STATISTICS CLASS 3

- Agenda for today's class
  - Correlation and causation
  - MGF
  - Central limit theorem
  - Law of large no:
  - ANOVA (One-way)

- Correlation and causation

$$\rho_{xy} = \text{corr}(x,y) = \frac{\text{cov}(x,y)]}{\sigma_x \cdot \sigma_y}$$

$$\rho_{xy} = \frac{E[(x-\bar{x})(y-\bar{y})]}{\sigma_x \, \sigma_y} = \underbrace{E\left[\overbrace{\left(\frac{x-\bar{x}}{\sigma_x}\right)}^{\tilde{x}} \cdot \overbrace{\left(\frac{y-\bar{y}}{\sigma_y}\right)}^{\tilde{y}}\right]}_{\text{standardize first and then get covar.}}$$

[Cauchy]

$\begin{bmatrix} \text{Schwarz} \\ \text{Inequality} \end{bmatrix}$ $\left(E[XY]\right) \leq E[X^2] \cdot E[Y^2]$

Properties of correlation:

(a) $\rho \in [-1, 1]$ dimensionless

(b) If X & Y are independent

$$\text{cov}(X,Y) = 0, \; \rho_{XY} = 0$$

But $\text{cov}(X,Y) = \rho_{XY} = 0 \not\Rightarrow$ Independence

(c) Correl. only captures linear relationship

$X \sim N(0,1)$, $Y = X^2$

$$\text{cov}(X,Y) = \text{cov}(X, X^2) = E[X \cdot X^2] - E[X] \cdot E[X^2]$$

$$E[\cancel{X^3}]^{\to 0} - E[\cancel{X}]^{\to 0} E[X^2] = 0$$

(d) $V[X+Y] = V[X] + V[Y] + 2 \cdot \text{cov}(X,Y)$

$$= V[X] + V[Y] + 2 \cdot \sigma_X \sigma_Y \cdot \rho_{XY}$$

$= V[X] + V[Y]$

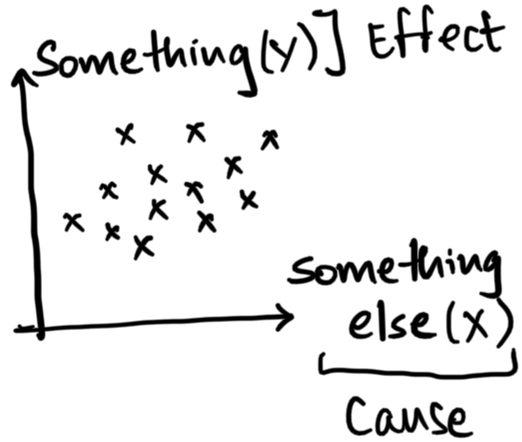
(e) $corr(X, aY) = corr(X, Y)$

(f) $corr(X, Y+c) = corr(X, Y)$

Studying causal relationships is the fundamental problem of econometric inference.

$\nwarrow y \sim x \nearrow$

Grades		Weather
Sun burn		Ice cream sales
Crime	$\rightleftarrows$	police spending



Wage	$\rightleftarrows$	Education]
Sales		Ad Exposure
inflation		unemployment

- Causation goes a step further than correl.

change $x \to$ change y ?

\- Establishing causality
- → Experiment (gives control)
- → Cause occurs before effect
- → If X occurs, Y also occurs
- → No alternate explanations

- Moment Generating Function (MGF)

$$\mu = E[X] = \int x \cdot f_X(x)\, dx$$

$$\sigma^2 = V[X] = E[X^2] - (E[X])^2$$

MGF for a RV X is defined as

$$M_X(t) = E[e^{tX}] = \begin{cases} \sum e^{tk} \cdot P(X=k) \\ \int_{-\infty}^{\infty} e^{tx} \cdot f_X(x) \cdot dx \end{cases}$$

Why does the MGF work?

