Sets 28 and 29: Sections 7.1, 7.2 - Inference for two samples

We now study the two sample problem where the data $X_1, \ldots, X_m$ iid $\text{Normal}(\mu_1, \sigma_1^2)$ is independent of $Y_1, \ldots, Y_n$ iid $\text{Normal}(\mu_2, \sigma_2^2)$. Initially, we make the unrealistic assumption that both σ_1 and σ_2 are known.

Under the above conditions, interest lies in the unknown parameter $\mu_1 - \mu_2$. The test statistic used in the construction of confidence intervals and hypothesis testing is

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \text{Normal}(0, 1)$$

Example: A sample of 100 recent University of Victoria grads and 80 recent UBC grads finds that the salaries for UVic grads has sample mean $\$31,000$ with standard deviation $\sigma_1 = 1000$, and the salaries for UBC grads has sample mean $\$28,800$ with standard deviation $\sigma_2 = 600$. Is there reason to believe the true mean salaries are different? Assume that salaries are normally distributed.

Example continued: Construct a 95% confidence interval for $\mu_1 - \mu_2$.

The significance of 'significance':

When we reject the null hypothesis H_0, we say that the result is *statistically significant.*

Discussion points:

- **always report the p-value**
- **keep in mind that $\alpha = 0.05$ is arbitrary**

- **statistical significance does not always mean importance**
- **p-values are related to sample size**

More on stat significance vs practical importance:

In an Austrian study of 507,125 military recruits, it was found that the average height of those born in the spring was $1/4$ **inch more than those born in the fall.**

In two sample problems, we can relax the normality assumption in the case of large samples.

Given $X_1, \ldots, X_m$ **iid independent of** $Y_1, \ldots Y_n$ **iid with** m **and** n **large (ie.** $m, n \geq 30$**), then the following statistic can be used for testing and the construction of confidence intervals.**

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/m + s_2^2/n}} \approx \text{Normal}(0, 1)$$

where μ_1 **and** μ_2 **are the respective means, and** s_1 **and** s_2 **are the respective sample std devs.**

Example: A college interviewed 1296 students wrt summer incomes. Based on the results in the following table, test whether there is a difference in earnings between male and female students.

Students	n	$\bar{X}$	s
male	675	\$1884.52	\$1368.37
female	621	\$1360.39	\$1037.46

Example: The test scores of first year students admitted to college directly from high school historically exceed the test scores of first year students with working experience by 10%. A recent sample of 50 first year students admitted directly from high school

has an average test score of 74.1% with std dev 3.8%. An indpt sample of 50 first year students with working experience yields an average test score of 66.5% with std dev 4.1%. Test whether a change has occurred.

Small samples: We consider another variation to the two sample problem. This time, the data are again normal. Realistically, σ_1 and σ_2 are unknown, but m and/or n are less than 30. We will need to make the additional assumption $\sigma_1 = \sigma_2$.

Given $X_1, \ldots, X_m$ iid $\text{Normal}(\mu_1, \sigma_1^2)$ independent of $Y_1, \ldots Y_n$ iid $\text{Normal}(\mu_2, \sigma_2^2)$ with $\sigma_1 = \sigma_2$, then the following statistic can be used for testing and the construction of confidence intervals.

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) s_p^2}} \sim t_{m+n-2}$$

where s_1 and s_2 are the respective sample std devs, and

$$s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m-1+n-1}$$

Example: The Chapin Social Insight Test* gave the following scores. Assuming normal data, test whether the mean score of males exceeds the mean score of females.

Group	n	$\bar{X}$	s
males	18	25.34	13.36
females	23	24.94	14.39

***This thirty-item test measures an individual's ability to diagnose a situation involving human interaction, recognize the dynamics underlying behavior, or choose the wisest course of action to resolve a difficulty.**

Example cont'd: Obtain a 95% CI for $\mu_1 - \mu_2$.

Example: We compare the lifespans of smart-phones produces by two companies. The average lifespan for Company A's $m = 15$ phones is 148 weeks with a standard deviation of 8.3 weeks, and Company B's $n = 6$ phones have an average lifespan of 153 weeks with a standard deviation of 5.1 weeks. Let μ_1, μ_2 be the mean lifespan of smartphones produced by Company A, B (respectively).

Is it reasonable to assume that $\sigma_1 = \sigma_2$?

A formal test of $H_0 : \sigma_1 = \sigma_2$ is beyond this course. Rule of thumb:

$$\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} \begin{cases} \leq 1.4 & \textbf{assume } \sigma_1 = \sigma_2, \\ > 1.4 & \textbf{do NOT assume that } \sigma_1 = \sigma_2. \end{cases}$$

In the second case, we use the following test statistic:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/m + s_2^2/n}} \sim \text{Student}(\nu)$$

$$\nu = \text{integer part}\left[\frac{(s_1^2/m + s_2^2/n)^2}{\dfrac{(s_1^2/m)^2}{m-1} + \dfrac{(s_2^2/n)^2}{n-1}}\right]$$

(a) **What is the estimated standard error of $\mu_1 - \mu_2$?**

$$\sqrt{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)}$$

$$= \sqrt{\frac{8.3^2}{15} + \frac{5.1^2}{6}} \approx 2.98792$$

(b) **What distribution (including degrees of freedom) is used in a hypothesis test on $\mu_1 - \mu_2$?**

$$\nu = \frac{(8.3^2/15 + 5.1^2/6)^2}{\frac{(8.3^2/15)^2}{14} + \frac{(5.1^2/6)^2}{5}} = 15.138$$

So, we use t_{15}.

(c) **Test $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 < 0$, at the significance level $\alpha = 0.1$.**

Observed value of the test statistic:
$(148 - 153 - 0)/2.98792 \approx -1.673$

P-value: $P(t_{15} \leq -1.673) = P(t_{15} \geq 1.673)$.

The p-value is between 0.05 and 0.1. Since the $p - value$ is less than α, we reject H_0.

Sample Data	Pivotal	Comments
paired data, $m = n$	take $D_i = X_i - Y_i$ and refer to single sample case	
non-paired, m, n large	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}} \sim \text{normal}(0,1)$	replace σ_i's with s_i's if σ_i's unknown
non-paired, m, n not large, data normal, σ_i's known	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}} \sim \text{normal}(0,1)$	unrealistic
non-paired, m, n not large, data normal, $\sigma_1 \approx \sigma_2$ unknown	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)s_p^2}} \sim \text{Student}(m+n-2)$	$s_p^2 = \frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}$ $\frac{\max\{s_1,s_2\}}{\min\{s_1,s_2\}} \leq 1.4$
non-paired, m, n not large, data normal, $\sigma_1 \neq \sigma_2$ but unknown	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\left(\frac{s_1^2}{m}+\frac{s_2^2}{n}\right)}} \sim \text{Student}(\nu)$	$\nu =$ integer part $\frac{(s_1^2/m+s_2^2/n)^2}{\frac{(s_1^2/m)^2}{m-1}+\frac{(s_2^2/n)^2}{n-1}}$
binomial data, m, n large, p_1, p_2 moderate	$\frac{\hat{p}_1-\hat{p}_2-(p_1-p_2)}{\sqrt{p_1(1-p_1)/m+p_2(1-p_2)/n}} \sim \text{normal}(0,1)$	replace p_i's with H_0 estimates or with $\hat{p}_i$'s in denominator for CI

Table 1: Summary of two-sample inference where $X_1, \ldots, X_m$ are iid with mean μ_1 and standard deviation σ_1, and $Y_1, \ldots, Y_n$ are iid with mean μ_2 and standard deviation σ_2.