Session 1 - Principles of Statistical Inference

Brian Leroux

Wednesday, January 9, 2019

Outline

Populations and Samples

The Standard Error of the Sample Mean

The Sampling Distribution of the Sample Mean (Central Limit Theorem)

Confidence Intervals for a Population Mean

Populations and Samples

What is Statistical Inference

Much of the practice of statistics is concerned with making inferences about a population of interest using a sample from that population. This is called *Statistical Inference*. The population can be either a *finite* or *infinite* population.

For example, consider the finite population of lengths (in miles) of 141 major North American rivers (see R data 'rivers'). Here is a summary of the population:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  135.0   310.0   425.0   591.2   680.0  3710.0 
```

Statistical methods are used when we are not able to observe the entire population. In this case, we draw a sample from the population and make inference about the population using the information in the sample.

How the sample is drawn is extremely important to the validity of the inferences drawn. Valid statistical inference requires a random sampling mechanism.

Simple Random Sample

A simple random sample is one for which all samples of a given size are equally likely. A simple random sample allows for valid inferences to be drawn using simple statistical methods. Other types of random sampling can yield valid inferences if specialized statistical methods are used. The key point is that the sample must be drawn using some type of *random mechanism* in order to justify statistical inference.

We can draw simple random samples from a population in R using the function 'sample'. ('set.seed' allows the results to be replicated if the code is re-run.)

```
set.seed(1)
x=sample(rivers,size=10,replace=F)
x
```

```
 [1]  850  281  618  301  407  310  620 1054  500  465
```

Suppose that we wanted to estimate the average river length and we only had access to the lengths of the above random sample of 10 rivers. We would use the sample mean: 'r mean(x)'.

This is not very close to the population mean value. We will return to this example to study the quality of the sample mean as an estimator of the population mean.

Sampling with and without replacement

Note that we sampled *without replacement*. Sampling without replacement means that we randomly select elements from the population one at a time, and each sampled element is **not** replaced in the population before the next element is sampled. This means that each member of the population can appear in the sample at most once (although if there are tied numerical values in the population you might see the same *numerical value* appear more than once).

In sampling *with replacement*, each sampled element is returned to the population before sampling the next one. This is helpful for thinking about sampling from probability distributions. We will also see this type of sampling in the bootstrap procedure.

Example: sampling with and without replacement

Let the population be the areas (in sq.mi.) of the 50 US states (R data set ‘state.area’).

Sample without replacement:

```
options(width=78)
population = state.area
sample(population, size=20, replace=F)
```

```
 [1]   6450  58560  52586  33215  69919  84068  49576 267339  82264  84916
[11]  45333  56154  69686 158693  58876  36291  51609  56400 110540  97914
```

Sample with replacement:

```
options(width=78)
set.seed(1)
sample(population, size=20, replace=T)
```

```
 [1] 36291 33215  9304 40815  6450  9609 24181 70665 49576 53104  6450 58560
[13] 41222 10577  1214 69686 69919 97914 10577  1214
```

Noe that in the second sample, some elements were selected more than once.

Sampling from an Infinite Population

Sometimes the target population is large enough so that it can be treated as effectively infinite, e.g., people living in the US, manufactured items in a large shipment, cars travelling on a busy highway in a month. When the population is effectively infinite, the sample size is a small fraction of the population size and there is effectively no difference between sampling with or without replacement.

Example: sample 20 elements with and without replacement from the population of the integers from 1 to 1000.

