LECTURE 03

SAMPLING DISTRIBUTIONS AND CENTRAL LIMIT THEOREM - II

Outline of today's lecture:

I. The Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) The mean of a random sample has a *sampling distribution* whose shape can be approximated by a Normal distribution. The larger the sample, the better the approximation will be.

Theorem 7.4 Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with $E(Y_i)=\mu$ and $V(Y_i)=\sigma^2<\infty$. Define

$$U_n = \left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right) \quad \text{where } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Then the distribution of U_n converges to a standard normal distribution function as $n \rightarrow \infty$.

- That is, probability statements about U_n can be approximated by corresponding probabilities for the standard normal random variable if n is large.
  - Usually a value of n greater than 30 will ensure that the distribution of U_n can be closely approximated by a normal distribution.
- The conclusion of the central limit theorem is often replaced with the simpler statement that $\overline{Y}$ is *asymptotically normally distributed* with mean μ and variance σ^2.

EXAMPLE 7.7 Achievement test scores of all high school seniors in a state have mean 60 and variance 64. A random sample of $n = 100$ students from one large high school had a mean score of 58. Is there evidence to suggest that this high school is inferior? (Calculate the probability that the sample mean is at most 58 when $n = 100$.)

Solution Let $\overline{Y}$ denote the mean of a random sample of $n = 100$ scores from a population with $\mu = 60$ and $\sigma^2 = 64$. We want to approximate $P(\overline{Y} \leq 58)$. We know from Theorem 7.4 that $\sqrt{n}(\overline{Y} - \mu)/\sigma$ is approximately a standard normal random variable, which we denote by Z. Hence, using Table 4, Appendix III, we have

$$P(\overline{Y} \leq 58) \approx P\left(Z \leq \frac{\sqrt{100}(58-60)}{\sqrt{64}}\right) = P(Z \leq -2.5) = .0062.$$

Because this probability is so small, it is unlikely that the sample from the school of interest can be regarded as a random sample from a population with $\mu = 60$ and $\sigma^2 = 64$. The evidence suggests that the average score for this high school is lower than the overall average of $\mu = 60$.

This example illustrates the use of probability in the process of testing hypotheses, a common technique of statistical inference that will be further discussed in Chapter 10.

EXAMPLE 7.8 The service times for customers coming through a checkout counter in a retail store are independent random variables with mean 1.5 minutes and variance 1.0. Approximate the probability that 100 customers can be served in less than 2 hours of total service time.

Solution If we let Y_i denote the service time for the ith customer, then we want

$$P\left(\sum_{i=1}^{100} Y_i \le 120\right) = P\left(\overline{Y} \le \frac{120}{100}\right) = P(\overline{Y} \le 1.20).$$

Because the sample size is large, the central limit theorem tells us that $\overline{Y}$ is approximately normally distributed with mean $\mu_{\overline{Y}} = \mu = 1.5$ and variance $\sigma^2_{\overline{Y}} = \sigma^2/n = 1.0/100$. Therefore, using Table 4, Appendix III, we have

$$P(\overline{Y} \le 1.20) = P\left(\frac{\overline{Y} - 1.50}{1/\sqrt{100}} \le \frac{1.20 - 1.50}{1/\sqrt{100}}\right)$$
$$\approx P[Z \le (1.2 - 1.5)\sqrt{100}] = P(Z \le -3) = .0013.$$

Thus the probability that 100 customers can be served in less than 2 hours is approximately .0013. This small probability indicates that it is virtually impossible to serve 100 customers in only 2 hours.

II. Sampling Distributions Related to Normal Distribution

A. Sampling Distribution of Mean

Theorem 7.1 Let Y_1, Y_2, …, Y_n be independent, normal random variables, each with mean μ and variance σ^2. Then:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is normally distributed with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$.

Example 7.1 and 7.2 A bottling machine can be regulated so that it discharges an average of μ ounces per bottle. It has been observed that the amount of fill dispensed by the machine is normally distributed with σ = 1.0 ounce. A sample of n = 9 filled bottles is randomly selected from the output of the machine on a given day (all bottled with the same machine setting) and the ounces of fill measured for each.

a) Find the probability that the sample mean will be within 0.3 ounce of the true mean μ for that particular setting.

b) How many observations should be included in the sample if we wish $\bar{Y}$ to be within 0.3 ounce of μ with probability 0.95?

