Chapter 7 Distributions of Sampling Statistics

Bibliography: Sheldon Ross(2014)

7.1 **DISTRIBUTIONS ARISING FROM THE NORMAL**

7.1.1 *The Chi-Square Distribution and Relation with Gamma Random Variable*

We know that if $X \sim N\left(\mu,\sigma^2\right)$, then the random variable $Y=\left(X-\mu\right)^2/\sigma^2$ is distributed as the gamma distribution $Y \sim \gamma\left(\frac{1}{2},\frac{1}{2}\right)$. Let $X_i \sim N\left(\mu_i,\sigma_i^2\right)$, $i=1,2,\ldots,n$, independent variables and define the new variable:

$$\chi_n^2=\sum_{i=1}^{n}\frac{\left(X_i-\mu_i\right)^2}{\sigma_i^2} \qquad (7.1)$$

χ_n^2 must be distributed as the gamma variate $\chi_n^2 \sim \gamma\left(\frac{1}{2},\frac{1}{2}n\right)$, which has the PDF

$$f\left(x\right)=\frac{\frac{1}{2}e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}, \quad x>0 \qquad (7.2)$$

$$E\left[\chi_n^2\right]=n \ , \quad V\left[\chi_n^2\right]=2n \qquad (7.3)$$

Another **definition**

If $Z_1,Z_2,\ldots,Z_n$ are independent *standard normal random variables*, then the random variable χ_n^2 defined by:

$$\chi_n^2=Z_1^2+Z_2^2+\ldots+Z_n^2 \qquad (7.4)$$

is said to have a *chi-square distribution* with *n* degrees of freedom.

The chi-square distribution has the *additive property* that if X_1 and X_2 are independent chi-square random variables with n_1 and n_2 degrees of freedom, respectively, then X_1+X_2 is chi-square with n_1+n_2 degrees of freedom. This can be shown by noting that $X_1 + X_2$ is the sum of squares of $n_1 + n_2$ independent standard normal and thus has a chi-square distribution with $n_1 + n_2$ degrees of freedom.

If X is a chi-square random variable with n degrees of freedom, then for any $\alpha \in (0,1)$, the quantity $\chi^2_{\alpha,n}$ is defined to be such that

$$P\left\{X \geq \chi^2_{\alpha,n}\right\} = \alpha \qquad (7.5)$$

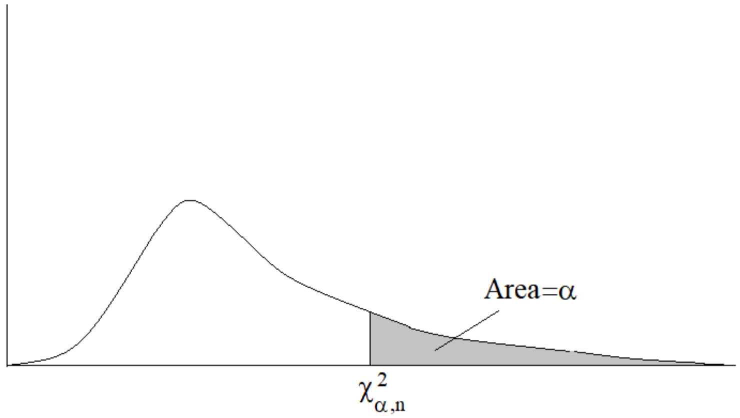


Figure 7.1

Table with values of $\chi^2_{\alpha,n}$ are available at the end of the statistical books. In addition, Program R can be used to obtain chi-square probabilities and the values of $\chi^2_{\alpha,n}$.

Exercise 1: Determine $P\left\{\chi^2_{26} \leq 30\right\}$ when χ^2_{26} is a chi-square random variable with 26 degrees of freedom.

