Section 3, 9/16/19

Normal Approximations

Given a set of measurements $X_1, \ldots, X_n$, we know $\mathbb{E}(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$. So for large $n < N$, the central limit theorem holds that

$$P\left(\frac{\bar{X}_n - \mu}{\sigma\sqrt{n}}\right) \to \Phi(z)$$

as $n \to \infty$ where Φ is the cdf of the standard normal. The limit only makes sense in the case of sampling with replacement, but as long as n is large but small relative to N, $\bar{X}$ is still approximately normally distributed.

Recall that $\sigma_{\bar{X}}$ converges to σ for n large but small relative to N. This lets us derives probabilistic bounds on the estimate of the mean and generate confidence intervals.

Ex. C

From the prior example C, we found for a sample of size 50 an estimated discharge proportion of $\hat{p} = 0.52$. Let the population proportion be 0.65 for a difference of 0.13. We wish to understand the probability of this difference occurring. We start by estimating the variance of our estimate as

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{1 - \frac{n-1}{N-1}} = 0.064$$

Then we can calculate the likelihood of this error, which is unfortunately "unlucky".

$$\begin{aligned} P(|p - \hat{p}| > 0.13) &= 1 - P(|p - \hat{p}| \leq 0.13) \\ &= 1 - P\left(\frac{|p - \hat{p}|}{\sigma_{\hat{p}}} \leq \frac{0.13}{\sigma_{\hat{p}}}\right) \\ &= 2[1 - \Phi(2.03)] = 0.4 \end{aligned}$$

Confidence Interval

The confidence interval of a population parameter θ is a random interval which contains θ with some probability $1 - \alpha$. This tells us the uncertainty of the estimate $\hat{\theta}$.

Let $z(\alpha)$ be a the z-score function whose value is such that for $Z \sim N(0, 1)$,

$$P(Z \leq z(\alpha)) = P(-z(\alpha) \leq Z) = 1 - \alpha$$

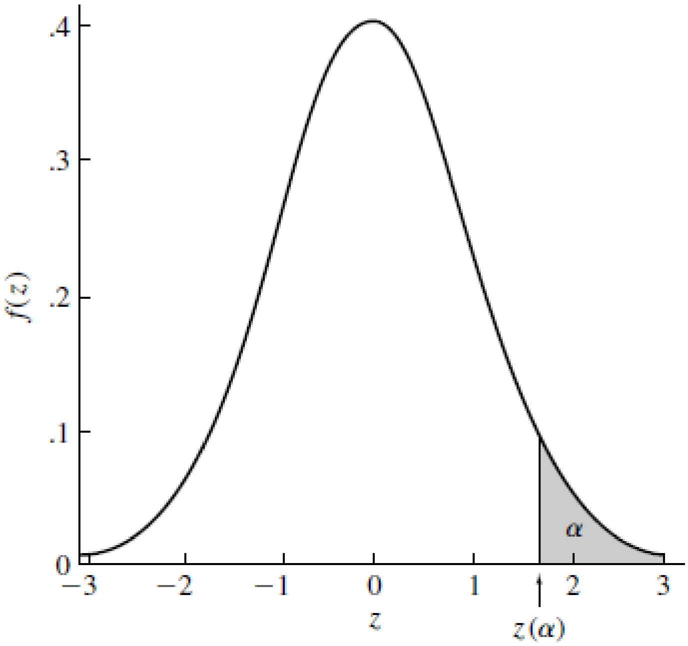

FIGURE 7.3 A standard normal density showing α and $z(\alpha)$.

So by the central limit theorem,

$$P\big(-z(\alpha/2) \le \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \le z(\alpha/2)\big) \approx 1-\alpha$$

Any thus it follows that the $\alpha-$**confidence interval** for the population parameter μ is

$$P\big(\bar{X} - \sigma_{\bar{X}} z(\alpha/2) \le \mu \le \bar{X} + \sigma_{\bar{X}} z(\alpha/2)\big) \approx 1-\alpha$$

Since $\sigma_{\bar{X}}$ is generally unknown, for large samples one can use the estimate $s_{\bar{X}}$.

Ratio Estimates

Suppose that for each member i of a population, two values x_i and y_i are measured. Thus these samples are matched, i.e. indexed by the same i as they correspond to the same individual. It may be that y is the acres of wheat planted and x is the total number of acres and we are interested in the percent of land used to plant wheat. Specifically, we are interested in the population **ratio**

$$r = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}$$

It is important to note that this is **not**

$$r \ne \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{x_i}$$

Our estimate of r is $R = \bar{Y}/\bar{X}$ but this is nonlinear and thus we cannot simply take its expected value and variance. So we use approximation techniques.

Approximation Methods

Step back and suppose we have some random variable X whose first and second moments are known. Suppose $Y = g(X)$, another random variable. If $g(X)$ is a linear function, then we would be able to calculate the first moments of Y. But if $g(X)$ is nonlinear, we need to find a linear approximation of $g(X)$ in the regions where X has high probability. We linearize using a Taylor Series expansion of g about μ_X.

Using a first order expansion,

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X)$$

By linearity of expectation,

$$\mu_Y \approx g(\mu_X)$$
$$\sigma_Y^2 \approx \sigma_X^2[g'(\mu_X)]^2$$

It doesn't make sense for $E[Y] = g(E[X])$ in general, a result of our approximation being too naive. Using a second order expansion instead,

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''(\mu_X)$$

And calculating the expectation get

$$\mu_Y \approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X)$$

The variance is not as trivial to compute.

For an nth order Taylor expansion centered about μ_X and evaluated at some x, there exists some $x^* \in [x, \mu_X]$ such that the error of the approximation is bounded by

$$R_n(x) = \frac{1}{(n+1)!}(x - \mu_x)^{n+1} f^{(n+1)}(x^*)$$