Computer Science

CSC380: Principles of Data Science

Course wrap-up 1

Xinchen Yu

Announcements

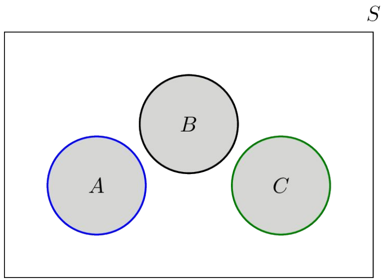
- Final exam
  - Time: Dec 13, 3:30 - 5:30pm
  - Location: C E Chavez Bldg, Rm 111 (same room)
  - What you can bring:
    - one letter size cheat sheet, you can use double sides
    - calculator (not necessary)

- Fill out SCS (https://scsonline.oia.arizona.edu/) – if 80% responses, will add 5 points to the homework with lowest grade (48% right now).

Probability

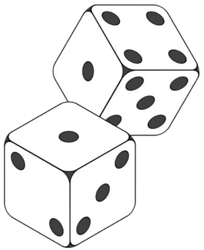
Probability

- Basic definitions: outcome space, events



- Probability P: maps events to [0, 1] values
  - Three axioms
  - Axiom 3: additivity

- Special case of P: each outcomes is equally likely

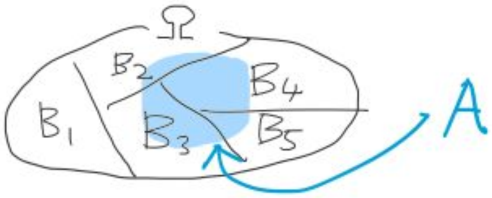


$$P(E) = \frac{|E|}{|\Omega|}$$

Number of elements in event set

Number of possible outcomes (36)

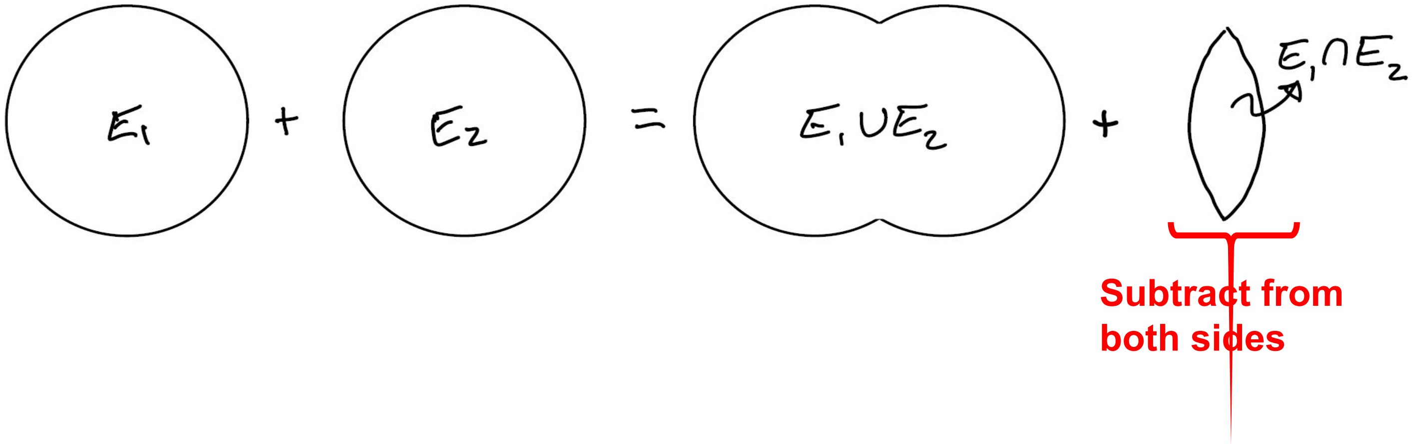
- distributive law, inclusion-exclusion rule; law of total probability



