PERCENTAGES: THE MOST USEFUL STATISTICS EVER INVENTED

Thomas R. Knapp

"Eighty percent of success is showing up."
- Woody Allen

“Baseball is ninety percent mental and the other half is physical.”
- Yogi Berra

"Genius is one percent inspiration and ninety-nine percent perspiration."
- Thomas Edison

Preface

You know what a percentage is. 2 out of 4 is 50%. 3 is 25% of 12. Etc. But do you know enough about percentages? Is a percentage the same thing as a fraction or a proportion? Should we take the difference between two percentages or their ratio? If their ratio, which percentage goes in the numerator and which goes in the denominator? Does it matter? What do we mean by something being statistically significant at the 5% level? What is a 95% confidence interval? Those questions, and much more, are what this book is all about.

In his fine article regarding nominal and ordinal bivariate statistics, Buchanan (1974) provided several criteria for a good statistic, and concluded: "The percentage is the most useful statistic ever invented…" (p. 629). I agree, and thus my choice for the title of this book. In the ten chapters that follow, I hope to convince you of the defensibility of that claim.

The first chapter is on basic concepts (what a percentage is, how it differs from a fraction and a proportion, what sorts of percentage calculations are useful in statistics, etc.) If you're pretty sure you already understand such things, you might want to skip that chapter (but be prepared to return to it if you get stuck later on!).

In the second chapter I talk about the interpretation of percentages, differences between percentages, and ratios of percentages, including some common mis-interpretations and pitfalls in the use of percentages.

Chapter 3 is devoted to probability and its explanation in terms of percentages. I also include in that chapter a discussion of the concept of "odds" (both in favor of, and against, something). Probability and odds, though related, are not the same thing (but you wouldn't know that from reading much of the scientific and lay literature).

Chapter 4 is concerned with a percentage in a sample vis-à-vis the percentage in the population from which the sample has been drawn. In my opinion, that is the most elementary notion in inferential statistics, as well as the most important. Point estimation, interval estimation (confidence intervals), and hypothesis testing (significance testing) are all considered.

The following chapter goes one step further by discussing inferential statistical procedures for examining the difference between two percentages and the ratio of two percentages, with special attention to applications in epidemiology.

The next four chapters are devoted to special topics involving percentages. Chapter 6 treats graphical procedures for displaying and interpreting

percentages. It is followed by a chapter that deals with the use of percentages to determine the extent to which two frequency distributions overlap. Chapter 8 discusses the pros and cons of dichotomizing a continuous variable and using percentages with the resulting dichotomy. Applications to the reliability of measuring instruments (my second most favorite statistical concept--see Knapp, 2009) are explored in Chapter 9. The final chapter attempts to summarize things and tie up loose ends.

There is an extensive list of references, all of which are cited in the text proper. You may regard some of them as "old" (they actually range from 1919 to 2009). I like old references, especially those that are classics and/or are particularly apt for clarifying certain points. [And I'm old too.]

Enjoy!

Table of Contents

Chapter 1: The basics

What is a percentage?

A percentage is a part of a whole. It can take on values between 0 (none of the whole) and 100 (all of the whole). The whole is called the base. The base must ALWAYS be reported whenever a percentage is determined.

Example: There are 20 students in a classroom, 12 of whom are males and 8 of whom are females. The percentage of males is 12 "out of" 20, or 60%. The percentage of females is 8 "out of" 20, or 40%. (20 is the base.)

To how many decimal places should a percentage be reported?

One place to the right of the decimal point is usually sufficient, and you should almost never report more than two. For example, 2 out of 3 is 66 2/3 %, which rounds to 66.67% or 66.7%. [To refresh your memory, you round down if the fractional part of a mixed number is less than 1/2 or if the next digit is 0, 1, 2, 3, or 4; you round up if the fractional part is greater than or equal to 1/2 or if the next digit is 5, 6, 7, 8, or 9.] Computer programs can report numbers to ten or more decimal places, but that doesn't mean that you have to. I believe that people who report percentages to several decimal places are trying to impress the reader (consciously or unconsciously).

Lang and Secic (2006) provide the following rather rigid rule:

> "When the sample size is greater than 100, report percentages to no more than one decimal place. When sample size is less than 100, report percentages in whole numbers. When sample size is less than, say, 20, consider reporting the actual numbers rather than percentages." (p. 5)

[Their rule is just as appropriate for full populations as it is for samples. And they don't say it, perhaps because it is obvious, but if the size of the group is equal to 100, be it sample or population, the percentages are the same as the numerators themselves, with a % sign tacked on.]

How does a percentage differ from a fraction and a proportion?

Fractions and proportions are also parts of wholes, but both take on values between 0 (none of the whole) and 1 (all of the whole), rather than between 0 and 100. To convert from a fraction or a proportion to a percentage you multiply by 100 and add a % sign. To convert from a percentage to a proportion you delete the % sign and divide by 100. That can in turn be converted to a fraction. For example, 1/4 multiplied by 100 is 25%. .25 multiplied by 100 is also 25%. 25% divided by 100 is .25, which can be expressed as a fraction in a variety of

ways, such as 25/100 or, in "lowest terms", 1/4. (See the excellent On-Line Math Learning Center website for examples of how to convert from any of these part/whole statistics to any of the others.) But, surprisingly (to me, anyhow), people tend to react differently to statements given in percentage terms vs. fractional terms, even when the statements are mathematically equivalent. (See the October 29, 2007 post by Roger Dooley on the Neuromarketing website. Fascinating.)

Most authors of statistics books, and most researchers, prefer to work with proportions. I prefer percentages [obviously, or I wouldn't have written this book!].

One well-known author (Gerd Gigerenzer) prefers fractions to both percentages and proportions. In his book (Gigerenzer, 2002) and in a subsequent article he co-authored with several colleagues (Gigerenzer, et al., 2008), he advocates an approach that he calls the method of "natural frequencies" for dealing with percentages. For example, instead of saying something like "10% of smokers get lung cancer", he would say "100 out of every 1000 smokers get lung cancer" [He actually uses breast cancer to illustrate his method]. Heynen (2009) agrees. But more about that in Chapter 3, in conjunction with positive diagnoses of diseases.

Is there any difference between a percentage and a percent?

The two terms are often used interchangeably (as I do in this book), but "percentage" is sometimes regarded as the more general term and "percent" as the more specific term. The AMA Manual of Style and the BioMedical Editor and Grammar Girl websites have more to say regarding that distinction. The Grammar Girl (Mignon Fogarty) also explains whether percentage takes a singular or plural verb, whether to use words or numbers before the % sign, whether to have a leading 0 before a decimal number that can't be greater than 1, and all sorts of other interesting things.

Do percentages have to add to 100?

A resounding YES, if the percentages are all taken on the same base for the same variable, if only one "response" is permitted, and if there are no missing data. For a group of people consisting of both males and females, the % male plus the % female must be equal to 100, as indicated in the above example (60+40=100). If the variable consists of more than two categories (a two-categoried variable is called a dichotomy), the total might not add to 100 because of rounding. As a hypothetical example, consider what might happen if the variable is something like Religious Affiliation and you have percentages reported to the nearest tenth for a group of 153 people of 17 different religions. If those percentages add exactly to 100 I would be terribly surprised.

Several years ago, Mosteller, Youtz, and Zahn (1967) determined that the probability (see Chapter 3) of rounded percentages adding exactly to 100 is perfect for two categories, approximately 3/4 for three categories, approximately 2/3 for four categories, and approximately $\sqrt{6}/c\pi$ for $c \geq 5$, where c is the number of categories and π is the well-known ratio of the circumference of a circle to its diameter (= approximately 3.14). Amazing!

[For an interesting follow-up article, see Diaconis & Freedman (1979). Warning: It has some pretty heavy mathematics!]

Here's a real-data example of the percentages of the various possible blood types for the U.S.:

O Positive	38.4%
A Positive	32.3%
B Positive	9.4%
O Negative	7.7%
A Negative	6.5%
AB Positive	3.2%
B Negative	1.7%
AB Negative	.7%

[Source: American Red Cross website]

Those add to 99.9%. The probability that they would add exactly to 100%, by the Mosteller, et al. formula, is approximately .52.

Can't a percentage be greater than 100?

I said above that percentages can only take on values between 0 and 100. There is nothing less than none of a whole, and there is nothing greater than all of a whole. But occasionally [too often, in my opinion] you will see a statistic such as "Her salary went up by 200%" or "John is 300% taller than Mary". Those examples refer to a comparison in terms of a percentage, not an actual percentage. I will have a great deal to say about such comparisons in the next chapter and in Chapter 5.

Why are percentages ubiquitous?

People in general, and researchers in particular, have always been interested in the % of things that are of a particular type, and they always will be. What % of voters voted for Barack Obama in the most recent presidential election? What % of smokers get lung cancer? What % of the questions on a test do I have to answer correctly in order to pass?

An exceptionally readable source about opinion polling is the article in the Public Opinion Quarterly by Wilks (1940a), which was written just before the entrance of the U.S. into World War II, a time when opinions regarding that war were diverse

and passionate. I highly recommend that article to those of you who want to know how opinion polls SHOULD work. S.S. Wilks was an exceptional statistician.

What is a rate?

A rate is a special kind of percentage, and is most often referred to in economics, demography, and epidemiology. An interest rate of 10%, for example, means that for every dollar there is a corresponding $1.10 that needs to be taken into consideration (whether it is to your advantage or to your disadvantage).

There is something called "The Rule of 72" regarding interest rates. If you want to determine how many years it would take for your money to double if it were invested at a particular interest rate, compounded annually, divide the interest rate into 72 and you'll have a close approximation. To take a somewhat optimistic example, if the rate is 18% it would take four years (72 divided by 18 is 4) to double your money. [You would actually have "only" 1.93877 times as much after four years, but that's close enough to 2 for government work! Those of you who already know something about compound interest might want to check that.]

Birth rates and death rates are of particular concern in the analysis of population growth or decline. In order to avoid small numbers, they are usually reported "per thousand" rather than "per hundred" (which is what a simple percent is). For example, if in the year 2010 there were to be six million births in the United States "out of" a population of 300 million, the ("crude") birth rate would be 6/300, or 2%, or 20 per thousand. If there were to be three million deaths in that same year, the (also "crude") death rate would be 3/300, or 1%, or 10 per thousand.

One of the most interesting rates is the "response rate" for surveys. It is the percentage of people who agree to participate in a survey. For some surveys, especially those that deal with sensitive matters such as religious beliefs and sexual behavior, the response rate is discouragingly low (and often not even reported), so that the results must be taken with more than the usual grain of salt.

Some rates are phrased in even different terms, e.g., parts per 100,000 or parts per million (the latter often used to express the concentration of a particular pollutant).

What kinds of calculations can be made with percentages?

The most common kinds of calculations involve subtraction and division. If you have two percentages, e.g., the percentage of smokers who get lung cancer and the percentage of non-smokers who get lung cancer, you might want to subtract one from the other or you might want to divide one by the other. Which is it better to do? That matter has been debated for years. If 10% of smokers get

lung cancer and 2% of non-smokers get lung cancer (the two percentages are actually lower than that for the U.S.), the difference is 8% and the ratio is 5-to-1 (or 1-to-5, if you invert that ratio). I will have much more to say about differences between percentages and ratios of percentages in subsequent chapters. (And see the brief, but excellent, discussion of differences vs. ratios of percentages at the American College of Physicians website.)

Percentages can also be added and multiplied, although such calculations are less common than the subtraction or division of percentages. I've already said that percentages must add to 100, whenever they're taken on the same base for the same variable. And sometimes we're interested in "the percentage of a percentage", in which case two percentages are multiplied. For example, if 10% of smokers get lung cancer and 60% of them are men, the percentage of smokers who get cancer and are male is 60% of 10%, or 6%. (By subtraction, the other 4% are female.)

You also have to be careful about averaging percentages. If 10% of smokers get lung cancer and 2% of non-smokers get lung cancer, you can't just "split the difference" between those two numbers to get the % of people in general who get lung cancer by adding them together and dividing by two (to obtain 6%). The number of non-smokers far exceeds the number of smokers (at least in 2009), so the percentages have to be weighted before averaging. Without knowing how many smokers and non-smokers there are, all you know is that the average lung cancer % is somewhere between 2% and 10%, but closer to the 2%. [Do you follow that?]

What is inverse percentaging?

You're reading the report of a study in which there is some missing data (see the following chapter), with one of the percentages based upon an n of 153 and another based upon an n of 147. [153 is one of my favorite numbers. Do you know why? I'll tell you at the end of this book.] You are particularly interested in a variable for which the percentage is given as 69.8, but the author didn't explicitly provide the n for that percentage (much less the numerator that got divided by that n). Can you find out what n is, without writing to the author?

The answer is a qualified yes, if you're good at "inverse percentaging". There are two ways of going about it. The first is by brute force. You take out your trusty calculator and try several combinations of numerators with denominators of 153 and 147 and see which, if any, of them yield 69.8% (rounded to the nearest tenth of a percent). OR, you can use a book of tables, e.g., the book by Stone (1958), and see what kinds of percentages you get for what kinds of n's.

Stone's book provides percentages for all parts from 1 to n of n's from 1 to 399. You turn to the page for an n of 153 and find that 107 is 69.9% of 153. (That is the closest % to 69.8.) You then turn to the page for 147 and find that 102 is

69.4% of 147 and 103 is 70.1% of 147. What is your best guess for the n and for the numerator that you care about? Since the 69.9% for 107 out of 153 is very close to the reported 69.8% (perhaps the author rounded incorrectly or it was a typo?), since the 69.4% for 102 out of 147 is not nearly as close, and the 70.1% is also not as close (and is an unlikely typo), your best guess is 107 out of 153. But you of course could be wrong.

What about the unit of analysis and the independence of observations?

In my opinion, more methodological mistakes are made regarding the unit of analysis and the independence of observations than in any other aspect of a research study. The unit of analysis is the entity (person, classroom, school,…whatever) upon which any percentage is taken. The observations are the numbers that are used in the calculation, and they must be independent of one another.

If, for example, you are determining the percentage male within a group of 20 people, and there are 12 males and 8 females in the group (as above), the percentage of male persons is 12/20 or 60%. But that calculation assumes that each person is counted only once, there are no twins in the group, etc. If the 20 persons are in two different classrooms, with one classroom containing all 12 of the males and the other classroom containing all 8 of the females, then the percentage of male classrooms is 1/2 or 50%, provided the two classrooms are independent They could be dependent if, to take an admittedly extreme case, there were 8 male/female twin-pairs who were deliberately assigned to different classrooms, with 4 other males joining the 8 males in the male classroom. [Gets tricky, doesn't it?]

One of the first researchers to raise serious concerns about the appropriate unit of analysis and the possibility of non-independent observations was Robinson (1950) in his investigation of the relationship between race and literacy. He found (among other things) that for a set of data in the 1930 U.S. Census the correlation between a White/Black dichotomy and a Literate/Illiterate dichotomy was only .203 with individual person as the unit of analysis (n = 97,272) but was .946 with major geographical region as the unit of analysis (n = 9), the latter being the so-called "ecological" correlation between % Black and % Illiterate. His article created all sorts of reactions from disbelief to demands for re-analyses of data for which something other than the individual person was used as the unit of analysis. It (his article) was recently reprinted in the International Journal of Epidemiology, along with several commentaries by Subramanian, et al. (2009a, 2009b), Oakes (2009), Firebaugh (2009), and Wakefield (2009). I have also written a piece about the same problem (Knapp, 1977a).

What is a percentile?

A percentile is a point on a scale below which some percentage of things fall. For example, “John scored at the 75th percentile on the SAT” means that 75% of the takers scored lower than he did and 25% scored higher. We don’t even know, and often don’t care, what his actual score was on the test. The only sense in which a percentile refers to a part of a whole is as a part of all of the people, not a part of all of the items on the test.

Chapter 2: Interpreting percentages

Since a percentage is simple to calculate (much simpler than, say, a standard deviation, the formula for which has 11 symbols!), you would think that it is also simple to interpret. Not so, as this chapter will now show.

Small base

It is fairly common to read a claim such as "66 2/3 % of doctors are sued for malpractice". The information that the claimant doesn't provide is that only three doctors were included in the report and two of them were sued. In the first chapter I pointed out that the base upon which a percentage is determined must be provided. There is (or should be) little interest in a study of just three persons, unless those three persons are very special indeed.

There is an interesting article by Buescher (2008) that discusses some of the problems with using rates that have small numbers in the numerator, even if the base itself is large. And in his commentary concerning an article in the journal JACC Cardiovascular Imaging, Camici (2009) advises caution in the use of any ratios that refer to percentages.

Missing data

The bane of every researcher's existence is the problem of missing data. You go to great lengths in designing a study, preparing the measuring instruments, etc., only to find out that some people, for whatever reason, don't have a measurement on every variable. This situation is very common for a survey in which questions are posed regarding religious beliefs and/or sexual behavior. Some people don't like to be asked such questions, and they refuse to answer them. What is the researcher to do? Entire books have been written about the problem of missing data (e.g., Little & Rubin, 2002). Consider what happens when there is a question in a survey such as "Do you believe in God?", the only two response categories are yes and no, and you get 30 yeses, 10 nos, and 10 "missing" responses in a sample of 50 people. Is the "%yes" 30 out of 50 (=60%) or 30 out of 40 (= 75%)? And Is the "%no" 10 out of 50 (=20%) or 10 out of 40 (=25%)? If it's out of 50, the percentages (60 and 20) don't add to 100. If it's out of 40, the base is 40, not the actual sample size of 50 (that's the better way to deal with the problem…"no response" becomes a third category).

Overlapping categories

Suppose you're interested in the percentages of people who have various diseases. For a particular population the percentage having AIDS plus the percentage having lung cancer plus the percentage having hypertension might very well add to more than 100 because some people might suffer from more

than one of those diseases. I used this example in my little book entitled Learning statistics through playing cards (Knapp, 1996, p. 24). The three categories (AIDS, lung cancer, and hypertension) could "overlap". In the technical jargon of statistics, they are not "mutually exclusive".

Percent change

Whenever there are missing data (see above) the base changes. But when you're specifically interested in percent change the base also does not stay the same, and strange things can happen. Consider the example in Darrell Huff's delightful book, How to lie with statistics (1954), of a man whose salary was $100 per week and who had to take a 50% pay cut to $50 per week because of difficult economic times. [(100-50)/100 = .50 or 50%.] Times suddenly improved and the person was subsequently given a 50% raise. Was his salary back to the original $100? No. The base has sifted from 100 to 50. $50 plus 50% of $50 is $75, not $100. [The illustrations by Irving Geis in Huff's book are hilarious!] There are several other examples in the research literature and on the internet regarding the problem of % decrease followed by % increase, as well as % increase followed by % decrease, % decrease followed by another % decrease, and % increase followed by another % increase. (See, for example, the definition of a percentage at the wordIQ.com website; the Pitfalls of Percentages webpage at the Hypography website; the discussion of percentages at George Mason University's STATS website; and the article by Chen and Rao, 2007.)

A recent instance of a problem in interpreting percent change is to be found in the research literature on the effects of smoking bans. Several authors (e.g., Lightwood & Glantz, 2009; Meyers, 2009) claim that smoking bans cause decreases in acute myocardial infarctions (AMI). They base their claims upon meta-analyses of a small number of studies that found a variety of changes in the percent of AMIs, e.g., Sargent, Shepard, and Glantz (2004), who investigated the numbers of AMIs in Helena, MT before a smoking ban, during the time the ban was in effect, and after the ban had been lifted. There are several problems with such claims, however:

1. Causation is very difficult to determine. There is a well-known dictum in research methodology that "correlation is not necessarily causation". As Sargent, et al. (2004) themselves acknowledged:

> "This is a "before and after" study that relies on historical controls (before and after the period that the law was in effect), not a randomised controlled trial. Because this study simply observed a change in the number of admissions for acute myocardial infarction, there is always the chance that the change we observed was due to some unobserved confounding variable or systematic bias." (p. 979)

2. Sargent, et al. found a grand total of 24 AMIs in the city of Helena during the six-month ban in the year 2002, as opposed to an average of 40 AMIs in other six-month periods just before and just after the ban. Those are very small numbers, even though the difference of 16 is "statistically significant" (see Chapters 4 and 5). They also compared that difference of 16 AMIs to a difference of 5.6 AMIs between 18 during and an average of 12.4 before and after for a "not Helena" area (just outside of Helena). The difference between those two differences of 16 and 5.6 was also found to be small but "statistically significant". But having a "not Helena" sample is not the same as having a randomly comparable group in a controlled experiment.

3. But to the point of this section, the drop from 40 to 24 within Helena is a 40% change (16 "out of" 40); the "rebound" from 24 to 40 is a 66 2/3% change (16 "out of" the new base of 24). To their credit, Sargent et al. did not emphasize the latter, even though it is clear they believe it was the ban and its subsequent rescission that were the causes of the decrease followed by the increase.

