CSC591: Foundations of Data Science
HW2: Probability distributions, Expectation, Maximum Likelihood Estimation, Sampling Distribution, Central Limit Theorem, Confidence Intervals, Hypothesis Testing.

Released: 9/23/15
Due: **10/4/15 (23:55pm);** (One day late: -25%; -100% after that).
Note: (R-project (code) can be submitted after mid-term, but by 10/11/15: 23.55pm)

Solution/Hints: Prepared By the Instructor and TA.

Notes

- Filename: Lastname_StudentID.pdf (only pdf).
- You can also submit scanned hand written solution (should be legible, TA's interpretation is final).
- This h/w is worth 5% of total grade.
- You can discuss with your friends, but solution should be yours.
- Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.
- All submission must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
- No makeups or bonus; for regarding policies, refer to syllabus and 1st day lecture slides.

Q#	Max Points	Your Score
1	5	
2	20	
3	10	
4	15	
5	20	
6	10	
7	20	

Q1. Simple statistics (5 points)

(a) List formulae for sample mean, mode, variance, standard deviation. (4 points).

(These are standard formulae that you can find form any standard Statistics books)

(b) Is the sample variance is an unbiased estimator of population variance? (1 point).

No, it underestimates the population variance by a factor of (N – 1)/N.

Q2. (Expected Values) (20 points)

Remember the following:

(i). If X and Y are two random variables with finite expected values,
then E(X+Y) = E(X) + E(Y).

(ii) If X and Y are independent, the E(XY) = E(X)E(Y).

(a) Define Expected Value of discrete (numerical) random variable. **(1 point)**

$$\mathrm{E}[X] = \sum_x xP(x)$$

(b) Suppose in an experiment a fair coin is tossed 3 times. Let X denotes the number of heads that appeared in the experiment. Then what is E(X). **(2 points)**

Outcome	Probability	Number of heads
HHH	1/8	3
HHT	1/8	2
HTH	1/8	2
HTT	1/8	1
THH	1/8	2
THT	1/8	1
TTH	1/8	1
TTT	1/8	0

E[X] = 1/8 * (3 + 3 * 2 + 3 * 1) = 1/8 * 12 = 3/2

(c) Recall the discussion on Bernoulli distribution (W2-C2 lecture). Let S_n be the number of success in n Bernoulli trails with probability p for success on each trail. Then what is $E(S_n)$. **(3 points)**

Let us denote X_i be a r.v. which has a value 1 if i^{th} outcome is a success

and 0 otherwise. Then for each X_i we have $E(X_i) = 0(1-p) + 1(p) = p$.

And, $S_n = X_1 + X_2 + \ldots X_n$

$E(S_n) = E(X_1) + E(X_2) + \ldots E(X_n) = np$.

(d) A coin is tossed twice. Let $X_i = 1$ if the i^{th} toss is heads and 0 otherwise. Then what is $E(X_1X_2)$? **(2 points)**

We know that X_1 and X_2 are independent. Therefore $E(X_1X_2) = E(X_1)E(X_2) = ½ \times ½ = ¼$

(e) Let X be a random variable with expected value $\mu = E(X)$. Then what is the Variance of X in terms of E(). **(1 point)**

$V(X) = E((X- \mu)^2)$

(f) Let X be a random variable with expected value $\mu = E(X)$, then show that the Variance, $V(X) = E(X^2) - \mu^2$. **(2 points)**

$V(X) = E((X- \mu)^2) = E(X^2 – 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2$

(g) Using the formulae in (f), compute the variance of the outcome of a roll of a die. **(3 points)**

$E(X^2) = 1(1/6) + 4(1/6) + 9(1/6) + 16(1/6) + 25(1/6) + 36(1/6) = 91/6$.
$V(X) = 91/6 – (7/2)^2 = 35/12$.

(h) Let X be an exponentially distributed r.v. with parameter λ. Then the density function of X is given by: $f_X(x) = \lambda e^{-\lambda x}$. Compute E(X) and V(X), where V stands for variance. **(3 + 3 = 6 points)**

$$
\begin{aligned}
E(X) &= \int_0^\infty x f_X(x)\,dx \\
&= \lambda \int_0^\infty x e^{-\lambda x}\,dx \\
&= -x e^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \\
&= 0 + \frac{e^{-\lambda x}}{-\lambda}\Big|_0^\infty = \frac{1}{\lambda}\,.
\end{aligned}
$$

$$
\begin{aligned}
V(X) &= \int_0^\infty x^2 f_X(x)\,dx - \frac{1}{\lambda^2} \\
&= \lambda \int_0^\infty x^2 e^{-\lambda x}\,dx - \frac{1}{\lambda^2} \\
&= -x^2 e^{-\lambda x}\Big|_0^\infty + 2\int_0^\infty x e^{-\lambda x}\,dx - \frac{1}{\lambda^2} \\
&= -x^2 e^{-\lambda x}\Big|_0^\infty - \frac{2x e^{-\lambda x}}{\lambda}\Big|_0^\infty - \frac{2}{\lambda^2} e^{-\lambda x}\Big|_0^\infty - \frac{1}{\lambda^2} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}\,.
\end{aligned}
$$

(3) Continuous Distributions (10 points)

(a) Let us assume that the life of pen drives before failure is normally distributed with mean = 10 years and a standard deviation of 2 years. Find the probability that the pen drive fails between 9 years and 11 years. (**3 points**)

Since data is normally distributed, we can compute Z-score = $(X - \mu)/\ \sigma$