Lemma: (inclusion-exclusion rule) *For any two events* E_1 *and* E_2,

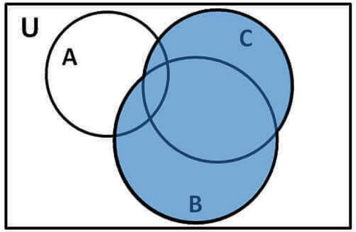
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Graphical Proof:

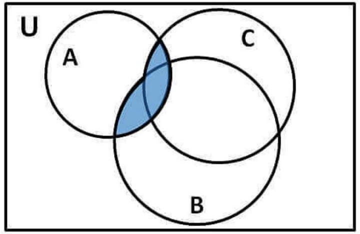


Probability

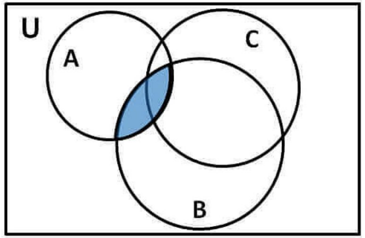
- $A \cap (B \cup C) = (A \cap \mathrm{B}) \cup (A \cap \mathrm{C})$ // distributive law



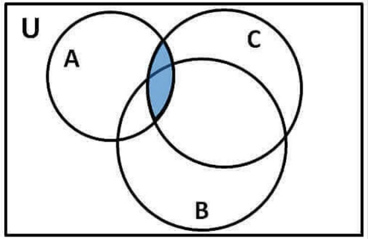
(B ∪ C)



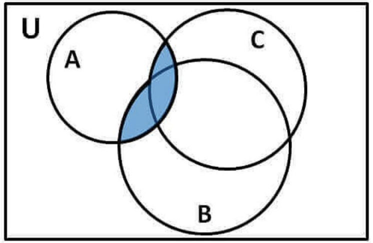
A ∩ (B ∪ C)



(A ∩ B)

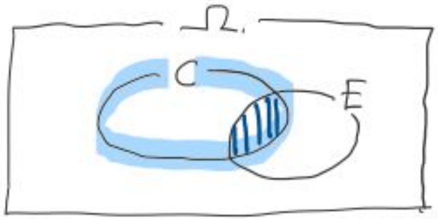


(A ∩ B)



(A ∩ B) ∪ (A ∩ C)

Probability



- $P(E \cap C) = P(E|C)P(C) = P(C|E)P(E)$

- Conditional probability
  - Chain rule

$$P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C)$$

  - Chain rule + law of total probability

$$P(A) = \sum_i P(A, B_i) = \sum_i P(B_i)P(A|B_i)$$

  - Bayes rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Independence of events:

$$P(A,B) = P(A)P(B)$$

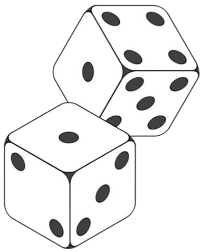
When we have two events A and B…

- Conditional probability: $P(A|B), P(A^c|B), P(B|A)$ etc.

- Joint probability: $P(A,B)$ or $P(A^c,B)$ or ...

- Marginal probability: $P(A)$ or $P(A^c)$

Probability



- Discrete random variable X (e.g., sum of two dice)
  - Representation of its distribution: probability mass function (PMF)
  - Tabular representation of joint distribution of 2 RVs (X,Y)
  - PMF of XY, X+Y given independence

- RVs: law of total probability, conditional probability, chain rule, bayes rule, independence, conditional independence

$$p(Y \mid Z) = \sum_x p(Y, X = x \mid Z)$$

$$p(X = x, Y = y \mid Z = z) = p(X = x \mid Z = z)p(Y = y \mid Z = z)$$

- Useful discrete distributions
  - Uniform
  - Bernoulli
  - Binominal

Probability

- Moments of random variables: expectation, variance, covariance

- Calculate mean (expectation) and variance of RVs
    - Linearity of expectation: E[X + cY] = E[X] + cE[Y] for constant c
    - E[X^2]
    - E[XY]
        - If independent: E[X]E[Y]
        - If not independent: $\mathrm{E}[XY] = \sum_{(x,y)} \sum xy \cdot p(x,y)$
    - E[X | Y = y]
    - Var[cX]
    - Var[X+Y] when independent