[Note: The StateMaster.com website cites the Helena study as an example of a "natural experiment". I disagree. In my opinion, "natural experiment" is an oxymoron. There is nothing natural about an experiment, which is admittedly artificial (the researcher intervenes), but necessary for the determination of causation. Sargent, et al. did not intervene. They just collected existing data.]

I recently encountered several examples of the inappropriate calculation and/or interpretation of percent change in a table in a newpaper (that shall remain nameless) on % increase or decrease in real estate sales prices. The people who prepared the table used [implicitly] a formula for % change of the form (Time 2 price minus Time 1 price)/ Time 1 price. One of the comparisons involved a median price at Time 1 of $0 and a median price at Time 2 of $72,500 that was claimed to yield a 0% increase, since the calculation of ($72,500 - 0)/ 0 was said to be equal to 0. Not so. You can't divide by 0, so the percent increase was actually indeterminate.

Percent difference

Percent change is a special case of percent difference. (It's change if it's for the same things, usually people, across time.) Both percent difference and the difference between two percentages (see Chapter 5) come up all of the time [but they're not the same thing, so be careful!].

The percent difference between two continuous quantities

The percent difference between two continuous quantities is also not the same as the difference between two percents. Cole (2000) suggests that it is better to use logarithms when interpreting the percent difference between two continuous quantities. He gives the example of a comparison between the average height of

British adult men (177.3 centimeters, which is approximately 69.8 inches, or slightly under 5'10") and the average height of British adult women (163.6 centimeters, which is approximately 64.4 inches). The usual formula for finding the percent difference between two quantities x_1 and x_2 is $100(x_2 - x_1)/x_1$. But which do you call x_1 and which do you call x_2? Working the formula one way (with x_1 = the average height of the women and x_2 = the average height of the men), you find that the men are 8.4% taller than the women. Working the formula the other way (with x_1 = the average height of the men and x_2 = the average height of the women) you find that the women are 7.7% shorter than the men (the numerator is negative). Cole doesn't like that asymmetry. He suggests that the formula be changed to $(100\log_e x_2 - 100\log_e x_1)$, where e = 2.1728… is the base of the natural logarithm system. If you like logarithms and you're comfortable working with them, you'll love Cole's article!

Comparisons between percentages that must add to 100

One annoying (to me, anyhow) tendency these days is to compare, by subtraction or division, the percentage of support for one candidate for political office with the percentage of support for another candidate when they are the only two candidates for that office. For example: "Smith is leading Jones by 80% to 20%, a difference of 60 points." Of course. If Smith has 80%, Jones must have 20% (unless there are missing data!), the difference must be 60%, and why use the word "points"?!

The situation is no better if the comparison takes the form "Smith has four times the support that Jones has". Again, of course. The only number that is necessary to report is EITHER the 80% for Smith or the 20% for Jones. Everything else follows automatically.

Ratios vs. differences of percentages

Consider an example (unlike the previous example) where it is reasonable to calculate the ratio of two percentages that don't have to add to 100. Suppose one-half of one percent of males in a population of 100 million males have IQ scores of over 200 and two percent of females in a population of 100 females have IQ scores of over 200. (There are approximately 100 million adult males and approximately 100 million adult females in the United States.) Should we take the ratio of the 2% to the .5% (a ratio of 4 to 1) and claim that the females are four times as smart?

No. There are at least two problems with such a claim. First of all, having a number less than 1 in the denominator and a number greater than 1 in the numerator can produce an artificially large quotient. (If the denominator were 0, the ratio couldn't even be calculated, since you can't divide by 0.) Secondly, does it really matter how large such a ratio is, given that both numerator and

denominator are small. Surely it is the difference between those two percentages that is important, not their ratio.

Although in general there are fewer problems in interpreting differences between percentages than there are in interpreting ratios of percentages, when subgroup comparisons are made in addition to an overall comparison, things can get very complicated. The classic case is something called Simpson's Paradox (Simpson, 1951) in which the differences between two overall percentages can be in the opposite direction from differences between their corresponding subgroup percentages. The well-known mathematician John Allen Paulos (2001) provided the following hypothetical (but based upon an actual lawsuit) example:

> "To keep things simple, let's suppose there were only two departments in the graduate school, economics and psychology. Making up numbers, let's further assume that 70 of 100 men (70 percent) who applied to the economics department were admitted and that 15 of 20 women (75 percent) were. Assume also that five out of 20 men (25 percent) who applied to the psychology department were admitted and 35 of 100 women (35 percent) were. Note that in each department a higher percentage of women was admitted.
>
> If we amalgamate the numbers, however, we see what prompted the lawsuit: 75 of the 120 male applicants (62.5 percent) were admitted to the graduate school as a whole whereas only 50 of the 120 female applicants (41.7 percent) were. "

How can that be? The "paradox" arises from the fact that there are unequal numbers of men and women contributing to the percentages (100 men and 20 women for economics; 20 men and 100 women for psychology). The percentages need to be weighted before they are combined into overall figures.

For additional discussions of Simpson's Paradox, see Malinas (2001) and Ameringer, Serlin, and Ward (2009).

Reporting of ranges in percentages across studies

In his editorial a few years ago, Cowell (1998) referred to an author's citing of the results of a successful surgical procedure as ranging from 43% to 100%, without mentioning that the 100% was for one successful procedure performed on one patient! He (Cowell) argued, as I have, that the base must always be given along with the percentage.

Other misunderstandings and errors in interpreting percentages

Milo Schield (2000) discussed a number of problems that people have when it comes to percentages in various contexts, especially rates. One example he

cites is the claim made by some people that “if X% of A are B, then X% of B are A”. No. If you don’t believe Schield or me, try various numbers or draw a “Venn diagram” for two overlapping circles, A and B. Schield is the director of the W.M. Keck Statistical Literacy Project at Augsburg College in Minneapolis. He has written other interesting articles regarding statistical literacy (or lack of same), one of which (Schield, 2005) deals largely with misunderstandings of percentages. He also put on the internet a test of statistical literacy (Schield, 2002). You can get to it by googling “statistical literacy inventory” and clicking on the first entry. Here are three of the questions on that test (as cited in The Washington Post on February 6, 2009):

> 1. True or False. If a stock decreases 50 percent and then increases by 50 percent, it will be back to its original value.
>
> 2. True or False. If a stock drops from $300 to zero, that is a 300 percent decrease.
>
> 3. A company has a 30 percent market share in the Western US and a 10 percent market share in the Eastern US. What is the company's overall market share in the entire US?

Do you know the answers?

The interesting book, Mathsemantics, by Edward MacNeal (1994), includes a chapter on percentages in which the author discusses a number of errors that he discovered when a group of 196 applicants for positions with his consulting firm were tested. Some of the errors, and some of the reasons that people gave for having made them, are pretty bizarre. For example, when asked to express .814 as a percentage to the nearest whole percent, one person gave as the answer “1/8 %”. One of my favorite examples in that chapter is to a person (fortunately nameless) who claimed that Richie Ashburn, a former baseball player with the Philadelphia Phillies, hit “three hundred and fifty percent” in one of his major league seasons. [I’m a baseball nut.] In case you’re having trouble figuring out what’s wrong with that, I’ll help you out. First of all, as you now know (if you already didn’t), percentages must be between 0 and 100. Secondly, baseball batting averages are “per thousand” rather than “per hundred”. Ashburn actually hit safely about 35% of the time, which, in baseball jargon, is “three fifty”, i.e., a proportion of .350.

Speaking of my love for baseball, and going back to Simpson’s Paradox, I once wrote an article (Knapp, 1985) in which I provided real data that showed Player A had a higher batting average than Player B against both right-handed and left-handed pitching but had a lower overall batting average. I later discovered additional instances of batting averages that constituted evidence for a transitive case of Simpson’s Paradox, i.e., one for which A > B > C against both right-handed and left-handed pitching, but for which A < B < C overall. (The symbol > means “greater than”; < means “less than”.)

Chapter 3: Percentages and probability

What do we mean by the probability of something?

There are several approaches to the definition of probability. The first one that usually comes to mind is the so-called "a priori definition" that is a favorite of teachers of statistics who use coins, dice, and the like to explain probability. In the "a priori" approach the probability of something is the number of ways that something can take place divided by the total number of equally likely outcomes. For example, in a single toss of a coin, the probability of "heads" is the number of ways that can happen (1) divided by the total number of equally likely outcomes (2…"heads" or "tails"), which is 1/2, .50, or 50%, depending upon whether you want to use fractions, proportions, or percentages to express the probability. Similarly, the probability of a "4" in a single roll of a die, is 1 (there is only one side of a die that has four spots) divided by 6 (the total number of sides), which is equal to 1/6, .167, or 16.7% (to three "significant figures").

But there are problems with that definition. In the first place, it only works for symmetric situations such as the tossing of fair coins and the rolling of unloaded dice. Secondly, it is actually circular, since it defines probability in terms of "equally likely", which is itself a probabilistic concept. Such concerns have led to a different definition, the "long-run empirical definition", in which the probability of something is the number of ways that something did happen (note the use of "did" rather than "can") divided by the total number of things that happened. This definition works for things like thumbtack tosses (what is the probability of landing with its point up?) as well as for coins, dice, and many other probabilistic contexts. The price one pays, however, is the cost of actually carrying out the empirical demonstration of tossing a thumbtack (or tossing a coin or rolling a die…) a large number of times. And how large is large??

There is a third (and somewhat controversial) "subjective definition" of probability that is used in conjunction with Bayesian statistics (an approach to statistics associated with a famous equation derived by the clergyman/mathematician Rev. Thomas Bayes who lived in the 18th century). Probability is defined as a number between 0 and 1 (for fractions and proportions) or between 0 and 100 (for percentages) that is indicative of a person's "strength of conviction" that something will take place (note the use of "will" rather than either "can" or "did").

An example that illustrates various definitions of probability, especially the subjective definition, is the question of the meaning of "the probability of rain". There recently appeared an article in The Journal of the American Meteorological Society, written by Joslyn, Nadav-Greenberg, and Nichols (2009), that was devoted entirely to that problem. (Weather predictions in terms of percentages and probabilities have been around for about a century---see, for example, Hallenbeck, 1920.)

I've already said that I favor the reporting of parts out of wholes in terms of percentages rather than in terms of fractions or proportions. I also favor the use of playing cards rather than coins or dice to explain probabilities. That should come as no surprise to you, since in the previous chapter I referred to my Learning statistics through playing cards book (Knapp, 1996), which, by the way, also concentrates primarily on percentages.

The probability of not-something

If P is the probability that something will take place, in percentage terms, then 100 – P is the probability that it will not take place. For example, if you draw one card from an ordinary deck of cards, the probability P that it's a spade is 13/52, or 1/4, or .25, or 25%. The probability that it's not a spade is 100 – 25 = 75%, which can also be written as 39/52, or 3/4, or .75.

Probabilities vs. odds

People are always confusing probabilities and odds. If P is the probability of something, in percentage terms, then the odds in favor of that something are P divided by (100 - P); and the odds against it are (100 - P) divided by P. The latter is usually of greater interest, especially for very small probabilities. For example, if you draw one card from an ordinary deck of cards, the probability P that it's a spade, from above, is 13/52, or 1/4, or .25, or 25%. The odds in favor of getting a spade are 25 divided by (100 – 25), or "1 in 3"; the odds against it are (100 – 25) divided by 25, or "3 to 1". [In his book, Probabilities and life , the French mathematician Emile Borel (1962) claims that we act as though events with very small probabilities never occur. He calls that "the single law of chance".]

There are actually two mistakes that are often made. The first is the belief that probabilities and odds are the same thing (so some people would say that the odds of getting a spade are 1/4, or 25%). The second is the belief that the odds against something are merely the reciprocal of its probability (so they would say that the odds against getting a spade are 4 to 1).

"Complex" probabilities

The examples just provided were for "simple" situations such as tossing a coin once, rolling a die once, or drawing one card from a deck of cards, for which you are interested in a simple outcome. Most applications of probability involve more complicated matters. If there are two "events", A and B, with which you are concerned, the probability that either of them will take place is the sum of their respective probabilities, if the events are mutually exclusive, and the probability that both of them will take place is the product of their respective probabilities, if the events are independent. Those are both mouthfuls, so let's take lots of examples (again using playing cards):

1. If you draw one card from a deck of cards, what is the probability that it is either a spade or a heart?

Since getting a spade and getting a heart are mutually exclusive (a card cannot be a spade and a heart), the probability of either a spade or a heart is the probability of a spade plus the probability of a heart, which is equal to 13/52 plus 13/52 = 26/52 = 1/2, or 50%. [It's generally easier to carry out the calculations using fractions, but to report the answer in terms of percentages.]

2. If two cards are drawn from a deck of cards, what is the probability that they are both spades?

This problem is a bit more difficult. We must first specify whether or not the first card is replaced in the deck before the second card is drawn. If the two "events", spade on first card and spade on second card, are to be independent (i.e., that the outcome of the second event does not depend upon the outcome of the first event) the first card must be replaced. If so, the desired probability is 1/4 for the first card times 1/4 for the second card, which is equal to 1/16 or 6.25%. If the first card is not replaced, the probability is 13/52 times 12/51 = 1/4 times 4/17 = 1/17 or 5.88%.

3. If two cards are drawn from a deck of cards, what is the probability that either of them is a spade?

This is indeed a complex problem. First of all, we need to know if the cards are to be drawn "with replacement" (the first card is returned to the deck before the second card is drawn) or "without replacement" (it isn't). Secondly, we need to specify whether "either" means "one but not both" or "one or both". Let us consider just one of the four combinations. (I'll leave the other three as exercises for the curious reader!)

If the drawing is with replacement and "either" means "one but not both", the possibilities that are favorable to getting a spade are "spade on first draw, no spade on second draw" and "no spade on first draw, spade on second draw". Those probabilities are, using the "or" rule in conjunction with the "and" rule, (1/4 times 3/4) plus (3/4 times 1/4), i.e., 3/16 + 3/16, or 6/16, or 3/8, or 37.5%.

4. In his other delightful book, How to take a chance, Darrell Huff (1959) discusses the probability of having two boys out of four children, if the probability of a boy and the probability of a girl are equally likely and independent of one another. Many people think the answer is 2/4 = 1/2 = .50 = 50%. Huff not only shows that the correct answer is 6/16 = 3/8 = .375 = 37.5%, but he (actually Irving Geis) illustrates each of the permutations. You can look it up (as Casey Stengel used to say). This is a conceptually different probability problem than the previous one. It just happens to have the same answer.

The birthday problem

There is a famous probability problem called “The Birthday Problem”, which asks: If n people are gathered at random in a room, what is the probability that at least two of them have the same birthday (same month and day, but not necessarily same year)? It turns out that for an n of 23 the probability is actually (and non-intuitively) greater than 50%, and for an n of 70 or so it is a virtual certainty! See, for example, the website www.physics.harvard.edu/academics/undergrad/probweek/sol46 and my favorite mathematics book, Introduction to finite mathematics (Kemeny, Snell, and Thompson, 1956). The best way to carry out the calculation is to determine the probability that NO TWO PEOPLE will have the same birthday (using the generalization of the “and” rule---see above), and subtract that from 100 (see the probability of not-something).

Risks

A risk is a special kind of percentage, and a special kind of probability, which is of particular interest in epidemiology. The risk of something, e.g., getting lung cancer, can be calculated as the number of people who get something divided by the total number of people who “could” get that something. (The risk of lung cancer, actually the “crude” risk of lung cancer, is actually rather low in the United States, despite all of the frightening articles about its prevalence and its admittedly tragic consequences.)

There is also an attributable risk (AR), the difference between the percentage of people in one group who get something and the percentage of people in another group who get that something. [N.B. "Attributable" doesn't necessarily mean causal.] In Chapter 1 I gave a hypothetical example of the percentage of smokers who get lung cancer minus the percentage of non-smokers who get lung cancer, a difference of 10% - 2% = 8%.

And then there is a relative risk (RR), the ratio of the percentage of people in one group who get something to the percentage of people in another group who get that something. Referring back again to smoking and lung cancer, my hypothetical example produced a ratio of 10%/ 2%, or “5 to 1”.

Risks need not only refer to undesirable outcomes. The risk of making a million dollars by investing a thousand dollars, for example, is a desirable outcome (at least for the “winner”).

The methodological literature is replete with discussions of the minimum value of relative risk that is worthy of serious consideration. The most common value is "2.00 or more", especially when applying relative risks to individual court cases. Those same sources often have corresponding discussions of a related concept, the probability of causality [causation], PC, which is defined as 1 - 1/RR. If the

RR threshold is 2.00, then the PC threshold is .50, or 50%. See Parascandola (1998); Robins (2004); Scheines (2008); and Swaen and vanAmelsvoort (2009) for various points of view regarding both of those thresholds.

Sensitivity and specificity

In medical diagnostic testing there are two kinds of probability that are of interest:

1. The probability that the test will yield a "positive" result (a finding that the person being tested has the disease) if the person indeed has the disease. Such a probability is referred to as the sensitivity of the test.

2. The probability that the test will yield a "negative" result (a finding that the person being tested does not have the disease) if the person indeed does not have the disease. Such a probability is referred to as the specificity of the test.

We would like both of those probabilities to be 100% (a perfect test). Alas, that is not possible. No matter how much time and effort go into devising diagnostic tests there will always be "false positives" (people who don't have the disease but are said to have it) and "false negatives" (people who do have the disease but are said to not have it). [Worse yet, as you try to improve the test by cutting down on the number of false positives you increase the number of false negatives, and vice versa.] Its sensitivity is the probability of a "true positive"; its specificity is the probability of a "true negative".

There is something called Youden's Index (Youden, 1950), which combines sensitivity and specificity. Its formula can be written in a variety of ways, the simplest being J = sensitivity + specificity – 100. Theoretically it can range from -100 (no sensitivity and no specificity) to 100 (perfect test), but is typically around 80 (e.g., when both sensitivity and specificity are around 90%). [A more interesting re-formulation of Youden's Index can be written as J = (100-sensitivity) – (100-specificity), i.e., the difference between the true positive rate and the false positive rate.]

For example (an example given by Gigerenzer, 2002, and pursued further in Gigerenzer et al., 2008 with slightly changed numbers), a particular mammography screening test might have a sensitivity of 90% and a specificity of 91% (those are both high probabilities, but not 100%). Suppose that the probability of getting breast cancer is 1% (10 chances in 1000). For every group of 1000 women tested, 10 of whom have breast cancer and 990 of whom do not, 9 of those who have it will be correctly identified (since the test's sensitivity is 90%, and 90% of 10 is 9). For the 990 who do not have breast cancer, 901 will be correctly identified (since the test's specificity is 91%, and 91% of 990 is 901). Therefore there will be 9 true positives, 901 true negatives, 89 false positives, and 1 false negative.

Gigerenzer goes on to point out the surprising conclusion that for every positive finding only about 1 in 11 (9 out of the 98 "positives"), or approximately 9%, is correct. He argues that if a woman were to test positive she needn't be overly concerned, since the probability that she actually has breast cancer is only 9%, with the corresponding odds of "89 to 9" (almost 10 to 1) against it. A further implication is that it might not be cost-effective to use diagnostic tests with sensitivities and specificities as "low" as those.

In his delightful book entitled Innumeracy (note the similarity to the word "illiteracy"), Paulos (1988) provides a similar example (p. 66) that illustrates how small the probability typically is of having a disease, given a positive diagnosis.

For another (negative) commentary regarding cancer screening, see the recent JNCI editorial by Woloshin and Schwartz (2009).

Probabilistic words and their quantification in terms of percentages

The English language is loaded with words such as "always", "never", "sometimes", "seldom", etc. [Is "sometimes" more often than "seldom"; or is it the other way 'round?] There is a vast literature on the extent to which people ascribe various percentages of the time to such words. The key reference is an article that appeared in the journal Statistical Science written by Mosteller and Youtz (1990; see also the several comments regarding that article in the same journal and the rejoinder by Mosteller and Youtz). They found, for example, that across 20 different studies the word "possible" received associated percentages throughout the entire scale (0% to 100%), with a median of 38.5%. (Some people didn't even ascribe 0% to "never" and 100% to "always". In an earlier article in the nursing research literature, Damrosch and Soeken (1983) reported a mean of 45.21% for "possible", a mean of 13.71% for "never" and a mean of 91.35% for "always".) Mosteller and Youtz quote former president Gerald R. Ford as having said that there was "a very real possibility" of a swine flu epidemic in 1976-77. In a previous article, Mosteller (1976) estimated the public meaning of "a very real possibility" to be approximately 29%, and Boffey (1976) had claimed the experts put the probability of a swine flu epidemic in 1976 -77 somewhat lower than that. Shades of concerns about swine flu in 2009!

There is a related matter in weather forecasting. Some meteorologists (e.g., Jeff Haby) have suggested that words be used instead of percentages. In a piece entitled "Using percentages in forecasts" on the weatherprediction.com website, he argues that probabilistic expressions such as "there is a 70% chance of a thunderstorm" should be replaced by verbal expressions such as "thunderstorms will be numerous". (See also the articles by Hallenbeck, 1920, and by Joslyn, et al., 2009, referred to above.) Believe it or not, there is an online program for doing so, put together by Burnham and Schield in 2005. You can get to it at the www.StatLit.org website.

