William Miller

Statistics and Measurement Concepts with OpenStat

EXTRA
MATERIALS
extras.springer.com

Springer

Statistics and Measurement Concepts with OpenStat

William Miller

Statistics and Measurement Concepts with OpenStat

Springer

William Miller
Urbandale, Iowa
USA

ISBN 978-1-4614-5742-8 ISBN 978-1-4614-5743-5 (eBook)
DOI 10.1007/978-1-4614-5743-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012952927

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

To the hundreds of graduate students and users of my statistics programs. Your encouragement, suggestions and patience have kept me motivated to maintain my interest in statistics and measurement.

To my wife who has endured my hours of time on the computer and wonders why I would want to create free material.

To my dogs Tuffy, Annie, Lacy and Heidi who sit under my desk for hours at a time while I pursue this hobby.

To Gary C. Ramseyer's First Internet Gallery of Statistics Jokes http://www.ilstu.edu/%7egcramsey/Gallery.html and Joachim Verhagen http://www.xs4all.nl/%7ejcdverha/scijokes

Urbandale, Iowa, USA William G. Miller

Contents

List of Figures

Chapter 1
Basic Statistics

It is proven that the celebration of birthdays is healthy. Statistics show that those people who celebrate the most birthdays become the oldest.

Introduction

This chapter introduces the basic statistics concepts you will need throughout your use of the OpenStat package. You will be introduced to the symbols and formulas used to represent a number of concepts utilized in statistical inference, research design, measurement theory, multivariate analyses, etc. Like many people first starting to learn statistics, you may be easily overwhelmed by the symbols and formulas—don't worry, that is pretty natural and does NOT mean you are retarded! You may need to re-read sections several times however before a concept is grasped. You will not be able to read statistics like a novel (don't we wish we could) but rather must "study" a few lines at a time and be sure of your understanding before you proceed.

Symbols Used in Statistics

Greek symbols are used rather often in statistical literature. (Is that why statistics is Greek to so many people?) They are used to represent both arithmetic types of operations as well as numbers, called parameters, that characterize a population or larger set of numbers. The letters you usually use, called Arabic letters, are used for numbers that represent a sample of numbers obtained from the population of numbers.

W. Miller, *Statistics and Measurement Concepts with OpenStat*,
DOI 10.1007/978-1-4614-5743-5_1, © Springer Science+Business Media New York 2013

Two operations that are particularly useful in the field of statistics that are represented by Greek symbols are the summation operator and the products operator. These two operations are represented by the capital Greek letters Sigma Σ and Pi Π. Whenever you see these symbols you must think:

$$\Sigma = \text{``The sum of the values :''}, \text{ or}$$

$$\Pi = \text{``The product of the values :''}$$

For example, if you see Y = Σ (1,3,5,9) you would read this as "the sum of 1, 3, 5 and 9". Similarly, if you see Y = Π(1,3,5,9) you would think "the product of 1 times 3 times 5 times 9".

Other conventions are sometimes adopted by statisticians. For example, as in beginning algebra classes, we often use X to represent any one of many possible numbers. Sometimes we use Y to represent a number that depends on one or more other numbers X1, X2, etc. Notice that we used subscripts of 1, 2, etc. to represent different (unknown) numbers. Lower case letters like y, x, etc. are also sometimes used to represent a deviation of a score from the mean of a set of scores. Where it adds to the understanding, X, and x may be italicized or written in a script style.

Now lets see how these symbols might be used to express some values. For example, we might represent the set of numbers (1,3,7,9,14,20) as X1, X2, X3, X4, X5, and X6. To represent the sum of the six numbers in the set we could write:

$$Y = \sum_{i=1}^{6} X_i = 1 + 3 + 7 + 9 + 14 + 20 = 54 \tag{1.1}$$

If we want to represent the sum of any arbitrary set of N numbers, we could write the above equation more generally, thus

$$Y = \sum_{i-1}^{N} X_i \tag{1.2}$$

represents the sum of a set of N values. Note that we read the above formula as "Y equals the sum of X subscript i values for the value of i ranging from 1 through N, the number of values".