Using Program R gives the result $P\left\{\chi^2_{26} \leq 30\right\} = 0.7325$

```
> pchisq(30, 26, lower.tail = TRUE)

[1] 0.732389
```

lower.tail logical; if TRUE (default), probabilities are $\Pr(X \leq x)$, otherwise, $\Pr(X > x)$.

```
> pchisq(30, 26)    [1] 0.732389

>qchisq(0.7325, df=26)   [1] 30.00268
```

Exercise 2: Find $\chi^2_{0.05,15}$. So, $\alpha = 0.05$ and using Program R:

```
qchisq(0.95, df=15)
[1] 24.99579
```

Exercise 3: Suppose that we are attempting to locate a target in three-dimensional space, and that the three coordinate errors (in meters) of the point chosen are independent normal random variables with mean 0 and standard deviation 2. Find the probability that the distance between point chosen and the target exceeds 3 meters.

If D is the distance, then

$$D^2 = X_1^2 + X_2^2 + X_3^2$$

where X_i is the error in the -*i*th coordinate. Since $Z_i = (X_i - 0)/2$, $i = 1,2,3$ are all standard normal random variables, it follows that

$$P\{D^2 > 9\} = P\{X_1^2 + X_2^2 + X_3^2 > 9\} = P\{4Z_1^2 + 4Z_2^2 + 4Z_3^2 > 9\}$$
$$= P\left\{Z_1^2 + Z_2^2 + Z_3^2 > \frac{9}{4}\right\} = P\left\{\chi_3^2 > \frac{9}{4}\right\} = 0.5222$$

where the final equality was obtained from R.

```
> pchisq(9/4, 3, lower.tail = FALSE)
[1] 0.5221672
```

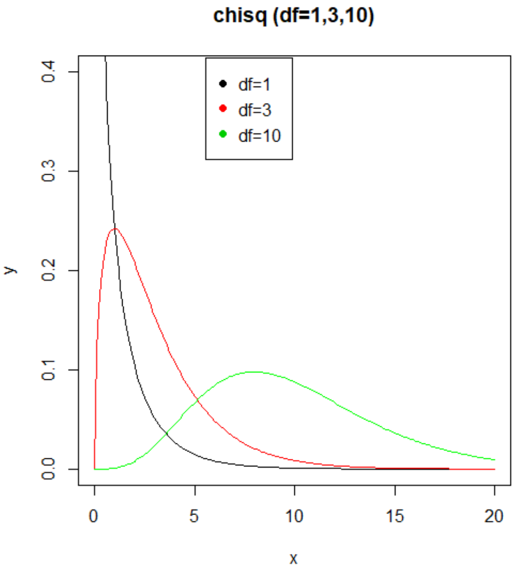


Figure 7.2 The chi-square density functions having 1, 3, and 10 degrees of freedom, respectively.

```
> x <- seq(0,20,by = 0.1)
> y <- dchisq(x,df=1)
> z <- dchisq(x,df=3)
> t <- dchisq(x,df=10)
> plot(x,y,main="chisq (df=1,3,10)",ylim=c(0,0.4),type='l')
> lines(x,z,type='l')
> lines(x,t,type='l')
```

Exercise 4: When we attempt to locate a target in two-dimensional space, suppose that the coordinate errors are independent normal random variables with mean 0 and standard deviation 2. Find the probability that the distance between the point chosen and the target exceeds 3.

If D is the distance and X_i, $i=1,2$, are the coordinate errors, then

$$D^2 = X_1^2 + X_2^2$$

Since $Z_i = (X_i - 0)/2$, $i=1,2$, are standard normal random variables, we obtain

$$P\{D^2 > 9\} = P\{X_1^2 + X_2^2 > 9\} = P\{4Z_1^2 + 4Z_2^2 > 9\}$$
$$= P\left\{Z_1^2 + Z_2^2 > \frac{9}{4}\right\} = P\left\{\chi_2^2 > \frac{9}{4}\right\} \approx 0.3247$$

```
pchisq(9/4, 2, lower.tail = FALSE)
[1] 0.3246525
```

