MATH 545
DR. MCLOUGHLIN'S CLASS
MEASURES OF CENTRAL TENDENCY OR LOCATION
HANDOUT II

Now, we are studying probability and statistics. We are considering random variables as defined in handout 2, 3, and 5 as well as the book and other books. What we have to determine when we are doing practical applications of the theory of probability and statistics is to what degree can we adequately justify the statistics used and that requires understanding the scale used. Returning to the baseball jersey discussion[1] (3 for Babe Ruth, 44 for Hank Aaron, 42 for Tom Seaver, 6 for Lou Gehrig) let us suppose we put baseball cards in an urn such that each was equiprobable with the aforementioned four players on each card. Let X be the player number. So, we have

X = x	3	8	42	44
Pr(X = x)	.25	.25	.25	.25

Now, calculating E[X] we get $\sum_x x \cdot p_x(x) = 3 \cdot \frac{1}{4} + 8 \cdot \frac{1}{4} + 42 \cdot \frac{1}{4} + 44 \cdot \frac{1}{4} = 24.25.$

This μ is meaningless. It is nonsense. One can calculate it but in the practical application work that most of you will do when you graduate (not *if,* **when** – get that idea and make it a part of your subconscious) you have to consider this *before* doing statistical analyses. So, a subject that one can study (post this course) is the types of measurement scales, what statistics are proper for which type of scale, etc.

So, we return to the contention that the main point to studying probability and statistics is the analysis of data. So, when one is studying variables it is important to understand that the types of measurement scales and what statistics are proper for which type of scale. So let us consider some statistics which are not a part of the course (per se) but we should be comfortable computing.

What we saw in the appendix to chapter 13, when we have a binomial mass function there is not a need to estimate E[X]. It is whatever it is for the distribution we are considering. So, when X ~ Bin(x, ¼, 3) we have the p.m.f. :

X = x	0	1	2	3
Pr(X = x)	0.421875	0.421875	0.140625	0.015625

Case close! Since that case is closed the so is it for E[X] (provided one has proven when X ~ Bin(x, p, n) E[X] = μ = np or one uses the definition of E[X] for a discrete random variable) μ = ¾ .

[1] Thanks to Mr. David for correcting me on Babe Ruth's jersey number.

Recall:
If X is a discrete random variable and the function given by $f(x) = \Pr(X = x)$ for each x in the domain of the function is the p. m. f. at x, then the expected value (or mean) of X is $E[X] = \sum_{x} x \cdot f(x)$.
There are estimators for μ and there are other measures of central tendency.

Definition 1: Let X be a random variable. The **mode of X** is the value of X, say x_0, where the p.d.f. or p.m.f. is the maximum value. Denote the mode as m_o.

Definition 2: Let X be a random variable. The **median of X** is the value of X, say x_1, where the $\Pr(X \geq x_1) = \Pr(X \leq x_1)$. Denote the median as m_d.

Definition 3: Let X be a random variable with a probability mass or density function. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X. The **arithmetic mean of the sample** is the value $\overline{X}$ where $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$.

Definition 4: Let X be a random variable with a probability mass or density function. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X. Let $w_1, w_2, w_3, \ldots, w_n$ be real numbers signifying the 'importance' of $X_1, X_2, X_3, \ldots, X_n$ respectively. The **weighted arithmetic mean of the sample** is the value $\overline{WX}$ where $\overline{WX} = \frac{\sum_{j=1}^{n} (w_j \cdot X_j)}{\sum_{j=1}^{n} w_j}$.

Definition 5: Let X be a random variable with a probability mass or density function. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X. The **mode of the sample** is the value of X which occurs most frequently such that there is at least one value that occurs with lesser frequency.

Definition 6: Let X be a random variable with a probability mass or density function. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X. The **median of the sample** is the value of X which occurs in the centre of ordered values of the sample if there are an odd number of values and it is the arithmetic mean of the centre two values if there are an even number of values.

Definition 7: Let X be a random variable with a probability mass or density function where $X > 0$ for all X where $X > 0$ for all X. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X. The **geometric mean of the sample** is the value G where $G = \sqrt[n]{\prod_{j=1}^{n} X_j}$.

Definition 8: Let X be a random variable with a probability mass or density function. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X where $X \neq 0$ for all X.. The **harmonic mean of the sample** is the value H where $H = \dfrac{n}{\sum_{j=1}^{n} \dfrac{1}{X_j}}$.

There are others kinds of means used in different applied areas:

Definition 9: Let X be a random variable with a probability mass or density function. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X where $X > 0$ for all X. The **quadratic mean of the sample** (or the **root mean squared**) is the value R where $R = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j^2)}$.