```r
set.seed(12)
population = 1:1000
```

Sample without replacement: 70, 817, 941, 269, 169, 34, 178, 638, 23, 9

Sample with replacement: 393, 814, 377, 381, 265, 440, 458, 541, 666, 113

With an effectively infinite population, we are unlikely to sample any element more than once; sampling with or without replacement are effectively equivalent.

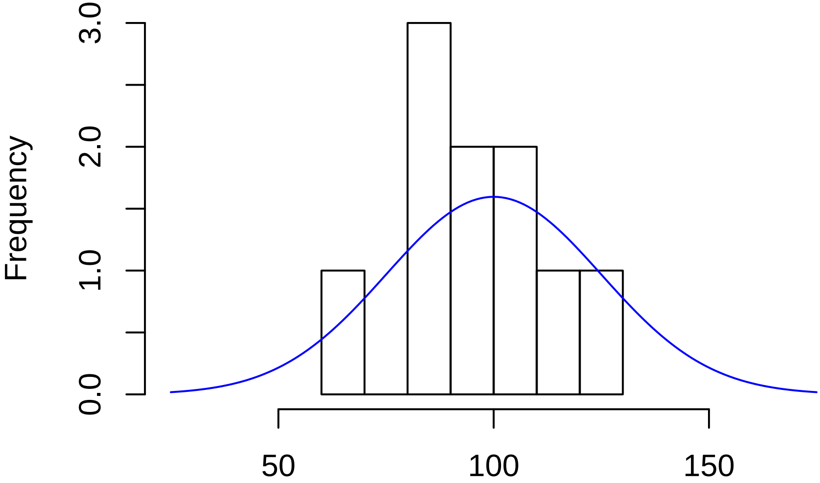
Biased Sampling

A critical aspect of sampling is that the sample be representative of the population. The only practical way to achieve this is with a random mechanism, such as simple random sampling. Examples of misleading conclusions from biased (non-representative) samples are very common.

We will consider how to deal with biased sampling later in the course. For now we will assume representative samples based on random sampling.

Sampling from a Continuous Population Distribution

We can also sample from a continuous probability distribution. A continuous probability distribution is a mathematical representation of an infinite population. Here we show a random sample of size 10 from a normal distribution with mean 100 and SD 25.

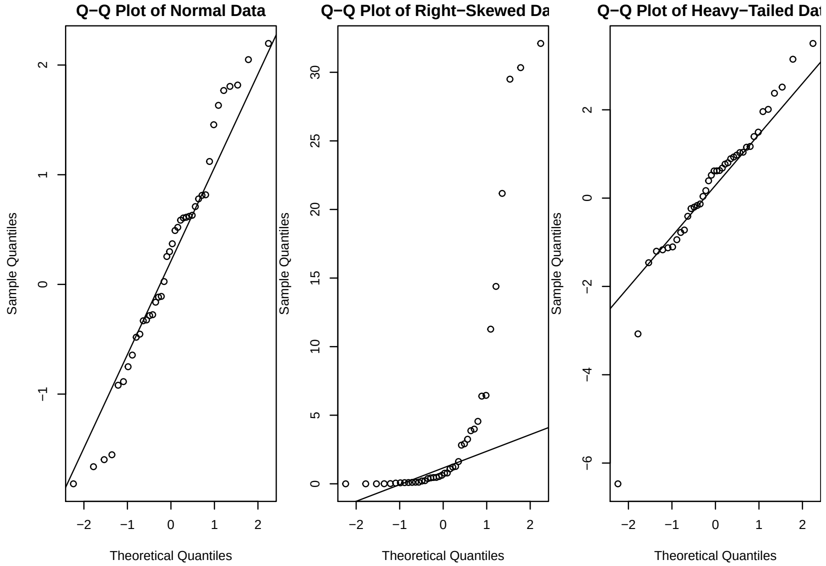


Is it normal?

The q-q plot is a useful graphical way to assess the normality of a sample distribution. The normal q-q plot compares the ordered data values to what would be expected if they had come from a normal distribution. (It is a plot of quantiles of the data compared to quantiles of the normal distribution, hence the name q-q plot.) If the data are normally distributed the q-q plot will form a straight line. (This is particularly useful as part of regression diagnostics.)

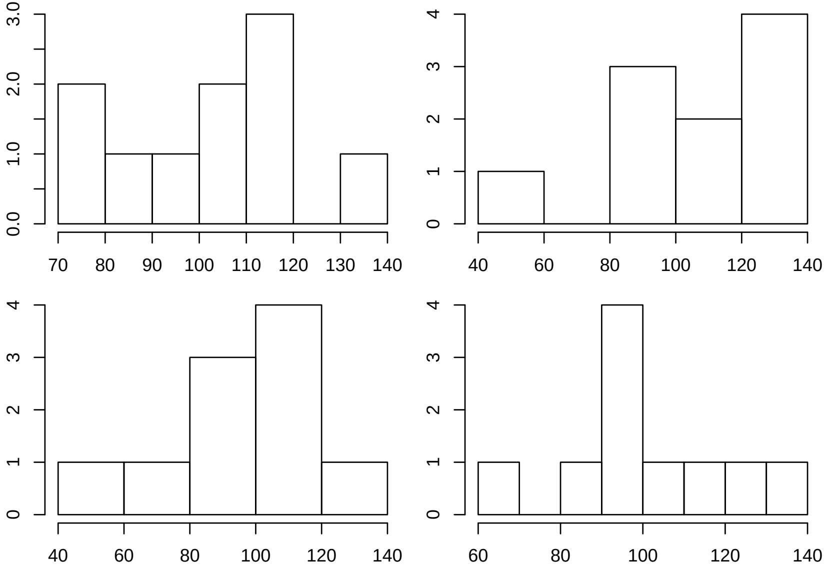
Different departures from normality give distinct patterns in the q-q plot.

Examples of Q-Q Plots

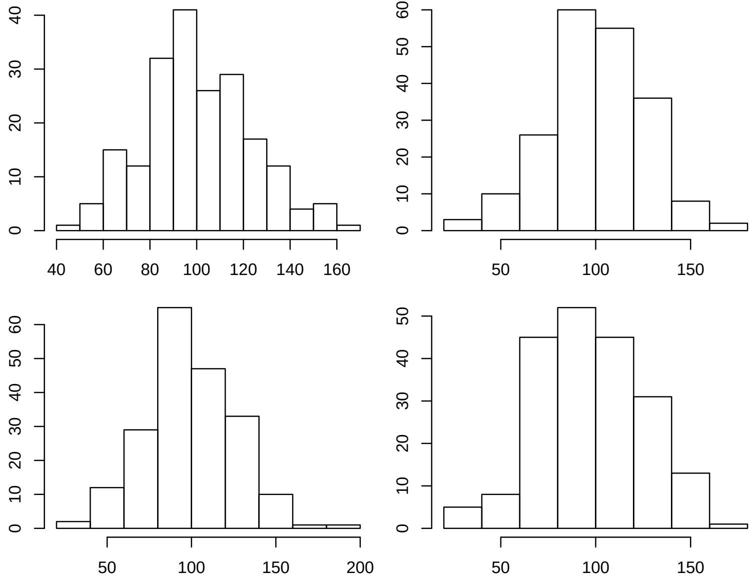


Effect of sample size on the distribution of the sample

For a small sample size a random sample might not resemble the population distribution at all. Below we show repeated samples of size 10 from the N(100,25) distribution.

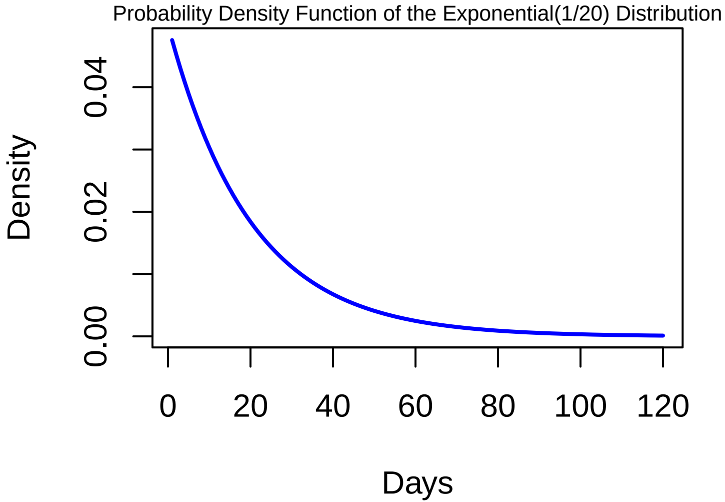


With a larger sample size (e.g., 200), the samples more closely resemble the population distribution.



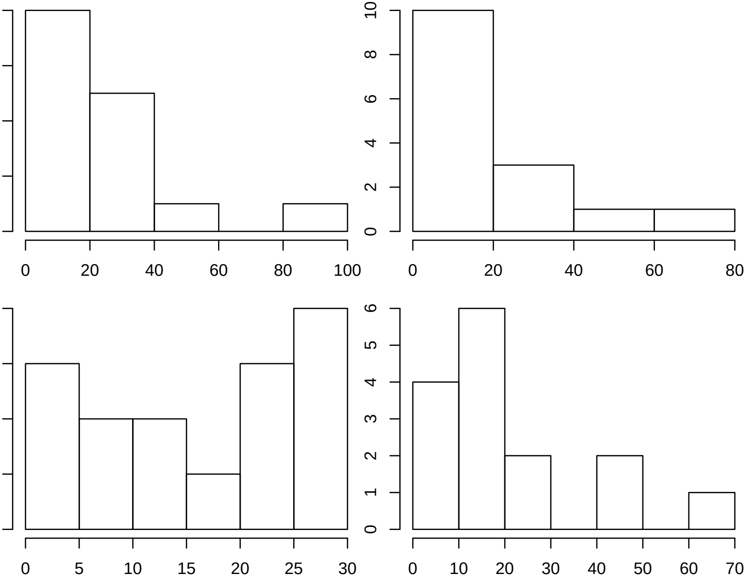
Sampling from the Exponential Distribution

Suppose that the distribution of time to a fault in a system can be described as an Exponential distribution with mean 20 days.

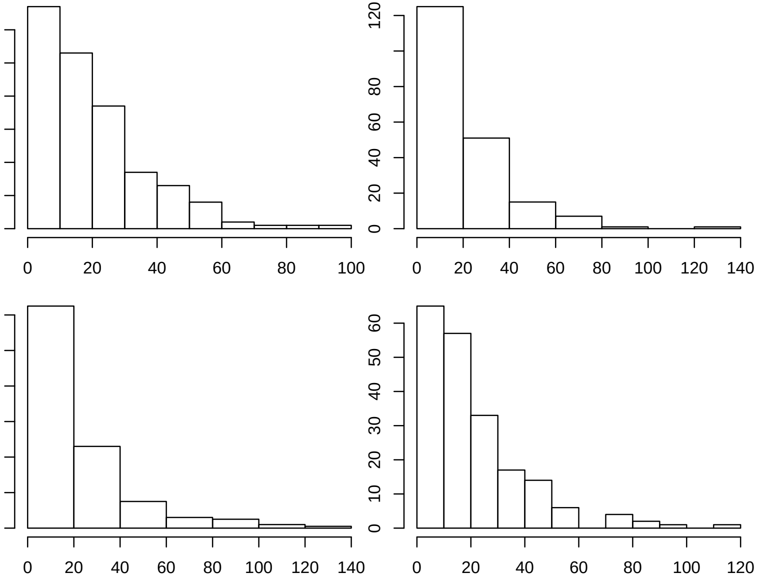


Note: the exponential distribution with parameter λ, denoted $\text{Exp}(\lambda)$, has mean $1/\lambda$. The exponential distribution with mean 20 is $\text{Exp}(1/20)$.

Sampling from Exp$(1/20)$ with $n = 15$



With a sample size of 200, the samples are similar to the population.



The Standard Error of the Sample Mean

Assessing the quality of the sample mean by repeated sampling

In the rivers example, the sample mean for a sample of size 10 was ‘r mean(x)’ which is not all that close to the population mean of 591.2. In practice, we don’t know the population mean, but we want to have some idea of how close our sample mean is to the population mean? We do this by studying how much the sample mean varies from sample to sample.

If we repeat the sampling we expect to get different values for the sample mean, we can see how much variability there is.