There is also an interesting controversy in the philosophical literature regarding the use of probabilistic words in the analysis of syllogisms, rather than the more usual absolute words such as “All men are mortal; Socrates is a man; therefore, Socrates is mortal”. It started with an article in the Notre Dame Journal of Formal Logic by Peterson (1979), followed by an article by Thompson (1982), followed by an unpublished paper by Peterson and Carnes, followed by another article by Thompson (1986), and ending (I think) with a scathing article by Carnes and Peterson (1991). The controversy revolves around the use of words like “few”, “many”, and “most” in syllogisms. An example given in Thompson’s second article (1986) is:

Almost 27% of M are not P.
Many more than 73% of M are S.
Therefore, some S are not P.

Is that a valid argument? (You decide.)

Chance success

I take an 80-item true-false test and I answer 40 of them correctly. Should I be happy about that? Not really. I could get around 40 (= 50%) without reading the questions, if the number of items for which “true” is the right answer is approximatley equal to the number of items for which “false” is the right answer, no matter what sort of guessing strategy I might employ (all true, all false, every other one true, etc.)

The scoring directions for many objective tests (true-false, multiple-choice, matching, etc.) often recommend that every score on such tests be corrected for chance success. The formula is R – W/(k-1), where R is the number of right answers, W is the number of wrong answers, and k is the number of choices. For the example just given, R = 40, W = 40, k = 2, so that my score would be 40 – 40/(2-1) = 40 – 40 = 0, which is what I deserve!

For more on chance success and the correction for guessing, see Diamond and Evans (1973).

Percentages and probability in the courtroom

As you might expect, probability (in terms of either percentages or fractions) plays an important role in jury trials. One of the most notorious cases was that of the famous professional football player and movie star, O.J. Simpson, who was accused in 1994 of murdering his wife, Nicole, and her friend, Ronald Goldman. There was a great deal of evidence regarding probabilities that was introduced in that trial, e.g., the probability that an individual chosen at random would wear a size 12 shoe AND have blood spots on the left side of his body. (Simpson wears size 12; the police found size 12 footprints nearby, with blood

spots to the left of the footprints. Simpson claimed he cut his finger at home.) For more on this, see the article by Merz and Caulkins (1995); the commentary by John Allen Paulos (1995---yes, that Paulos), who called it a case of "statisticide"; and the letters by defense attorney Alan Dershowitz (1995, 1999). [Simpson was acquitted.]

Several years prior to the Simpson case (in 1964), a Mrs. Juanita Brooks was robbed in Los Angeles by a person whom witnesses identified as a white blonde female with a ponytail, who escaped in a yellow car driven by a black male with a mustache and a beard. Janet and Malcolm Collins, an inter-racial couple who fit those descriptions, were arrested and convicted of the crime, on the basis of estimates of the following probabilities for persons drawn at random:

P(yellow car) = 1/10 = 10%
P(male with mustache) = 1/4 = 25%
P(female with hair in ponytail) = 1/10 = 10%
P(female with blonde hair) = 1/3 = 33 1/3 %
P(black male with beard) = 1/10 = 10%
P(inter-racial couple in car) = 1/1000 = .1%

Product of those probabilities = 1/12,000,000 = .00000833%
[The convictions were overturned because there was no empirical evidence provided for those probabilities and their independence. Oy.]

There was another interesting case, Castenada v. Partida, involving the use of percentages in the courtroom, which was cited in an article by Gastwirth (2005). It concerned whether or not Mexican-Americans were discriminated against in the jury-selection process. (They constituted only 39% of the jurors, although they constituted 79.1% of the relevant population and 65% of the adults in that population who had some schooling.)

<u>My favorite percentages and probability example</u>

Let me end this chapter by citing my favorite example of misunderstanding of probabilities, also taken from Paulos (1988):

> "Later that evening we were watching the news, and the TV weather forecaster announced that there was a 50 percent chance of rain for Saturday and a 50 percent chance for Sunday, and concluded that there was therefore a 100 percent chance of rain that weekend." (p. 3)

I think that says it all.

Chapter 4: Sample percentages vs. population percentages

Almost all research studies that are concerned with percentages are carried out on samples (hopefully random) taken from populations, not on entire populations. It follows that the percentage in the sample might not be the same as the percentage in the population from which the sample is drawn. For example, you might find that in a sample of 50 army recruits 20 of them, or 40%, are Catholics. What percentage of all army recruits is Catholic? 40%? Perhaps, if the sample "mirrors" the population. But it is very difficult for a sample to be perfectly representative of the population from which it is drawn, even if it is randomly drawn.

The matter of sampling error, wherein a sample statistic (such as a sample percentage) may not be equal to the corresponding population parameter (such as a population percentage) is the basic problem to which statistical inference is addressed. If the two are close, the inference from sample to population is strong; if they're not, it's weak. How do you make such inferences? Read on.

Point estimation

A "single-best" estimate of a population percentage is the sample percentage, if the sample has been drawn at random, because the sample percentage has been shown to have some nice statistical properties, the most important of which is that is "unbiased". "Unbiased" means that the average of the percentages for a large number of repeated samples of the same size is equal to the population percentage, and therefore it is a "long-run" property. It does NOT mean that you'll hit the population percentage on the button each time. But you're just as likely to be off "on the high side" as you are to be off "on the low side". How much are you likely to be off? That brings us to the concept of a standard error.

Standard error

A standard error of a statistic is a measure of how off you're likely to be when you use a sample statistic as an estimate of a population parameter. Mathematical statisticians have determined that the standard error of a sample percentage P is equal to the square root of the product of the population percentage and 100 minus the population percentage, divided by the sample size n, if the sample size is large. But you almost never know the population percentage (you're trying to estimate it!). Fortunately, the same mathematical statisticians have shown that the standard error of a sample percentage is APPROXIMATELY equal to the square root of the product of the sample percentage and 100 minus the sample percentage, divided by the sample size n; i.e.,

$$S.E. \approx \sqrt{P(100 - P)/n}$$ [the symbol ≈ means approximately equal to]

For example, if you have a sample percentage of 40 for a sample size of 50, the standard error is $\sqrt{}$ 40(60)/50, which is equal to 6.93 to two decimal places, but let's call it 7. So you would be likely off by about 7% (plus or minus) if you estimate the population percentage to be 40%.

Edgerton (1927) constructed a clever "abac" (mathematical nomogram) for reading off a standard error of a proportion (easily convertible to a percentage), given the sample proportion and the sample size. [Yes, that was 1927… 82 years ago!] It's very nice. There are several other nomograms that are useful in working with statistical inferences for percentages (see, for example, Rosenbaum, 1959). And you can even get a business-card-size chart of the standard errors for various sample sizes at the www.gallup-robinson.com website.

Interval estimation (confidence intervals)

Since it is a bit presumptuous to use just one number as an estimate of a population percentage, particularly if the sample size is small (and 50 is a small sample size for a survey), it is recommended that you provide two numbers within which you believe the population percentage to lie. If you are willing to make a few assumptions, such as sample percentages are normally distributed around population percentages, you should "lay off" two (it's actually 1.96, but call it two) standard errors to the right and left of the sample percentage to get a "95% confidence interval" for the population percentage, i.e., an interval that you are 95% confident will "capture" the unknown population percentage. (Survey researchers usually call two standard errors "the margin of error",) For our example, since the standard error is 7%, two standard errors are 14%, so 40% ± 14%, an interval extending from 26% to 54%, constitutes the 95% confidence interval for the population percentage. 40% is still your "single-best" estimate, but you're willing to entertain the possibility that the population percentage could be as low as 26% and as high as 54%. It could of course be less than 26% or greater than 54%, but you would be pretty confident that it is not.

Since two standard errors = 2 $\sqrt{}$ P (100 - P) / n and P(100 - P) is close to 2500 for values of P near 50, a reasonably good approximation to the margin of error is 100/ $\sqrt{}$n .

I said above that the formula for the standard error of a percentage is a function of the population percentage, but since that is usually unknown (that's what you're trying to estimate) you use the sample percentage instead to get an approximate standard error. That's OK for large samples, and for sample and population percentages that are close to 50. A situation where it very much does matter whether you use the sample percentage or the population percentage in the formula for the standard error is in the safety of clinical trials for which the number of adverse events is very small. For example, suppose no adverse events occurred in a safety trial for a sample of 30 patients. The sample P = 0/30

= 0%. Use of the above formula for the standard error would produce a standard error of 0, i.e., no sampling error! Clearly something is wrong there. You can't use the sample percentage, and the population percentage is unknown, so what can you do? It turns out that you have to ask what is the worst that could happen, given no adverse events in the sample. The answer comes from "The rule of 3" (sort of like "The rule of 72" for interest rates; see Chapter 1). Mathematical statisticians have shown that the upper 95% confidence bound is 3/n in terms of a proportion, or 300/n in terms of a percentage. (See Jovanovic & Levy, 1997, and van Belle, 2002 regarding this intriguing result. The latter source contains all sorts of "rules of thumb", some of which are very nice, but some of the things that are called rules of thumb really aren't, and there are lots of typos.) The lower 95% confidence bound is, of course, 0. So for our example you could be 95% confident that the interval from 0% to 10% (300/30 = 10) "captures" the percentage of adverse events in the population from which the sample has been drawn.

There is nothing special about a 95% confidence interval, other than the fact that it is conventional. If you want to have greater confidence than 95% for a given sample size you have to have a wider interval. If you want to have a narrower confidence interval you can either settle for less confidence or take a larger sample size. [Do you follow that?] But the only way you can be 100% confident of your inference is to have an interval that goes from 0 to 100, i.e., the entire scale!

One reason why many researchers prefer to work with proportions rather than percentages is that when the statistic of interest is itself a percentage it is a bit awkward to talk about a 95% confidence interval for a %. But I don't mind doing that. Do you?

In Chapter 1 I cited an article in the Public Opinion Quarterly by S.S. Wilks (1940a) regarding opinion polling. In a supplementary article in that same issue he provided a clear exposition of confidence intervals for single percentages and for differences between two percentages (see the following chapter for the latter matter). An article two years later by Mosteller and McCarthy (1942) in that journal shed further light on the estimation of population percentages. [I had the personal privilege of "TAing" for both Professor Mosteller and Professor McCarthy when I was doing my doctoral study at Harvard in the late 1950s. Frederick Mosteller was also an exceptional statistician.]

For a very comprehensive article concerning confidence intervals for proportions, see Newcombe (1998a). He actually compared SEVEN different methods for getting confidence intervals for proportions, all of which are equally appropriate for percentages.

Hypothesis testing (significance testing)

Another approach to statistical inference (and until recently far and away the most common approach) is the use of hypothesis testing. In this approach you start out by making a guess about a parameter, collect data for a sample, calculate the appropriate statistic, and then determine whether or not your guess was a good one. Sounds complicated, doesn't it? It is, so let's take an example.

Going back to the army recruits, suppose that before you carried out the survey you had a hunch that about 23% of the recruits would be Catholic. (You read somewhere that 23% of adults in the United States are Catholic, and you expect to find the same % for army recruits.) You therefore hypothesize that the population percentage is equal to 23. Having collected the data for a sample of 50 recruits you find that the percentage Catholic in the sample is 40. Is the 40 "close enough" to the 23 so that you would not feel comfortable in rejecting your hypothesis? Or are the two so discrepant that you can no longer stick with your hypothesis? How do you decide?

Given that "the margin of error" for a percentage is two standard errors and for your data two standard errors is approximately 14%, you can see that the difference of 17% between the hypothesized 23% and the obtained 40% is greater than the margin of error, so your best bet is to reject your hypothesis (it doesn't reconcile with the sample data). Does that mean that you have made the correct decision? Not necessarily. There is still some (admittedly small) chance that you could get 40% Catholics in a sample of 50 recruits when there are actually only 23% Catholics in the total population of army recruits.

We've actually cheated a little in the previous paragraph. Since the population percentage is hypothesized to be 23, the 23 should be used to calculate the standard error rather than the 40. But for most situations it shouldn't matter much whether you use the sample percentage or the hypothesized population percentage to get the standard error. [$\sqrt{40(60)/50} = 6.93$ is fairly close to $\sqrt{23(77)/50} = 5.95$, for example.]

<u>The jargon of hypothesis testing</u>

There are several technical terms associated with hypothesis testing, similar to those associated with diagnostic testing (see the previous chapter):

The hypothesis that is tested is often called a null hypothesis. (Some people think that a null hypothesis has to have zero as the hypothesized value for a parameter. They're just wrong.)

There is sometimes a second hypothesis that is pitted against the null hypothesis (but not for our example). It is called, naturally enough, an alternative hypothesis.

If the null hypothesis is true (you'll not know if it is or not) and you reject it, you are said to have made a Type I error.

If the null hypothesis is false (you'll not know that either) and you fail to reject it, you are said to have made a Type II error.

The probability of making a Type I error is called the level of significance and is given the Greek symbol α.

The probability of making a Type II error doesn't usually have a name, but it is given the Greek symbol β.

1 – β is called the power of the hypothesis test.

Back to our example

Null hypothesis: Population percentage = 23

If the null hypothesis is rejected, the sample finding is said to be "statistically significant". (Hypothesis testing is often called significance testing.) If the null hypothesis is not rejected, the sample finding is said to be "not statistically significant".

Suppose you reject that hypothesis, since the corresponding statistic was 40, but it (the null hypothesis) is actually true. Then you have made a Type I error (rejecting a true null hypothesis).

If you do not reject the null hypothesis and it's false (and "should have been rejected") then you would have made a Type II error (not rejecting a false null hypothesis).

The level of significance, α, should be chosen before the data are collected, since it is the "risk" that one is willing to run of making a Type I error. Sometimes it is not stated beforehand. If the null hypothesis is rejected, the researcher merely reports the probability of getting a sample result that is even more discrepant from the null hypothesis than the one actually obtained if the null hypothesis is true. That probability is called a p value, and is typically reported as $p < .05$ (i.e., 5%), $p < .01$ (i.e., 1%), or $p < .001$ (i.e., .1%) to indicate how unlikely the sample result would be if the null hypothesis is true.

β and/or power (= 1 – β) should also be stated beforehand, but they depend upon the alternative hypothesis, which is often not postulated. [In order to draw the "right" sample size to test a null hypothesis against an alternative hypothesis, the alternative hypothesis must be explicitly stated. Tables and formulas are available (see, for example, Cohen, 1988) for determining the "optimal" sample size for a desired power.]

The connection between interval estimation and hypothesis testing

You might have already figured out that you can do hypothesis testing for a percentage as a special case of interval estimation. It goes like this:

1. Get a confidence interval around the sample percentage.

2. If the hypothesized value for the population percentage is outside that interval, reject it; if it's inside the interval, don't reject it.

[Strictly speaking, you should use the sample percentage to get the standard error in interval estimation and you should use the hypothesized population percentage to get the standard error in hypothesis testing--see above--but let's not worry about that here.]

Neat, huh? Let's consider the army recruits example again. The sample percentage is 40. The 95% confidence interval goes from 26 to 54. The hypothesized value of 23 falls outside that interval. Therefore, reject it. (That doesn't mean it's necessarily false. Remember Type I error!)

It's all a matter of compatibility. The sample percentage of 40 is a piece of real empirical data. You know you got that. What you don't know, but you wish you did, is the population percentage. Percentages of 26 to 54 are compatible with the 40, as indicated by the 95% confidence you have that the interval from 26 to 54 "captures" the population percentage. 23 is just too far away from 40 to be defensible.

van Belle (2002) takes an idiosyncratic approach to interval estimation vs. hypothesis testing. He claims that you should use the hypothesis testing approach in order to determine an appropriate sample size, before the study is carried out; but you should use interval estimation to report the results after the study has been carried out. I disagree. There are the same sorts of sources for the determination of sample size in the context of interval estimation as there are for the determination of sample size in the context of hypothesis testing. (See the reference to Walker & Lev, 1953 in the following section.) In my opinion, if you have a hypothesis to test (especially if you have both a null and an alternative hypothesis), you should use hypothesis-testing procedures for the determination of sample size. If you don't, go the interval estimation route all the way.

Caution: Using interval estimation to do hypothesis testing can be more complicated than doing hypothesis testing directly. I will provide an example of

such a situation in Chapter 5 in conjunction with statistical inferences for relative risks.

Sample size

In all of the foregoing it was tacitly assumed that the size of the sample was fixed and the statistical inference was to be based upon the sample size that you were "stuck with". But suppose that you were interested in using a sample size that would be optimal for carrying out the inference from a sample percentage to the percentage in the population from which the sample had been drawn. There are rather straightforward procedures for so doing. All you need do is to decide beforehand how much confidence you want to have when you get the inferred interval, how much error you can tolerate in making the inference, have a very rough approximation of what the population percentage might be, and use the appropriate formula, table, or internet routine for determining what size sample would satisfy those specifications.

Let's take an example. Suppose you were interested in getting a 95% confidence interval (95% is conventional), you don't want to be off by more than 5%, and you think the population percentage is around 50 (that's when the standard error is largest, so that's the most "conservative" estimate). The formula for the minimum optimal sample size is:

$n \approx 4z^2 P(100-P)/W^2$ [see, for example, Walker and Lev (1953, p. 70)]

where P is your best guess, W is the width of the confidence interval (the width is twice the margin of error), and z is the number of standard errors you need to "lay off" to the right and to the left of the sample P (z comes from the normal, bell-shaped sampling distribution). Substituting the appropriate values in that formula (z is approximately equal to 2 for 95% confidence) you find that n is equal to $4(2)^2 50(100-50)/10^2 = 400$. If you draw a sample of less than 400 you will have less than 95% confidence when you get the sample P and construct the interval. If you want more confidence than 95% you'll need to lay off more standard errors and have a larger n (for three standard errors you'll need an n of about 900). If you want to stay with 95% confidence but you can tolerate more error (say 10% rather than 5%, so that W = 20), then you could get away with an n of about 100.

The Dimension Research, Inc. website actually does all of the calculations for you. Just google "dimension research calculator", click on the first entry that comes up, and click on "sample size for proportion" on the left-hand-side menu . Then select a confidence interval, enter your "best-guess" P and your tolerable ½ W, click the Calculate button, and Shazam! You've got n.

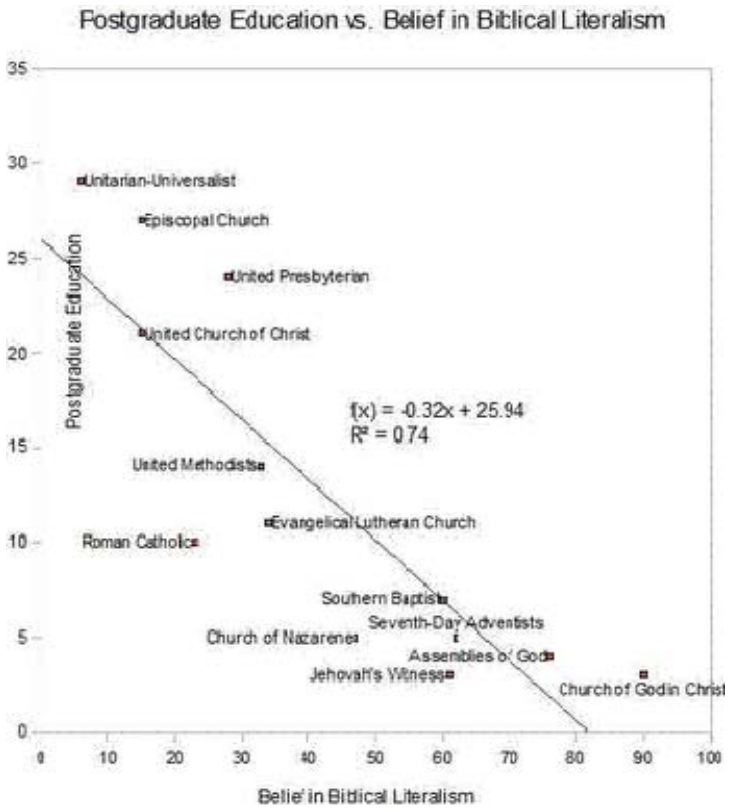
van Belle (2002) claims that you should have a sample size of at least 12 when you construct a confidence interval. He provides a diagram that indicates the precision of an interval is very poor up to an n of 12 but starts to level off thereafter.

