MATH 2310 Lab 2 - Discrete Distributions

Jakob Balkovec

2024-04-10

Activity One

In the general population, about 10% of people are left-handed.

[A] - Suppose I randomly pick 200 people. What is the chance I will see fewer than 15 people who are left-handed?

```
# Sample size: 200 people
# Probability: 0.1 (10%)
# Number of lefties: 15

# >>> Using Binomial Distribution

# Define our observations as separate variables
sample_size <- 200
probability <- 0.1
num_of_lefties <- 15

# Calculate new probability
p_less_than_15 <- pbinom(num_of_lefties, size=sample_size, prob=probability)

# Round the result so it can be printed in a neat way
rounded_probability <- as.character(round(p_less_than_15, digits = 4))

# Answer
print(paste("Answer [A]:", rounded_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [A]: 0.1431"
```

`"Answer [A]: 0.1431"`

[B] - Suppose I randomly pick 300 people. What is the chance I will see at least 40 people who are left-handed?

```
# Sample size: 300 people
# Probability: 0.1 (10%)
# Number of lefties: 40

# >>> Using Binomial Distribution

# Re-Define our observations as separate variables
sample_size <- 300
num_of_lefties <- 40

# Calculate new probability
p_at_least_40 <- 1 - pbinom(num_of_lefties - 1, size = sample_size, prob = probability)

# Round the result so it can be printed in a neat way
rounded_probability <- as.character(round(p_at_least_40, digits = 4))

# Answer
print(paste("Answer [B]:", rounded_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [B]: 0.0378"
```

`Answer [B]: 0.0378"`

[C] - The Seattle Mariners have 21 pitchers. Of those 21 pitchers, 4 are left-handed. If we assume the probability of a pitcher being left-handed is the same as the probability of any randomly selected person from the general population being left-handed, what would be the probability of seeing at least 4 left-handed pitchers out of 21?

```
# Go M's!!!

# Sample size: 21 pitchers
# Probability: 0.1 (10%) as per text
# Adjusted probability: [4/21 -> 0.19 (19%)]
# Number of lefties: 4

# >>> Using Binomial Distribution

# Re-Define our observations as separate variables
sample_size <- 21
num_of_lefties <- 4

# Calculate new probability
p_at_least_4 <- 1 - pbinom(num_of_lefties - 1, size = sample_size, prob = probability) # (num_of_lefties / sample
_size)
p_at_least_4_adjusted <- 1 - pbinom(num_of_lefties - 1, size = sample_size, prob = (num_of_lefties / sample_siz
e))

# Round the result so it can be printed in a neat way
rounded_probability <- as.character(round(p_at_least_4, digits = 4))
rounded_adjusted_probability <- as.character(round(p_at_least_4_adjusted, digits = 4))

# Answer
print(paste("Answer [C]:", rounded_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [C]: 0.152"
```

```
print(paste("Answer [D]:", rounded_adjusted_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [D]: 0.5874"
```

`"Answer [C]: 0.152"`
`"Answer [D]: 0.5874"`

[D] - Based on your answer from part **[C]**, does our assumption about the probability of a pitcher being left-handed seem reasonable? If not, how does the probability differ from the probability for the general population?

Answer: The assumption about the probability of a pitcher being left-handed seems reasonable to some extent. In part [C.a], the probability of seeing at least 4 left-handed pitchers out of 21, assuming the general population's left-handedness rate, was approximately 15%. However, considering the Seattle Mariners' pitchers as part of the general population, this assumption holds.

To provide a more nuanced analysis, I also calculated the probability using the occurrence rate of left-handed pitchers within the Seattle Mariners' team. I adjusted the probability, assuming an occurrence rate of $\frac{4}{21} \rightarrow$ 0.19 or 19%, and I found it to be approximately 58%. This suggests a higher prevalence of left-handedness among Seattle Mariners' pitchers compared to the general population.

Thus, while the assumption about the probability of a pitcher being left-handed holds when considering the Seattle Mariners' pitchers as part of the general population, but a closer examination revealed a divergence from the general population's left-handedness rate within the specific context of the Seattle Mariners'.

