

MATH 180A: INTRO TO PROBABILITY (FOR DATA SCIENCE)

www.math.ucsd.edu/~tkemp/180A

Today: § 9.2 - 9.3

Next: REVIEW

CAPEs! Please fill them out.

Lab 7 Due Wednesday, Dec 4.

Homework 8 Due Friday, Dec 6.

Final Exam: Monday, Dec 9, 11:30a-2:30p
in **REC GYM**

Reminder: Chebyshev's Inequality

9.1

For any random variable X with finite

$$\mathbb{E}(X) = \mu \quad \text{Var}(X) = \sigma^2$$

$$\boxed{\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}} \Rightarrow \mathbb{P}(|X - 1000| \geq 3 \cdot 15) \leq \frac{1}{3^2} = \frac{1}{9}$$

We proved this using the fact that \mathbb{E} is monotone:

$$X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$$

Pf. $\mathbb{E}(Y) - \mathbb{E}(X) = \mathbb{E}(X) + \mathbb{E}(-Y) \rightarrow X - Y \geq 0$
 $= \mathbb{E}(X - Y) \rightarrow \mathbb{E}(Y - X) \geq 0$

$\rightarrow = \sum_{k \geq 0} \dots$
or $\int_0^{\infty} \dots$

E.g. Ramen Menya Ultra has, on average, **1000** customers/day, with a standard deviation of **15**. Estimate the probability that today they will have between 956 and 1044 customers. #customers = X

$$\begin{aligned} \mathbb{P}(956 \leq X \leq 1044) &= \mathbb{P}(955 < X < 1045) \rightarrow = \mathbb{P}(|X - 1000| < 3 \cdot 15) && 88.9\% \\ &= \mathbb{P}(-45 < X - 1000 < 45) = 1 - \mathbb{P}(|X - 1000| \geq 3 \cdot 15) \geq 1 - \frac{1}{9} = \frac{8}{9} \end{aligned}$$

(Weak) Law of Large Numbers

9.2

Let $X_1, X_2, X_3, \dots, X_n, \dots$ be an infinite sequence of iid. random variables, each with $E(X_j) = \mu < \infty$ and $\text{Var}(X_j) = \sigma^2 < \infty$ finite.

Let $S_n = X_1 + X_2 + \dots + X_n$. Then for any fixed $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = O\left(\frac{1}{n}\right)$$

Pf. Let $\bar{X}_n = S_n/n$.

$$\begin{aligned} E(\bar{X}_n) &= E\left(\frac{S_n}{n}\right) = \frac{1}{n} E(X_1 + \dots + X_n) \\ &= \frac{1}{n} (E(X_1) + E(X_2) + \dots + E(X_n)) \\ &= \frac{1}{n} (\mu + \mu + \dots + \mu) \\ &= \mu. \end{aligned}$$

$$\begin{aligned} P(|\bar{X}_n - \mu| \geq \varepsilon) & \text{ want } \varepsilon = k \sqrt{\text{Var}(\bar{X}_n)} \\ & = k \sqrt{\frac{\sigma^2}{n}} \\ \therefore k &= \varepsilon \sqrt{\frac{n}{\sigma^2}} \\ P(|\bar{X}_n - \mu| \geq k \cdot \frac{\sigma}{\sqrt{n}}) & \leq \frac{1}{k^2} \\ & = \frac{1}{\left(\varepsilon \sqrt{\frac{n}{\sigma^2}}\right)^2} = \frac{\sigma^2}{\varepsilon^2 n} \end{aligned}$$

$$\begin{aligned} \therefore \text{S.D.}(\bar{X}_n) &= \sqrt{\frac{\sigma^2}{n}} \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \text{Var}\left(\frac{1}{n} S_n\right) = \frac{1}{n^2} \text{Var}(S_n) \\ &= \frac{1}{n^2} \text{Var}(X_1 + X_2 + \dots + X_n) \\ &= \frac{1}{n^2} (\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)) \\ &= \frac{1}{n^2} (\sigma^2 + \sigma^2 + \dots + \sigma^2) \end{aligned}$$

E.g. The Large Hadron Collider was built to detect and measure the mass of the Higgs Boson. Call the mass M .

For theoretical reasons, it is known that $M \leq 1.78 \times 10^{-23} \text{ g}$.

How many trials do the LHC physicists need to do to estimate the correct mass (via sample mean) within $10^{-24} \text{ g} = \varepsilon$, with probability $\geq 95\%$?

Trial measurements $M_1, M_2, M_3, \dots, M_n, \dots$ $\bar{M}_n = \frac{M_1 + \dots + M_n}{n}$

want $P(|\bar{M}_n - \mu| < \varepsilon) \geq 95\%$

WLLN: $P(|\bar{M}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \cdot \frac{1}{n}$.

$\therefore P(|\bar{M}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2} \cdot \frac{1}{n} \geq 95\%$ (want)

$\therefore \frac{\sigma^2}{\varepsilon^2} \cdot \frac{1}{n} \leq 5\%$

$n \geq \frac{\sigma^2}{\varepsilon^2} \cdot 20$.

