Binomial Distribution and Confidence Interval

Taehoon Ha

Contents

1 binomial random variable

Recall that if Y = number of success in n independent trials, then Y is a binomial random variable with π = probability of success. So $Y \sim \mathrm{B}(n, \pi)$. The PMF of Y is

$$p_Y(y) = \sum_{n=0}^{n} \binom{n}{k} \pi(1-\pi) \quad y = 0, 1, \ldots, n$$

The expected value of Y, $\mathrm{E}(Y)$ is determined by

$$p_Y(y) = \sum_{n=0}^{n} y \times \binom{n}{k} \pi(1-\pi) = n\pi$$

2 Sample proportion

The sample proportion is defined as $P = \frac{Y}{n}$, which is a random variable.

We know that

$$E(P) = E\Big(\frac{Y}{n}\Big) = \frac{E(Y)}{n} = \frac{n\pi}{n} = \pi$$

This is an example of the *method of moments*. The method of moments is a way of estimating parameters, based on matching a moment of the data-generating distribution with the related moment of the empirical distribution.

It works well in a variety of settings, but it can sometimes lead to biased estimators.

2.1 The Maximum Likelihood Estimator

A better way of estimating parameters is by using the *maximum likelihood estimator* (MLE). As the name suggest the MLE is the quantity that maximizes the *likelihood function*.

2.2 The Likelihood Function

The likelihood function is the probability mass function or density evaluated at the data $X_1, \ldots, X_n$, viewed as a *function of the parameter*. Assume we have a set of discrete i.i.d. r.v.'s, $X_1, \ldots, X_n$ whose distribution depends on a parameter θ.

Denote with $p(x|\theta)$ the PMF of each X_i, the likelihood function is then

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

It is often more useful to compute the log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta).$$

Note that we view this as a function of the parameter, so it is important to define the *parameter space*, or the set of values that a parameter can take.

2.3 Example: Likelihood of a Binomial sample

Suppose that we want to test whether a coin is fair, i.e., if the probabilities that it lands on "heads" or "tails" are the same. We can flip the coin a few times, say $n = 15$ and see how many times it gives "heads" ($x = 1$) or "tails" ($x = 0$). Then $Y = X_i + X_2 + \cdots + X_n$ is a binomial random variable, $Y \sim B(n, \pi)$.

More formally, we have a series of i.i.d. Bernoulli random variables (which you can think of as a Binomial with n=1), $X_1, \ldots, X_n$, such that

$$X_i \sim B(1, \pi)$$

Note that we don't know π because we don't know if the coin is fair, but we know the observed values of x_i because we performed the experiment.

We can compute the likelihood function of the n Bernoulli trials:

$$L(\pi) = \prod_{i=1}^{n} \pi^{x_i} (1-\pi)^{1-x_i}, \quad x_i \in \{0, 1\}$$

And the log-likelihood function

$$
\begin{aligned}
l(\pi) &= \sum_{i=1}^{n} x_i \log \pi + (1 - x_i) \log(1 - \pi) \\
&= \log(1 - \pi) + (\log \pi - \log(1 - \pi)) \sum_{i=1}^{n} x_i
\end{aligned}
$$

Assume we have observed these data.

```
x <- rbinom(15, size = 1, prob = 0.5)
x
```

```
##  [1] 0 1 1 1 0 1 1 0 1 1 0 0 1 0 1
```

```
loglik <- function(pi, data) {
   sum(log(dbinom(data, size = 1, prob = pi)))
 }

loglik(pi = 0.5, data = x) %>% round(3)
```

