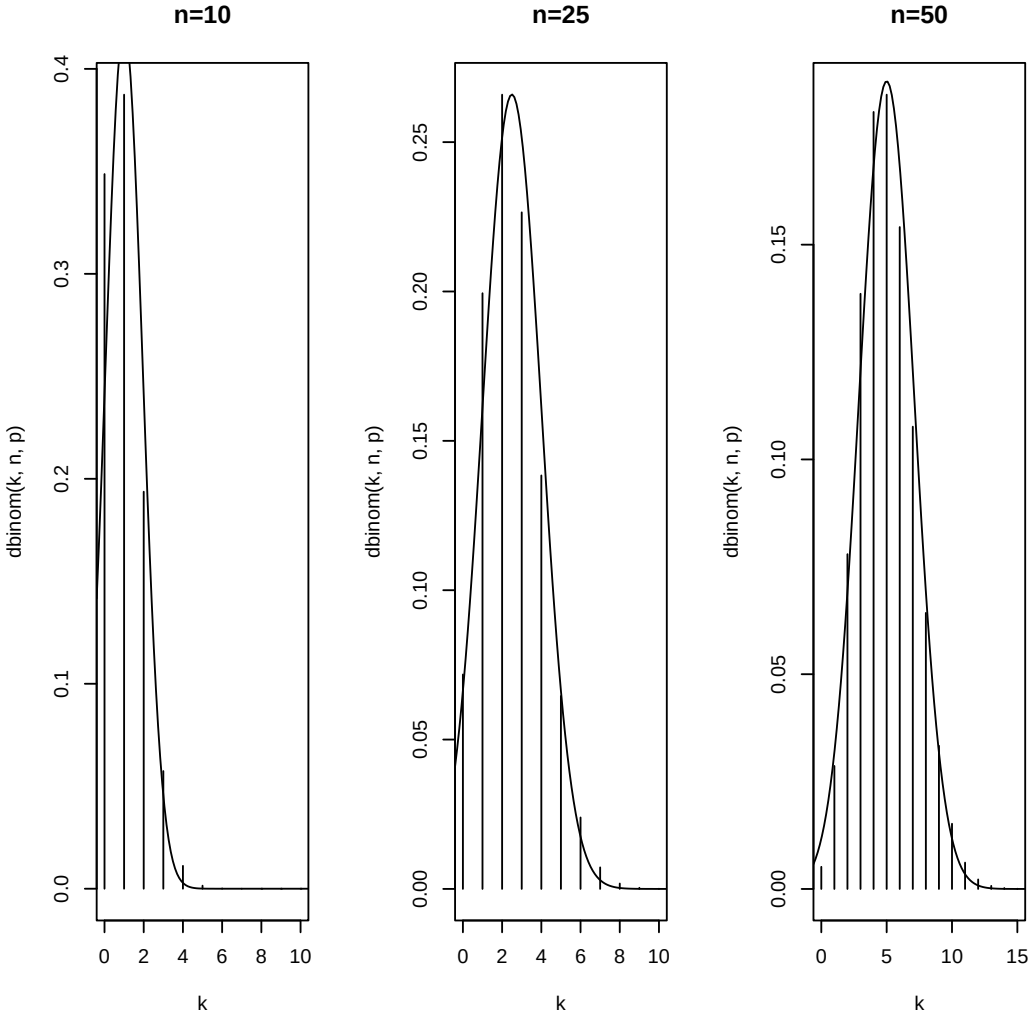
One of the key facts used in this class is that the binomial distribution can be approximated using the normal distribution. This is a consequence of the **central limit theorem** which states the following.

If $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and standard deviation σ then the sample mean, $\bar{X}$ is approximately normal for large n, in particular

$$P(\bar{X} \leq b) \approx P(Z \leq \frac{b-\mu}{\sigma/\sqrt{n}})$$

The normal approximation to the binomial uses $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. The book uses the continuity correction of adding $1/2$ to on the right and subtracting $1/2$ on the left.

