UNDERSTANDING STATISTICS THROUGH PROBABILITY

Mayer Shawer

This paper will explore the place of probability in statistics, the need for probability in statistics and how probability can be linked with statistics in the classroom.

There is a trend, today, in the teaching of statistics that attempts to eliminate or ignore the subject of probability. Some of the newer books for the first course in statistics have removed most instruction in probability. This is true especially with the relationships between testing the hypothesis and probability theory. As a result modern authors of mathematical texts reduce the statistical inference to recipes.

The science of statistics consists of three parts, (1) collecting data, (2) arranging the data in the form of statistical tables or graphs and, (3) utilizing and interpreting the data. It is the first and the last of these parts, collecting data and utilizing and interpreting the data, where probability is essential.

An example of the first part of the science of statistics, collecting data, is shown if one tries to estimate population parameters. We must control the quantity of information of the sample unit we select and the methods used to collect the data. This requires that randomness be built into the sampling design, so that properties of the estimators can be assessed probabilistically. With proper randomness of the sampling, one should be able to state that the estimate is unbiased. Suppose θ is an estimator to a population parameter θ, we should specify a "bound of error β", in advance, so that the error of estimation is $|\theta - \theta| \leq \beta$. We must also state, in advance, a probability, $(1-\alpha)$, that specifies the fraction of times in repeated sampling that require the error of estimation to be less than or equal to β, $[P(|\theta-\theta|\leq\beta) = 1-\alpha]$. We can now make the statement that our estimate is unbiased and we are now $(1-\alpha)100$ confident that our estimate is within β of the true population parameter.

The literature is full of examples where an investigator's estimate is worthless because he used non-probability sampling, which is a well-known problem in collecting data.

A concrete example of failure to adequately represent the population is shown from a study to estimate the amount of money lost each year from shoplifting.[1] An investigator began by making inquiry of 26 super markets of a certain grocery store chain. He found that the chain lost $30,000.00 per year from shoplifting. He then forecast the losses due to shoplifting for the entire nation by multiplying 30,000 by the number of food chains in the fifty states. He concluded that shoplifting was a million dollar a year racket. The grocery chain that was investigated was not representative either in size of the store nor in susceptibility of shoplifting in all grocery chains in the country. Hence the investigator's estimate was worthless.

In order to avoid this situation, several steps must be taken. Before we select the sample, we should determine what kind of sampling design should be chosen so the sample will be adequate to represent the population. Also the sample must be large enough to be able to make a conclusion about the population using information we have in the sample

The third part of statistics that depends on probability, utilizing and interpreting the data includes testing the hypothesis. To test statistical hypothesis there are a few steps that should be taken. State the hypothesis in the null form [H_0] e.g. [distribution of the number of heads in a coin toss is no different from the expectation based on random probabilities:

Collect the data required;

[1] The Rocky Mountain News (Denver, Colorado), Dec 1. 1968

Estimate the probability [p value] of obtaining the data if the null hypothesis is true (the objective of a statistics test is to find the p value);

If the p value is small enough, e.g. less than α where α equals the "probability of falsely rejecting the null hypothesis" we reject H_0 otherwise we fail to reject H_0.

To understand the relationship between statistics and probability, the writer asks his classes in the first course of statistics to perform the following experiment. Each student is asked to toss one coin 50 times; repeat the experiment with two coins, three coins and then four coins. The student then determines the total number of heads from each count and represents the total number of heads in a frequency and relative frequency table. The outcome will look like the following where H = total number of heads:

H	F	RF
0	950	. 475
1	1050	.525

H	F	RF
0	460	.23
1	1040	.52
2	500	.25

H	**F**	**RF**
0	240	.012
1	750	.375
2	720	.360
3	290	.145

H	**F**	**RF**
0	100	.05
1	560	.28
2	760	.38
3	540	.27
4	49	.02

Figure 1

If we define the sample space (S) as all the possible outcomes resulting from the experiment and the event (A) as the subset of the sample space then the probability of A or $P(A)=\frac{n(A)}{n(S)}$, where n(A) is the number of elements in A and n(S) is the no. of elements of S.

Using the above definitions one can find the sample spaces and the number of heads for the probability distribution for one coin, two coins, three coins and four coins. We discover immediately, that the number of heads in each category forms a Pascal's triangle as shown in figure 2. The probability table is constructed by dividing the number of elements of the table by the number of elements in the sample space (total number of rows). In the probability table we assume that each simple event is equally likely which means that the coins are fair.

N (number of coins) Total No. of rows

H T 1

1 1 2

HH TH TT

HT

2 1 2 1 4

HHH HHT TTH TTT

THT THT

HTH HTT

3 1 3 3 1 8

HHHH HHHT HHTT TTTH TTTT

HHTH HTHT TTHT

HTHH HTTH THTH

THHH TTHH HTTH

THTH

THHT

4 1 4 6 4 1 16

5 1 5 10 10 5 1 32

Figure 2

The following is the statistical tables and the probability tables as constructed from the values in figure 1:

STATISTICAL TABLES

One coin

H	F	RF
0	950	.475
1	1050	.525

Two coins

H	F	RF
0	460	.230
1	1040	.520
2	500	.250

Three coins

H	F	RF
0	240	.120
1	750	.375
2	720	.360
3	290	.145

PROBABILITY TABLES

One coin

H	n(A)	P(H)
0	1	.500
1	1	.500

Two coins

H	n(A)	P(H)
0	1	.250
1	2	.500
2	1	.250

Three coins

H	n(A)	P(H)
0	1	.125
1	3	.375
2	3	.375
3	1	.125

Four coins

H	F	RF
0	100	.05
1	560	.28
2	760	.38
3	540	.27
4	40	.02

Four coins

H	n(A)	P(H)
0	1	.0625
1	9	.2500
2	6	.3750
3	4	.2500
4	1	.0625

Figure 3

Figure 3 shows the closeness of the relative frequency in the statistical table with the p(h) of the probability table. If the coins are fair then the statistical table will be close to the probability table. Also the comparison holds when the observation falls into one and only one category. The outcome for each observation in the sample (random sampling with replacement) is independent.

Testing the null hypothesis that the four coin tosses are fair is equivalent to testing that the statistical distribution and the probability distribution of the following tables are no different based on random probabilities.

Pursuant to the collected data: Probability table if H_o is true

H	OF[2]
0	100
1	560
2	760
3	540
4	40

H	P(H)	EF[3]
0	.0625	125
1	.250	500
2	.375	750
3	.250	500
4	.0625	125

Figure 4

Furthermore, we can calculate the p value from a multinomial distribution as follows:

Pvalue = P(of obtaining the observed distribution $| H_0$) =

$$\frac{2000!}{100!560!760!540!40!}(.0625)^{100}(.25)^{560}(.376)^{760}(.25)^{540}(.0625)^{40}$$

To find the exact probability from a multinomial probability distribution for each possible sample requires a staggering amount of calculation especially when the sample size and the number of categories is large. To avoid the staggering calculations the statistician looks for an approximation for the calculations.

If we compare the observed frequency o^i with the expected frequency E_i which we can obtain by multiplying n times the probability of each category (nP_i) as shown in the probability table above. Expected frequency for the number of one head (1h), in the table = 2000(.25) = 500. The sum of the squared difference of the observed and the expected frequencies $\Sigma(O_i-E_i)^2$ begins to reflect the extent of the disagreement. This quantity will be 0 only when the fit between the observed and expected frequencies is perfect.

[2] Observed Frequency distribution for 2000

[3] Expected Frequency distribution for 2000

An even better index which reflects the disagreement will be $\Sigma\frac{(O_i-E_i)^2}{E_i}$ which is known as the "Pearson's X^2 statistics" after it's inventor Karl Pearson so that $X^2 = \Sigma\frac{(O_i-E_i)^2}{E_i}$

Given n is large the probabilities calculated by using the X^2 statistics test with degree of freedom = j-1), j being the number of categories, are approximately or the same as the exact probabilities calculated before for the p value from the multinomial distribution.

In our case:

$$X^2 = \Sigma\frac{(O_i-E_i)^2}{E_i} = \frac{(100-125)^2}{125} + \frac{(560-500)^2}{500} + \frac{(760-750)^2}{750} + \frac{540-500)^2}{500} + \frac{(40-125)^2}{125} = 5 + 7.2 + .13 + 3.2 + 57.8 = 73.33 \quad p \text{ value} < .001$$

If you consider $\alpha = .01$, since p value $< \alpha$, we reject H_0 and conclude the coins are not fair. When we use any statistics test the assumption underlying the test must be satisfied. In our case the following assumptions were satisfied.

Each sample observation falls into one and only one category;

The outcome for n respective observations in the sample (rather than with replacement) are independent;

The size of the sample is large.

The first two assumptions stem from the multinomial sampling distribution and the third assumption comes from using X^2 to approximate the multinomial probabilities distribution.

In conclusion, the student should understand the following:

The methods we used to test the statistical hypothesis is based on a comparison of the actual frequency distribution with an ideal or probability distribution such as binomial, passion, normal, etc.;

Any statistics test, such as Z, t, X^2, etc., is developed to estimate the exact probability (p value) of obtaining data if the null hypothesis is true. If we reject the hypothesis that the four coins are fair and the actual distribution does not fit the probability distribution, then the calculated p value could be very wrong. If we reject the null hypothesis because the p value is low, it could be from the failure to meet the assumption rather than from the improbability of the results when the assumption is true.

Knowing the relationship between probability and testing statistical hypothesis will give the student a clear understanding about hypothesis testing.

The relationship between probability and statistics should be taught in a concrete way, such that the student in the first course of statistics can fully understand the entire concept of statistics through probability.

Bibliography

Ahlgren, Andrew and Joan Garfield. "Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research." Journal for Research in Mathematics Education, 19. 1988.

Aliago, Martha and Gunderson, Brenda. "Interactive Statistics." Prentice Hall, 1999.

Beaver, B., Beaver, J. and Mendenhall, W. Introduction to Probability and Statistics." Duxbury Press, 1990.

Blunan, Allan. "Elementary Statistics– A Brief Version." McGraw Hill, 2000.

Campbell, Stephen K. "Flows and Fallacies in Statistical Thinking." Prentice Hall, Inc., 1973

Hays, William L. "Statistics for Social Sciences." Holt, Teinhart, and Winston, Inc., 1973.

Kimble, Gregory a. "How to Use (and Misuse) Statistics." Prentice-Hall, Inc, 1978.

Levinson, Horace C. "Chance, Luck and Statistics." Dover Publications, Inc. New York, 1963.

Mardew, Jr., Staut, William F, and Travers, Kenneth J. "Statistics: Making Sense of the Data."

Mendenhall, William and Scheaffer, Richard L. "Elementary Survey Sampling." Fifth Edition. Duxbury Press 1996

Moore, David and Steen Arthur-Lynn, Editor. "On the Shoulders of the Giants, New Approaches to Numeracy." National Academy Press.