DATA2002

Chi-squared tests

Garth Tarr
The University of Sydney

THE UNIVERSITY OF SYDNEY

In this lecture

- Hypothesis testing (recap)
- No linkage model
- Linkage model

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. Researchers classified 400 offspring by phenotype:

Phenotype	AB	Ab	aB	ab
Count	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
  - AB means a plant with pink flowers and smooth leaves
  - Ab means a plant with pink flowers and wrinkled leaves
  - aB means a plant with yellow flowers and smooth leaves
  - ab means a plant with yellow flowers and wrinkled leaves

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

Genetic linkage

Under the *no linkage* model (), the four phenotypes are equally likely. So we would expect to see equal proportions for each phenotype:

Phenotype	AB	Ab	aB	ab
Expected proportion	0.25	0.25	0.25	0.25

If linkage is in the *coupling phase* (), the probabilities of the four phenotypes are given by

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

where p is the **recombination fraction**. We will need to estimate p as the overall proportion of observed Ab and aB.

Hypothesis testing (recap)

Hypothesis

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

Assumptions

- Observation are generally assumed to have been chosen at random from a population and so they are *iid* (independently and identically distributed).
- Each test we consider will have its own set of assumptions.

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 to 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.

Decision

An observed *large* positive or negative test statistic 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 $\alpha = 0.05$).

No linkage model

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 i th phenotype $i = AB, Ab, aB, ab$.

Under the null hypothesis (i.e. assuming that the no linkage model is correct) the counts are **uniformly distributed** across the 4 categories,

$$p_i = 0.25 \text{ for all } i.$$

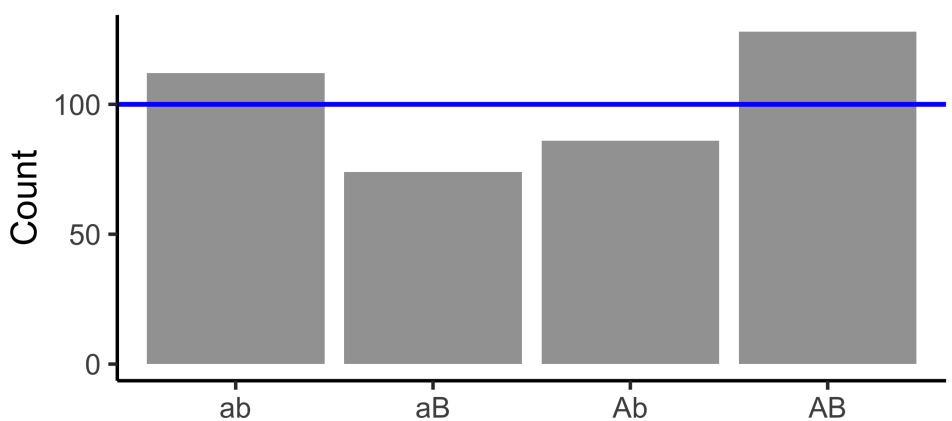
```
library(tidyverse)
df = tibble(
  phenotype = c("AB","Ab","aB","ab"),
  # observed counts
  y = c(128, 86, 74, 112),
  # hypothesised proportions
  p = c(1/4, 1/4, 1/4, 1/4),
  # expected counts
  e = sum(y) * p
)
df
```

```
# A tibble: 4 × 4
  phenotype     y     p     e
  <chr>     <dbl> <dbl> <dbl>
1 AB          128  0.25   100
2 Ab           86  0.25   100
3 aB           74  0.25   100
4 ab          112  0.25   100
```

No linkage model

> Is the no linkage model a good fit for the observed data?

```
df |> ggplot() +
  aes(x = phenotype, y = y) +
  geom_col(alpha = 0.6) +
  geom_hline(yintercept = 100,
             colour = "blue",
             size = 2) +
  labs(x = "", y = "Count")
```



