**Problem Set #3**

**Geog 2000: Introduction to Geographic Statistics**

**Instructor: Dr. Paul C. Sutton**

**#1) Sampling Design**

|  |
| --- |
| 1. **Define and provide an example of the following terms: Population, Sample, Parameter, Random Selection, Sampling Frame**
 |
|  |
| 1. **Define and provide an example of the following sampling approaches: Simple Random Sampling, Stratified Random Sampling, Cluster Sampling, and Systematic sampling.**
 |
|  |
| 1. **Explain the two important aspects of Simple Random Sampling: *Unbiasedness* and *Independence***
 |
|  |
| 1. **Suppose you wanted to estimate the fraction of people of voting age in Colorado that believe the landmark Supreme Court decision regarding abortion (Roe v. Wade) should be overturned. (BTW: An interesting factoid I discovered in the Literature review for my Master’s Thesis was this – Only 30% of American adults could answer the question: “*Roe vs. Wade was a landmark Supreme Court decision regarding what?*”). You use all the names and phone numbers in all the yellow page phone books of the state. You randomly sample 1,000 people from this these phonebooks and ask them if they want Roe v. Wade overturned. Is this sampling approach a good one? What is the sampling frame? Is this an unbiased approach to sampling? Explain why or why not.**
 |
|  |

**#2) Condom manufacturing as a Bernoulli Trial**

|  |
| --- |
| **Take a statistical approach to characterize the ‘effectiveness’ of your condoms based on your test results. Define your population and unknown parameter(s); find a statistic that estimates this parameter (an estimator) and the theoretical sampling distribution, mean, and standard deviation of this estimator; and use the data above as a random sampling of your product; finally, report and interpret your results and the statistical or sampling error associated with it.** |
|  |
| 1. **How will your estimate of the failure rate of your condoms change if you only sampled 100 condoms?**
 |
|  |
| 1. **Explain why your actual ‘condom failure’ in real use (e.g. someone gets pregnant) might be significantly higher or lower than your estimates of the fraction of your product that has a hole in it.**
 |
|  |
| 1. **Suppose you made 100 condoms 1,000 times and estimated ‘p-hat’ 1,000 times (once for each batch of 100). What would the probability mass function (pmf) look like for ‘p-hat’ where ‘’p-hat’ = # of condoms with holes / 100?**
 |
|  |
| 1. **What if your batches were of size N= 10; how would the pmf change? What if your batch sizes were N=10,000; how would the pmf change?**
 |
|  |
| 1. **How does Sample Size improve parameter estimation of ‘p’?**
 |
|  |
| 1. **What assumptions are you making with respect to estimation of ‘p’?**
 |
|  |
| **G)** **Provide the formulas (i.e. estimators) for your estimates of both ‘p’ and σ(p).**  |
|  |
|  **H) Use the idea of ‘Sampling Error’ to explain exactly what it would mean to have an estimate of p (i.e. ‘p-hat’) = 0.67 and an estimate of σ(p) = .07.**  |
|  |

**#3) Black Velvet, Wildfire, and Mr. Ed were not ordinary horses**

|  |  |
| --- | --- |
| 1. **What is the true population mean (μ)?**
 |  |
| 1. **What is the true population variance (σ2)?**
 |  |
| 1. **What is the true population standard deviation (σ)?**
 |  |
| 1. **What was your estimate of the mean (x-bar)?**
 |  |
| 1. **What was your estimate of the variance (s2)?**
 |  |
| 1. **What probability distribution function (pdf) would best describe A sampling distribution of the mean for this random variable Assuming a sample size of N = 100?**
 |  |
| 1. **What pdf would best characterize the sampling**

**distribution of The mean of this random variable** **if the sample size was N = 10?** |  |
| 1. **What estimators did you use for μ and σ2 ?**
 |  |
| 1. **Did your estimates of these population parameters seem reasonable Or are they wrong enough to suggest your sampling methodology was biased? Explain.**
 |  |

**#4) The Central Limit Theorem**

|  |
| --- |
| 1. **Summarize the Central Limit Theorem in your own words**
 |
|   |
| **B) Given the pdf of a continuous random variable that looked like the figure** **Below w/ known μ = 100 and known σ2  = 25 , what would the sampling**  **Distribution of this distribution’s mean look like for samples of N = 49?** |
|  |
|  **C) Given the pdf of a continuous random variable that looked like the figure** **Below w/ known μ = 100 and known σ2  = 25 , what would the sampling**  **Distribution of this distribution’s mean look like for samples of N = 49?** |
|  |

