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

Released: 9/20/18

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 **4**% of total grade.
- **Separate R mini project will be released after 1st midterm (to facilitate more time for exam preparation)**.
- 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.
- You are encouraged to do research, study online materials; discuss with fellow students; BUT ANSWERS SHOULD BE YOUR OWN. Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.

Q1. Simple statistics (**20 points**)

(a)

List formulae for sample mean, mode, variance, standard deviation.

Solution:

Sample mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Sample mode:
The mode is the most frequently occurring score or value.

$$\arg\max_{x_i} f(x_i) \qquad x_i \in X$$

Where $f(x_i)$ is the frequency of x_i

Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Is the sample variance is an unbiased estimator of population variance? Why or why not?

Solution:

Yes. That is the reason why the sample variance use n-1 rather than n, so that the sample variance is an unbiased estimator of population variance. Many proofs can be found online.

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

Solution:

Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2 / n as n, the sample size, increases.

Assumptions:

1. Independence: It is assumed that the samples are independent of each other
2. Sample size should not be more than 10% if sampling without replacement
3. For non-normal distribution, the sample size must be at least 30.

Q2 (Expected Values) **(10 points)**

Remember the following (i) and (ii) as you may need them for answering some of the questions:

(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).

Answer (a) – (h).

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

Solution:

Definition: The expectation of a discrete random variable X taking the values $a_1, a_2, \ldots$ and with $P(X=a_i)$

$$E[X] = \sum_i a_i P(X = a_i)$$

(b) Suppose in an experiment a fair coin is tossed 4 times. Let X denotes the number of tails that appeared in the experiment. Then what is E(X). **(2 points)**

Solution:

Since the coin is fair, the probability that head appears is P(head)=0.5.

$$E[X] = 4 * (1 * 0.5 + 0 * 0.5) = 2$$

(c) Recall the discussion on Bernoulli distribution. 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)**

Solution:

From Bernoulli distribution function:

$$f(k,p) = p \quad if\ k = 1\ ;\ 1-p \quad if\ k = 0$$
$$E[S_n] = nE[S_1] = n(1 * p + 0 * (1-p)) = np$$

(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)**

Solution:

Since X_1 and X_2 are independent,

$$E\left[X_1X_2\right] = E\left[X_1\right]E\left[X_2\right] = (1 * 0.5 + 0 * 0.5)^2 = 0.25$$

If you did not assert that P(head)=0.5, then your answer should be $P(head)^2$.

(e) 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)**

Solution:

$$\begin{aligned} V(X) &= E[(X-\mu)^2] \\ &= E[X^2 - 2X\mu + \mu^2] \\ &= E[X^2] - 2E[X]\mu + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned}$$

Q3 (Expected Values) **(10 points)**

(a) 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)**

Solution

From $f(x) = \lambda e^{-\lambda x}$ $\quad x \in [0, +\infty]$

$$E[X] = \int_{-\infty}^{+\infty} x f(x)\, dx$$

$$= \int_0^{+\infty} \lambda x e^{-\lambda x} dx$$

$$= -\int_0^{+\infty} x\, de^{-\lambda x}$$

$$= -x e^{-\lambda x}\Big|_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x} dx$$

$$= -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^{+\infty}$$

$$= \frac{1}{\lambda}$$

$$E[X^2] = \int_{-\infty}^{+\infty} x^2 f(x)\, dx$$

$$= \int_0^{+\infty} \lambda x^2 e^{-\lambda x} dx$$

$$= -x^2 e^{-\lambda x}\Big|_0^{+\infty} + \int_0^{+\infty} 2x e^{-\lambda x} dx$$

$$= 2\left(-\frac{1}{\lambda} x \cdot e^{-\lambda x}\Big|_0^{+\infty} + \frac{1}{\lambda}\int_0^{+\infty} e^{-\lambda x} dx\right)$$

$$= 2 \cdot \frac{1}{\lambda} \cdot \left(-\frac{1}{\lambda}\right) \cdot e^{-\lambda x}\Big|_0^{+\infty}$$

$$= \frac{2}{\lambda^2}$$

$$\therefore V(X) = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

(b) Let us say an insurance company pays $500 for lost luggage or a cancelled flight. Historical data shows that the company ends up paying 1 out of 100 policies it sells. What premium should the company charge in order to make profit? (**4 points**)

Solution:

Let X be a random variable representing the insurance company compensation

E[X] = 500 * 0.01 + 0 * 0.99 = 5

So when the company set premium > $5, they will make profit.