[E] - The Seattle Mariners have 3 catchers. All 3 of them are right-handed. If we assume the probability of a catcher being left-handed is the same as the probability of any randomly selected person from the general population being left-handed, what would be the probability of seeing no left-handed catchers out of 3?

```
# Go M's!!!

# Sample size: 3 catchers
# Probability: 0.1 (10%)
# Number of lefties: 0

# >>> Using Binomial Distribution

# Re-Define our observations as separate variables
sample_size <- 3
num_of_lefties <- 0

# Calculate new probability
p_no_lefties <- 1 - dbinom(num_of_lefties, size = sample_size, prob = probability)
p_no_lefties_adjusted <- 1 - pbinom(num_of_lefties, size = sample_size, prob = 0)

# Round the result so it can be printed in a neat way
rounded_probability <- as.character(round(p_no_lefties, digits = 4))
rounded_probability_adjusted <- as.character(round(p_no_lefties_adjusted, digits = 4))

# Answer
print(paste("Answer [E]:", rounded_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [E]: 0.271"
```

```
print(paste("Answer [F]:", rounded_probability_adjusted, sep = " ", collapse = NULL))
```

```
## [1] "Answer [F]: 0"
```

`"Answer [E]: 0.271"`
`"Answer [F]: 0"`

[F] - Based on your answer from part **[E]**, does our assumption about the probability of a catcher being left-handed seem reasonable? If not, how does the probability differ from the probability for the general population?

Answer: It doesn't seem reasonable, given that all 3 catchers are right-handed. With no left-handed catchers observed in the sample, the probability of a left-handed catcher should logically be 0%. Therefore, the assumption of the probability of a catcher being left-handed being the same as the general population's left-handedness rate is not supported by the observed data.

[G.a] - Suppose we were to look at a larger sample of catchers. If every catcher we sampled were right-handed, how many catchers would we need to sample before you would conclude that the true probability of a catcher being left-handed is less than 10%?

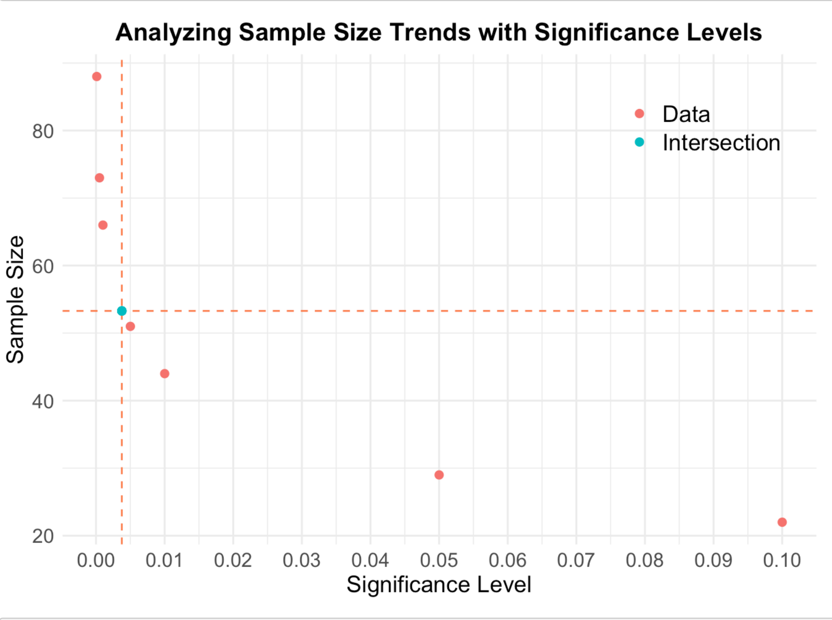
```
# Hypothesis ->
# [Null] H0: The true probability of a cacther being a left-handed is 0.1 (10%)
# [Alternative] H1: The true probability of a catcher being a left-handed is less that 0.1 (x < 10%)

# Probability of a catcher being a left-handed : 0.1 (10%)
# Significance Level: [0.0001 | 0.01 | 0.05 | 0.1, ...]

# Define our observations as separate variables
probability <- 0.1
sample_size <- 0 # Initial Value


# brief: Calculates the sample size needed for a binomial test given a significance level.
#
# param:
#   - significance_level: The desired significance level for the binomial test.
#
# returns:
#   The calculated sample size needed to achieve the specified significance level.
calculate_sample_size <- function(significance_level) {
  while (pbinom(0, size = sample_size, prob = probability) >= significance_level)
  { sample_size <- sample_size + 1 }
  return (sample_size)
}

# brief: Calculates the significance level for a binomial test given a sample size.
#
# param:
#   - sample_size: The size of the sample used in the binomial test.
#
# retunrs:
#   The significance level achieved by the given sample size.
calculate_significance_level_from_sample_size <- function(sample_size) {
  significance_level <- pbinom(0, size = sample_size, prob = probability)
  return(significance_level)
}

# Define the significance levels as a vector
significance_levels <- c(0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001)

sample_sizes <- list()

# Calculate sample size for each significance level in the vector
for (level in significance_levels) {

  # Construct variable name
  variable_name <- paste("sample_size_", level, sep = "")
  sample_sizes[[variable_name]] <- calculate_sample_size(level)
}
```

