Healthcare Statistics for People Who Can't Do Math Good

Introduction

There are two major problems with statistics in healthcare. The first is perpetrated by the researchers and authors of the articles. Research is a high stakes game, in which careers are built and lost based on publications. Certain types of results are more likely to be published than others, so even if the research is solid, authors are often tempted to "cook" their statistics to make them seem more than they are. The fancier the statistics (and the more of them), the smarter the authors must be. (Lies, Damned Lies, and Statistics"—Benjamin Disraeli.) The second problem with healthcare statistics concerns the reader or consumer of the statistics. Most of us took at least one statistics course but apparently it was such a traumatic experience that we have blocked out any remembrance of it.

If you are one of those poor traumatized souls, this chapter is written with you in mind. We do need to talk some basic statistics, but you will not be doing any number crunching. You simply need to have an understanding of what the various statistics mean.

Review of Parameters and Statistics

As previously stated in Chapter x, a parameter summarizes a population of interest, and a statistic summarizes a study sample. The same name is used for both the parameter and the statistic. Each response variable has a specific parameter and statistic associated with it. In addition, each response variable also has a distribution associated with it. See Table 1.

Response Variable	**Statistic and parameter**	**Distribution**
Nominal	Proportion	Binomial, multinomial
Ordinal	Median	
Continuous	Mean and standard deviation	Normal (bell curve)

Nominal Response Variable Statistics

The statistic and parameter used to describe and summarize nominal response variables is the proportion —usually described as a percentage, but other units are also possible. For example, rare diseases are sometimes reported as the number of cases per 100,000. If there are only two possible responses for a nominal response variable, only one proportion needs to be calculated, as the proportion of the other response can be calculated by subtracting the first from one.

For example, a study measures the number of people who die in a years time, the proportion is calculated as:

$$\frac{number\ of\ people\ who\ died}{total\ number\ of\ people} * 100 = Proportion\ of\ those\ who\ died$$

This above proportion is often called mortality. To calculate the proportion of those who did not die, simply subtract the mortality from 100. This proportion is often called survival.

Ordinal Response Variable Statistics

Ordinal response variables are tricky creatures. The best correct statistic to use is the median, which is the middle response. However, ordinal response variables are more often treated as either nominal or numerical response variables when it comes to statistics. Treating them as nominal response variables is technically more correct than using them as numerical.

Numerical Response Variable Statistics

Numerical response variables require two parameters or statistics—the mean and standard deviation. Warning: Technical Stat-babble ahead! Because of the law of large numbers (large being more than 30 in this case), samples that measure numerical response variables with resemble a normal distribution (often called a bell curve when graphed). In a normal distribution the three measures of "central tendency" (mean, mode, median) all happen to be the same number, so you only need one of them—the mean. In addition to a measure of central tendency, you also need a measure of variance—the standard deviation. Standard deviation is simply a measure of how far from the mean the curve is shaped. Large standard deviations describe fat curves (with a lot of variation), and small standard deviations describe thin curves (with a small amount of variation).

Generally speaking, a mean is calculated as follows:

$$\frac{sum\ of\ all\ measurements}{number\ of\ measurements} = mean$$

Standard deviation is calculated by finding all the deviations (the difference of each measurement minus the mean). The deviations are then squared (multiplied by themselves). The squared deviations are then added together and divided by the number of measurements minus one (don't ask me why). Then take the square root of that number, and that is the standard deviation.

If that seems like a lot of work, it is, and I think that exercises like it are part of the emotional scarring that took place during statistics class. The civilized way to take a standard deviation is to enter the numbers in a spreadsheet and then enter the formula: "=stdev(cells with data)".

Additional note: One of the interesting things about normal distributions is that ~68% of all the responses will fall within one standard deviation from the mean (±1 standard deviation), ~95% of all responses within two standard deviations, and ~99.7% within three standard deviations. This seems to be all that many people remember from statistics class. The important thing to remember is that this is a characteristic of the normal distribution, not the definition of the standard deviation itself.

Many a student has wondered why normal lab values vary from hospital to hospital. The reason should be apparent in the name—normal. Normal lab values are traditionally defined as values within 2 standard deviations of the mean of measurements. Geographical and calibration issues result in minor differences in the mean and standard deviation at each lab, causing slight but annoying differences (for the student tasked with memorizing them).

Turning Numerical Response Variables Into Ordinal

A researcher may want to turn numerical response variables into categories. For example, the researcher may want to divide ages into ranges of 0 – 10, 11 – 20, 21 – 30, etc. There are several advantages to this approach. For example, the researcher can compare one group to another quite easily this way.

There are two basic approaches to converting numerical responses into categories. One is predefined

categories based on the possible values of the response variable. The age example above is one such example. The grading scale in the nursing program is another example. A certain percentage in the course is translated to an ordinal grading scale that is predefined based on the values recorded. The other approach is to allow the responses themselves to provide the categories. For example, the data could be divided into four groups, each containing a quarter of the responses. Or the data could be divided up into 10 groups, each group containing a tenth of the responses. This approach is often named according to the number of groups into which the data is divided. See table below:

Name	**Number of Groups**	**Proportion in each group**
Quartile	4	25%
Quintile	5	20%
Decile	10	10%
Percentile	100	1%

Generally speaking, numerical response variables should not be categorized for analysis (inference) purposes, but it is may be useful categorized for reporting (descriptive) purposes. Another purpose may be to use a numerical response variable as a qualitative factor in a study with two research objectives. For example, study may have two 2 objectives—one studying the effect of an intervention on blood pressure, and another studying the effect of blood pressure on stroke. The researchers may want to categorize the blood pressures into the JNC 7 classifications of normal, prehypertension, and hypertension 1 and 2.

