STAT230

PROBABILITY

Chapter 10

UNIVERSITY OF WATERLOO

Normal Approximations and Moment generating Functions Chapter 10 :Chapter objectives

- The Central Limit Theorem.

- Normal approximation to the Poisson distribution.

  If $\mu \geq 5$ then it can be shown that X~ Poisson (μ)has approximately a N(μ, μ) distribution.

- Normal approximation to the Binomial distribution.

  Use Normal Approximation to Binomial if $np \geq 5$.

- Continuity Correction.

UNIVERSITY OF WATERLOO

Please Do Chapter 10 Problems: 10.1-10.6

Use of the Normal Distribution in Approximations

- The major reason that the Normal distribution is so commonly used is that it tends to approximate probabilities for linear combinations of variables having non-Normal distribution,

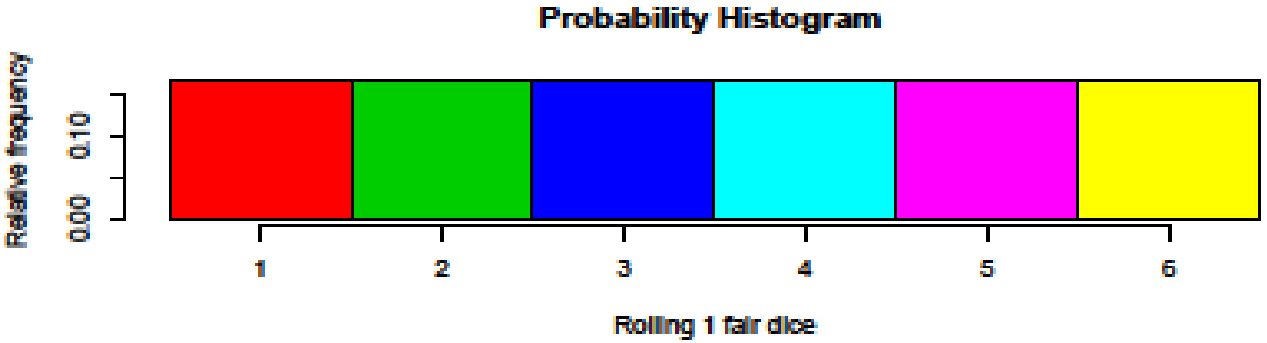
(under certain condition).

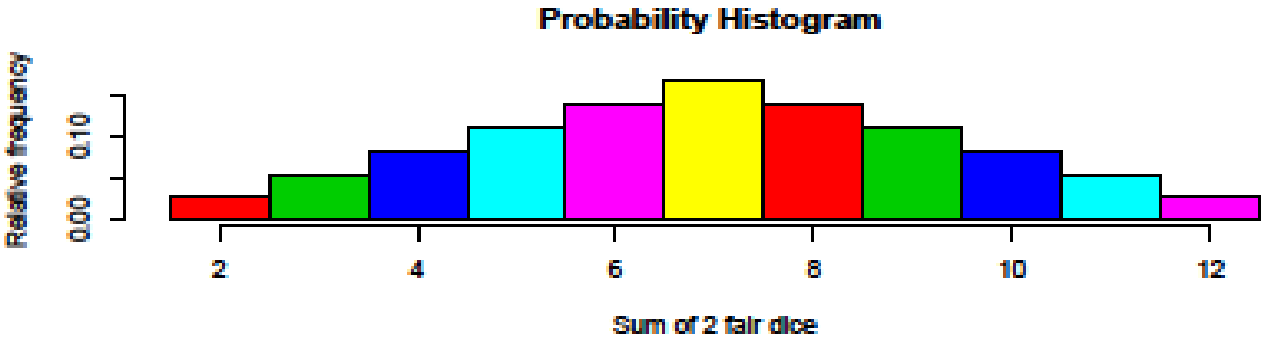
- This property follows from the Central Limit Theorem.

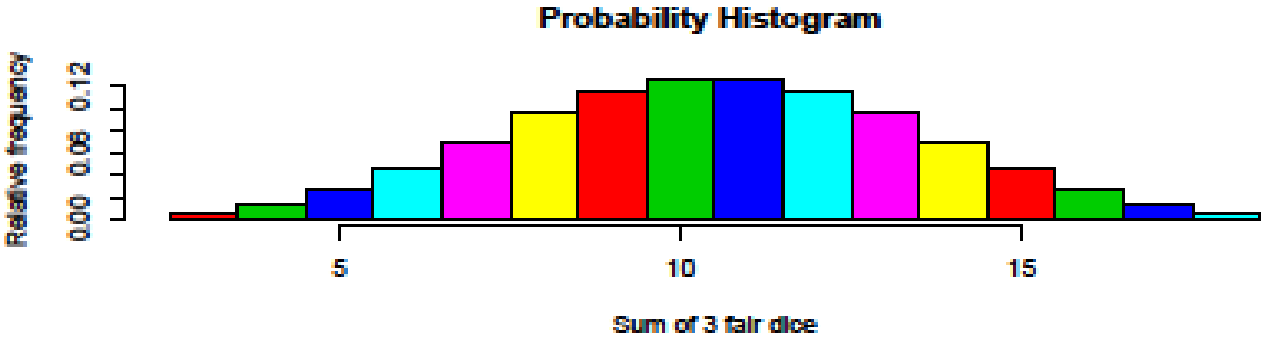
- There are actually several versions of the Central Limit Theorem.

Example

- If we throw n fair dice and S_n is the sum of the outcomes, what is the distribution of S_n?

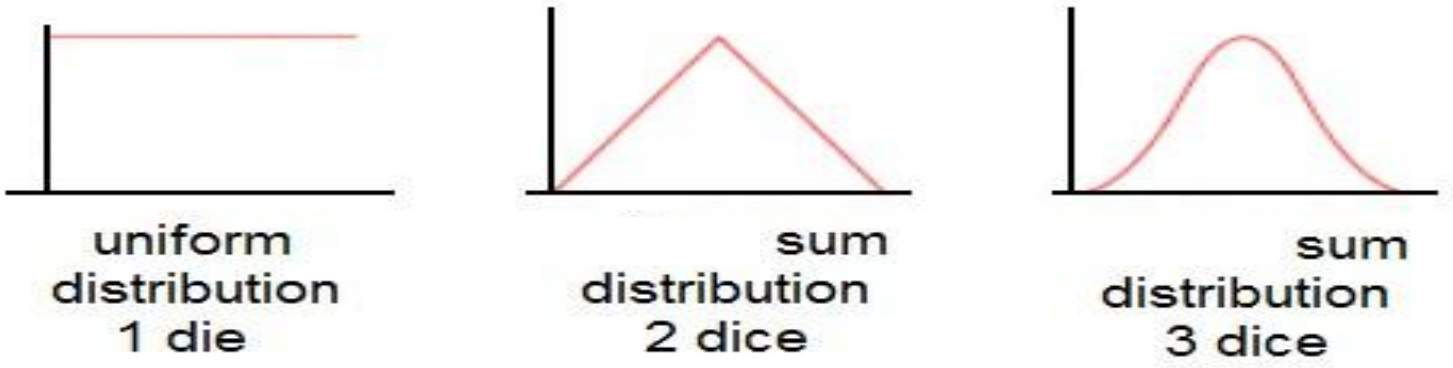






UNIVERSITY OF WATERLOO

- Rolling one die, the probability distribution is simply the discrete uniform distribution.



