CSE 544: Probability and Statistics for Data Science (Spring 2018)

Assignment 4 Solutions

Solution 1 **[5 points]**

The log-likelihood function is:

$$l(\theta) = \sum_{i=1}^{n} logf(X_i|\theta) = \sum_{i=1}^{n}(log\theta + \theta logx_0 - (\theta+1)logX_i)$$
$$= nlog\theta + n\theta logx_0 - (\theta+1)\sum_{i=1}^{n} logX_i$$

Set the derivative w.r.t θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{n}{\theta} + nlogx_0 - \sum_{i=1}^{n} logX_i = 0$$

Solving the equation gives the MLE for θ:

$$\hat{\theta}_{MLE} = \frac{1}{\overline{logX} - logx_0}$$

Solution 2 **[5 points]**

The MLE is $\hat{\theta} = Z_{max}$, the max data point. Not that $\hat{\theta} \leq \theta$. Thus,

$$\begin{aligned}
P(|\hat{\theta}-\theta| > \epsilon) &= P((\theta - Z_{max}) > \epsilon) && (1)\\
&= P(Z_{max} < (\theta-\epsilon)) && (2)\\
&= P(Z_1 < (\theta-\epsilon)\text{and}Z_2 < (\theta-\epsilon)\text{...till n}) && (3)\\
&= \left(\frac{\theta-\epsilon}{\theta}\right)^n && (4)\\
&= \left(1-\frac{\epsilon}{\theta}\right)^n && (5)
\end{aligned}$$

which $\rightarrow 0$ as n $\rightarrow \infty$.

Solution 3 **[13 points]**

(a) First we know that the probability for Binomial distribution is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

For $X_1, \cdots, X_n$ iid Poisson random variables will have a joint frequency function that is a product of the marginal frequency functions. Thus, the likelihood will be:

$$L(p) = \Pi_i \binom{n}{x_i} p^{x_i} (1-p)^{(n-x_i)} = \Pi_i \binom{n}{x_i} \cdot p^{\sum_i x_i} (1-p)^{(n^2 - \sum_i x_i)}$$

The log likelihood will then be

$$l(p) = \log p \sum_i x_i + (n^2 - \sum_i x_i) \log(1-p) + \sum_i log \binom{n}{x_i}$$

We need to find the maximum by finding the derivative:

$$l'(p) = \frac{1}{p} \sum_i x_i - \frac{1}{1-p} (n^2 - \sum_i x_i) = 0$$

which implies that the MLE should be

$$\hat{\lambda} = \frac{\sum_i x_i}{n^2}$$

(b) Let $X_1, \cdots, X_n \sim N(\theta, 1)$. The MLE for θ is $\hat{\theta} = \bar{X}$. Let $\delta = E[I_{X_1 > 0}]$. Thus,

$$\begin{aligned}
\delta &= E[I_{X_1>0}] \\
&= P(X_1 > 0) \\
&= 1 - P(X_1 \leq 0) \\
&= 1 - F_{X_1}(0) \\
&= 1 - \Phi(\frac{0-\mu}{\sigma}) \\
&= \Phi(\frac{\mu}{\sigma}) \\
&= \Phi(\frac{\theta}{1}) \\
&= \Phi(\theta)
\end{aligned}$$

So the MLE for δ is $\Phi(\bar{X}) = \Phi(\frac{\sum_i X_i}{n})$

(c) The PDF can be written as:

$$f(x|\theta) = \begin{cases} 1, & \text{for } \theta \geq x \leq \theta + 1 (i = 1, ..., n). \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The condition that $\theta \leq x_i$ for i = 1, ..., n is $\leq min(x_1, ..., x_n)$. Similarly, the condition $x_i \leq \theta + 1$ for i = 1, ..., n is equivalent to the condition that $\theta \geq max(x_1, ..., x_n) - 1$. Likelihood function can be written as:

$$L(\theta) = \begin{cases} 1, & \text{for } max(x_1, ..., x_n) - 1 \geq \theta \leq min(x_1, ..., x_n). \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Thus, we can select any value in the interval $[max(x_1, ..., x_n) - 1 \geq \theta \leq min(x_1, ..., x_n)]$ as the MLE for θ.

Solution 4 **[5 points]**

The sample proportion is:

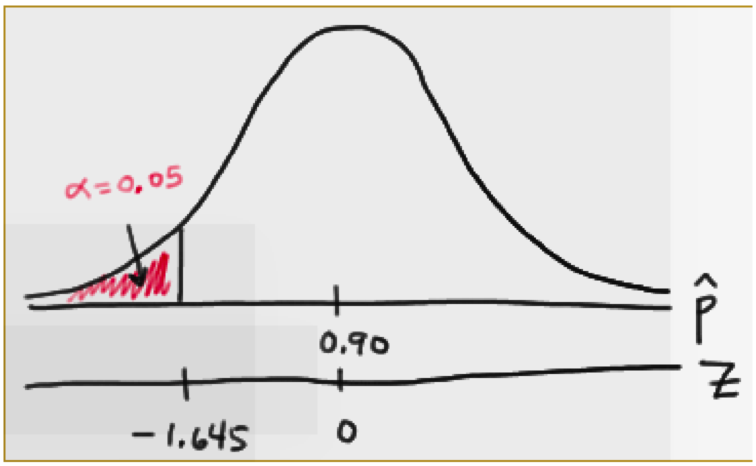
$$\hat{p} = \frac{128}{150}$$

Let $H_0 : p = 0.90$ and $H_1 : p < 0.90$

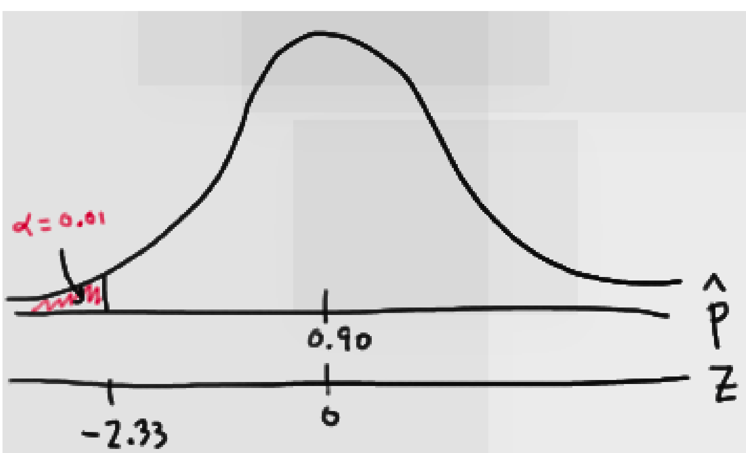
The test statistic is,

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = -1.92$$

(a) Since the test statistic is $Z = -1.92 < -1.645$, we reject the null hypothesis. There is sufficient evidence at the $= 0.05$ level to conclude that the rate has been reduced.



(b) Since our test statistic $Z = -1.92 > -2.33$, we do not reject the null hypothesis. There is insufficient evidence at the $= 0.01$ level to conclude that the rate has been reduced.



You can look up the Z_α values from below. Note that since it is one tailed you neeed z_α and not $z_{\frac{\alpha}{2}}$.

	Lower Tailed	Upper Tailed	Two Tailed
alpha = .10	z < -1.28	z > 1.28	z < -1.645 or z > 1.645
alpha = .05	z < -1.645	z > 1.645	z < -1.96 or z > 1.96
alpha = .01	z < -2.33	z > 2.33	z < -2.575 or z > 2.575

Solution 5 **[11 points]**

(a) The probability density function of a $Uniform(0, 3)$ random variable X. Thus,

$$f(x) = \frac{1}{3}$$

for $0 < x < 3$. Therefore, the $P[X \leq x] = \frac{1}{3}x$.

Now we could set up the hypothesis that,

- H_0: $F(x) = F_0(x)$
- H_1: $F(x) \neq F_0(x)$

where $F(x)$ is the (unknown) CDF from which our data were sampled and $F_0(x)$ is the CDF of $Uniform(0, 3)$.

Based on the function of probability, we could get the table below, which provides all the values for the KS test.

The largest of the values in last 2 columns is 0.137. For $\alpha = 0.05$, the critical value is 0.41. So, we can not reject the claim that the data were sampled from $Uniform(0, 3)$.

(b) This time, we could set up the hypothesis that,

- H_0: $F(x) = F_0(x)$
- H_1: $F(x) \neq F_0(x)$

	A	B	C	D	E	F
1	0.02	0	0.1	0.00667	0.00667	0.09333
2	0.65	0.1	0.2	0.21667	0.11667	0.01667
3	0.93	0.2	0.3	0.31	0.11	0.01
4	0.99	0.3	0.4	0.33	0.03	0.07
5	1.09	0.4	0.5	0.36333	0.03667	0.13667
6	1.5	0.5	0.6	0.5	0	0.1
7	1.78	0.6	0.7	0.59333	0.00667	0.10667
8	2.01	0.7	0.8	0.67	0.03	0.13
9	2.33	0.8	0.9	0.77667	0.02333	0.12333
10	2.87	0.9	1	0.95667	0.05667	0.04333

where $F(x)$ is the (unknown) CDF from which our data were sampled and $F_0(x)$ is the CDF of Normal distribution with mean 1.5.

Since the sample mean $\bar{X} = \sum_i x_i/n = 14.17/10 = 1.417$ and the sample variance $S_n^2 = \sum_i (x_i - \bar{X})^2/n = 0.650141$.

The t-test is

$$T = \frac{(\bar{X} - \mu)\sqrt{n}}{S_n} = \frac{(1.417 - 1.5)\sqrt{10}}{\sqrt{0.650141}} = -0.32551748$$

For $\alpha = 0.05$, the critical value is 2.228 and $|T| = 0.326 < t_{n-1,\alpha}$. Thus, we can not reject the claim that the data were sampled from Normal distribution.

(c) Consider testing

$$H_0 : p = \frac{1}{2} \text{ vs } H_1 : p \neq \frac{1}{2}$$

The size α Wald test is: reject H_0 when $|W| > z_{\alpha/2}$ where $W = \frac{\hat{\theta} - \theta_0}{\hat{se}}$.

