Lecture 10: Continuous Distributions

For each question below, write the R code you would use and record the result when appropriate. Answer in complete sentences when asked to explain.

Part 1: Discrete vs Continuous Probability

1. In your own words, explain the difference between a discrete random variable and a continuous random variable.

2. For each situation below, identify whether the random variable is discrete or continuous.

   (a) The number of students absent from class.

   (b) The height of a randomly selected student.

   (c) The amount of time until the next phone call arrives.

   (d) The number of Heads in 20 coin flips.

   (e) The distance a penny lands from the center of a table.

3. Explain why $P(X = 15)$ is usually not useful for a continuous random variable, but $P(X < 15)$ or $P(10 < X < 15)$ can be useful.

Part 2: Penny Tosses

4. You decide to run an experiment where people toss a penny on a table and you mark down the distance it is from the center. You find that every possible distance between 10 and 33.5 inches is equally likely.

   (a) What kind of distribution should we use?

   (b) Why is this a continuous distribution instead of a discrete distribution?

   (c) What are the minimum and maximum values?

5. Generate a random sample of 100 observations and visualize it.

```
penny_100 <- runif(100, min=10, max=33.5)
hist(penny_100, main="Penny Toss Distances", xlab="Distance")
```

   (a) Describe the shape of the histogram. Does it look perfectly flat? Explain why or why not.

   (b) Calculate the $P(X < 15)$ and interpret it in context.

   ```
   sum(penny_100 < 15)/100
   ```

   (c) Calculate the $P(17 \leq X \leq 23.7)$ and interpret it in context.

   ```
   mean(penny_100 >= 17 & penny_100 <= 23.7)
   ```

6. Repeat the simulation with 2000 observations.

```
penny_2000 <- runif(2000, min=10, max=33.5)
hist(penny_2000, main="Penny Toss Distances", xlab="Distance")

mean(penny_2000 < 15)
sum(penny_2000 >= 17 & penny_2000 <= 23.7)/2000
```

   (a) How do the values change when the sample size increases to 2000?

   (b) How does the histogram change when the sample size increases to 2000?

Part 3: Amusement Park Arrivals and the Gamma Distribution

7. You are the Head of Marketing at an amusement park ride and would like to give the 200^{th} person a lifetime pass. People enter the park in the morning at a rate of about 12 people per minute. You want to determine the length of time it takes for 200 people to enter.

   (a) What does the random variable represent in this situation?

   (b) What are the shape and rate values for the Gamma distribution?

   (c) Why does a Gamma distribution make sense here?

8. Generate a random sample of 100 observations and visualize it.

```
arrival_100 <- rgamma(100, shape=200, rate=12)
hist(arrival_100, main="Time Until 200th Person", xlab="Minutes")
```

   (a) Describe the shape of the histogram.

   (b) What does one value in `arrival_100` represent?

9. Estimate the following probabilities using your sample of 100.

   (a) Calculate the $P(X < 16 \text{ minutes})$ and interpret it in context.

   (b) Calculate the $P(12 \text{ minutes} \leq X \leq 18 \text{ minutes})$ and interpret it in context.

10. Repeat the simulation with 2000 observations.

```
arrival_2000 <- rgamma(2000, shape=200, rate=12)
hist(arrival_2000, main="Time Until 200th Person", xlab="Minutes")
```

   (a) Record the new estimates.

   (b) How do the values change when the sample size increases to 2000?

(c) Does the histogram look smoother with 2000 observations? Explain briefly.

Part 4: Component Failure and the Exponential Distribution

11. You work in Quality Control at a large electronics manufacturer and are tasked with investigating the failure rate for a component. An electronic component fails roughly every 10 days. So, the mean time to failure follows an Exponential distribution with rate $\lambda = \frac{1}{10}$.

    (a) What does the random variable represent in this situation?

    (b) What is the rate value?

    (c) Why does an Exponential distribution make sense here?

12. Generate a random sample of 100 observations and visualize it.

```
failure_100 <- rexp(100, rate=1/10)
hist(failure_100, main="Time Until Component Failure", xlab="Days")
```

    (a) Describe the shape of the histogram.

    (b) What does one value in `failure_100` represent?

13. Estimate the following probabilities using your sample of 100.

    (a) Calculate the $P(X < 10 \text{ days})$ and interpret it in context.

    (b) Calculate the $P(3 \text{ days} \leq X \leq 7 \text{ days})$ and interpret it in context.

    (c) Calculate the $P(X > 30 \text{ days})$ and interpret it in context.

14. Repeat the simulation with 2000 observations.

```
failure_2000 <- rexp(2000, rate=1/10)
hist(failure_2000, main="Time Until Component Failure", xlab="Days")
```

    (a) How do the values change when the sample size increases to 2000?

    (b) How does the histogram change when the sample size increases to 2000?

Part 5: Exam Scores and the Normal Distribution

15. You are in charge of the national Data Science exam and would like to visualize the data. Assume the scores follow a bell-shaped distribution with a mean of 77 and a standard deviation of 9.

    (a) What does the mean tell us in this situation?

(b) What does the standard deviation tell us in this situation?

(c) Why might a Normal distribution be reasonable for exam scores?

16. Generate a random sample of 100 exam scores and visualize it.

```
exam_100 <- rnorm(100, mean=77, sd=9)
hist(exam_100, main="Data Science Exam Scores", xlab="Score")

mean(exam_100)
sd(exam_100)
```

(a) Record the sample mean.

(b) Record the sample standard deviation.

(c) Does the histogram look roughly bell-shaped? Explain briefly.

17. Estimate the following probabilities using your sample of 100.

(a) Calculate the $P(X < 60)$ and interpret it in context.

(b) Calculate the $P(75 < X < 79)$ and interpret it in context.

(c) Calculate the $P(X > 90)$ and interpret it in context.

18. Repeat the simulation with 2000 exam scores.

```
exam_2000 <- rnorm(2000, mean=77, sd=9)
hist(exam_2000, main="Data Science Exam Scores", xlab="Score")
```

(a) How do the values change when the sample size increases to 2000?

(b) How does the histogram change when the sample size increases to 2000?

Part 6: The 68-95-99.7 Rule

19. Using the sample of 2000 exam scores from Part 5, estimate the percentage of data within 1, 2, and 3 standard deviations of the mean.

```
m <- mean(exam_2000)
s <- sd(exam_2000)

mean(exam_2000 > m - s & exam_2000 < m + s)
mean(exam_2000 > m - 2*s & exam_2000 < m + 2*s)
mean(exam_2000 > m - 3*s & exam_2000 < m + 3*s)
```

(a) Record the percentage within 1, 2, and 3 standard deviations of the mean.

(b) Are these values close to 68%, 95%, and 99.7%? Explain briefly.