Lecture 10: Weak Law of Large Numbers and Central Limit Theorem

Ye Tian

Department of Statistics, Columbia University
Calculus-based Introduction to Statistics (S1201)

July 26, 2022

Recap: Discrete distributions

Bernoulli trials: (like coin tossing)

- Each trial has two outcomes: denoted as S (success) and F (failure).
- For each trial, $\mathbb{P}(S) = p$ and $\mathbb{P}(F) = q = 1 - p$.
- The trials are independent.

Bernoulli distribution: $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$, denoted by $X \sim \text{Bernoulli}(p)$

- $\mathbb{E}X = p$, $\text{Var}(X) = p(1 - p)$

Binomial distribution: **the number of successes** in a Bernoulli process with n trials and probability of success p

- X_i: the Bernoulli r.v. corresponding to i-th Bernoulli trial, then the Binomial r.v. $X = \sum_{i=1}^{n} X_i$.
- pmf: $p(k) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$, where $x = 0, \ldots, n$.
- cdf: $\mathbb{P}(X \leq x) = \sum_{i=0}^{x} \mathbb{P}(X = x)$
- $\mathbb{E}X = np$, $\text{Var}(X) = np(1 - p)$

Recap: Discrete distributions

Geometric distribution: The number of Bernoulli trials until we first get an "S", denoted by $X \sim \text{Geometric}(p)$.

- pmf: $p(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\underbrace{\text{F.........F}}_{(x-1)\text{ F's}}\ \text{S}\}) = q^{x-1}p$, where $x = 1, 2, \ldots$
- cdf: $F(x) = 1 - \mathbb{P}(X > x) = 1 - \mathbb{P}(\{\underbrace{\text{F.........F}}_{x\text{ F's}}\}) = 1 - q^x$
- $\mathbb{E}X = 1/p$, $\text{Var}(X) = 1/p^2$

Hyper-geometric distribution: There are N products in total with M defect ones and $N - M$ good ones. We sample n from these N products **without replacement**. X be **the number of defect products**

- pmf: $\mathbb{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$, if $\max\{0, n - N + M\} \leq x \leq \min\{n, M\}$ and $\mathbb{P}(X = x) = 0$ elsewhere.
- $\mathbb{E}X = n \times \frac{M}{N}$.

Recap: Continuous distribution

Uniform distribution: (on $[A, B]$) if:

- pdf is $f(x) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{elsewhere} \end{cases}$
- cdf is $F(x) = \begin{cases} 0, & x \leq A \\ \frac{x-A}{B-A}, & A < x \leq B \\ 1, & x > B \end{cases}$

Normal distribution (or Gaussian distribution): $X \sim N(\mu, \sigma^2)$

- $N(0, 1)$: **standard normal distribution**, pdf ϕ, cdf Φ
- pdf: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < +\infty$
- $\mathbb{E}X = \mu$, Var$(X) = \sigma^2$.
- pdf f is symmetric around μ: $f(x - \mu) = f(x + \mu)$ for any x
- (Standardization) $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.
- ϕ is symmetric around 0: $\phi(x) = \phi(-x)$ for any x.
- $\Phi(-x) = 1 - \Phi(x)$ for any x

Recap: Sum of independent normal variables

Suppose $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and $X \perp\!\!\!\perp Y$.
Theorem: $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Warning: This only holds when X and Y are independent!
Counter-example when $X \not\perp\!\!\!\perp Y$: $Y = -X$

Consequence:

- $X_i \sim N(\mu_i, \sigma_i^2)$. Then $\sum_{i=1}^n X_i \sim N(\sum_i^n \mu_i, \sum_{i=1}^n \sigma_i^2)$
- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Then $\frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n)$

Proof: By induction: $\sum_{i=1}^n X_i \sim N(\tilde{\mu}, \tilde{\sigma}^2)$. Next we will find $\tilde{\mu}$ and $\tilde{\sigma}^2$.
Since X_i's are independent: $\tilde{\mu} = \mathbb{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mu_i$.
$\tilde{\sigma}^2 = \text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i) = \sum_{i=1}^n \sigma_i^2$.

Review the "linearity" of expectation and variance from last lecture:

Suppose $X_1, \ldots, X_n$ are independent. $a_1, \ldots, a_n, b$ are constants.

