Testing goodness of fit of a distribution

Issue: Suppose we have data $X_1, \ldots, X_n$ on a **continuous** / discrete variable X. How to assess whether $X \sim F_0$, a given probability distribution?

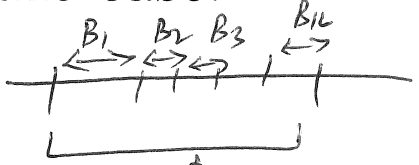
Approach 1: Perform a chi-square test.

Hypotheses: $H_0 : X \sim F_0$ versus H_1 : This is not the case.

completely known.

↓ Assumed model is correct.

Steps:



range of values of X (finite / infinite)

- Divide the support of F_0 into k bins, $B_1, \ldots, B_k$.
- $O_i = \#$ of observations that fall in B_i, $i = 1, \ldots, k$.
- $p_{i,0} = P(X \in B_i | H_0 \text{ is true})$, $i = 1, \ldots, k$. ($\sum_i p_{i,0} = 1$)
- $E_i = n p_{i,0}$
- $\chi^2 = \sum_i (O_i - E_i)^2 / E_i$
- Null distribution: Approximately χ^2_{k-1} when n is large.
- Rule of thumb: Have 5 to 8 categories, each with $E_i \geq 5$.

Modification if F_0 is not completely known: If F_0 involves **unknown** model parameters, then

- estimate them using maximum likelihood, and use the estimated values to get $\hat{p}_{i,0}$ — estimated $p_{i,0}$. $\Rightarrow \hat{E}_i = n\hat{p}_{i,0}$ ↑ estimated.
- The degrees of freedom in this case is $k - 1 -$ # parameters estimated under H_0.

$$\chi^2 = \sum_i (O_i - \hat{E}_i)^2 / \hat{E}_i \approx \chi^2_{k-1-\#\text{parameters under } H_0}.$$

Approach 2: Use a test procedure designed specifically for a given distribution

- `nortest` library in `R` for testing normality — Anderson-Darling test (`ad.test`), Shapiro-Wilk test (`shapiro.test`), etc.
- `pearson.test` to perform the above chi-square test

Approach 3: Use a Q-Q plot. In general, a quantile-quantile (Q-Q) plot is a graphical technique to determine if two datasets (x and y) come from the same distribution. It plots sample quantiles of x data on x-axis and those of y data on y-axis. If the points approximately fall along a 45^o line, then the two datasets may have the same distribution.

↑ most popular approach in practice for assessing goodness of fit.

- Use `qqplot` function in `R` for a general Q-Q plot

To see whether a dataset (x) comes from a given distribution F_0, use Q-Q plot with

- y simulated from F_0, or
- directly compute quantiles of F_0 and use them on y-axis.
- Use `qqnorm` in `R` when F_0 is normal, and get the reference line using `qqline`

Ex: Network load data in Exercise 8.2 on page 234. A network provider investigates the load of its network. The number of concurrent users is recorded at fifty locations (thousands of people). The data are stored in a column in a file called `load.txt`.

Is normality reasonable?

On page 310, the book performs a chi-square test with 6 bins and gets 1.07 as the test statistic. The p-value is `1-pchisq(1.07, 3) = 0.784`.