- The distribution of sums 2 dice is symmetric triangular (a piecewise linear function) with a peak at $6 + 1$. The probabilities increase linearly from 2 to $6 + 1$, and decrease linearly from $6 + 1$ to $2(6)$.

- The distribution for the sums on 3 dice is shaped like a symmetric bell that is a piecewise function of quadratic polynomials.

UNIVERSITY OF WATERLOO

Notes

We will use the Central Limit Theorem when n is large, but finite to approximate the distribution of S_n or $\overline{X}$ by a Normal distribution.

- $S_n = \sum_{i=1}^{n} X_i$ has approximately a $N(n\mu, n\sigma^2)$
- $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ has approximately a $N(\mu, \sigma^2/n)$

Central Limit Theorem

First formulation

- Let $X_1, X_2 ..., X_n$ are i.i.d. with $E(X_i) = \mu$ and finite $Var(X_i) = \sigma^2$
- Let $S_n = X_1 + X_2 + ..., + X_n = \sum_{i=1}^{n} X_i$. Then, the c.d.f of the r.v.

$$\frac{\sum^{n} X_i - n\mu}{\sigma \sqrt{n}} = \frac{S_n - n\mu}{\sigma \sqrt{n}} \qquad \text{as } n \to \infty$$

approaches the N(0, 1) c.d.f.

$\sum X_i \sim N(n\mu, n\sigma^2)$

UNIVERSITY OF WATERLOO

Second formulation

Suppose $X_1, X_2 ..., X_n$ are i.i.d. with $E(X_i) = \mu$ and finite $Var(X_i) = \sigma^2$. Then, as $n \to \infty$ the c.d.f of

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

approaches the N(0,1) c.d.f.

$\bar{X} \sim N(\mu, \sigma^2/n)$

UNIVERSITY OF WATERLOO

Note

As $n \to \infty$, both distribution $N(n\mu, n\sigma^2)$ and $N(\mu, \sigma^2/n)$ fail to exist.

Because both $n\mu$ and $n\sigma^2 \to \infty$.

$\sigma^2/n \to 0$

Notes

(1) This theorem works for essentially all distributions which X_i could have. The only exception occurs when X_i has a distribution whose mean or variance don't exist. There are such distributions, but they are rare.

Example (Cauchy Distribution)

The simplest Cauchy distribution is called the **Standard Cauchy distribution**. It is the distribution of a random variable that is the ratio of two independent Standard Normal variables and has the probability density function

$$f(x;0,1) = \frac{1}{\pi(1+x^2)}. \qquad x \in \mathbb{R}$$

Both its mean and its variance are undefined.

(2) We will use the Central Limit Theorem to approximate the distribution of sum $\sum_{i=1}^{n} X_i$ or averages $\overline{X}$.

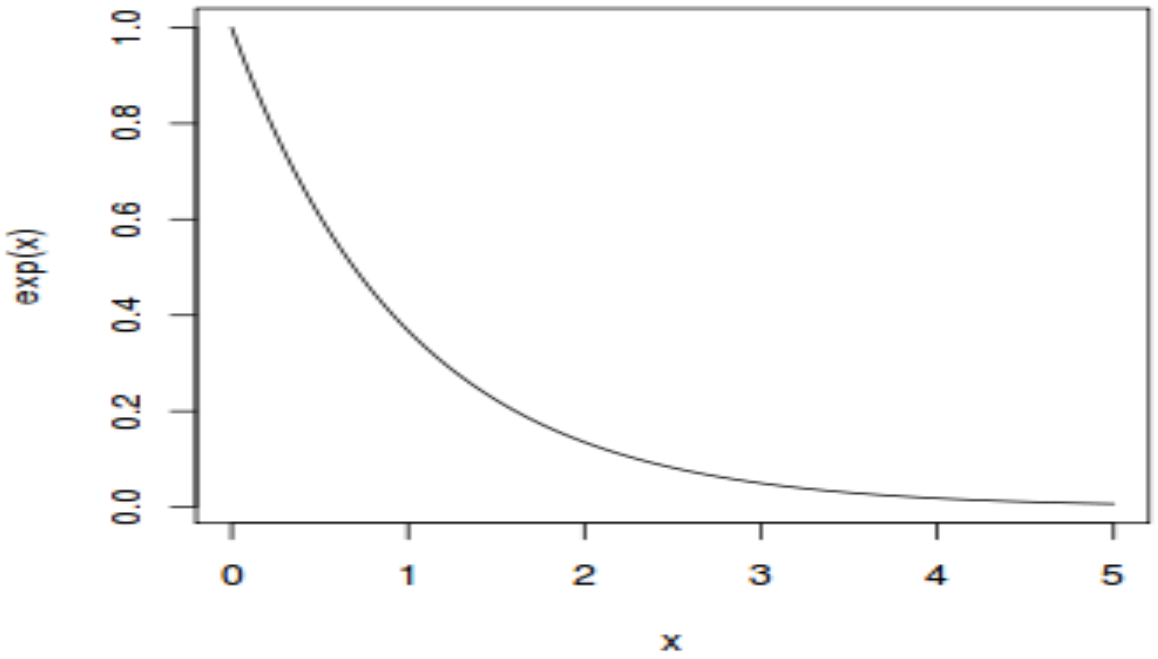
The accuracy of the approximation depends on

- n (bigger is better).
- The approximation works better for small n when the shape of the p.f / p.d.f. of X is close to symmetric.
- If the shape of the probability (density) function of X_i is very non-Normal shaped then the approximation will be poor for small n.

Example

Exponential distribution.

The Exponential distribution with $\mu = 1$.



(3)

➡ If the X_i's themselves have a Normal distribution, then $\sum_{i=1}^{n} X_i$ and $\bar{X}$ have exactly Normal distributions for all values of n.

➡ If the X_i's do not have a Normal distribution themselves, then $\sum_{i=1}^{n} X_i$ and $\bar{X}$ have approximately Normal distributions when n is large.