- Expectation and variance of useful distributions
    - Bernoulli
    - Gaussian

Probability

- Expectation

$$E[X] = \sum_x x \cdot p(X = x)$$

- Properties

$$E[X+Y] = E[X] + E[Y]$$
$$E[cX] = cE[X]$$
$$E[c] = c$$

c is a constant

- Conditional expected value

$$E[X|Y = y] = \sum_x x \cdot p(X = x|Y = y)$$

- Variance

$$Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

- Properties

$$Var[cX] = c^2 Var[X]$$

- Covariance

$$\begin{aligned} Cov(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

$$Cov(X,X) = E[X^2] - E[X]E[X] = Var(X)$$

- Variance of $X + Y$

$$Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$$

Theorem: *If* $X \perp Y$ *then* $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$ *.*

<u>Comparison</u>: $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ regardless of independence!

Theorem: *If* $X \perp Y$ *then* $\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$$

$$\mathbf{Cov}(X,Y) = 0$$

Probability

Find the Marginal PMFs of X and Y.

$$P_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_X(1) = \frac{2}{5} + 0 = \frac{2}{5},$$
$$P_Y(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_Y(1) = \frac{2}{5} + 0 = \frac{2}{5}.$$

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{1}{5}$	$\frac{2}{5}$
$X = 1$	$\frac{2}{5}$	0

Probability

Find the conditional PMF of X given Y=0 and Y=1

	$Y=0$	$Y=1$
$X=0$	$\frac{1}{5}$	$\frac{2}{5}$
$X=1$	$\frac{2}{5}$	0

$$P_{X|Y}(0|0) = \frac{P_{XY}(0,0)}{P_Y(0)}$$

$$= \frac{\frac{1}{5}}{\frac{3}{5}} = \frac{1}{3}.$$

$$P_{X|Y}(1|0) = 1 - \frac{1}{3} = \frac{2}{3}.$$

$$X|Y=0 \sim \text{Bernoulli}\left(\frac{2}{3}\right).$$

$$P_{X|Y}(0|1) = 1,$$
$$P_{X|Y}(1|1) = 0.$$

Probability

Let Z=E[X|Y], find the PMF of Z.

	$Y=0$	$Y=1$
$X=0$	$\frac{1}{5}$	$\frac{2}{5}$
$X=1$	$\frac{2}{5}$	0

$$Z = E[X|Y] = \begin{cases} E[X|Y=0] & \text{if } Y=0 \\ E[X|Y=1] & \text{if } Y=1 \end{cases}$$

$$E[X|Y=0] = \frac{2}{3}, \qquad E[X|Y=1] = 0,$$

$$Z = E[X|Y] = \begin{cases} \frac{2}{3} & \text{with probability } \frac{3}{5} \\ 0 & \text{with probability } \frac{2}{5} \end{cases}$$

$$P_Z(z) = \begin{cases} \frac{3}{5} & \text{if } z = \frac{2}{3} \\ \frac{2}{5} & \text{if } z = 0 \\ 0 & \text{otherwise} \end{cases}$$

Probability

Let Z=E[X|Y], find E[Z].

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{1}{5}$	$\frac{2}{5}$
$X = 1$	$\frac{2}{5}$	0

$$P_Z(z) = \begin{cases} \frac{3}{5} & \text{if } z = \frac{2}{3} \\ \frac{2}{5} & \text{if } z = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[Z] = \frac{2}{3} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{2}{5}.$$

Probability

Let Z=E[X|Y], find var(Z).

