

Using Resampling to Compare Two Proportions

KEYWORDS:

Teaching;
Randomization test;
Resampling;
Fisher's exact test.

W. Robert Stephenson
Iowa State University, Iowa, USA
e-mail: wrstephe@iastate.edu

Amy G. Froelich
Iowa State University, Iowa, USA
e-mail: amyf@iastate.edu

William M. Duckworth
Creighton University, Nebraska, USA
e-mail: williamduckworth@creighton.edu

Summary

This article shows that when applying resampling methods to the problem of comparing two proportions, students can discover that whether you resample with or without replacement can make a big difference.

INTRODUCTION

Resampling methods have been around for over 50 years dating back to the jackknife of Tukey (1958) and Quenouille (1949, 1956). Almost 30 years ago Efron (1979) laid out the theory for the bootstrap, a method for resampling from a given sample of data. A quick search of the recent literature turns up over 1000 articles about resampling methods since 2000.

Although statisticians have embraced resampling methods for their own uses, it has only been more recently that resampling methods have been included in the teaching of statistics. Julian Simon and Peter Bruce (Bruce 1992; Simon 1992; Simon and Bruce 1991) were early proponents of using resampling to teach statistics.

Several articles in *Teaching Statistics* have dealt with the use of resampling and the bootstrap in teaching statistics. Ricketts and Berry (1994) discuss using resampling to teach hypothesis testing. They compare the means of two independent samples using the Resampling Stats program developed by Simon and Bruce (c.f. <http://www.resample.com/content/about.shtml>). Taffe and Garnham (1996) discuss resampling to estimate the population mean based on a single sample. They also provide a Minitab macro for the comparison of means for two independent samples. Johnson (2001) presents bootstrap methods for estimating standard errors and constructing confidence intervals. Butler, Rothery and Roy (2003) provide many Minitab macros to address a variety of statistical problems using resampling. Their macros are available at

www.ceh.ac.uk/products/software/minitab/download.asp . Christie (2004) provides a way to use Excel to implement many of the resampling methods mentioned in the earlier articles, updating considerably the Willemain (1994) article on resampling using Lotus 1-2-3. Arnholt (2007) shows how to use the free statistical software program R to do the same analyses as those found in Christie (2004). In addition to these papers, Hesterberg (1998) provides a nice review of simulation and bootstrapping in the teaching of statistics. He gives practical advice on software for simulation and bootstrapping and provides an extensive list of references.

The above-mentioned references discuss topics that are core to introductory statistics courses, e.g. inference for a single mean, comparison of two means, correlation, etc. One topic that is not addressed separately is the comparison of two proportions. Perhaps this is due to the fact that a test of equality of proportions can be viewed as a test of equality of means with Bernoulli variables. Simon (1997) touches on the comparison of two proportions in Chapter 15 but only examines resampling with replacement for one example. Similarly, Hesterberg, Monaghan, Moore, Clipson and Epstein (2003) have an exercise on the comparison of two proportions. The comparison of two proportions provides an opportunity for students to use resampling and to discover that how you resample, with or without replacement, can make a big difference.

IS YAWNING CONTAGIOUS?

Our example for introducing the use of resampling for comparing two proportions involves data from an experiment conducted on the Discover Channel television show MythBusters. The idea of the show is to use modern-day science to explore what is real and what is fiction. On Episode 28, the MythBusters crew carried out an experiment to see if people are more likely to yawn when someone else yawns than if no one yawns. The experiment was set up with a yawning treatment consisting of an experimental subject in a room with another person who yawns. The control condition consisted of an experimental subject in a room with another person who does not yawn. Thirty-four individuals experienced the yawning treatment while sixteen experienced the control. A hidden camera behind a two-way mirror was used to record the subjects. The results of the experiment are summarized in a 2×2 contingency table displayed in table 1.