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \frac{t^3}{3!} \cdot X^3 + \cdots$$

$$E[e^{tX}] = 1 + t\, E[X] + t^2\, E[X^2] + \cdots \qquad \frac{t^n}{n!} = 1$$

$$E[e^{tx}] = 1 + \underline{t\,E[x]} + \frac{t^2}{2!}E[x^2] + \cdots$$

$$\mu = \frac{\partial}{\partial t} E[e^{tx}] \Big|_{t=0} = E[x] + \frac{t}{1!}E[x^2] + \cdots \Big|_{t=0} = E[x] + 0$$

$$\frac{\partial^2}{\partial t^2} E[tx] \Big|_{t=0} = E[x^2] + t \cdot E[x^3] + \cdots \Big|_{t=0} = E[x^2]$$

Ex: $X \sim \text{Binom}(n, p)$ $\quad\Big|\quad (x+y)^n = \sum {}^nC_k\, x^k y^{n-k}$

$$M_X(t) = E[e^{tx}] = \sum_{x=0}^{n} e^{tx} \cdot {}^nC_x \cdot p^x (1-p)^{n-x}$$

$$= \sum {}^nC_x (p \cdot e^t)^x (1-p)^{n-x} = \underbrace{(pe^t + \overset{1-p}{q})^n}$$

$$\mu = \frac{\partial}{\partial t} M_t(x) = \underbrace{n \cdot (pe^t + q)^{n-1} \cdot p\cancel{e^t}}\Big|_{t=0} = n \cdot p$$

Properties of MGF:

(a) If $M_X(t) = M_Y(t)$ then $F_X(x) = F_Y(x)$

(b) If $Y = X_1 + X_2 + \cdots + X_n$ } independent RVs

$$M_Y(t) = E[e^{tY}] = E[e^{t(X_1 + X_2 + \cdots + X_n)}]$$

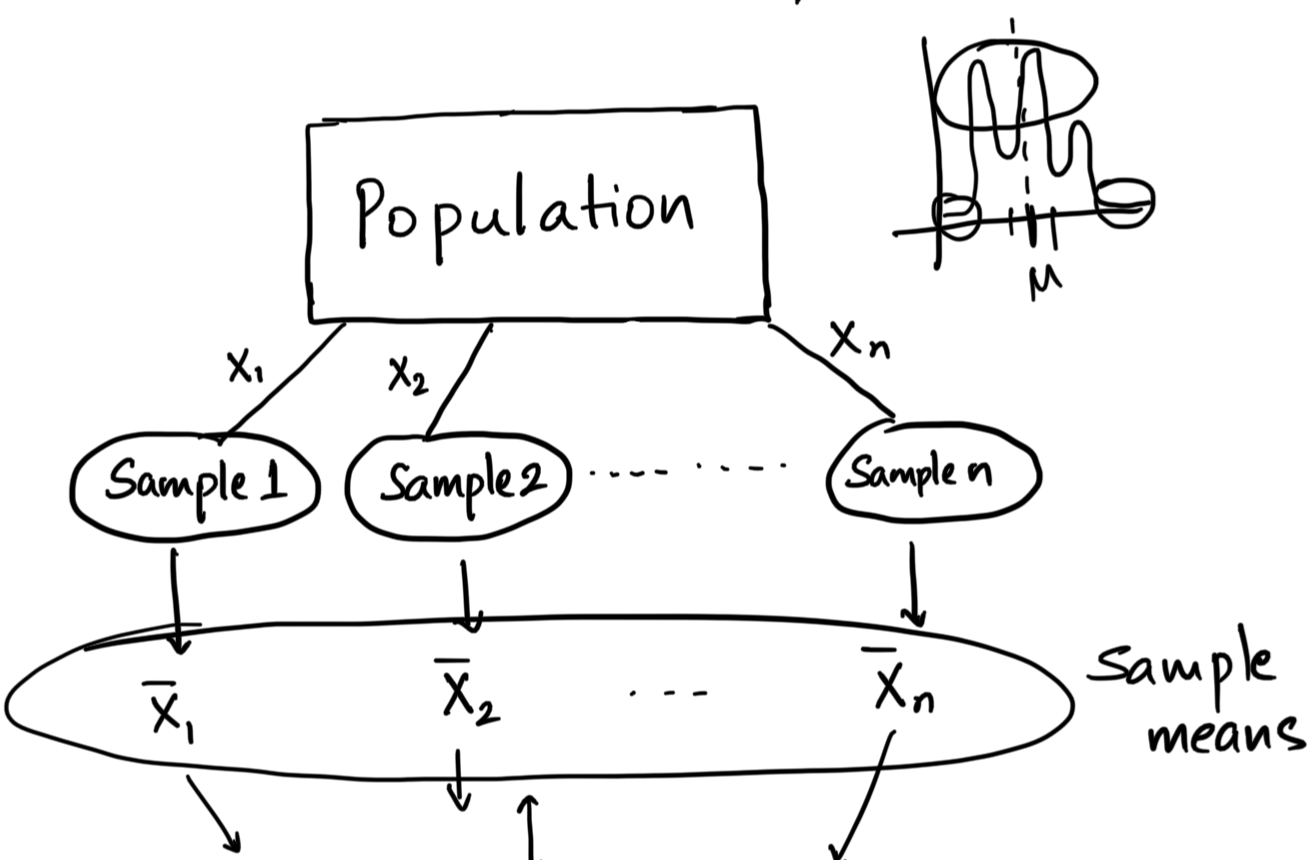
$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdots M_{X_n}(t)$$

