Data Analytics (CS40003)

Practice Set IV
(Topic: Probability and Sampling Distribution)

I. Concept Questions

1. Give an example of a random variable in the context of "Drawing a card from a deck of cards".

2. What is the difference between "Probability distribution" and "Sampling distribution"?

3. Decide the parameter(s), which you should decide for each of the following probability discrete distribution functions
   a) Discrete uniform distribution
   b) Binomial distribution
   c) Poisson's distribution

4. Decide the parameter(s), which you should decide for each of the following continuous probability distribution functions
   a) Continuous uniform normal distribution
   b) Normal distribution
   c) Standard normal distribution
   d) Chi-squared distribution
   e) Gamma distribution

5. Write the full expression for each of the following discrete probability distribution functions, you should clearly mention each term in them.
   a) Binomial probability distribution
   b) Multinomial probability distribution
   c) Hyper geometric probability distribution
   d) Multivariate hyper geometric probability distribution
   e) Poisson's distribution

6. Write the full expression for each of the following continuous probability distribution functions, you should clearly mention each term in them.

a) Normal distribution
b) Standard normal distribution
c) Chi-squared distribution
d) Gamma distribution

7. Find the mean and variance for each of the following probability distribution functions.

a) Discrete uniform distribution
b) Binomial distribution
c) Poisson's distribution
d) Continuous uniform distribution
e) Normal distribution
f) Chi-squared distribution
g) Gamma distribution
h) Weibull distribution

8. Give an example of populations of the following probability distribution function, so that the population satisfies the distribution.

a) Normal distribution
b) Binomial distribution
c) Poisson's distribution
d) Chi-squared distribution
e) Weibull distribution
f) Gamma distribution
g) Hyper geometric distribution

9. Point out the major differences between each pair of the probability distributions

a) Normal distribution, Standard Normal distribution
b) Binomial distribution, Multinomial distribution
c) Binomial distribution, Hyper geometric distribution
d) Hyper geometric distribution, Multivariate Hyper geometric distribution.

10. Suppose, X is normally distributed with $\mu = 10$ and $\sigma^2 = 20$.

a) What is P(X>15) ?
b) What is P(5<X<15) ?
c) What is P(5<X<10) ?

[Hint: Use table of standard normal distribution]

11. Let X is the random variable representing the distribution of grades in the Data Analytics course. It is observed that grades are approximately normally distributed with $\mu = 75$ and $\sigma = 10$. If the instructor wants more than 10% of the class to get an EX, what should be the cut-off grade?

12. A supplier supplies 8 pcs to a retail outlet, which contains 3 of them are defective. If an office makes a random purchase of 2 of these computers, find the probability distribution for the number of defectives.

13. Classify the following random variables as discrete or continuous:
X: the number of automobile accidents per year in Virginia.
Y: the length of time to play 18 holes of golf.
M: the amount of milk produced yearly by a particular cow.
N: the number of eggs laid each month by a hen.
P: the number of building permits issued each month in a certain city.
Q: the weight of grain produced per acre.

14. Determine the value *c* so that each of the following functions can serve as a probability distribution of the discrete random variable *X*:
(a) $f(x) = c(x^2 + 4)$, for $x = 0, 1, 2, 3$;
(b) $f(x) = c\binom{2}{x}\binom{3}{3-x}$, for $x = 0, 1, 2$.

15. The shelf life, in days, for bottles of a certain prescribed medicine is a random variable having the

density function

$$f(x) = \begin{cases} \dfrac{20{,}000}{(x+100)^3}, & x > 0 \\ 0, & elsewhere \end{cases}$$

Find the probability that a bottle of this medicine will have a shell life of

(a) at least 200 days;

(b) anywhere from 80 to 120 days.

16. The total number of hours, measured in units of 100 hours, that a family runs a vacuum cleaner over a period of one year is a continuous random variable X that has the density function

$$f(x) = \begin{cases} x, 0 < x < 1 \\ 2 - x, 1 < x < 2 \\ 0, elsewhere \end{cases}$$

Find the probability that over a period of one year, a family runs their vacuum cleaner

(a) less than 120 hours;
(b) between 50 and 100 hours.

