ex. $E(X) = 1$, X non-negative $\rightarrow$

$P(X \geq 100) \leq \frac{1}{100}$ The inequality is sharp, meaning that the upper bound $\frac{E(X)}{t}$ on $P(X \geq t)$ is attainable, ⊛

ex. $E(X) = 1$, X - nonnegative $\rightarrow$ put probability 0.99 on $X = 0$ and 0.01 on 100

⊛ but most of the time (i.e., for most distributions) it's a crude upper bound.

(30 May 19)

Can apply Markov inequality to the r.v. $Y = (X - E(X))^2$ to get

Chebyshev Inequality → X r.v. with $V(X)$ existing

for every $t \geq 0$,

$$P\left[|X - E(X)| \geq t\right] \leq \frac{V(X)}{t^2}$$

(attributed to Pafnuty Chebyshev (1821 - 1894), also a Russian mathematician, one of whose Ph.D. students was Markov)

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

Chebyshev says $P\left[\left|\frac{X-\mu}{\sigma}\right| \geq 3\right] \leq \frac{1}{3^2} = \frac{1}{9}$,

so no more than $\frac{1}{9} = 11\%$ of the probability in any distribution with finite variance can be more than 3 SDs away from the mean (recall for Normal dist. this prob. is 0.3%)

This upper bound is also 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 $t > 0$

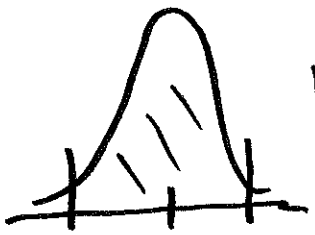
This can be rewritten $P(|\bar{X}_n - \mu| < t) \geq 1 - \frac{\sigma^2}{nt^2}$

PDF of $\bar{X}_n$

$n=1$

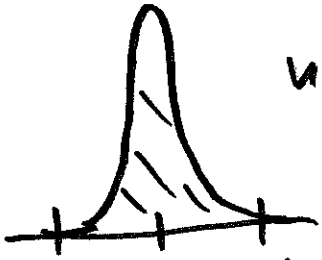
$\mu-t \quad \mu \quad \mu+t$

$n=10$



$\mu-t \quad \mu \quad \mu+t$

$n=100$



$\mu-t \quad \mu \quad \mu+t$

$\vdots$

This suggests a way to quantify how close a r.v. like $\bar{X}_n$ is to a constant like μ:

Def. A sequence $Z_1, Z_2, \ldots$ of r.v. 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 $Z_n \xrightarrow{P} b$.

This definition is immediate consequence of Chebyshev's

(Weak) Law of Large Numbers

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

($\bar{X}_n$ is consistent for μ)

$\rightarrow \bar{X}_n \xrightarrow{P} \mu$

This result has a long history: the Italian mathematician Geralamo Cardano (1501-1576) asserted it without proof; Jacob Bernoulli (1655-1705) proved it for $(X_i | \theta) \overset{IID}{\sim}$ Bernoulli(θ) (it took him 20 years to find a correct proof, published posthumously in 1713; Bernoulli thought that this theorem proved the existence of God); Siméon Denis Poisson named it the Law of Large Numbers in 1837.

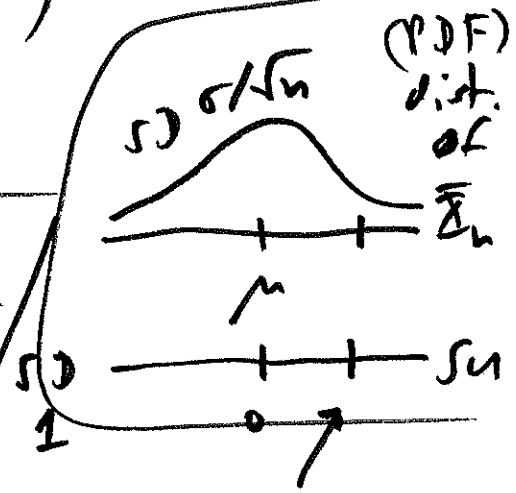
Corollary If $Z_n \xrightarrow{P} b$ and $g(z)$ is continuous at $z = b$ then $g(Z_n) \xrightarrow{P} g(b)$.

Central Limit Theorem (CLT)

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

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

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



Q: Does something like this work for other choices of $X_i \overset{IID}{\sim}$?

