Causal Graphs: Probability Cheat Sheet

Julian Schuessler, Universität Konstanz

April 15, 2018

1 Probability

- We always should make explicit the population we are interested in, e.g. all countries in the world in 2018, or every German citizen in 2018
- These "units" (countries/citizens) have features or "variables" Y and X like GDP or age. In this course, we mostly deal with discrete variables; this simplifies everything, but generalizations to continuous variables are usually very straightforward
- An "event" or "realization" is an assignment of values to one or multiple variables, like "Y = 1 and X = 23"
- $P(Y = y)$ is the ("marginal") probability that a feature Y takes on the value y, or the share of units in the population with $Y = y$ (frequency interpretation), or our belief that Y will be y (Bayesian interpretation). It is a number between 0 and 1
- $\sum_y P(Y = y) = 1$, where $\sum_y$ is shorthand for "sum over all possible values y of Y"
- $P(Y)$ is the probability distribution or probability mass function (PMF), a function that outputs probabilities for different $Y = y$
- $P(Y = y, X = x)$ is the share/(joint) probability/belief that a unit has both $Y = y$ and $X = x$. Example: Share of people of age $Y = 23$ and gender $X = 1$.
- $P(Y|X = x)$ is the probability distribution of Y conditional on $X = x$. Conditioning is like filtering: You throw away all units which have $X \neq x$. In the Bayesian interpretation, conditioning is "learning new information": You are being told that $X = x$.
- At the end of the day, conditional probabilities are just probabilities. $\sum_y P(Y = y|X = x) = 1$
- $P(Y, X) = P(Y|X) \cdot P(X) = P(X|Y) \cdot P(Y)$. If you want to know the joint probability of X and Y, consider the probability that X occurs, and then the probability of Y conditional on that X. Also works starting with Y.
- It follows that $P(Y|X) = \dfrac{P(Y, X)}{P(X)}$ (Bayes' Law)
- Independence (very important): Y and X are independent if $P(Y|X) = P(Y)$. That is, knowing X is not useful for learning about $P(Y)$. Or: If you filter the population along X, the resulting distribution of Y looks just like before

- We also write independence as $Y \perp\!\!\!\perp X$.
- Just like variables can be (marginally) independent, they can also be conditionally independent: $P(Y|X,Z) = P(Y|Z)$. We say "Y is independent from X, conditional on Z". Shorthand notation: $Y \perp\!\!\!\perp X|Z$.
- Bayesian interpretation of conditional independence: Once we know Z, additionally learning X does not change our belief about Y
- Law of total probability: $P(Y=y) = \sum_x P(Y=y, X=x) = \sum_x P(Y=y|X=x)P(X=x)$ AKA "divide and conquer", or "subset and average"
- With the LoTP, you can also further decompose conditional probabilities like $P(Y|X)$. Since the latter apply to a population filtered along X, all decompositions are also conditional on X. So $P(Y|X) = \sum_z P(Y|X, Z=z)P(Z=Z|X)$

2 Expected Values

- $E[Y] = \sum_y yP(Y=y)$ is the expectation or expected value or mean of Y. This is a number and a property of the population
- Law of the unconscious statistician: $E[f(Y)] = \sum_y f(y)P(Y=y)$
- The expected value of Y, given X, is often called $E[Y|X]$ and is defined as $\sum_y yP(Y=y|X)$. This is a function of the random variable X; and it will vary for different realizations x of X. $E[Y|X=x]$ is a number. The same filtering logic as for conditional probabilities applies
- "Expectations are linear": a, b constants, then $E[a+bY] = a + b \cdot E[Y]$
- If X and Y are independent, $E[Y|X] = E[Y]$ and $E[X|Y] = E[X]$
- "Law of Iterated Expecations": $E[Y|X]$ is random through X, so it has a mean. $E[E[Y|X]] = E[Y]$.
- $E[E[Y|X]|X]$ filters twice along the same variable, so one filter is unncessary. So $E[E[Y|X]|X] = E[Y|X]$
- $E[E[Y|X]|Z]$ filters along Z, then in those subgroups filters along X, and finaly averages over X. So $E[E[Y|X]|Z] = E[Y|Z]$. We could also add the superfluous Z-filter inside: $E[E[Y|X]|Z] = E[E[Y|X,Z]|Z]$
- $E[Y|X] = f(X)$. For binary X, $E[Y|X] = E[Y|X=0] + (E[Y|X=1] - E[Y|X=0])X$. Adding Y two both sides, rearrangement and renaming gives the linear regression $Y = \alpha + \beta X + \epsilon$
- $\epsilon = Y - E[Y|X]$. $E[\epsilon|X] = E[Y - E[Y|X]|X] = E[Y|X] - E[E[Y|X]|X] = E[Y|X] - E[Y|X] = 0$ using arguments from above
- So ϵ and X are mean-independent by construction if X is discrete; OLS is consistent for the regression coefficients under perfect measurement and random sampling; the regression coefficients by definition are just differences in means and by themselves have no causal meaning