Distribution of Variance and the Chi square (χ^2) distribution

The distribution of Variance

Given $Y \sim \mathrm{N}(\mu, \sigma^2)$

$\mathrm{E}(S^2) = \sigma^2$

where;

S^2 = SS / df and if we let df = γ,

for a sample γ = n–1

$$S^2 = {}^{SS}\!/_{\gamma} = {}^{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}\!\Big/_{\gamma}$$

or for a population γ = N and

$$\sigma^2 = {}^{SS}\!/_{\gamma} = {}^{\sum_{i=1}^{n}(Y_i - \mu)^2}\!\Big/_{N}$$

This is the structure of variance for any variable.

Recall, $Z_i = \left(\dfrac{Y_i - \mu}{\sigma}\right) \sim \mathrm{N}(0, 1)$

We now define a new distribution, Z_i^2,

$$\sum_{i-1}^{n} Z_i^2 = \sum_{i-1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2 = \sum_{i-1}^{n}\frac{(Y_i - \mu)^2}{\sigma^2} = \frac{\text{Sum of Squares}}{\sigma^2} = \frac{\text{SS}}{\sigma^2}$$

Recall, $\sigma^2 = {}^{\sum_{i=1}^{n}(Y_i - \mu)^2}\!\Big/_{N} = {}^{SS}\!/_{\gamma}$

So $SS = \gamma\sigma^2$ or more generally, SS = γVar, because the variance will not always be σ^2

Therefore, $\sum_{i-1}^{n} Z_i^2 = {}^{SS}\!/_{\sigma^2} = {}^{\gamma Var}\!/_{\sigma^2}$

Finally

E(Var) = σ^2

$$E\left(\sum_{i-1}^{n} Z_i^2\right) = \left({}^{SS}\!/_{\sigma^2}\right) = E\left({}^{\gamma Var}\!/_{\sigma^2}\right) = {}^{\gamma\sigma^2}\!/_{\sigma^2} = \gamma$$

where,

γ = N for a population or sample with known σ^2

γ = n–1 for a sample using $\overline{Y}$ to get deviations from the mean and variance

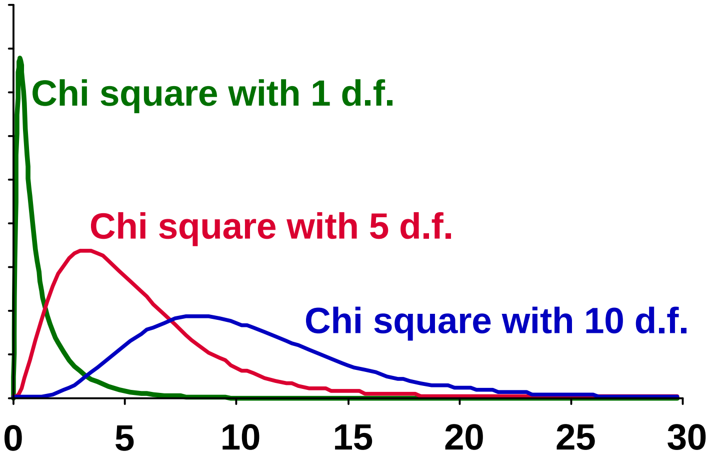
Use of the expected value tells us that if all possible samples of size n are drawn from $Y \sim$ N(μ, σ^2), then on the average, $\sum_{i-1}^{n} Z_i^2$ will take the value of the d.f. (γ).

This is the distribution of variance, $S^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{\text{SS}}{d.f.} = \frac{\text{SS}}{\gamma}$

and $E\left(\sum_{i-1}^{n} Z_i^2\right) = \gamma$ is a new distribution is called the Chi square (χ^2). The most useful form of this distribution for hypothesis testing is SS /σ^2, and the distribution centers on γ.

Properties of the Chi Square distribution

- The distribution has only one parameter, γ
- For every γ there is a different distribution
- The distribution is non-negative (positive values only), ranging from 0 to $+\infty$
- The distribution is asymmetrical, but it approaches symmetry as γ increases



The Chi square tables

- The left side gives the degrees of freedom, γ. Each degree of freedom is a different distribution, given in the rows as with the t-table.
- The probability in the upper TAIL of the distribution is given in the row at the top of the table.
- The distribution is NOT symmetric, so the probabilities at the top must be used for both upper and lower limits.