UNIVERSITY OF WATERLOO

Example

Hamburger patties are packed 8 to a box, and each box is supposed to have 1 kg of meat in it. The weights of the patties vary a little because they are mass produced, and the weight X of a single patty is actually a random variable with mean $\mu = 0.128$ kg and standard deviation $\sigma = 0.005$ kg.

Find the probability a box has at least 1 kg of meat, assuming that the weights of the 8 patties in any given box are independent.

UNIVERSITY OF WATERLOO

- Let $X_1, \ldots, X_8$ be the weights of the 8 patties in a box, and $S_8 = X_1 + \cdots + X_8$ be their total weight.

- By the Central Limit Theorem, S_8 is approximately $N(8\mu, 8\sigma^2)$; we'll assume this approximation is reasonable even though $n = 8$ is small since $X \sim N(\mu, \sigma^2)$

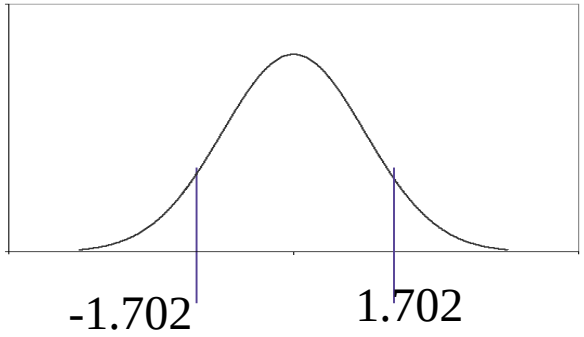
- Thus $S_8 \sim N(1.024, 0.0002)$ and

$$P(S_8 > 1) = P\left(Z > \frac{1-1.024}{\sqrt{0.0002}}\right)$$

$$= P(Z > -1.702)$$

$$= P(Z \leq 1.702)$$

$$= 0.9554$$



- We see that only about 96% of the boxes actually have 1 kilogram or more of hamburger.

UNIVERSITY OF WATERLOO

Example

Suppose we flip a fair coin 900 times. What is the probability to get at least 465 heads? $P(S_n \geq 465) = ?$

$\mu = 0.5, \quad \sigma^2 = p(1-p) = 0.25$

$n\mu = 0.5(900) = 450, \quad \sigma\sqrt{n} = 15$

$$P\left(\frac{S_{900} - n\mu}{\sigma\sqrt{n}} \geq \frac{465 - 450}{15}\right) = P(Z \geq 1)$$

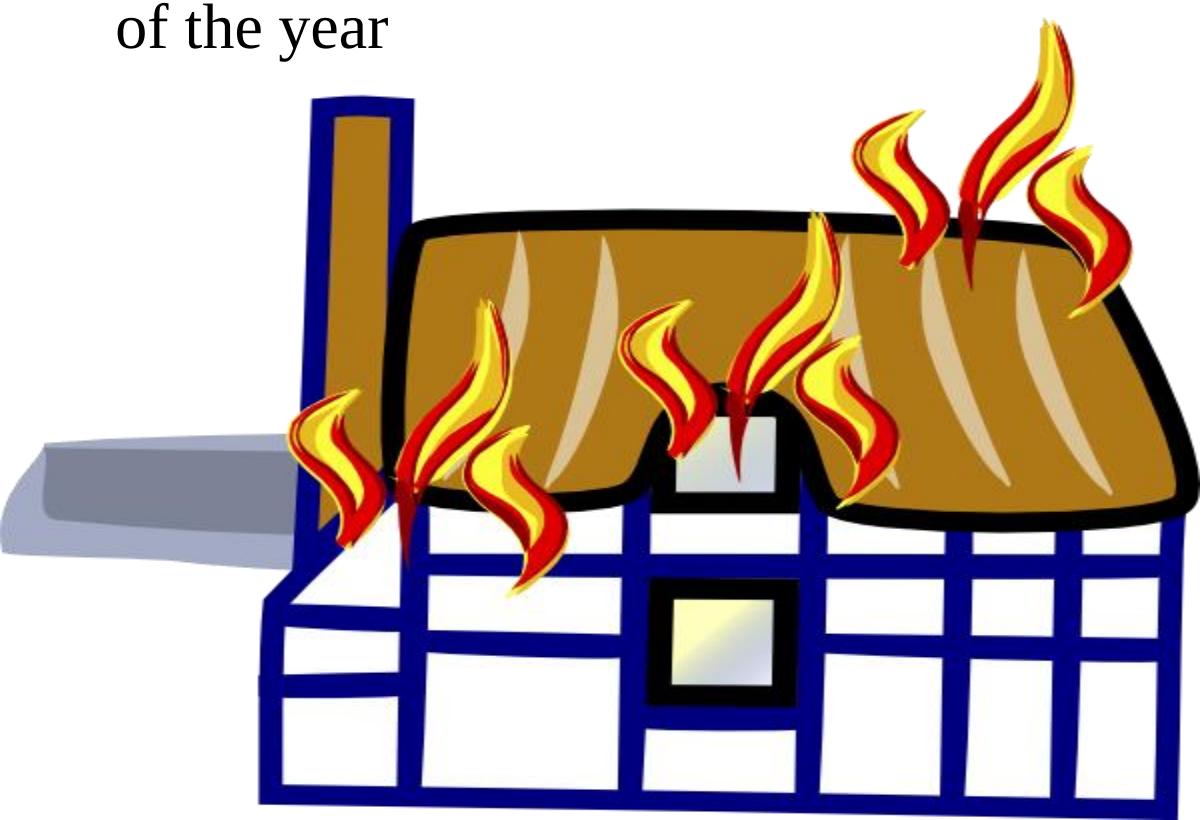
$$= 1 - P(Z \leq 1)$$

$$= 1 - 0.84134$$

$$= 0.15866$$

Example

Suppose fires reported to a fire station satisfy the conditions for a Poisson process, with a mean of 1 fire every 4 hours. Find the probability the 500^{th} fire of the year is reported on the 84^{th} day of the year



- Let X_i be the time between the $(i-1)^{st}$ and i^{th} fires (X_1 is the time to the 1st fire).

- X_i has an Exponential distribution with $\theta = 1/\lambda = 4$ hours, or $\theta = 1/6$ day.

- Let $S_{500} = \sum_{i=1}^{500} X_i$ be the time until the 500th fire.