Solution

a)

If $Y_1, Y_2, \ldots, Y_9$ denote the ounces of fill to be observed, then we know that the Y_i are normally distributed with mean μ and variance $\sigma^2 = 1$ for i = 1, 2, . . . , 9.

Therefore $\bar{Y}$ possesses a normal sampling distribution with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma^2_{\bar{Y}} = \dfrac{\sigma^2}{n} = \dfrac{1}{9}$.

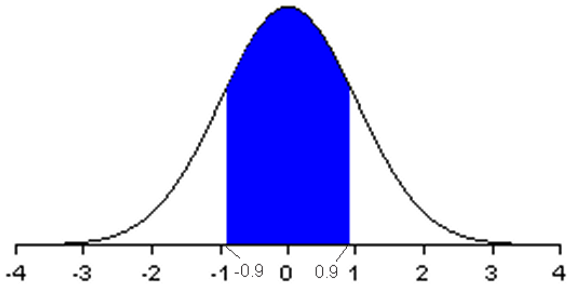
In the problem, we are asked to find:

$$P\left(\left|\bar{Y} - \mu\right| \leq 0.3\right) = P\left(-0.3 \leq \bar{Y} - \mu \leq 0.3\right)$$

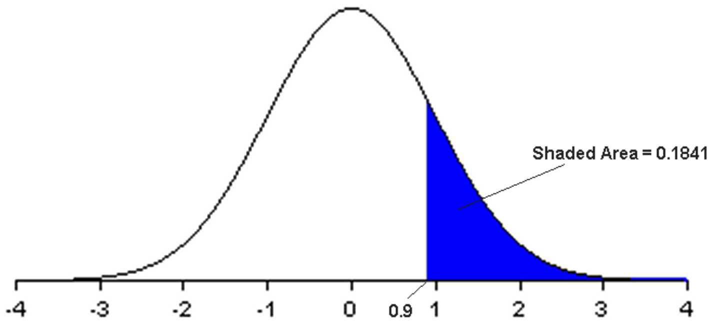
$$P\left(\left|\bar{Y}-\mu\right|\le 0.3\right)=P\left(-\frac{0.3}{\sigma_{\bar{Y}}}\le\underbrace{\frac{\bar{Y}-\mu}{\sigma_{\bar{Y}}}}_{Z}\le\frac{0.3}{\sigma_{\bar{Y}}}\right)$$

$$P\left(\left|\bar{Y}-\mu\right|\le 0.3\right)=P\left(-\frac{0.3}{1/3}\le Z\le\frac{0.3}{1/3}\right)$$

$$P\left(\left|\bar{Y}-\mu\right|\le 0.3\right)=P\left(-0.9\le Z\le 0.9\right)$$



$$P\left(\left|\bar{Y}-\mu\right|\le 0.3\right)=1-2.\underbrace{P\left(Z>0.9\right)}_{0.1841\ from\,table}$$



$$P\left(\left|\bar{Y}-\mu\right|\le 0.3\right)=1-2.\left(0.1841\right)=0.6318$$

Thus the chance is only 0.6318 that the sample mean will be within 0.3 ounce of the true population mean.

b) Now we want:

$$P\left(\left|\bar{Y}-\mu\right| \le 0.3\right) = P\left[-0.3 \le \bar{Y}-\mu \le 0.3\right] = 0.95$$

$$P\left(\left|\bar{Y}-\mu\right| \le 0.3\right) = P\left(-\frac{0.3}{\frac{\sigma}{\sqrt{n}}} \le \underbrace{\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}}_{Z} \le \frac{0.3}{\frac{\sigma}{\sqrt{n}}}\right) = 0.95$$

It is given that σ=1, hence we can write:

$$P\left(\left|\bar{Y}-\mu\right| \le 0.3\right) = P\left(-0.3\sqrt{n} \le Z \le 0.3\sqrt{n}\right) = 0.95$$

From Standard Normal Distribution table, we know that:

$$P\left(-1.96 \le Z \le 1.96\right) = 0.95$$

It implies that $0.3\sqrt{n}$ =1.96. Hence

$n = \left(\frac{1.96}{0.3}\right)^2 = \left(6.5\bar{3}\right)^2 = 42.68$, which is impractical for sample size.

