INTRODUCTION TO DATA AND DATA ANALYSIS
May 2016

This document is part of several training modules created to assist in the interpretation and use of the Maryland Behavioral Health Administration Outcomes Measurement System (OMS) data. This module provides a brief overview of data and data analysis terminology. It is recommended reading for those who have little background in this area but who are interested in the OMS data.

Why use data?

People turn to data because they have a story to tell or a problem to solve. Most people start with a question, then look to data for answers. In a service setting, questions might include, "who is receiving services?" and "who does best in treatment?"

What if you do not have a question to begin with? Exploring data without a defined question, sometimes referred to as **"data mining"**, can sometimes reveal interesting patterns in the data that are worth exploring. Regardless of what leads you to look at data, thinking about your audience (your staff, supervisor, Board members, etc.) is helpful to shape the story and guide your thinking about the data.

Whenever you look at data, it is important to be open to unexpected patterns, explanations, and unusual results. Sometimes the most interesting stories to be told with data are not the ones you set out to tell.

What is data? What types of data exist?

Data is used to describe things by assigning a value to them. The values are then organized, processed, and presented within a given context so that it becomes useful. Data can be in different forms: qualitative and quantitative:

Qualitative data

"Qualitative data" is data that uses words and descriptions. Qualitative data can be observed but is subjective and therefore difficult to use for the purposes of making comparisons. Descriptions of texture, taste, or an experience are all examples of qualitative data. Qualitative data collection methods include focus groups, interviews, or open-ended items on a survey. The OMS questionnaires do not collect qualitative data, but it is helpful to be aware of the differentiation.

Quantitative data

"Quantitative data" is data that is expressed with numbers. Quantitative data is data which can be put into categories, measured, or ranked. Length, weight, age, cost, rating scales, are all examples of quantitative data. Quantitative data can be represented visually in graphs and tables and be statistically analyzed. The OMS questionnaires collect quantitative data.

The qualitative data that describes this cup of coffee are that it has a strong taste and robust aroma. The quantitative data that describes the cup of coffee is that it is 12 ounces, 150 degrees Fahrenheit, and costs $1.50.

There are two types of quantitative data: categorical and continuous.

Categorical data

"Categorical data" is data that has been placed into groups. An item cannot belong to more than one group at a time. Examples of categorical data within OMS would be the individual's current living situation, smoking status, or whether he/she is employed. As discussed in more detail later, the type of analysis used with categorical data is the Chi-square test.

Continuous data

"*Continuous data*" is numerical data measured on a continuous range or scale. In continuous data, all values are possible with no gaps in between. Examples of continuous data are a person's height or weight, and temperature. There are a few examples of continuous data in the OMS Datamart, such as scores calculated for the BASIS-24® and the Youth Short Symptom Index (the symptom scales used in the Adult and Child and Adolescent Questionnaires, respectively). As discussed in more detail later, many types of analysis can be used with continuous data, including effect size calculations.

SUMMARY

- **Qualitative data involves words and descriptions.**
- **Quantitative data is data expressed with numbers.**
  - **Categorical data is a type of quantitative data that involves grouping things.**
  - **Continuous data is a type of quantitative data where values fall along a continuous scale.**

How is data analyzed?

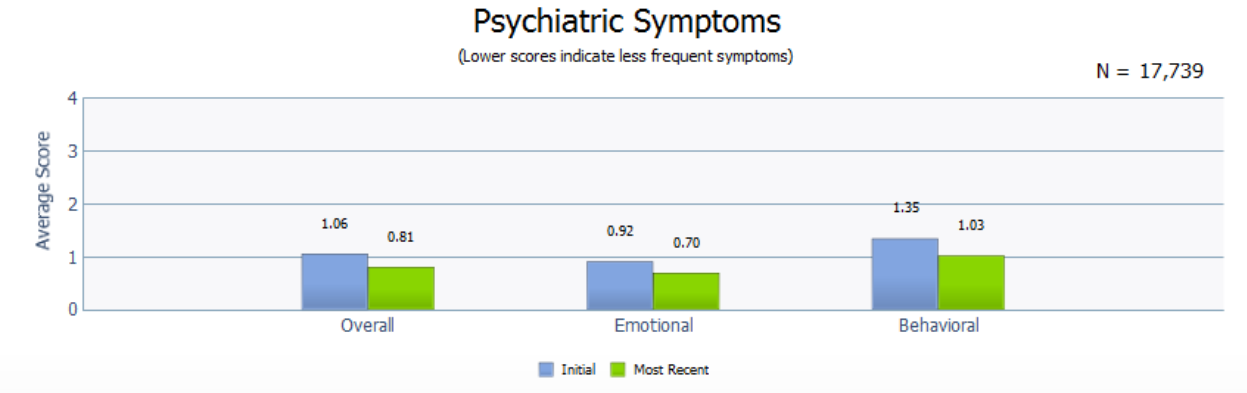
Data is analyzed using statistics. Essentially statistics can be classified in two ways: some help describe data and some help compare data.