Differences between observed counts and expected counts:

```
df = df |>
  mutate(d = y - e)
df
```

```
# A tibble: 4 × 5
  phenotype     y     p     e     d
  <chr>     <dbl> <dbl> <dbl> <dbl>
1 AB          128  0.25   100    28
2 Ab           86  0.25   100   -14
3 aB           74  0.25   100   -26
4 ab          112  0.25   100    12
```

```
df |>
  summarise(avg_diff = mean(d))
```

```
# A tibble: 1 × 1
  avg_diff
     <dbl>
1        0
```

Test statistic

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).

```
df = df |> mutate(
  squared_discrepency = (y-e)^2,
  contribution = (y-e)^2/e
)
t0 = sum(df$contribution)
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.

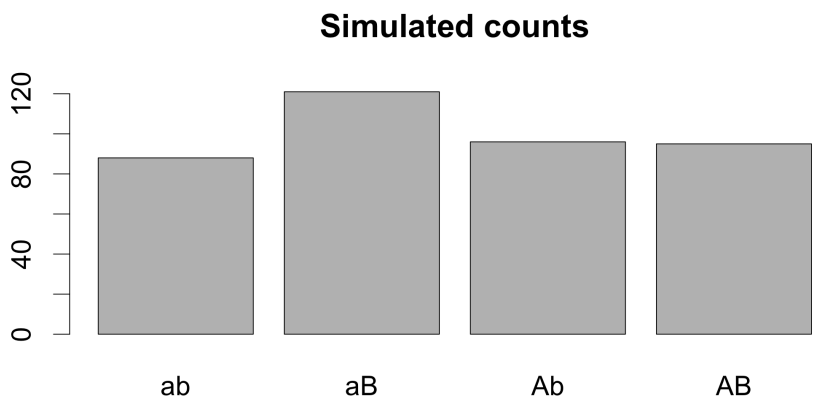
Fixing the sample size at $n = 400$ we can **simulate** data assuming the null hypothesis is true.

```
n = 400
phenotype = c("AB", "Ab", "aB", "ab")
no_link_p = c(1, 1, 1, 1)/4
e = n * no_link_p
set.seed(1)
sim1 = sample(
  x = c("AB", "Ab", "aB", "ab"),
  size = n,
  replace = TRUE,
  prob = no_link_p
)
```

```
table(sim1)
```

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

```
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
```

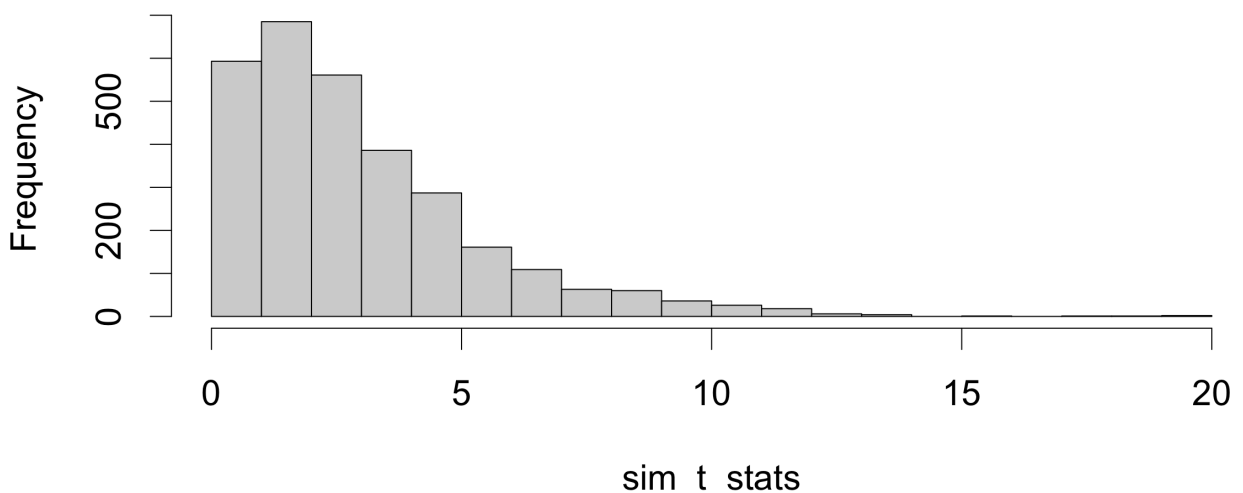
which is a lot smaller than what we observed on our **actual** data:

```
t0
```

```
[1] 18
```

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

```
B = 3000
sim_t_stats = vector(mode = "numeric", length = B)
for(i in 1:B){
  sim = sample(x = phenotype, size = n,
               replace = TRUE, prob = no_link_p)
  sim_y = table(sim)
  sim_t_stats[i] = sum((sim_y - e)^2/e)
}
hist(sim_t_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?

```
# sum(sim_t_stats >= t0)/B
mean(sim_t_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!

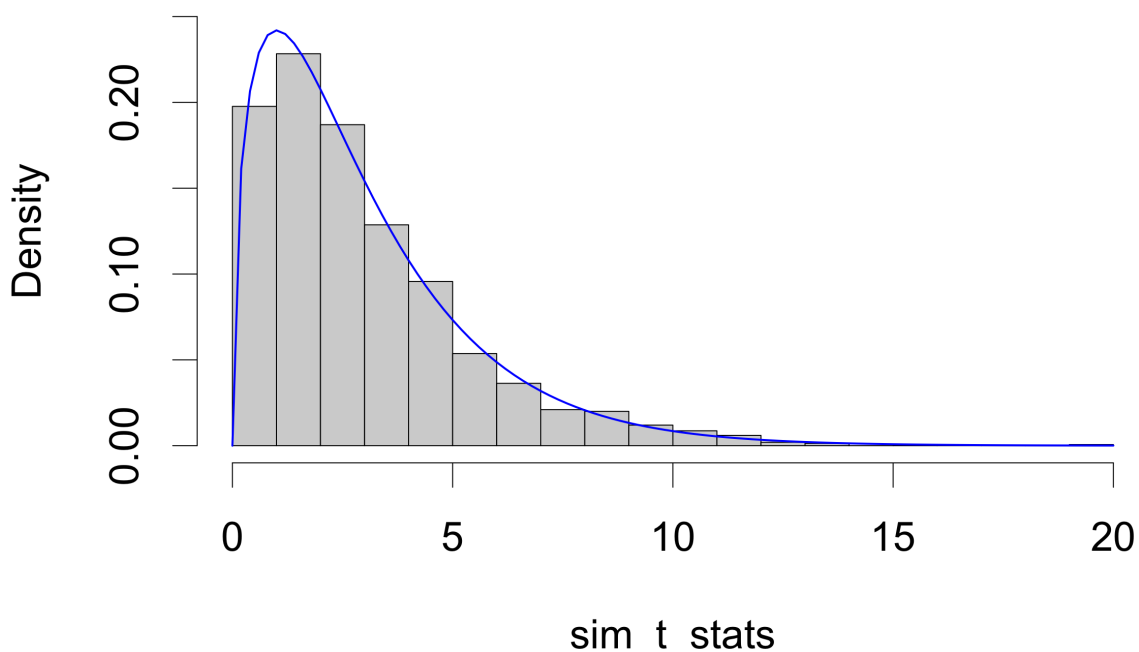
In this example the test statistic,

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

approximately, where k is the number of groups.

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