Statistics in Research

There are two basic types of statistics reported in research: descriptive and inference statistics. Descriptive statistics simply describe the data collected, i.e., the sample. Inference statistics assist the researcher (and reader) in drawing a conclusion about the data and the population of interest. There are two types of inference statistics: estimation and hypothesis testing. Only inference statistics will be discussed in this chapter. **Basic assumption:** Inference statistics can only lead to correct conclusions to the degree that the sample is representative of the population of interest.

Estimation

Components of Estimation

Estimation is the simpler of the two inference statistics and will be discussed first. The purpose of estimation is to estimate the population parameter using the sample statistic. For example, a sample of a student population might be used to estimate the average age of the entire student population.

You are, in fact, already quite familiar with estimation, although you do not yet know it. In your studies or even your reading of popular magazines, you have come across a statement that asserts something like, "The expected weight loss is 5 ± 2 pounds." That is an estimation.

There are three components to estimation. The first is the point estimate which represents the best guess based on the data collected. The next component is call the bound on error and is usually denoted by the ± symbol. (The point estimate is the researcher's best guess, but it could be anywhere within the ± bound on error.) The final component is the confidence coefficient which has nothing to do with being

confident. The confidence coefficient represents the degree of certainty that the true parameter falls within the bound on error. Put another way, if the exact same study was done over and over again, the confidence coefficient is the percentage of time the true parameter would fall within the bound on error.

Confidence coefficient and bound on error are inversely related. The more precise (smaller) the bound on error, the less confidence the study will have. The more confidence desired, the less precise (larger) the bound on error must be. In order to increase both precision (smaller bound on error) and confidence, the research must increase the sample size. There is one additional factor that must be taken into account—the inherent variability of the response variable. If large variation is expected, a larger sample must be studied to obtain the same bound on error and confidence coefficient when compared to a response variable with a smaller expected variability. In healthcare research, a confidence coefficient of 95% is encountered most often, but 90% and 99% may also be seen regularly.

The key to these components is that confidence and bound on error should be decided upon *a priori*, .i.e., before the research takes place. The researcher should decide on an appropriate bound on error and confidence coefficient before recruiting data collection begins. The desired bound on error and confidence along with the inherent variability of the response variable will together influence the needed sample size. If the researcher does not think that it is feasible to recruit the needed sample size, then the researcher must decide whether to relax either the bound on error or confidence coefficient, or whether to change the study or not conduct the study at all.

Calculating estimation

Calculating estimations is a fairly simple affair. The researcher calculates the appropriate statistic which becomes the point estimate. The researcher then calculates standard error (SE), which is a measure of variability adjusted for sample size. The larger the sample size, the smaller the standard error. The bound on error for a 95% confidence coefficient is then calculated as 2 times the SE.

Estimation: Point estimate ± 2SE

It is important to note the distinction between standard deviation and standard error. Standard deviation is a measure of variability for continuous response variables. Standard error is the variability adjusted for sample size. (Standard error is always smaller than standard deviation.) When you see standard deviation reported in a research article, it is *descriptive*. When you see standard error reported in an article, it is usually for *inference* purposes.

The formulas for calculating standard error are listed in the Table below for illustration only. Note the effect that sample size has on standard error. This effect is more fully explored below in the the example.

Statistic (Response Variable)	**Standard Error Calculation**
Proportion (nominal)	$\sqrt{\frac{proportion*(1-proportion)}{sample}}$
Standard Deviation (continuous)	$\frac{standard\ deviation}{\sqrt{sample}}$

Estimation example

Estimation is actually quite simple for numerical response variables as long as you understand that concept

of standard error. Standard error (SE) for continuous response variables is the standard deviation divided by the square root of the sample size (see table above). For example, imagine you conducted a study of 36 students and measured their weight and obtained the following results:

Mean:	145 lbs
Standard Deviation	18 lbs

In this case, the standard error is calculated:

$$\frac{18}{\sqrt{36}} = 3$$

The estimation 95% confidence estimate of weight then becomes 145 ± 6 lbs (2 times SE). If you did the study again, this time with 81 students, the SE would be 2, so the estimation would be 145 ± 4 lbs. As you can see, increasing the sample size decreased bound on error, because the 95% confidence bound on error is defined as 2 times the standard error, and the standard error is reduced by the square root of the sample size. The table below shows the calculation of 90%, 95%, and 99% confidence levels.

Confidence coefficient	Bound on error	Example using data from above example (SE = 3)
90%	1.645 * SE	Bound on error = 1.645 * 3 = ± 4.937
95%	2 * SE	Bound on error = 2 * 3 = ± 6
99%	2.57 * SE	Bound on error = 2.57 * 3 = ± 7.1

As can be seen above, the higher the confidence, the larger the bound on error (less precise). In order to have a more precise bound on error at a given confidence coefficient, the researcher can either increase the sample size or reduce the variability of the response variable usually by excluding subjects who may have more inherent variability (inclusion/exclusion criteria).

