Lecture 25, Mar 15, 2023

Tolerance Intervals

Definition

Given n identically and independently distributed observations $X_1, \cdots, X_n$, the tolerance interval is defined as

$$[\bar{x} - ks, \bar{x} + ks]$$

with k chosen such that a fraction $1 - \alpha$ of the population is within the interval; that is,

$$\lim_{n \to \infty} P(-kS \leq X \leq kS) = 1 - \alpha$$

where X is an observation of the population, $\bar{X}$ is the sample mean and S^2 is the sample variance

- The tolerance interval n doesn't shrink with n
- Values of k can be obtained from tables

Summary

3 types of intervals:

1. Confidence intervals: $1 - \alpha$ chance of the true mean μ being in this interval around $\bar{x}$; with increasing n, this shrinks to 0, because $\bar{x}$ approaches the true mean
2. Prediction intervals: $1 - \alpha$ chance of the next observation x_0 being in this interval around $\bar{x}$; with increasing n, the interval shrinks to a fixed value $\bar{x} \pm z_{\alpha/2}\sigma$
3. Tolerance limits: For large n, fraction $1 - \alpha$ of all measurements will be in this interval around $\bar{x}$; k does not change with n, but as n increases the relationship becomes more precise

Two Samples, Known Variance

- Consider 2 samples with sizes n_1, n_2, each having mean μ_1, μ_2 and known variances σ_1^2, σ_2^2
- $\bar{X}_1 - \bar{X}_2$ is normal with mean $\mu_1 - \mu_2$, variance $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$ since $\bar{X}_1, \bar{X}_2$ are normal by the CLT
- Then $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ follows the standard normal by CLT

Two Samples, Unknown Variance

- First consider the case where the variances are equal but unknown
  - If $n_1, n_2 > 30$ then we can use s_1, s_2 as estimates of σ_1, σ_2 and use the CLT as usual
  - If $n_1, n_2 < 30$, we have to use the t-distribution and assume normal population
  - Use variance $S_p^2 = \dfrac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$
    * This pooled variance estimate is a sample size-weighted average of the two individual sample variances
  - Let $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
  - T has a t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom
- If $\sigma_1 \neq \sigma_2$ and both are unknown
  - Let $T' = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
  - T' approximately has a t-distribution

- $v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$
- If this is not an integer, round it down to the nearest integer (Satterthwaite approximation)