```
for(i in 1:5) print(mean(sample(rivers,size=10,replace=F)))
```

```
[1] 765.4
[1] 336.4
[1] 1046.2
[1] 512
[1] 1051.2
```

The sample mean varies quite a lot from sample to sample.

Sampling Distribution of the Sample Mean

By repeating this process many times we can illustrate the variability of the sample mean from sample to sample. The distribution of values we get is called the **sampling distribution** of the sample mean.

Here is a summary of 200 values from the sampling distribution.[1]

```
set.seed(1)
sample.means=rep(NA,200)
for(i in 1:200)sample.means[i]=mean(sample(rivers,size=10,replace=F))
summary(sample.means)
```

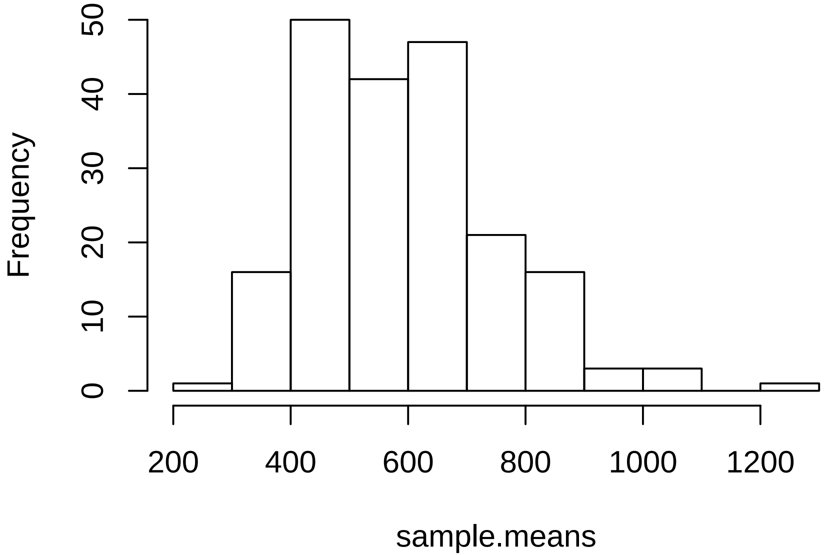
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  285.9   478.9   570.4   595.2   683.3  1216.0 
```

The average value of the sample means is 595.244 which is quite close to the mean of the population. However, the sample means are quite variable, ranging from 285.9 to 1215.6. Although this range is less than the range of values in the population it is large enough to limit the usefulness of the sample mean from any given sample. We can learn more by examining the distribution of the sample means, and then by seeing how the distribution changes as we change the sample size.

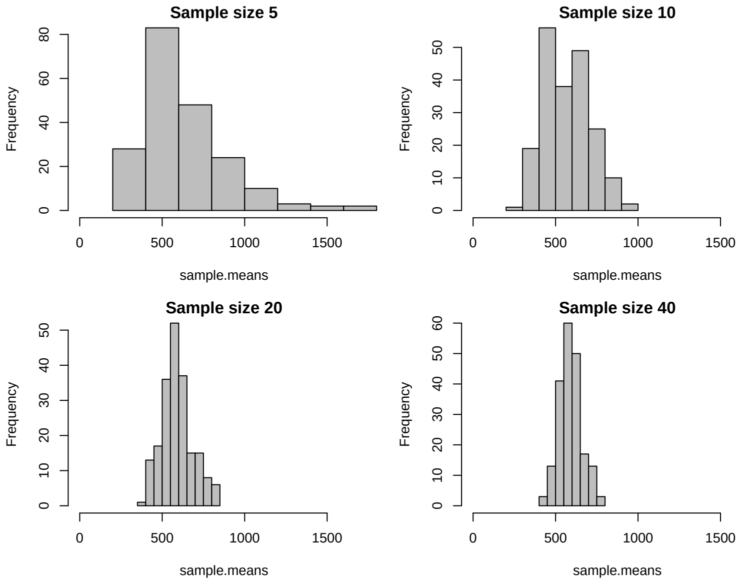
[1] Note: 200 is somewhat arbitray but is chosen to be big enough to yield a reaonably good picture of the distribution.

Histogram Illustrating the Sampling Distribution of the Sample Mean

```r
hist(sample.means, main="")
```



Effect of sample size on the sampling distribution of the sample mean



Effect of sample size (continued)

There are two effects of an increase in sample size:

1. Increasing the sample size reduces the variability of the sample mean.
2. Increasing the sample size makes the distribution of the sample mean more symmetrical and more like a normal distribution.

The first property can be explained in terms of the *standard error* while the second one is explained by the Central Limit Theorem. We will first consider the standard error.

The Standard Error of the Sample Mean

The standard error reflects the amount of variability in the sample mean from sample to sample. Specifically, the standard error is the *standard deviation of the sampling distribution of the sample mean*. From probability theory the formula for the variance of the sample mean is

$$\text{var}(\bar{x}) = \sigma^2/n$$

where σ^2 is the population variance and n is the sample size. Therefore, the standard error is the square root of this

$$\text{SE}(\bar{x}) = \sqrt{\text{var}(\bar{x})} = \sigma/\sqrt{n}$$

where σ is the population standard deviation.

Derivation: From probability theory, $\text{var}(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} \text{var}(x_i) = n\sigma^2$ because the x_i are independent, each with variance σ^2. Therefore, using properties of the variance, $\text{var}(\bar{x}) = \text{var}(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n^2}\text{var}(\sum_{i=1}^{n} x_i) = \sigma^2/n$.

Illustration of SE($\bar{x}$) for an Exponential Distribution

Consider sampling from the Exp(1/20) distribution. The variance of the Exp(λ) distribution is $1/\lambda^2$, (this is a property of the exponential distribution, i.e., that the variance is the square of the mean). For the income distribution $\lambda = 1/20$; hence the population variance of income is $\sigma^2 = (20)^2 = 400$ and the population standard deviation is $\sigma = 20$.

Here are the values of the SE for various sample sizes.

```r
sigma=20
n=c(10,20,50,100,200)
se=sigma/sqrt(n)
data.frame(n,se)
```

```
    n       se
1  10 6.324555
2  20 4.472136
3  50 2.828427
4 100 2.000000
5 200 1.414214
```

Demonstration of SE($\bar{x}$) by simulation

```
set.seed(8888)
sigma=20
n.list=c(10,20,50,100,200)
k=length(n.list)
SE.theoretical=rep(NA,k)
SE.estimated=rep(NA,k)
for(j in 1:k){
  n=n.list[j]
  SE.theoretical[j]=sigma/sqrt(n)
  sample.means=rep(NA,1000)
  for(i in 1:1000)sample.means[i]=mean(rexp(n,rate=1/20))
  SE.estimated[j]=sd(sample.means)
}
data.frame(n.list,SE.theoretical,SE.estimated)
```

```
  n.list SE.theoretical SE.estimated
1     10       6.324555     6.023386
2     20       4.472136     4.445155
3     50       2.828427     2.767786
4    100       2.000000     1.936146
5    200       1.414214     1.411880
```

Note that we simulated a large number (1000) of data sets for each value of n in order to limit random simulation error and get a good match to the theory.

SE for a sample proportion

A sample proportion is a sample mean so the SE formula applies to sample proportions as well. For example, if we take a random sample of size n from the Bernoulli(p) distribution, the sample mean is the sample proportion, i.e., $\bar{x} = \hat{p} = \text{Number of 1s}/n$.

The formula for SE($\bar{x}$) that we saw before also applies to a Bernoulli distribution; we only need to specify the value of σ in the formula. For the Bernoulli(p) distribution, the variance is $p(1-p)$.[2] Therefore, $\sigma = \sqrt{p(1-p)}$, and the SE of $\hat{p}$ is

$$\text{SE}(\hat{p}) = \sqrt{p(1-p)/n}$$

.

