CHAPTER 9

INTRODUCTION TO STATISTICS AND RESEARCH DESIGN

Author: Chad Knoderer, PharmD, FPPAG

There will be interactions that pharmacists have with other providers that will require the pharmacist to be able to definitively and confidently support their position or perspective with evidence, or published medical literature. Along with the ability to utilize drug information databases and effectively searching medical literature, pharmacists must also be able to understand and discuss research and research findings. Collectively these skills enable the pharmacist to practice evidence-based medicine.

Not all research is the same. Understanding the different types of research will help the pharmacist when advocating for the patient. Generally, the type of research that pharmacists will use will be clinical research, or research that is examining questions related to clinical practice, outcomes, or both.[1] This is different from laboratory research, or basic research, which may be experimental or exploratory in nature. Basic research is conducted in a controlled laboratory setting and does not involve human subjects. Clinical research focuses on applications to patient care and uses a systematic process to answer clinical questions.[2] The goal of clinical research is to draw inference from the study sample to a larger population. A population would include every person with the specific disease or illness being studied and it's not realistic or feasible to include everyone in a study, so instead a sample is used. A sample is a group of subjects in the study that represent the population.

Generally clinical research will either directly involve human subjects by having humans participate in the study or use data from human subjects. In the latter, the data already exists, but humans aren't actively enrolled in the study. Instead human subjects' data is used in the research. To highlight this difference, Fuller and colleagues implemented a chronic obstructive pulmonary disease (COPD) clinic in a community pharmacy setting and designed a study to determine if pharmacists can accurately perform spirometry screenings.[3] Some rationale behind the study was that if pharmacists could accurately perform the screening, then they could improve their education of the patient about COPD risk factors and potentially increase patient participating in smoking cessation programs. A total of 175 patients were enrolled and completed a spirometry session.[3] In contrast, Chen and colleagues conducted a study to determine the effectiveness of a pharmacist-managed telephone

tobacco cessation clinic in a Veterans Affairs (VA) health system.[4] This study looked at data from patients who had participated in the clinic, and the data was obtained through the health system's electronic medical records. Both of these studies are examples of clinical research that aimed to answer a clinical question about smoking/tobacco cessation. A difference was that the study by Fuller actively included patients whereas the study by Chen looked at patient data, but didn't actually enroll a single human.

RESEARCH DESIGN

The way a study is designed is based on the question the study sets out to answer. Some questions can't be answered with certain study designs. Pharmacists will use information from studies during their interactions with other providers, and a good starting place in using the information is to determine what research question was being asked. The research question is developed by researchers by starting with a broad idea or problem, and this might be something observed in everyday practice. This larger idea or problem is then narrowed down into a smaller and specific question. Sometimes several specific research questions might come from one larger idea, and this would result in multiple studies. Once the researchers have finalized their research question, they will rephrase the question and write it into the study as the study objective. As pharmacists read over a study's objective, they can consider what the original question might have been and if the study was designed appropriately to answer the question.

The studies by Fuller and Chen offer a look at different study designs and studies with different perspectives. There are two perspectives a study might take. A prospective study progresses forward in time and a retrospective study looks at data or information or events from the past. Fuller and colleagues started at one point in time and progressively enrolled patients and followed those patients into the future. Chen and colleagues looked back in time at a specific time period and collected data that was obtained during that time period. There are strengths and limitations of each perspective. Retrospective studies might be easier to conduct and cost less, but are limited by confounding factors. Prospective studies provide the research with more control, which will be discussed later in the chapter, but are potentially limited by the cost or time they take. Another limitation of prospective studies is patient enrollment, especially for studies looking at less commonly encountered conditions. It's important for pharmacists to understand these differences when discussing clinical research with providers.

Other aspects of research design will be discussed throughout this chapter. Aspects such as control types refer to what type of control, if any, is used in the study. Some studies may not

include a control group, whereas other designs might include a placebo, historical, crossover (where the subject serves as their own control), or active (e.g. standard of care) control.[5] Studies should be designed to increase both internal and external validity. Validity relates to how well a test is measuring what is intended to be measured.[6] Internal validity, then, is how well the study was designed to minimize errors within the study. Systematic and random error can impact internal validity.[2] Systematic errors include bias and confounding and random errors include chance.[2] Chance is what it sounds like and is the likelihood that a finding is due to random error. Statistical tests are used to quantify or determine the level of chance.[2,5] Bias occurs when there is a systemic error in subject selection or the measurement/collection of observations that leads to a false conclusion of association.[2] Confounding is commonly heard when thinking about variables. For example, a study might be looking to determine if a variable is a risk factor for an outcome variable. Confounding variables are those that are associated with the predictor variable, but are also themselves predictors of the outcome independent of the variable being studied. But, the confounding variable is not part of the relationship between the study variable and outcome of interest.[7] Pharmacists evaluate internal validity of studies by reading through the methods and analyses and determine if they were appropriate to answer the research question. Evaluation of the results, to determine if they are accurate based on the methods, and reading through the conclusions to make sure conclusions are supported by the data are additional ways for pharmacists to assess internal validity.

External validity is slightly different and refers to how the study is applied to a wider population. External validity can be thought of as related to generalizability.[2] The ability to apply the findings from the study and make inference to a wider population is generalizability and this is an area that is sometimes difficult to navigate in interactions with practitioners. Issues of how easily findings from a study can be applied to a cohort within a certain practice arise commonly and these are instances where pharmacists can be pivotal in these conversations. Ultimately, however, if there were errors leading to a decrease of internal validity, then there will be decreased external validity. Without internal validity, there can be no external validity. Some methods to improve both internal and external validity include reducing confounding variables, improve subject selection, blinding, using a control group, and using objective data and validated measurement techniques.