Calculating a Confidence Interval

In our example above, we estimated the average weight of the student body as 145 ± 6 lbs with 95% confidence. The point estimate is 145lbs, and the bound on error is ±6 lbs. If we do the ±, we will obtain the confidence interval. So add 6 to 145 and subtract 6 from 145 to obtain the interval: 139 – 151 lbs. Let us assume for a moment, that we had calculated that our study has a confidence coefficient of 95%. This would mean that if we repeated our study over and over again, 95% of the time, the true average would fall between 139 and 151 pounds.

The confidence interval helps to serve as a true world test of the research in question. If the confidence interval includes the number 0 (zero), it means the research is meaningless. If the confidence interval is too big, it means the research is meaningless. How big is too big? It depends on the what is being studied. You as the reader probably already have some threshold number in your head. In the case of weight, most people seem to think that ± 2.5 or ±3 is meaningful, and that numbers above it are increasingly less meaningful. This brings us to an extremely important concept. **All of the statistics in the world still boil down to a subjective decision, "Is it meaningful or meaningless?"**

Hypothesis Testing

Hypothesis testing is, by far, more complicated than estimation in its mechanics, but its interpretation is actually rather simple. Unfortunately, you have to have a basic understanding of the mechanics.

At its core, hypthothesis testing asks the question, "Is there a difference between these two (or more) groups." Imagine for a moment that you ran a business selling weightloss counseling. You want to test a new program, so you measure a group's weight before they start the program and then again after the program ends. Your data is as follow:

Measurement	Average	Standard Deviation
Baseline	235	18
Post study	229	15

Obviously 229 is smaller than 235, but the question is whether the distributions of weights is different enough for the two measurements to be considered different. So here are the steps to find out the answer.

Step	**Task**	**Example**
1	Write the null hypothesis. It is always stated as one group equals another.	Baseline = Post study
2	Write the alternative hypothesis. It's always an inequality, but there's a choice (1 or 2 tailed; see below)	Baseline ≠ Post study
3	Decide on an alpha level and calculate power (beta) (see below)	$\alpha = 0.05$
4	Collect data, run the statistic, and get a p value.	t test
5	Interpret the results	

Throughout this process, there were several choices to be made.

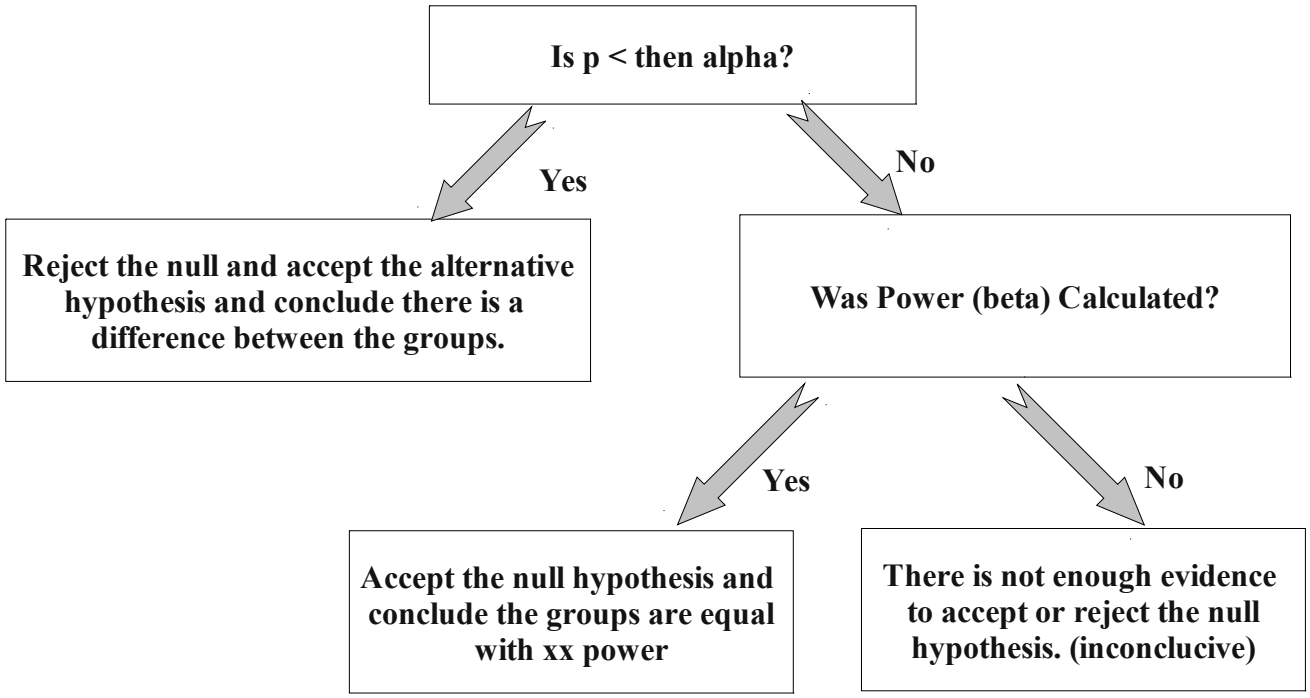
- The first choice to be made is whether our test should be a one tailed test or a two tailed test. A two tailed test tests the possibility that our post study group actually weighs more than the baseline. The decision should not be based on whether you want it to come out a certain way or think it should come out a certain way, but on what would happen if it comes out the other way. In this case, you would want to know if people actually gained weight during the study so that you could modify or eliminate the new program (unless of course the new weight is rock hard abs). Practically, speaking the only reason to choose a one tailed test is because you can use smaller sample sizes to get the same result. Studies done with one tailed tests should be viewed with suspicion. (You may be asking yourself why this matters, because anyone can see that the post study weight was smaller than the baseline weight. The reason is that the decision to use a one or two tailed test should be made before data collection occurs.)
- The second choice is the determination of alpha. It is the acceptable level of risk in being wrong in rejecting the null hypothesis. It is often called the risk of a type I error. Beta is the risk of error in accepting the null hypothesis. Power is 1 – beta expressed as a percentage. (For example, a beta of .2 would be 1 - .2 = .8 = 80%.) Power calculation is a gamble. It takes a much larger sample size to increase power than to to decrease alpha.