Percentage transformations

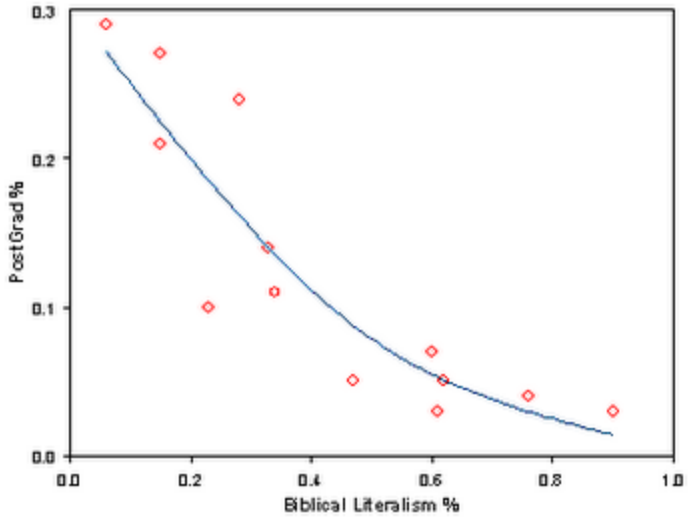
One of the problems when carrying out statistical inferences for percentages is the fact that percentages are necessarily "boxed in" between 0 and 100, and often have rather strange distributions across aggregates for which they have been computed. There can't be less than 0% and there can't be more than 100%, so if most of the observations are at the high end of the scale (large percentages) or at the low end of the scale (small percentages) it is almost impossible to satisfy the linearity and normal distribution assumptions that are required for many inferential tests.

Consider the following example taken from the Ecstathy website:

You have data regarding % Postgraduate Education and % Belief in Biblical Literalism for members of 13 religious denominations (Unitarian-Universalist, Episcopal Church, United Presbyterian, United Church of Christ, United Methodist, Evangelical Lutheran Church, Roman Catholic, Southern Baptist, Seventh Day Adventist, Church of Nazarene, Assemblies of God, Jehovah's Witness, and Church of God in Christ), and you're interested in the relationship between those two variables. You plot the data as depicted in the following scatter diagram (which includes the "best-fitting" line and the regression statistics:

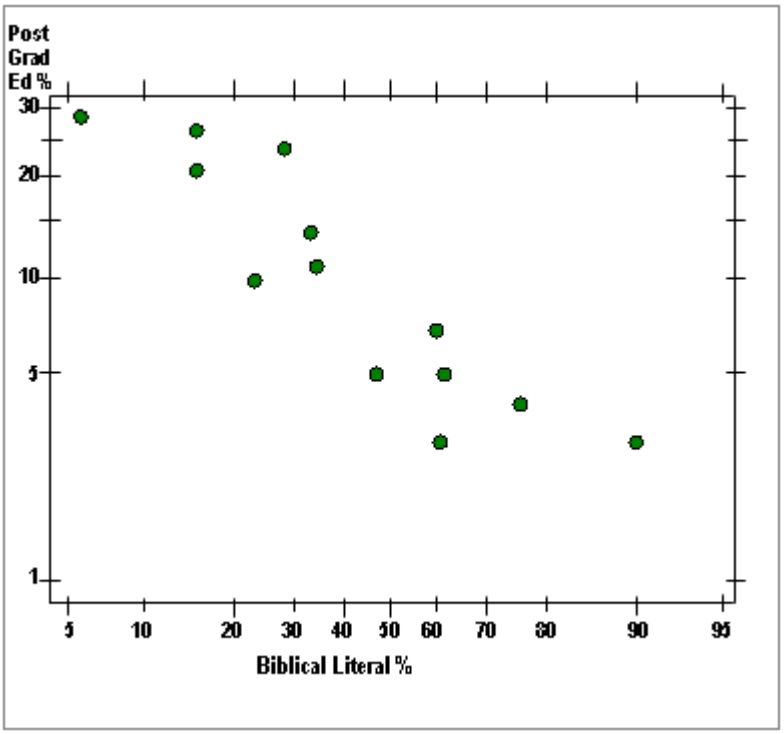


Here is the plot without the names of the religions superimposed (and with proportions rather than percentages, but that doesn't matter):



You would like to use Pearson's product-moment correlation coefficient to summarize the relationship and to make an inference regarding the relationship in the population of religious denominations from which those 13 have been drawn (assume that the sample is a simple random sample, which it undoubtedly

was not!). But you observe that the plot without the names is not linear (it is curvilear) and the assumption of bivariate normality in the population is also not likely to be satisfied. What are you to do? The recommendation made by the bloggers at the website is to transform both sets of percentages into logits (which are special types of logarithmic transformations), plot the logits, and carry out the analysis in terms of the logits of the percentages rather than in terms of the percentages themselves. It works; here's the plot (this one looks pretty linear to me):



There are transformations of percentages other than logits that have been recommended in the methodological literature--see, for example, the articles by Zubin (1935), by Finney (1947; 1975), and by Osborne (2002). Zubin even provided a handy-dandy table for converting a percentage into something he called t or T (not the t of the well-known t test, and not the T of T scores). Nice.

<u>The classic case of inferences regarding single percentages</u>

You manufacture widgets to be sold to customers. You worry that some of the widgets might be defective, i.e., you are concerned about "quality control". What should you do? If the widgets are very small objects (such as thumbtacks) that are made by the thousands in an assembly-line process, the one thing you can't afford to do is inspect each and every one of them before shipping them out. But you can use a technique that's called acceptance sampling, whereby you take a random sample of, say, 120 out of 2000 of them, inspect all of the widgets in the

sample, determine the percentage of defectives in the sample, and make a judgment regarding whether or not that percentage is "acceptable".

For example, suppose you claim (hope?) that your customers won't complain if there are 2% (= 40) or fewer defectives in the "lot" of 2000 widgets that they buy. You find there are 3 defectives (1.67%) in the sample. Should you automatically accept the lot (the population) from which the sample has been drawn? Not necessarily. There is some probability that the lot of 2000 has more than 2% defectives even though the sample has ony 1.67%. This is the same problem that was discussed in a different context (see above) regarding the percentage of army recruits that is Catholic. Once again, you have three choices: (1) get a point estimate and use it (1.67%) as your single-best estimate; (2) establish a confidence interval around that estimate and see whether or not that interval "captures" the tolerable 2%; or (3) directly test the 2% as a null hypothesis.

There is an excellent summary of acceptance sampling available at myphliputil.pearsoncmg.com/student/bp_heizer...7/ct02.pdf. For the problem just considered, it turns out that the probability of acceptance is approximately .80 (i.e., an 80% probability). I used the same numbers that they do, their "widgets" are batteries, and they take into account the risk to the customer (consumer) as well as the risk to the manufacturer (producer).

A great website for inferences regarding percentages in general

The West Chester University website has an excellent collection of discussions of statistical topics. Although that website is intended primarily for students who are taking college courses online, any interested parties can download any of the various sections. Section 7_3 is concerned with the finite population correction that should be used for inferences regarding percentages for samples drawn from "small", i.e., finite, populations. See also Krejcie & Morgan, 1970; Buonaccorsi, 1987; and Berry, Mielke, and Helmericks, 1988 for such inferences. vanBelle (2002) argues that the correction can usually be ignored.) The website's name is:
http://courses.wcupa.edu/rbove/Berenson/CD-ROM%20Topics/Section 7_3

Chapter 5: Statistical inferences for differences between percentages and ratios of percentages

In the previous chapter I talked about statistical inferences for a single percentage. Such inferences are fairly common for survey research but not for other kinds of research, e.g., experimental research in which two or more "treatments" are compared with one another. The inference of greatest interest in experimental research is for the difference between two statistics or the ratio of two statistics, e.g., the percentage of people in "the experimental group" who do (or get) something and the percentage of people in "the control group" who do (or get) something. The "something" is typically a desirable outcome such as "passed the course" or an undesirable outcome such as "died".

Differences between percentages

Just as for a single percentage, we have our choice of point estimation, interval estimation, or hypothesis testing. The relevant point estimate is the difference between the two sample percentages. Since they are percentages, their difference is a percentage. If one of the percentages is the % in an experimental group that "survived" (they got the pill, for example), and the other percentage is the % in a control group that "survived" (they didn't get the pill), then the difference between the two percentages gives you an estimate of the "absolute effect" of the experimental condition. If 40% of experimental subjects survive and 30% of control subjects survive, the estimate of the experimental effect is 10%.

But just as for a single percentage, it is better to report two numbers rather than one number for an estimate, i.e., the endpoints of a confidence interval around the difference between the two percentages. That necessitates the calculation of the standard error of the difference between two percentages, which is more complicated than for the standard error of a single percentage. The formula for two independent samples ("unmatched") and the formula for two dependent samples (matched by virtue of being the same people or matched pairs of people) are different.

The independent samples case is more common. The formula for the standard error of the difference between two independent percentages is:

$$S.E. \approx \sqrt{\ P_1(100 - P_1)/n_1 \ + \ P_2(100 - P_2)/n_2}$$

where the P's are the two percentages and the n's are the two sample sizes. It often helps to display the relevant data in a "2 by 2" table:

	Sample 1	Sample 2
"Success"	P_1	P_2
"Failure"	$100 - P_1$	$100 - P_2$

where "Success" and "Failure" are the two categories of the variable upon which the percentages are taken, and need not be pejorative.

Edgerton's (1927) abac can be used to read off the standard error of the difference between two independent percentages, as well as the standard error of a single sample percentage. Later, Hart (1949) and Lawshe and Baker (1950) presented quick ways to test the significance of the difference between two independent percentages. Stuart (1963) provided a very nice set of tables of standard errors of percentages for differences between two independent samples for various sample sizes, which can also be used for the single-sample case. And Fleiss, Levin, and Paik (2003) provide all of the formulas you'll ever need for inferences regarding the difference between percentages. They even have a set of tables (pp. 660-683) for determining the appropriate sample sizes for testing the significance of the difference between two percentages. (See also Hopkins & Chappell, 1994.)

The formula for dependent samples is a real mess, involving not only the sample percentages and the sample sizes but also the correlation between the two sets of data (since the percentages are for the same people or for matched people). However, McNemar (1947) provided a rather simple formula that is a reasonable approximation to the more complicated one:

$$S.E. \approx 100/n \sqrt{(b + c)}$$

where n (= n_1 = n_2 , since the people are paired with themselves or with their "partners"), b is the number of pairs for which the person in Sample 1 was a "success" and the partner in Sample 2 was a "failure"; and c is the number of pairs for which the person in Sample 1 was a "failure" and the partner in Sample 2 was a "success".

Sample 1	Sample 2 "Success"	Sample 2 "Failure"	
"Success"	[a]	[b]	$P_1 = (a+b)/n$
"Failure"	[c]	[d]	
	$P_2 = (a+c)/n$		

a is the number of pairs for which both members were “successes”, and
d is the number of pairs for which both members were “failures”; but, rather surprisingly, neither a nor d contributes to the standard error.

If a researcher is concerned with change in a given sample from Time 1 to Time 2, that also calls for the dependent-samples formula.

An example to illustrate both the independent and the dependent cases

You are interested in the percentage of people who pass examinations in epidemiology. Suppose there are two independent samples of 50 students each (50 males and 50 females) drawn from the same population of graduate students, where both samples take an epidemiology examination. The number of males who pass the examination is 40 and the number of females who pass is 45.

Displaying the data as suggested above we have:

	Males	Females
Passed	40/50 = 80%	45/50 = 90%
Failed	10/50 = 20%	5/50 = 10%

The standard error of the difference between the two percentages is

S.E. ≈ $\sqrt{\ }$ 80(20)/50 + 90(10)/50 = 7.07 (rounded to two places)

On the other hand, suppose that these students consist of 50 married couples who take the same course, have studied together (within pairs, not between pairs), and take the same epidemiology examination. Those samples would be dependent. If in 74% of the couples both husband and wife passed, in 6% of the couples wife passed but husband failed, in 16% of the couples husband passed but wife failed, and in 2 couples both spouses failed, we would have the following “2 by 2” table:

Husband	Wife: Passed	Wife: Failed	
Passed	37 [a]	3 [b]	40 (= 80%)
Failed	8 [c]	2 [d]	10
	45 (= 90%)		

S.E. ≈ 100/50√ (3 + 8) = 6.63

In both cases 80% of the males passed and 90% of the females passed, but the standard error is smaller for matched pairs since the data for husbands and wives are positively correlated and the sampling error is smaller. If the correlation between paired outcomes is not very high, say less than .50 (van Belle, 2002) the pairing of the data is not very sensitive. If the correlation should happen to be NEGATIVE, the sampling error could actually be WORSE for dependent samples than for independent samples!

Would you believe that there is also a procedure for estimating the standard error of the difference between two "partially independent, partially dependent" percentages? In the husbands and wives example, for instance, suppose there are some couples for which you have only husband data and there are some couples for which you have only wife data. Choi and Stablein (1982) and Thompson (1995) explain how to carry out statistical inferences for such situations.

Interval estimation for the difference between two independent percentages

As far as the interval estimation of the difference between two independent population percentages is concerned, we proceed just as we did for a single population percentage, viz., "laying off" two S.E.'s to the right and to the left of the difference between the two sample percentages in order to get a 95% confidence interval for the difference between the two corresponding population percentages.

The sample difference is 90% - 80 % = 10% for our example. The standard error for the independent case is 7.07%. Two standard errors would be 14.14%. The 95% confidence interval for the population difference would therefore extend from 10% - 14.14% to 10% + 14.14%, i.e., from -4.14% to 24.14%. You would be 95% confident that the interval would "capture" the difference between the two population percentages. [Note that the -4.14% is a difference, not an actual %.] Since Sample 2 is the wives sample and Sample 1 is the husbands sample, and we subtracted the husband % from the wife %, we are willing to believe that in the respective populations the difference could be anywhere between 4.14% "in favor of" the husbands and 24.14% "in favor of" the wives.

I refer you to an article by Wilks (1940b) for one of the very best discussions of confidence intervals for the difference between two independent percentages. And in the previous chapter I mentioned an article by Newcombe (1998) in which he compared seven methods for determining a confidence interval for a single proportion. He followed that article with another article (Newcombe, 1998b) in which he compared ELEVEN methods for determining a confidence interval for the difference between two independent proportions!

Hypothesis testing for the difference between two independent percentages

In the previous chapter I pointed out that except for a couple of technical details, interval estimation subsumes hypothesis testing, i.e., the confidence interval consists of all of the hypothesized values of a parameter that are "not rejectable". For our example any hypothesis concerning a population difference of -4.14 through 24.14 would not be rejected (and would be regarded as "not statistically significant at the 5% level"). Any hypotheses concerning a population difference that is outside of that range would be rejected (and would be regarded as "statistically significant at the 5% level").

In a very interesting article concerning the statistical significance of the difference between two independent percentages (he uses proportions), the late and ever-controversial Alvan R. Feinstein (1990) proposed the use of a "unit fragility index" in conjunction with the significance test. This index provides an indication of the effect of a "switch" of an observation from one category of the dependent variable to the other category (his illustrative example had to do with a comparison of cephaloridine with ampicillin in a randomized clinical trial). That index is especially helpful in interpreting the results of a trial in which the sample is small. (See also the commentary by Walter, 1991 regarding Feinstein's index.)

Feinstein was well-known for his invention of methodological terminology. My favorite of his terms is "trohoc" [that's "cohort" spelled backwards] instead of "case-control study". He didn't like case-control studies, in which "cases" who have a disease are retrospectively compared with "controls" who don't, in an observational non-experiment.

There is an advantage of interval estimation over hypothesis testing that I've never seen discussed in the methodological literature. Researchers often find it difficult to hypothesize the actual magnitude of a difference that they claim to be true in the population (and is not "null"). The theory underlying their work is often not far enough advanced to suggest what the effect might be. They are nevertheless eager to know its approximate magnitude. Therefore, instead of pitting their research (alternative) hypothesis against a null hypothesis and using power analysis for determining the appropriate sample size for testing the effect, all they need to do is to specify the magnitude of a tolerable width of a confidence interval (for a margin of error of, say, 3%), use that as the basis for the determination of sample size (see the appropriate formula in Fleiss, et al., 2003), carry out the study, and report the confidence interval. Nice; straightforward; no need to provide granting agencies with weak theories; and no embarrassment that often accompanies hurriedly-postulated effects that far exceed those actually obtained.

Two examples of lots of differences between percentages

Peterson, et al. (2009) were interested in testing the effectiveness of a particular invention designed to help teenagers to stop smoking. Using a rather elaborate design that had a tricky unit-of analysis problem (schools containing teenage smokers were randomly assigned to the experimental treatment and to the control treatment, rather than individual students). Their article is loaded with both confidence intervals for, and significance tests of, the difference between two percentages.

Sarna, et al. (2009) were also interested in stopping smoking, but for nurses rather than teenagers. Like Peterson, et al., their article contains several tables of confidence intervals and significance tests for the differences between percentages. But it is about a survey, not an experiment, in which nurses who quit smoking were compared to nurses who did not quit smoking, even though all of them registered at the Nurses QuitNet website for help in trying to do so.

If you're interested in smoking cessation, please read both of those articles and let me know (tknapp5@juno.com) what you think of them.

The difference between two percentages that have to add to 100

In Chapter 2 I said that I don't care much for the practice of taking differences between two percentages that have been calculated on the same base for the same variable, e.g., the difference in support for Candidate A and Candidate B for the same political office. I am even more opposed to making any statistical inferences for such differences. If you care about that sort of thing, I refer you to Richard Lowry's fine VassarStats website.

Ratios of percentages

Now for the "biggie" in epidemiological research. We've already discussed the difference between absolute risk, as represented by the difference between two percentages, and relative risk, as represented by the ratio of two percentages. Relative risk tends to be of greater importance in epidemiology, since the emphasis is on risks for large populations of people having one characteristic compared to risks for equally large populations of people having a contrasting characteristic.

The classic example is smokers vs. non-smokers and the relative risk of getting lung cancer. But let's take as a simpler example the relationship between maternal age and birthweight. Fleiss, et al. (2003) provide a set of hypothetical data for that problem. Here are the data:

	Maternal age	
Birthweight	≤ 20 years	> 20 years
≤ 2500 grams	10	15
> 2500 grams	40	135

The ratio of interest is the percentage of younger women whose baby is of low birthweight (10/50, or 20%) divided by the percentage of older women whose baby is of low birthweight (15/150, or 10%). The relative risk of low birthweight is therefore 20%/10%, or 2.00. If these data are for a random sample of 200 women, what is the 95% confidence interval for the relative risk in the population from which the sample has been drawn? Is the relative risk of 2.00 statistically significant at the 5% level? Although the first question is concerned with interval estimation and the second question is concerned with hypothesis testing, the two questions are essentially the same, as we have already seen several times.

I shall give only a brief outline of the procedures for answering those questions. The determination of an estimate of the standard error of the ratio of two percentages is a bit complicated, but here it is (Fleiss, et al., 2003, p. 132):

S.E. $\approx$ r $\sqrt{}$ (n_{12} / n_{11} $n_{1.}$ + n_{22} / n_{21} $n_{2.}$), where

n_{11} is the number in the upper-left corner of the table (10 in the example)

n_{12} is the number in the upper-right corner (40)

n_{21} is the number in the lower-left corner (15)

n_{22} is the number in the lower-right corner (135)

$n_{1.}$ is the total for the first row (50)

$n_{2.}$ is the total for the second row (150)

Substituting those numbers in the formula for the standard error, we get

S.E. = .75 (to two decimal places)

Two standard errors would be approximately 1.50, so the 95% confidence interval for the population ratio would be from .50 to 3.50. Since that interval includes 1 (a relative risk of 1 is the same risk for both groups), the obtained sample ratio of 2.00 is not statistically significant at the 5% level.

Fleiss, et al. (2003) actually recommend that the above formula for estimating the standard error not be used to get a confidence interval for a ratio of two percentages. They suggest instead that the researcher use the "odds ratio" instead of the relative risk (the odds ratio for those data is 2.25), take the logarithm of the odds ratio, and report the confidence interval in terms of "log odds". [Here we go with logarithms again!] I don't think that is necessary, since everything is approximate anyhow. If those data were real, the principal finding is that younger mothers do not have too much greater risk for having babies of low birthweight than do older mothers. Fleiss et al. arrive at the same conclusion by using the logarithmic approach.

Another "hybrid" inferential problem

Earlier in this chapter I referred to procedures derived by Choi and Stablein (1982) and by Thompson (1995) for estimating the standard error of the difference between two percentages where the samples were partially independent and partially dependent, due to missing data. There is another interesting situation that comes up occasionally where you would like to test the difference between two independent percentage gains, i.e., where each gain is the difference between two dependent percentages. (A loss is treated as a negative gain.) Building upon the work of Marascuilo and Serlin (1979) [see also Levin & Serlin, 2000], Howell (2008) discussed a hypothetical example where a change from fall (42/70 = 60%) to spring (45/70 = 64.3%) for an intervention group is compared with change from fall (38/70 = 54.3%) to spring (39/70 = 55.7%) for a control group. The difference between the 4.3% gain for the intervention group and the 1.4% gain for the control group was not statistically significant, which is not surprising since the "swing" is only about 3%. Those of you who are familiar with the classic monograph on experimental design by Campbell and Stanley (1966) might recognize Howell's example as a special case of Campbell and Stanley's True Experimental Design #4, i.e., the Pretest/Posttest Control Group Design. (See also the article by Vickers, 2001 in which he discusses four different ways for analyzing the data for such a design.)

