Library statistics without Fear

Ellen Maleszewski and Michael Bailou Huang

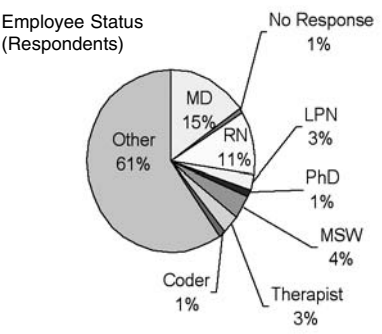
Health Sciences Center Library
Stony Brook University
8034 SUNY, HSC Level 3, Room 136
Stony Brook, NY 11794-8034, USA
emaleszewski@optonline.net
Michael.B.Huang@stonybrook.edu

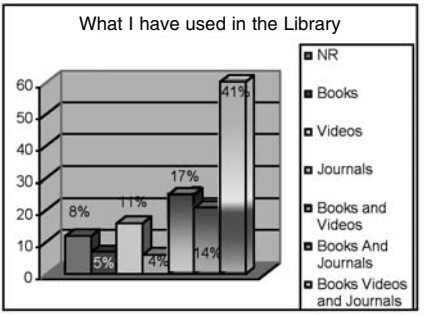
I. Introduction

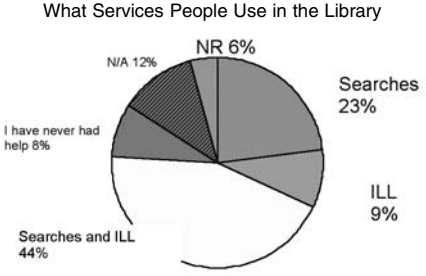
In today's world, the concept of calculating statistics is still frightening, yet the actual calculating part is now automatic. With programs such as Microsoft Excel or SPSS, the "number crunching" side of statistics is no longer arduous. However, the concepts of statistics are still difficult for some to grasp. Most often, the truly grueling statistical functions are only necessary when an experimental research study has occurred. For most daily functions, the basic statistical functions or descriptive statistics are all that are needed.

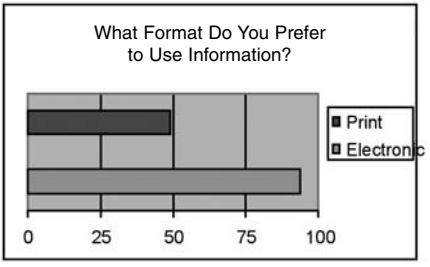
Knowing how what statistics are used is essential for all librarians. They are an important aspect of budget rationales, usage reports, patron trends and other indicators that librarians would use to manage the library. They are required when summarizing results from surveys. They are a vital part of "benchmarking" or comparing from library to library. Research that is reported in library literature or for any professional literature will require an understanding of statistics. Reference librarians should be able to grasp results of professional literature reporting statistics, so that they can guide patrons to the information that they are seeking.

Examples:

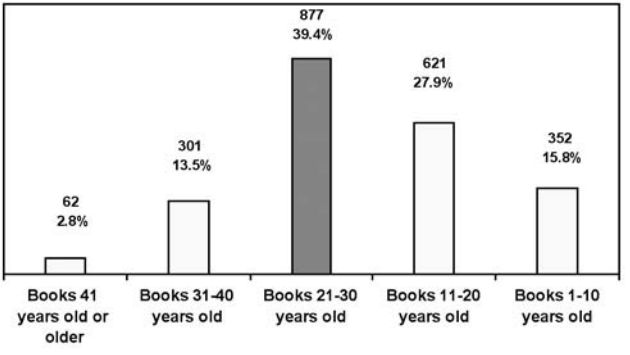








Publication date analysis of the 2,226 Public Health Books in the Library: FY 2002/03



II. Concepts of Statistics

Descriptive Statistics: Why use them?

Statistics are facts and figures. The purpose of statistics is to utilize rules and methods, to organize and interpret observations. Generally, there are four types of statistical tests; Descriptive, Parametric test for means and mean differences, Nonparametric tests, and Measures of relationship between two variables Descriptive statistics is the most well known statistical tests, as they are the foundation of all types of statistics. The purpose of descriptive statistics is to summarize data. They are the best type to use for summarizing surveys, trends in purchasing, trends in usage, and other "non-experimental" data collection that libraries do on a daily basis. Hypotheses are not required! The 4 different types of descriptive statistics tests are: (Gravetter)

1. Frequency Distribution tables and graphs
   Frequency tables and graphs are easy to create and give a snapshot of all your observations. The different types of tables and graphs are; the frequency table, the stem and leaf display, a pie chart, line graph, bar graph and histogram.
2. Measures of central tendency
   Some of the Measures of central tendency are the mean, median, and the mode. The purpose of the central tendency is to find one number that represents the entire set of data by being a measurement of the most typical score. So the "summary" is one number representing many. This works well with large numbers of data because it can be too complex at times to graph enormous data sets.
3. Measures of variability
   The measure of variability is also one number that represents an entire set of data. What it measures is how the data varies from score to score. Standard deviation is an example of a measure of variability. Standard Deviation is the average distance of each score from the mean. Range is another example, it tells you the length of your score set (minimum to maximum score)
4. Z-scores-
   The z-score is a measure of each individual score. Where as the Central Tendency is one score that represents many, the purpose of the Z-score is to measure where an individual score lands in a distribution of scores. For example is a score below the mean or above the mean?

