Section 5.1 - Sampling Distributions for Counts and Proportions

Statistics 104

Autumn 2004

Copyright ©2004 by Mark E. Irwin

Distributions

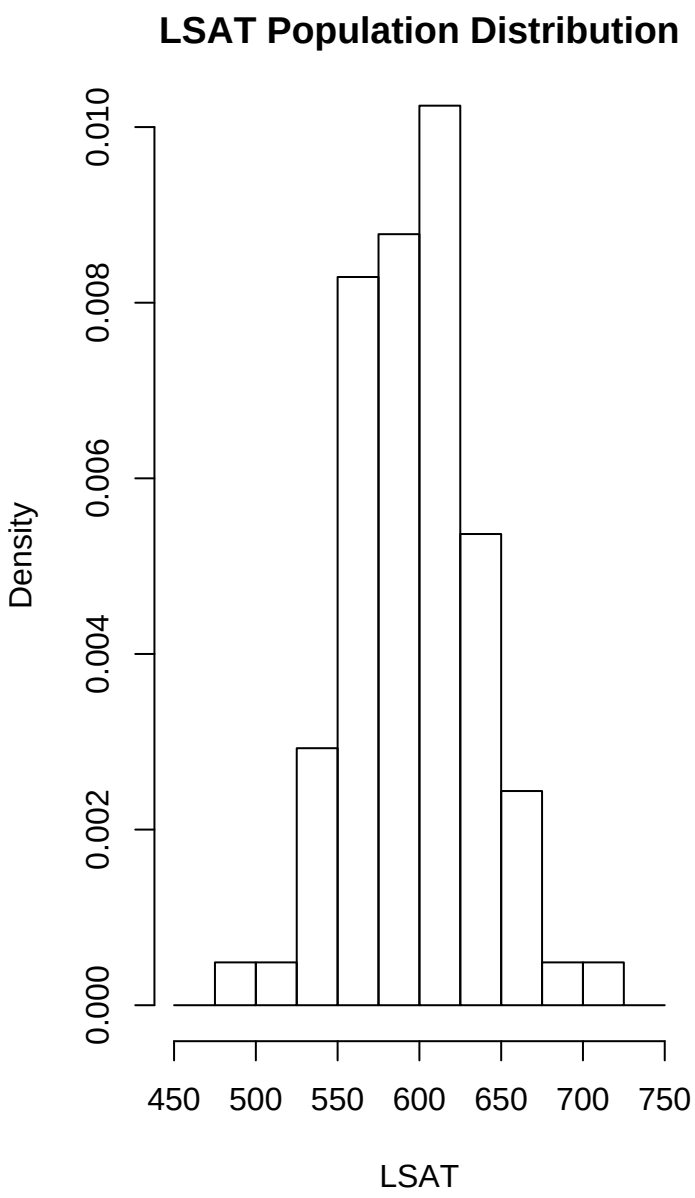
When dealing with inference procedures, there are two different distributions that you need to keep track of

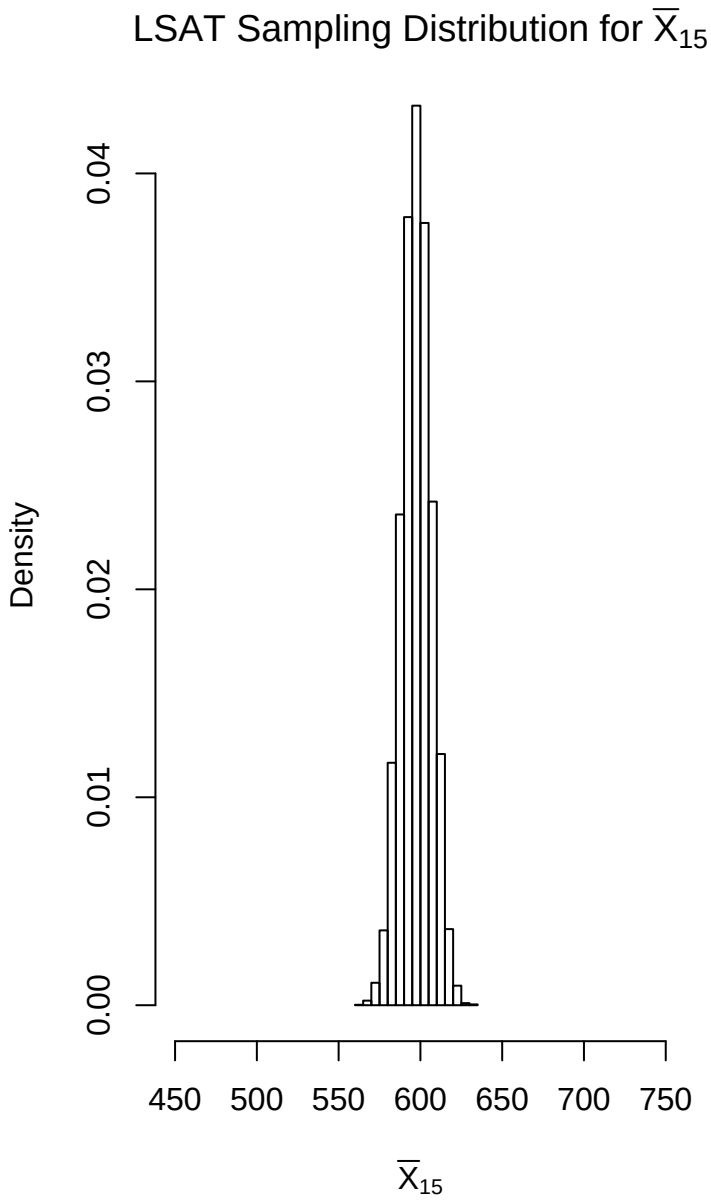
Population Distribution

The population distribution of a variable is the distributions of its values for all members of the population. The population distribution is also the probability distribution of the variable when we choose one individual from the population at random.