Sample size

In the previous chapter I talked about a handy-dandy internet calculator that determined the optimal sample size for a confidence interval for a single percentage. The situation for determining the optimal sample sizes for confidence intervals for the difference between two percentages or the ratio of two percentages (for either independent samples or for dependent samples) is much more complicated. (See Fleiss, et al., 2003, for all of the gory details. And the PASS2008 software is particularly good for carrying out all of the calculations for you [it is available for a 7-day free trial].)

Non-random samples and full populations

Suppose you have a non-random sample of boys and a non-random sample of girls from a particular school and you want to compare the percentage of boys in the boy sample who think that President Obama is doing a good job with the percentage of girls in the girl sample who think that President Obama is doing a good job. Would a confidence interval or a significance test of the difference between, or the ratio of, the two percentages be appropriate? Suppose you have percentages for the entire population of boys and the entire population of girls? Would a confidence interval or a significance test be appropriate there?

You can't imagine how controversial both of those matters are! The opinions range from "very conservative" to "very liberal". The very conservative people argue that statistical inferences are appropriate only for probability samples, of which "simple" random samples are the most common type (everybody has an equal and independent chance of being drawn into the sample) and not for either non-random samples or entire populations. Period. End of discussion. The very liberal people argue that they are appropriate for both non-random samples and for entire populations, since they provide an objective basis for determining whether or not, or to what extent, to get excited about a finding. The people in-between (which from a cursory glance at the scientific literature are the majority) argue that for a non-random sample it is appropriate to use statistical inferential procedures in order to generalize from the non-random sample to a hypothetical population of people "like these"; and/or it might be appropriate to use statistical inferential procedures for an entire population in order to generalize from a finding now to findings for that population at other times. As one of those "very conservative people" (we meet in a telephone booth every year), those last two arguments blow my mind. I don't care about hypothetical populations (do you?) and hardly anybody studies populations by sampling them across time.

In his article, Desbiens (2007) did a review of the literature and found that many authors of research reports in medical education journals use statistical inferences for entire populations. He claims that they shouldn't. I agree.

More than two percentages

The previous discussion was concerned with procedures for statistical inferences when comparing the difference between, or the ratio of, two percentages. It is natural to ask if these procedures generalize to three or more percentages. The answer is "sort of".

If you're interested in testing the significance of the difference AMONG several percentages (e.g., the percentage of Catholics who voted for Obama, the percentage of Protestants who voted for Obama, and the percentage of Jews who voted for Obama), there are comparable (and more complicated) formulas for so doing (see Fleiss, et al., 2003). Confidence intervals for the more-than-

two case, however, are much more awkward to handle, primarily because there are three differences (A-B, A-C, B-C) to take into consideration. [There might also be those same three differences to take into consideration when carrying out the significance testing, if you care about pairwise differences as well as the overall difference. It's just like the problem of an overall F test vs. post hoc comparisons in the analysis of variance, if that means anything to you!]

The situation for ratios is even worse. There is no appropriate statistic for handling A/B/C, for example, either via significance testing or confidence intervals.

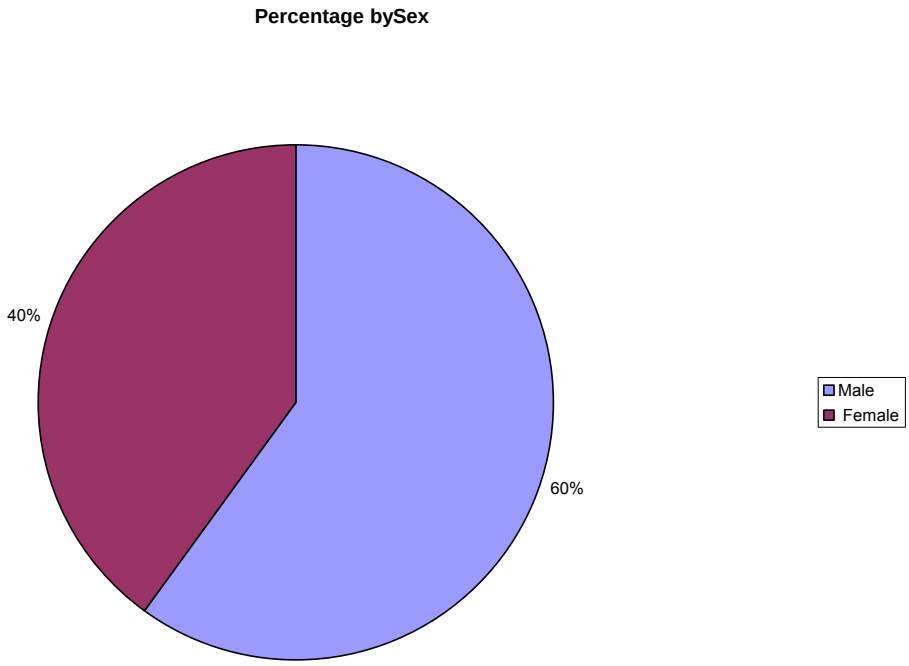
Chapter 6: Graphing percentages

I've never cared much for statistical graphics, except for scatter diagrams that facilitate the understanding of the form and the degree of the relationship between two variables. (See the scatter diagrams that I used in Chapter 4 to illustrate data transformations for percentages.) I also don't always agree with the often-stated claim that "a picture is worth a thousand words". (I like words.) But I realize that there are some people who prefer graphs to words and tables, even when it comes to percentages. I therefore decided to include in this book a brief chapter on how to display percentages properly when graphical techniques are used. You may want to adjust your "zoom" view for some of these graphs, in order to get a better idea of the information contained therein.

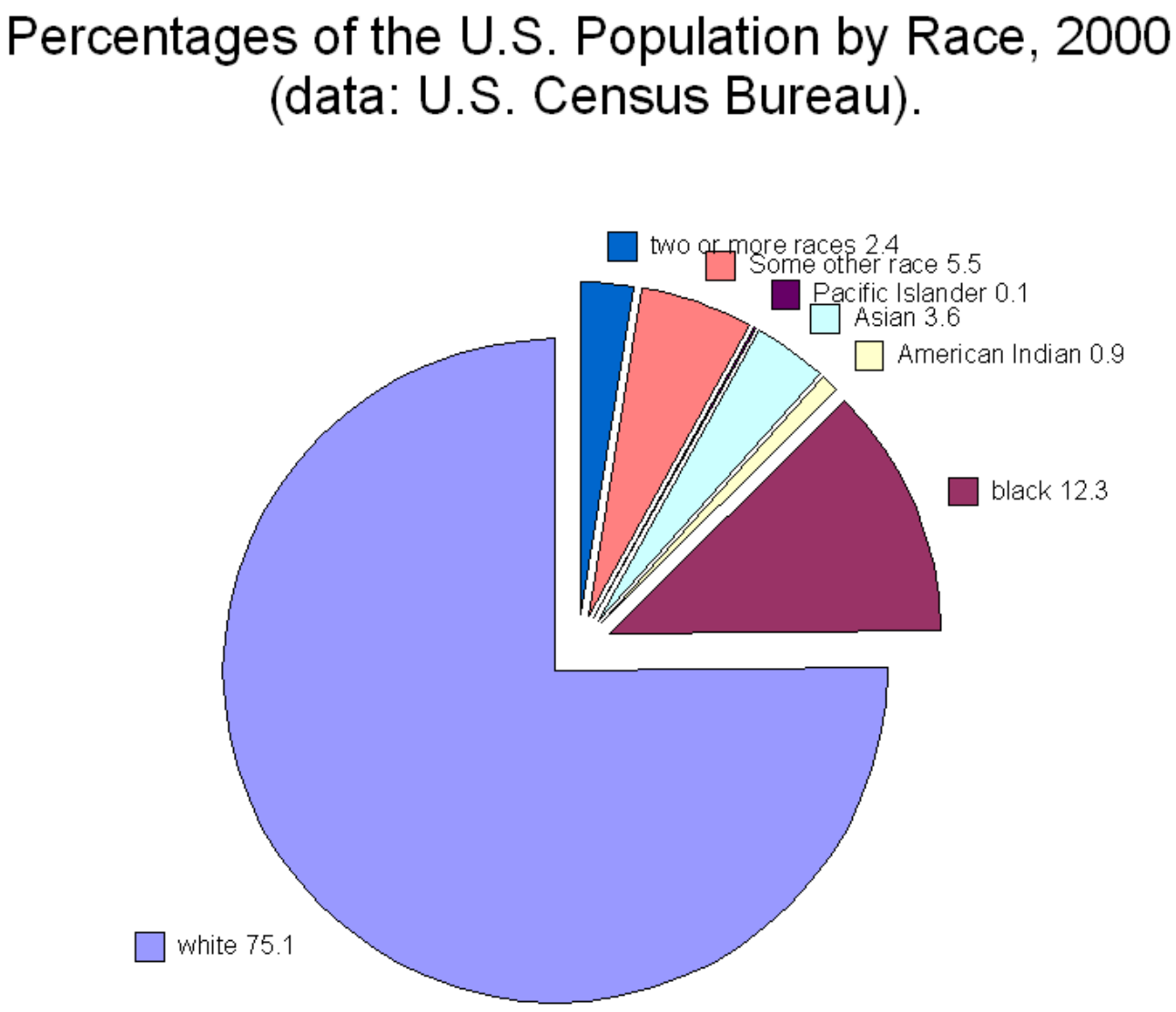
Pie charts

Far and away the most common way to show percentages is the use of pie charts, with or without colors. For example, if one of the findings of a survey is that 60% of cigarette smokers are males and 40% of cigarette smokers are females, that result could be displayed by using a "pie" (circle) divided into two "slices", a blue slice constituting 60% of the pie (216 of the 360 degrees in the circle) labeled MALES, and a red slice constituting the other 40% of the pie (the other 144 degrees) labeled FEMALES. There is absolutely nothing wrong with such charts, but I think they're unnecessary for summarizing two numbers (60 and 40)---actually only one number (60 or 40)---since the other follows automatically. If the variable has more than two categories, pie charts are somewhat more defensible for displaying percentages, but if the number of categories is too large it is difficult to see where one slice ends and another slice begins.

The software EXCEL that is part of Microsoft Office has the capability of constructing pie charts (as well as many other kinds of charts and graphs), and it is fairly easy to "copy and paste" pie charts into other documents. Here's one for the 60% male, 40% female example.



Here's another, and more complicated, pie chart that illustrates one way to handle "small slices". The data are for the year 2000.

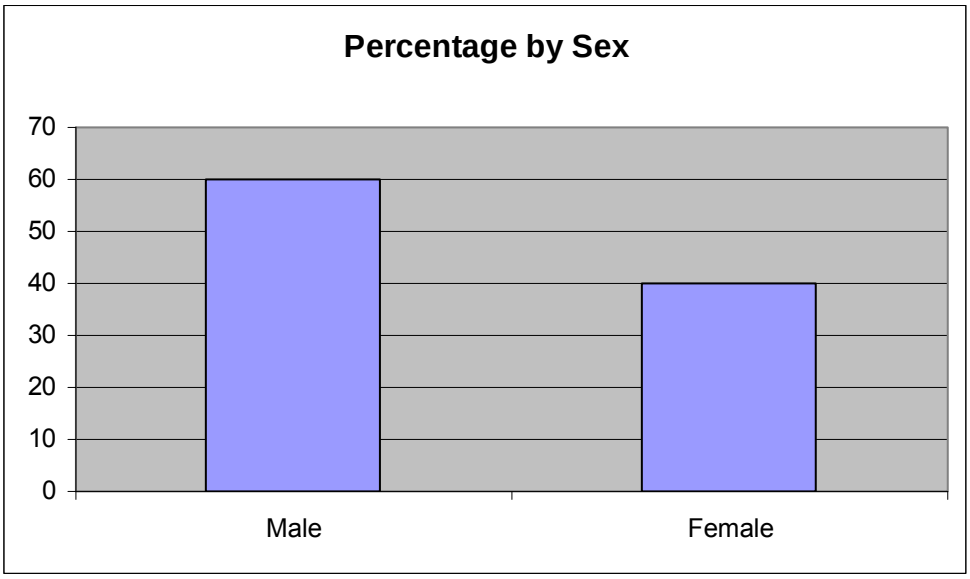


Some people are adamantly opposed to the use of pie charts for displaying percentages (van Belle, 2002, p. 160, for example, says "Never use a pie chart"), but Spence and Lewandowsky (1991) supported their use. They even provided data from experiments that showed that pie charts aren't nearly as bad as the critics claim.

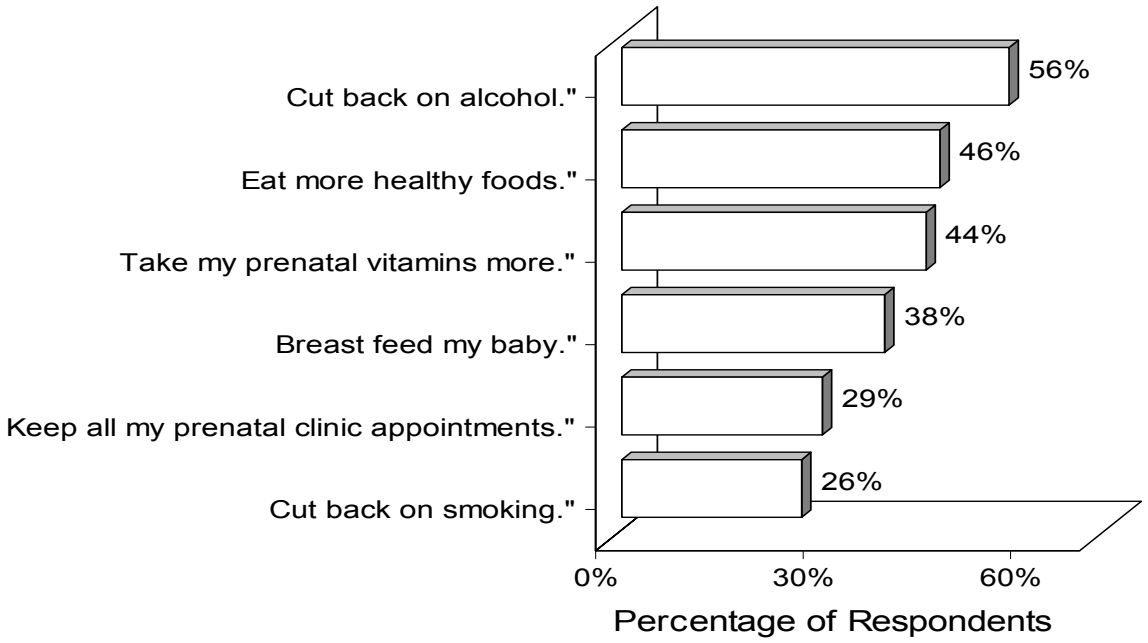
Bar graphs

Bar graphs are probably the second most common way to display percentages. (But van Belle, 2002, doesn't like them either.) The categories of the variable are usually indicated on the horizontal (X) axis and the percentage scale usually constitutes the vertical (Y) axis, with bars above each of the categories on the X

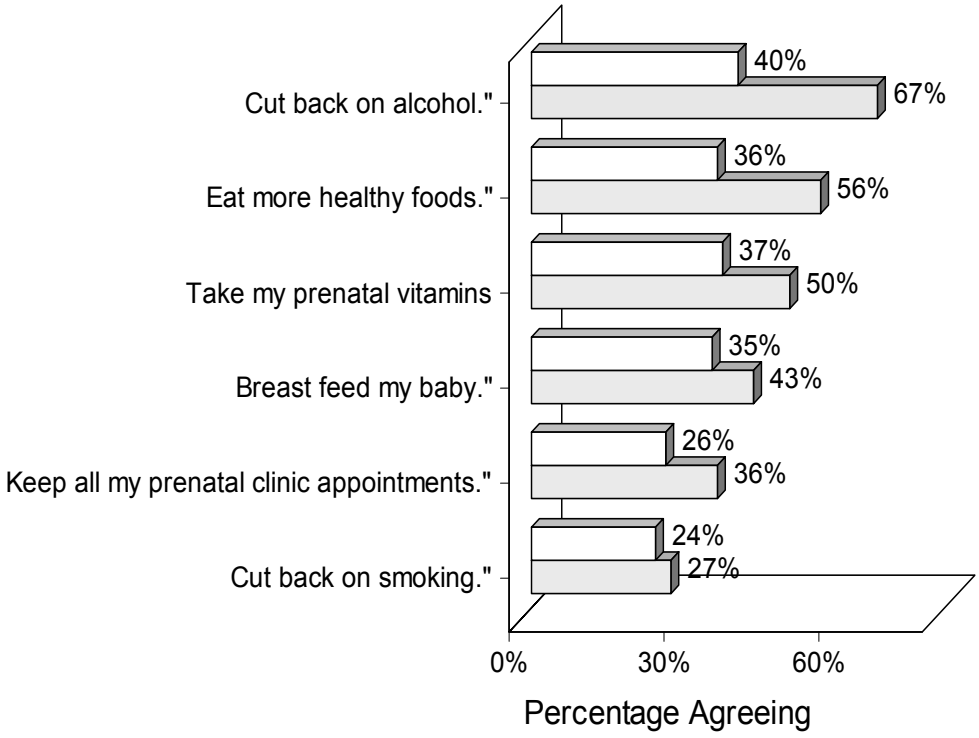
axis extending to a height corresponding to the relevant percentage on the Y axis. The categories need not be in any particular order on the X axis, if the variable is a nominal variable such as Religious Affiliation. But if the variable is an ordinal variable such as Socio-economic Status, the categories should be ordered from left to right on the X axis in increasing order of magnitude. Here's the 60%, 40% example as a bar graph:



Here's a bar graph for more than two categories. The data are percentages of responses by pregnant mothers to the question "Reading the [Preparing to Parent] newsletters helped convince me to...". Note that the bars are horizontal rather than vertical and the percentages do not add to 100 because more than one response is permitted.



Here's a more complicated (but readable) bar graph for the "breakdown" of responses of two groups of pregnant women (those at risk for complications and those not at risk) in that same study:



For other examples of the use of bar graphs, see Keppel et al. (2008).

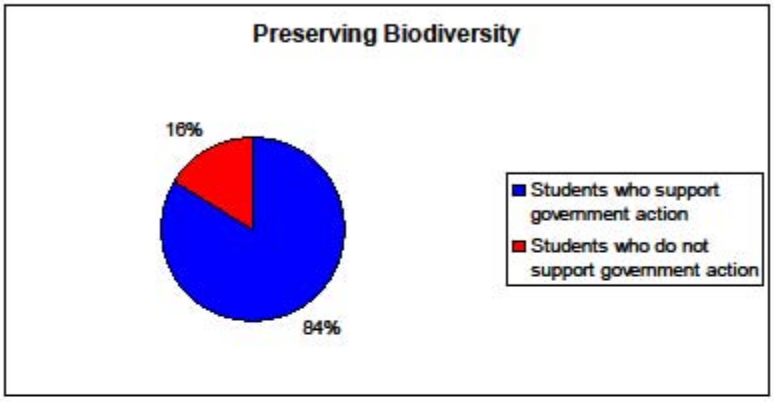
One of the least helpful percentage bar graphs I've ever seen can be downloaded from the StateMaster.com website. It is concerned with the percent of current smokers in each of 49 states, as of the year 2004. It lists those percents in decreasing order (from 27.5% for Kentucky to 10.4% for Utah; it also lists the District of Columbia, Puerto Rico, and the U.S. Virgin Islands, but not my home state of Hawaii!). Each percent is rounded to one place to the right of the decimal point, and there is a bar of corresponding horizontal length right next to each of those percents. It is unhelpful because (a) the bars aren't really needed (the list of percents is sufficient); and (b) rounding the percents to one decimal place resulted unnecessarily in several ties (since the number of current smokers in each of the states and the population of each state are known or easily estimable, all of those ties could have been broken by carrying out the calculations to two decimal places rather than one).

A research example that used both a pie chart and a bar graph

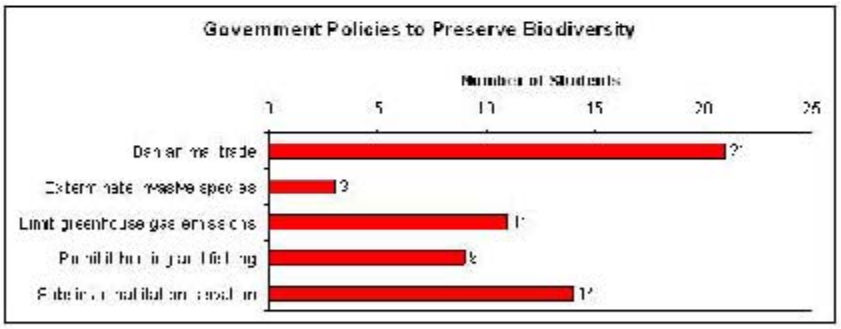
On its website, the Intel©Technology Initiative provides the following example of the use of a pie chart and a bar graph for displaying counts and percentages obtained in a survey regarding attitudes toward biodiversity. (Note that the bars in the bar graph are horizontal rather than vertical. It doesn't really matter.)