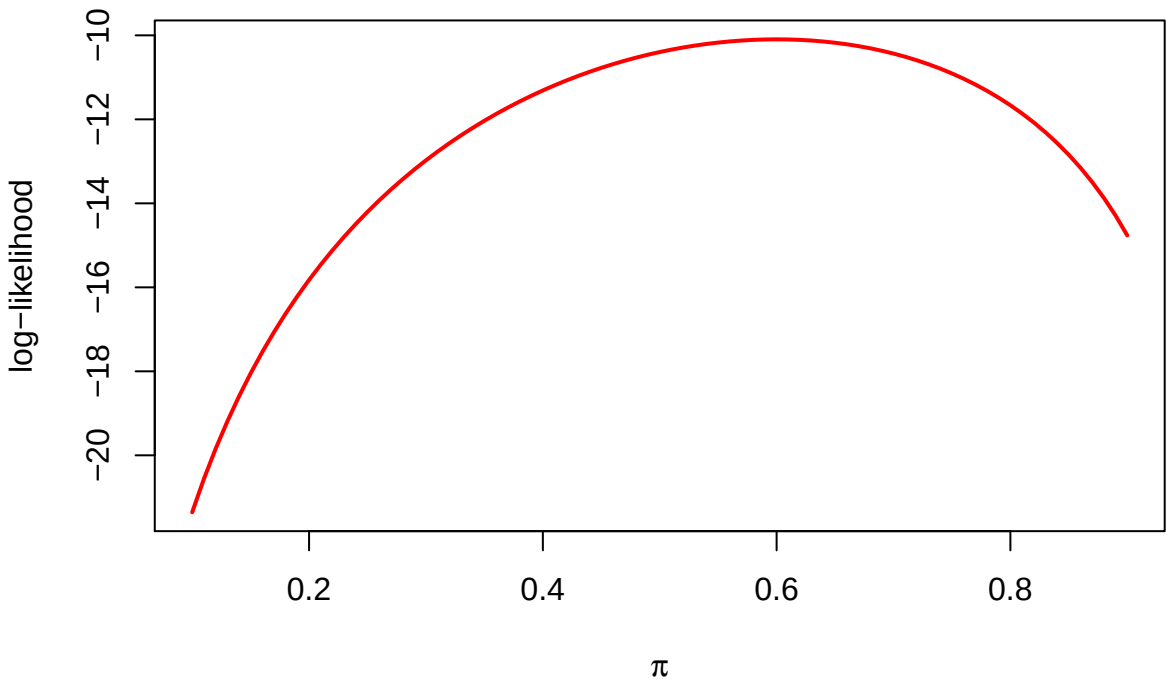
```
## [1] -10.397
```

```
loglik(pi = 0.4, data = x) %>% round(3)
```

```
## [1] -11.312
```

Now let's plot the likelihood over a range of π values

```
pis <- seq(0.1, 0.9, by=0.01)
ll <- sapply(pis, loglik, data=x)
plot(pis, ll, type='l', col=2, lwd=2, xlab=expression(pi),
     ylab = "log-likelihood")
```