[2] Derivation: From probability theory, for any random variable x, $\text{var}(x) = E(x^2) - [E(x)]^2$. If x is binary, then $\text{var}(x) = p - p^2 = p(1-p)$.

Demonstration of the SE of a sample proportion by simulation

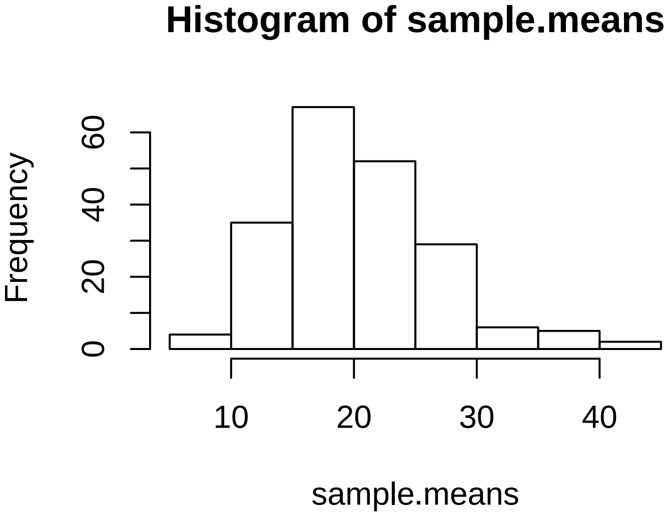
```r
set.seed(4)
p=0.5
n.list=c(50,100,200,500,2000)
k=length(n.list)
SE.theoretical=rep(NA,k)
SE.estimated=rep(NA,k)
for(j in 1:k){
  n=n.list[j]
  SE.theoretical[j]=sqrt(p*(1-p)/n)
  sample.proportions=rep(NA,1000)
  for(i in 1:1000)sample.proportions[i]=mean(rbinom(n,size=1,prob=p))
  SE.estimated[j]=sd(sample.proportions)
}
data.frame(n.list,SE.theoretical,SE.estimated)
```

```
  n.list SE.theoretical SE.estimated
1     50     0.07071068   0.07111422
2    100     0.05000000   0.04900875
3    200     0.03535534   0.03428286
4    500     0.02236068   0.02187590
5   2000     0.01118034   0.01078830
```

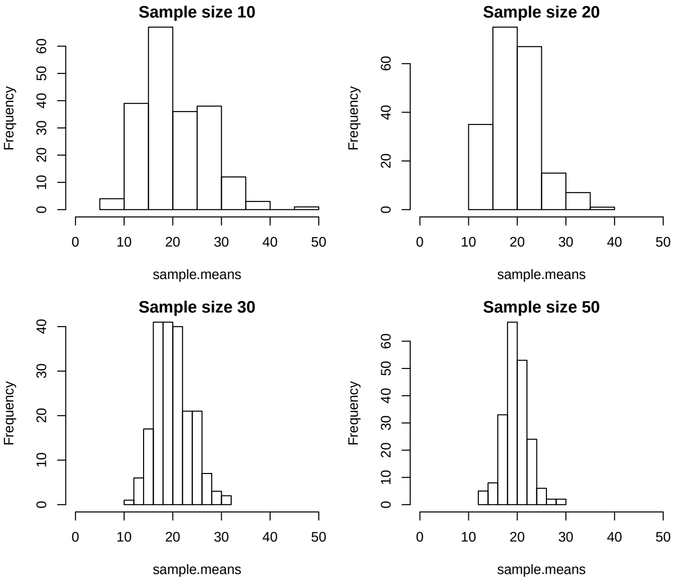
The Sampling Distribution of the Sample Mean (Central Limit Theorem)

Distribution of the sample mean for an exponential distribution

```r
set.seed(3333)
sample.means=rep(NA,200)
for(i in 1:200)sample.means[i]=mean(rexp(10,rate=1/20))
hist(sample.means,cex=0.8)
```



Effect of sample size on the distribution of the sample mean for an exponential distribution



The Standard Error and the Central Limit Theorem

Just as we saw for the sampling from the population of rivers, the variability of the sample mean becomes smaller as the sample size increase. Also, as the sample size increases, the distribution of the sample mean becomes more symmetric (and more like a normal distribution).

There are two concepts in probability theory that explain these observations: the *Standard Error*, and the *Central Limit Theorem*.

Sampling from a Binary Variable

In political polls the target population is the voting intention of all registered voters, which we conceptualize as a set of 1s (for those voting for Part A) and 0s (for those voting for Party B). The target parameter is *p* the proportion of voters for Party A.

Suppose the population has 50,000 voters, with 25,000 for Party A and 25,000 for Party B. Then, sampling with or without replacement are almost equivalent for a sample of size 2000.

```
options(width=72)
set.seed(3)
population = c(rep(1,25000),rep(0,25000))
sample.proportions=rep(NA,200)
for(i in 1:1000)sample.proportions[i]=mean(sample(population,size=2000,replace=
summary(sample.proportions)
```

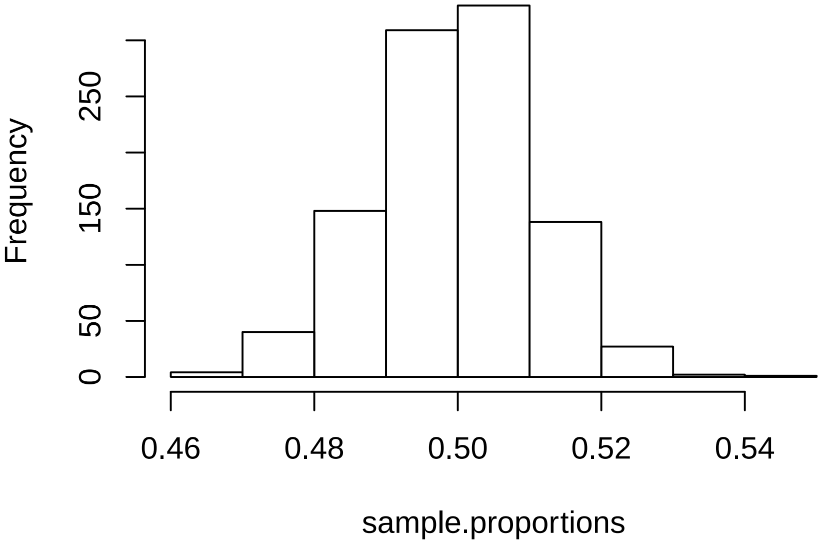
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4670  0.4925  0.4995  0.5000  0.5075  0.5375 
```

```
for(i in 1:1000)sample.proportions[i]=mean(sample(population,size=2000,replace=
summary(sample.proportions)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4645  0.4925  0.5000  0.4997  0.5070  0.5450 
```

Sampling Distribution of the Sample Proportion

```r
hist(sample.proportions,main="")
```



Sampling from a Bernoulli Distribution

The sampling of voter intentions with replacement can be thought of as sampling from a Bernoulli distribution. The Bernoulli distribution is a Binomial distribution with number of trials $= 1$, denoted as Binomial$(1, p)$ or Bernoulli(p).

Simulate sampling $n = 2000$ voters from an infinite population with true proportion $p = 0.5$.

```r
set.seed(2222)
sample.means=rep(NA,200)
for(i in 1:200)sample.means[i]=mean(rbinom(2000,size=1,prob=0.5))
summary(sample.means)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4730  0.4950  0.5010  0.5011  0.5085  0.5310 
```

Note that this gives similar results to those obtained from sampling either with or without replacement from the finite population of 50,000 voters.

The Central Limit Theorem

The simulations of sampling distributions illustrate how the distribution of sample means and sample proportions look like normal distributions when the sample size is large.

The probablity theory behind this property is called the *Central Limit Theorem* (CLT). The CLT says that for almost any probability distribution for the population[3], the sample mean will have an approximate normal distribution if the sample size is large. The same result holds for a sample proportion, which is a sample mean based on sampling from a Bernoulli distribution.

[3] The exceptions are technical oddities rarely applicable in practice.

The CLT for sampling from the Bernoulli distribution

Consider a study of the effect of on-line advertising on the probability of a sale. For each individual viewing the ad the outcome is a Bernoulli(p) random variable with values 0 (no sale) or 1 (sale) where p is the probability of a sale. Suppose that $p = 0.0001$ and we observe 10,000 trials. We could simulate sampling n=10,000 individuals from the Bernoulli(0.0001) distribution and tabulate the results.

