Quantitative Methods

Human Sciences
Syllabus review, 2020–2021

1. Probability

▲ Describe the three things we need to define a probability:

1. A sample space S.
2. A class of well-defined events: A, B, A^{c}, $A - B$, etc.
3. A probability function $\mathbb{P} : S \to [0, 1]$.

Exercise: You are a medical researcher trying to better understand the causes of a (non-infectious) disease. The disease can manifest in one of two ways: mildly (the patient exhibits certain mild but harmless symptoms) or strongly (the patient becomes very ill). You hypothesise that ingesting a particular substance causes the disease, but you have no evidence for this. Since you have no sense of ethics, you decide to pick one of your patients at random and make them ingest the chosen substance to see what happens. (Assume that the final outcome of the experiment will be certain after a fixed amount of time.) Describe the sample space corresponding to this experiment and the set of events in which you (the researcher) are most interested.

▲ Describe and justify the three axioms of probability:

1. $\mathbb{P}(A) \geq 0$ for any event A.
2. $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(S) = 1$.
3. If $A_1, \ldots, A_n$ are mutually exclusive events, then $\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$.

Exercise: I assign a probability $\mathbb{P}(A) \in [0, 1]$ to some event A based on my subjective degree of belief in how likely A is to occur. I issue a certificate that reads: *The owner of this certificate can redeem it for £1,000 if A occurs.* I am willing to sell or buy such a certificate for £1000 × $\mathbb{P}(A)$, as I consider this to be a fair price. However, I refuse to accept the third axiom of probability. In other words, I posit that for any two disjoint events A and B, $\mathbb{P}(A \cup B) \neq \mathbb{P}(A) + \mathbb{P}(B)$. Nonetheless, I keep buying and selling such certificates. Explain how, after a sufficient amount of time has passed (assuming that the certificates are in fact legally binding and that I have a fixed amount of money in my bank account), I am guaranteed to have lost all my money.

▲ Be able to count, and explain counting identities using story proofs.

Exercise: Without doing any mathematical calculations, prove the following identity:

$$\binom{n}{k} = \binom{n}{n-k}.$$

▲ Define and explain the following:

- ⋆ Conditional probabilities: $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.
- ⋆ Bayes's Rule: $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$.
- ⋆ The Law of Total Probability: $\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B \mid A_1) + \cdots + \mathbb{P}(A_n)\mathbb{P}(B \mid A_n)$, for a partition $A_1, \ldots, A_n$ of the sample space.

Exercise: Consider a disease that exhibits X chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X chromosome is affected with certainty, whereas a female carrying the gene on only one of her two X chromosomes is not affected. Assume the disease is fatal at birth in males with one such gene and in females with two such genes. Consider a woman named Sarah who has an affected brother.

(a) Given this information, what conclusion can we draw about Sarah's mother?

(b) Our event of interest, call it G, is the outcome of a binary variable measuring whether Sarah has the gene in question or not (we have either G or G^{c}). Given only the above information, and based on your answer to part (a), what is the prior probability that Sarah carries the gene? That is, what is $\mathbb{P}(G)$?

(c) We are now given a piece of additional information: Sarah herself has two (non-identical) sons, neither of whom is affected. We denote the event that both sons are unaffected by S^{c}. Calculate $\mathbb{P}(S^{\mathsf{c}} \mid G)$ and $\mathbb{P}(S^{\mathsf{c}} \mid G^{\mathsf{c}})$.

(d) Use Bayes' Rule and the Law of Total Probability to calculate $\mathbb{P}(G \mid S^{\mathsf{c}})$.

▲ Describe and explain the following:

- ⋆ (In)dependence: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A) = \mathbb{P}(B)\mathbb{P}(A \mid B)$.
- ⋆ Conditional (in)dependence: $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid A \cap C) = \mathbb{P}(B \mid C)\mathbb{P}(A \mid B \cap C)$.

Exercise: A family has three children, named Alpha, Beta, and Gamma.

(a) Is the event where Alpha is older than Beta independent of the event where Alpha is older than Gamma?

(b) Imagine Gamma is the youngest and is very fond of playing chess. On any given day, Gamma's two older siblings might independently offer to play a round of chess with her. Given that Gamma will only play a single game of chess tomorrow, is the event that she will play against Alpha independent of the event that she will play against Beta?

(c) On odd-numbered days of the month, Gamma likes to play chess in an unorthodox way: all of her odd-numbered moves are independent of her even-numbered move(s). Conditioning on the day of the month, what is the dependence structure between the event A that Gamma's nth move is x and the event B that her $(n+1)$st move is y?

▲ Define a (discrete or continuous) random variable as a function $X : S \to \mathbb{R}$ that assigns a numerical value to each possible outcome of an experiment and identify (in)dependence structures between random variables.

Exercise: If X, Y, and Z are random variables such that X and Y are independent and Y and Z are independent, does it follow that X and Z are independent?

▲ Define and describe probability distributions associated with random variables, especially:

- ⋆ Bernoulli, Binomial, Poisson, Uniform, Normal distributions.
- ⋆ Recognise the two things needed for a function f to count as a probability distribution:
  1. $f(x) \geq 0$ for all x.
  2. $\sum f(x) = 1$ (discrete) or $\int f(x)dx = 1$ (continuous).

Exercise: Let X be a Binomial random variable with n denoting the number of 'trials', k denoting the number of 'successes', and p denoting the probability of success in each trial.
(a) By going through each component of the following formula, explain why the probability distribution of X is given by

$$f(n,k,p) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

(b) Consider the case where $n = 2$ and $p = 0.5$. Verify that $f(2, k, 0.5)$ is a valid probability distribution.

▲ Describe and compute (see below) the expectation of and (co)variance/correlation between random variables (with respect to their probability distributions):

- ⋆ $\mathbb{E}(X) = \sum_i x_i \mathbb{P}(X = x_i)$ (discrete) or $\mathbb{E}(X) = \int x f(x)dx$ (continuous).
- ⋆ $\mathsf{Var}(X) = \mathbb{E}\big[X - \mathbb{E}(X)\big]^2$.
- ⋆ $\mathsf{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
- ⋆ $\mathsf{Corr}(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}$.

Exercise: Consider a population in which life expectancy at birth is 80 years. John is 79 years old. Is his remaining life expectancy 1 year?

▲ Describe the Law of Large Numbers and the Central Limit Theorem and their implications.

Exercise: We have a fair coin. We define 'success' as obtaining heads when flipping the coin. Let p denote the probability of 'success' for a given trial.
(a) What are the mean and variance of the binary random variable X indicating whether we obtain heads or tails? (Hint for calculating the variance: note that X^2 is still binary for a binary variable X.)
(b) We flip the coin 10 times, getting 8 heads and 2 tails. Without knowing that it is in fact a fair coin, we might draw the wrong conclusion about p. With reference to the Law of Large Numbers, explain what we should do in order to compute the true value of p.
(c) One way of expressing the Central Limit Theorem is to say that, for a large sample size n, the distribution of the sample mean $\overline{X}_n$ is approximately $\mathcal{N}(\mu, \sigma^2/n)$, where μ is the true expectation of X and σ^2 is the true variance of X. Using this result, describe the distribution of $\overline{X}_n$ for large n.

2. Inference

▲ Understand and describe the key problem of statistical inference involving the transition from $\mathbb{P}(\text{data} \mid \text{model})$ to $\mathbb{P}(\text{model} \mid \text{data})$.

Exercise: Describe how Bayes' Rule can be used as a tool to address the key problem of statistical inference.

▲ Understand and describe the likelihood theory of inference.

Exercise: Describe the principle of maximum likelihood inference.

▲ Understand and describe the Bayesian theory of inference and its relation to the likelihood theory of inference.