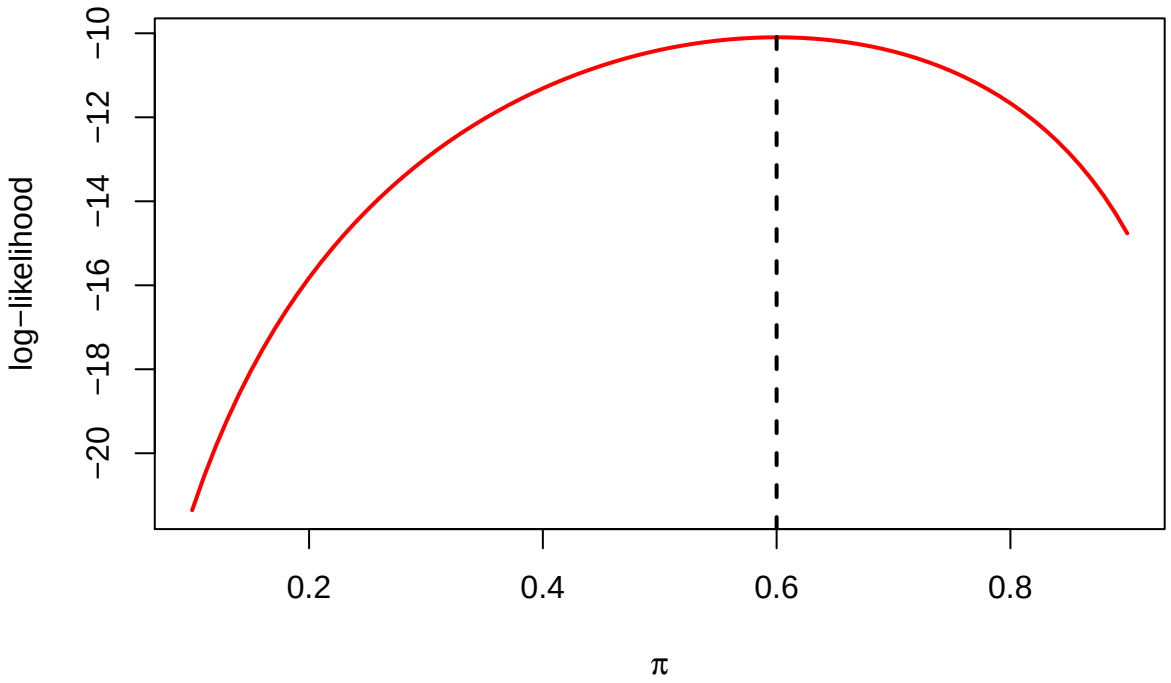
The maximum likelihood estimator (MLE) is the parameter value that maximizes the likelihood (or log-likelihood) function. To compute the MLE we start by taking the derivative of the log-likelihood function with respect to the parameter and we find the value of the parameter for which the derivative is zero. We have already computed the log-likelihood function, we can now take the derivative with respect to θ:

$$\begin{aligned}\frac{dl(\pi)}{d\pi} &= \frac{1}{\pi(1-\pi)}\sum_{i=1}^{n} x_i - \frac{n}{1-\pi} \\ &= \frac{\sum_{i=1}^{n} x_i - n\pi}{\pi(1-\pi)} \\ &= 0\end{aligned}$$

Which implies

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

```
plot(pis, ll, type='l', col=2, lwd=2, xlab=expression(pi),
     ylab = "log-likelihood")
abline(v=mean(x), lty=2, lwd=2)
```



3 Properties of the estimators

3.1 Properties of estimators: Bias

Remember that an estimator $\hat{\theta}$ is a r.v. for which we can compute mean and variance.

We can define the *bias of an estimator* as the following quantity:

Definition

The **bias of an estimator** is defined as

$$\text{Bias} = E[\hat{\theta}] - \theta.$$

An estimator is *unbiased* when

$$E[\hat{\theta}] = \theta$$

3.2 Mean Squared Error

When we compare two estimators, we don't care only about bias but also about variance. We may prefer a biased estimator over an unbiased one, if the bias estimator has smaller variance. We sometimes refer to this as the *bias-variance tradeoff*.

Definition

A measure to compare two estimators is the **mean squared error** (MSE), which combines bias and variance:

$$MSE[\hat{\theta}] = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

In general we want estimators to have *small bias* and *small variance*, which implies smaller MSE.

3.3 Consistency

Another important property is *consistency*.

Definition

An estimator $\hat{\theta}$ is **consisten**t if, as n goes to infinity, it *converges in probability* to the true parameter value θ. for any $\varepsilon > 0$.

One method is to show that

$$\lim_{n\to\infty} Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

Another method is to show that

$$\lim_{n\to\infty} MSE(\hat{\theta}_n) = 0$$

3.4 Example: sample proportion

We have already seen that $E(P) = \pi$, so the sample proportion is an unbiased estimator of π.) The MSE of the sample proportion, P, can be determined. Recall the $Var(P) = \frac{\pi(1-\pi)}{n}$ so

$$MSE(P) = Var(P) + Bias(P)^2 = \frac{\pi(1-\pi)}{n} + 0$$

This implies that an estimator of the population proportion based on a larger sample size will have a smaller MSE than one based on a smaller sample size.

To determine whether the sample proportion is a consistent estimator we note that as n goes to infinity,

$$MSE(P_n) = \frac{\pi(1-\pi)}{n} \to 0$$

and so the sample proportion is consistent.

