Lecture 31, Mar 29, 2023

Hypothesis Testing

- A *hypothesis* is a conjecture made about a population
  - e.g. $H_0 : P(H) = 0.5$ for a coin; an alternative hypothesis is $H_1 : P(H) > 0.5$
- In general, we have a *null hypothesis* H_0 (the status quo, or what we believe before the experiment), and an *alternative hypothesis* H_1; with each sample tested, we either reject H_0 for H_1, or we fail to reject H_0 and nothing changes
  - In the case where we fail to reject H_0, we can't make conclusions because it's still possible that H_0 is rejected in a later experiment
  - e.g. in a drug trial, H_0 is the drug has no effect, H_1 is the drug having an effect

Type I and II Errors

- *Type I* errors are rejections of H_0 when it is true (i.e. false positives); α is the probability of a type 1 error
  - These result from oversensitive tests
- *Type II* errors are failures to reject H_0 when it is false (i.e. false negatives); β is the probability of a type 2 error
  - These results from undersensitive tests
- Example: testing for mean tensile strength of a new alloy
  - $H_0 : \mu = 1000\text{MPa}$
  - $H_1 : \mu \neq 1000\text{MPa}$
  - Suppose we have $n = 25$ with $\sigma = 50$, to use the CLT
  - First, $P(\mu = 1000) = 0$ since μ is continuous; we therefore need to define a range where we don't reject H_0, e.g. $990 \leq \bar{x} \leq 1010$
    - The *critical region* is the complement of this range (i.e. the range that results in rejection of H_0)
  - This is a confidence interval; we want to compute the confidence
  - $\alpha = P(\text{Type I})$
    $$\begin{aligned} &= P(\bar{X} < 990 \cup \bar{X} > 1010 \mid \mu = 1000) \\ &= 1 - P(990 \leq \bar{X} \leq 1010 \mid \mu = 1000) \\ &= 1 - P\left(\frac{990 - 1000}{\frac{50}{\sqrt{25}}} \leq Z \leq \frac{1010 - 1000}{\frac{50}{\sqrt{25}}}\right) \\ &= 1 - P(-1 \leq Z \leq 1) \\ &= 0.32 \end{aligned}$$
    - The chance of a false positive is 32%, which is not good
    - To reduce α, we can either increase n, or widen the range where H_0 is accepted
    - If we do the latter however, that makes our test less sensitive and increases the probability of a type II error
- The exact size of the critical region is often ad-hoc, but consistency is key – do many tests with the same critical region, and compare results