7.1.2 *The t -Distribution*

If Z and χ_n^2 are independent random variables, with Z having a standard normal distribution and χ_n^2 having a chi-square distribution with n degrees of freedom, then the random variable T_n defined by

$$T_n = \frac{Z}{\sqrt{\chi_n^2 / n}} \qquad (7.6)$$

is said to have a *t-distribution with n degrees of freedom*. A graph of the probability density function of T_n is given in Figure 7.3 for n = 1, 3, and 10.

```
x <- seq(-5,5,by = 0.1)
> y <- dt(x,df=1)
> z <- dt(x,df=3)
> t <- dt(x,df=10)
> plot(x,y,main="t-distribution (df=1,3,10)",ylim=c(0,0.4),type='l')
> lines(x,z,type='l')
> lines(x,t,type='l')
```

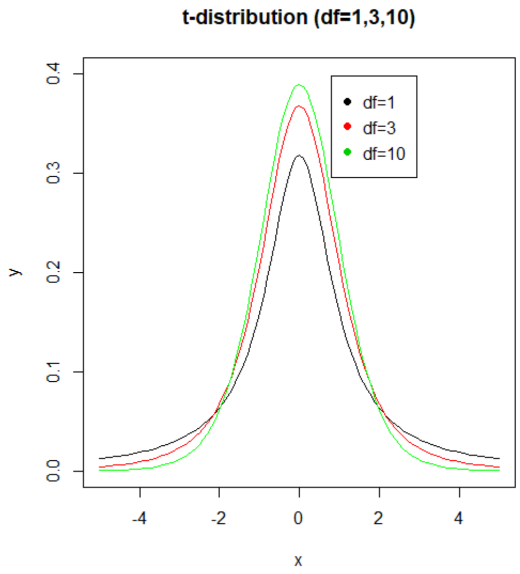


Figure 7.3

Like the standard normal density, the *t*-density is symmetric about zero. In addition, as *n* becomes larger, it becomes more and more like a standard normal density.

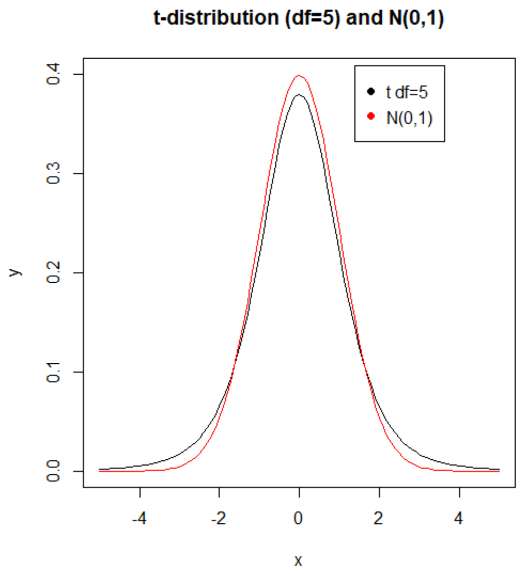


Figura 7.4

```
x <- seq(-5,5,by = 0.1)
> y <- dt(x,df=5)
> z <- dnorm(x,mean=0,sd=1)
> plot(x,y,main="t-distribution (df=5) and normal standard",ylim=c(0,0.4),typ
e='l')
> lines(x,z,type='l')
```

Figure 7.4 shows a graph of the *t*-density function with 5 degrees of freedom compared with the standard normal density. Notice that the *t*-density has thicker "tails," indicating *greater variability*, than does the normal density.

The mean and variance of T_n can be shown to equal

$$E\left[T_n\right]=0\ ,\ n>1 \tag{7.7}$$

$$Var\left(T_n\right)=\frac{n}{n-2}\ ,\quad n>2 \tag{7.8}$$

Thus the variance of T_n decreases to 1 — the variance of a standard normal random variable — as *n* increases to ∞.