Sampling Distribution

A statistic from a random sample or randomized experiment is a random variable. The probability distribution of the statistic is its sampling distribution.


LSAT Population Distribution
Density
0.000
0.002
0.004
0.006
0.008
0.010
450
500
550
600
650
700
750
LSAT


LSAT Sampling Distribution for $\overline{X}_{15}$
Density
0.00
0.01
0.02
0.03
0.04
450
500
550
600
650
700
750
$\overline{X}_{15}$

Binomial Distribution

Example: Did you attend church of synagogue in the previous week?

Sampled 1785 and 550 said yes. This gives a sample proportion of

$$\hat{p} = \frac{550}{1785} = 0.42$$

What is the sampling distribution of $\hat{p}$?

This can be modelled with the Binomial Distribution.

Binomial Distribution

1. Fixed number of observations n

2. Each of the n observations are independent

3. Each observation falls into one of two categories, which for convenience get called "Success" and "Failure"

4. The probability of successes (call it p), is the same for each observation

Interested in the number of successes (call it X).

X is said to have a binomial distribution with parameters n and p.
$(X \sim Bin(n, p))$.

Binomial or not?

1. Flip a coin 20 times and count the number of heads.

   Yes. $Bin(n = 20, p = 0.5)$ if its a fair coin.

2. Draw 5 cards from a standard deck of cards and count the number of black cards.

   No. The draws are not independent which implies that the probabilities change as you go through the draws.

$$\begin{aligned} P[1^{st} \text{ card black}] &= \frac{1}{2} \\ P[2^{nd} \text{ card black} | 1^{st} \text{ card black}] &= \frac{25}{51} \\ P[2^{nd} \text{ card black} | 1^{st} \text{ card red}] &= \frac{26}{51} \end{aligned}$$

3. Number of faulty switches out of 6 from one company. $P[\text{Faulty}] = 0.2$

   Probably ok.

4. The number of successful field goals that Adam Vinatieri will kick in Sunday's Patriots game.

   No. n, the number of kicks is random and currently unknown.

5. Take a simple random sample of 1000 voters. Count the number who say that they voted to re-elect President Bush.

   Close, but not quite. Its similar to the deck example.

When the population is much larger that the sample size, the count of successes in a SRS of size n has approximately a $Bin(n, p)$ distribution if the population proportion of successes is p.

Rule of thumb for the approximation to be ok

$$\text{Population size} > 10n$$

Lets suppose that we have a population of 100,000 individuals and that 20% are "successes"

$$\begin{aligned} P[\text{Success on draw 1}] &= 0.2 \\ P[\text{Success on draw 2}|\text{Success on draw 1}] &= \frac{19999}{99999} = 0.199992 \\ P[\text{Success on draw 2}|\text{Failure on draw 1}] &= \frac{20000}{99999} = 0.200002 \end{aligned}$$

The success probabilities won't change much as the various units get sampled.

Now suppose that the population size is 5, still with a 20% "success" rate

$$\begin{aligned} P[\text{Success on draw 1}] &= 0.2 \\ P[\text{Success on draw 2}|\text{Success on draw 1}] &= \frac{0}{4} = 0 \\ P[\text{Success on draw 2}|\text{Failure on draw 1}] &= \frac{1}{4} = 0.25 \end{aligned}$$

Calculating binomial probabilities

The probability of exactly k successes when $X \sim Bin(n, p)$ is

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is the number of ways of choosing k items from n. Its often pronounced n choose k for this reason.

Motivation:

For each trial $P[\text{Success}] = p$; $P[\text{Failure}] = 1 - p$

Assume that k successes are followed by $n - k$ failures.

This has probability

$$\underbrace{p \times p \times \ldots \times p}_{k} \times \underbrace{(1-p) \times (1-p) \times \ldots \times (1-p)}_{n-k} = p^k (1-p)^{n-k}$$

Now each other possibility with k successes has exactly the same probability, which implies

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Why is $\binom{n}{k}$ the number of ways of choosing k items from n?

You have n ways of picking the first success, then $n-1$ ways of picking the second success after the first one, and so on down to $n-k+1$ ways of picking the kth success.

Multiplying these together gives

$$n \times (n-1) \times (n-2) \times \ldots \times (n-k+1) = \frac{n!}{(n-k)!}$$

Now the order of the successes doesn't matter. Given k items there is $k!$ different ways of ordering them. You have k choices for the list item in the list, which leaves $k-1$ choices for the 2nd item in the list, and so. Combining this with the above gives

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

One way of getting probabilities involving binomials is to work with the earlier probability formula.

For example, if $X \sim Bin(6, 0.2)$

$$\begin{aligned} P[X > 4] &= P[X = 5] + P[X = 6] \\ &= \binom{6}{5} 0.2^5 0.8^1 + \binom{6}{6} 0.2^6 0.8^0 \\ &= 0.0016 \end{aligned}$$

Another option is to work with binomial probability tables (Table C in Moore and McCabe)

This table gives binomial probabilities for certain choices of n and p.