	Y = 0	Y = 1
X = 0	$\frac{1}{5}$	$\frac{2}{5}$
X = 1	$\frac{2}{5}$	0

$$P_Z(z) = \begin{cases} \frac{3}{5} & \text{if } z = \frac{2}{3} \\ \frac{2}{5} & \text{if } z = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathrm{Var}(Z) &= E[Z^2] - (EZ)^2 \\ &= E[Z^2] - \frac{4}{25}, \end{aligned}$$

$$E[Z^2] = \frac{4}{9} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{4}{15}.$$

$$\begin{aligned} \mathrm{Var}(Z) &= \frac{4}{15} - \frac{4}{25} \\ &= \frac{8}{75}. \end{aligned}$$

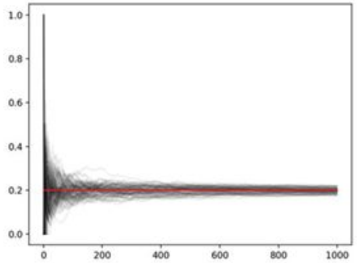
Statistics

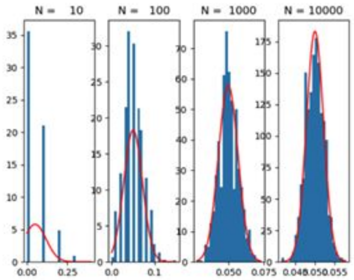
Statistics

- Statistics: make statements about data generation process based on data seen; reverse engineering

- Point estimation
  - Given iid samples $X_1, \ldots, X_n \sim \mathcal{D}_\theta$, estimate θ by constructing *statistics* $\hat{\theta}_n$
  - Basic estimators: sample mean, sample variance
  - Performance measures: unbiasedness, consistency, MSE (efficiency)
  - Bias-variance decomposition:
    - $\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}(\hat{\theta})^2 + \mathrm{var}(\hat{\theta})$

- Useful probability tools:
  - Law of Large Numbers
  - Central Limit Theorem





Statistics

- Sample mean, sample variance

- Sample variance
  - biased version
  - unbiased version
  - how to determine an estimator is biased or unbiased?

- MSE, Bias, Variance
  - how to calculate expectation and variance if there are more than 1 random variable -- use what we learned in probability lecture 5 & 6

Statistics

- Calculate bias and variance

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_n) &= \mathbf{E}[(\hat{\theta}_n - \theta)^2] \\ &= \left(\mathbf{E}[\hat{\theta}] - \theta\right)^2 + \mathbf{E}[(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2] \\ &= \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})\end{aligned}$$

Important properties of Gaussian

- Closed under additivity:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under affine transformation (a and b constant):

$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$

Midterm solutions

9. Given a distribution D with unknown mean μ and variance σ^2, and a set of n iid samples $X_1, \ldots, X_n$ drawn from it. Define $\tilde{\mu}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ as an estimator of μ.
(a) (4 points) Is $\tilde{\mu}_n$ an unbiased estimator of μ? Justify your answer.
(b) (6 points) Let $n = 4$. What is the bias, variance, and Mean Square Error (MSE) of $\tilde{\mu}_4$, respectively? Note: For variance, you can compute $Var[\tilde{\mu}_4]$, in other words, $Var[\frac{X_1+X_2+X_3+X_4}{3}]$.
(You can have μ, σ^2 or numbers in the results).

Lecture statistics 3, page 7

$$\tilde{\mu}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i \qquad E[\tilde{\mu}_n] = E\left[\frac{1}{n-1}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E[X_i]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \mu \quad = \frac{n\mu}{n-1}$$

$\tilde{\mu}_n$ **is not an unbiased estimator of** μ**.**

9. Given a distribution D with unknown mean μ and variance σ^2, and a set of n iid samples $X_1, \ldots, X_n$ drawn from it. Define $\tilde{\mu}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ as an estimator of μ.
(a) (4 points) Is $\tilde{\mu}_n$ an unbiased estimator of μ? Justify your answer.
(b) (6 points) Let $n = 4$. What is the bias, variance, and Mean Square Error (MSE) of $\tilde{\mu}_4$, respectively? Note: For variance, you can compute $Var[\tilde{\mu}_4]$, in other words, $Var[\frac{X_1+X_2+X_3+X_4}{3}]$. (You can have μ, σ^2 or numbers in the results).

