The t-Distribution, t-Tests, and Simulation
Part B

(William Sealy Gosset published work on the t-distribution while working at the Guinness Brewery in 1908)
Last Time

• Want to compare two samples
  \[ X_1, X_2, X_3, \ldots, X_{N_x} \sim N(\mu_X, \sigma^2) \]
  \[ Y_1, Y_2, Y_3, \ldots, Y_{N_Y} \sim N(\mu_Y, \sigma^2) \]

• Null Hypothesis:
  \[ \mu_X = \mu_X \]

• Known \( \sigma^2 \):
  \[ \frac{(\bar{X} - \bar{Y})}{\sqrt{2} \frac{\sigma}{\sqrt{n}}} \sim N(\mu_X - \mu_Y, 1) \]

• 95% CI for \( \mu_X - \mu_Y \):
  \[ P \left( (\bar{X} - \bar{Y}) - z_{.975} \frac{\sigma}{\sqrt{N}} \leq \mu_X - \mu_Y \leq (\bar{X} - \bar{Y}) + z_{.975} \frac{\sigma}{\sqrt{N}} \right) = 0.95 \]
Last Time

• Unknown $\sigma$:
  - Estimate $s_X^2 = \frac{\sum(x_i - \bar{X})^2}{N_X-1}$, $s_Y^2 = \frac{\sum(y_i - \bar{Y})^2}{N_Y-1}$
  - Pooled estimate: $s_{pooled} = \sqrt{\frac{(N_X-1)s_X^2 + (N_Y-1)s_Y^2}{(N_X+N_Y-2)}}$
  - $SD(\bar{X} - \bar{Y}) = s_{pooled} \sqrt{(\frac{1}{N_X} + \frac{1}{N_Y})}$
  - CI:
  \[
  (\bar{X} - \bar{Y}) \pm t_{.975}(N_X + N_Y - 2) \cdot s_{pooled} \sqrt{(\frac{1}{N_X} + \frac{1}{N_Y})}
  \]
Two-Sample t-Test

• (in R)
One-sided or Two-sided p-Values?

• One-sided: \(P(T > t);\) or \(P(T < t)\)
  • Appropriate if it’s completely implausible that \(T < 0\) (resp. \(T > 0\))
  • Example: does speeding let you get home faster?

• Two-sided \(P(|T| > |t|)\)
  • Appropriate if we are not completely sure whether the effect should be positive or negative
  • More conservative

• No general consensus on clear rules
t-Test Robustness

• Robustness: a statistical procedure is robust to departures from a particular assumption if it is valid even when the assumption is not satisfied

• (Experiments in R)

• Basically: if the sample size is large enough, the sample mean will still be normally distributed, so the t-Test is fine

• Consider transforming the data (e.g. with a log transformation) if that will help with the normality assumption
Cloud Seeding Data

• Cumulus clouds were seeded/injected with silver iodide on some days, and data was collected on “seeded” days and “unseeded” days.
Sampling Distribution of the Sample Variance

- The sample variance $s_X^2$, which we used before in order to get at the sampling distribution of the mean, is itself a random variable
- Reminder: $s_X^2 = \frac{1}{N_x-1} \sum (X_i - \bar{X})^2$
- By definition, if $Z_1, Z_2, ..., Z_N \sim N(0, 1)$ are iid, then $\sum Z_i^2 \sim \chi^2(N)$
- $\frac{(N-1)s_X^2}{\sigma^2} = \sum \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(N - 1)$
- (This is *not easy* to prove, but follows the intuition that we have one less degree of freedom because we have to estimate the mean)