What would be the result of the formula below if we used the same set of numbers (1,3,7,9,14,20) but each were multiplied by five ?

$$Y = \sum_{i-1}^{N} 5X_i = 5\sum_{i-1}^{N} X_i = 270 \tag{1.3}$$

To answer the question we can expand the formula to

$$\begin{aligned} Y &= 5X_1 + 5X_2 + 5X_3 + 5X_4 + 5X_5 + 5X_6 \\ &= 5(X_1 + X_2 + X_3 + X_4 + X_5 + X_6) \\ &= 5(1 + 3 + 7 + 9 + 14 + 20) \\ &= 5(54) = 270 \end{aligned} \tag{1.4}$$

In other words,

$$Y = \sum_{i-1}^{N} CX_i = C \sum_{i-1}^{N} X_i \tag{1.5}$$

We may generalize multiplying any sum by a constant (C) to

$$Y = \sum_{i-1}^{N} CX_i = C \sum_{i-1}^{N} X_i \tag{1.6}$$

What happens when we sum a term which is a compound expression instead of a simple value? For example, how would we interpret

$$Y = \sum_{i-1}^{N} (X_i - C) \tag{1.7}$$

where C is a constant value?

We can expand the above formula as

$$Y = (X_1 - C) + (X_2 - C) + \ldots + (X_N - C) \tag{1.8}$$

(Note the use of ... to denote continuation to the Nth term).

The above expansion could also be written as

$$Y = (X_1 + X_2 + \ldots + X_N) - NC \tag{1.9}$$

$$\text{Or } Y = \sum_{i=1}^{N} X_i - NC \tag{1.10}$$

We note that the sum of an expression which is itself a sum or difference of multiple terms is the sum of the individual terms of that expression. We may say that the summation operator distributes over the terms of the expression!

Now lets look at the sum of an expression which is squared. For example,

$$Y = \sum_{i=1}^{N} (X_i - C)^2 \tag{1.11}$$

When the expression summed is not in its most simple form, we must first evaluate the expression. Thus

$$
\begin{aligned}
Y &= \sum_{i=1}^{N}(X_i - C)^2 = \sum_{i=1}^{N}(X_i - C)(X_i - C) = \sum_{i=1}^{N}\left[X_i^2 - 2CX_i + C^2\right] \\
&= \sum_{i=1}^{N}X_i^2 - \sum_{i=1}^{N}2CX_i + \sum_{i=1}^{N}C^2 \\
\text{or } Y &= \sum_{i=1}^{N}X_i^2 - 2C\sum_{i=1}^{N}X_i + NC^2 = \sum_{i=1}^{N}X^2 - 2CN\overline{X} - NC^2 \\
&= \sum_{i=1}^{N}X^2 - CN(2\overline{X} - C) \qquad (1.12)
\end{aligned}
$$

Probability Concepts

Maybe, possibly, could be, chances are, probably are all words or phrases we use to convey uncertainty about something. Yet all of these express some belief that a thing or event could occur or exist. The field of statistics is concerned about making such statements based on observations that will lead us to correct "guesses" about an event occuring or existing. The field of study called "statistics" gets its name from the use of samples that we can observe to estimate characteristics about the population that we cannot observe. If we can study the whole population of objects or events, there is no need for statistics! Accounting methods will suffice to describe the population. The characteristics (or indexes) we observe about a sample from a population are called *statistics*. These indexes are estimates of population characteristics called *parameters*. It is the job of the statistician to provide indexes (statistics) about populations that give us some level of *confidence* that we have captured the true characteristics of the population of interest.

When we use the term *probability* we are talking about the *proportion* of objects in some population. It might be the proportion of some discrete number of heads that we get when tossing a coin. It might be the proportion of values within a specific range of values we find when we observe test scores of student achievement examinations.

In order for the statistician to make useful observations about a sample that will help us make confident statements about the population, it is often necessary to make *assumptions* about the *distribution* of scores in the population. For example, in tossing a coin 30 times and examining the outcome as the number of heads or tails, the statistician would assume that the distribution of heads and tails after a very large number of tosses would follow the *binomial* distribution, a theoretical distribution of scores for a binary object. If the population of interest is the relationship between beginning salaries and school achievement, the statistician may have to assume that the measures of salary and achievement have a *normal* distribution and that the relationship can be described by the *bivariate-normal* distribution.