	<i>Treatment</i>	<i>Control</i>	
<i>Yawned</i>	10	4	14
<i>Did not yawn</i>	24	12	36
	34	16	50

Table 1. Yawning experiment results

Overall, 28% of all subjects in the experiment yawned. However, those exposed to the yawning treatment had a higher percentage of subjects who yawned (29.4%) compared to the control subjects (25%). Is this difference in observed proportions ($0.2941 - 0.2500 = 0.0441$) attributable to the treatment condition versus the control or could such a difference have arisen by chance? A statistical hypothesis test could answer this question.

To set up the appropriate hypotheses we need some notation. Specifically, let p_T represent the proportion of the population that would yawn if subjected to the yawning treatment. Similarly, let p_C represent the proportion of the population that would yawn if subjected to the control. We wish to test the null hypothesis of no difference in population proportions against a one-sided alternative that p_T is greater than p_C .

$$H_0 : p_T = p_C$$

$$H_A : p_T > p_C$$

The one-sided alternative is appropriate because the experimental hypothesis is that yawning is contagious, i.e. that being exposed to the yawning treatment increases the likelihood that an experimental subject will yawn.

We propose to use resampling methods to perform the test of hypothesis and answer the question, “Is yawning contagious?”

RESAMPLING

The essence of resampling is to use only the sample data and to resample from that data to create different realizations of the experimental results. In our example, we have the data from the MythBusters experiment consisting of 34 treated subjects and 16 control subjects. Of those 50 experimental subjects, 14 yawned while 36 did not yawn. So how does resampling for these data work?

First we take the sample information and create the framework for resampling. For each of the 50 subjects, we will place a ball in a box. Because 14 of the subjects yawned, 14 of the balls will have the number 1. The remaining 36 balls will have the number 0. Figure 1 gives a visual representation of the sample.

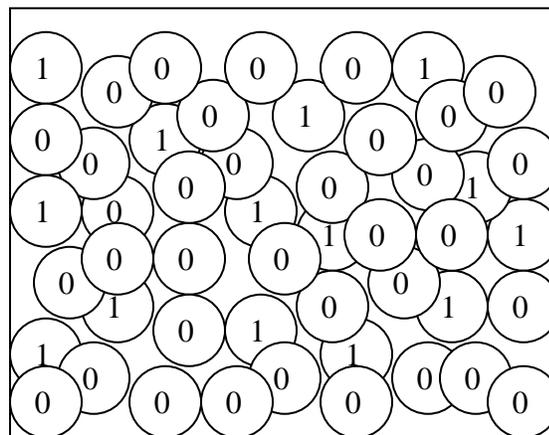


Fig. 1. Representation of the sample as balls in a box

Resampling with replacement

The 50 balls, 14 ones and 36 zeroes, should be thoroughly mixed. A resample of 34 balls is taken at random, with replacement, to make a new realization of the treatment group. Because the balls are replaced, the sample remains the same throughout the resampling process with 28% ones (yawners) and 72% zeroes (non-yawners). Similarly, a resample of 16 balls is taken at random, with replacement, to make a new realization of the control group. This process creates a new realization of the experiment under the condition that the two groups, treatment and control, came from populations with the same proportion who would yawn. We can summarize the treatment resample by counting the number of ones in the resample and computing a treatment resample proportion that yawned. Similarly, we can summarize the control resample to get a control resample proportion that yawned. The difference in resample proportions can be used to compare the treatment resample to the control resample.

This resampling procedure must be repeated a number of times in order to see the variation that results from the random resampling. Figure 2 shows the results of 1000 repeated resamples, with replacement, for the yawning experiment data. A Minitab macro and an R program for doing this resampling are included at the end of this article.