For the $X \sim Bin(6, 0.2)$ example, we need to look at the block with $n = 6$ and $p = 0.2$.

TABLE C Binomial probabilities (*continued*)

Entry is $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

		p								
n	k	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5		.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0937
	6			.0001	.0002	.0007	.0018	.0041	.0083	.0156

The table doesn't have anything for $p > 0.5$. This is not a problem as we can just switch the definition of "success" and "failure" to fit the problem.

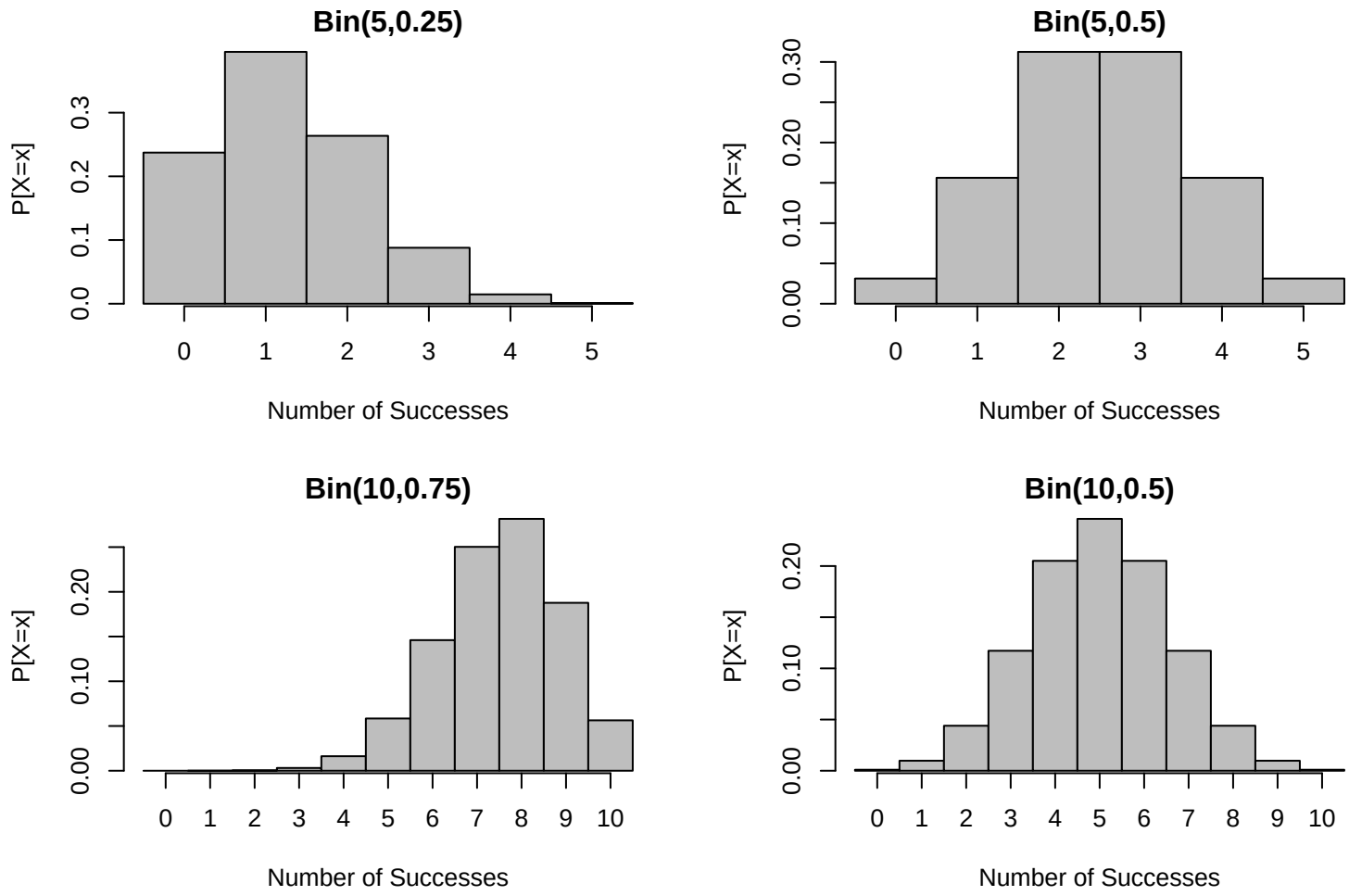
Let $X \sim Bin(n, p)$ and $Y \sim Bin(n, 1-p)$. Then

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k} = P[Y = n - k]$$

Most stat packages, Excel, scientific calculators can also be used to get binomial probabilities. There is one big advantage to using software: n and p are not restricted. For example, if $X \sim Bin(11, 0.78)$,

$$P[X = 7] = 0.1358$$

which isn't available from the table.



The binomial distribution is always unimodal, but can be symmetric or skewed. It is symmetric if $p = 0.5$, skewed left if $p < 0.5$ and skewed right if $p > 0.5$

Mean and Variance of a Binomial

$$\mu_x = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

by the definition of the mean for a discrete random variable. However this is somewhat ugly, though can be solved with a little algebra. The variance is even worse (though still solvable this way)

$$\sigma_x^2 = \sum_{x=0}^{n} (x-\mu_x)^2 \binom{n}{x} p^x (1-p)^{n-x}$$

There is an easier way to get a handle on this though.