- $\mathbb{E}[\sum_{i=1}^n (a_i X_i) + b] = \sum_{i=1}^n (a_i \mathbb{E} X_i + b)$
- $\text{Var}[\sum_{i=1}^n (a_i X_i) + b] = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$

Recap: Continuous distribution

Exponential distribution: $X \sim \mathsf{Exp}(\lambda)$

- pdf: $f(x) = \lambda \exp(-\lambda x), \quad x \geq 0$
- cdf: $F(x) = 1 - \exp(-\lambda x), \quad x \geq 0.$
- $\mathbb{E}X = \lambda^{-1}$, $\mathsf{Var}(X) = \lambda^{-2}$
- Memoryless property: $\mathbb{P}(X > a + x | X > a) = \mathbb{P}(X > x)$

Summary: some commonly used distributions

Discrete distributions:

- Bernoulli distribution
- Binomial distribution
- Geometric distribution
- Hyper-geometric distribution
- Poisson distribution

Continuous distributions:

- Uniform distribution
- Normal distribution
- Exponential distribution
- χ^2-distribution
- t-distribution
- F-distribution

Weak Law of Large Numbers

Recap: Population universe and sample universe

Sample universe (what we see)

Data ⇒

- Sample mean
- Sample variance
- Sample standard deviation
- Sample quantiles (sample median, quartiles...)
- Sample maximum/minimum
- Sample covariance/correlation
- Bar chart (discrete r.v.)
- Histogram (continuous r.v.)

Population universe (inaccessible)

Probability distribution ⇒

- Expectation/Mean
- Variance
- Standard deviation
- Quantiles (median, quartiles ...)
- Maximum/minimum
- Covariance/correlation
- pmf (discrete r.v.)
- pdf (continuous r.v.)

Rationale: When we have enough samples and our model assumption is close to the actual distribution, then the sample-version data summaries will be close to the underlying population-version ones.

Weak law of large numbers

Theorem: X_1, ..., X_n are i.i.d. r.v.'s following the same distribution of X. If $\mathbb{E}X$ exists[1], then $\bar{X} = \frac{1}{n}\sum_{i=1} X_i$ satisfies that for any $\epsilon > 0$,

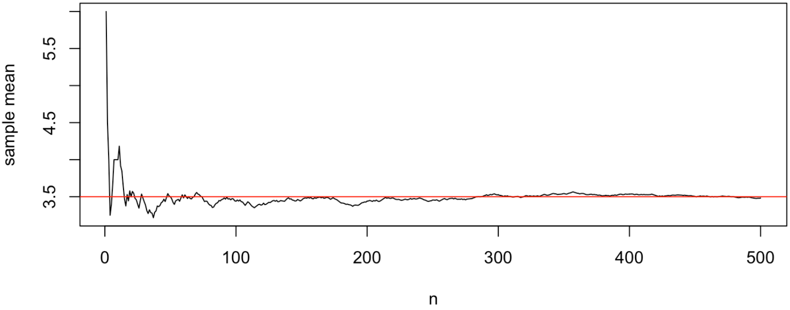
$$\mathbb{P}(|\bar{X} - \mathbb{E}X| > \epsilon) \approx 0, \quad \text{when } n \text{ is large.}$$

- If you don't understand this probability, no worries! You can understand it as $\bar{X} \approx \mathbb{E}X$ when the number of samples n is very large.
- This implies that sample mean is close to the population mean (expectation) when n is large.
- This also implies that sample variance and standard deviation are close to the population ones when n is large.
  To see this, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\bar{X}^2 \approx \mathbb{E}X^2 - (\mathbb{E}X)^2 = \text{Var}(X)$. And $s \approx \sqrt{\text{Var}(X)} = \text{SD}(X)$.
- If X is discrete, then relative frequency at $X = x$ equals $\frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i = x) \approx \mathbb{E}[\mathbb{1}(X = x)] =$ the pmf at $X = x$ i.e. $\mathbb{P}(X = x)$

[1] This holds for all the distributions we have seen, e.g. Bernoulli, normal, exponential, binomial, uniform, ...

