# BIOE: Biostatistics Course Fall 2017

## Mid Term: Roughly 60 mins

### 2017, v1.0

**Put your name on the top of the answer sheet. Total points 76.**

This is an an open book test. Answers on separate sheets (provided).

You may use your crib sheet(s), text book and the internet. You may **not** cut and paste python code from the internet.

When doing the Python questions, write the python code on your own computer, tablet etc, and then write out the answers you get onto the answer sheet. If the question asks for a graph via Python, sketch the graph as accurately as you can on your answer sheet.

Show all your workings for each question (except for the python based questions)

If you want your crib sheet graded (A to C), hand that in with your answers.

This exam has 26 questions, for a total of 76 points.

1. (1 point) Name each of the following symbols:
    1. $\sigma$
    2. $\bar{x}$
    3. $\sigma^2$
    4. $s$
    5. $s^2$

**Solution:**

1. $\sigma$: standard deviation of population
2. $\bar{x}$: mean of a sample
3. $\sigma^2$ : Variance of a population
4. $s$: Standard deviation of a sample
5. $s^2$: Variance of a sample

2. (2 points) Write out the formula for computing the:

    1. Sample mean
    2. Population standard deviation
    3. Sample standard deviation
    4. Population variance.

**Solution:**

1. $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$
3. $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
4. $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

3. (1 point) Spell out the following acronyms:

    1. PMF
    2. PDF
    3. CDF
    4. SE

**Solution:**

1. Probability Mass Function
2. Probability Density Function
3. Cumulative Density Function
4. Standard Error

4. (2 points) Based on the first assignment you did, give the intuition why we divide the sample standard deviation by $n-1$ instead of $n$? Don't given the answer in terms of degrees of freedom.

> **Solution:** The reason we divide by $n-1$ is to take into account the inaccuracy in the sample average which we tend to assume is roughly the same as the population means. For small samples however, the sample mean is a very rough approximation which means that the computed standard deviation is also affected. In fact the sample deviation tends to be lower, to adjust for this we divide by $n-1$ which tends to increase the estimate closer to the true value.

5. (1 point) Given a fair six sided die what is the probability of throwing a 1 or 2 in the next roll?

> **Solution:** $1/6 + 1/6 = 2/6$

6. (1 point) What is the chance of not rolling a 2?

> **Solution:** $5 \times 1/6 = 5/6$

7. (1 point) Consider rolling two fair die. What is the chance of getting a 1 in the first roll and a 2 in the second roll?

> **Solution:** $1/6 \times 1/6 = 1/36$

8. (1 point) Write out the general probability addition rule.

> **Solution:** $P(A \cup B) + P(A) + P(B) - P(A \cap B)$
>
> or
>
> $P(A \text{ or } B) + P(A) + P(B) - P(A \text{ and } B)$

9. Given the pack of playing card shown in the figure below:


Representations of the 52 unique cards in a deck.

(a) (2 points) What is the probability that a randomly selected card is a diamond?

**Solution:** 0.25

(b) (2 points) What is the probability that a randomly selected card is a face card?

**Solution:** 0.231

10. (2 points) How many different ways can the letters of the word 'LEADING' be arranged in such a way that the vowels always come together?

    Vowels: a, e, i, o, u

**Solution:** The word 'LEADING' has 7 different letters.

When the vowels EAI are always together, they can be supposed to form one letter.

Then, we have to arrange the letters LNDG (EAI).

Now, 5 (4 + 1 = 5) letters can be arranged in 5! = 120 ways.

The vowels (EAI) can be arranged among themselves in 3! = 6 ways.

Required number of ways = (120 x 6) = 720.

11. (2 points) How many ways can the word 'PEPPER' be arranged?