A: Yes: it's the most famous result in all of probability:

Central Limit Theorem (CLT)

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

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

Careful statement

Def. $X_1, X_2, \ldots$ a sequence of r.v.; 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 people say that $X_n \xrightarrow{D} F^*$ ("X_n converges in distribution to F^*")

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

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

The CLT

---

also has a long history: it was

first demonstrated for $X_i \overset{IID}{\sim} \text{Bernoulli}(\theta)$ (p. 95) by the French/British mathematician Abraham de Moivre (1667 - 1754) in 1733; almost forgotten until revived by the French mathematician Pierre-Simon de Laplace (1749 - 1827) in 1812; almost forgotten again until 1901, when the Russian mathematician Aleksandr Lyapunov gave a more general proof; even more general proof provided by JW Lindeberg (Finnish mathematician (1876 - 1932)) and independently by Paul Lévy (French mathematician (1886 - 1971)) in the early 1920s. [CLT name due to Hungarian-American mathematician (1887-1985) George Pólya in 1920

Example Contaminated water supply:

X = arsenic concentration

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

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

(proportion of contamination due to lead)

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

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

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

By definition

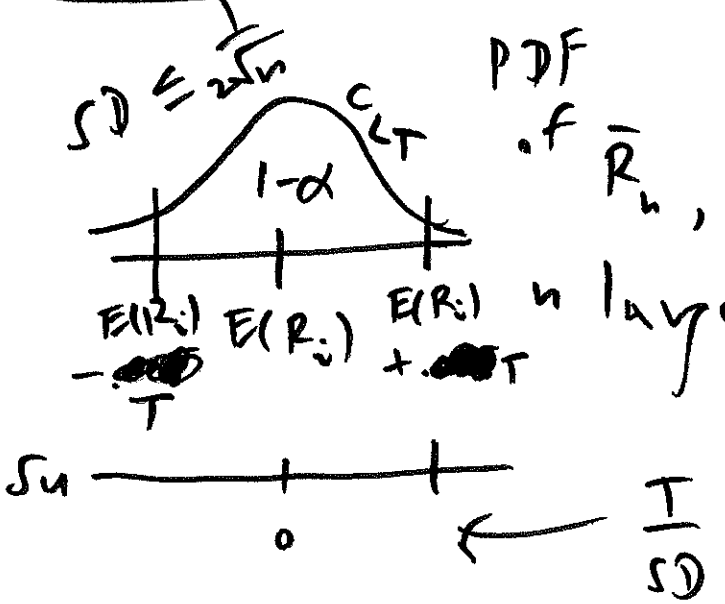
$$|R_i| = \left|\frac{I_i}{\Sigma_i + I_i}\right| \le 1$$; can show that

as a result $V(R_i) \le \frac{1}{4}$.

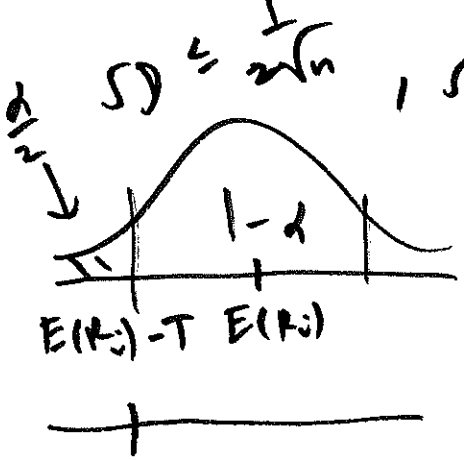
CLT

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

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



$SD \le \frac{1}{2\sqrt{n}}$, so $\frac{1}{SD} \ge 2\sqrt{n}$ and



$$\frac{-T}{SD} \le 2T\sqrt{n}$$

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

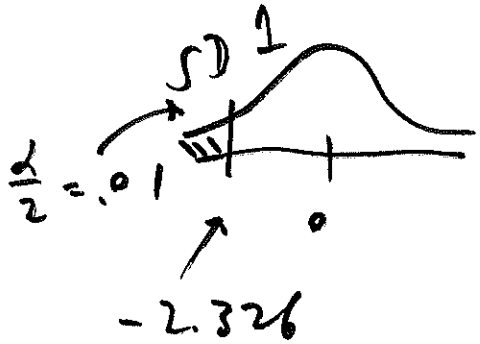
from which $n \ge \left[\frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{2T}\right]^2$

For instance, set $T = 0.005$ ($\frac{1}{2}$ of 1%)

and $\alpha = .02$ to get

$$n \ge \left[\frac{-2.326}{2(.005)}\right]^2 \doteq 54,119$$

simulation replications needed

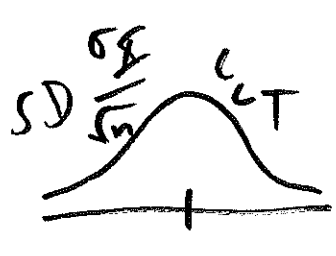


Case study: Escalators in the London Underground ()

