Lessons 035 - 036
Two Sample t Test and Paired Data

Wednesday, November 29

Two Sample t Test

- We previously considered two sample hypothesis tests with known variances or with a large enough sample size to use normal approximations.
- If we have small sample sizes, in a normal population, with unknown variances we can use the trick from single sample testing

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \sim t_\nu.$$

Sampling Distribution

- The number of degrees of freedom is given by

$$\nu = \frac{\left(\dfrac{s_1^2}{n} + \dfrac{s_2^2}{m}\right)^2}{\dfrac{s_1^4/n^2}{n-1} + \dfrac{s_2^4/m^2}{m-1}}$$

- In practice, we will simply use $\nu = \min\{m-1,\ n-1\}$ almost always.

Two Sample Procedures

- Once the test statistic is formed, hypothesis testing proceeds exactly as we have previously seen.

- This procedure will apply to any $m, n > 1$ when the population is normal.

- Previously, we required either a known variance or else a large sample.

Pooled Variance Estimation

- Sometimes there is a reason to believe that, theoretically, $\sigma_1^2 = \sigma_2^2$ even when we do not know the value.
- In this case, var$(\overline{X} - \overline{Y}) = \sigma^2 \left(\frac{1}{n} + \frac{1}{m} \right)$.
- We can estimate σ^2 using information from both samples.

$$\widehat{\sigma}^2 = S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

Sampling Distribution (Pooled Estimator)

- If we use S_p^2 to scale the previous estimator we get

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{m+n-2}\,.$$

- We should only use this when there is **good reason to believe we can pool the variance.**

Paired Data

- So far we have assumed that our two samples are independent of one another.
- This is often not the case.
- A common type of dependence observed is data "pairing".
  - Two treatments on the same individual.
  - Before-and-after testing.
  - Studies on twins.

Paired Data: Formally

- In the case of paired data we have two sets of n observations, say $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$.
- For each $j = 1, \ldots, n$, X_j and Y_j are dependent on each other.
- For each non-paired variate, the data remain independent.
- Instead of averaging and then differencing, we will difference first.

$$D_i = X_i - Y_i$$

Single Sample from Paired Data

- Once we have formed $D_1, \ldots, D_n$ we have a single sample.
- We will have $E[D_i] = \mu_1 - \mu_2$ and we will have

$$\text{var}(D_i) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}.$$

- Here, σ_{12} is the **covariance** between X_i and Y_i.

$$\text{cov}(X, Y) = E\left\{(X - E[X])(Y - E[Y])\right\}.$$

A Brief Aside on Covariance:

- We have not *really* discussed the tools to find covariances directly.
- If $X \perp Y$ then cov$(X, Y) = 0$. For paired observations typically cov$(X, Y) > 0$.
- The **correlation** is simply a scaled version of covariance.
- We require the **joint probability density function** which is a multivariate extension of the PDF.
- If population covariances are required, you will be told them.

Testing with Paired Data

- With D_i formed, then we can take the standard test statistic and use the standard cases for the sampling distribution.

$$T = \frac{\overline{D} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})/n}}$$

- If X and Y are both normal, so too is D.
- If n is large enough, this will behave approximately normally.
- If n is small, D is normal, can use the t_{n-1} sampling distribution.

The assumption that data are paired should be used sparingly.