A variety of indexes (statistics) have been developed to estimate characteristics (measurements) of a population. There are statistics that describe the *central tendency* of the population such as the mean (average), median and mode. Other statistics are used to describe how variable the scores are. These statistics include the variance, standard deviation, range, semi-interquartile range, mean deviation, etc. Still other indices are used to describe the relationship among population characteristics (measures) such as the product–moment correlation and the multiple regression coefficient of determination. Some statistics are used to examine differences among samples from possibly different populations to see if they are more likely to be samples from the same population. These statistics include the "t" and "z" statistic, the chi-squared statistic and the F-Ratio statistic.

The sections below will describe many of the statistics obtained on samples to make inferences about population parameters. The assumed (theoretical) distribution of these statistics will also be described.

Additive Rules of Probability

Formal aspects of probability theory are discussed in this section. But first, we need to define some terms we will use. First, we will define a *sample space* as simply a set of points. A point can represent anything like persons, numbers, balls, accidents, etc. Next we define an *event*. An event is an observation of something happening such as the appearance of "heads" when a coin is tossed or the observation that a person you selected at random from a telephone book is voting Democrat in the next election. There may be several points in the sample space, each of which is an example of an event. For example, the sample space may consist of 5 black balls and 4 white balls in an urn. This sample space would have 9 points. An event might be "a ball is black." This event has 5 sample space points. Another event might be "a ball is white." This event has a sample space of 4 points. We may now say that the probability of an event E is the ratio of the number of sample points that are examples of E to the total number of sample points provided all sample points are equally likely. We will use the notation P(E) for the probability of an event. Now let an event be "A ball is black" where the sample space is the set of 9 balls (5 black and 4 white.) There are 5 sample points that are examples of this event out of a total of 9 sample points. Thus the probability of the event P(E) = 5/9. Notice that the probability that a ball is white is 4/9. We may also say that the probability that a ball is red is 0/9 or that the probability that the ball is both white and black is 0/9. What is the probability that the ball is either white OR black? Clearly this is (5 + 4)/9 = 1.0.

In our previous example of urn balls, we noticed that a ball is either white or black. These are mutually exclusive events. We also noted that the sum of exclusive events is 1.0. Now let us add 3 red balls to our urn. We will label our events as B, W or R for the colors they represent. Our sample space now has 12 points. What is the probability that two balls selected are either B or W? When the events are exclusive we may write this as P(B U A).

Since these are exclusive events, we can write: P(B U W) = P(B) + P(W) = 5/12 + 4/12 = 9/12 = 3/4 = 0.75.

It is possible for a sample point to be an example of two or more events. For example if we toss a "fair" coin three times, we can observe eight possible outcomes:

1. HHH 2. HHT 3. HTH 4. HTT 5. TTT 6. TTH 7. THT and 8. THH

If our coin is fair we can assume that each of these outcomes is equally likely, that is, has a probability of 1/8. Now let us define two events: event A will be getting a "heads" on flip 1 and flip 2 of the coin and event B will be getting a "heads" on flips 1 and 3 of the coin. Notice that outcomes 1 and 2 above are sample points of event A and that outcomes 1 and 3 are events of type B. Now we can define a new event that combines events A *and* B. We will use the symbol A ∩ B for this event. If we assume each of the eight sample points are equally likely we may write P(A ∩ B) = number of sample points that are examples of A ∩ B/total number of sample points, or

P(A ∩ B) = 1/8. Notice that only 1 of the points in our sample space has heads on both flips 1 and 2 and on 2 and 3 (sample point 1.) That is, the probability of event A *and* B is the probability that both events A and B occur.

When events may not be exclusive, we are dealing with the probability of an event A or Event B or both. We can then write

$$\mathrm{P(A\ U\ B) = P(A) + P(B) - P(A \cap\ B)} \tag{1.13}$$

Which, in words says, the probability of events A or B equals the probability of event A plus the probability of event B minus the probability of event A and B. Of course, if A and B are mutually exclusive then the probabilty of A and B is zero and the probability of A or B is simply the sum of P(A) and P(B).

