DATA2002

Chi-squared tests

Garth Tarr

THE UNIVERSITY OF
SYDNEY

Hypothesis testing

Chi-squared test for categorical data

Genetic linkage

Genetic linkage

In a backcross[1] experiment to investigate the **genetic linkage** between two genes A and B in a species of flower. Some researchers classified 400 offspring by phenotype as follows:

AB	Ab	aB	ab
128	86	74	112

- A might be pink flowers and a might be yellow flowers
- B might be smooth leaves and b might be wrinkled leaves

1. Under the *no linkage* model (🔗), the four phenotypes are equally likely.
2. If linkage is in the *coupling phase* (🔗), the probabilities of the four phenotypes are given by

AB	Ab	aB	ab
$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

   where p is the *recombination fraction* and is estimated by the overall proportion of Ab and aB.

1. **Backcrossing** is a crossing of a hybrid with one of its parents or an individual genetically similar to its parent, in order to achieve offspring with a genetic identity which is closer to that of the parent. For a more detailed discussion see here.

No linkage model

- **Null hypothesis:** each of the phenotypes are equally likely.
- **Alternative hypothesis:** the phenotypes are not equally likely.

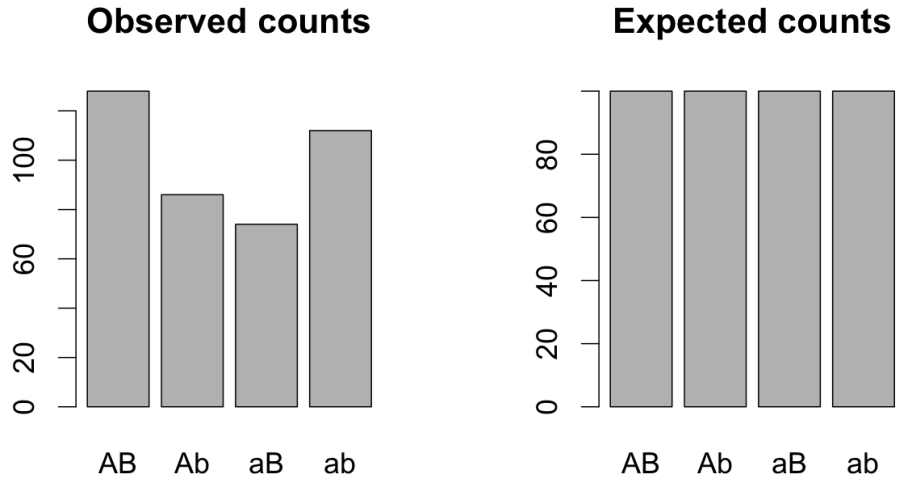
Let p_i be the probability of being in the ith phenotype $i = AB, Ab, aB, ab$.

Under the null hypothesis $p_i = 0.25$ for all i.

```
# observed counts
y = c(128, 86, 74, 112)
n = sum(y)
# hypothesised proportions
p = c(1/4, 1/4, 1/4, 1/4)
# expected counts
e = p*n
```

What does this look like visually?

```
names = c("AB", "Ab", "aB", "ab")
par(mfrow = c(1, 2), cex = 1.5) # set up a grap
barplot(y, names.arg = names,
        main = "Observed counts")
barplot(e, names.arg = names,
        main = "Expected counts")
```

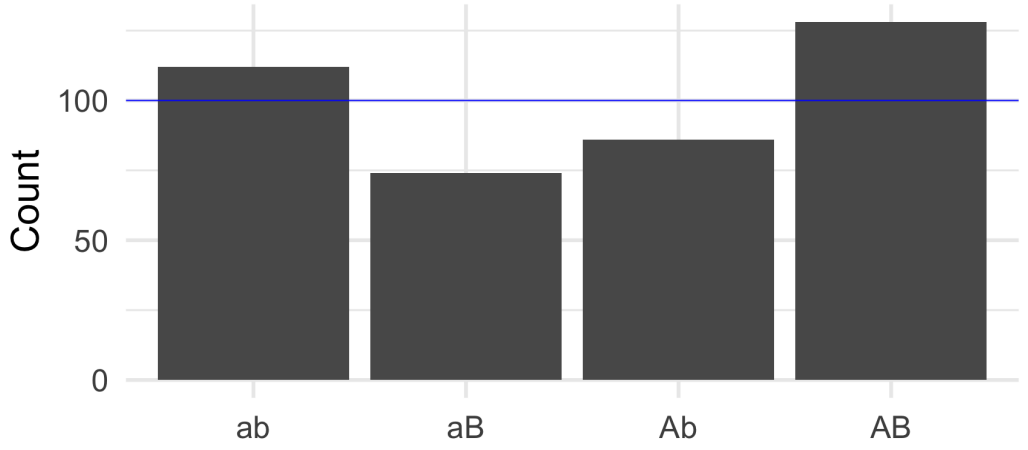


These are base graphics, not **ggplots**, but they do the job, even though they're not pretty.