For $\alpha\in\left(0,1\right)$, let $t_{\alpha,n}$ be such that

$$\Pr\left(T_n\geq t_{\alpha,n}\right)=\alpha \tag{7.9}$$

It follows from the symmetry about zero of the *t*-density function that $-T_n$ has the same distribution as T_n, and so

$$\alpha=\Pr\left(-T_n\geq t_{\alpha,n}\right)=\Pr\left(T_n\leq -t_{\alpha,n}\right)=1-\Pr\left(T_n>-t_{\alpha,n}\right)$$

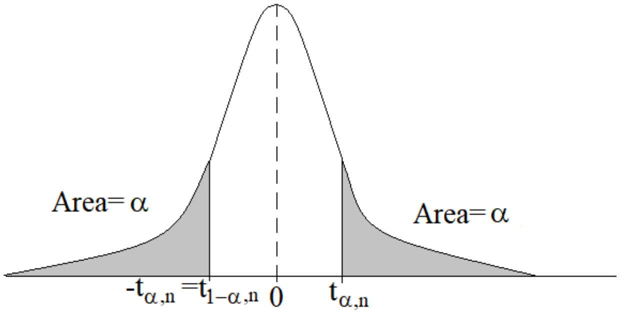


Figure 7.5

Therefore,

$$\Pr\left(T_n\geq -t_{\alpha,n}\right)=1-\alpha$$

Dar, $\Pr\left(T_n\geq t_{1-\alpha,n}\right)=1-\alpha$ leading to the conclusion that

$$-t_{\alpha,n}=t_{1-\alpha,n} \tag{7.10}$$

The values of $t_{\alpha,n}$ for a variety of values of *n* and *α* are tabulated. In addition, Programs *R* compute the *t*-distribution function and the values $t_{\alpha,n}$.

Exercise 5: Find (a) $\Pr(T_{12} \leq 1.4)$ and (b) $t_{0.025,9}$.

Run Programs *R* to obtain the results: (a)0 .9066 (b) 2.2625

```
> pt(1.4,df=12)
[1] 0.9065835

> qt(1-0.025,df=9)
[1] 2.262157
```

7.1.3 *The F-Distribution*

If χ_n^2 and χ_m^2 are independent chi-square random variables with *n* and *m* degrees of freedom, respectively, then the random variable $F_{n,m}$ defined by

$$F_{n,m} = \frac{\chi_n^2 / n}{\chi_m^2 / m} \tag{7.11}$$

is said to have an *F-distribution with n and m degrees of freedom*.

For any $\alpha \in (0,1)$, let $F_{\alpha,n,m}$ be such that

$$\Pr\left(F_{n,m} > F_{\alpha,n,m}\right) = \alpha \tag{7.12}$$

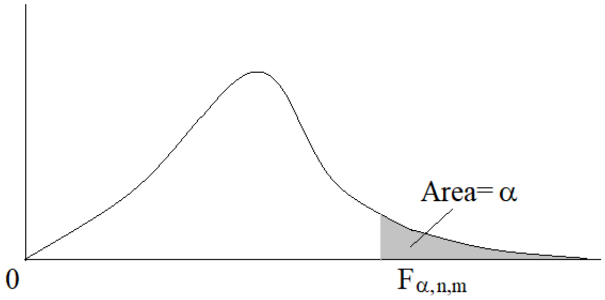


Figure 7.6

The quantities $F_{\alpha,n,m}$ are tabulated for different values of *n, m,* and $\alpha \leq 1/2$. If $F_{\alpha,n,m}$ is desired when $\alpha > 1/2$, it can be obtained by using the following equalities:

$$\alpha = \Pr\left(\frac{\chi_n^2 / n}{\chi_m^2 / m} > F_{\alpha,n,m}\right) = \Pr\left(\frac{\chi_m^2 / m}{\chi_n^2 / n} < \frac{1}{F_{\alpha,n,m}}\right)$$

$$=1-\Pr\left(\frac{\chi_m^2/m}{\chi_n^2/n}\geq\frac{1}{F_{\alpha,n,m}}\right)$$

or, equivalently,

$$\Pr\left(\frac{\chi_m^2/m}{\chi_n^2/n}\geq\frac{1}{F_{\alpha,n,m}}\right)=1-\alpha$$