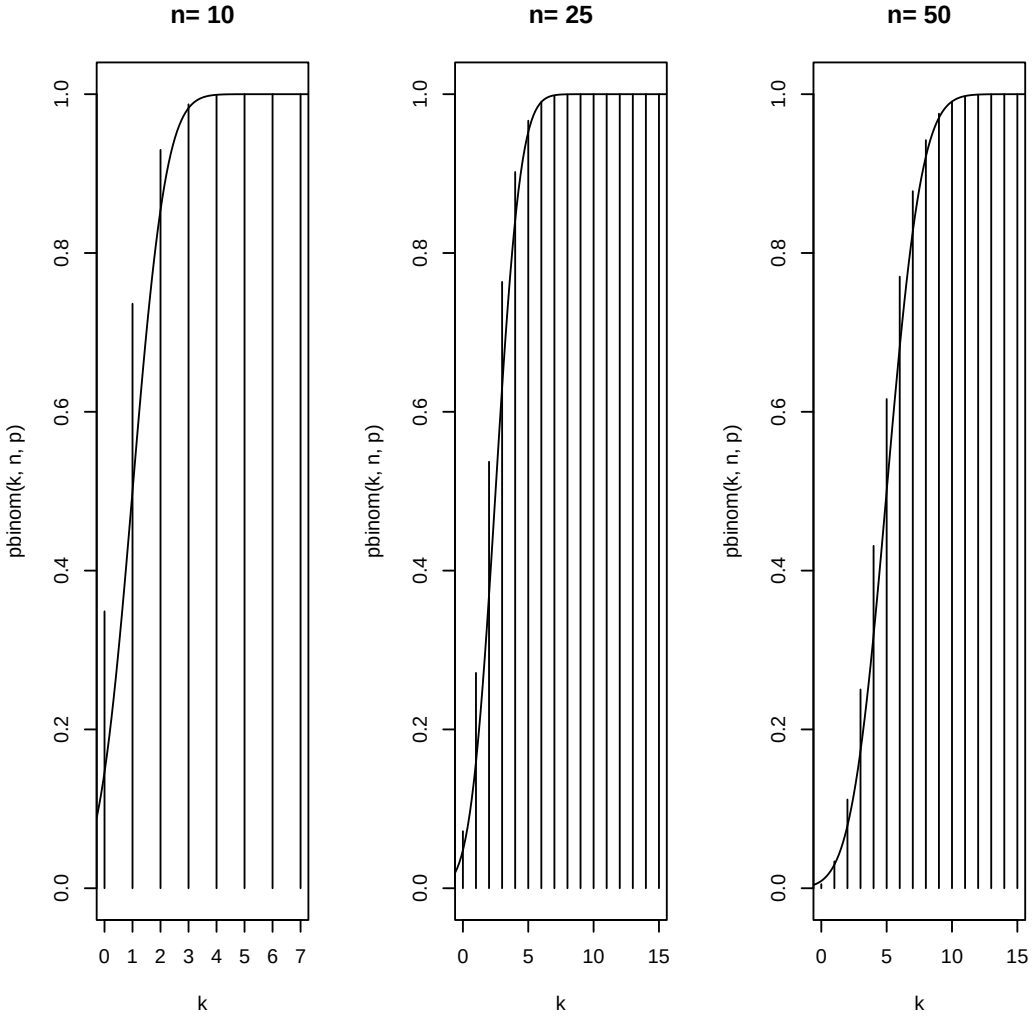
First, lets make some graphs to visualize the binomial distribution and the normal approximation. The figure shows 3 different binomial distributions plotted with vertical bars and the corresponding normal density. The values are for $n = 10, 25$ and 50 and $p = 1/10$.



Question 1: A rule of thumb is that the normal curve fits the binomial well if both np and $n(1-p)$ are 5 or more. Do the graphs bear this out? Explain.

The approximation says the probabilities are similar, not just the graphs. To check this we can use the `pbinom()` and `pnorm()` functions. These add up all the probability to the left. So the graphs look different than above as the area is accumulated. The spikes on the graph give $P(X \leq k)$, the smooth curve gives $P(Y \leq b)$ where X is binomial and Y is normal.

These graphs show the same as above with $n = 10, 25, 50$ and $p = 1/10$.



Question 2: For the graphs do any of them seem to match the probabilities well. Explain.

The importance of this is when we apply it to do calculations not just with the binomial but also the sample mean.

1 taxes

As it is getting near tax season and I have taxes to do, I'll ask some questions about tax returns. For example,

> Suppose the IRS historically has had 10% of filers cheat on their tax returns. If a random sample of 100 is taken from the population of returns, what is the probability 5 or more are from cheaters?

If we have a random sample, then it makes sense to consider the number of cheaters, X, to be binomially distributed with $n = 100$ and $p = 1/10$. The problems asks us to find $P(X \geq 5)$. This is answered exactly with

```
> n = 100
> p = 1/10
> k = 5:100
> sum(dbinom(k, n, p))
```

```
[1] 0.9763
```

Or we can use the normal approximation. This is done with

```
> mu = n * p
> sigma = sqrt(n * p * (1 - p))
> 1 - pnorm(5, mu, sigma)
```

```
[1] 0.9522
```

The continuity correction would subtract 1/2 from the value of 5. This would give

```
> 1 - pnorm(5 - 1/2, mu, sigma)
```

```
[1] 0.9666
```

As you can see all three numbers are fairly close.

Question 3: Suppose the IRS samples 1000 returns, what is the probability that 80 or fewer are from cheaters? Use the normal approximation.

Question 4: Historically, 3 out of 10 returns is done using the 1040EZ form. In a random sample of 100 forms, what is the probability that more than 40 are done using the 1040ez

Question 5: The IRS finds that of suspected cheaters, 35% have fudged on the home office deduction. If they take a random sample of 300 suspected cheaters, what is the probability they find between 90 and 100 who have fudged on their home office deduction? Use the normal approximation.

The central limit theorem applies to more than binomial, in particular to the sample mean. For example,

> Workers in certain sector have tax refunds that average $1000 with a standard deviation of $1000. A random sample from this population of size 25 is taken. What is the probability their average refund is more than $1200?

We use the fact that if n is large enough that $\bar{X}$ is normally distributed with mean μ, the population mean, and standard deviation $\sigma/\sqrt{n}$, where σ is the population standard deviation. That is

$$P(\bar{X} > 1200) = P(Z > \frac{1200 - \mu}{\sigma/\sqrt{n}})$$

So our answer can be found as

```
> n = 25
> mu = 1000
> sigma = 1000
> 1 - pnorm((1200 - mu)/(sigma/sqrt(n)))
```

```
[1] 0.1587
```

Or using the arguments for the mean and sd rather than finding the z-score, we have

```
> 1 - pnorm(1200, mean = mu, sd = sigma/sqrt(n))
```

```
[1] 0.1587
```

Question 6: A sample of size 100 is taken from the same population, what is the probability the average return is more than 1200? Is this the same as the example?

Question 7: For a certain class of returns the amount of charitable deductions as a percentage of deductions averages 10% with a standard deviation of 7%. In a random sample from the same population of size 40 the sample average was 15%. What is the probability it would be 15% or more?

Question 8: The average tax paid for a certain class of workers dropped by 1% with a standard deviation of 5%. If a random sample of 200 is taken from this population, what is the probability the sample average amount saved was negative?