No linkage model

Is this a good fit?

```
library(tidyverse)
df = tibble(names,y,p,e)
df %>% ggplot() +
  aes(x = names, y = y) +
  geom_bar(stat="identity") +
  geom_hline(yintercept = 100, colour = "blue") +
  theme_minimal(base_size = 32) +
  labs(x = "", y = "Count")
```



Consider the differences between observed counts and expected counts:

```
d = y-e
d
```

```
## [1]  28 -14 -26  12
```

```
mean(d)
```

```
## [1] 0
```

Test statistic

Considering the average of the differences doesn't tell us much.

Let's take the squared differences, and "normalise" by dividing by the expected cell counts:

$$t_0 = \sum_{i=1}^{k} \frac{(y_i - e_i)^2}{e_i}$$

where k is the number of categories (groups).

```
(y-e)^2
```

```
## [1] 784 196 676 144
```

```
(y-e)^2/e
```

```
## [1] 7.84 1.96 6.76 1.44
```

```
t0 = sum((y-e)^2/e)
t0
```

```
## [1] 18
```

Is this evidence for or against the null hypothesis?

Simulate

Under the null hypothesis, the counts are *uniformly* distributed across the 4 categories.

Fixing the sample size at $n = 400$ we can **simulate** data assuming the null hypothesis is true.

```
n = 400
names
```

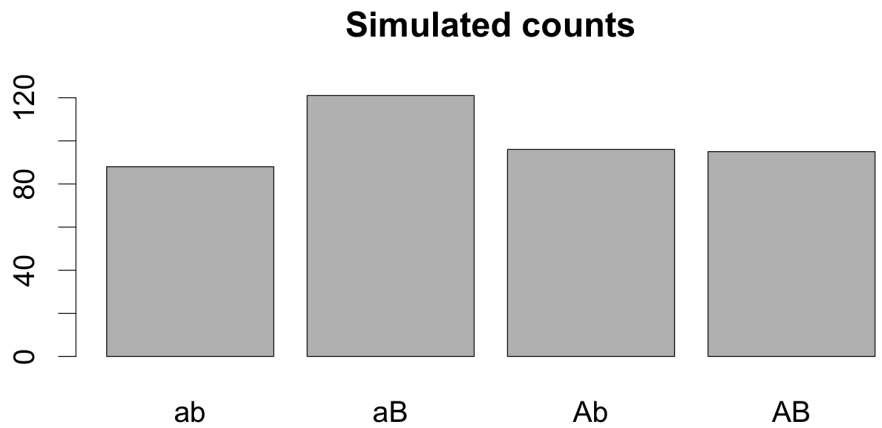
```
## [1] "AB" "Ab" "aB" "ab"
```

```
set.seed(1)
sim1 = sample(x = names,
              size = n,
              replace = TRUE,
              prob = c(0.25,0.25,0.25,0.25))
```

```
table(sim1)
```

```
## sim1
##  ab  aB  Ab  AB
##  88 121  96  95
```

```
par(cex=2)
barplot(table(sim1), main = "Simulated counts")
```



Simulate

Our test statistic for that simulated sample is:

```
sim_y = table(sim1)
sum((sim_y - e)^2/e)
```

```
## [1] 6.26
```

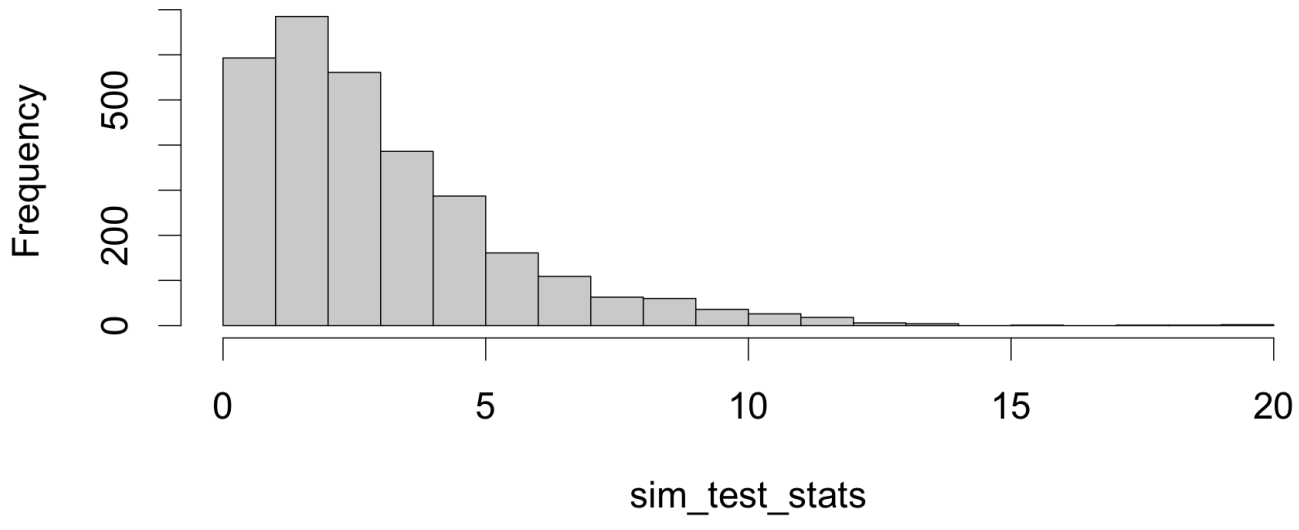
which is a lot smaller than what we observed on our **actual** data:

```
t0
```

```
## [1] 18
```