```r
hist(sim_t_stats, main = "", breaks = 20,
     probability = TRUE, ylim = c(0, 0.25))
curve(dchisq(x, df = 3), add = TRUE,
      col = "blue", lwd = 2)
```



The symbol, χ, is the Greek letter chi. Pronunciation: . Read χ^2_a as: chi-squared distribution with a degrees of freedom.

χ^2 test degrees of freedom

- The **degrees of freedom** 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 takes the form,

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

  where q is the number of parameters that needs to be estimated from the sample.
- In the no linkage example, $q = 0$ because we do not need to estimate any parameters (i.e. we don't need to estimate the hypothesised proportions, all $p_{i0} = 0.25$).
- 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 and want to see if it follows a hypothesised distribution.

- **Hypothesis:** H_0: the proportions in each category (p_i) are equal to the corresponding hypothesised proportions (p_{i0}), i.e. $p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$ vs H_1: at least proportion is not equal to the hypothesised proportion, i.e. 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

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

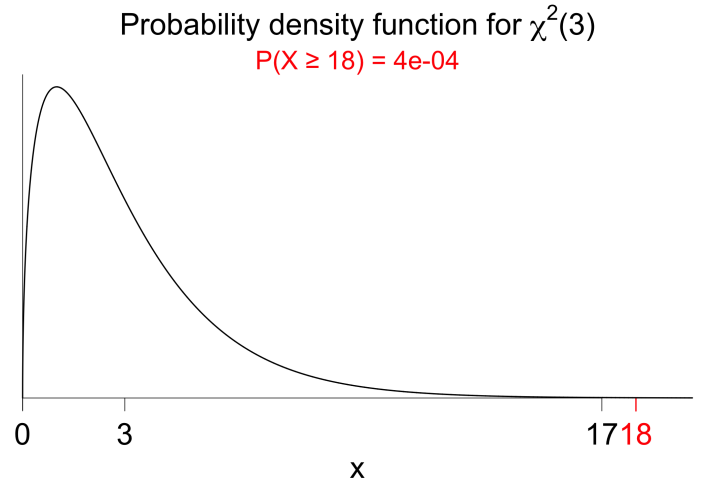
- y_i are the observed counts
- p_{i0} are the hypothesised probabilities
- e_i are the expected counts *assuming the null hypothesis is true*
- t_0 is the observed test statistic

No linkage model

Under the *no linkage* model we assume that the observations are uniformly distributed across the four categories, i.e. the hypothesised probabilities are $p_{i0} = \frac{1}{4}$ for $i = 1, 2, 3, 4$:

Type	i	y_i	$e_i = np_{i0}$	$y_i - e_i$	$\frac{(y_i - e_i)^2}{e_i}$
AB	1	128	$400 \times \frac{1}{4} = 100$	$128 - 100 = 28$	$\frac{(28)^2}{100} = 7.84$
Ab	2	86	$400 \times \frac{1}{4} = 100$	$86 - 100 = -14$	$\frac{(-14)^2}{100} = 1.96$
aB	3	74	$400 \times \frac{1}{4} = 100$	$74 - 100 = -26$	$\frac{(-26)^2}{100} = 6.76$
ab	4	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:** The null hypothesis is a no linkage model with uniform probabilities, H_0: $p_{AB} = p_{Ab} = p_{aB} = p_{ab} = \frac{1}{4}$. The alternative is anything other than a no linkage model, 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:** $T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i}$. 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 much smaller than 0.05, there is strong evidence in the data against H_0. Hence the four phenotypes are not equally likely and we reject the no linkage model.




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

```
[1] 0.0004398497
```

No linkage model

```
y = df$y
y
```

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

```
no_link_p
```

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

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

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

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

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

```
all(ey >= 5)
```

```
[1] TRUE
```

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

```
[1] 18
```

```
1 - pchisq(t0, df = 3)
```

```
[1] 0.0004398497
```

The `chisq.test()` function in R can do it all for us. We give it the vector of observed counts and the vector of hypothesised probabilities:

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