The Law of Large Numbers

Assume again that you have an urn of 5 black balls and 4 white balls. You stir the balls up and draw one from the urn and record its color. You return the ball to the urn, again stir the balls vigourously and again draw a single ball and record its color. Now assume you do this 10,000 times, each time recording the color of the ball. Finally, you count the number of white balls you drew from the 10,000 draws. You might reasonably expect the proportion of white balls to be close to 4/9 although it is likely that it is not exactly 4/9. Should you continue to repeat this experiment over and over, it is also reasonable to expect that eventually, the proportion would be extremely close to the actual proportion of 4/9. You can see that the larger the number of observations, the more closely we would approximate the actual value. You can also see that with very small replications, say 12 draws (with replacement) could lead to a very poor estimate of the actual proportion of white balls.

Multiplication Rule of Probability

Assume you toss a fair coin five times. What is the probability that you get a "heads" on all five tosses? First, the probability of the event P(E) = 1/2 since the sample space has only two possible outcomes. The multicative rule of probability states that the probability of five heads would be 1/2 * 1/2 * 1/2 * 1/2 * 1/2 or simply (1/2) to the fifth power (1/32) or, in general, $P(E)^n$ where n is the number of events E.

As another example of this rule, assume a student is taking a test consisting of six multiple-choice items. Each item has five equally attractive choices. Assume the student has absolutely no knowledge and therefore guesses the answer to each item by randomly selecting one of the five choices for each item. What is the probability that the student would get all of the items correct? Since each item has a probability of 1/5, the probability that all items are answered correctly is $(1/5)^6$ or 0.000064. What would it be if the items were true-false items?

Permutations and Combinations

A *permutation* is an arrangement of n objects. For example, consider the letters A, B, C and D. How many permutations (arrangements) can we make with these four letters? We notice there are four possibilities for the first letter. Once we have selected the first letter there are 3 possible choices for the second letter. Once the second letter is chosen there are two possibilities for the third letter. There is only one choice for the last letter. The number of permutations possible then is 4 × 3 × 2 × 1 = 24 ways to arrange the four letters. In general, if there are N objects, the number of permutations is N × (N−1) × (N−2) × (N−3) × ... (1). We abbreviate this series of products with an exclamation point and write it simply as N! We say "N factorial" for the product series. Thus 4! = 24. We do, however, have to let 0! = 1, that is, by definition the factorial of zero is equal to one. Factorials can get very large. For example, 10! = 3,628,800 arrangements. If you spent a minute examining one arrangement of 12 guests for a party, how long would it take you to examine each arrangement? I'm afraid that if you worked 8 hours a day, 5 days a week for 52 weeks a year you (and your descendants) would still be working on it for more than a 1,000 years!

A *combination* is a set of objects without regard to order. For example, the combination of A, B, C and D in any permutation is one combination. A question arises however concerning how many combinations of K objects can be obtained from a set of N objects. For example, how many combinations of 2 objects can be obtained from a set of 4 objects. In our example, we have the possibilities of A + B, A + C, A + D, B + C, B + D and C + D or a total of 6 combinations. Notice that the combination AB is the same as BA because order is not considered. A formula may be written using permutations that gives us a general formula for combinations. It is

$$\mathrm{N!\ /\ [\ K!\ (N-K)!]} \tag{1.14}$$

In our example then, the number of combinations of 2 things out of 4 is 4!/[2! (4−2)!] which might be written as

$$\frac{4 \text{ x } 3 \text{ x } 2 \text{ x } 1}{(2 \text{ x } 1) \text{ x } (2 \text{ x } 1)} = \frac{24}{4} = 6 \tag{1.15}$$

A special mathematics notation is often used for the combination of k things out of N things. It is

$$\binom{N}{K} = \frac{N!}{K!(N-K)!} \tag{1.16}$$

You will see the use of combinations in the section on the binomial distribution.

Conditional Probability

In sections above we defined the additive law for mutually exclusive events as the sum of the invidual probabilities. For example, for a fair die the probability of each of the faces is 1/6 so the probability of getting a 1 in two tosses (toss A and a toss B) is P(A) + P(B) = 1/6 + 1/6 = 1/3. Our multiplicative law for independent events states that the probability of obtaining event A *and* event B is P(A) × P(B). So the probability of getting a 1 on toss A of a die 1 *and* toss B of the die is P(1) × P(2) = 1/6 × 1/6 = 1/36. But what if we don't know our die is a "fair" die with equal probabilties for each face on a toss? Can we use the prior information from toss A of the die to say what the probability if for toss B?