Define Z_i to be the result of trial i where

$$Z_i = \begin{cases} 1 & \text{trial } i \text{ is a success} \\ 0 & \text{trial } i \text{ is a failure} \end{cases}$$

Therefore $X = Z_1 + Z_2 + \ldots + Z_n$, the sum of n independent random variables. So we need to figure out μ_z and σ_z^2.

These are easy, as

$$\begin{aligned}
\mu_z &= 0 \times (1-p) + 1 \times p = p \\
\sigma_z^2 &= (0-p)^2 \times (1-p) + (1-p)^2 \times p = p(1-p)
\end{aligned}$$

These give

$$\begin{aligned}
\mu_x &= \mu_{z_1} + \mu_{z_2} + \ldots + \mu_{z_n} \\
&= p + p + \ldots + p = np \\
\sigma_x^2 &= \sigma_{z_1}^2 + \sigma_{z_2}^2 + \ldots + \sigma_{z_n}^2 \\
&= p(1-p) + p(1-p) + \ldots + p(1-p) = np(1-p) \\
\sigma_x &= \sqrt{np(1-p)}
\end{aligned}$$

So for the switch example ($Bin(6, 0.2)$)

$$\begin{aligned}
\mu_x &= 6 \times 0.2 = 1.2 \\
\sigma_x^2 &= 6 \times 0.2 \times 0.8 = 0.96 \\
\sigma_x &= \sqrt{6 \times 0.2 \times 0.8} = \sqrt{0.96} = 0.9798
\end{aligned}$$

Sample Proportions

$$\hat{p} = \frac{\#\text{ successes}}{\text{sample size}} = \frac{X}{n}$$

So if we know X we know $\hat{p}$, and vice versa.

Probability Calculations

We can use this one to one relationship between sample proportions and counts to do probability calculations

Example: Switch example ($Bin(6, 0.2)$)

$$\begin{aligned} P[\hat{p} \geq 0.5] &= P[X \geq 3] \\ &= P[X = 3] + P[X = 4] + P[X = 5] + P[X = 6] \\ &= 0.0989 \end{aligned}$$

We can also use this idea to get means and variances for proportions.

$$\begin{aligned}
\mu_{\hat{p}} &= \frac{1}{n}\mu_x = \frac{1}{n}np = p \\
\sigma^2_{\hat{p}} &= \frac{1}{n^2}\sigma^2_x = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n} \\
\sigma_{\hat{p}} &= \sqrt{\sigma^2_{\hat{p}}} = \sqrt{\frac{p(1-p)}{n}}
\end{aligned}$$

This is based on the rules discussed earlier for linear transformations of random variables.

So for the switch example

$$\begin{aligned}
\mu_{\hat{p}} &= 0.2 \\
\sigma^2_{\hat{p}} &= \frac{0.2 \times 0.8}{6} = 0.02667 \\
\sigma_{\hat{p}} &= \sqrt{\frac{0.2 \times 0.8}{6}} = \sqrt{0.02667} = 0.1633
\end{aligned}$$

Notice that as n increases,

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

decreases. This implies that with a larger sample size, you are more likely to have your sample proportion close to the true population proportion.

Its also a justification of using long run frequencies to motivate probabilities. With a little more work (take Stat 110 to see it), you can show that

$$\hat{p}_n \to p$$

as $n \to \infty$.

Example: Flip a coin 100 times. Count the number of heads. What is $P[\hat{p} \geq 0.6]$? Similarly for 1000 flips.

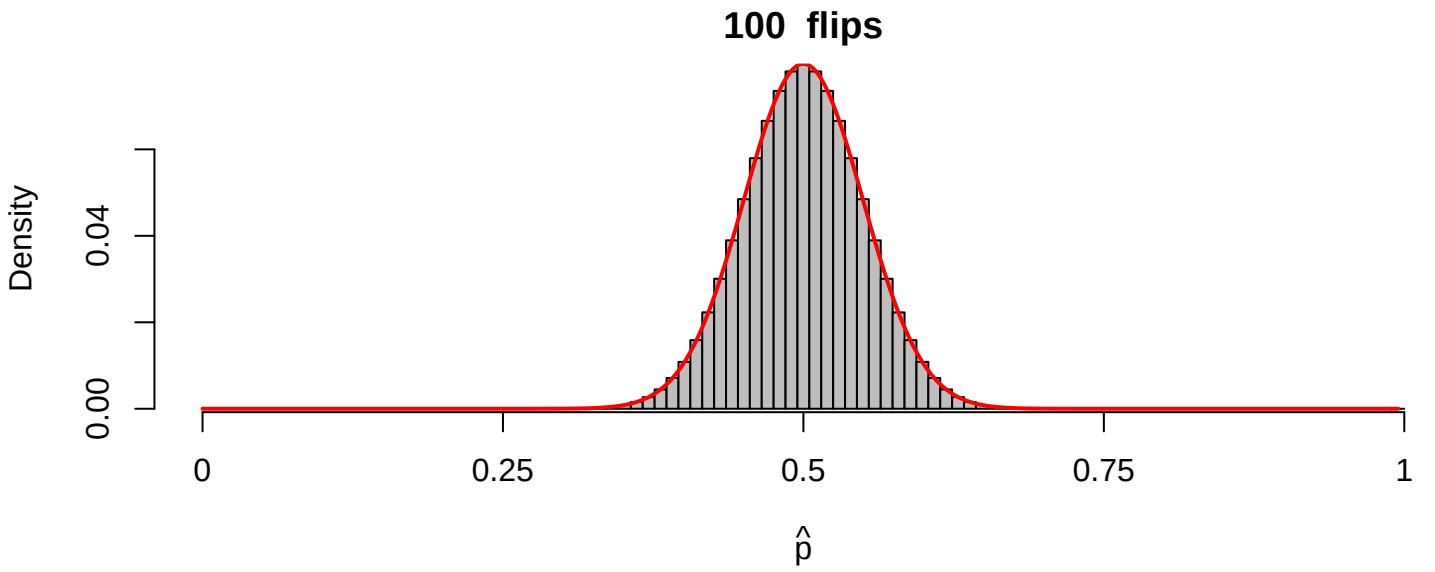
100 flips:

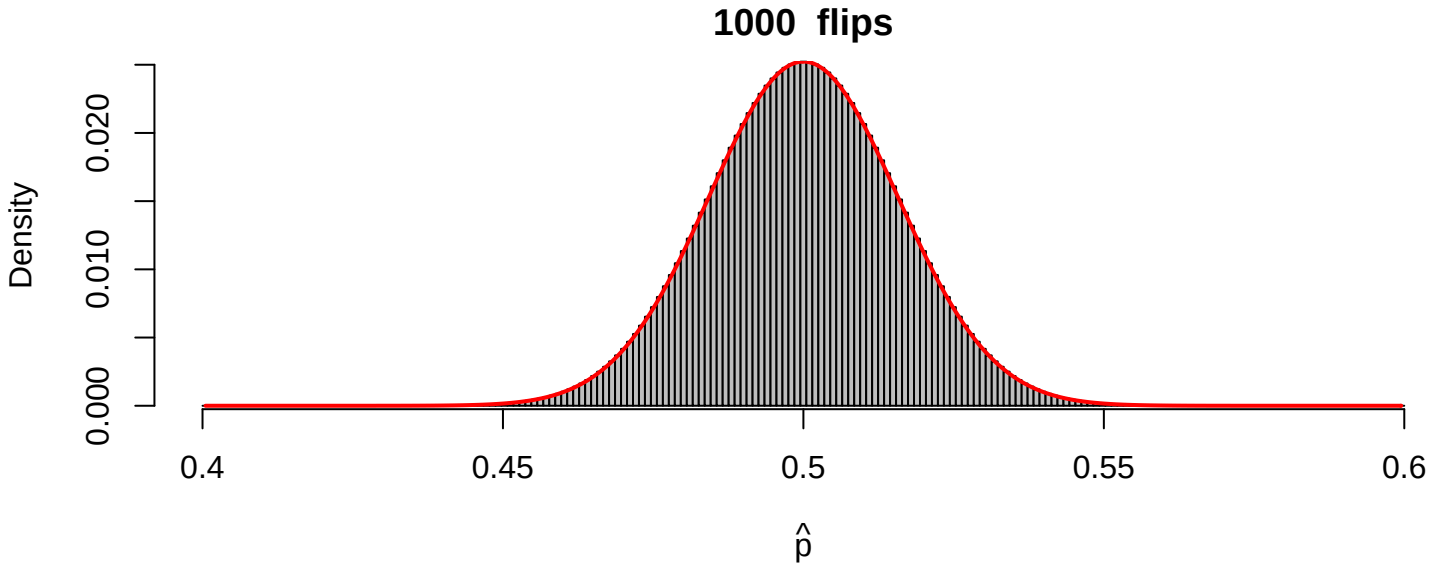
$$\begin{aligned} P[\hat{p} \geq 0.6] &= P[X \geq 60] \\ &= P[X = 60] + P[X = 61] + \ldots + P[X = 100] \end{aligned}$$

1000 flips:

$$\begin{aligned} P[\hat{p} \geq 0.6] &= P[X \geq 600] \\ &= P[X = 600] + P[X = 601] + \ldots + P[X = 1000] \end{aligned}$$

In theory its easy to get the answer – just add up a whole bunch of terms. In fact its easy in Stata as there is a function (`Binomial(n,k,p)`) which gives probabilities of the form $P[X \geq x]$. Other packages have similar functions though most are based on $P[X \leq x]$, the Binomial CDF.


100 flips
Density
0.00
0.04
0
0.25
0.5
0.75
1
p̂


1000 flips
Density
0.000
0.010
0.020
0.4
0.45
0.5
0.55
0.6
p̂

Both of these cases are symmetric and unimodal. In fact, both are close to normal distributions.

