STAT165/265 HW 8

March 6, 2024

Due Tuesday, March 12, 2024 at 11:59pm

Deliberate Practice: Invalidating Considerations

Expected completion time: 120 minutes
Graded on accuracy
This exercise follows the activity in last week's discussion. We will use the following simplified model for invalidating considerations:

- Suppose you want an 80% confidence interval around some quantity
- You have an initial distribution p_0 for what happens in a "normal" world. For example, maybe this is the distribution that you get by looking at a reference class.
- After brainstorming invalidating considerations, you realize that there is a small probability ϵ that you are instead in a "crazy" world, in which case your distribution is instead p_1.
- Your new probability distribution is therefore the mixture $(1-\epsilon)p_0 + \epsilon p_1$.

For each of the following examples of p_0, p_1, and ϵ, do the following:

- Give an 80% confidence interval when accounting for the ϵ probability of a "crazy" world. Your interval should be centered, i.e. your lower and upper bounds should be the 10th and 90th percentiles of the mixture distribution.
- Confirm your reasoning by simulation: using Python, draw 1000 samples from $(1-\epsilon)p_0 + \epsilon p_1$, and use them to estimate the 10th and 90th percentiles of this distribution. Sampling from a mixture happens in two steps:
  1. Sample from a Bernoulli(ϵ) to choose between p_0 and p_1.
  2. Sample from the chosen distribution (in this exercise the distributions will be uniform).

1. Non-overlapping uniform distributions
    - $p_0 = \text{Uniform}(0, 2)$
    - $p_1 = \text{Uniform}(2.5, 3)$
    - (a) $\epsilon = 0.05$, (b) $\epsilon = 0.1$, (c) $\epsilon = 0.2$

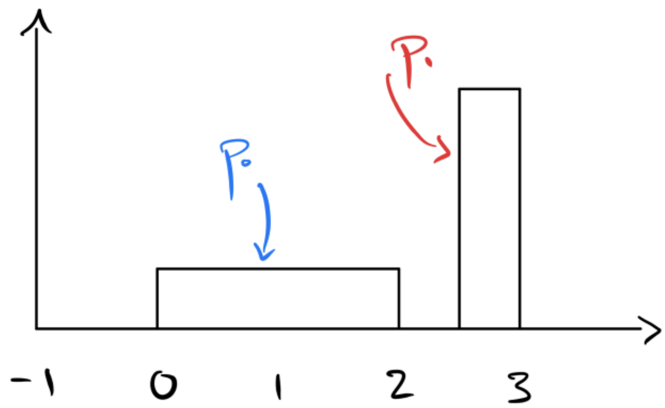

Figure 1: Non-overlapping uniform distributions

2. Overlapping uniform distributions
    - $p_0 = \text{Uniform}(0, 2)$
    - $p_1 = \text{Uniform}(1, 3)$
    - (a) $\epsilon = 0.05$, (b) $\epsilon = 0.1$, (c) $\epsilon = 0.2$

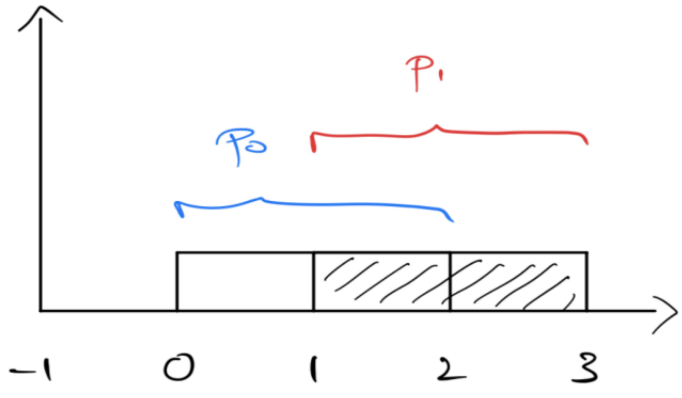

Figure 2: Overlapping uniform distributions

On Gradescope, please also submit the time it took to complete this exercise. Please note that we have a separate assignment set up for this, which is worth 1 point for this question.

Deliberate Practice: Numerical Sensitivity

Expected completion time: 180 minutes
Graded on accuracy

Consider the formula for the number of days T until the peak of Omicron that we used in the *Turning Considerations into Probabilities* lecture:

$$T = \log_2 (N/2N_0) \cdot t + \Delta_0 + \Delta_1,$$

where:

- N is the total number of future UK Omicron cases
- N_0 is the current number of UK Omicron cases
- t is the Omicron doubling time
- Δ_0 is the lag between case peak and hospital peak
- Δ_1 is the lag between single-day hospital peak and 7-day average hospital peak

Using simulations, we will assess the sensitivity of this formula to variations of its five inputs.

1. Suppose the inputs are sampled independently from Normal and LogNormal distributions. For this, sample α, β, γ, and δ independently from Normal$(0, 1)$, and define the inputs N, N_0, t, and Δ as follows:

   - $N = \exp(15.57 + 0.30 \cdot \alpha)$
   - This means that N follows a LogNormal distribution with mean 6.7×10^6 and standard deviation 4×10^6.
   - $N_0 = \exp(12.18 + 0.06 \cdot \beta)$
   - This means that N_0 follows a LogNormal distribution with mean 0.2×10^6 and standard deviation 0.05×10^6.
   - $t = 2.4 + 0.5 \cdot \gamma$
   - $\Delta = \Delta_0 + \Delta_1 = 12 + 3 \cdot \delta$.