Design Types

In addition to the study perspective, there are different types of study design types that could be used to answer the research question. Broad types include observational and

interventional studies, which are commonly referred to as clinical trials. As the name indicates, there is no intervention in an observational study, rather events or patients are observed either prospectively or retrospectively. Descriptive studies describe findings from observational studies. Although statistics are commonly used to describe the findings, there won't be any statistical comparisons of patient groups. Analytical studies make use of statistical tests or analysis to determine presence of any association among variables and to make inference. There are several different designs within observational studies. A cohort study, makes observations of a group over time. These can be retrospective or prospective and can be descriptive or analytical. Cohort studies can evaluate independent variables (or risk factors) for development of diseases or conditions. This design is related to exposure. The Framingham Heart Study is a well-known and classic example of a prospective cohort study. Investigators prospectively followed over 5000 residents of Framingham, Massachusetts in order to get a better understanding of risk factors for cardiovascular disease.[8] Nearly 70 years and hundreds of subsequent studies later, information obtained from the original Framingham cohort has shaped the way practitioners approach patient care. A recent publication by Bolesta and Kong describes a retrospective and analytical cohort study to determine the impact of the 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase inhibitors (also known as statins) on post-operative atrial fibrillation.[9]

Case-control studies are different from cohort studies in that there will be 2 distinct groups of subjects, but the design first finds cases of subjects with the outcome of interest (the condition or disease) and compares those subjects with a control group of subjects to determine exposure to an independent variable (risk factor). These are retrospective and the design starts with the disease or condition of interest. For example, a 2013 study by Barletta and colleagues is a case-control study aimed at determining any associations between proton-pump inhibitor usage and Clostridium difficile infection (CDI).[10] With this design the investigators first identified patients who had a hospital-acquired CDI. This made up the case group. The investigators then identified patients within the same time period who did not have CDI and used those patients to form the control group.

Cross-sectional studies collect data or make observations at a single point in time. Surveys are a common type of cross-sectional studies. Surveys will be distributed and the collected responses will reflect data from the respondent at a single time point. These study designs can be efficient and relatively cheap to conduct. The designs can be used for different target recipients (ie. patients or practitioners), but they are limited by the observations being at one point in time. Significant consideration must be made to the development of the survey questions and the types of response scales used in order to get the most accurate and unbiased

information from the target sample. The biggest limitation is participation, in that subjects may not participate. Survey response rate is important to consider when reviewing these types of studies. Owenby and colleagues distributed surveys to VA pharmacy clinical coordinators to characterize pharmacy services within VA emergency departments.[11] In contrast, Li and colleagues surveyed patients with recent medical visits to assess factors related to their search of online health information.[12] Both of these examples are cross-sectional studies, but with different target recipients.

Clinical trials are interventional studies that are designed to evaluate the efficacy and/or safety of an intervention. Parallel designs, where subjects receive only the study intervention or the control throughout the study, are common. Crossover designs, where subjects serve as their own controls may also be used. Randomized controlled trials (RCTs) are the gold standard of clinical trials because of the degree of experimental control utilized in the study. These are also going to be the most expensive to conduct. To pay for RCTs, researchers may seek federal or organizational grants to pay for the study. In other situations, and pharmaceutical or medical device company may pay to conduct the study.

To better understand RCTs, it's important to understand all of the aspects that go into one. Randomization is one key component of RCTs, and this is done at the beginning of the study to make the study groups similar. The goal of randomization is to make any observed differences in the study group due solely to chance. The group differences ultimately are balanced through randomization. One way to think of randomization is to think of flipping a coin. If a study has 2 study groups (also sometimes referred to as arms), a treatment and control, then a researcher might flip a coin to randomize assignment. Subjects with heads on the coin flip could be assigned to the treatment group and those with tails on the coin flip could be assigned to the control group. There are many more sophisticated randomization methods, but at the core randomization is intended to assure that subjects have an equal chance of being assigned to each study group, and baseline confounding variables are eliminated. Blinding is a technique used to hide group assignments. The intent is to reduce or eliminate information bias from the participant or observer/researcher. Blinding can be single or double. Single is where only one side is blinded (subject or researcher) and double is where both subject and researcher are unaware of the group assignment. It can be difficult to blind some interventions. For example, it might be difficult in a study comparing liquid clindamycin and amoxicillin because liquid clindamycin is well known for having a bad taste and amoxicillin is more palatable.

Understanding RCT and being able to discuss RCT findings with practitioners also requires

knowledge of the analysis technique utilized in the study. There are two general types, intention to treat (ITT) and per protocol (PP). Either may be used and the methods section of the article should state which was utilized. With the ITT analysis, outcomes are compared based on subjects' assignments and no data is eliminated. If subjects drop out of the study or are lost to follow-up their data is still included, but the outcome is characterized as no outcomes. This provides a conservative estimate of efficacy difference between groups, but a more liberal estimate of toxicity differences. Per protocol is different in that it compares outcomes based on subjects who followed the study protocol and eliminates data of subjects who did not complete study. This provides higher degree of efficacy difference between groups for efficacy outcomes, but lower degree of difference for toxicity outcomes. For example, a study comparing drug A and drug B is evaluating which drug has better treatment success. If 20 patients are included in each study group, outcomes will be compared. Let's say for drug A, 15 subjects have success, 4 have failure, and 1 drops out. For drug B, 15 subjects have success, 1 has failure, and 4 drop out. Success rates for drug A and drug B would be equal with the ITT technique (ie. both at 75%), but would be 79% (15 of 19 completing protocol) vs. 94% (15 of 16 completing protocol), respectively, with the PP technique. The opposite would be true if a toxicity were an outcome of interest. There are studies that will perform analyses based on both the ITT and per protocol to evaluate if there were different findings based on analysis strategy.

There are different types of RCT. Table 9-1 highlights some differences between superiority, equivalence, and non-inferiority designs.[13] When interacting with practitioners about equivalence or non-inferiority studies, it should be noted that a fundamental concept in these studies is that they are based off of a superiority RCT that has been previously conducted. It would be important, then, to read through the original superiority RCT before discussing the merits of the non-inferiority or equivalence study. A couple of terms will continually be noted for both equivalence and non-inferiority studies. The delta (Δ) notes the least relevant difference and should be considered a marker of clinical significance or importance. This will be established by researchers and should generally be based on existing literature. The non-inferiority margin is the pre-determined maximum difference between "intervention" and "control", and is another term for delta. The delta and non-inferiority margin will be utilized for interpreting study results.