Normal Approximation to the Binomial

When n is large, $\hat{p}$ is approximately normally distributed with

$$\begin{aligned} \mu_{\hat{p}} &= p \\ \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

and X is also approximately normal with

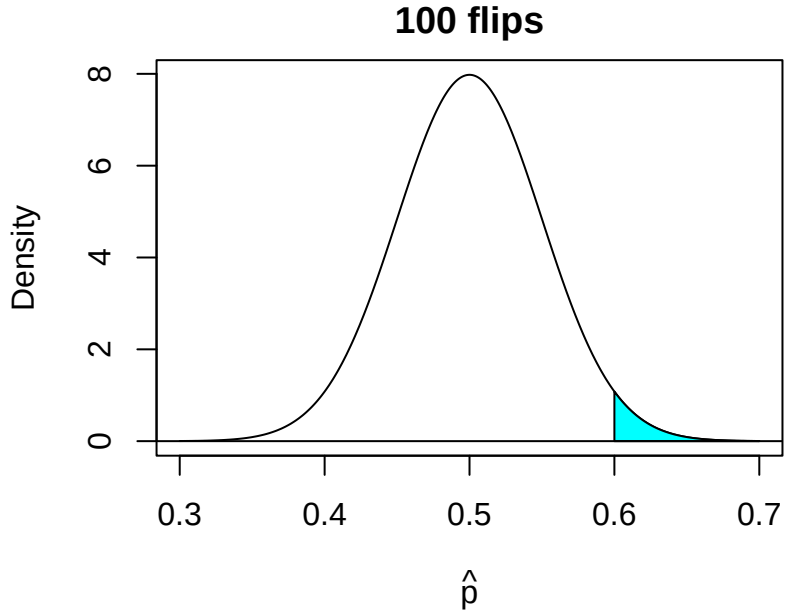
$$\begin{aligned} \mu_x &= np \\ \sigma_x &= \sqrt{np(1-p)} \end{aligned}$$

For $n = 100$ flips

$$\begin{aligned} \mu_{\hat{p}} &= 0.5 \\ \sigma_{\hat{p}} &= \sqrt{\frac{0.5 \times 0.5}{100}} = 0.05 \\ Z &= \frac{\hat{p} - 0.5}{0.05} \text{ is approximately } N(0,1) \end{aligned}$$

$$\begin{aligned} P[\hat{p} \geq 0.6] &= P\left[\frac{\hat{p} - 0.5}{0.05} \geq \frac{0.6 - 0.5}{0.05}\right] \\ &= P[Z \geq 2] \approx 0.0228 \end{aligned}$$

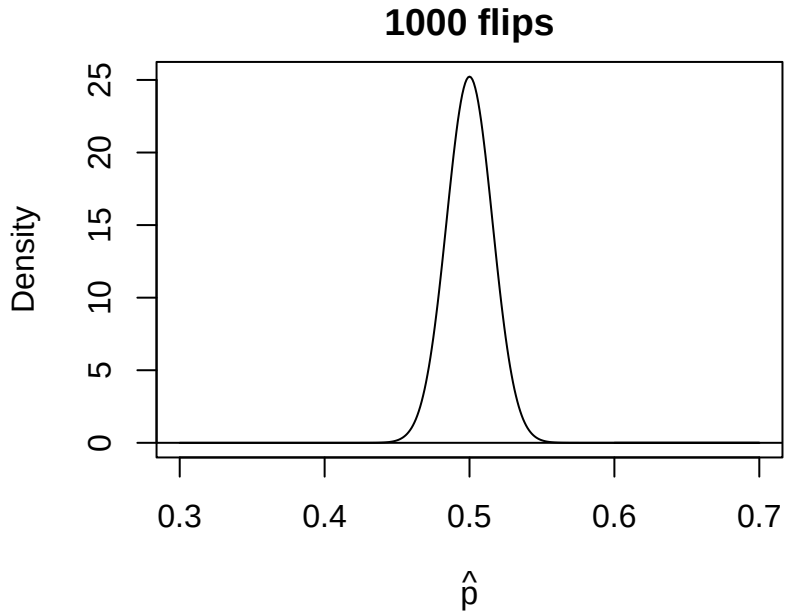
The true probability is 0.0284.



For $n = 1000$ flips

$$\begin{aligned} \mu_{\hat{p}} &= 0.5 \\ \sigma_{\hat{p}} &= \sqrt{\frac{0.5 \times 0.5}{1000}} = 0.0158 \end{aligned}$$

$$\begin{aligned} P[\hat{p} \geq 0.6] &= P\left[\frac{\hat{p} - 0.5}{0.0158} \geq \frac{0.6 - 0.5}{0.0158}\right] \\ &= P[Z \geq 6.329] \\ &\approx 1.234 \times 10^{-10} \end{aligned}$$



The true probability is 1.364×10^{-10}.

Should John Kerry have conceded Ohio while the provisional and absentee ballots still needed to be counted?

Assumptions:

- Kerry is behind by 140,000 votes (its slightly less than this).
- There are 200,000 valid ballots still to be counted (probably a bit higher than actually the case)
- For each ballot, $P[\text{Kerry}] = \frac{2}{3}, P[\text{Bush}] = \frac{1}{3}$ (this is the split in Cuyahoga county, the county John Kerry his highest percentage in Ohio)

For John Kerry to win Ohio, he needs to get over 170,000 (85%) of the 200,000 votes to be counted.

