11. Null Hypothesis Significance Testing

- In NHST the goal of inference is to decide whether a particular value of a parameter can be rejected.
- The p value is the probability of getting an outcome from the null hypothesis that is at least as extreme as the actual outcome *when using the intended sampling and testing procedures*
- This reasoning depends on defining a space of all possible outcomes for the null hypothesis. This space is based on how we intend to collect the data.
  - E.g. was the intention to flip the coin N times? Then space of outcomes is all possible sequences of N flips
  - Or was it to flip until zth head appeared? Different set of possible outcome sequences
  - Or flip for a fixed duration?
- A good experiment (or survey) is founded on the principle that the data are insulated from the experimenter's intentions. The essential constraint on the stopping rule for flipping for example is that it should not bias the data that are obtained.

11.1 Paved with Good Intentions

Flip coin N=24, z=7

11.1.1 Definition of p value

- Null hypotheses example: $\theta = 0.5$
- To derive a p value from the null hypothesis, we must also specify how to generate full samples of data.
- The likelihood function defines the probability for a single measurement.
- The intended sampling process defines the cloud of possible sample outcomes.
- The null hypothesis is the likelihood function with its specific value for paramter θ
- Cloud of possible samples is defined by the stopping and testing intentions denoted I.

11.1.2 With intention to fix N

- **Sampling distribution**: the probability distribution of the possible sample outcomes.
  - Sample outcome for example might be z/N
- **Frequentist** methods are ones that rely on sampling distributions. A particular application of frequentist methods is NHST.
- Sampling distribution is a binomial distribution.
- $p = 0.032$, so we do not reject the null hypothesis that $\theta = 0.5$

11.1.3 With intention to fix z

- Sampling distribution is negative binomial
- $p = 0.017$, so we reject the null hypothesis.
- This stopping rule is triggered by the data itself and can produce a biased sample.

11.1.4 With intention to fix duration

- N distributed like a Poisson distribution. For comparison choose λ to be what N was when we fixed N.
- Overall distribution is a weighted mixture of binomial distributions.
- $p = 0.024$, so we reject the null hypothesis - *marginally significant*.
- Real-world sampling distributions are often complex.

11.1.5 With intention to make multiple tests

- Toss two coins, assumed independent, and test whether either are biased.
- p value for the first coin depends on the intention to filp the second coin N_2 times.

11.1.6 Soul searching

- Difficult to determine true intention of researcher. Was something really conceived a priori or post hoc?
- We must carefully design experiments to insulate coins from the intentions of the experimenter.

11.1.7 Bayesian analysis

- Does not rely on intention of researcher, only operates with the actual data.

11.2 Prior Knowledge

- Flipping a nail vs flipping a coin example.

11.2.1 NHST analysis

11.2.2 Bayesian analysis

11.3 Confidence Interval and Highest Density Interval

- Confidence intervals made by taking the p value at every possible outcome

11.3.1 CI depends on intention

11.3.2 Bayesian HDI

11.4 Multiple Comparisons

11.4.1 NHST correction for experimentwise error

11.4.2 Just one Bayesian posterior no matter how you look at it

11.4.3 How Bayesian analysis mitigates false alarms

11.5 What a Sampling Distribution IS Good For

11.5.1 Planning an experiment

11.5.2 Exploring model predictions (posterior predictive check)