INTRODUCTION TO BIOSTATISTICS

1 Confidence Intervals of the mean

Often in statistical inference it is very important to look into a narrow range of values, that an unknown population parameter may take. Most parameters can take possibly an infinite number of values, we will only be partially *confident* about the chance that the "true" value of this parameter in the population actually falls within this interval. Thus, every confidence interval we propose will have a degree of certainty associated with it.

1.1 Confidence interval of a single mean

Similarly to the case of one-sample hypothesis testing, we may be involved in estimating the mean of a single population.

1.1.1 Confidence Intervals with σ known

In the extraordinary case where the population standard deviation σ_x is known, the following intervals can be calculated (based on the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and its associated standard deviation $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$). Note that z_α is the right $\alpha\%$ tail of the standard normal distribution.

- **Two-sided** $(1-\alpha)\%$ **confidence intervals**

$$\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{1}$$

- **One-sided** $\alpha\%$ **confidence intervals**
  If we are only interested in a "cut-off" value, and the confidence is associated with the true mean being above (below) this value, we have what are called **upper** (**lower**) confidence intervals.

  - Upper $\alpha\%$ confidence interval

$$\mu \le \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \tag{2}$$

  - Lower $\alpha\%$ confidence interval

$$\mu \ge \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \tag{3}$$

1.1.2 Confidence Intervals with σ unknown

Usually the population standard deviation will be unknown. The estimate we use to "guess" its value is, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. The confidence intervals derived below are based on the t distribution with $n-1$ degrees of freedom rather than the standard normal distribution. Note that $t_{n-1;\alpha}$ is the right $\alpha\%$ tail of a t distribution with $n-1$ degrees of freedom.

- **Two-sided $\alpha\%$ confidence intervals**

$$\left(\bar{x} - t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) \tag{4}$$

- **One-sided $\alpha\%$ confidence intervals**

  – Upper $\alpha\%$ confidence interval

$$\mu \leq \bar{x} + t_{n-1;\alpha}\frac{s}{\sqrt{n}} \tag{5}$$

  – Lower $\alpha\%$ confidence interval

$$\mu \geq \bar{x} - t_{n-1;\alpha}\frac{s}{\sqrt{n}} \tag{6}$$

1.2 Confidence intervals of the *difference* between two means

There are several cases where the difference between two means will be of interest. These cases correspond to the two-sample (independent) t test. We will only consider the case where the two populations have equal (but unknown) variances, with a pooled common estimate $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

- **Two-sided $\alpha\%$ confidence intervals**

$$\left((\bar{x_1} - \bar{x_2}) - t_{n_1+n_2-2;\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{x_1} - \bar{x_2}) + t_{n_1+n_2-2;\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \tag{7}$$

- **One-sided $\alpha\%$ confidence intervals**

  – Upper $\alpha\%$ confidence interval

$$\mu \leq (\bar{x_1} - \bar{x_2}) + t_{n_1+n_2-2;\alpha}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{8}$$

  – Lower $\alpha\%$ confidence interval

$$\mu \geq (\bar{x_1} - \bar{x_2}) - t_{n_1+n_2-2;\alpha}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{9}$$

2 Confidence intervals for the population proportion

2.1 Confidence intervals for single proportions

When attempting to guess the proportion of "successes" in a single population, we use as our estimate the sample proportion $\hat{p} = \frac{x}{n}$ of the number of successes x out of n items sampled from the population. Since the population proportion p is unknown, so is the population standard deviation $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$. We use as an estimate, $\hat{\sigma}_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The confidence intervals for the population proportion are as follows:

- **Two-sided α% confidence intervals**

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \tag{10}$$

- **One-sided α% confidence intervals**

  - Upper α% confidence interval

$$p \le \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{11}$$

  - Lower α% confidence interval

$$p \ge \hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{12}$$

2.2 Confidence intervals of the *difference* of two proportions

Similarly to the two-sample hypothesis testing for proportions, we are frequently interested in calculating a confidence interval of the difference between two proportions. Our thinking proceeds similarly with that of the hypothesis testing, with the important difference that we do not need to assume that the two population proportions p_1 and p_2 are equal. This impacts on the estimate of the standard deviation which is $\hat{s} = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$. The confidence intervals for the population proportion are as follows:

- **Two-sided α% confidence intervals**

$$\left((\hat{p_1} - \hat{p_2}) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}\,(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}\,(1-\hat{p_2})}{n_2}},\ (\hat{p_1} - \hat{p_2}) + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}\,(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}\,(1-\hat{p_2})}{n_2}}\right) \tag{13}$$

- **One-sided α% confidence intervals**

  - Upper α% confidence interval

$$(p_1 - p_2) \le (\hat{p_1} - \hat{p_2}) + z_{\alpha}\sqrt{\frac{\hat{p_1}\,(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}\,(1-\hat{p_2})}{n_2}} \tag{14}$$

  - Lower α% confidence interval

$$(p_1 - p_2) \ge (\hat{p_1} - \hat{p_2}) - z_{\alpha}\sqrt{\frac{\hat{p_1}\,(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}\,(1-\hat{p_2})}{n_2}} \tag{15}$$