Chapter 1

Chi-Squared χ^2 Tests

Definition 1.1

A random variable X is said to follow a χ^2-distribution, with degree of freedom ν, iff its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

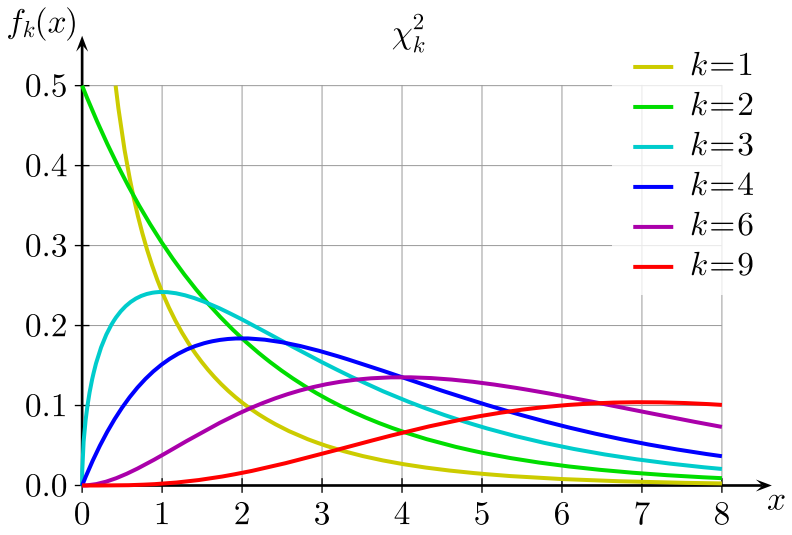


Figure 1.1: Illustration of how the $\chi^2_{(\nu)}$ distribution looks with increasing degree of freedom ν.

General Information

- Properties of chi-squared distributions.
  - $\mathrm{E}(X) = v$ and $\mathrm{Var}(X) = 2\nu$.
  - The $\chi^2_{(\nu)}$ distribution tends to a normal distribution as $\nu \to \infty$.
  - Suppose $Z_i \sim \mathrm{N}(0,1)$ are independent. Then, $Z_1^2 + \cdots + Z_n^2 \sim \chi^2_{(n)}$.
  - If $X \sim \chi^2_{(\nu)}$ and $Y \sim \chi^2_{(v)}$, then $X + Y \sim \chi^2_{(\nu+v)}$.
- A goodness-of-fit test.
  1. Let [X in context].

2. | Test H_0: [X follows the distribution in context]
against H_1: [X does not follows the distribution in context]
at the $100\alpha\%$ significance level. |

3.

x	x_1	x_2	$\cdots$	x_n
f_i	f_1	f_2	$\cdots$	f_n
e_i	e_1	e_2	$\cdots$	e_n
$\frac{(f_i-e_i)^2}{e_i}$	$\frac{(f_1-e_1)^2}{e_1}$	$\frac{(f_2-e_2)^2}{e_2}$	$\cdots$	$\frac{(f_n-e_n)^2}{e_n}$

Table 1.1: Observed and expected frequencies for a goodness-of-fit test

4. Check whether $e_i \geq 5$ for each of the n classes. If it isn't, we need to combine *just enough* adjacent classes, till they do. Working-wise, use some underbraces/overbraces to indicate the combined values.
5. Under H_0, the test statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(F_i - E_i)^2}{E_i} \sim \chi^2_{(\nu)}.$$

   Here, $n \coloneqq$ #classes and $\nu =$ (#classes − #estimated parameters) − 1.
6. Continue as per usual, calculating the critical region $\chi^2_{(\nu)} > \chi^2_{(\nu,1-\alpha)}$ or the p-value.

G.C. Skills

- To find the value of $\chi^2_{(\nu,1-\alpha)}$, which satisfies $\mathrm{P}\left(X > \chi^2_{(\nu,1-\alpha)}\right) = \alpha$, we use the table in the MF26 formula sheet (Page 9). Unfortunately, there is no inverse χ^2 function available.
- For the p-value:

$$\texttt{stat} \Longrightarrow \texttt{TESTS} \Longrightarrow \texttt{D:}\chi^2\texttt{GOF-Test...}$$

Note

If X follows a *discrete* uniform distribution, we must state it out in words. We cannot write $X \sim \mathrm{U}(\mu, \sigma^2)$ as this would denote that X is a *continuous* random variable. But if $X \sim \mathrm{B}(n, p)$ (or $X \sim \mathrm{Po}(\lambda)$, etc), then we can just denote it as such.

Example 1.1: #estimated parameters = 0

Given $X \sim \mathrm{N}(0, 1)$ (note how the *population parameters* that define the distribution are *known*), the degree of freedom $\nu =$ #estimated parameters $\coloneqq n$.

Example 1.2: #estimated parameters = 1

Consider when $X \sim \mathrm{B}(m, p)$, such that the expected frequency for each of the n classes is at least 5, but we do not know the exact value of p. So, we *estimate* it according to the sample given. Then, the degree of freedom is $\nu = n - 1 - 1 = n - 2$.

Example 1.3: #estimated parameters = 2

Similarly, suppose $X \sim \mathrm{N}(\mu, \sigma^2)$, such that the expected frequency of each of the n classes is at least 5, and the true value of μ and σ^2 are unknown. In this case, the degree of freedom

$\nu = n - 2 - 1 = n - 3$.

Note

Suppose we are given a question of the following form.

Some context...

x_i	x_1	x_2	$\cdots$	x_n
f_i	f_1	f_2	$\cdots$	f_n

Table 1.2: Some data.

(a) Show, at the $100\alpha\%$ significance level, that the data does not support the hypothesis of $X \sim \text{Geo}(p)$ with $p = 0.5$.

(b) State how the test in (a) would have to be amended to test the hypothesis of a geometric distribution for an *unspecified value of p*.

Then, for (ii), two main changes have to be made:

1. Estimate the value of p by computing the sample mean $\bar{x}$ and letting $p = \frac{1}{\bar{x}}$.
2. Adjust the degree of freedom from 4 to $4 - 1 = 3$, as there is one more restriction, that the mean must agree.

(The phrasing is similar for gof tests for other distributions; simply use the appropriate estimators for the unknown population parameters.)

Tests of independence.

1. Let [X in context].
2. Test H_0: [X in context] is independent of [Y in context]
   against H_1: [X in context] is dependent on [Y in context]
   at the $100\alpha\%$ significance level.
3. *Note.* Unless the question asks for it, we do not need to write $\left[\frac{(f_i - e_i)^2}{e_i}\right]$ or its corresponding values, in the following table.

f_i (e_i) $\left[\frac{(f_i-e_i)^2}{e_i}\right]$		X				
		x_1	x_2	$\cdots$	x_n	Total
Y	y_1					t_{r_1}
	y_2					t_{r_2}
	$\vdots$					$\vdots$
	y_m					t_{r_m}
	Total	t_{c_1}	t_{c_2}	$\cdots$	t_{c_n}	$\sum t_{r_i} + \sum t_{c_i}$

Table 1.3: *Expected* frequencies for a test of independence.

4. Under H_0, the test statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(F_i - E_i)^2}{E_i} \sim \chi^2_{(\nu)}.$$

Here, $n \coloneqq \#\text{cols}$ and $\nu = (\#\text{rows} - 1)(\#\text{cols} - 1)$.

5. Continue as per usual, calculating the critical region $\chi^2_{(\nu)} > \chi^2_{(\nu,1-\alpha)}$ or the p-value.

G.C. Skills

Key in the matrix of *observed frequencies* (not Table 1.2 of *expected* frequencies):

$$\texttt{2nd} \implies \texttt{x}^{-1} \implies \texttt{EDIT} \implies \texttt{[A]}.$$

Then, conduct the test for independence:

$$\texttt{stat} \implies \texttt{TESTS} \implies \texttt{C:}\chi^2\texttt{-Test...}$$

Note

If the question says to "use an approximate χ^2-statistic...", then we must use the critical region method. It is incorrect to use the p-value.

Note

Consider we are asked to state which cells correspond to the highest contributions to the test statistic, and relate that back to the context of the question. Then:

1. State the cells in the form (___, ___). E.g. (High, Good) and (Low, Good).
2. In table 1.3, add an asterisk to each of these cells. E.g. $\boxed{1\ (5)\ [10.1]^*}$.
3. Use words that imply correlation and *not* causation. E.g. directly associated, correlates with, etc.

Note

On a similar note, if the question asks "Can it can be concluded that...", but is unclear about whether it's implying correlation or causation, it may be safer to explain both ways. i.e. what correlation is there and why is there no causation.

Note

Explain why we cannot conclude any casual relationships from a test of independence.

No, the above test does not reflect the actual casual relationship between the two factors, if it exists. Rather, it merely suggests that they are not independent.

Note

Explain why we cannot apply a χ^2-test for independence using the data given.

The expeceted frequency for (___, ___) is ___ < 5. If we combine the columns, the degree of freedom $\nu = 1 \cdot 0 = 0$. If we combine the rows, $\nu = 0 \cdot 1 = 0$. Thus, we cannot apply a χ^2-test for independence.

Chapter 2

Correlation and Linear Regression

Note

A good scatter diagram should follow the guidelines below.

- The relative position of each point on the scatter diagram should be clearly shown.
- The range of values for the set of data should be clearly shown by marking out the extreme x and y values on the corresponding axis.
- The axes should be labeled clearly with the variables.

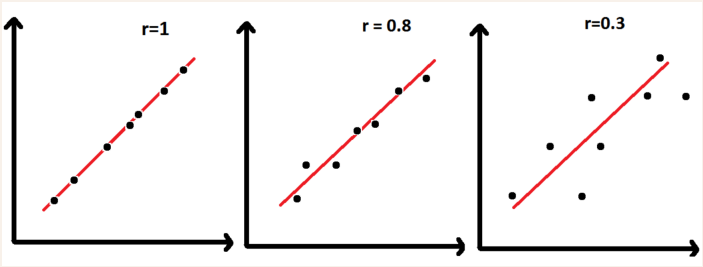
General Information

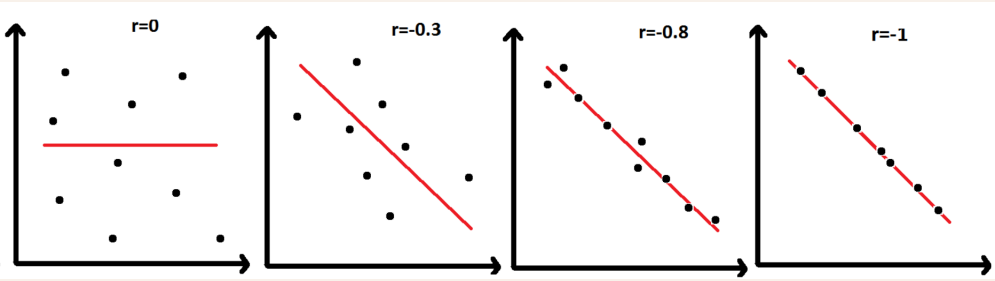
- The Product Moment Correlation Coefficient is a measure of the linear correlation between two variables. It is defined by

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}},$$

  which takes on a value from 0 to 1.

- When $r = 0$, there is no linear relationship. But, a nonlinear relationship may be present. Additionally, the regression lines are perpendicular.
- The closer the value of r is to 1 (or -1), the stronger the positive (or negative) linear correlation. Furthermore, the regression lines coincide.





- The regression line of y on x minimises the sum of squares deviation (error) in the y-direction. (i.e. we are assuming x is the independent variable whose values are known exactly.) It is given by

$$y = \bar{y} + b(x - \bar{x}), \qquad \text{where} \qquad b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}}.$$

- The regression lines of y on x and x on y intersect at $(\bar{x}, \bar{y})$.
- Say we are given the value of one variable, and asked to approximate the the value of the other variable. Then, we should always use the line of the *dependent* variable on the *independent*.
- Estimations should not be taken for data outside the range of the sample provided, even if the value of r is close to 1.