```r
options(width=72)
set.seed(2)
sample.proportions=rep(NA,200)
for(i in 1:200)  sample.proportions[i] = mean(rbinom(10000,size=1,prob=0.0001))
table(sample.proportions)
```

```
sample.proportions
    0 1e-04 2e-04 3e-04 4e-04 5e-04 
   84    63    35    12     2     4 
```

We actually don't need to use simulation here – probability theory can tell us what the distribution is. Note that the various values of the sample proportions correspond to the number of sales out of n=10,000 trials being equal to 0, 1, 2, etc. The probabilities of these values can be calculated from the binomial distribution.

Using the Binomial distribution

If X is the total number of sales (i.e., X/n is the sample proportion), then X follows a Binomial(n, p) distribution which has probability mass function

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \; x = 0, 1, 2, \ldots, n,$$

The distribution for $n = 10000$, $p = 0.0001$ is tabulated below along with the relative frequencies from the simulation.

```r
options(width=72)
n=0:5
Relative.Frequency = as.numeric(table(sample.proportions)/200)
data.frame(Number=n, Proportion=n/10000, Probability=dbinom(n,size=10000,prob=0
```

```
  Number Proportion Probability Relative.Frequency
1      0      0e+00 0.367861046              0.420
2      1      1e-04 0.367897836              0.315
3      2      2e-04 0.183948918              0.175
4      3      3e-04 0.061310174              0.060
5      4      4e-04 0.015324478              0.010
6      5      5e-04 0.003063976              0.020
```

These relative frequencies from the simulation match the theoretical probabilities quite well. If we increased the number of simulations, e.g., 2000 instead of 200, we would get an even better match.

Normal Approximation to the Binomial Distribution

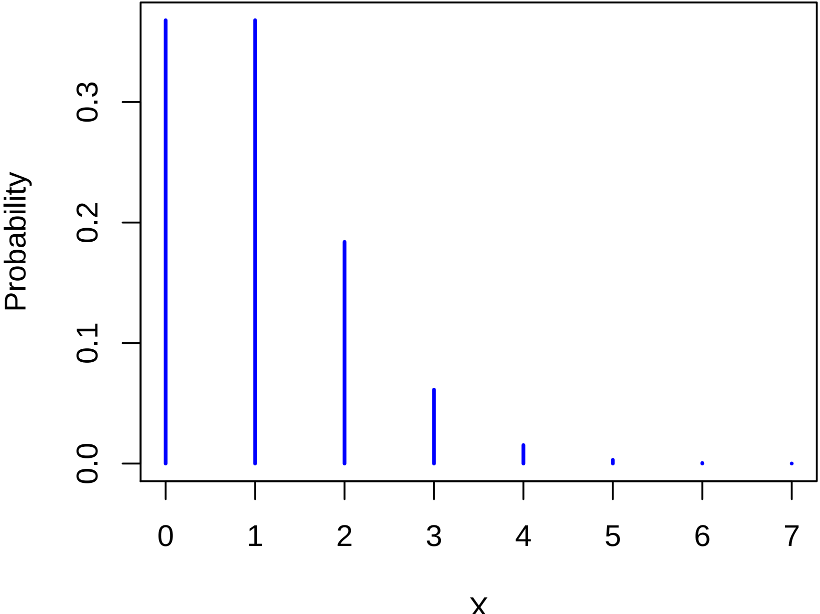
The probabilities are graphed on the next slide.

The mean of the distribution is $np = 10000 \times 0.0001 = 1$ and the SD is $\sqrt{np(1-p)} = 1$.

Note that the distribution is not well approximated by a normal distribution. Although the sample size is quite large (n=10,000), it is not large enough for the CLT because of the small probability, $p = 0.0001$.

Probability mass function of Binomial(10000,0.0001)

```
plot(0:7,dbinom(0:7,10000,0.0001),type="h",col="blue",lwd=2,ylab="Probability",
```



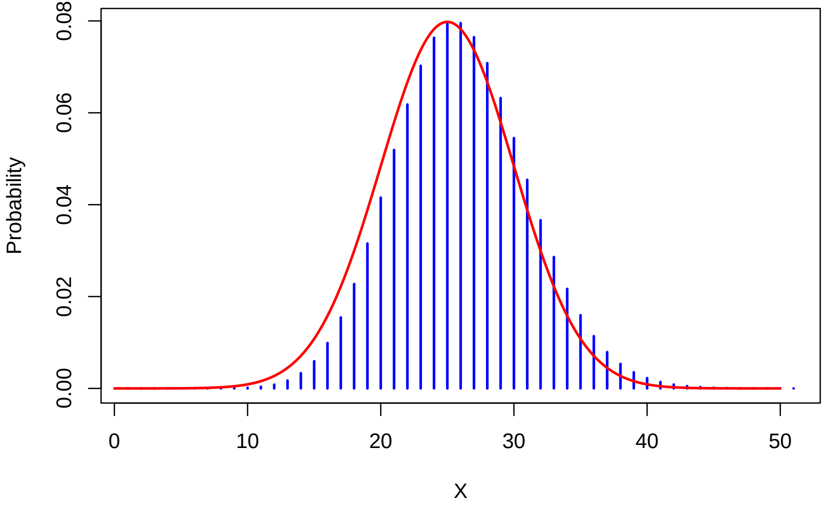
How large does the sample size need to be for the CLT?

As we have seen, the sample size requirement for the CLT is very much dependent on the context. For some distributions, relatively small sample sizes are sufficient.

The problem with the current example is that the probability p is very small. In this case, an extremely large sample size is needed to make the distribution of the sample proportion approximate a normal distribution.

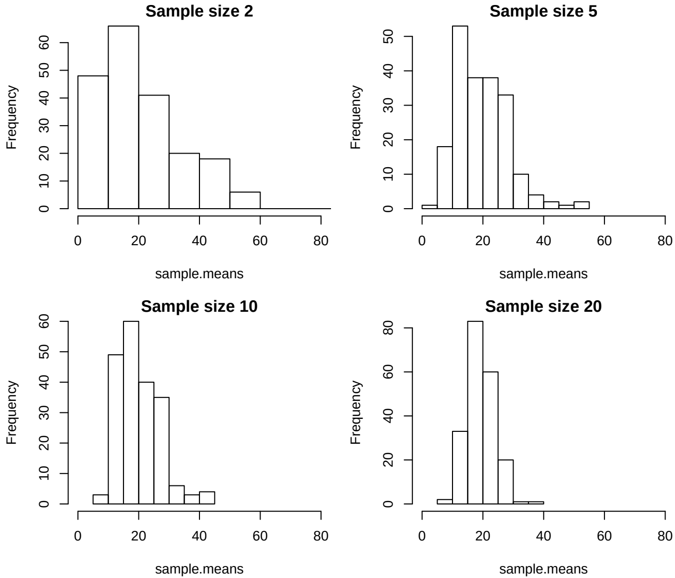
Let's try n=250,000 (next slide). Now we get a reasonably good approximation. A "continuity correction" improves the approximation by shifting the normal curve to better match the binomial distribution (see later).

Binomial(250000,0.0001) with approximating Normal distribution



The CLT for Sampling from an Exponential Distribution

Let's try the exponential distribution, Exp(1/20), for a range of sample sizes.



The CLT for an Exponential Distribution

For $n = 2$ the sampling distribution of the sample mean is very skewed and in fact does not look too much different from the population distribution itself Exp(1/20). (see slide 17)

For $n = 5$ the sampling distribution is still quite skewed but for $n = 10$ the skewness is mostly gone and for $n = 20$ the distribution is looking quite a bit like a normal distribution.

In practice, it can be difficult to know how large a sample size is needed in order to obtain valid inference while using the normal approximation. The answer is context dependent. We will come back to this question when discussing specific types of methods, e.g., confidence intervals, t-tests, linear regression, etc.

Using the CLT for sampling from a Poisson distribution

Consider the example of the manufacturer wanting to estimate the total number of defects in a large shipment of items which are produced in 500 batches of 100 items each. Suppose that past experience has shown that the number of defectives in a batch follows a *Poisson* distribution with mean 0.4. The probability mass function for the Poisson(λ) distribution with mean λ is given by $P(X = x) = \lambda^x e^{-\lambda}/x!$.

Probability mass function for the Poisson(0.4) distribution

