**Tor Thogersen**

**DA 460 – Fall 2017**

**Lab 6 - Handout 6 R and Handout 6 SAS**

**Part 6 – R Handout - Inference for numerical data**

**Exercise 1:**

1. What are the cases in this data set?
   1. North Carolina Births

> names(nc)

[1] "fage" "mage" "mature" "weeks"

[5] "premie" "visits" "marital" "gained"

[9] "weight" "lowbirthweight" "gender" "habit"

[13] "whitemom"

1. How many cases are there in our sample?
   1. 1000

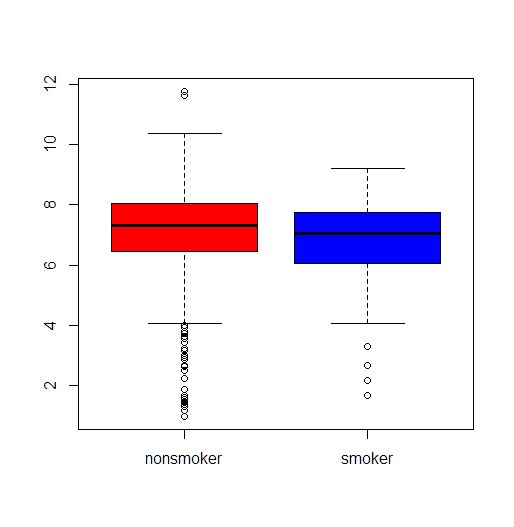
> dim(nc)

[1] 1000 13

**Exercise 2:**

1. Make a side-by-side boxplot of habit and weight.

> boxplot(nc$weight~nc$habit, col=c("red", "blue"))



> by(nc$weight, nc$habit, mean)

nc$habit: nonsmoker

[1] 7.144273

----------------------------------------------------------

nc$habit: smoker

[1] 6.82873

1. What does the plot highlight about the relationship between these two variables?
   1. Plot shows how the means birth weight of the two dist, compares shows difference in birth weight between non-smoker and smokers, it shows the avg mean birth weight for non-smoker is higher than the avg mean birth weight for a smoker.

**Exercise 3:**

1. Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same by command above but replacing mean with length.
   1. Yes, based the by function we compared the birth length of non-smokers to smokers and avg length for non-smoker was significantly higher than comparable smokers length of the child at birth.

> by(nc$weight, nc$habit, length)

nc$habit: nonsmoker

[1] 873

----------------------------------------------------------

nc$habit: smoker

[1] 126

**Exercise 4:**

1. Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.
   1. Ho: Birth weight of Non-smoker = Birth Weight of Smoker

Ha: Birth weight of Non-smoker ≠ Birth Weight of Smoker

Are birth rates of new born in NC the same for mothers that did not smoke during pregnancy versus mothers that did smoke during pregnancy.

We reject the Ho because the p-value = 0.0184 > alpha .05, thus significant evedice that the mean values for birth of smokers and nonsmokers are not equal.

> inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

Response variable: numerical, Explanatory variable: categorical

Difference between two means

Summary statistics:

n\_nonsmoker = 873, mean\_nonsmoker = 7.1443, sd\_nonsmoker = 1.5187

n\_smoker = 126, mean\_smoker = 6.8287, sd\_smoker = 1.3862

Observed difference between means (nonsmoker-smoker) = 0.3155

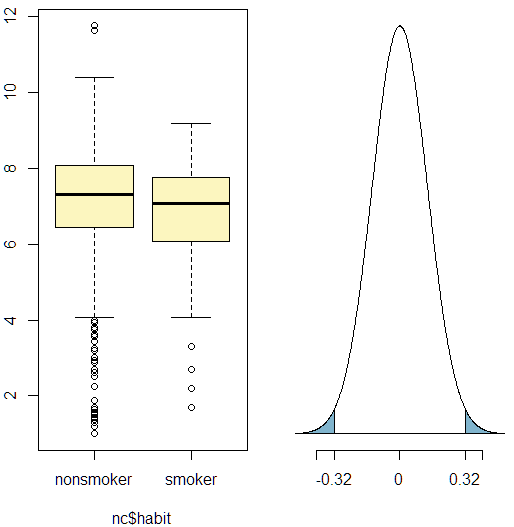
H0: mu\_nonsmoker - mu\_smoker = 0

HA: mu\_nonsmoker - mu\_smoker != 0

Standard error = 0.134

Test statistic: Z = 2.359

p-value = 0.018



**Exercise 5:**

1. Change the type argument to "ci" to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

> inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0, alternative = "twosided", method = "theoretical", order = c("smoker", "nonsmoker"))

Response variable: numerical, Explanatory variable: categorical

Difference between two means

Summary statistics:

n\_smoker = 126, mean\_smoker = 6.8287, sd\_smoker = 1.3862

n\_nonsmoker = 873, mean\_nonsmoker = 7.1443, sd\_nonsmoker = 1.5187

Observed difference between means (smoker-nonsmoker) = -0.3155

Standard error = 0.1338

95 % Confidence interval = ( -0.5777 , -0.0534 )

**Part 6 – R Handout - Inference for categorical data**

**Exercise 1:**

1. In the first paragraph, several key findings are reported. Do these percentages appear to be sample statistics (derived from the data sample) or population parameters?
   1. The data looks to come from sample data that represents the population, the data was gathered by Win-Gallup International.