$$\tilde{\mu}_4 = \frac{1}{3}(X_1 + X_2 + X_3 + X_4)$$

$$\mathbf{Bias}(\tilde{\mu}_4) = E[\tilde{\mu}_4] - \mu$$
$$= \frac{4\mu}{3} - \mu$$
$$= \frac{\mu}{3}$$

$$\mathbf{Var}[\tilde{\mu}_4] = \mathbf{Var}\left[\frac{1}{3}(X_1 + X_2 + X_3 + X_4)\right]$$
$$= \frac{1}{9}\mathbf{Var}[X_1 + X_2 + X_3 + X_4]$$

Since the X_i are iid:

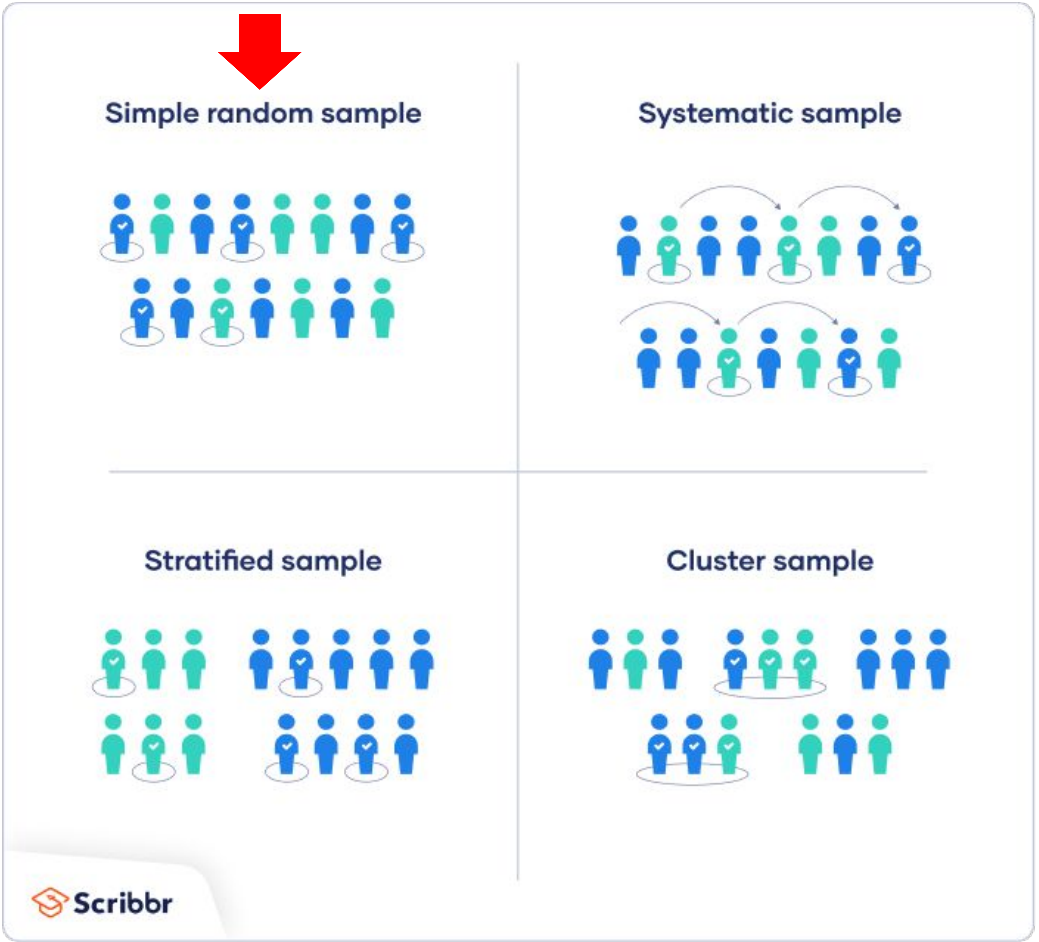
$$= \frac{1}{9}(\mathbf{Var}[X_1] + \mathbf{Var}[X_2] + \mathbf{Var}[X_3] + \mathbf{Var}[X_4])$$
$$= \frac{1}{9}(4\sigma^2)$$

