

Please do these in Stata and R. Do not use built-in functions for any of these except the standard normal CDF.

1. Generate a single, well-labeled figure overlaying the normal, logistic, and cauchy PDFs.
2. Generate a single, well-labeled figure overlaying normal, logistic, and cauchy CDFs (see "comment" section below for guidance.)
3. Generate a figure overlaying Poisson PDFs at  $\lambda$  values of .5, 1.2, 3, and 5.  $\lambda$  is the mean or location parameter of the Poisson.

Data Please do these in Stata and R. Use the "beats.csv" data o. These data, scraped from Spotify's API, contain information about every Taylor Swift and Rolling Stones song in Spotify's database. You'll see variables indicating characteristics of songs like "danceability," "valence," etc. (definitions provided in the "comment" section below). The goal here is to analyze the "danceabilty" of Taylor Swift and Rolling Stones songs in two ways - are their songs more/less danceable over time, and are the two artists' songs different from one another?

1. write code to clean the data so it's ready for analysis. Annotate your code.
2. present a professional table summarizing key features of the data including means, medians, variances, other interesting percentiles, etc. Comment briefly on features you find interesting.
3. Does the danceability of Stones songs change over time? Estimate a linear regression predicting danceability over time - think carefully about how to do this. Present the model results in a professional table.
4. Interpret your estimates.
5. Do the same for Taylor Swift songs - does danceability change over time?
6. Present and comment on a box plot evaluating danceability by year for Taylor Swift.
7. Do the same for the Stones.
8. Generate a single figure comparing danceability for the two artists - what does it show?
9. Is danceability statistically different between the Stones and Taylor Swift? Explain. How did you evaluate this? Present your result.

## Simulation

You all have the sense that our estimates (of both  $\beta$  and se) change as sample size changes.

1. if you wanted to see exactly how N affects those estimates, what would you do? Write out, in English, specific steps you'd take to find this out.
2. Sketch out code in both Stata and R to do this - this code doesn't have to work (yet), but you should definitely develop a sense of process and some things Stata can do.

## Comments

In section 1, do not use built-in commands for any of these figures except the Normal CDF. Part of the goal here is for you to connect notional distributions to their mathematical statements to data. So look up the mathematical statements of these distributions. The Normal CDF is an integral, difficult to write - use built-in commands for it. For labeling in Stata, you'll find "scatteri" and "pcarrow" useful. In R, using "ggplot", you'll find "annotate" useful.

In section 2, be creative - I'm trying to get you to work through some things, and to see what you know, what you don't. In section 3, you'll be best served by thinking in English first, not starting in code. Simulation is a powerful tool and one we'll use a lot, so getting comfortable with thinking about how to structure such things is important. Here are descriptions of Spotify's song features:

- **Acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **Danceability:** Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **Energy:** Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
- **Instrumentalness:** Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal." The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- **Liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- **Speechiness:** Speechiness detects the presence of spoken words in a track. The more exclusively speechlike the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including

such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

- Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)