Associations Between Variables

School of Information Studies
Syracuse University

Learning Topics for This Week

Introduction to associations/covariance

Cross-products/Pearson product moment correlation

Reading a correlation matrix of the Iris data set

Inferential reasoning about the correlation coefficient

Categorical associations

The Chi-Square distribution and the Chi-Squared test

Modeling proportions with Bayesian posterior distributions

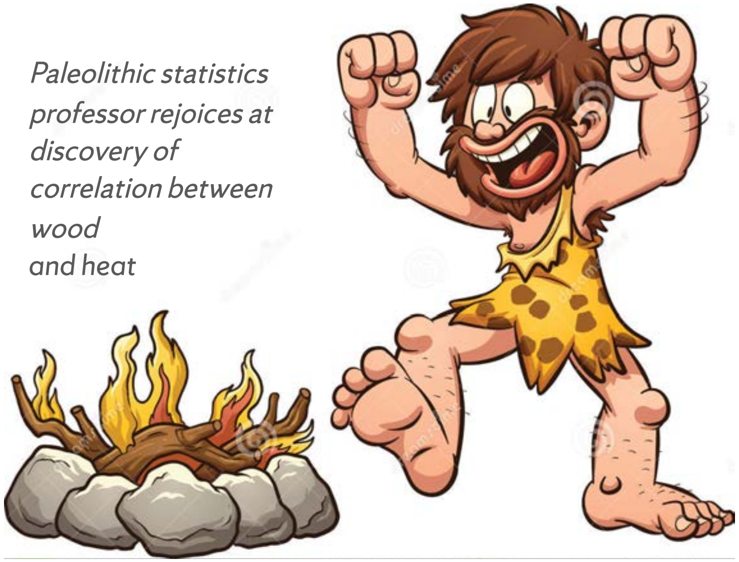
Introduction to Associations/Covariance

The Nature of Associations

More wood on the fire makes more heat, and less wood on the fire makes less heat—the amount of wood and the amount of heat are associated

Associations vary in their "strength"—some associations are strong, other associations are weak and difficult to spot, but may nonetheless be meaningful

We can *partition* the variance of heat and wood variables into two components: a shared component and an independent component; the ratio of common variance—*covariance*—vs. independent variance is the correlation between the two variables



Paleolithic statistics professor rejoices at discovery of correlation between wood and heat

Image credit: Dreamstime.com

Does Wood Cause Heat?

In this instance probably yes, but. . .

Other Examples of Correlations

The height of adolescents correlates positively with their age in years

The number of stories in a building correlates positively with its market value

The amount of milk poured into a bowl of cereal correlates negatively with ratings of how crispy it is

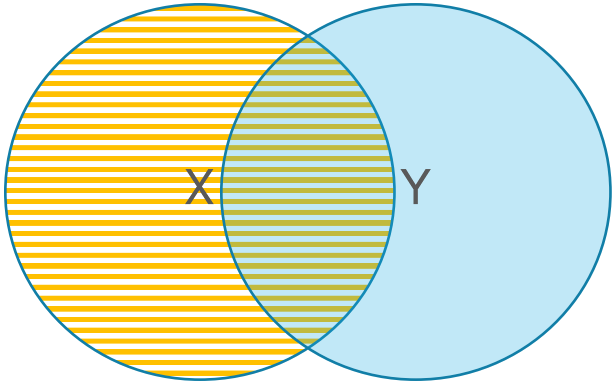
The amount of water poured on a fire correlates negatively with how hot the fire is

Preview of Next Week

Correlations and covariances are the raw ingredients in many other statistics

Next week we will examine "linear multiple regression," which translates a set of covariances into a linear prediction model

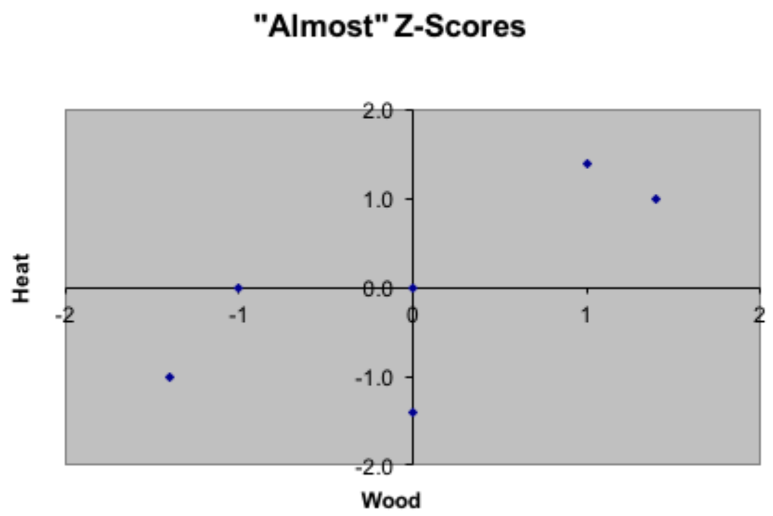
Of particular importance is the notion of common variance: some like to think of it as a Venn diagram:



Cross-Products/Pearson Product Moment Correlation

"Almost" z-Scores: Strong Positive Correlation

Cross-products of "almost" z-scores				
	Wood	Heat		CP
	-1	0		0
	-1.4	-1		1.4
	0	0		0
	0	-1.4		0
	1.4	1		1.4
	1	1.4		1.4
Mean	0	0		**0.7**
Stdevp	0.99	0.99		



"Almost" z-Scores: Strong Negative Correlation

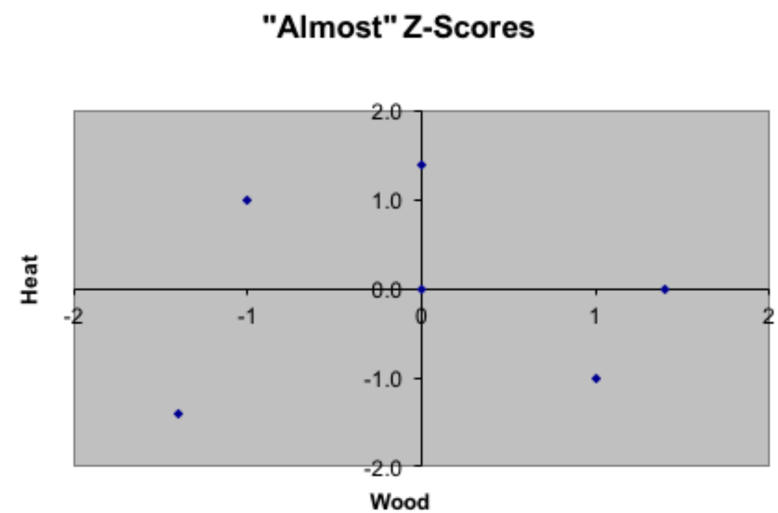
Cross-products of "almost" z-scores				
	Wood	Heat		CP
	1	0		0
	1.4	-1		-1.4
	0	0		0
	0	-1.4		0
	-1.4	1		-1.4
	-1	1.4		-1.4
Mean	0	0		-0.7
Stdevp	0.99	0.99		

"Almost" z-Scores: Small Negative Correlation

Cross-products of "almost" z-scores				
	Wood	Heat		CP
	-1	0		0
	1.4	-1		-1.4
	0	0		0
	0	-1.4		0
	-1.4	1		-1.4
	1	1.4		1.4
Mean	0	0		-0.233
Stdevp	0.99	0.99		

"Almost" z-Scores: Nearly Zero Correlation

Cross-products of "almost" z-scores				
	Wood	Heat		CP
	-1.4	-1.4		1.96
	-1	1		-1
	0	0		0
	0	1.4		0
	1	-1		-1
	1.4	0		0
Mean	0	0		-0.007
Stdevp	0.99	0.99		



School of Information Studies
Syracuse University

Reading a Correlation Matrix of the Iris Data Set

Introducing the Iris Database

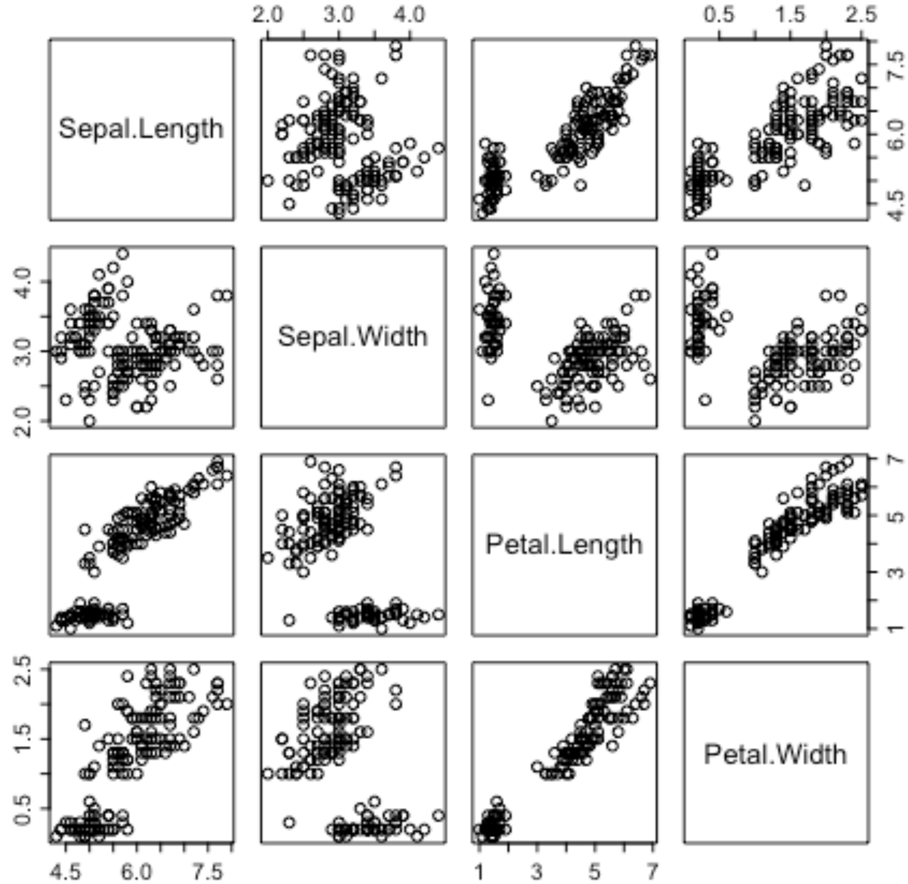
> ?iris

"This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

The data were collected by Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2-5."

The Pairs() Command Shows Scatterplots for Every Pair of Vars

```
pairs(iris[,1:4])
```



Reading a Correlation Matrix

```
cor(iris[,1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Inferential Reasoning About the Correlation Coefficient

Inferential Reasoning About “r”

r is a sample statistic, so an imperfect representation of the population, always somewhat off the mark

The population value of the correlation is “rho”: ρ

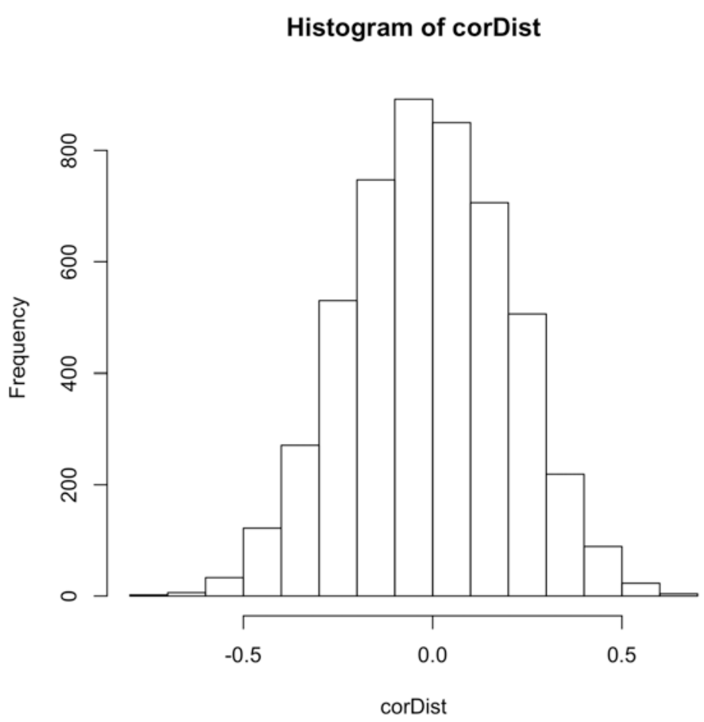
Each time we draw a sample from a population, we will get a different result

A common question among researchers: Given the characteristics of this sample, is the actual value of rho in the population (very close to) zero?

Similar to when we examined samples for mean differences: Inferential test on r gives information about whether or not the population value of r is (very close to) zero

Inferential Reasoning About "r"

```
set.seed(12345)
wood <- rnorm(2400)
heat <- rnorm(2400)
fireDF <- data.frame(wood, heat)
# Generate 5000 samples, calculate "r" for each one
corDist <- replicate(5000,cor(fireDF[sample(nrow(fireDF), 24), ])[1,2])
hist(corDist)
```



School of Information Studies
Syracuse University

Significance Test on “r” Is “t”

```
wood <- rnorm(24)
heat <- rnorm(24)
cor.test(wood,heat)

        Pearson’s product-moment correlation

data: wood and heat
t = -0.2951, df = 22, p-value = 0.7707
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4546764 0.3494514
sample estimates:
      cor
-0.06279774
```

Bayesian Test on “r”

2. Quantiles for each variable:

2.5%	25%	50%	75%	97.5%
-0.4105	-0.1622	-0.0420	0.0724	0.3140

Bayes factor analysis:

[1] rhoNot0 : 0.385294 ±0%

Against denominator: intercept only

Bayes factor type: BFlinearModel, JZS

The 2.5% and 97.5% quantiles define the HDI for this distribution

This is the point estimate of ”r” from the MCMC analysis

This is the Bayes factor: the odds in favor of the alternative hypothesis

?

Categorical Associations

Categorical Associations

The Pearson Product-Moment Correlation, "r," represents the covariance between two metric variables

But there are other kinds of associations

We've seen another kind before in the form of a contingency table: Does membership in a certain row have any connection to membership in a certain column?

Categorical Associations

The research question is about the independence of the rows and columns. Does the type of topping make any difference with respect to whether the toast lands up or down? In other words, is up/down independent of jelly/butter.

	Down	Up	Row totals
Jelly	20	10	30
Butter	30	40	70
Column totals	50	50	100

The Null Hypothesis for a Categorical Association

What is it about this table that makes it the "null hypothesis" (in other words that topping type and landing type are independent)?

	Down	Up	Row totals
Jelly	15	15	30
Butter	35	35	70
Column totals	50	50	100

Calculate the *expected value* for any cell by multiplying its row and column totals and dividing by the grand total

The Chi-Square Distribution and the Chi-Squared Test

Chi-Squared Represents the Differences Between Actual and Expected Cell Values

Actual (Observed) Values

	Down	Up
Jelly	20	10
Butter	30	40

Expected Values

	Down	Up
Jelly	15	15
Butter	35	35

	Down	Up
Jelly	$(20-15)^2/15$	$(10-15)^2/15$
Butter	$(30-35)^2/35$	$(40-35)^2/55$

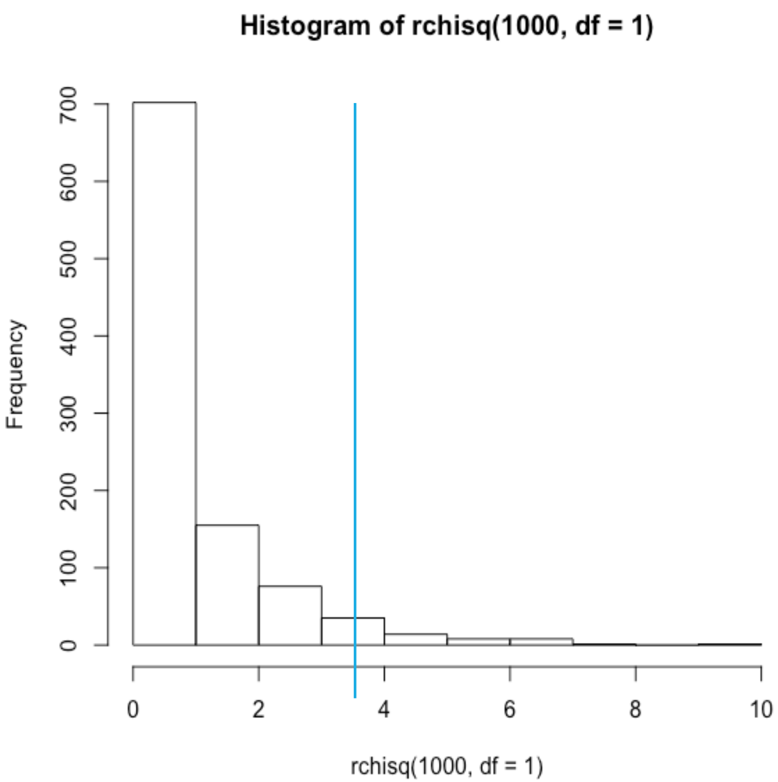
Sum these squared differences to get the chi-squared value

Chi-Squared Is a Random Distribution, Just Like "t" and "F"

Under the assumption of the null hypothesis (i.e., independence between rows and columns), chi-squared is a long-tailed distribution with an expected value = df.

The diagram at the right shows 1000 randomly sampled values of chi-squared for df=1. Why does the two-by-two contingency table have just one degree of freedom? (See next slide)

Looks at the blue vertical line at 3.84 on the x-axis. That is the point that divides this distribution into 95% on the left and 5% in the tail. Why is that position/value important?



School of Information Studies
Syracuse University

Calculating df in a Contingency Table

One must calculate the marginal totals to formulate the expected values and the chi-squared values

One degree of freedom is lost within each row and column for calculation of the marginal totals

The general formula is (rows-1)*(cols-1)

Try it here: with the marginal totals in place, only one cell is free to vary

	Down	Up	Row totals
Jelly	15		30
Butter			70
Column totals	50	50	100

Modeling Proportions With Bayesian Posterior Distributions

Proportions Instead of Cell Sizes

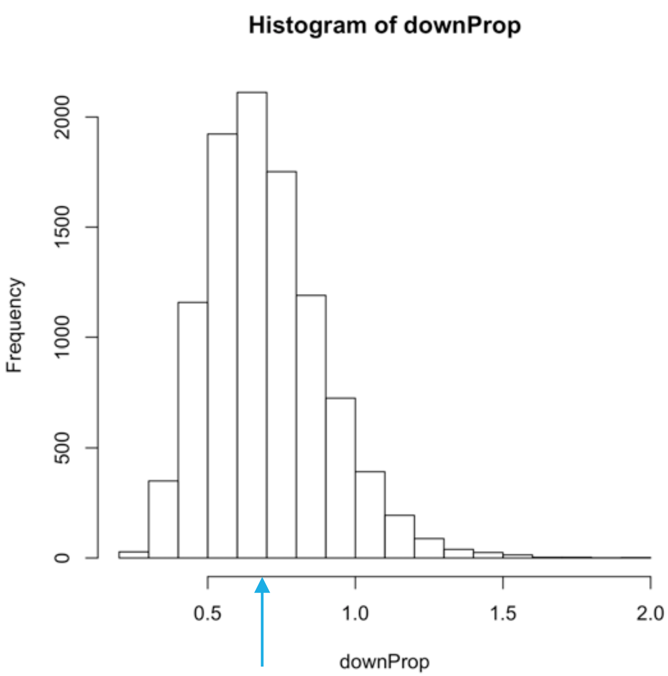
	Down	Up	Row totals
Jelly	20	10	30
Butter	30	40	70
Column tot	50	50	100

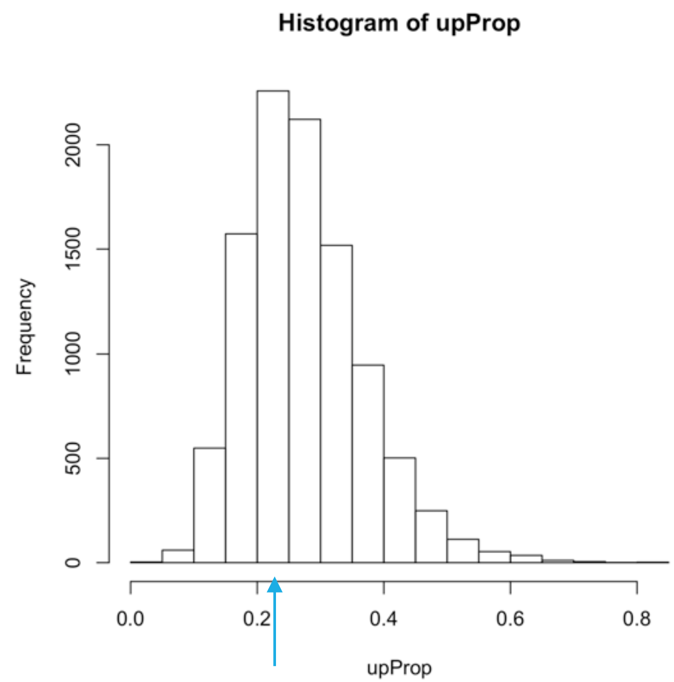
Observed ratio in this sample: 2/3 = 0.667

Observed ratio in this sample: 1/4 = 0.25

The Bayesian Approach: Model Proportions Across Conditions

Posterior distribution of 10,000 proportions from MCMC.
Left: ratio of JellyDown/ButterDown. Right: ratio of JellyUp/ButterUp.



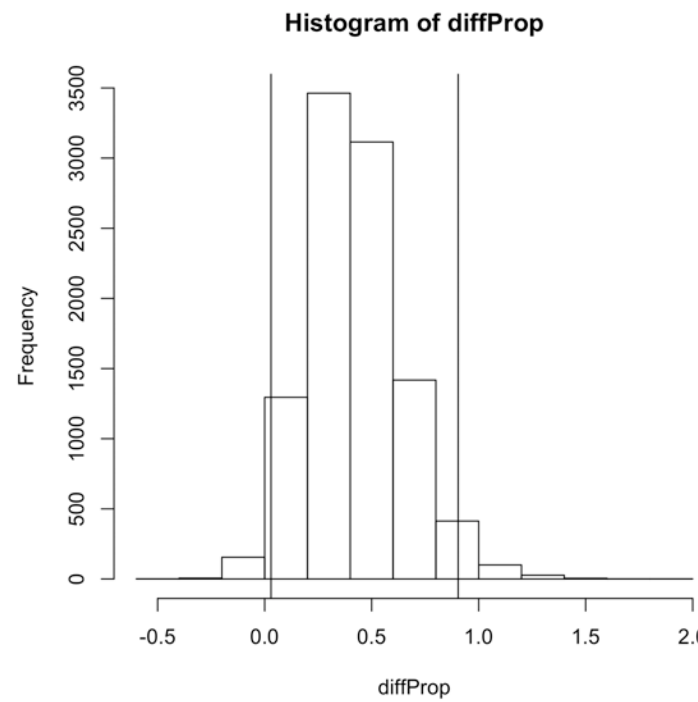


Making Difference in Proportions Visible

The figure to the left contains a histogram of the posterior distribution of differences in proportions between the two columns.

To put this idea into different words, this is how much the Jelly:Butter ratio decreases as we switch columns from down-facing toast (left column) to up-facing toast (right column).

The center of this distribution is a difference in proportions of 0.42. I've used abline() to put in vertical lines marking off the 95% HDI. The low end of the HDI is just barely above zero, while the top end of the HDI is just below one.



Repeat With Titanic Analysis

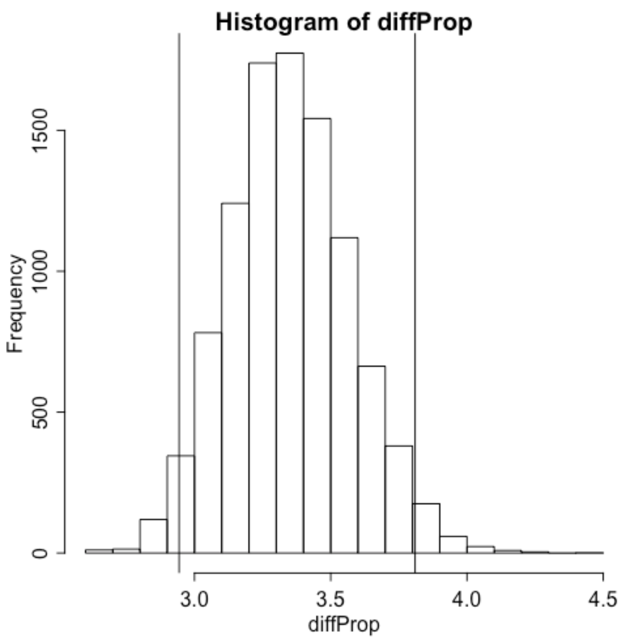
set.seed(314)

badBoatMF <- ftable(Titanic, row.vars=2, col.vars="Survived")

ctBFout <- contingencyTableBF(badBoatMF,sampleType="poisson", posterior=FALSE)

ctMCMCout <- contingencyTableBF(badBoatMF,sampleType="poisson" posterior=TRUE,iterations=10000)

maleProp <- ctMCMCout[,"lambda[1,1]"]/ctMCMCout[,"lambda[1,2]"]



School of Information Studies
Syracuse University

Repeat With Titanic Analysis

femaleProp <- ctMCMCout[,"lambda[2,1]"]/ctMCMCout[,"lambda[2,2]"]

diffProp <- maleProp - femaleProp

hist(diffProp)

abline(v=quantile(diffProp,c(0.025)), col="black")

abline(v=quantile(diffProp,c(0.975)), col="black")