Lecture 35

Today we begin the topic of inference for two samples. As with one sample problems we will use confidence intervals and hypothesis tests as our primary tool.

Basic situation

Population 1

mean μ_1

variance σ_1^2

SRS of size m

$X_1, \ldots, X_m$

sample mean $\bar{X} = \dfrac{\sum X_i}{m}$

sample standard deviation $s_X = \sqrt{\dfrac{\sum (X_i - \bar{X})^2}{m-1}}$

Population 2

mean μ_2

variance σ_2^2

SRS of size n

$Y_1, \ldots, Y_n$

sample mean $\bar{Y} = \dfrac{\sum Y_i}{n}$

$s_Y = \sqrt{\dfrac{\sum (Y_i - \bar{Y})^2}{n-1}}$

Idea: use sample data $\bar{X}, \bar{Y}, s_X, s_Y$ to make inferences about μ_1, μ_2 (or σ_1^2, σ_2^2) and possible differences between them.

Typically we examine (or want to know about)

$$\mu_1 - \mu_2$$

the difference between the two means. Note that if μ_1 and μ_2 are equal ie $\mu_1 = \mu_2$ then $\mu_1 - \mu_2 = 0$.

Note that

$\bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$

and ie $E[\bar{X} - \bar{Y}] = \mu_1 - \mu_2$

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\Rightarrow \sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If we assume that both populations are large then by the CLT $\bar{X}$ and $\bar{Y}$ should have approximately normal distributions. This means $\bar{X}-\bar{Y}$ should also have approximately normal distribution. Therefore

$$Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$$ is a random variable with approx $N(0,1)$ distribution.

Confidence intervals

Following similar derivations to what has been used in the past it can be shown that a $100(1-\alpha)\%$ CI for $\mu_1-\mu_2$ is given by

$$\bar{x}-\bar{y} \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_1}}$$

when σ_1^2, σ_2^2 are known. (Note exact when populations are both normal, approximate otherwise)

If both ~~population~~ sample sizes n and m are large then it is reasonable to use s_1^2 and s_2^2 in place of σ_1^2 and σ_2^2 (if unknown) in the prior formula.

It can be shown that if σ_1^2 and σ_2^2 are at least approximately known and equal sized samples are to be taken then to get a CI of width w take

$$m = n = \frac{4 z_{\alpha/2}^2 (\sigma_1^2 + \sigma_2^2)}{w^2}$$

round up to the nearest whole integer.

Hypothesis Testing

Assume that the true difference between μ_1 and μ_2 is Δ_0 (some constant value). Often $\Delta_0 = 0$

This means that

$$Z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ will have $N(0,1)$

distribution if both populations are normally distributed

using these assumptions leads us to ...

$\boxed{\text{See next lecture}}$