Parametric tests for means and mean differences and nonparametric tests are utilized for experimental data, when comparing populations, samples and a hypothesis has

been stated. The term parametric refers to the concept that one is creating parameters or a number which dictates whether your experiment proves or disproves your hypothesis. It is then plausible to infer the results of the experiment from the sample (that you experimented with) to the population that your hypothesis is about. The parameters utilized are; the mean, the standard deviation and z-scores. Nonparametric tests infer results in the same manner; they are just utilizing data that cannot use the mean and other quantitative measurements because the data is non numerical, it is categorical. (You cannot average gender for example). This is known as inferential statistics.

You can use statistics to measure and describe relations. The data for these statistics involve two observations for each individual – one observation for each of the variables being examined. This is known as regression or correlations. Measures of relationships are well known to most. Correlation and regression are easy to calculate and great for "observation" studies. Just remember they do not mean causation.

The biggest problem for most people is to know when to use what. The usage of statistics depends on two factors:

1. What are your variables? How are you measuring them? What are the scales of measurement? The four scales of measurement are nominal, ordinal, interval and ratio.
2. What are you trying to demonstrate? If you are trying to compare two samples of data, z scores might be your route. If you are trying to show the change over time, then a line graph would be a good choice. When trying to get across the "big picture", choose graphical or tabular statistics. When trying to get a handle on the typical score, or if you are in need to show one score in representation of many, choose central tendency.

III. How to do the statistics

What Are Your Variables?

To choose the statistics you need to know what your variables will be. If you are running an experiment, choosing your variables depends upon your hypothesis. Most often, librarians utilize statistics to demonstrate usage or trends of usage in the past. Statistics are often a rational for budget increases, weeding collections or other day to day tasks. Quite often a report will be run from the OPAC, or perhaps tallies of what is done will be kept by hand on a piece of paper; in other words data will be collected in some fashion. It will be the librarian's job to report that information so that action can be taken.

How the data is summarized and reported will be dependent on what type of data it is. As previously mentioned, there are four types of data; nominal, ordinal, interval and ratio. The data types are what dictate the type of statistics utilized. The nominal scale is not actually a scale as one would ordinarily think of one. Actually, data that is nominal, is not numerical, it cannot be "measured". Nominal scales are just groups of data sorted and categorized. An Example is gender (who is male, who is female). Eye color is another nominal scale. If you work in an academic library, and you keep track of who uses the library, faculty, administrative staff, or students, each of those are categories and this would be known as nominal data. Each category that you utilize to catalog information could be nominal data.

Ordinal scales exist for ranking. In an ordinal scale, items are ranked according to their size or magnitude. The "numbering" that is utilized is not measurable other than with your data. For example most surveys will utilize a Likert scale or the other similar scales, the scale utilized with that system will be an ordinal scale.

Please circle the appropriate response.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Do you feel you have learned from this class?	1	2	3	4	5

The "1, 2, 3,4,5" really means nothing other than that "1" means something lower than "2" the words could very well be used rather than the numbers, they are interchangeable.

Interval scales have no absolute zero, however they do demonstrate magnitude. What does this mean? In the above example, there is no real difference in "distance" between the 1 and the 2. The numbers used could have easily been 1, 3, 5, 7 and 9. Instead, look at the temperature scale. There is a difference between 1? and 2?. In fact the difference is 1?. And that difference is the same as the difference between 2? and 3?. This scale has a zero, you can go below or above the zero. However, we created the zero.

Ratio scales are the scales that are the easiest to measure. They are numerical and measurable. Weight, time, and distance are all ratio scales. If you are keeping track of the number of days an item is allowed to circulate, the number of days would be a variable that would be a ratio scale. The weight of a book is a ratio scale. The number of minutes computers are used for internet searches is a ratio scale.

The importance of your scales is as easy as this: you cannot do numerical calculations to data that is nominal or ordinal. To work with that data, you need to do frequency tables, frequency graphs or the mode. Interval and ratio data can be multiplied, subtracted etc, in other words you can do mathematical calculations with data that is quantitative. You can average the number of searches you do per day!