(4) 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. (**4 points**)

Solution:

Let x presents the life of the pen drive

$P(9 <= x <= 11) = P(x <= 11) - P(x <9)=P(z\leq \frac{11-10}{2})-P(z< \frac{9-10}{2}) = 0.6915-0.3085=0.383$

(b). (**6 points**) Let us assume that CSC-591 FDS class final numerical grades (maximum 100) are values of a continuous r.v. X that follows a normal distribution with mean 75 and s.d. 15. Students are assigned letter grades as following: A (X $\geq$90); $B(80 \leq X < 90); C(70 \leq X < 80); D(60 \leq X < 70),$ *and* $F(X < 60)$. Answer following:

(i) If a student is chosen at random then compute the probability that the student earns a given letter grade

Solution:

Let Y presents the grade of the student

$P(Y=A)=P(X>=90) = 1 – P(X<90) = 1 – P(z< \frac{90-75}{15}) = 1 – 0.8413 = 0.1587$

$P(Y=B)=P(80 \leq X < 90) = P(X<90) – P(X<80)= P(z< \frac{90-75}{15})-P(z< \frac{80-75}{15})=0.2108$

$P(Y=C)=P(70 \leq X < 80) = P(X<80) – P(X<70)= P(z< \frac{80-75}{15})-P(z< \frac{70-75}{15})=0.2611$

$P(Y=D)=P(60 \leq X < 70) = P(X<70) – P(X<60)= P(z< \frac{70-75}{15})-P(z< \frac{60-75}{15})=0.2108$

$P(Y=F)=P(X < 60)= 0.1587$

(ii) Compute the expected proportion of students in each letter grade

Solution:

From the result we get from (i), we can conclude that

E(proportion of X>=90) = 15.87%;

E(*proportion of* $80 \leq X < 90$)=21.08%

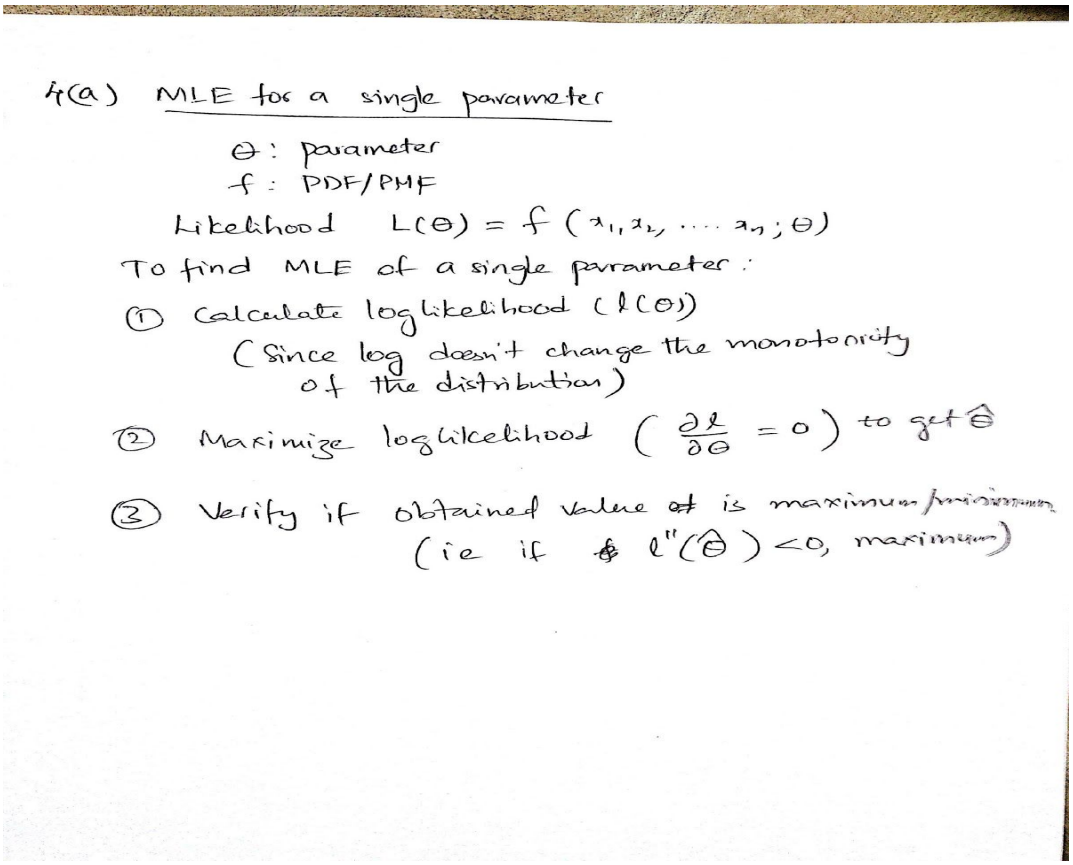
E(*proportion of* $70 \leq X < 80$)=26.11%

E(*proportion of* $60 \leq X < 70$)=21.08%

E(proportion of $X < 60$)=15.87%

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

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



4(a) MLE for a single parameter

θ : parameter
f : PDF/PMF

Likelihood $L(\theta) = f(x_1, x_2, \ldots x_n; \theta)$

To find MLE of a single parameter :