**#5) William Gosset, The Guinness Brewery, and the Student’s ‘t’ distribution**

|  |
| --- |
| **The student’s ‘t’ (aka ‘t’ distribution of ‘t-test’ fame) looks a lot like the standard Normal (aka N(0,1) or ‘Z’) distribution. It looks very similar, symmetric, bell shaped etc. but it’s tails are a little ‘fatter’. Explain in your own words the student’s ‘t’ distribution. In your explanation provide a mathematical formula for the ‘t’ distribution, define and explain it’s parameter or parameters, and answer the following questions: 1) A ‘t’ distribution with how many degrees of freedom is identical to a N(0,1) distribution? And 2) Is the ‘t’ distribution more leptokurtic or platykurtic than the normal distribution?** |
|  |

**#6) Dream Job: Logistical Statistician for Marborg Disposal - NOT!**

|  |
| --- |
| **A) Estimate the mean and the 95% Confidence Interval for the mean amount of trash (in lbs) your average customer produces in a given week. Interpret your result.** |
|  |
| 1. **How many customers would you schedule for each truck if you wanted to be 99% confident that any given truck would not exceed it’s weight limit? Note #1: When you add random variables (see pp 68-71 in Cartoon Guide) the variances add also. Note #2: Is this a one-tailed or two-tailed test?**
 |
|  |
| 1. **What kind of geographic and socio-economic-demographic information might suggest you tweak the # of customers per truck on a route-based basis in order for You to be most efficient? Also, do you think this statistical approach is really Reasonable for a trash hauling company?**
 |
|  |

**#7) Dead Men Tell No Tales**

|  |
| --- |
| **Suppose over the years you meet 100 people who have opinions about passing cars by driving over the double yellow lines. Seventy of them think it’s crazy. Thirty claim they do it all the time and think its fine. How might this be a biased sample?** |
|  |

**#8) Alpha Inflation and Stockbroker Scams?**

|  |
| --- |
| **A stockbroker identifies 2,000 wealthy potential customers and sends them a newsletter telling ½ of them to buy stock ‘X’ and the other half to sell stock ‘X’. Stock ‘X’ then goes up. He then sends another newsletter to the ½ he told to by stock ‘X’. He sends 500 of them advice to buy stock ‘Y’ and sell advice to the other 500. He keeps doing this until he is down to about 10 or so customers and moves in for the kill. ‘Look at what a track record I have?’ etc. Google the statistical concept of ‘alpha inflation’ and relate it to this story and Explain. (An interesting book by Nicholas Nassim Taleb called: The Black Swan: The high impact of low probability events explores these ideas in many interesting ways).** |
|  |

**#9) Capital Crimes and Type I and Type II Errors**

|  |
| --- |
| **Imagine a murder trial. The defendant may or may not be guilty. The jury may or may not convict him or her. What are all the possibilities? Make an analogy between these possibilities and a typical statistical test involving hypothesis testing and explain type I and Type II error. Be sure to incorporate α and β into your answer.** |
|  |

**#10) Hypothesis Testing and flipping a Bogus (?) Coin?**

|  |
| --- |
| **Suppose someone hands a strange looking coin to you and claims it has a 50 / 50 chance of landing heads or tails (i.e. it’s a ‘fair’ coin). Create a Null Hypothesis (H0) and an Alternative Hypothesis (Ha) for this claim and test the coin with 10 flips. What will your 95% confidence decision rule be (i.e. what results of this experiment will suggest that you reject the Null Hypothesis)?** |
|  |

**#11) Wetlands, Wal-Mart, and Weasels**

|  |
| --- |
| 1. **Do either of these wetlands meet the standard with 95% Confidence? Explain**
 |
|  |
| 1. **Are the tests you are conducting 1–tailed or 2-tailed?**
 |
|  |
| 1. **What is better to use in this case a ‘t’ distribution or a normal?**
 |
|  |

**#12) Reading the Standard Normal Table revisited…**

|  |
| --- |
| 1. **For what values of ‘Z’ (assume they are equal with one being negative and the other positive) will the white area under the curve depicted below be equal to 0.95?**
 |
|  |
| 1. **How does this value of ‘Z’ relate to the creation of 95% Confidence Intervals? Explain.**
 |
|  |
|  **C) What would ‘Z’ have to be for the same area to be equal to 0.99?** |
|  |

|  |
| --- |
|  **D) The dark area on the figure below spans the N(0,1) curve from a value of 0 to 1.33. What is the area of this dark space?**  |
|  |
|  **E) What is the area of the curve from Z = 1.5 to infinity as depicted in the image below?** |
|  |
|  **F) What is the area of the curve from Z = 1 to Z = 2 as depicted in the image below?** |
|  |