Now we have $X_1, \cdots, X_{100} \sim Bernoulli(p)$. Also, we could know that for both MLE and MME, the estimator of the probability is

$$\hat{p} = \frac{\sum_i X_i}{n} = 46/100$$

Thus,

$$W = \frac{\hat{\theta} - \theta_0}{\hat{se}} = \frac{0.46 - 0.5}{\sqrt{0.46(1 - 0.46)/100}} = -0.8026$$

and

$$p - value = 0.4238.$$

$z_{\frac{\alpha}{2}}$ is 1.97 (using z score table). Since $|W|$ is not greater than 1.97 we accept H_0.

Now, consider testing

$$H_0 : p = 0.7 \text{ vs } H_1 : p \neq 0.7$$

Thus,

$$W = \frac{\hat{\theta} - \theta_0}{\hat{se}} = \frac{0.46 - 0.7}{\sqrt{0.46(1 - 0.46)/100}} = -4.816$$

and

$$p - value = \approx 0$$

$z_{\frac{\alpha}{2}}$ is 1.97 (using z score table). Since $|W|$ is greater than 1.97 we accept H_1.

Solution 6 **[9 points]**

(a)

$$W = \frac{\hat{\theta} - 0}{\hat{se}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{99.67 - 105.77}{\sqrt{\frac{9.99^2}{1000} + \frac{20.11^2}{1000}}} = -8.59$$

$$\because |W| > Z_{\frac{\alpha}{2}} \quad \therefore Reject\ H_0$$

(b) Discarded.

(c)

$$T = \frac{\hat{d} - 0}{\frac{\hat{\sigma_d}}{\sqrt{n}}} = \frac{-6.09}{\frac{22.20}{\sqrt{1000}}} = -8.675$$

$$\because\ |T| > t_{n-1, \frac{\alpha}{2}} \quad \therefore Reject\ H_0$$

Both Wald test and t-test are applicable.

Solution 7 **[6 points]**

H_0 :not different and H_1 :different

(a) The Wald statistic is:
$Z = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{s_1^2}{10}+\frac{s_2^2}{10}}} = 4.6$

The p-value is 0.0001 and C.I. is $\bar{X} - \bar{Y} \pm 2\sqrt{\frac{s_1^2}{10} + \frac{s_2^2}{10}} = (0.01, 0.03)$.
Since $|W|$ is greater than 1.97 we reject H_0.

(b) Permutation test on the absolute difference of means gives a p-value of ≈ 0. The provides evidence against the H_0.

Solution 8 **[5 points]**

We know that

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{8}$$

and $X_1, X_2, ..., X_n$ are all IID samples. Based on this, we know that (ignoring constants)

$$f(x^n|\theta) \sim \exp\{-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\}$$

$$g(\theta) \sim \exp\{-\frac{(\theta-a)^2}{2b^2}\}$$

Thus, the posterior of θ is

$$
\begin{aligned}
h(\theta|x^n) &= f(x^n|\theta)g(\theta) \\
&\sim \exp\{-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\}\exp\{-\frac{(\theta-a)^2}{2b^2}\} \\
&\sim \exp\{-\frac{n(\bar{x}-\theta)^2}{2\sigma^2} - \frac{(\theta-a)^2}{2b^2}\} \\
&\sim \exp\{-\frac{nb^2(\bar{x}-\theta)^2 + \sigma^2(\theta-a)^2}{2\sigma^2 b^2}\} \\
&\sim \exp\{-\frac{(nb^2+\sigma^2)\theta^2 - 2(a\sigma^2 + n\bar{x}b^2)\theta + nb^2\bar{x}^2 + a^2\sigma^2}{2\sigma^2 b^2}\} \\
&\sim \exp\{-\frac{(b^2+se^2)\theta^2 - 2(ase^2 + \bar{x}b^2)\theta + b^2\bar{x}^2 + a^2 se^2}{2se^2 b^2}\} \\
&\sim \exp\{-\frac{\theta^2 - 2\frac{(ase^2+\bar{x}b^2)}{(b^2+se^2)}\theta + \frac{b^2\bar{x}^2+a^2se^2}{(b^2+se^2)}}{2\frac{se^2b^2}{(b^2+se^2)}}\} \\
&\sim \exp\{-\frac{(\theta - \frac{ase^2+\bar{x}b^2}{b^2+se^2})^2}{2\frac{se^2b^2}{b^2+se^2}}\}\cdot \text{constant} \\
&\sim Normal(\frac{ase^2+\bar{x}b^2}{b^2+se^2}, \frac{se^2b^2}{b^2+se^2}) \\
&\sim Normal(x, y^2)
\end{aligned}
$$

where $x = \frac{ase^2+\bar{x}b^2}{b^2+se^2}$, $y = \frac{se^2b^2}{b^2+se^2}$, $\bar{x} = \frac{1}{n}\sum_i^n X_i$ and $se^2 = \frac{\sigma^2}{n}$

Solution 9 **[11 points]**

(a) and (b) Use the formulas given in Q8 to calculate these values.

(c) Data with low variance is more useful for convergence (move away from prior). With high variance posterior remains close to prior.