Stat 252 (Winter 2007)
Confidence Intervals for a Population Proportion

Theorem (Central Limit Theorem). *Suppose that $Y_1, Y_2, \ldots$ is a collection of independent, and identically distributed L^2 random variables with $\mathbb{E}(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$ for each i. For each n, let Z_n be the random variable defined by*

$$Z_n := \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \quad \textit{where} \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Then, for $z \in \mathbb{R}$, it follows that as $n \to \infty$, $P(Z_n \le z) \to \Phi(z) := \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt$. That is, $Z_n \to Z$ in distribution as $n \to \infty$ where $Z \sim \mathcal{N}(0,1)$.

Since the limiting distribution of the random variable Z_n is normal *no matter what the underlying distribution is*, we can argue that for a *large* sample size n, a normal approximation can be used. In fact, if $Y_1, \ldots, Y_n$ (with n large) are i.i.d. with a common (non-normal) distribution depending on a parameter θ, and $\hat{\theta}$ is an unbiased estimator of θ, then

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

where $\sigma_{\hat{\theta}}$ denotes the standard error of the estimator $\hat{\theta}$. We will use this approximation in much more generality later in the course when we discuss maximum likelihood estimation. For now, we will use this only for estimating a population proportion.

Estimating a Population Proportion

Suppose that we are interested in estimating a population proportion p. We collect a random sample from the population and let

$$Y_i = \begin{cases} 1, & \text{if } i\text{th individual has the characteristic of interest,} \\ 0, & \text{if not.} \end{cases}$$

That is, $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli(p) random variables. Since

$$\mathbb{E}(Y_i) = 1 \cdot P(Y_i = 1) + 0 \cdot P(Y_i = 0) = p,$$

we find that $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is an unbiased estimator of p. Furthermore,

$$\mathbb{E}(Y_i^2) = 1^2 \cdot P(Y_i = 1) + 0^2 \cdot P(Y_i = 0) = p,$$

so that $\mathrm{Var}(Y_i) = p - p^2 = p(1-p)$. Therefore,

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}.$$

In fact, much more can be said about the distribution of $\overline{Y}$. Using moment generating functions you showed in Stat 251 that if $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli(p) random variables, then

$$n\overline{Y} = \sum_{i=1}^{n} Y_i \sim \mathrm{Binomial}(n,p).$$

Remark 1. It is traditional when estimating a population proportion to use $\hat{p}$ as the notation for the estimator. That is, if $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli(p) random variables, then

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

satisfies $\mathbb{E}(\hat{p}) = p$ and $\sigma_{\hat{p}} := \sqrt{\mathrm{Var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$. Furthermore, $n\hat{p} \sim \mathrm{Binomial}(n,p)$. From this point, we will use the $\hat{p}$ notation when estimating population proportions.

Since the exact sampling distribution of $\hat{p}$ is known, it is possible to use the pivotal method from Lecture #12 to construct exact confidence intervals. However, it is extremely tedious to manipulate the summations of the binomial distribution. In fact, it is impossible *even for extremely fast computers* to calculate $n!$ for large n such as $n = 1\ 400\ 000$. (This is the actual sample sizes that are being considered by geneticist analyzing the human genome.)

Therefore, in order to construct confidence intervals for p we will use the approximation based on the Central Limit Theorem. That is,

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{approx}}{\sim} \mathcal{N}(0, 1).$$

The problem, of course, is that $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$ depends on the parameter of interest p. When we encountered this in Lecture #10 our solution was to replace the variance with the estimated variance. Therefore, we consider

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{\text{approx}}{\sim} t(n-1).$$

We are now able to find an *approximate* $1 - \alpha$ confidence interval for p based on $\hat{p}$ as follows:

$$1 - \alpha \approx P\left(-t_{\alpha/2,n-1} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq t_{\alpha/2,n-1}\right) = P\left(\hat{p} - t_{\alpha/2,n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + t_{\alpha/2,n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

Thus, the required approximate $1 - \alpha$ confidence interval is

$$\left[\hat{p} - t_{\alpha/2,n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\, \hat{p} + t_{\alpha/2,n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]. \qquad (*)$$

Remark 2. In first undergraduate courses (like Stat 151) it is more common to see the formula

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

where $z_{\alpha/2}$ is the critical value corresponding to the normal distribution. (That is, if $Z \sim \mathcal{N}(0, 1)$, then $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$.) However, there is no contradiction here with $(*)$. In order for the normal approximation to be valid, the sample size must be sufficiently large. For large values of n, the $t(n-1)$ distribution and the normal distribution are approximately equal, and the critical values $t_{\alpha/2,n-1}$ and $z_{\alpha/2}$ are equal to three or four decimal places. This can clearly be seen from Tables 5 and 6.

Remark 3. The conventional wisdom has historically been that this normal approximation is valid provided that $n \geq 30$. There has never been, however, documented research to justify this arbitrary value of 30. The original reason was simply that in the 1960s it was computationally impractical to compute values of the $t(n-1)$ statistic for $n \geq 30$. Therefore, accuracy to only three decimal places was considered acceptable. The advent of modern processors has rendered this choice of 30 obsolete since more complete t-tables are now available. For instance, a listing of t-values for $n = 1, \ldots, 100$ is available at

`http://davidmlane.com/hyperstat/t-table.html`

and a java calculator able to compute t-values accurate to four decimal places for arbitrarily large degrees of freedom may be found at

`http://statpages.org/pdfs.html`.

Furthermore, recently published research has demonstrated just how poor the z-approximation actually is. For these reasons, we will use $(*)$ as the approximate $1 - \alpha$ confidence interval for p in our calculations for this course.