Reporting Percentages

Charts can display survey data. The following chart shows that 21 out of 25 students, or 84 percent of students, support government action to preserve biodiversity. A pie chart is a good way to show percentages.



You could also report percentages for each of the five policies. The following bar graph shows how many students reported supporting each policy. You can use the numbers to compute percentages. A bar graph is a good way to report the number of items in selected categories.



For example, 44 percent (=11/25*100) of the students supported regulating greenhouse gas emissions. In comparison, the policy most frequently supported by students was tax breaks for electric and hybrid cars. Overall, 84 percent (=21/25*100) of the students reported that they support such tax breaks. The least frequently supported policy was building more nuclear power plants. Only 12 percent (3/21*100) of the students supported more nuclear power.

In their article, Spence and Lewandowsky (1991) reported results that indicated that bar graphs and pie charts were equally effective in displaying the key features of percentage data. They provide various bar graphs for displaying four percentages (A = 10%; B = 20%; C = 40%; D = 30%). Nice.

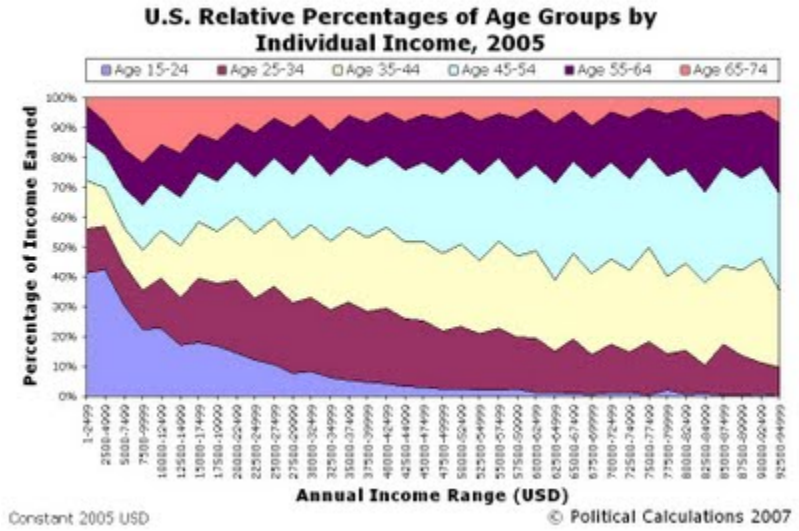
In Chapter 2 I referred to Milo Schield and the W.M. Keck Statistical Literacy Project at Augsburg College. In a presentation he gave to the Section on Statistical Education of the American Statistical Association, Schield (2006) gave a critique of pie and bar percentage graphs that appeared in the newspaper USA Today. I've seen many of those graphs; some are really bad.

Other graphical techniques for displaying percentages

Kastellec and Leoni (2007) provided several arguments and a great deal of evidence supporting the use of graphs to improve the presentation of findings in political science research. In their article they include real-data examples for which they have converted tables into graphs. Some of those examples deal with percentages or proportions presented through the use of mosaic plots, dot plots, advanced dot plots, or violin plots (those are their actual names!). Rather than trying to explain those techniques here, I suggest that you read the Kastellec and Leoni article and see for yourself. (They're not just applicable to political science.) Their article also has an extensive set of references pro and con the use of graphs.

If you skipped Chapter 1 and you're having difficulty distinguishing among percentages, proportions, and fractions, I suggest that you take a look at the British website www.active-maths.co.uk/.../fracdec_index.html, which lays out nicely how each relates to the others.

And here's another example of the graphing of percentages (taken from the Political Calculations website):



That graph contains a lot of interesting information (e.g., that the percentage of people aged 65-74 who have very high incomes is almost as high as the percentage of people aged 25-34--read along the right-hand edge of the graph), but I personally find it to be too “busy”, and it looks like Jaws!

Chapter 7: Percentage overlap of two frequency distributions

One of the things that has concerned me most about statistical analysis over the years is the failure by some researchers to distinguish between random sampling and random assignment when analyzing data for the difference between two groups. Whether they are comparing a randomly sampled group of men with a randomly sampled group of women, or a randomly assigned sample of experimental subjects with a randomly assigned sample of control subjects (or, worse yet, two groups that have been neither randomly sampled nor randomly assigned), they invariably carry out a t-test of the statistical significance of the difference between the means for the two groups and/or construct a confidence interval for the corresponding "effect size".

I am of course not the first person to be bothered by this. The problem has been brought to the attention of readers of the methodological literature for many years. [See, for example, Levin's (1993) comments regarding Shaver (1993); Lunneborg (2000); Levin (2006); and Edgington & Onghena (2007).] As I mentioned in an earlier chapter of this book, some researchers "regard" their non-randomly-sampled subjects as having been drawn from hypothetical populations of subjects "like these". Some have never heard of randomization (permutation) tests for analyzing the data for the situation where you have random assignment but not random sampling. Others have various arguments for using the t-test (e.g., that the t-test is often a good approximation to the randomization test); and still others don't seem to care.

It occurred to me that there might be a way to create some sort of relatively simple "all-purpose" statistic that could be used to compare two independent groups no matter how they were sampled or assigned (or just stumbled upon). I have been drawn to two primary sources:

1. The age-old concept of a percentage.

2. Darlington's (1973) article in Psychological Bulletin on "ordinal dominance" (of one group over another). [The matter of ordinal dominance was treated by Bamber (1975) in greater mathematical detail and in conjunction with the notion of receiver operating characteristic (ROC) curves, which are currently popular in epidemiological research.]

My recommendation

Why not do as Darlington suggested and plot the data for Group 1 on the horizontal axis of a rectangular array, plot the data for Group 2 on the vertical axis, see how many times each of the observations in one of the groups (say Group 1) exceeds each of the observations in the other group, convert that to a percentage (he actually did everything in terms of proportions), and then do with that percentage whatever is warranted? (Report it and quit; test it against a hypothesized percentage; put a confidence interval around it; whatever).

Darlington's example [data taken from Siegel (1956)]

The data for Group 1: 0, 5, 8, 8, 14, 15, 17, 19, 25 (horizontal axis)
The data for Group 2: 3, 6, 10, 10, 11, 12, 13, 13, 16 (vertical axis)

The layout:

16							x	x	x
13					x	x	x	x	x
13					x	x	x	x	x
12					x	x	x	x	x
11					x	x	x	x	x
10					x	x	x	x	x
10					x	x	x	x	x
6			x	x	x	x	x	x	x
3		x	x	x	x	x	x	x	x
	0	5	8	8	14	15	17	19	25

The number of times that an observation in Group 1 exceeded an observation in Group 2 was 48 (count the x's). The percentage of times was 48/81, or .593, or 59.3%. Let's call that P_e for "percentage exceeding". [Darlington calculated that proportion (percentage) but didn't pursue it further. He recommended the construction of an ordinal dominance curve through the layout, which is a type of cumulative frequency distribution similar to the cumulative frequency distribution used as the basis for the Kolmogorov-Smirnov test.]

How does this differ from other suggestions?

Comparing two independent groups by considering the degree of overlapping of their respective distributions appears to have originated with the work of Truman Kelley (1919), the well-known expert in educational measurement and statistics at the time, who was interested in the percentage of one normal distribution that was above the median of a second normal distribution. [His paper on the topic was typographically botched by the Journal of Educational Psychology and was later (1920) reprinted in that journal in corrected form.] The notion of

distributional overlap was subsequently picked up by Symonds (1930), who advocated the use of biserial r as an alternative to Kelley's measure, but he was taken to task by Tilton (1937) who argued for a different definition of percentage overlap that more clearly reflected the actual amount of overlap. [Kelley had also suggested a method for correcting percentage overlap for unreliability.] Percentage overlap was subsequently further explored by Levy (1967), by Alf and Abrahams (1968), and by Elster and Dunnette (1971).

In their more recent discussions of percentage overlap, Huberty and his colleagues (Huberty & Holmes, 1983; Huberty & Lowman, 2000; Hess, Olejnik, & Huberty, 2001; Huberty, 2002) extended the concept to that of "hit rate corrected for chance" [a statistic similar to Cohen's (1960) kappa] in which discriminant analysis or logistic regression analysis is employed in determining the success of "postdicting" original group membership. (See also Preese, 1983; Campbell, 2005; and Natesan & Thompson, 2007.)

There is also the "binomial effect size display (BESD)" advocated by Rosenthal and Rubin (1982) and the "probability of superior outcome" approach due to Grissom (1994). BESD has been criticized because it involves the dichotomization of continuous variables (see the following chapter). Grissom's statistic is likely to be particularly attractive to experimenters and meta-analysts, and in his article he includes a table that provides the probabilistic superiority equivalent to Cohen's (1988) d for values of d between .00 and 3.99 by intervals of .01.

Most closely associated with the procedure proposed here (the use of P_e) is the work represented by a sequence of articles beginning with McGraw and Wong (1992) and extending through Cliff (1993), Vargha and Delaney (2000), Delaney and Vargha (2002), Feng and Cliff (2004), and Feng (2006). [Amazingly--to me, anyhow--the only citation to Darlington (1973) in any of those articles is by Delaney and Vargha in their 2002 article!] McGraw and Wong were concerned with a "common language effect size" for comparing one group with another for continuous, normally distributed variables, and they provided a technique for so doing. Cliff argued that many variables in the social sciences are not continuous, much less normal, and he advocated an ordinal measure d (for sample dominance; δ for population dominance). [This is not to be confused with Cohen's effect size d, which is appropriate for interval-scaled variables only.] He (Cliff) defined d as the difference between the probability that an observation in Group 1 exceeds an observation in Group 2 and the probability that an observation in Group 2 exceeds an observation in Group 1. In their two articles Vargha and Delaney sharpened the approach taken by McGraw and Wong, in the process of which they suggested a statistic, A, which is equal to my P_e if there are no ties between observations in Group 1 and observations in Group 2, but they didn't pursue it as a percentage that could be treated much like any other percentage. Feng and Cliff, and Feng, reinforced Cliff's earlier arguments for preferring δ and d, which range from -1 to +1. Vargha and Delaney's A

ranges from 0 to 1 (as do all proportions) and is algebraically equal to (1 + d)/2, i.e., it is a simple linear transformation of Cliff's measure. The principal difference between Vargha and Delaney's A and Cliff's d, other than the range of values they can take on, is that A explicitly takes ties into account.

Dichotomous outcomes

The ordinal-dominance-based "percentage exceeding" measure also works for dichotomous dependent variables. For the latter all one needs to do is dummy-code (0,1) the outcome variable, string out the 0's followed by the 1's for Group 1 on the horizontal axis, string out the 0's followed by the 1's for Group 2 on the vertical axis, count how many times a 1 for Group 1 appears in the body of the layout with a 0 for Group 2, and divide that count by n_1 times n_2, where n_1 is the number of observations in Group 1 and n_2 is the number of obervations in Group 2. Here is a simple hypothetical example:

The data for Group 1: 0, 1, 1, 1
The data for Group 2: 0, 0, 1, 1, 1

The layout:

1				
1				
0		x	x	x
0		x	x	x
0		x	x	x
	0	1	1	1

There are 9 instances of a 1 for Group 1 paired with a 0 for Group 2, out of 4X5 = 20 total comparisons, yielding a "percentage exceeding" value of 9/20, or .45, or 45%.

Statistical inference

For the Siegel/Darlington example, if the two groups had been simply randomly sampled from their respective populations, the inference of principal concern might be the establishment of a confidence interval around the sample P_e . [You get tests of hypotheses "for free" with confidence intervals for percentages, as I pointed out in Chapter 4.] But there is a problem regarding the "n" for P_e. In that example the sample percentage, 59.3, was obtained with $n_1 \times n_2 = 9 \times 9 = 81$ in the denominator. 81 is not the sample size (the sum of the sample sizes for the two groups is only 9 + 9 = 18). This problem had been recognized many years ago in research on the probability that Y is less than X, where Y and X are vectors of length n and m, respectively. In articles beginning with Birnbaum and McCarty (1958) and extending through Owen, Craswell, and Hanson (1964), Ury

(1972), and others, a procedure for making inferences from the sample probabilities to the corresponding population probabilities was derived.

The Owen, et al. and Ury articles are particularly helpful in that they include tables for constructing confidence intervals around a sample P_e . For the Siegel/Darlington data, the confidence intervals are not very informative, since the 90% interval extends from 0 (complete overlap in the population) to 100 (no overlap), because of the small sample size.

If the two groups had been randomly assigned to experimental treatments, but had not been randomly sampled, a randomization test is called for, with a "percentage exceeding" calculated for each re-randomization, and a determination made of where the observed P_e falls among all of the possible P_e's that could have been obtained under the (null) hypothesis that each observation would be the same no matter to which group the associated object (usually a person) happened to be assigned.

For the small hypothetical example of 0's and 1's the same inferential choices are available, i.e., tests of hypotheses or confidence intervals for random sampling, and randomization tests for random assignment. [There are confidence intervals associated with randomization tests, but they are very complicated. See, for example, Garthwaite (1996).] If those data were for a true experiment based upon a non-random sample, there are "9 choose 4" (the number of combinations of 9 things taken 4 at a time) = 126 randomizations that yield P_e 's ranging from 0.00 (all four 0's in Group 1) to 80 (four 1's in Group 1 and only one 1 in Group 2). The 45 is not among the 10% least likely to have been obtained by chance, so there would not be a statistically significant treatment effect at the 10% level. (Again the sample size is very small.) The distribution is as follows:

P_e	frequency
.00	1
.05	22
.20	58
.45	40
.80	5
	126

To illustrate the use of an arguably defensible approach to inference for the overlap of two groups that have been neither randomly sampled nor randomly assigned, I turn now to a set of data originally gathered by Ruback and Juieng (1997). They were concerned with the problem of how much time drivers take to leave parking spaces after they return to their cars, especially if drivers of other cars are waiting to pull into those spaces. They had data for 100 instances when other cars were waiting and 100 instances when other cars were not waiting. On his statistical home page, Howell (2007) has excerpted from that data set 20

instances of "someone waiting" and 20 instances of “no one waiting”, in order to keep things manageable for the point he was trying to make about statistical inferences for two independent groups. Here are the data (in seconds):

Someone waiting (Group 1)

49.48 43.30 85.97 46.92 49.18 79.30 47.35 46.52 59.68 42.89
49.29 68.69 41.61 46.81 43.75 46.55 42.33 71.48 78.95 42.06

No one waiting (Group 2)

36.30 42.07 39.97 39.33 33.76 33.91 39.65 84.92 40.70 39.65
39.48 35.38 75.07 36.46 38.73 33.88 34.39 60.52 53.63 50.62

Here is the 20x20 dominance layout (I have rounded to the nearest tenth of a second in order to save room and not bothered to order each data set):

36.3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
42.1	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x		
40.0	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
39.3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
33.8	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
33.9	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
39.7	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
84.9			x			x															
40.7	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
39.7	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
39.5	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
35.4	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
75.1			x			x														x	
36.5	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
38.7	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
33.9	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
34.4	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
60.5			x			x						x						x	x		
53.6			x			x			x			x						x	x		
50.6			x			x			x			x						x	x		
	49.5	43.3	86.0	46.9	49.2	79.3	47.4	46.5	59.7	42.9	49.3	68.7	41.6	46.8	43.8	46.6	42.3	71.5	79.0	42.1	

For these data P_e is equal to 318/400 = 79.5%. Referring to Table 1 in Ury (1972) a 90% confidence interval for π_e is found to extend from 79.5 – 36.0 to 79.5 + 36.0, i.e., from 43.5 to 100. A "null hypothesis" of a 50% proportion overlap in the population could not be rejected.

Howell actually carried out a randomization test for the time measures, assuming something like a natural experiment having taken place (without the random assignment, which would have been logistically difficult if not impossible to carry out). Based upon a random sample of 5000 of the 1.3785×10^{11} possible re-randomizations he found that there was a statistically significant difference at the 5% level (one-tailed test) between the two groups, with longer times taken when there was someone waiting. He was bothered by the effect that one or two outliers had on the results, however, and he discussed alternative analyses that might minimize their influence.

Disadvantages of the "percentage exceeding" approach

The foregoing discussion was concerned with the postulation of P_e as a possibly useful measure of the overlap of the frequency distributions for two independent groups. But every such measure has weaknesses. The principal disadvantage of P_e is that it ignores the actual magnitudes of the $n_1 \times n_2$ pairwise differences, and any statistical inferences based upon it for continuous distributions are therefore likely to suffer from lower power and less precise confidence intervals. A second disadvantage is that there is presently no computer program available for calculating P_e . [I'm not very good at writing computer programs, but I think that somebody more familiar with Excel than I am would have no trouble dashing one off. The layouts used in the two examples in this paper were actually prepared in Excel and "pasted" into a Word document.] Another disadvantage is that it is not (at least not yet) generalizable to two dependent groups, more than two groups, or multiple dependent variables.

A final note

Throughout this chapter I have referred to the 10% significance level and the 90% confidence coefficient. The choice of significance level or confidence coefficient is of course entirely up to the researcher and should reflect his/her degree of willingness to be wrong when making sample-to-population inferences. I kinda like the 10% level and 90% confidence for a variety of reasons. First of all, I think you might want to give up a little on Type I error in order to pick up a little extra power (and give up a little precision) that way. Secondly, as illustrated above, more stringent confidence coefficients often lead to intervals that don't cut down very much on the entire scale space. And then there is my favorite reason that may have occurred to others. When checking my credit card monthly statement (usually by hand, since I like the mental exercise), if I get the units (cents) digit to agree I often assume that the totals will agree. If they agree, Visa's "null hypothesis" doesn't get rejected when perhaps it should be rejected.

If they don't agree, if I reject Visa's total, and if it turns out that Visa is right, I have a 10% chance of having made a Type I error, and I waste time needlessly re-calculating. Does that make sense?

Chapter 8: Dichotomizing continuous variables: Good idea or bad idea?

A very bad idea, or at least so say Cohen (1983); Hunter and Schmidt (1990); MacCallum, Zhang, Preacher, and Rucker (2002); Streiner, 2002; Owen and Froman (2005); Royston, Altman, and Sauerbrei (2006); Altman and Royston (2006); Taylor, West, and Aiken (2006); and others. [2006 was a good year for anti-dichotomization articles!] But it's done all the time. Is there no good defense for it? In what follows I'll try to point out some of its (admittedly few) advantages and its (unfortunately many) disadvantages.

Here are a few advantages:

Simplicity of description

When it comes to investigating the relationship between two variables X and Y, nothing is simpler than dichotomizing both variables at their medians and talking about what % were above the median on X and Y, what % were below the median on X and Y, what % were above on X but below on Y, and what % were below on X but above on Y. Having to plot the continuous data, trying to figure out whether or not the plot is "linear enough" to use Pearson r, worrying about outliers, etc., is a pain.

Simplicity of inference

Percent of agreement, i.e., the % for both above plus the % for both below, can be treated just like a simple percentage (see Chapter 4). The "single-best" point estimate of the population percentage of agreement is the sample percentage of agreement, the confidence interval for the population percentage is straight-forward, and so is the hypothesis test.

Applicability to "crazy" distributions

There are some frequency distributions of continuous or "near-continuous" variables that are so unusual that dichotomization is often used in order to make any sense out of the data. In the following sections I would like to consider two of them.

Number of cigarettes smoked per day

When people are asked whether or not they smoke cigarettes and, if so, approximately how many they smoke each day, the frequency distribution has a big spike at 0, lesser spikes at 20 (the one-pack-a-day people), 40 (two packs), and 60 (three packs), but also some small spikes at 10 (half pack), 30 (pack and a half), etc. Some people smoke (or say they smoke) just one cigarette per day, but hardly anyone reports 3, 7, 11 or other non-divisors of 20. In Table 1, below,

I have provided a frequency distribution for 5209 participants in the well-known Framingham Heart Study at Time 7 (which is Year 14--around 1960--of that study). I have also included some descriptive statistics for that distribution, in order to summarize its central tendency, variability, skewness, and kurtosis. (The distribution and all calculations based upon it were carried out in Excel, carried out to far too many decimal places!)

Note some of the interesting features. The distribution exhibits the "crazy" pattern indicated in the previous paragraph, with several holes (particularly at the high end) and with heapings at observations ending in 0 and 5. It has a mode and a median of 0; a mean of about 9 1/2; standard deviation of about 13; skewness between 1 and 2; and kurtosis of approximately that same magnitude. [At first I thought that the 90 (4 1/2 packs per day!) was an error and should have been 9, but there was more than one such observation in the full data set.]

I am of course not the first person to study the frequency distribution of number of cigarettes smoked per day (see, for example, Klesges, Debon, & Ray, 1995 and the references they cite).