- $S_{500} \sim N(500\,\mu, 500\,\sigma^2)$

- $\mu = \theta = 1/6 \Rightarrow n\mu = 500/6 = 83.3, \quad \sigma^2 = \theta^2 = 1/36$

- $\Rightarrow n\sigma^2 = 500/36 = 13.89 \Rightarrow \sqrt{n\,\sigma^2} = 3.7$

UNIVERSITY OF WATERLOO

$$P(\ 83 < S_{500} \leq 84) \approx P(\ \ (83 - 83.3)/\ 3.7 < \ Z\ \ \leq\ (84 - 83.3)/\ 3.7)$$

$$= P(\ -0.09 < \ Z\ \leq 0.18)$$

$$= P(\ Z \leq 0.18) - P(\ Z \leq -0.09)$$

$$= P(\ Z \leq 0.18) - [\ 1 - P(\ Z \leq 0.09)\ ]$$

$$= 0.57142 + 0.53586 - 1$$

$$= 0.10728$$

Normal approximation to the Poisson Distribution

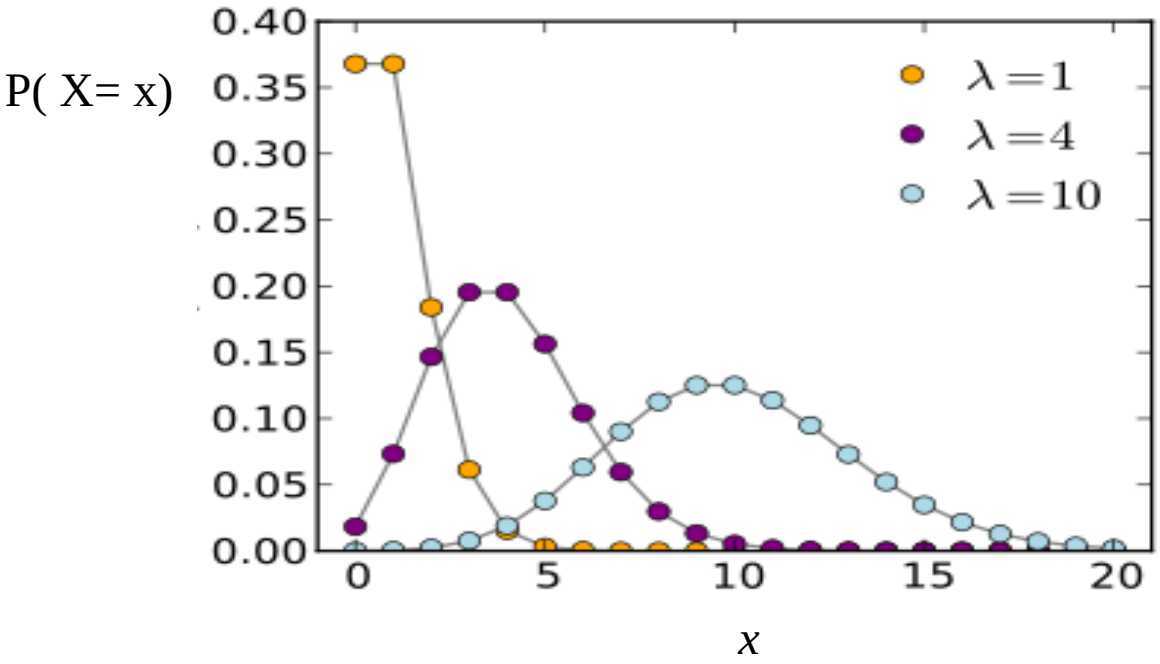
- Let X be a random variable with a Poisson(μ) distribution and suppose μ is large (E(X) = μ)

- For the moment suppose that μ is an integer.

- Suppose X ~ Poisson(μ), then the c.d.f of the standardized r.v.

$$Z = \frac{X - \mu}{\sqrt{\mu}}$$

Approaches that of a standard Normal random variable as $\mu \to \infty$



- The connecting lines are only guides for the eye.

Note

- **Suppose X ~ Poisson(μ).**
- **Then E(X) = μ and Var(X) = μ.**

If μ ≥ 5 then it can be shown that X has approximately a N(μ, μ) distribution.

Example

Assume that the number of asbestos particles in a squared meter of dust on a surface follows a Poisson distribution with mean of 1000. If squared meter of dust is analyzed, what is the probability that at most 950 particles are found?

$$P(X \leq 950) = \sum_{x=0}^{950} \frac{e^{-1000}1000^{x}}{x!}$$

The computational difficulty is clear. Using the Normal approximation since $\boldsymbol{\mu = 1000 \geq 5}$

$$P(X \leq x) = P(X \leq 950) = P\left(Z \leq \frac{950 - 1000}{\sqrt{1000}}\right)$$

$$= P(Z \leq -1.58) = 0.0571$$

UNIVERSITY OF WATERLOO

Continuity Correction

Example

- In an apple orchard, suppose the number X of worms in an apple has probability function:

x	0	1	2	3
f (x)	0.4	0.3	0.2	0.1

Find the probability a basket with 250 apples in it, has between 225 and 260 (inclusive) worms in it.

$$P\left(225 \leq \sum_{i=1}^{250} X_i \leq 260\right) \ ?$$

By the central limit theorem, $\sum_{i=1}^{250} X_i$ has approximately a $N(250\mu, 250\sigma^2)$ distribution, where X_i is the number of worms in the ith apple.

$$\mu = E(X) = \sum_{x=0}^{3} x\, f(x) = 1$$

$$E(X^2) = \sum_{x=0}^{3} x^2\, f(x) = 2$$

$$\sigma^2 = E(X^2) - \mu^2 = 1$$

$$\sum_{i=1}^{250} X_i \sim N(250, 250)$$

$$P\left(225 \le \sum_{i=1}^{250} X_i \le 260\right) = P\left(\frac{225 - 250}{\sqrt{250}} \le Z \le \frac{260 - 250}{\sqrt{250}}\right)$$

$$= P(-1.58 \le Z \le 0.63)$$

$$= 0.67850$$

Note

While this approximation is adequate, we can improve its accuracy, as follows:

$$P\left(224.5 \leq \sum_{i=1}^{250} X_i \leq 260.5\right) = P\left(\frac{224.5 - 250}{\sqrt{250}} \leq Z \leq \frac{260.5 - 250}{\sqrt{250}}\right)$$

