PH142 Midterm II SOLUTIONS

October 22, 2018

First and last name (print clearly): _____

Student number (print clearly): _____

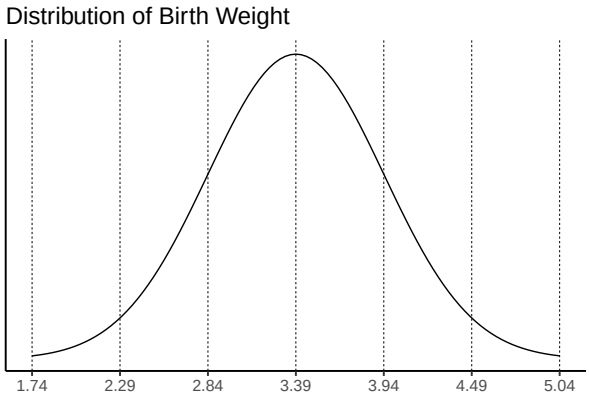
Question	points
1	7
2	3
3	14
4	3
5	8
6	6
7	4
Total	45

Notes:

- The last page of the exam includes the list of relevant R functions. Do not tear off this (or any) page.
- You can use the back of each page as scratch paper.
- No points will be given for answers on the back of each page.
- Cellphones and computers must be stored and on silent.
- You must show your student ID when you submit your test.

Question 1 [7 points total]

The following curve represents the birth weight of 3,226 newborn babies studied in O'Cathain et al (2002). The mean birthweight was 3.39 kilograms. The dashed lines on the plot represent standard deviation-sized intervals. Answer the following questions about this curve.



1.1 [2 points] What is the variance of the birthweight? What are the units for variance?

(3.940-3.39)^2 = 0.55^2 = 0.3025 kilograms squared. units for variance: kg^2

1.2 [1 point] What is the probability that a baby weighs exactly 3.39 kilograms at birth?

P(W = 3.39) = 0, because for continuous distributions the probability of seeing an exact amount always equals 0.

1.3 [1 point] What is the approximate probability that a baby weighs more than 4.49 kilograms at birth?

P(W > 4.49) = 2.5%. This is because 4.49 is at +2SD. We know that approximately 95% of the data is between +/-2SD, which implies that 5% of the data is above or below +/-2SD, or 2.5% above 4.49 grams.

1.4 [1 point] Write R code to calculate the probability that a baby's birth weight is between 2 and 3 kilograms.

```
pnorm(q = 3, mean = 3.39, sd = 0.55) - pnorm(q = 2, mean = 3.39, sd = 0.55)
```

1.5 [2 points] Write R code to calculate the probability that a baby's birth weight is less than 3 kilograms. *Only use function(s) that assume a standard Normal distribution (i.e.,* $N(0,1)$*).*

$P(W < 3) = P(Z < \frac{3-\mu}{\sigma}) = P(Z < \frac{3-3.39}{0.55} = P(Z < -0.71))$

Thus the function is:

`pnorm(q = -0.71, mean = 0, sd = 1)` or simply `pnorm(q = -0.71)`

Question 2 [3 points total]

Sickle-cell anemia is thought to protect against malaria. A study in malaria-endemic African countries tested 742 children for the sickle-cell trait and also for malaria infection. In all, 23% of the children had the sickle-cell trait, and 6.6% of the children had both the sickle-cell trait and malaria. Overall, 39.2% of the children had malaria.

2.1 [1 point] Are the events 'sickle-cell trait' and 'malaria' independent? Support your answer using calculations.

If these events were independent then $P(SC \cap M) = P(SC) \times P(M)$.

P(SC)=0.23

P(M)=0.392

–> P(SC)*P(M) = 0.09016. However, $P(SC \cap M) = 0.066$, implying that these events are not independent.

2.2 [1 point] What is the probability that a given child has neither malaria nor sickle-cell trait?

1 - P(SC or M) = 1 - [P(SC) + P(M) - P(SC & M)] = 1 - [0.23 + 0.392 - 0.066] = 0.444 = 44.4%

2.3 [1 point] What is the probability that a given child has malaria but does not have sickle-cell trait?

