hw1

Mohammad Mottahedi

January 13, 2016

```
set.seed(1234)
```

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

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

answer: Binary variable.

(b) Age range in a clinical trial?

answer: Continuous variables grouped into small number of categories.

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

answer: Nominal.

(d) Michelin stars?

answer: Ordinal.

Problem 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?

$p(y) = \frac{e^{-\mu}\mu^y}{y!} = \frac{e^{-0.2}0.2^5}{5!} = 2.183282 \times 10^{-6}, \mu = 0.2, y = 5$

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

There are 100 (independent) 10x10 km^2 region in a 100x100 km^2 region, number of ships in each of the 100 regions has possion distrubiton $\mu = 0.2$ (X_i). The sum $X_1 + ... + X_{100}$ has the poisson distribuiton with mean $\mu_{total} = \mu_1 + ... + \mu_{100} = 20$.

$p(y) = p(y = 0) + p(y = 2) + p(y = 3) + p(y = 4) + p(y = 5) = 7.1908841 \times 10^{-5}$

Problem 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. Note: Some helpful R and SAS code is provided on ANGEL. See hw1. R and hw1_ help. SAS .

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

estimated population proportion: $\hat{p} = \frac{344}{1170} = 0.2940171$

estimated standard error: $\hat{SE} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = 0.0133196$

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

$H_0 : p <= 0.5$

$H_A : p > 0.5$

$Z = \frac{\overline{p}-0.5}{\hat{SE}} = \frac{0.2940171-0.5}{0.0133196} = -15.4646461$

$p - value = 1 > 0.05$

it's very likely to see value as low as 0.29 so we can't reject the null hypothesis.

(c) Construct and interpret 99% confidence interval for the population proportion who would say Yes. Here you should construct a classical confi- dence interval, i.e. based on Wald test; later we will see some problems with this approach

$\hat{p} = \frac{344}{1170} = 0.2940171$

$\hat{SE} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = 0.0133196$

(0.2630311 , 0.3250031)

Problem 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 randomly.

(a) Specify the distribution of the students number of correct answers.

is binomial with probability mass fucntion $p(y) = \frac{100!}{[y!(100-y)!]}0.25^y(0.75)^{100-y}$

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

$E(Y) = n\pi = 100(\frac{1}{4}) = 25$, $\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{100\frac{1}{4}(1-\frac{1}{4})} = 4.330127$

Hypothesis testing:

$H_0 : \mu \geq 50$, $H_A : \mu < 50$

$Z = \frac{25-50}{4.330127/\sqrt{(100)}} = -57.7350272$

$p - value = 0 < 0.05$, yes it's suprising and it's unlikely to see 25 correct answers if the null hypothesis is correct, so we reject the null hypothesis.

(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.

Multinomial distrubiton

$p(n_1, n_2, n_3, n_4) = (\frac{n!}{n_1!n_2!n_3!n_4!})\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3}\pi_4^{n_4}$

(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 .

$E(n_j) = n\pi_j = 100(1/4) = 25$

$var(n_j) = n\pi_j(1 - \pi_j) = 18.75$

$$\mathbf{cov} = \begin{bmatrix} 18.75 & -6.25 & -6.25 & -6.25 \\ -6.25 & 18.75 & -6.25 & -6.25 \\ -6.25 & -6.25 & 18.75 & -6.25 \\ -6.25 & -6.25 & -6.25 & 18.75 \end{bmatrix}$$

Problem 5 (30 points). Let X be the number of Chargers (San Diegos NFL team) fans observed in a random sample of n = 30 graduate students at Penn State. It is reasonable to assume that X ~ Bin(30; p). The true proportion p is unknown, but it is likely to be small (especially in comparison to potential number of Pittsburgh Steelers fans). Note: Some helpful R and SAS code is provided on ANGEL; see hw1. R and hw1_ help. sas . For plotting likelihood there is only R code.

