A Brief Primer

on

Basic Medical Statistics

by

Kevin P White, MD, PhD (Epidemiology & Biostatistics, 1998)

CEO, ScienceRight Editing & Publishing

http://scienceright.com/

Contents

Basics in Medical Statistics

1. Introduction

Would you like to feel confident choosing the right statistical test or tests to analyze your data? If so, this primer was written specifically for you, especially if you currently have minimal to no background in medical biostatistics.

But it also was written for you to provide others with this knowledge: perhaps medical students or residents on your service, or a Masters or PhD student working with you, who otherwise will have little to no formal teaching in statistics.

It is primarily intended for those in the clinical sciences, though all the principles that will be presented are applicable to those who do wet bench work as well. Why can't the subjects analyzed be rats or guinea pigs? Or even cells? Statistical analysis is not species specific.

Because my background and experience are largely in clinical studies — primarily patient and general population surveys, and clinical trials — the examples I provide here will all be clinically based.

I have tried to write this as clearly and simply as possible and, in doing so, will not discuss all the potential statistical tests that are available. This primer discusses the ones that are most commonly used, tests that together encompass virtually all the clinical-research, data-analysis scenarios you are likely to find yourself in. They are:

Again, the goal of this primer is to teach you when to select and how to use these tests, not how to perform them. I also will include a few essential "rules" of analysis. Frankly, the specifics of how to perform all these tests depends on the statistical software package you are using (SPSS, SAS, Minitab, etc.). I will provide you with a link to inform you about the various statistical packages that are out there and how user-friendly they are, in case you are (as they say) in the market for one.

Further primers will follow on specifics, like how to use your new-found knowledge of statistical analysis to markedly improve your studies and chances of receiving grant funding to support them; how to

perform meta-analyses, including how to draw Fox Plots in Excel; and how to convert a single case report or small case series of just a few patients into multiple scientific publications.

2. Why do statistical analysis?

There are many different reasons to do statistical analyses, but I will only address the two that are, by far, the most common reasons in clinical research:

(1) Inter-group comparisons

The first, extremely common purpose of statistical tests in clinical trials or surveys is to test whether two or more groups are statistically different, in terms of a given baseline value/characteristic or post-treatment outcome. These groups can be patients receiving different treatments (e.g., an active drug versus placebo; or surgery versus no surgery), or people with some other distinguishing characteristic (e.g. age < 40 versus 40-59 versus ≥ 60 years).

(2) Within-group comparisons

A second very common purpose of statistical testing is to compare a particular measurement performed more than once. One example is between observers, to see how closely, on average, the two measurements agree. Another example is multiple measurements over time; for example, you could test whether patients experience a statistically-significant reduction in the severity of a given symptom or disease marker over time, as in comparing pain severity after, relative to before patients have started a particular drug or undergone a specific surgical procedure.

Note that, for most (if not all) clinical trials, it would be prudent to do both inter- and within-group comparisons. For example, if you want to know if a given drug is more effective than placebo at reducing the severity of a patient's pain, you likely will want to perform **inter-group comparisons** of these two patient groups (active drug versus placebo) to see how the two groups compare with respect to their baseline characteristics, and their final outcome(s). However, you also may want to perform **within-group comparisons** to see if either group statistically improves or worsens with treatment, versus how the patients' pain and other symptoms were at baseline.

What about doing both at the same time: not only seeing if either group changes versus baseline, but also if the two groups are different in their extent of change from baseline? One way you might think to do this is just to see if one group statistically changes versus baseline, while the other does not. But this, in fact, does NOT mean the two groups are statistically different in their extent of improvement. An example of this can be seen in Figure 1 (next page). In this example, pain graphically declines in both groups, with the two lines almost parallel. However, the difference between baseline and final follow-up JUST achieves statistical significance for the active drug, and just fails to achieve borderline statistical significance (borderline statistical significance is often considered p between 0.05 and 0.10) for the placebo. Could anyone say, with confidence, that this particular active drug is better than the placebo?

Unless the two treatment groups start with exactly the same level of pain, just comparing the level of pain at final follow-up also would tell you little, since this would not reflect the change in pain over time, as shown in Figure 2 (below). Again, the two lines indicating change are almost parallel. The only reason the average level of pain was different in the two groups at final follow-up is because the two groups also were quite different at baseline (albeit, again just failing to achieve borderline statistical significance).

Figure 1

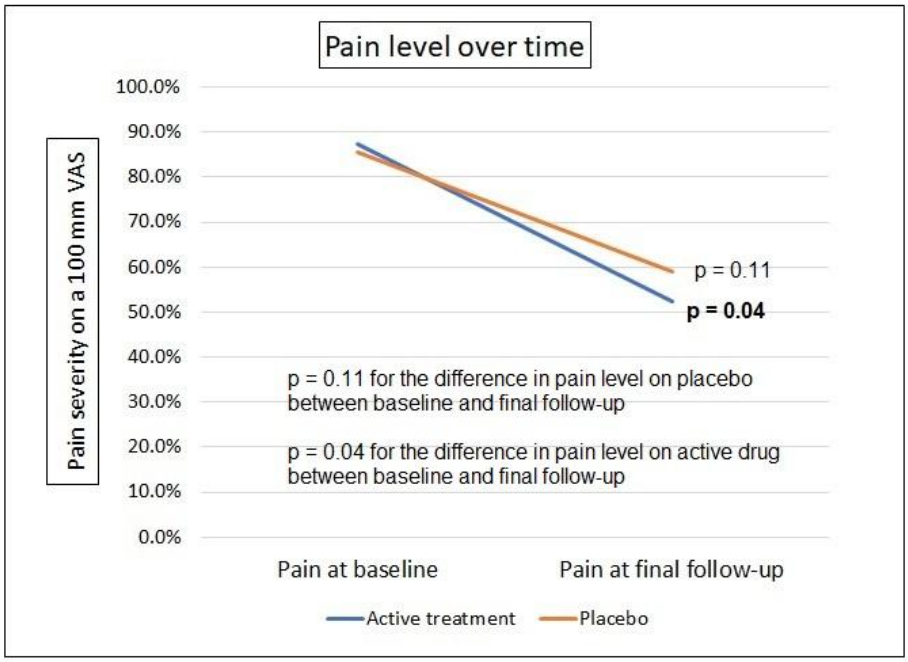
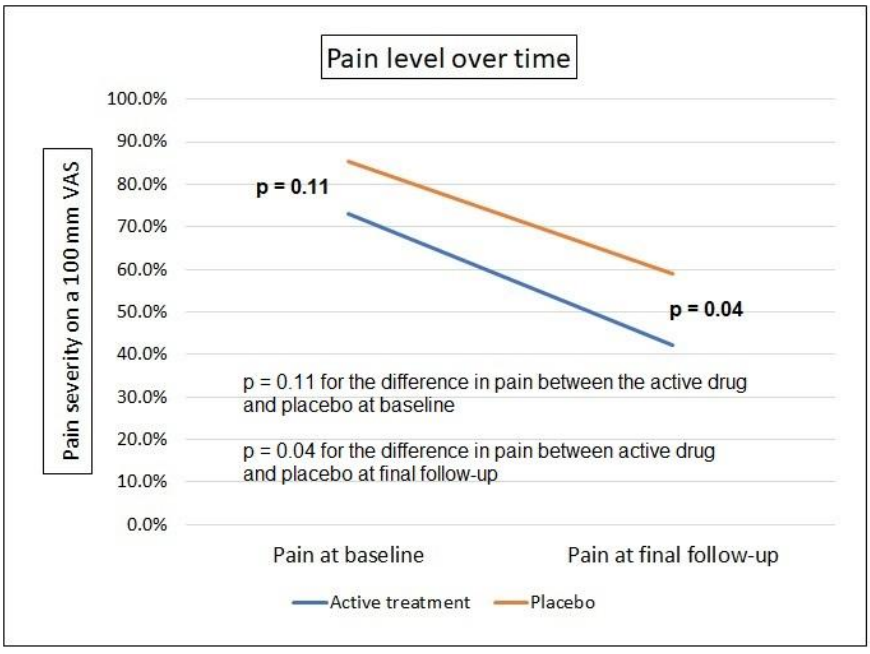