**Solution:**

$\frac{6!}{3!2!} = 60$

12. The average time it takes you to cycle to and from UW has a mean of 20 minutes and standard deviation of 5 minutes. Assume that your travel time on each day of the week is independent. Over an entire 5 day week what is:

    (a) (1 point) The total time taken to cycle to and from UW

**Solution:** $X_1 + X_2 + X_3 + X_4 + X_5$

5 times 20 = 100 minutes

    (b) (2 points) The standard deviation of the total weekly commute time

**Solution:** $Var(\sum aX) = \sum a^2 Var(X)$

$Var = 1^2\ 5^2 + 1^2\ 5^2 + 1^2\ 5^2 + 1^2\ 5^2 + 1^2\ 5^2 = 125$

$sd = \sqrt{125} = 11.18$

(c) (3 points) On the Monday you had to go back home at lunch time to pick up some books you'd forgotten. This meant you cycled to and from our house twice on the Monday. Recompute the total time traveled and the standard deviation for this new situation.

**Solution:** $2X_1 + X_2 + X_3 + X_4 + X_5$

2 times 20 + 4 times 20 = 120 minutes

$Var(\sum aX) = \sum a^2 Var(X)$

$Var = 2^2\ 5^2 + 1^2\ 5^2 + 1^2\ 5^2 + 1^2\ 5^2 + 1^2\ 5^2 = 200$

$sd = \sqrt{200} = 14.14$

13. What is the mean and standard deviation for the following distributions:

(a) (2 points) Binomial Distribution

**Solution:** $\mu = np$

$\sigma = \sqrt{np(1-p)}$

(b) (2 points) Poisson Distribution

**Solution:** $\mu = \lambda$

$\sigma = \sqrt{\lambda}$

14. Let X represent a random variable from $N(\mu = 3, \sigma = 2)$, and suppose we observe $x = 5.19$.

(a) (2 points) Find the $z$-score of $x$.

**Solution:** 1.095

(b) (2 points) Use the $z$-score to determine the area under the normal curve to the **right** of the z-score.

**Solution:** $z = (5.19 - 3)/2 = 1.1$

$1 - 0.86433 = 0.136$

15. (2 points) GRE scores can be approximated by a normal distribution, $N(\mu = 1500, \sigma = 300)$.

(a) (2 points) What is the probability that a student will get at least a score of 1630?

**Solution:** $z = (1630 - 1500)/300 = 0.433333$

0.668

(b) (2 points) What is the percentage of students who get an GRE score between 900 and 1200?

**Solution:** $z_{900} = 0.02275$ $z_{1200} = 0.158655$ $P\% = 15.8655 - 2.275 =$ 13.6 0.135

16. What is the area (**to four decimal places**) that falls between:

(a) (1 point) $z = -1$ and $z = 1$

**Solution:** 0.6827

(b) (1 point) $z = -2$ and $z = 2$

**Solution:** 0.9545

(c) (1 point) $z = -3$ and $z = 3$

**Solution:** 0.9973

17. (3 points) List the values **and** draw on the provided graph paper a Binomial distribution where $p = 0.2$ and $n = 8$. You may use any means to compute the probabilities, eg you could use Python, calculator etc.

**Solution:** 0.1677, 0.33554, 0.2936, 0.1468, 0.04588, 0.00918, 0, 0, 0

18. (3 points) What are the four conditions necessary to claim that a given random variable is distributed Binomially?

**Solution:** 1. Trials are independent
2. The number of trials, $n$, is fixed
3. Each trial outcome can be classified as a success or failure
4. The probability of success, $p$, is the same for each trial

19. The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0.3.

Select four smokers at random.

(a) (2 points) Is the binomial model an appropriate distribution to describe the prevalence of lung condition in the four smokers? Justify your answer.

**Solution:**

1. Whether a smoker gets a lung disease or not is independent of other four smokers
2. We selected 4 smokers at random, therefore number of trials is fixed
3. The outcome is lung disease or no lung disease
4. The biggest assumption is that the likelihood of getting a lung condition is equally likely in all four smokers, $p = 0.3$

(b) (2 points) What is the probability that none of the four smokers will develop a severe lung condition?