```
n=0:5
data.frame(n,Prob=round(dpois(n,lambda=0.4),4))
```

```
  n   Prob
1 0 0.6703
2 1 0.2681
3 2 0.0536
4 3 0.0072
5 4 0.0007
6 5 0.0001
```

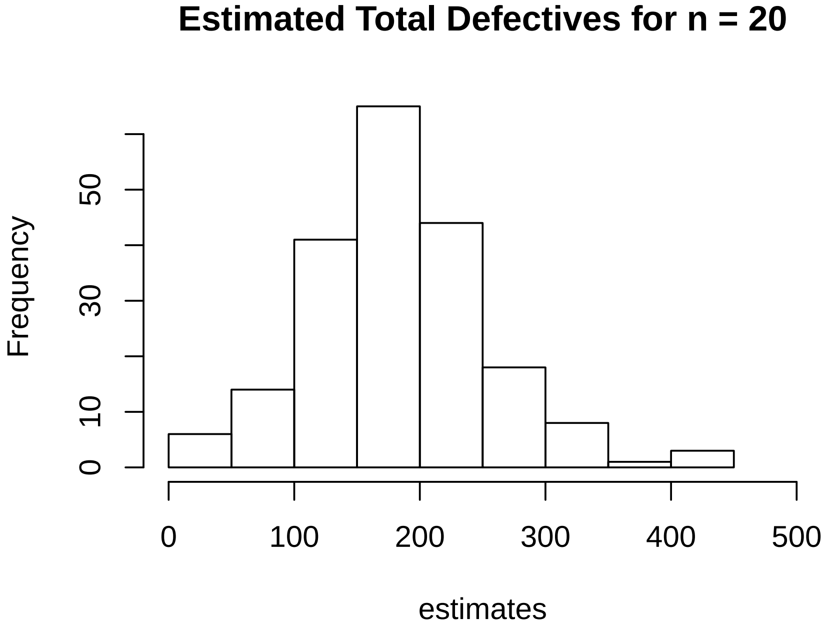
The sampling distribution of the mean number of defectives

The Poisson(λ) distribution has mean λ and variance also equal to λ. Therefore, a sample mean of n batches will have expected value λ and SE $= \sqrt{\lambda/n}$. If $\lambda = 0.4$ then a sample size of $n = 20$ will yield an SE of $\sqrt{0.4/20} = 0.141$.

The estimate of the total number of defectives in the 500 batches is just 500 times the sample mean, i.e., $500 \times \bar{x}$. Therefore, its SE is just 500 times the SE of the sample mean, i.e., its SE is $500 \times \mathrm{SE}(\bar{x}) = 500 \times \sqrt{0.4/20} = 70.7$.

In order to decide if this is too much uncertainty we can simulate the process to see the range of distribution of estimates that we might get. If this has too much variability we would choose a larger sample size.

The Sampling Distribution of the Estimate of Total Defectives



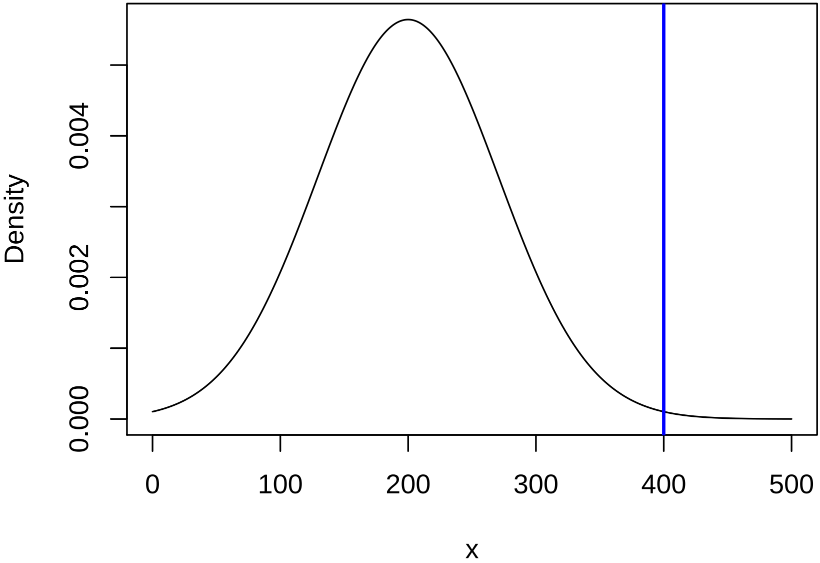
Using the CLT to refine our inferences

The CLT says that the distribution of the sample mean (and hence the estimate of the total) has an approximate Normal distribution if n is sufficiently large. The histogram above suggests that $n = 20$ is large enough (in this situation) to provide a reasonable approximation. The approximating normal distribution has a mean of 200 and standard deviation of 70.7.

With the normal approximation we can now make probability statements about what to expect from our sampling. Suppose that the shipment would be halted if the number of defectives is estimated to be larger than 400. Then we would want to know the chance of getting an estimate this large, if the true number of defectives is actually 200 (500×0.4).

Under the assumption that the true number of defectives is 200 our estimate of the total number of defectives will be approximately normally distributed with mean 200 and SD 70.7. The probability of getting an estimate greater than 400 is the tail probability to the right of 400 in this Normal distribution (see figure on next slide).

The tail probability in the Normal distribution for the number of defectives



Calculating the tail probability

To calculate the tail probability, first calculate the z-score associated with 400:

$$z = (400 - 200)/70.7 = 2.83.$$

This means that 400 is 2.83 SDs above the mean. This is equivalent to the value 2.83 in the *standard normal distribution* (which has mean 0 and SD 1). Then we look up the tail probability associated with 2.83 in the standard normal distribution to give the answer 0.0023274. This is very small, which means that false alarms (halting shipments) will be rare.

Using an Estimate of the Population Variance

For sampling from a Bernoulli distribution we used the variance for a Bernoulli distribution $p(1-p)$ which is the theoretical value for the Bernoulli distribution. A binary variable must have a Bernoulli distribution so there is no risk in making an invalid assumption.

However, for the estimation of the number of defectives we assumed that the number of defectives in a batch has a Poisson distribution which means that the variance is equal to the mean. If this assumption is incorrect then our calculated SE value will be invalid. A better approach in this situation is to use an estimate of the population variance from the sample data. Specifically, we would use the sample variance. In this case our *estimated* SE for the sample mean becomes

$$\text{Estimated SE}(\bar{x}) = \sqrt{\text{Sample Variance}/n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}$$

Using the Sample Variance to Estimate the SE of the Sample Mean

Suppose that we collected the following data on the number of defectives:

```
x=c(0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,2,2,3)
mean(x)
```