Table 1: Data for Year 14 of the Framingham Heart Study

# Cigs	Frequency
0	2087
1	73
2	44
3	65
4	41
5	32
6	30
7	26
8	19
9	15
10	130
11	15
12	26
13	6
14	6
15	111
16	9
17	13
18	19
19	6
20	581
21	0
22	10
23	1
24	2
25	53
26	0

Mean	9.477795
Median	0
Mode	0
Standard Dev.	13.18546
Variance	173.8564
Kurtosis	1.291503
Skewness	1.334515
Range	90
Minimum	0
Maximum	90
Sum	37134
Count	3918
Missing	1291

27	9
28	0
29	1
30	225
31	0
32	0
33	0
34	0
35	22
36	0
37	0
38	0
39	0
40	193
41	0
42	1
43	0
44	0
45	6
46	0
47	0
48	0
49	0
50	19
51	0
52	0
53	0
54	0
55	2
56	0
57	0
58	0
59	0
60	17
61	0
62	0
63	0
64	0
65	0
66	0
67	0
68	0
69	0
70	1
71	0
72	0
73	0
74	0
75	0
76	0

77	0
78	0
79	0
80	1
81	0
82	0
83	0
84	0
85	0
86	0
87	0
88	0
89	0
90	1

So what? That distribution fairly cries out to be dichotomized. But where to cut? The obvious place is between 0 and 1, so that all of the people who have "scores" of 0 can be called "non-smokers" and all of the people who have "scores" from 1 to 90 can be called "smokers". For the data in Table1 there were 2087 non-smokers out of 2727 non-missing-data persons, or 76.5%, which means there were 640 smokers, or 23.5%. [As usual, "missing" causes serious problems. I don't know why there were so many participants who didn't respond to the question. Can you speculate why?]

Klondike

Just about every computer that has Microsoft Windows as an operating system includes as part of a free software package the solitaire game of Klondike. It has a number of versions, but the one that is most interesting (to me) is the "turn-one, single pass through the pack" version. The object is to play as many cards as possible on the foundation piles of ace through king of each of the four suits. The possible "scores" (number of cards played to those piles) range from 0 to 52. Of considerable interest (again, to me, anyhow) is the frequency distribution of those scores. One would hope that the distribution could be derived mathematically, but since there is a deterministic aspect (skill) to the game (in addition to the stochastic aspect) and things can get very complicated very quickly, all such efforts to do so appear to have been unsuccessful. As the authors of a recent paper on solitaire (Yan, et al., 2005) put it: " It is one of the embarrassments of applied mathematics that we cannot determine the odds of winning the common game of solitaire." (p. 1554) Some probabilities have been mathematically derived for some versions of Klondike, e.g., the probability of being unable to play a single card in the "turn three, unlimited number of passes" version (see Latif, 2004).

I recently completed 1000 games of "turn-one, single-pass" Klondike [we retired professors have lots of time on our hands!], and the distribution of my scores is displayed in Table 2, below (summary descriptive statistics have been added, all

from Excel). Note the long tail to the right with small frequencies between 19 and 31, a big hole between 31 and 52, and heaping on 52. (Once you're able to play approximately half of the deck on the foundation piles you can usually figure out a way to play the entire deck.) I won (got a score of 52) 36 times out of 1000 tries, for a success rate of 3.6%. [Note also the paucity of scores of 0. I got only two of them in 1000 tries. It's very unusual to not be able to play at least one card on the foundation piles. And it's positively re-inforcing each time you play a card there. B.F. Skinner would be pleased!]

Table 2: Results of 1000 games of Klondike

Score	*Frequency*
0	2
1	23
2	37
3	66
4	89
5	117
6	88
7	110
8	72
9	68
10	64
11	40
12	34
13	34
14	18
15	22
16	11
17	21
18	11
19	6
20	4
21	6
22	6
23	2
24	1
25	3
26	3
27	0
28	2
29	2
30	1
31	1
32	0
33	0
34	0
35	0
36	0

Mean	9.687
Median	7
Mode	5
Standard Dev.	9.495114
Variance	90.15719
Kurtosis	11.5888
Skewness	3.242691
Range	52
Minimum	0
Maximum	52
Sum	9687
Count	1000

37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	0
45	0
46	0
47	0
48	0
49	0
50	0
51	0
52	36

Again, so what? This distribution also cries out to be dichotomized, but where? If all you care about is winning (being able to play all 52 cards on the foundation piles) the obvious place to cut is just below 52, call the winners (36 of them) 1's and the losers 0's, and talk about the percentage of winners (or, alternatively, the probability of winning), which is approximately 4%. Another reasonable possibility is to dichotomize at the median (of 7), with half of the resulting scores below that number and the other half above that number. Klondike is occasionally played competitively, so if you are able to play 7 or more cards you have approximately a 50% chance of beating your opponent.

[I just finished another 1000 games, with essentially the same results: 41 wins (4.1%), a mean of about 10; etc.]

Although he is generally opposed to dichotomizing, Streiner (2002) referred to situations where it might be OK, e.g., for highly skewed distributions such as the above or for non-linearly-related variables. [I love the title of his article!]

Now for a few of the disadvantages:

Loss of information

The first thing that's wrong with dichotomization is a loss of information. For the original variable, "number of cigarettes smoked per day", we have a pretty good idea of the extent to which various people smoke, despite its "crazy" distribution. For the dichotomy, all we know is whether or not they smoke.

Inappropriate pooling of people

For the "smoker vs.non-smoker" dichotomy there is no distinction made between someone who smokes one cigarette per day and someone who smokes four or more packs per day. Or, switching examples from smoking to age (above or

below age 21, say), height (above or below 5'7"), or weight (above or below 130#), the problem could be even worse.

Decreased precision or power

The principal objective of interval estimation is to construct a rather tight interval around the sample statistic so that the inference from statistic to corresponding parameter is strong. Confidence intervals for percentages derived from dichotomization are generally less precise than their counterparts for continuous variables. The situation for hypothesis testing is similar. If the null hypothesis is false you would like to have a high probability of rejecting it in favor of the alternative hypothesis, i.e., high power. The power for dichotomies is generally lower than the power for continuous variables. (But see Owen & Froman, 2005 for a counter-example.)

You will find discussions of additional disdvantages to dichotomization in the references cited at the beginning of this chapter.

So what's a researcher to do?

There is no substitute for common sense applied to the situation in hand.
A good rule to keep in mind is "when tempted to dichotomize, don't", UNLESS you have one or more "crazy" continuous distributions to contend with.

Chapter 9: Percentages and reliability

"Reliability and validity" are the "Rosencranz and Guildenstern" of scientific measurement. In Shakespeare's Hamlet people couldn't say one name without saying the other, and the two of them were always being confused with one another. Similarly, in discussing the properties of good measuring instruments, "reliability and validity" often come out as a single word; and some people confuse the two.

What is the difference between reliability and validity?

Simply put, reliability has to do with consistency; validity has to do with relevance. An instrument might yield consistent results from "measure" to "re-measure", yet not be measuring what you want it to measure. In this chapter I shall concentrate on reliability, in which I am deeply interested. Validity, though more important (what good is it to have a consistent instrument if it doesn't measure the right thing?), ultimately comes down to a matter of expert judgment, in my opinion, despite all of the various types of validity that you read about.

How do percentages get into the picture?

In the previous chapter I referred to a couple of advantages of dichotomies, viz., their simplicity for description and for inference. Consider the typical classroom spelling test for which 65% is "passing", i.e., in order to pass the test a student must be able to spell at least 65% of the words correctly. (We shall ignore for the moment why 65%, whether the words are dictated or whether the correct spelling is to be selected from among common misspellings, and the like. Those matters are more important for validity.)

Mary takes a test consisting of 200 words and she gets 63% right (126 out of the 200). You're concerned that those particular 200 words might contain too many "sticklers" and she really deserved to get 65% or more (at least 130 out of the 200; she only missed the "cutoff" by four words). Suppose that the 200 words on the test had been randomly drawn from an unabridged dictionary. You decide to randomly draw another set of words from that same dictionary and give Mary that "parallel form". This time she gets 61% right. You now tell her that she has failed the test, since she got less than 65% on both forms.

Types of reliability

The example just presented referred to parallel forms. That is one type of reliability. In order to investigate the reliability of a measuring instrument we construct two parallel forms of the instrument, administer both forms to a group of people, and determine the percentage of people who "pass" both forms plus the percentage of people who "fail" both forms: our old friend, percent agreement.

Percent agreement is an indicator of how consistently the instrument divides people into “passers” and “failers”.

But suppose that you have only one form of the test, not two. You can administer that form twice to the same people and again determine the % who pass both times plus the % who fail both times. This test/re-test approach is not quite as good as parallel forms, since the people might “parrot back” at Time 2 what they say at Time 1, therefore endowing the instrument with artificially high reliability.

Or suppose that you’re interested in the reliability of rating essays. You administer the essay test just once, but you ask the teacher to rate the students’ essays twice (so-called intra-rater reliability) or ask two different teachers to rate the students’ essays once each (inter-rater reliability). Percent agreement is again a good way to determine the extent to which the two sets of ratings agree. Robinson (1957) discussed the advantages and disadvantages of percent agreement vs. traditional Pearson correlations for measuring intra-rater or inter-rater reliability.

Got the idea?

Kappa

There is a strange (again, in my opinion) statistic called kappa (Cohen, 1960), which is percent agreement corrected for chance. Its formula is:

$\kappa = (P - P_c)/(100 - P_c)$

where P is actual percent agreement and P_c is the percent agreement that is expected “by chance”. So if two raters of essays have 80% agreement using a four-point rating scale, and if they were both susceptible to occasional random ratings (without reading the essay itself?), they could have (1/4)(1/4) = 1/16 = 6.25% agreement “by chance”. That would be P_c. Therefore, κ would be (80 – 6.25)/(100 – 6.25) = 78.67%.

There are two reasons why I think kappa is strange. First of all, I don’t think raters rate “by chance”. Secondly, even if they do, a researcher need only demand that the percent agreement be higher in order to compensate for same. [Hutchinson (1993) presented an argument for the use of tetrachoric correlation rather than kappa.] Landis and Koch (1977) claim that a kappa of 61% to 80% , for example, is indicative of “substantial” agreement. Why not up those numbers by 10% and define percent agreement of 71% to 90% as “substantial”? But kappa is VERY commonly used; see Fleiss et al. (2003) and some of the references that they cite.

One very interesting non-reliability use of kappa is in the detection of possible cheating on an examination (Sotaridona, 2006). Now there's a context in which there is indeed liable to be a great deal of "chance" going on!

Criterion-referenced vs. norm-referenced measurement

The previous section described various ways for determining the reliability of an instrument where there is some sort of cutoff point above which there is "success" and below which there is "failure". Such instruments are called criterion-referenced. On the other hand, instruments such as the SAT or the GRE do not have cutoff points; they are not "passed" or "failed". Scores on those tests are interpreted relative to one another rather than relative to a cutoff point. They're called norm-referenced. [Be careful not to confuse norms with standards. Norms are what are; standards are what should be.]

There are several other contributions in the criterion-referenced measurement literature regarding the use of percentages as indicators of the reliability of such instruments. For example, in building upon the work of Hambleton and Novick (1973), Subkoviak (1976), and others, Smith (2003) and, later, Walker (2005) advocated the use of the standard error of a percentage in the estimation of the reliability of a classroom test (a potentially different reliability for each student). The formula for the standard error becomes $\sqrt{P(100-P)/k}$, where P is the % of items answered correctly and k is the number of items (the "item sample size", analogous to n, the traditional "people sample size"). For example, if John answered correctly 16 out of 20 items, his P is 80%, and his standard error is $\sqrt{80(100-80)/20}$, which is about 9%. If Mary answered correctly 32 out of 40 items correctly (not necessarily items on the same test), her P is also 80% but her standard error is $\sqrt{80(100-80)/40}$, which is about 6 1/3%. Therefore the evidence is more reliable for Mary than for John. The problem, however, is that the traditional formula for the standard error of a percentage assumes that the number of observations that contribute to the percentage (people, items,…, whatever) are independent of one another. That is much more defensible when people are sampled than when items are sampled.

Chase (1996) went one step further by discussing a method for estimating the reliability of a criterion-referenced instrument test before it's ever administered!

Miscellany

There have been a number of other contributions in the literature regarding the uses of percentages in conjunction with the estimation of the reliability of a measuring instrument. Here are a few examples:

Barnette's (2005) Excel program for computing confidence intervals for various reliability coefficients includes the case of percentages.

Feldt (1996) provides formulas for confidence intervals around a proportion of mastery.

Guttman (1946) discussed a method for determining a lower bound for the reliability of an instrument that produced qualitative (nominal or ordinal) data.

I (Knapp, 1977b) proposed a technique for determining the reliability of a single test item that has been dichotomously scored.

Much later I (Knapp, 2009) I put together a whole book on reliability, some of which was concerned with the use of percentages as indicators of the reliability of a measuring instrument.

Chapter 10: Wrap-up

In this book I have tried to explain why I think that percentages are "the most useful statistics ever invented". I hope you agree. But even if you don't, I hope you now know a lot more about percentages than you did when you started reading the book.

I also said I would tell you why 153 is one of my favorite numbers. It comes from the New Testament in a passage that refers to a miracle that Jesus performed when he made it possible for his apostles to catch a boatload of fish after they had caught nothing all day long. The evangelists claim that the catch consisted of 153 large fish. Who counted them? Was it exactly 153 fish?

I would like to close with a brief annotated bibliography of references that I did not get an opportunity to cite in the previous nine chapters. Here it is (the full bibliographical information can be found in the References section that follows the conclusion of this chapter):

Aiken, et al. (2003). This article in the Journal of the American Medical Association about the relationship between nurse educational level and patient mortality has tons of percentages in its various tables. (Hospital was the unit of analysis; n = 168 of them.) There were several letters to the editor of that journal in early 2004 regarding the article. I suggest that you read the article, the letters, and the rejoinder by Aiken et al., and make your own judgment. As they say on the Fox News Channel, "I report, you decide".

Azar (2004, 2007, 2008) has written several papers on "percentage thinking". Economists claim that many people behave irrationally when making shopping saving decisions by focusing on percentage saving rather than absolute saving. He cites the classic example (Thaler, 1980; Darke and Freedman, 1993) of a person who exerts more effort to save $5 on a $25 radio than on a $500 TV. It's the same $5. (See also Chen & Rao, 2007, for comparable examples.) Fascinating stuff.

Freedman, Pisani, & Purves (2007). This is far and away the best statistics textbook ever written (in my opinion), the illustrations are almost as hilarious as those in Darrell Huff's books, and there is some great stuff on percentages. (My favorite illustration is a cartoon on page 376 in which a prospective voter says to a politician "I'm behind you 100 percent, plus or minus 3 percent or so" .) Check it out!

Gonick and Smith (1993). If you want to learn statistics on your own, and have a lot of laughs in the process, this book is for you. Through a combination of words, formulas, and cartoons (mostly cartoons, by Gonick) the authors summarize nicely most of the important concepts in statistics, both descriptive

and inferential. My favorite cartoon in the book is the one on page 2 picturing a statistician dining with his date. He says to her: "I'm 95% confident that tonight's soup has probability between 73% and 77% of being really delicious!" They even discuss the probability of a disease given a positive diagnosis (pp. 46-50) and the estimation of confidence intervals for percentages--actually proportions (pp. 114-127) that we talked about in Chapters 3 and Chapters 5, respectively, in this book (but without the great illustrations that Gonick provides).

Paulos (2008). In this companion to his Innumeracy book (he really has a way with words!), Paulos claims that the arguments for the existence of God don't add up, and he closes the book with the tongue-in-cheek claim that "96.39 per cent" of us want to have a world that is closer to a heaven on earth than it is now. Amen.

Resis (1978). In what must be one of the most important applications of percentages known to mankind, Resis described a meeting in 1944 in which Winston Churchill suggested to Josef Stalin a way of dividing up European spheres of influence between Britain and Russia. On page 368 he cited Churchill's actual words, as follows:

"Let us settle about our affairs in the Balkans. Your armies are in Rumania and Bulgaria. We have interests, missions, and agents there. Don't let us get at cross-purposes in small ways. So far as Britain and Russia are concerned, how would it do for you to have ninety per cent predominance in Rumania, for us to have ninety per cent of the say in Greece, and go fifty-fifty about Yugoslavia?" While this was being translated I wrote out on a half-sheet of paper:

Rumania	
Russia	90%
The others	10%
Greece	
Great Britain (in accord with U.S.A.)	90%
Russia	10%
Yugoslavia	50–50%
Hungary	50–50%
Bulgaria	
Russia	75%
The others	25%

I pushed this across to Stalin, who by then had heard the translation. There was a slight pause. Then he took his blue pencil and made a large tick upon it, and passed it back to us. It was all settled in no more time than it takes to set down.

For some additional interesting information regarding this matter, just google "percentages agreement" [not to be confused with "percent agreement", which is a way of determining reliability]).

Robbins & Robbins (2003a and 2003b). This pair of articles represents one of the strangest, yet interesting, applications of percentages I have ever seen. The

authors have collected data for estimating the percentage of people (both men and women) who have hair of various lengths! Read both articles. You'll like them.

Thibadeau (2000). It's hard to know whether Thibadeau is serious or not when he presents his arguments for doing away with all taxes and replacing all paper money and coins with electronic currency. But this is a delightful read (free, on the internet) and he has several interesting comments regarding percentages. My favorite one is in the section on sales taxes, where he says:

> " …sales tax is almost always a strange percentage like 6% or 7%. If something costs $1, we have to take the time to figure out whether the guy is giving the proper change on $1.07 for the five dollar bill. Most people don't check. " (p. 20)

Some great websites that I haven't previously mentioned:

1. RobertNiles.com was developed by Robert Niles and is intended primarily for journalists who need to know more about mathematics and statistics. He has a particularly nice discussion of percentages.

2. Dr. Ray L. Winstead's website has a "Percentage metric time" clock that tells you at any time of any day what percentage of the day (to four decimal places!) has transpired. How about that?!

3. The website for the physics department at Bellevue College (its name is scidiv.bellevuecollege.edu/Physics/.../F-Uncert-Percent.html) calculates for you both the "absolute percentage certainty" and the "relative percentage certainty" of any obtained measurement. All you need do is input the measurement and its margin of error. Nice.

4. The Healthy People 2010 website has all sorts of percentages among its goals for the year 2010. For example, it claims that 65% of us are presently exposed to second-hand smoke [I think that is too high]; its goal is to reduce that to 45%.

5. The CartoonStock website has some great percentage cartoons. Here are two of the best (be sure to "zoom" in at 200%):

6. There is a downloadable file called Baker's Percentage (just google those words) that provides the ingredients for various recipes as percentages of the weight of the principal ingredient (usually flour). Unfortunately (in my opinion) all of the weights of the ingredients are initially given in grams rather than in ounces.

7. www.StatPages.org is John Pezzullo's marvelous website, which will refer you to sources for calculating just about any descriptive statistic you might be interested in, as well as carry out a variety of inferential procedures.

It's been fun for me. I hope it has been for you also.

References

Aiken, L.H., Clarke, S.P., Cheung, R.B., Sloane, D.M, & Silber, J.H. (2003). Education levels of hospital nurses and surgical patient mortality. Journal of the American Medical Association, 290 (12), 1617-1623.

Alf, E., & Abrahams, N.M. (1968). Relationship between per cent overlap and measures of correlation. Educational and Psychological Measurement, 28, 779-792.

Altman, D.G., and Royston, P. (2006). The cost of dichotomising continuous variables. British Medical Journal (BMJ), 332, 1080.

Ameringer, S., Serlin, R.C., & Ward, S. (2009). Simpson's Paradox and experimental research. Nursing Research, 58 (2), 123-127.

Azar, O.H. (2004). Do people think about dollar or percentage differences? Experiments, pricing implications, and market evidence. Working paper, Northwestern University.

Azar, O.H. (2007). Relative thinking theory. Journal of Socio-Economics, 36 (1), 1-14.

Azar, O.H. (2008). The effect of relative thinking on firm strategy and market outcomes: A location differentiation model with endogenous transportation costs. Journal of Economic Psychology, 29, 684-697.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology, 12, 387-415.

Barnette, J.J. (2005). ScoreRel CI: An Excel program for computing confidence intervals for commonly used score reliability coefficients. Educational and Psychological Measurement, 65 (6), 980-983.

Berry, K.J., Mielke, P.W., Jr., & Helmericks, S.G. (1988). Exact confidence limits for proportions. Educational and Psychological Measurement, 48, 713-716.

Birnbaum, Z.W., & McCarty, R.C. (1958). A distribution-free upper confidence bound for P{Y<X}, based on independent samples of X and Y. The Annals of Mathematical Statistics, 29 (2), 558-562.

Boffey, P.M. (1976). Anatomy of a decision how the nation declared war on swine flu. Science, 192, 636-641.

Borel, E. (1962). Probabilities and life. New York: Dover.

Buchanan, W. (1974). Nominal and ordinal bivariate statistics: The practitioner's view. American Journal of Political Science, 18 (3), 625-646.

Buescher, P.A. (2008). Problems with rates based on small numbers. Statistical Primer No. 12, Division of Public Health, North Carolina Department of Health and Human Services, State Center for Health Statistics, pp. 1-6.

