YZV 231E

20.12.2021

Probability Theory & Stats

Gü.

Recap: Conditional Expectations

Given the value of y of an r.v. Y:

$E[X \mid Y = y]$ becomes a fn. of Y, $\Rightarrow g(Y)$.

$E[X \mid Y] \rightarrow$ is an r.v.

$$E[E[X \mid Y]] = E[X]$$

Conditional Variance: $Var(X \mid Y)$ & its expectations

$Var(X \mid Y = y)$

& $Var(X \mid Y)$ is an r.v.

&

Total Variance Law: $Var(X) = E[Var(X \mid Y)] + Var(E[X \mid Y])$

Ex: A class of 30 students w/ 2 sections,

They take a quiz: 2 r.v.s of interest

→ X: quiz score

Y: section number → $y=1$ (section 1) (10 students)

$y=2$ section 2 (20 students)

Given statistics; # students in each section & quiz statistics

Quiz average in section 1 : 90

Quiz average " 2 : 60

$E[X]=?$

$$\begin{cases} y=1: \frac{1}{10}\sum_{i=1}^{10} x_i = 90 \\ y=2: \frac{1}{20}\sum_{i=11}^{30} x_i = 60 \end{cases}$$

$$E[X] \overset{\sim}{=} \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{90 \cdot 10 + 60 \cdot 20}{30} = 70.$$

→ let's look at conditional expectation.

$E[X \mid Y=1] = 90$

$E[X \mid Y=2] = 60$

$E[X|Y]$ is a random variable → calculate its variance

$$E[X|Y] = \begin{cases} 90, & \text{w.p. } \frac{1}{3} \\ 60, & \text{w.p. } \frac{2}{3} \end{cases}$$

take its expectation

$$\rightarrow E[\underbrace{E[X|Y]}] = 90 \cdot \frac{1}{3} + 60 \cdot \frac{2}{3} = 70 = E[X] \checkmark$$

$$\to \mathrm{Var}(E[X|Y]) = \frac{1}{3}\underbrace{(90-70)^2} + \frac{2}{3}(60-70)^2 = 200$$

Suppose we are given variances in each section

$$\underbrace{\frac{1}{10}\sum_{i=1}^{10}(x_i-90)^2 = 10}, \quad \underbrace{\frac{1}{20}\sum_{i=11}^{30}(x_i-60)^2 = 20}$$

$\boxed{\mathrm{Var}(X|Y)}$ r.v. $\to$ $\mathrm{Var}(X|Y=1) = 10$, $\mathrm{Var}(X|Y=2) = 20$

The r.v $\mathrm{Var}(X|Y) = \begin{cases} 10 & \text{w.p. } \frac{1}{3} \\ 20 & \text{w.p. } \frac{2}{3} \end{cases}$

$$\to E[\mathrm{Var}(X|Y)] = \frac{1}{3}\cdot 10 + \frac{2}{3}\cdot 20 = \frac{50}{3}$$

$= \frac{50}{3} + 200$

$$\to \underbrace{\mathrm{Var}(X)}_{\text{Total Variance}} = \underbrace{E[\mathrm{Var}(X|Y)]}_{\text{AVERAGE VARIABILITY WITHIN SECTIONS}} + \underbrace{\mathrm{Var}(E[X|Y])}_{\text{Variability BETWEEN sections}}$$

1 + 2

$E[\mathrm{Var}(X|Y)]$

WITHIN

(1) Average of (σ_i^2) : average variance within sections

$\mathrm{Var}(E[X|Y])$: how widely spread the means μ_i are for each group.

(2) Variability of (μ_1, μ_2, μ_3) :

BETWEEN

* Sum of a random number of independent r.v.-s:

N : # stores visited ≥ 0 :

X_i : money spent in store i.

— X_i : are i.i.d. (independent identically distrib)

* X_i, independent of N.

N : is a non-negative integer r.v.

Let $Y = X_1 + \dots + X_N$; sum of money you spent

⟹ Q. $E[Y] = ?$, $Var(Y) = ?$

$E[Y \mid N=n] = E[X_1 + \dots + X_n \mid N=n]$ due indep.

$= E[X_1 + \dots + X_n]$

$= E[X_1] + \dots + E[X_n] = \underbrace{n}_{\text{given number}} \cdot \underbrace{E[X]}_{\text{a number}}$

take this again to an r.v.