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



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

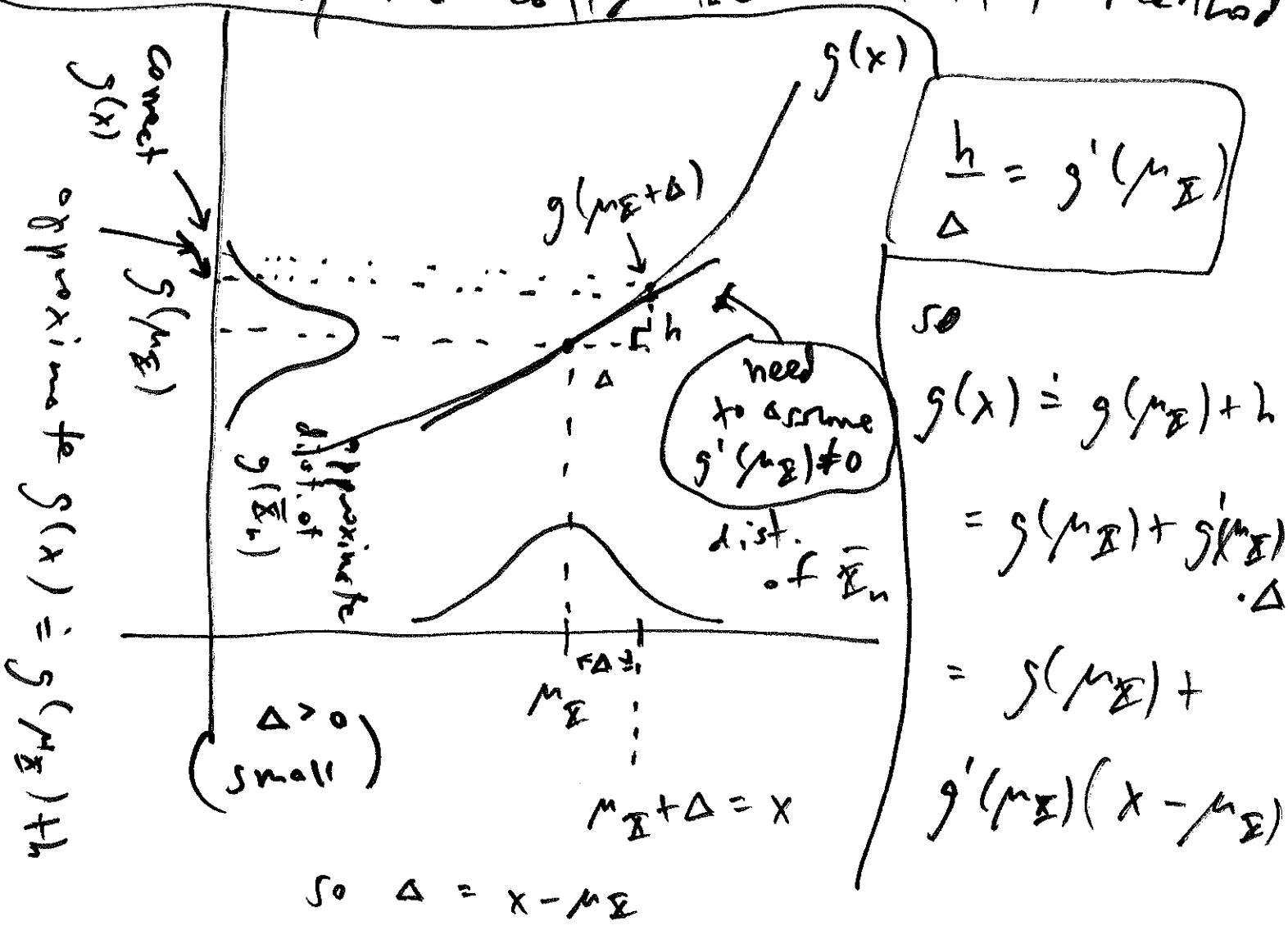
Question: If $g(x)$ is a sufficiently "nice" function, is there a comparable result for $g(\bar{X}_n)$?

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 $x = \mu_X$:

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

this is why it's called the Δ (Delta) - method



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

(constant; r.v.; constant; constant)

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

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

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

and

(constant; constant; r.v.)

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

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

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

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

There's one hidden assumption in this calculation: $g'(\mu_{\bar{X}}) \neq 0$.