4 Why Confidence Intervals?

4.1 Confidence Intervals

We now know how to estimate the value of an unknown parameter. Statistics is not only about estimating the unknown quantities in the population, but also about estimating the *uncertainty* of the estimates. Once we have an estimate of our parameter of interest, say $\hat{\theta}$, we want to construct a *range of plausible values* for the true value of θ. We want to be confident, say at the 95% level, that the true parameter lies within a certain range (or interval).

One way to do that is to use *the quantiles of the sampling distribution* to compute the probability that the parameter lies within within the interval.

We know that

$$Pr(q_{0.025} \leq \hat{\theta} \leq q_{0.975}) = 0.95.$$

We also often know, either exactly or approximately, the distribution of $\hat{\theta}$. Hence, we can compute the quantiles to obtain the interval.

4.2 Exact 95% CI for the sample proportion

Assume that we have Y binomial random variable such that $Y \sim B(n, \pi)$.

We can use the quantiles of the sampling distribution of P to compute the confidence interval.

$$\begin{aligned} 0.95 &= Pr(q_{0.025} \leq P \leq q_{0.975}) \\ &= Pr(q_{0.025} \leq \frac{Y}{n} \leq q_{0.975}) \\ &= Pr\left(nq_{0.025} \leq Y \leq nq_{0.975}\right) \end{aligned}$$

We can use R to compute the quantiles for the binomial. The lower limit would be computed by `n*qbinom(0.025, n, \pi)`. Here we run into a problem because we need to know π to determine this value but we are trying to get an estimate of π, somewhat circular reasoning.

Another attempt is to realize that we know the number of observed successes, y. So want we need is to determine the lower limit of the confidence interval, p_L and the upper limit, p_U, by solving the equations.

To find the upper bound, we would solve the equation below for P_U:

$$\sum_{k=0}^{y} \binom{n}{k} p_U^k (1 - p_U^{n-k}) = \frac{0.05}{2} = 0.025$$

To find the lower bound, we would solve the equation below for p_L:

$$\sum_{k=0}^{y-1} \binom{n}{k} p_L^k (1 - p_L^{n-k}) = 1 - \frac{0.05}{2} = 0.975$$

The interval (p_L,p_U) is an exact $100(1-\alpha)$% confidence interval for P.

The equations above that determine p_L and p_U can be solved using available functions. The steps for calculating a 95% confidence interval for the probability of success in a binomial, π are as follow.

Step One: Initalize constants
$\alpha = 0.05$
y = observed number of successes
n = number of trials

Step Two: Define a function for the upper limit and lower limit
$f_u = F(y, p_u, n) - \alpha/2$ (upper limit)

$f_l = F(y-1, p_l, n) - (1-\alpha/2)$ (lower limit)

F is the cumulative density function for the binomial distribution.

Step Three: Solve equations
Find the value of p_u that corresponds to $f_u = 0$ and the value of p_l that corresponds to $f_l = 0$ using software to find the roots of a function.

Here is R code for calculating exact binomial confidence intervals

```
ciLimits<- function(y, n, alpha)
{
  fl <- function(p){pbinom(y-1,n,p) - (1-alpha/2)}
  fu <- function(p){pbinom(y,n,p) - alpha/2}
  pl <- uniroot(fl,c(.01,.99))
  pu = uniroot(fu,c(.01,.99))
  return(c(pl$root, pu$root))
}
```

Common practice in the statistics literature is to refer to the method given here as the Wilson method. There is a similar, but different, method described in Brown, Cai, and DasGupta as the Agresti-Coull method (the Agresti-Coull paper refers to this as the "adjusted Wald" method).

4.3 Example: Number of heads in 15 coin flips

Suppose we are interested in determining whether a coin is fair and we flip it 15 times. We observe that there were 4 heads observed. Based on this data, our estimate for the probability of heads would be $p = \frac{y}{n} = \frac{4}{15} = 0.267$ This would be the *point estimate* for the probability of getting heads with this coin. What would be the 90% confidence interval? Note that in this case α is equal to $\alpha = 1 - 0.90 = 0.10$.