But let's do this a lot of times rather than just once.

```
B = 3000
sim_test_stats = vector(mode = "numeric", length = B)
for(i in 1:B){
  sim = sample(x = names,  size = n, replace = TRUE,
               prob = c(0.25,0.25,0.25,0.25))
  sim_y = table(sim)
  sim_test_stats[i] = sum((sim_y - e)^2/e)
}
par(cex = 2, mar = c(4,4,0.5,0.5))
hist(sim_test_stats, main = "", breaks = 20)
```



Simulate

- Now we have a pretty good idea about the shape of the **distribution** of the test statistic when the null hypothesis is true.
- We can compare the test statistic that we calculated on the original data to the "null distribution".
- One way to do this is to ask the question:

Given that the null hypothesis is true, how likely is it that we observe a test statistic as or more extreme than that we calculated from our original sample.

```
mean(sim_test_stats >= t0)
```

```
## [1] 0.001
```

In 0.1% of samples when the null hypothesis is true, we got a simulated sample that was "more extreme" than our original sample.

What does this tell us about the agreement between the null hypothesis and our sample of data?

Is there way to do it without simulation?

Yes! A χ^2 test!

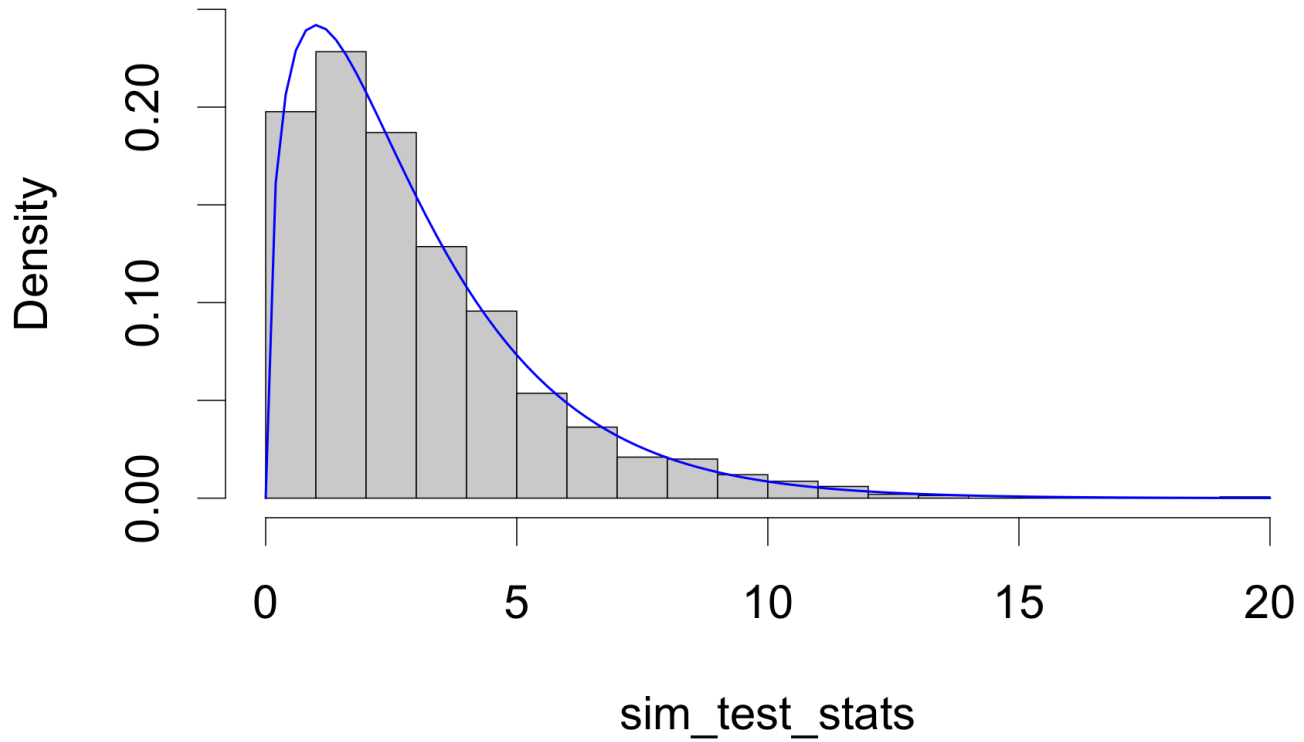
It is known that our test statistic,

$$T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i} \sim \chi^2_{k-1},$$

approximately, where k is the number of groups.

Let's compare this distribution to the simulated test statistic distribution.

```
par(cex = 2.5, mar = c(4,4,0.5,0.5))
hist(sim_test_stats, main = "", breaks = 20,
     probability = TRUE, ylim = c(0, 0.25))
curve(dchisq(x, df = 3), add = TRUE, col = "blue", lwd = 2)
```



A χ^2 test!

- The degrees of freedom from the sample is $k-1$ because the first $k-1$ observations y_i contain all the information and the last observation is fixed by $y_k = n - \sum_{i=1}^{k-1} y_i$ adding no extra information.