$\boxed{E[Y \mid N] = N \cdot E[X]}$ r.v.

Now this is a r.v.

$E[E[Y \mid N]] = E[Y] = E[N \cdot \underbrace{E[X]}_{\text{this is a number}}] = E[N] \cdot E[X]$

* Variance of a sum of a Random Number of Independent r.v.s.

$$Var(Y) = \underbrace{E[Var(Y|N)]}_{①} + \underbrace{Var(E[Y|N])}_{②}$$

① $Var(\underbrace{Y}_{(X_1+\dots+X_n)} | N=n) = n . Var(X)$ $\Big\} \rightarrow \underbrace{Var(Y|N) = N . Var(X)}_{r.v.}$

$\rightarrow E[N . Var(X)] = E[N] . Var(X)$

② $E[Y|N] = N . \underbrace{E[X]}$ (found on prev page)

$Z = a . N$
$Var(Z) = a^2 Var(N)$

$Var(E[Y|N]) = (E[X])^2 . Var(N)$

$$\rightarrow Var(Y) = \underbrace{E[N] . Var(X)}_{\text{within variability}} + \underbrace{(E[X])^2 . Var(N)}_{\text{variability between.}}$$

LIMIT THEOREMS :

1) Weak Law of Large Numbers (Averages)
2) Central Limit Theorems : (Distribution)

Want to come up w/ an expected value

Height of Students in Turkey : $X_1, \dots, X_n$ i.i.d ☆

$$M_n = \frac{X_1 + \dots + X_n}{n} \; ; \quad n \to \infty \; ?$$

Sample Mean : is a $\boxed{\text{random variable}}$ ← M_n

as $n \to \infty$ $\quad M_n \xrightarrow{n\to\infty} E[X]$?

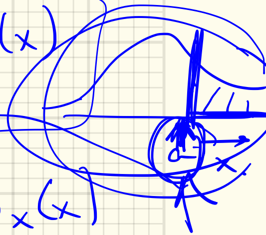
in what sense ? ↗ in prob ↘ in distrib.

To study these convergences let's first study a few tools :

Markov Inequality :

- Say we have an r.v X , say $X \geq 0$, assume it's discrete r.v.

$$E[X] = \underbrace{\sum_{x}}_{\text{whole sum}} \underset{\geq 0}{x} \cdot \underbrace{p_X(x)}_{\geq 0} \geq \boxed{\sum_{x \geq a} x \cdot p_X(x)}$$



$$\geq \underbrace{\sum_{x \geq a} a \cdot p_X(x)}$$

$$= \underbrace{a \sum_{x \geq a} p_X(x)}$$

$$= a \cdot P(X \geq a)$$

Markov Inequality :

$$\boxed{E[X] \geq a \cdot P(X \geq a)}$$

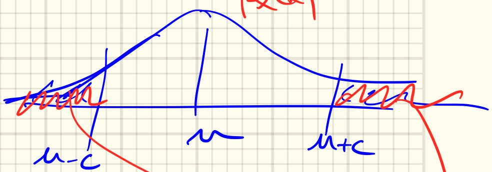
apply this to $(X-\mu)^2$ r.v

smallness of expected values $\longleftrightarrow$ smallness of probabilities

$$E[(X-\mu)^2] \geq a^2 \cdot P((X-\mu)^2 \geq a^2)$$

$$= \operatorname{Var}(X) \geq a^2 P((X-\mu)^2 \geq a^2)$$

→ Another proof: R.V. X w/ finite mean μ & variance σ^2:



$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p_X(x)\,dx$$

$$\geq \int_{-\infty}^{\mu-c} \underbrace{(x-\mu)^2}_{\geq c^2} p_X(x)\,dx + \int_{\mu+c}^{\infty} \underbrace{(x-\mu)^2}_{\geq c^2} p_X(x)\,dx$$

$$\geq c^2 \int_{-\infty}^{\mu-c} p_X(x)\,dx + \int_{\mu+c}^{\infty} p_X(x)\,dx$$

$$= c^2 P(|X-\mu| \geq c)$$

$$\boxed{\sigma^2 \geq c^2 P(|X-\mu| \geq c)}$$

→ Chebysheff Inequality :