17. A continuous random variable X that can assume values between $x = 2$ and $x = 5$ has a density function given by $f(x) = 2(1 + x)/27$. Find

(a) $P(X < 4)$;

(b) $P(3 \leq X < 4)$.

18. Suppose it is known from large amounts of historical data that X, the number of cars that arrive at a specific intersection during a 20-second time period, is characterized by the following discrete probability function:

$$f(x) = e^{-6}\frac{6^x}{x!}, for\ x = 0,1,2, ...$$

(a) Find the probability that in a specific 20-second time period, more than 8 cars arrive at the intersection.

(b) Find the probability that only 2 cars arrive.

19. According to *Chemical Engineering Progress* (November 1990), approximately 30% of all pipework failures in chemical plants are caused by operator error.

(a) What is the probability that out of the next 20 pipework failures at least 10 are due to operator error?

(b) What is the probability that no more than 4 out of 20 such failures are due to operator error?

(c) Suppose, for a particular plant, that out of the random sample of 20 such failures, exactly 5 are due to operator error. Do you feel that the 30% figure stated above applies to this plant? Comment.

20. A nationwide survey of college seniors by the University of Michigan revealed that almost 70% disapprove of daily pot smoking, according to a report in *Parade*. If 12 seniors are selected at random and asked their opinion, find the probability that the number who disapprove of smoking pot daily is

(a) anywhere from 7 to 9;

(b) at most 5;

(c) not less than 8.

21. A traffic control engineer reports that 75% of the vehicles passing through a checkpoint are from within the state. What is the probability that fewer than 4 of the next 9 vehicles are from out of state?

22. Three people toss a fair coin and the odd one pays for coffee. If the coins all turn up the same, they are tossed again. Find the probability that fewer than 4 tosses are needed.

23. On average, a textbook author makes two word processing errors per page on the first draft of her textbook. What is the probability that on the next page she will make

(a) 4 or more errors?

(b) no errors?

24. A certain area of the eastern United States is, on average, hit by 6 hurricanes a year. Find the probability that in a given year that area will be hit by

(a) fewer than 4 hurricanes;

(b) anywhere from 6 to 8 hurricanes.

25. Find the mean, median, and mode for the sample whose observations, 15, 7, 8, 95, 19, 12, 8, 22, and 14, represent the number of sick days claimed on 9 federal income tax returns. Which value appears to be the best measure of the centre of these data? State reasons for your preference.

26. For the sample of reaction times in Exercise 8.3, calculate

(a) the range;

(b) the variance, using the formula of form (8.2.1).

27. The random variable X, representing the number of cherries in a cherry puff, has the following probability distribution:

x	4	5	6	7
$P(X = x)$	0.2	0.4	0.3	0.1

(a) Find the mean μ and the variance σ^2 of X.

(b) Find the mean $\mu(X)$ and the variance $\sigma^2(X)$ of the X for random samples of 36 cherry puffs.

(c) Find the probability that the average number of cherries in 36 cherry puffs will be less than 5.5.

28. Show that the variance of $S2$ for random samples of size n from a normal population decreases as n becomes large. [*Hint*: First find the variance of $(n - 1)S^2/\sigma^2$.]

29. A normal population with unknown variance has a mean of 20. Is one likely to obtain a random sample of size 9 from this population with a mean of 24 and a standard deviation of 4.1? If not, what conclusion would you draw?