Conditional probability is the probability of an event given that another event has already occurred. We would write

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1.17}$$

If A and B are independent then

$$P(B|A) = \frac{P(A)P(B)}{P(A)} = P(B) \tag{1.18}$$

or the probability of the second toss is 1/6, the same as before.

Now consider two events A and B: for B an individual has tossed a die four times with outcomes E1, E2, E3 and E4; For A the event is the tosses with outcomes E1 and E2. The events might be the toss results of 1, 3, 5 and 6. Knowing that event A has occurred, what is the probabilty of event B, that is, P(A|B)? Intuitively you might notice that the probabilty of the B event is the sum of the individual probabilities or 1/6 + 1/6 + 1/6 + 1/6 = 2/3, and that the probability of the A event is 1/6 + 1/6 = 1/3 or half the probability of B. That is, P(A)/P(B) = 1/2.

A more formal statement of conditional probability is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.19}$$

Thus the probability of event A is conditional on the prior probability of B. The result P(A|B) is sometimes called the posterior probability. Notice we can rewrite the above equation as:

$$P(A|B)P(B) = P(A \cap B) \tag{1.20}$$

and

$$P(B|A)P(A) = P(A \cap B) \tag{1.21}$$

Since both equations equal the same thing we may write

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.22}$$

The above is known as Bayes Theorem for events.

Now consider an example. In a recent poll in your city, 40% are registered Democrats and 60% are registered Republicans. Among the Democrats, the poll shows that 70% feel that invading Iraq was a mistake and 20% feel it was justified. You have just met a new neighbor and have begun a conversation over a cup of coffee. You learn that this neighbor feels that invading Iraq was a mistake. What is the probability that the neighbor is also a Democrat? Let A be the event that the neighbor is Democrat and B be the event that she feels the invasion was a mistake. We already know that the probability of A is P(A) = 0.6. We also know that the probability of B is P(B $|$A) = 0.7 . We need to compute P(B), the probability the neighbor feels the invasion was a mistake. We notice that the probability of B can be decomposed into two exclusive parts: P(B) = P(B and A) and P(B and *not* A) where the probability of *not* A is 1—P(A) or 0.4, the probability of not being a democrat. We can write

$$P(B \cap notA) = P(notA)P(B|A) \tag{1.23}$$

$$\text{or } \mathrm{P(B) = P(B\ and\ A) + P(}not\mathrm{A)P(B|}not\mathrm{A)} \tag{1.24}$$

$$\text{or } \mathrm{P(B) = P(B|A)P(A) + P(}not\mathrm{A)\ P(B|}not\mathrm{A)} \tag{1.25}$$

Now we know P(A) = 0.4, P(*not* A) = 1−.4 = 0.6, P(B|A) = 0.7 and P(B| *not* A) = 0.2. Therefore,

$$\mathrm{P(B) = (0.7)\ (0.4) + (0.6)(0.2) = 0.40}$$

Now knowing P(B) we can compute P(A|B) using Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{(0.7)(0.4)}{0.4} = 0.7 \tag{1.26}$$

is the probability of the neighbor being Democrat.

Bayesian Statistics

In the previous section we explored Bayes Theorem. In that discussion we had prior information P(A) and sought posterior probabilities of A given that B occurred. In general, Bayesian statistics follows this core:

$$\begin{aligned}&\text{Prior Probabilities, e.g. P(A)} + \text{New Information,}\\&\quad\text{e.g. appossed to invading Iraq P(B)} = \text{Posterior}\\&\quad\text{Probability P(A|B).}\end{aligned}$$

The above example dealt with specific events. However, Bayesian statistics also can be generalized to situations where we wish to develop a posterior distribution by combining a prior distribution with a distribution of new information. The Beta distribution is often used for prior and posterior distributions. This text will not attempt to cover Bayesian statistics. The reader is encouraged to find text books specific to this topic.

Maximum Liklihood (Adapted from S. Purcell, http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_1.html)

Model-Fitting

If the probability of an event X dependent on model parameters *p* is written

$$\mathrm{P\ (X\ |\ p)}$$