But because $\left(\chi_m^2/m\right)/\left(\chi_n^2/n\right)$ has an F-distribution with degrees of freedom *m* and *n* , it follows that

$$\Pr\left(\frac{\chi_m^2/m}{\chi_n^2/n}\geq F_{1-\alpha,m,n}\right)=1-\alpha$$

implying, from Equation (3), that

$$\frac{1}{F_{\alpha,n,m}}=F_{1-\alpha,m,n} \qquad (7.13)$$

For instance, $F_{0.9,5,7}=\frac{1}{F_{0.1,7,5}}=\frac{1}{3.37}=0.296$

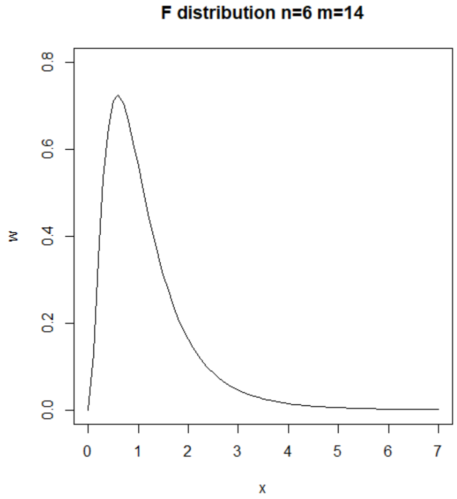
```
qf(1-0.1,7,5)      [1] 3.367899
> qf(1-0.9,5,7)    [1] 0.296921

> pf(3.37,7,5) [1] 0.9001106
```

Exercise 6: Determine $\Pr\left(F_{6,14}\leq 1.5\right)$.

Run R to obtain the solution 0.7518.

```
> pf(1.5,6,14)  [1] 0.7515004
```



7.2 SAMPLING STATISTICS

INTRODUCTION

The science of statistics deals with drawing conclusions from observed data. For instance, a typical situation arises when one is confronted with a large collection, or population, of items that have measurable values associated with them. By suitably sampling from this collection, and then analyzing the sampled items, one hopes to be able to draw some conclusions about the collection as a whole.

To use sample data to make inferences about an entire population, it is necessary to make some assumptions about the relationship between the two. One such assumption, is that *there is an underlying (population) probability distribution* such that *the measurable values* of the items in the population can be thought of as being independent random variables having this distribution. If the sample data are then chosen in a random fashion, then it is reasonable to suppose that they too are independent values from the distribution.

Definition

If $X_1, X_2, \dots, X_n$ are independent random variables having a common distribution *F*, then we say that they constitute a *sample* (sometimes called a random sample) from the distribution *F*.

In most applications, the population distribution *F* will not be completely specified and one will attempt to use the data to make inferences about *F*. Sometimes it will be supposed that *F* is specified up to some unknown parameters (for instance, one might suppose that *F* was a normal distribution function having an unknown mean and variance, or that it is a Poisson distribution function whose mean is not given), and at other times it might be assumed that almost nothing is known about *F* (except maybe for assuming that it is a continuous, or a discrete, distribution). Problems in which the form of the underlying distribution is specified up to a set of unknown parameters are called *parametric* inference problems, whereas those in which nothing is assumed about the form of *F* are called *nonparametric* inference problems.

EXAMPLE Suppose that a new process has just been installed to produce computer chips, and suppose that the successive chips produced by this new process will have useful lifetimes that are independent with a common unknown distribution *F*. Physical reasons sometimes suggest the parametric form of the distribution *F*; for instance, it may lead us to believe that *F* is a normal distribution, or that *F* is an exponential distribution. In such cases, we are confronted with a *parametrical* statistical problem in which we would want to use the observed data to estimate the parameters of *F*. For instance, if *F* were assumed to be a normal distribution, then we would want to estimate its mean and variance; if *F* were assumed to be exponential, we would want to estimate its mean. In other situations, there might not be any physical justification for supposing that *F* has any particular form; in this case the problem of making inferences about *F* would constitute a *nonparametric* inference problem.