Buonaccorsi, J.P. (1987). A note on confidence intervals for proportions. The American Statistician, 41 (3), 215-218.

Camici, P.G. (2009). Absolute figures are better than percentages. JACC Cardiovascular Imaging, 2, 759-760.

Campbell, D.T., & Stanley, J.C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Campbell, T.C. (2005). An introduction to clinical significance: An alternative index of intervention effect for group experimental designs. Journal of Early Intervention, 27 (3), 210-227.

Carnes, R.D., & Peterson, P.L. (1991). Intermediate quantifiers versus percentages. Notre Dame Journal of Formal Logic, 32 (2), 294-306.

Chase, C. (1996). Estimating the reliability of criterion-referenced tests before administration. Mid-Western Educational Researcher, 9 (2), 2-4.

Chen, H., & Rao, A.R. (2007). When two plus two is not equal to four: Errors in processing multiple percentage changes. Journal of Consumer Research, 34, 327-340.

Choi, S.C., & Stablein, D.M. (1982). Practical tests for comparing two proportions with incomplete data. Applied Statistics, 31 (3), 256-262.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin, 114 (3), 494-509.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1983). The cost of dichotomization. Applied Psychological Measurement, 7, 249–253.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Ed.). Hillsdale, NJ: Erlbaum.

Cole, T.J. (2000). Sympercents: Symmetric percentage differences on the 100 $\log_e$ scale. Statistics in Medicine, 19, 3109-3125.

Cowell, H.R. (1998). The use of numbers and percentages in scientific writing. The Journal of Bone and Joint surgery, 80-A, 1095-1096.

Damrosch, S.P., & Soeken, K. (1983). Communicating probability in clinical reports: Nurses' numerical associations to verbal expressions. Research in Nursing and Health, 6, 85-87.

Darke, P.R., & Freedman, J.L. (1993). Deciding whether to seek a bargain: Effects of both amount and percentage off. Journal of Applied Psychology, 78 (6), 960-965.

Darlington, R.B. (1973). Comparing two groups by simple graphs. Psychological Bulletin, 79 (2), 110-116.

Delaney, H.D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality of ordinally scaled variables and small to moderate sized samples. Psychological Methods, 7 (4), 485-503.

Dershowitz, A.M. (January 15, 1995). Letter to the editor, Los Angeles Times. Cited in Merz & Caulkins (1995)…see below.

Dershowitz, A.M. (May 30, 1999). Letter to the editor, New York Times. Quoted in Chance News 8.05.

Desbiens, N.A. (2007). The reporting of statistics in medical educational studies. BMC Medical Research Methodology, 7, 35-37.

Diaconis, P., & Freedman, D. (1979). On rounding percentages. Journal of the American Statistical Association, 74 (366), 359-364.

Diamond, J., & Evans, W. (1973). The correction for guessing. Review of Educational Research, 43 (2), 181-191.

Edgerton, H.A. (1927). An abac for finding the standard error of a proportion and the standard error of the difference of proportions. Journal of Educational Psychology, 18 (2), 127-128 and 18 (5), 350. [The abac itself was inadvertently omitted from the former article but was reprinted in the latter article.]

Edgington, E.S., & Onghena, P. (2007). Randomization tests (4th. ed.). London: Chapman & Hall.

Elster, R.S., & Dunnette, M.D. (1971). The robustness of Tilton's measure of overlap. Educational and Psychological Measurement, 31, 685-697.

Feinstein, A.R. (1990). The unit fragility index: An additional appraisal of "statistical significance" for a contrast of two proportions. Journal of Clinical Epidemiology, 43 (2), 201-209.

Feldt, L.S. (1996). Confidence intervals for the proportion of mastery in criterion-referenced measurement. Journal of Educational Measurement, 33, 106-114.

Feng, D. (2006). Robustness and power of ordinal d for paired data. In S. S. Sawilowsky (Ed.), Real data analysis (pp. 163-183). Greenwich, CT : Information Age Publishing.

Feng, D., & Cliff, N. (2004). Monte Carlo evaluation of ordinal d with improved confidence interval. Journal of Modern Applied Statistical Methods, 3 (2), 322-332.

Finney, D.J. (1947). The estimation from individual records of the relationship between dose and quantal response. Biometrika, 34 (3&4), 320-334.

Finney, D.J. (1975). Numbers and data. Biometrics, 31 (2), 375-386.

Firebaugh, G. (2009). Commentary: 'Is the social world flat? W.S. Robinson and the ecologic fallacy'. International Journal of Epidemiology, 38, 368-370.

Fleiss, J.L., Levin, B., & Paik, M.C. (2003). Statistical methods for rates and proportions (3rd ed.). New York: Wiley.

Freedman, D., Pisani, R., & Purves, R. (2007) Statistics (4th. ed.). New York: Norton.

Garthwaite, P.H. (1996). Confidence intervals from randomization tests. Biometrics, 52, 1387-1393.

Gigerenzer, G. (2002). Calculated risks. New York: Simon & Schuster.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L.M., & Woloshin, S. (2008). Psychological Science in the Public Interest, 8 (2), 53-96.

Gonick, L., & Smith, W. (1993). The cartoon guide to statistics. New York: Harper Perennial.

Grissom, R.J. (1994). Probability of the superior outcome of one treatment over another. Journal of Applied Psychology, 79 (2), 314-316.

Guttman, L. (1946). The test-retest reliability of qualitative data. Psychometrika, 11, 81-95.

Hallenbeck, C. (November, 1920). Forecasting precipitation in percentages or probability. Monthly Weather Review, 645-647.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and methods for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Hart, H. (1949). A rapid test of significance for differences between percentages. Social Forces, 27 (4), 401-408.

Hess, B., Olejnik, S., & Huberty, C.J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and non-normality. Educational and Psychological Measurement, 61, 909-936.

Heynen, G. (April 22, 2009). Risk assessment: the illusion of certainty. Retrieved from the internet on August 1, 2009.

Hopkins, K.D., & Chappell, D. (1994). Quick power estimates for comparing proportions. Educational and Psychological Measurement, 54 (4), 903-912.

Howell, D.C. (2007). Randomization test on two independent samples. www.uvm.edu/~dhowell/StatPages/StatHomePage.html .

Howell, D.C. (August 28, 2008). Testing change over two measurements in two independent groups. Available on the internet by googling "David Howell testing proportions" and clicking on the first entry that comes up.

Huberty, C.J. (2002). A history of effect size indices. Educational and Psychological Measurement, 62, 227-240.

Huberty, C.J., & Holmes, S.E. (1983). Two-group comparisons and univariate classification. Educational and Psychological Measurement, 43, 15-26.

Huberty, C.J., & Lowman, L.L. (2000). Group overlap as a basis for effect size. Educational and Psychological Measurement, 60, 543-563.

Huff, D. (1954). How to lie with statistics. New York: Norton.

Huff, D. (1959). How to take a chance. New York: Norton.

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. Journal of Applied Psychology, 75, 334–349.

Hutchinson, T.P. (1993). Kappa muddles together two sources of disagreement: Tetrachoric correlation is preferable. Research in Nursing & Health, 16, 313-315.

Jovanovic, B.D., & Levy, P.S. (1997). A look at the rule of three. The American Statistician, 51 (2), 137-139.

Kastellec, J.P., & Leoni, E.L. (2007). Using graphs instead of tables in political science. Perspectives in Politics, 5 (4), 755-771.

Kelley, T.L. (1919). Measurement of overlapping. Journal of Educational Psychology, 10, 229-232.

Kelley, T.L. (1920). Measurement of overlapping [corrected version]. Journal of Educational Psychology, 11, 458-461.

Kemeny, J.G., Snell, J.L., & Thompson, G.L. (1956). Introduction to finite mathematics. Englewood Cliffs, NJ: Prentice-Hall.

Keppel, K, Garcia, T, Hallquist, S, Ryskulova, A, & Agress L. (2008). Comparing racial and ethnic populations based on Healthy People 2010 objectives. Healthy People Statistical Notes, no 26. Hyattsville, MD: National Center for Health Statistics.

Klesges, R.C., Debon, M., & Ray, J.W. (1995). Are self-reports of smoking rates biased? Evidence from the second National Health And Nutrition Survey. Journal of Clinical Epidemiology, 48 (10), 1225-1233.

Knapp, T.R. (1977a). The unit-of-analysis problem in applications of simple correlation analysis to educational research. Journal of Educational Statistics, 2, 171-196.

Knapp, T.R. (1977b). The reliability of a dichotomous test item: A correlationless approach. Journal of Educational Measurement, 14, 237-252.

Knapp, T.R. (1985). Instances of Simpson's Paradox. The College Mathematics Journal, 16, 209-211.

Knapp, T.R. (1996). Learning statistics through playing cards. Thousand Oaks, CA: Sage. Now available free of charge at www.tomswebpage.net.

Knapp, T.R. (2009). The reliability of measuring instruments. Available free of charge at www.tomswebpage.net.

Krejcie, R.V., & Morgan, D.W. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607-610.

Landis, J.R., & Koch, G.C. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Lang, T.A., & Secic, M. (2006). How to report statistics in medicine. Philadelphia: American College of Physicians.

Latif, U. (2004). The probability of unplayable solitaire (Klondike) games. Retrieved from the TechUser.Net website.

Lawshe, C.H., & Baker, P.C. (1950). Three aids in the evaluation of the significance of the difference between percentages. Educational and Psychological Measurement, 10, 263.

Levin, J.R. (1993). Statistical significance testing from three perspectives. Journal of Experimental Education, 61 (4), 378-382.

Levin, J.R. (2006). Randomization tests: Statistical tools for assessing the effects of educational interventions when resources are scarce. In S.S. Sawilowsky (Ed.), Real data analysis. (Chapter 7, pp. 115-123.) Charlotte, NC: Information Age Publishing.

Levin, J.R., & Serlin, R.C. (2000). Changing students' perspectives of McNemar's test of change. Journal of Statistics Education, 8 (2),

Levy, P. (1967). Substantive significance of significant differences between two groups. Psychological Bulletin, 67 (1), 37-40.

Lightwood, J.M., & Glantz, S.A. (2009). Declines in acute myocardial infarction after smoke-free laws and individual risk attributable to secondhand smoke. Circulation, 120, 1373-1379.

Little, R.J.A., & Rubin, D.B. (2002). Statistical analysis with missing data (2nd. Ed.). New York: Wiley.

Lunneborg, C.E. (2000). Random assignment of available cases: Let the inference fit the design. http://faculty.washington.edu/lunnebor/Australia/randomiz.pdf

MacCallum, R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. Psychological Methods, 7 (1), 19-40.

MacNeal, E. (1994). Mathsemantics. New York: Penguin.

Malinas, G. (2001). Simpson's Paradox: A logically benign, empirically treacherous Hydra. The Monist, 84 (2), 265-283.

Marascuilo, L. A., & Serlin, R. C. (1979). Tests and contrasts for comparing change parameters for a multiple sample McNemar data model. British Journal of Mathematical and Statistical Psychology, 32, 105-112.

McGraw, K.O., & Wong, S.P. (1992). A common language effect size statistic. Psychological Bulletin, 111 (2), 361-365.

Merz, J.F., & Caulkins, J.P. (1995). Propensity to abuse---propensity to murder. Chance, 8 (3).

Meyers, D.G., Neuberger, J.S., & He, J. (2009). Cardiovascular effects of bans on smoking in public places: A systematic review and meta-analysis. Journal of the American College of Cardiology, 54 (14), 1249-1255.

Mosteller, F. (1976). Swine flu: Quantifying the "possibility". Science, 192, 1286, 1288.

Mosteller, F., & McCarthy, P.J. (1942). Estimating population proportions. Public Opinion Quarterly, 6 (3), 452-458.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. Statistical Science, 5 (1), 2-12.

Mosteller, F., Youtz, C., & Zahn, D. (1967). The distribution of sums of rounded percentages. Demography, 4, 850-858.

Natesan, P., & Thompson, B. (2007). Extending improvement-over-chance I-index effect size simulation studies to cover some small-sample cases. Educational and Psychological Measurement, 67, 59-72.

Newcombe, R.G. (1998a). Two-sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in Medicine, 17, 857-872.

Newcombe, R.G. (1998b). Interval estimation for the difference between independent proportions: Comparison of eleven methods. Statistics in Medicine, 17, 873-890.

Oakes, J.M. (2009). Commentary: Individual, ecological and multilevel fallacies. International Journal of Epidemiology, 38, 361-368.

Osborne, J.W. (2002). Notes on the use of data transformations. Practical Assessment, Research and Evaluation (PARE), 8 (6).

Owen, D.B., Craswell, K.J., & Hanson, D.L. (1964). Nonparametric upper confidence bounds for Pr {Y<X} and confidence limits for Pr {Y<X} when X and Y are normal. Journal of the American Statistical Association, 59 (307), 906-924.

Owen, S.V., & Froman, R.D. (2005). Why carve up your continuous data? Research in Nursing & Health, 28, 496-503.

Parascandola, M. (1998). What's wrong with the probability of causation? Jurimetrics Journal, 39, 29-44.

Paulos, J.A. (1988). Innumeracy. New York: Hill and Wang.

Paulos, J.A. (October 15, 1995). Murder he wrote. Op-ed piece in The Philadelphia Inquirer. Retrievable by googling "paulos murder he wrote".

Paulos, J.A. (June 1, 2001). Average paradoxes that went to college. Retrievable at abcnews.go.com/Technology/WhosCounting/story?id=98444.

Paulos, J.A. (2008). Irreligion. New York: Hill and Wang.

Peterson, A.V., Kealey, K.A., Mann, S.L., Marek, P.M., Ludman, E.J., Liu, J, & Bricker, J.B. (2009). Group-randomized trial of a proactive, personalized telephone counseling intervention for adolescent smoking cessation. Journal of the National Cancer Institute, 101 (20), 1378-1392.

Peterson, P.L. (1979). On the logic of "few", "many", and "most". Notre Dame Journal of Formal Logic, 20 (1), 155-179.

Preece, P.F.W. (1983). A measure of experimental effect size based on success rates. Educational and Psychological Measurement, 43, 763-766.

Resis, A. (1978). The Churchill-Stalin "percentages" agreement on the Balkans, Moscow, 1944. The American Historical Review, 83 (2), 368-387.

Robbins, C., & Robbins, M.G. (2003a). Scalp hair length. I. Hair length in Florida theme parks: An approximation of hair length in the United States of America. Journal of Cosmetic Science, 54, 53-62.

Robbins, C., & Robbins, M.G. (2003b). Scalp hair length. II. Estimating the percentages of adults in the USA and larger populations by hair length. Journal of Cosmetic Science, 54, 367-378.

Robins, J. (May 4, 2004). Should compensation schemes be based on the probability of causation or expected years of life? Web essay.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. American Sociological Review, 15 (3), 351-357. Reprinted in 2009 in the International Journal of Epidemiology, 38, 337-341.

Robinson, W.S. (1957). The statistical measurement of agreement. American Sociological Review, 22 (1), 17-25.

Rosenbaum, S. (1959). A significance chart for percentages. Journal of the Royal Statistical Society. Series C (Applied Statistics), 8 (1), 45-52.

Rosenthal, R., & Rubin, D.B. (1982). A simple, general purpose display of experimental effect. Journal of Educational Psychology, 74 (2), 166-169.

Royston, P., Altman, D.G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. Statistics in Medicine, 25, 127-141.

Ruback, R.B., & Juieng, D. (1997). Territorial defense in parking lots: Retaliation against waiting drivers. Journal of Applied Social Psychology, 27, 821-834.

Sargent, R.P., Shepard, R.M., & Glantz, S.A. (1994). Reduced incidence of admissions for myocardial infarction associated with public smoking ban: before and after study. British Medical Journal (BMJ), 328, 977-980.

Sarna, L., Bialous, S., Wewers, M.E., Froelicher, E.S., Wells, M.J., Kotlerman, J., & Elashoff, D. (2009). Nurses trying to quit smoking using the internet. Nursing Outlook, 57 (5), 246-256.

Scheines, R. (2008). Causation, truth, and the law. Brooklyn Law Review, 73, 3, 959-984.

Schield, M. (2000). Statistical literacy: Difficulties in describing and comparing rates and percentages. Paper presented at the annual meeting of the American Statistical Association. Retrievable at www.augsburg.edu/ppages/~schield.

Schield, M. (2002). Reading and interpreting tables and graphs involving rates and percentages. Retrievable at the Augsburg StatLit website.

Schield, M. (2005). Statistical prevarication: Telling half truths using statistics. Retrievable at the Augsburg StatLit website.

Schield, M. (2006). Percentage graphs in USA Today Snapshots Online. Paper presented at the annual meeting of the American Statistical Association. Retrievable at www.StatLit.org/pdf/2006SchieldASA.pdf.

Shaver, J.P. (1993). What statistical significance testing is and what it is not. Journal of Experimental Education, 61 (4), 293-316.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society. Series B, Methodological, 13 (2), 238-241.

Smith, J.K. (2003). Reconsidering reliability in classroom assessment and grading. Educational Measurement: Issues and Practice, 22 (4), 26-33.

Sotaridona, L., van der Linden, W.J., & Meijer, R.R. (2006). Detecting answer copying using the kappa statistic. Applied Psychological Measurement, 30 (5), 412-431.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. Applied Cognitive Psychology, 5, 61-77.

Stone, C.L. (1958). Percentages for integers 1 to 399. Station Circular 341, Washington Agricultural Experiment Stations, Institute of Agricultural Sciences, State College of Washington.

Streiner, D.L. (2002). Breaking up is hard to do: The heartbreak of dichotomizing continuous data. Canadian Journal of Psychiatry, 47, 262-266.

Stuart, A. (1963). Standard errors for percentages. Journal of the Royal Statistical Society. Series C (Applied Statistics), 12 (2), 87-101.

Subkoviak, M.J. (1976). Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276.

Subramanian, S.V., Jones, K., Kaddour, A., & Krieger, N. (2009a). Revisiting Robinson: The perils of individualistic and ecologic fallacy. International Journal of Epidemiology, 38, 342–360.

Subramanian, S.V., Jones, K., Kaddour, A., & Krieger, N. (2009b). Response: The value of a historically informed multilevel analysis of Robinson's data. International Journal of Epidemiology, 38, 370-373.

Swaen, G., & Amelsvoort. (2009). A weight of evidence approach to causal inference. Journal of Clinical Epidemiology, 62, 270-277.

Symonds, P.M. (1930). A comparison of statistical measures of overlapping with charts for estimating the value of bi-serial r. Journal of Educational Psychology, 21, 586-596.

Taylor, A.B., West, S.G., & Aiken, L.S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. Educational and Psychological Measurement, 66 (2), 228-239.

Thibadeau, R. (2000). No taxes. Accessible free of charge at www.notaxesbook.com.

Thompson, B.E.R. (1982). Syllogisms using “few”, “many” , and “most”. Notre Dame Journal of Formal Logic, 23 (1), 75-84.

Thompson, B.E.R. (1986). Syllogisms with statistical quantifiers. Notre Dame Journal of Formal Logic, 27 (1), 93-103.

Thompson, P.C. (1995). A hybrid paired and unpaired analysis for the comparison of proportions. Statistics in Medicine, 14, 1463-1470.

Tilton, J.W. (1937). The measurement of overlapping. Journal of Educational Psychology, 28, 656-662.

Ury, H.K. (1972). On distribution-free confidence bounds for Pr {Y<X}. Technometrics, 14 (3), 577-581.

vanBelle, G. (2002). Statistical rules of thumb. New York: Wiley.

Vargha, A., & Delaney, H.D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. Journal of Educational and Behavioral Statistics, 25 (2), 101-132.

Vickers, A.J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. BMC Medical Research Methodology, 1:6.

Wakefield, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. International Journal of Epidemiology, 38, 330–336.

Walker, D.A. (2005). The standard error of a proportion for different scores and test length. Practical Assessment, Research, & Evaluation (PARE), 10 (5).

Walker, H.M., & Lev, J. (1953). Statistical inference. New York: Holt.

Wilks, S.S. (1940a). Representative sampling and poll reliability. Public Opinion Quarterly, 4 (2), 261-269.

Wilks, S.S. (1940b). Confidence limits and critical differences between percentages. Public Opinion Quarterly, 4 (2), 332-338.

Woloshin, S., & Schwartz, L. (2009). Numbers needed to decide. Journal of the National Cancer Institute, 101 (17), 1163-1165.

Yan, X., Diaconis, P., Rusmevichientong, P., & Van Roy, B. (2005). Solitaire: Man versus machine. In L.K. Saul, Y. Weiss, & L. Bottou (Eds.), Advances in Neural Information Processing Systems 17. Cambridge, MA: MIT Press. (Pp. 1553-1560)

Youden, W.J. (1950). Index for rating diagnostic tests. Cancer, 3, 32-35.

Zubin, J. (1935). Note on a transformation function for proportions and percentages. Journal of Applied Psychology, 19, 213-220.