Presentation of an Entire Distribution

1. Frequency Tables and Graphs

A frequency table is best used when you are trying to present an entire distribution. A frequency distribution is organized tabulation of the number individuals located in each category on the scale of measurement. One column lists the range of scores, from the maximum number to the minimum. The second column lists how many times that score occurs (the frequency).

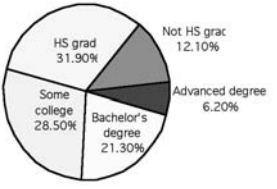
An Example: Library has a training class. A quiz is given and you want to keep track of the scores. The grade is the best out of 10.
Below are the class grades.
Score 8,9,8,7,10,9,6,4,9,8,7,8,10,9,8,6,9,7,8,8.
To the right is a tabular representation of the data. It answers many questions such as: 1. What is the lowest score? 2. What is the highest score? 3. What is the most commonly occurring score? 4. Did most of the test takers pass? 5. How many failed? Etc.

X	F
10	2
9	5
8	7
7	3
6	2
5	0
4	1

Frequency Graphs

Which graph you use is dependent on the type of data that you collect. If you have collected nominal or ordinal data then the graph should be a bar graph or a pie chart. Interval or ratio data should be a histogram.
a.) To create a pie chart, you need to multiply the percentage by 360?
You survey your library patrons to do a "needs assessment". One of the questions in your survey asks them to tell you their level of education. You would like to offer more education for adults in library services. You need to present this information to the board. One way to display it is to use the pie chart.
Your data is as follows:

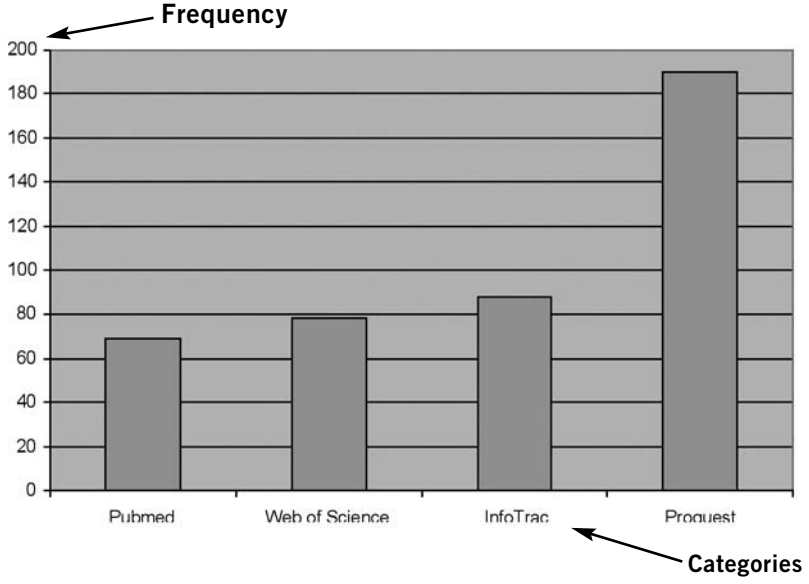
12.1 % have not graduated from High School
31.9 % are High School graduates
28.5 % have some college
21.3% have a bachelor's degree
6.2% have an advanced degree



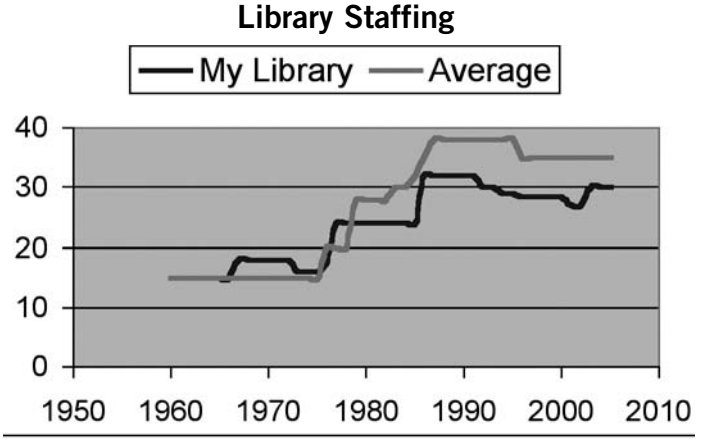
To create a pie chart, you convert the percentage into a decimal and multiply it by 360?. Example: 21.3% have bachelor degrees, so 21.3%=0.213. 0.213 X 360? =77?
(Example Modified from Moore 176-177)

b.) To create a bar graph, you would be working with nominal or ordinal data. The height of each bar represents the number of times that category occurred, or the frequency. It can be easier to draw then a pie chart, and can be better to use when comparing items. Pie charts have to equal 100%, where as bar graphs, if you add up the columns do not have to be part of a whole (or be equal to 100%). The vertical axis should be the frequency (either a count or a percentage is fine) and the horizontal axis should be the categories. In bar graphs, the bars do not touch.