**#13) Polls, Pols, and the Binomial Distribution**

|  |
| --- |
| 1. **The true fraction of registered democrats that would state in a phone survey on that day that they would vote for Barack Obama is unknown (an unknown population parameter that could only be ascertained by asking every registered democrat in the state of Texas – we’ll get to know that value on election day). However, this survey has produced an estimate of that fraction. What is the estimate in percentage terms and what is your 95% Confidence Interval regarding that estimate? Explain. Show your formulas and calculations.**
 |
|  |
| 1. **When the actual vote takes place it is very unlikely that Barack Obama would get 54.7% of the vote. List 5-10 distinct reasons why the actual vote may turn out differently.**
 |
|  |
| 1. **The TV weatherperson claims there’s a 75% chance of snow tomorrow. Tomorrow comes and it doesn’t snow. Homer Simpson says the TV weatherman is a moron! Was the weatherman wrong? Explain the meaning of confidence intervals and the meaning of % chance of ‘X’ predictions.**
 |
|  |

**#14) Avocado Agriculture, Urea, and a Difference of Means Test**

|  |
| --- |
| 1. **Given the data presented above make a conclusion as to whether or not the urea improved the productivity of your avocado trees. (The numbers are kilograms of avocado per tree).**
 |
|  |
| 1. **Is this a one-sample or a two-sample test?**
 |
|  |

**# 15) Weight Watchers vs. The Atkins Diet**

|  |
| --- |
| **Perform and explain the results of the appropriate t-test after considering group variances.** |
|  |

**Computer Problems (use JMP and/or Excel for these exercises)**

**#16) Are Binomials always symmetric?**

|  |
| --- |
| **Generate a Bin(.5, 100) w/ N=500 and Generate a Bin(.95, 100) with N=500. Compare the histograms. How do hard edges (e.g. you can’t go below zero but you can go way more than twice the mean – (e.g. Rainfall data for a given month) influence the symmetry of some commonly measured distributions (e.g. housing prices, rainfall data, personal incomes, personal net worth)? Explain.** |
|  |

**#17) Sponge Bob Square Pants and Blood Pressure**

|  |
| --- |
| **Suppose someone hypothesized that watching ‘Sponge Bob Square Pants’ cartoons was an effective means of reducing someone’s blood pressure. An experiment is designed in which fifteen people are subjected to the McNeil Lehrer News Hour on the first day of the trial and their blood pressure is measured when they watch the show. They are then subjected to an hour’s worth of ‘Sponge Bob Square Pants’ cartoons and the BP is measured. On the third day they simply have their blood pressure measured without watching any TV. Statistically analyze the data below and come to a conclusion.** |
|  |

**#18) Skewed Distributions: Housing Prices & the Sampling Distribution of the mean**

|  |
| --- |
| **Generate your own set of 1,000 numbers that you think are a reasonable sample of housing prices throughout the United States. (Note: This should be skewed high – e.g. you have a big bump centered around ~$250,000 (?) that doesn’t go very far below $20,000 but it goes way up to multi-million dollar properties. Create and interpret a histogram of this dataset you create – Note the relative size of the median and the mean. (You might want to play around with the random number generating capability of JMP here ☺). Now – figure out a way to ‘sample’ 100 records from your dataset – 100 times (It’s ok if you sample some records more than once between these samples). Now, calculate the mean of each of your 100 samples. Plot and characterize (definitely use a histogram) the distribution of your 100 sampled means. Explain what you see.** |
|  |

|  |
| --- |
| **#17) Write a 4 to 7 sentence summary of Chapter 9: *How to statisticulate*** |
|  |
| **# 18) Myron Dorkoid did a study in which he randomly sampled 1,000 Americans of all ages, genders, etc. He found that there is a strong positive correlation between shoe size and personal annual income. Is this correlation possible? Can you explain it? Myron went off to a podiatrist to explore the idea of having his feet stretched in order to increase his income. Explain why his reasoning (regarding the foot stretching idea) is flawed and tie it to one of the principles described in Darrell Huff’s *How to Lie With Statistics*.** |
|  |
| **#19) Brenda ‘Bucket’ O’hare just had a baby that was born with a weight of 7 lbs and 7 ozs. She plotted the weight of her baby for the first three months of its life (note: these numbers are not that unusual for most babies):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Month** | **0** | **1** | **2** | **3** |
| **Weight** | **7lbs 7ozs** | **8 lbs 10 ozs** | **10 lbs 4 ozs** | **14 lbs 14 ozs** |

**She calculated that she’d have a 240 lb three year old and that her child was a freak of either nature or that Alien that she was probed by. Consequently she gave up her child for adoption. Explain why her reasoning is flawed and tie it to one of the principles described in Darrell Huff’s *How to Lie With Statistics*.** |
|  |