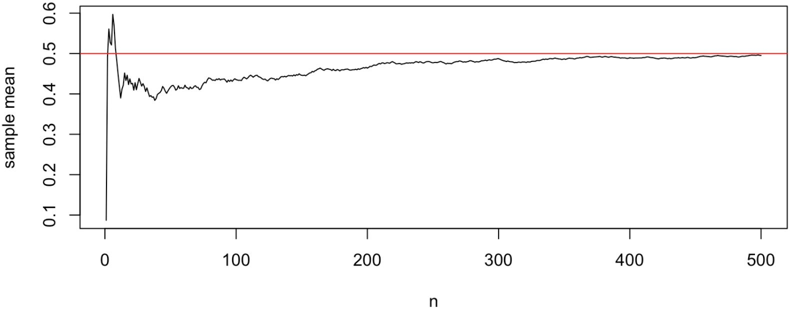
Review: Explanation of the previous examples

Example 1: Rolling a die. What is the expected number of spots X on the top?
$\mathbb{P}(X=1)=\cdots=\mathbb{P}(X=6)=\frac{1}{6}\Rightarrow$
$\mathbb{E}X=1\times\frac{1}{6}+2\times\frac{1}{6}+\cdots+6\times\frac{1}{6}=3.5.$



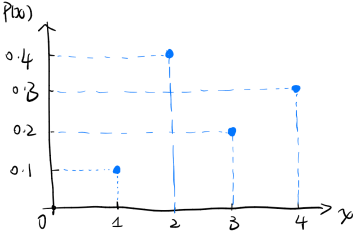
Review: Explanation of the previous examples

Example 2: X follows a uniform distribution on $[0, 1]$. (i.e. pdf $f(x) = 1$ on $[0, 1]$ and 0 elsewhere)
$\mathbb{E}X = \int_0^1 xf(x)dx = \int_0^1 xdx = \frac{1}{2}x^2 \mid_0^1 = \frac{1}{2}$.

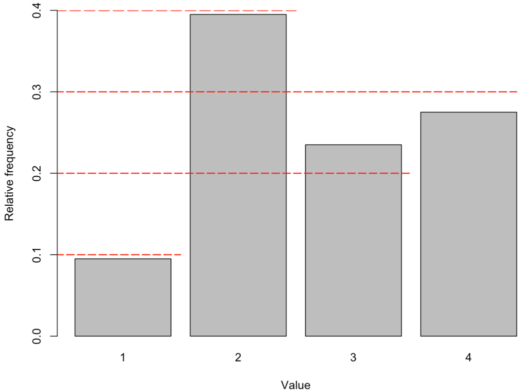


Review: Relative frequency bar chart and pmf

Suppose a r.v. X has distribution with this pmf. I sampled X_1, X_2, ..., X_{1000} **independently** from this distribution.



pmf

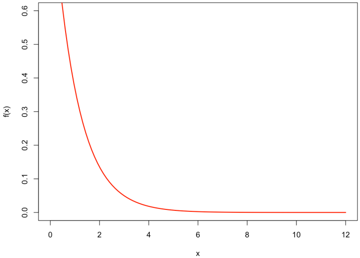


histogram

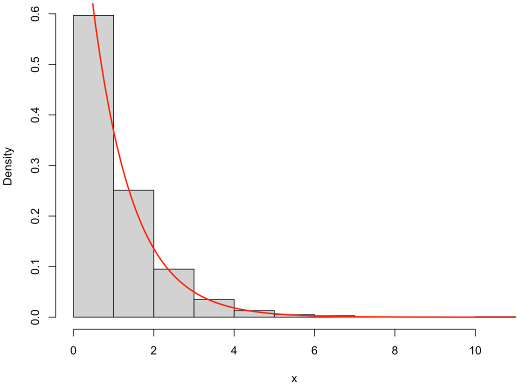
- Empirical relative frequency is an approximation of pmf.
- If we sample **infinite** points, the relative frequency will equal pmf.
- ~~We will discuss more on this next week.~~ Now you know it's due to ...

Review: Density histogram and pdf

Suppose a r.v. X has distribution with this pmf. I sampled X_1, X_2, ..., X_{1000} **independently** from this distribution.