To answer this, we will use the functions defined above to compute this.

```
answer <- ciLimits(4,15,0.10) %>% round(3)
```

We can see that the 90% confidence interval is 0.097, 0.511. Since this interval contains 0.5, the results are consistent with a fair coin.

You are obviously thinking that there must be a function already defined in `R` that can compute this for us. You are correct. The function is `binom.test(y,n)`. The `y` argument is the number of successes you observed and the `n` argument is the number of trials. Let's use `binom.test()` to get our answer.

```
answer2 <- binom.test(4,15,conf.level=0.90)
```

This yields the following 90% confidence interval: 0.097, 0.511, which matches the result above.

4.4 Interpretation of a confidence interval

Although the confidence interval involves the computation of a probability, we have to be careful in interpreting it! Because π is a parameter, not a random variable! The randomness comes from P, the sample proportion, which means that the *boundaries of the interval are random.* Hence we say that, there is a 95% chance the interval contains π and NOT there is a 95% chance that π is in the interval. Beware, this is a subtle use of

language. The item that is random is the interval and NOT π and so the probability is associated with the interval and not with π because π is fixed (but we just do not know its value).

This is worth repeating another way. A 95% confidence interval is a *random interval* generated by `binom.test()` that has probability 0.95 of containing the population proportion π. This means that if we repeat the sampling many times, 95% of the times, on average, the interval will contain the population proportion. Once we observe P, the sample proportion computed for our sample, the interval is known and it either *does or does not* contain the population proportion. Hence, it is *incorrect* to say that there is 95% probability that the population proportion is in the observed interval. This is incorrect because the population proportion is a parameter, not a random variable! Specifically, the value of the population proportion does not vary from sample to sample, only P and the confidence interval varies from sample to sample.

4.5 One-sided confidence intervals

Although 95% is the value that is often used in practice for the confidence level of a confidence interval, it does not have any special meaning. The confidence level could be anything you would want. Other commonly used intervals are the 90% interval and the 99% interval. In general, a confidence interval is denoted as $100(1-\alpha\%)$ so for a 99% interval, $\alpha = 0.01$.

The intervals we generated here are two-sided intervals, that have both a lower bound and an upper bound. Two sided intervals usually assign half the α value to the lower side and half the α to the upper side. We can also generate one-sided intervals where we are only interested in either the lower bound or the upper bound. In this case, all of α goes to one-side. In addition, we have generated a two-sided interval meaning there is both an upper bound and a low boundary.

Suppose we wanted to know whether a coin was biased so that the probability of heads is greater than 0.50. As before, we flip the coin 15 times and observe 4 heads. In this case, we would be interested in having a lower bound for our confidence interval so we can see if the lower bound is greater than 0.50. The `R` command is

```
answer3 <- binom.test(4,15,conf.level=0.90, alternative="greater")
```

The resulting one-sided interval is 0.122, 1. We can see that this interval contains 0.50 and this is consistent with the coin being fair; there is no evidence that the probability of heads is greater than 0.50.

Suppose we wanted to know whether a coin was biased so that the probability of heads is less than 0.50. As before, we flip the coin 15 times and observe 4 heads. In this case, we would be interested in having an upper bound for our confidence interval so we can see if the upper bound is less than 0.50. The `R` command is

```
answer4 <- binom.test(4,15,conf.level=0.90, alternative="less")
```

The resulting one-sided interval is 0, 0.464. We can see that this interval does **NOT** contains 0.50 and this is inconsistent with the coin being fair; there is evidence that the probability of heads is less than 0.50.

Do you know why the boundaries for these one-sided intervals differ from what we got when we obtained the two-sided interval?