```
    Chi-squared test for given probabilities

data:  y
X-squared = 18, df = 3, p-value = 0.0004398
```

Linkage model

Linkage model

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

Phenotype	AB	Ab	aB	ab
Observed count	128	86	74	112
Hypothesised proportion	$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

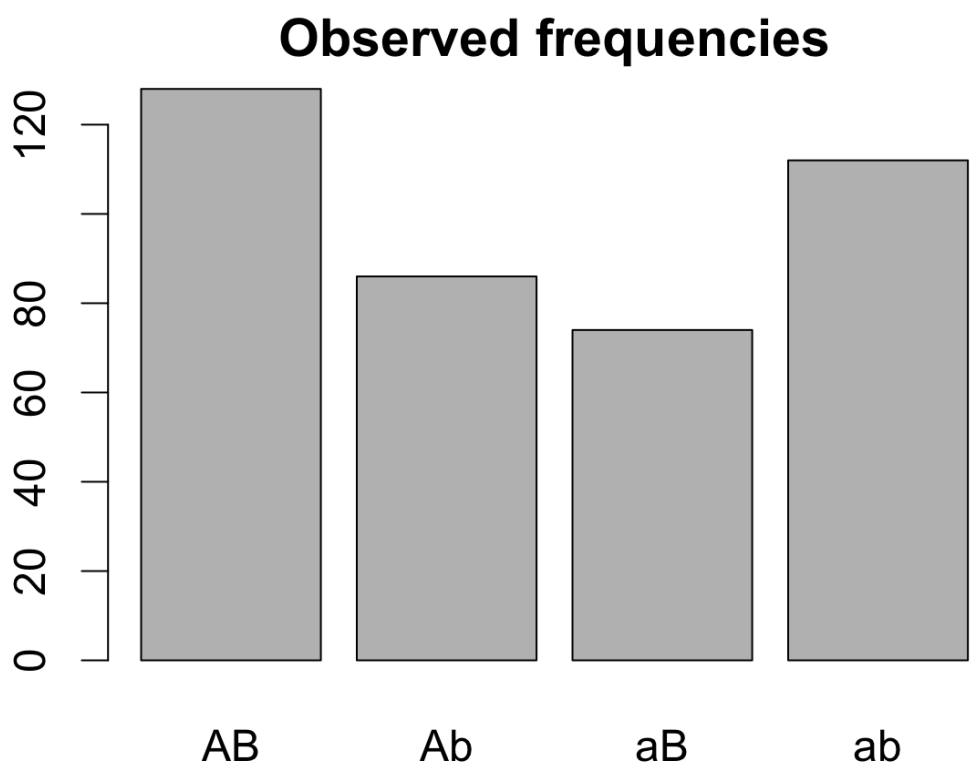
To test the linkage model hypothesis, we need to estimate the parameter p, the recombination fraction from the observed data: $\hat{p}$ is the proportion of observed offspring with phenotype Ab or aB,

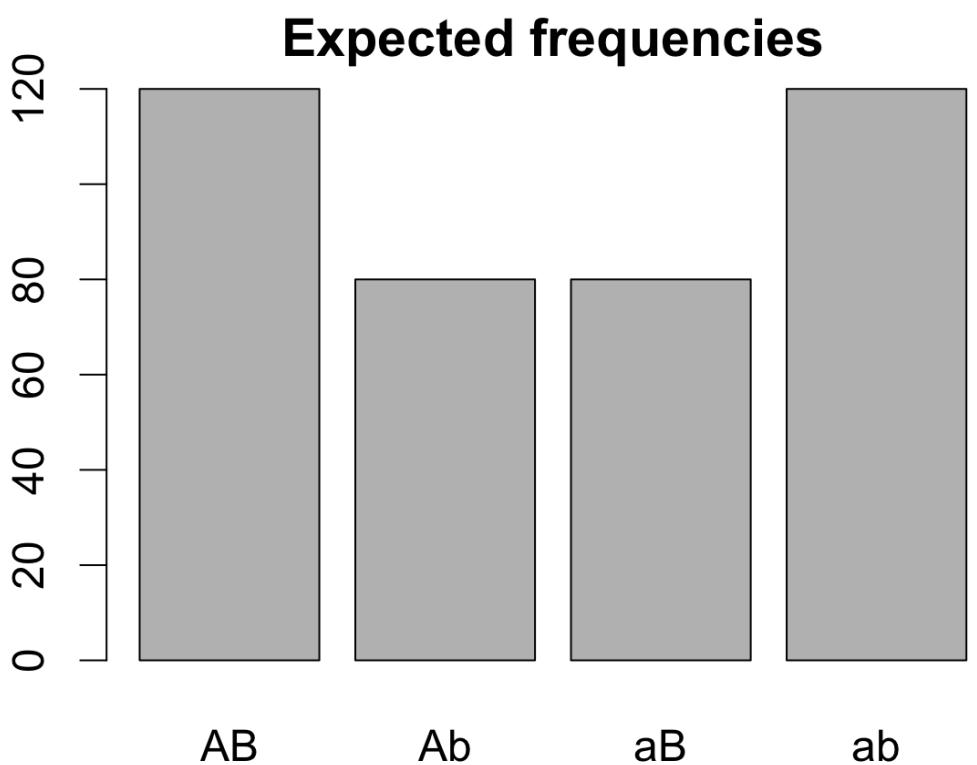
$$\hat{p} = \frac{86 + 74}{400} = 0.4.$$

Hence the four (estimated) hypothesised probabilities are,

Linkage model