$$= P(-1.61 \leq Z \leq 0.66)$$

$$= 0.69167$$

Note

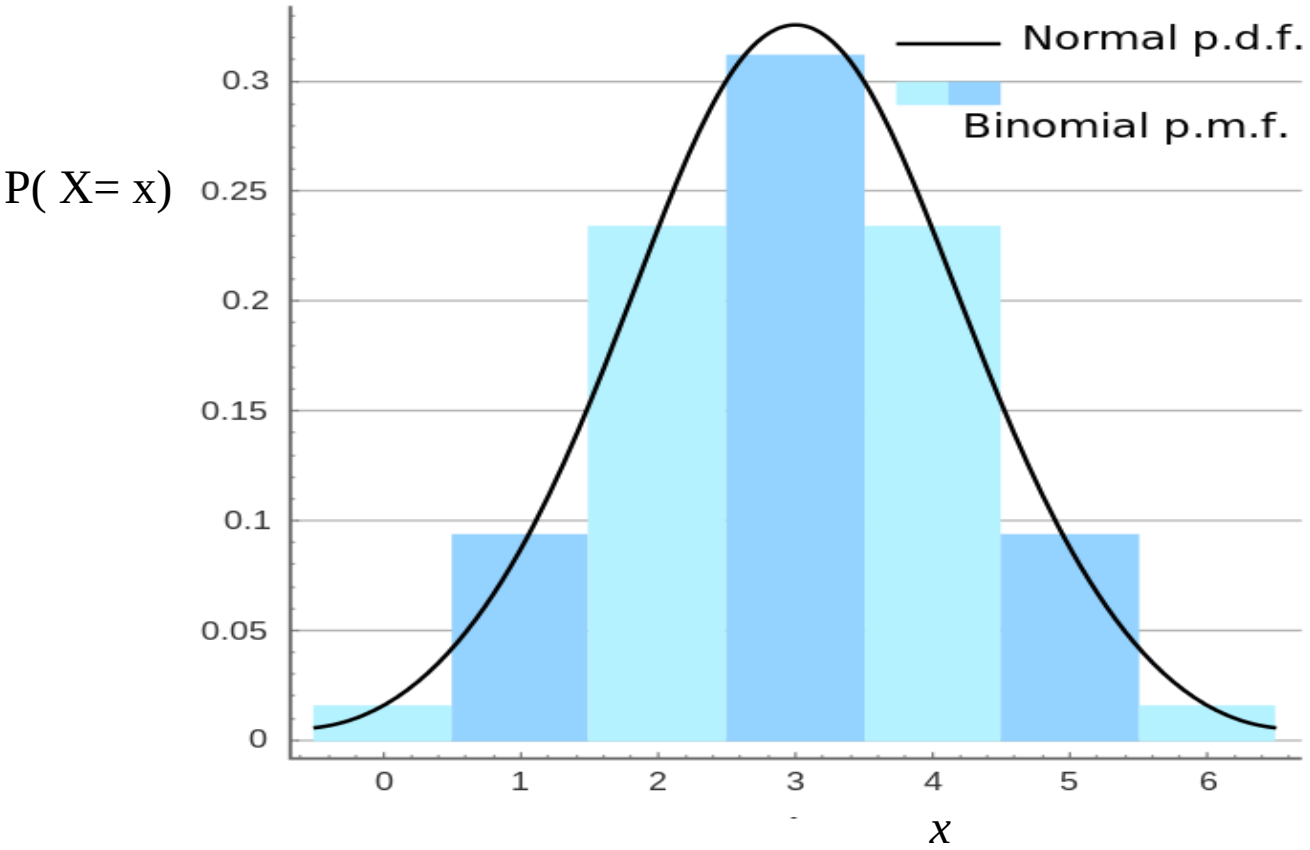
It is preferable to make this "**continuity correction**"

Note

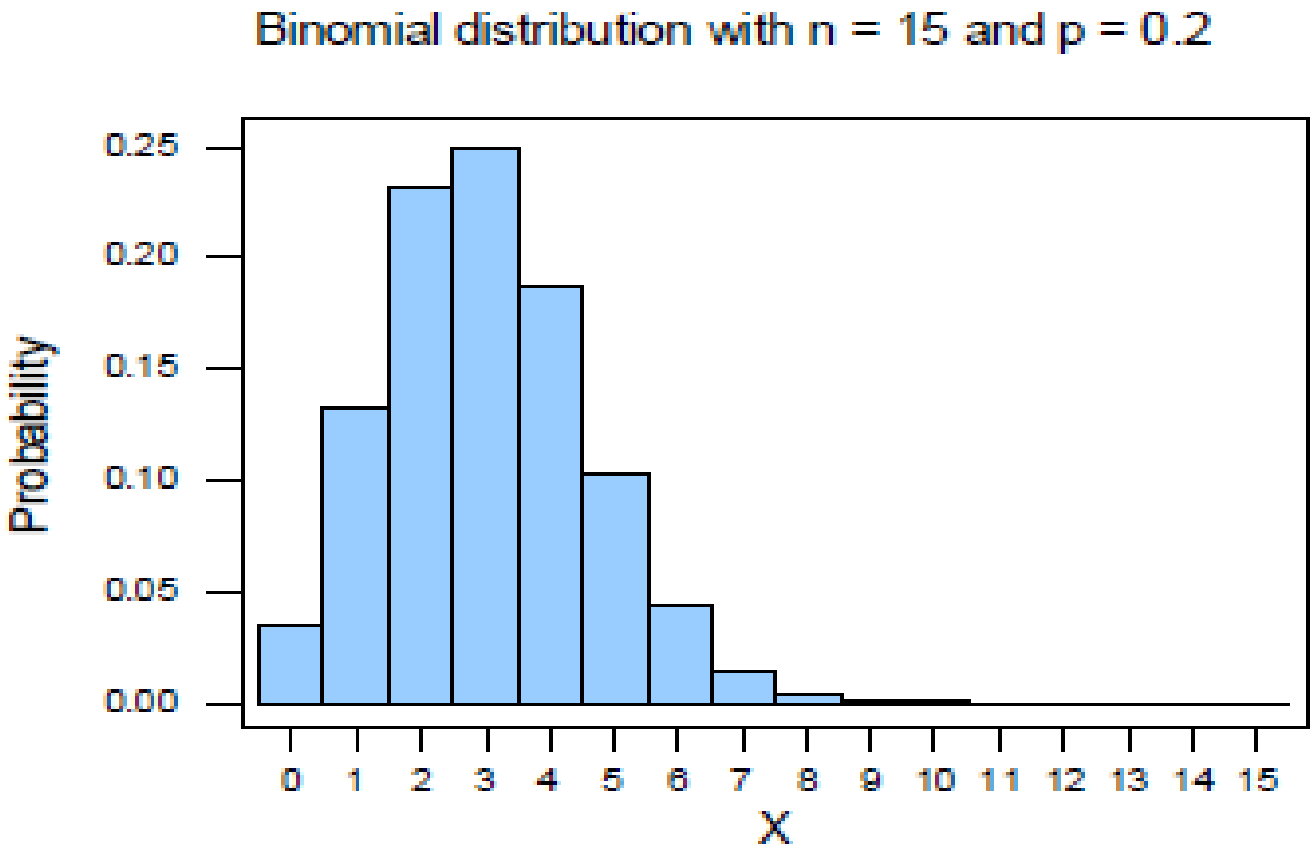
A continuity correction should not be applied when approximating a continuous distribution by the Normal distribution. Since it involves going halfway to the next possible value of x, there would be no adjustment to make if x takes real values.