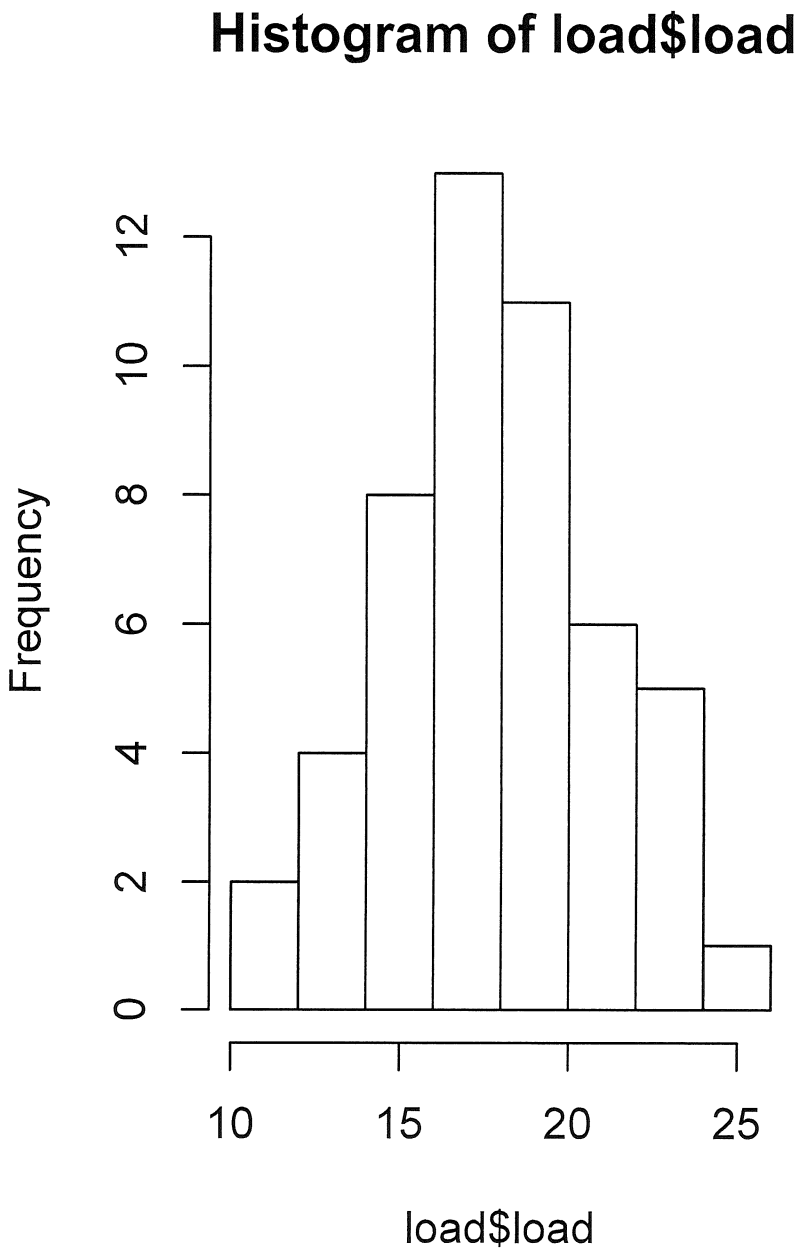
Let's use `R` to analyze these data.

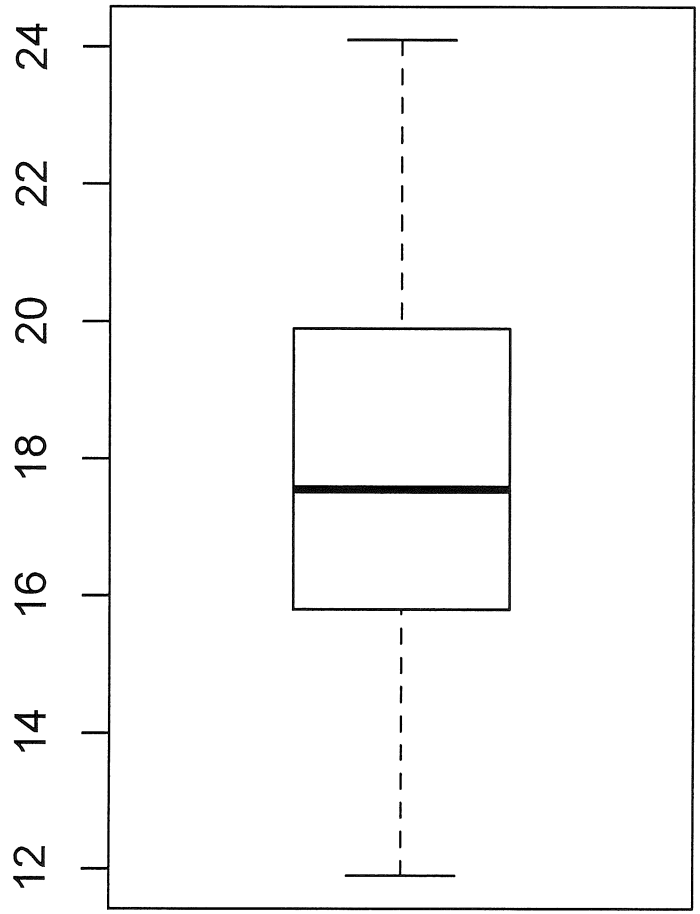
```
# Network load data from Exercise 8.2

load <- read.table(file="load.txt", header=T)

> head(load)
  load
1 17.2
2 24.1
3 13.5
4 15.4
5 19.7
6 22.1
>


# Histogram and boxplot
```

```
par(mfrow=c(1,2)) # 2 plots in 1 row

hist(load$load)
boxplot(load)

par(mfrow=c(1,1))

# Normal QQ plot

qqnorm(load$load)
qqline(load$load)

# Do these data come from a uniform distribution?

x <- load$load
y <- runif(100, min=min(x), max=max(x))

qqplot(x,y)
```

