Lecture 32

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 we 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.

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}}} \sim t_{m+n-2}$$

where s_1 and s_2 are the respective sample std devs.

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

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

There are actually lots of testing methodologies corresponding to different data scenarios. We will study one more situation (a common one involving paired data) but keep in mind that the principles that we have studied carry over to more complex situations.

Suppose in the paired data situation, we have $X_1, \ldots, X_n$ iid arising from a population with mean μ_1, and $Y_1, \ldots, Y_n$ iid arising from a population with mean μ_2. Furthermore, assume that the data are paired such that X_i corresponds to Y_i. This natural pairing implies that there is a dependence between X_i and Y_i.

To carry out inference (testing and the construction of CI's), we define a new random variable, the difference $D_i = X_i - Y_i$. Our interest concerns the unknown parameter

$$\begin{aligned} \mathrm{E}(D_i) &= \mathrm{E}(X_i - Y_i) \\ &= \mathrm{E}(X_i) - \mathrm{E}(Y_i) \\ &= \mu_1 - \mu_2. \end{aligned}$$

Our analysis proceeds as in the single sample case based on the data $D_1, \ldots, D_n$.

Example: Suppose scores measuring jitteriness are normally distributed . We believe that scores increase after drinking coffee. Let X_i be the before drinking coffee score and let Y_i be the the after drinking coffee score for the i-th individual. Based on $\alpha = 0.01$, test the hypothesis.

x_i	y_i	d_i
50	56	
60	70	
55	60	
72	70	
85	82	
78	84	
65	68	
90	88	

Example cont'd: Obtain a 95% CI for the mean difference in jitteriness scores.

Example cont'd: Suppose we have the same data but the experiment involves 16 people where 8 people were measured without having coffee and 8 other people where measured after drinking coffee. How does the analysis differ?

Example cont'd: Suppose now that the 16 people involve 8 pairs of twins such that X_i and Y_i are twins. How should the analysis proceed?

Example cont'd: Assume the same conditions as above but the data are no longer normal. How should the analysis proceed?

Pairing is a special case of *blocking* (read in text). Blocking attempts to reduce variation by grouping data that are similar, and this hopefully leads to *more sensitive* tests (ie. tests that reject H_0 more often when H_0 is false).

Example: To illustrate the above, consider five before and after measurements involving a drug where there are big differences in responses between people but there is small variation in the D_i's. Assuming normal data, we carry out a paired analysis and a non-paired analysis.

x_i	y_i	d_i
25	29	-4
46	50	-4
30	33	-3
75	78	-3
19	25	-6

Two Sample Testing - Summary

Assume $X_1, \ldots, X_m$ iid with mean μ_1 and std dev σ_1, and $Y_1, \ldots, Y_n$ iid with mean μ_2 and std dev σ_2.

Data	Test Statistic	Comments
paired data, $m = n$	take $D_i = X_i - Y_i$ and refer to single sample case	
non-paired, m, n large	$\frac{X - Y - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \mathrm{N}(0,1)$	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	$\frac{X - Y - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \mathrm{N}(0,1)$	unrealistic
non-paired, m, n not large, data normal, $\sigma_1 \approx \sigma_2$ but unknown	$\frac{X - Y - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) s_p^2}} \sim t_{m+n-2}$	$s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$
binomial data, m, n large, p_1, p_2 moderate	$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/m + p_2(1-p_2)/n}} \sim \mathrm{normal}(0,1)$	replace p's with estimates in denominator