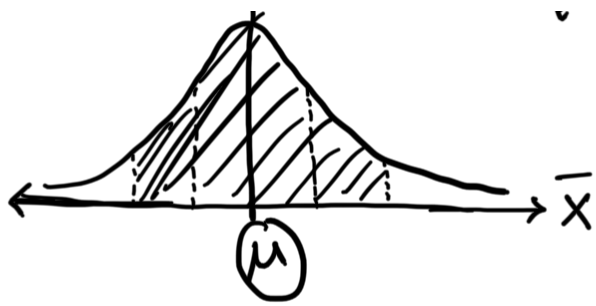
- Law of Large No: and Central Limit Thm.

If $X_1, X_2, \ldots$ are iid, from some population P with a defined mean and variance.

[LOLN]: $\bar{X}_n \to \mu$ as $n \to \infty$ with prob. 1

[CLT]: $\sqrt{n}\left(\dfrac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$





95% confident $\bar{X}_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]$

(OR) $\bar{X}_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]$ with 95% probability

$$\bar{X}(n) = \frac{\bar{X}_1 + \bar{X}_2 + \cdots + \bar{X}_n}{n}$$

LOLN says $\bar{X}(n) \xrightarrow{p} \mu$ as $n \to \infty$ with prob. 1

CLT says $\bar{X}(n) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\bar{X}(n)$ converges to μ at rate $1/\sqrt{n}$

$$E\left[\bar{X}(n)\right] = \frac{\overbrace{E[\bar{X}_1] + E[\bar{X}_2] + \cdots + E[\bar{X}_n]}^{\mu + \mu + \cdots + \mu}}{n} = \frac{n \cdot \mu}{n} = \mu$$

$$V\left[\bar{X}(n)\right] = \frac{\overbrace{V[\bar{X}_1] + V[\bar{X}_2] + \cdots + V[\bar{X}_n]}^{\sigma^2 + \cdots + \sigma^2}}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$E[\bar{X}(n)] = \mu \quad , \quad SD[\bar{X}(n)] = \sigma/\sqrt{n}$$

$$(\text{Standardization}): Z = \frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}}$$

why does central limit theorem work?

- $\bar{X}_i = \mu + \overbrace{\varepsilon_i}^{\text{Error}}$

  When we avg $\bar{X}_i$ the ε_i's cancel

- Low prob. values are not drawn freq.