Answer

A two-sided 90% interval is such that the proportion of values less than the lower boundary is 0.05 and the proportion of values greater than the upper boundary is 0.05. When we do a one-sided interval that has an upper boundary only (the lower value is 0), this means that the proportion of values greater than that boundary is 0.10. Likewise, when we do a one-sided boundary that has a lower boundary only (the upper value is 1), the proportion of values less than it is 0.10. This is why the boundary values differ between a two-sided interval and the corresponding one-sided intervals. How do you think the one-sided boundaries for a 95% confidence interval would compare to the boundaries

for a two-sided 90% boundary? Check it out.

5 Normal distribution approximation of a binomial distribution

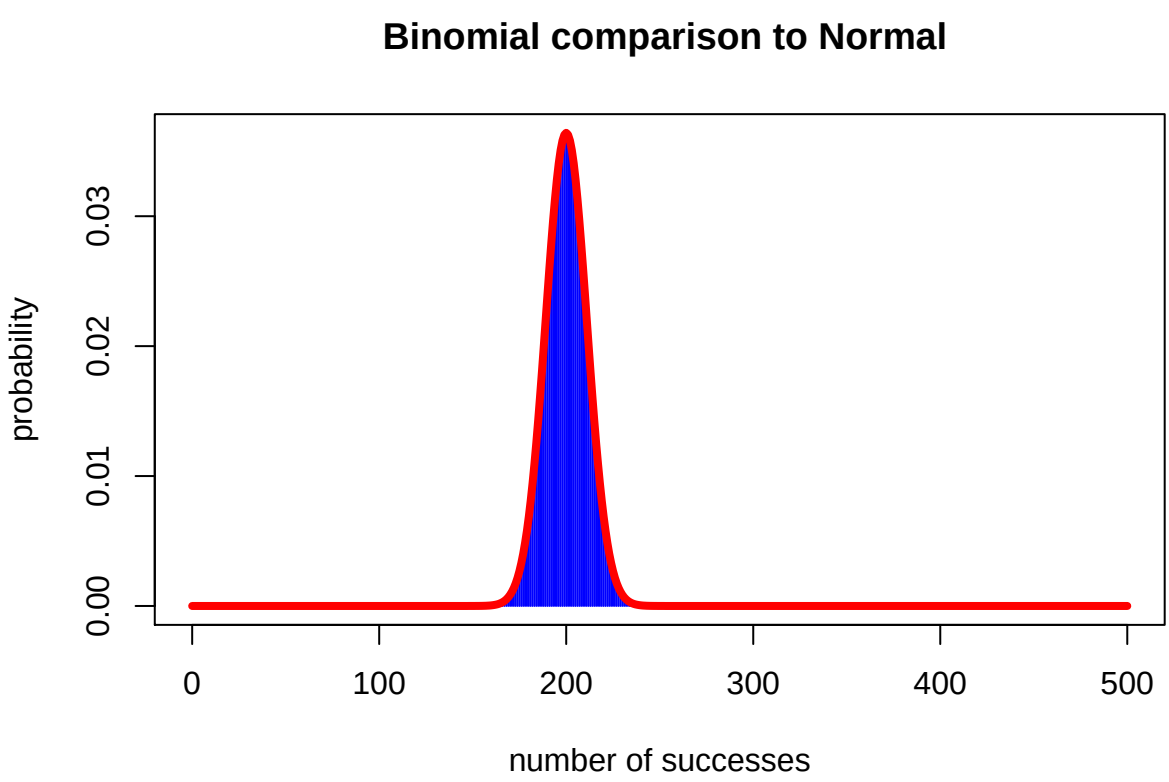
Remarkably, when n, $n\pi$ and $n(1-\pi)$ are large, then the binomial distribution is well approximated by the normal distribution. Specifically,

$$B(n,\pi) \sim N\Big(n\pi, \sqrt{n\pi(1-\pi)}\Big)$$

Note that this says the when n, $n\pi$ and $n\pi(1-\pi)$ are large, the binomial distribution is well approximated with a normal distribution with mean $n\pi$ (the same mean as the binomial) and and standard deviation $\sqrt{n\pi(1-\pi)}$ (the same standard deviation as the normal distribution).