- In general, the test statistic $T \sim \chi^2_{k-1-q}$ where q is the number of parameters need to be estimated from the sample. In the no linkage example, $q = 0$ as we do not need to estimate any parameters.
- The approximation will only be accurate if *no expected frequency* is too small, as a rule of thumb we require all $e_i \geq 5$. Otherwise, we need to pool adjacent categories so that the expected frequencies are always ≥ 5.

Workflow: Chi-squared goodness of fit test

- one categorical variable from a single population
- want to see if it follows a hypothesised distribution

- **Hypothesis:** H_0: $p_1 = p_{10},\ p_2 = p_{20}, \cdots,\ p_k = p_{k0}$ vs H_1: at least one equality does not hold.
- **Assumptions:** independent observations and $e_i = np_{i0} \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i}$. Under H_0, $T \sim \chi^2_{k-1-q}$ approximately where k is the number of groups and q is the number of parameters that needs to be estimated from the data.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{k} \frac{(y_i - e_i)^2}{e_i}$.
- **P-value:** $P(T \geq t_0) = P(\chi^2_{k-1-q} \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$, otherwise do not reject H_0.

Table for calculating the test statistic

If you were doing this manually, the calculations can be summarised in the following table:

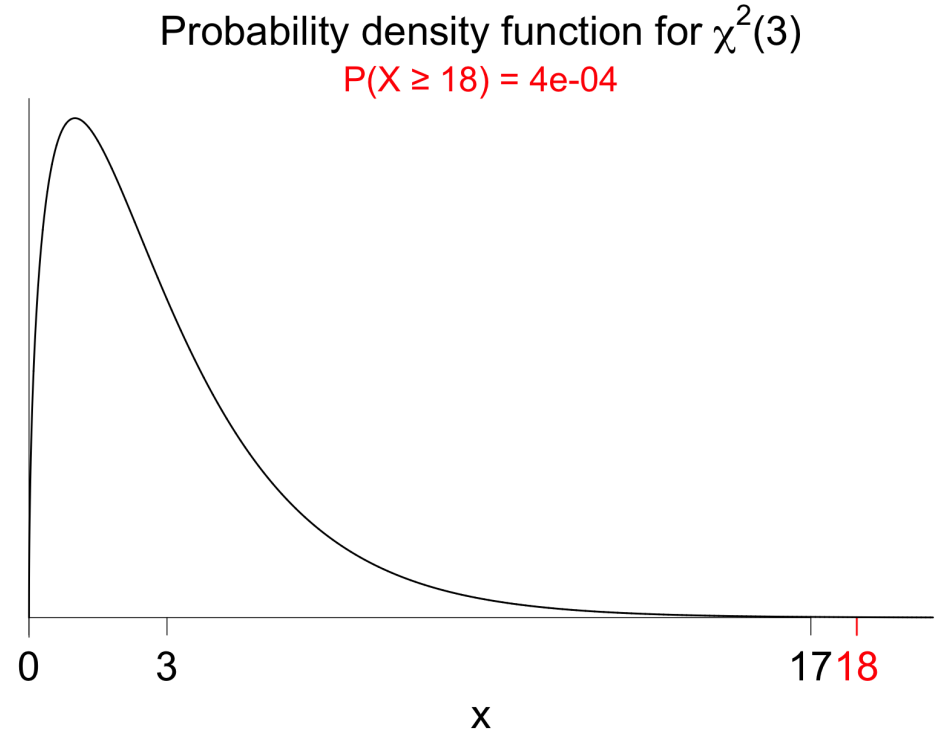
Group i	y_i	p_{i0}	$e_i = np_{i0}$	$y_i - e_i$	$\frac{(y_i - e_i)^2}{e_i}$
1	y_1	p_{10}	np_{10}	$y_1 - e_1$	$(y_1 - e_1)^2/e_1$
2	y_2	p_{20}	np_{20}	$y_2 - e_2$	$(y_2 - e_2)^2/e_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	y_k	p_{k0}	np_{k0}	$y_k - e_k$	$(y_k - e_k)^2/e_k$
Sum	n	1	n	0	t_0