STATISTICAL ANALYSIS

Even the term statistical analysis has a tendency to bring out a variety of uncomfortable responses from practitioners, but a common response is one of confusion and hesitancy.

Type	Method	Goal
Superiority	Compares investigational treatment to control (ie. placebo, standard of care)	Show that investigational treatment is superior to control or standard treatment
Equivalence		Show that the treatments differ by an amount that is NOT clinically important (ie. difference between – Δ and + Δ)
Non-inferiority		Show that investigational treatment is NOT clinically inferior to control (ie. can be worse, but no worse than – Δ)

Table 9-1. Randomized controlled trial types[a]
a = Table modified from reference #47, Δ = delta

Many practitioners find statistics intimidating, almost like a foreign language. Regardless of the response, the reality is that a basic understanding of how to apply statistical analysis techniques is needed if pharmacists are to have meaningful interactions with practitioners and optimize patient care. The majority of pharmacists working in patient care roles won't be doing statistical analysis, just like they won't be doing research. As was highlighted previously, a good understanding of research design and all of the elements that go into doing research will allow pharmacists to be smarter consumers leading to more effective interactions with other practitioners and more optimal patient advocacy. This is the case for statistical analysis as well.

The statistical analysis section of a published study may seem like a good area to pass over quickly. Perhaps some authors may not even spend much time writing this section. Spending some time evaluating the statistical analysis section of a published study may allow the reader to discover that researchers used inappropriate statistical methods or potentially manipulate the findings. Understanding statistics can help a reader connect the dots in a published study. That stated, statistical analysis is another piece of the overall research design. It is connected with the research design, and ultimately the research question. For example, the goal of clinical research is to draw inference from a study sample to a larger population, but the validity of that inference is dependent upon the appropriate design being married with the

appropriate statistical analysis. If one of the two is not appropriate, then the study's validity may be called into question. If the validity is questionable, then the study findings may not be appropriate to utilize to drive patient care decisions. All of this is complicated. Knowing where to start and the process to use to apply knowledge of statistical analysis to patient care situations is crucial in overcoming the intimidating barrier that can sometimes arise from statistics.

Data Types and Scales

The type of statistics that are used depends on the type of data that is collected, and there are different types of data and data scales for discussion. Remember that some practitioners will use the term variables or parameters or even outcomes, but all of these terms are also terms for data, and data can be considered either categorical or continuous. Categorical are qualitative and can be placed into categories or buckets. Examples of categorical data include hair color or grade level. Continuous data are quantitative and can take any range of possibilities. Examples of continuous data include age, weight, or blood pressure. Data are further considered on scales, and for the purposes of the medical literature there are 4 data scales (**Table 9-2**).

Scale	Characteristic	Statistics	Example
Categorical or Discrete			
Nominal	Unordered categories	Counts, frequencies, percentages	Gender, diagnosis
Ordinal	Ordered categories	Above + medians	Likert scales, pain scales
Continuous			
Interval	Arbitrary zero	Above + means, standard deviations	Temperature (C, F)
Ratio	Absolute zero		Weight, age

Table 9-2. Data scales

The nominal data scale is the most basic data scale. Data falling onto this scale can be classified into two or more categories that are mutually exclusive from one another and are exhaustive. The categories are unordered and have no relation to each other. Dichotomous data are data that can be placed into one of two categories like alive or dead, pregnant or non-

pregnant. Dichotomous data is a specific type of nominal data. The ordinal scale includes ordered categories with some rank and relation to another. Data derived from surveys using Likert-scale questions (ie. strongly agree, agree, disagree, strongly disagree) or measurement scales such as pain rating scales are examples of ordinal data. While ordinal data may be reported as a number (ie. pain score of 7 or Likert score of 4) the number corresponds to a category, and there is no consistent degree of difference between each category (ie. 3 is not 3 times greater than 1). Continuous variables can take on an infinite number of possibilities. With continuous data there is both an order to the values and a consistent degree of difference between each value. Continuous data can be placed on either the interval or ratio scale with the difference being that ratio data has real zero (ie. heart rate) compared to an arbitrary zero (ie. temperature on fahrenheit scale) of for interval data.[14] Any continuous data could be turned into either nominal or ordinal categorical data, but the opposite is not true. Categorical cannot be converted to continuous data.[14]

Descriptive Statistics

Statistics should be considered as two broad categories, descriptive and inferential. As the name implies, descriptive statistics describes study observations. These are very useful to research consumers because descriptive statistics allow the reader an understanding of the subjects that were included in the study. Data within these data scales can take on a number of possible values, making it important to understand data distribution. The binomial and Poisson distributions are examples of distributions of categorical data.[15] The normal distribution is a distribution of continuous data. With this, data are equally distributed around the mean and the values of the mean, median, and mode are all equal. The next paragraph describes mean, median, and mode. The curve of a normal distribution takes on a bell appearance, and it's often referred to as a bell-shaped curve. **Figure 10-1** illustrates a normal distribution curve of age that came from a dataset containing 229 patient observations and in addition to the shape of the curve, the distribution can provide a lot of useful information to providers. Some examples of these will follow.