$\geq \frac{(1.78 \times 10^{-23})^2}{(10^{-24})^2} \cdot 20$

$= 6337$

Fact: (HW)

If $|X| \leq C$,

$\Rightarrow \sqrt{\text{var}(X)} \leq C$

Strong Law of Large Numbers

Let $X_1, X_2, X_3, \dots, X_n, \dots$ be an infinite sequence of iid. random variables each with $E(X_j) = \mu$.

Let $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$. Then

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1,$$

Beautiful Proof uses existence of $E(X_n^4) < \infty$,

The Law of Large Numbers says that if X_1, X_2, X_3, \dots are iid. with mean μ , and $S_n = X_1 + \dots + X_n$, then

$$\frac{S_n}{n} \rightarrow \mu \quad \text{i.e.} \quad \frac{S_n - n\mu}{n} \rightarrow 0$$

$$\text{i.e.} \quad S_n - n\mu = o(n).$$

At what exact rate n^α does $S_n - n\mu = O(n^\alpha)$?

Central Limit Theorem

Let $X_1, X_2, \dots, X_n, \dots$ be iid random variables with $E(X_j) = \mu$, $\text{Var}(X_j) = \sigma^2$.

Then $S_n - n\mu = O(\sqrt{n})$, and for $-\infty < a \leq b < \infty$,

$$\lim_{n \rightarrow \infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \in [a, b]\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(b) - \Phi(a).$$
$$\sigma\sqrt{n} = \sqrt{\sigma^2 n} = \sqrt{\text{Var}(S_n)} = \sqrt{\text{Var}(S_n - n\mu)}$$

Proof. Let $Y_n = (S_n - n\mu) / \sigma\sqrt{n}$. We will show that

$$M_{Y_n}(t) \rightarrow M_{N(0,1)}(t) = e^{t^2/2} \text{ for all } t \in \mathbb{R}.$$

$$M_{Y_n}(t) = E(e^{tY_n}) = E(e^{t \cdot \frac{1}{\sigma\sqrt{n}} (S_n - n\mu)})$$

$$= E(e^{\frac{t}{\sigma\sqrt{n}} (\overset{\circ}{X}_1 + \overset{\circ}{X}_2 + \dots + \overset{\circ}{X}_n)})$$

$$= E(e^{\frac{t}{\sigma\sqrt{n}} \overset{\circ}{X}_1} e^{\frac{t}{\sigma\sqrt{n}} \overset{\circ}{X}_2} \dots e^{\frac{t}{\sigma\sqrt{n}} \overset{\circ}{X}_n})$$

↑ ↑ ... ↑
independent.

$$= E(e^{\frac{t}{\sigma\sqrt{n}} \overset{\circ}{X}_1}) \dots E(e^{\frac{t}{\sigma\sqrt{n}} \overset{\circ}{X}_n})$$

$$= M_{\overset{\circ}{X}_1}(\frac{t}{\sigma\sqrt{n}}) \dots M_{\overset{\circ}{X}_n}(\frac{t}{\sigma\sqrt{n}})$$

$$= M_{\overset{\circ}{X}_1}(\frac{t}{\sigma\sqrt{n}})^n$$

$$\begin{aligned} S_n - n\mu &= \underbrace{(X_1 - \mu)}_{\overset{\circ}{X}_1} + \underbrace{(X_2 - \mu)}_{\overset{\circ}{X}_2} + \dots + \underbrace{(X_n - \mu)}_{\overset{\circ}{X}_n} \\ &= \overset{\circ}{X}_1 + \overset{\circ}{X}_2 + \dots + \overset{\circ}{X}_n \end{aligned}$$

$$\overset{\circ}{X}_j = X_j - \mu$$

i.i.d., $E(\overset{\circ}{X}_1) = 0$

$$E(\overset{\circ}{X}_1^2) = E((X_1 - \mu)^2)$$

$$= \text{Var}(X_1) = \sigma^2$$

$$\begin{aligned} M_{\overset{\circ}{X}_1}(s) &= 1 + \frac{E(\overset{\circ}{X}_1)}{1!} s + \frac{E(\overset{\circ}{X}_1^2)}{2!} s^2 + \frac{E(\overset{\circ}{X}_1^3)}{3!} s^3 + \dots = 1 + \frac{\sigma^2}{2} s^2 + O(s^3) \\ M_{\overset{\circ}{X}_1}(\frac{t}{\sigma\sqrt{n}}) &= 1 + \frac{\sigma^2}{2} \frac{t^2}{\sigma^2 n} + O(\frac{1}{n^{3/2}}) \Rightarrow M_{\overset{\circ}{X}_1}(\frac{t}{\sigma\sqrt{n}})^n = (1 + \frac{t^2}{2n} + O(n^{-3/2}))^n \end{aligned}$$