6/4/19 Lecture 19

Extra lecture webcasted tomorrow evening

Daily office hours will start on Thursday June 6 and will continue through Sunday June 16 (~7 at night)

Make sure you fill out the class evaluations because they truly are important!

Large Random Samples

You draw an IID random sample $X_1, \ldots, X_n$ from a population, with the goal of estimating the population mean $\mu = E(X_i)$

From a root mean squared error point of view, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the best you can do

(in the absence of prior information)

Markov Inequality

Suppose X is a nonnegative rv, i.e. $P(X \geq 0) = 1$

then for all real $t > 0$, $P(X \geq t) \leq \frac{E(X)}{t}$

Ex: $E(X) = 1$ X non-negative

$P(X \geq 100) \leq \frac{1}{100}$

Chebyshev Inequality

X rv with $V(X)$ existing

for every $t \geq 0$, $P[|X - E(X)| \geq t] \leq \frac{V(X)}{t^2}$

Ex: $E(X) = \mu$ $V(X) = \sigma^2$

$P\left[\left|\frac{X-\mu}{6}\right| \geq 3\right] \leq \frac{1}{3^2} = \frac{1}{9}$ so no more than 11% of the probability

in any distribution with finite variance can be more than 3 SDs away from the mean

This upper bound is sharp, but for most distributions it's (also crude) (as with the Markov bound)

Back to $\bar{X}_n$

$X_i \overset{IID}{\sim}$ some dist. with mean $E(X_i) = \mu$ $(i = 1, \ldots, n)$ and variance $V(X_i) = \sigma^2 < \infty$

We have already shown that if $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

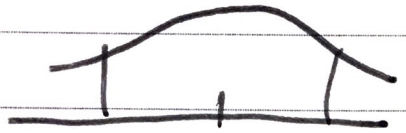
Then $E(\bar{X}_n) = \mu$ and $V(\bar{X}_n) = \frac{\sigma^2}{n}$ for all $n = 1, 2, \ldots$

Chebyshev then gives $P(|\bar{X}_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2}$ for all $t > 0$

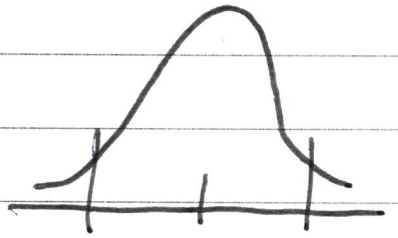
This can be rewritten $P(|\bar{X}_n - \mu| < t) \geq 1 - \frac{\sigma^2}{nt^2}$

This suggests a way to quantify how close a rv like $\bar{X}_n$ is to a constant like μ

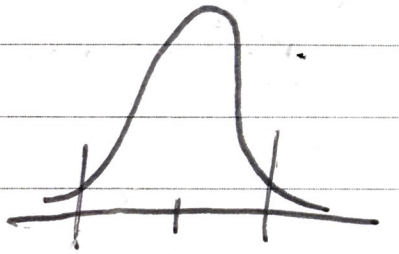
PDF of $\bar{X}_n$ $n=1$



PDF of $\bar{X}_n$ $n=10$



PDF of $\bar{X}_n$ $n=100$



Def: A sequence $Z_1, Z_2, \ldots$ of rv is said to converge in probability to a constant b, if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|Z_n - b| < \epsilon) = 1$$

This is denoted

$$Z_n \xrightarrow{P} b$$

Weak Law of Large Numbers

$\bar{X}_i$ iid a dist. with mean μ and variance $\sigma^2 < \infty$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} \bar{X}_i \quad \rightarrow \quad \bar{X}_n \xrightarrow{P} \mu$$

$\bar{X}_n$ is consistent for μ

Corollary

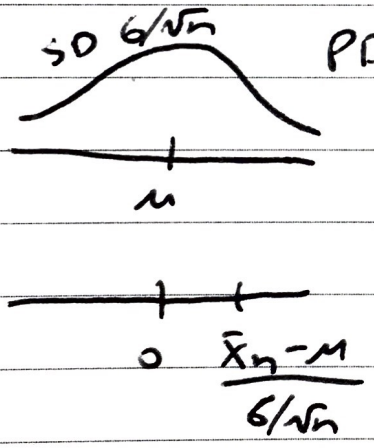
If $Z_n \xrightarrow{p} b$ and $g(z)$ is continuous at $z=b$ then $g(Z_n) \xrightarrow{p} g(b)$

Central Limit Theorem

Ex: $X_i \overset{IID}{\sim} N(\mu, \sigma^2)$, $\sigma < \infty$ $(i = 1, \ldots, n)$

We know that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has mean μ, variance $\frac{\sigma^2}{n}$ and is normally distributed, so that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ for all } n = 1, 2, \ldots$$



This works for other choices of $X_i \overset{IID}{\sim}$ [?]

CLT

$X_i \overset{IID}{\sim}$ any dist with mean μ and finite variance $0 < \sigma^2 < \infty$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow$ for large n, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$

Def:

Let F_n be the CDF of X_n if there exists a CDF F^* such that $\lim_{n\to\infty} F_n(x) = F^*(x)$ for all x at which $F^*(x)$ is continuous, then $X_n \xrightarrow{D} F^*$

CLT

$X_i \overset{IID}{\sim}$ any dist with mean μ and variance $0 < \sigma^2 < \infty$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$\rightarrow \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1)$$

Ex: Contaminated water supply:

X = arsenic concentration

Y = lead concentration (same units) (both ≥ 0)

Interest focuses on $R = \frac{Y}{X+Y}$

(proportion of contamination due to lead)

$E(R) = E\left(\frac{Y}{X+Y}\right)$ difficult to calculate

Simulation approach

Randomly sample n pairs (X_i, Y_i) from the joint PDF of (X, Y)

Calculate $R_i = \frac{X_i}{X_i + Y_i}$ and $\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$

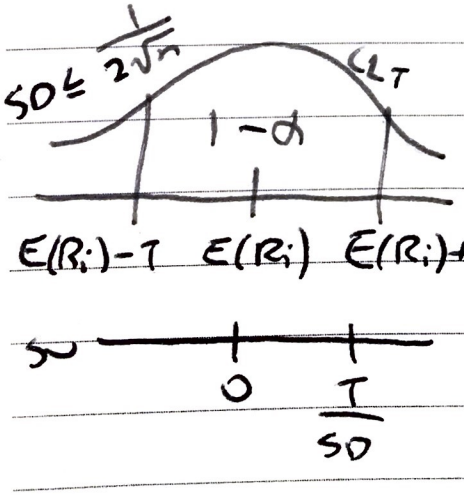
$\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$ ← good Monte Carlo (simulation) estimate of $E(R)$

Q: How big does n need to be to achieve desired accuracy target?

By definition $|R_i| = \left|\frac{Y_i}{X_i + Y_i}\right| \leq 1$

Can show that as a result $V(R_i) \leq \frac{1}{4}$

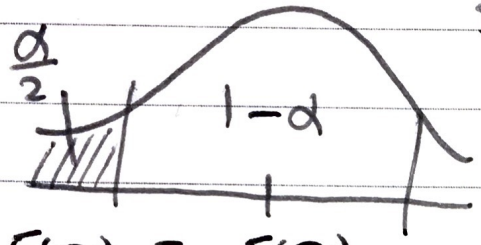
CLT says that dist. of $\bar{R}_n$ will be close to Normal for large n, with mean $E(R_i)$ and variance $\frac{V(R_i)}{n} \leq \frac{1}{4n}$



PDF of $\bar{R}_n$, n large

Suppose we want $\bar{R}_n$ to differ from $E(R_i)$ by no more than some tolerance T with probability at least $(1-\alpha)$

$SD \leq \frac{1}{2\sqrt{n}}$



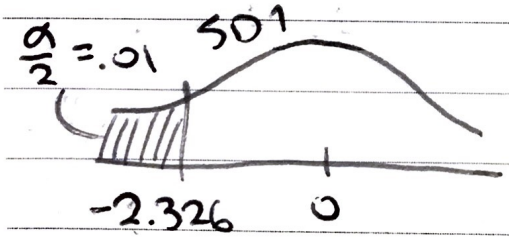
so $\frac{1}{SD} \geq 2\sqrt{n}$ and $\frac{-T}{SD} \leq 2T\sqrt{n}$

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) = \frac{[E(R_i)-T] - E(R_i)}{SD} = \frac{-T}{SD} \leq 2T\sqrt{n}$$

from which $n \geq \left[\frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{2T}\right]^2$

For instance, set $T = 0.005$ ($\frac{1}{2}$ of 1%) and $\alpha = .02$ to get

$$n \geq \left[\frac{-2.326}{2(.005)}\right]^2 = 54,119 \text{ simulation replications}$$



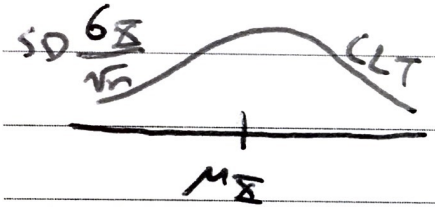
The Delta Method

The CLT says that if $X_i \overset{IID}{\sim}$ any dist. with finite mean μ_X and finite variance σ_X^2, then the distribution of $\dfrac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}}$ for large n is

approximately standard normal, where

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This is equivalent to saying that



PDF of $\bar{X}_n$ $\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$

Question:

If $g(x)$ is a sufficiently "nice" function, is there a comparable result for $g(\bar{X}_n)$?

Answer:

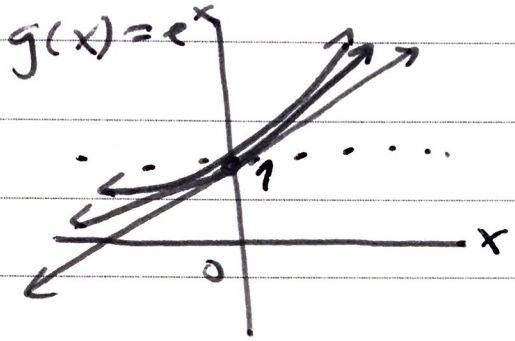
Yes, via a Taylor-series-based approach called the Delta Method

$\bar{X}_n$ should be close to μ_X for large n (that's the (weak) Law of Large Numbers)

This suggests making a two-term Taylor expansion of $g(\bar{X}_n)$ around the point $x = \mu_X$

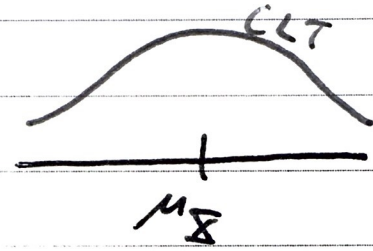
Expand $g(x)$ around $x = x_0$

$$g(x) \doteq g(x_0) + g'(x_0)(x - x_0)$$



$g(x) = 7$

$\bar{X}_n$ has mean μ_X



Expand $g(\bar{X}_n)$ around μ_X:

$$g(\bar{X}_n) \doteq g(\mu_X) + g'(\mu_X)(\bar{X}_n - \mu_X)$$

constant constant rv constant

$$E(g(\bar{X}_n)) \doteq E[g(\mu_X) + g'(\mu_X)(\bar{X}_n - \mu_X)]$$

$$= g(\mu_X) + g'(\mu_X)[\underbrace{E(\bar{X}_n) - \mu_X}_{0}]$$

so $E(g(\bar{X}_n)) \doteq g(\mu_X) = g(E(\bar{X}_n))$

$$V[g(\bar{X}_n)] \doteq V[g(\mu_X) + g'(\mu_X)(\bar{X}_n - \mu_X)]$$

$$= [g'(\mu_X)]^2 \cdot V(\bar{X}_n - \mu_X)$$

so $V[g(\bar{X}_n)] \doteq [g'(\mu_X)]^2 V(\bar{X}_n)$

$$V[g(\bar{X}_n)] \doteq [g'(\mu_X)]^2 \frac{\sigma_X^2}{n}$$

There's one hidden assumption in this calculation: $g'(\mu_X) \neq 0$

This works for any rv with finite variance, not just $\bar{X}_n$:

V any rv with finite variance σ_V^2 (and therefore finite mean μ_V), $W = g(V)$
$\rightarrow E(W) \doteq g(\mu_V)$ and $V(W) \doteq [g'(\mu_V)]^2 \sigma_V^2$
(Δ Method Part 1) provided $g'(v)$ is continuous and $g'(\mu_V) \neq 0$

Moreover, if $V \sim$ Normal then $W = g(V) \sim$ Norma
(Δ Method Part 2)

Ex: A bank typically has a single queue (line) at which customers arrive to transact banking business

Let $\bar{X}_i$ = time customer i waits from reaching the head of the queue until served.

To be completely realistic, the dist of $\bar{X}_i$ would vary by day of week and time of day, so pick

a single time slot (e.g. Tue 10-10:15am) and observe the X_i from week to week only in that time slot

Now the $\{X_i, i = 1, 2, \ldots\}$ form a stationary stochastic process with fixed (non-time-varying) finite $E(X_i) = \mu_X > 0$ and fixed (non-time-varying) finite $V(X_i) = \sigma_X^2$

Gather data over many weeks and form $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ for large n.

The rate of service is defined to be $g(\mu_X) = \frac{1}{\mu_X}$ which would naturally be estimated by

$$g(\bar{X}_n) = \frac{1}{\bar{X}_n}$$

$$E(\bar{X}_n) = \mu_X \qquad V(\bar{X}_n) = \frac{\sigma_X^2}{n}$$

$$g(x) = \frac{1}{x} = x^{-1} \qquad g'(x) = -\frac{1}{x^2}$$

$$g'(\mu_X) = -\frac{1}{\mu_X} \qquad \bar{X}_n \sim \text{Normal by CLT}$$

So Δ method says $g(\bar{X}_n) = \frac{1}{\bar{X}_n} \sim$ Normal

with mean $g(\mu_X) = \frac{1}{\mu_X}$ and variance $\sigma_X^2/(n\mu_X^4)$

$$\left[g'(\mu_{\bar{X}})\right]^2 = \frac{1}{\mu_{\bar{X}}^4} \neq 0$$

Specific calculation

Under some plausible assumptions, we've seen that $(\bar{X}_i \mid \lambda) \overset{IID}{\sim} \text{Exponential}(\lambda)$ may be a reasonable model for waiting times.

$$E(\bar{X}_i) = \frac{1}{\lambda}, \quad V(\bar{X}_i) = \frac{1}{\lambda^2}$$
$$= \mu_{\bar{X}} \qquad = \sigma^2_{\bar{X}}$$

$(\bar{X}_i \mid \lambda)$ has PDF $f_{\bar{X}_i}(x_i \mid \lambda) = \lambda e^{-\lambda x_i} I(x_i > 0)$

So $\frac{1}{\bar{X}_n}$ should (for large n) be approximately Normal with mean $\frac{1}{\frac{1}{\lambda}} = \lambda$

and SD $\frac{\sigma_{\bar{X}}}{\mu_{\bar{X}}^2 \sqrt{n}} = \frac{\frac{1}{\lambda}}{(\frac{1}{\lambda})^2 \sqrt{n}} = \frac{\lambda}{\sqrt{n}}$

Fancy version of Δ method

$\bar{Y}_1, \bar{Y}_2, \ldots$ sequence of discrete or continuous rv's

F^* continuous CDF; θ a real number;

$a_1, a_2, \ldots \uparrow \infty$ positive sequence

$g(\cdot)$ a real-valued function of a real variable such that $g'(\cdot)$ is continuous and $g'(\theta) = 0$

Then if $a_n(\bar{Y}_n - \theta) \xrightarrow{D} F^*$, $a_n\left[\frac{g(\bar{Y}_n) - g(\theta)}{|g'(\theta)|}\right] \xrightarrow{D} F^*$

Typical application: $\bar{X}_1, \bar{X}_2, \ldots$ IID

$\bar{Y}_n = \bar{\bar{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \bar{X}_i$; $\quad a_n = \frac{\sqrt{n}}{\sigma_{\bar{X}}}$;

$\theta = \mu_{\bar{X}}$;

$F^* = \Phi$, the standard normal CDF

In this context the theorem says that if

$\frac{\bar{\bar{X}}_n - \mu_{\bar{X}}}{\sigma_{\bar{X}}/\sqrt{n}} \sim N(0,1)$ then $\frac{g(\bar{\bar{X}}_n) - g(\mu_{\bar{X}})}{|g'(\mu_{\bar{X}})|\,\sigma_{\bar{X}}/\sqrt{n}} \sim N(0,1)$

More about the continuity correction

T-S case study revisited

$\bar{X}$ = # T-S babies in family of $n=5$ children

Both parents carriers so that

$P(\text{T-S baby}) = \frac{1}{4} = p$

$\bar{X} \sim \text{Binomial}(n, p)$

$T_i = \begin{cases} 1 & \text{if child } i \text{ is T-S baby} \\ 0 & \text{else} \end{cases}$ $i = 1, \ldots, 5 = n$

Then $(T_i) \overset{IID}{\sim} \text{Bernoulli}(p)$ and $\bar{X} = \sum_{i=1}^{n} T_i$
$(i = 1, \ldots, n)$

So by the CLT the dist of $\bar{X}$ should be approximately Normal with mean $\mu_{\bar{X}} = E(\bar{X}) = np = 1.25$

and SD $\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \sqrt{np(1-p)} \doteq 0.98$

On Day 1 of this class we worked out that

$P(1 \text{ or more T-S babies}) = P(\bar{X} \geq 1)$

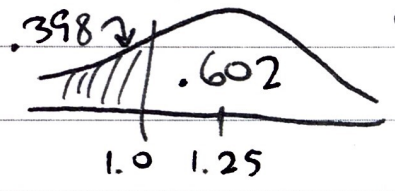
$1 - P(\text{no T-S babies}) = 1 - P(\bar{X} = 0)$

$= 1 - (1-p)^n$

$\doteq 0.76$

Naive Normal approximation from CLT:

SD 0.98



PDF of $\bar{X}'$

$\frac{1.0 - 1.25}{0.98} \doteq -0.2$

$P(\bar{X} \geq 1) \doteq 1 - P(\bar{X}' < 1)$

$= 1 - 0.398$

$\doteq 0.602$ (quite a bad approximation)

Improved approximation obtained by paying attention to the edges of the histogram (PMF) bars

Normal approximation with continuity correction

$P(\bar{X} \geq 1) = 1 - P(\bar{X}' < 0.5)$

$= 1 - .219$

$= 0.781$ (correct answer 0.76; much better approx.)