pdf



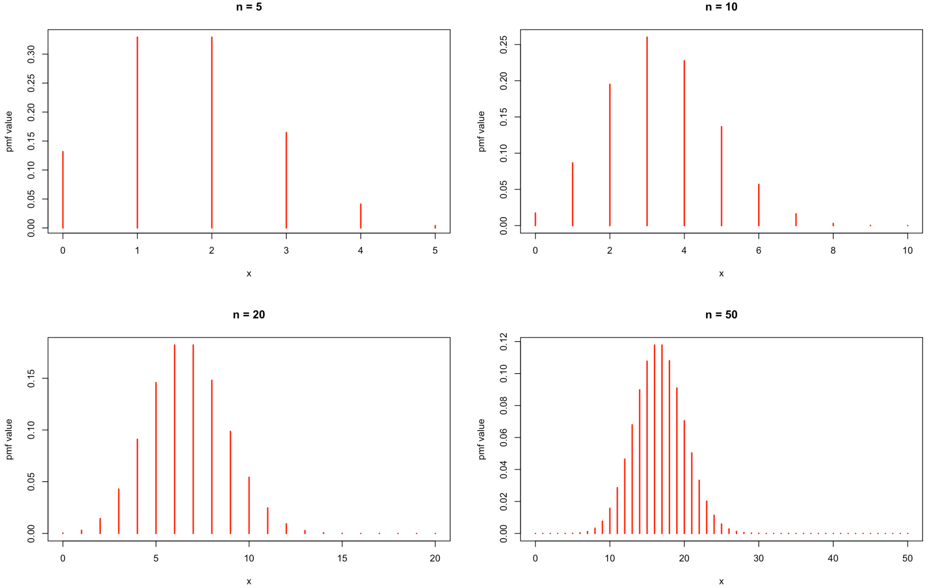
histogram

- Empirical density histogram is an approximation of pdf.
- If we sample **infinite** points and the bin width is **infinitely small**, the density histogram will be the same as the pdf curve.

Central Limit Theorem

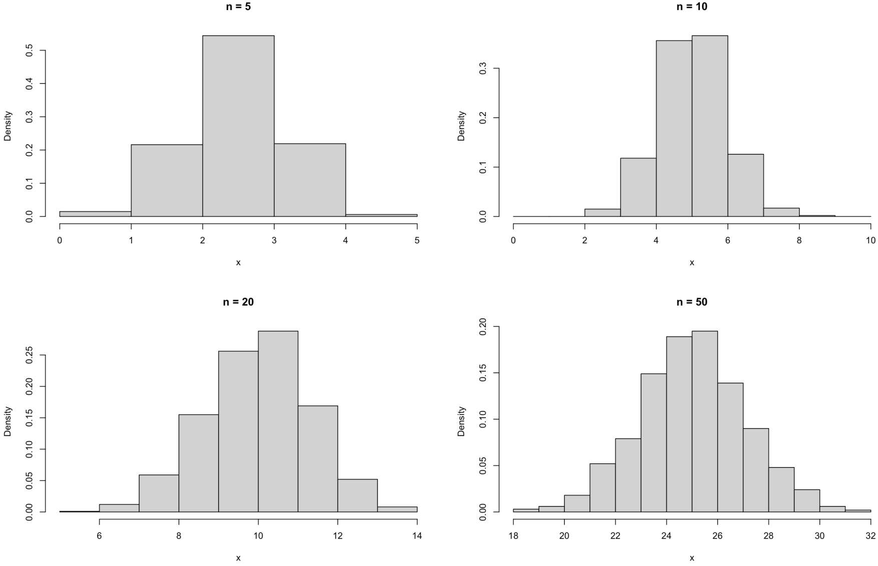
Binomial pmf when n is large

pmf of Bin$(n, 1/3) = \sum_{i=1}^{n} X_i$, where $X_i \overset{i.i.d.}{\sim}$ Bernoulli$(1/3)$



Histogram of sum of uniform r.v's

Histogram of $\sum_{i=1}^{n} X_i$, where $X_i \overset{i.i.d.}{\sim} \text{Unif}(0,1)$



Central limit theorem

Our guess: the sum of i.i.d. (independently and identically distributed) r.v.'s is approximately normal distributed.

The guess is true!!!

Theorem: X_1, ..., X_n are i.i.d. r.v.'s following the same distribution of X. If $\mathbb{E}X = \mu$ and Var$(X) = \sigma^2$ exists[2], then $\bar{X} = \frac{1}{n}\sum_{i=1} X_i$ satisfies

$$\mathbb{P}\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq x\right) \approx \Phi(x), \quad \text{when } n \text{ is large,}$$

where Φ is the cdf of standard normal distribution $N(0,1)$.

- This is saying that $\bar{X}$ approximately follows $N(\mu, \sigma^2/n)$ when n is large Recall that $\mathbb{E}(\bar{X}) = \mu$, Var$(\bar{X}) = \sigma^2/n$.
- When X_1, ..., $X_n \sim N(\mu, \sigma^2)$, "$\approx$" will be "$=$" (by properties of normal distribution).
- Equivalently, $\sum_{i=1}^n X_i$ approximately follows $N(n\mu, n\sigma^2)$.

