1 Introduction

In the homework for week 6, you explored the beta distribution. The beta distribution is a continuous distribution that is used to model a random variable X that ranges from 0 to 1, making it useful for modeling proportions, probabilities, or rates. The beta distribution is also known for being remarkably flexible with regards to its shape – it can be left-skewed, right-skewed, or symmetric depending on the value of the parameters that define its shape: $\alpha > 0$ and $\beta > 0$.

2 Task One: Describe The Population Distribution

The beta distribution's probability density function is given by

$$f_X(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma\alpha\Gamma\beta} x^{\alpha-1}(1-x)^{\beta-1} I(x \in [0,1]),$$

where $I(x \in [0,1]) = 1$ when $x \in [0,1]$ and 0 otherwise. Note that this simply makes the probability density function zero where x cannot possibly take on those values, i.e., everywhere outside of [0,1].

Because the distribution's shape is defined by its parameters, the population-level characteristics are also described by those parameters.

$$E(X) = \frac{\alpha}{\alpha + \beta} \qquad \text{(The Mean)}$$

$$var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{(The Variance)}$$

$$skew(X) = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}} \qquad \text{(The Skewness)}$$

$$kurt(X) = \frac{6\left[(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)\right]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} \qquad \text{(The Excess Kurtosis)}$$

Note: The kurtosis formula was updated after we realized the one on the homework was incorrect!

Consider four cases:

- Beta($\alpha = 2$, $\beta = 5$)
- Beta($\alpha = 5$, $\beta = 5$)
- Beta($\alpha = 5$, $\beta = 2$)
- Beta($\alpha = 0.50$, $\beta = 0.50$)

Plot these four distributions. Compute the mean, variance, skewness, and kurtosis for all four cases. Provide these values in a table in your write-up and reference them when describing the figure.
Hint: To be more code efficient, you can write a function to complete these tasks that takes α and β as arguments. Otherwise, you can copy-paste-edit.

3 Task Two: Compute the moments

The mean, variance, skewness and kurtosis are computing using the moments of a distribution. The kth uncentered moment of a continuous distribution is

$$E(X^k) = \int_{\mathcal{X}} x^k f_X(x) dx,$$

that is, a weighted continuous average of X^k using the probability density functions to compute the weights. The kth centered moment of a continuous distribution is

$$E\left[(X-\mu_X)^k\right] = \int_{\mathcal{X}} (x-\mu_X)^k f_X(x)dx,$$

where $\mu_X = E(X)$ that is, a weighted continuous average of $(X-\mu_X)^k$ using the probability density functions to compute the weights.

We can write the population-level characteristics using these moments.

$$\mu_X = E(X) \qquad \text{(The Mean)}$$

$$\sigma_X^2 = var(X) = E\left[(X-\mu_X)^2\right] \qquad \text{(The Variance)}$$

$$skew(X) = \frac{E\left[(X-\mu_X)^3\right]}{E\left[(X-\mu_X)^2\right]^{3/2}} \qquad \text{(The Skewness)}$$

$$kurt(X) = \frac{E\left[(X-\mu_X)^4\right]}{E\left[(X-\mu_X)^2\right]^2} - 3 \qquad \text{(The Excess Kurtosis)}$$

Write a function called `beta.moment()` that takes `alpha`, `beta`, `k`, and a logical argument `centered` as arguments. Use the `integrate()` function to compute and return the centered (when `centered=T`) and uncentered (when `centered=F`) moments. Confirm your function works by computing the population-level characteristics using the formulas above and comparing them to the values you obtained in Task One.
Hint: For the centered moments, you'll need to compute $\mu_X = E(X)$ and then compute $E\left[(X-\mu_X)^k\right]$

4 Task Three: Do Data Summaries Help?

When summarizing data, our goal is to approximate what the population distribution might be. In this lab, we will see how our graphical and numerical summaries connect to the actual population distribution. We can do this because we can randomly generate data from a known distribution, perform the same summary tasks, and compare the results to the known distribution.

Generate a sample of $n = 500$ from each distribution. Use `set.seed(7272)` to ensure we all work with the same samples. Plot a histogram for each sample, include the estimated density using `geom_density()` and the true probability density function for each sample. Compute a numerical summary for these data using the `summarize()` function from `dplyr`, part of the `tidyverse` library (Wickham et al., 2023, 2019). Make sure to include the mean, variance, skewness, and excess kurtosis. Compare these sample values to the population-level quantities.

5 Task Four: Is Sample Size Important?

Compute the cumulative numerical summaries (mean, variance, skewness, and kurtosis) for the beta($\alpha = 2$, $\beta = 5$) data. Use the `cumstats` package (Erdely and Castillo, 2017) instead of building your own functions. Plot a 2×2 grid of plots using `geom_line()` to plot the cumulative statistics in four separate plots.In each, add a y-intercept line at the true values that describe the actual population distribution that you calculated in Task One.
Hint: You can either create the individual plot objects and combine using `patchwork` (Pedersen, 2024), or you can pivot your data so you can use `facet_wrap()`. Note that if you use the latter, you'll have to create a tibble that contains the y-intercepts for each statistic, and you may find the `scales="free"` argument to be helpful because the statistics are on rather different scales.

When you have that working for the original sample, write a `for()` loop to simulate new data ($n = 500$). Make sure to use `set.seed(7272+i)` to ensure we all work with the same samples. At each iteration (`2:50`),

add a line for the cumulative statistics calculated on the new data. Note that you can specify color with a number (e.g., `color=i`).
Note: I tell you to use `2:50` because `color=1` is black.
Hint: You can add a line to a ggplot object as follows.

```
plot.on.og.sample <- plot.on.og.sample +
  geom_line(data=new.data, aes(x=x, y=y), color=i)
```

6 Task Five: How can we model the variation?

Write a `for()` loop to simulate new data ($n = 500$) from the beta($\alpha = 2$, $\beta = 5$) distribution and compute the mean, variance, skewness, and excess kurtosis. Make sure to use `set.seed(7272+i)` to ensure we all work with the same samples. At each iteration (`1:1000`), compute and store the statistics of interest. The result is a sample of $n = 1000$ means, variances, skewnesses, and excess kurtosises. Just as we summarize data, we can summarize statistics to see their distribution – we will call these sampling distributions after break.

Plot a histogram for each statistic, and include the estimated density using `geom_density()`. What do you notice about the distributions?

7 Optional Coding Challeneges

The `gganimate` package for `R` (Pedersen and Robinson, 2024) enables us to create animations using `ggplot2` (Wickham, 2016).

1. Create a plot that shows how the beta distribution as you increase α from 1 to 10, where $\beta = 5$
2. Create a plot that shows how the beta distribution as you increase β from 1 to 10, where $\alpha = 5$.
3. Consider the $beta(\alpha = 1, \beta = 1)$ distribution. This special case is equivalent to another distribution we discussed.
4. Create plots that show how the mean, variance, skewness and kurtosis change with α and β. Use `geom_raster()` to plot α on the x-axis, β on the y-axis, and let each statistic determine the `fill` color.
5. Create a plot showing how the sample mean histogram changes with sample size. That is, redo task four for sample sizes on `seq(1,500,10)`.

References

Erdely, A. and Castillo, I. (2017). *cumstats: Cumulative Descriptive Statistics*. R package version 1.0.

Pedersen, T. L. (2024). *patchwork: The Composer of Plots*. R package version 1.2.0.

Pedersen, T. L. and Robinson, D. (2024). *gganimate: A Grammar of Animated Graphics*. R package version 1.0.9.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.