This works for <u>any</u> r.v. with finite variance, not just $\bar{X}_n$:

V any r.v. 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

$g'(\mu_V) \neq 0$

Δ method part 2

Moreover, if V is Normal then $W = g(V)$ is Normal also.

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

Let X_i = time customer i waits from reaching the head of the queue until served. To be completely realistic, the dist. of 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 $g(\bar{X}_n) = \frac{1}{\bar{X}_n}$.

complications: seasonal effects (ignored here)

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

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

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

$g'(\mu_X) = -\frac{1}{\mu_X^2}$ | $\bar{X}_n \dot{\sim}$ Normal by CLT

so Δ-method says $g(\bar{X}_n) = \frac{1}{\bar{X}_n} \dot{\sim}$ Normal also, with mean $g(\mu_X) = \frac{1}{\mu_X}$ and variance $\sigma_X^2/(n\mu_X^4)$

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

Specific Calculation

Under some plausible assumptions, we'll see that $(X_i|\lambda) \overset{\text{IID}}{\sim} \text{Exponential}(\lambda)$ may be a reasonable model for waiting times.

$E(X_i) = \frac{1}{\lambda}$, $V(X_i) = \frac{1}{\lambda^2}$ $(X_i|\lambda)$ has PDF

$= \mu_X$ $= \sigma_X^2$

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

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

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

(discrete or continuous)

Fancy version of Δ-method

$Y_1, Y_2, \ldots$ sequence of r.v.;

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 $g'(\theta) \neq 0$; then if $a_n(T_n - \theta) \xrightarrow{D} F^*$,

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

Typical application: $X_1, X_2, \ldots$ IID

$T_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$; $\theta = \mu_X$; $a_n = \frac{\sqrt{n}}{\sigma_X}$;

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

In this context the theorem says that

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

(28 Aug 17)

A little bit more about the continuity correction

Tay-Sachs case study, revisited

X = # T-S babies in family of $n = 5$ children, both parents carriers so that

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

But also let $T_i = \begin{cases} 1 & \text{if child } i \text{ is T-S baby} \\ 0 & \text{else} \end{cases}$

$i = 1, \ldots, n = 5$

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

So by the CLT the dist. of X should be approximately Normal with mean

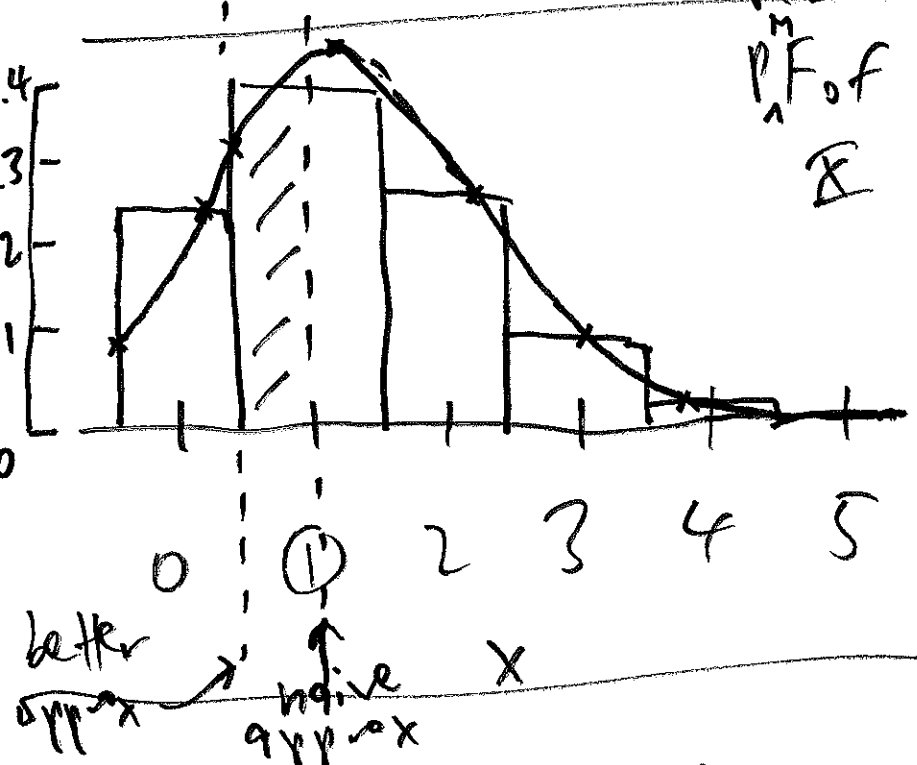
$\mu_X = E(X) = np = 1.25$ and SD

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

On day 1 of this class we worked out that $p(\text{1 or more T-s boiling}) = P(X \geq 1)$