**Exercise 2:**

1. The title of the report is “Global Index of Religiosity and Atheism”. To generalize the report’s findings to the global human population, what must we assume about the sampling method? Does that seem like a reasonable assumption?
   1. Random polling of the population to gather the required sample
   2. Sample must be large enough to meaning represent the population true mean
   3. Questions should be similar for each country represented in the study

**Exercise 3:**

1. What does each row of Table 6 correspond to?
   1. Country of Origin, sample size for country, response in the survey question: religious, not religious, atheist, no response
2. What does each row of atheism correspond to?
   1. Country, response, year

**Exercise 4:**

1. Using the command below, create a new dataframe called us12 that contains only the rows in atheism associated with respondents to the 2012 survey from the United States. Next, calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?
   1. Yes, because we are sampling the same sample data so it should have the same porportion

> us12 <- subset(atheism, nationality == "United States" & year == "2012")

> summary(us12$response)

atheist non-atheist

50 952

> 50/952

[1] 0.05252101

**Exercise 5:**

1. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met
   1. Sample must be greater than 30 with a normally distributed sample

> inference(us12$response, est = "proportion", type = "ci", method = "theoretical",

+ success = "atheist")

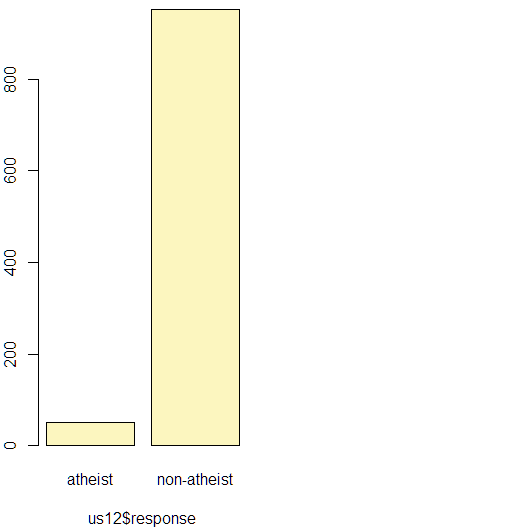
Single proportion -- success: atheist

Summary statistics: p\_hat = 0.0499 ; n = 1002

Check conditions: number of successes = 50 ; number of failures = 952

Standard error = 0.0069

95 % Confidence interval = ( 0.0364 , 0.0634 )



**Exercise 6:**

1. Based on the R output, what is the margin of error for the estimate of the proportion of the proportion of atheists in US in 2012?
   1. ME: 0.013524

> inference(us12$response, est = "proportion", type = "ci", method = "theoretical",

+ success = "atheist")

Single proportion -- success: atheist

Summary statistics: p\_hat = 0.0499 ; n = 1002

Check conditions: number of successes = 50 ; number of failures = 952

Standard error = 0.0069

95 % Confidence interval = ( 0.0364 , 0.0634 )

> 1.96\*.0069

[1] 0.013524

**Exercise 7:**

1. Using the inference function, calculate confidence intervals for the proportion of atheists in 2012 in two other countries of your choice, and report the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create new data sets for each of the two countries first, and then use these data sets in the inference function to construct the confidence intervals.
   1. Spain:
      1. 95 % Confidence interval = ( 0.0734 , 0.1065 )
      2. ME: > 1.96\*0.0085 [1] 0.01666
   2. India:
      1. 95 % Confidence interval = ( 0.0201 , 0.0404 )1.96
      2. ME: > 1.96\*0.0052 [1] 0.010192

> spain12 <- subset(atheism, nationality == "Spain" & year == "2012")

> india12 <-subset(atheism, nationality == "India" & year == "2012")

> View(spain12)

> summary(spain12)

nationality response year

Spain :1145 atheist : 103 Min. :2012

Afghanistan: 0 non-atheist:1042 1st Qu.:2012

Argentina : 0 Median :2012

Armenia : 0 Mean :2012

Australia : 0 3rd Qu.:2012

Austria : 0 Max. :2012

(Other) : 0

> p <-103/1042

> p

[1] 0.09884837

> inference(spain12$response, est = "proportion", type = "ci", method = "theoretical",

+ success = "atheist")

Single proportion -- success: atheist

Summary statistics: p\_hat = 0.09 ; n = 1145

Check conditions: number of successes = 103 ; number of failures = 1042

Standard error = 0.0085

95 % Confidence interval = ( 0.0734 , 0.1065 )

> india12 <-subset(atheism, nationality == "India" & year == "2012")

> View(india12)

> summary(india12)

nationality response year

India :1092 atheist : 33 Min. :2012

Afghanistan: 0 non-atheist:1059 1st Qu.:2012

Argentina : 0 Median :2012

Armenia : 0 Mean :2012

Australia : 0 3rd Qu.:2012

Austria : 0 Max. :2012

(Other) : 0

> p <-33/1059

> inference(india12$response, est = "proportion", type = "ci", method = "theoretical", success = "atheist")

Single proportion -- success: atheist

Summary statistics: p\_hat = 0.0302 ; n = 1092

Check conditions: number of successes = 33 ; number of failures = 1059

Standard error = 0.0052

95 % Confidence interval = ( 0.0201 , 0.0404 )

**Exercise 8:**

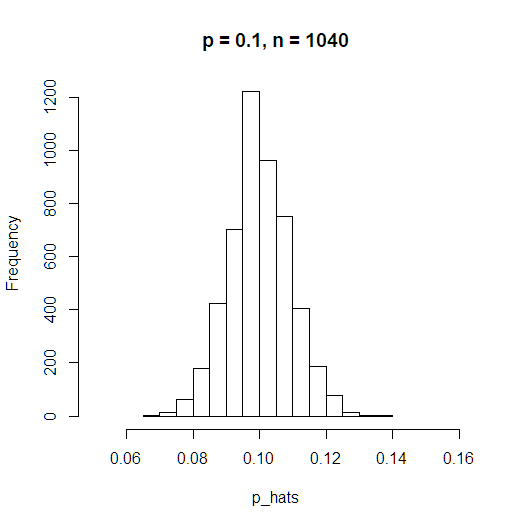
1. Describe the relationship between p and me.
   1. As proportion goes up the me will decrease and as proportion goes down me will increase

**Exercise 9:**

1. Describe the sampling distribution of sample proportions at n=1040 and p=0.1. Be sure to note the center, spread, and shape. Hint: Remember that R has functions such as mean to calculate summary statistics.
   1. Sample dist is approximately normally dist, with mean (center) at 0.1000902, the distribution has a bell shaped curve that is evenly spread from center, the spread is from .06 to .14 from center (0.1000902).
2. > p <- 0.1
3. > n <- 1040
4. > p\_hats <- rep(0, 5000)
5. >
6. > for(i in 1:5000){
7. + samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))
8. + p\_hats[i] <- sum(samp == "atheist")/n
9. + }
10. >
11. > hist(p\_hats, main = "p = 0.1, n = 1040", xlim = c(0.05, 0.16))

> mean(p\_hats)

[1] 0.1000902



**Exercise 10:**

1. Repeat the above simulation three more times but with modified sample sizes and proportions: for n=400 and p=0.1, n=1040 and p=0.02, and n=400 and p=0.02. Plot all four histograms together by running the par(mfrow = c(2, 2)) command before creating the histograms. You may need to expand the plot window to accommodate the larger two-by-two plot. Describe the three new sampling distributions. Based on these limited plots, how does n appear to affect the distribution of p^? How does p affect the sampling distribution?
   1. Increasing the sample size did not really change the distribution of the model much, both samples are still approximately normally distribution, with a bell shape curve, the spread is evenly dispersed from center.
   2. Changing the P^ makes a huge difference in the range from center, for both models. When changing the from .1 to .2 you see the range min and max both increase.

> par(mfrow = c(1, 2))

> p <- 0.1

> n <- 400

> p\_hats <- rep(0, 5000)

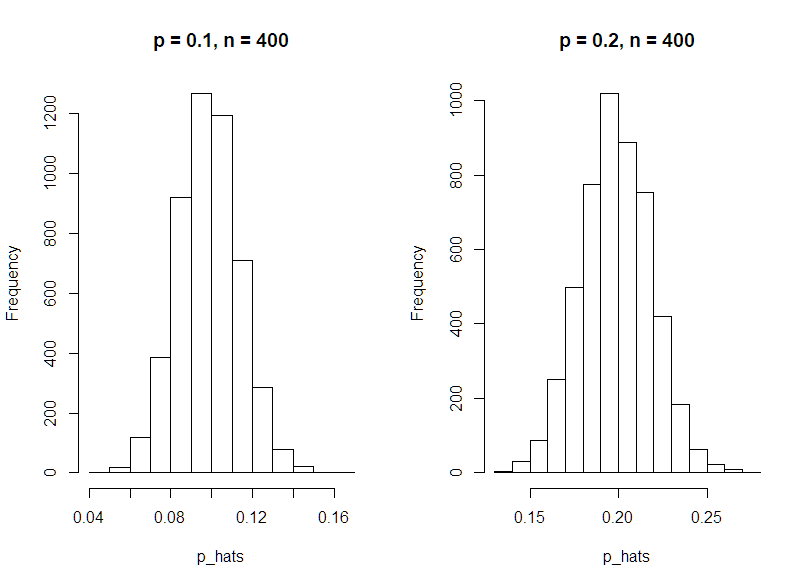
> for(i in 1:5000){

+ samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))

+ p\_hats[i] <- sum(samp == "atheist")/n

+ }

> hist(p\_hats, main = "p = 0.1, n = 400")



> p <- 0.2

> n <- 400

> p\_hats <- rep(0, 5000)

> for(i in 1:5000){

+ samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))

+ p\_hats[i] <- sum(samp == "atheist")/n

+ }

> hist(p\_hats, main = "p = 0.2, n = 400")

> par(mfrow = c(1, 2))

> p <- 0.1

> n <- 1040

> p\_hats <- rep(0, 5000)

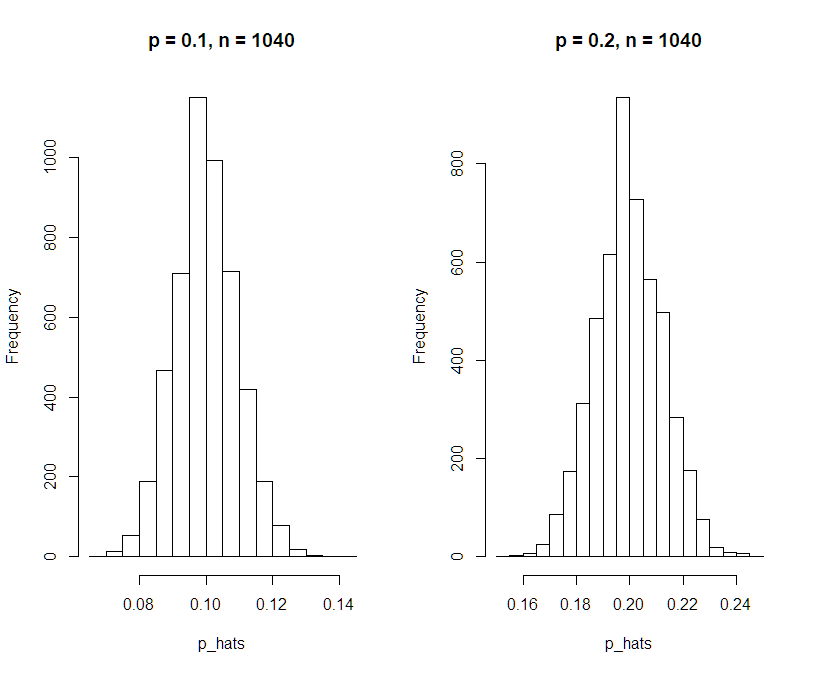
> for(i in 1:5000){

+ samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))

+ p\_hats[i] <- sum(samp == "atheist")/n

+ }

> hist(p\_hats, main = "p = 0.1, n = 1040")



> p <- 0.2

> n <- 1040

> p\_hats <- rep(0, 5000)

> for(i in 1:5000){

+ samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))

+ p\_hats[i] <- sum(samp == "atheist")/n

+ }

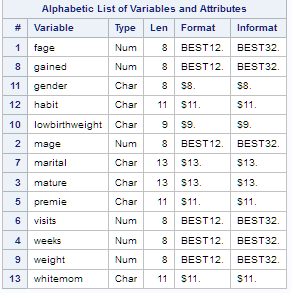
> hist(p\_hats, main = "p = 0.2, n = 1040")

**Exercise 11:**

1. If you refer to Table 6, you’ll find that Australia has a sample proportion of 0.1 on a sample size of 1040, and that Ecuador has a sample proportion of 0.02 on 400 subjects. Let’s suppose for this exercise that these point estimates are actually the truth. Then given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference and report margin of errors, as the reports does?
   1. No, because the sample ME will be way off based on the sample proportions

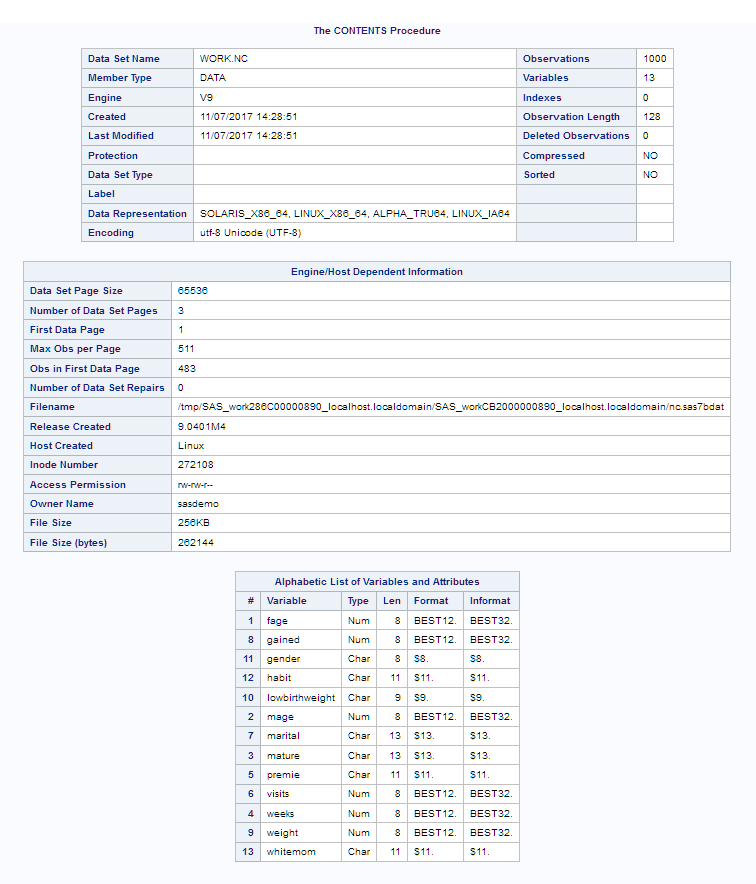
**Part 6 - SAS Handout - Inference for Numerical Data**

**Exercise 1:**

1. What are the characteristics of the babies represented in this data set? How many cases are there in our sample?
   1.  1000

proc contents data=nc;

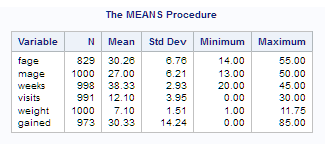
run;

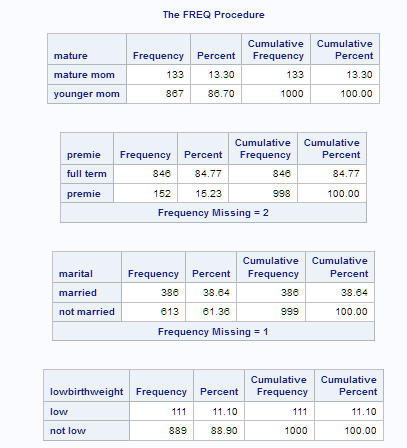
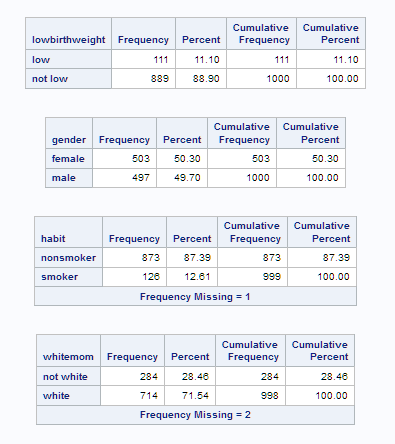


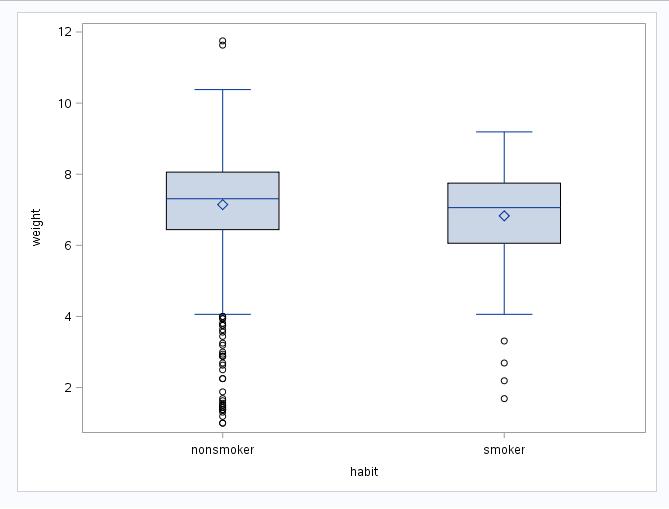
**Exercise 2:**

1. Create a side-by-side box plot of habit and weight. What does the plot highlight about the relationship between these two variables?
   1. Relationship between the two variables nonsmoker vs smoker means

proc means data=nc maxdec=2;  
 var fage mage weeks visits weight gained;  
run;

  
  
proc freq data=nc;  
 tables mature premie marital lowbirthweight gender habit whitemom;  
run;

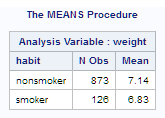
   
  
proc sgplot data=nc;  
 vbox weight/category=habit;  
run;



**Exercise 3:**

1. Check whether the conditions necessary for inference are satisfied.
   1. Yes, sample size is larger than 30 and sample population is approximately normally distributed

proc means data=nc mean maxdec=2;  
 class habit;  
 var weight;  
run;



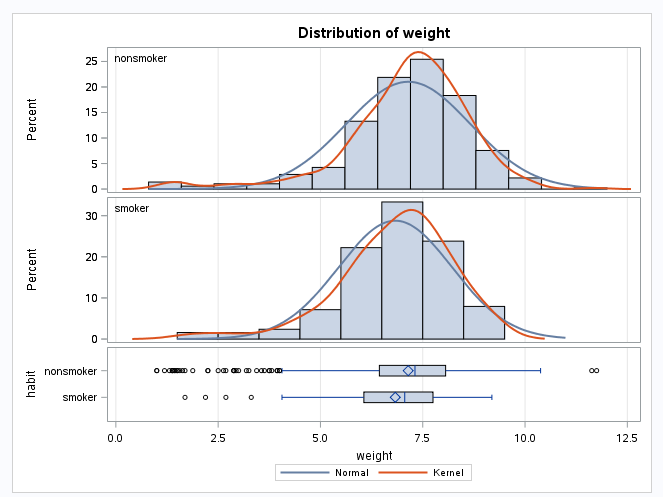
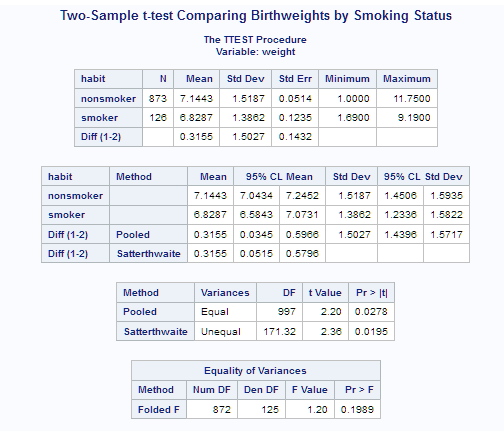
**Exercise 4:**

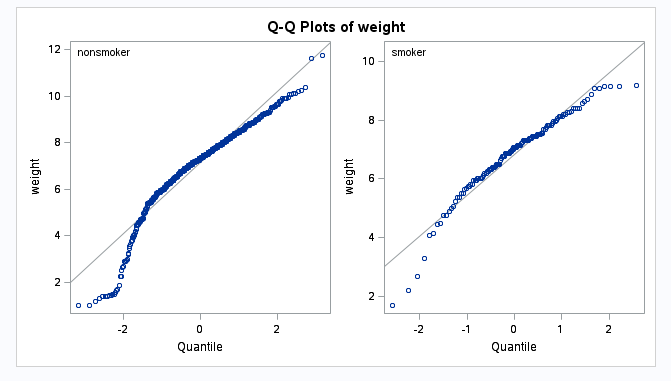
1. Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different. State the conclusion of the hypothesis test and report a 95% confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.
   1. Ho: Birth weight of Non-smoker = Birth Weight of Smoker

Ha: Birth weight of Non-smoker ≠ Birth Weight of Smoker

Are birth rates of new born in NC the same for mothers that did not smoke during pregnancy versus mothers that did smoke during pregnancy.

proc ttest data=nc sides=2 H0=0;  
 class habit;  
 var weight;  
 title "Two-Sample t-test Comparing Birthweights by Smoking Status";  
run;





**Part 6 - SAS Handout - Inference for categorical data**

**Exercise 1:**

1. In the first paragraph, several key findings are reported. Do these percentages appear to be sample statistics (derived from the data sample) or population parameters?
2. The data looks to come from sample data that represents the population, the data was gathered by Win-Gallup International.

**Exercise 2:**

1. The title of the report is “Global Index of Religiosity and Atheism.” To generalize the report’s findings to the global human population, what must we assume about the sampling method? Does that seem like a reasonable assumption?
2. The data looks to come from sample data that represents the population, the data was gathered by Win-Gallup International.
3. Random polling of the population to gather the required sample
4. Sample must be large enough to meaning represent the population true mean
5. Questions should be similar for each country represented in the study

**Exercise 3:**

1. What does each row of Table 6 correspond to?
   1. Country of Origin, sample size for country, response in the survey question: religious, not religious, atheist, no response
2. What does each row of atheism correspond to?
   1. Country, response, year

**Exercise** 4:

1. Using the DATA step below, create a new data set named us12 that contains only the rows in atheism that are associated with respondents to the 2012 survey from the United States. Next, calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?
   1. Yes, because we are sampling the same sample data so it should have the same porportion

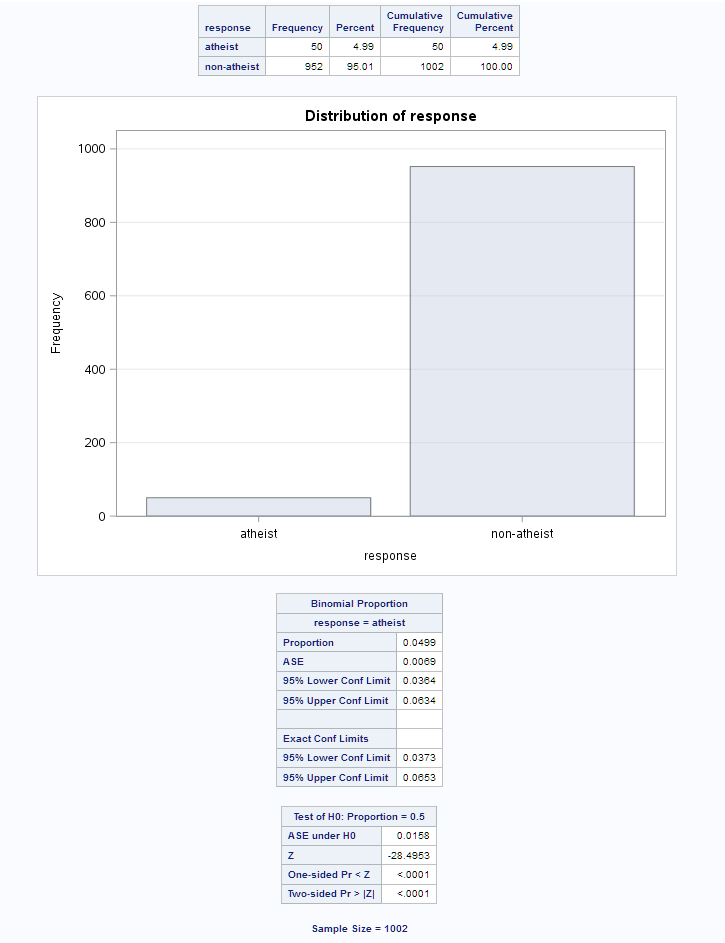
**Exercise 5:**

1. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?
   1. Sample must be greater than 30 with a normally distributed sample
   2. Sample is random and independence is assumed

**Exercise 6:**

1. Based on the SAS output, what is the margin of error for the estimate of the proportion of atheists in the US in 2012?
   1. 1.96\*.0069 = 0.013524

ods graphics;  
proc freq data=work.us12;  
 tables response / binomial(level="atheist") plots=freq;  
RUN;

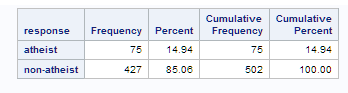


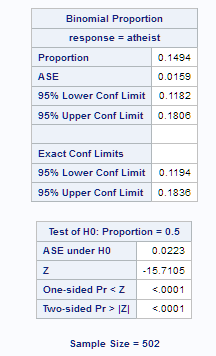
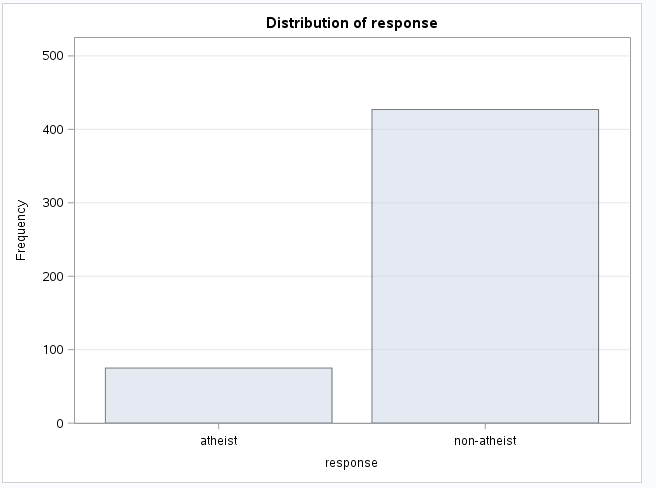
**Exercise 7:**

1. Using PROC FREQ, calculate confidence intervals for the proportion of atheists in 2012 in two other countries of your choice and report the associated margins of error. Be sure to note whether the conditions for inference are met. It might be helpful to create new data sets for each of the two countries first, and then use these data sets in PROC FREQ to construct the confidence intervals.
   1. Germany - Yes, sample sizes are larger than 30 and sample is random
   2. Japan - Yes, sample sizes are larger than 30 and sample is random

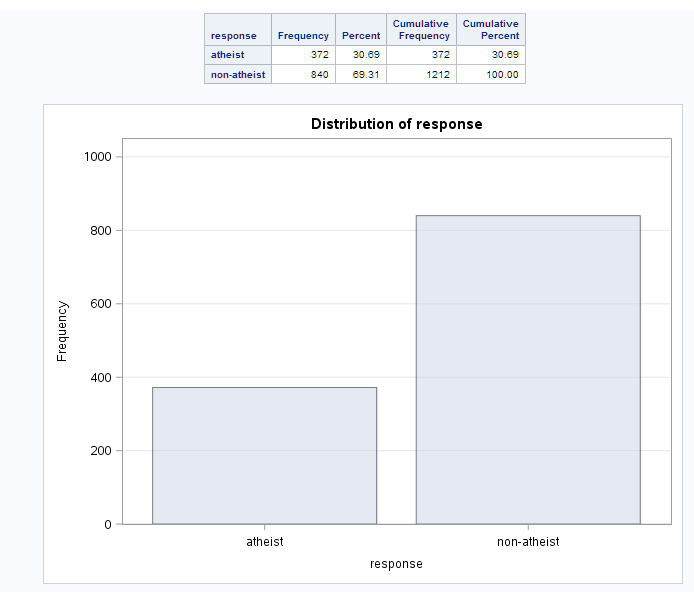
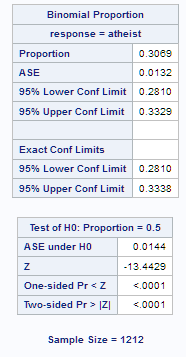
data work.germany12;  
 set work.atheism;  
 if nationality="Germany" and year=2012;  
run;

ods graphics;  
proc freq data=work.germany12;  
 tables response / binomial(level="atheist") plots=freq ;  
RUN;





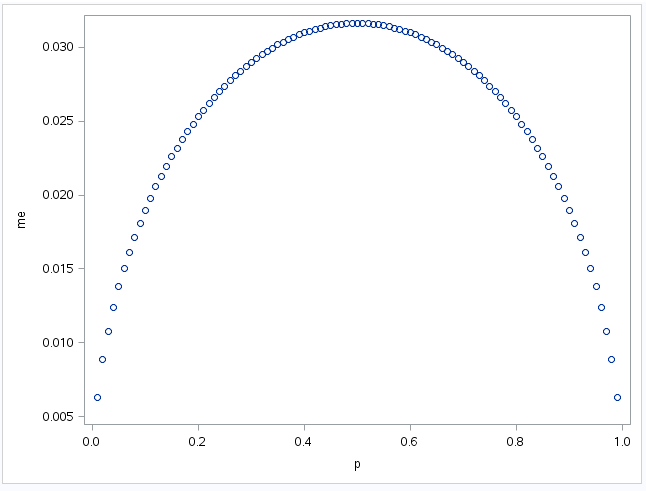
data work.japan12;  
 set work.atheism;  
 if nationality="Japan" and year=2012;  
run;  
  
ods graphics;  
proc freq data=work.japan12;  
 tables response / binomial(level="atheist") plots=freq ;  
RUN;

**Exercise 8:**

1. Describe the relationship between p and me.
   1. As proportion increases the margin of error increases likewise is you decrease the margin of error your proportions would need to decrease.

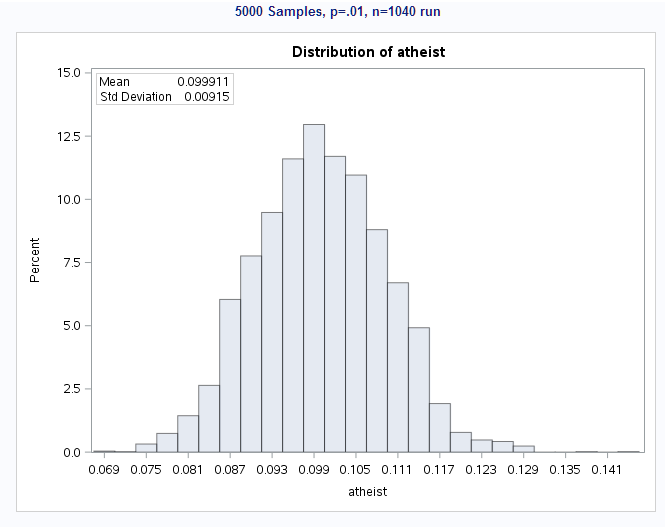
data work.proportions;  
 do p=0.01 to 0.99 by .01;  
 me = 2 \* sqrt(p\*(1-p)/1000);  
 output;  
 end;  
run;  
  
proc sgplot data=work.proportions;  
 scatter x=p y=me;  
run;



**Exercise 9:**

1. Describe the sampling distribution of sample proportions at n = 1040 and p = 0.1. Be sure to note the center, spread, and shape.
   1. Dist is normally distributed with center at .099911, the spread is bell shaped with evenly dispersed distribution ranging from .06 to .130

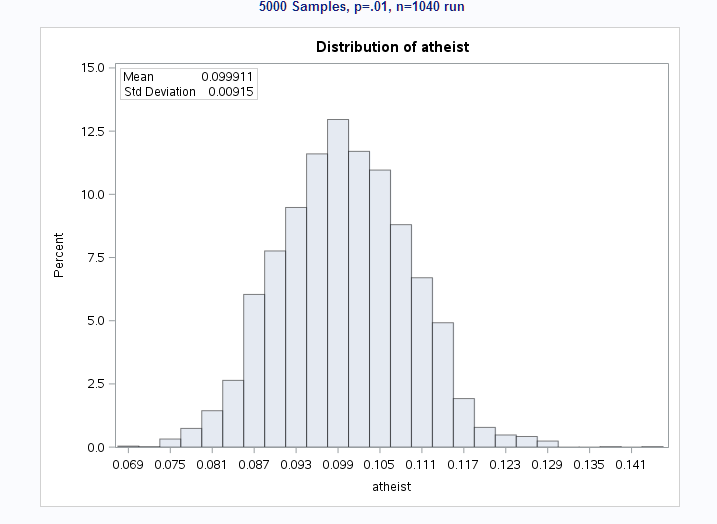
data work.multiple\_samples;  
 call streaminit (27513);  
 do sample = 1 to 5000;  
 do obs=1 to 1040;  
 if RAND("UNIFORM") < 0.1 then atheist=1;  
 else atheist=0;  
 output;  
 end;  
 end;  
run;  
  
proc means data=work.multiple\_samples noprint nway;  
 class sample;  
 var atheist;  
 output out=work.means mean=;  
run;  
  
ods select histogram;  
proc univariate data=work.means;  
 histogram atheist;  
 inset mean std;  
 title "5000 Samples, p=.01, n=1040"  
run;

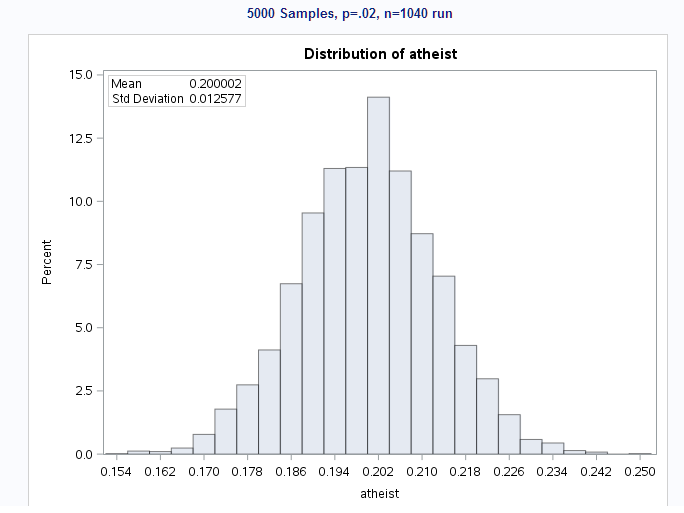


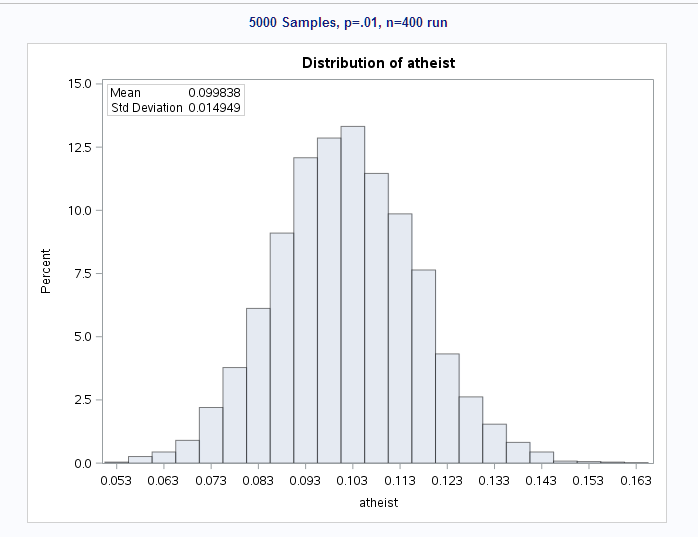
**Exercise 10:**

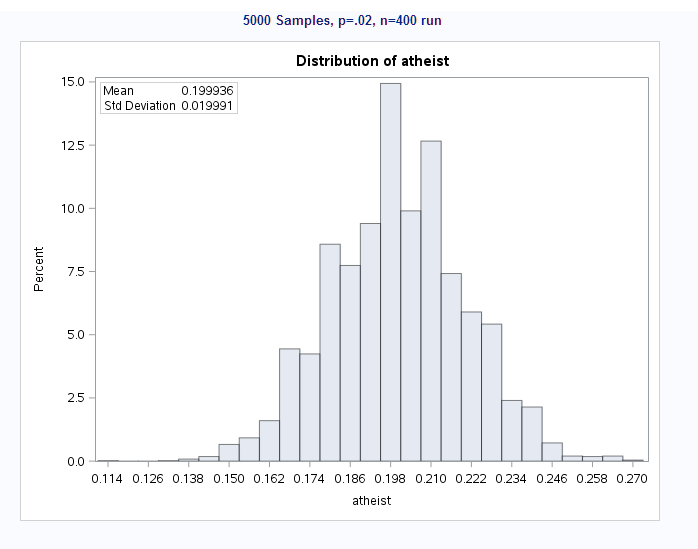
1. Replicate the above simulation three more times but with modified sample sizes and proportions: for n = 400 and p = 0.1, n = 1040 and p = 0.02, and n = 400 and p = 0.02. Plot all three histograms. Describe the three new sampling distributions. Based on these limited plots, how does n appear to affect the distribution of pˆ? How does p affect the sampling distribution?
   1. Increasing the sample size did not really change the distribution of the model much, both samples are still approximately normally distribution, with a bell shape curve, the spread is evenly dispersed from center.
   2. Changing the P^ makes a huge difference in the range from center, for both models. When changing the from .1 to .2 you see the range min and max both increase.

data work.multiple\_samples;  
 call streaminit (27513);  
 do sample = 1 to 5000;  
 do obs=1 to 1040;  
 if RAND("UNIFORM") < 0.1 then atheist=1;  
 else atheist=0;  
 output;  
 end;  
 end;  
run;  
  
proc means data=work.multiple\_samples noprint nway;  
 class sample;  
 var atheist;  
 output out=work.means mean=;  
run;  
  
data work.multiple\_samples;  
 call streaminit (27513);  
 do sample = 1 to 5000;  
 do obs=1 to 1040;  
 if RAND("UNIFORM") < 0.2 then atheist=1;  
 else atheist=0;  
 output;  
 end;  
 end;  
run;  
  
proc means data=work.multiple\_samples noprint nway;  
 class sample;  
 var atheist;  
 output out=work.means2 mean=;  
run;  
  
data work.multiple\_samples;  
 call streaminit (27513);  
 do sample = 1 to 5000;  
 do obs=1 to 400;  
 if RAND("UNIFORM") < 0.1 then atheist=1;  
 else atheist=0;  
 output;  
 end;  
 end;  
run;  
  
proc means data=work.multiple\_samples noprint nway;  
 class sample;  
 var atheist;  
 output out=work.means3 mean=;









**Exercise 11:**

1. If you refer to Table 6, you’ll find that Australia has a sample proportion of 0.1 on a sample size of 1040, and that Ecuador has a sample proportion of 0.02 on 400 subjects. Let’s suppose for this exercise that these point estimates are actually the truth. Then given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference and report margin of errors, as the reports does?
   1. No, because the sample ME will be way off based on the sample proportions