7.2.1 **THE SAMPLE MEAN**

Consider a population of elements, each of which has a numerical value attached to it. For instance, the population might consist of the adults of a specified community and the value attached to each adult might be his or her annual income, or height, or age, and so on. We often suppose that the value associated with any member of the population can be regarded as being the value of a random variable having expectation μ and variance σ^2. The quantities μ and σ^2 are called the population mean and the population variance, respectively. Let $X_1, X_2, \ldots, X_n$ be a sample of values from this population. The sample mean is defined by

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \tag{7.14}$$

Since the value of the sample mean $\bar{X}$ is determined by the values of the random variables in the sample, it follows that $\bar{X}$ is also a random variable. Its expected value and variance are obtained as follows:

$$E\left[\bar{X}\right] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]$$

$$= \frac{1}{n}\left(E\left[X_1\right] + E\left[X_2\right] + \ldots + E\left[X_n\right]\right) = \mu \tag{7.15}$$

$$V\left[\bar{X}\right] = V\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]$$

$$= \frac{1}{n^2}\left(V\left[X_1\right] + V\left[X_2\right] + \ldots + V\left[X_n\right]\right) \quad \text{by independence}$$

$$V\left[\bar{X}\right]=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n} \tag{7.16}$$

where μ and σ^2 are the population mean and variance, respectively. Hence, the expected value of the sample mean is the population mean μ whereas its variance is 1/n times the population variance. As a result, we can conclude that $\bar{X}$ is also centered about the population mean μ, but its spread becomes more and more reduced as the sample size increases.

7.2.2 THE CENTRAL LIMIT THEOREM

This theorem asserts that the sum of a large number of independent random variables has a distribution that is approximately normal. In its *simplest form*, the central limit theorem is as follows:

The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for *n* large, the distribution of

$$X_1+X_2+\ldots+X_n \tag{7.17}$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

It follows from the central limit theorem that

$$\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}} \tag{7.18}$$

is approximately a standard normal random variable; thus, for *n* large,

$$\Pr\left(\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}<x\right)\approx\Pr\left(Z<x\right)=\Phi\left(x\right) \tag{7.19}$$

where *Z* is a standard normal random variable.

Exercise 1: An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

Let *X* denote the total yearly claim. Number the policy holders, and let X_i denote the yearly claim of policy holder *i*. With $n = 25000$, we have from the central limit theorem that $X = \sum_{i=1}^{n} X_i$ will have approximately a normal distribution with mean $320 \times 25000 = 8 \times 10^6$ and standard deviation $540\sqrt{25000} = 8.5381 \times 10^4$.

Therefore,

$$P\left\{X > 8.3\times 10^6\right\} = P\left\{\frac{X - 8\times 10^6}{8.5381\times 10^4} > \frac{8.3\times 10^6 - 8\times 10^6}{8.5381\times 10^4}\right\}$$

$$= P\left\{\frac{X - 8\times 10^6}{8.5381\times 10^4} > \frac{0.3\times 10^6}{8.5381\times 10^4}\right\}$$

$$\approx P\left\{Z > 3.51\right\} \approx 0.00023 \quad \text{where } Z \text{ is a standard normal}$$

Thus, there are only 2.3 chances out of 10,000 that the total yearly claim will exceed 8.3 million.

One of the most important applications of the central limit theorem is in regard to binomial random variables. Since such a random variable *X* having parameters (n, p) represents the number of successes in *n* independent trials when each trial is a success with probability *p*, we can express it as

$$X = X_1 + X_2 + \ldots + X_n \qquad (7.20)$$

$$X_i = \begin{cases} 1 & \text{if the i-th trial is a succes} \\ 0 & \text{otherwise} \end{cases} \qquad (7.21)$$