Most statistical tests, no matter how they are calculated will result in a p value. To interpret the test, the p value is compared to the alpha. If the p value is less than alpha, the null hypothesis is rejected and the alternative hypothesis is accepted. If the situation becomes more complicated if the p value is larger than alpha. In that case, one must consider beta (or power). If power was not considered, then the null

hypothesis can neither be accepted nor rejected. If power was considered, then the null hypothesis can be accepted with the degree of confidence of the power. This thought process is represented in graphical form below. The interpretation from our weight loss example above is presented in the table below.

Situation	Test result	Interpretation
Situation 1	p < alpha	We reject the null hypothesis, accept the alternative hypothesis and **conclude** that there is a ***significant*** difference between baseline and post study weight.
Situation 2	p > alpha; power 80%	We accept the null hypothesis and **conclude** that there is ***no*** difference between baseline and post study weights with 80% confidence.
Situation 3	P > alpha; no power calculated	There is not enough evidence to accept or reject the null hypothesis.

Graphical Representation of Interpretation of Hypothesis Testing



As can be seen, it is advantageous to the researchers for a study to *show significance*, i.e., show a statistical difference (p < alpha). Such studies do not require any worrying about power and are far more likely to be published than studies that do not show significance. Even if power is calculated, the study may be considered *underpowered*, meaning that the degree of confidence in accepting the null hypothesis is rather low. The worse possible situation for a researcher is for the study not to show significance and not have calculated power. "There is not enough evidence..." is tantamount to an admission of incompetence and of having wasted the money of whoever funded the research. Researchers in this position may choose not to publish their results, call the study a pilot or feasibility study, or use creative statistics to try and make the research look more interesting.

Hypothesis Testing Considerations

When most English speaking people use the word significant, they think of words such as important or large. When statisticians and researchers use the word significant, they mean statistical significance—that p was less than alpha. It says nothing about whether the difference was large or small. Just as with estimation, hypothesis testing also boils down to a subjective decision, "What is a meaningful difference?" Imagine for a moment that you were going to try a new diet or weightloss plan. How much weight would you need to lose before you attributed the weightloss to the new diet and not simply fluctuations in water balance and the precision of the scale?

Researchers must go through a similar process when planning a study for hypothesis testing. If not enough subjects are recruited, the researcher may find himself in the embarrassing "not enough evidence" situation. If the sample size is too large, the researcher may find that the results are "significant" statistically speaking, but are meaningless practically speaking. The Table below shows an example of the effect of sample size on hypothesis testing. The first blood pressure study measures 500 subjects in each group and finds a highly significant statistical difference between the two groups even though the actual means are only 0.1 mm Hg apart. The researcher must then explain why even though the p value was "significant" the actual difference was negligible. Said another way, the research must explain the difference between statistical significance and clinical (practical) significance. Additionally, the researcher will have wasted time and money that could have been used in other ways. Conversely, the second blood pressure study shows no statistical significance despite blood pressure means that are 25.2 mm Hg apart. In this case, although the results are clinically meaningful, because there is no difference statistically, no conclusions can be drawn. Both kinds of studies may be used to dupe unsophisticated readers, especially when reported in the mainstream media.

Effect of Sample Size on Statistical Significance

Sample size (per group)	Drug A Mean Systolic Pressure	Drug B mean Systolic Pressure	Difference in pressure	p value
500	140.2	140.3	0.1 mm Hg	.001
10	124.6	149.8	25.2 mm Hg	0.31

Determining the proper sample size for hypothesis testing is an important step in planning any research. The determination with depend upon the alpha and power (beta) levels desired, one or two tailed test, the inherent variability of the response variable, and the minimum clinical (practical) difference. In the end, interpreting hypothesis testing depends on a subjective determination of significance.