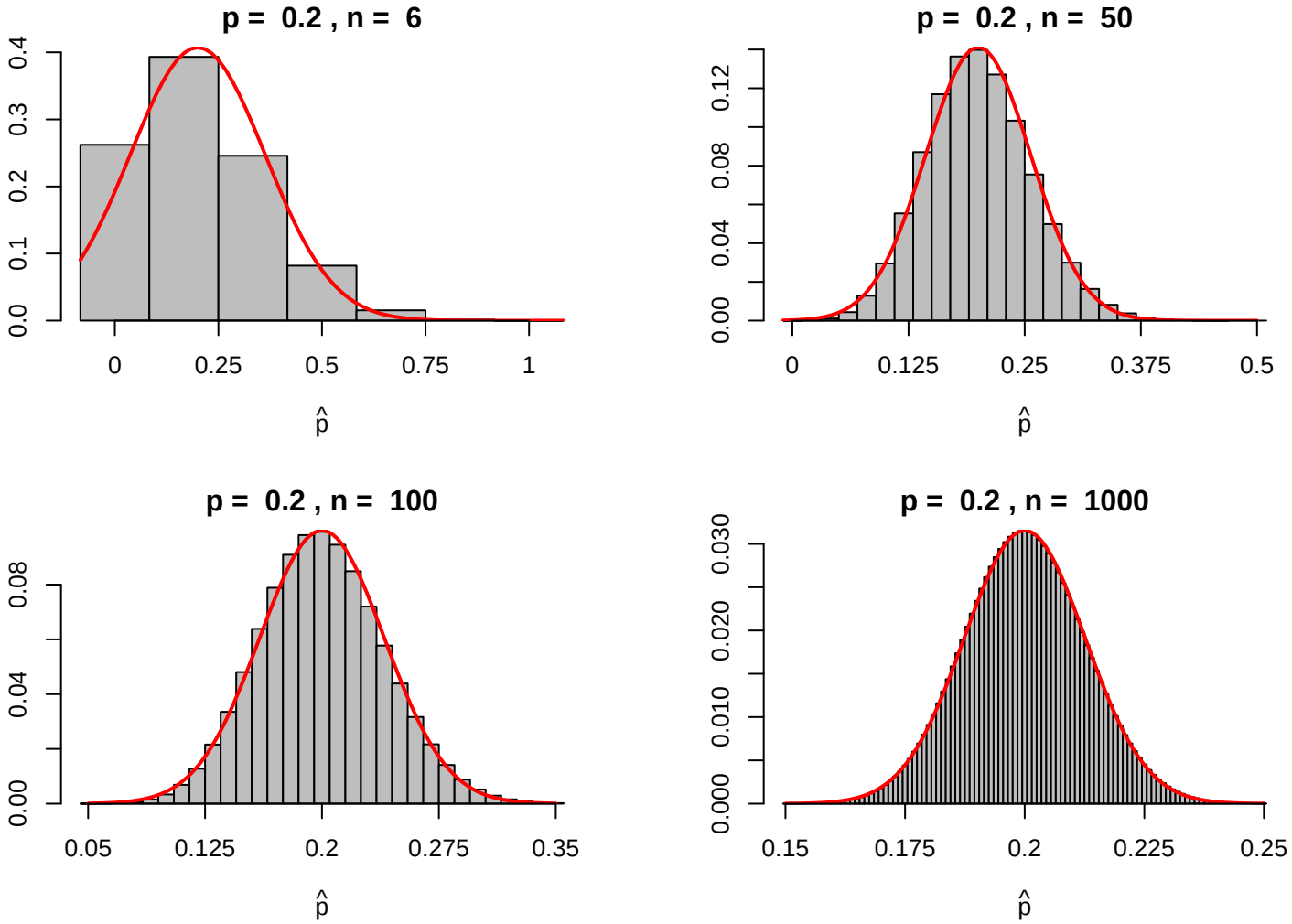
Assuming that these ballots can be considered by a Binomial model with the probabilities given above, what is the probability that John Kerry would get enough votes?

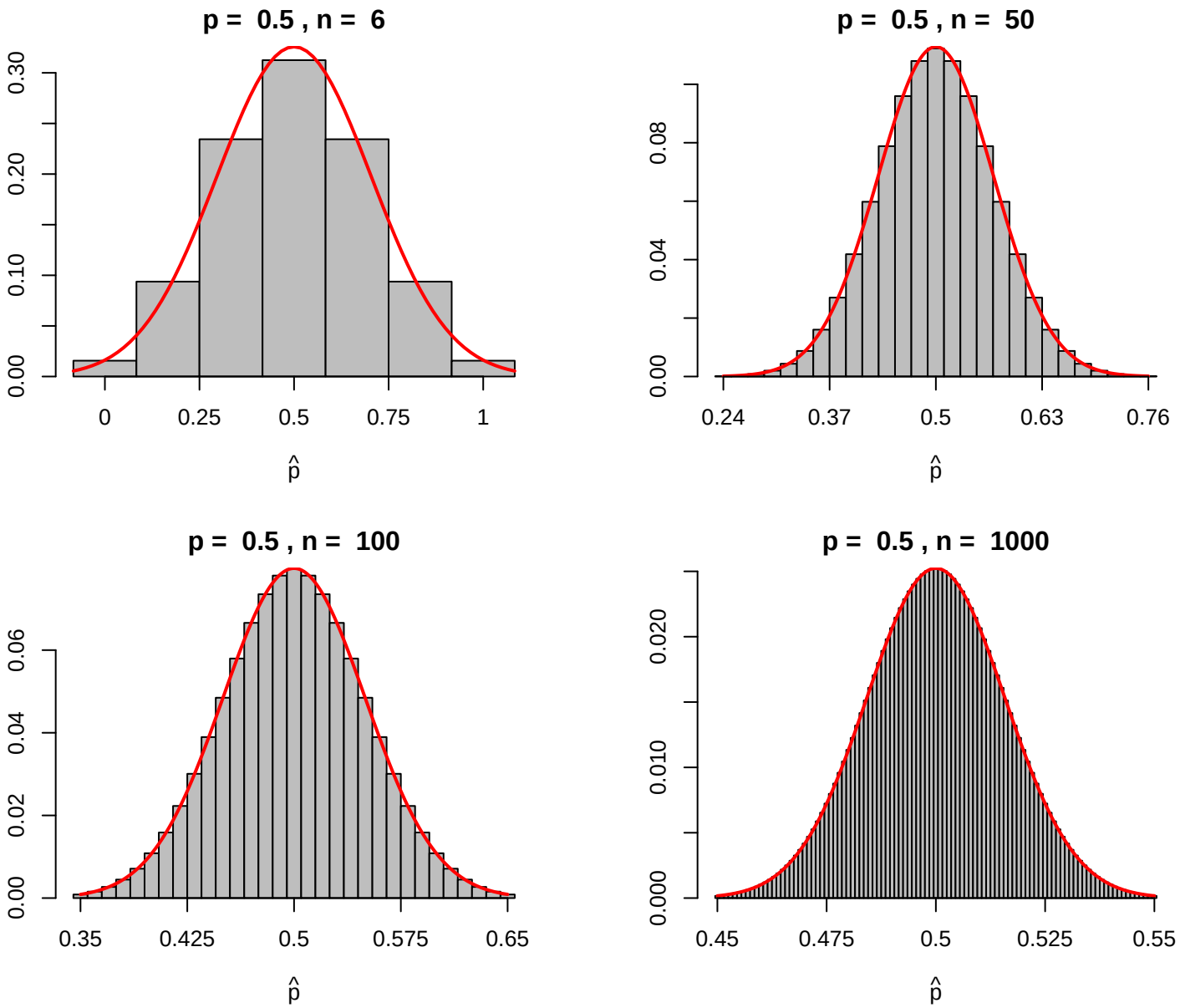
$$\begin{aligned}
\mu_x &= 200000 \times \frac{2}{3} = 133333.3 \\
\sigma_x &= \sqrt{200000 \times \frac{2}{3} \times \frac{1}{3}} = 210.82
\end{aligned}$$

