LECTURE - 12

Agenda:

1. Two population problem : Known variance
2. Two population problem: Unknown variance

A very standard problem a statistician is asked to solve is the following. Suppose we have two populations

Population 1: Mean $= \mu_1$, Variance $= \sigma_1^2$.
Population 2: Mean $= \mu_2$, Variance $= \sigma_2^2$.

Let us assume that σ_1^2 and σ_2^2 are known. Suppose we have a random sample $X_1, X_2, ---, X_{n_1}$ from Population 1 and a random sample $Y_1, Y_2, ---, Y_{n_2}$ from Population 2.

GOAL1: Estimate the differente between the population means, i.e, estimate $\mu_1 - \mu_2$.

For example, this situation arises when we are trying to compare between the performance of two drugs.

We have already seen that the standard unbiased estimate for $\mu_1 - \mu_2$ is given by $\bar{X} - \bar{Y}$ $\left(\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \ \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i \right)$

GOAL 2: Suppose both the populations are normal.
Provide a confidence interval for $\mu_1 - \mu_2$.

(1) Note that we have to construct a pivotal quantity in order to obtain a confidence interval.

$$\boxed{\text{RESULT: } \quad \overline{X} - \overline{Y} \text{ is Normal}\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right).}$$

Hence, $\dfrac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ is Normal(0, 1).

(2) Find a and b such that

$$P\left(a \le \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le b \right) = 1 - \alpha$$

We have seen that if $a = -z_{1-\frac{\alpha}{2}}$ and $b = z_{1-\frac{\alpha}{2}}$,

then, the requirement is satisfied.

RECALL THAT $z_{1-\alpha}$ is defined as the $(1-\alpha)^{th}$ quantile of the Normal(0, 1) distribution.

The notation we are using is different than the one in the textbook. They use z_α to denote the $(1-\alpha)^{th}$ quantile of the Normal $(0,1)$ distribution.

(3)
$$-z_{1-\frac{\alpha}{2}} \le \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \le z_{1-\frac{\alpha}{2}}$$

$$\Leftrightarrow (\bar{X}-\bar{Y})-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le (\mu_1-\mu_2) \le (\bar{X}-\bar{Y})+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Hence, a confidence interval for $(\mu_1-\mu_2)$ with confidence level $1-\alpha$ is given by

$$\left[(\bar{X}-\bar{Y})-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ (\bar{X}-\bar{Y})+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right].$$

WHAT IF σ_1 and σ_2 ARE UNKNOWN?

In this case, we make a simplifying assumption that $\sigma_1^2=\sigma_2^2=\sigma^2$. Let S_1^2 and S_2^2 be the adjusted sample variances for the two populations. Then

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}$$

is known as the POOLED SAMPLE VARIANCE.

Note that

$$E[S_p^2] = E\left[\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}\right]$$

$$= \frac{(n_1-1)\,E[S_1^2] + (n_2-1)\,E[S_2^2]}{(n_1+n_2-2)}$$

$$= \frac{(n_1-1)\,\sigma_1^2 + (n_2-1)\,\sigma_2^2}{(n_1+n_2-2)}$$

$$= \sigma^2.$$

Hence S_p^2 is an unbiased estimator for σ^2.

RESULT: $\dfrac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ has a t-distribution with n_1+n_2-2 degrees of freedom.

Hence $\dfrac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ can be used as a pivotal quantity.

Using the standard techniques, a confidence interval for $\mu_1 - \mu_2$ is given by

$$\left[(\bar{X}-\bar{Y}) - t_{n_1+n_2-2,\,1-\frac{\alpha}{2}}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ (\bar{X}-\bar{Y}) + t_{n_1+n_2-2,\,1-\frac{\alpha}{2}}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

where $t_{n_1+n_2-2,\,1-\alpha}$ is the notation for the $(1-\alpha)^{th}$ quantile of the t-distribution with n_1+n_2-2 degrees of freedom.