① Calculate log likelihood ($l(\theta)$)
(Since log doesn't change the monotonicity of the distribution)

② Maximize log likelihood $\left(\frac{\partial l}{\partial \theta} = 0\right)$ to get $\hat{\theta}$

③ Verify if obtained value ~~of~~ is maximum/minimum
(ie if $l''(\hat{\theta}) < 0$, maximum)

(b) The Pareto distribution is sometimes used to model heavy tailed distributions. Consider a Pareto distribution with density function given by:
$f(x;\theta) = (\theta - 1)x^{-\theta}$ *if* $\theta > 2$ *and* $1 \leq x < \infty$

If $X_1, X_2, X_3, \ldots, X_n$ are i.i.d with density function given by $f(x;\theta)$, calculate MLE for θ. **(8 points)**

Solution:

$$f(x) = (\theta - 1) x^{-\theta}$$

$$L(\theta) = (\theta - 1)^n \prod_i x_i^{-\theta}$$

$$l(\theta) = n \log(\theta - 1) - \theta \sum_{i=1}^{n} \log x_i$$

$$\frac{dl}{d\theta} = 0 \Rightarrow \frac{d}{d\theta}\left(n \log(\theta - 1) - \theta \sum_{i=1}^{n} \log x_i = 0\right)$$

$$\theta = \frac{n}{\sum \log x_i} + 1$$

(5) CLT, and CI (**15 points**)

(a) Define Confidence Interval for population mean (2.5 points)

Solution:

Interval around the population mean covering possible range of values.

Computed using $\bar{x} + / - z * \frac{s}{\sqrt{n}}$

Where $\bar{x}$ is sample mean, z is z-score, s is sample std. dev. and n is sample size.

(b) Outline the procedure for finding C.I. (2.5 points)

Solution:

Procedure for finding CI:

1. Identify sample statistic
2. Select confidence level
3. Compute margin of error

Margin of error = Critical value * Standard deviation of statistic

Or

Margin of error = Critical value * Standard error of statistic

1.1 Critical value computation:
Compute alpha: $\alpha = 1 - \frac{Confidence\ Level}{100}$
Compute critical probability $p^* = 1 - \frac{(\alpha)}{2}$
Compute z score from z-table using this critical probability

2. CI = Sample statistic +/- Margin of Error

(c) The following data for a sample of 40 users from a social media site shows number of friends for each user. Compute the 97% CI for the point estimate of mean, and margin of error. (5 points)

28 32 45 28 65 45 29 31 23 34
35 31 23 54 34 25 23 15 65 38
64 65 46 56 36 45 67 65 54 66
45 56 57 45 38 48 25 26 34 36

Solution:

1. sample statistic:

mean = 41.925

2. Select confidence level: 95%
3. Computer margin of error:

$$\alpha = 1 - \frac{Confidence\ Level}{100} = 0.03$$

$$p^* = 1 - \frac{(\alpha)}{2} = 0.985$$

$$z = 2.17$$

Margin of error = $2.17 * \frac{s}{\sqrt{n}} = 2.17 * \frac{14.93299}{\sqrt{40}} = 5.123616$

4. CI = 41.925 ± 5.123616 => [36.80138, 47.04862]

(6) Hypothesis testing (5x3 = 15 points)

(a) A student researcher claims that the average cost of an engineering book is less than $80. He selects a random sample of 36 books from University engineering book stores, where cost of each book in $s is listed below:

50 95 120 85 45 90 70 60 70 50 40 80 70 90 75 60 90 90 75 85 80 60 110 65 80 85 85 45 60 95 110 70 75 55 80 55.

Assume $\sigma = 19.2$. Is there enough evidence to support the student researchers claim at $\propto = 0.10$?

Since we know that $\sigma = 19.2$ and the sample size n >= 30, we use the z test.

Step 1: State the hypothesis

$H_0 : \mu = 80$

$H_1 : \mu < 80$ (claim)

Step 2: Find the critical value. Since $\propto = 0.10$, and the test is left-tailed, the critical value is z' = -1.28

Step 3: Compute the test value

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 75$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{75 - 80}{19.2/\sqrt{36}} = -1.5625$$

Step 4: Make the decision

Since the test value -1.5625 is less than the critical value and is in the critical region, the decision is to reject the null hypothesis.

Step 5: There is enough evidence to support the claim that the

average cost is less than 80.

(**b**) The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.

(i) state null and alternative hypothesis

$H_0: \mu \leq 31, H_1: \mu > 31$ (right-tail test)

(ii) is data significant at 1% levlel?

$\propto = 0.01$ (given)

p = 0.0264 (given)

Since p = 0.0264 > 0.01 (right tailed test), so we don't reject the null hypothesis. That is, data is not significant.

P-value interpretation: p value is the the probability of getting a sample mean of 32 given that the population mean being at most 31 is true.