[2]This holds for all the distributions we have seen, e.g. Bernoulli, normal, exponential, binomial, uniform, ...

Central limit theorem

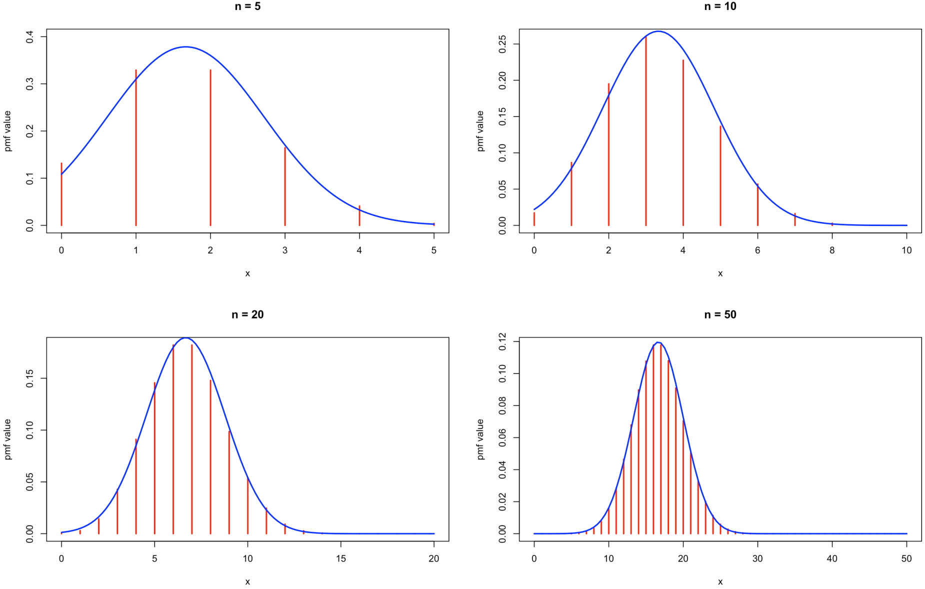
Example 1: X_1, ..., $X_n \overset{i.i.d.}{\sim}$ Bernoulli(p). Then
$n\bar{X} = \sum_{i=1}^n X_i \sim \text{Bin}(n,p)$. And $\mathbb{E}X_i = p$, $\text{Var}(X_i) = p(1-p)$.
By CLT,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \overset{d}{\approx} N(p, p(1-p)/n),$$
$$\sum_{i=1}^n X_i \overset{d}{\approx} N(np, np(1-p)).$$

Notes: "$\overset{d}{\approx}$" means "approximately follows ... distribution when n is large"

Binomial pmf when n is large

pmf of $\text{Bin}(n, 1/3) = \sum_{i=1}^{n} X_i$, where $X_i \overset{i.i.d.}{\sim}$ Bernoulli$(1/3)$



