جامعة المستقبل

الاحصاء الحيوي /المحاضرة السادسة و السابعة

الفيزياء الطبية /المرحلة الثانية

The Sampling Distribution of the Sample Mean

دلائل سعد عبد الزهرة –صفا محمد هادي

The Sampling Distribution of the Sample Mean

Suppose that a variable x of a population has mean, μ and standard deviation, Then, for samples of size n,

1) The mean of $\bar{x}$ equals the population mean, μ, in other words: $\mu\bar{x} = \mu$

2) The standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the sample size, in other words: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

3) If x is normally distributed, so is $\bar{x}$, regardless of sample size

4) If the sample size is large (n > 30), $\bar{x}$ is approximately normally distributed, regardless of the distribution of x.

Examples:

1) The times that college students spend studying per week have a distribution that is right skewed with a mean of 8.4 hours and a standard deviation of 2.7 hours. Suppose a random sample of 45 students is selected.

a) What is the sampling distribution of the mean number of hours these 45 students spend studying per week? (i.e., What is the sampling distribution of $\bar{x}$?)
$n =$ _____ $\mu =$ _____ $\sigma =$ _____

b) Find the probability that the mean time spent studying per week is between 8 and 9 hours. Find the probability that the mean time spent studying per week is greater than 9.5 hours.

2) At an urban hospital the weights of newborn babies are normally distributed, with a mean of 7.2 pounds and standard deviation of 1.2 pounds. Suppose a random sample of 30 is selected.

a) What is the sampling distribution of the mean weight of these newborn babies? (i.e. What is the sampling distribution of $\bar{x}$?)

b) Find the probability the mean weight is less than 6.9 pounds?

c) Find the probability the mean weight is between 6.5 and 7.5 pounds?

d) Find the probability the mean weight is greater than 8 pounds?

3) A battery manufacturer claims that the lifetime of a certain battery has a mean of 40 hours and a standard deviation of 5 hours. A simple random sample of 100 batteries is selected.

a) What is the sampling distribution of the mean life of the batteries?
(i.e. What is the sampling distribution of $\bar{x}$?)

b) What is the probability the mean life is less than 38.5? Would this be unusual?

WEIGHTED MEANS AND MEANS AS WEIGHTED SUMS

In the Speeds Problem we saw that there is more than one kind of "average." In this handout, we will explore this topic further. The ordinary mean is sometimes called the "arithmetic mean" to distinguish it from other types of means.

** The most common way to think of the average (arithmetic mean) of numbers is to add them up and divide by the total number of summands:

e.g., the average of $1,1,2,3,4,4,4$ *is* $(1+1+2+3+4+4+4)/7$

But we could write these two other ways:

1. **"Distributing" the denominator gives**

$$(1/7)1 + (1/7)1 + (1/7)2 + (1/7)3 + (1/7)4 + (1/7)4 + (1/7)4.$$

Thus, we have the mean as a sum of coefficients times the original numbers in the list. *Note that the sum of the coefficients is 1*.

2. **Collecting like terms gives**

$$(2x1 + 2 + 3 + 3x4)/7 = (2/7)1 + (1/7)2 + (1/7)3 + (3/7)4.$$

Now we have a sum of coefficients times the *distinct* values (not allowing repetitions) in the original list of numbers. The coefficient of a value is the *proportion* of that value in the original list of numbers. We still have the coefficients adding to 1, but they are no longer all the same. We now see the mean as a *weighted sum* of the *distinct values*, where each value is weighted according to its proportion in the total list of numbers. This perspective prompts two generalizations of the arithmetic mean.

A. Weighted Means

To form a *weighted mean* of numbers, we first multiply each number by a number ("weight") for that number, then add up all the weighted numbers, then divide by the sum of the weights. We often do this in computing course

grades – e.g., weighting the final exam twice as much as a midterm exam. The ordinary (arithmetic) mean is a weighted mean with all weights equal to 1. Another way to describe a *weighted mean* of a list of numbers is a sum of coefficients times the numbers, where the coefficients add up to 1. In this case, the **coefficients** are called the *weights*. (*Note the ambiguity in the use of "weight".)* If all the weights are the same, we get the ordinary arithmetic mean.

***Why are these two descriptions equivalent*?**

Examples of weighted means:

1. The discussion above shows that the ordinary (arithmetic) mean can also be considered as a *weighted mean* of the *distinct values* being averaged, with the weight of a value being its proportion in the original list of numbers being averaged.