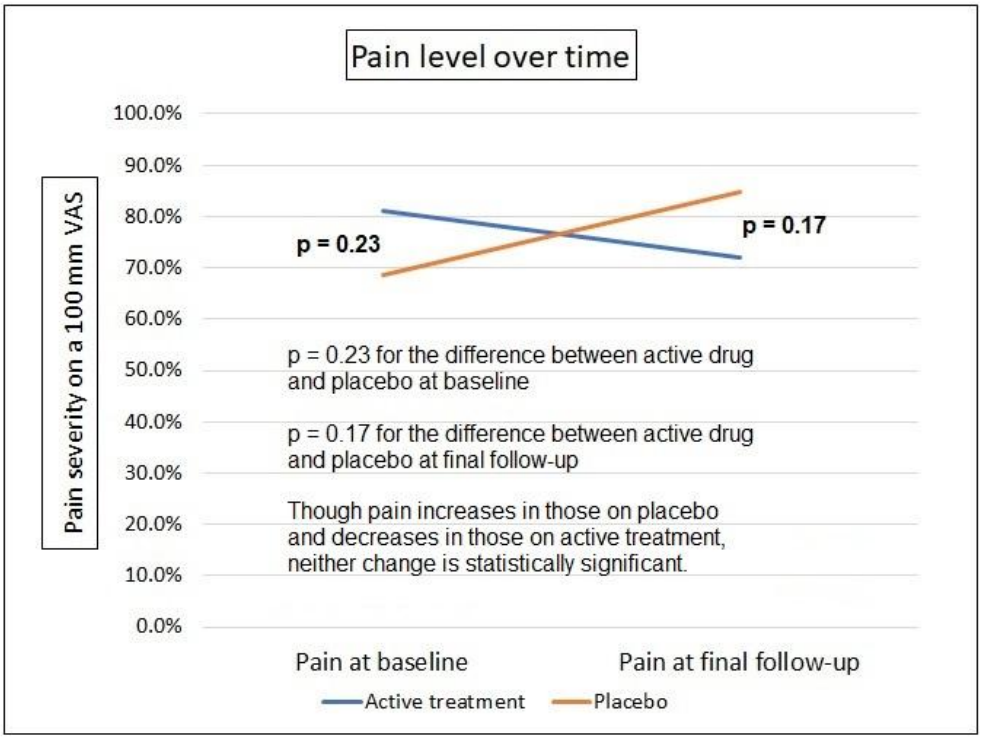
Figure 2



What you need to do is somehow combine your inter- and within-group differences into a single test. One way to do this would be by creating a new variable, called 'change_from_baseline', which you calculate as the level of pain at final follow-up MINUS the level of pain at baseline. For the data in Figure 1, you'd be comparing a change of -34.7% in the active drug group against a change of -26.5% in the placebo group. For the data in Figure 2, you'd be comparing -31.1% against -26.5%. In both instances, you'd arrive at the conclusion that — even though the change from baseline was statistically significant for the active drug, but not the placebo (Figure 1); or the final level of pain was statistically less with the active drug than with placebo at final follow-up (Figure 2) — there really was no statistically-significant difference in the effectiveness of the two treatments.

It also is possible that two treatments (e.g., active drug and placebo) could be statistically different from each other even when neither drug results in meaningful improvement and there is no significant difference in the outcome of interest, either at baseline or follow-up. How? Because patients on one or the other treatment could actually get worse, as depicted in Figure 3.

Figure 3



In Figure 3, there is slight improvement in patients on active treatment, but not enough to even approach statistical significance. There is slight worsening on placebo but, again, not enough to be statistically significant. And there is no statistically-significant inter-group difference in the outcome of interest at either baseline or final follow-up. However, there is greater than a 25% difference between the two treatment groups in the percentage of change between baseline and final follow-up (an 11% increase vs. 14% reduction in pain), which could be highly statistically significant.
In this last example, then, the only way to show that one treatment was better than the other is to look at the change from baseline to final follow-up, in this instance by creating a variable that encompassed

the percentage of change in the outcome of interest between the baseline and final follow-up data points.

You will learn an even BETTER way of doing this (detecting statistically-significant change) later in this primer (see pages 18-19). My point so far is:

Rule #1: In clinical trials or any prospective study in which subject groups are followed over time, performing both inter-group and within-group comparisons often leads to a much better understanding of the "truth" than performing either form of comparison alone.

This being said, it is crucial to understand that proving that two groups or two values over time are statistically different is only half the story. What also is important is whether any statistically-significant differences detected are also clinically-significant. For example, picture that there is a particular disease that is considered 100% fatal within one year, despite treatment. Now picture one patient receiving Drug A and remaining alive three years post diagnosis. This patient's highly-unanticipated survival is clearly of clinical significance; but it can't be tested statistically, given that there is only one such patient. If you were that patient's doctor, could you seriously deny giving your next patient with the same disease the option of that same drug? In fact, there have been instances where a very small of number of treatment successes, or catastrophic failures, have resulted in ethics review boards refusing to approve even a single, small controlled study involving the implicated treatment, claiming that it would be unethical for anyone to either deny or approve any patient's use of the active drug.

Similarly, if, as part of a huge survey, you determined that there was a statistically-significant difference of just one percent in one-year survival rate among patients given one drug versus another, could you strongly recommend that first drug as "superior" to your patients? Statistically, yes, perhaps. But clinically, they are close enough that other factors, like cost and side effects, should have more of a role in decision-making than the 1% "superiority" of the one drug.

Consequently, my second rule of statistical analysis is:

Rule #2: Statistical significance and clinical significance are BOTH important.

In other words, for a given research finding to be considered meaningful, in most instances it needs to be BOTH clinically and statistically significant. Recognize that, almost ANY trial that is large enough will find a statistically-significant difference between subject groups. However, demonstrating that one treatment has a statistically-significant, but clinically-meaningless benefit over another, itself is meaningless and without merit. That's why I sometimes cringe when I hear people say: "If the study ONLY had been a bit larger, we might have found something." Investigators who know what they are doing decide beforehand what size of difference is necessary to be clinically meaningful, and design their studies to be large enough to detect that difference.

However, because this is a primer on basic statistical analysis, the rest of this primer will only discuss statistical significance.

3. What can you compare?

When performing statistical analysis, whether you are comparing different groups or comparing measurements collected from one or more groups over time, you can compare:

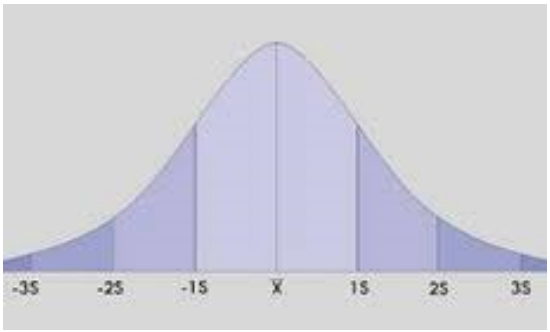
(1) Mean values

For example, is the mean systolic blood pressure in one patient group different than the mean systolic blood pressure in another group? Note that systolic blood pressure is a continuous value (in other words, there is a continuous scale of blood pressures, ranging from a low of maybe 80 to a high of perhaps 180, with every value in between possible) If you were to create a table to describe the two groups, it could look something like this:

	Group A	Group B
Mean Systolic Blood Pressure	145.7 mmHg	153.2 mmHg

Means can be compared when the potential values of a particular variable of interest are (1) continuous (e.g., 0 to 100), and (2) normally distributed (meaning that, when plotted on a graph, they roughly produce a bell-shaped curve), like this:

Figure 4



(2) The proportion with one variable option versus another

For example, what percentage of females versus males have a high systolic blood pressure, defined as SBP ≥ 140. Note that, though blood pressure is a continuous variable, here it is being treated as a binomial (either/or) variable (SBP ≥ 140; SBP < 140). Each patient's blood pressure (BP) is either 'high' or 'not high'. If you created a table to describe the two groups, it would look something like this, showing the number (and percentage) of subjects in each group with a high versus a normal SBP:

	Females, n (%)	Males, n (%)
Systolic BP normal	45 (75.0%)	32 (53.3%)
Systolic BP high	15 (25.0%)	28 (46.7%)

(3) The ranks of a particular value or measurement

Rather than comparing mean blood pressures or the proportion with high versus normal blood pressures, you may want to rank blood pressures. For example, if the systolic blood pressures in Group A are, from high to low, 183, 162, 150, 145, 138, 128, 110, and 108; and in Group B 185, 178, 152, 146, 126, 125, 114, and 106, you'd have a table that looks like this, with the relative ranks of these SBP measurements in parentheses:

Group A systolic BP and (rank)	Group B systolic BP and (rank)
183 (2)	185 (1)
162 (4)	178 (3)
150 (6)	152 (5)
145 (8)	146 (7)
138 (9)	126 (11)
128 (10)	125 (12)
110 (14)	114 (13)
108 (15)	106 (16)

Note that Group B has blood pressures that are the highest, 3rd highest, 5th highest, 7th highest, 11th through 13th highest, and 16th highest.

Rank tests are used in analysis when the measured values for the variable of interest are not normally distributed across the study groups (there is no bell-shaped curve), since a bell-shaped distribution is an essential assumption of all so-called *parametric tests*. Rank tests are *non-parametric*, which makes them somewhat more conservative (less sensitive to detecting differences) than parametric tests, thereby compensating for the relative uncertainty associated with non-normally distributed data (discussed later). For this reason, non-parametric tests should generally NOT be used on normally-distributed data, because your analysis will be prone to missing true differences between groups.