   Sample the inputs in this way 1000 times, computing each time the corresponding value of T using the formula above. Plot the histogram of the distribution of T: what are its 10th, 25th, 50th, 75th, and 90th percentiles?

2. Increase the standard deviation of N while leaving the other distributions constant: for this, sample α from Normal$(0, 4)$ rather than Normal$(0, 1)$. Do the same but for Δ instead: double the standard deviation of δ, while leaving the other distributions constant. Which of these leads to the greatest increase in the variance of T?

3. Replace the Normal distributions by Student's t distributions: instead of sampling α, β, γ and δ from Normal$(0, 1)$, sample them from Student$(\nu = 5)$. Plot the histogram of the distribution of T: how did using a t distribution instead of a Normal change the quantiles and variance of T?

4. Repeat your simulation with $\nu = 3, \nu = 7$ and $\nu = 9$: how does the degrees of freedom parameter ν affect the distance between the 10th and 90th percentiles of T?

5. Suppose now that instead of being independent, the inputs are correlated. That is, sample $[\alpha, \beta, \gamma, \delta]^\top$ from a multivariate Gaussian distribution:

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} \sim \text{Normal}(\mathbf{0}, \mathbf{\Sigma}), \text{ with } \mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}, \text{ and } \rho \in \left(-\frac{1}{3}, \frac{1}{3}\right).$$

   Draw 1000 samples, and plot a histogram of the values of T. Do the 10th and 90th quantiles get closer together or farther apart as the correlation parameter ρ increases from 0 to $\frac{1}{3}$? as ρ decreases from 0 to $-\frac{1}{3}$?

On Gradescope, please also submit the time it took to complete this exercise. Please note that we have a separate assignment set up for this, which is worth 1 point for this question.

Predictions

Expected completion time: 90 minutes
Graded on accuracy as part of the class forecasting competition

Make and submit predictions to the questions on this Google Form:
https://forms.gle/vcgtDXYaN8Tv8c6E7.

Be sure to follow the format described at the top of the form. For each question, you will submit a mean and inclusive 80% confidence interval or a probability (whichever the question asks for). We provide cells on the Google form for you to type out your reasoning (1-2 paragraphs), which you should submit to Gradescope with the rest of this assignment. For questions 1-3, your prediction (but not the explanation) will appear on the public leaderboard.

[STAT 265 only] Sums of correlated variables

Expected completion time: 60 min
Graded on accuracy

Consider random variables $X_1, ..., X_k$ each with standard deviation σ. Let X^* denote their sum. That is,

$$X^* = X_1 + ... + X_k$$

In this problem we will explore how different levels of correlation between each X_i change the standard deviation of X^*, which we denote as σ^*.

Let

$$\vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Suppose each X_i has correlation ρ with all others, so we have the following correlation matrix:

$$\text{Corr}(\vec{X}) = \begin{bmatrix} 1 & \rho & \rho & ... & \rho \\ \rho & 1 & \rho & ... & \rho \\ \rho & \rho & 1 & ... & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & ... & 1 \end{bmatrix}$$

1. Derive a formula for σ^* in terms of k, σ, ρ.
   You may use the property $\text{Var}(\vec{v}^\top X) = \vec{v}^\top \text{Cov}(X)\vec{v}$ without proof.
   Hint: You can express X^* as $\mathbf{1}\vec{X}$ where $\mathbf{1}$ is a vector of ones.

2. For each correlation ρ, calculate the value of σ^* in terms of k, σ.

   a. $\rho = 0$ (the X_i are uncorrelated)
   b. $\rho = 1$ (the X_i are perfectly positively correlated)
   c. $\rho = \frac{1}{2}$ (the X_i are halfway to being perfectly positively correlated)
   d. $\rho = \frac{-1}{k-1}$ (the X_i are as negatively correlated as possible)
   e. $\rho = \frac{-1}{2(k-1)}$ (the X_i are halfway to being as negatively correlated as possible)

3. Reflect on what happened to σ^* for different values of ρ. When is σ^* the largest? Smallest? How does σ^* scale as a function of k in each of the cases above?

4. In this next part, we will explore a creatively modified and simplified version of what led to the 2008 Financial Crisis.

   Suppose each X_i represents a home mortgage and X^* is a financial instrument whose value is the sum of all the individual mortgage values. Suppose $X_i = 0.9Y_i + 0.1Y_0$ where $Y_1, \ldots, Y_n$ are independent (for $i \neq j, Y_i \perp\!\!\!\perp Y_j$), but Y_0 is the same for all of them. In other words, Y_i is different for each home mortgage, but Y_0 is a global, variable, say, one that depends on the national interest rate. Let $k = 10000$, and assume for simplicity that $\text{Var}(Y_i) = 1$.

a. Compute $\text{Corr}(\vec{X})$.
b. Compute the standard deviation of X^*.
c. Compute the standard deviation if people had failed to model the Y_0 component (i.e. let $X_i = Y_i$).
d. What do you observe about the difference in standard deviations (1-2 sentences)?