Lectures 01/27/20 and 01/29/20

Dr. Syring

Confidence Intervals

Suppose we have iid data $X_1, \ldots, X_n$ from a distribution P . Suppose θ is a parameter of P . A $100(1-\alpha)\%$ confidence interval for θ is the interval (a, b) , composed of endpoints a and b such that

$$P(a \leq \theta \leq b) \geq 1 - \alpha.$$

If this "coverage probability" is onlay approximate

$$P(a \leq \theta \leq b) \approx 1 - \alpha$$

then we call (a, b) an approximate CI for θ .

Example with derivation

Suppose $X_1, \ldots, X_n$ iid $N(\theta, 1)$. By Theorem 3.5.2 we know that $\bar{X} \sim N(\theta, \sigma^2 = 1/n)$, and that $Z = \frac{\bar{X} - \theta}{\sqrt{\frac{1}{n}}} \sim N(0, 1)$.

Then,

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

where z_α denotes the $100\alpha\%$ quantile of the standard normal distribution, e.g. $\Phi(z_\alpha) = \alpha$.

Using a little algebra and the symmetry of the normal distribution,

$$1 - \alpha = P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2})$$

$$= P\left(z_{\alpha/2} \leq \frac{\bar{X} - \theta}{\sqrt{\frac{1}{n}}} \leq z_{1-\alpha/2} \right)$$

$$= P\left(z_{\alpha/2}\sqrt{\frac{1}{n}} \leq \bar{X} - \theta \leq z_{1-\alpha/2}\sqrt{\frac{1}{n}} \right)$$

$$= P\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{1}{n}} \geq \theta \geq \bar{X} - z_{1-\alpha/2}\sqrt{\frac{1}{n}}\right)$$

$$= P\left(\bar{X} - z_{1-\alpha/2}\sqrt{\frac{1}{n}} \leq \theta \leq \bar{X} + z_{1-\alpha/2}\sqrt{\frac{1}{n}}\right)$$

Other examples

1. t-interval for one sample: Suppose $X_1, \ldots, X_n$ iid $N(\theta, \sigma^2)$. Then, by Student's Theorem $T = \frac{\bar{X}-\theta}{s/\sqrt{n}} \sim t(n-1)$ and by similar algebra/symmetry

$$1-\alpha = P\left(\bar{X} - t^{n-1}_{1-\alpha/2}\sqrt{\frac{s^2}{n}} \le \theta \le \bar{X} + t^{n-1}_{1-\alpha/2}\sqrt{\frac{s^2}{n}}\right).$$

2. approximate z-, t-intervals for one sample mean: If $X_1, \ldots, X_n$ iid from a distribution with a finite mean and variance then the CLT implies that for moderate n

$$1-\alpha \approx P\left(\bar{X} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \le \theta \le \bar{X} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

and

$$1-\alpha \approx P\left(\bar{X} - t^{n-1}_{1-\alpha/2}\sqrt{\frac{s^2}{n}} \le \theta \le \bar{X} + t^{n-1}_{1-\alpha/2}\sqrt{\frac{s^2}{n}}\right).$$

For large n , convergence of the t-distribution MGF implies

$$1-\alpha \approx P\left(\bar{X}-z_{1-\alpha/2}\sqrt{\frac{s^2}{n}} \le \theta \le \bar{X}+z_{1-\alpha/2}\sqrt{\frac{s^2}{n}}\right).$$

3. approximate z-intervals for population proportion: If $X_1, \ldots, X_n$ iid Bernoulli(p) then the CLT implies

$$1-\alpha \approx P\left(\hat{p}-z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \theta \le \hat{p}+z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

4. Two-sample CI for difference of means: Suppose $X_1, \ldots, X_n$ iid $N(\theta_X, \sigma^2)$, and $Y_1, \ldots, Y_m$ iid $N(\theta_Y, \sigma^2)$. Then, by Student's Theorem $T = \frac{\bar{X}-\bar{Y}-(\theta_X-\theta_Y)}{s_P\sqrt{1/n+1/m}} \sim t(n+m-2)$ where s_P is the pooled sample standard deviation, $S_P^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$. Then,

$$1-\alpha = P\left(\bar{X}-\bar{Y}-t_{1-\alpha/2}^{n+m-2}s_P\sqrt{1/n+1/m} \le \theta_X-\theta_Y \le \bar{X}-\bar{Y}+t_{1-\alpha/2}^{n+m-2}s_P\sqrt{1/n+1/m}\right).$$

5. One-sample interval for a population variance: Suppose $X_1, \ldots, X_n$ iid $N(\theta, \sigma^2)$. Then,

$$1-\alpha = P\left(\frac{(n-1)s^2}{\chi^{2,n-1}_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^{2,n-1}_{\alpha/2}}\right).$$

Derivation of Two-sample CI for difference of means, Pooled

Suppose $X_1, \ldots, X_n$ iid $N(\theta_X, \sigma^2)$, and $Y_1, \ldots, Y_m$ iid $N(\theta_Y, \sigma^2)$.