So, my third rule of statistical analysis is:

Rule #3: If the measurement values of a continuous variable of interest (e.g., blood pressure) are normally distributed (i.e., roughly form a bell-shaped curve), use parametric tests. If they are not, use non-parametric (e.g., rank) tests.

Hint: In the widely-used statistical software called SPSS, the normality of numerical data (e.g., blood pressure and pain ranking on a 0-100 scale) can be tested by clicking on ANALYSIS, then DESCRIPTIVE STATISTICS, then EXPLORE. Once you select your variable(s) of interest for the dependent variable box, click on PLOTS and make sure that the box for NORMALITY PLOTS WITH TESTS, as well as the box for HISTOGRAM are checked. Then click CONTINUE, then OK. Then check your OUTPUT file to find the p value for the Shapiro-Wilk test*. If the p value is < 0.05, then your data are non-normal, and you should select non-parametric versus parametric tests for those data (more on this later). (*Ignore the results for the Kolmogorov-Smimov test, since this has been proven inaccurate for testing the normality of data.)

4. Understanding Your Variables

To understand which statistical test to use, you must understand your variables, relative to six terms defined below, all of which refer to different types of variable. These terms are: *dependent variable; independent variable; continuous variable; categorical variable; nominal variable*; and *ordinal variable*.

With this in mind, there are essentially two questions you need to ask about each of your data variables before using them in analysis, so you understand both how to use them, and which tests to select:

(1) Do you intend to use this variable as a dependent or as an independent variable?
(2) Is the variable continuous, ordinal or nominal?

(1) Is the variable dependent OR independent?

Let's start by defining these terms:

Dependent variable: This is the main variable of interest that you are measuring. In the first three examples given in this primer (on pages 7 and 8), the dependent variable was systolic blood pressure, as you were comparing blood pressure between different groups.

Independent variable: This is the variable that defines the different subject groups you are looking at. In the first and third examples on pages 7 and 8, the independent variable was 'Group' (Group A versus Group B); in the second example, the independent variable was 'patient gender' (male versus female).

Analyses generally require both a dependent and independent variable. For example, if you wanted to see if brain cancer patients who undergo subtotal resection live longer than those who don't undergo surgery, the dependent variable would be survival time (e.g., measured in whatever units you choose: days, weeks, months, years), while the independent variable would be treatment group (subtotal versus no subtotal tumor resection).

Note also that any given variable can be used as a dependent variable in some analyses, and as an independent variable in others. In other words, whether a variable is dependent or independent is entirely dependent on which you choose it to be.

For example, where systolic blood pressure (SBP) was used as the dependent variable in the three examples listed on pages 7 and 8, you also could ask the question: is SBP a predictor of stroke? In this instance, the 'presence versus absence of stroke' would be the dependent variable (the variable of primary interest), while SBP would be independent.

(2) Is the variable continuous, ordinal or nominal?

Now, whether they are being used as dependent or independent, all variables are, by their very nature, continuous, ordinal, OR nominal.

Continuous variables: A continuous variable is one where the range of potential measured values is continuous; for example, systolic blood pressure when measured in mmHg; or patient height in inches

or centimeters, or weight in pounds or kilograms; or a patient's level of pain, when rated on a scale from 0 to 100, where zero is no pain and 100 is extreme pain.

A second distinguishing characteristic of all continuous variables is that the incremental relationship between adjacent numbers remains constant, no matter where you are on the scale. For example, on a 0-100 scale rating pain severity, the distance between a patient's pain rating of 10 and one of 11 is the same as the distance between the two ratings 90 and 91; ditto for 45 and 46, and for 79 and 80. As such, it is appropriate and standard to summarize continuous variables by reporting their means.

If a variable meets these two criteria, it essentially is continuous. Note that it doesn't need to have an infinite or even a large number of options. If you ask someone how many days of the week they floss their teeth, and give them the option of entering anything from zero to 7, THAT is a continuous scale. Once you have a variable with five or fewer options (like, how many non-weekend days do you exercise in the average week?), most agree that it's time to start treating a variable as ordinal, even if it meets the two criteria of being continuous and having a consistent relationship between all options. Why is this? I don't know; but it may have something to do with the reduced likelihood of normalized data, the fewer response options individuals have, and you lose statistical power if you analyze non-normally distributed data with tests designed for normalized data (more about this later).

The remaining two categories, nominal and ordinal, both fit under the umbrella of **categorical variables**. Contrary to continuous variables, categorical variables are those where all the data are assigned to different categories, which either can be or cannot be logically ranked.

The easier of these to understand is the nominal variable.

Nominal variables: A nominal (named) variable is one where there is no logical ranking or order of the various categories. For example: male versus female; Group A receiving active drug vs. Group B receiving a placebo; Caucasians vs. Blacks vs. Hispanics vs. Asians vs. other. In each case, the assumption is that no group is better or should be more highly ranked than another. This is true even if you assign each category a number to aid in data analysis (e.g., 0 = no diabetes; 1 = juvenile-onset diabetes; 2 = adult-onset diabetes; 3 = gestational diabetes; 4 = other).

Ordinal variables are a bit more complicated, because you may sometimes be tempted to call an ordinal variable continuous; though **it usually is crucial that you don't.**

Ordinal variables: An ordinal variable is one in which (1) each subject is assigned a value or rating that encompasses a range of individual measurements; and (2) there is a logical order to those values. For example, if you categorize systolic blood pressure into level 1: SBP < 140; level 2: SBP between 140 and 159; level 3: SBP between 160 and 179; and level 4: SBP ≥ 180, you are using an ordinal scale, with every subject receiving a SBP rating between one and four. Hence, if a given subject's SBP is 145, they would be allocated to the 'SBP between 140 and 159' category, and assigned the value of '2'; if their SBP is 182, they would be assigned a value of '4'. Note that it makes no sense to calculate a mean value for an ordinal scale like this one, since it is conceivable that, in one group of subjects, a disproportionately-high percentage might have SBPs at the lower range of each level while, in the other group, the reverse

might be true. Calculating group means would, therefore, be quite misleading and, hence, uninterpretable.

You can divide patient age this way too; for example, (1) patients under age 20; (2) patients 20-39 years old; (3) patients 40-59; (4) patients 60-79; and (5) patients > 80 years old, with every subject assigned an age rating between one and five. In fact, ANY continuous variable can be sub-categorized into an ordinal scale, IF you choose to do this.

There are some instances in which converting a continuous variable into an ordinal one could be of interest.

(1) Identifying a threshold value: The first instance is when you are trying to identify some cut-off value or threshold at which something changes. For example, statistically demonstrating that someone's risk of stroke rises as their SBP increases, is definitely worth knowing. But it doesn't tell you when a doctor should start thinking about finding ways to lower a patient's blood pressure. On the other hand, determining that the incidence of stroke doubles once one's SBP reaches 140mmHg DOES aid in the decision of when to initiate treatment. In the same study in which you show that the risk of stroke increases with increasing SBP (as a continuous variable), you also might convert your continuous SBP scale into an ordinal scale; for example, 1 = SBP < 120mmHg; 2 = SBP 120-139mmHg; 3 = SBP 140-159mmHg; and so on. If the incidence of stroke is roughly the same in groups 1 and 2, but doubles in group 3, and continues to increase in subsequent groups, this suggests that 140mmHg is a level of SBP at which initiating treatment might be advisable.

(2) When the answer given is an estimate: A second reason to convert a continuous scale into an ordinal scale is when the value you are asking about has somehow been estimated. For example, if you ask someone what their annual income is, and they say $35,000, it likely isn't EXACTLY $35,000. More likely, their salary is APPROXIMATELY $35,000. This is very different than the level of accuracy you would expect measuring someone's systolic blood pressure. For this reason, it makes much more sense to allocate different subjects into different annual income ranges (e.g., < $20,000; $20-39,000; $40-59,000... etc.) before conducting any analyses using this variable.

Can the reverse also be true? Can an ordinal scale be converted into one that is continuous? I don't mean reconverting an ordinal scale back into a continuous scale when you already possess the continuous data: like converting SBP into categories, but then re-using the already-obtained continuous data for certain further analyses. I mean actually converting an ordinal scale into a continuous one.

I've already mentioned how, when you only have a few response options (e.g., how many non-weekend days in the average week do you go for a walk outside?) where you might consider a continuous variable ordinal? But is the reverse be true? When you have enough response options, should you consider an ordinal variable continuous?

For example, in a very large population survey, you could assign values from 1 to 12 to individuals based upon their level of education, with 1 = less than elementary school; 2 = finished elementary school; 3 = some high school; 4 = graduated from high school; 5 = some time in junior college; 6 = graduated from junior college; 7 = some college/university; 8 = an undergraduate degree; 9 = some Masters degree work… up to 12, a doctoral degree.

Given that you have potential ratings continuously from 1 through 12, could you safely consider this a continuous scale? The short answer is 'No'. And here's why.