Using Statistics to Describe Data

A **"frequency table"** presents aggregate data and tells us the number of times each particular data value occurred. For example, the OMS Datamart provides frequency tables including both the number of adults providing each response as well as the percentages. In this example, the categorical data values are "No" and "Yes":

Answer Options	# of Clients	%
No	100	20.0%
Yes	400	80.0%
Total	**500**	**100.0%**

A **"measure of central tendency"** is a single value that attempts to describe a set of data by identifying the central position within that set of data. The **"mean"** (also known as the **"average")** is the most well-known measure of central tendency that describes data. To obtain the mean, you add up all the numbers and then divide by the number of observations (interviews, people, etc.). An important property of the mean is that it includes every value in the data set as part of the calculation, although it is not necessarily one of the actual values that is in the data set. For example, in the OMS Datamart, the mean is used to describe data for the psychiatric symptom scales (both for Children and Adolescents and Adults). The means of the initial symptom scores are reported, along with the means for the most recent symptom scores. See below for an example of this using data from OMS Child and Adolescent Questionnaires:



Another commonly used statistic to describe data is the **"standard deviation"**. Although standard deviations are not presented directly on the OMS Datamart, they are used to calculate some of the information presented. Standard deviations describe the variability of the data or how close or far away the data is to the average of the group. For example, assume the shirts hanging below are arranged by their different sizes and spaced evenly apart (one shirt per size, average size in the middle). They would have a larger standard deviation, because their size values are spread out. But, if they were replaced with shirts that were almost all the same size, and then clustered together according to size, they would have a smaller standard deviation because there is less variability in their size value.

SUMMARY

- **Some statistics are used to describe data.**
- **Frequencies tell you how many times the value/answer has occurred.**
- **The "mean" or "average" is a measure of central tendency.**
- **The "standard deviation" represents variability in the data.**

Using Statistics to Compare Data

Some statistics allow us to compare groups to one another in order to determine if the differences are "**statistically significant."** Statistical significance generally refers to the probability that the results are not due to chance. It is important to remember that a statistically significant difference only means there is mathematical evidence that there is a difference. It does not mean the difference is necessarily large, important, or meaningful in terms of the utility of the finding.

For example, when comparing two groups on employment, the test may indicate that two groups are significantly different from one another statistically. However, if one group has 40.1% employed and the other has 40.5% employed, is this difference really meaningful? The interpretation of this difference will depend on other contextual information, such as current employment trends, history of the employment data for each group, special initiatives , etc.

There are many factors that influence statistical significance, which is why it is important to be careful with how much importance is placed on a finding, even when it is statistically significant. These factors include the number of data points involved in the test (for example, number of people answering the employment question), the number of statistical tests being conducted, and more. When conducting tests of statistical significance, the tester determines a **"probability level"** (often called **"alpha"**) to be used in determining significance. The alpha indicates the level of confidence that the person can have that any statistically significant results found are not simply due to chance. Alpha levels of .05 and .01 are used in the OMS Statistical Significance Workbooks (these are described in another module). Using an alpha of .05 means that the tester can be 95% confident that the results are not just a coincidence (.01 would equal 99% confidence). The smaller the alpha, the more confidence the tester can have. If an alpha of .10 were to be used, then the tester can only be 90% confident that the results were not simply due to chance.

SUMMARY

- **Some statistics are used to compare data. "Statistical significance" is the term used to indicate that differences between groups are not due to chance.**
- **Many factors can affect statistical significance, including sample size.**
- **"Alphas" indicate the level of confidence that statistically significant results are not simply due to chance.**

One of the statistics used to compare groups (and therefore determine statistical significance) is the "**Chi-square."** The Chi-square is often the "go-to" statistic for categorical data, such as yes/no items or level of agreement (very much, quite a bit, etc.). It is used to compare the

pattern of answer options between two groups. For example, in the OMS Statistical Significance Workbooks, a Chi-square is used to compare the responses to the homeless question. The example below is drawn from the Adult PIT - OMS Statistical Significance Workbook. It includes sample data for the item about whether or not the person was homeless in the past six months.

Worksheet for Calculating Chi Square for Outcome Measurement Results						
COMPLETE ONLY SHADED CELLS - ALL OTHERS WILL AUTOMATICALLY POPULATE						
	Orange Area: Fill in time frame and filter(s)					
	Green Area: Fill in names of the two groups (e.g., agencies, jurisdictions, etc.)					
	Yellow Area: Fill in data					
Question:	**Q3. Have you been homeless at all in the past six months? (COT)**					
Adult/Child:	Adult					
PIT/COT:	Initial Compared to Most Recent Interview (COT)					
Time Frame:	FY2014					
Filter(s):	All ages, genders, races, time in treatment					
	All yellow cells must have data for accurate Chi-square result					
	Group 1	**Group 2**			Percentage Distribution	
Outcome	Statewide	My Agency	Total	Outcome	Statewide	My Agency
Gained housing	2250	20	2270	Gained housing	9.3%	9.0%
Not homeless either interview	19951	172	20123	Not homeless either interview	82.2%	77.8%
Homeless both interviews	1137	12	1149	Homeless both interviews	4.7%	5.4%
Lost housing	944	17	961	Lost housing	3.9%	7.7%
Total	24282	221	24503	Total	100.0%	100.0%
Interpretation:	**Distributions differ at the .05 level**					

The phrase "**Distributions differ at the .05 level**" means that the two groups are statistically different from one another and that there is less than a 5% chance that this result was due to chance.