$$\mathbf{MSE}(\tilde{\mu}_4) = \mathbf{Var}[\tilde{\mu}_4] + \mathbf{Bias}(\tilde{\mu}_4)^2 = \frac{4\sigma^2 + \mu^2}{9}$$

Lecture statistics 3, page 8 and 18

Probability Sampling

Probability Sampling



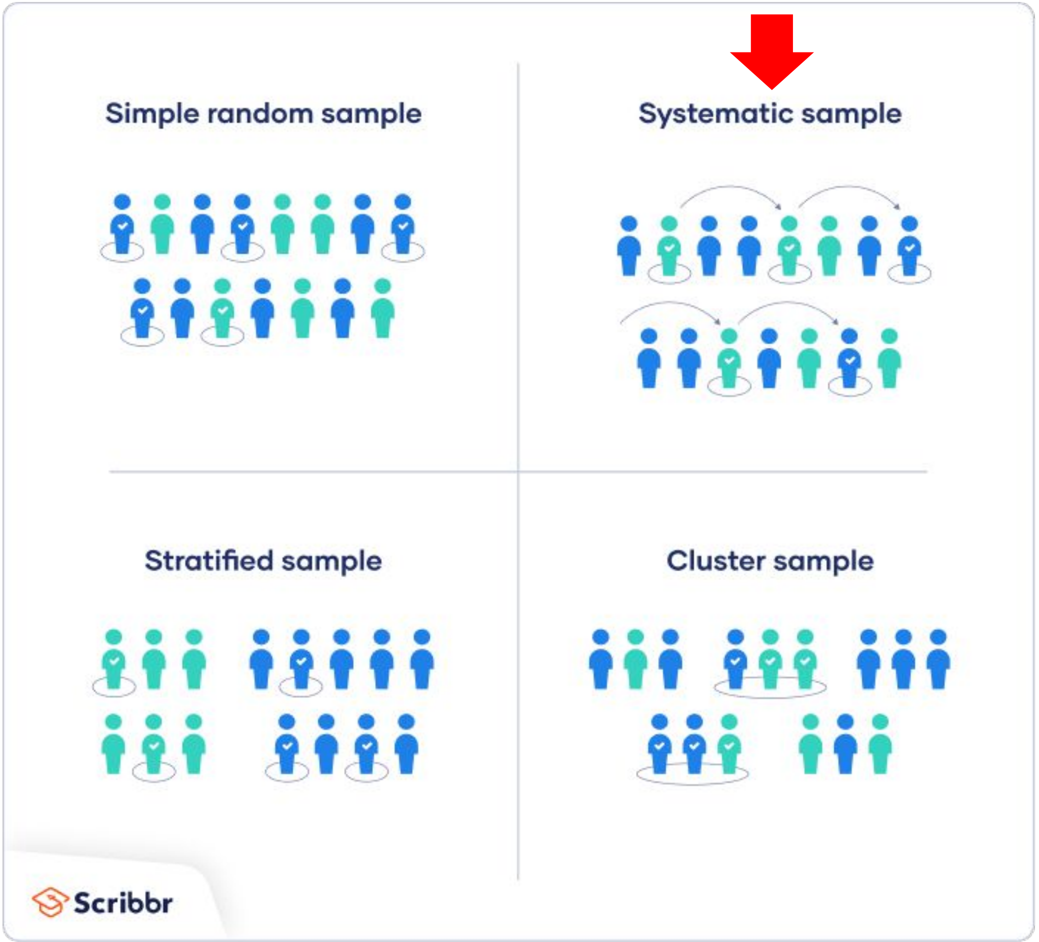
Simple Random Sample (SRS)

Each member of the population has the *same chance* of being selected (i.e., uniform over the population)

Example : American Community Survey (ACS)

Each year the US Census Bureau use *simple random sampling* to select individuals in the US. They follow those individuals for 1 year to draw conclusions about the US population as a whole.