Exercise: Using the equation $\mathbb{P}(\theta \mid y) \propto \mathbb{P}(\theta)\mathbb{P}(y \mid \theta)$, where y is an observed data point and θ is some parameter of interest (e.g., the mean of a distribution), explain how the Bayesian theory of inference differs from the likelihood theory of inference. Illustrate by giving a simple (real or fictional) example.

▲ Understand and describe the principles of ordinary least squares estimation.

Exercise: You are given a scatterplot with observations of a random variable X on the X-axis and observations of a random variable Y on the Y-axis. You are interested in describing the association between X and Y. Explain how a computer programme would estimate this association using ordinary least squares.

▲ Understand and describe what constitutes a statistical model, its systematic and stochastic components, and its key assumptions.

▲ Understand, compute, and interpret regression models of the form $y_i = \alpha + x_i\beta + \epsilon_i$ and assess the distributional assumptions surrounding the error term.

Exercise: You are interested in the association between parental wealth (X) and children's educational attainment (Y). You posit that, for an individual child i, $y_i = \alpha + x_i\beta + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
(a) Describe the key assumptions of this model.
(b) What is an alternative way of expressing the same model?
(c) How does the model relate to the conditional expectation given by $\mathbb{E}(Y \mid X)$? Give a substantive interpretation of this conditional expectation using the given example.
(c) In an intuitive but clear way, explain why $\beta = \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(X)}$.

▲ Assess when a model parameter β might have a causal interpretation by reasoning in terms of counterfactuals (or potential outcomes).

Exercise: Using the same example as above, under what conditions might the estimated association between parental wealth and children's educational attainment be interpreted as causal?

▲ Describe the fundamental problem of causal inference: only one potential outcome is observed for any individual.

▲ Define the population average causal effect as a contrast between expectations (or probabilities) of counterfactual outcomes Y_0 and Y_1:

$$\mathbb{E}(Y_1) - \mathbb{E}(Y_0) \quad \text{or} \quad \mathbb{P}(Y_1) - \mathbb{P}(Y_0).$$

▲ Describe the key characteristics of, compare and contrast, and critically assess the strengths and weaknesses of randomised controlled trials and observational studies.

Exercise: What is the main characteristic that distinguishes a randomised controlled trial from an observational study?

Exercise: 'Unlike observational studies, randomised controlled trials do not suffer from systematic bias and therefore provide rigorous and reliable causal effect estimates'. Discuss.

▲ Define and visualise using causal graphs the three main forms of systematic bias — confounding, selection, and measurement bias — that can undermine causal inferences.

Exercise: Using a causal graph and concrete example, describe the phenomenon of confounding.

3. Data analysis in R

▲ Be able to import data into R.

▲ Be able to tidy data in R.

▲ Be able to simulate from and compute key features of the most important probability distributions (e.g., `rnorm()`, `mean()`, etc.).

▲ Be able to specify linear regression models using `lm()` (or equivalent) and interpret model outputs. Any presentation of results should:

- ⋆ describe the underlying data and the key assumptions made about the data-generating process,
- ⋆ present key quantities of interest (not merely those things that your software happens to print out) in a clear, easily interpretable way,
- ⋆ assess model fit (noting the difference between absolute and relative measures of model fit),
- ⋆ discuss strengths and limitations of the model.

▲ Be able to visualise data using `ggplot()` or `plot()`. Any graphic should

- ⋆ have a clear purpose and convey a simple message (avoid clutter or too much information at the same time),
- ⋆ be clearly annotated to aid the reader (meaningful axis scales and labels, as well as informative captions),
- ⋆ be easy to interpret (both theoretically and substantively). Pretty colours or aesthetic tweaks are secondary, though welcome — unless they divert attention from the substantive matter at hand.

▲ Write clean and replicable code in R (annotate your code and make sure that if you, or anyone else, runs the same code in two months, they will get the same results).

Elias Nosrati
February 2021