Normal Approximation to the Binomial Distribution

- The Binomial distribution is symmetric (like the Normal distribution) whenever p = 0 .5

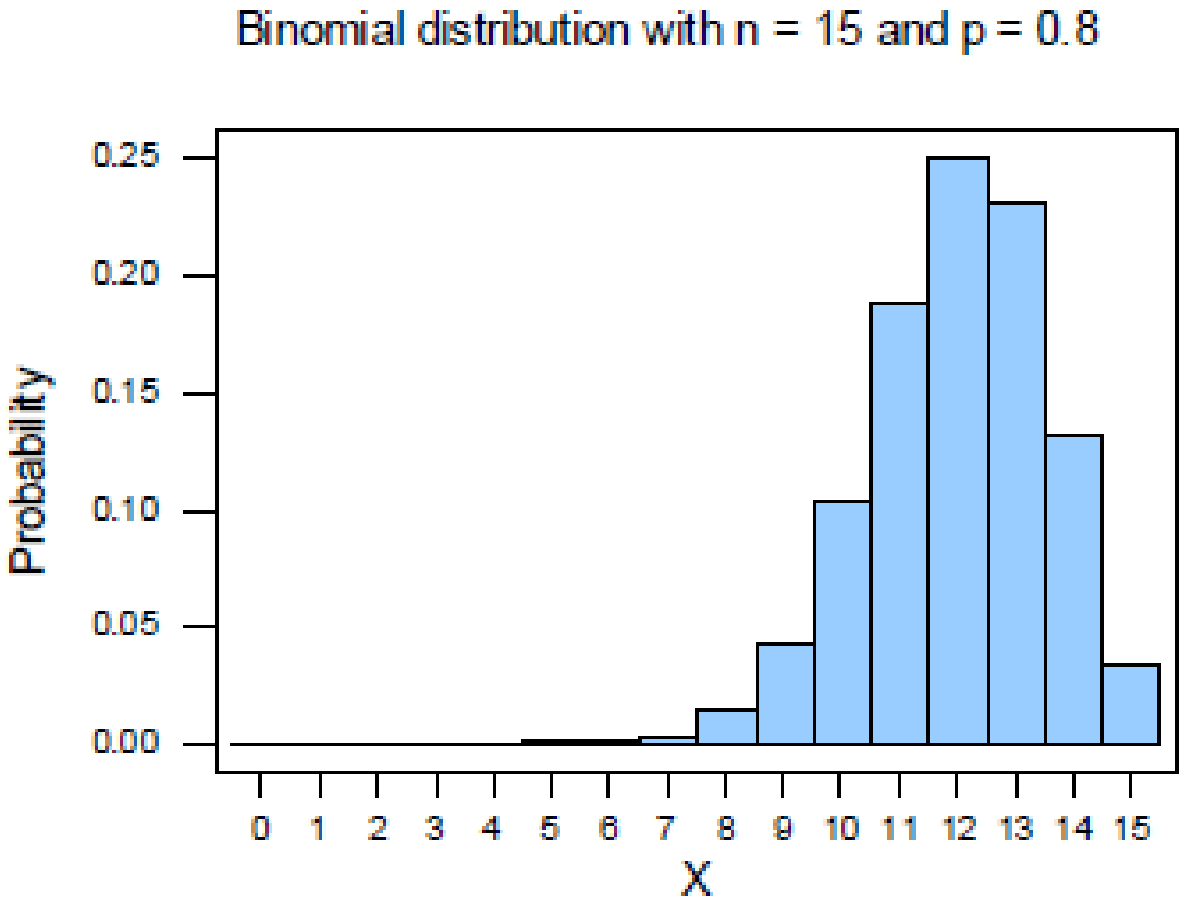


- For **small *p*** and **small *n***, the Binomial distribution is what we call **skewed right**