Z1 = (9-10)/2 = -0.5
Z2 = (11-10)/2 = 0.5

P(X ≥ 9) = -0.3085
P(X ≤ 11) = 0.6915

P(9 ≤ X ≤ 11) = 0.6915 – 0.3085 = 0.383

(b). Let us assume that your instructor assigns a letter grade of "Pass = C" to final score of 75 ± d. It is known that students' scores are normally distributed with mean of 75 and a standard deviation of 5. Find the value of d such that C's covers 95% of scores. (**3 points**)

We have N(75, 5), and α = 0.05, and z-score = 1.96; Therefore, d = 1.96 x 5 = 9.8

(4) Maximum Likelihood Estimation (MLE) (10 points)

(a) Concisely describe MLE procedure for single parameter **(2 points)**

MLE involves writing out the likelihood function L(Θ) = P(X; Θ) (or usually its log) and maximizing with respect to the parameters Θ by differentiating and setting to zero. The solution gives the parameters for which the probability of the data is the highest.

(b) Let X be a continuous random variable with p.d.f. for $\lambda > 0$, is defined as

$$f(x;\lambda) = \begin{cases} \lambda x^{\lambda-1} & if\ 0<x<1 \\ 0 & otherwise \end{cases}$$

Then find MLE for the parameter λ. **(8 points)**

$$L(\theta) = \log f(X;\ \lambda) = \log\lambda + (\lambda - 1)\sum_i \log x_i$$

$$\frac{\delta}{\delta\lambda}L(\theta) = \frac{1}{\lambda} + \sum_i \log x_i = 0$$

$$\lambda = -\frac{N}{\sum_{i=1}^{N} \log x_i}$$

(5) CLT, CI (15 points)

(a) Define Central Limit Theorem and state assumptions (5 points)

The CLT says that for a sample from an arbitrary distribution (not just normal) with a finite mean μ and variance σ^2, as the sample size increases the sampling distribution of the mean approaches a normal distribution with mean μ and variance σ^2/N.

It assumes that the observations are independent, and for skewed distributions N must be large (approximately > 30) for the sampling distribution to be normal.

(b) Define Confidence Interval for population mean (2 points)

The confidence interval is the range of values in which the true population mean lies with a certain probability (that is specified when giving the C.I.).

(c) Find the critical value for 95% C.I. (1 point)

The values outside of a 95% C.I. are the top and bottom 2.5%, which gives critical values of ±1.96.

(d) Outline the procedure for finding C.I. (2 points)

See slides 22-24 of W4C2 lecture.

(e) A sample of 30 students is drawn from CSC-522 population of 100. Average weight of sample is 150 pounds and standard deviation is 20 pounds. Compute the 95% CI. (5 points)

Margin of Error: $\pm 1.96*20/\sqrt{30} = \pm 7.1569$
C.I = 150 ± 7.1569

(6) Hypothesis testing fundamentals (20 points)

(a) Define null and alternate hypothesis (**2 points**)

Null hypothesis: what is assumed to be true unless evidence is seen to the contrary (a drug doesn't work, a discrepancy between expected and observed values isn't significant, etc.)

Alternate hypothesis: an alternative explanation for an observed phenomenon if the null hypothesis is proven to be false – generally something that would be somewhat surprising

(b) State null and alternate hypothesis for left--, right, and two-tailed tests (**3 points**)

Let x* be the true (unobserved) value.

Left-tailed:
Null: $x^* = x_1$
Alternate: $x^* < C$
Right-tailed:
Null: $x^* = C$
Alternate: $x^* > C$
Two-tailed:
Null: $x^* = C$
Alternate: $x^* \neq C$

(c) State type-1 and type-2 errors (**2 points**)

Type-1: incorrectly rejecting the null hypothesis

Type-2: incorrectly not rejecting the null hypothesis

(d) Define level of significance and list 3 widely accepted significance levels. (**2 points**)

The level of significance is the probability of a type 1 error. Commonly used levels are 0.1, 0.05, 0.01.

(d) How do you reduce type-1 errors (**3 points**)

By reducing the significance level. However, please note that once a significance level is chosen, one should not change it as an after the fact.

(e) Define critical value (**1 point**)

A critical value is the value of the test statistic at the threshold between accepting and rejecting the null hypothesis.

(f) Write down the general hypothesis testing procedure (**4 points**)

State the hypothesis and choose a significance level.

Calculate the variance of the test statistic and use that to determine the critical values.

Compute the test statistic and decide whether to accept or reject the null hypothesis based on whether it falls in the critical region or not.

(g) Write down z-test (**2 points**)

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

(h) Which test do you use when population standard deviation is not known (**1 point**)

t-test

(7) Hypothesis testing (10 points)

Average weight last year CS graduate students is 162.5lb, with a standard deviation of 6.9 lb. A sample of 50 students of this year is 165.2lb with same standard deviation. Answer the following: (1) Is there a reason to believe that there is a change in average weight of current batch of students? State your conclusion using traditional hypothesis testing using significance level of 0.05; (2) Compute P-value and state your conclusion.

(i) State Hypothesis:

$H_0 : \mu = 162.5$

$H_0 : \mu \neq 162.5$

(ii) For $\alpha = 0.05$, from z-table, we get critical value of 1.96.

(iii) Compute Z.

Z = (165.2 – 162.5)/(6.9/sqrt(50)) = 2.77

(iv) We see that Z-score > critical value, therefore reject null hypothesis.

(v) Write summary statement: (use the flow chart given slides).

(2)Using P-Value Method:

Using z-table, find area for z-value of 2.77, which is 0.9972. Therefore, P-value = 2(1-0.9972) = 0.0056.

As P-Value < 0.05, we reject null hypothesis

Write summary statement (use the flow chart given slides).