Remember that one of the essential characteristics of any continuous variable is that the relationship between successive values always remains the same. What this means is that, for any scale that starts at either 0 or 1, the mean value of category '2' MUST be exactly twice the mean value of category '1'; the mean value of category '3' MUST be exactly three times the mean value of category '1'; the mean value of category '4' must be exactly 4 times the mean value of category '1' and exactly twice the mean value of category '2'; and so on. As explained below, this rarely happens.

In the previous example, can you confidently call completing elementary school (rating = 1) exactly twice the value of only finishing some elementary school (rating = 2)? Can you call having an undergraduate degree (rating 8) exactly twice the value of completing high school (rating 4)? The answer to both these questions is no. Clearly, having a high-school diploma means having more years of formal education than just having completed elementary school, but of how much greater value one is relative to the other can only be speculated, and varies from individual to individual.

And how could you ever interpret a mean value of, for example, 5.9? Does that mean that the average person in your sample has a mean level of education just shy of a junior college degree? Clearly, such an interpretation would be inappropriate.

This problem converting an ordinal variable into a continuous one is even more obvious if the available responses are something like: (1) more than once daily, (2) ≤ once per day, (3) ≤ once per week, (4) ≤ once per month, or (5) ≤ once per year. In this second example, not only is each response an estimate within a range, but the size of each category ranges vastly – from one day for response options (1) more than once daily, and (2) ≤ once per day, to 365+ days for response option (5) ≤ once per year.

Consider that one person who reports having fainting spells within the '≤ once per month' category might only average two spells per year. Meanwhile, another person selecting the same categorical frequency of spells might have spells an average of 10 times per year. Both frequencies are too many to be considered ≤ once per year and two few to be considered ≤ once per week. Hence, there is only one category into which both subjects will fit (≤ once per month). However, one of them actually has about five times as many spells as the other. And again, how could you possibly interpret a mean spell frequency value of, for example, 3.9? Almost one a month?

Hence, when you have an ordinal scale that merely estimates the frequency of a given symptom or event, like this one does (into broad categories), it makes no sense to consider it continuous, no matter how many potential ratings there are.

This problem of inconsistent relationships between adjacent values even occurs when the ordinal categories relate to numerical values. For example, what if patients, rather than rating their pain from 0 to 100, are asked to rate their pain as (1) between 0 and 20 out of 100; (2) between 21 and 40; (3) between 41 and 60; (4) between 61-80 category 4; or (5) between 81 and 100 out of 100. As with asking them to rate their pain from 0 to 100 on a 100mm line, this ordinal scale clearly contains the full spectrum of potential values, from 0 to 100. However, note again that each rating is an estimate. Also, note that the mean values of these five categories are 10.0, 30.5, 50.5, 70.5, and 90.5, respectively. Clearly, 30.5 is NOT exactly twice 10.0; it is 3.05 times 10.0. And 70.5 is neither four times 10.0, nor twice 30.5. So, even though the various categories span the full range of pain, from 0 to 100, the numerical relationship between the various categories changes. You cannot safely assign someone who rates their pain between 20 and 39 a mid-point value of 30 and compare it to the midpoint value of 10 you might assign to someone who rated their pain between 0 and 19.

As a final example, lets compare the following two questions and response options:

(A) Considering that you spend at least one day surfing the Internet per week, in the average week, how many days (1 to 7) do you spend at least SOME time on the Internet?
(B) How much would you agree with the following statement: “I spend too much time surfing the Internet”? (1) Strongly disagree — (2) Disagree — (3) Partially disagree — (4) Neither agree nor disagree — (5) Partially agree — (6) Agree — (7) Strongly Agree.

Note that both questions have seven response options. However, with Question A, answering four days per week indicates exactly four times the frequency of answering once weekly. On the other hand, answering ‘(4) Neither agree nor disagree’ is clearly NOT four times the value of (1) Strongly disagree.

As a final way to clarify this issue, Table 1, below, gives several examples of each type of variable:

Table 1: Examples of nominal, ordinal and continuous variables

Type of variable	Examples
Nominal	Gender; Race; Country of origin Treatment group (e.g., active treatment vs. placebo) Employment status: full-time, part-time; student; unemployed; retired; other Diabetes (yes/no); Survived (yes/no) Outcome at 30 days follow-up: died; remained hospitalized; discharged to home
Ordinal	Level of satisfaction with treatment: 1 = very dissatisfied; 7 = very satisfied Years in the workforce: zero-5; 6-10; 11-15; 16-20; over 20 Dose of medication: zero (placebo); 5mg/day; 10mg/day; 20mg/day Number of children: 1, 2, 3, 4, 5 or more How often do you read a book of fiction? Never; Occasionally; Often; Daily
Continuous	Systolic blood pressure on a continuous scale Age in years; Months pregnant; Gestational age (in weeks) Average number of days per month you leave town Score (20-80) for each section of the State-Trait Anxiety Scale Average lymphocyte count per microscopic field

I hope, with all this, I have clarified this issue. The reason I have spent considerable time trying to clarify this issue is that it is crucial for you to realize that tests designed for continuous variables should not, under most circumstances, be used for ordinal scales, even when the ordinal scale involves a large number of seemingly-continuous numbers.

Rule #4: Continuous variables CAN be converted to ordinal scales, and sometimes this is useful. Rarely, if EVER, can you safely convert an ordinal scale into one that is continuous.

The fifth rule is the reason for this entire past section, and is a segue for everything that follows:

Rule #5: To determine which test you are going to use, you must consider whether your dependent variable is continuous, nominal or ordinal; and then do the same with your independent variable.

The following table then can be used to select which test you need:

Table 2: Choosing the right statistical test for your data

	----------Dependent Variable----------		
Independent Variable	Nominal	Ordinal	Continuous
Nominal	Pearson chi-square (χ2) analysis Cohen's Kappa (κ)	Pearson chi-square (χ2) analysis Non-parametric rank test (If 2 groups: Wilcoxen rank sum test, or Mann-Whitney U test) (If > 2 groups, Kruskal Wallis H test)	Student's t-test (if 2 groups) ANOVA (if > 2 groups) Non-parametric rank test
Ordinal	Pearson chi-square (χ2) analysis	Pearson chi-square (χ2) analysis Non-parametric rank test (Kruskal Wallis H test)	Student's t-test (if 2 groups) ANOVA (if > 2 groups) Non-parametric rank test
Continuous	Logistic (binary) regression analysis Multinomial regression	Ordinal regression analysis	Pearson Correlation Analysis Linear regression Analysis

5. Understanding the tests

The primary objective of virtually all the tests listed above and described below is to calculate a number that then can be used to estimate the probability that a specific hypothesis is true; for example: are the mean values of a specific variable different between two or more groups? Did the mean value of a specific measurement change over time; e.g., pre- to -post-treatment? Are two measurements correlated with each other; i.e.., how closely/accurately does the value of one predict the value of the other? The number that is calculated by each statistical test is called the **'test statistic'**, and each test has its own specific test statistic. For Student's t tests, for example, the test statistic is the 't value'. For Pearson $\chi 2$ (chi square) analysis, it is the value of $\chi 2$. For ANOVA, it is the 'F statistic'; and so on. Below is a short table (Table 3) listing the most common tests and the test statistic that each test generates.

Table 3: Statistical tests and their test statistics

Statistical test	Test statistic	Symbol	Possible range
Student's t test	t value	t	Any + or – value
Pearson $\chi 2$ analysis	Chi square ($\chi 2$)	$\chi 2$	Any positive value
Analysis of variance (ANOVA)	F value	F	Any positive value
Pearson correlation analysis	Pearson correlation coefficient	r	-1.00 to +1.00
Regression analysis	Regression coefficient	β (or R)	Any + or - value

Note that all the test statistics are calculated, not only based upon the measured values, but also the number of subjects, the degree of variance between measurements, values that would be expected if the various groups were identical, and so on. For some tests, calculating these test statistics by hand is possible, though tedious (e.g., Student's t tests, Pearson $\chi 2$ analysis). However, it is virtually impossible for others (e.g., regression analysis). All these test statistics are calculated automatically with the click of the mouse within any of a long list of statistical software programs, which include the three programs I learned during my PhD studies — SPSS, SAS and Minitab. Of the ones I learned, I find SPSS the most user-friendly; but I must admit to not having used SAS or Minitab in over a decade, so they may be much more user-friendly now. I can't comment.

If you are considering purchasing a statistical software program for you or your research team, a list (with 5-star rankings) of all the various major statistical software programs currently available can be found at https://www.capterra.com/statistical-analysis-software/. However, besides checking out that page, I would STRONGLY urge you to find out what your colleagues and/or (potential) collaborators are using, as having everyone using the same program can be very helpful, and sometimes is necessary. Data files can sometimes be converted from one program to another, but not always directly or easily.

Whatever test statistic you do calculate is then used to determine a p value, utilizing a table of p values for each given value of the test statistic; in statistics software programs, this is done automatically. What follows are brief descriptions of the tests that are, by far, the most widely used.

