LARGE-SAMPLE PIVOT QUANTITIES

So far we have only considered pivot quantities which arise when sampling from a Normal distribution.

Recall: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. Then

(i) $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0,1)$

$$\Rightarrow \quad \bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{is a } (1-\alpha)^* 100\% \text{ C.I. for } \mu.$$

(ii) $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

$$\Rightarrow \quad \left(\frac{(n-1)S_n^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2}} \right) \quad \text{is a } (1-\alpha)^* 100\% \text{ C.I. for } \sigma^2.$$

(iii) $\dfrac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$

$$\Rightarrow \quad \bar{X}_n \pm t_{\alpha/2} \frac{S_n}{\sqrt{n}} \quad \text{is a } (1-\alpha)^* 100\% \text{ C.I. for } \mu.$$

(iv) If $X_1, \ldots, X_{n_1} \overset{iid}{\sim} \text{Normal}(\mu_1, \sigma_1^2)$, $S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X}_n)^2$

$Y_1, \ldots, Y_{n_2} \overset{iid}{\sim} \text{Normal}(\mu_2, \sigma_2^2)$, $S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y}_n)^2$

then $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$.

$$\Rightarrow \left(\frac{S_2^2}{S_1^2} F_{n_1-1, n_2-1, 1-\alpha/2}, \frac{S_2^2}{S_1^2} F_{n_1-1, n_2-1, \alpha/2} \right) \text{ is a } (1-\alpha)^* 100\% \text{ C.I. for } \frac{\sigma_2^2}{\sigma_1^2}.$$

CRUCIAL: If the random samples do NOT come from a Normal distribution, NONE of the above holds.

Question: What if we draw a random sample from a non-normal distribution and wish to make a C.I. for the mean?

E.g. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, p unknown.

How do we build a $(1-\alpha)^* 100\%$ for p?

E.g. Let $X_1, \dots, X_n$ be a random sample from some right-skewed distribution with unknown mean μ.

How do we build a $(1-\alpha)^* 100\%$ for μ?

The main results in this lecture will be the following:

If we draw a random sample $X_1, \dots, X_n$ from a non-Normal distribution with mean μ and variance $\sigma^2 < \infty$, then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad \frac{\bar{X}_n - \mu}{\hat{\sigma}_n/\sqrt{n}},$$

where $\hat{\sigma}_n$ is a consistent estimator of σ, behave more and more like $\text{Normal}(0,1)$ random variables as $n \to \infty$.

This means that for large n,

$$\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{X}_n \pm z_{\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}}$$

are approximate $(1-\alpha)^* 100\%$ C.I.s for μ.

We begin by formalizing what it means for a random variable to "behave more and more like" another. The following definition concerns a sequence of random variables indexed by n, and we may think of the sequence of rvs as the random values of a function computed on the sample for each sample size n.

Defn: A sequence of rvs $Y_1, Y_2, \ldots$ with cdfs $F_{Y_1}, F_{Y_2}, \ldots$ is said to <u>converge in distribution</u> to the random variable $Y \sim F_Y$ if

$$\lim_{n \to \infty} F_{Y_n}(y) = F_Y(y)$$

for all $y \in \mathbb{R}$.

Remark: If Y_n converges in distribution to Y, we write $Y_n \xrightarrow{D} Y$.

We refer to the distribution with cdf F_Y as the <u>asymptotic distribution</u> of Y_n.

In words, if Y_n converges in distribution to Y, the cdf of Y_n approaches the cdf of Y at all values $y \in \mathbb{R}$ as n goes to infinity.

So convergence in distribution is a sense in which the random variables $Y_1, Y_2, \ldots$ can be said to behave more and more like another random variable Y.

Example of convergence in distribution:

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exponential}(\lambda)$ and let $Y_n = \frac{1}{\lambda}\left(X_{(n)} - \lambda \log n\right)$.

Moreover, let $Y \sim F_Y(y) = e^{-e^{-y}}$ for $y \in \mathbb{R}$.

Show that $Y_n \xrightarrow{D} Y$.

(i) Find the cdf of $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

We have

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda} \mathbb{1}(x > 0) \quad \text{and} \quad F_X(x) = \begin{cases} 1 - e^{-x/\lambda}, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

so that the cdf of $X_{(n)}$ is given by

$$F_{X_{(n)}}(x) = \begin{cases} \left(1 - e^{-x/\lambda}\right)^n & , \quad x > 0 \\ 0 & , \quad x \le 0 \end{cases}$$

(ii) Find the cdf of $Y_n = \frac{1}{\lambda}\left(X_{(n)} - \lambda \log n\right)$

First note that $Y_n \in (-\log n, \infty)$. For $y \in (-\log n, \infty)$,

$$\begin{aligned} F_{Y_n}(y) &= P_{Y_n}\left(Y_n \le y\right) \\ &= P_{X_{(n)}}\left(\frac{1}{\lambda}\left(X_{(n)} - \lambda \log n\right) \le y\right) \\ &= P_{X_{(n)}}\left(X_{(n)} \le \lambda\left(y + \log n\right)\right) \\ &= F_{X_{(n)}}\left(\lambda\left(y + \log n\right)\right) \\ &= \left[1 - e^{-\frac{\lambda(y+\log n)}{\lambda}}\right]^n \\ &= \left[1 - \frac{e^{-y}}{n}\right]^n . \end{aligned}$$

$e^{-\log n} = e^{\log n^{-1}} = n^{-1}$

So we have

$$F_{Y_n}(y) = \begin{cases} \left[1 - \frac{e^{-y}}{n}\right]^n & , \quad y > -\log n \\ 0 & , \quad y \le -\log n \end{cases} .$$

(iii) Find the limit of the cdf of $Y_n = \frac{1}{\lambda}\left(X_{(n)} - \lambda \log n\right)$ as $n \to \infty$.

We have $\lim_{n\to\infty} F_{Y_n}(y) = e^{-e^{-y}} \, \mathbb{1}(-\infty < y < \infty)$.

So $Y_n \xrightarrow{D} Y$.

The asymptotic dist. of Y_n is the standard Gumbel distribution:

Gumbel, E.J. (1935). Les valeurs extrêmes des distributions statistiques. Ann. Inst. Henri Poincaré, 5(2), 115-158.

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

et

$$\frac{d\mathrm{W}^{\mathrm{N}}(x)}{dx} = \mathrm{N}w(\bar{u}_m)e^{-y_m - e^{-y_m}}. \tag{30'}$$

Pour $m = 1$, on obtient d'après (12') la distribution finale de la plus grande valeur

$$\mathfrak{w}_1 = \mathrm{N}w(\bar{u}_1)e^{-y_1 - e^{-y_1}}. \tag{31'}$$

La probabilité pour que la dernière valeur soit inférieure à x sera

$$\mathfrak{W}_1 = e^{-e^{-y_1}}, \tag{32'}$$

formule déduite par R. A. Fisher (8).

Oui oui!

Note that the asymptotic distribution of Y_n is fully known; it does not depend on the unknown parameter λ. So we may refer to Y_n as a large-sample or asymptotic pivot quantity.

THE CENTRAL LIMIT THEOREM

The following theorem, called the central limit theorem (CLT), is a very important theorem in statistics. In very many statistics research papers, some version of the CLT is invoked.

Theorem (CLT): Let $X_1, \dots, X_n$ be a random sample from a dist. with mean μ and variance $\sigma^2 < \infty$ and for which the mgf $M_X(t)$ is defined for t in some neighborhood of zero. Let $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z, \quad \text{where } Z \sim \text{Normal}(0,1).$$

Proof: We show $\lim_{n\to\infty} M_{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}(t) = e^{t^2/2}$, where $e^{t^2/2}$ is the Normal(0,1) mgf.

First rewrite

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{1}{n}\left[\left(\frac{X_1-\mu}{\sigma}\right) + \dots + \left(\frac{X_n-\mu}{\sigma}\right)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i,$$

where $Y_i = \frac{X_i - \mu}{\sigma}$, $i = 1, \dots, n$.

Denote by M_Y the common mgf of $Y_1, \ldots, Y_n$.

If $M_X(t)$ is defined for all t such that $|t| < h$, for some $h > 0$,

$$M_Y(t) = M_{\frac{X-\mu}{\sigma}}(t) = e^{-\mu t} M_X\left(\tfrac{1}{\sigma} t\right)$$

is defined for all t such that $|t| < \sigma h$.

Now we have

$$M_{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}(t) = M_{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i}(t) = \prod_{i=1}^n M_{Y_i}\left(\tfrac{1}{\sqrt{n}} t\right) = \left[M_Y\left(\tfrac{1}{\sqrt{n}} t\right)\right]^n.$$

By Taylor expansion, we may write

$$M_Y\left(\tfrac{1}{\sqrt{n}} t\right) = M_Y(0) + M_Y^{(1)}(0)\left(\tfrac{1}{\sqrt{n}} t - 0\right) + \tfrac{1}{2} M_Y^{(2)}(0)\left(\tfrac{1}{\sqrt{n}} t - 0\right)^2 + R_Y\left(\tfrac{1}{\sqrt{n}} t\right),$$

where $M_Y^{(k)}(0) = \left(\frac{d}{dt}\right)^k M_Y(t)\Big|_{t=0}$

and $R_Y\left(\tfrac{1}{\sqrt{n}} t\right) = \sum_{k=3}^{\infty} \frac{M^{(k)}(0)}{k!}\left(\tfrac{1}{\sqrt{n}} t - 0\right)^k$

We know that

$$M_Y(0) = \mathbb{E} e^{Y_1 \cdot 0} = 1,$$

$$M_Y^{(1)}(0) = \mathbb{E} Y_1 = 0,$$

$$M_Y^{(2)}(0) = \mathbb{E} Y_1^2 = \operatorname{Var} Y_1 + (\mathbb{E} Y_1)^2 = 1.$$

$$\operatorname{Var} Y_1 = \operatorname{Var}\left[\frac{X_1 - \mu}{\sigma}\right] = \frac{1}{\sigma^2} \operatorname{Var} X_1 = \frac{1}{\sigma^2}\sigma^2 = 1$$

And, concerning the remainder term $R_Y\left(\tfrac{1}{\sqrt{n}} t\right)$, we have

$$\lim_{n \to \infty} \frac{R_Y\left(\tfrac{1}{\sqrt{n}} t\right)}{\left(\tfrac{1}{\sqrt{n}} t\right)^2} = 0$$

for any fixed $t \neq 0$, which gives

$$\lim_{n \to \infty} n R_Y\left(\tfrac{1}{\sqrt{n}} t\right) = 0.$$

$$\boxed{\begin{array}{l} \underline{\text{Taylor's Thm:}} \text{ If } f \text{ has derivatives of order } K, \text{ and} \\ T_K(x; x_0) = \sum_{k=1}^{K} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \\ \text{then} \\ \lim_{x \to x_0} \frac{f(x) - T_K(x; x_0)}{(x - x_0)^K} = 0. \end{array}}$$

So we get

$$\lim_{n\to\infty}\left[M_Y\left(\tfrac{1}{\sqrt{n}}t\right)\right]^n = \lim_{n\to\infty}\left[1 + \frac{1}{2}\frac{t^2}{n} + R_Y\left(\tfrac{1}{\sqrt{n}}t\right)\right]^n$$

$$= \lim_{n\to\infty}\left[1 + \frac{1}{n}\left(\frac{t^2}{2} + \underbrace{n R_Y\left(\tfrac{1}{\sqrt{n}}t\right)}_{\to 0 \text{ as } n\to\infty}\right)\right]^n,$$

$$= e^{t^2/2}.$$

Lemma: Let $a_1, a_2, \ldots$ be a sequence of numbers such that $\lim_{n\to\infty} a_n = a$. Then

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a$$

$\square$

Remark: We do not actually need the assumption that the mgf of the population distribution exists, but it simplifies the proof. Asymptotic Normality of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ can be established assuming only that $\sigma^2 < \infty$.

Application of CLT: Let $X_1, \ldots, X_n$ be a random sample from a distribution which is non-Normal and has mean μ and variance $\sigma^2 < \infty$. Then

$$\lim_{n\to\infty} P\left(-z_{\alpha/2} < \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < z_{\alpha/2}\right) = 1-\alpha,$$

$\Rightarrow$

$$\lim_{n\to\infty} P\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

so that for $\overbrace{\text{large}}^{\text{Rule of thumb: } n \geq 30 \text{ is "large"}}$ n,

$$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

is an approximate $(1-\alpha)^*100\%$ C.I. for μ.

What about $\dfrac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$, where S_n replaces the unknown σ?

Theorem: (Corollary of Slutzky's Thm) Let $X_1, \dots, X_n$ be a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$. Moreover, let $\hat{\sigma}_n$ be a consistent estimator of σ based on $X_1, \dots, X_n$. Then

$$\frac{(\bar{X}_n - \mu)}{\hat{\sigma}_n / \sqrt{n}} \xrightarrow{D} Z, \quad \text{where} \quad Z \sim \text{Normal}(0,1).$$

Remark: The sample variance S_n^2 is a consistent estimator of σ^2 if the 4^{th} moment μ_4 of the population distribution is finite. If S_n^2 is consistent for σ^2, then

$$S_n = \sqrt{S_n^2} \xrightarrow{P} \sigma = \sqrt{\sigma^2},$$

that is, S_n is consistent for σ, since $\sqrt{x}$ is a continuous function for $x > 0$. This gives

$$\left(\frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \right) \xrightarrow{D} Z, \quad Z \sim \text{Normal}(0,1)$$

if $\mu_4 < \infty$.