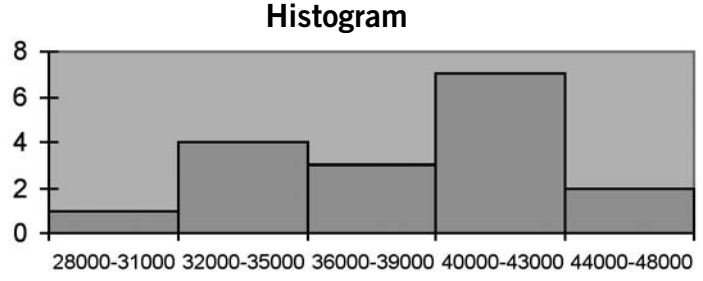
Example:
You are tracking what databases your users utilize in your library – below is your data:
PubMed 69 (Medical Database) InfoTrac 88 (General Academic)
Web of Science 78 (Science Database) Proquest 190 (General Academic)
** Each database is part of the whole for the specific discipline they cover.



c.) A line graph is used mainly to demonstrate change over time. One axis is the measure of change, the other axis is time. Example: You have done a survey of the number of staff in your library. You are comparing it to benchmark data, which demonstrates the average of staff through the years.



d.) Histogram-
Histograms are used when your data is interval or ratio (numerical). Again, they are a frequency graph, and one axis is the frequency (or percentage). Unlike in a bar graph, the histogram bars do touch. Think of using a histogram when you are using central tendency numbers as well. It will give you the big picture plus the central number. For example, you survey your area for household salaries. The following histogram displays your results:



e) Stem and Leaf Display

Stem and leaf displays are another graphical representation. You use this for numerical data. You can use a stemplot when you have small data sets. It's easier to draw then a histogram, and gives you actual values. It gives the same information. To make one

1. Separate each observation into a stem consisting of all but the final most digit and a leaf is the final digit. Stems may have many digits but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top and draw a vertical line at the right of this column write each leaf in the row to the right of its stem in increasing order from the stem.

Example:

Data			Stem and Leaf Display	
83	82	63	3	23
62	93	78	4	26
71	68	33	5	2679
76	52	97	6	238
85	42	46	7	1344668
32	57	59	8	1235
56	73	74	9	37
74	81	56		

2. Central Tendency

Trying to decide between mean, median and mode can be difficult. They are all basically defined as the center number, representing the entire distribution.

You utilize the mean for numerical data (scales of measurement that are ratio or interval) and for distributions that are normal. Your data has to be numerical, considering the mean (or average) is a sum of all the scores divided by N or the number of scores.

- For a population of N=4 scores, 3,7,4,6
  Mean = Ì = ™X / N 3+7+6+4 = 20 20/4 = 5
- A group of six students decide to earn some extra money one weekend picking vegetables at a local farm. At the end of the weekend the students discovered

that their aver income was Ì = $30. If they decide to pool their money for a party how much will they have?

You don't know how much money each student earned. But you do know that the mean is $30. This is the amount that each student would have if the total were divided equally. For each of six students to have $30 you must start with 6 X $30 = $180. The total, ™ = $180

To check this answer, use the formula for the mean:

Ì = ™ X = $180 = $30
N 6

(Examples taken from Graveteer)

When data is categorical (nominal or ordinal) you should use the mode. The mode is the category with the highest frequency. Do not make the mistake that most make, and state that the mode is equal to the frequency. The mode must be the category. Utilizing our example our example from I.b.(the bar graph on the databases that are used the most by our users) the Mode in this example is Proquest-because that is the database with the highest frequency.

- Suppose for example you ask a sample of 100 students on campus to name their favorite restaurants in town.

Restaurant	Frequency
College Grill	5
George & Harry's	16
Luigi's	42
Oasis Diner	18
Roxbury Inn	7
Sutters Mill	12

What is the mode? Luigi's Because Luigi's has the most students who claimed it was their favorite (42)

What is the Median? The median is the middle score. You utilize the median when you have numerical data and the distribution is skewed. Or, you utilize the median when you have ordinal data (a type of categorical data). The median is most often used when the data is skewed. Example: You work in a public library that has a mostly poor population. Yet, in a survey that you have mailed out, you request to know the salary range of your constituents. There might be one area, which is very wealthy, but the majority of the population is not. Your survey results are as follows:

35,000	37,500	100,000	30,675	31,003
35,000	36,400	100,000	37,251	
28,000	40,850	85,000	36,223	
40,000	29,750	33,750	35,313	
80,000	30,605	33,750	33,113	

The average or mean would be $45,199, yet the median would be $35,000. That is a substantial difference. However, if I was to remove the data that is out of the ordinary (the 80-100,000) then my average would be $34,364 and my median would be $35,000. How did I calculate that? To get the mean I added all the salaries and divided them by 21 (there are 21 scores).
To get the median I put all the numbers in order from lowest to highest and counted 10 from each side.

28,000	33,113	35,313	40,000	100,000
29,750	33,750	36,223	40,850	
30,605	33,750	36,400	80,000	
30,675	35,000	37,251	85,000	
31,003	35,000	37,500	100,000	

Those 4 salaries that stand out are known as outliers. They are data that stand away from the norm. If your distribution is skewed like this, you would choose to utilize the median.

- Consider the following: 3, 5, 8, 10, 11. N = 5 scores.
  And the middle score is X=8. So the median is equal to 8.0
- 3,3,4,5,7,8 Now select the middle score (4 and 5) add them and divide them by 2 = 9/4 = 4.5

3. Measures of Variability: Standard Deviation

Standard deviation is the most commonly used and the most important measure of variability. Standard deviation uses the mean of the distribution as a reference point and measures variability by considering the distance between each score and the mean. It determines whether the scores are generally near or far from the mean. That is, are the score clustered together, or scattered? In simple terms the standard deviation approximates the standard distance from the mean.

Although the concept of standard deviation is straightforward, the actual equations will appear complex.

Scores: 10,7,6,10,6,15 Find the variance and the standard deviation
What did we say was the formula for variance?
Compute the Sum of squares for SS we will use the definitional formula
SS = ™(X-Xbar)2
Step 1: Calculate the Sample mean Xbar = Sum of X/n = 54/6 = 9 Xbar = 9
Step2: Compute the deviation scores (X-Xbar) for every ex value

X	**X – Xbar**	**=**
10	10-9	1
7	7-9	2
6	6-9	-3
10	10-9	1
6	6-9	-3
15	15-9	6

Step 3:

X	X – Xbar	=	(X-Xbar)2
10	10-9	1	1
7	7-9	2	4
6	6-9	-3	9
10	10-9	1	1
6	6-9	-3	9
15	15-9	6	36

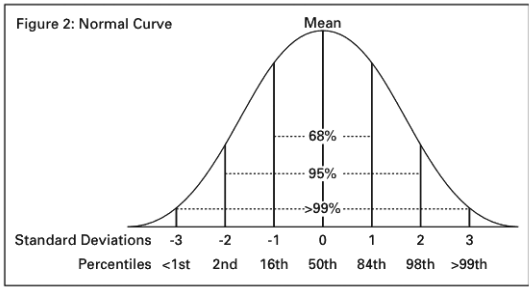
Step 4:
Sum the squared deviation scores to obtain the value for SS
SS = Sum(X-Xbar)sssquared =1+4+9+1+9+36=60
Compute the sample variancee ss/n-1 = 60/5 =12

Step 5:
Compute the sample standard deviation SD=Square root of ss/n-1 =Square root of 12=3.46

How is this useful in the library? Let's go back to our example of the average age of a book in the library. The average age of the books in that example is somewhere between 21 and 30. Let's choose the number 23. If the mean is 23, and there is a SD of 10 years where does that leave us? That would mean that the majority of our data would have to be with 1 standard deviation or with 10 years (+/-) of 23 (13 or 33 years old). If the standard deviation is 5 years, then the data would look totally different. Looking at that chart, that tells us that this is the case. How the data varies around the mean makes a big difference.

Let's look at the example talking about salaries. If we include the outliers our Mean is 45,199 with a standard deviation of 20,791 and if we don't include the outliers our Mean is 34,364 with a standard deviation of 4932. The dropping of the outliers demonstrates that without them, the data is extremely close. The standard deviation shows that the data does not vary much from the mean. Including the outliers causes the SD (Standard deviation) to jump up in size, because the data varies much more.

4. Z Scores
Z scores describe one individual score within a set of data. They consist of a sign and a number. The sign (+/-) tells you whether the score is above or below the mean. The numerical part indicates how many standard deviations away from the mean your score is located. It is an excellent choice for comparing individual scores. Let us look at our staffing example again. For the year 2005, the average number of staff in libraries similar to ours was 35. Our Library had a staff size of 30 people. Is that a tremendously large difference? It is if the standard deviation is 2 people. Z= (x-xbar)/SD = (35-30)/2 =-5/2=-2.5. This means our score is less than the mean, and 2.5 standard deviations away from the mean. If most of your data in a normal curves lies within 1 standard deviation, this makes for a large difference.



Taken from http://www.afb.org/images/celeb_sol-figure2.gif

If our sister school has 34 staff members, that makes their z score -.5 and they are not even one standard deviation away from the mean. Some people look at z scores as if they are a conversion to a standard deviation score. Now let's compare to another department in your institution. Perhaps HR. The average number of employees in benchmark data for HR is 48. Your HR department has 38 with a standard deviation of 8. Who has a lower "deficit"? HR's Z score is (38-48)/8= -10/8= -1.25. The library has a more significant staffing issue!

Using Excel to calculate your statistics.

You have several options on how to utilize Excel for your statistics.

1. To create charts click on the chart button located on the tool bar. There is a four-step wizard that will walk you through the process. Some tips for create charts are:

- Once a chart is created, you can always double click on an element of the chart (a column, point or axis) and change the item after it is created.
- When making a line graph, utilize the XY scatterplot vs. the line.
- Make sure you let the wizard know if your data is organized into columns or rows.

2. To create a Histogram or utilize other packages that excel has pre-packaged for you, you can use the data analysis option. It is located in the tools menu. However, if you do not see it at first you will have to install it (a one time thing). To install it, select Tools, Add-ins and choose Analysis Tool Pak by click on the box to the left of it (a check mark will appear)

3. Utilize the paste function.

The button for the paste function is located on your toolbars-
Once you choose this button you have access to a variety of functions.
To do the mean, look under average. Correlation is correl etc.

Tips:

- The histogram has a "bin" option. That is the intervals that you set up for your data. In the example listed above for histogram, 28-31000 is an interval. If you don't put a bin in, then excel will. I find it easier to leave it blank and let excel decide the bin. Then I go into the bin and change it.
- The histogram in excel, appears without the bars touching it (really!). To fix that simply double click on the series (the vertical bars) and click on the options tab. Look for overlap and gap. Change gap to 0. Then you have a correct histogram.

- Descriptive statistics in the data analysis option is your best choice for most of the statistical work you need to do. It gives you all of your central tendency, your minimum and maximum score and more.
- For Correlation and regression graphs, use the scatter plot chart. To calculate r utilize the paste function option and go to correl.

Statistically Significant.

Statistically significance is based on a value of probability. What is the probability that the answer you give will be wrong? That is the "layman's" definition of it. Statistical significance has to do with hypothesis testing. When you test something you say either the null hypothesis is true (nothing happens) or the null hypothesis is not true (meaning some change has taken effect). Think of it in a doctor-patient setting. Does aspirin cure a headache? The null hypothesis is no, it does not (no change occurs). If you say the null hypothesis is not true (called rejecting the null hypothesis) then aspirin does work (change has occurred due to your intervention). By mistake, you can reject the null hypothesis, even though it is true! Maybe, aspirin has no effect, but everyone taking the aspirin automatically relaxed and the headache went away. Maybe the results were recorded wrong. For whatever ever reason, you reject the null hypothesis (you state that change occurred) when actually the answer should have been null. What are the chances of that? If you say that you have a 90% chance making that mistake, then your experiment is not statistically significant! If you lower that to 10%, it will have a greater degree of significance. 5% and 1% are the most widely used significance levels. They have not the best degree, but some of the highest degrees of significance. Statistical significance in symbolic form is known as alpha or · . So using the symbolic form, · =.01 or .05 (.01 = 1%).

The question most people have is where does the alpha come from? In school, alpha is always given to you. In the real world, the person conducting the study chooses the alpha. You of course would want to select a good alpha, so 90% would be out of the question. I like to choose my alpha based on my sample, if I have a really large sample, .05 is ok, smaller samples should have a .01.

Reading Statistical Literature

Try to keep in mind, that not everybody using statistics really understands it. I always look for the statistical significance, the variables and the size of the sample. I do that for a variety of reasons. If the statistical significance is not outstanding, I become extremely critical. If the sample is too small, then I may look at this like it is preliminary trial of research on a topic, but I won't look at the study as if it is the end all be all. I also looks to see what type of data they collected (are my variables categorical or numerical). Then I read the study. What is the format of the paper

I am reading? Is it in a structured format? If not, it will make deciphering information very difficult. I then look at the study. When considering polls, surveys et, I consider the following points: (1997 Moore.)

1. What was the population? Do they say?
2. What type of experiment, survey, study etc. did they use?
3. What was the intervention (independent variable)?
4. Do you really think that the variables will demonstrate the theory?
5. How were the subjects chosen?
6. How did they decide that the intervention worked?

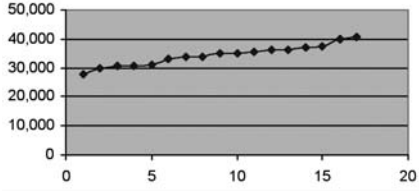
Some information about Polls Online/ Online Polls
Mr. Poll www.Mrpoll.com
Current Population Survey: http://www.bls.gov/cps/
Gallop Poll. www.gallup.com/poll

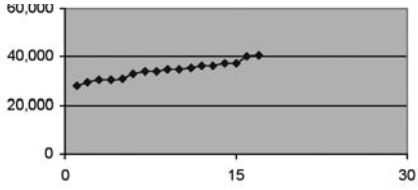
Graphs and Charts:
1. Are they labeled properly?
2. Are they scaled properly?
3. Sate what units you used.
4. What is the source of the data?
5. Are the wedges, bars etc the right size?
6. Look for patterns.

Furthermore, I look for the data. You should be able to replicate their charts and statistical representations based on the data that they provide. Knowing that the data is included, it easy for you to draw your own conclusions before you read theirs. Don't let an author bias your own reading capabilities. Know the difference between categorical data and numerical data. Consider what could have biased the researcher or their results. What errors could have changed the answers? Always question the methods that they used. Be wary of correlational studies as they only indicate that a relationship has occurred. They do not tell you cause and effect. Take note of overall trends, data sources and patterns.

Making statistics work for you.
Remember to label everything and to state your source of data. But if you are utilizing statistics for your advantage (fighting for more money, more staff) then consider the following tactics.

1. Try to utilize graphs, bar graphs, line graphs, and histograms, something people can see easily. It works to your advantage to use fewer words!
2. Don't have too many categories in your bar graphs; if your survey has many just highlight the top 5.
3. Try to utilize color, be sure not to make pastel, and if you need something to stand out make that item the brightest color.
4. Change intervals in axis to make your charts stand out more. See example below.

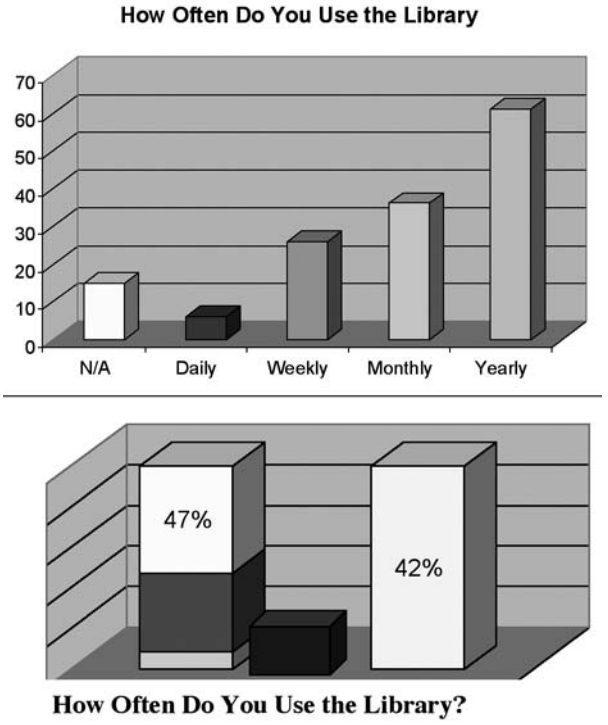




5. Be careful when creating questions in surveys. Example: How often do you go to the library? Most people will not answer daily! But if your choices are :
   a. daily b. weekly c. monthly d. yearly
   most people will choose monthly- it would be better to have
   a. daily b. once a week c. every other week d. monthly d. every other month e. yearly f. not at all
   You need to be careful how you word things. However if you have done something such as daily weekly monthly yearly not at all, and you want to show that the library is utilized, try combining the daily, weekly and monthly numbers. See below:

   If Yes How Often Have you used the Library

Daily	Weekly	Monthly	Yearly	N/A	
6	26	36	61	15	144
4%	18%	25%	42%	10%	100%



6. Know your audience.
7. Benchmark! Comparisons to other libraries not only help you learn what you need and don't need, but they also help your audience.

Conclusion

In conclusion, most of us for our day to day use of statistics can utilize the fundamentals or basic statistic option. Once you understand the theory behind the basic statistic, you can simply utilize a software package to do the math! The most important things to keep in mind is what you are trying to do with the data and what type of data do you have. Try to use your descriptive statistics to make the work you do in your library stand out. Working with more complicated statistics such as inferential can be difficult and you should try to avoid that unless you have a strong understanding of statistics. However, many librarians are daunted by statistics because they feel they cannot do math. The more statistics is utilized as an application, the more you understand them. To understand more about statistics,

a class is always an excellent idea. Statistics are always offered in library schools, however any school that offers behavioral science courses will have a statistics course as well and this meshes very well with library science.

Glossary: *Many of these definitions are taken directly from the texts (electronic and paper format) that are mentioned in the bibliography. I simply searched for the clearest definition.*

Central Tendency:
A statistical measure that identifies a single score as a representative for an entire distribution. The goal is to find the single score that is most typical or most representative of the entire group.

Control Group:
Is a condition of the independent variable that does not receive the experimental treatment. The control group is the baseline for comparison with the experimental group.

Dependent Variable:
Is one that is observed for changes in order to assess the effect of the treatment.

Hypothesis Testing:
Hypothesis Testing is an inferential procedure that uses sample data to evaluate the credibility of a hypothesis about a population.

Independent Variable:
The variable that is manipulated or controlled by the researcher.

Inferential statistics consist of techniques that allow us to study sample and then make generalizations about the populations from which they were selected.

Level of Confidence: The level of confidence says what percent of all possible samples satisfy the margin of error.

Margin of Error: The margin of error says how close the sample statistic lies to the population parameter.

The mean is the most commonly used measure. It is the central number of a normal distribution. It is equaled to the sum of all the scores divided by the number of scores.

The Median: exactly 50% have values less than equal to 50th percentile.

The mode is the score with the greatest frequency. The mode is used when the scores consist of measurements on a nominal scale.

N is the number of scores in a data set (number of observations …etc)

The range is the distance from the lowest to the highest score in a data set. The range is the difference between the upper real limit for the largest X Value and the lower real limit of the smallest.

Raw score:
Is an original measurement or observed value.

p-value
The p-value represents the probability of error that is involved in accepting our observed result as valid, that is, as "representative of the population." For example, a p-value of .05 (i.e.,1/20) indicates that there is a 5% probability that the relation between the variables found in our sample is a "fluke.

Parameter:
Is a value, usually numeric, that describes a population.

Population:
The entire group of individuals that a researcher wishes to study.

Probability
In a situation where several different outcomes are possible we define probability for any particular outcome as a fraction or a proportion.

$$\text{The probability of Event A} = \frac{\text{\# of outcomes classified}}{\text{Total number of outcomes}}$$

$$\text{Or , for a clinical setting} = \frac{\text{\# of times the outcome occurs}}{\text{\# trials}}$$

Sample:
Is a set of individuals selected from the population, usually intended to represent the population in a study.

Sampling error is the discrepancy or amount of error, between a sample statistic, and its corresponding population parameter.

Sigma is a Greek letter and it stands for sum. (Summation)

The Standard Deviation is a measurement of the standard distance from the mean.

Statistics: are facts and figures.

Statistically Significant – Findings are said to be statistically significant when the null hypothesis has been rejected. In other words the findings were not due to chance, the intervention/treatment is the cause of the results.

Type 1 Errors:
It is possible to reject the null hypothesis when in reality the treatment has no effect. The outcome of the experiment could be different from what Ho predicted just by chance. This is a type 1 error. A Type 1 Error consists of rejecting the null hypothesis when Ho is actually true.

Type 2 Errors:
A Type 2 Error the investigator fails to reject a null hypothesis that is really false.

Variability: describes how spread out the values of the sample statistic are when we take many samples. Large variability means that the result of sampling is not repeatable

Variance: Is the mean squared deviation

Z scores: The sign of a z score indicates whether it is above or below the mean. The numerical value of the z score. The numerical value of the z-score indicates how many standard deviations are between the score and the mean.

Bibliography

Bailey, K.D. Methods of Social Research, 4th edition. Macmillan, NY, 1994
Current Population Survey: http://www.bls.gov/cps/
Electronic Statistics Textbook. http://www.statsoftinc.com
General Social Survey. www.norc.org
Gallop Poll. www.gallup.com/poll
Gravetter, F.J. Satistics for the Behavioral Sciences, 3rd Edition. West Publishing Co., NY, 1992
Nielsen Media Research. www.nielsenmedia.com
Norman, G.R. PDQ Statistics, 2nd Edition, Mosby, St. Louis, 1997
Shaughnessy, John J., Research methods in psychology, 2nd Edition, McGraw Hill, NY 1990
Statistical Abstract of the United States. http://www.census.gov/statab/www/
Statistics: Concepts and Controversies (5th edition) 1997 David. S. Moore WH Freeman & Co.
Statistics: The Standard Deviants. Video 1,2 and 3. Cerebellum Corporation. www.cerebellum.com
US Census Bureau: http://www.census.gov/

Abstract

Many librarians who don't have a mathematical or statistical background are often intimidated by a wide spectrum of statistics. A basic understanding of statistics would help librarians answer reference inquiries; write successful proposals for funding of libraries, and to conduct research. This paper will address the following topics: (1) what are basic descriptive statistical methods and how they can be utilized; (2) how to understand "statistically significant" and how to make statistics work using graphs and charts; and (3) how to understand, interpret, and utilize statistics and a variety of the professional research literature to help manage your library.