Probability Sampling



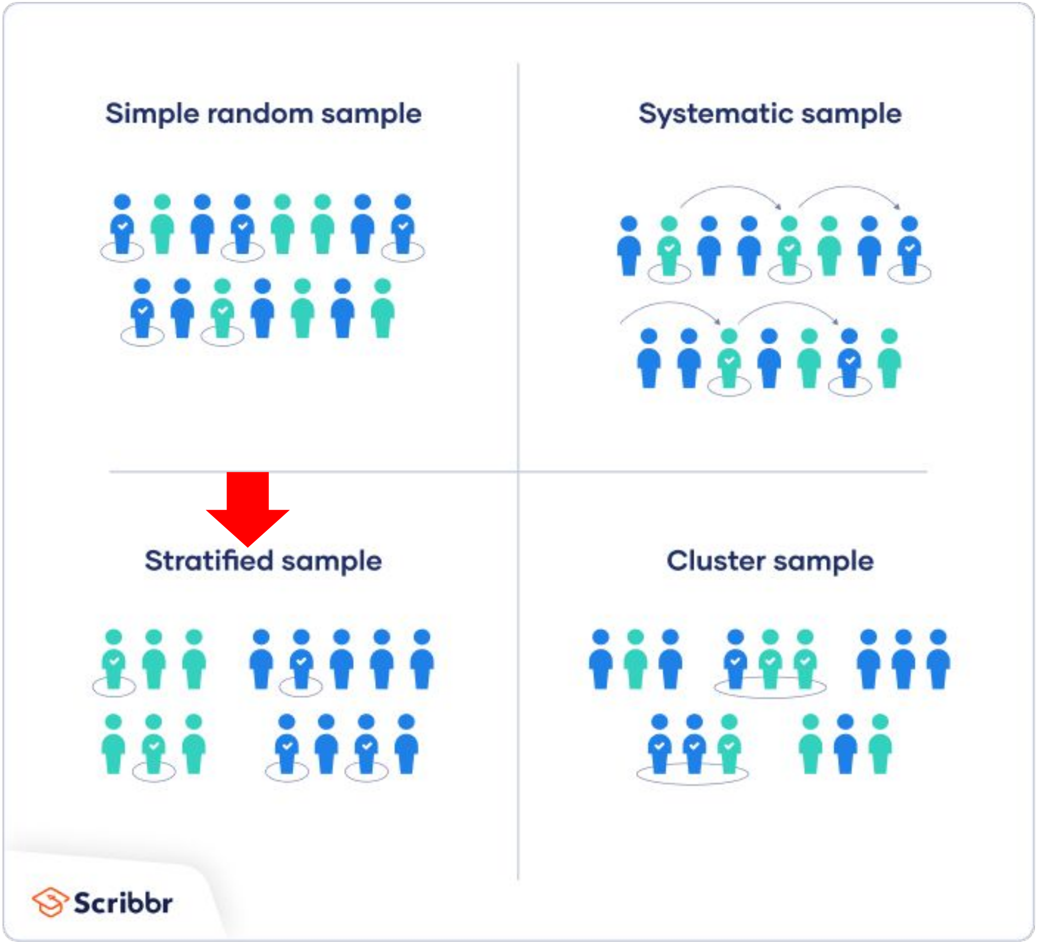
Systematic Sample

Select members of population at a regular interval, determined in advance

Example You own a grocery store and want to study customer satisfaction. You ask *every 20th customer* at checkout about their level of satisfaction.