Application: Let $X_1, \dots, X_n$ be a random sample from a distribution which is non-Normal and has mean μ, variance $\sigma^2 < \infty$, and 4^{th} moment $\mu_4 < \infty$. Then

$$\lim_{n \to \infty} P\left(-z_{\alpha/2} < \frac{(\bar{X}_n - \mu)}{S_n / \sqrt{n}} < z_{\alpha/2} \right) = 1 - \alpha$$

$$\Rightarrow \lim_{n \to \infty} P\left(\bar{X}_n - z_{\alpha/2} \frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2} \frac{S_n}{\sqrt{n}} \right) = 1 - \alpha$$

so that for $\underbrace{\text{large}}_{\text{Rule of thumb: } n \geq 30 \text{ is "large"}}$ n,

$$\bar{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

is an approximate $(1-\alpha)*100\%$ C.I. for μ.

Application: Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, p unknown, and let $\hat{p}_n = \frac{1}{n}(X_1 + \dots + X_n)$.

We have $\mu = p$ and $\sigma^2 = p(1-p)$.

Recall that $\hat{p}_n(1-\hat{p}_n) \xrightarrow{p} p(1-p)$, so $\sqrt{\hat{p}_n(1-\hat{p}_n)} \xrightarrow{p} \sqrt{p(1-p)}$.

Therefore

$$\lim_{n\to\infty} P\left(-z_{\alpha/2} < \frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} < z_{\alpha/2}\right) = 1-\alpha$$

$\Rightarrow$

$$\lim_{n\to\infty} \left(\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} < p < \hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) = 1-\alpha,$$

so that for large n,

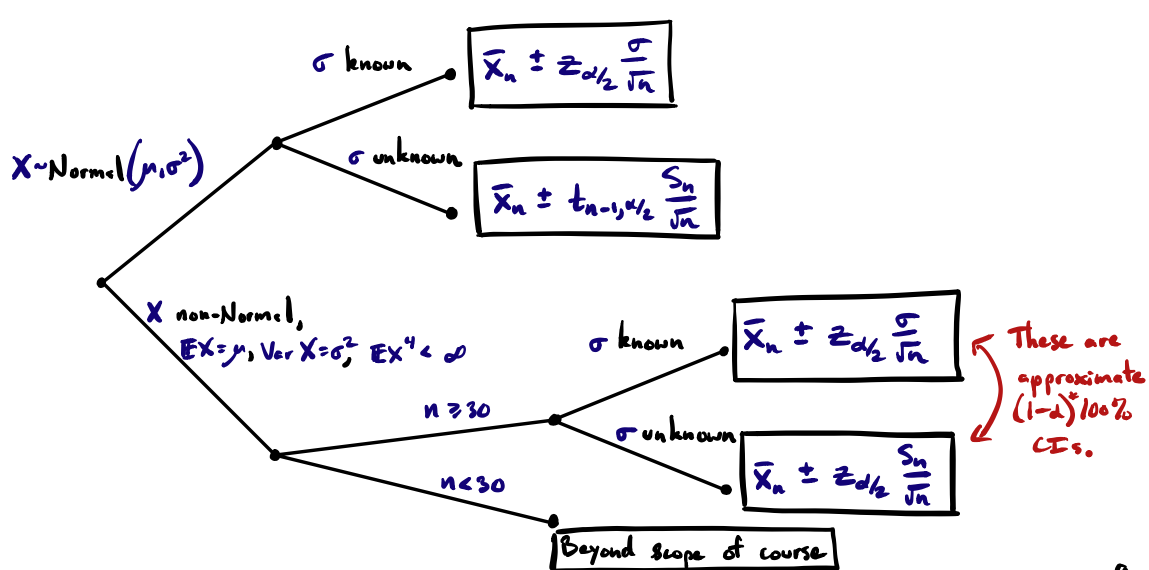
Rule of thumb: require $\min\{n\hat{p}_n, n(1-\hat{p}_n)\} \geq 15$

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

is an approximate $(1-\alpha)^* 100\%$ C.I. for p.

SUMMARY OF C.I.s FOR THE MEAN

Let $X_1, \dots, X_n$ be independent r.v.s with the same distribution as X.



ADDENDUM: SLUTZKY'S THEOREM

Theorem: If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} 1$, then

(i) $X_n Y_n \xrightarrow{D} X$

(ii) $X_n + Y_n - 1 \xrightarrow{D} X$