30. Construct a normal quantile-quantile plot of these data, which represent the diameters of 36 rivet heads in 1/100 of an inch:

6.72	6.77	6.82	6.70	6.78	6.70	6.62
6.75	6.66	6.66	6.64	6.76	6.73	6.80
6.72	6.76	6.76	6.68	6.66	6.62	6.72
6.76	6.70	6.78	6.76	6.67	6.70	6.72
6.74	6.81	6.79	6.78	6.66	6.76	6.76
6.72						

II Objective Questions

1. Which of the following probability distributions function belong to discrete probability distribution?

(a) Binomial distribution
(b) Poisson's distribution
(c) Hypergeometric distribution
(d) Weibull distribution

2. If μ and σ denote the mean and standard deviation of a population, then the standard normal distribution is better described as

(a) $f(x: A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$

(b) $f(x: \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$

(c) $f(z: 0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty < z < \infty$

(d) $f(x: \mu, \sigma) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$

3. Which of the following statements is/are **not** correct?

(a) $\frac{1}{\sigma^2}\sum(x_i - \mu)^2$ is a Chi-square distribution with *n*-degrees of freedom

(b) $\frac{(n-1)S^2}{\sigma^2}$ is a Chi-square distribution with (*n*-1) degrees of freedom

(c) $\frac{(\overline{x}-\mu)^2}{\sigma^2/n}$ is Chi-square distribution with 1 degree of freedom

(d) None of the above

4. Which of the following statement is correct?

(a) χ^2-distribution is used to describe the sampling distribution of S^2
(b) t-distribution is used when population mean μ is known and standard deviation of sample S is known
(c) F distribution is used when variance of populations ${\sigma_1}^2$, ${\sigma_2}^2$ and samples ${S_1}^2$, ${S_2}^2$ are known

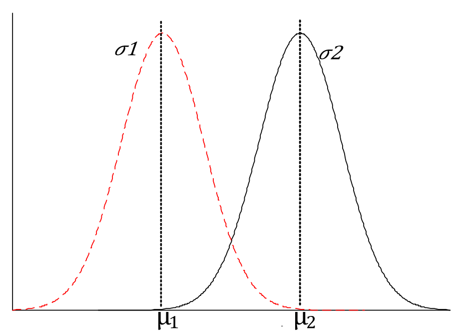
(d) All of the above

5. In the following table, Column A lists some sampling distributions, whereas Column B lists the names of sampling distributions. All symbols bear their usual meanings. The matching from Column A and Column B is

(A)	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	(W)	Normal distribution
(B)	$\frac{\bar{X}-\mu}{S/\sqrt{n}}$	(X)	Chi-squared distribution
(C)	$\frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2}$	(Y)	t-distribution
(D)	$\frac{(n-1)S^2}{\sigma^2}$	(Z)	F distribution

(a)

6. With reference to the following figure, which option correctly represents the two normal distributions?



(a) $\sigma_1 \leq \sigma_2$, $\mu_1 = \mu_2$
(b) $\sigma_1 = \sigma_2$, $\mu_1 \geq \mu 2$
(c) $\sigma_1 \leq \sigma_2$, $\mu_1 = \mu 2$
(d) $\sigma_1 = \sigma_2$, $\mu_1 \leq \mu 2$

7. In the following context, which denotes a random variable?
There is a box containing 100 balls: 30 red, 20 blue and 50 black balls.
(a) Probability that we draw two blue balls or two red balls from the box.
(b) Drawing any 5 balls at random.
(c) The number of red, blue and black balls drawn from the box.
(d) Drawing five red and six black balls drawn from the box.

8. Which of the following statement(s) is(are) not true in the context of any continuous probability distribution functions?
(a) $a \leq f(x) \leq 1$ for all $x \in R$

(b) $\sum_{i=1}^{n} f(x_i) = 1$
(c) $\mu = \int_{-\propto}^{\propto} xf(x)dx$
(d) $\sigma^2 = \int_{-\propto}^{\propto} (x-\mu)^2 f9x0dx$

9. Central limit theorem is applicable to

(a) Only continuous probability distributions
(b) Only discrete probability distributions
(c) Any probability distribution
(d) Only normal distribution

10. How factorial of a fraction say $\frac{1}{2}$ can be calculated?

(a) It cannot be calculated for a fractional number
(b) Is known as a universal constant
(c) It can be calculated using Gamma function.
(d) Factorial of a fraction or negative number does not bear any physical significance.