0.1 Difference Of Two Means

In order to construct a confidence interval, we are going to make three assumptions:

- The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
- For the time being, we will use this assumption. Later on in the course, we will discuss the validity of this assumption for two given samples.
- The populations are normally distributed.
- Each value is sampled independently from each other value.

0.2 Computing the Confidence Interval

- As always the first step is to compute the point estimate. For the difference of means for groups X and Y, the point estimate is simply the difference between the two means i.e. $\bar{x} - \bar{y}$.
- As we have seen previously, sample size has a bearing in computing both the quantile and the standard error. For two groups, we will use the aggregate sample size $(n_x + n_y)$ to compute the quantile. (For the time being we will assume, the aggregate sample size is large $(n_x + n_y) > 30$.)
- Lastly we must compute the standard error $S.E.(\bar{x} - \bar{y})$. The formula for computing standard error for the difference of two means, depends on whether or not the aggregate sample size is large or not. For the case that the sample size is large, we use the following formula (next slide).

0.3 Computing the Confidence Interval

Standard Error for difference of two means (large sample)

$$S.E.(\bar{x} - \bar{y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

- s_x^2 and s_x^2 is the variance of samples X and Y respectively.
- n_x and n_y is the sample size of both samples.

- For small samples, the degrees of freedom is $df = n_x + n_y - 2$. If the sample size $n \leq 32$, we can find appropriate $t-$quantile, rather than assuming it is a $z-$quantile.

0.4 CI for Difference in Two Means

A research company is comparing computers from two different companies, X-Cel and Yellow, on the basis of energy consumption per hour. Given the following data, compute a 95% confidence interval for the difference in energy consumption.

Type	sample size	mean	variance
X-cel	17	5.353	2.743
Yellow	17	3.882	2.985

Remark: It is reasonable to believe that the variances of both groups is the same. Be mindful of this.

- Point estimate : $\bar{x} - \bar{y} = 1.469$
- Standard Error: 0.5805

$$S.E.(\bar{x} - \bar{y}) = \sqrt{\frac{2.743}{17} + \frac{2.985}{17}} = \sqrt{0.33698}$$

- Quantile : 1.96 (Large sample, with confidence level of 95%.)

$$1.469 \pm (1.96 \times 0.5805) = (0.3321, 2.607)$$

This analysis provides evidence that the mean consumption level per hour for X-cel is higher than the mean consumption level per hour for Yellow, and that the difference between means in the population is likely to be between 0.332 and 2.607 units.

0.5 Computing the Confidence Interval

Standard Error for difference of two means (small aggregate sample)

$$S.E.(\bar{x} - \bar{y}) = \sqrt{s_p^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}$$

Pooled Variance s_p^2 is computed as:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}$$

0.6 CI for Difference in Two Means

From the previous example (comparing X-cel and Yellow) lets compute a 95% confidence interval when the sample sizes are $n_x = 10$ and $n_y = 12$ respectively. (Lets assume the other values remain as they are.)

Type	sample size	mean	variance
X-cel	10	5.353	2.743
Yellow	12	3.882	2.985

The point estimate $\bar{x} - \bar{y}$ remains as 1.469. Also we require that both samples have equal variance. As both X and Y have variances at a similar level, we will assume equal variance.

0.7 Computing the Confidence Interval

- Pooled variance s_p^2 is computed as:

$$s_p^2 = \frac{(10-1)2.743 + (12-1)2.985}{(10-1)+(12-1)} = \frac{57.52}{20} = 2.87$$

- Standard error for difference of two means is therefore

$$S.E.(\bar{x}-\bar{y}) = \sqrt{2.87\left(\frac{1}{10}+\frac{1}{12}\right)} = 0.726$$

- The aggregate sample size is small i.e. 22. The degrees of freedom is $n_x + n_y - 2 = 20$. From Murdoch Barnes tables 7, the quantile for a 95% confidence interval is 2.086.

- The confidence interval is therefore

$$1.469 \pm (2.086 \times 0.726) = 1.4699 \pm 1.514 = (-0.044, 2.984)$$

0.7.1 Difference in Two means

For this calculation, we will assume that the variances in each of the two populations are equal. This assumption is called the assumption of homogeneity of variance.

The first step is to compute the estimate of the standard error of the difference between means ().

$$S.E.(\bar{X}-\bar{Y}) = \sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

- s_x^2 and s_x^2 is the variance of both samples.

- n_x and n_y is the sample size of both samples.

The degrees of freedom is $n_x + n_y - 2$.

0.7.2 Confidence Interval for a Mean (Small Sample)

- The mean operating life for a random sample of $n = 10$ light bulbs is $\bar{x} = 4,000$ hours, with the sample standard deviation $s = 200$ hours.

- The operating life of bulbs in general is assumed to be approximately normally distributed.

- We estimate the mean operating life for the population of bulbs from which this sample was taken, using a 95 percent confidence interval as follows:

$$4,000 \pm (2.262)(63.3) = (3857, 4143)$$

- The point estimate is 4,000 hours. The sample standard deviation is 200 hours, and the sample size is 10. Hence

$$S.E(\bar{x}) = \frac{200}{\sqrt{10}} = 63.3$$

- From last slide, the t quantile with $df = 9$ is 2.262.