Note We cannot itemize the whole population in this example, so SRS is not possible.

Probability Sampling

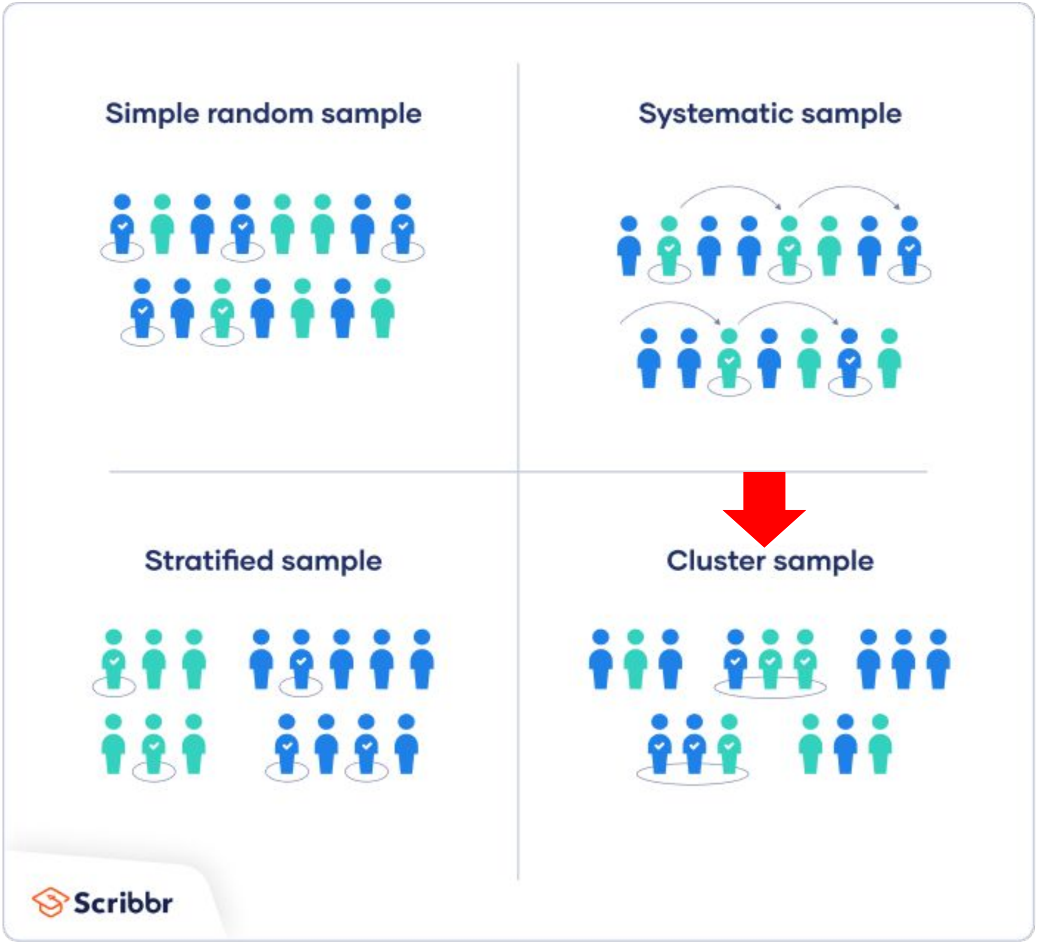


Stratified Sample

Divide population into *homogeneous* subpopulations (strata). Probability sample the strata.

Example We wish to solicit opinions of UA CS freshman by asking 100 of them, but they are about 14% women. SRS could easily fail to capture adequate number of women. We divide into men / women and perform SRS within each group.

Probability Sampling



Cluster Sample

Divide population into subgroups (clusters). Randomly select entire clusters.

Example We wish to study the average reading level of *all 7th graders in the city* (population). Create a list of all schools (clusters) then randomly select a subset of schools and test every student.