Pearson χ2 test

The Pearson χ2 test (also called Pearson chi-square analysis) is used whenever you construct a 2 by 2 (2 x 2) table like the one below, where both the dependent and independent variable are either nominal or ordinal. In a nutshell, you are testing whether the proportions (or percentages) of subjects in the various 'cells' are the same in the two groups.

	Females (number, n =)	Males (n =)
Systolic BP normal	45	32
Systolic BP high	15	28

The Pearson χ2 test also can be used if you have a 2 x 3 or 2 x 4, or even a 3 x 3 or 6 x 8 table, if both the dependent and independent variable are either nominal or ordinal. For example:

----------Gender----------

	Females (number, n =)	Males (n =)
Systolic BP < 120 mmHg	33	18
Systolic BP 120 – 139 mmHg	12	14
Systolic BP ≥ 140 mmHg	15	28

----------Age----------

Systolic blood pressure (SBP)	30 — 39 years	40 — 49 years	50 — 59 years	60 — 69 years	≥ 70 years
SBP < 120	121	113	93	68	59
SBP 120 - 139	72	81	95	86	121
SBP ≥ 140	36	52	71	88	63

Since χ2 analysis deals only with categorical (nominal or ordinal) data, there is no need to estimate means and, therefore, no issue about normal distribution; hence, no issue about whether to use parametric or non-parametric analysis. With Pearson χ2 analysis, the issue of data distribution does not apply.

Cohen's Kappa

A somewhat special statistic, in that it has a very specific application, is Cohen's Kappa. Cohen's Kappa is very similar to Pearson χ2 analysis, in that it is a test for which both the dependent and independent variable are categorical. Where it differs is that, whereas with Pearson χ2 analysis, either the dependent or independent variable, or both, can be ordinal, to generate a Kappa statistic, both must be nominal. Another difference is that, unlike Pearson's χ2 analysis, you can only test a 2 x 2 table of results, and not the 2 x 3 or 5 x 3 tables of results given as examples above.

What Cohen's Kappa analysis specifically tests is whether two individuals (or, less commonly, groups) AGREE on something. For example, do two different doctors all come up with the same diagnosis when

examining a series of patients? Do two different labs come up with the same result (an abnormal result or a normal result) when testing a series of samples?

As for Pearson χ2 analysis, the results must be applicable to some sort of n x n table; specifically, a 2 x 2 table, as explained above. For example, if you were testing whether two different radiologists agree that a given X-ray image shows a normal or abnormal result in a series of patients, you would generate a table like the one below:

		Radiologist A	
		Normal	Abnormal
Radiologist B	Normal	25	5
	Abnormal	10	40

The Kappa statistic (depicted as κ) essentially is calculated to reflect the combined proportions of X-rays read by both radiologists as normal, and read by both radiologists as abnormal, and will end up being some number between 0 (no agreement) and 1.00 (full agreement). A kappa value κ = 0.63 essentially means that there was 63% agreement, overall, between the two radiologists.

Note that you COULD use Pearson's χ2 analysis on the exact same data to generate a table like the one immediately below. But you'd insert different numbers (not the 25, 5, 10 and 40 shown above). For example, from the above numbers, we know that Radiologist A rated a total of 45 X-rays as abnormal (40 in agreement with Radiologist B, and 5 in disagreement with Radiologist B), and 35 as normal (25 in agreement with Radiologist B and 10 in disagreement). Meanwhile, using the same logic, Radiologist B rated 50 X-rays as abnormal and 30 as normal.

	Number of patients (n =) with a normal X-ray	Number of patients (=) with an abnormal X-ray
Radiologist A	35	45
Radiologist B	30	50

From these numbers, you COULD calculate a Pearson χ2 statistic and see if the two Radiologists are statistically different, but this doesn't reflect the percentage of agreement, which renders it quite different than Cohen's Kappa. Could BOTH tests be worth doing in a given study? Most certainly, depending on what you want to find out.

What do you do if you have more than two radiologists, or physicians making a diagnosis, or labs providing a test result? For such cases, there is something called Fleiss' Kappa, which works much the same way.

Student's t tests

The Student's t test is a test that compares two groups in terms of some normally-distributed continuous variable, like systolic blood pressure measured in mmHg or weight measured in kilograms. Going back to the first example on page 7, it would be a very appropriate test to use a Student's t test. Using this test for the example below, you are trying to determine if the mean value of 145.7 is statistically lower than the mean value of 153.2 mmHg.

	Group A	Group B
Mean Systolic Blood Pressure	145.7 mmHg	153.2 mmHg

Note that Student's t tests can be either paired or unpaired. **Unpaired tests** are used when you are comparing two groups, which is called an **inter-group comparison** (e.g., patients on active drug vs. patients on placebo). **Paired tests** are used when you are comparing the same variable at two different times in one group, which is called a **within-group comparison** (e.g., before vs. after treatment). A more detailed description of inter-group versus within-group analyses is provided back on page 3 of this primer.

Unlike $\chi 2$ analysis — since t tests, by their very nature, deal with continuous data — you must check whether your data are normally or non-normally distributed prior to selecting this test. If the latter, you need to consider a non-parametric test, like the Wilcoxen rank sum test or the Matt-Whitney U test.

Analysis of Variance (ANOVA)

Analysis of Variance is very much like a t-test, except that you are comparing MORE than 2 groups. For example:

	Group A	Group B	Group C
Mean Systolic Blood Pressure	145.7 mmHg	153.2 mmHg	167.4 mmHg

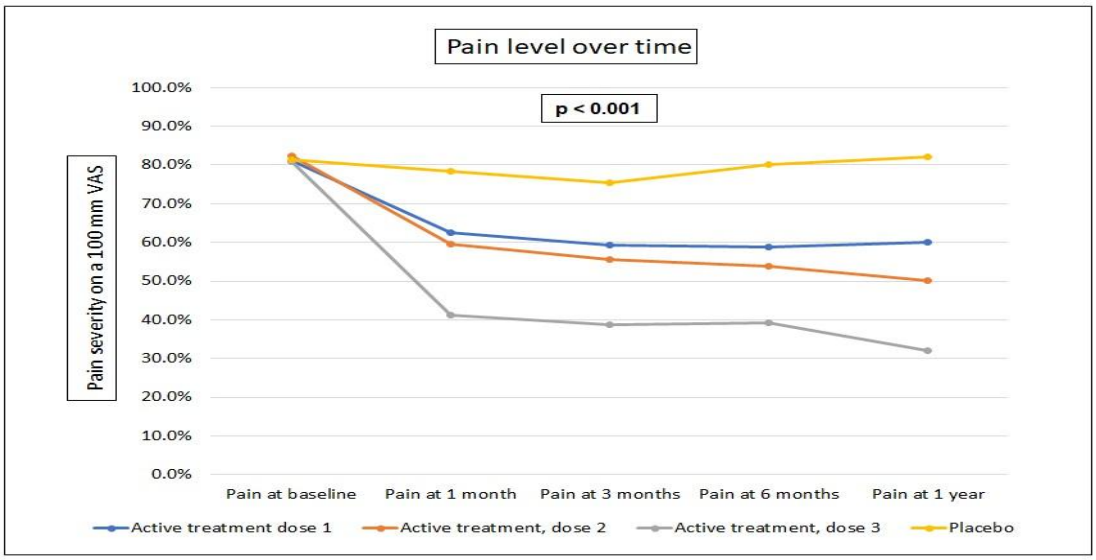
There are variants of ANOVA, like analysis of co-variance (ANCOVA), which compare a continuous variable mean between > 2 groups, all the while also looking at some co-variable, like gender. In other words, if there is a difference between the three groups, can it be explained on the basis of differences in one gender versus the other? For example, consider mean SBP across the three groups to be almost identical in women (e.g., 149.2, 150.0, 148.7 mmHg), but quite different in men (e.g., 143.9, 156.2, 171.8 mmHg).

Note that, as with Student's t-tests, ANOVAs can perform both paired and unpaired comparisons. For unpaired comparisons (e.g., comparing three or more different groups), one-way or two-way ANOVA works (depending on whether or not you are looking at just one or two categorical independent variables).

For paired analyses (e.g., looking at the same variable several times over time), use repeated-measures ANOVA, which is the "better way" to compare the degrees of improvement in different treatment

groups that I promised to provide you, way back on page 6. What repeated-measures ANOVA allows you to do is to compare outcome measures at multiple different times between multiple different groups, as in Figure 5, below.

Figure 5



In the above example, repeated-measures ANOVA would simultaneously test for the effects of time and subject group on the dependent variable (e.g., pain severity). Given the results of the figure above, almost certainly both effects would be highly significant (e.g., $p < 0.001$), with pain clearly decreasing over time overall (especially in the three active treatment groups), but also differing between the four subject groups.