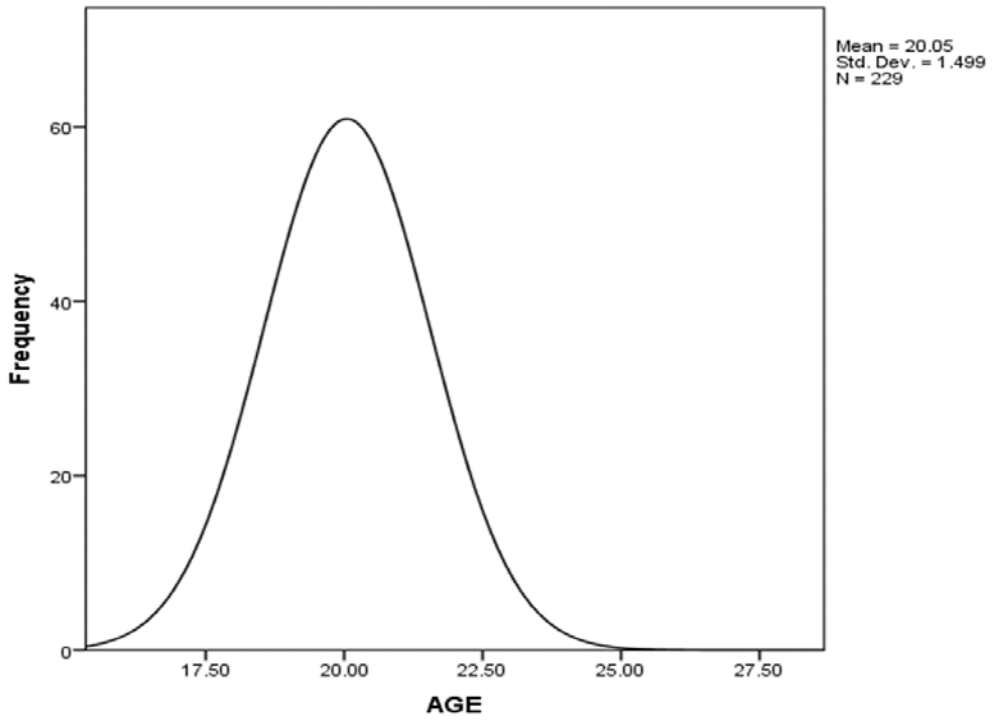

Figure 9-1. Normal distribution curve

Measures of central tendency and measures of variability are types of descriptive statistics. Measures of central tendency include mean, median, and mode and describe the main observations of the dataset while measures of variability are used to describe the amount of uncertainty in the dataset. The mean is the average of the values in the dataset and can be calculated for continuous data. The mean is highly affected by outliers making it suboptimal to describe data that aren't normally distributed. Because there is inconsistent degree of difference between ordinal categories, the mean is not useful to describe ordinal data. The median is the middle measurement (ie. 50^{th} percentile) in a dataset and can be used for ordinal data or continuous data that aren't normally distributed. There are an equal number of values above and below the median making it resistant to the influence of outlying variables. The mode is the value occurring most often in a dataset. The standard deviation (SD) is a measure of variability commonly used with the mean because it represents deviation from the dataset's mean. The SD is the square root of variance which is derived from sum of squares. The mean of a dataset is subtracted from each data observation, and all of those deviations are added together to make the sum of squares. Neither the variance nor sum of squares are used commonly in the medical literature, but the SD is and knowing its origins helps make sense of what information the SD provides practitioners. Because the mean is used in the calculation

of the SD, this variability measure is most appropriate for normal or near normally distributed data. Using the SD, a reader can apply the empirical rule for additional information about the dataset. The empirical rule states that for normally distributed data, 68% of observations will fall between 1 SD of the mean, 95% of observations will fall between 2 SD, and 99% of observations will fall between 3 SD. **Figure 9-2** provides an illustration. Applying this rule to the mean and SD within **Figure 10-1** and learn that 68% of the patients had ages of 18.5 – 21.5 years, 95% were 17.1 – 23 years, and 99% were 15.5 – 24.5 years. This type of application of descriptive statistics is particularly useful as a provider and when discussing with other providers because it allows for some generalization. Using data from a study where 95% of the patients were 17.1 – 23 years old wouldn't be useful or appropriate for providers who predominantly care for elderly patients.

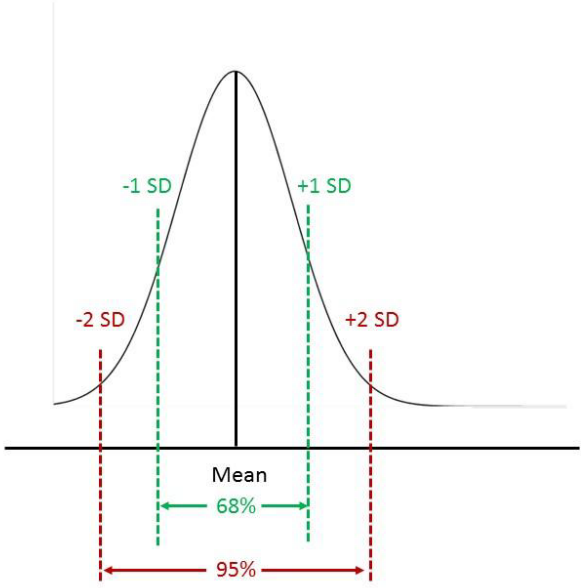

Figure 9-2. Normal distribution curve with 68, 95 rule

The standard error of the mean (SEM) may be commonly confused with or used in place of SD. However, the SEM quantifies certainty in the estimate of the true population mean where the SD is providing an estimate of variability around a sample mean. The two are not interchangeable, but the SEM will always be smaller leading some researchers to report it because of the feeling that less variability is better. A reader can always determine the SD

from the SEM by taking the SD divided by the square root of the number of observations:

$(SEM = SD \div \sqrt{n})$.

Using the data from **Figure 10-1**, the SEM would be calculated as:

$1.499 \div \sqrt{229} = 0.098$

which is obviously smaller than the reported SD of 1.499. This equation can be used to determine the SD when the SEM is inappropriately reported. Standard errors are used in the calculation of confidence intervals (CI) which estimate a range of values that is likely to contain a population value a certain percentage of time. These are commonly reported in the literature as 95% confidence intervals. Confidence intervals will be further explained later in the chapter. The interquartile range (IQR) represents data between the 25th and 75th percentiles and contains the middle 50% of observations. This measure of variability is commonly reported as 2 numbers with the highest representing the 75th percentile and lowest representing the 25th percentile. The IQR is particularly useful for non-normally distributed continuous data, along with the median, or for ordinal data. Other types of descriptive statistics include frequencies and percentages or visual data displays such as histograms, scatter plots, and box plots (**Figures 9-3 - 9-5**).