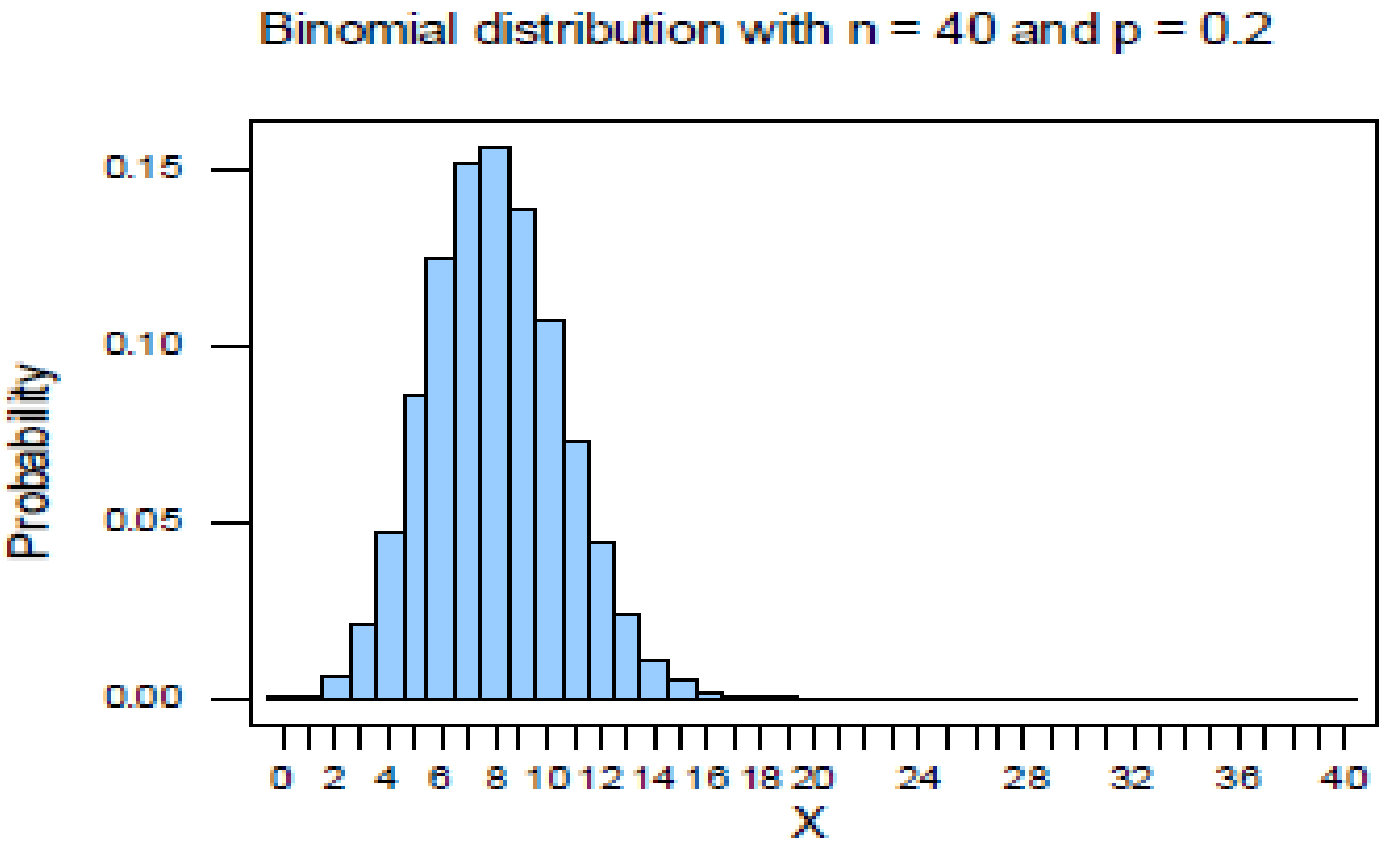
UNIVERSITY OF WATERLOO

- For **large *p*** and **small *n***, the Binomial distribution is what we call **skewed left**.



UNIVERSITY OF WATERLOO

- For **small *p*** and **large *n***, the Binomial distribution **approaches symmetry**.



- On the other hand, the larger the number of observations in the sample, the more tedious it is to compute the exact probabilities of success

- Fortunately, though, whenever the sample size is large, the Normal distribution can be used to approximate the exact probabilities of success.

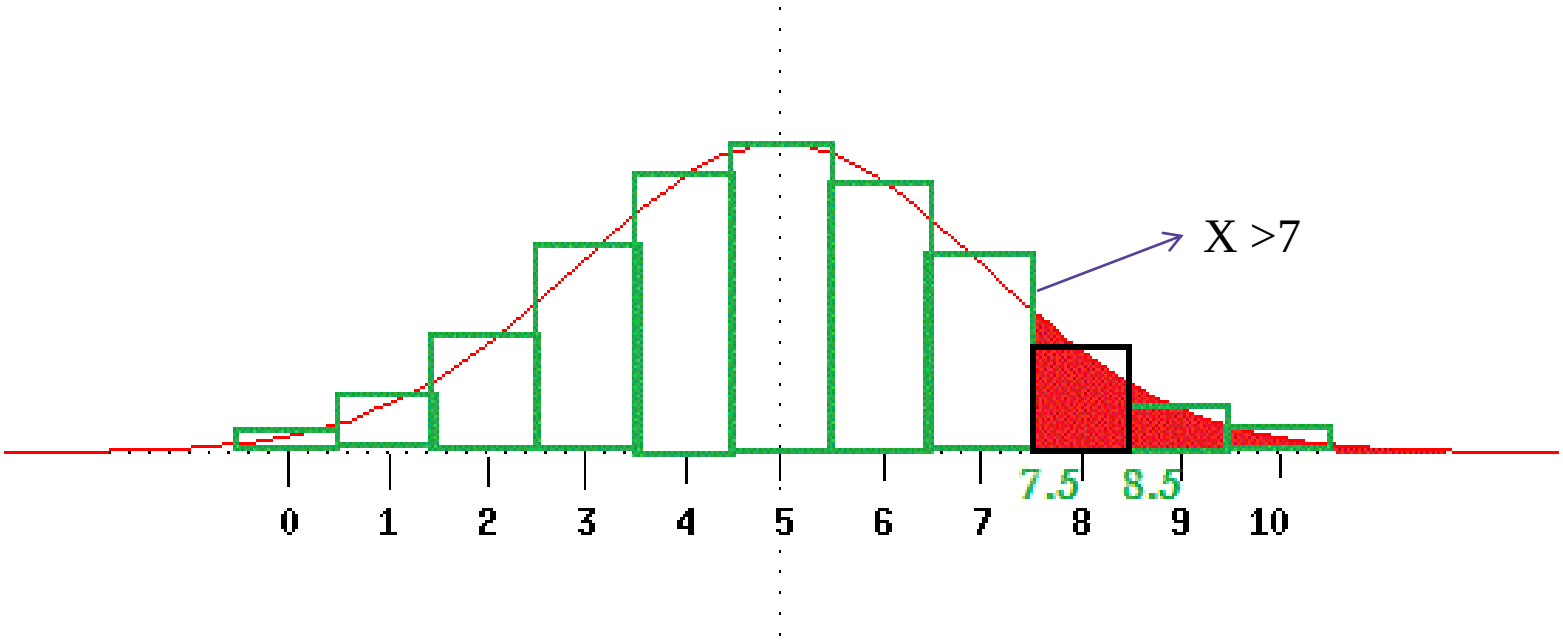
Note

Use Normal Approximation to Binomial if $np \geq 5$.

Note: The Continuity Correction

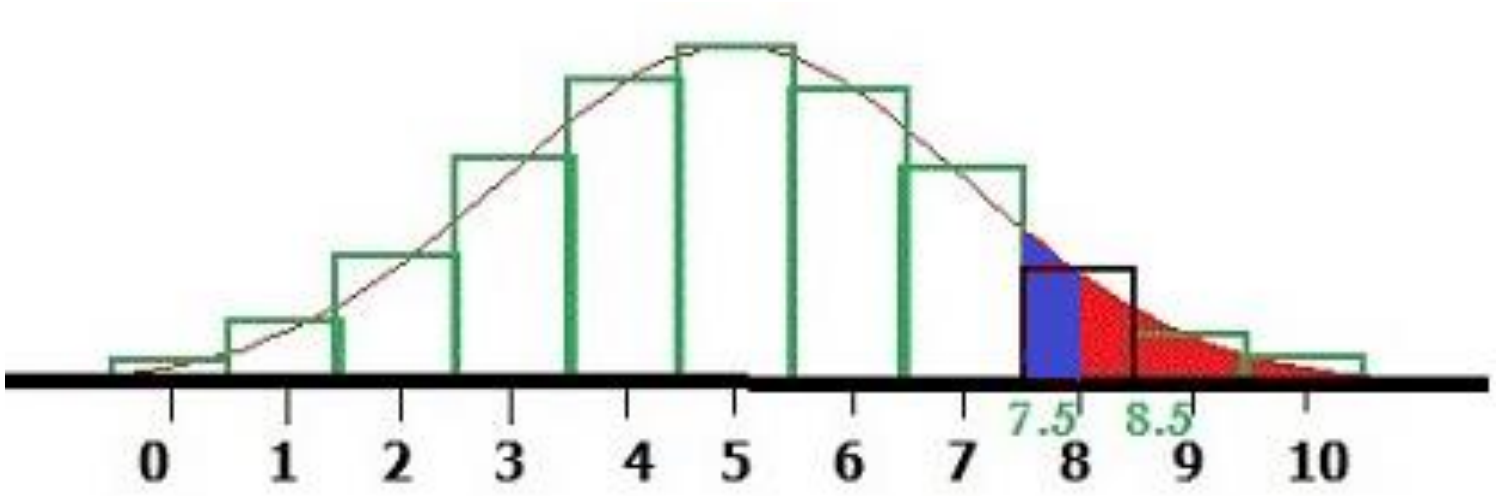
The addition of 0.5 is an example of the continuity correction which is intended to refine the approximation by accounting for the fact that the Binomial distribution is discrete while the Normal distribution is continuous.

