Introduction

Some years ago I separated from my spouse of 29 years and my friends, concerned about my singular status, got into the habit of arranging the "dinner party". Low and behold I would turn up to the event to find sitting opposite me someone of the opposite sex in a similar singular state to myself. Clearly it was my responsibility in this arrangement to determine whether or not I liked the person (I invariably did, but of course I could not afford to be too choosy!) and whether or not they liked me. Over the course of dinner I would watch and listen to the person's reactions to what I had to say. It was an information collection exercise and at some stage I had to arrange what I had observed, synthesize it if you like, into a form that would enable me to decide whether or not she liked me. All I could do was infer whether or not she liked me and of course my inference could have been completely off the mark. Essentially in this problem I faced four possible outcomes and I was by no means indifferent between them. I could infer that she liked me when she really did like me, which of course was a good outcome (we would probably ride off happily into the sunset!). Alternatively I could infer that she did not like me when she really did not like me, which too is a good outcome (we could enjoy the rest of the meal and go our separate ways). Another possibility is that I would decide that she did not like me when she really did (this is bad news and corresponds to an opportunity for a lifetime of happiness foregone!). Finally I could decide that she liked me when really she did not (this too is bad news, the worst outcome in my mind, it was just plain embarrassing!). Consequently I would require more than just "on balance" evidence to convince myself that she liked me in order to avoid the fourth outcome.

So what has this got to do with statistics? The answer is…everything! The dinner party was the very essence of the statistical process. Statistics is about dealing with uncertainty, it is about using things we observe and know to learn and understand more about things that we cannot observe, do not know and about which we are uncertain. It is about handling information, deciding what is relevant and what is not, understanding the nature of the probability of events, synthesizing the relevant material and arranging it into a fashion that is useful for making choices in an uncertain environment. Ultimately it is about making those choices in a way that holds the chance of making least preferred mistakes at some acceptable level.

The things we know and are going to use come in the form of theory and data. The theory, which will be developed throughout the course, is based upon ideas about the probability of events. It will tell us about the properties of the techniques we are going to use to synthesize the data, it will tell us how to best organize the decision process, it will tell us the nature of uncertainty under certain circumstances and it will tell us how to best collect and organize the data so that it will be informative with respect to the object of our interest. Information or data comes to us in many shapes and sizes, much of what the social scientist uses is what is referred to as qualitative data, it is essentially non-numeric. Someone's gender, the schooling level they reached, their hair or eye color are all non-numeric characteristics that constitute data on that person. Typically for analysis purposes we need to convert non-numeric data into numeric data. Numeric data comes in two fundamental forms, discrete (or integer) form and continuously measured form. This is a slight nuisance since we shall need to know a bit about the mathematics of cumulation that is appropriate for each form (essentially it is the mathematics of summation for discrete data and the mathematics of integration for continuous data).

Data is organized in terms of **observations** on **variables**. Part of my current research has to do with the relationship between a husband's and wife's educational attainments. The object of observation is a married couple and the characteristic for each spouse I'm interested in is the stage at which they ceased schooling which is represented by an integer (e.g. grade 10 = 1, grade 12 = 2, college 1st year = 3, college completed = 4, post graduate education = 5..etc.). In this study the data is the information on a collection of married couples, the variables are the educational attainment of the husband and the educational attainment of his wife and an observation is the attainment values for a particular husband and wife pairing. Sometimes problems involve just one variable sometimes they involve more than one variable. In the first part of the course we shall just deal with problems involving one variable, later we will deal with multivariate problems.

Data Summary

When presented with a large amount of data in the form of a collection of numbers relevant to a problem, we need ways of describing various aspects of those numbers as a way of summarizing what the data are telling us. Statisticians use summary statistics to describe features of, or summarize, the data that may be of interest in a particular problem. We shall concern ourselves primarily with two types of summary or descriptive measures that will be the main focus of this course, namely Location measures and Dispersion measures. Note however that there are many more, especially when we consider problems that concern more than one variable, for then we shall need summary statistics that describe relationships between variables these will be introduced later in the course when we deal with multivariate problems.

