Random Variables, Probability Distributions, and Large Sample Theorems

Introduction to Quantitative Social Science

Kosuke Imai
Harvard University / University of Tokyo

Summer 2022

Random Variables and Probability Distributions

- What is a random variable?: assigns a number to an event
  1. Coin flip: head $= 1$ and tail $= 0$
  2. Gambling: win $= \$100$ and lose $= -\$10$
  3. Voting: vote $= 1$ and not vote $= 0$
  4. Survey response: strongly agree $= 4$, agree $= 3$, disagree $= 2$, and strongly disagree $= 1$

- Race prediction example:
  1. race: black $= 1$, white $= 2$, hispanic $= 3$, etc.
  2. residence: lives in precinct $1 = 1$, lives in precinct $2 = 2$, etc.

- Probability distribution: Probability of an event that a random variable takes a certain value
  - $P(\text{race})$: $P(\text{race} = 1)$, $P(\text{race} = 2)$ etc.
  - $P(\text{race} \mid \text{residence})$: $P(\text{race} = 1 \mid \text{residence} = 2)$ etc.

Race Prediction Revisited

- Surname data: $P(\text{surname}),\ P(\text{race} \mid \text{surname})$
- Demographic data: $P(\text{race} \mid \text{residence})$
- We want to compute: $P(\text{race} \mid \text{surname}, \text{residence})$
- Bayes' rule:

$$\begin{aligned} & P(\text{race} \mid \text{surname}, \text{residence}) \\ = \ & \frac{P(\text{surname} \mid \text{race}, \text{residence}) P(\text{race} \mid \text{residence})}{P(\text{surname} \mid \text{residence})} \end{aligned}$$

- Assumption: $P(\text{surname} \mid \text{race}, \text{residence}) \ = \ P(\text{surname} \mid \text{race})$
- Law of total probability:

$$\begin{aligned} & P(\text{surname} \mid \text{residence}) \\ = \ & \sum_{\text{race}} P(\text{surname} \mid \text{race}, \text{residence}) P(\text{race} \mid \text{residence}) \\ = \ & \sum_{\text{race}} P(\text{surname} \mid \text{race}) P(\text{race} \mid \text{residence}) \end{aligned}$$

Probability Model as a Data Generating Process

1. Probability density function (PDF): $f(x)$
   - How likely does X take a particular value?
   - Probability mass function (PMF): When X is discrete, $f(x) = P(X = x)$

2. Cumulative distribution function (CDF): $F(x) = P(X \leq x)$
   - What is the probability that a random variable X takes a value equal to or less than x?
   - Area under the density curve
   - Non-decreasing

Binomial Distribution

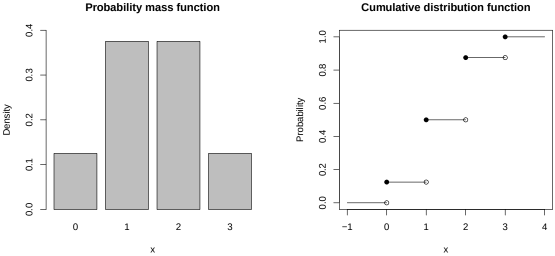
- PMF: for $x \in \{0, 1, \ldots, n\}$,

$$f(x) \;=\; P(X = x) \;=\; \binom{n}{x} p^x (1-p)^{n-x}$$

- CDF: for $x \in \{0, 1, \ldots, n\}$

$$F(x) \;=\; P(X \leq x) \;=\; \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

- Example: flip a fair coin 3 times



People v. Collins Revisited

A purse snatching in which witnesses claimed to see a young women with blond hair in a ponytail fleeing from the scene in a yellow car driven by a black young man with a beard. A couple meeting the description was arrested a few days after the crime, but no physical evidence was found. The probability that a randomly selected couple would possess the described characteristics was estimated to be about one in 12 million. Faced with such overwhelming odds, the jury convicted the defendants. Given that there was already one couple who met the description, what is the conditional probability that there was also a second couple such as the defendants?

1. p: proportion of couples with the characteristics in the
2. $n = 8$ million: total number of couples in the population population
3. A: the event that at least one couple has the characteristics
4. B: the event that at least two couples have the characteristics
5. C: the event that exactly one couple has the characteristics

- What is $P(B \mid A)$?
- Compute $P(A)$, $P(C)$, and then $P(B \mid A)$.

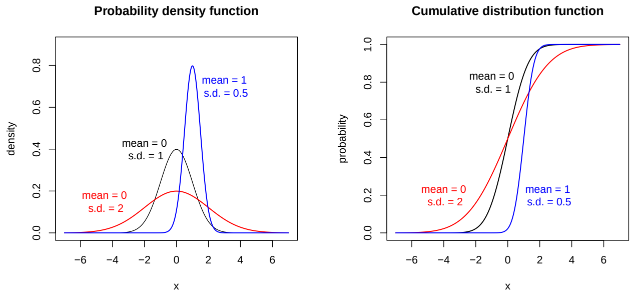
Normal Distribution

- Normal distribution with mean μ and standard deviation σ
- PDF:

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- CDF (no simple formula. use **R** to compute it):

$$F(x) \;=\; P(X \leq x) \;=\; \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)dt$$



