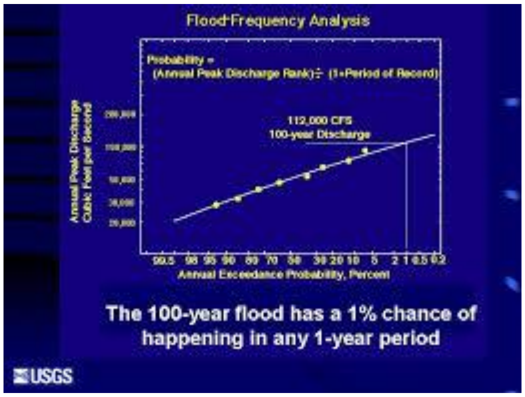
SESSION 6 ANALYSIS OF DATA AND GRAPHIC OF FREQUENCY DISTRIBUTIONS



The 100-year flood has a 1% chance of happening in any 1-year period

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. According to Shamoo and Resnik (2003) various analytic procedures “provide a way of drawing inductive inferences from data and distinguishing the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data”..

While data analysis in qualitative research can include statistical procedures, many times analysis becomes an ongoing iterative process where data is continuously collected and analyzed almost simultaneously. Indeed, researchers generally analyze for patterns in observations through the entire data collection phase (Savenye, Robinson, 2004). The form of the analysis is determined by the specific qualitative approach taken (field study, ethnography content analysis, oral history, biography, unobtrusive research) and the form of the data (field notes, documents, audiotape, videotape).

An essential component of ensuring data integrity is the accurate and appropriate analysis of research findings. Improper statistical analyses distort scientific findings, mislead casual readers (Shepard, 2002), and may negatively influence the public perception of research. Integrity issues are just as relevant to analysis of non-statistical data as well.

Considerations/issues in data analysis

There are a number of issues that researchers should be cognizant of with respect to data analysis. These include:

•Having the necessary skills to analyze

• Concurrently selecting data collection methods and appropriate analysis

• Drawing unbiased inference

•Inappropriate subgroup analysis

•Following acceptable norms for disciplines

• Determining statistical significance

• Lack of clearly defined and objective outcome measurements

• Providing honest and accurate analysis

•Manner of presenting data

•Environmental/contextual issues

• Data recording method

•Partitioning 'text' when analyzing qualitative data

• Training of staff conducting analyses

• Reliability and Validity

• Extent of analysis

Having necessary skills to analyze

A tacit assumption of investigators is that they have received training sufficient to demonstrate a high standard of research practice. Unintentional 'scientific misconduct' is likely the result of poor instruction and follow-up. A number of studies suggest this may be the case more often than believed (Nowak, 1994; Silverman, Manson, 2003). For example, Sica found that adequate training of physicians in medical schools in the proper design, implementation and evaluation of clinical trials is "abysmally small" (Sica, cited in Nowak, 1994). Indeed, a single course in biostatistics is the most that is usually offered (Christopher Williams, cited in Nowak, 1994).

A common practice of investigators is to defer the selection of analytic procedure to a research team 'statistician'. Ideally, investigators should have substantially more than a basic understanding of the rationale for selecting one method of analysis over another. This can allow investigators to better supervise staff who conduct the data analyses process and make informed decisions

Concurrently selecting data collection methods and appropriate analysis

While methods of analysis may differ by scientific discipline, the optimal stage for determining appropriate analytic procedures occurs early in the research process and should not be an afterthought. According to Smeeton and Goda (2003), "Statistical advice should be obtained at the stage of initial planning of an investigation so that, for example, the method of sampling and design of questionnaire are appropriate".

Drawing unbiased inference

The chief aim of analysis is to distinguish between an event occurring as either reflecting a true effect versus a false one. Any bias occurring in the collection of the data, or selection of method of analysis, will increase the likelihood of drawing a biased inference. Bias can occur when recruitment of study participants falls below minimum number required to demonstrate statistical power or failure to maintain a sufficient follow-up period needed to demonstrate an effect (Altman, 2001).

Inappropriate subgroup analysis

When failing to demonstrate statistically different levels between treatment groups, investigators may resort to breaking down the analysis to smaller and smaller subgroups in order to find a difference. Although this practice may not inherently be unethical, these analyses should be proposed before beginning the study even if the intent is exploratory in nature. If it the study is exploratory in nature, the investigator should make this explicit so that readers understand that the research is more of a hunting expedition rather than being primarily theory driven. Although a researcher may not have a theory-based hypothesis for testing relationships between previously untested variables, a theory will have to be developed to explain an unanticipated finding. Indeed, in exploratory science, there are no a priori hypotheses therefore there are no hypothetical tests. Although theories can often drive the processes used in the investigation of qualitative studies, many times patterns of behavior or occurrences derived from analyzed data can result in developing new theoretical frameworks rather than determined a priori (Savenye, Robinson, 2004).