```
y = c(128, 86, 74, 112)
n = sum(y)
link_p = c(0.3, 0.2, 0.2, 0.3)
ey = link_p * n
phenotype = c("AB", "Ab", "aB", "ab")
barplot(y, names.arg = phenotype, main = 'Observed frequencies')
barplot(ey, names.arg = phenotype, main = 'Expected frequencies')
```





Linkage model test statistic

The form of the observed test statistic is the same as in the no linkage model,

$$t_0 = \sum_{i=1}^{4} \frac{(y_i - e_i)^2}{e_i},$$

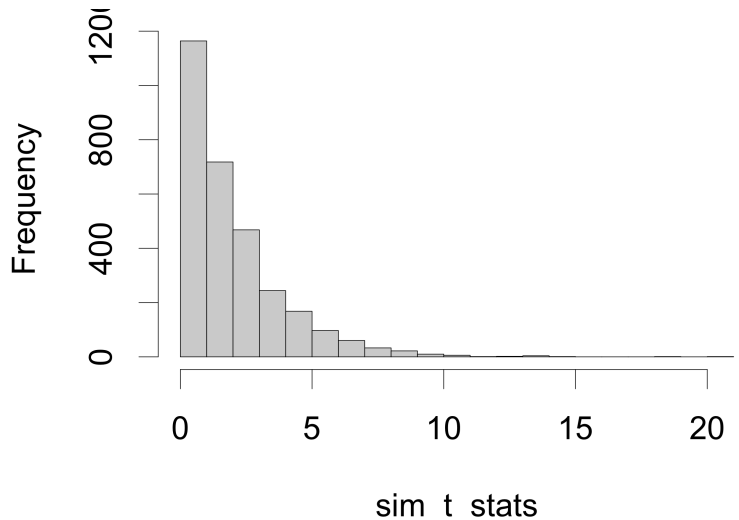
but now the e_i are now calculated using the new hypothesised proportions.

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	$t_0 = 1.96$

Linkage model simulated null distribution

```
B = 3000
sim_t_stats = vector(mode = "numeric",
                     length = B)
for(i in 1:B){
  sim = sample(x = phenotype, size = n,
               replace = TRUE, prob = link_p)
  sim_y = table(sim)
  # estimate recombination fraction
  p_e = sum(table(sim)[2:3])/n
  # calculate expected cell counts
  e = 400 * c(1 - p_e, p_e, p_e, 1 - p_e)/2
  sim_t_stats[i] = sum((sim_y - e)^2/e)
}
```

```
hist(sim_t_stats, main = "",
     breaks = 20)
```



Important

We re-estimate the recombination fraction and expected cell counts *within* the loop replicating what we did to calculate the original test statistic. This is important because it changes the shape of the null distribution. Try for yourself to see what happens when the hypothesised proportions and expectations are fixed.

Linkage model significance