2. In part (b) of the Average Speeds Problem (Problem 2 in the handout "What Do You Mean by Average?"), the average speed can be written as a weighted mean:

$$Average\ speed = \frac{a_1v_1 + a_2v_2 + \ \ldots\ + a_nv_n}{a_1 + a_2 + \cdots + a_n}$$

$$= \frac{a_1}{a_1 + a_2 + \cdots + a_n}\ v_1 + \frac{a_2}{a_1 + a_2 + \cdots + a_n}\ v_2 + \cdots + \frac{a_2}{a_1 + a_2 + \cdots + a_n}\ v_2$$

$$w_1v_1 + w_2v_2 + \quad \ldots \ + w_nv_n$$

where

$$w_i = \frac{a_1}{a_1 + a_2 + \cdots + a_n}$$

Note that the sum of the w_i's is 1. Thus, the answer to part b in the Average Speeds Problem can be seen as *a weighted mean of the original speeds, with the weight of each speed being the fraction (proportion) of the total number of intervals that are traveled at that speed.*

3. Another place where weighted means are important is when the purpose of the study is to compare means of two groups, but the two groups are appreciably different in size. Consider for example, a study whose purpose is to compare the educational and workforce experiences of male and female

electrical engineers. There are many fewer women in electrical engineering than men, so a simple random sample of all engineers in the population would include very few women, and therefore not give as good estimates for the women as the men. Instead, the researchers would use a *stratified* sample – they might, for example, sample 200 men and 200 women. But then if they want a *sample* "average" that estimates the average for *all* electrical engineers, including both men and women, they need to take a weighted average.

***Example**:* Suppose that the total number of men in the *population* being studied (e.g., *all* electrical engineers) is N_M and the total number of women in the *population* is N_w. If 200 of each sex are sampled, what would be appropriate weights for calculating an average of some variable (e.g., salary; number of years in the profession) on which data are collected, if the intent is to estimate the average for the *entire population* of interest (e.g., all electrical engineers)?

B. **Means of Random Variables** Viewing the mean of a list of (not necessarily distinct) numbers (e.g., exam scores) as a weighted mean of the distinct values occurring in the list prompts us to define the *mean of a discrete numerical random variable* as

$$Mean\ of\ X\ = \sum x\, f(x)$$

where the sum is over all values that X can take on.

***Example*:** You put two more dots on the "one" side of a fair die to make it into a "three". X is the random variable "number that comes up when you roll the die." Then the mean of X is:

$$(1/6)2\ +\ (2/6)3\ +\ (1/6)4\ +\ (1/6)5\ +\ (1/6)6$$

We can extend this idea to continuous random variables by using an integral instead of a sum: The *mean of a continuous random variable* as

$$Mean\ of\ X\ = \int_{-x}^{x} xfX\,(x)dx$$

The mean of a random variable is also called the *expected value* or the *expectation* of the random variable, and denoted $E(X)$.

C. **Harmonic Means and Weighted Means**

i. In part (c) of the Average Speeds Problem, you showed that if you travel a certain distance at speed v_1, then the same distance at speed v_2, and so forth, finishing by traveling that same distance at speed v_n, your average speed for the whole trip is:

$$v_n = \frac{n}{\frac{1}{v_1} + \frac{1}{v_2} + \cdots + \frac{1}{v_n}}$$

The expression on the right is called the *harmonic mean* of $v1, v2, \ldots, vn$. The harmonic mean of a set of numbers can be described as: The reciprocal of the arithmetic mean of the reciprocals of the numbers. Usually, it is just defined for positive numbers. (WHY?)

The name "harmonic" presumably comes from the fact that in the harmonic series 1 + ½ + 1/3 + ¼ + …, each term is the harmonic mean of the preceding and the following term. (Check it out algebraically!). The harmonic series is so named because it is connected to what is called the "harmonic series" .

Problems:

1) a. Calculate and compare the arithmetic and harmonic means of some numbers in several cases and form a conjecture as to whether the harmonic mean is always less than, always greater than, or sometimes less than and sometimes greater than the arithmetic mean.

b. Prove your conjecture in the case n = 2 (that is, means of just two numbers). [Note: The general case is harder; we may come back and do it later.]

2) The arithmetic and harmonic means can both be considered "measures of center." The median is another measure of center. Recall that the *median* of a set of numbers is the number in the middle when the numbers are listed in order (or the average of the two middle numbers if the number is even.) So, the median of the numbers 1,1,2,3, 4,4,4 is 3 (and would still be 3 if the numbers were rearranged), and the median of the numbers 1,1,2,3, 4,4 is (2+3)/2 = 2.5 (and would be the same if the numbers were rearranged). So, the median has equal numbers (of the numbers in the original list, counting each one as many times as it occurred in the original list) on either side of it. This suggests how to define the *median of a continuous random variable*: The number with equal probability of the random variable occurring on either side of that number. That is, the median M is a number with the property

$P(X<M) = P(X>M)$.

a. Use a symmetry argument to find the median of the uniform and normal distributions.

b. Figure out the median of distribution #3 in part I.

c. Estimate the median of the empirical distribution whose histogram is shown: