08/28/19

Lecture 14 AMS 131

Contaminated water supply:

X = arsenic concentration
Y = lead concentration
] both ≥ 0

$R = \frac{Y}{X+Y}$ proportion of contamination due to lead.

difficult

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

randomly sample n pairs (X_i, Y_i) from the joint PDF of (X, Y) and calculate:

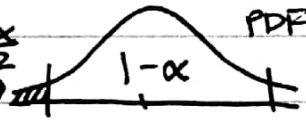
$R_i = \frac{Y_i}{X_i + Y_i}$ and $\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$ good Monte Carlo estimate (simulation) of $E(R)$

How big does n need to be to achieve a desired accuracy?

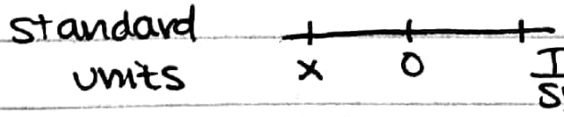
$|R_i| = \left|\frac{Y_i}{X_i + Y_i}\right| \leq 1 \rightarrow V(R_i) \leq \frac{1}{4}$

CLT says that the dist. of $\bar{R}_n$ will be close to normal for large n, with mean $E(R_i)$ and variance:

$\frac{V(R_i)}{n} \leq \frac{1}{4n}$ $\quad SD \leq \frac{1}{2\sqrt{n}}$



n large



suppose we want $\bar{R}_n$ to differ from $E(R_i)$ by no more than tolerance T with probability $(1-\alpha)$, so

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

use inverse calculator

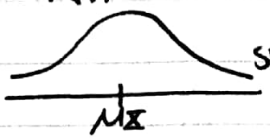
$x = \Phi^{-1}\left(\frac{\alpha}{2}\right) = \frac{[E(R_i) - T] - E(R_i)}{SD} = \frac{-T}{SD} \leq 2T\sqrt{n}$

$n \geq \left[\frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{2T}\right]^2$ say $\alpha = 0.02$ and $T = 0.005$ $\quad n \geq \left[\frac{-2.326}{2(0.005)}\right]^2 \doteq 54,119$

↑ simulation replications

AMS 131 Lecture 14 (cont.)

The Delta Method: the CLT says that if $X_i \sim$ IID any dist. with finite mean and variance, then the dist of $\frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}}$ for large n is approx. standard normal

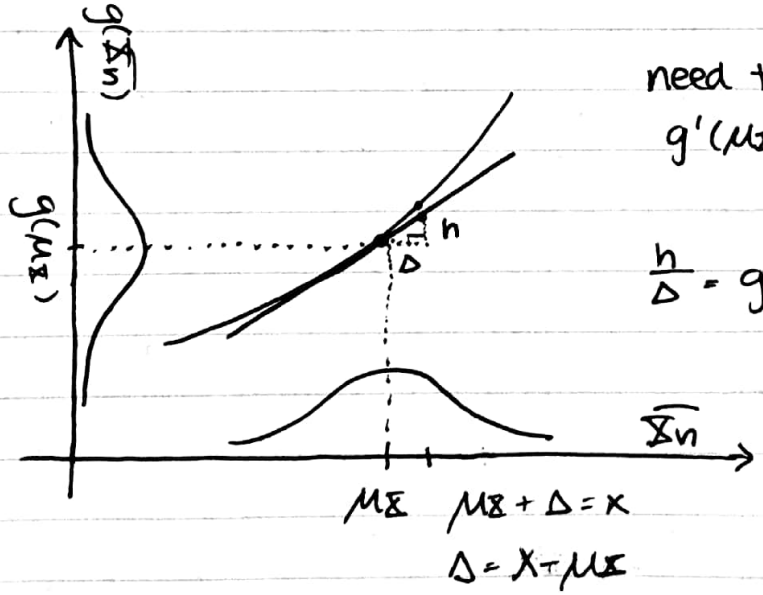
$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$  SD $\frac{\sigma_X}{\sqrt{n}}$ $\bar{X}_n \doteq N(\mu_X, \frac{\sigma_X^2}{n})$

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

↳ Taylor series based approach: delta method

$\bar{X}_n$ should be close to μ_X for large n (the weak law of large numbers)

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



need to assume $g'(\mu_X) \neq 0$

$\frac{h}{\Delta} = g'(\mu_X)$ $g(x) \doteq g(\mu_X) + h$

$= g(\mu_X) + g'(\mu_X) \cdot \Delta$

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

$\Delta = X - \mu_X$

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

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

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

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

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

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

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

This works for any R.V. with $V < \infty$ as $g'(\mu_X) \neq 0$

V any R.V. with finite variance σ_V^2 (and ∴ finite mean μ_V), $W = g(V) \rightarrow$

$E(W) \doteq g(\mu_V)$ and $V(W) \doteq [g'(\mu_V)]^2 \sigma_V^2$

provided $g'(v)$ is continuous and $g'(\mu_V) \neq 0$

if $V \sim$ Normal, then $W = g(V) \sim$ Normal also

X_i = time customer i waits from reaching the front of the line until served

- usually varies by day of week and time of day, so pick a single time slot (ex: M 10:00-10:15 AM) and observe X_i week to week in that time slot

i indexes week, so $\{X_i, i = 1, 2 \ldots n\}$ forms a stochastic process with fixed $E(X_i) = \mu_X > 0$

· stationary: non-time-varying mean $V(X_i) = \sigma_X^2$ non-time-varying variance

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ for large n rate of service: $g(\mu_X) = \frac{1}{\mu_X}$, $g(\bar{X}_n) = \frac{1}{\bar{X}_n}$

$E(\bar{X}_n) = \mu_X \quad g(x) = \frac{1}{x} \quad g'(x) = -\frac{1}{x^2}$

$V(\bar{X}_n) = \frac{\sigma_X^2}{n} \quad g'(\mu_X) = -\frac{1}{\mu_X^2} \quad \bar{X}_n \doteq$ Normal by CLT

Δ-method says: $g(\bar{X}_n) = \frac{1}{\bar{X}_n} \doteq$ Normal with,

$g(\mu_X) = \frac{1}{\mu_X}$ variance $= \frac{\sigma_X^2}{n\mu_X^4} \quad [g'(\mu_X)]^2 = \frac{1}{\mu_X^4} \neq 0$

AMS 131 Lecture 14 (cont.)

$(X_i \mid \lambda) \overset{IID}{\sim}$ Exponential(λ) reasonable for waiting times

$$E(X_i) = \mu_X = \frac{1}{\lambda} \qquad V(X_i) = \sigma_X^2 = \frac{1}{\lambda^2}$$

PDF: $f_{X_i}(x_i \mid \lambda) = \lambda e^{-\lambda x_i} I(x_i > 0)$

So $\frac{1}{\overline{X}_n}$ should be approx. Normal for large n with:

mean $\frac{1}{\frac{1}{\lambda}} = \lambda$ and SD $\frac{\sigma_X}{\mu_X^2 \sqrt{n}} = \frac{(\frac{1}{\lambda})}{(\frac{1}{\lambda})^2 \sqrt{n}} = \frac{\lambda}{\sqrt{n}}$

$Y_1, Y_2 \ldots$ sequence of R.V., F^* continuous CDF

θ real number, $a_1, a_2 \to \infty$ positive sequence

$g(\cdot)$ a real-valued function of a R.V. such that $g'(\cdot)$ is continuous and $g'(\theta) \neq 0$, then if $a_n (Y_n - \theta) \xrightarrow{D} F^*$,

$$a_n \left[\frac{g(Y_n) - g(\theta)}{|g'(\theta)|} \right] \xrightarrow{D} F^* \text{ also}$$

typical application: $X_1, X_2 \ldots$ IID

$$\overline{Y}_n = \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \theta = \mu_X \qquad a_n = \frac{\sqrt{n}}{\sigma_X}$$

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

if $\frac{\overline{X}_n - \mu_X}{\sigma_X / \sqrt{n}} \dot{\sim} N(0,1) \longrightarrow$ then $\frac{g(\overline{X}_n) - g(\mu_X)}{|g'(\mu_X)| \sigma_X / \sqrt{n}} \dot{\sim} N(0,1)$ also

pop. sample

$\begin{bmatrix} ? \end{bmatrix} \xrightarrow{IID} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \begin{matrix} I_1 \\ \cdots \\ I_n \end{matrix}$ $\uparrow n \downarrow$

with unknown mean, SD, and PDF of the population

Known: mean $\bar{I}_n$ ($\bar{Y}_n$)

This is statistical inference: we want to draw inferential conclusions about μ

(Based on the data, μ is around ___1.___ give or take ___2.___ and I'm highly confident μ is between ___3.___ and ___4.___.

confidence interval (CCI)

1. $\bar{I}_n$ ($\bar{Y}_n$) 2. $\frac{\sigma}{\sqrt{n}}$ 3. $\bar{I}_n - 2\frac{\sigma}{\sqrt{n}}$ 4. $\bar{I}_n + 2\frac{\sigma}{\sqrt{n}}$

(or t^{-1} number)

pop sample

$\begin{bmatrix} \end{bmatrix} \xrightarrow{IID} \begin{bmatrix} \end{bmatrix}$ new hypothetical sample with a different $\bar{I}_n$

mean: $\bar{I}_n$

How probable is it that $\bar{I}_n$ will differ from μ by more than ____?