**Solution:** Assume a binomial model, with $p = 0.3$ and $n = 4$.
0.2401

(c) (2 points) What is the probability that one of the four smokers will develop a severe lung condition?

**Solution:** 0.4116

(d) (2 points) What is the probability that more than one will develop a severe lung condition?

**Solution:** $0.2646 + 0.0756 + 0.0081 = 0.3483$

20. (2 points) What are the Poisson distribution assumptions?

**Solution:**

1. Events are independent
2. Probability of two events occurring at the same time is negligible
3. The probability of an even is proportional to the length of the interval.

21. (3 points) Give three examples of processes (at least one must be biological) which could be described using a Poisson distribution.

**Solution:** Biological examples: Number of cell distributed on a hemocytometer per $mm^2$; the number of trees in an area of land; number of DNA mutations in a gene over a period of time;

22. (3 points) In a sample of 100 bunny rabbits the average diameter of a bunny's tail is 2.5 cm with a standard deviation of 0.5 cm. Compute the 95% confidence limits for the average diameter of a tail.

**Solution:** 0.098

23. (1 point) What type of distribution is recommended when the sample size is below 30?

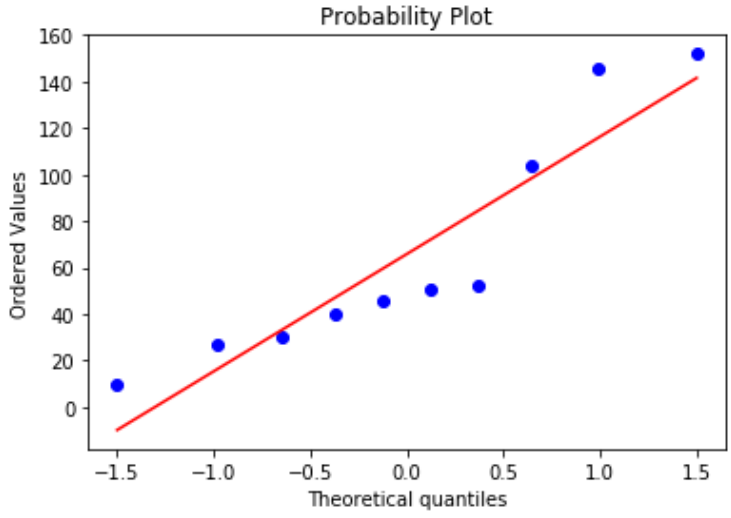
**Solution:** t distribution

24. The following set of data were collected from 10 petri dish plates. The data records the number of colonies per plate.

    10, 27, 30, 40, 46, 51, 52, 104, 146, 152

    (a) (2 points) Plot a Q-Q plot of the raw data. Sketch the scatter plot you obtain. Explain what you expect to see if the data were normally distributed. Suggest whether the data is normally distributed or not.

**Solution:**




    (b) (3 points) Calculate the 95% confidence limits using a t distribution.

**Solution:**
Mean = 65.8

Sample SD = 50.210

$65.8 \pm 2.26 \times 15.88 = 65.8 \pm 35.918$

(c) (3 points) Use a bootstrap technique to estimate the 95% confidence limits. Sample the data 50,000 times.

**Solution:** For 50,000 samplings:

2.5% and 95% percentiles with 50,000 samplings:
38.397.0

25. (2 points) Give an example of a null hypothesis and its alternative, $H_1$.

**Solution:**
$H_o$ : Nothing happens

$H_1$ : Something happens

26. (2 points) A new drug is being developed to treat migraine. The pharmaceutical company will set up a trail to test its effectiveness on a sample population. There is strong evidence that the drug will reduce the number of migraines but by how much is not known.

The company will determine the mean number of migraines in a sample population, let this mean be represented by $\mu$ migraines per week.

The company will administer the drug to the same sample population and again determine the mean number of migraines per week, this will be denoted by $\mu_o$ migraines per week.

State the null and alternative hypotheses in this context.

**Solution:**
$H_o : \mu_o = \mu$

$H_1 : \mu_o < \mu$