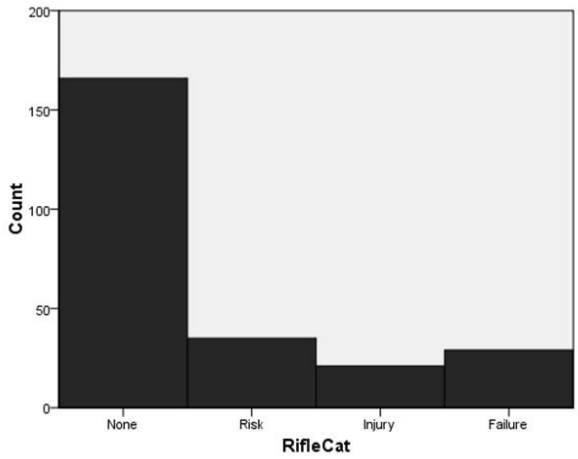

Figure 9-3. Histogram

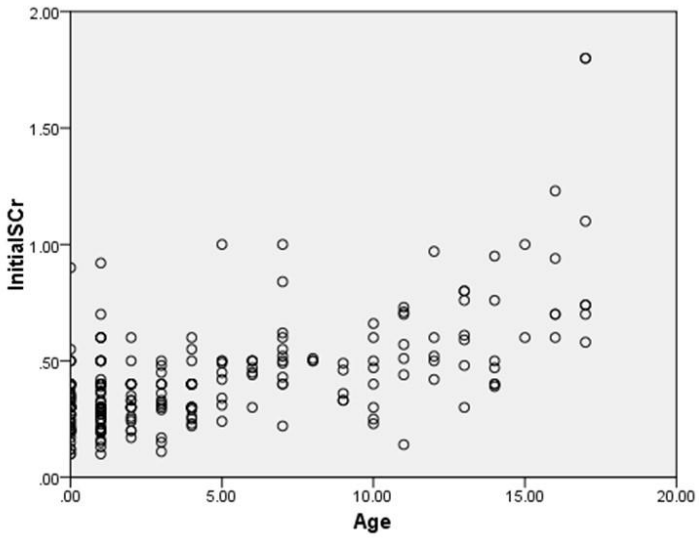

Figure 9-4. Scatterplot

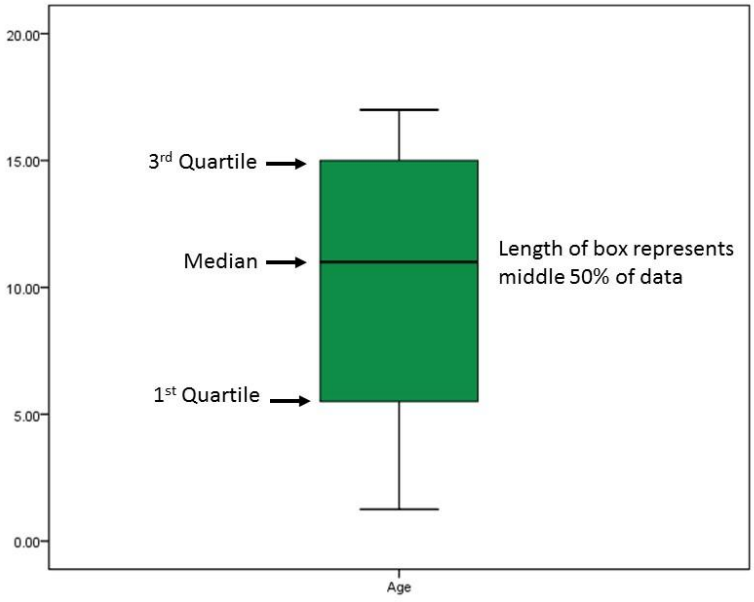

Figure 9-5. Box plot

Inferential Statistics, Hypothesis Testing, and Decision Errors

Inferential statistics are used to make a generalization about a population from the results of a study and practitioners will use study finding to make decisions on how to manage the patients in their care. The foundation of inferential statistics is with hypothesis testing, and statistical tests are used in hypothesis testing. Not all studies will be testing hypotheses, so not every study will include inferential statistics, but all studies that are reporting findings will include some element of descriptive statistics. With hypothesis testing, statistical tests are used to test the null hypothesis (H_0) which states that there is no difference between predictor and outcome variables. In other words, if a study were comparing two interventions (Intervention A and Intervention B) the H_0 would state that Intervention A = Intervention B. The alternate hypothesis (H_1) would then state that Intervention A ≠ Intervention B. The H_1 cannot be directly tested, so it is accepted by default in instances where the H_0 is rejected. The H_0 is rejected based on the results of a statistical test. In general, for studies which pharmacists will be reading and using, hypothesis testing is two-sided. In other words, the researcher is interested in a difference in either direction. It might be important if Intervention A is better than or worse than Intervention B. Occasionally, inferential statistics will be used for one-sided hypothesis testing. In a one-sided hypothesis, a direction of difference is stated and tested. For example, the H_0 might be written as Intervention A ≥ Intervention B with the H_1 written as Intervention A < Intervention B.

For statistical testing, researchers determine which statistical test should be used because there are different tests for different types of data or the number of groups being compared. Once the researchers decide on the statistical test, the level of significance (also referred to as alpha or α) is selected. This is generally set at 0.05 and it essentially means there is a 5% probability of incorrectly rejecting the null hypothesis (ie. making a Type 1 error). Once the alpha is set, a critical value will be determined based on what statistical test will be used. The critical value corresponds to the pre-set α. A common place where statistical testing becomes confusing for providers is what it all means. The p-value is the probability of making an observation as extreme or more extreme than the observed if the H_0 were true.[17] The statistical test is then done and the test statistic is compared to the critical value and a p-value is calculated. P-values less than the preset alpha are considered to be statistically significant, the H_0 is rejected, and by default the H_1 is accepted.