```
[1] 0.6
```

```
var(x)
```

```
[1] 0.7789474
```

Then the mean and variance are 0.6 and 0.779, respectively. Note that the variance is somewhat larger than the mean, suggesting that the Poisson distribution assumption might not be correct. The estimated SE is calculated as $\mathrm{SE}(\bar{x}) = \sqrt{0.779/20} = 0.1973575$. The estimated SE for the total number of defectives is 98.6787718.

Therefore, the estimated SE is larger than that obtained under the Poisson distribution assumption. *In general, it is not wise to make unnecessary distributional assumptions.*

Confidence Intervals

The Confidence Interval for a Population Mean

So far, we have seen how to calculate an estimate from a sample as well as how to assess the uncertainty in the estimate (its SE). We have also seen how to use the CLT to make simple probability statements (e.g., what is the chance that our estimate of total defectives is greater than 400?). In practice, we would like to make more precise inferential statements about target parameters of interest.

For example, we would like to be able to give a range of plausible values for the number of defectives in the shipment based on our observed sample. Similarly, in political polling we would like to give a plausible range for the proportion of voters intending to vote Democratic. This idea of a "plausible range" is made mathematically precise by the concept of a *confidence interval*.

A confidence interval is an interval calculated from the data using a rule which ensures that the interval has a certain pre-specified probability (often 95%) of *containing the true value of the target parameter*.

Confidence interval for a proportion

Suppose that we have voting intentions (either Democrat or Republican) from a random sample of 2000 registered voters. Assume that 1100 said they were voting Democratic. Then the sample proportion of Democratic voters is $\hat{p} = 1100/2000 = 0.55$. Note that the formula for the SE of a sample proportion is $\text{SE}(\hat{p}) = \sqrt{p(1-p)/n}$, which depends on the true proportion in the population which we don't know (that is why we are collecting the data!). How can we calculate an SE?

Up to now we have been starting with assumptions about the population distribution and simulation the sample mean or sample proportion to see how it behaves. For example, we explored the sampling distribution of the sample proportion when the true population proportion was $p = 0.5$ (slide 15) or $p = 0.0001$ (slide 28).

In practice, when we don't know the true p we substitute our sample estimate. Therefore, we estimate the SE of the sample proportion using the formula

$$\text{Estimated SE}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$$

Using the estimated SE to make probability statements

In our polling example, the estimated SE is $\sqrt{0.55 \times 0.45/2000} = 0.011$. The result might be reported as an estimate of 0.55 with a margin of error of 0.011. (More commonly percentages are used, ie, 55% with a margin of error of 1.1%.) This statement is meant to convey the fact that the true proportion is not too far from the estimated value of 0.55.

However, caution must be used in interpretation. For example, the results are sometimes written as 0.55 ± 0.011. This notation is fine but it can be misintepreted to mean that the estimation error is *at most 0.011*, so that the true proportion must be between 0.539 and 0.561. This is wrong! The SE is the **typical** error in the estimate, not the **maximum** error.

The CLT tells us that, whatever the true proportion, the sample estimate will be distributed around that true value approximately like a normal distribution with a standard deviation equal to the SE of 0.011. Therefore, it would not be implausible that the true proportion is equal to 0.53 because in that case an estimate of 0.55 would not be that unusual (less than two standard deviations from the mean). The confidence interval is a way of making precise statements about plausible values for the true proportion.

Rationale for the Confidence Interval for a Proportion

The CLT tells us that the sample proportion will be distributed like a normal distribution around the true proportion with a standard deviation equal to the SE. By the properties of the normal distribution, this means that

> *the sample proportion will be at most 1.96 SEs away from the true proportion with probability 0.95*

If we interchange "sample proportion" and "true proportion", the statement remains true:

> *the true proportion will be at most 1.96 SEs away from the sample proportion with probability 0.95*

In other words, the interval of values centered around the sample proportion, at most 1.96 times the SE on either side, has a 95% chance of *containing the true proportion*. This is the definition of a confidence interval.

Formula for the Confidence Interval for a Proportion

Assume a random sample of size n with an estimated proportion of $\hat{p}$.

A 95% confidence interval for the true population proportion p is the following interval:

$$(\hat{p} - 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}).$$

Example: If $n = 2000$ and $\hat{p} = 0.55$, then SE $= \sqrt{0.55 \times 0.45/2000} = 0.011$ and the 95% confidence interval is $(0.55 - 1.96 \times 0.011, 0.55 + 1.96 \times 0.011)$ or (0.52844, 0.57156), which would be rounded to (0.53, 0.57).

Illustration of the confidence interval for a proportion

We can illustrate how a confidence interval for a population proportion works by using simulation. For this, we go back to specifying values for the population parameters and then seeing how the confidence interval performs on simulated data from that population.

Consider a population of trees in a large forest with the target parameter of interest being the proportion of trees infected with a fungus. Suppose that the true proportion of infected trees is $p = 0.15$. Let's use a random sample of $n = 150$ trees to calculate a confidence interval for p. We will repeat the sampling process 100 times to see what happens.

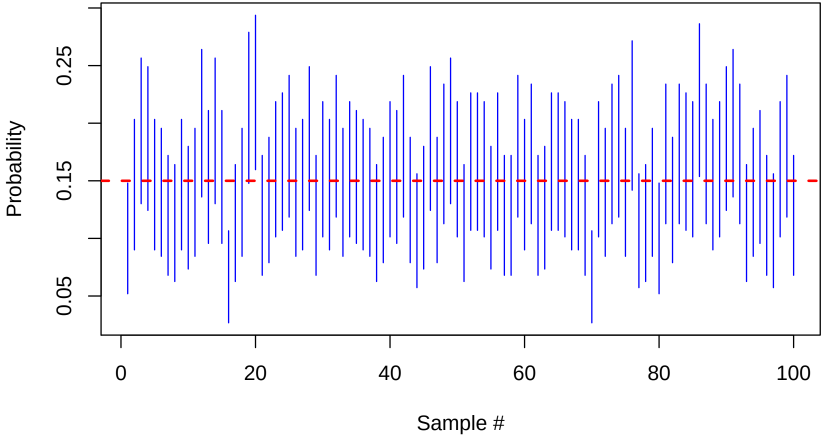
Simulation of the confidence interval for a proportion with $n = 150$

```
set.seed(3)
n=150
p=0.15
phat=rep(NA,100)
for(i in 1:100)phat[i]=mean(rbinom(n,size=1,prob=p))
se = sqrt(phat*(1-phat)/n)
lower = phat - 1.96*se
upper = phat + 1.96*se
results=data.frame(phat,se,lower,upper)
summary(results)
```

```
      phat                se               lower              upper       
 Min.   :0.06667   Min.   :0.02037   Min.   :0.02675   Min.   :0.1066  
 1st Qu.:0.12667   1st Qu.:0.02716   1st Qu.:0.07344   1st Qu.:0.1799  
 Median :0.14667   Median :0.02889   Median :0.09005   Median :0.2033  
 Mean   :0.15033   Mean   :0.02897   Mean   :0.09356   Mean   :0.2071  
 3rd Qu.:0.17333   3rd Qu.:0.03091   3rd Qu.:0.11276   3rd Qu.:0.2339  
 Max.   :0.22667   Max.   :0.03418   Max.   :0.15966   Max.   :0.2937  
```

Illustration of 100 confidence intervals with $n = 150$

```r
matplot(rbind(1:100,1:100),t(results[,c("lower","upper")]),type="l",lty
abline(h=p,lty=2,col=2,lwd=2)
```



The Coverage Probability of a Confidence Interval

Inspection of the plot of the 100 confidence intervals indicates that most (about 95) of them cover the true proportion of 0.15. Let's check.