[G.b] - Plot

This plot visualizes the relationship between significance levels and sample sizes. The orange points represent the actual data points. The dashed coral lines mark the mean sample size and the intersection point between the trend line and the mean sample size. Overall, the plot provides a clear depiction of how sample size varies with different significance levels.

```
library(ggplot2)

# Combine data/create a data frame
sample_size_values <- unlist(sample_sizes)
data <- data.frame(significance_levels, sample_size_values)

# Get mean for the x axis intercept
mean_sample_size <- mean(sample_size_values)
x_intercept <- calculate_significance_level_from_sample_size(round(mean_sample_size))

# Plot the sample sizes wrt significance levels
ggplot(data, aes(x = significance_levels, y = sample_size_values)) +
  geom_point(aes(color = "Data"), shape = 19, size = 2) +

  # Probably not a Linear Model >>> probably LOESS or SPLINE
  geom_smooth(method = "spline") +
  geom_hline(yintercept = mean_sample_size, linetype = "dashed", color = "coral") +
  geom_vline(xintercept = x_intercept, linetype = "dashed", color = "coral") +
  geom_point(aes(x = x_intercept, y = mean_sample_size, color = "Intersection"), shape = 19, size = 2) +

  labs(x = "Significance Level", y = "Sample Size", title = "Analyzing Sample Size Trends with Significance Level
s",
       color = "Legend") +

  scale_x_continuous(breaks = seq(0, 0.1, by = 0.01)) +

  theme_minimal() +  theme(
        plot.title = element_text(size = 15, face = "bold", hjust = 0.5),
        axis.title = element_text(size = 14, hjust = 0.5),
        axis.text = element_text(size = 12),  # Decrease font size for axis text
        legend.title = element_blank(),
        legend.text = element_text(size = 14),
        legend.position = c(0.85, 0.85))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# ...Questionable...
# Answer (!Assuming 0,05 significance level!)
sig_lvl <- 0.05

print(paste("Answer [F]:", round(calculate_sample_size(sig_lvl)), sep = " ", collapse = NULL))
```

```
## [1] "Answer [F]: 29"
```

`"Answer [F]: 29"`

Activity Two

Suppose you are designing a computer server for students to log into to work remotely. You know that on average you will see ten students logging into the server per hour.

[H] - What is the chance that more than 15 students will log into the server in a particular hour?

```
# Lambda: 10 students per hour
# Sample size: 15 students

# Calculate probability that we see more than 15 students log into the server
# P(X > 15) = 1 - P(X <= 15)

# Using >>> Poisson distribution

lambda <- 10
sample_size <- 15
p_more_than_15 <- ppois(sample_size, lower.tail = FALSE, lambda)

# Round the result so it can be printed in a neat way
rounded_probability <- as.character(round(p_more_than_15, digits = 4))

# Answer
print(paste("Answer [H]:", rounded_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [H]: 0.0487"
```