A recent study by Hammond and colleagues sought to compare the incidence of acute kidney injury (AKI) in patients who were receiving vancomycin plus either piperacillin/tazobactam or cefepime.[18] In this example, the H_0 would be AKI with PT = AKI with cefepime. This study

had two groups (PT vs. cefepime) and the primary variable being compared was AKI which was a nominal variable. Patients either had AKI or didn't making this a categorical and nominal variable. Because the researchers were comparing a nominal variable between two groups, the chi-square test was most appropriate. More on this test will follow. The researchers state that the pre-determined $\alpha = 0.05$, which meant that the corresponding critical value of the chi-square test = 3.84. In general, textbooks of statistical analysis will include supplementary tables containing critical values for statistical tests, so this information is easy to obtain. With a critical value of 3.84, any test statistic greater than 3.84 would correspond to a p-value of < 0.05 (**Figure 9-6**). Hammond and colleagues reported that there was no difference in AKI between the PT (32.7%) and cefepime (28.8%) patients and that the p-value = 0.761. The actual test statistic was not reported, and it typically is not in medical literature, but a reader can assume from the reported p-value of 0.761 that the test statistic resulting from the chi-square test was less than the critical value of 3.84. With this example, because the test value was less than the critical value, the researchers could not show a difference in AKI between the drug regimens and would ultimately conclude that there is no AKI difference between patients receiving PT and those receiving cefepime.

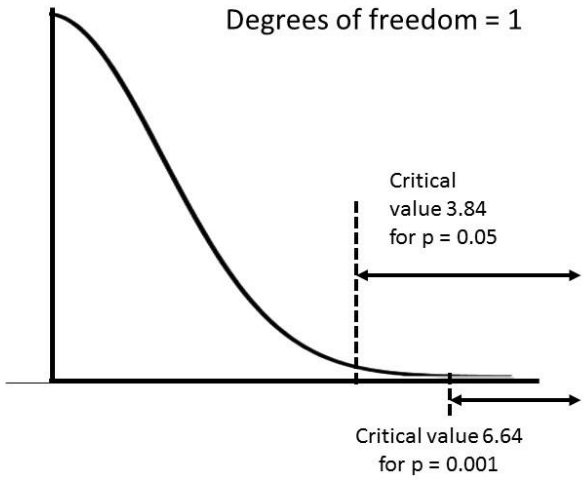

Figure 9-6. Chi-square critical value and p-value

With inferential statistics, there is always a chance of error and the two types of error that can be encountered are Type 1 (false-positive) and Type 2 (false-negative). These errors are due to chance variability or bias and the likelihood of encountering them may decrease with increasing sample sizes. With a type 1 error the researcher rejects an H_0 that is actually true.[19] The probability of making a type 1 error is signified by pre-set alpha (α) and the p-value that is determined from the test provides a more precise probability of making a type 1 error. A type 2 error occurs when a researcher fails to reject an H_0 that is actually not true.[18] Beta (β) is the probability of making a type 2 error and this typically ranges from 0.1 – 0.2. There are several ways to think about decision errors so that they make more sense. One way to consider is related to that of a jury decision.[20] An innocent person did not commit a crime and a guilty person did commit a crime. If a jury convicts the criminal and acquits the innocent, then the correct decision was made. But, if the jury convicts the innocent person, a type 1 error has been made because the jury has concluded that there is an association when in reality there is not. Likewise, if a jury acquits the criminal they have made a type 2 error by concluding that there is no association when there actually is one.[20] The popular belief of those who followed or were familiar with the O.J. Simpson murder trial was that Simpson did commit the murders although the jury acquitted him (ie. making a type 2 error).[21] Contrast that with the case of Mark Schand who spent 27 years in prison after a jury wrongfully convicted him of committing murder (ie. making a type 1 error).[22] In medicine, type 1 errors may result in drugs or treatment being introduced to market or practice when they are truly no different than standard treatments. Concerning are the potential downfalls of the new treatments which likely are more expensive than standard care, but also may have more side effects.

Power is a term that is used quite a bit by practitioners who might say that a particular study was underpowered or that a study had adequate power. Power is the probability of correctly rejecting the H_0. This might also be considered as the probability to detect a difference when the difference exists. The equation for power is:

$$Power = 1 - \beta$$

and it's dependent upon several things. The predefined α, sample size, estimated size of difference between outcomes (ie. the difference that researchers are trying to detect), outcome variability, and the statistical test can all influence power. Sample size and power are connected in that while sample size can influence power, power is also considered when determine the required sample size. Hammond and colleagues determined that 122 patients were needed to detect an approximate 20% difference in AKI between PT and cefepime.[18]

Researchers established the α at 0.05 and β at 0.2 for a power of 80%. The sample size was based on AKI estimates of 36.5% in the PT groups vs. 15% in the cefepime group. While beyond the scope of this text, there are resources that would allow the reader to compute the actual power of this particular study.