Partial Chi square table

d.f.	0.995	0.99	0.975	0.95	0.5	0.05	0.025	0.01	0.005
1	0.00	0.00	0.00	0.00	0.45	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	1.39	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	2.37	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	3.36	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	4.35	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	5.35	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	6.35	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	7.34	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	8.34	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	9.34	18.31	20.48	23.21	25.19
20	7.43	8.26	9.59	10.85	19.34	31.41	34.17	37.57	40.00
30	13.79	14.95	16.79	18.49	29.34	43.77	46.98	50.89	53.67
100	67.33	70.06	74.22	77.93	99.33	124.34	129.56	135.81	140.17

Hypothesis testing

We will be able to use this distribution to test hypotheses about the variance.

1) H_0: $\sigma^2 = \sigma_0^2$ (note we are testing hypotheses about variances)

2) H_1: $\sigma^2 \neq \sigma_0^2$ (directional alternatives can also be tested).

3) Assume independence and normality*

* some types of Chi square test do not require the assumption of normality.

4) Set α (at say 0.05 or 0.01 as before) and we will need to learn to set critical limits.

5) Draw a sample of size n and calculate an estimate of the variance (S^2 for a sample).

We know for the Chi square that

$$\chi^2 = \Sigma Z_i^2 = \left(\sum \left(Y_i - \bar{Y} \right)^2 \Big/ \sigma^2 \right) = SS \Big/ \sigma^2$$

And if the null hypothesis is true, $\chi^2 = SS \Big/ \sigma_0^2$ will have a Chi square distribution and will center on γ (the degrees of freedom).

6) Compare the critical limits to the calculated statistic, and

7) Draw our conclusions and interpret the results.

Critical limits from the Chi square distribution

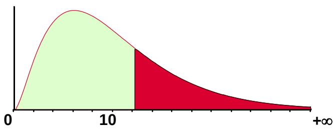
The tables are similar to the t-table in that (1) each row is a different distribution and (2) selected probabilities are given at the top.

The tables are different from the t-table in that the tables (1) are not symmetric and (2) do not center on a single value (zero) like the t-table, but rather each distribution centers on its d.f.

Examples of using the Chi square tables

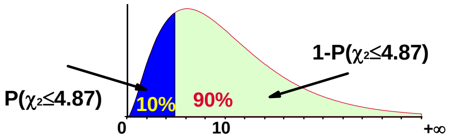
Given γ = d.f. = 10, find $P(\chi^2 \geq \chi_0^2) = 0.25$

This is an area in the tail of the distribution, consistent with our tables. We look up the value in the tables and find for d.f. = 10 the value that leaves 25% in the upper tail is 12.549.



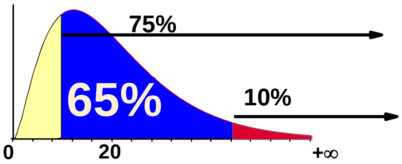
Given γ = d.f. = 10, find $P(\chi^2 \leq 4.87) = ?$

This is an area in the LOWER tail of the distribution. This value is not given in our tables, but the area under the curve is still equal to one, so $P(\chi^2 \leq 4.87) = 1 - P(\chi^2 \geq 4.87)$.



Given γ = d.f. = 20, find $P(15.5 \leq \chi^2 \leq 28.4)$

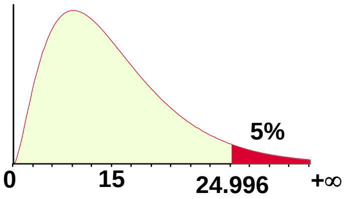
This is an area in the center of the distribution. There are two ways to get this area, find the tails. $P(\chi^2 \leq 15.5)$ and $P(\chi^2 \geq 28.4)$, and subtract them from 1.



The other way is to find the probability that $P(\chi^2 \geq 15.5)$ and $P(\chi^2 \geq 28.4)$ and subtract the first from the second. This is easier given the way our tables are set up.

Given γ = d.f. = 15, find $P(\chi^2 \geq \chi_0^2) = 0.05$

This is the area in the upper tail provided by our tables. This value can be red directly from the tables.



Given γ = d.f. = 15, find $P(\chi_1^2 \leq \chi^2 \leq \chi_2^2) = 0.95$

This is the area in the middle of our tables. We must assume symmetry or there are an infinite number of answers. Since a=1–0.95 = 0.05, we calculate $\alpha/2 = 0.025$. The upper tail can be read directly from the table, χ_2^2 =27.488. The value that would leave 2.5% in the lower tail would leave 97.5 (0.975) in the upper tail, so χ_1^2 =6.262.

Hypothesis testing

1) State the null hypothesis. $H_0: \sigma^2 = \sigma_0^2$

e.g. $H_0: \sigma^2 = 10$, where $\sigma_0^2 = 10$

2) $H_1: \sigma^2 \neq \sigma_0^2$. This could also have been a one tailed alternative, $H_1: \sigma^2 < \sigma_0^2$ or $H_1: \sigma^2 > \sigma_0^2$.

3) Assume independence and normality. Some other Chi square tests do not require the assumption of normality.

4) Set α, say 0.05 (0.01 would be another common choice) and determine the critical limits for the test.

We have a two-tailed test given $H_1{:}\sigma^2 \neq \sigma_0^2$, and want α = 0.05 for two tails, 0.025 in each tail. Given that n = 20 and d.f. = γ = 19 we want to find upper and lower limits so that P($\chi_1^2 \le \chi^2 \le \chi_2^2$)=0.95.

P($\chi_1^2 \le \chi^2 \le \chi_2^2$)

P($\chi^2 \le \chi_1^2$) = 0.025 or P($\chi^2 \ge \chi_1^2$) = 0.975, χ_1^2 = 8.91

P($\chi^2 \ge \chi_2^2$) = 0.025, χ_2^2 = 32.9

5) Draw a sample of size n and calculate an estimate of the variance (S^2 for a sample).

$\chi^2 \le SS/\sigma_0^2 = \sum(Y_i - \bar{Y})^2/\sigma_0^2$. In this case χ^2 = 400/10 = 40 with 19 d.f.

6) Compare the critical limits from the Chi square table (8.91 and 32.9) to the calculated test statistic (χ^2 = 40). The calculated value exceeds the upper limit in the area of rejection.

7) Since the calculated value exceeds the upper limit we would reject the null hypothesis and conclude the results were consistent with the alternate hypothesis. Since we have rejected the null hypothesis, there is a 5% possibility that we made a type I error.

Numerical example

Lobsters are to be used in a growth experiment. Weight gain will be studied, and it is important that there be little variation in the initial weights. Based on previous experience, we know that an initial standard deviation of NO MORE than 0.5 oz. would be adequate. Determine if this tolerance is met. Less variation is no problem, only exceeding the 0.5 oz. value.

1) $H_0{:}\sigma^2 = \sigma_0^2$, in this case $\sigma_0^2 = (0.5)^2 = 0.25$

2) $H_1{:}\sigma^2 > \sigma_0^2$

3) Assume the sample is independent (randomly sampled) and the lobster weights are normally distributed.

4) State the level of significance (α). We will use α = 0.01 since the validity of our experiment depends on this one and a type II error only means we draw a different sample. Determine the critical limit. Given that this is a one tailed test ($H_1{:}\sigma^2 > \sigma_0^2$) and the value of α = 0.01 and that the sample size is 12 and degrees of freedom are γ = df = 12 – 1 = 11 then find $P\left(\chi^2 \ge \chi_0^2\right) = 0.01$. From the table this value is 24.7.

6) Draw a sample and compute χ^2 value. The new sample values, where n = 12,

Y_i = 11.9, 11.8, 12.7, 12.3, 12.1, 11.3, 12.6, 11.5, 11.9, 12.0, 11.8, 12.1

$$\sum_{i=1}^{n} Y_i^2 = 1729.80, \quad \sum_{i=1}^{n} Y_i = 144$$

$$SS = \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2 \Big/ n = 1729.80 - (144)^2/12 = 1.8$$

$$\chi^2 = SS/\sigma_0^2 = 1.8/0.25 = 7.2 \quad \text{with 11 df}$$

6) Compare the calculated test statistic value (7.2) to the critical limit from the table (24.7). In this case the test statistic does not occur in the region of rejection.

7) Since the calculated value (7.2) is less than the critical limit value, and we would fail to reject the null hypothesis and conclude that our results are consistent with the null huypothsis. There is a chance of a Type II error.

We never actually finished our calculation of the variance. The SS was 1.8, so S^2 = 1.8 / 11 = 0.1636 and S = √0.1636 = 0.4045. Had we noted earlier that the calculated value was actually smaller than our hypothesized value, we could have stopped. It couldn't be in the upper tail.

Review

For a population, $E\left(\sum_{i-1}^{n} Z_i^2\right) = \sum(Y_i-\mu)^2/\sigma^2 = \left(SS/\sigma^2\right) = E\left(\gamma Var/\sigma^2\right) = \gamma\sigma^2/\sigma^2 = \gamma$

All of these follow a chi square distribution

For a sample, $E\left(\sum_{i-1}^{n} Z_i^2\right) = \sum(Y_i-\bar{Y})^2/\sigma^2 = \left(SS/\sigma^2\right) = E\left((n-1)S^2/\sigma^2\right) = \gamma$

All of these follow a chi square distribution

$\chi^2 = SS/\sigma^2$ is the form used for hypothesis testing

Hypothesis tests covered so far.

$$Z = \frac{\left(\bar{Y}-\mu_0\right)}{\sigma_{\bar{Y}}}$$

$$t = \frac{\left(\bar{Y}-\mu_0\right)}{S_{\bar{Y}}}$$

$$\chi^2 \le SS/\sigma_0^2 = \sum(Y_i-\bar{Y})^2/\sigma_0^2$$

The last distribution we will discuss is the F distribution. This distribution will allow us to test H_0: $\sigma_1^2 = \sigma_2^2$, and some other tests about the equality of means for more than two means.

Which test to use?

To test means	
H_0: $\mu = \mu_0$, σ^2 known	Z test
H_0: $\mu = \mu_0$, σ^2 not known	t test
To test variances	
H_0: $\sigma^2 = \sigma_0^2$	χ2 test
H_0: $\sigma_1^2 = \sigma_2^2$ (covered later)	F test

Other applications of the Chi Square distribution

There are two other broad applications of the Chi square distribution.

Chi square test of independence

Chi square test of goodness of fit

Both tests are based on the general formula

$$\chi^2 = \sum (\text{Observed - Expected})^2 \Big/ \text{Expected}$$

Assumptions

These are non-parametric tests, they do not require an assumption of normality.

The assumption for previous tests was that the distribution being examined was a "Normally and Independently Distributed random variable", often abbreviated NID r.v.

The assumption for these applications of the Chi square is that the distribution being examined is an "Independently and Identically Distributed random variable", often abbreviated IID r.v.

Chi Square Test of Independence

In an experiment we have observed the number of individuals of 3 species of mosquitoes at a particular site during 3 times of the day. Do the mosquito species occur with different frequencies during different periods of the day, or is the type of mosquito observed independent of the time of day?

	Mosquito species				
Time of day	A	B	C	Row total	Row percent
Mid morning	8	16	7	31	20.67
Mid Afternoon	9	5	9	23	15.33
Dusk	23	9	64	96	64
Column total	40	30	80	150	
Column percent	26.67	20	53.33		

Expected values are calculated as the (row total * column total / grand total). The first cell would be calculated as 40*31/150 = 8.27.

	Mosquito species		
Time of day	A	B	C
Mid morning	8.2700	6.2000	16.5300
Mid Afternoon	6.1300	4.6000	12.2700
Dusk	25.6000	19.2000	51.2000

Individual Chi square values are calculated as the [(Observed - Expected)^2/Expected]. The first cell would be calculated $(8–8.27)^2$ / 8.27=0.01.

	Mosquito species		
Time of day	A	B	C
Mid morning	0.0088	15.4900	5.5000
Mid Afternoon	1.3400	0.0300	0.8700
Dusk	0.2600	5.4200	3.2000

The Chi square test statistic is the sum of the chi square values in each cell.

$$\sum(\text{Observed}-\text{Expected})^2 \big/ \text{Expected} = 32.12335 \quad \text{with d. f.} = (\text{rows}–1)^*(\text{col}–1) = 2^*2 = 4$$

The steps of the hypothesis test

1) Mosquito species occurrence is independent of time of day

2) Mosquito occurrence is NOT independent

3) Assume IID r.v.

4) $\alpha = 0.05$, the critical value with (r–1)(c–1) = 4 d.f. for $\alpha = 0.05$ is 9.4877.

5) From the sample we get a calculated χ^2=32.12335

6) The calculated test statistic exceeds the tabular value with (r–1)(c–1) = 4 d.f.. The critical tabular values are;

$\alpha = 0.05$,	$\chi^2 = 9.4877$	statistically significant
$\alpha = 0.01$,	$\chi^2 = 13.2767$	highly significant
$\alpha = 0.001$,	$\chi^2 = 18.4662$	Wow!

Using this terminology we see we have highly significant departure in this case.

So what is the "P value", the chance of finding a value of 32.1335 or greater by random chance? This is P(>χ^2)=0.0000018052, about 2 in a million.

7) So we conclude that the occurrence of mosquito species is not independent of the time of day. It appears to be very dependent.

SAS example 3a of the Chi square Test of Independence in SAS

Two varieties of a particular moth species occur in two colors (brown and white). A biologist in North Carolina wants to know if the distribution of the two varieties varies with the area of the state. He collects individuals from each region of the state and note the number of each variety.

See computer output