Like any statistical test, the Chi-square has certain criteria that must be met in order for the test to be conducted. A Chi-square test may not work if the samples sizes are small or if there is a small number of individuals providing a certain response (for example "Yes" or "No") to the OMS question.

It is also important to note that a Chi-square does not tell you which group was bigger or smaller than any other. Therefore, the Chi-Square indicates that the pattern of six-month homelessness responses is different between groups, but it does not indicate which group (statewide or agency) showed greater homelessness. In order to figure that out, the user will have to visually compare the numbers in the two tables.

SUMMARY

- **Chi-square is a statistical test used to compare the pattern of responses between two groups with categorical data.**
- **Chi-square tests cannot indicate which group did better, only that the pattern of responses between the two is different.**
- **Chi square tests can be affected by the number of individuals in each group and the number of groups.**

"Effect sizes" are used to compare groups of data to determine how much change has occurred within a group. For example, an effect size might be calculated to evaluate the impact an intervention had. Effect sizes are used with continuous data, such as psychiatric symptom scale scores in the OMS Datamart. They are like the mean, but are comparative in nature. They are calculated by taking the difference between the means then dividing by the standard deviation.

The benefit of effect sizes are that they standardize the differences in data so that they can be compared with the differences in other data. Another advantage of effect sizes is that they go beyond simple questions of "is there a difference" and give answers such as "the groups differ by X amount." While effect sizes are reported as a single number, it can help to put that number in context such as a "large" effect size versus a "small" effect size. Depending on the effect size calculation, different conventions are used to categorize effect sizes. Effect sizes in the social sciences tend to be in the small to medium range (.20 to .79).

SUMMARY

- **Effect sizes can be used to assess the amount of change or the impact of an intervention.**
- **Effect sizes in social sciences tend to be in the small to moderate range.**

Other Commonly Used Statistical Terms and Topics:

Outliers

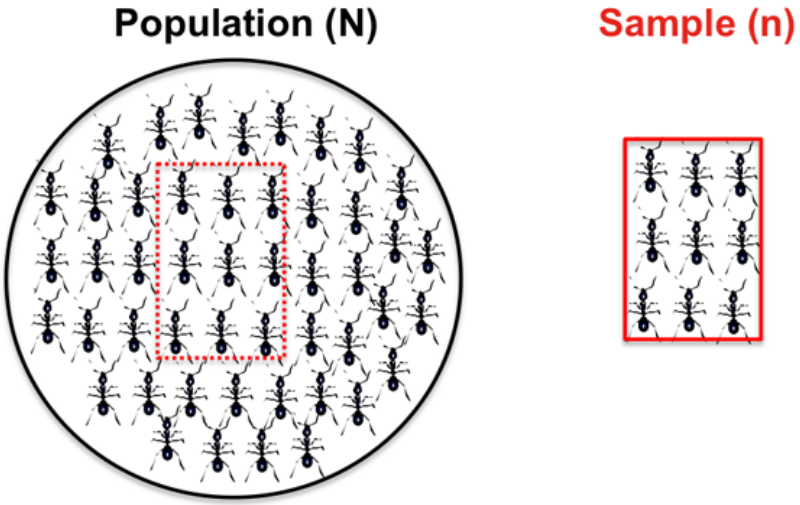
"Outliers" are values that are far from the mean; an outlier is a value that is very large or very small compared to the rest of the data set. Such values can have a large impact on the mean if the sample is small, but are less of an issue when working with a larger sample.

Small sample sizes

Small sample sizes can affect data analysis and interpretation as well. For example, if you have a group of ten individuals who answered an item, each individual represents 10% of the data and therefore heavily affects the percentage of responses. However, if you had a group of 100 individuals, then each person is only 1% of the data so any change in their reporting has much less of an impact.

Populations and Samples

Data is often collected to make statements or tell a story about a group or "**population"** of interest. However, often a population is so large that we cannot possibly measure the variables of interest for every person in the population. The alternative is to select a **"sample"** from the population of interest. The tricky part of taking a sample is ensuring that the sample is representative of the larger population about which you would like to make statements. The illustration below may be helpful in understanding this:



SUMMARY

- **"Outliers" are values that are far from the mean; they may be very large or very small; in smaller samples they can have a disproportionate impact on the mean.**
- **Small sample sizes can affect data analysis and interpretation, particularly when using percentages.**
- **A "population" is the larger group you want to learn about; a "sample" is a subset of that population that is examined in order to generalize to the larger group.**