Then, by Theorem 3.5.2 (or Student's Theorem)

$$\frac{\bar{X} - \bar{Y} - (\theta_X - \theta_Y)}{\sigma\sqrt{1/n + 1/m}} \sim N(0, 1).$$

And, By Corollary 3.3.1 (or Student's Theorem)

$$\frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(n + m - 2).$$

Further, these two r.v.'s are independent by Student's Theorem. So, the ratio

$$\left[\frac{\bar{X} - \bar{Y} - (\theta_X - \theta_Y)}{\sigma\sqrt{1/n + 1/m}}\right] \Big/ \left[\sqrt{\left(\frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}\right)/n + m - 2}\right] \sim t(n + m - 2)$$

Further, this fraction can be algebraically simplified to

$$T = \frac{\bar{X} - \bar{Y} - (\theta_X - \theta_Y)}{s_P\sqrt{1/n + 1/m}}.$$

Then, to compute the coverage probability (derive the CI) we have

$$1-\alpha = P(t_{\alpha/2}(n+m-2) < T < t_{1-\alpha/2}(n+m-2))$$

$$= P(t_{\alpha/2}(n+m-2) < \frac{\bar{X}-\bar{Y}-(\theta_X-\theta_Y)}{s_P\sqrt{1/n+1/m}} < t_{1-\alpha/2}(n+m-2))$$

$$= P(t_{\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m} < \bar{X}-\bar{Y}-(\theta_X-\theta_Y) < t_{1-\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m})$$

$$= P(-(\bar{X}-\bar{Y})+t_{\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m} < -(\theta_X-\theta_Y) < -(\bar{X}-\bar{Y})+t_{1-\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m})$$

$$= P(\bar{X}-\bar{Y}-t_{\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m} > \theta_X-\theta_Y > \bar{X}-\bar{Y}-t_{1-\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m})$$

$$= P(\bar{X}-\bar{Y}+t_{\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m} < \theta_X-\theta_Y < \bar{X}-\bar{Y}+t_{1-\alpha/2}(n+m-2)s_P\sqrt{1/n+1/m})$$

where the last step follows by symmetry of the t distribution.

Derivation of two-sample CI for ratio of variances

Suppose $X_1, \ldots, X_n$ iid $N(\theta_X, \sigma_X^2)$, and $Y_1, \ldots, Y_m$ iid $N(\theta_Y, \sigma_Y^2)$. By Student's Theorem $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1)$ and $\frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(m-1)$. Further, these two random variables are independent simply because they come from different populations. By the definition of an F random variable (see 3.6.2) we have

$$F := \left[\frac{(n-1)S_X^2}{(n-1)\sigma_X^2}\right] / \left[\frac{(m-1)S_Y^2}{(m-1)\sigma_Y^2}\right] = \frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2} \sim F(n-1, m-1).$$

Therefore,

$$1-\alpha = P(F_{\alpha/2}(n-1, m-1) < F < F_{1-\alpha/2}(n-1, m-1))$$

$$= P(F_{\alpha/2}(n-1, m-1) < \frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2} < F_{1-\alpha/2}(n-1, m-1))$$

$$= P(F_{\alpha/2}(n-1, m-1)S_Y^2/S_x^2 < \frac{\sigma_Y^2}{\sigma_X^2} < F_{1-\alpha/2}(n-1, m-1)S_Y^2/S_x^2)$$

or equivalently,

$$= P\left(S_X^2/[F_{1-\alpha/2}(n-1, m-1)S_Y^2] < \sigma_X^2/\sigma_Y^2 < S_X^2/[F_{\alpha/2}(n-1, m-1)S_Y^2]\right).$$

$$= P\left(S_X^2/[F_{1-\alpha/2}(n-1, m-1)S_Y^2] < \sigma_X^2/\sigma_Y^2 < S_X^2 F_{1-\alpha/2}(m-1, n-1)/S_Y^2\right).$$

where the last line is the most common form and uses the fact that $F_\alpha(u, v) = 1/F_{1-\alpha}(v, u)$.

Approximate CIs

If we say a CI is approximate we mean that the coverage probability is not exactly $1-\alpha$. For example,

If $X_1, \ldots, X_n$ iid Bernoulli(p) then the CLT implies

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \overset{\cdot}{\sim} N(0, 1).$$

So, for large n (like $n \geq 30$, or more precisely $np,\ n(1-p) \geq 8$) we have a good approximation:

$$1 - \alpha \approx P(z_{\alpha/2} < Z < z_{1-\alpha/2})$$

$$= P(\hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}).$$