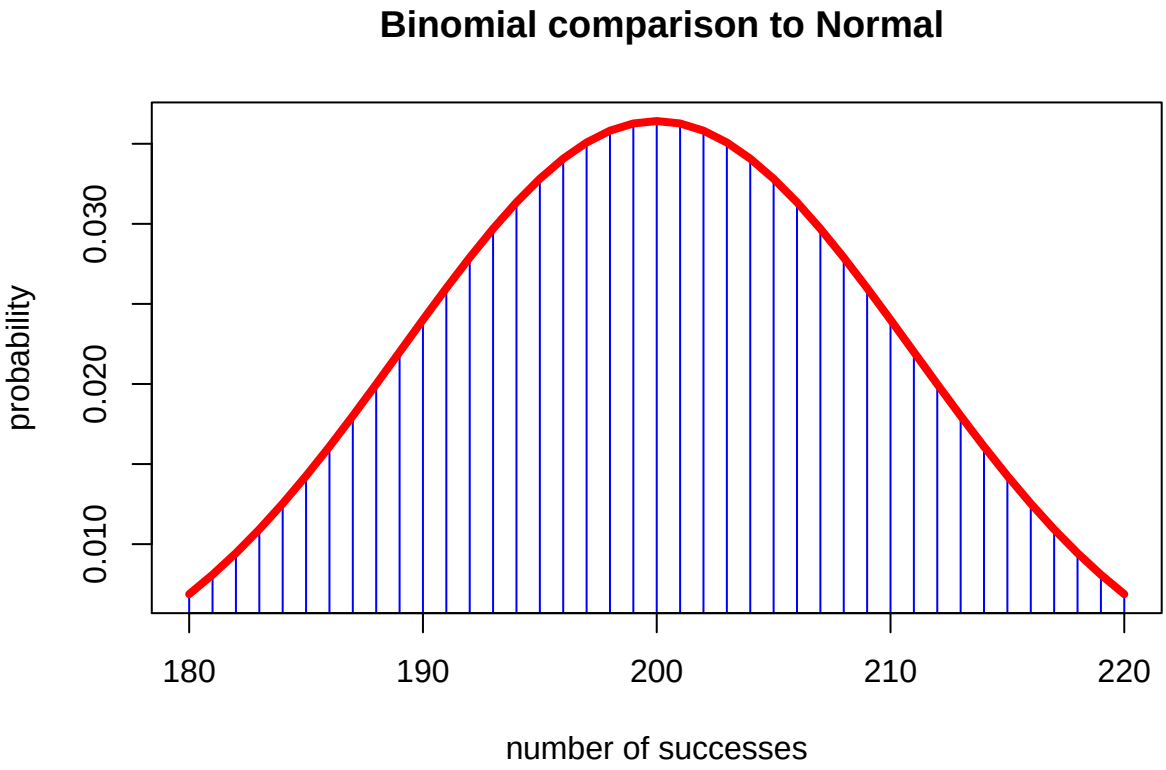
Let's see if this is believable. Suppose Y has a binomial distribution with $n = 500$ and $\pi = 0.40$. Let's plot the PMF for Y and superimpose the PDF for a normal distribution with mean $= 500 \times 0.4 = 200$ and standard deviation $= \sqrt{500 \times 0.4 \times 0.6} = 10.955$.

```
plot(0:500, dbinom(0:500,500,0.4), type = "h", col="blue", xlab = "number of successes",
     ylab="probability", main=" Binomial comparison to Normal")
lines(0:500, dnorm(0:500, 200, sqrt(500*0.4*0.6)),lwd=4, col="red")
```



This looks really good. Let's take a zoom into a range with the bulk of the values lie to see if it still looks good.

```
plot(180:220, dbinom(180:220,500,0.4), type = "h", col="blue", xlab = "number of successes",
     ylab="probability", main=" Binomial comparison to Normal")
lines(180:220, dnorm(180:220, 200, sqrt(500*0.4*0.6)),lwd=4, col="red")
```



Again, the fit looks amazing. The closer π is to 0.5, the better the normal approximation will be. If $n \leq 50$, the approximation will not be so good. A reasonable rule of thumb is that n is large if $n\pi(1-\pi) \geq 10$.