$$1 - p(\text{no T-s boiling}) = 1 - (1-p)^n \doteq 0.76$$

$$= 1 - P(X=0)$$

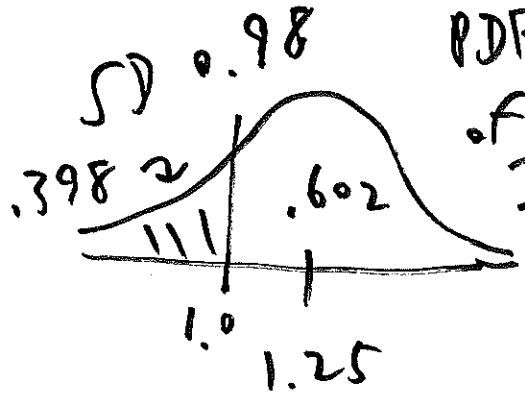


Naive Normal approximation, from CLT:

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

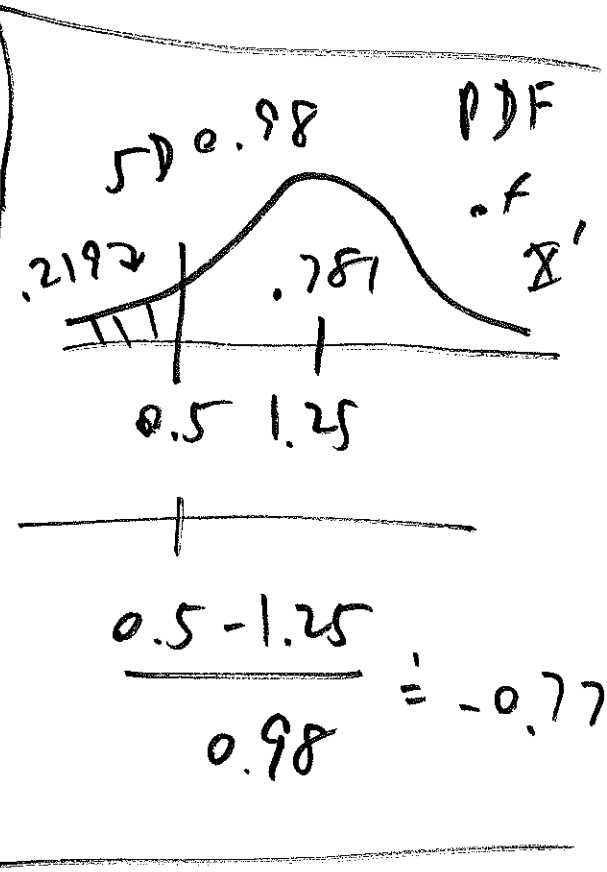
$$= 1 - 0.398$$

$$\doteq \boxed{0.602}$$ (quite a bad approximation)



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

Improved approximation obtained by paying attention to the edges of the histogram ($\hat{F}^M$) bars:



Normal approximation with continuity correction

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

$\doteq 1 - .219$

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

(4 Jan 19)

Markov Chains

Recall the definition of a stochastic process:

Def. A sequence of rvs $X_1, X_2, \ldots$ is called a stochastic process with discrete time parameter $t = 1, 2, \ldots$.

X_1 is the initial state of the process;

X_n, $n \geq 1$ is the state of the process at time $t = n$.

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

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

Q: Does this process have a memory?

Example, revisited: Machine with θ dial from 0 to 1, produces IID Bernoulli(θ) trials X_i. Recall that the process $(X_1, X_2, \ldots)$ does have a memory for you if θ is unknown to you: the information that 17 out of the first 20 trials were successes helps you to predict X_{21}, because it's reasonable to conclude from $X_1, \ldots, X_{20}$ that θ is around $\frac{17}{20} \doteq 0.85$, so X_{21} will probably be a success. But the process $\{(X_i | \theta), i = 1, 2, \ldots\}$ has no memory once θ is known: information about

the first n trials is irrelevant to your prediction of X_{n+1} if you know

Def. An IID process $(X_i | \theta) \overset{IID}{\sim}$ is called a white-noise (stochastic) process or a white noise time series.

Q: What's the next level of complexity (for discrete-time stochastic processes) up from white noise? **A:** Allow X_{n+1} to depend on X_n but not on $X_{n-1}, X_{n-2}, \ldots$ (i.e., let the process have a short-term memory, (1) time period back in the past).