No linkage model

Under the *no linkage* model, we complete the following table:

Type	y_i	$e_i = np_{i0}$	$y_i - e_i$	$\frac{(y_i - e_i)^2}{e_i}$
AB	128	$400 \times \frac{1}{4} = 100$	$128 - 100 = 28$	$\frac{(28)^2}{100} = 7.84$
Ab	86	$400 \times \frac{1}{4} = 100$	$86 - 100 = -14$	$\frac{(-14)^2}{100} = 1.96$
aB	74	$400 \times \frac{1}{4} = 100$	$74 - 100 = -26$	$\frac{(-26)^2}{100} = 6.76$
ab	112	$400 \times \frac{1}{4} = 100$	$112 - 100 = 12$	$\frac{(12)^2}{100} = 1.44$
Total	400	400	0	$t_0 = 18.00$

- **Hypothesis:**
  H_0: $p_{AB} = p_{Ab} = p_{aB} = p_{ab} = \frac{1}{4}$ vs H_1: at least one equality does not hold.
- **Assumptions:** independent observations and $e_i = np_{i0} \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i}$. Under H_0, $T \sim \chi^2_3$ approx.
- **Observed test statistic:** $t_0 = 18$
- **P-value:**
  $P(T \geq t_0) = P(\chi^2_3 \geq 18) = 0.0004$
- **Decision:** Since the p-value is < 0.05, there is strong evidence in the data against H_0. Hence the four phenotypes are not equally likely.




```
1 - pchisq(18, df = 3)
```

```
## [1] 0.0004398497
```

No linkage model

```
y
```

```
## [1] 128  86  74 112
```

```
p
```

```
## [1] 0.25 0.25 0.25 0.25
```

```
(ey = n * p)  # expected counts
```

```
## [1] 100 100 100 100
```

```
ey >= 5  # test e_i >= 5
```

```
## [1] TRUE TRUE TRUE TRUE
```

```
chisq.test(y, p = p)
```

```
##
##     Chi-squared test for given probabilities
##
## data:  y
## X-squared = 18, df = 3, p-value = 0.0004398
```

Linkage model

AB	Ab	aB	ab
128	86	74	112

Under the *coupling phase* linkage model, the probabilities of each of the four phenotype outcomes are given by

AB	Ab	aB	ab
$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

We can estimate the parameter p, the recombination fraction, as the proportion of observed offspring in categories Ab or aB,

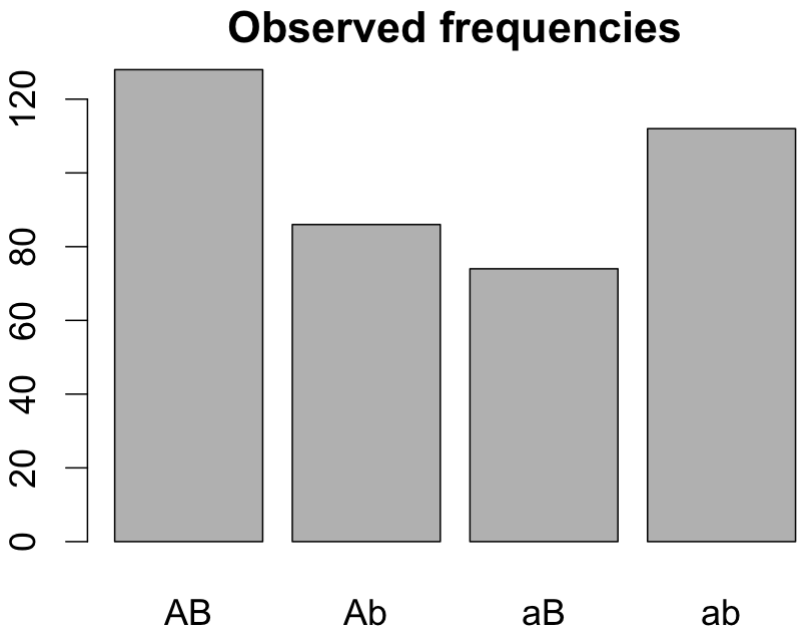
$$\hat{p} = \frac{86 + 74}{400} = 0.4.$$

