STAT 504 ANALYSIS OF DISCRETE DATA
Homework #1 WITH SOLUTIONS

1. (16 points) Which scale of measurement (nominal or ordinal) is the most appropriate for the following variables?

   (a) Treatment group (treatment or control) in a clinical trial?.

   (b) Age range in a clinical trial?

   (c) Geographic location (North America, Southeast Asia, Europe)?

   (d) Michelin stars?

   Solution:

   (a) Nominal

   (b) Interval or possibly Ordinal

   (c) Nominal

   (d) Ordinal

2. (16 points) Suppose in a certain area, the number of ships per 10 by 10 kilometer region of the ocean has a Poisson distribution with parameter 0.2.

   (a) What is the probability of observing 5 ships in one 10 by 10 kilometer region?

   (b) What is the probability of observing no more than 5 ships in a 100 by 100 kilometer region?

   (a) What is the probability of observing 5 ships in one 10 by 10 kilometer region?

   Solution Let X be the number of ships in a per 10 by 10 kilometer region. Then $X \sim \text{Poisson}(\lambda)$, where $\lambda = 0.2$. The probability of observing 5 ships in one 10 by 10 kilometer region is
   $P(X = 5) = \frac{e^{-0.2} \cdot 0.2^5}{5!} = 2.18 \times 10^{-6}$.
   In R

   ```
   dpois(5,0.2)
   [1] 2.183282e-06
   ```

   (b) What is the probability of observing no more than 5 ships in a 100 by 100 kilometer region?

   Solution Let X be the number of ships in a per 100 by 100 kilometer region. Then $X \sim \text{Poisson}(\lambda)$, where $\lambda = 0.2 \times 100 = 20$ (there are one hundred 10×10 squares in a 100 by 100 region). Then the probability of observing 5 ships or less in a 100 by 100 kilometer region is
   $P(X \leq 5) = \sum_{k=0}^{5} \frac{e^{-20} \cdot 20^k}{k!} = 7.19 \times 10^{-5}$.
   In R

   ```
   ppois(5,2)
   [1] 7.190884e-05
   ```

3. (18 points) When the 2000 General Social Survey asked subjects whether they would be willing to accept cuts in their standard of living to protect the environment, 344 of 1170 subjects said "Yes".

   (a) Estimate the population proportion who would say "Yes".

   Solution: The binomial maximum likelihood estimate of the proportion of who would be willing to accept a cut in their standard of living to prevent the environment is $\hat{p} = \frac{344}{1170} = 0.294$

(b) Conduct significance test to determine whether a majority of the population would say "Yes." Report and interpret the p-value.

Solution: The hypothesis to be tested, whether a majority would accept a cut, is $H_0 : \pi \leq 0.5$ v.s $H_1 = \pi > 0.5$. This is a one-sided test. The test statistic is:

$$Z = \frac{0.294 - 0.5}{\sqrt{0.5 \cdot 0.5/1170}} = -14.10$$

From the R/SAS output, we see that the p-value is 1. Thus, we cannot reject the null hypothesis. We conclude that a minority of the population would accept a cut.

R CODE

```
>  prop.test(344,1170,alternative='greater',conf.level=0.95,correct=FALSE)
1-sample proportions test without continuity correction

data:  344 out of 1170, null probability 0.5
X-squared = 198.5675, df = 1, p-value = 1
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.2726037 1.0000000
sample estimates:
        p
0.2940171
```

SAS CODE

```
data hw1_3b;
input response $3.
weight;
 datalines;
no 826
yes 344 ;

proc freq data=hw1_3b;
tables response / binomial(level="yes" p=.5);
weight weight;
exact binomial;
run;
```

The SAS System

The FREQ Procedure

response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	826	70.60	826	70.60
yes	344	29.40	1170	100.00

Binomial Proportion	
response = yes	
Proportion (P)	0.2940
ASE	0.0133
95% Lower Conf Limit	0.2679
95% Upper Conf Limit	0.3201
Exact Conf Limits	
95% Lower Conf Limit	0.2680
95% Upper Conf Limit	0.3210

Test of H0: Proportion = 0.5	
ASE under H0	0.0146
Z	-14.0914
One-sided Pr < Z	<.0001
Two-sided Pr > |Z|	<.0001
Exact Test	
One-sided Pr <= P	<.0001
Two-sided = 2 * One-sided	<.0001

Sample Size = 1170

(c) Construct and interpret 99% confidence interval for the population proportion who would say "Yes." This is the classical confidence interval, i.e. based on the Wald test.

Solution:

The 99% Wald confidence interval for the binomial parameter p is

$$\hat{p} \pm 2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

which is $0.2940 \pm 2.58\sqrt{\frac{0.2940(1-0.2940)}{1170}} = (0.2596, 0.3284)$.

The R output gives the 99% Wald confidence interval for the population proportion p which is (0.26,0.33) while the SAS code below gives the 99% Score (or Wilson) confidence interval for p which is (0.2609,0.3294). Notice that if we want

the Wald CI from SAS, which is the default, we can simply say "binomial(p=.5 level="yes")" or if we want both the Wald and Score we can use "binomial(p=.5 level="yes" wald wilson)". Since the sample size is large enough the Wald and Score confidence intervals are the same. We can conclude that the majority of people would not accept a cut in their standard of living. More specifically, we can be 99% confident that the population proportion (or the probability) of people who would accept a cut in their standard of living is between 0.26 and 0.33.

R CODE

```
>  prop.test(344,1170,,conf.level=0.99,correct=FALSE)
1-sample proportions test without continuity correction

data:  344 out of 1170, null probability 0.5
X-squared = 198.5675, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.2609468 0.3294104
sample estimates:
        p
0.2940171
```

SAS CODE

```
proc freq data=hw1_3b;
tables response / binomial(p=.5 level="yes" wald wilson) alpha=0.01;
weight weight;
exact binomial;
run;
```

The SAS System

The FREQ Procedure

response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	826	70.60	826	70.60
yes	344	29.40	1170	100.00

Binomial Proportion	
response = yes	
Proportion (P)	0.2940
ASE	0.0133

Confidence Limits for the Binomial Proportion		
Proportion = 0.2940		
Type	99% Confidence Limits	
Wald	0.2597	0.3283
Wilson	0.2609	0.3294

Test of H0: Proportion = 0.5	
ASE under H0	0.0146
Z	-14.0914
One-sided Pr < Z	<.0001
Two-sided Pr > |Z|	<.0001
Exact Test	
One-sided Pr <= P	<.0001
Two-sided = 2 * One-sided	<.0001

Sample Size = 1170

4. (20 points) Each of 100 multiple-choice questions on an exam has four possible answers, one of which is correct. For each question, a student guesses by selecting an answer uniformly at random.

   (a) Specify the distribution of the student's number of correct answers.

   Solution: Let X be the number of correct questions in out of 100 questions on an exam. Then $X \sim \text{Bin}(100, \frac{1}{4})$. Here, while the distribution of answers in terms of (a,b,c,d) is multinomial, we have grouped these into a "true" and "false" group, so we get a binomial distribution.

   (b) Find the mean and standard deviation of that distribution. Would it be surprising if the student made at least 50 correct responses? Why?

   Solution: The mean is $E(X_i) = np = 100(1/4) = 25$ and the standard deviation is sd=$\sqrt{np(1-p)} = 4.3$. Yes, it would be surprising if the student made at least 50 correct responses since using R the probability that the student made

at least 50 correct responses $P(X \geq 50) = 2.13 \times 10^{-8}$.

R CODE

```
1-pbinom(50,100,0.25)
[1] 2.131192e-08
```

(c) Specify the distribution of (n_1, n_2, n_3, n_4) where n_j is the number of times the student picked choice j.

Solution: The distribution of (n_1, n_2, n_3, n_4) is multinomial with parameters $n = 100$ and $p = (1/4, 1/4, 1/4, 1/4)$, $X \sim \text{Mult}(100, (1/4, 1/4, 1/4, 1/4))$.

(d) Find the mean of n_j, the variance of n_j, the covariance between n_j and n_k and the correlation between n_j and n_k.

Solution: $E(X_i) = np_i$, $i = 1, 2, 3, 4$. So $E(X_1) = E(X_2) = E(X_3) = E(X_4) = 100 \cdot 0.25 = 25$. The variance of n_j is $np_j(1 - p_j)$. The covariance between n_j and n_k is $\text{cov}(n_j, n_k) = -np_jp_k = -100(1/4)(1/4) = -6.25$ and the correlation between n_j and n_k is $\text{corr}(n_j, n_k) = \frac{-np_jp_k}{\sqrt{np_j(1-p_j)np_k(1-p_k)}} = \frac{-6.25}{18.75} = -0.333$.

5. (30 pts) Let X be the number of Chargers (San Diego's NFL team) fans observed in a random sample of $n = 30$ graduate students at Penn State. It's reasonable to assume that $X \sim Bin(30, p)$. The true proportion p is unknown, but it's likely to be small (especially in comparison to potential number of Pittsburgh Steelers fans).

(a) Assume for now that $p = 0.04$. Find the mean, the variance, and the standard deviation of X. Find the probabilities of the events $X = 0, X = 1, X = 2, X = 3$, and $X \geq 4$. You can do this by hand, or use either SAS or R.

Solution:

$$E(X) = np = 30 \times 0.04 = 1.2,$$
$$Var(X) = np(1-p) = 30 \times 0.04 \times 0.96 = 1.152$$
$$sd(X) = \sqrt{var(X)} = \sqrt{1.152} = 1.073$$
$$P(X = k) = \left(\begin{array}{c} 30 \\ k \end{array} \right) 0.04^k \times (1 - 0.04)^{30-k}, k = 0, \ldots, 30$$
$$P(X = 0) = 0.294, P(X = 1) = 0.367,$$
$$P(X = 2) = 0.222, P(X = 3) = 0.0086,$$
$$P(X \geq 4) = 1 - \sum_{k=0}^{3} P(X = k) = 0.031.$$