Applications
- → Hypothesis testing
- → Elections
- → Bootstrapping

- Analysis of Variance (ANOVA)

  ANOVA $\Leftrightarrow$ Dummy variable regression

  Intuition of ANOVA is total variance can

Intuition of ANOVA is

be decomposed into within & across group variation.

$$\chi^2_{n-1} \sim \underbrace{S^2}_{} = \frac{\sum (x_i - \bar{x})^2}{(n-1)}, \quad \underbrace{SST}_{} = \sum (x_i - \bar{x})^2$$

$$F(X) = \left\{ \underbrace{1, 1, 4}_{C1}, \underbrace{4, 5, 6}_{C2}, \underbrace{7, 7, 10}_{C3} \right\}$$

10% 15% 20%

$$F(Y) = \left\{ \underbrace{3, 4, 2}_{C2 - C1}, \underbrace{6, 6, 4}_{C3 - C1} \right\} \quad \mu = \frac{25}{6} = 4.16$$

$$SST = 8.77$$

$C2 - C1 = \{3, 4, 2\}$ $\qquad$ $C3 - C1 = \{6, 6, 4\}$

$\mu_1 = 3$ $\qquad$ $\mu_2 = 16/3$

$SS21 = (-1)^2 + 1^2 = 2$ $\qquad$ $SS31 =$

$$\left. \begin{aligned} \mu_1 &= \mu_2 = \mu_M \\ \mu_1 - \mu_M &= \mu_2 - \mu_M \end{aligned} \right\}$$

Q ANOVA asks – is there a diff. b/w the 3 classes?

$$\text{SS Total} = \text{SS Within} + \text{SS Across}$$

$\mu = \frac{45}{9} = 5$, $SST = (-4)^2 \times 2 + 1^2 \times 2 + 0 + 1^2 + 2^2 \times 2 + 5^2 = 68$

$C1 = \{1, 1, 4\}$ $\quad C2 = \{4, 5, 6\}$ $\quad C3 = \{7, 7, 10\}$

$\mu_1 = 2$ $\quad \mu_2 = 5$ $\quad \mu_3 = 8$

$SS1 = (-1)^2 + (-1)^2 + 2^2$ $\quad SS2 = (-1)^2 + 0 + 1^2$ $\quad SS3 = 1^2 + 1^2 + 2^2$

$SS1 = 6$ $\quad SS2 = 2$ $\quad SS3 = 6$] SSW

$$\text{SS Across} = \sum n_i (\mu_i - \mu)^2$$

$$= \underbrace{3\overset{2-5}{(-3)^2} + 3 \times 0 + 3 \times \overset{8-5}{3^2}} = 54$$

$\text{SS Total} = 68$

$\text{SS Within} = 6 + 2 + 6 = 14$

$\text{SS Across} = 54$

$\rightarrow \chi^2$

$$F = \frac{\boxed{\text{Mean Sq. Across}}}{\boxed{\text{Mean sq. Within}}} = \frac{'SSA' / (k-1)}{\underbrace{SSW}_{} / (n-k)}$$

χ^2

F-stat

10^6

$$H_0 : \underbrace{\mu_1 \neq \mu_2}_{} \overbrace{= \cdots = \mu_n}^{=0} \Big] \quad F = 0$$

$$\mu_1 = \mu_2 = \mu_M$$

$$\mu_1 - \mu_M = \mu_2 - \mu_M$$

Ratio of 2 χ^2 distributions (iid) has F-dist.

$$\underbrace{U_1 \sim \chi^2_n}_{}, \quad \underbrace{U_m \sim \chi^2_m}_{}$$

$$F = \frac{U_1/n}{U_2/m} \sim \text{F-dist. with } \underbrace{\text{n and m deg.}}_{} \text{ of freedom}$$

$$F = \frac{54/3-1}{14/\underbrace{9-3}_{}} \approx \underbrace{11.5}_{} \quad \Big| \quad F = 1$$

Larger F value, more variation across groups

σ^2

(a) Random sampling from pop^n

SS Across << SS Within

$$\Rightarrow 0 \leq \underbrace{F \leq 1}$$

(b) Ex: here would be comparing diff. portfolio returns!

How diff. are means of diff. portfolios?