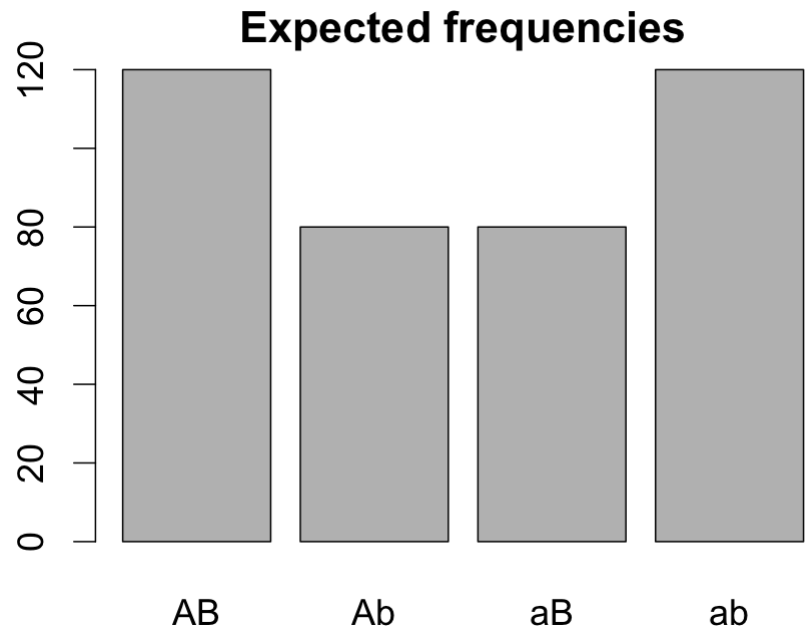
Hence the four (estimated) hypothesised probabilities are,

$$p_{10} = \frac{1}{2}(1-0.4) = 0.3,\ p_{20} = \frac{1}{2}0.4 = 0.2,\ p_{30} = \frac{1}{2}0.4 = 0.2 \text{ and } p_{40} = \frac{1}{2}(1-0.4) = 0.3.$$

Linkage model

```
y = c(128, 86, 74, 112)
p = c(0.3, 0.2, 0.2, 0.3)
names = c("AB", "Ab", "aB", "ab")
# set up a graphics device with 1 row and 2 columns
par(mfrow = c(1, 2), cex = 1.5, mar = c(2,2,2,1))
barplot(y, names.arg = names, main = 'Observed frequencies')
barplot(p * sum(y), names.arg = names, main = 'Expected frequencies')
```

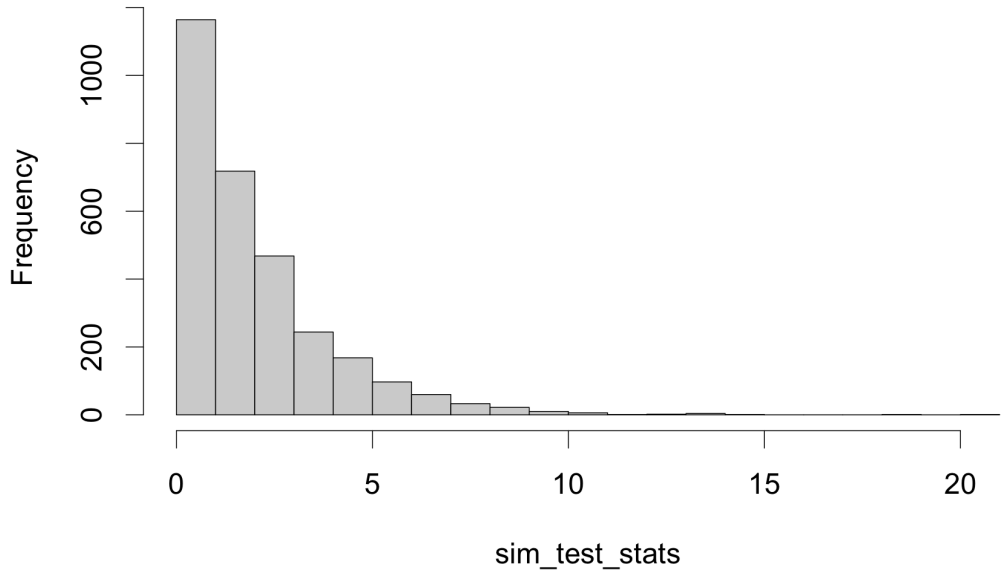




Linkage model simulation

```
n = 400
hyp_probs = c(0.3, 0.2, 0.2, 0.3)
B = 3000
sim_test_stats = vector(mode = "numeric",
                        length = B)
for(i in 1:B){
  sim = sample(x = names,
               size = n,
               replace = TRUE,
               prob = hyp_probs)
  sim_y = table(sim)
  # estimated probability
  p_e = sum(table(sim)[2:3])/n
  # expected values using estimated probabilit
  e = 400*c(1 - p_e, p_e, p_e, 1 - p_e)/2
  sim_test_stats[i] = sum((sim_y - e)^2/e)
}
```

```
par(cex = 2, mar = c(4,4,0.5,0.5))
hist(sim_test_stats, main = "", breaks = 20)
```



Calculate observed test statistic

Type	y_i	$e_i = np_{i0}$	$y_i - e_i$	$\frac{(y_i - e_i)^2}{e_i}$
AB	128	$400 \times \frac{3}{10} = 120$	$128 - 120 = 8$	$\frac{(8)^2}{120} = 0.53$
Ab	86	$400 \times \frac{2}{10} = 80$	$86 - 80 = 6$	$\frac{(6)^2}{80} = 0.45$
aB	74	$400 \times \frac{2}{10} = 80$	$74 - 80 = -6$	$\frac{(-6)^2}{80} = 0.45$
ab	112	$400 \times \frac{3}{10} = 120$	$112 - 120 = -8$	$\frac{(-8)^2}{120} = 0.53$
Total	400	400	0	$t_0 = 1.96$

Calculate observed test statistic

```
n = 400
hyp_probs = c(0.3, 0.2, 0.2, 0.3)
expected_counts  = hyp_probs * n
t0 = sum((y-expected_counts)^2/expected_counts)
t0
```

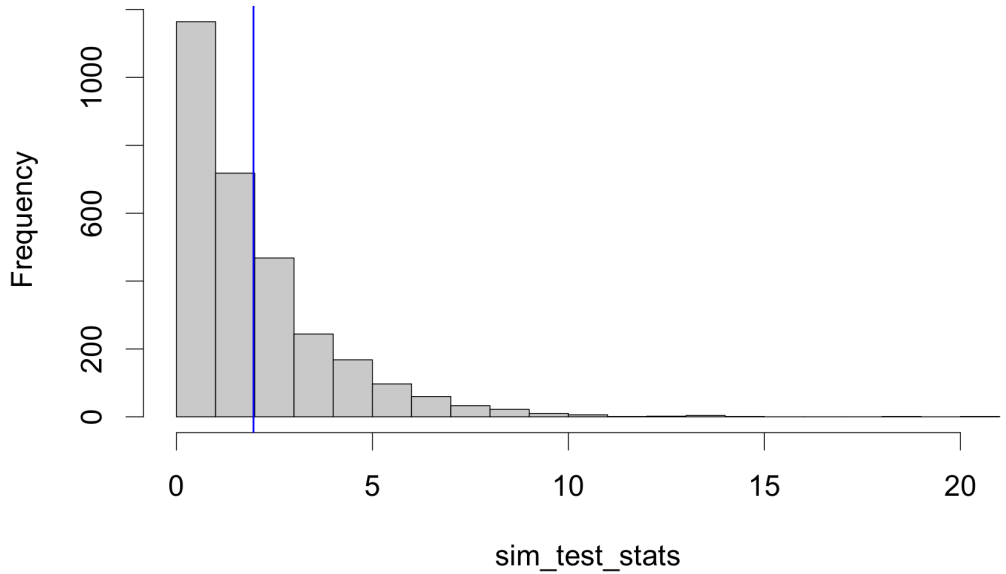
```
## [1] 1.966667
```

Let's compare this with the distribution of test statistics that we simulated assuming the null hypothesis is true.

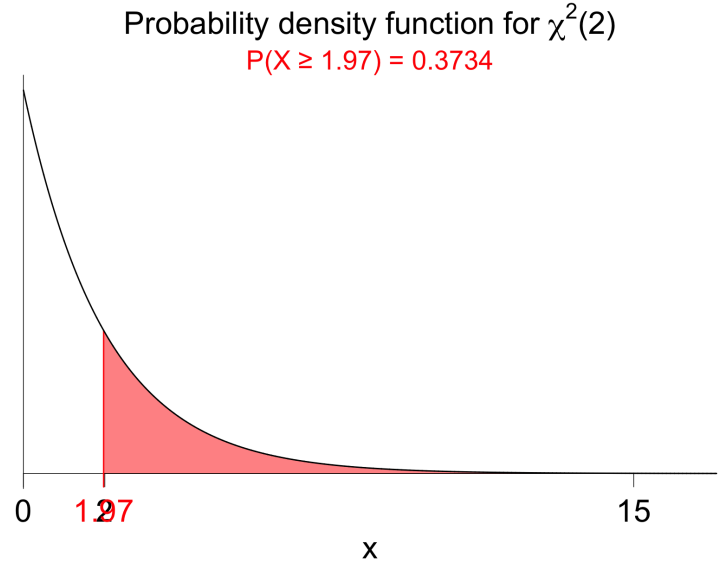
```
mean(sim_test_stats >= t0)
```

```
## [1] 0.382
```

```
par(cex = 2, mar = c(4,4,0.5,0.5))
hist(sim_test_stats, main = "", breaks = 20)
abline(v = t0, col = "blue", lwd = 2)
```



- **Hypothesis:** The null hypothesis is a coupling phase linkage model, H_0: $p_{AB} = p_{ab} = (1-p)/2$ and $p_{Ab} = p_{aB} = p/2$. The alternative hypothesis is that the proportions do not follow a coupling phase linkage model, i.e. H_1: at least one equality does not hold.
- **Assumptions:** independent observations and expected cell counts at least 5, $e_i = np_{i0} \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{k} \frac{(Y_i - np_{i0})^2}{np_{i0}}$. Under H_0, $T \sim \chi^2_2$ approximately. We have estimated one parameter, $\hat{p}$, so the degrees of freedom are $4 - 1 - 1 = 2$.
- **P-value:** $P(T \geq t_0) = P(\chi^2_2 \geq 1.97) = 0.37$
- **Decision:** Since the p-value is much larger than 0.05, the data are consistent with the "coupling phase" linkage model.



Defining parameters:

- p_{AB} is the probability of an offspring having phenotype AB,
- p_{ab} is the probability of an offspring having phenotype ab,

and similarly for p_{Ab} and p_{aB}.

In R

```
chisq.test(y, p = p)  # Note the incorrect degrees of freedom!
```

```
##
##     Chi-squared test for given probabilities
##
## data:  y
## X-squared = 1.9667, df = 3, p-value = 0.5794
```

```
n = sum(y)
k = length(y)
(ey = n*p)
```

```
## [1] 120  80  80 120
```

```
ey >= 5 # check e_i >= 5
```

```
## [1] TRUE TRUE TRUE TRUE
```

```
(t0 = sum((y - ey)^2/ey))
```

```
## [1] 1.967
```

```
(pval = 1 - pchisq(t0, k - 1 - 1))
```

```
## [1] 0.3741
```

Recap

Hypothesis

- The statement against which you search for evidence is called the null hypothesis, and is denoted by H_0. It is generally a "no difference" statement.
- The statement you claim is called the alternative hypothesis, and is denoted by H_1 (or sometimes you'll see H_A)

Assumptions

- Each observation are generally assumed to have been chosen at random from a population.
- We say that such random variables are *iid* (independently and identically distributed).
- Each test we consider will have its own set of assumptions.

Recap

Test statistic

- Since observations vary from sample to sample we can never be sure whether H_0 is true or not.
- A test statistic is a function of the observations, $T = f(X_1, \ldots, X_n)$, such that the distribution of T is known assuming H_0 is true. It can be used test if the data are consistent with H_0.
- The **observed test statistic**, t_0, is where we plug our observed data into the formula for the test statistic.
- Large (positive or negative depending on H_1) observed test statistic values is taken as evidence of poor agreement with H_0.

Significance

The p-value is defined as the probability of getting a test statistic, T, *as or more extreme* than the value we observed, t_0, *assuming* that H_0 is true.

Recap

Decision

An observed *large* positive or negative value of t_0 and hence small p-value is taken as evidence of poor agreement with H_0.

- If the p-value is small, then either H_0 is true and the poor agreement is due to an unlikely event, or H_0 is false. **The smaller the p-value, the stronger the evidence against the null hypothesis.**
- **A large p-value does not mean that there is evidence that the null hypothesis is true.**
- The level of significance, α, is the strength of evidence needed to reject H_0 (often $\alpha = 0.05$).

R packages and functions

- `chisq.test()` chi-squared test given probabilities
- `pchisq()` probability of getting outcomes from a chi-squared distribution
- `length()` number of elements in a vector
- `sum()` add elements in a vector
- `barplot()` for creating bar plots in base graphics

References

For further details see Larsen and Marx (2012), sections 10.3 and 10.4.

Griffiths A. J. F., J. H. Miller, D.T. Suzuki, et al. (2000). *An Introduction to Genetic Analysis.* 7th ed. New York: W. H. Freeman. Chi-square test for linkage. https://www.ncbi.nlm.nih.gov/books/NBK22084/

Larsen, R. J. and M. L. Marx (2012). *An Introduction to Mathematical Statistics and its Applications*. 5th ed. Boston, MA: Prentice Hall. ISBN: 978-0-321-69394-5.