R CODE

```
>  dbinom(0,30, 0.04)
[1] 0.2938576
>  dbinom(1,30, 0.04)
[1] 0.3673221
>  dbinom(2,30, 0.04)
[1] 0.2219237
>   dbinom(3,30, 0.04)
[1] 0.08630368
>   1-pbinom(3,30, 0.04)
[1] 0.03059288
```

The SAS code to compute the binomial probabilities is:

SAS CODE

```
OPTIONS NODATE;
DATA BINOMIAL;
 LENGTH x $5.;
 i=0;
 x=put(i,1.);
 PROB=PROBBNML(.04,30,0);
 OUTPUT;
 DO i= 1 to 3;
 PROB = (PROBBNML(.04,30,i)-PROBBNML(.04,30,i-1));
 x=put(i,1.);
 OUTPUT;
 END;
 PROB=1-PROBBNML(.04,30,3);
 x=put(i,1.);
 x="GE 4";
 OUTPUT;
 drop i;

RUN;
```

(b) The classic approximate 95% confidence interval for p given in the most textbooks is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} = \frac{x}{n}$. When does this interval become degenerate (i.e., when the lower and the upper bounds are the same)? If the true p was actually 0.04, could this interval actually cover the true parameter 95% of the time? Why or why not? (Hint: based on part (a), how often would the interval become degenerate at zero?)

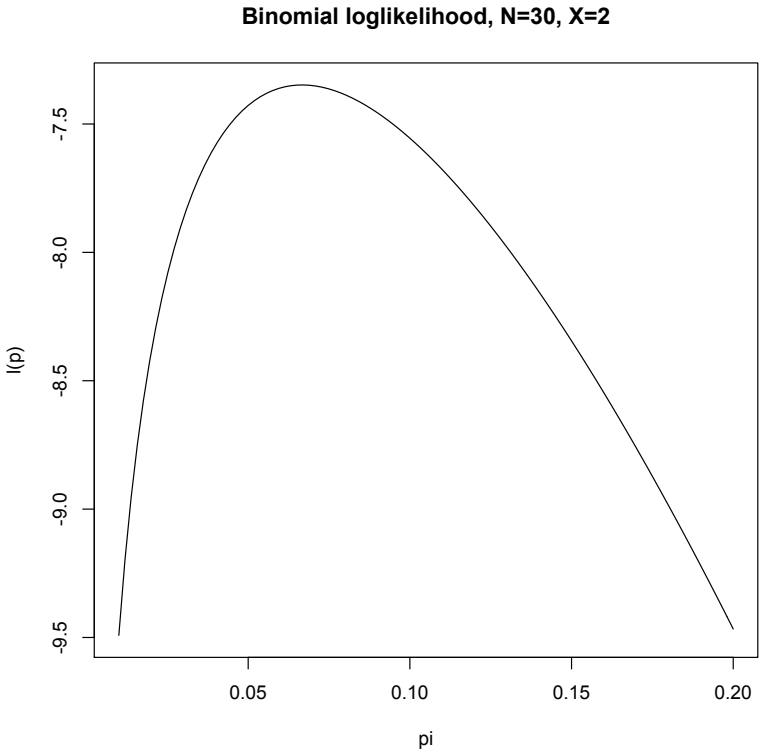
Solution: When $\hat{p} = 0$ or 1, i.e., when $x = 0$ or $x = n$, the upper bound and the lower bound are the same, $\hat{p}$.
From (a), we have $P(X = 0) = 0.294$ when $p = 0.04$. That is to say, the interval become degenerate at zero for about 29.4 percent of the time. Thus the degenerate possibility is $0.294 + P(X = 30) > 0.294$. This implies that the interval only covers the true parameter with probability smaller than 1-0.294=0.706.

(c) Suppose that we observe two Chargers fans in the sample. Plot the loglikelihood function for p over a range of values from $p = 0.01$ to $p = 0.20$. Find the ML estimate $\hat{p}$ and the value of the loglikelihood at $\hat{p}$. Calcluate the approximate 95% confidence interval for p based on the expected information. You may need to read the Supplemental Materials to do this part, and look at 1.6.1 and 1.6.3 in the online notes.

Solution:
R code

```
> loglik=function(p) 2*log(p)+28*log(1-p)
> plot(loglik,0.01,0.2,xlab="pi",ylab="l(p)",main="Binomial loglikelihood, N=30, X=2")
> p.mle<-2/30
> loglik(p.mle)
```

```
[1] -7.347901
```



SAS CODE

```
data for_plot;
do x=0.01 to 0.2 by 0.01;
y=2*log(x)+28*log(1-x);   *the log-likelihood function;
output;
end;
run;

proc gplot data=for_plot;
plot y*x / haxis=axis1 / hvref=-9.42;   ;
run;

quit;
```

From the plot we observe that the log-likelihood is not symmetric but it is skewed. The MLE is $\hat{p} = k/n = 2/30 \approx 0.0667$, and the value of the log likelihood at the MLE is -1.273. The approximated 95% confidence interval for p is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (-0.0226, 0.1559).$$

Note that the lower bound is less than 0, which has no practical meaning since it strays outside the parameter space . Thus the adjusted confidence interval is $(0, 0.1559)$.