Choosing an Appropriate Statistical Test

This is an extremely complex topic and a detailed discussion is far beyond the scope of this text. The purpose of this section is simply to impart an appreciation for some of the considerations that go into choosing a statistical test. The table below lists some of the more common statistical tests

Response variable	Number of Groups to be compared	Statistical test
Nominal	2 or multiple	Chi-squared (X^2)
Continuous	2	t-test ($n < 30$)
Continuous	2	z-test ($n > 30$)

Continuous	3 or more	ANOVA

When choosing the appropriate statistical test, researchers often have multiple choices. Some of the considerations involved include:

- Response variable
- Assumptions of the data
- Robustness of the test (will it give good results even if the assumptions are violated)
- Sample size
- Number of factors
- Number of levels within a factor

Other Types of Inference Statistics

In addition to hypothesis testing and estimation, you may see other types of statistics used in healthcare research. These include risk estimation (just an applied form of estimation), correlations, and regression analysis. Each of these will be examined in this section.

Risk Calculation

In healthcare research, risk refers to the probability that an individual may experience some outcome (often negative, although statistically, risk can be calculated for positive outcomes as well). There are two kinds of risks that can be calculated—absolute and relative risk—and their meanings and interpretations are very different. Both kinds of risk are essentially applied estimations. The actual risk is the point estimate, and should be accompanied by either a bound on error or a confidence interval.

By definition, risk calculations must use categorical (nominal and ordinal) response variables. Numerical response variables must be converted to ordinal.

An important characteristic of risk calculations is that they should also be reported with a time frame and population of interest. For example, a study that observes mortality over a one year period in elderly adults who smoke can only be used to calculate risk for elderly patients who smoke over one year—not young patients who smoke, and not risk over five years.

Absolute Risk

Absolute risk is the probability that any one member of the population will experience the studied outcome within a certain time frame. Its calculation is straightforward. The risk is the number of experimental units affected divided by the total number of experimental units (the sample size). The resulting decimal may be reported as such or multiplied by 100 and reported in percent or multiplied by larger numbers such as 100,000 when dealing with very small risks.

For example, researcher may want to study the effect of smoking on the new diagnosis of lung cancer over a five year period. The researcher obtains the following data (warning: made up statistics ahead).

Calculating Absolute Risk

	Smokers	Non-Smokers	Difference in Risk
Total Number studied	1000	1000	
Number of new diagnoses of lung cancer	100	10	
Absolute risk calculation	$\frac{100}{1000} * 100$	$\frac{10}{1000} * 100$	
Absolute risk	10%	1%	10% – 1% = 9%

To reiterate, absolute risk is the probability that any one member of a population will experience the studied outcome. It compares the individual to the individual's population of interest.

Relative Risk

Relative risk compares the probability of an individual experiencing the studied outcome with the probability of another group. For example, rather than simply calculating the risk of a smoker being diagnosed with lung cancer, relative risk calculates the risk of a smoker being diagnosed with lung cancer *compared* to the risk of someone who is not a smoker. Relative risk is calculated as the number of exp units experiencing the outcome in one population divided by the number who experienced the outcome in a different population. Using the same example as above, we see.

Calculating Relative Risk Ratios

	Smokers	**Non-Smokers**	**Difference in Risk**
Total Number studied	1000	1000	
Number of new diagnoses of lung cancer	100	10	
Absolute risk calculation	$\frac{100}{1000} * 100 = 10\%$	$\frac{10}{1000} * 100 = 1\%$	10% – 1% = 9%
Relative Risk			$RR = \frac{(10\%)}{(1\%)} = 10$

In the example above, the Relative Risk Ratio (RR) is calculated as 10% (absolute risk of people who smoked who were diagnosed with cancer) divided by 1% (absolute risk of people who were diagnosed with cancer who did not smoke). The RR of 10 indicates that smokers are 10 times more likely to develop lung cancer than non-smokers. Note that the RR amplifies the risk when compared to Absolute risk. If you were designing an ad for the Truth campaign (anti-smoking) which would you prefer to report? "Smokers have a 9% greater (absolute) risk of being diagnosed with lung cancer" or "smokers have 10 times the (relative) risk of non smokers?" And if you are a tobacco executive, which would you rather have reported?

This brings us to an interesting observation. In healthcare research benefits of interventions, drugs, and treatments are almost always reported as relative risk to amplify the perceived benefit. Meanwhile, adverse effects are almost always reported as absolute risk in order to minimize their impact.

Trivial Not: Hazard ratios (HR) are essentially relative risk ratios that represt risk of death (or other negative event). They are sometimes referred to as Cox Hazard Ratios (David Cox was a statitician who devised a method of calculating HR). Relative risk ratios also have a cousin, the **odds ratio**, which is calculated differently and somewhat more difficult to interpret. Generally speaking, relative risk ratios are used for prospective studies and Odds ratios (OR) are used for retrospective studies (especially Case Control). However, many authors, especially in foreign journals (and the mainstream media) seem to use the terms interchangeably.

For a more thorough (but basic discussion, you may consider the following websites:

- http://www.childrensmercy.org/stats/journal/oddsratio.asp
- http://itre.cis.upenn.edu/~myl/languagelog/archives/004767.html

Relative Risk vs. Absolute Risk

The most important thing to keep in mind is that relative risk by itself is meaningless. It has to be accompanied by absolute risk. In the example above, the five year risk of lung cancer for smokers was ten times higher than for non smokers, but does this number in and of itself tell us anything? Before you answer consider the following examples.

Imagine for a moment that the study that estimated the RR of lung cancer for smokers had the following data:

Relative Risk vs Absolute Risk: Example 1

	Smokers	**Non-Smokers**	**Difference in Risk**
Total Number studied	100,000	100,000	
Number of new diagnoses of lung cancer	100	10	RR = 100/10 = 10
Absolute risk calculation	$\frac{100}{100000} * 100 = 0.1\%$	$\frac{10}{100000} * 100 = 0.01\%$	0.1% – 0.01% = 0.09%
Absolute risk			$RR = \frac{(0.1\%)}{(0.01\%)} = 10$

In the case above, although the relative risk of cancer is 10 times greater for smokers than non-smokers, the absolute risk is so small that most smokers are likely not to care. Now imagine another scenario:

Relative Risk vs Absolute Risk: Example 1

	Smokers	**Non-Smokers**	**Difference in Risk**
Total Number studied	100	100	
Number of new diagnoses of lung cancer	100	10	RR = 100/10 = 10
Absolute risk calculation	$\frac{100}{100} * 100 = 100\%$	$\frac{10}{100} * 100 = 10\%$	100% – 10% = 90%
Absolute risk			$RR = \frac{(100\%)}{(10\%)} = 10$

In this case, the Relative risk is exactly the same as in the example above, but now the absolute risk is 100%. That would make a compelling reason for smokers to stop smoking. Interestingly, many studies only report relative risk and not absolute risk. In this author's opinion, that amounts to academic malpractice, and any such research and author should be viewed with the highest suspicion.

One of the most egregious examples of this kind of omission is the NIH Workshop Summary: Scientific Evidence on Condom Effectiveness for Sexually Transmitted Disease (STD) Prevention (2000). Forty-

nine pages of text culminates with this statement, "From the two incidence estimates, consistent condom use decreased the risk of HIV/AIDS transmission by approximately 85%. These data provide strong evidence for the effectiveness of condoms for reducing sexually transmitted HIV (p. 17)."

The relative risk of becoming infected with HIV is 85% less when using a condom consistently compared to...actually, they forgot to tell the comparison group. Is the comparison to couples never using a condom? Or couples using a condom inconstistently? More importantly, however, the Absolute risk is not reported. The government spent millions of dollars on condom research to issue a report that lacks the one piece of information that would allow intelligent citizens to make an informed decision in their personal lives.

In your research critiques, if you encounter relative risk without absolute risk, or at least the raw data to compute it on your own, show them no mercy.

Estimating Relative Risk as Opposed to Hypothesis Testing

Traditionally, if a researcher wanted to know if a study factor made a difference, hypothesis testing would have selected as the statistical method. A relatively new (more than 20 years now) and growing trend is to use relative risk and confidence intervals to instead of hypothesis testing. Some studies will use both techniques. A brief explanation follows.

The most important thing to remember is that a Relative Risk ratio of one (1) represents no increase or decrease in risk, i.e., the two groups are the same. For example, if one wanted to study the effect of drinking Dasani vs Pepsi on the incidence of GI upset in one month, one might find the following information:

	Dasani	**Pepsi**	**Difference in Risk**
Total Number studied	100	100	
Reports of GI upset	10	10	$RR = \frac{(10\%)}{(10\%)} = 1$

One can easily see that if the proportions are equal, that the the relative risk equal one. Now combine that knowledge with the confidence intervals discussed above in the estimation chapter. If the 95% Confidence Interval (CI 95) includes 1, then the relative risk is considered non-significant. If the CI 95 is greater than 1, it represents greater risk and is sometimes called a positive risk factor. If the CI 95 is less than 1 it represents lower risk and is sometimes called a negative risk factor. In addition to reporting the relative risk and 95% confidence intervals, some studies will also report p values representing a test for statistical significance.

Visual Representations of Relative Risk

Relative risk and confidence intervals are often simply listed in a table as shown below (Risk Ratio Table). However, these can be cumbersome to interpret. Visual representations allow reader to quickly interpret the results. There are two visual representations that are often used to graphically display relative risk: a specialized bar graph and the Kaplan Meier Curve.

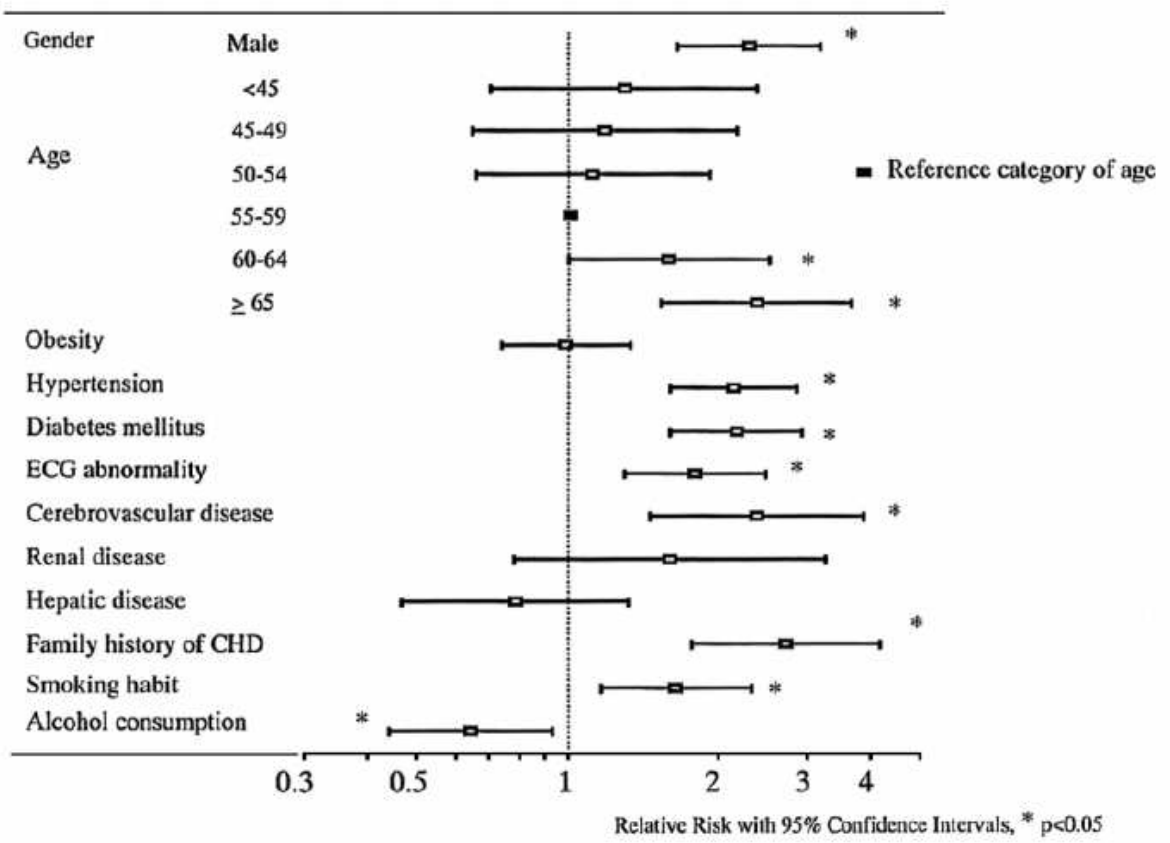
The specialized bar graph makes interpreting many risk ratios very quick and simple. The vertical line represents a risk ratio of 1 (equal risk). Any risk ratio confidence interval that touches or crosses the line is

considered non-significant. Any confidence interval that is to the left of the line represents decreased risk. Any confidence interval to the right of the line represents increased risk. In the example below, being male, above age 60, hypertension, diabetes, ECG abnormalities, family history, and smoking all increased risk of myocardial infarction. Alcohol consumption decreased risk of myocardial infarction. Being less than 54, obesity, renal disease, and hepatic disease were all non significant.

Risk Ratio Table

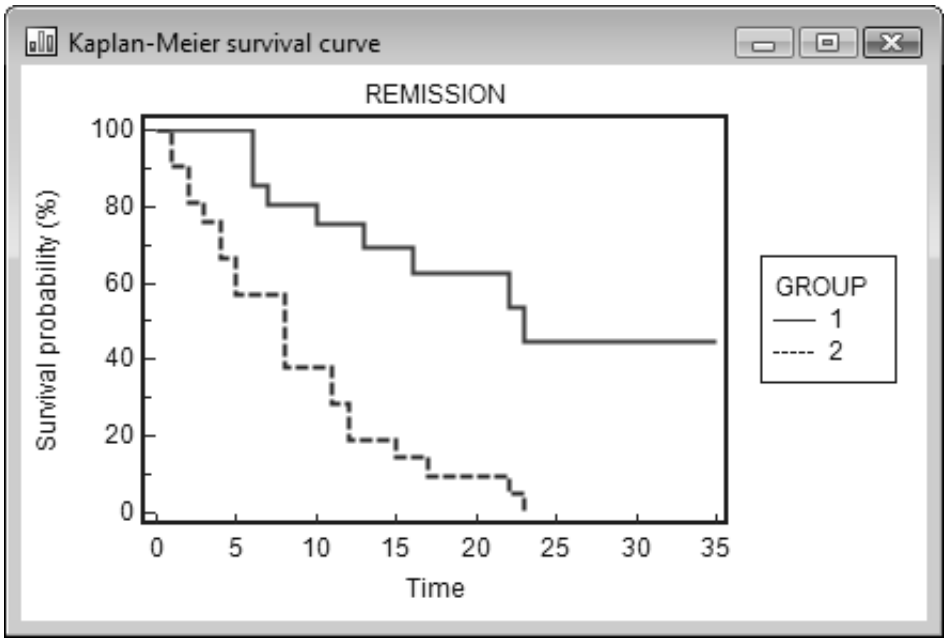
Risk factor	Adjusted* HR†	95% CI†	Infections/person-years at risk
Cumulative sex partners (continuous)	1.1	1.03, 1.1	168/1,056
Condom use with new partners			
Always	0.8	0.5, 1.2	144/938
Not always	1.0		24/118
Sex partner's no. of other partners‡			
None	1.0		79/790
≥1	5.2	1.3, 21.2	80/250
Unknown	8.0	1.8, 36.5	9/18
Time having known partner before sex (months)			
≥8	1.0		58/151
<8	1.8	1.2, 2.7	110/906
Current smoker			
No	1.0		135/931
Yes	1.5	1.0, 2.3	33/126
Currently using oral contraceptives			
No	1.0		76/553
Yes	1.4	1.01, 1.8	92/503

Confidence Interval Bar Graph



The Kaplan-Meier Survival Curve simply plots the percentage of surviving study members over time. It allows the reader to visually appreciate the difference in risk. Additionally, it gives a time appreciation for the data, showing when the differences occur. In the figure below clearly shows that group 2 begins dying much more quickly than group 2, and that 50% of group 2 is still alive after all of group 1 has died.

Kaplan-Meier Survival Curve



Correlation

Correlation is the statistical linear relationship between two variables. The correlation coefficient is represented by the lower case letter "r" and represents the slope of the line. Positive correlations mean the variables both increase together. Negative correlations mean one variable increases as the other decreases. The strongest correlations are r=1 and r=-1. The weakest correlation is 0. What constitutes a "strong correlation" varies by discipline. Biological sciences consider correlations of .9 and higher as strong, whereas psychology has typically accepted correlations of .3 and higher as strong.

An important thing to keep is mind is that the strength of a correlation can strongly be affected by outliers. (I'll need pictures to explain this one, so if you want to know more, ask me at school.)

In addition to the correlation coefficient, correlations are also described by a p value that represents the statistical significance of of the relationship. Low p values indicate that there is a small chance that the data represents randomness and not a correlative relationship, while large p values indicate that there is a greater likelihood that the data represents randomness. Interpreting correlative p values follows the same rules as hypothesis testing (see above).

The key thing to remember is that correlation does not equal causation and in fact the two variables may have no direct relationship to one another. A classic example is the finding that there is a positive correlation between ice cream sales and rape. The unsophisticated user of statistics might infer that one is caused by the other. In fact, both are co-correlates with at least two other variables: temperature and number of daylight hours. (This needs lots of development.)

Co-correlates in research.

When collecting data for research, it is important to reduce the amount of co-correlates. For example, weight is related to height and also to age. If trying to study the effect of age on weight, height must also be accounted for. This is the reason for using Body Mass Index (BMI), as it takes into account both height and weight.

Regression

Regression is the use of statistical models to explain the variation in data. The regression model is represented by the letter r-squared and represents the amount of variation that is explained by the model. Imagine for a moment that you were studying the effect of age on height in children. You measured multiple subjects as they aged, and recorded their height and age. Then you create a model to explain the variations in height. If the regression model's r-squared is 0.83, that means that 83% of the variations in your data are explained by your age model. The remaining variation is unexplained. The value of regression models is that they can not only be used to explain existing data but also to predict uncollected data with a known degree of confidence or statistical certainty. The model also has a p value. The best models will have high r-squared and low p values.

The simplest regression is simple linear regression (of which there is nothing simple nor linear). Recall from Algebra II that parabolas are considered linear even though they curve. What makes it "simple linear" is that the model contains only one variable and may be represented by a single line (or curve) on a graph.

Suppose for a moment that you realize that there is more to height than simply age. Perhaps you want to include other variables such as average daily calories and parental height. You now have a more complicated model and have entered the realm of multiple regression.

How do you know you have the best model? You don't. Generally speaking, regression modeling is a trial and error process. The researcher must trade off the accuracy of the model with the amount of time and energy it takes to collect the additional data and compute the various models.

The **two key sins of regression** are not using enough data points and extrapolation. The first sin is usually caused by budget and time constraints. Each study is different, but generally, at least 30 data points should be used for numerical response variables. The more complicated the model and the greater the variation in data, the greater the number of data points that should be used. A power analysis should be done before the study begins to determine the appropriate number of data points (subjects). The second sin is far more insidious and surrounds us daily.

Imagine for a moment, that with your study of age and height in children, you were able to come up with a model that has a very high r-squared and a very low p value. Using your model you have been able to successfully predict the height of all kinds of children not in your original study. At what point should you start using your model to predict the height of 30 year-olds? or 60 year olds? Using this common example, it is quite easy to see that the age-height relationship does not survive past lat adolescence, and using a model based in child height is not helpful to predict adult height at given ages. Making predictions beyond the data that went into the model is called extrapolation.

This same fallacy is seen every day in science and health care. It is seen in Global Warming models. It is seen in recommendations that 20 year olds take statins. Predictions based on extrapolation may or may not be accurate, but there is simply no way to judge *a priori* whether they are. Serious researchers should take every effort not to base their findings, predictions, or recommendations on extrapolation. And if they do

base their recommendations on extrapolation, they should take extra care to alert others to the potential for error.

Logistical regression

Logistical regression is nothing more than regression modeling for nominal variables. The result of logistical regression is an odds ratio (not the same as risk ratio*). Logistic regression requires more data points than linear regression or multiple regression.

*See page 11.