```
n = 400
y = c(128, 86, 74, 112)
n = sum(y)
link_p = c(0.3, 0.2, 0.2, 0.3)
ey = link_p * n
t0 = sum((y - ey)^2/ey)
t0
```

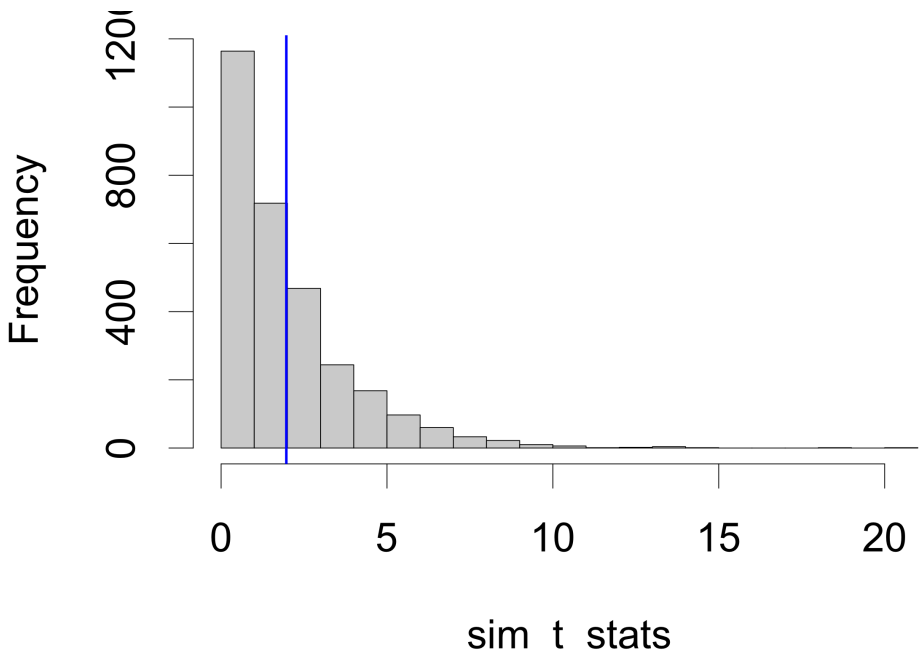
```
[1] 1.966667
```

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

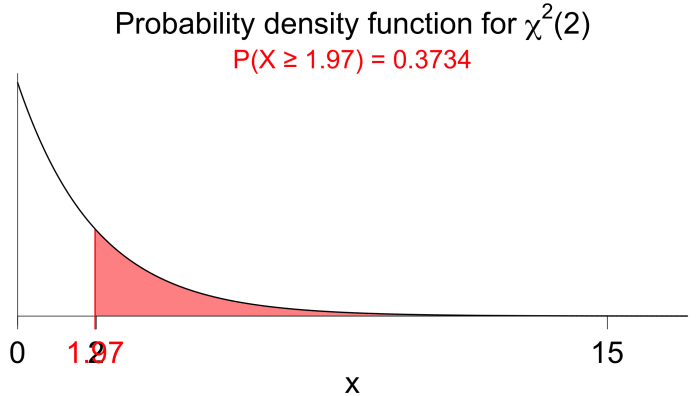
```
mean(sim_t_stats >= t0)
```

```
[1] 0.382
```

```
hist(sim_t_stats, main = "",
     breaks = 20)
abline(v = t0, col = "blue", lwd = 3)
```



- **Hypothesis:** The null 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.
- **P-value:** $P(T \geq t_0) = P(\chi^2_2 \geq 1.97) = 0.37$
- **Decision:** The p-value is much larger than 0.05, so we do not reject the null hypothesis and conclude that the data are consistent with the "coupling phase" linkage model.



For the linkage model, we estimated one parameter, the recombination fraction, $\hat{p}$, so the degrees of freedom are $4 - 1 - 1 = 2$. In contrast, for the no linkage model we didn't need to estimate any parameters, so the degrees of freedom there were $4 - 1 = 3$.

In R

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

```
    Chi-squared test for given probabilities

data:  y
X-squared = 1.9667, df = 3, p-value = 0.5794
```

> ⓘ **Note the incorrect degrees of freedom!**

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

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 & Marx (2012), sections 10.3 and 10.4.

Larsen, R.J., & Marx, M.L. (2012). *An introduction to mathematical statistics and its applications* (5th ed.). Boston, MA: Prentice Hall.