The Fields Institute for Research in Mathematical Sciences

SIMMER

The University of Toronto Mathematics Network

April 1998 Presentation Topic
Hypothesis Testing and the χ^2 Test of Independence

Alison Gibbs and Martin Van Driel

The Logic of Hypothesis Testing— Beyond a Reasonable Doubt

Steps for Carrying out a Statistical Hypothesis Test:

- Identify the null hypothesis. Often, the goal is to show that the null hypothesis is false.
- Collect the data.
- Calculate a **test statistic**. A test statistic is a number, calculated from the data, which has a known statistical distribution *assuming the null hypothesis is true.*
- From the distribution of the test statistic, calculate the probability of getting the value we got or a more extreme value. This is the **p-value**.
- If the p-value is "small", we've observed data values that are very unlikely. So there must be something wrong with our assumptions. We have evidence that our null hypothesis is false.
- What's "small" enough? Our definition of small is called the **significance level** of our test. Commonly used values are 0.05 and 0.01.

Trial by Jury	**Statistical Hypothesis Testing**
Prosecutor	Statistician
Trial	Collection of Data
Jury decides on the verdict	Statistical test
Assume defendant is innocent	Assume the null hypothesis is true
Weigh the evidence provided by testimony and exhibits *assuming defendant is innocent*	Assess the evidence provided by the data (as summarized in the test statistic) *assuming null hypothesis is true*
Evidence against the defendant *assuming defendant is innocent*	Calculate a p-value for the test statistic *assuming null hypothesis is true*
Defendant found guilty *beyond a reasonable doubt*	Reject the null hypothesis if *p-value less than the significance level*

The Heart of Statistics

Law of Large Numbers

The **mean** of a random observation is defined as the expected outcome, based on the distribution. For example, if we toss a fair coin four times, then the mean number of heads is 2. Now suppose we repeat the coin tossing experiment 5 times and observe on each experiment 4,2,0,1 and 2 heads respectively. Then the **sample mean** based on these 5 observations is

$$\frac{4+2+0+1+2}{5} = \frac{9}{5} = 1.8 \text{ heads.}$$

The Law of Large Numbers states that **as the sample size (number of observations) increases, the sample mean will approach the actual mean.**

Normal Distribution

For a population with a standard deviation σ and mean μ, we say that the data has a **Normal Distribution** if 95% of the observations are within 2 standard deviations of the mean, and 68% of the observations are within one standard deviation of the mean, and the mean is also the **median**.

Central Limit Theorem

Given observations $x_1, x_2, \ldots, x_n$ from a common distribution, for the sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

the Central Limit Theorem states that **as the sample size increases, the distribution of $\overline{\mathbf{x}}$ becomes closer to a normal distribution.** Also, the distribution of the sum of the random observations, $\sum_{i=1}^{n} x_i$ becomes closer to a normal distribution.

The χ^2 Test of Independence

We will be analyzing *count data*, for example, the number of women present this evening.

The Background Theory

- A fact from probability theory: If A and B are independent, then the probability of **both** A **and** B is the product of the probability of A and the probability of B.

- A random variable can be standardized by subtracting its mean and then dividing by its standard deviation. A standardized normal random variable has normal distribution with mean 0 and standard deviation 1.

- The square of a standard normal random variable has a χ^2 distribution with one degree of freedom. The sum of the squares of k standard normal random variables has a χ^2 distribution with k degrees of freedom. The number of degrees of freedom is a parameter of the distribution. The higher the degrees of freedom, the flatter the distribution.

- A count can be viewed as the sum of a (binomial) observation that assigns 1 to the observation if it possesses the feature we're interested in, and 0 otherwise. So by the Central Limit Theorem, a count has approximately a normal distribution.

The Calculations Behind the Test

- Suppose our counts are grouped in a format such as the following, called a *Two-way Contingency Table*:

	Preferred Newspaper		
Gender	Globe and Mail	Toronto Star	Toronto Sun
Male	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$
Female	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$

where $O_{i,j}$ is the observed count that falls into category (i, j). Let n be the total number of people polled (so $n = \sum_i \sum_j O_{i,j}$). Assume that there is *no relationship* between gender and newspaper preference. Then applying our fact from probability theory, the probability of a male preferring the Globe and Mail is the proportion of males times the proportion of Globe readers; the expected number of male Globe readers is n times that. Call this expected count in category (i, j): $E_{i,j}$.

- Calculate

$$X^2 = \sum_i \sum_j \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

If gender and newspaper are truly independent, X^2 has a χ^2 distribution on

$$rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$$

degrees of freedom, where r is the number of rows in our table and c is the number of columns.

Note 1: We lose a degree of freedom each time we treat something as fixed, for example, the total number of males, the total number of Sun readers, etc.

Note 2: The distribution of X^2 follows from the above distribution theory, plus some calculation. See, for example, *Mathematical Statistics with Applications*, by Mendenhall, Wackerly, and Scheaffer.