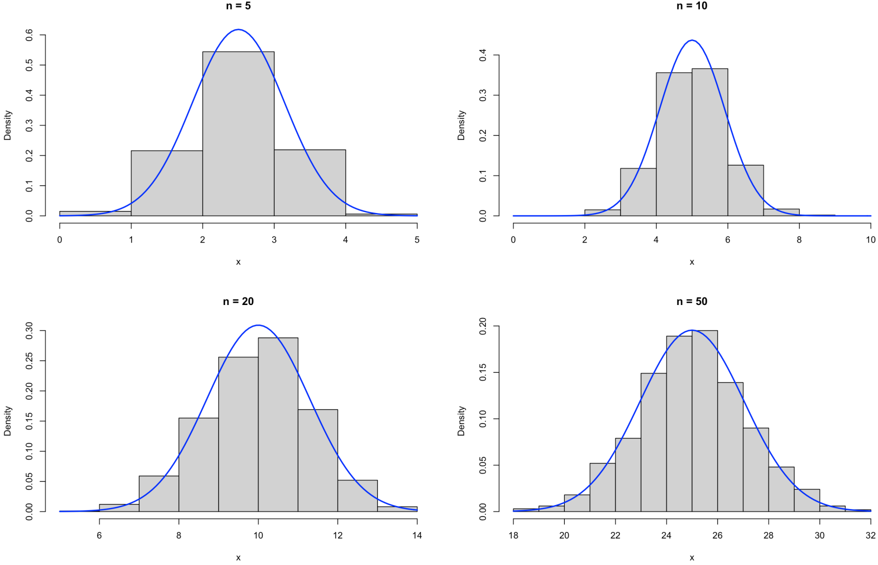
Central limit theorem

Example 2: X_1, ..., $X_n \overset{i.i.d.}{\sim} \text{Unif}(0,1)$. And $\mathbb{E}X_i = 1/2, \text{Var}(X_i) = 1/12$.
By CLT,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{d}{\approx} N\left(\frac{1}{2}, \frac{1}{12n}\right),$$
$$\sum_{i=1}^{n} X_i \overset{d}{\approx} N\left(\frac{n}{2}, \frac{n}{12}\right).$$

Histogram of sum of uniform r.v's

Histogram of $\sum_{i=1}^{n} X_i$, where $X_i \overset{i.i.d.}{\sim}$ Unif$(0, 1)$



Normal approximation to the binomial

By CLT, $X \sim \text{Bin}(n, p)$ can be approximated by $N(np, np(1-p))$.

When $n \geq 30$, $n \min(p, 1-p) \geq 10$, we can use the following normal approximations to calculate binomial probability:

- Without continuity correction:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}\left(\frac{a - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{b - np}{\sqrt{np(1-p)}}\right) \approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right)$$

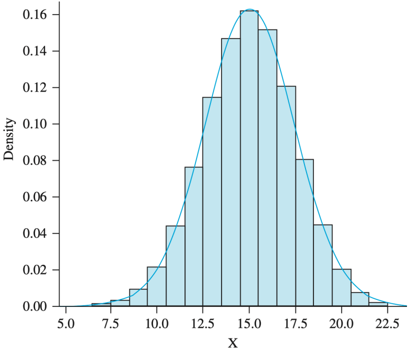
- With continuity correction:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a - 0.5 \leq X \leq b + 0.5) = \mathbb{P}\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Normal approximation to the binomial

When $n \geq 30$, $n \min(p, 1-p) \geq 10$, we can approximate the binomial probability with continuity correction:

- $\mathbb{P}(X = x) = \mathbb{P}(x - 0.5 \leq X \leq x + 0.5) = \Phi\left(\frac{x+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x-0.5-np}{\sqrt{np(1-p)}}\right)$
- $\mathbb{P}(a \leq X \leq b) = \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right)$
- $\mathbb{P}(a < X \leq b) = \mathbb{P}(a + 1 \leq X \leq b) =$ similar to above
- $\mathbb{P}(a \leq X < b) = \mathbb{P}(a \leq X \leq b - 1) =$ similar to above



Normal approximation to the binomial

Suppose that 25% of all students at a large public university receive financial aid. Let X be the number of students in a random sample of size 50 who receive financial aid, so that $p = 0.25$. Calculate:

- the probability that at most 10 students receive aid
- the probability that between 5 and 15 (inclusive) of the selected students receive aid

---

Check: $n = 50 \geq 30, n\min(p, 1-p) = 12.5 > 10$. Hence

$X \sim \text{Bin}(50, 0.25) \overset{d}{\approx} N(np, np(1-p)) = N(12.5, 9.375)$

- $\mathbb{P}(X \leq 10) = \mathbb{P}\left(\frac{X-12.5}{\sqrt{9.375}} \leq \frac{10+0.5-12.5}{\sqrt{9.375}}\right) \approx \Phi(-0.65) = 0.2578$
- $\mathbb{P}(5 \leq X \leq 15) = \mathbb{P}\left(\frac{5-0.5-12.5}{\sqrt{9.375}} \leq \frac{X-12.5}{\sqrt{9.375}} \leq \frac{15+0.5-12.5}{\sqrt{9.375}}\right) \approx$
  $\Phi(2.61) - \Phi(-0.98) = 0.9955 - 0.1635 = 0.8320$

Reading list (optional)

- "Probability and Statistics for Engineering and the Sciences" (9th edition):
  - Chapter 5.3 (skip examples of joint distribution), 5.4 and 5.5
- "OpenIntro statistics" (4th edition, free online, download [here]):
  - Chapter 5.1.3-5.1.6

Many thanks to

- Yang Feng
- Joyce Robbins
- Chengliang Tang
- Owen Ward
- Wenda Zhou
- And all my teachers in the past 25 years