It is conceivable that multiple statistical tests could yield a significant finding by chance alone rather than reflecting a true effect. Integrity is compromised if the investigator only reports tests with significant findings, and neglects to mention a large number of tests failing to reach significance. While access to computer-based statistical packages can facilitate application of increasingly complex analytic procedures, inappropriate uses of these packages can result in abuses as well.

Following acceptable norms for disciplines

Every field of study has developed its accepted practices for data analysis. Resnik (2000) states that it is prudent for investigators to follow these accepted norms. Resnik further states that the norms are ‘…based on two factors:

(1) the nature of the variables used (i.e., quantitative, comparative, or qualitative),

(2) assumptions about the population from which the data are drawn (i.e., random distribution, independence, sample size, etc.). If one uses unconventional norms, it is crucial to clearly state this is being done, and to show how this new and possibly unaccepted method of analysis is being used, as well as how it differs from other more traditional methods. For example, Schroder, Carey, and Vanable (2003) juxtapose their identification of new and powerful data analytic solutions developed to count data in the area of HIV contraction risk with a discussion of the limitations of commonly applied methods.

If one uses unconventional norms, it is crucial to clearly state this is being done, and to show how this new and possibly unaccepted method of analysis is being used, as well as how it differs from other more traditional methods. For example, Schroder, Carey, and Vanable (2003) juxtapose their identification of new and powerful data analytic solutions developed to count data in the area of HIV contraction risk with a discussion of the limitations of commonly applied methods.

Determining significance

While the conventional practice is to establish a standard of acceptability for statistical significance, with certain disciplines, it may also be appropriate to discuss whether attaining statistical significance has a true practical meaning, i.e., ‘clinical significance’. Jeans (1992) defines ‘clinical significance’ as “the potential for research findings to make a real and important difference to clients or clinical practice, to health status or to any other problem identified as a relevant priority for the discipline”.

Kendall and Grove (1988) define clinical significance in terms of what happens when “… troubled and disordered clients are now, after treatment, not distinguishable from a meaningful and representative non-disturbed reference group”. Thompson and Noferi (2002) suggest that readers of counseling literature should expect authors to report either practical or clinical significance indices, or both, within their research reports. Shepard (2003) questions why some authors fail to point out that the magnitude of observed changes may too small to have any clinical or practical significance, “sometimes, a supposed change may be described in some detail, but the investigator fails to disclose that the trend is not statistically significant ”.

Lack of clearly defined and objective outcome measurements

No amount of statistical analysis, regardless of the level of the sophistication, will correct poorly defined objective outcome measurements. Whether done unintentionally or by design, this practice increases the likelihood of clouding the interpretation of findings, thus potentially misleading readers.

Provide honest and accurate analysis

The basis for this issue is the urgency of reducing the likelihood of statistical error. Common challenges include the exclusion of outliers, filling in missing data, altering or otherwise changing data, data mining, and developing graphical representations of the data (Shamoo, Resnik, 2003).

Manner of presenting data

At times investigators may enhance the impression of a significant finding by determining how to present derived data (as opposed to data in its raw form), which portion of the data is shown, why, how and to whom (Shamoo, Resnik, 2003). Nowak (1994) notes that even experts do not agree in distinguishing between analyzing and massaging data. Shamoo (1989) recommends that investigators maintain a sufficient and accurate paper trail of how data was manipulated for future review.

Environmental/contextual issues

The integrity of data analysis can be compromised by the environment or context in which data was collected i.e., face-to face interviews vs. focused group. The interaction occurring within a dyadic relationship (interviewer-interviewee) differs from the group dynamic occurring within a focus group because of the number of participants, and how they react to each other’s responses. Since the data collection process could be influenced by the environment/context, researchers should take this into account when conducting data analysis.

Data recording method

Analyses could also be influenced by the method in which data was recorded. For example, research events could be documented by:

a. recording audio and/or video and transcribing later

b. either a researcher or self-administered survey

c. either closed ended survey or open ended survey

d. preparing ethnographic field notes from a participant/observer

e. requesting that participants themselves take notes, compile and submit them to researchers.

While each methodology employed has rationale and advantages, issues of objectivity and subjectivity may be raised when data is analyzed.

Partitioning the text

During content analysis, staff researchers or ‘raters’ may use inconsistent strategies in analyzing text material. Some ‘raters’ may analyze comments as a whole while others may prefer to dissect text material by separating words, phrases, clauses, sentences or groups of sentences. Every effort should be made to reduce or eliminate inconsistencies between “raters” so that data integrity is not compromised.

