Approximations for discrete distributions, Section 3.7

Grethe Hystad

October 2, 2012

The normal Approximation to the Binomial distribution

- Let $X_1, X_2, ..., X_n$ are n independent, identically distributed **Bernoulli** random variables with mean p and variance $p(1-p)$, where $0 < p < 1$. Then

$$Y = \sum_{i=1}^{n} X_i$$

  is $b(n, p)$ with mean $E(Y) = np$ and variance $\text{Var}(Y) = np(1-p)$.
- By the Central limit theorem, the distribution of

$$Z_n = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

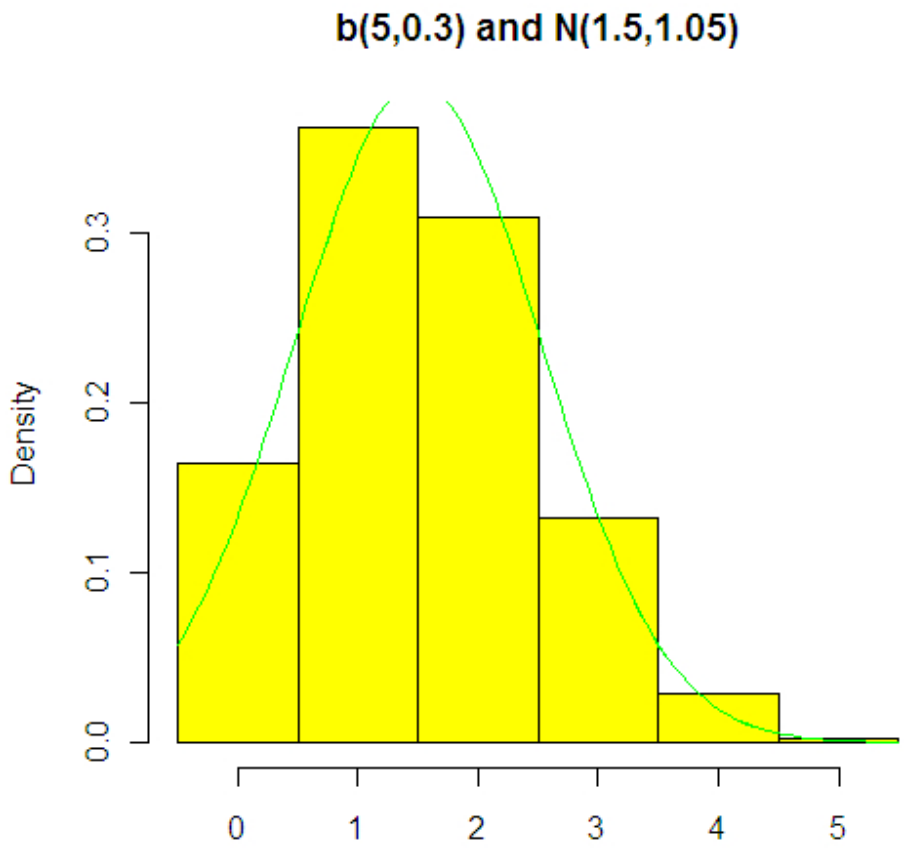
  is standard normal, $N(0, 1)$, in the limit as $n \to \infty$.

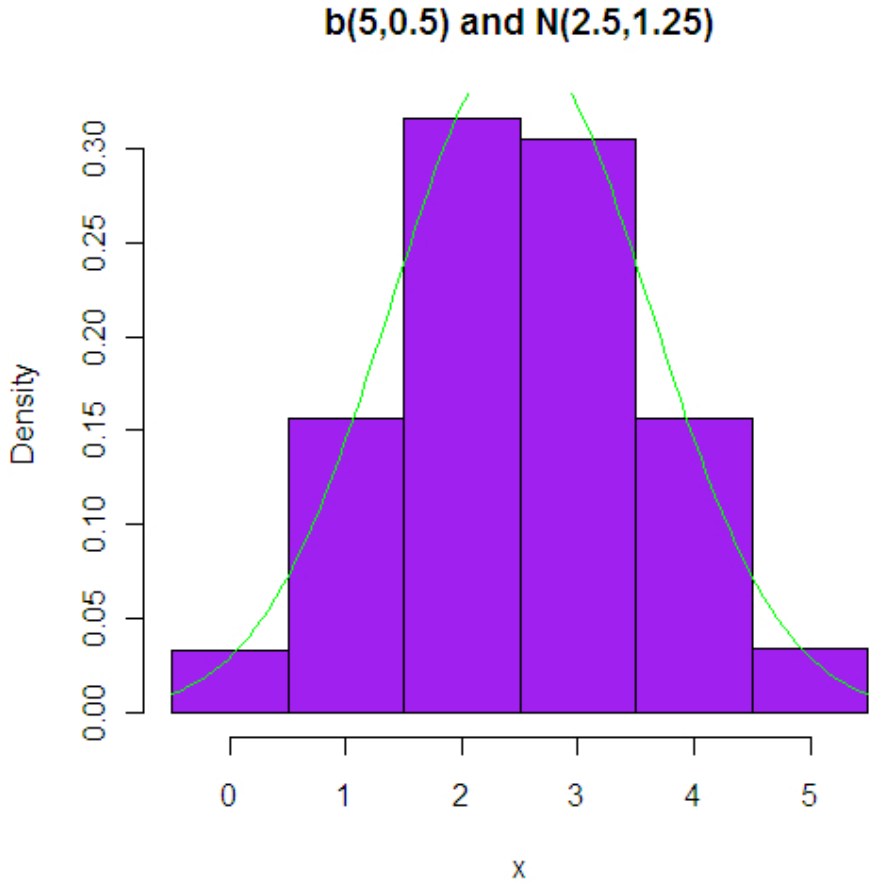
The normal Approximation to the Binomial distribution