possible $\bar{I}_n$

$\begin{bmatrix} \bullet \\ \bullet \\ \cdots \end{bmatrix} \begin{matrix} \uparrow \\ M \\ \downarrow \end{matrix}$ repeated sampling data set

as $M \to \infty$, long term mean: $E(\bar{I}_n)$

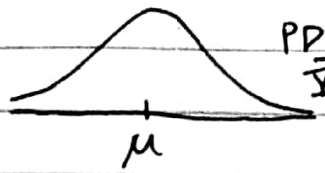
by (weak) law of large numbers $E(\bar{I}_n) = \mu$ ✓

as $M \to \infty$, long term SD: $\sqrt{V(\bar{I}_n)} \to V(\bar{I}_n) = \frac{\sigma^2}{n}$

long term PDF:

if $\hat{\theta}_n$ estimates θ, $SD(\hat{\theta}_n) \triangleq$ standard error (SE) of $\hat{\theta}_n$

$SE(\bar{I}_n) = SD(\bar{I}_n)$



PDF of $\bar{I}_n$

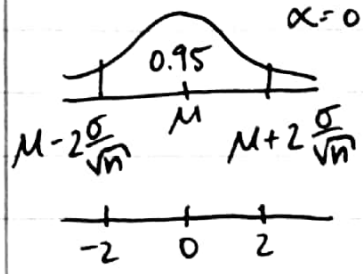
n large enough for CLT to apply

SE $\frac{\sigma}{\sqrt{n}}$

Lecture 14 (cont.)

pick $0 < \alpha$, but small

pretend σ known (step 1)



SE $\frac{\sigma}{\sqrt{n}}$

frequentist.

$P_F\left(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{Y}_n < \mu + 2\frac{\sigma}{\sqrt{n}}\right) \doteq 95\%$.

Neyman's confidence trick

$\mu < \bar{Y}_n + 2\frac{\sigma}{\sqrt{n}}$ $\quad$ $\mu > \bar{Y}_n - 2\frac{\sigma}{\sqrt{n}}$

$P_F\left(\bar{Y}_n - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y}_n + 2\frac{\sigma}{\sqrt{n}}\right) \doteq 95\%$.

$\therefore$ lets define $\left(\bar{Y}_n \pm 2\frac{\sigma}{\sqrt{n}}\right)$ to be a 95% confidence interval for μ

$\bar{Y}_n = 10$

$n = 25$

$\sigma = 4$

$\bar{Y}_n \pm 2\frac{\sigma}{\sqrt{n}} = 10 \pm 2\left(\frac{4}{5}\right) = (8.4, 11.6)$

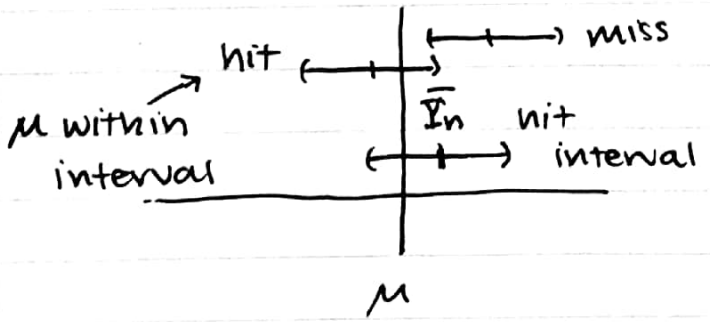
(95% CI for μ)

8.4 10 11.6

easy to conclude that $P_F(8.4 < \mu < 11.6) \doteq 95\%$.

but this is wrong, μ is a fixed unknown constant

$\therefore$ no frequentist probability statements can be made about it: $P_F(8.4 < \mu < 11.6) =$ undef.



95% confidence $\longleftrightarrow$ 95% to hit across many CIs

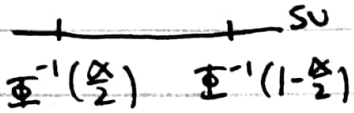
We will not know if <u>our</u> 95% CI is a hit or miss, only way to know is if you know μ - if you knew μ, you wouldn't need a CI in the first place.

our confidence is in the <u>process</u> of building the CI, not the outcome (the CI itself)

$100(1-\alpha)\%$ CI:

$$\bar{X}_n \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$$

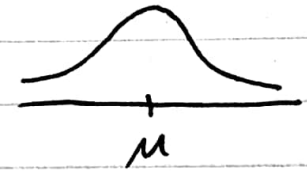
$100(1-\alpha)\%$ CI have a $100\cdot\alpha\%$ false-positive (miss) rate