Training of Staff conducting analyses

A major challenge to data integrity could occur with the unmonitored supervision of inductive techniques. Content analysis requires raters to assign topics to text material (comments). The threat to integrity may arise when raters have received inconsistent training, or may have received previous training experience(s). Previous experience may affect how raters perceive the material or even perceive the nature of the analyses to be conducted. Thus one rater could assign topics or codes to material that is significantly different from another rater. Strategies to address this would include clearly stating a list of analyses procedures in the protocol manual, consistent training, and routine monitoring of raters.

Reliability and Validity

Researchers performing analysis on either quantitative or qualitative analyses should be aware of challenges to reliability and validity. For example, in the area of content analysis, Gottschalk (1995) identifies three factors that can affect the reliability of analyzed data:

- stability , or the tendency for coders to consistently re-code the same data in the same way over a period of time
- reproducibility , or the tendency for a group of coders to classify categories membership in the same way
- accuracy , or the extent to which the classification of a text corresponds to a standard or norm statistically

The potential for compromising data integrity arises when researchers cannot consistently demonstrate stability, reproducibility, or accuracy of data analysis

According Gottschalk, (1995), the validity of a content analysis study refers to the correspondence of the categories (the classification that raters' assigned to text content) to the conclusions, and the generalizability of results to a theory (did the categories support the study's conclusion, and was the finding adequately robust to support or be applied to a selected theoretical rationale?).

Extent of analysis

Upon coding text material for content analysis, raters must classify each code into an appropriate category of a cross-reference matrix. Relying on computer software to determine a frequency or word count can lead to inaccuracies. "One may obtain an accurate count of that word's occurrence and frequency, but not have an accurate accounting of the meaning inherent in each particular usage" (Gottschalk, 1995). Further analyses might be appropriate to discover the dimensionality of the data set or identity new meaningful underlying variables.

Whether statistical or non-statistical methods of analyses are used, researchers should be aware of the potential for compromising data integrity. While statistical analysis is typically performed on quantitative data, there are numerous analytic procedures specifically designed for qualitative material including content, thematic, and ethnographic analysis. Regardless of whether one studies quantitative or qualitative phenomena, researchers use a variety of tools to analyze data in order to test hypotheses, discern patterns of behavior, and ultimately answer research questions. Failure to understand or acknowledge data analysis issues presented can compromise data integrity.

Student's T Critical Values

Conf. Level	**50%**	**80%**	**90%**	**95%**	**98%**	**99%**
One Tail	**0.250**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
Two Tail	**0.500**	**0.200**	**0.100**	**0.050**	**0.020**	**0.010**
df = 1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499

Conf. Level	50%	80%	90%	95%	98%	99%
One Tail	**0.250**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
Two Tail	**0.500**	**0.200**	**0.100**	**0.050**	**0.020**	**0.010**
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
50	0.679	1.299	1.676	2.009	2.403	2.678
60	0.679	1.296	1.671	2.000	2.390	2.660
70	0.678	1.294	1.667	1.994	2.381	2.648
80	0.678	1.292	1.664	1.990	2.374	2.639
90	0.677	1.291	1.662	1.987	2.368	2.632
100	0.677	1.290	1.660	1.984	2.364	2.626
z	0.674	1.282	1.645	1.960	2.326	2.576

The values in the table are the areas critical values for the given areas in the right tail or in both tails.

Statistics: Frequency Distributions & Graphs

Definitions

Raw Data

Data collected in original form.

Frequency

The number of times a certain value or class of values occurs.

Frequency Distribution

The organization of raw data in table form with classes and frequencies.

Categorical Frequency Distribution

A frequency distribution in which the data is only nominal or ordinal.

Ungrouped Frequency Distribution

A frequency distribution of numerical data. The raw data is not grouped.

Grouped Frequency Distribution

A frequency distribution where several numbers are grouped into one class.

Class Limits

Separate one class in a grouped frequency distribution from another. The limits could actually appear in the data and have gaps between the upper limit of one class and the lower limit of the next.

Class Boundaries

Separate one class in a grouped frequency distribution from another. The boundaries have one more decimal place than the raw data and therefore do not appear in the data. There is no gap between the upper boundary of one class and the lower boundary of the next class. The lower class boundary is found by subtracting 0.5 units from the lower class limit and the upper class boundary is found by adding 0.5 units to the upper class limit.

Class Width

The difference between the upper and lower boundaries of any class. The class width is also the difference between the lower limits of two consecutive classes or the upper limits of two consecutive classes. It is not the difference between the upper and lower limits of the same class.