→ just sees the μ, σ^2 of the r.v.

$$\boxed{P\left(|X-\mu| \geq c\right) \leq \frac{\sigma^2}{c^2}}$$

Ties probabilities & expectations.

another version:

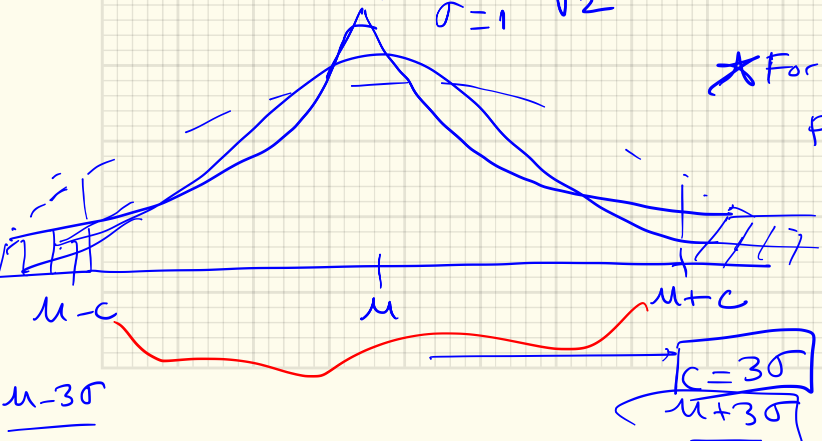
$$\left(P\left(|X-\mu| \geq k.\sigma\right) \leq \frac{1}{k^2} \right)$$

eg. $X \sim N(0,1)$

$X \sim$ Laplace $\left(\Rightarrow \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \right) \rightarrow E[X] = 0$

$\sigma = 1$ $\quad Var(X) = 1$

Chebysheff bound is very conservative : considers worst-case distrib. (not a tight bound)



* For any distrib w/ mean μ & var σ^2.

$$P(|X-\mu|> 3\sigma) \leq \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9} = 0.111$$

* For Gaussian, we can calculate an exact value

$$P(|X-\mu|\geq 3\sigma) = 1 - \int_{-3\sigma}^{3\sigma} \frac{1}{\sqrt{2\pi}} e^{-x^2/2\sigma^2} dx$$

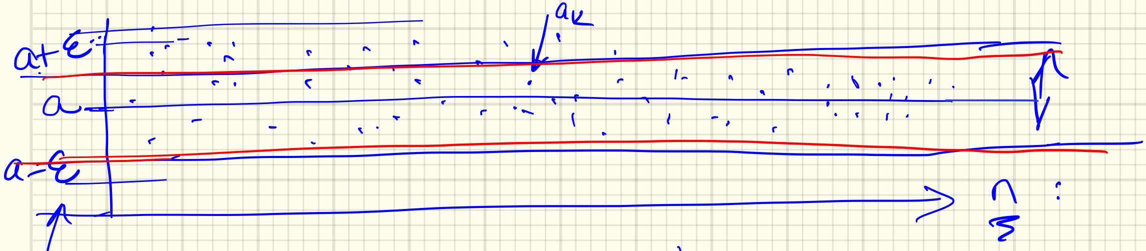
$$= \underset{\uparrow \text{normal cdf}}{\Phi(\mu-3\sigma)} + \underbrace{1 - \Phi(\mu+3\sigma)} = 0.002$$

very tight

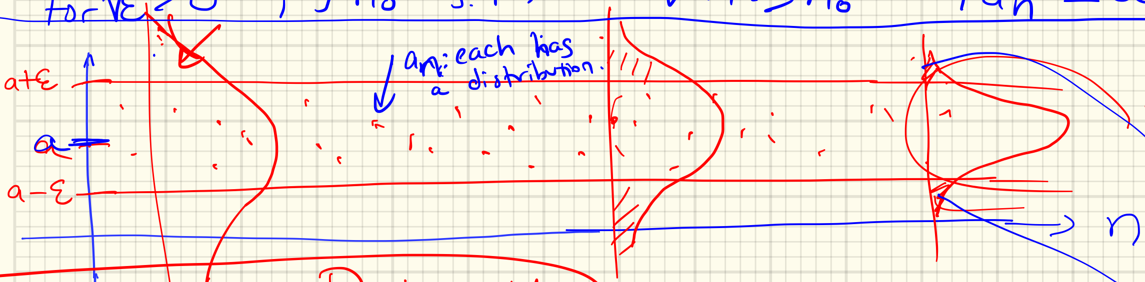
Convergence:

$a_n \longrightarrow a$; $\lim_{n\to\infty} a_n = a$.

from calculus. standard convergence.