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

$E(X) = np = 1.2$, $var(X) = np(1 - p) = 1.152$

$p(X = 0) = 0.2938576$

$p(X = 1) = 0.3673221$

$p(X = 2) = 0.2219237$

$p(X = 3) = 0.0863037$

$p(X \geq 4) = 0.0305929$

(b) The classic (Wald) 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 $\hat{p} = X/n$. When does this interval become degenerate (i.e. when are 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?)

The interval is degenerate when $\hat{p}$ is close to 0 or 1. $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0 \rightarrow \hat{p}(1 - \hat{p}) = 0$.

with $p = 0.04$ the 95% confidence interval is $(0.0042229, 0.0042229)$ and does not cover the 0.04.

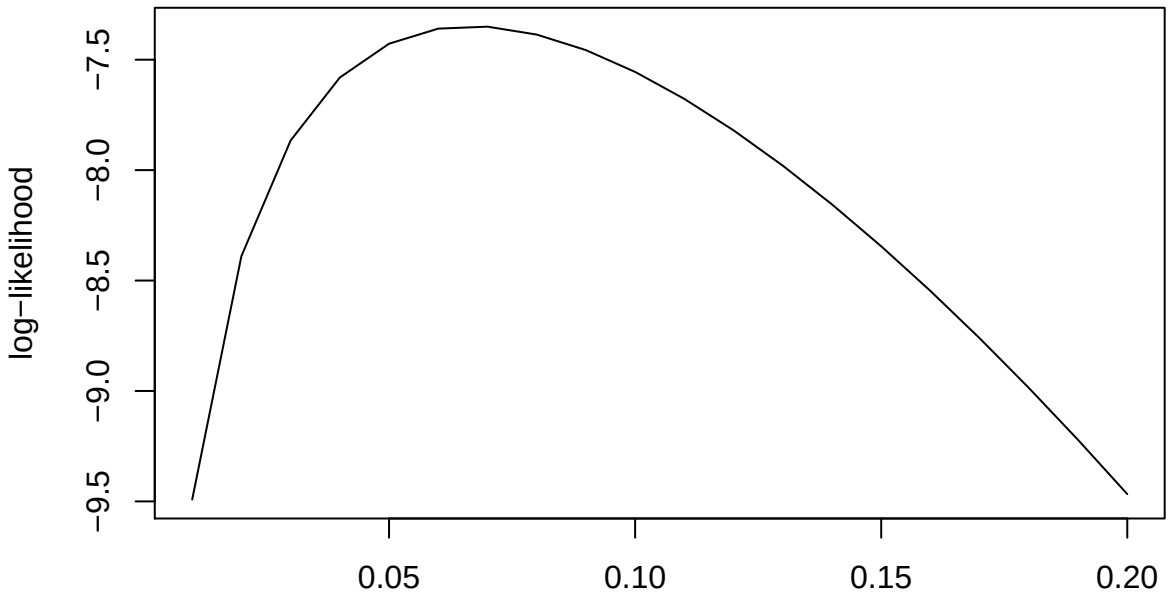
(c) Suppose that we observe two Chargers fans in the sample. Plot the log-likelihood function for p over a range of values from p = 0.01 to p = 0.20. Find the ML estimate and the value of the loglikelihood at $\hat{p}$. Calculate the approximate 95% confidence interval for p based on the expected information.

$l(p|x) = xlog(p) + (n - x)log(1 - p)$

$X = 2, n = 30$

$l(p|x) = 2log(p) + (28)log(1 - p)$

```
p = seq(from=0.01, to=0.2, by=0.01)
y = 2 * log(p) + 28 * log(1-p)
plot(p, y, xlab="", ylab="log-likelihood", type="l")
```



log-likelihood at $l(\hat{p} = 0.0666667) = -7.3479008$

$I(p) = \frac{n}{p(1-p)}$

95% confidence interval for p: $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (-0.0225957, 0.155929)$

(d) Plot the log-likelihood function with respect to the transformed parameter over a range of values from. Comment on this plot in comparison to the one in part (b).

(e) What is the 95% confidence interval for based on the expected infor- mation, and then transform the end points back to scale of p.

(f) Now find an approximate 95% confidence interval based on the LR method, and compare it to the two other intervals calculated above. Which one is the most appropriate and why?