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

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

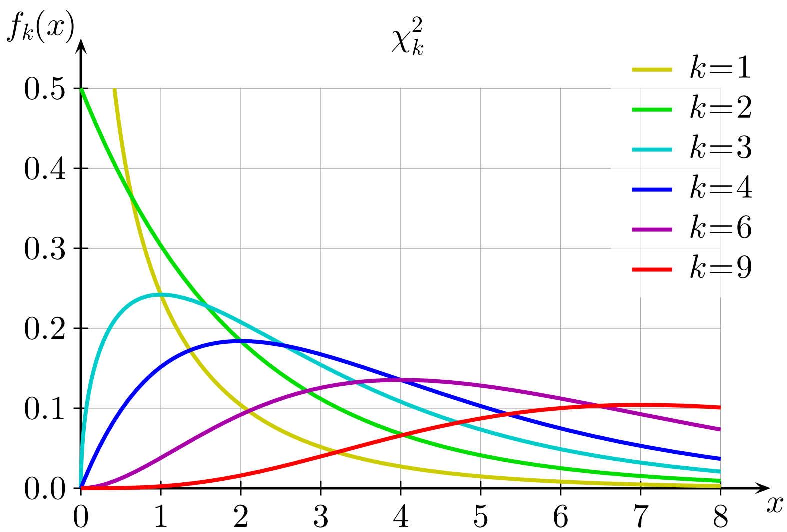


Figure 1.1: Illustration of how the $\chi^2_{(\nu)}$ distribution looks with increasing degree of freedom ν.

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 $Y \sim \chi^2_{(\upsilon)}$, then $X + Y \sim \chi^2_{(\nu+\upsilon)}$.
- 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
Observed frequency f_i	f_1	f_2	$\cdots$	f_n
Expected frequency e_i	e_1	e_2	$\cdots$	e_n

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

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

Note

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

Note

The expected frequency for each of the n classes should be at least 5. If it isn't, we need to combine *just enough* adjacent classes, till they do.

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 $:= n$.

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

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:

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

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.

		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 1.2: *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, $n :=$ #cols and $\nu =$ (#rows $-$ 1)(#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:

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

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.

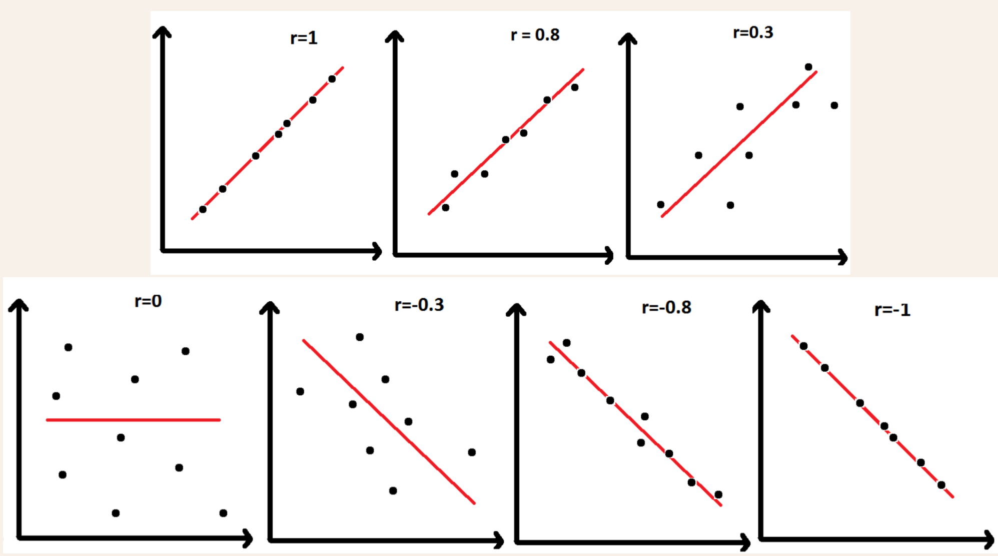
General Information

- 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

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

- The point $(\bar{x}, \bar{y})$ always lies on both the regression lines of y on x, and x on 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.