So we can decide to n=43, since if n=43, $P\left(\left|\bar{Y}-\mu\right| \le 0.3\right)$ slightly exceeds 0.95.

B. Chi-Square Distribution

In statistics, the sampling distribution of the *sum of the squares* of independent, standard normal random variables is also widely used.

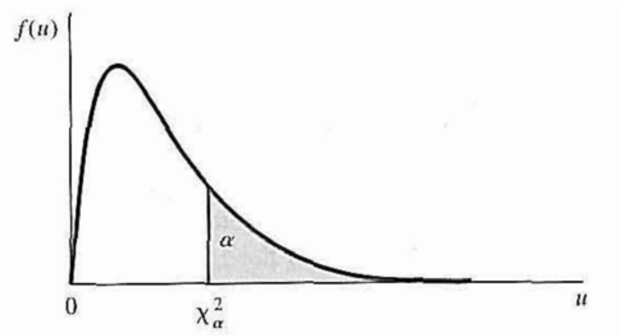
Theorem 7.2 Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)$ has a Chi-square, χ^2, distribution with n degrees of freedom.

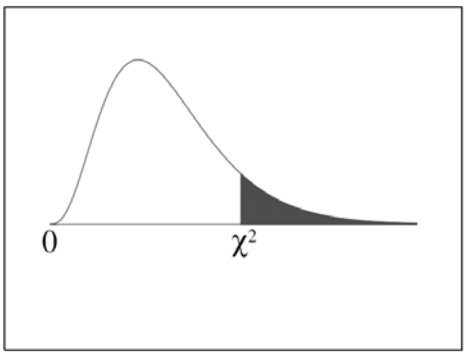
Note: We can denote a chi-square distribution variable with n degrees of freedom (df) for α level of significance as $\chi^2_{\alpha,df}$.

From Chi-square tables, we can find values χ^2_α so that:

$$P\left(\chi^2 > \chi^2_\alpha\right) = \alpha$$

for random variables with χ^2 distributions.



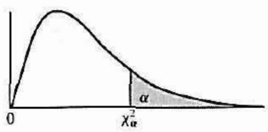


A distribution χ^2 showing upper-tail area α

Attention! The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$

For example, if the χ^2 random variable of interest has 10 degrees of freedom, we can use χ^2 Tables to find $\chi^2_{0.90}$. We look in the row labeled 10 df. and the column headed $\chi^2_{0.90}$ and read the value 4.86518. Therefore if Y has a χ^2 distribution with df=10: $P\left(\chi^2 > 4.86518\right) = 0.10$

Table 6. Percentage points of the χ^2 distributions



d.f.	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.168[illegible]
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150

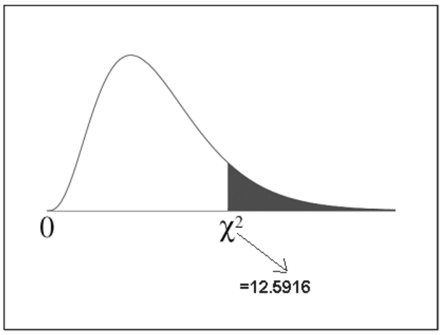
Example 7.3 If Z_1, Z_2, …Z_6 denotes a random sample from the standard normal distribution, find a number *b* such that: $P\left(\sum_{i=1}^{6} Z_i^2\right) = 0.95$

Solution

$\sum_{i=1}^{6} Z_i^2$ has a χ^2 distribution with 6 degrees of freedom (df) *since it represents the summation of 6 squared normally distributed random variables*, *Z*. Equivalently:

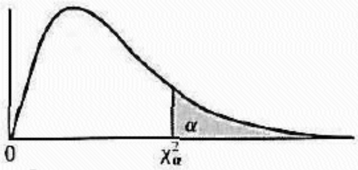
$$P\left(\sum_{i=1}^{6} Z_i^2\right) = 0.95$$

From Chi-square table, we find that:



$$P\left(\sum_{i=1}^{6} Z_i^2 > b\right) = 0.05 \Rightarrow b = 12.5916$$

Table 6. Percentage points of the χ^2 distributions



$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$	d.f.
2.70554	3.84146	5.02389	6.63490	7.87944	1
4.60517	5.99147	7.37776	9.21034	10.5966	2
6.25139	7.81473	9.34840	11.3449	12.8381	3
7.77944	9.48773	11.1433	13.2767	14.8602	4
9.23635	11.0705	12.8325	15.0863	16.7496	5
10.6446	12.5916	[illegible]94	16.8119	18.5476	6
12.0170	14.0671	[illegible]28	18.4753	20.2777	7
13.3616	15.5073	17.5346	20.0902	21.9550	8
14.6837	16.9190	19.0228	21.6660	23.5893	9
15.9871	18.3070	20.4831	23.2093	25.1882	10
17.2750	19.6751	21.9200	24.7250	26.7569	11
18.5494	21.0261	23.3367	26.2170	28.2995	12
19.8119	22.3621	24.7356	27.6883	29.8194	13

The χ^2 distribution plays an important role in many inferential procedures. For example, suppose that we wish to make an inference about the population variable σ^2 based on a random sample Y_1, Y_2, …, Y_n from a normal population. A good estimator for *population variance* σ^2 (we will see in Ch. 8) is the *sample variance,* S^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

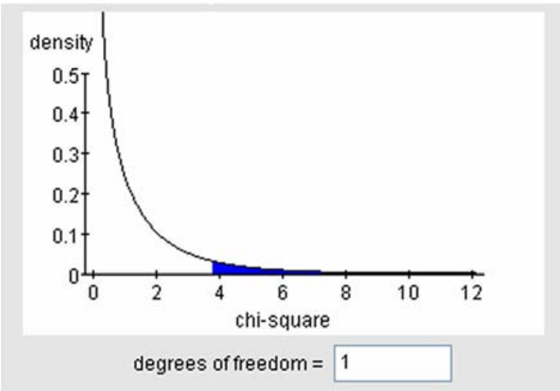
The following theorem gives the probability distribution for a function of the statistic S^2:

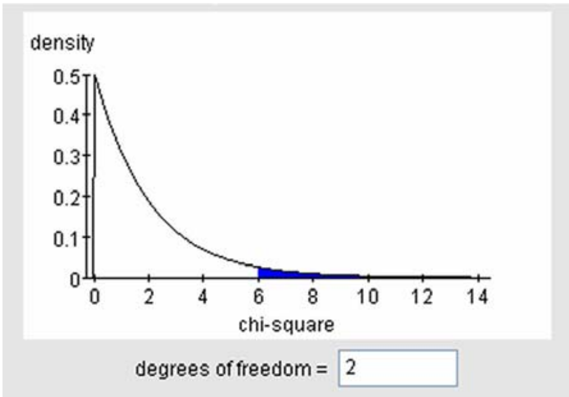
Theorem 7.3 Let Y_1, Y_2,…,Y_n be a random sample from a normal distribution with mean μ and variance σ^2. Then:

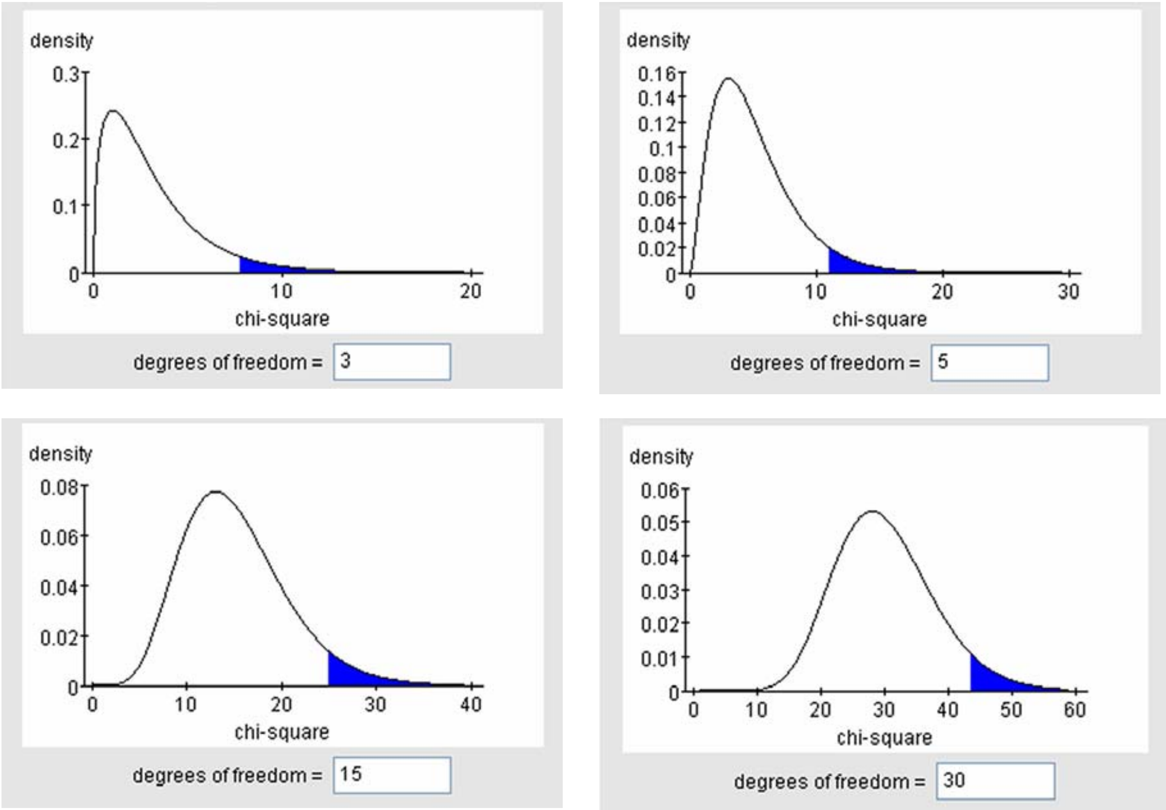
$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

has a χ^2 distribution with (n-1) degrees of freedom (df). Hence $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{\alpha,(n-1)}$. Also $\bar{Y}$ and S^2 are independent random variables.

Impact of Degrees of Freedom in the χ^2 Distribution







Note: Shaded areas in the graphs represent 0.05.
Source of Graph Applet: http://www.stat.tamu.edu/~west/applets/chisqdemo.html

Example 7.4 Recall the bottle example, where the ounces of fill from the bottling machine are assumed to have a normal distribution with $\sigma^2 = 1$.

- Suppose that we plan to select a random sample of 10 bottles and measure the amount of fill in each bottle. If these ten observations are used to calculate S^2, it might be useful to specify an interval of values that will include S^2 with a high probability.
- Find numbers b_1 and b_2 such that:

$$P\left(b_1 \leq S^2 \leq b_2\right) = 0.90$$

Solution

For our case, α=0.90 and degrees of freedom, *df*=10-1=9.

$$P\left[b_1 \leq S^2 \leq b_2\right] = P\left[\frac{(n-1)b_1}{\sigma^2} \leq \underbrace{\frac{(n-1)S^2}{\sigma^2}}_{\chi^2} \leq \frac{(n-1)b_2}{\sigma^2}\right] = 0.90$$

We are given that $\sigma^2 = 1$ so it follows that $(n-1)S^2$ has a χ^2 distribution with *n-1=10-1=9* degrees of freedom (*df*):

$$P\left[b_1 \leq S^2 \leq b_2\right] = P\left[\underbrace{(n-1)b_1}_{a_1} \leq (n-1)S^2 \leq \underbrace{(n-1)b_2}_{a_2}\right] = 0.90$$

From a chi-square table we can find two numbers a_1 and a_2 such that:

$$P\left[a_1 \leq (n-1)S^2 \leq a_2\right] = 0.90$$

One method of doing this to find the value of a_2 that represents an area of 0.05 in the upper tail and the value of a_1 that represents 0.05 in the lower tail (0.95 in the upper tail).