```r
sum((results$lower <= p)&(results$upper >= p))
```

```
[1] 94
```

We would say that the observed *coverage probability* is 94%. Note that this is very close to what the theory predicts! We cannot always expect results to match the theory so well. A simulation study will tell us if our statistical method is working well enough to use.

Let's see what happens when we reduce the number of trees sampled to 50.

Simulation of a confidence interval for a proportion with $n = 50$

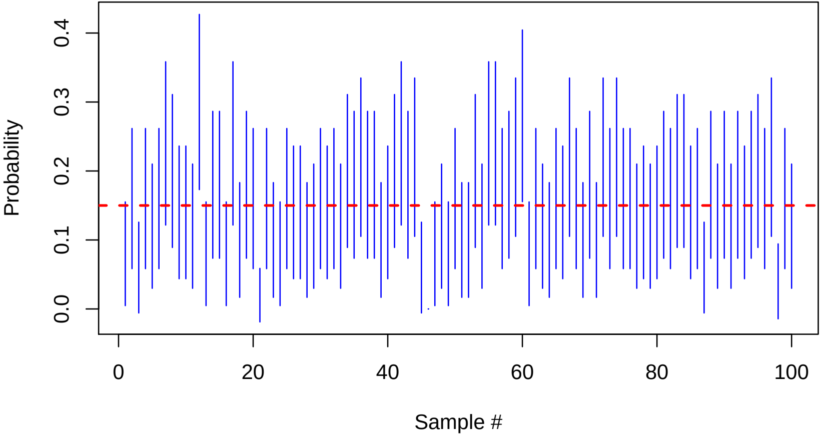
```
      phat              se              lower               upper       
 Min.   :0.0000   Min.   :0.00000   Min.   :-0.01881   Min.   :0.0000  
 1st Qu.:0.1200   1st Qu.:0.04596   1st Qu.: 0.02993   1st Qu.:0.2101  
 Median :0.1600   Median :0.05185   Median : 0.05838   Median :0.2616  
 Mean   :0.1512   Mean   :0.04924   Mean   : 0.05469   Mean   :0.2477  
 3rd Qu.:0.1800   3rd Qu.:0.05433   3rd Qu.: 0.07351   3rd Qu.:0.2865  
 Max.   :0.3000   Max.   :0.06481   Max.   : 0.17298   Max.   :0.4270  
```

The coverage probability is not as good in this case because of the smaller sample size.

```
sum((results$lower <= p)&(results$upper >= p))
```

```
[1] 92
```

Confidence intervals for a proportion with $n = 50$



In addition to a coverage probabilty less than 95% we see that the confidence intervals are much wider due to the smaller n.

The Confidence Interval for a Mean

The concept of a confidence interval for a proportion carries over to forming a confidence interval for a mean in almost the same way. In general a 95% confidence interval for a population mean has the form

$$(\bar{x} - 1.96 \times \text{Estimated SE},\ \bar{x} + 1.96 \times \text{Estimated SE}).$$

The only difference is in how we calculate the estimated SE. In place of $\hat{p}(1-\hat{p})$ we need to use an estimate of the population variance. There are two main approaches to estimating the variance.

1. Assuming a specific distributional form for the population distribution.
2. Not assuming a specific distributional form for the population distribution.

The latter is the preferred approach in most applications. However, it is helpful to see how the first approach works. We will illustrate it using a Poisson distribution.

A confidence interval for the mean of a Poisson distribution

For example, consider the problem of estimating the number of defectives. If we are confident in the assumption that the number of defectives in a batch follows a Poisson distribution then we might want to take advantage of the fact that for a Poisson distribution the variance and mean are equal. This means that we can use the sample mean as an estimate of the variance as well as an estimate of the mean.

For the previous example, with a sample mean number of defectives equal to $\bar{x} = 0.4$ from a sample of $n = 20$ batches, we would calculate the SE as follows:

$$\text{Estimated SE}(\bar{x}) = \sqrt{\bar{x}/n}$$

The result is Estimated SE $= \sqrt{0.4/20} = 0.1414$. Thus, the 95% confidence interval for the mean number of defectives per batch is

$$(0.4 - 1.96 \times 0.1414,\ 0.4 + 1.96 \times 0.1414)$$

or (0.12, 0.68).

A confidence interval for the total number of defectives

If we instead want to report a confidence interval for the estimated total number of defectives in the shipment of 500 batches we simply multiply both ends of the confidence interval by 500, giving (60, 340). The decision of whether or not to halt the shipment would be based on whether or not 340 is an acceptable number of defective items.

Confidence Intervals for a Mean Without Distributional Assumptions

It is common statistical practice to not use distributional assumptions when they are unnecessary. For the sample mean we have already seen how to use the sample variance to calculate the SE (slide 42):

$$\text{Estimated SE}(\bar{x}) = \sqrt{\text{Sample Variance}/n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}$$

This formula makes no assumption of a Poisson (or any other) distribution for the number of defectives in a batch. From the data used earlier, we have sample variance $= 0.779$ and the estimated SE $= \sqrt{.779/20} = 0.197$. Therefore, the 95% confidence interval for the true mean number of defectives is $(0.4 - 1.96 \times 0.197,\ 0.4 + 1.96 \times 0.197)$ or (0.014,0.786). The confidence interval for the total number of defectives is (7, 393), which is considerably wider than the confidence interval based on the Poisson assumption.

In general, it is not advisable to make unnecessary distributional assumptions.

Simulation of the confidence interval for a mean of a Poisson distribution

```r
set.seed(4)
n=50
lambda=1
xbar=rep(NA,100)
v=rep(NA,100)
for(i in 1:100){
  x=rpois(n,lambda)
  xbar[i]=mean(x)
  v[i]=var(x)
}
se = sqrt(v/n)
lower = xbar - 1.96*se
upper = xbar + 1.96*se
results=data.frame(xbar,se,lower,upper)
summary(results)
```

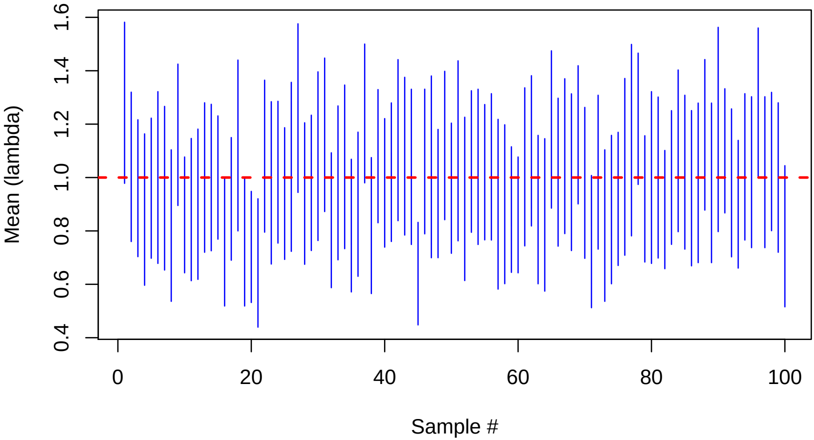
```
      xbar             se              lower            upper      
 Min.   :0.640   Min.   :0.09798   Min.   :0.4395   Min.   :0.832  
 1st Qu.:0.900   1st Qu.:0.12923   1st Qu.:0.6513   1st Qu.:1.170  
 Median :1.000   Median :0.14228   Median :0.7200   Median :1.280  
 Mean   :0.991   Mean   :0.14079   Mean   :0.7151   Mean   :1.267  
 3rd Qu.:1.060   3rd Qu.:0.15045   3rd Qu.:0.7722   3rd Qu.:1.349  
 Max.   :1.280   Max.   :0.19523   Max.   :0.9998   Max.   :1.582  
```

```r
sum((results$lower <= lambda)&(results$upper >= lambda))
```

```
[1] 97
```

Illustration of 100 confidence intervals of a mean

```r
matplot(rbind(1:100,1:100),t(results[,c("lower","upper")]),type="l",lty
abline(h=lambda,lty=2,col=2,lwd=2)
```



Summary - Principles of Statistical Inference

simple random sample

sampling with replacement

sampling without replacement

sampling from a probability distribution

sampling distribution of a sample mean

sampling distribution of a sample proportion

standard error of a sample mean

standard error of a sample proportion

central limit theorem

confidence interval

coverage probability

the role of distributional assumptions