$$\begin{aligned}
P[X \geq 170000] &= P\left[\frac{X - 133333.3}{210.82} \geq \frac{170000 - 133333.3}{210.82}\right] \\
&= P[Z \geq 173.92] \\
&\approx 0 \qquad (< 10^{-6570})
\end{aligned}$$

This is the most extreme z-score I have ever seen. Remember that the table in the book only goes up to 3.49. Kerry has no chance of passing Bush, assuming everything is on the up and up in Ohio.

Now lets look at different combinations of n and p to see how well the approximation works. Let $p = 0.2$ and 0.5 and $n = 6, 49, 100, 1000$.




p = 0.5 , n = 6
0.00
0.10
0.20
0.30
0
0.25
0.5
0.75
1
p̂
p = 0.5 , n = 50
0.00
0.04
0.08
0.24
0.37
0.5
0.63
0.76
p̂
p = 0.5 , n = 100
0.00
0.02
0.04
0.06
0.35
0.425
0.5
0.575
0.65
p̂
p = 0.5 , n = 1000
0.000
0.010
0.020
0.45
0.475
0.5
0.525
0.55
p̂

The approximation appears to work better when n is bigger and when p is close to 0.5.

Rule of Thumb:

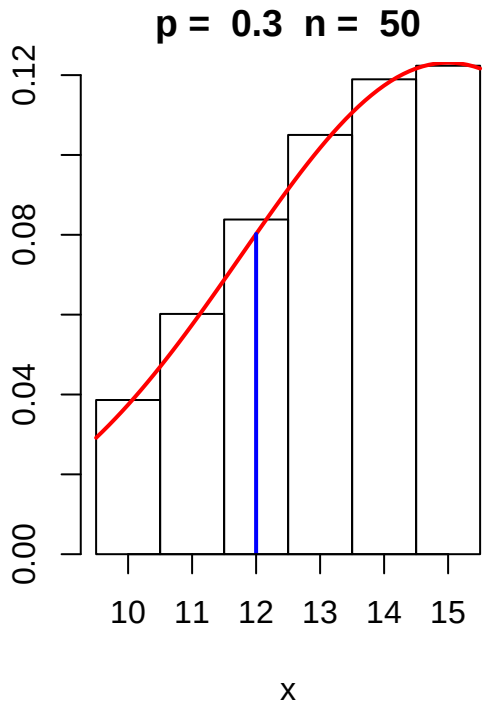
> The approximation is ok if
>
> $$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$
>
> e.g. the expected number and successes and failures are both at least 10.

So the closer p gets to 0 or 1, the bigger n needs to be

So for $p = 0.2$, what is $P[\hat{p} \leq 0.1]$ for various sample sizes

n	Normal Approximation	True Probability
10	0.21460	0.37581
50	0.03855	0.04803
100	0.00621	0.00570
200	0.00020	0.00011

Continuity correction



Suppose we want to get $P[X \leq 12]$ by using the normal approximation. Notice that for the bar corresponding to $X = 12$, the normal curve picks up about half the area, as the bar gets drawn from 11.5 to 12.5.

The normal approximation for this problem can be improved if we ask for the area under the normal curve up to 12.5

True Prob $= 0.2229$

Estimated Prob (no correction) $= 0.1773$

Estimated Prob (correction) $= 0.2202$

While this does give a better answer, for many problems, I recommend ignoring it. If the correction makes an important difference, you probably want to be doing an exact probability calculation instead.