For $\forall \varepsilon > 0$, $\exists n_0$ s.t. $\forall n \geq n_0$ $|a_n - a| \leq \varepsilon$.



Converge in Probability

Sequence of r.v.s X_n converges in probability to a number a,

For every $\varepsilon > 0$, $\lim_{n\to\infty} \underbrace{P(|X_n - a| \geq \varepsilon)} = 0$

tail probabilities go to zero as $n \to \infty$.

Weak Law of Large Numbers (WLLN)

$X_1, X_2, \dots$ i.i.d. r.v.s w/ finite mean μ, finite variance σ^2

Sample mean: $\boxed{M_n = \frac{X_1 + X_2 + \dots + X_n}{n}}$ r.v.

$$E[M_n] = \frac{E[X_1] + \dots + E[X_n]}{n} = \frac{n \cdot \mu}{n} = \mu.$$

How big is the variance of M_n?

$$\mathrm{Var}(M_n) = \mathrm{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

Use Chebysheff ineq:

$$P(|M_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(M_n)}{\varepsilon^2} = \frac{\sigma^2}{n \cdot \varepsilon^2} \xrightarrow[n \to \infty]{} 0.$$

$$\boxed{M_n \xrightarrow[\text{in prob.}]{} \mu}$$

convergence in prob.

WLLN

WLLN:

→ Sample mean converges in probability to the true mean (μ)

M_n

Typical Ex: Polling: What fraction of the population prefers something?

"soğanlı menemen" vs "soğansız menemen".

80 million ; 30 million prefers smt : $\frac{3}{8} = p$.

a number f: fraction of population that " "

ith randomly selected person polled, we record it

$$X_i = \begin{cases} 1, & \text{yes (soğanlı)} \\ 0, & \text{no} \end{cases}$$ Bernoulli r.v. w/ p

$$M_n = \left(\frac{X_1 + \dots + X_n}{n}\right) \longrightarrow \hat{p}$$

Goal: 95% confidence ≤ 1% error →

$$\rightarrow \quad P\Big(|M_n - p| \geq \underbrace{0.01}_{\text{accuracy that we want}}\Big) \leq \underbrace{0.05}_{1-\text{confidence.}}$$

Q. How large n, sample size, you should be using in order to satisfy the $\underbrace{\text{specs}}_{\text{accuracy \& confidence.}}$ given by the pollsters.

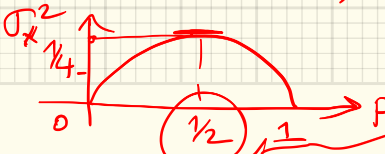
<u>Note:</u> Importance of sampling "uniformly"; eg. don't sample only from the relatives of a candidate!

let's use Chebysheff:

$$P(|M_n - p| \geq 0.01) \leq \frac{\sigma_{M_n}^2}{(0.01)^2} = \frac{\sigma_x^2}{n.(0.01)^2}$$

want this to be ≤ 0.05

Q. What is σ_x^2? for Bernouilli; $\sigma_x^2 = p(1-p)$: Variance of a Bernouilli r.v. → maximize this? Find p^* s.t. σ_x^2 is maximized?



∴ We'll use worst case $\sigma_x^2 = \max \sigma_x^2 = \frac{1}{4}$

$$P(|M_n - p| \geq 0.01) \leq \frac{\sigma_x^2}{n(0.01)^2} \leq \frac{1}{4n(0.01)^2} \leq 0.05$$

$\sigma_x^2 \leq 1/4$

solve for n.

$$n \geq \frac{1}{4 \cdot 10^{-4} \cdot 5 \cdot 10^{-2}} = 50K.$$

If $n = 50{,}000$

$$P(|M_n - p| \geq 0.01) \leq 0.05$$

We'll continue & finish Limit Theorems next time.