- Our statistical test:
  The null hypothesis: Gender and newspaper preference are independent.
  The test statistic:

$$X^2 = \sum_i \sum_j \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

  The distribution of the test statistic assuming the null hypothesis is true: χ^2 with $(r-1)(c-1)$ degrees of freedom.
  The conclusion: If the probability of getting an X^2 that is as large or larger than what we got is small, we have evidence that our null hypothesis is false.

References

[1] The Chance Database: `www.dartmouth.edu/~chance/`

[2] Mendenhall, W., Wackerly, D. and Schaeffer, R. *Mathematical Statistics with Applications*, 4th edition. PWS-Kent Publishing Company, Boston, 1990.

[3] Moore, D. and McCabe. G. *Introduction to the Practice of Statistics*, 2nd edition. W.H. Freeman and Company, New York, 1993.

[4] Morgan, Larry. *Statistics Handbook for the TI-83.* Texas Instruments Inc., 1997.

[5] Paulos, John Allen. *Innumeracy: Mathematical Illiteracy and its Consequences.* Hill and Wang, New York, 1988.

[6] Rice, John A. *Mathematical Statistics and Data Analysis*, 2nd edition. Wadsworth, Belmont, California, 1995.

[7] The SIMMS Project (Systemic Initiative for Montana Mathematics and Science). *What Did You Expect, Big Chi?*. Simon and Schuster, Houston. `www.math.montana.edu/mathed/simms/`

Problems

1. A life insurance company sells a term insurance policy to a 21-year-old male. The policy pays $100,000 if the insured dies within the next 5 years. The company collects a premium of $250 each year. There is a high probability that the man will live, and the insurance company will gain $1250 in premiums. But if he were to die, the company would would lose almost $100,000! Why would the insurance company want to take on this much risk?

2. In advertising for a study guide, the producers claim that students that use it do significantly better ($p < 0.05$) than students who don't. What does this mean? Is there any reason you many not want to trust the producers' claim?

3. A researcher is looking for evidence of extra-sensory perception. She tests 500 subjects, 4 of whom do significantly better ($p < 0.01$) than random guessing. Should she conclude that these 4 have ESP?

4. Did the baseball player Reggie Jackson earn the title "Mr. October"? In his 21-year career he had 2584 hits in 9864 regular season at-bats. During the World Series, he had 35 hits in 98 at bats. Is the improvement in his batting average during the World Series statistically significant?

5. Does gender influence newspaper preference? Test the hypothesis that there is not relationship between gender and preferred Toronto daily newspaper for the data we've collected:

	Newspaper		
Gender	Globe	Star	Sun
Male			
Female			

6. Here is some more data on Jane Austen and her imitator (from J. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed.). The following table gives the relative frequency of the word *a* preceded by (PB) and not preceded by (NPB) the word *such*, the word *and* followed by (FB) or not followed by (NFB) *I*, and the word *the* preceded by and not preceded by *on*.

Words	*Sense and Sensibility*	*Emma*	*Sanditon I*	*Sandition II*
a PB *such*	14	16	8	2
a NPB *such*	133	180	93	81
and FB *I*	12	14	12	1
and NFP *I*	241	285	139	153
the PB *on*	11	6	8	17
the NPB *on*	259	265	221	204

Was Austen consistent in these habits of style from one work to another? Did her imitator successfully copy this aspect of her style?

7. Was there block judging in the ice dance competition at the Olympics? Claims have been made that the decision had been determined by the judges before the games even started. In particular, it has been claimed that the judges from the five Eastern bloc countries (Russia, Ukraine, Lithuania, Poland and the Czech Republic) agreed to support each other's competitors. Some also claim that France was part of the conspiracy. Could we test this judging irregularity statistically? How? (Hint: it's not a χ^2 test!)

Answers to Problems

1. Hopefully lots of 21-year-olds buy policies from the insurance company. The law of large numbers guarantees that only a few will die, so premiums collected will more than cover pay-outs.

2. Assuming that there is no difference between the two groups of students, the probability of seeing a difference as great or greater than that observed is less than 0.05. Of course, we've no indication how the producers of the study guide found students who used it, and students who didn't. Perhaps their claim says more about the students who buy study guides.

3. One percent of the time we'd expect a person who is guessing randomly to do that well. So in a group of 500, it wouldn't be surprising if 5 people did that well. It's not likely that the 4 really do have ESP.

4. We can test this by performing a χ^2 test of independence on the following table:

	Hit	No hit
Regular season	2584	7280
World Series	35	63

   The test statistic has value 4.536 and has a χ^2 distribution with 1 degree of freedom under the hypothesis of no relationship. The p-value is 0.033. Whether or not the null hypothesis should be rejected depends on the significance level. Assuming there's no relationship between Jackson's batting average and whether or not it's a World Series game, observing a difference as greater or greater than what Jackson accomplished would happen 3% of the time. Do you consider that highly unusual?

6. The test of Austen with herself (taking just the first three columns) has a test statistic of 23.287 which, under the null hypothesis of no relationship between work and word distribution has a χ^2 distribution with 10 degrees of freedom and a p-value of 0.0097. So it appears that Austen was not consistent in the use of these word combinations! So does it matter what the imitator did?

7. An open question!