Stat100 Concepts

DJM

As of: 2020-04-22

Out of how many Numbers are magnitudes. It often helps to interpret if we convert these to ratios. For example, rather than knowing the number of murders in a city it may help to compare to the population of the city (murders per 100,000 residents, say).

Compare with rates, but be careful Related to "out of how many", it's often more useful to examine rates. But these can be counterintuitive if "how many" is relatively small.

Clumping/streaks Pure random process produce streaks or clumps. This isn't necessarily a sign of some thing real. Think a coin that lands heads 5 times in a row out of 10 or a roulette wheel that lands on "odds" 6 times.

Independence/Dependence It's easy to calculate the probabilities of independent events all occurring: just multiply. But be careful to make sure the events are independent!

Sunk costs Be wary of making decisions based on past decisions you can't change. Don't bet in poker based on past losses. Don't stay with your significant other just because you've been together for a while.

Decision making with uncertainty Try to determine the utility you'll earn for each possible outcome. How likely are those outcomes? Make the decision with a higher expected utility.

Regression to the mean Other things equal, performance below the average is likely to be followed by relatively higher performance. Performance above the average is likely to be followed by relatively lower performance.

Color scaling Be aware of the point you're trying to make. Is it apparent? Be aware of colorblindness (about 10% of men and 4% of women).

Expected value Mathematically, the sum of all potential values weighted by the probability of seeing those values. For example, the expected number of heads in 2 flips of a fair coin is

$$EV = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

Law of large numbers As you repeat a process over and over and over forever, the running average will get closer and closer to the expected value.

Simpson's paradox a trend appears in several different groups of data but disappears or reverses when these groups are combined

Hypothesis test if the *null hypothesis* is true, would we expect to see data like this? If yes, then stick with the status quo. If no, then we reject the null in favor of the *alternative*.

Null hypothesis The current state of our understanding. The presumed truth. The thing we want to falsify.

Alternative hypothesis The thing we wish to prove. Think *null* = innocent, *alternative* = guilty.

P-value Given that the null is true, what is the probability of our observed data, or something more extreme

Statistical significance if the p-value is "small enough", we call the effect "statistically significant"

Margin of error we have an estimate of something (say a poll that shows 30% support), but how much might that estimate vary? This range is the "margin of error"

Graphics scaling, coloring, axes, percentages, etc. Does this graphic make the point clearly? Does it obscure alternative interpretations to manipulate?

Proportionality principle If various alternatives are equally likely, and then some event is observed, the updated probabilities for the alternatives are proportional to the probabilities that the observed event *would* have occurred under those alternatives.

Sample space An enumeration of the possible outcomes of a process.

Foxy forecasting drawing on a wide variety of experiences and evidence to make predictions rather than using the lens of a single guiding idea

Reproducible research Work which can be easily replicated by another researcher or group by taking in the original data and running a series of well-described codes

Peer review The process by which other experts evaluate and ascertain whether the results described can be widely accepted within a research community as valid

Mean vs Median The mean or average is the sum of all the observations divided by the number of observations. The median is the value that splits the data into two equal sized parts. The mean is influenced by very large or small values, outliers, while the median is not. The median is more *robust* to abberant values.

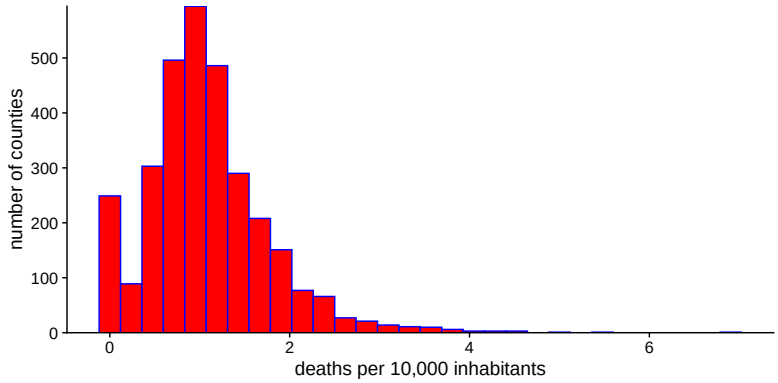
Forecast horizon How far in the future you're making predictions. Generally, the farther ahead you try to predict, the more uncertain you are.

Forecast uncertainty When you make a prediction, you want to think about sources and sizes of possible errors. How accurate is your prediction?

Model averaging Combining information from lots of different sources to make a prediction. This is "foxy".

Histogram A graphical display of possible values along with the observed frequencies of those values. For example:

```
## Warning: `expand_scale()` is deprecated; use `expansion()` instead.
```



Bias The difference between the truth and your estimate, on average.

Variance How much does your estimate vary around the truth, on average.

Over-fitting and Under-fitting Using too complex (over) or too simple (under) a method of analysis for the amount of data you have.

Power law a relationship wherein a relative change in one quantity results in a proportional relative change in the other quantity. For example, the side-length and the area of a square follow a power law relationship.

Correlation vs. Causation correlation measures the strength of a linear association between two quantities. Contrast this with the statement that manipulating one quantity results in a meaningful change in the other

Hidden common causes It is difficult to determine if one thing causes another purely through observation. We need to manipulate one to watch for an effect. If we simply observe, it is hard to rule out hidden factors that may cause both to change.

Bayesian Using "prior" information to think about predictions and uncertainty. As we collect data, we update our existing beliefs to incorporate this new evidence into our understanding.

Sampling How are the data collected? Some different types of samples are the *random sample* or the *convenience sample*.

Sampling Frame The collection of all possible individuals (or objects or . . .) that can be selected to be part of a sample. The Census draws *every* person in the frame (all people in the US).

Confirmation bias The tendency to search for, interpret, favor, and recall information in a way that confirms or strengthens one's prior personal beliefs or hypotheses.

Algorithm a computational workflow. A computer program or collection of mathematical formulas that

take some inputs to produce outputs based on a collection of rules

Mathematical model (p18) an abstract representation of a real-world process. A simplified mathematical description of the world that possibly ignores some complexities

Weapon of math destruction (initial definition, p31) a mathematical model that posesses the following 3 features. (1) opacity: only, possibly, the users understand how the model produces predictions; (2) scale: the model has an outsize impact on large swaths of the population; and (3) damage, that impact results in negative economic, social, or other outcomes.

Feedback loop (throughout) a WMD often has the side effect that it's predictions lead to negative outcomes that then *induce* the prediction. For example, the recidivism model leads to longer sentences but longer sentences tend to produce recidivism.

Proxy measure (many, eg p52, p146) frequently, the traits one would want to use in a model are unavailable or unmeasured, so WMDs instead use something to stand in for those traits

Machine learning (p76) Allowing a machine to process many examples of features and associated outcomes to find a plausible method of predicting outcomes for future examples

Calibration (p110) we discussed a different definition earlier in the context of weather forecasts. here the author means that data scientists turn some knobs in a model until the predictions match their beliefs. This is done without feedback that would allow them to determine if those predictions are accurate

Training and Testing Data (p116) training data is the information used to create the model. Testing data is the new data where you want to make predictions. Say I look at the last 10 years of IU basketball games to learn when they're likely to win. That's training data. Testing data would be the future games whose outcomes I'd like to predict.

Operations research (p127) A field of math/engineering/business that historically focused on optimizing things like supply chain management (how do we get the inputs our widgets need while sending out our widgets in the most efficient way possible). But the modern economy is much more of a service economy, so OR has morphed to manage the human interactions as well.

False negatives and False positives (p133) A hugely important idea at the moment. Suppose you take a test for COVID19. There are four outcomes (1) true positive: the test says you have it and you actually do (2) true negative: the test says you don't have it and you don't (3) false positive: test says you have it, but you have a cold (4) false negative: you have it, but test says no. (4) here can be a real problem because they send you home with no instructions to quarantine and you pass it to others. The same idea happens whenever we try to predict one of two outcomes (eg. potential criminal, bad teacher)

Stratification (p162) See also Simpson's paradox above. The idea is to make analyses with smaller pieces of the whole population as well as the whole. Sometimes, small slices drive the overall effect, but other slices behave differently. Sometimes, each slice and the overall population behave the same. Simpson's paradox is when all the slices go one way, but the overall average goes the opposite. Ideally, we don't want to mistake any of these situations for the other.