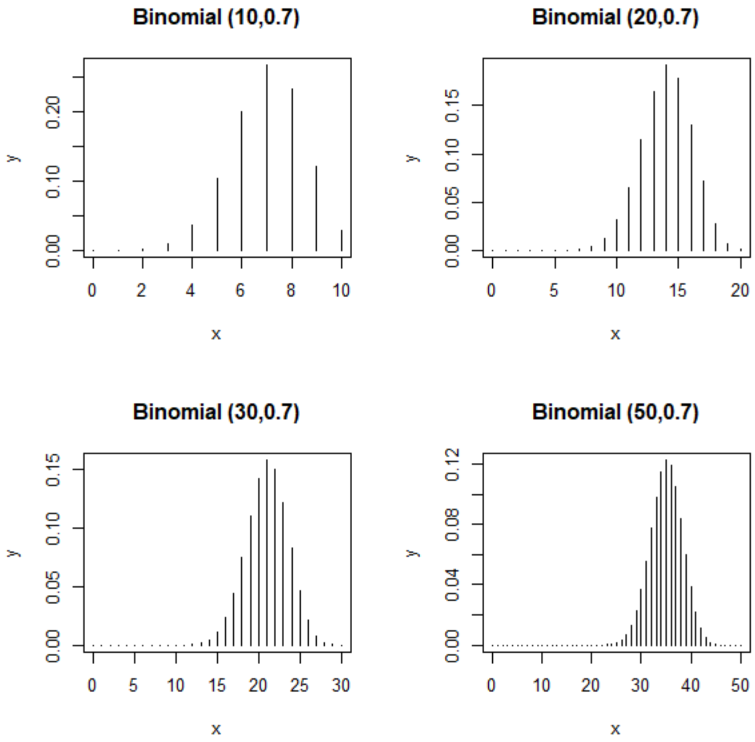
Because

$$E\left[X_i\right] = p, \qquad V\left[X_i\right] = p\left(1-p\right) \qquad (7.22)$$

it follows from the central limit theorem that for *n* large

$$\frac{X - np}{\sqrt{np\left(1-p\right)}} \qquad (7.23)$$

will approximately be a standard normal random variable [see Figure, which graphically illustrates how the probability function of a binomial (n, p) random variable becomes more and more “normal” as n becomes larger and larger].



```
> x <- seq(0,10,by = 1)
> y <- dbinom(x,10,0.7)
> plot(x,y,main="Binomial (10,0.7)",type='h')
> x <- seq(0,20,by = 1)
> y <- dbinom(x,20,0.7)
> plot(x,y,main="Binomial (20,0.7)",type='h')
> x <- seq(0,30,by = 1)
> y <- dbinom(x,30,0.7)
> plot(x,y,main="Binomial (30,0.7)",type='h')
> x <- seq(0,50,by = 1)
> y <- dbinom(x,50,0.7)
```

Exercise 2: The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

Let X denote the number of students that attend; then assuming that each accepted applicant will independently attend, it follows that X is a binomial random variable with parameters $n = 450$ and $p = 0.3$. Since the binomial is a discrete and the normal a continuous distribution, it is best to compute $\Pr(X = i)$ as $\Pr(i - 0.5 < X < i + 0.5)$ when applying the normal approximation (the continuity correction). This yields the approximation

$$P(X > 150.5) = P\left\{\frac{X - 450 \times 0.3}{\sqrt{450 \times 0.3 \times 0.7}} \geq \frac{150.5 - 450 \times 0.3}{\sqrt{450 \times 0.3 \times 0.7}}\right\}$$

$$\approx P(Z > 1.59) = 1 - P(Z < 1.59) = 0.06$$

```
> 1-pnorm(1.59,0,1)
[1] 0.0559174
```

Hence, only 6 percent of the time do more than 150 of the first 450 accepted actually attend.

Approximate Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean

$$\bar{X} = \sum_{i=1}^{n} X_i / n \qquad (7.24)$$

From the central limit theorem $\bar{X}$ will be approximately normal when the sample size *n* is large. Since the sample mean has expected value μ and standard deviation $\sigma / \sqrt{n}$, it then follows that

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \qquad (7.25)$$

has approximately a standard normal distribution.

Exercise 3: An astronomer wants to measure the distance from her observatory to a distant star. However, due to atmospheric disturbances, any measurement will not yield the exact distance *d*. As a result, the astronomer has decided to make a series of measurements and then use their average value as an estimate of the actual distance. If the astronomer believes that the values of the successive measurements are independent random variables with a mean of *d* light years and a standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within ±0.5 light years?