Definition 10: Let X be a random variable with a probability mass or density function. Let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X where $X > 0$ for all X. The **p-power mean of the sample** (there are many for different values of p (or the **Hölder mean**)) is the value P where $P = \sqrt[p]{\frac{1}{n}\sum_{j=1}^{n}(X_j^p)}$.

Notice for p = 2 it is the root mean squared.
Notice for p = 1 it is the arithmetic mean.
Notice for p = 0 it is the geometric mean.
Notice for p = -1 it is the harmonic mean.

Definition 11: Let X be a random variable with a probability mass or density function where $X > 0$ for all X. Let X_1, X_2 be a random sample ($X_1 \neq X_2$) of size 2 for X where $X > 0$ for all X. The **Heronian mean of the sample** is the value H_E where $H_E = \frac{1}{3}(X_1 + \sqrt{X_1 \cdot X_2} + X_3)$.

Definition 12: Let X be a random variable with a probability mass or density function where $X > 0$ for all X. Let X_1, X_2 be a random sample ($X_1 \neq X_2$) of size 2 for X where $X > 0$ for all X. The **logarithmic mean of the sample** is the value L where

$$L = \frac{X_1 - X_2}{\ln(X_1) - \ln(X_2)} = \frac{X_1 - X_2}{\ln(\frac{X_1}{X_2})}.$$

Definition 13: Let X be a random variable with a probability mass or density function where $X > 0$ for all X. Let X_1, X_2 be a random sample ($X_1 \neq X_2$) of size 2 for X where $X > 0$ for all X. The **identric mean of the sample** is the value I where

$$I = \frac{1}{e} \cdot \left(\frac{(X_1)^{X_1}}{(X_2)^{X_2}} \right)^{\frac{1}{X_1 - X_2}}$$

Definition 14: Let X be a random variable with a probability mass or density function where $X > 0$ for all X. Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample (all distinct) for X where $X > 0$ for all X. Seppo Mustonen of the Department of Statistics at the University of Helsinki defines the **logarithmic mean of the sample** is the value L where

$$L = (n-1)! \sum_{i=1}^{n} \left(\frac{X_i}{\prod_{\substack{j=1 \\ j \neq i}}^{n} \ln(\frac{X_i}{X_j})} \right).$$

PHEW! Ridiculous. Needless to say, there are many estimators.

Computational Examples:

Suppose the X_i are given by $X_1 = 0.1$, $X_2 = 0.1$, $X_3 = 0.1$, and $X_4 = 10$.

$\overline{X} = 2.575$, $G = \sqrt[4]{\frac{1}{100}}$, $G \approx 0.316227766$, $H = \frac{40}{301}$, $H \approx 0.13$, $m_o = 0.1$, $m_d = 0.1$, etc.

Suppose the X_i are given by $X_1 = 0.1$, $X_2 = 10$, $X_3 = 10$, and $X_4 = 10$.

$\overline{X} = 7.525$, $G = \sqrt[4]{100}$, $G \approx 3.16227766$, $H = \frac{40}{103}$, $H \approx 0.39$ $m_o = 10$, $m_d = 10$, etc.

Suppose the X_i are given by $X_1 = 1$, $X_2 = 2$, $X_3 = 3$, and $X_4 = 4$.

$\overline{X} = 2.5$, $G = \sqrt[4]{24}$, $G \approx 2.2133638394$, $H = 0.48$, m_0 does not exist, $m_d = 2.5$, etc.

The geometric, logarithmic, and harmonic means are used by Chemists and Chemical Engineers. The geometric mean is used when averaging numbers that involve growth data. The mean relative volatility between two components in a distillation column is based on the geometric mean. The logarithmic mean is used in the heat transfer to determine the mean temperature difference between hot and cold streams flowing in a heat exchanger. The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. It is used, for example, to calculate the mean overall heat transfer coefficient, V_o, in a heat exchanger.

"Many wastewater dischargers, as well as regulators who monitor swimming beaches and shellfish areas, must test for and report fecal coliform bacteria concentrations. Often, the data must be summarized as a "geometric mean" (a type of average) of all the test results obtained during a reporting period. Typically, public health regulations identify a precise geometric mean concentration at which shellfish beds or swimming beaches must be closed.

A geometric mean, unlike an arithmetic mean, tends to dampen the effect of very high or low values, which might bias the mean if a straight average (arithmetic mean) were calculated. This is helpful when analyzing bacteria concentrations, because levels may vary anywhere from 10 to 10,000 fold over a given period." - Dr. Joe Costa, Buzzards Bay Project

I notice on the State Department web-site there was a tutorial on bio-terrorism which mentions the geometric mean.

The Heronian arises in the determination of the volume of a pyramidal frustum.

For further investigation:

Bullen, P. S.; Mitrinovic, D. S.; and Vasic, P. M. (1988). *Means and Their Inequalities.*

Dordrecht, Netherlands: Reidel.

Last revised 2 Oct. 2006