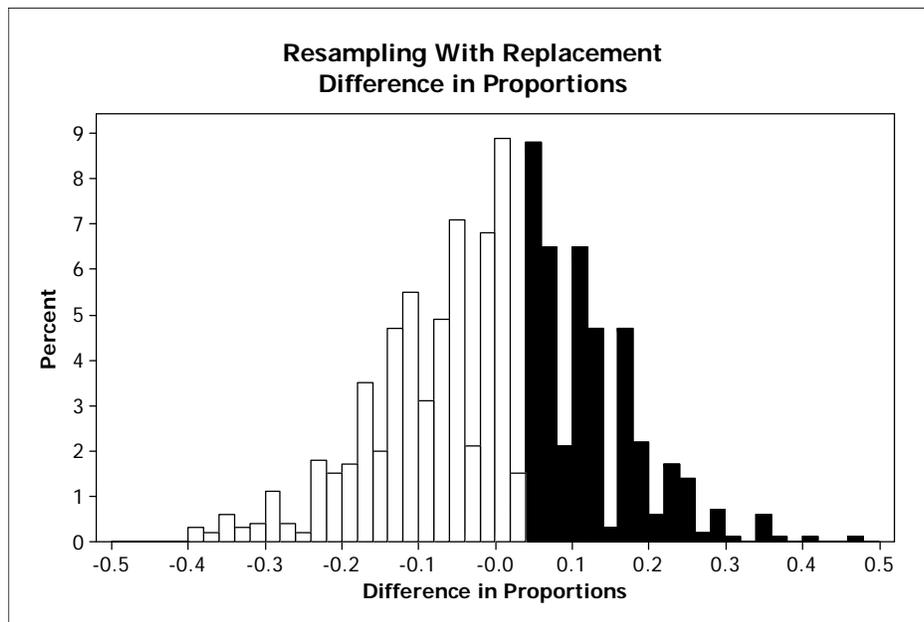


Fig. 2. 1000 resamples, with replacement, from the MythBusters experiment data

In order to evaluate whether a difference in proportions as large as the one from the actual MythBusters experiment (0.0441) is likely, we simply note the relative frequency of the 1000 resamples where the difference in resample proportions is greater than or equal to the observed difference of 0.0441. This gives a one-tailed resampling with replacement p -value of 393/1000 or 0.393, depicted by the dark shading in the histogram.

Interpreting this p -value we would say that it is fairly likely one would observe a difference in sample proportions as large, or larger, than the one produced by the MythBusters experiment even when there is no treatment effect. This large resampling p -value supports the claim that there is no difference in the proportion of people who would yawn under the two conditions, treatment and control. Thus there is little statistical evidence that yawning is contagious.

Resampling without replacement

With the same set up as depicted in figure 1, we could also resample without replacement. Random selection, without replacement, of 34 balls from the sample is used to create a new realization of the treatment group. There will be 16 balls remaining after this selection and they become the control group. Note that this method of resampling does not produce independent groups.

Again, the resampling must be repeated a number of times in order to quantify the variation introduced by the resampling. Figure 3 displays the results of 1000 repeated resamples, without replacement, from the yawning experiment sample data. A Minitab macro and R program to do this resampling are provided at the end of this article.

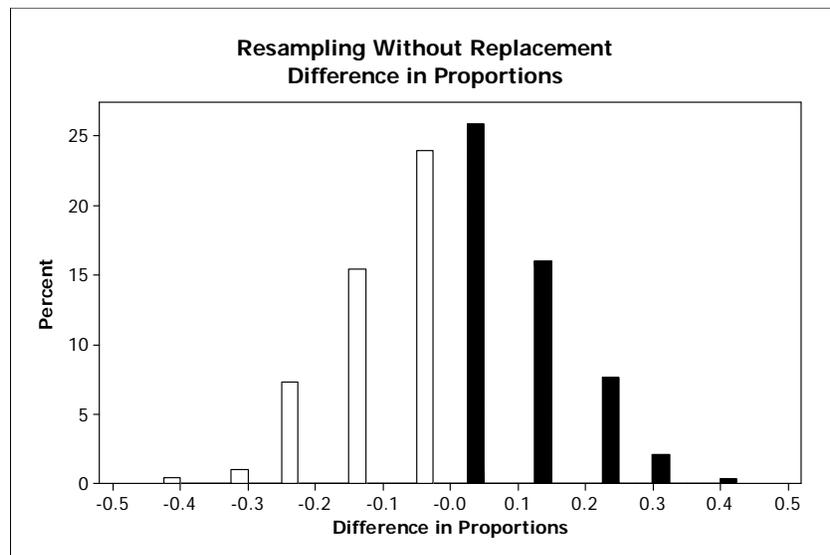


Fig. 3. 1000 resamples, without replacement, from the MythBusters experiment data

Resampling, without replacement, produces a one-tail resampling p -value of 0.519 depicted by the dark shaded bars in the histogram. This large p -value indicates that there is little statistical evidence that yawning is contagious.

Although resampling with and without replacement leads to the same conclusion that yawning is not contagious, the resampling p -values are quite different. The resampling with replacement p -value is below 0.4 while the resampling without replacement p -value is above 0.5. While the distributions in figures 2 and 3 are quite similar in shape, the

values obtained when sampling with replacement (figure 2) more closely follow a continuous distribution. That is, relatively fewer values are possible for the difference in proportions when sampling without replacement. This provides one explanation for the difference in p -values. A better explanation for the difference, however, comes from looking at more traditional methods of conducting our hypothesis test.

SAMPLING FROM A FINITE POPULATION

Figure 1 could also represent a finite population from which we can sample, with or without replacement. Sampling with replacement from this finite population and counting the number of ones (yawners) can be modelled by a binomial random variable. Therefore the number of yawners in a treatment sample, x_T is a binomial random variable with $n_T = 34$ and $p_T = 0.28$. Similarly, the number of yawners in a control sample, x_C is a binomial random variable with $n_C = 16$ and $p_C = 0.28$. The two binomial random variables are independent. The basis for the test of hypothesis is the difference between the sample proportions $\hat{p}_T - \hat{p}_C = \frac{x_T}{n_T} - \frac{x_C}{n_C}$. The distribution of the difference in sample proportions is a discrete distribution. For small sample sizes you can enumerate the different possible values for the difference. When sample sizes are equal and the null hypothesis is true, the distribution of the difference is symmetric. For larger sample sizes the number of different possible values becomes very large and the discrete distribution begins to look almost continuous. The usual test statistic for the difference between two independent proportions is

$$z = \frac{(\hat{p}_T - \hat{p}_C)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}} \quad \text{where } \hat{p} = \frac{x_T + x_C}{n_T + n_C}$$

Carrying out this test on the MythBusters' data yields,

$$z = \frac{(10/34 - 4/16)}{\sqrt{0.28(1-0.28)\left(\frac{1}{34} + \frac{1}{16}\right)}} = 0.3241$$

As this approximately follows a standard normal distribution the one-tail p -value is estimated to be 0.373. This is comparable to the resampling with replacement p -value of 0.393.

The resampling with replacement distribution displayed in figure 2 is essentially a simulation of the distribution of the difference in two proportions, the numerator of the z -test statistic. One could create the resampling distribution of the test statistic z rather than the difference in two proportions if one chooses. Note also that the square of the z

statistic is equivalent to the usual χ^2 statistic calculated on the 2×2 contingency table data provided in table 1.

Sampling without replacement from a finite population creates dependencies among subsequent selections because the population changes with each selection. In particular, the selection of the treatment sample also determines, by what remains in the population, the make up of the control sample. Note that every table produced using resampling without replacement will have exactly the same marginal totals as the original data table. There will always be 14 who yawned and 36 who did not yawn in each of the resamples without replacement. This restriction limits the number of possible differences in the two sample proportions.

The resampling without replacement distribution displayed in figure 3 is essentially a simulation of Fisher's exact test. Fisher's exact test can be used to test for the difference in two proportions based on data presented in a 2×2 contingency table, e.g. table 1. Fisher's test does not rely on the normal approximation, as the z test or equivalent χ^2 test do, and so is appropriate even when sample sizes and expected values are small. The p -value in this case may be taken to be the chance of 10 or more yawners in the treatment group for the given marginal totals. In particular,

$$p\text{-value} = \frac{\binom{14}{10}\binom{36}{24}}{\binom{50}{34}} + \frac{\binom{14}{11}\binom{36}{23}}{\binom{50}{34}} + \dots + \frac{\binom{14}{14}\binom{36}{20}}{\binom{50}{34}} = 0.513$$

The Web applet at <http://faculty.vassar.edu/lowry/fisher.html> performs this calculation and gives further background on Fisher's exact test. Note that resampling without replacement gives a very similar p -value.

As to which method should be used in general - the z or χ^2 test or Fisher's exact test, we refer the reader to Little (1989).

SUMMARY

Using resampling to compare two proportions provides a way for students to investigate the difference between resampling with and without replacement. The different resampling p -values produced by the two methods can be explained by revisiting ideas from elementary probability and sampling with and without replacement from a finite population. This in turn leads to the discovery of Fisher's exact test as an alternative to the usual z or χ^2 test when sample sizes and expected counts are small.

Acknowledgement

The idea for this article and many of the references came from Duckworth and Stephenson (2003).

References

- Arnholt, A. T. (2007). Resampling with R. *Teaching Statistics*, **29**(1), 21-26.
- Bruce, P. C. (1992). Resampling as a complement to 'Against All Odds'. *American Statistical Association Proceedings of the Section on Statistical Education*, 85-93.
- Butler, A., Rothery, P. and Roy, D. (2003). Minitab macros for resampling methods. *Teaching Statistics*, **25**(1) 22-25.
- Christie, D. (2004). Resampling with Excel. *Teaching Statistics*, **26**(1), 9-14.
- Duckworth, W. M. and Stephenson, W. R. (2003). Resampling methods: Not just for statisticians anymore. *American Statistical Association Proceedings of the Section on Teaching Statistics in Health Sciences*, 1280-1285.
- Efron, B. (1979). Bootstrap Methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.
- Hesterberg, T. C. (1998), Simulation and bootstrapping for teaching statistics. *American Statistical Association Proceedings of the Section on Statistical Education*, 44-52.
- Hesterberg, T., Monaghan, S., Moore, D., Clipson, A., and Epstein, R. (2003). Bootstrap Methods and permutation Tests. W.H. Freeman, NY, http://bcs.whfreeman.com/pbs/cat_160/PBS18.pdf.
- Johnson, R. W. (2001). An introduction to the bootstrap. *Teaching Statistics*, **23**(2), 49-54.
- Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician*, **43**(4), 283-288.
- Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, **11**, 68-84.
- Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353-360.
- Ricketts, C. and Berry, J. (1994). Teaching statistics through resampling. *Teaching Statistics*, **16**(2), 41-44.

Simon, J. L. (1992). Resampling and the ability to do statistics. *American Statistical Association Proceedings of the Section on Statistical Education*, 78-84.

Simon, J. L. (1997). Resampling: The New Statistics, 2nd ed. <http://www.resample.com/content/text/index.shtml>.

Simon, J. L. and Bruce, P. (1991). Resampling: A tool for everyday statistical work. *Chance*, **4**(1), 22-32.

Taffe, J. and Garnham, N. (1996). Resampling, the bootstrap and Minitab. *Teaching Statistics*, **18**(1), 24-25.

Tukey, J. W. (1958). Bias and confidence in not quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 614.

Willemain, T. R. (1994). Bootstrap on a shoestring: Resampling using spreadsheets. *The American Statistician*, **48**(1), 40-42.

Minitab Macros

MACRO

Pearson nreps n1 n2 obsx1 obsx2

Mconstant nreps n1 n2 obsx1 obsx2 obsdiff x n nx Rcount OneTail Count TwoTail I x1 x2
Mcolumn sample trmt1 trmt2 diffprop

Let obsdiff = obsx1/n1 - obsx2/n2

Let x = obsx1 + obsx2

Let n = n1 + n2

Let nx = n - x

Set sample

x(1)

nx(0)

End

Let RCount = 0

Let Count = 0

Do I = 1:nreps

Sample n1 sample trmt1;

Replace.

Let x1 = sum(trmt1)

Sample n2 sample trmt2;

Replace.

Let x2 = sum(trmt2)

Let diffprop[I] = x1/n1 - x2/n2

If diffprop[I] GE obsdiff

```

Let RCount = RCount + 1
EndIf
If abs(diffprop[I]) GE obsdiff
Let Count = Count + 1
EndIf

EndDo

Let OneTail = RCount/nreps
Let TwoTail = Count/nreps

Print OneTail
Print TwoTail

Stem-and-Leaf diffprop.

Histogram diffprop;
  Percent;
  Bar;
  Title "Resampling With Replacement";
  Title "Difference in Proportions".

ENDMACRO

```

Fig 4. Minitab macro for resampling with replacement

MACRO

Fisher nreps n1 n2 obsx1 obsx2

Mconstant nreps n1 n2 obsx1 obsx2 x n nx obsdiff Rcount OneTail Count TwoTail I x1 x2
Mcolumn sample trmt1 diffprop
Let obsdiff = obsx1/n1 - obsx2/n2

```

Let x = obsx1 + obsx2
Let n = n1 + n2
Let nx = n - x
Set sample
x(1)
nx(0)
End

```

```

Let RCount = 0
Let Count = 0

```

```

Do I = 1:nreps
Sample n1 sample trmt1
Let x1 = sum(trmt1)
Let x2 = x - x1
Let diffprop[I] = x1/n1 - x2/n2

```

```

If diffprop[I] GE obsdiff
Let RCount = RCount + 1
EndIf

```

```

If abs(diffprop[I]) GE obsdiff

```

```

Let Count = Count + 1
EndIf

EndDo

Let OneTail = RCount/nreps
Let TwoTail = Count/nreps

Print OneTail
Print TwoTail

Stem-and-Leaf diffprop.

Histogram diffprop;
  Percent;
  Bar;
  Title "Resampling Without Replacement";
  Title "Difference in Proportions".

ENDMACRO

```

Fig 5. Minitab macro for resampling without replacement

R Code

```

withreplacement<- function(nt, nc, xt, xc, numresamp){

n<- nt + nc
number1s<- xt+xc
number0s<- n - number1s
pop<- c(rep(1,number1s), rep(0,number0s))

obspthat<- xt/nt
obspchat<- xc/nc
obsdiff<<- obspthat - obspchat

simdiffs<- rep(0, numresamp)

for(i in 1:numresamp){
resampt<- sample(pop, nt, replace = T)
resampc<- sample(pop, nc, replace = T)
simdiffs[i]<- sum(resampt)/nt - sum(resampc)/nc
}

onesided.pvalue<- sum(ifelse(simdiffs >= obsdiff, 1, 0))/numresamp
twosided.pvalue<- sum(ifelse(abs(simdiffs) >= obsdiff, 1, 0))/numresamp

cat("One-Sided p-value = ", onesided.pvalue, "\n")
cat("Two-Sided p-value = ", twosided.pvalue, "\n")

hist(simdiffs, prob = T, main = "Resampling With Replacement, Difference in Proportions", xlab =
"Difference in Proportions", ylab = "Percent")
}

```

Fig 6. R code for resampling with replacement

```
withoutreplacement<- function(nt, nc, xt, xc, numresamp){  
  
  n<- nt + nc  
  number1s<- xt+xc  
  number0s<- n - number1s  
  
  pop<- c(rep(1,number1s), rep(0,number0s))  
  
  obspthat<- xt/nt  
  obspchat<- xc/nc  
  obsdiff<<- obspthat - obspchat  
  
  simdiffs<- rep(0, numresamp)  
  
  for(i in 1:numresamp){  
    resampt<- sample(pop, nt, replace = F)  
    resampc<- number1s - sum(resampt)  
    simdiffs[i]<- sum(resampt)/nt - resampc/nc  
  }  
  
  onesided.pvalue<- sum(ifelse(simdiffs >= obsdiff, 1, 0))/numresamp  
  twosided.pvalue<- sum(ifelse(abs(simdiffs) >= obsdiff, 1, 0))/numresamp  
  
  cat("One-Sided p-value = ", onesided.pvalue, "\n")  
  cat("Two-Sided p-value = ", twosided.pvalue, "\n")  
  
  hist(simdiffs, prob = T, main = "Resampling Without Replacement, Difference in Proportions", xlab =  
  "Difference in Proportions", ylab = "Percent")  
}
```

Fig 7. R code for resampling without replacement