Class Mark (Midpoint)

The number in the middle of the class. It is found by adding the upper and lower limits and dividing by two. It can also be found by adding the upper and lower boundaries and dividing by two.

Cumulative Frequency

The number of values less than the upper class boundary for the current class. This is a running total of the frequencies.

Relative Frequency

The frequency divided by the total frequency. This gives the percent of values falling in that class.

Cumulative Relative Frequency (Relative Cumulative Frequency)

The running total of the relative frequencies or the cumulative frequency divided by the total frequency. Gives the percent of the values which are less than the upper class boundary.

Histogram

A graph which displays the data by using vertical bars of various heights to represent frequencies. The horizontal axis can be either the class boundaries, the class marks, or the class limits.

Frequency Polygon

A line graph. The frequency is placed along the vertical axis and the class midpoints are placed along the horizontal axis. These points are connected with lines.

Ogive

A frequency polygon of the cumulative frequency or the relative cumulative frequency. The vertical axis the cumulative frequency or relative cumulative frequency. The horizontal axis is the class boundaries. The graph always starts at zero at the lowest class boundary and will end up at the total frequency (for a cumulative frequency) or 1.00 (for a relative cumulative frequency).

Pareto Chart

A bar graph for qualitative data with the bars arranged according to frequency.

Pie Chart

Graphical depiction of data as slices of a pie. The frequency determines the size of the slice. The number of degrees in any slice is the relative frequency times 360 degrees.

Pictograph

A graph that uses pictures to represent data.

Stem and Leaf Plot

A data plot which uses part of the data value as the stem and the rest of the data value (the leaf) to form groups or classes. This is very useful for sorting data quickly.

Descriptive statistics describe the main features of a collection of data quantitatively.[1] Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to summarize a data set quantitatively without employing a probabilistic formulation,[2] rather than use the data to make inferences about the population that the data are thought to represent. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented. For example in a paper reporting on a study involving human subjects, there typically appears a table giving the overall sample size, sample sizes in important subgroups (e.g., for each treatment or exposure group), and demographic or clinical characteristics such as the average age, the proportion of subjects of each sex, and the proportion of subjects with related comorbidities.

Inferential statistics

Inferential statistics tries to make inferences about a population from the sample data. We also use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one, or that it might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Use in statistical analyses

Descriptive statistics provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of quantitative analysis of data.[

Descriptive statistics summarize data. For example, the shooting percentage in basketball is a descriptive statistic that summarizes the performance of a player or a team. This number is the number of shots made divided by the number of shots taken. A player who shoots 33% is making approximately one shot in every three. One making 25% is hitting once in four. The percentage summarizes or describes multiple discrete events. Or, consider the scourge of many students, the grade point average. This single number describes the general performance of a student across the range of their course experiences.

Statistical treatment of data is essential in order to make use of the data in the right form. Raw data collection is only one aspect of any experiment; the organization of data is equally important so that appropriate conclusions can be drawn. This is what statistical treatment of data is all about.

Don't have time for it all now? No problem, save it as a course and come back to it later.

here are many techniques involved in statistics that treat data in the required manner. Statistical treatment of data is essential in all experiments, whether social, scientific or any other form. Statistical treatment of data greatly depends on the kind of experiment and the desired result from the experiment.

For example, in a survey regarding the election of a Mayor, parameters like age, gender, occupation, etc. would be important in influencing the person's decision to vote for a particular candidate. Therefore the data needs to be treated in these reference frames.

An important aspect of statistical treatment of data is the handling of errors. All experiments invariably produce errors and noise. Both systematic and random errors need to be taken into consideration.

Depending on the type of experiment being performed, Type-I and Type-II errors also need to be handled. These are the cases of false positives and false negatives that are important to understand and eliminate in order to make sense from the result of the experiment.

Treatment of Data and Distribution

Trying to classify data into commonly known patterns is a tremendous help and is intricately related to statistical treatment of data. This is because distributions such as the normal probability distribution occur very commonly in nature that they are the underlying distributions in most medical, social and physical experiments.

Therefore if a given sample size is known to be normally distributed, then the statistical treatment of data is made easy for the researcher as he would already have a lot of back up theory in this aspect. Care should always be taken, however, not to assume all data to be normally distributed, and should always be confirmed with appropriate testing.

Statistical treatment of data also involves describing the data. The best way to do this is through the measures of central tendencies like mean, median and mode. These help the researcher explain in short how the data are concentrated. Range, uncertainty and standard deviation help to understand the distribution of the data. Therefore two distributions with the same mean can have wildly different standard deviation, which shows how well the data points are concentrated around the mean.

Statistical treatment of data is an important aspect of all experimentation today and a thorough understanding is necessary to conduct the right experiments with the right inferences from the data obtained.