University of Guelph Numeracy Project

About Proportions

TABLE OF CONTENTS

About Distributions 1
What is a DISTRIBUTION? 1
Introduction 1
Three Characteristics to Describe Distributions 1
Normal Distribution 4
Characteristics of a Normal Distribution 4
Empirical Rule 5
Importance of a Normal Distribution 5
Central Limit Theorem 5
Skewed Distributions 6
Glossary 7
References 9

About Proportions

What is a PROPORTION?

- Proportions are a series of counts, as opposed to measurements.

- A common example of a proportion is an opinion poll. These polls generally ask the public to respond either "yes" or "no" to a question. The numbers of "yes" and "no" are counted to produce a proportion.

Introduction

- Inferences about proportions indicate that we wish to use proportions to make conclusions about the population.

  Example:
  A recent poll conducted on 1000 Canadian adults in Ottawa indicated that 60% of the people feel "hopeful" about Prime Minister Stephen Harper's leadership. How was this estimate derived?

- We can use statistics derived from the sample to make estimates about population parameters.

  p = population proportion
  (or the true parameter for the population)

  p^ = sample proportion
  (the point estimator to estimate p)

  The sample proportion is calculated using the following equation:

  $$\hat{p} = X / n$$

  where X is a random variable
  n is the number of observations

- Sample proportions, as discussed previously, are based on the binomial distribution.

  This is where the random variable X is the number of successes in n trials, for a given probability of success, called p. Recall that p is the parameter for a population which is estimated using p^.

  Binomial distributions are essentially a form of probability distributions.

Hypothesis Testing for p

Estimation of Population Parameters

- Generally, we assume that p^ = p (this is a similar assumption that we use when estimating population means).

- Since we are using p^ to estimate p, we will need the standard error of p^. In other words, how much p^ varies from sample to sample.

  Standard error is an essential value used in hypothesis calculations.

  $$\text{Standard Error (SE) of } \hat{p} = \{ [p (1 - p)] / n \}^{1/2}$$

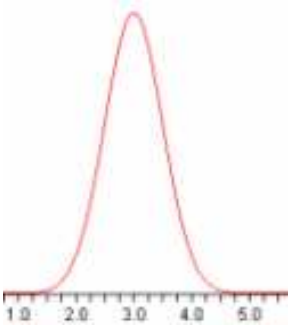
  The SE depends on the true value of p. This can be problematic because we do not know the true value of p. However, the SE is important when looking at the distribution of p^.

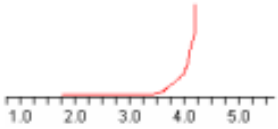
  p^ follows a normal distribution, as long as n is large enough. This approximation is best if p^ is close to 0.5. The sampling distribution of p^ is:

  $$N (p, \{ [p (1 - p)] / n \}^{1/2})$$

How does p^ follow a normal distribution?

- The following 2 graphs show how the sampling distribution of p^ can follow a normal distribution.

A) 

B) 

Graph A shows a normal distribution where either n is large or p^ is close to 0.5.

Graph B shows a skewed distribution where either n is small or p^ grows farther away from 0.5

Skewness will start to diminish (i.e., graph B will begin to look like graph A) when n increases and p^ gets closer to 0.5.

- Bringing everything together, we know:

p is approximately normal with a sampling distribution of

$$N\left(p, \left\{ [p(1-p)] / n \right\}^{1/2}\right)$$

the sample proportion p is an unbiased estimator of the parameter p, or p = p^

Overall, the ***SE of a statistic is the standard deviation of its sampling distribution.***

Hypothesis Tests for p

- The definitions used in hypothesis testing are:

  H_o = null hypothesis; statement that is used when applying a test of significance

  H_a = alternative hypothesis; statement given in a test of significant that we suspect is true rather than Ho

  The same tests are used in proportions:

  H_o: $p = p_o$ $\quad$ H_a : $p > p_o$

  $p < p_o$

  p does not equal p_o

- Z-scores is a critical concept to grasp, as there are only Z-distributions for proportions.

Confidence Intervals for p

- A confidence interval is a range of values where we believe our true value of p is located. The confidence interval includes our statistic p and a margin of error:

$$p\hat{} \pm Z^* \{ [p\hat{} (1 - p\hat{})] / n\}^{1/2}$$

Issues with Confidence Intervals in Proportions

- Confidence intervals used in proportions are based on the normal distribution of p. However, since we do not know the true value of p, we must replace it with the normal distribution of p^

- There is some skewness in the distribution of p^. Instead of taking on a value of 0.5, the value of p^ tends to move towards 0 or 1, resulting in a skewed distribution

- In order to move the sample proportion p^ away from 0 and 1, a simple adjustment to our proportion calculations needs to be made. An improvement in the confidence interval can be achieved by using the **Wilson estimate.**

The Wilson Estimate

- The Wilson estimate allows us to adjust p^ so it will shift its value closer to 0.5. We do this by 'pretending' that there are 2 more successes and 2 more failures in the number of counts for p^. This allows for greater accuracy in the confidence intervals.

  The following is the adjusted equation for p using the Wilson estimate:

  $$p` = X + 2 / n + 4$$

  where p` is the new statistical estimator of p^
  X + 2 is the new random variable
  n + 2 is the new sample size

  We can now use the new p` value to get an adjusted confidence interval:

  $$p` +/- Z^* \{ [ p` (1 – p`)] / n + 4\}^{1/2}$$

  Often, both methods of finding a confidence interval will yield similar values. However, the Wilson estimate is a valuable tool to use with smaller sample sizes because it adjusts the values by adding a success and failure to each observation.

Comparing Two Proportions

The steps for hypothesis testing with two proportions are as followed:

1) Draw two independent samples to determine p^ for each
2) Using the two p^ values, determine the pooled proportion
3) Generate a hypothesis test

Glossary

Proportion:	Counts as opposed to measurements.
Standard Error of p^:	Variation of p^ from sample to sample.
H_o :	Null hypothesis; statement that is used when applying a test of significance.
H_a :	Alternative hypothesis; statement given in a test of significance that we suspect is true rather than H_o.
Confidence interval:	A range of values where we believe our true value (or parameter) lies. The confidence interval includes a statistical mean and a margin of error.
Wilson Estimate:	Method used when calculation proportions that takes into account the need for 2 extra "successes" and 2 extra "failures" to improve calculations of confidence intervals.

References

Mccabe, George P. & Moore, David S. Introduction to the Practice of Statistics, Fourth Edition. 2003. W.H. Freeman and Company, New York.