5.1 Implications for the confidence interval for the sampling proportion

Recall there is a relationship between the PMF for Y, the number of successes in n trials with probability of success, π, and the sample proportion P, the proportion of successes in n trials. The difference is the scale on which we are working, the Y scale or the $P = \frac{Y}{n}$ scale. It turns out that as n, $n\pi$, and $n\pi(1-\pi)$ become large, the sampling distribution for the sample proportion can be approximated with a normal distribution with mean π and standard deviation of $\sqrt{\frac{\pi(1-\pi)}{n}}$. So for large n, say $np(1-p) \geq 10$, an approximate $100(1-\alpha)\%$ confidence interval for π, can be determined with a normal distribution.

5.2 Example: Approximaton of 95% CI for the population proportion

We can use the quantiles of the standard normal to compute the confidence interval. Recall that if we have a normal random variable with mean μ and standard deviation σ, we can transform it into a standard normal by subtracting the mean from the random variable and dividing by its standard deviation.

$$0.95 = Pr\left(q_{0.025} \leq \frac{p-\pi}{\sqrt{p(1-p)/n}} \leq q_{0.975}\right)$$

NOTE: we substituted the estimate for π in the formulation for the standard deviation. The quantity $\frac{p-\pi}{\sqrt{p(1-p)/n}}$ has a standard normal distribution.

We can use R to compute the quantiles of the standard normal.

```
qnorm(0.025)
```

```
## [1] -1.959964
```

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Which usually get rounded to -1.96 and 1.96.

An alternative way of writing the interval is the following.

$$0.95 = Pr\left(p - q_{0.975}\sqrt{p(1-p)/n} \leq \pi \leq p + q_{0.975}\sqrt{p(1-p)/n}\right)$$

Or, by substituting the value of $q_{0.975}$

$$0.95 = Pr\left(p - 1.96\sqrt{p(1-p)/n} \leq \pi \leq p + 1.96\sqrt{p(1-p)/n}\right)$$

Let's try to generate the 90% confidence interval two ways. Suppose we have a random sample of 500 individuals from a population and that 212 of them are obese (e.g. have a BMI > 30). What is an estimate for the proportion of obese individuals in this population?

Answer

The estimate for the proportion of obese individuals in this population would be $p = \frac{212}{500} = 0.424$.

What is the exact 90% confidence interval for the proportion of obese individuals?

```
answer5 <- binom.test(212,500,conf.level=0.90)
```

Answer

The exact 90% confidence interval is:

0.387, 0.462

What is the 90% confidence interval obtained from a normal approximation?

From above we see that the lower bound would be $p - q_{0.95}\sqrt{p(1-p)/n}$ because now we want a 90% confidence interval so 5% of the values are below the lower boundary and 5% are above the upper boundary. Likewise, the upper bound would be $p + q_{0.95}\sqrt{p(1-p)/n}$. The quantile we need is

```
qnorm(0.95)
```

```
## [1] 1.644854
```

Answer

The lower bound of the 90% confidence interval would be $.424 - 1.645 \times \sqrt{\frac{.424\ times 0.576}{500}} = 0.388$ and the upper bound would be $.424 + 1.645 \times \sqrt{\frac{.424 \times 0.576}{500}} = 0.460$. Note that these are good approximations for the boundaries produced by the exact interval.

There is an R function that will give the confidence interval for the population proportion based on the normal approximation. It is called prop.test. Let's see what it produces.

```
answer6 <- prop.test(212,500,conf.level=0.90)
```

The 90% confidence interval is 0.387, 0.462. This also yields the same boundaries as the exact method. The reason the R function gives more accurate estimates is that it uses a correction for continuity. Specifically the binomial distribution is discrete and the normal distribution is continuous so a correction is needed to assign the area under the curve to each mass of the binomial (see the Biostatistics textbook.)