- Thus, the distribution of Y is approximately $N(np, np(1-p))$ for sufficiently large n.
- Rule of thump: n is sufficiently large if $np \geq 5$ and $n(1-p) \geq 5$.
- As p deviates more and more from 0.5, the Bernoulli distribution would be more and more skewed, so a larger value of the sample size, n, is needed.

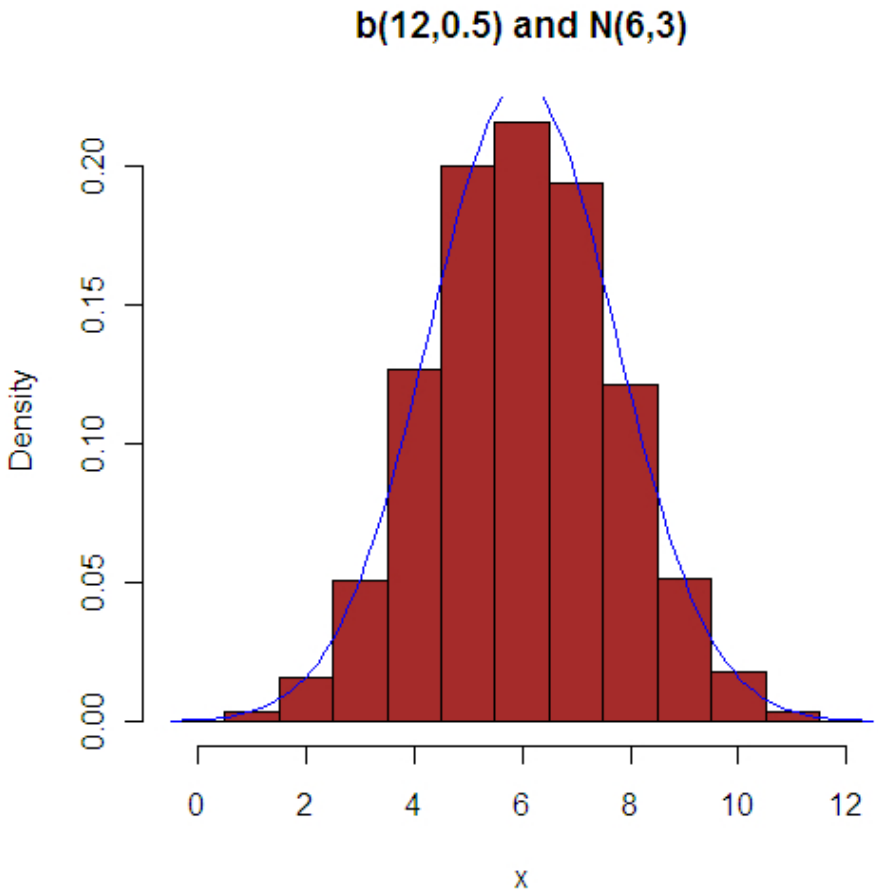
- Graphically, we can represent this probability as the area of a rectangle with a base of length one, centered at k (base goes from $k - \frac{1}{2}$ to $k + \frac{1}{2}$) and with height $f(k)$.
- The sum of these rectangles for $k = 0, 1, ..., n$ is then a Riemann sum and is the probability distribution for Y.
- We use the normal distribution to approximate the binomial distribution, that is the area under the curve of the p.d.f. for the normal distribution will approximate the areas of the rectangles in the probability histogram for the binomial distribution.
- Thus, to approximate $P(Y = k)$, we use the value of the area under the $N(np, np(1-p))$ p.d.f. between $k - \frac{1}{2}$ and $k + \frac{1}{2}$.

Here the normal approximation is not a good approximation of the binomial distribution since the binomial distribution is skewed here. Since p here deviates from 0.5, we need a larger sample size in order for the normal approximation to be a good approximation of the binomial distribution.




b(5,0.5) and N(2.5,1.25)
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1
2
3
4
5
x

Normal approximation to the Binomial distribution



Example

Let Y be $b(60, 0.3)$. Use the Central limit theorem to find the following:
(A) $P(16 \leq Y < 19)$
(B) $P(17 < Y \leq 20)$
(C) $P(Y = 18)$. Compare the value of this obtained from the normal approximation to the value obtained from the Binomial formula.

Solution

We have $E(Y) = np = (60)(0.3) = 18$ *and* $\text{Var}(Y) = np(1-p) = (60)(0.3)(0.7) = 12.6$ *Thus both* $np \geq 5$ *and* $np(1-p) \geq 5$. *Define* $Z_{60} = \frac{Y-18}{\sqrt{12.6}}$. *We have that* Z_{60} *is approximately standard normal.*

(A) Using the central limit theorem, we obtain,

$$\begin{aligned} P(16 \leq Y < 19) &= P(15.5 \leq Y \leq 18.5) \\ &= P\left(\frac{15.5-18}{\sqrt{12.6}} \leq \frac{Y-18}{\sqrt{12.6}} \leq \frac{18.5-18}{\sqrt{12.6}}\right) \\ &= P(-0.704 < Z_{60} < 0.141) \\ &\approx P(-0.704 < Z < 0.141) \\ &= \Phi(0.141) - \Phi(-0.704) = 0.315. \end{aligned}$$

Solution continue

Solution

(B) Using the central limit theorem, we obtain,

$$\begin{aligned} P(17 < Y \leq 20) &= P(17.5 \leq Y \leq 20.5) \\ &= P\left(\frac{17.5-18}{\sqrt{12.6}} \leq \frac{Y-18}{\sqrt{12.6}} \leq \frac{20.5-18}{\sqrt{12.6}}\right) \\ &\approx P(-0.141 < Z < 0.704) \\ &= \Phi(0.704) - \Phi(-0.141) = 0.315. \end{aligned}$$

Solution continue

Solution

(C)

$$
\begin{aligned}
P(Y = 18) &= P(\leq 17.5 \leq Y \leq 18.5) \\
&= P\left(\frac{17.5 - 18}{\sqrt{12.6}} \leq \frac{Y - 18}{\sqrt{12.6}} \leq \frac{18.5 - 18}{\sqrt{12.6}}\right) \\
&\approx \Phi(0.1408590) - \Phi(-0.1408590) = 0.1120187
\end{aligned}
$$

Using the Binomial formula:

> dbinom(18,60,0.3)

[1] 0.1118036

We see that the normal approximation gives a value very close to the value computed by the Binomial formula.

In general, for large n, we have for $k = 0, 1, 2, ..., n$:

- $$P(Y \leq k) \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$
- $$P(Y < k) \approx \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$
- $$P(Y \geq k) = 1 - P(Y < k) \approx 1 - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$
- $$P(Y > k) = 1 - P(Y \leq k) \approx 1 - \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

Problem

About 8% *of males are color blind. In a sample of* 100 *randomly selected males, determine the approximate probability that at least* 12 *of them are color blind?*

Solution given in class.

Sample sum

- Suppose $X_1, X_2, ..., X_n$ are independent and identically distributed random variable with mean μ and variance σ^2.
- Define the sample sum $S_n = X_1 + X_2 + ... + X_n$. We have
- $E(S_n) = n\mu$.
- $\text{Var}(S_n) = n\sigma^2$.
- standard deviation of S_n: $\sigma\sqrt{n}$.
- S_n is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.
- Its standardized variable is $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.
- Then $P(S_n \leq y) = P(Z_n \leq \frac{y-n\mu}{\sigma\sqrt{n}}) \approx \Phi(\frac{y-n\mu}{\sigma\sqrt{n}}) = \Phi(z)$, where $z = \frac{y-n\mu}{\sigma\sqrt{n}}$ and $\Phi(z)$ is the standard normal distribution.

Sample mean

- Suppose $X_1, X_2, ..., X_n$ are independent and identically distributed random variable with mean μ and variance σ^2.
- Define the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We have
- $E(\overline{X}) = \mu$.
- $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$.
- standard deviation of $\overline{X}$: $\frac{\sigma}{\sqrt{n}}$.
- $\overline{X}$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- Its standardized variable is $Z_n = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$.
- Then $P(\overline{X} \leq y) = P(Z_n \leq \frac{y-\mu}{\frac{\sigma}{\sqrt{n}}}) \approx \Phi(\frac{y-\mu}{\frac{\sigma}{\sqrt{n}}}) = \Phi(z)$, where $z = \frac{y-\mu}{\frac{\sigma}{\sqrt{n}}}$ and $\Phi(z)$ is the standard normal distribution.

Sample proportions

- Suppose $X_1, X_2, ..., X_n$ are independent and identically distributed Bernoulli random variable with parameter p.
- Define the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We have
- $E(\overline{X}) = p$.
- $\text{Var}(\overline{X}) = \frac{p(1-p)}{n}$ and standard deviation of $\overline{X}$: $\sqrt{\frac{p(1-p)}{n}}$.
- $\overline{X}$ is approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.
- Its standardized variable is $Z_n = \frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}}$.
- Then $P(\overline{X} \leq y) = P\left(Z_n \leq \frac{y-p}{\sqrt{\frac{p(1-p)}{n}}}\right) \approx \Phi(\frac{y-p}{\sqrt{\frac{p(1-p)}{n}}}) = \Phi(z)$, where $z = \frac{y-p}{\sqrt{\frac{p(1-p)}{n}}}$ and $\Phi(z)$ is the standard normal distribution.