With inferential statistics, there is always a degree of uncertainty and p-values will tell one aspect of the story, which is about statistical significance. But, the p-values themselves don't provide information about the clinical impact or size of any difference, if present. Confidence intervals are useful to determine the size of the potential difference because confidence intervals give a range of possible estimates including the observed value. All values within the CI are statistically possible. For examples, authors report no difference in mean ± SD vancomycin dosing (mg/kg/day) in pediatric patients with (47.5 ± 14.6) and without AKI [41.2 ± 16.6) with a p-value = 0.102.[23] This finding is not statistically significant, so the researchers fail to reject the H_0 that the dose in those with AKI = dose in those without AKI. But, is this finding clinically meaningful? Some practitioners might say yes, leading to the suggestion that dose is an important consideration with AKI, while other practitioners may disagree. The actual difference between the doses was -6.3 mg/kg/day, but if the 95% CI of the difference would have been included, it would have been -13.8 to 1.25. Remembering that the CI is the range of values that could contain the true population value, one could interpret this 95% CI by saying there is a 95% certainty that the true dosing difference lies between -13.8 and 1.25 mg/kg/day. This perspective provides more clinical importance to the finding. Some practitioners might agree that a dosing difference of nearly -14 mg/kg/day is clinically important for this drug, but they may also to say that a difference of 1.25 mg/kg/day is not important at all. It should be noted that within this 95% CI is 0, meaning that the true dosing difference could be also zero. This is how CIs can also be used to test hypotheses. If the H_0 is that the dose in those with AKI = dose in those without AKI, or there is 0 difference in dose, then the 95% CI tells a reader that the H_0 should not be rejected because 0 is contained within the CI. A more simplified approach to determining statistical significance from reading a 95% CI is this. If estimating a difference between continuous (ie. means) or categorical variables (ie. percentage or proportions), CIs containing zero (0) are NOT statistically significant (ie. the p-value > 0.05), and the H_0 should not be rejected. If estimating an odds ratio (OR), relative risk (RR), or hazard ratio, CIs containing one (1) are NOT statistically significant, (ie. p-value > 0.05), and the H_0 should not be rejected.

Common Statistical Tests

Statistical tests are widely reported in both pharmacy and medical literature.[24,25] Common

statistical tests can be divided into parametric and non-parametric categories. Assumptions for parametric tests are that the data is continuous with a normal or near-normal distribution, data is randomly obtained, the observations are independent of one another, and the variances between groups are equal. Parametric tests include t-tests and analysis of variance (ANOVA) and are considered more powerful than non-parametric tests. The t-test compares means between no more than 2 groups. These will be used to test an H_0 that a mean in one group is the same as the mean dose in another group, similar to the previous example of mean vancomycin dose and AKI. There are different types of t-tests. A one-sample compares the mean in a study group to that of a known population. Independent samples t-tests (also called Student t-tests) compare means of two unrelated groups (ie. Group 1 vs. Group 2), where paired t-tests compare the means of dependent observations (ie. measurement 1 vs. measurement 2 in one group). A paired t-test would be commonly used in a before-after type of study design. Analysis of variance compares means of 3 or more groups. The ANOVA is more powerful than a t-test when there are 3 or more groups because the alpha is held constant. There are a different types of ANOVA: one-way (ie. mean in Group 1 vs. Group 2 vs. Group 3), two-way (ie. comparing two factors in Group 1 vs. Group 2 vs. Group 3), and repeated measures (ie. one group but comparing the mean of Measurement 1 vs. Measurement 2 vs. Measurement 3). With ANOVA, the test result will correspond to a p-value that tells the researcher or reader if a significant difference exists, but the test doesn't signify where the difference exists. If an ANOVA yields a significant p-value, researchers must perform further tests, called post-hoc tests, to determine precisely where the difference lies. Pearson correlation is a parametric procedure used to examine the direction and strength of relationship between 2 normally distributed continuous variables. The Pearson correlation coefficient (r) ranges from -1 to +1 with values closer to 1 representing a stronger relationship between the variables. Positive numbers indicate direct or positive relationships with negative numbers representing inverse or negative relationships. In general r values of < 0.3 are considered weak, 0.3 – 0.5 are moderate, and > 0.5 are considered strong relationships.[26] The coefficient of determination (r^2) is the percentage of variance in the dependent variable that is explained by the other.[26] Lee and colleagues evaluated the relationship of serum creatinine concentrations on antifactor Xa concentrations and found an $r = -0.262$.[27] This can be interpreted as a weak negative relationship. The r^2 value = 0.0688 which means that 7% of the variance in antifactor Xa concentrations can be explained by serum creatinine concentrations.

Non-parametric tests are used when the assumptions of parametric tests are violated. Several analogies can be made to parametric tests. The Wilcoxan rank sum and Mann-Whitney U are analogous to the independent samples t-tests and are used for two independent samples of

non-normally distributed continuous data or ordinal data. The Wilcoxan signed rank test is analogous to the paired samples t-test and would be used for data types above from related samples. The Kruskal-Wallis test can be considered analogous to ANOVA, and would be used for non-normally distributed continuous data or ordinal data from 3 or more groups. For nominal data, Chi-square analysis can be used to compare proportions between 2 or more independent groups. An assumption of the Chi-square analysis is that of sufficiently large expected frequencies, so when this is violated (eg. when any cell has less than 5 expected frequencies) the Fisher exact test is more appropriate. The McNemar test can be used to compare proportions in paired samples. Lastly, Spearman correlation is analogous to Pearson correlation and is used for continuous variables that aren't normally distributed or for nominal and ordinal data. Range of values from the Spearman rho (r_s), which is the output from a Spearman correlation, are negative 1 to positive 1 and interpretation is similar as Pearson. **Table 9-3** describes common statistical tests and their appropriate uses.

Regression determines how one variable predicts another variable, but it shouldn't be considered as establishing a cause-effect relationship. Simple regression is a relationship between a single dependent and independent variable; multiple regression is a single dependent with multiple independent variables. There are 2 types of regression. Linear regression is used when dependent variable is continuous and normally distributed, and the independent variable is continuous (simple regression) or either continuous or categorical (multiple regression). The output from linear regression is the coefficient of determination (r^2) which explains how clearly the model describes the relationship and beta (β) which describes the change in the dependent variable caused by a 1-unit change in the independent variable. Logistic regression differs in that the dependent variable is categorical while the independent variables could be continuous or categorical. The result of logistic regression is an odds ratio that can be interpreted as previous described.