If the astronomer makes *n* measurements, then $\bar{X}$, the sample mean of these measurements, will be approximately a normal random variable with mean *d* and standard deviation $2/\sqrt{n}$. Thus, the probability that it will lie between $d \pm 0.5$ is obtained as follows:

$$\Pr\left(d-0.5<\bar{X}<d+0.5\right)=\Pr\left(\frac{-0.5}{2/\sqrt{n}}<\frac{\bar{X}-d}{2/\sqrt{n}}<\frac{0.5}{2/\sqrt{n}}\right)$$

$$\approx \Pr\left(-\sqrt{n}/4<Z<\sqrt{n}/4\right)=2\Pr\left(Z<\sqrt{n}/4\right)-1$$

where *Z* is a standard normal random variable.

Thus, the astronomer should make *n* measurements, where *n* is such that

$$2\Pr\left(Z<\sqrt{n}/4\right)-1\geq 0.95$$

$$\Pr\left(Z<\sqrt{n}/4\right)\geq 0.975$$

Since $\Pr\left(Z<1.96\right)=\Phi\left(1.96\right)=0.975$, it follows that *n* should be chosen so that

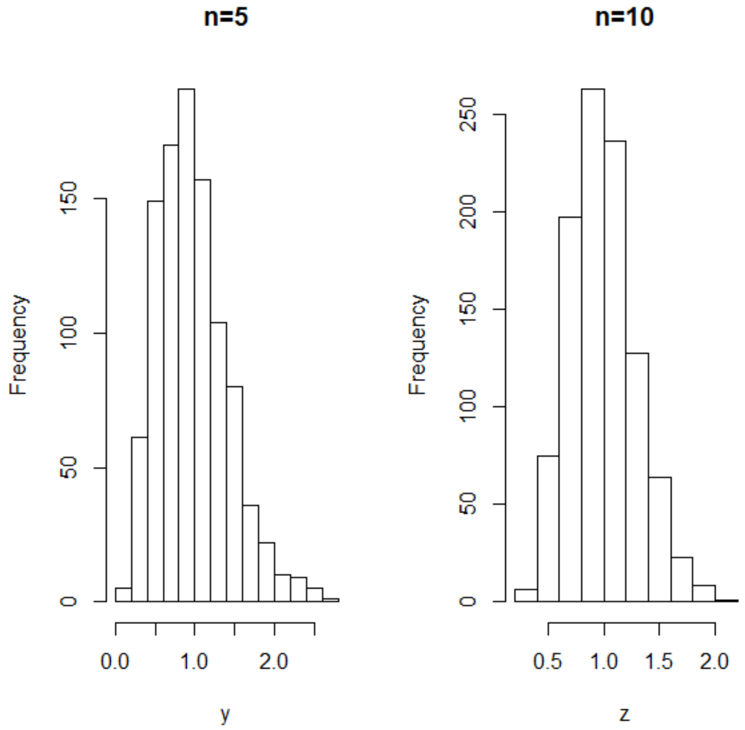
$$\sqrt{n}/4\geq 1.96$$

That is, at least 62 observations are necessary.

How Large a Sample Is Needed?

The central limit theorem leaves open the question of how large the sample size *n* needs to be for the normal approximation to be valid, and indeed the answer depends on the population distribution of the sample data. For instance, if the underlying population distribution is normal, then the sample mean $\bar{X}$ will also be normal regardless of the sample size. A general rule of thumb is that one can be confident of the normal approximation whenever the sample size *n* is at least 30. That is, practically speaking, no matter how nonnormal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal. In most cases, the normal approximation is valid for much smaller sample sizes. Indeed, a sample of size 5 will often suffice for the approximation to be valid.

Figure presents the distribution of the sample means from an exponential population distribution for samples of sizes $n = 5, 10$.



```
> par(mfrow=c(1,2))
> y <- vector("numeric",length=1000)
> for (k in 1:1000 ){
+ x<-rexp(5, rate = 1)
+ y[k]<-mean(x)
+ }
> mean(y)
[1] 0.9704635
> var(y)
[1] 0.1930445
> hist(y,main = 'n=5')

> z <- vector("numeric",length=1000)
> for (k in 1:1000 ){
+   x<-rexp(10, rate = 1)
+   z[k]<-mean(x)
+ }
> mean(z)
[1] 0.98889
> var(z)
[1] 0.0886846
```