Probability Models

Review

Normality

- Normality is a powerful assumption
  Obtain probabilities from mean and SD…via the empirical rule.

- Not all data are normal, and most deviate in some way or another
  Discreteness in GMAT scores, skewness in exec compensation.

- Diagnose normality via quantile plots
  Normality implies an expected "spacing" among the ordered data, and quantile plots use this spacing to expose deviations from normality.

- Simulated normal data
  As we grow the size of the simulated data set, we can see the effect on the quantile plot (the bands narrow).

Examples

- GMAT and stocks returns, both normally distributed

- Log transformations and skewness
    change in "natural" log –> % change

- Grouping as a source of variation in both
    Executive Compensation and cookies counting

JMP Tricks

Using JMP-IN with other software

Cutting and pasting with MS Word.

Scripting (i.e., programming) JMP-IN

Key Application

Estimating the chance for a rare event

– "Hurricane" bonds
– Movement in financial markets (options)
– In both cases, it is hard to predict what will happen since the event is so rare. Data are often not so helpful, even with lots of data, and we'll need a model to help us, *but which model?*

Definitions

Random variable

– Symbol denoting the numerical outcome of a random experiment
– Used in *thought experiments*, especially in what-if scenarios
– Correspondence (table or function) assigning probabilities to outcomes.
– Coin-tossing example: do you really need to toss the coin?

Expected value, variance and SD of random variable

– Analogous to properties of data, but computed precisely using assumed probabilities that describe the behavior of the random events.
– Expected value of a random variable = population mean
– Avoid the need for explicit data.

Student's t model (distribution)

– Alternative distribution or probability models
– Student's t has fat tails (i.e., readily produces outliers)
    "DF" parameter controls "fatness" of tails
– The larger the DF term, the closer the t-distribution comes to normal.

Quantile plots

– How they work in the case of normal data
    How small would you expect the smallest in a sample from a normal population to be? Use the empirical rule.
– Generalize idea to other models than the normal.
– Identifying a model: Animate graphics using JMP-IN scripts

Concepts

Extrapolation

– Prediction, frequency of rare events

– Using a model to predict in situations where you have no data.

Diagnostic

– How good is that extrapolation?

– Without data, is there a way to check the sensitivity of your conclusions to the assumptions that you have made in your model?

Discussion

Subjectivity in statistics

– What's a large deviation from normality in the quantile plot?

– What question are you trying to answer?
  • Place/ranking among Wharton GMAT scores
  • Pricing an option

Pace of class

– Starting to get outside the range of ideas explored in many undergraduate statistics classes.

– We'll continue to pick up speed…

Examples for Today

Variation by industry in executive salary (page 45, review)

– What industries have the high-paid execs?

– Using boxplots for comparison

– Value of the log transformation
Can't see differences among industries without this transformation.
Once transformed, you can try to assess the role of industry as a
component of variation, as a source of variation.

– Use of cookies as time permits.

Pricing a digital option

– Cost for insuring against big move in financial markets

How much to insure that SP500 does not drop by 10% or more next
month? Interesting if we have a big need for the money coming.

– Gambling and the notion of expected value and variance
• Computing costs using expected value (coin-toss context)
• Distinction from sample values

– Probability models (like normal) give answers that data do not.
• 10% drop is 2.863 SDs below sample mean … pretty rare
• Normality suggests only happens with probability 0.0021.

– *Sensitivity* of our calculations to normality (we're way out in tails)
• Maybe the data is not normal… maybe it comes from t-distribution
• Quantile plots suggest t is a better model than the normal.
• t-distribution suggests probability is about 10 times higher.

– Bernstein, *Against the Gods*, fate of Long Term Cap. Management.