Lecture 27, Mar 20, 2023

Paired Observations

- *Paired observations* are when we have 2 populations and 2 samples of the same size (1 from each sample); in this case we can take one sample from each population and pair them up
  - e.g. measuring a chemical reactor, taking measurements at the inlet and outlet at the same time; a medical experiment where we measure before and after for each person
- Let $(X_i, Y_i), i = 1, \cdots, n$ be the paired samples; we're interested in the difference $D_i = X_i - Y_i$
  - $\text{var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$
  - If we assume X and Y to be negatively correlated, we expect $\sigma_{XY} > 0$; this reduces the total variance
    - Compared to treating X and Y as independent and taking the difference of means, this gives lower variance
    - Due to lower variance this gives tighter confidence intervals

Note

The gain in quality of the confidence interval of pairing vs. not pairing will be the greatest when there is homogeneity within units (strong correlation between two observations in a pair) and large differences between different units.
Pairing effectively reduces the number of degrees of freedom, so it may actually be counterproductive if the reduction in variance is small.

Confidence Intervals for Binomial Distributions

- Consider n IID trials, with $P(Y_i = 1) = p, P(Y_i = 0) = 1 - p$ for $i = 1, \cdots, n$; $X = \sum_{i=1}^{n} Y_i$ is the number of 1s, giving the binomial PMF $b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$, $\mu = np, \sigma^2 = np(1-p)$
- Let $Z = \dfrac{X - np}{\sqrt{np(1-p)}}$ then by CLT as $n \to \infty$, the PDF of Z becomes $n(x; 0, 1)$
  - This is because X is the sum of a series of Bernoulli RVs so the CLT applies
- Can we estimate $p = P(Y_i = 1)$?
  - Use $\hat{P} = \dfrac{X}{n}$ as the estimator
  - $\mu_{\hat{P}} = E\left[\dfrac{X}{n}\right] = \dfrac{np}{n} = p$ so the estimator is unbiased
  - $\sigma_{\hat{P}}^2 = \text{var}\left(\dfrac{X}{n}\right) = \dfrac{1}{n^2}\text{var}(X) = \dfrac{1}{n^2}np(1-p) = \dfrac{p(1-p)}{n}$
  - Let $Z = \dfrac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$ and by CLT this approaches the standard normal
  - Confidence interval is more challenging because it's harder to isolate for p
- $1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$

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

  - We have 2 choices: solve quadratics for p to get an exact confidence interval, or if p is large, approximate $p = \hat{p} = \dfrac{x}{n}$

- $1-\alpha = P\left(-z_{\alpha/2} \le \frac{\hat{P}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le z_{\alpha/2}\right)$
  $$= P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$
- Here $\hat{p} = \frac{x}{n}$ is not a random variable, but $\hat{P}$ is
- This approximation relies on p not being too close to 0 or 1; as a heuristic, both $n\hat{p}$ and $n\hat{q}$ should be at least 5

* How big does n need to be to have $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \delta$?
  - Solving for n, we get $n > \frac{z_{\alpha/2}^2\hat{p}(1-\hat{p})}{\delta^2}$, but $\hat{p} = \frac{x}{n}$
    * If we can get a crude estimate of p, we can use that to first determine n
    * n should be rounded up
  - $\hat{p}(1-\hat{p})$ is bounded by $1/4$ since $\hat{p} \le 1$
  - We can have a safe lower bound by $n \ge \frac{z_{\alpha/2}^2}{4\delta^2}$