Easiest to draw a Venn diagram for those one. Then see it is 39.2 - 0.6 = 32.6%

39.2 - 6.6

= 32.6%

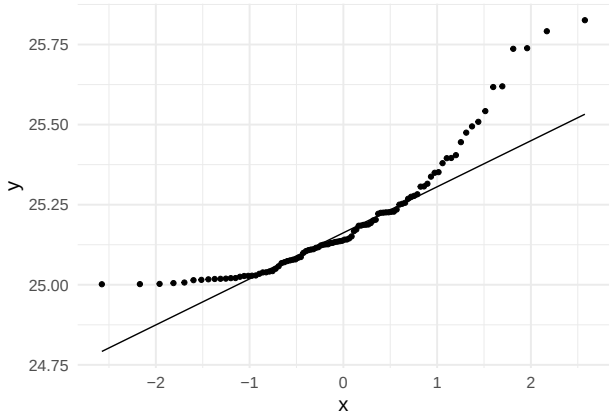
Question 3 [14 points total]

Circle the correct answer(s):

3.1 [1 point] A collaborator approaches you with data from a study on childhood obesity. The participants are part of a random sample that is representative of the children and adolescents between the ages of 8 and 18 living in Berkeley, California. Your collaborator needs help determining if the participants' BMI scores are normally distributed. Based on the following qqplot, can we conclude that `child_bmi` is Normally distributed?

```
i. Yes, ‘child_bmi‘ appears approximately Normal
ii. No, ‘child_bmi‘ does not appear approximately Normal
iii. This plot does not help determine Normality
```

Solution: (ii)



3.2 [1 point] Two disjoint events cannot be independent.

```
i) True
ii) False
```

Solution: i) True (disjoint events must be dependent)

3.3 [1 point] Any particular normal distribution is completely specified by its mean.

```
i) True
ii) False
```

Solution: (ii) False (you also need to know the sd)

3.4 [1 point] Any particular Poisson distribution is completely specified by its mean.

```
i) True
ii) False
```

Solution: i) True

3.5 [1 point] Which of the following properties does the Normal distribution have? Circle all that apply.

```
i) approximately 95% of the data lie within 1 standard deviation of the mean.
ii) The distribution is symmetric.
iii) It is used to model rare events.
iv) The total area under the curve is equal to 1.
```

Solution: ii) and iv). i) is wrong (should be 68%), iii) is wrong (that is Poisson)

3.6 [1 point] Suppose that you have an unbiased estimator of a parameter. Under what conditions is the mean of the sampling distribution equal to the parameter? Pick only one option.

```
i) If the sample size $n$ is large enough for the Central Limit Theorem to apply.
ii) If the parameter is Normally distributed.
iii) Any conditions. This is true by the definition of an unbiased estimator.
```

Solution: iii) Any conditions. This is true by the definition of an unbiased estimator.

Fill in the blank in the following statements:

3.7 [1 point] Let X be an observation generated by a standard normal distribution. The probability that $X > 0$ is _____%.

Solution: 50%

3.8 [1 point] A sample is to a statistic, like a population is to a _____.

Solution: parameter

3.9 [1 point] _____ is a method for drawing conclusions about a population from a sample.

Solution: Inference

3.10 [2 points] The standard normal distribution has a mean of _____, a variance of _____, and a standard deviation of _____.

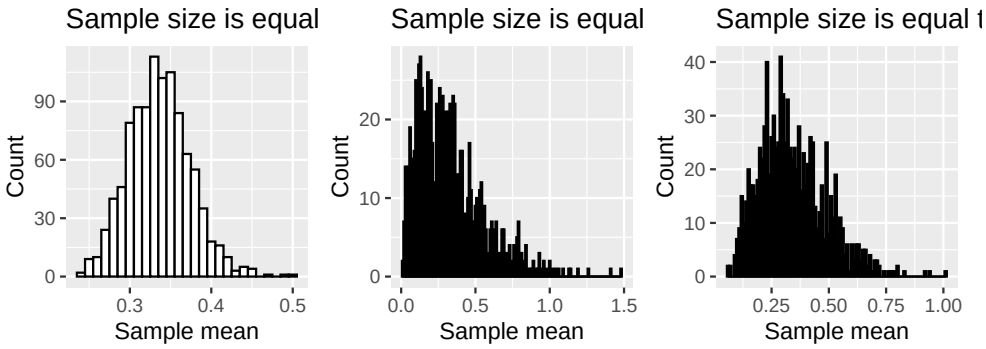
Solution: 0, 1, 1

3.11 [1 point] Suppose you took 20 samples from a population and created a 90% confidence interval around each sample mean $\bar{x}$. How many confidence intervals would you expect to not include the true value for μ from the underlying population? _____

Solution: 2

3.12 [1 point] Each of the plots below correspond to the approximate sampling distribution for the mean for the same underlying population distribution. The only difference is their underlying sample sizes. Those sample sizes are 2, 5, and 20. Label each plot with its corresponding sample size.