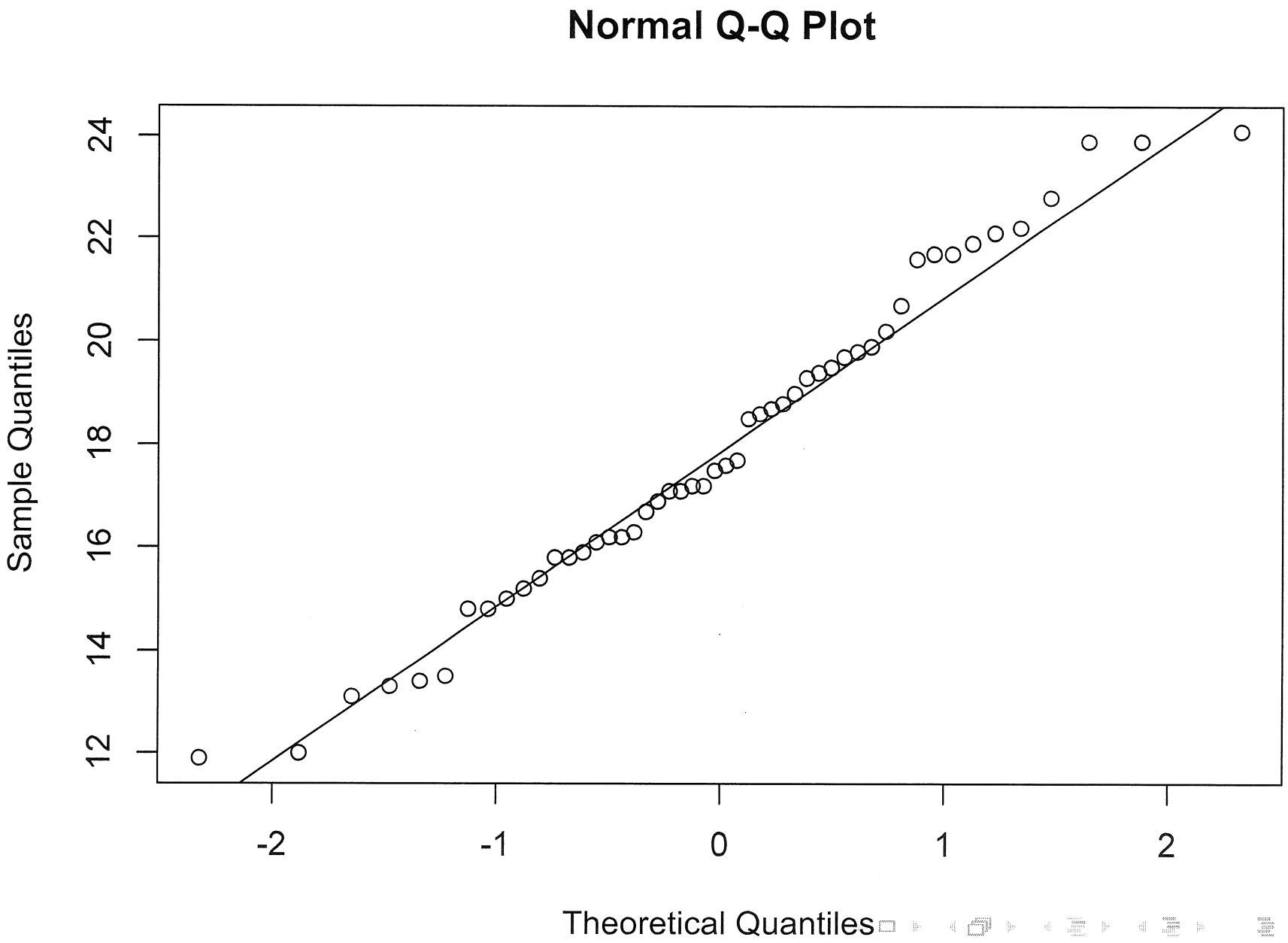
```
abline(a=0,b=1)

# Testing normality

# First download and install "nortest" packages using
# install.packages("nortest")

# Load the package in R

library(nortest)

> shapiro.test(x)

Shapiro-Wilk normality test

data:  x
W = 0.9782, p-value = 0.4787
```

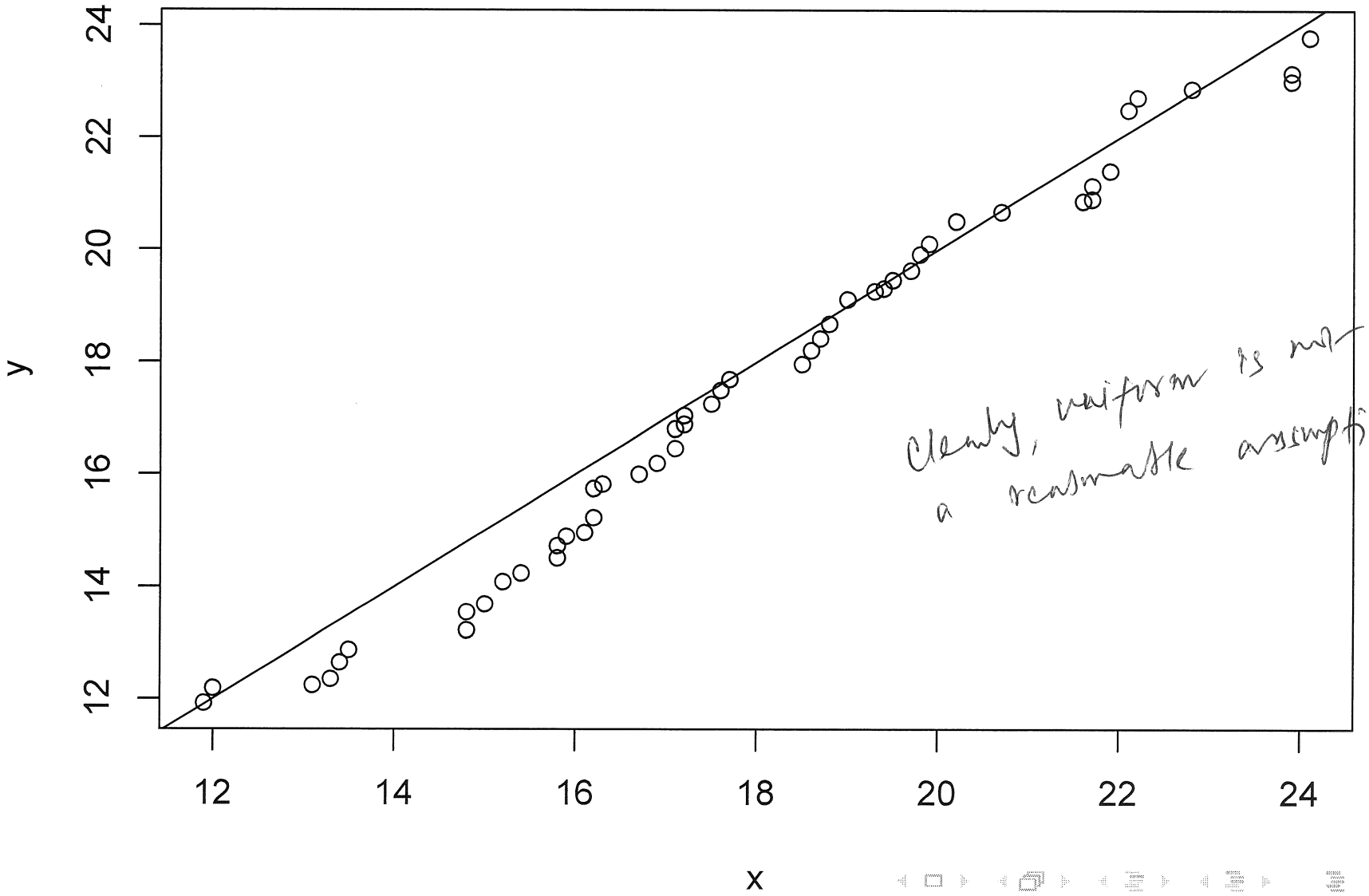
```
>

> pearson.test (x)

Pearson chi-square normality test

data:  x
P = 2.4, p-value = 0.9344

>

# The book got a  p-value of 0.78. Investigate what
# causes the difference by reading how pearson.test
# forms the bins
```


Histogram of load$load
Frequency
0
2
4
6
8
10
12
10
15
20
25
load$load


12
14
16
18
20
22
24

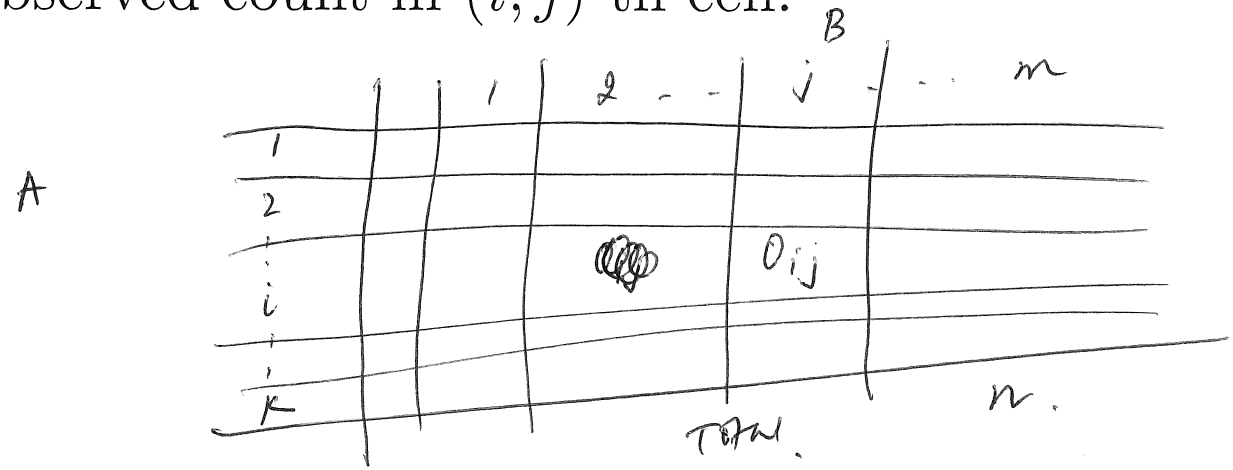
normality assumption seems reasonable.




y
12
14
16
18
20
22
24
x
12
14
16
18
20
22
24
Clearly, uniform is not
a reasonable assumption.

Chi-square test of independence

Set up: Count data on two categorical variables (or factors) A and B obtained from a sample of n subjects. Suppose the categories of A are $i = 1, \ldots, k$, and the categories of B are $j = 1, \ldots, m$. The data are arranged in a $k \times m$ table. Let O_{ij} = observed count in (i, j)-th cell.



Hypotheses: H_0 : A and B are independent (i.e., are not associated), vs., H_1 : A and B are not independent (i.e., are associated). If there is an association, the value one variable depends (at least to some extent) on the value of the other.

Example: The table below shows 695 children under 15 years of age are cross-classified according to ethnic group and hemoglobin level. Is hemoglobin level associated (related) to ethnicity?

	Hemoglobin Level (g/100 ml)				
Ethnic Group	≥ 10	9.0 - 9.9	< 9.0	Total	Proportion
A	80	100	20	200	
B	99	190	96	385	
C	70	30	10	110	
Total	249	320	126	695	
Proportion					

- If He level is not associated to ethnicity, then the proportion of subjects in population that fall a He group does not depend on ethnicity, i.e., it is the same for each ethnicity group, and vice versa.

To do a chi-square test, we need the expected counts E_{ij} assuming that H_0 is true. Let X and Y indicate respective categories of A and B in which a randomly selected subject from the population falls. When A and B are independent,

$$P(X = i, Y = j) = P(X = i)P(Y = j) \text{ for all } i, j.$$

- $P(X = i)$ is estimated as
- $P(Y = j)$ is estimated as
- Assuming independence, $P(X = i, Y = j)$ is estimated as
- Assuming independence, E_{ij} is estimated as

Test statistic:

Degrees of freedom: