

# Introduction

This page has a collection of 33 articles on topics in statistics that would be of interest to students who are taking Statistics after having had a course in Calculus. The articles range from one to 25 pages that attempt to consider some of the important principles of statistics from a more mathematical and deeper perspective than is typically seen in a first course in Statistics. The papers are listed in alphabetical order below with a short description of the topic and a note on the mathematical requirements.

## **A Closer Look at Blocking**

This six page paper uses a vector/matrix approach to ANOVA to explain in what way “blocking reduces variation”. The analysis of a randomized complete block is described along with the Least Significant Difference procedure for deciding which means differ when the null hypothesis has been rejected.

Students should be familiar with matrix representations and vector arithmetic.

## **ANOVA (One and Two Factor)**

This 25 page paper covers the basic concepts of experimental design with a development of the techniques of ANOVA. Randomized complete block designs, Latin square designs, one and two-factor designs with interactions are described with examples. Much of **A Closer Look at Blocking** is revisited in this paper.

Students should be familiar with matrix representations and vector arithmetic.

## **ANOVA for Regression**

This three page paper explains how to read and interpret the ANOVA table that is part of the computer output when a regression equation is computed.

The paper on **ANOVA (One and Two Factor)** should be read prior to considering these ideas.

## **Bernoulli to Binomial and Geometric**

This one page paper begins with a single Bernoulli trial, then finds the mean and variance of the Binomial distribution as repeated Bernoulli trials. The mean and variance of a binomial random variable are derived. Extending these results, the geometric distribution is defined and some clever use of differential calculus and series are required to derive the mean and variance formulas. This uses the expected value theorems found in the paper **Expected Value Theorems**.

### **Big Bang Problem**

This is an investigation for students that considers the data Edwin Hubble used in 1929 to develop the Big Bang theory. For the big bang theory to be supported by the data, the intercept of linear regression must not be significant. In Hubble's data it is. Students are asked to derive a formula for the least squares estimate of slope in the no-intercept form.

Students should understand the concept of a confidence interval for slope and intercept and be able to solve an optimization problem in a single variable.

### **Binomial To Normal Derivation**

This three page paper develops the normal distribution as an approximation of the binomial. Consider a situation in which you are trying to generate the value  $X$ . In the process, there are  $2n$  random errors each of size  $\delta$  being either positive or negative with equal probability, so the actual value generated, then, will vary from  $X - 2n\delta$  to  $X + 2n\delta$ . We investigate the probability of an error of size  $R = 2r\delta$ .

Heavy use of algebra and Sterling's approximation. No calculus is used in the derivation, but taking limits as  $n \rightarrow \infty$  and Sterling's approximation are essential. You must read **Calculus, Factorials, and Sterling's Approximation** before looking at this derivation.

### **Calc Derivation of Normal Dist**

This five page paper provides a derivation of the normal distribution as the distribution of errors using a continuous model.

Very heavy use of calculus. Integration by parts and separation of variables are required. The derivation is appropriate for BC calculus students.

### **Calculus, Factorials, and Sterling's Approximation**

This five page paper provides a generalization of the factorial as  $F(n) = \int_0^\infty x^n e^{-x} dx$  is derived and used to compute the half-factorials in the pdfs for the  $t$ , chi-square, and  $F$  distributions ( $3.5! \approx 11.6317$ ). This generalization is then used to derive Sterling's approximation,  $n! \approx \sqrt{2\pi n} n^n e^{-n}$ .

Heavy use of calculus, but the calculus is not difficult.

### **Capture-Recapture (Estimating the Size of a Population)**

This four page paper gives some formulas for estimating the variance in the capture-recapture setting. The variance formulae are not derived. Inverse sampling is also discussed.

No special mathematics is used.

### **Central Limit Theorem**

This six page paper presents a proof of the Central Limit Theorem using generating functions. Generating functions are presented in **Normal + Normal**, which should be read first.

In addition to moment generating functions, polynomial expansions, Taylor series, and limits at infinity are required.

### **Challenger Revisited**

This two page paper uses logistic regression to develop a model to estimate of the probability of an O-ring failure based on the results of previous launches of the space shuttle.

Logistic regression is used so that section of **Re-expression and Regression** should be read first.

### **Chi-Square Analyses**

This seven page paper describes the mathematical model leading to the Chi-square computational formula, uses simulations to illustrate why each expected cell count must be larger than 5, and compares the structure of a test of independence to a test of homogeneity. Degrees of freedom are also discussed.

No mathematics is needed.

### **Chi-Square and Proportion Tests**

This three page paper shows the mathematical equivalence in Chi-square tests and one and two proportion tests.

No calculus, but a very heavy use of algebra, factoring, and regrouping with rational expressions.

### **Estimating Parameters (Simple, Stratified, and Cluster Samples)**

This 19 page paper gives the computational formulas for estimating means and proportions from simple, stratified, and cluster samples. Some formulae are derived, others given. Ratio estimation is also included.

No calculus, but very challenging algebra based on expected value theorems.

### **Estimators and the Taxi Problem**

This five page paper presents a student investigation that requires student to devise an estimator for the largest in a list. Simulations are used to show biased and unbiased estimators and the variability of an estimator.

No calculus needed but an understanding of a uniform distribution is required.

### **Expected Value Theorems**

This four page paper is a theorem-proof development of the basic theorems of expected values. For example, we prove that  $E(aX) = aE(X)$  and that  $Var(aX) = a^2Var(X)$ . Of importance to the course are the proofs that  $Var(X + Y) = Var(X) + Var(Y)$ , that  $Var(\bar{X}_n) = \frac{Var(X)}{n}$ . The paper culminates in a proof that  $E(s^2) = \sigma^2$ , which explains why  $n - 1$  is the proper divisor for the sample variance.

No calculus is needed, but students need to be able to follow a mathematical argument.

### **Hypergeometric to Binomial**

This three page paper shows that we can approximate the hypergeometric distribution (sampling without replacement) with the binomial distribution (sampling with replacement).

No calculus is needed, but students need to be able to follow a mathematical argument.

### **Hypothesis Testing and Power using the Binomial Distribution**

This five page paper presents the concepts of hypothesis testing and power in the context of a binomial setting, so the actual probabilities can be computed and compared.

No calculus, but binomial probabilities are used throughout along with conditional probability.

### **Chebychev's Theorem and the Law of Large Numbers**

This four page paper proves the Weak Law of Large Numbers by first proving Chebychev's Theorem and applying it to a sample mean. It also proves Markov's Inequality using the same technique.

This uses no calculus, but students should recognize an integral as an area under a curve and understand limits at infinity.

### **Log(Odds) and Quality Control**

This six page paper uses tree diagrams, conditional probability, and the rules of logarithms to describe a quality control process used in industry.

Students must be comfortable using trees and computing conditional probabilities.

### **Normal + Normal**

This four page paper proves that the sum and difference of two normally distributed random variables is normally distributed. This is an essential in many of the inference computations. We define a moment generating function, use it to find the moment generating function for  $N(\mu, \sigma)$ , and then prove the result.

Requires differential and integral calculus and lots of algebra (completing the square and rules of exponents), but no difficult techniques.

### **PDF's**

This one page paper gives the formulas for the probability distribution or density functions for the standard distributions: binomial, geometric, normal,  $t$ , chi-square, and  $F$ .

### **$r$ and $r^2$**

This nine page paper gives a mathematical treatment of the interpretations of the correlation coefficient,  $r$ , and the coefficient of determination  $r^2$ .

No calculus is needed but there is lots of messy algebra with summations and squares.

### **Re-expression and Regression**

This sixteen page paper is in two parts. The first 10 pages considers the reasons for re-expressing data and techniques for comparing two competing models. Techniques for finding vertical shifts are also considered. Logistic regression and a sequential multivariable fit to the Wind Chill Table are also presented.

Students must be comfortable with function transformations from Precalculus.

### **Sampling Methods and Planning a Survey**

This nine page paper gives the terminology and components of a planned survey. The process used by the New York Times and the Gallop poll are described. Probability samples and sources of error are described.

No mathematics.

### **Scope of Inference**

This four page paper gives examples of the kinds of inference that are possible based on the application of random selection and random allocation in a study. The selection of units can be done by random sampling or not. The allocation of units to treatment groups can be done by randomization or not. The four resulting structures resulting different scopes of inference and assessment of cause and effect.

No mathematics is used in this paper.

### **Sensitivity and Specificity**

The sensitivity (the probability that the test says a person has the disease when in fact they do have the disease) and the specificity (the probability that the test says a person does not have the disease when in fact they are disease free) of a test is explained with an example.

No calculus or difficult mathematics.

### **Standard Errors for Regression**

This five page paper derives the formulas for the standard error of the slope, the intercept, as well as prediction intervals for the mean response for a given  $x$  and individual values in regression.

No calculus, but students must be comfortable with the manipulations and results in the paper **Expected Value Theorems**.

### **t converges to N**

This five page paper uses L'Hopital's Rule to show that the pdf for the  $t$ -distribution converges to the pdf for the standard normal distribution as the degrees of freedom increases without bound.

Students need to know L'Hopital's rule for the forms  $1^\infty$  and  $0 \cdot \infty$ .

### **Tests of Hypotheses**

This twelve page paper uses some simulation results to develop the central limit theorem and to present an overview of the concepts of a confidence interval for proportions. The condition of  $np > 5$  and  $n(1-p) > 5$  is addressed. These concepts are modified to consider hypothesis tests using the  $t$  distribution, including matched pairs and two-sample tests.

No special mathematical background is needed. Some familiarity with summation notation is helpful.

### **Two Formula for Correlation**

This four page paper shows how to derive the formula in terms of  $z$ -scores,  $r = \frac{z_x z_y}{n-1}$ ,

from the formula related to the regression line.  $r = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ .

An optimization problem is solved along with a lot of algebra with sums.

### **Vector Approach to Regression**

This nine page paper derives the regression coefficients for linear regression using a vector or geometric approach. The correlation coefficient is seen in terms of the angle between vectors. The results are generalized to multiple regression for both a single variable and several variables.

Students must be comfortable with vector addition and vector multiplication. They should be able to write a system of linear equations in matrix form and solve the system using a matrix inverse.

### **Wind Chill Table Investigation**

This is a student investigation that derives a multivariable model for the relationship between the wind velocity and the actual temperature and the wind chill temperature. First we model the relationship between actual temperature and wind chill temperature for fixed wind velocities with linear equations. Then we determine the relationships between the wind velocity and the slopes and intercepts of those lines.

No calculus, but students need to be able to work with data that has a vertical shift, so **Re-expression and Regression** should be covered first.