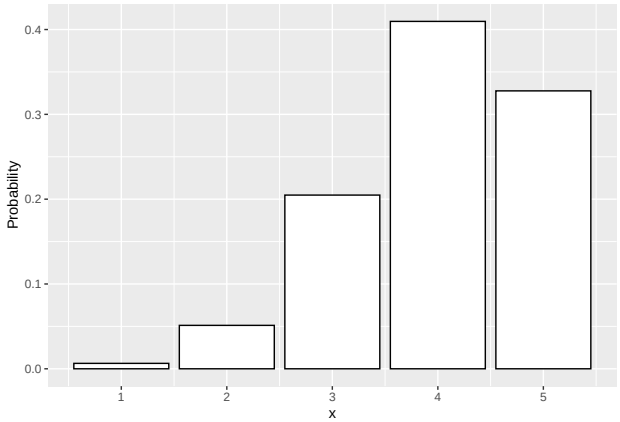
Solution: 20, 2, 5



3.13 [1 point] Using your pen or pencil, shade the bar(s) that correspond to this piece of R code.

```
dbinom(x = 2, size = 5, prob = 0.8)
```

Solution: Shade the X=2 bar only.



Question 4 [3 points total]

Suppose there is a new medication to help prevent against a certain disease. In a sample of 100 people, 65 took the medication and 5 of those people got the disease. The other 35 did not take the medication, and 25 of those people got the disease.

4.1 [1 point] Using calculations involving probabilities or absolute frequencies, find the probability that a person who takes the medication will not get the disease. Remember that probabilities range from 0 to 1 (i.e, Do not report the answer in number of people).

Solution: $60/65 = 92.31\%$

4.2 [2 points] Using calculations involving probabilities or absolute frequencies, what is the probability someone who did not get the disease took the medication?

Solution: $60/70 = 85.71\%$

Question 5 [10 points total]

Senioritis, a disease known to cause mediocre grades and odd sleeping habits, sweeps UC Berkeley seniors every year. A statistician in Evans Hall modeled the probability of k senior-standing students with the following function. Let X be the number of infected Berkeley seniors.

$$P(X = k) = \binom{n}{k}(0.67)^k(0.33)^{n-k}$$

5.1 [2 points] What is the name of this distribution? What does n represent?

Binomial. n is the sample size of the number of seniors.

5.2 [1 point] According to the distribution, what is the probability that any given senior has senioritis?

67%.

5.3 [1 point] What did the statistician assume when making the probability model? Circle all that apply.

```
i)  Dependence between seniors.
ii) The event of senioritis has a binary outcome.
iii) The number of seniors is finite and equal to $n$.
iv) The chance that a senior gets senioritis is equal amongst all Berkeley seniors.
```

Solution: ii), iii), iv)

5.4 [1 point] Write a probability statement for the exact probability that half of the seniors at UC Berkeley get senioritis. Assume that the number of Berkeley seniors total is 10,000.

$P(X = 500) = \binom{10000}{5000}(0.67)^{5000}(0.33)^{5000}$

5.5 [1 point] Use an R function to calculate the probability that more than 6700 of the seniors are sick with senioritis. You don't need to perform the calculation.

```
1 - pbinom(q=6700, n = 10000, p = 0.67)
```

or

```
pbinom(q=6700, n = 10000, p = 0.67, lower.tail=F)
```

5.6 [3 points] We can approximate the value from part (e) with a distribution, under certain conditions. List those conditions and specify the distribution and its mean and standard deviation.