`"Answer [H]: 0.0487"`

[I] - What is the chance of seeing exactly 10 students log into the server in a particular hour?

```
# Lambda: 10 students per hour
# Sample size: 10 students

# Calculate probability that we see exactly 10 students log into the server
# P(X = 10 )

# Using >>> Poisson distribution

lambda <- 10
sample_size <- 10
p_exactly_10 <- dpois(sample_size, lambda)

# Round the result so it can be printed in a neat way
rounded_probability <- as.character(round(p_exactly_10, digits = 4))

# Answer
print(paste("Answer [H]:", rounded_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [H]: 0.1251"
```

`"Answer [H]: 0.1251"`

[J] - What is the chance of fewer than 15 students logging into the server in a two-hour period?

```
# Lambda: 10 students per hour
# Sample size: 15 students

# Calculate probability that we see more than 15 students log into the server
# P(X < 15)

# Using >>> Poisson distribution

lambda <- 10
sample_size <- 14
p_less_than_15 <- ppois(sample_size, lower.tail = TRUE, lambda)

# Round the result so it can be printed in a neat way
rounded_probability <- as.character(round(p_less_than_15, digits = 4))

# Answer
print(paste("Answer [H]:", rounded_probability, sep = " ", collapse = NULL))
```

```
## [1] "Answer [H]: 0.9165"
```

`"Answer [H]: 0.9165"`

[K.a] - In designing the server, you must decide the maximum number of students that it can accommodate at one time. The more students you allow it to accommodate, the more expensive it will be. But if more students attempt to log in during a single hour than it can accommodate, it will crash. How many students should you design it to accommodate if you want there to be at most a 1% chance that it will crash during any particular hour?

```
# Lambda: 10 students per hour
# Initial Threshold: 0

# Using >>> Poisson distribution

lambda <- 10
threshold <- 0

while(ppois(threshold, lambda, lower.tail = FALSE) >= 0.01) {
  threshold <- threshold + 1
}

# Answer
print(paste("Answer [K]:", threshold))
```

```
## [1] "Answer [K]: 18"
```

`"Answer [K]: 18"`

[K.b] - Plot

The plot visualizes the cumulative probability distribution of a server crashing due to an overload of student log-ins per hour. With an average of ten students logging in per hour, the plot demonstrates the increasing likelihood of a crash as the maximum number of students the server can accommodate rises. The dashed coral line represents the 0.01 (1%) crash probability threshold, indicating the maximum number of students the server should accommodate to maintain a low risk of crashing. The brown point marks the intersection where the cumulative probability drops below 1%, highlighting the optimal threshold for server capacity at **18** students.

```
# library(ggplot2) already imported above

# Observations as variables
lambda <- 10
thresholds <- seq(0, 25, by = 1)

# Calculate the cumulative probabilities
cumulative_probabilities <- ppois(thresholds, lambda, lower.tail = FALSE)

# Note: 18 for the plot, 17 because we have to round down
data <- data.frame(threshold = thresholds, cumulative_prob = cumulative_probabilities)
intersection_threshold <- max(data$threshold[data$cumulative_prob > 0.01]) + 1
intersection_data <- data.frame(threshold = intersection_threshold,
                                cumulative_prob = 0.01)

# Plot the cumulative probability distribution
ggplot(data, aes(x = threshold, y = cumulative_prob)) +
  geom_line(color = "dodgerblue") +
  geom_hline(yintercept = 0.01, linetype = "dashed", color = "coral") +
  geom_vline(xintercept = intersection_threshold, linetype = "dashed", color = "coral") +
  geom_point(data = intersection_data, aes(x = threshold, y = cumulative_prob),
             color = "brown", shape = 19, size = 2) +
  labs(x = "Threshold (Number of Students)", y = "Cumulative Probability",
       title = "Cumulative Probability Distribution of Server Crash") +
  theme_minimal() +
  theme(plot.title = element_text(size = 15, face = "bold", hjust = 0.5),
        axis.title = element_text(size = 14, hjust = 0.5),
        axis.text = element_text(size = 12)) +
  scale_x_continuous(breaks = seq(0, 25, by = 5)) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))
```