REFERENCES

1. Portney LG, Watkins MP. A concept of clinical research. In: Portney LG, Watkins MP, eds. *Foundations of Clinical Researcher: Applications to Practice.* 3rd edition. Upper Saddle River, NJ: Pearson Prentice Hall; 2009:p3-31.
2. Hartung DM, Touchette D. Overview of clinical research design. *Am J Health Syst Pharm* 2009; 66(4): 398-408. https://doi.org/10.2146/ajhp080300 .
3. Fuller L, Conrad WF, Heaton PC, Panos R, Eschenbacher W, Frede SM. Pharmacist-managed chronic obstructive pulmonary disease screening in a community setting. *J Am Pharm Assoc* 2012; 52:e59-66. https://doi.org/10.1331/JAPhA.2012.11100 .
4. Chen T, Kazerooni R, Vannort EM, et al. Comparison of an intensive pharmacist-managed telephone clinic with standard of care for tobacco cessation in a veteran population. *Health Promot Pract* 2014; 15(4): 512-20. https://doi.org/10.1177/1524839913509816 .
5. Hulley SB, Newman TB, Cummings SR. Getting Started: The Anatomy and Physiology of Clinical Research. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB, eds. *Designing Clinical Research*. 3rd edition. Philadelphia, PA: Lippincott Williams and Wilkins; 2007:p3-15.
6. Portney LG, Watkins MP. Reliability of Measurements. In: Portney LG, Watkins MP, eds. *Foundations of Clinical Researcher: Applications to Practice*. 3rd edition. Upper Saddle River, NJ: Pearson Prentice Hall; 2009:p77-96.
7. Portney LG, Watkins MP. Validity in Experimental Design. In: Portney LG, Watkins MP, eds. *Foundations of Clinical Researcher: Applications to Practice*. 3rd edition. Upper Saddle River, NJ: Pearson Prentice Hall; 2009:p161-191.
8. Mahmood SS, Vasan RS, Wang TJ. The Framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet.* 2014 Mar 15; 383(9921): 999-1008. https://doi.org/10.1016/S0140-6736(13)61752-3
9. Bolesta S, Kong F. Effect of statins on the incidence of postoperative atrial fibrillation after cardiac valve surgery. *Pharmacotherapy* 2015; 35:998-1006. https://doi.org/10.1002/phar.1655. Epub 2015 Nov 2.
10. Barletta JF, El-Ibiary SY, Davis LE, Nguyen B, Raney CR. Proton pump inhibitors and the risk for hospital-acquired *Clostridium difficile* infection. *Mayo Clin Proc* 2013;88:1085-90. https://doi.org/10.1016/j.mayocp.2013.07.004 .
11. Owenby RK, Brown JN, Kemp DW. Evaluation of pharmacy services in emergency departments of Veterans Affairs Medical Centers. *Am J Health Syst Pharm* 2015;72(suppl2):S110-4. https://doi.org/10.2146/sp150019 .
12. Li N, Orrange S, Kravitz RL, Bell RA. Reasons for and predictors of patients' online health information seeking following a medical appointment. *Fam Pract* 2014;31:550-6. doi: 10.1093/fampra/cmu034.
13. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. *J Gen Intern Med* 2010;26:192-6.
14. Gaddis ML, Gaddis GM. Introduction to biostatistics: part 1, basic concepts. *Ann*

Emerg Med 1990;19:86-9.
15. Daniel DW. Probability Distributions. In: Daniel DW, ed. *Biostatistics: A Foundation for Analysis in the Health Sciences*. 9th edition. Hoboken, NJ: John Wiley and Sons, Inc; 2009:p93-134.
16. DeMuth JE. Overview of biostatistics used in clinical research. *Am J Health Syst Pharm* 2009;66:70-81. https://doi.org/10.2146/ajhp070006 .
17. Glantz SA. The Special Case of Two Groups: The t Test. In: Glantz SA, ed. *Primer of Biostatistics*. 7th edition. San Francisco, CA: McGraw-Hill;2012:p49-72.
18. Hammond DA, Smith MN, Painter JT, Meena NK, Lusardi K. Comparative incidence of acute kidney injury in critically ill patients receiving vancomycin with concomitant piperacillin-tazobactam or cefepime: a retrospective cohort study. *Pharmacotherapy* 2016;36:463-71. **https://doi.org/10.1002/phar.1738** . Epub 2016 Apr 1.
19. Glantz SA. What Does "Not Significant" Really Mean? In: Glantz SA, ed. *Primer of Biostatistics*. 7th edition. San Francisco, CA: McGraw-Hill. 2012; p101-124.
20. Browner WS, Newman TB, Hulley SB. *Getting Ready to Estimate Sample Size: Hypotheses and Underlying Principles*. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB, eds. *Designing Clinical Research*. 3rd edition. Philadelphia, PA: Lippincott Williams and Wilkins; 2007:p51-63.
21. Ross P. Nicole Brown Simpson house murder anniversary: what we still don't know about June 12, 2994 case. *International Business Times*. June 11, 2016. Media and Culture. https://perma.cc/WN7B-Q8UP Accessed May 29, 2018.
22. Brown K. Life after wrongful conviction. *The New York Times*. May 28, 2016. Opinion. https://perma.cc/BW32-7NAX Accessed May 28, 2018.
23. Knoderer CA, Gritzman AL, Nichols KR, Wilson AC. Late occurring vancomycin-associated acute kidney injury in children receiving prolonged therapy. *Ann Pharmacother* 2015; 49:1113-19.https://doi.org/10.1177/1060028015594190
24. Lee CM, Soin HK, Einarson TR. Statistics in the pharmacy literature. *Ann Pharmacother* 2004;38:1412-8.https://doi.org/10.1345/aph.1D493.
25. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA* 2007; 298: 1010-22.
26. Overholser BR, Sowinski KM. Biostatistics primer: part 2. *Nutr Clin Pract* 2008;23:76-84. https://doi.org/10.1177/011542650802300176.
27. Lee YR, Vega JA, Duong HN, Ballew A. Monitoring enoxaparin with antifactor Xa levels in obese patients. *Pharmacotherapy* 2015;35:1007-15. https://doi.org/10.1002/phar.1658.