Conditions: n large enough such that $np > 10$ and $n(1-p) > 10$

If so, can approximate with a Normal distribution with $N(\mu = np, \sigma = \sqrt{np(1-p)}$

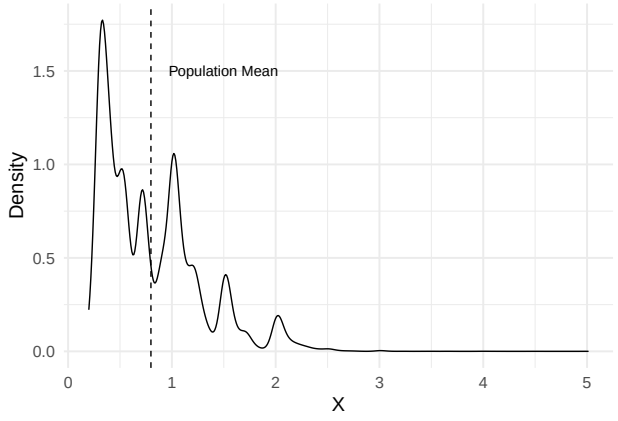
$= N(10000 \times 0.67, \sqrt{10000 \times 0.67 \times 0.33}) = N(6700, 47.02)$

5.7 [1 point] Write R code that would calculate the probability of observing 6700 or less events using the distribution from (f) and applying the continuity correction.

```
pnorm(q = 6700.5, mean = 6700, sd = 47.02)
```

Question 6 [6 points total]

Suppose a random variable X has the following population distribution, with mean $\mu = 0.8$ and standard deviation $\sigma = 0.47$.



6.1 [3 points] Draw the sampling distribution of the mean $\bar{X}$ with sample size $n = 1000$. What are the mean and standard deviation of this distribution? Label the mean on your drawing.

Draw a Normal shaped curve that is centred at $\mu = 0.8$. The sd of the distribution is $\sigma/\sqrt{n} = 0.47/\sqrt{1000} = 0.1486271$

6.2 [1 point] What would happen to the shape and/or location of the distribution from 6.1 if n were increased?

As n increases the width (standard deviation) decreases and the distribution becomes narrower.

6.3 [1 point] What would happen to the shape and/or location of the distribution from 6.1 if μ were increased?

If μ increases the sampling distribution shifts to the right (to be centered at the new μ).

6.4 [1 point] What would happen to the shape and/or location of the distribution from 6.1 if σ were increased?

Solution: If σ increases then $\sigma/\sqrt{n}$ also increases and the distribution becomes wider.

###Question 7 [4 points]

Many counties in the Bay Area have banned the sale of flavored tobacco (found in cigarettes and e-cigarettes) with the goal of reducing tobacco addiction, especially among historically marginalized groups. Suppose you were interested in estimating the impact of these bans on tobacco addiction and examined rates of addiction in the 2 years preceding and 2 years after the bans using two simple random samples (one before the ban, one after).

a) [1 point] If you saw a sizable reduction in addiction, can you attribute it to the ban? That is, what is one problem with the study design, that increasing sample size cannot solve?

Solution: No. Because it could be that something else caused the reduction in addiction. The problem is confounding. [Students don't have to say confounding specifically, but just describe that the cause of the reduction could have been something else.]

b) [1 point] Suppose physicists discovered a parallel Earth, that is the same as our planet except for the fact that these counties did not have a tobacco ban. What would you call this parallel universe?

Solution: Counterfactual.

c) [1 point] Suppose another country is considering a flavored tobacco ban. The country has 100 islands and randomly assigns 50 of them to undergo bans and the other 50 to not undergo bans. After two years, the researchers look at tobacco addiction on all the islands. What design does this remind you of?

Solution: Randomized controlled trial or a controlled experiment.

d) [1 point] If these island researchers find a statistically significant difference, and conclude that the ban reduces tobacco addition by x %, is it safe to assume that the same effect would be found in the Bay Area? Why or why not?

Solution: Not safe to make these assumption because these people can be very different. The result may not be externally valid/representative. [Students can say any of these things.]

Relevant R Functions

```
pnorm(q = , mean = , sd =  , lower.tail = )
qnorm(p = , mean = , sd = , lower.tail = )
rnorm(n = , mean = , sd = )
pbinom(q = , size = , prob = , lower.tail = )
qbinom(p = , size = , prob = , lower.tail = )
dbinom(x = , size = , prob = )
rbinom(n = , size = , prob = )
ppois(q = , lambda = , lower.tail = )
qpois(p = , lambda = , lower.tail = )
dpois(x = , lambda = )
rpois(n = , lambda = )
```