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{npq}}\right)$$

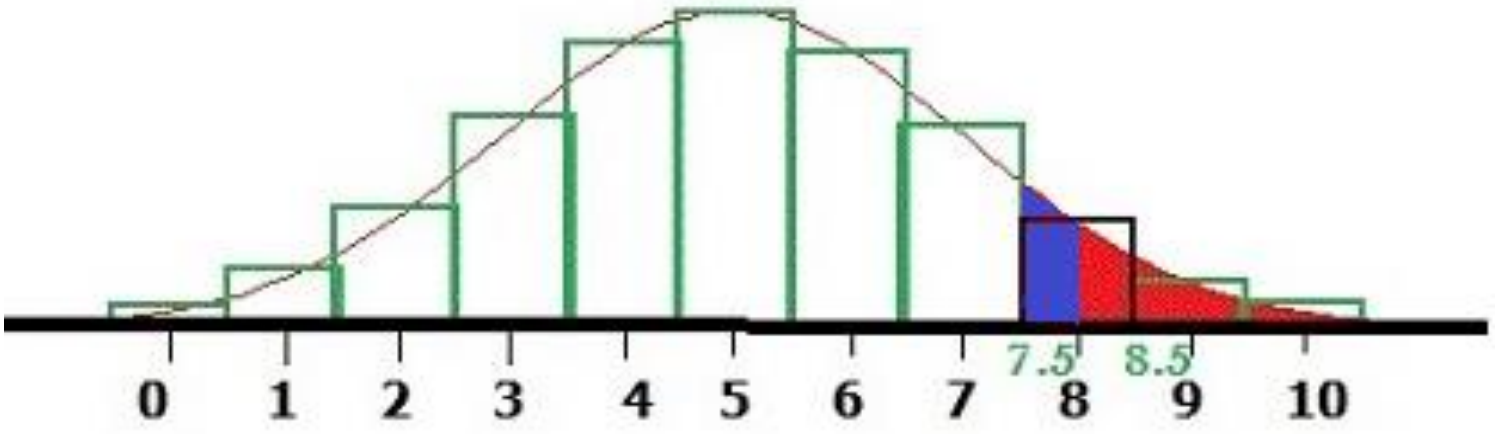


- The above image shows the Binomial plotted on a normal distribution graph.

- The red region shows that if the Binomial probability is greater than 7 P(X >7), then we want to include all of the red region — a continuous area that starts at 7.5



- If we want $P(X \geq 8)$, then without a continuity correction it would only get the red area. To include the blue area as well, we need $P(X > 7.5)$.



- If everything on the graph above was multiplied by 10 say (so the variable could only take values 0, 10, 20, ..., 90, 100), then to find $P(X \geq 80)$ we would want $P(X > 75)$, since we go half way to the next value of X.

Continuity Correction Factor Table When the Increment is 1

If $P(X = x)$ use $P(x - 0.5 < X < x + 0.5)$

If $P(X > x)$ use $P(X > x + 0.5)$

If $P(X \geq x)$ use $P(X > x - 0.5)$

If $P(X < x)$ use $P(X < x - 0.5)$

If $P(X \leq x)$ use $P(X < x + 0.5)$

Not always 0.5, halfway to the next possible value of x.

For example, if the discrete random variable X takes on values
a, a+d, a+2d, ...
then the continuity correction is d/2.

Example

At a particular small college the pass rate of Intermediate Algebra is 72%. If 500 students enroll in a semester determine the probability that at most 375 students pass.

$$\mu = np = 500(.72) = 360$$

$$\sigma = \sqrt{npq} = \sqrt{500(.72)(.28)} \approx 10$$

$$P(X \leq 375) \approx \Phi\left(\frac{375.5 - 360}{10}\right) = \Phi(1.55)$$

$$= 0.9394$$

Example

Let p be the proportion of Canadians who think Canada should adopt the US dollar.

a) Suppose 400 Canadians are randomly chosen and asked their opinion. Let X be the number who say yes. Find the probability that the proportion, X/ 400 , of people who say yes is within 0.02 of p, if p = 0 .20.

n ↘ ↗ p

$$X \sim \text{Binomial}(400, 0.2)$$

X approximately has Normal with mean $np = (400)(0.2) = 80$ and variance $np(1-p) = (400)(0.2)(0.8) = 64$

Using the Normal approximation since $np \geq 5$

$$P\left(\left|\frac{X}{400} - 0.2\right| \leq 0.02\right) = P(|X - 80| \leq 8) = P(|X - 80| \leq 8.5)$$

$$= P(|Z| \leq 8.5 / \sqrt{64})$$
$$= P(|Z| \leq 1.06)$$
$$= 2P(Z \leq 1.06) - 1$$
$$= 2(0.85543) - 1$$
$$= 0.711$$

b) Find the number, n, who must be surveyed so there is a 95% chance that X/n lies within 0.02 of p. Again suppose p is 0.20.

Since n is unknown, it is difficult to apply a continuity correction, so we omit it in this part.

By the Normal approximation,

$$X \sim N(\mu = np = 0.2n,\ \sigma^2 = np(1-p) = 0.16n).$$

$$P\left(\left| \frac{X}{n} - 0.2 \right| \leq 0.02 \right) \geq 0.95$$

$$P(|X - 0.2n| \leq 0.02n) \geq 0.95$$

$$P\left(\frac{|X - 0.2n|}{\sqrt{.16n}} \leq \frac{0.02n}{\sqrt{.16n}} \right) \geq 0.95$$

$$P (|Z| \leq 0.05\sqrt{n}) \geq 0.95$$

$$P (|Z| \leq 0.05\sqrt{n}) \geq 0.95$$

$$2P (Z \leq 0.05\sqrt{n}) - 1 \geq 0.95$$

$$P (Z \leq 1.96) \geq 0.975$$

Therefore,

$0.05\sqrt{n} \geq 1.96$

$n = 1536.64 = 1537$

UNIVERSITY OF WATERLOO