Because there are *n-1* degrees of freedom, the Chi-square table gives a_2=16.919 and a_1=3.32511.

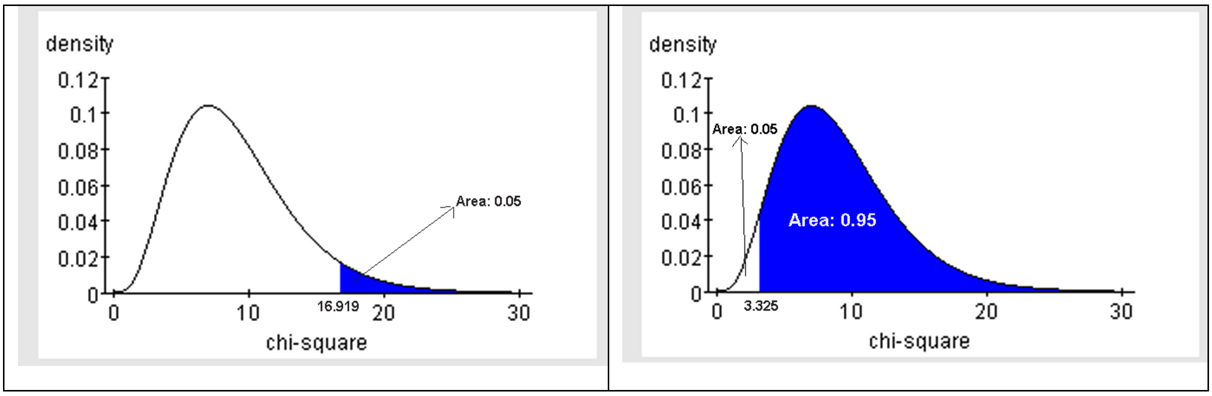
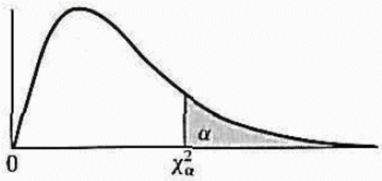
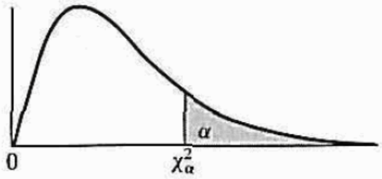


Table 6. Percentage points of the χ^2 distributions



$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$	d.f.
2.70554	3.84146	5.02389	6.63490	7.87944	1
4.60517	5.99147	7.37776	9.21034	10.5966	2
6.25139	7.81473	9.34840	11.3449	12.8381	3
7.77944	9.48773	11.1433	13.2767	14.8602	4
9.23635	11.0705	12.8325	15.0863	16.7496	5
10.6446	12.5916	14.4494	16.8119	18.5476	6
12.0170	14.0671	16.0128	18.4753	20.2777	7
13.3616	15.5073	17.5346	20.0902	21.9550	8
14.6837	16.9190	[illegible]	21.6660	23.5893	9
15.9871	18.3070	[illegible]	23.2093	25.1882	10
17.2750	19.6751	21.9200	24.7250	26.7569	11
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Table 6. Percentage points of the χ^2 distributions



d.f.	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	[illegible]
10	2.15585	2.55821	3.24697	3.94030	[illegible]
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Consequently, values for b_1 and b_2 are given by:

- $a_1 = \frac{(n-1)b_1}{\sigma^2} = \frac{9b_1}{1} = 3.32511 \Rightarrow b_1 = \frac{3.32511}{9} \approx 0{,}369$
- $a_2 = \frac{(n-1)b_2}{\sigma^2} = \frac{9b_2}{1} = 16.9190 \Rightarrow b_2 = \frac{16.9190}{9} \approx 1{,}880$

Thus, if we want to have an interval that will include S^2 with probability 0.90, this interval is given by $(0.369, 1.880)$. Note that this interval is quite wide.