Stat100 Concepts

DJM

As of: 2020-02-15

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.