Regression Towards the Mean Revisited

- Linear Regression Model:

$$y_i \ = \ 15 + 0.8x_i + \epsilon_i$$

1. y_i: Second take home exam score for student i (percent)
2. x_i: First take home exam score for student i (percent)
3. ϵ_i: error term

- Suppose ϵ_i is normally distributed with mean $= 0$ and s.d. $= 5$

- Two group of students: $x_1 = 60$ and $x_2 = 80$
- Which group of students is likely to do better in the final?
  - when compared with the other group
  - when compared with their own midterm score

Examples Using Normal Distribution

- If X and Y are normal random variables, then $aX + bY$ is also normally distributed with appropriate mean and variance

- z-score:

$$Z = \frac{X - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)}} \sim \mathcal{N}(0, 1)$$

- Sum: X_i is independently distributed as $\mathcal{N}(\mathbb{E}(X), \mathbb{V}(X))$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mathbb{E}(X), n\mathbb{V}(X))$$

- Sample mean:

$$\overline{X} \sim \mathcal{N}\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$

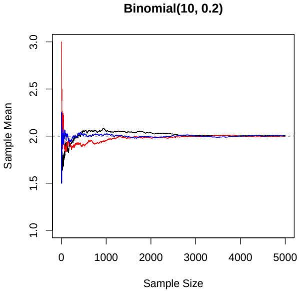
- Regression: $Y_i = -15 + 1.2X_i + \epsilon_i$ with $X_i \sim \mathcal{N}(60, 16)$ and $\epsilon_i \sim \mathcal{N}(0, 25)$
  1. $Y_i \sim \mathcal{N}(57, 48.04)$
  2. Y_i given $X_i = 60$ is $\sim \mathcal{N}(57, 25)$

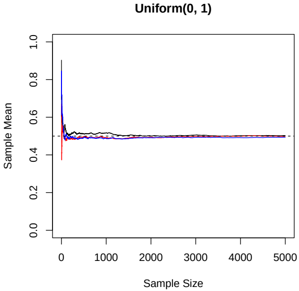
Law of Large Numbers

- As the sample size increases, the sample average of a random variable approaches to its expected value

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}(X)$$

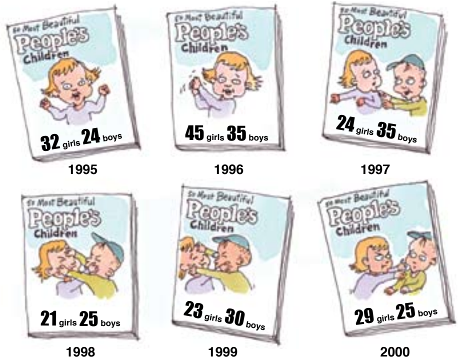
- Example:
  1. flip a coin 10 times and count # of heads
  2. repeat it many times and compute the sample mean





Do Beautiful People Have More Girls?

- In *Journal of Theoretical Biology*,
  1. "Big and Tall Parents have More Sons" (2005)
  2. "Engineers Have More Sons, Nurses Have More Daughters" (2005)
  3. "Violent Men Have More Sons" (2006)
  4. "Beautiful Parents Have More Daughters" (2007)



- Law of Averages in action
  1. 1995: 57.1%
  2. 1996: 56.6
  3. 1997: 51.8
  4. 1998: 50.6
  5. 1999: 49.3
  6. 2000: 50.0
- No dupilicates: 47.7%
- Population frequency: 48.5%

Gelman & Weakliem, *American Scientist*

Central Limit Theorem

- What is the distribution of sample mean $\overline{X}_n$ when X is not normally distributed?
- Polling example: repeated (often hypothetical) polls
- The approximate (asymptotic) distribution of $\overline{X}_n$ is still normal!
- In particular, when n is large, we have

$$\overline{X}_n \ \sim \ \mathcal{N}\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$

- Theorem: As the sample size increases, the distribution of the z-score for the sample mean,

$$Z \ = \ \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{\mathbb{V}(\overline{X})}} \ = \ \frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)/n}}$$

approaches to the standard Normal distribution $\mathcal{N}(0,1)$

Election Polls

- Hypothetically repeated polls with sample size n
- $X_i = 1$ if supports Obama, $X_i = 0$ if supports McCain
- Probability model: $\sum_{i=1}^{n} X_i \sim \mathrm{Binom}(n, p)$

- Obama's support rate: $\overline{X}_n = \sum_{i=1}^{n} X_i / n$
- LLN: $\overline{X}_n \longrightarrow p$ as n tends to infinity
- CLN: $\overline{X}_n \overset{\text{approx.}}{\sim} \mathcal{N}\left(0, \frac{p(1-p)}{n}\right)$ for a large n

- Margin of victory: $\delta = p - (1 - p) = 2p - 1$
- Estimate: $\hat{\delta}_n = 2\overline{X}_n - 1$
- LLN: $\hat{\delta}_n \longrightarrow \delta$ as n tends to infinity
- CLN: $\hat{\delta}_n \overset{\text{approx.}}{\sim} \mathcal{N}\left(0, \frac{4p(1-p)}{n}\right)$ for a large n