Irrespective of the type of ANOVA being performed (e.g., repeated-measures versus one-way or two-way ANOVA), if significant effects are identified, the next step is to perform what is called a ***post-hoc* test** to identify which groups are different from which. At one-month follow-up in Figure 5, for example, the average level of pain among those on the 3rd (highest) dose of the active drug would certainly be lower than that among those in placebo. Perhaps the same would be true comparing doses 1 and 2 versus placebo. But it is unlikely that doses 1 and 2 differ from each other. And the pain reported at those doses might or might not be higher than for dose 3. *Post-hoc* testing would clarify all of this. Post-hoc testing also could be done to identify differences between baseline and the various follow-up appointments.

The most-commonly used *post-hoc* test is **Tukey's test**, which can be performed on SPSS merely by clicking the box marked 'Tukey's test' when setting up your ANOVA. It then will be done automatically. There are numerous other post-hoc tests; but remember Tukey's test and that should be enough for almost any situation.

As with Student's t tests, ANOVA (a parametric test) requires the assumption of normally-distributed data. Hence, if your data are not normally distributed, you cannot use ANOVA.

Fortunately, in addition to there being a non-parametric test that you can use instead of a Student's t-test when you are comparing two groups (see page 15), there also is a non-parametric test that can be used when you have more than two subject groups. This non-parametric test for use comparing more than two groups is called the **Kruskal-Wallis test**. It is similar to the Wilcoxen test (a non-parametric test for comparing two groups, described above on page 15), but is specifically designed to handle MORE than two groups. Table 4, below, summarizes all this.

Table 4: Comparing continuous variables between two vs. more than two groups

If you are comparing two groups:

- If data are normally-distributed → Student's t test
- If data are NOT normally-distributed → Wilcoxen rank sum test or Matt-Whitney U test

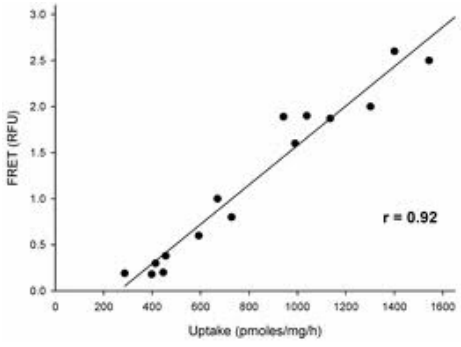
If you are comparing three or more groups:

- If data are normally-distributed → Analysis of variance (ANOVA)
- If data are NOT normally-distributed → Kruskal-Wallis test

Correlation analysis

Correlation analysis looks to see if two continuous variables are correlated; in other words, does one of the variables change in some consistent way relative to changes in the other. A simple way to picture this is to ask two questions: (1) If I plot the two variables for each and every subject on a graph, how well does a line connect the values? And (2) Does this line have either a positive or negative (i.e., a non-zero) slope?

Figure 6: A correlation plot

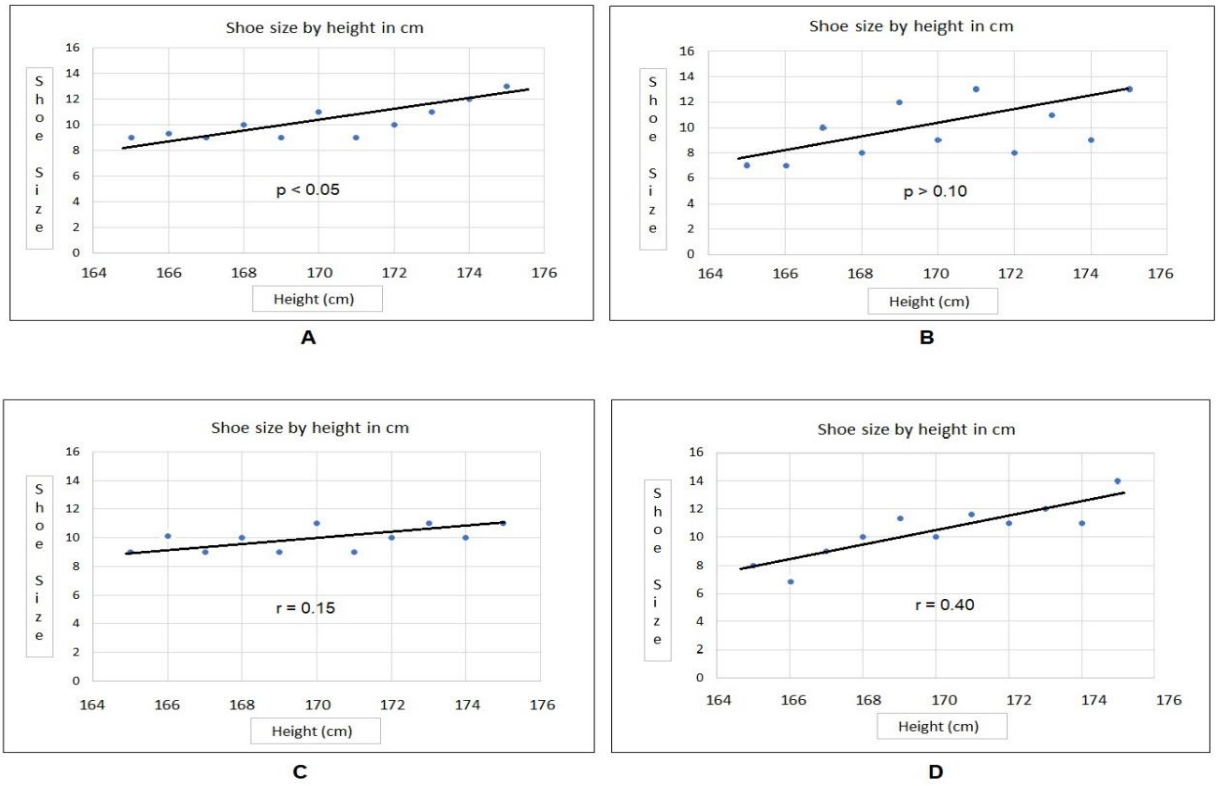


If the slope is positive (as one goes up, the other goes up, as in Figure 6, above), you say that the variables are positively or directly correlated; if the slope is negative (as one goes up, the other goes down), they are negatively or inversely correlated. The two variables in the graph above are highly, directly correlated, with r = 0.92 (maximum possible r = 1.00).

Pearson correlation analysis is the most common correlation test used. It gives you two values: a correlation coefficient (r), and a p value. It is crucial to remember that these two values tell you very different things. The p value tells you if the two variables of interest are **statistically correlated** (if p < 0.05, they are); it says NOTHING about how strong the degree of correlation is. **The strength of correlation is what the r value tells you**. R values range from -1.00 to +1.00. If r is more than +0.70, the two values are generally considered strongly positively (directly) correlated. If r is between 0.50 and 0.70, most will say that the two values are moderately positively correlated. If r is between 0.30 and 0.50, you can say the correlation is weak. Anything below 0.30 is considered a very weak correlation. And the same is true of negative r values. If r is less than minus 0.70 (-0.70), the two variables are strongly negatively (inversely) correlated, and so on.

Another albeit somewhat over-simplified way to think about the difference between p and r values in correlation analyses is to see how they are reflected in correlation plots (graphs); as in Figure 7, below:

Figure 7: Examining p and r values in correlation plots



The first two plots (A and B) demonstrate how p values are reflected when continuous dependent and independent data are graphed. Note in A, for which $p < 0.05$, how all the data points hug (i.e., are very close to) the slope line. Conversely, in plot B, for which p is > 0.10 (indicating no significant correlation), many of the data points are quite distant from the slope line. In this comparison, where the two slope lines are almost identical, the value of p is largely a reflection of how close the data are to the slope line. The slope of that line, meanwhile, is an indicator of the value of r.

Meanwhile, plots C and D exhibit slope lines with very different slopes, with the positive slope considerably greater in plot D than in plot C. Hence, the value of r also is greater in D than in C.

There is one more point to make about r values. As it happens, r^2 is the proportion of variance in one variable that is explained by the other. So, if r = +0.50, this means that $(0.50)^2$, or 25% of the variance in one is explained by the other. That is why r = 0.70 is considered a threshold for strong correlation. If r = 0.70, it means that roughly half (49%) of the variance in one is explained by the other. In Figure 6 (previous page), where r = 0.92, almost 85% (84.6%) of the variance in one is explained by the other, which would make the correlation depicted there a very strong correlation. For plot 7C, above, r^2 would be $(0.15)^2$ = 2%; for plot D (above) $r^2 = (0.40)^2$ = 16%. In other words, the percentage of variance in shoe size explained by shoe size would be 2% and 16% in the two plots, respectively, the former indicative of only a very weak correlation, and the latter a weak correlation.

With respect to the issue of data normality, Pearson correlation coefficients can be calculated for either normally- or non-normally-distributed data, so the issue of parametric versus non-parametric tests doesn't apply.

Caution!!! Be careful NOT to use the word 'correlation' casually in your discussion, papers or grant applications, because anyone who knows statistics will think you've done or are planning to do a Pearson correlation analysis, which ONLY IS POSSIBLE if you have two continuous variables. In all other instances, instead of saying that two things are correlated, say that they are 'related', 'associated', or even 'linked'. For example, a history of cigarette smoking is associated with (or linked to) an increased risk of lung cancer. But since cigarette smoking and the presence of lung cancer are both binary variables (yes/no), they cannot be statistically correlated. (Question: what test WOULD you use to show an association between these two binary variables? See the answer in the box at the top of the next page.)

Regression Analysis

Regression analysis is something that many people think is very complicated; but it really isn't. What regression analysis does is examine if changes in one dependent variable (whether it is continuous, ordinal or nominal) can be predicted by any of a series of other variables. For example, can systolic blood pressure be predicted by someone's age, weight, and/or gender, and to what degree? Like correlation analysis, R values are generated. However, because they are distinct from the r values generated for correlation analysis, some prefer to call them β (beta) values. One huge advantage of

regression analysis over all the other tests discussed in this primer, is that it allows you to examine the influences of a sizeable number of independent variables on a given dependent variable simultaneously.

If you actually wrote a regression equation down on paper, it might look like this:

$$\text{Systolic blood pressure} = (\beta_1 \times \text{Age}) + (\beta_2 \times \text{Weight}) + (\beta_3 \times \text{Gender}) + \beta_4$$

Answer to the question on page 19 (above): The most appropriate test to use with a binary dependent and binary independent variable would be Pearson $\chi 2$ analysis.

… where β_1, β_2, and β_3 are the magnitude of influence each of the variables (age, weight and gender) has on systolic blood pressure, and β_4 is the residual not explained by these variables. Regression analysis is essentially just 'solving' for β1, β2 and β3 (in the equation above, β4 is the constant), just like you used to solve for x, y and z in algebra. If you recall your algebra, you could solve for x, y and z if you had three equations with each variable in it. For example:

$$3x + 5y + 10z = 12$$
$$5x - 2y - 7z = 4$$
$$2x - 3y - 3z = 4$$

This example is easy, because you just add the three formulas together, and y and z both drop out. You end up with 10x = 20, so you know that x = 2. You then insert the value of 2 for x into any two of the equations to solve for y, and finally for z.

Regression analysis is essentially the same; except that, if you have data on 60 subjects, you have 60 equations you can use to estimate the values of β1, β2, β3 and the constant β4. This increased number of equations allows you not only to determine what the values of β1, β2, β3 and β4 are, but also to determine if any of these values are statistically different than zero. If, for example, the 95% confidence limits for β1 overlap zero (e.g., -0.21 ≤ β ≤ +0.39), then the independent variable associated with β1 (from the example above, age) has no statistically-significant effect on systolic blood pressure. On the other hand, if the confidence limits for β1 do NOT overlap zero statistically (e.g., +0.06 ≤ β ≤ +0.52), then age is assumed to have exerted some influence on SBP; and the size and direction (direct or inverse) of that influence is determined by the magnitude and sign (+ or -) of β1. And so on, for every other independent variable (for β2, β3…) in your equation.

Note that there are different types of regression analysis, based upon whether your dependent variable is nominal, ordinal or continuous. **Linear regression analysis** is used when the dependent variable (like systolic blood pressure) is a continuous variable. **Ordinal regression analysis** is used if the dependent variable is ordinal (e.g., working full-time, working part-time, not working). **Logistic regression analysis** is used if the dependent variable is both nominal and binary (e.g., disease present = 1; disease absent = 0): and **multinomial regression analysis** is used if your variable is nominal but has more than two categories (e.g., Caucasian, African American, Hispanic, other). Table 5 summarizes all this.

Table 5: Types of regression analysis

Dependent variable	Example(s)	Statistical test
Continuous	Systolic BP; 0-100 pain rating; survival time in months	Simple linear regression
Ordinal	Return to work (full-time, part-time, no)	Ordinal regression
Multinomial	City of residence (Toronto, New York, London)	Multinomial regression
Binary (dichotomous)	Stroke/no stroke; death/survival	Binary logistic regression

Let me make one further crucial point about regression analysis, which essentially applies to any and all forms of regression. The point is this: you generally require 10 to 15 subjects per independent variable tested to have enough statistical power to find anything statistically significant.

Let's say you want to identify potential predictors of one-year survival after a stroke. Since death versus survival is binary (e.g., only two options: patient lived, or patient died), you would need to do logistic regression, with 'survived/died' (survived = 1, died = 0) as your dependent variable. If the independent variables you wanted to test in the model included patient age, patient gender, presence/absence of co-morbid illness, family history of stroke, systolic BP, diastolic BP, other cardiovascular disease, diabetes, and four characteristics of the stroke itself (e.g., the extent of paralysis; +/- altered speech), you would have twelve independent variables of interest, which would mean you need a minimum of 120 subjects, and preferably 180 or more to have enough statistical power to be confident in your analysis.

But what do you do if you only have 80? One answer is hierarchical (or stepwise) regression analysis which, in the case of a binary logistic variable like 'survived/died', would be hierarchical binary logistic regression analysis. To do this, start by entering just a few variables into the model (e.g., patient age, patient gender, family history) and see which ones remain as significant predictors of survival, as a first step. Let's say just patient age remains in the model. For step 2, you create a second model, entering patient age and three more variables — e.g., co-morbid illness, cardiovascular disease, diabetes — and retesting the model to see which variables remain… and so on until you have entered every independent variable of interest at least once. One caution is that two variables that essentially measure the same thing (e.g., SBP with a BP cuff, SBP by intra-arterial catheter) MAY cancel each other out; so, don't enter such variables together. Let's finish this section by talking about the tests you might use if your data are not normally distributed.

Non-Parametric (rank) Tests

Don't worry too much about Rank tests, except to know that they are also called non-parametric tests, and are used whenever a continuous or ordinal dependent variable is not normally distributed. Most commonly used are the Wilcoxen rank sum test and Mann-Whitney U test, for use when you have two subject groups, and the Kruskal-Wallis H test when you have more than two subject groups.

In other words, if you wanted to do a Student's t test, but can't because you discover that your data are not normally distributed, do either the Wilcoxen rank sum test or the Mann-Whitney U test. If, on the other hand, you were planning to do ANOVA, but can't because your data are not normally distributed, then select the Kruskal-Wallis H test. This is all summarized succinctly in Table 4, on page 20.

An example of how measurements are ranked (given earlier in this primer on page 8) is repeated here. In this example, systolic blood pressure (SBP) was measured in all subjects in two groups, but the data

were non-normally distributed. To compensate for this, rather than calculating mean SBP for the two groups, all the individual readings were ranked, from highest to lowest, as shown in the table below. Note again how a rank, from highest to lowest systolic blood pressure (BP), is provided in parentheses next to each measurement, and that each rank is across the entire data set (both subject groups).

Group A systolic BP and (rank)	Group B systolic BP and (rank)
183 (2)	185 (1)
162 (4)	178 (3)
150 (6)	152 (5)
145 (8)	146 (7)
138 (9)	126 (11)
128 (10)	125 (12)
110 (14)	114 (13)
108 (15)	106 (16)

Rank tests use these rankings to identify if the two groups are different, rather than group means. And that's all I think most people need to know about non-parametric tests.

Summarizing the tests

Table 6 summarizes all the various tests that have been described in this primer:

Table 6: A summary of common statistical tests

Statistical test	Dependent variable	Independent variable	Number of groups	Data distribution	Test statistic
Students t test	Continuous	Categorical*	2	Normal	t
ANOVA	Continuous	Categorical*	More than 2	Normal	F
Pearson χ2	Categorical	Categorical*	Any number	n/a	χ2
Pearson correlation analysis	Continuous	Continuous	n/a	n/a	r, r^2 **
Regression analysis	Either***	Both****	n/a	n/a	β, R, R^2
Wilcoxen rank sum test	Continuous or ordinal	Categorical*	2	Non-normal	W1 & W2
Mann-Whitney U test	Continuous or ordinal	Categorical*	2	Non-normal	U1 & U2
Kruskal-Wallis H test	Continuous or ordinal	Categorical*	More than 2	Non-normal	χ2

* Categorical = nominal or ordinal

** r^2 = the percentage of variance in one variable predicted by the other variable.

*** If the dependent variable is binary, do logistic regression; if multinomial, multinomial regression; if ordinal, ordinal regression; if continuous, simple linear regression.

**** Nominal, ordinal and continuous variables can all be entered into the same regression model.

Note that, because it is so specific, I excluded Cohen's Kappa from the above table. To review Cohen's Kappa, see page 16 of this primer.

6. Making Sense of Your Results

The fifth and final rule I'll give you is one you almost certainly already know:

Rule #5: Generally, p = 0.05 is set as the threshold for statistical significance.

But what does this really mean?

When reporting p values, note that a p value of 0.04 means that you are 96% sure that there is a difference between groups. A p value of 0.001 means that you are 99.9% sure. Obviously then, $p < 0.05$ means that you are MORE than 95% sure that the difference you've detected is real.

It makes little sense to report that a p value is 0.00023. Anything less than 0.001 should just be reported as $p < 0.001$. If you are more than 99.9% certain of something, that is MORE than enough to report. There's no need to say you are 99.99977% sure. Besides, unless you have an enormous subject sample (e.g., a clinical or general population survey of over 1000 subjects), you almost certainly don't have the numbers to justify reporting ANYTHING beyond 2 or 3 decimal places.

You may, at some point, hear of a p value being adjusted for multiple comparisons.

Here is what that means. If you set your threshold for statistical significance as $p < 0.05$, any p value under 0.05 indicates that you are more than 95% sure that your conclusion is correct. However, this means conceding that, when $p = 0.05$, there also is a 5% chance that the conclusion that you have drawn is wrong. In fact, one out of 20 tests, on average, will be a false positive, purely by chance.

Consider now that you are doing twenty statistical tests within a given study. Statistically, just by chance, one of these twenty tests should yield an incorrect result. This could be a false positive result (you identify a difference between two groups when no true difference exists; so-called **type 1 error**), or a false negative result (you fail to identify a difference between two groups when the two groups actually are different; so-called **type 2 error**). Either way, setting your p value threshold for statistical significance as $p < 0.05$, you will theoretically come to an incorrect conclusion once every twenty tests. And that test might just be the most important one (i.e., your primary outcome).

How can you adjust for this?

The answer is simple: set your p-value threshold for statistical significance at some value less than 0.05. If, for example, you set your p-value threshold as $p < 0.01$, this means you are 99% sure that a given result is accurate, meaning that you believe there is only a 1% chance any given conclusion is wrong. If you set the p-value threshold as 0.02, you would be 98% sure, 2% unsure, and would theoretically only expect one erroneous result in 50 tests.

One formalized, and very stringent way to adjust the p for multiple comparisons is by dividing the standard p threshold, of 0.05, by the number of tests you are planning to do. For example, if you are doing 10 tests, you divide 0.05 by 10, which gives you a new threshold for statistical significance of $p <$

0.005. With this p-value threshold, you should be 99.5% certain and just 0.5% uncertain of each result. You would therefore be 10 x 0.05%, or 5% certain that ALL your conclusions are correct.

This method of adjusting your p value, by dividing 0.05 by the number of statistical comparisons you intend to do, is called a **Bonferroni adjustment for multiple comparisons**.

And that is a quick and dirty summary of basic medical statistics.

Note that I have enlarged the six tables included in the text, so your can have them readily available. Feel free to copy them for ready use.

And please feel free to contact me if I can be of ANY further assistance to you.

Kevin P. White, MD, PhD (Epidemiology & Biostatistics, 1996)
CEO, ScienceRight Editing & Publishing
http://scienceright.com/
dr.kpwhite@scienceright.com
519-200-8441

- Remember that the following pages include the six tables provided earlier in this primer, all on individual pages for quick reference, and a glossary of terms.

DR. KEVIN WHITE (MD, PHD)

Independent research consultant, science editor and writer
Best-selling author; Award-winning teacher
Winner of SEVEN international book awards

Credentials

- ✓ Bachelor of Science, Chemistry, CMC, California
- ✓ M.D., University of California, Davis, California
- ✓ Residency in Internal Medicine, Stanford University, U.S.A.
- ✓ Fellowship in Rheumatology, Western University, Ontario, Canada;
- ✓ Ph.D. in Epidemiology, Western University
- ✓ Winner of Collip Medal, as top graduating doctoral student at UWO, 1998
- ✓ Winner of SEVEN international book awards
- ✓ Winner, Hippocratic Council Teaching Award, UWO, 2002-03
- ✓ 14 years experience as a scientific editor, writer and research consultant, designing clinical and population-based studies and drafting grants, papers, books and book chapters.
- ✓ A proven record of success obtaining grants, both in clinical and basic science, publishing in all venues
- ✓ National and international recognition for research, teaching, and writing.
- ✓ Regular, long-term clients worldwide

Services provided

- ✓ Proofreading and editing of clinical and basic science manuscripts, including research papers and abstracts, books and book chapters, and Master's and Doctoral degree dissertations;
- ✓ Assistance writing papers and abstracts, books and book chapters;
- ✓ Survey development and editing;
- ✓ Setting up statistical databases (in SPSS);
- ✓ Data analysis;
- ✓ Creation and editing of Power Point presentations;
- ✓ Literature reviews and literature summaries;
- ✓ Creating reference databases (Endnote and Reference Manager);
- ✓ Supervision of research by students, residents and fellows;
- ✓ Group and individual teaching of basic statistics
- ✓ Will attend scientific rounds to provide feedback and assistance
- ✓ Other related services.

Table 1: Examples of nominal, ordinal and continuous variables

Type of variable	Examples
Nominal	Gender; Race; Country of origin Treatment group (e.g., active treatment vs. placebo) Employment status: full-time, part-time; student; unemployed; retired; other Diabetes (yes/no); Survived (yes/no) Outcome at 30 days follow-up: died; remained hospitalized; discharged to home
Ordinal	Level of satisfaction with treatment: 1 = very dissatisfied; 7 = very satisfied Years in the workforce: zero-5; 6-10; 11-15; 16-20; over 20 Dose of medication: zero (placebo); 5mg/day; 10mg/day; 20mg/day Number of children: 1, 2, 3, 4, 5 or more How often do you read a book of fiction? Never; Occasionally; Often; Daily
Continuous	Systolic blood pressure on a continuous scale Age in years; Months pregnant; Gestational age (in weeks) Average number of days per month you leave town Score (20-80) for each section of the State-Trait Anxiety Scale Average lymphocyte count per microscopic field

Table 2: Guide to choosing the right statistical test for your data

	Dependent Variable		
Independent Variable	Nominal	Ordinal	Continuous
Nominal	Pearson chi-square (χ2) analysis Cohen's Kappa (κ)	Pearson chi-square (χ2) analysis Non-parametric rank test (If 2 groups: Wilcoxen rank sum test, or Mann-Whitney U test) (If > 2 groups, Kruskal Wallis H test)	Student's t-test (if 2 groups) ANOVA (if > 2 groups) Non-parametric rank test
Ordinal	Pearson chi-square (χ2) analysis	Pearson chi-square (χ2) analysis Non-parametric rank test (Kruskal Wallis H test)	Student's t-test (if 2 groups) ANOVA (if > 2 groups) Non-parametric rank test
Continuous	Linear regression analysis Multinomial regression	Ordinal regression analysis	Pearson Correlation Analysis Linear regression Analysis

Table 3: Common statistical tests and their test statistics

Statistical test	Test statistic	Symbol	Possible range
Student's t test	t value	t	Any + or – value
Pearson χ2 analysis	Chi square (χ2)	χ2	Any positive value
Analysis of variance (ANOVA)	F value	F	Any positive value
Pearson correlation analysis	Pearson correlation coefficient	r	-1.00 to +1.00
Regression analysis	Regression coefficient	β	Any + or - value

Table 4: Comparing continuous variables between two vs. more than two groups

If you are comparing two groups:

- If data are normally-distributed → Student's t test
- If data are NOT normally-distributed → Wilcoxen rank sum test or Matt-Whitney U test

If you are comparing three or more groups:

- If data are normally-distributed → analysis of variance (ANOVA)
- If data are NOT normally-distributed → Kruskal-Wallis test

Table 5: Types of regression analysis

Dependent variable	Example(s)	Statistical test
Continuous	Systolic BP; 0-100 pain rating; survival time	Simple linear regression
Ordinal	Return to work (full-time, part-time, no)	Ordinal regression
Multinomial	City of residence (Toronto, New York, London)	Multinomial logistic regression
Binary	Stroke/no stroke; death/survival	Binary logistic regression

Table 6: A summary of common statistical tests

Statistical test	Dependent variable	Independent variable	Number of groups	Data distribution	Test statistic
Students t test	Continuous	Categorical*	2	Normal	t
ANOVA	Continuous	Categorical*	More than 2	Normal	F
Pearson χ2	Categorical	Categorical*	Any number	n/a	χ2
Pearson correlation analysis	Continuous	Continuous	n/a	n/a	r, r^2 **
Regression analysis	Either***	Both****	n/a	n/a	β, R, R^2
Wilcoxen rank sum test	Continuous or ordinal	Categorical*	2	Non-normal	W1 & W2
Mann-Whitney U test	Continuous or ordinal	Categorical*	2	Non-normal	U1 & U2
Kruskal-Wallis H test	Continuous or ordinal	Categorical*	More than 2	Non-normal	χ2

* Categorical = nominal or ordinal

** r^2 = the percentage of variance in one variable predicted by the other variable.

*** If dependent variable binary, logistic regression; multinomial, multinomial regression; ordinal, ordinal regression; continuous, simple linear regression.

**** Nominal, ordinal and continuous variables can all be entered into the same regression model.

INDEX OF TERMS