STAT 165/265 HW 6

Feb 21, 2024

Due Tuesday, February 27 at 11:59pm

Deliberate Practice: Common Probability Distributions

Expected completion time: 50 minutes
See note below for how this question will be graded.

What distribution would you expect the following quantities to follow: normal, log-normal, power law, or other? Write down a few sentences of considerations that you took into account and your final answer.

1. The size of raindrops when it rains (in a fixed location).
2. The citation count of different research papers in a single research field (e.g. physics or computer science).
3. Standardized test scores (e.g. for the SAT or GRE) in the United States.
4. The amount of time (in minutes) spent playing a game of chess.
5. The estimated cost effectiveness (in dollars per life saved) of different global health charities/interventions.

Note: we will primarily grade you on the quality of your considerations (for example, you won't be penalized for saying log-normal when the answer is power law, as long as you provide a plausible justification for it being log-normal instead).

Lab

Expected completion time: 60 minutes
Graded on accuracy

Link to Jupyter notebook.

Please follow the instructions in the notebook to print out your code and answers and submit to Gradescope. You may use languages other than Python, although we will generally be providing starter code in Python.

On Gradescope, please also submit the time it took to complete this exercise.

Predictions

Expected completion time: 60 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/jkrYWx3o2xqrn3Bu5.

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] Combining Confidence Intervals

Expected completion time: 60 min
Graded on accuracy
Recommended reading: Chapter 10.3 of Prof. Steinhardt's *Forecasting*
Supplemental reading: Wikipedia article on mixture distributions

We will borrow the setup and notation from the textbook chapter linked above. Suppose we want to forecast a numerical outcome event. We have n forecasts $[a_1, b_1]$, ..., $[a_n, b_n]$ that we wish to combine. Think of each of these confidence intervals as a probability distribution p_i with mean μ_i and standard deviation σ_i such that a_i and b_i are lower and upper percentiles of p_i. We take the average, $\bar{p}$:

$$\bar{p} = \frac{p_1 + ... + p_n}{n}$$

where the sum denotes the *mixture* of distributions (see Wikipedia link above), from which we wish to find a confidence interval.

a. Compute $\bar{\mu}$ and $\bar{\sigma}$, the mean and standard deviation of $\bar{p}$.

b. Assume $\mu_i = \frac{a_i+b_i}{2}$, $\sigma_i = \frac{b_i-a_i}{2}$, and $\bar{p} \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$. Show that taking $[\bar{\mu} - \bar{\sigma}, \bar{\mu} + \bar{\sigma}]$ as our combined interval yields the LB* and UB* estimates from discussion 5 (copied below for convenience). You may also assume further that the means and variances are uncorrelated.

$$\text{LB}^* = \bar{\mu} - \sqrt{\frac{\sum_{i=1}^{n} (\bar{\mu} - a_i)^2}{n}}$$

$$\text{UB}^* = \bar{\mu} + \sqrt{\frac{\sum_{i=1}^{n} (b_i - \bar{\mu})^2}{n}}$$