AnSc 5403

Biometry

Lecture Notes 2

I. Some things about sums

A. Addition is a very common operation in statistics

1. Shortcut notation is used because of the frequent use of addition (summation)

a. In statistical operations, the Greek capital symbol $\sum$ is used as the summation sign

(a) Example - $\sum x$ = the summation of x

(b) However, this notation needs further definition to avoid confusion – the specific x's that are to be summed need to be defined

b. Suppose we want to add $x_1 + x_2 + x_3 + x_4 + \ldots x_n$

(a) The shorthand notation for this summation is: $\sum_{i=1}^{n}$

(i) This notation is read as – "the sum of x_i, where i ranges from 1 to n."

1. i =1 below the $\sum$ symbol defines the first value of x_i

2. The n written above the $\sum$ symbol defines the last value of x_i

B. Properties of $\sum$

1. Property 1 - $\sum_{i=1}^{n} \quad = \sum_{i=1}^{n} \quad + \quad \sum_{i=1}^{n}$

2. Property 1a - $\sum_{i=1}^{n} \quad = \sum_{i=1}^{n} \quad - \quad \sum_{i=1}^{n}$

3. Property 2 - $\sum_{i=1}^{n} \quad =$, where c is a constant

4. Property 3 - $\sum_{i=1}^{n} \quad = (\sum_{i=1}^{n} \quad) +$, where c is a constant

5. Property 4 - $\sum_{i=1}^{n} \quad = \quad \sum_{i=1}^{n}$, where c is a constant

C. Double sums

1. Suppose we have the following data:

   a. Body weight data by treatment and replicate

	Treatments				
Replicate	A	B	C	D	E
1	750	752	755	748	749
2	740	742	747	742	741
3	749	750	751	755	754
4	755	744	748	752	757
5	748	747	745	744	749

   b. These data are classified in two ways (replicate and treatment)

   c. To indicate the manner in which summations are to be done , we need to use a system of double subscripts – one to identify the rows (replicates) and one to identify the columns (treatments)

      (a) The letters i and j are commonly used as subscripts with a double summation

         (i) The letter i is used to designate rows and the letter j is used to designate columns

            1. The general notation for any data point would be x_{ij}

            2. In the data table above, $x_{12} = 752$, and $x_{55} = 749$

   d. If we sum all the points in the data table, the notation for this double sum is written as:

      (a) $\sum_{i=1}^{r} \sum_{j=1}^{c}$

   e. The same principle could be applied to additional (e.g., three-way) classifications of data

II. Descriptive statistics – Measures of central tendency

A. Groups of observations are often conveniently described by a single number

1. Example –

B. Central values, most frequent values, and middle values are most commonly used because they would most likely represent the typical value in a group of observations

C. Measures of central tendency –

1. Most common -

2. Less common –

D. Arithmetic mean

1. Probably the most commonly used measure of central tendency

2. Calculated as the sum of all observations divided by the number of observations

a. Typically designated as - $\bar{x}$

b. Calculated as – $(\sum_{i=1}^{n} x_i)/n$, where n = the total number of observations

c. Example data – 4, 5, 7, 9, 2

(a) $\bar{x}$ = 27/5 = 5.4

E. Median

1. Definition –

2. Example data – 4, 5, 7, 9, 2

a. Ordered by size – 2, 4, 5, 7, 9 (Don't forget to order the values!)

b. Value in the middle position = median = 5

3. What happens when we have an even number of observations?

a. Example data set with an even number of observations – 2, 2, 4, 5, 7, 9

(a) There is no middle observation; there are two middle observations (4 and 5)

(b) Median in this case = the value halfway between the two middle observations

(i) In this example, the median = *4.5*

F. Mode

1. Definition – the most frequently occurring observation in a set of observations

2. Example data – 2, 2, 4, 5, 7, 9

   a. Mode = 2

   b. But for the data set – 2, 4, 5, 7, 9 – there is no mode!

   c. For the data set 2, 2, 4, 4, 5, 7, 9 –

G. Comparison of the mean, median, and mode

1. Data set – 2, 2, 4, 5, 7, 9

   a. Mean = 4.833; Median = 4.5

   b. If we change the last data point from 9 to 18 –

      (a) The mean is sensitive to extremes – the median is not

2. For some data sets, the mean can give a misleading picture of the observations

   a. Example – 2, 2, 2, 2, 17– Mean (5) is not very representative of the data set

3. The median sometimes ignores potentially useful information because only the middle value (or two middle values) affect the median

4. The mode can sometimes be useful, but it tends to characterize individuals more than groups

   a. Example use of the mode – survey of consulting nutritionists regarding crude protein levels in diets

      (a) – The mode might be more meaningful than the mean to describe nutritional practices of the majority of consultants

III. Other measures of central tendency

A. Mid-range – the value midway between he smallest and largest values in the data set

1. Example data set - 2, 4, 5, 7, 9 – mid-range = (2 + 9)/2 = 5.5

B. Geometric mean – the n^{th} root of the product of the observations in the data set

1. Example data set – 1 3, 9 = geometric mean = *3*

   a. A property of the geometric mean is that the $\log_{10}$ of the geometric mean equals the mean of the logs of the individual observations

      (a) Often used in microbiology and immunology for computing the average dilution of titers

      (b) Often used to describe particle size distributions

C. Harmonic mean – the reciprocal of the arithmetic mean of the reciprocals of the observations

1. Formula – $n/(\sum_{i=1}^{n} 1/x_i)$

2. How is this used?

   a. Example – a car travels 20 miles at 40 mph and 20 more miles at 50 mph

      (a) Average velocity = total distance/total time (40 miles/[0.5 + 0.4] = 44.44 mph), which does not equal the simple average of 40 and 50 mph (e.g., 45 mph)

      (b) The harmonic mean = 2/(1/40+1/50) = 44.44 mph

D. Weighted mean – the arithmetic mean adjusted for the weights of observations

1. Formula - $\sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$, where w_i is the weight for each observation

2. Example – grade point average

3. Example – Dry matter intake by cattle in two 28-d periods and one 14-d period was 20, 21, and 22 lb/d. What is the average dry matter intake for the 70-d period?

   a. Answer =