95% CI: too low for good, replicable science

(0.05) false positive rate → (0.005), 99.5% CI

σ unknown (step 2)



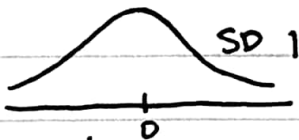
PDF of $\bar{X}_n$ (CLT) SE $\frac{\sigma}{\sqrt{n}}$

$$\bar{X}_n \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$$ ← now use S_n, the SD of the sample, not normal anymore

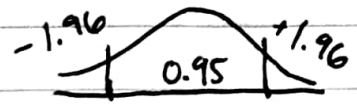
long term SD $\sqrt{\hat{V}(\bar{X}_n)} = \frac{S_n}{\sqrt{n}}$

n large → $\Phi^{-1}(1-\frac{\alpha}{2})$ approx. ok

n small



$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ CLT



↓ t curve

t_{n-1} degrees of freedom

$n\uparrow$, PDF → normal, $t_n \to \Phi$ standard normal

AMS 131 Lecture 14 (cont.)

T-S case study

X = # of T-S babies in family of 5, both parents are carriers

$P(\text{T-S baby}) = \frac{1}{4} = p$ $\quad X \sim \text{Binomial}(n, p)$

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

then $T_i \overset{IID}{\sim} \text{Bernoulli}(p)$ $\quad i = 1, \ldots, n$ and

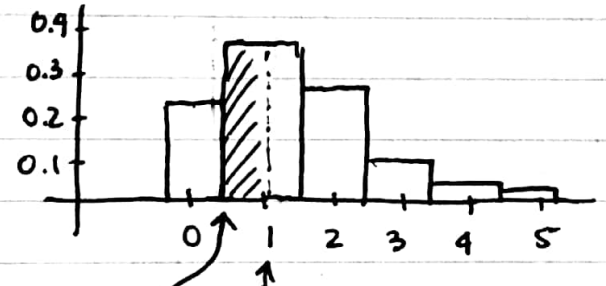
$X = \sum_{i=1}^{n} T_i$ by CLT, X should be approx. normal

$\mu_X = E(X) = np = 1.25$

$\sigma_X = \sqrt{V(X)} = \sqrt{np(1-p)} \doteq 0.98$

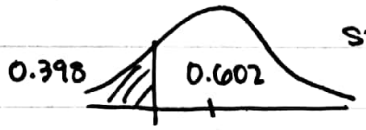
$P(1 \text{ or more T-S}) = P(X \geq 1)$

$1 - P(\text{no T-S}) = 1 - (1-p)^n \doteq 0.76 \quad = 1 - P(X = 0)$



better approx. naive approx.

naive normal approx. from CLT



$\frac{1.0 - 1.25}{0.98} \doteq -0.26$

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

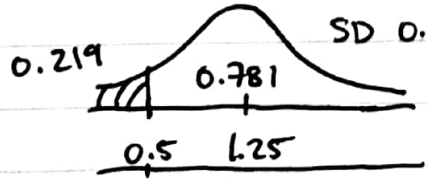
$= 1 - 0.398$

$\doteq 0.602$ bad approx. (vs. 0.76)

not good approx. because $n = 5$

Continuity correction

now look at the edges of the histogram bars



$\frac{0.5 - 1.25}{0.98} \doteq -0.77$

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

$= 1 - 0.219$

$\rightarrow 0.781$

much closer to 0.76

Markov Chains

A sequence of R.V.s $X_1, X_2 \ldots$ is called a stochastic process with discrete time parameter $t = 1, 2 \ldots$ X_1 is the <u>initial state</u> of the process X_n, $n \geq 1$ is the <u>state</u> of the process at $t = n$

The simplest possible discrete time stochastic process is an IID sequence of R.V.s $(X_1, X_2 \ldots)$

Suppose that there's a parameter θ such that $(X_i \mid \theta) \sim$ IID from some dist. depending on θ

↳ machine with θ dial produces IID Bernoulli(θ)

- the process $(X_1, X_2 \ldots)$ does have memory for you if θ is unknown: the information of previous trials helps you predict future ones
- the process $(X_i \mid \theta)$ has no memory if θ known

An IID process $(X_i \mid \theta)$ is called a white-noise (stochastic) process or white noise time series

What's the next level of complexity?

Allow X_{n+1} depend on X_n, but not $X_{n-1}, n-2 \ldots$ the process has short term memory, 1 time period

A discrete-time stochastic process is a (first order) Markov chain if for $n = 1, 2 \ldots$; b, any real number; and for all possible sequences of states $x_1, x_2 \ldots$

$$P(X_{n+1} \leq b \mid X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} \leq b \mid X_n = x_n)$$

The only thing you need to simulate where the Markov chain is going next is where it is now.