Location Measures

Location measures focus on representing where the observations are centered in some sense. There are three primary location measures, the mean, the median and the mode. All of these location measures are in the same units of measurement as X. For formulaic purposes suppose we have n observations X_i, i=1, …, n that for convenience have been arranged in rank order so that $X_1 \leq X_2 \leq \ldots \leq X_n$.

The Mean

Perhaps the most common measure of location, the mean is the arithmetic average of a collection of numbers. The first thing students ask after a test is "what was the class average?" largely because it gives a sense of how the class did overall and, given their own mark, how well they did relative to the class. For example in a class of five students whose grades were 45, 55, 60, 70 and 80 the mean grade would be 62. Formally:

The Mean:

$$\overline{X} = \tfrac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

The Median

In a very real sense the median is "the middle number" in a collection of numbers, in words it is a value for which 50% of the numbers in a collection of numbers are less than equal to it and 50% of them are greater than or equal to it. Formally the median, denoted $X_{med,}$ is a value such that for n odd $X_{med} = X_{(n+1)/2}$ and for X even $X_{(n/2)} < X_{med} < X_{(n/2+1)}$. (the latter case obviously results in a range of numbers, usually X_{med} is set to $(X_{(n/2)} + X_{(n/2+1)}/2)$. Unlike the mean the median is much less influenced by extreme values in the collection of X's. For example in the set of 5 class marks {45, 55, 60, 70, 80}, the median would be 60 (recall the mean was 62), now suppose the 80 was remarked and became a 90, the median remains 60 however the mean has now become 64. The median is a particular case of a quantile value of X, in this case the 50th percentile, a value such that 50% of the X's are less than or equal to that value.

The Mode

The center of the range of most frequently observed X. Arranging the data in the form of a histogram by splitting up the range of values of the X's into segments or cells and counting the proportion falling in each cell, the modal cell corresponds to the one with the largest proportion in it. It is a popular measure with marketers who deal with products sold by size or shape where the X's correspond to a list of product sold since it indicates the most frequently sold size or shape. We shall not be using the mode much in our analysis.

Theoretically, for large collections of data with one interior modal point, the three location measures obey one of two inequalities, either 1) Mean ≤ Median ≤ Mode or 2) Mean ≥ Median ≥ Mode. Which inequality prevails will depend upon how the data is arranged around the mean. If data above the mean are bunched close to it and data below the mean are spread out we say the data are Left Skewed and inequality 1) will hold. Similarly if data below the mean are bunched closely to it and data above the mean are spread out we say the data are Right Skewed, in this case inequality 2) will hold. When data below the mean are reflective of the way data above the mean are spread out we say the data are symmetrically spread, in this case the mean, median and mode will all be equal. Diagrams at the end of these notes illustrate the point.

Weighted Data.

Sometimes data come to us with weights associated with the individual observations reflecting their importance in the data set. Suppose for example the object of interest is income per capita in Canada and our observations are the income per capita in each of the provinces. We know that the population in each province is quite different (as is the income per capita) so that, in attaching equal weight to each observation, the mean in (1) above attaches too much weight to the lowly populated provinces (e.g. PEI) and not enough weight to the highly populated provinces (e.g. Ontario). Letting the population in province “i” be “w_i” and the income per capita in province “i” be “X_i” the provincial incomes per capita can be re-weighted with weights w^*_i as follows:

$$\overline{X}_w = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i}{\overline{w}}X_i = \frac{1}{n}\sum_{i=1}^{n} w_i^* X_i$$

Where:

$$\overline{w} = \sum_{i=1}^{n} w_i \quad and \quad w_i^* = \frac{w_i}{\overline{w}}$$

The weight for province “i” is the ratio of its population to the average provincial population so that the weights w^*_i have the effect of exaggerating the contribution of the per capita income of highly populated provinces and diminishing the impact of the low population provinces in calculating the overall average.

To calculate the Median with weighted data (recalling that the X_i’s are in rank order) the average of the X_j and X_{j+1} for which:

$$\frac{1}{n}\sum_{i=1}^{j} w_i^* < 0.5 \quad and \quad \frac{1}{n}\sum_{i=1}^{j+1} w_i^* > 0.5$$

gives a reasonable approximation to the median.

Dispersion Measures

These measures describe how spread out or widely dispersed the collection of number is. The Range, Inter-Quartile Range, variance, Standard Deviation and Coefficient of variation are introduced.

The Range

This is simple the difference between the highest and lowest values in the data set X_n – X_1. Like the mean its unit of measurement is that of X, it is also susceptible to the varieties of extreme values in the data set.

The Inter Quartile Range

Defining $X_{0.25}$ to be the 25th percentile, a value such that for index value j where j ≤ 0.25n ,j+1 $X_j \leq X_{0.25} \leq X_{j+1}$ (if either j or j+1 = 0.25 then the corresponding value of X is taken as $X_{0.25}$ otherwise $(X_j + X_{j+1})/2$ is used). Similarly defining $X_{0.75}$ to be the 75th percentile, a value such that for index value k where k ≤ 0.75n ≤ k+1 $X_k \leq X_{k+1}$ then the Inter Quartile range is given by $X_{0.75}$ - $X_{0.25.}$ Like the range its unit of measurement is the same as that of X however unlike the Range it is much less susceptible to the distortions of extreme values in the data set.

The Variance

The variance is usually denoted σ_x^2 and is defined as:

The Variance:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \tag{2}$$

If the mean represents a central value of the collection of numbers then $X_i - \overline{X}$ (the deviation from the mean of the i'th value of X) corresponds to the distance from it of the i'th value, this will be negative for values of X_i below x and positive for values above. If the denominator were n, then the variance would correspond to the average value of the squared deviation from the mean, why make it n-1?

Since, from the definition of the mean, the sum of the X's is equal to n times the mean, it is easily shown that the sum of all deviations from the mean is 0.

$$\sum_{i-1}^{n}\left(X_i - \overline{X}\right) = \sum_{i-1}^{n} X_i - \sum_{i-1}^{n} \overline{X} = \sum_{i-1}^{n} X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$$

What this implies is that any particular deviation from the mean is always equal to minus the sum of all other deviations from the mean.

$$\left(X_j - \overline{X}\right) = -\sum_{i=1, i\neq j}^{n}\left(X_i - \overline{X}\right)$$

So that in the collection of n deviations from the mean there are only n-1 independent pieces of information. Now we can see that using n-1 rather than n in the denominator of the variance is simply adjusting for the number of independent pieces of information employed in the calculation.

Finally why use the mean in the calculation, the median or the mode could have been used? Consider the sum of squared deviations from some arbitrary value A and let us call it SSA so that:

The Sum of Squared Deviation:

$$\text{SSA} = \sum_{i=1}^{n}(X_i - A)^2 \qquad (3)$$

Elementary application of the calculus shows us that the value of A that minimizes SSA is the mean as follows:

$$\frac{\partial SSA}{\partial A} = -2\sum_{i=1}^{n}(X_i - A) = -2\left(\left(\sum_{i=1}^{n} X_i\right) - nA\right) = 0 \rightarrow A = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$$

and

$$\frac{\partial^2 SSA}{\partial A^2} = 2n > 0$$

which shows the sum of squared deviations SSA to be a minimum when A equals the mean of the X's.

Note with weighted data the variance would be calculated as:

$$\sigma_w^2 = \frac{1}{n-1}\sum_{i=1}^{n} w_i^*(Xi - \overline{X}_w)^2$$

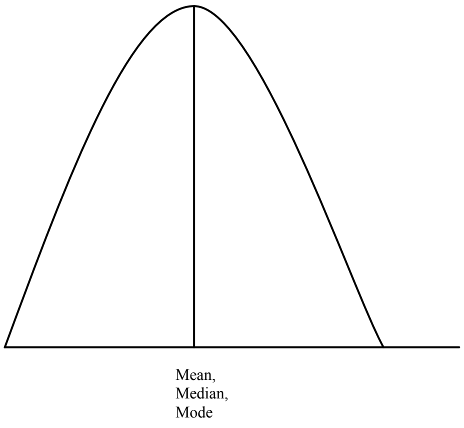
The unit of measurement of σ_x^2 is the squared unit of measurement of X which makes for some inconvenience when the degree of dispersion relative to location is required. For this reason Standard Deviation σ_x (=$^+\sqrt{\sigma_x^2}$) is often employed. Obviously σ_x is measured in the same units as X and, when the mean is not zero, can be compared to the mean. In fact such a comparison gives rise to the Coefficient of Variation which is defined as $\sigma_x/|\,\overline{X}\,|$. The Coefficient of Variation is unit free and allows the comparison of different entities. For example, suppose one collection of X's corresponded to incomes in Russia and another collection of X's corresponded to incomes in Bangladesh, comparison of the respective Coefficient of Variations would permit the assessment of which collection of incomes was more spread out relative to their respective means. Notice such a comparison could not be made by comparing the respective σ_x's directly since they are measured in different units.

Other Summary Statistics

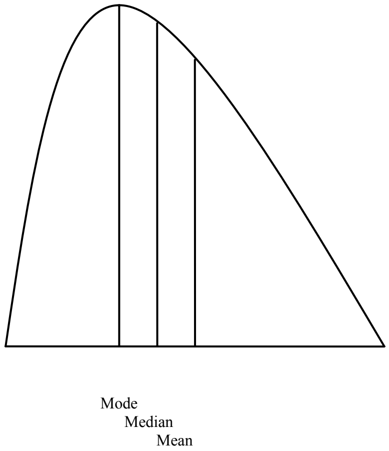
Through they are beyond the scope of this book, the existence of other summary statistics should be mentioned for completeness and to show that many things can be done to describe the nature of a collection of data, the only bound is the extend of our creativity.

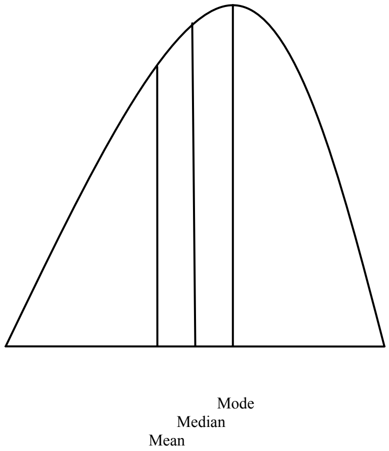
Measures of how “Lop-sided”, asymmetrically spread or **Skewed** a collection of numbers is can be assessed by a variety of Skewness measures. Because skewness affects the relationship between the mean, median and mode (see above) and the inequalities consistently reflect the nature of the skewness the difference between any two of the location measures, divided by a dispersion measure calibrated in the same units (any one of the range, interquartile range, or standard deviation) will give an indication of the nature and degree of skewness. Measures of how “peaked”, “pointed” or **Kurtotic** a collection of numbers is can be obtained by comparing the relative magnitudes of sums of squared deviations from the mean with sums of squared deviations from the mean. These however are matters for study in a more advanced course.

For Symmetric distributions the mean, median and mode will be the same.



For Asymmetric distributions the location measure relationship depends upon the skewness.


Mode
Median
Mean


Mode
Median
Mean

Appendix. Summation Notation

In statistics it is often necessary to sum a collection of numbers or functions of numbers. The basic rules and notation for doing this are outlined here. The operation of summation is denoted by the capital Greek letter sigma (∑). The things to be summed are sometimes indexed with a subscript “i” and the upper and lower limits of the range of summation are written above and below the summation sign so that:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$$

should be read as “the sum from i=1 to n of the values X_i”. When functions of the X_i’s are to be summed, for example the sum of the probabilities of the X_i’s denoted $P(X_i)$, it is written as:

$$\sum_{i=1}^{n} P(X_i) = P(X_1) + P(X_2) + \ldots + P(X_n)$$

It should be stressed that the nature of the function P( ) does not change across the summation notation only the value of the argument changes. Sometimes, for convenience, the objects of summation are not indexed so that if the outcomes in a collection of outcomes A are individually denoted o_i, the sum of the probabilities of all of the outcomes in A would be written as:

$$\sum_{\forall\, o \in A} P\left(o_i \right)$$

The summation operation is what mathematicians call a linear operation, it has the nice property that a sum of a linear function of the X_i’s is the same linear function of the sum of the X_i’s so that:

$$\sum_{i=1}^{n} (a + bX_i) = (a + bX_1) + (a + bX_2) + \ldots + (a + bX_n) = na + b\sum_{i=1}^{n} Xi$$

Breaking this down it can be seen that the sum of n constants is equal to n times the constant and the sum of b times the X_i’s is b times the sum of the X_i’s.

Male-Female and Union-Non-Union wage rates.

The following table reports four random samples of 12 wage rates of unionized and non-unionized female and male workers culled from The Statistics Canada’s Survey of Labour and Income Dynamics for 1994. Calculate the mean, median, range, interquartile range, variance and coefficient of variation for each category of each gender. What do your calculations indicate about the differences in male and female wage rates.

Females		Males	
non-union	union	non-union	union
22.020000	12.250000	23.290000	11.980000
26.980000	20.740000	5.9600000	10.000000
9.8600000	28.570000	18.870000	10.670000
12.450000	15.650000	20.400000	22.000000
15.500000	13.940000	42.670000	17.000000
24.580000	17.990000	21.600000	34.790000
32.030000	15.740000	38.380000	21.600000
9.8500000	10.560000	24.000000	21.800000
10.560000	25.000000	33.600000	23.780000
12.000000	10.000000	23.550000	17.280000
25.950000	17.810000	14.400000	16.710000
10.990000	19.200000	15.200000	21.000000

Homogeneous Matching.

The correlation between spouses educational levels has attracted the interest of Sociologists, Demographers and Economists alike. A complementarity view of spousal roles argues for differences in educational attainment levels. Maximizing earning power suggests a preference for mates with the highest possible educational attainment level and, since higher earning power attracts higher earning power, similarities in educational attainment levels. Random samples of married or cohabiting couples are taken from the 1960 and 1990 USA Censuses of Population. Individual schooling attainment is ranked into five educational categories, 1 - No High School, 2 - No college 3 - No More than 1 year of College, 4 - No More than 4 years of College and 5 - More than 4 years of college and the male educational attainment level is subtracted from his spouses to give an educational difference index.Calculate the mean, median, range, interquartile range, variance and coefficient of variation for each year. What do your results suggest about changes in spousal educational attainment difference over 30 years.

1960

0, 1, 0, 1, 0, -1, -3, 1, -1, -1, -1, 1, 0, 1, 0.

1990

-2, -1, 0, -2, -2, 0, 2, 0, 2, 0, 1, -2, 0, 0, 0, -1, 0, -1, -1, 0, 1, 0, 0.

Weighted Data.

The following table reports the per capita gross domestic product for the provinces of Canada together with their respective populations for the year 1991.

	GDP per capita	Population (1000's)
Newfoundland	15878.396	578.20700
Prince Edward Island	16078.723	130.48300
Nova Scotia	19083.999	912.33500
New Brunswick	18161.636	743.21500
Quebec	22119.575	7033.3630
Ontario	26880.663	10359.231
Manitoba	20800.434	1106.2750
Saskatchewan	20637.584	1002.3460
Alberta	27726.149	2571.7960
British Columbia	24398.375	3338.4600
Yukon and NWT	32662.259	87.869000

Calculate the population weighted and un-weighted average GDP per capita together with the weighted and unweighted variances and coefficients of variation.