the likelihood of the occurrence of some event

An event can be any outcome (e.g. verbal, mathematical or graphical)

Some rules of Probability

If an event (A) is certain to occur, the probability is 1 (one, unity), so P(A) = 1

If the event is certain to NOT occur, the probability is 0 (zero, null), so P(A) = 0

The probability of an event will always range be between 0 and 1 (inclusive). $0 \leq P(A) \leq 1$