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

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

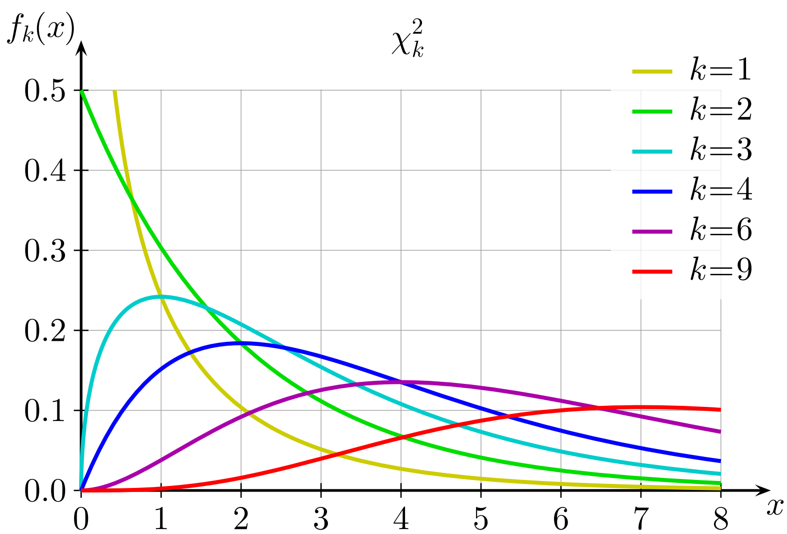


Figure 1.1: Illustration of how the $\chi^2_{(\nu)}$ distribution looks with increasing degree of freedom ν.

General Information

- Properties of chi-squared distributions.
  - $\mathrm{E}(X) = \nu$ 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. *Note.* Use a pen to draw any necessary tables.

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
$\dfrac{(f_i - e_i)^2}{e_i}$	$\dfrac{(f_1 - e_1)^2}{e_1}$	$\dfrac{(f_2 - e_2)^2}{e_2}$	$\cdots$	$\dfrac{(f_n - e_n)^2}{e_n}$

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, $n :=$ #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:

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

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 $=: n$.

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 values of μ and σ^2 are unknown. In this case, the degree of freedom $\nu = n - 2 - 1 = n - 3$.

Note

Consider when we are testing

Test	$H_0\colon X \sim \mathrm{N}(\mu, \sigma^2)$
against	$H_1\colon X \not\sim \mathrm{N}(\mu, \sigma^2)$
at the $100\alpha\%$ significance level.	

So, we want to fill up the values of e_i below.

x	$a_1 \le x_1 \le a_2$	$a_2 \le x_2 \le a_3$	$\cdots$	$a_n \le x_n \le a_{n+1}$
f_i	f_1	f_2	$\cdots$	f_n
e_i	e_1	e_2	$\cdots$	e_n

Table 1.2: Observed and expected frequencies when testing goodness-of-fit with a normal distribution.

Let the sample size $\sum f_i$ be m. Then, we should calculate $e_1 = m\,\mathrm{P}(-\infty < X \le a_2)$ and $e_n = m\,\mathrm{P}(a_n \le X < \infty)$, instead of $e_1 = m\,\mathrm{P}(a_1 \le X \le a_2)$ or $e_n = m\,\mathrm{P}(a_n \le X \le a_{n+1})$. Similarly, for goodness-of-fit tests with Poisson and Geometric distributions, we must also be careful in ensuring that we account for *all* possible values which X can take on, in calculating e_i.

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 1.3: Some data.

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

(ii) State how the test in (i) 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 $p = 1/\bar{x}$.
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 1.4: *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, $n \coloneqq$ #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):

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

Then, conduct the test for independence:

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

Note

If it's unclear as to what is to be stated as independent/dependent in the hypotheses, consider the expected values and how they relate to the context.

Example 1.4

Consider the following context:

Statement	Independent/Dependent?
There is consistency in the marking of the two T.A.s.	?
There is no consistency in the marking of the two T.A.s.	?

Table 1.5: Two statements on the relationship between the marks awarded and the T.A. marking.

Then, under H_0 — the independence claim — the expected frequencies are as stated below.

e_{ij}		Grade A	B	C
T.A.	X	a	b	c
	Y	a	b	c

Table 1.6: Expected frequencies.

Since $e_{1j} = e_{2j}$ for all $1 \leq j \leq 3$, we infer the following.

Statement	Independent/Dependent?
There is consistency in the marking of the two T.A.s.	Independent
There is no consistency in the marking of the two T.A.s.	Dependent

Table 1.7: Which statement corresponds to independence and which coresponds to dependence.

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 when 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 1.4, 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.

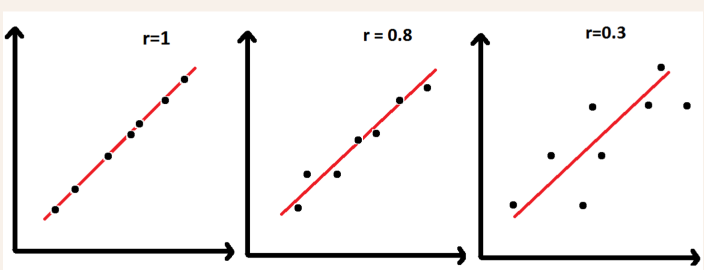
General Information

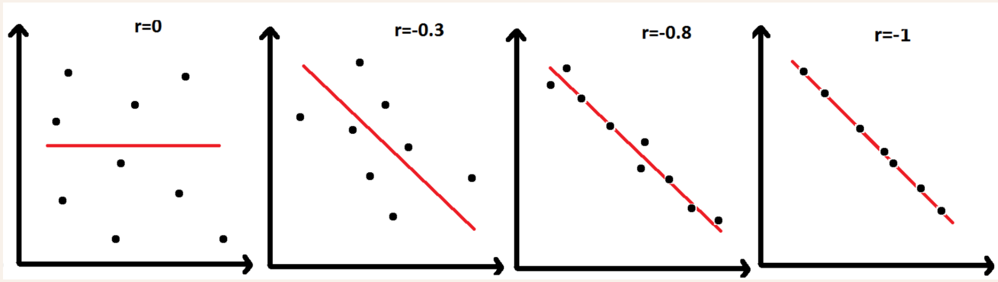
- The Product Moment Correlation Coefficient is a measure of the linear correlation between two variables. It is defined by

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

  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

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

- 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.

Chapter 3

Non-Parametric Tests

General Information

- A *sign test*.

  1. Let m be the population median of $D =$ ________ $-$ ________.
  2. Test $H_0\colon m = m_0$ against $H_1\colon$ (a) $m < m_0$, (b) $m \neq m_0$, or (c) $m > m_0$, at the $100\alpha\%$ significance level.
  3.

[label in context]	1	2	3	$\cdots$	n
Sign	$+$	0	$-$	$\cdots$	$+$

Table 3.1: The signs of $d_1, d_2, \ldots, d_n$, for a sign test. Instead of $1, 2, \ldots, n$ the labeling/column headers can differ in the given context. E.g. $A, B, \ldots, K$. Similarly, the signs here are mere examples; the ith sign cell should be filled with $+$ $(-)$ $[0]$ if $\operatorname{sgn}(d_i) = 1$ $(= -1)$ $[= 0]$.

  4. Let X_+ be the number of '$+$'. Under H_0, $X_+ \sim \mathrm{B}(n, 1/2)$, $x_+ = 11$. (Alternatively, X_- can also be used.)
  5. Since p-value $=$ ______ $< 100\alpha\%$ $(\geq 100\alpha\%)$, there is sufficient (insufficient) evidence, at the $100\alpha\%$ significance level, to conclude that [H_1 in context].

- *Note.* The p-value for a sign test is given by

H_1	$m < m_0$	$m > m_0$	$m \neq m_0$
X_+	$\mathrm{P}(X_+ \leq x_+)$	$\mathrm{P}(X_+ \geq x_+)$	$2\min\{\mathrm{P}(X_+ \geq x_+), \mathrm{P}(X_+) \leq x_+\}$
X_-	$\mathrm{P}(X_- \geq x_-)$	$\mathrm{P}(X_- \leq x_-)$	$2\min\{\mathrm{P}(X_- \geq x_-), \mathrm{P}(X_-) \leq x_-\}$

Table 3.2: The p-value for a sign test.

Note

Sign test. Suppose we have $H_1\colon m \neq m_0$. To find the range of values of x_+ that result in the rejection of H_0, use GC to compute the following tables.

x_+	$\alpha/2 - 2\,\mathrm{P}(X_+ \leq x_+)$
$n-1$	___ > 0
n	___ > 0
$n+1$	___ < 0

x_+	$\alpha/2 - 2\,\mathrm{P}(X_+ \geq x_+)$
$m-1$	___ < 0
m	___ < 0
$m+1$	___ > 0

Then, we conclude that $x_+ \leq n$ or $x_+ \geq m$.

- A Wilcoxon matched-pairs signed rank test.
  1. Let m be the population median of $D =$ ________ $-$ ________.
  2. Test $H_0: m = 0$ against H_1: (a) $m < 0$, (b) $m \neq 0$, or (c) $m > 0$, at the $100\alpha\%$ significance level.
  3.

[label in context]	1	2	3	$\cdots$	n
D	d_1	0	d_3	$\cdots$	d_n
Rank	1		5	$\cdots$	2

Table 3.3: The value of the differences $d_1, d_2, \ldots, d_n$, which are then ranked according to their absolute size $|d_i|$. If $d_i = 0$, simply leave the corresponding cell, for rank, blank.

  4.
     - $t_- =$ ___ $+$ ___ $+ \cdots +$ ___ $=$ ___
     - $t_+ =$ ___ $+$ ___ $+ \cdots +$ ___ $=$ ___
     - The test statistic is $T := \min\{T_-, T_+\} =$ ___.
     - Reject H_0 if $T =$ ___. (see table 3.4)
  5. Since $t =$ ___ $\square$ ___, there is sufficient/insufficient evidence, at the $100\alpha\%$ significance level, to conclude that [H_1 in context].

- The test statistics T_+ and T_- can also be used, depending on our preference.
- The critical regions for a Wilcoxon test, for each alternative hypothesis and test statistic T_- or T_+. The value of c is obtained from MF26*.

  Note. the value of c may differ for a one-tail vs a two-tail test, so look at the table carefully, to obtain the correct value.

H_1	$m < m_0$	$m > m_0$	$m \neq m_0$
T_+	$T_+ \leq c$	$T_+ \geq \frac{n(n+1)}{2} - c$	$T_+ \leq c$ or $T_+ \geq \frac{n(n+1)}{2} - c$
T_-	$T_- \geq \frac{n(n+1)}{2} - c$	$T_- \leq c$	$T_- \leq c$ or $T_- \geq \frac{n(n+1)}{2} - c$
T	$T \leq c$[1]		$T \leq c$ or $T \geq \frac{n(n+1)}{2} - c$

Table 3.4: The critical regions for Wilcoxon tests.

[1]Assuming $T_- \geq T_+$ for $m < m_0$, and $T_+ \geq T_-$ for $m > m_0$.

- For large sample sizes $n \geq 21$, we use the approximation

$$T \sim \mathrm{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

  and conduct a one/two-tailed z-test.

Note

The value of n in both tests should be the total number of columns minus the number of columns with $d = 0$. i.e.

$$n \coloneqq \#\text{cols} - \#\{i \mid d_i \neq 0\}.$$

Note

If we need to use both the sign test and a Wilcoxon test on the same sample, then consider creating just a single table, as shown below.

[label in context]	1	2	3	$\cdots$	n
D	d_1	0	d_3	$\cdots$	d_n
Sign	$+$	0	$-$	$\cdots$	$+$
Rank	1		5	$\cdots$	2

Table 3.5: Combined table for both the sign test and Wilcoxon test.

Note

How do you improve the Wilcoxon test used in [the previous part]?

Increase the sample size for the test.

Note

State the circumstances under which a non-parametric test would be used rather than a parametric test.

We use a non-parametric test, rather than a parametric test, when:

1. The population is not known to be normally distributed.
2. The population mean is not the best way to measure tendency.
3. The measurement scale has no predetermined rank or ordering.

Note

Why is it not appropriate to use a paired-sample t-test?

There is no contextual evidence to support the assumption that $D_1, D_2, \ldots, D_n$ are normally distributed. So, conducting a paired-sample t-test may result in unreliable results, given our small sample size n.

Note

State the precautions that should be taken to avoid (statistical) bias.

Choose any approperate ones.

1. The test should be *'blind'*. [Testers in context] should not know which of the [two variations involved in the test, in context] they are [tasting/wearing/etc, in context]. If the [testers] knew, their preconceptions may affect ____________.
2. Pick a random sample of n [testers].
3. The *order* of the test — whether the [first variation] or [second variation] comes first — should be randomised.
4. The [testers] should not communicate with each other.
5. There should be sufficient rest time between the two runs, so that the running timing of the second run would not be affected due to fatigue.

Note

Explain why it is better to conduct a Wilcoxon test than a sign test.

While a sign test only considers the sign of the differences, a Wilcoxon test takes into account both the sign and *magnitude* of the differences. Therefore, a Wilcoxon test is more reliable, as it incorporates more information about the data.

Note

Explain why a sign test is more suitable/a Wilcoxon test is inappropriate.

Choose any approperiate ones

1. The data here is non-numeric and is not meausred on an ordinal scale. Hence, it is inapproperiate to conduct a Wilcoxon test. A sign test is better, as the data can still be represented by positie and negative responses — denoting ______ and ______, respectively.
2. The magnitude of the differences is irrelevant because ______. So, a sign test — which only accounts for the sign of the differences — is more appropriate.
3. In this case, the data has too many *tied ranks*. Thus, the conclusion obtained from a Wilcoxon test may not be reliable.