Topic 8 - Chi squared

Chi-Squared distribution:

1. **Test of homogeneity / proportion**

   > Preconceived ideas about proportion

   H0: Each population has the same proportion of observations

   H1: At least one of the population has a different proportions of observations

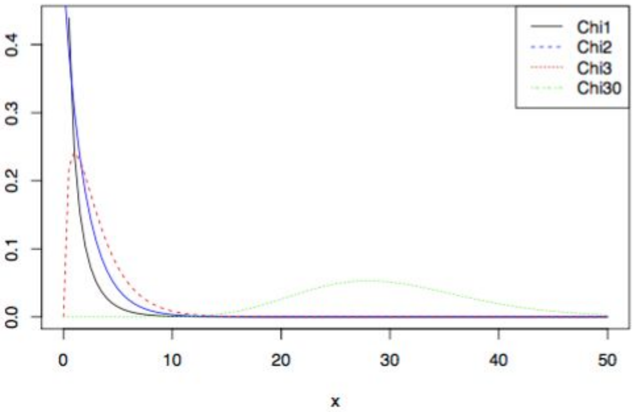
2. **Test of independence**

   > To determine if there is an association (X perceived ideas)

   H0: The variables of interests are independent (no asso)

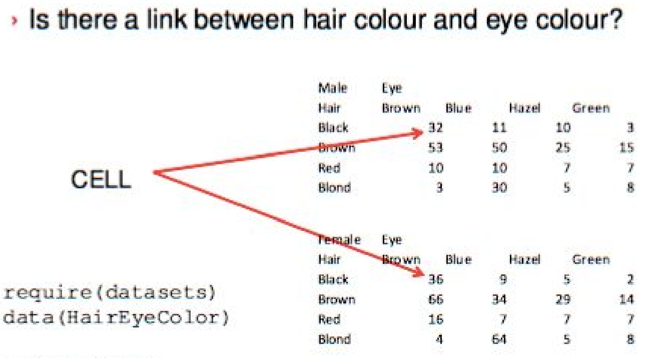
   H1: The variables of interests are not independent (asso)

- Models a variable which can only take positive values
- Skewed in distribution
- $X \sim X_n^2$, with n degrees of freedom
- **Contingency Table**: A table that is used to record relationships between categorical variables
- **Expected frequencies**: The number of times that a category is expected to appear.
- **Observed frequencies** (sample frequencies): The number of times that a category appears in the data.
- **Goodness-of-fit Test:** A test of how well observed data matches a specified, expected probability function.



Contingency Table (4x4)

- A table that is used to record relationships between categorical variables


› Is there a link between hair colour and eye colour?

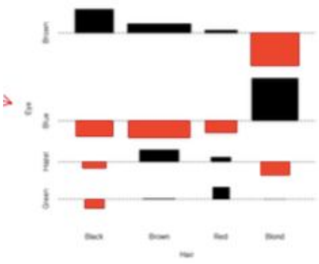
```
Male    Eye
Hair    Brown   Blue    Hazel   Green
Black      32     11       10       3
Brown      53     50       25      15
Red        10     10        7       7
Blond       3     30        5       8

Female  Eye
Hair    Brown   Blue    Hazel   Green
Black      36      9        5       2
Brown      66     34       29      14
Red        16      7        7       7
Blond       4     64        5       8
```

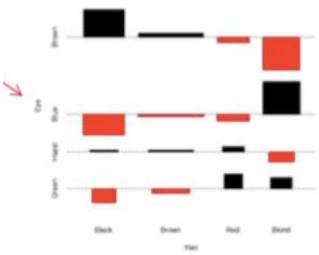
```
require(datasets)
data(HairEyeColor)
```

Mosaic plots

- To detect whether datasets are independent or not
- Block with similar widths along y axis -> independent (no asso.)
- Block with different widths along y axis -> dependent (asso.)

Association plots





Test of independence / proportion

1. Hypothesis
   Test of independence:
   H0: The variables of interests are independent (no asso.)
   H1: The variables of interests are not independent (asso.)

   Test of proportion:
   H0: Each population has the same proportion of observations
   H1: At least one of the population has a different proportions of observations
2. Level of significance $\alpha = 0.5$
3. Check assumptions:
   - No cell has expected frequencies <1
   - No more than 20% of cells have expected frequencies <5
   In the case of above then the probability of a type I error occurring will increase
   We may combine cells to ensure these assumptions are met
4. Calculate Expected frequency = $\frac{\textit{Row total x Column total}}{\textit{Grand total}}$ (Create a new table)
5. Calculate Test statistic (X^2_{ob})

   $$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

   Where O is the observation, E is the expected frequency, n is the number of cells
6. Calculate d.f.
   d.f. = (rows-1)(columns-1)
7. Obtain X^2_{cri} from the table using d.f. & $\alpha = 0.5$
8. Compare X^2_{ob} *and* X^2_{cri}
   If $X^2_{ob} < X^2_{cri}$ → Fail to reject Null
   If $X^2_{ob} > X^2_{cri}$ → Reject Null
   OR
   Obtain p-value from X^2_{ob} and d.f.
   P-value < 0.05 → Reject Null Hypothesis
   P-value > 0.05 → Fail to reject Null Hypothesis

9. Statistical conclusion
10. Biological conclusion

□□

e.g.

Is there an asso between general/advanced and pass/fail? **(Test of independence)**

	General Maths	Advanced Maths	Total Row
Pass	7	19	26
Fail	8	6	14
Total Column	15	25	40

- α = 0.5
- H0: no asso
- H1: asso
- Expected Frequency = $\frac{(26+14)\ x\ (15+25)}{40}$
    - = 40

	General	Advanced
Pass	$\frac{26\ x\ 15}{40} = 9.75$	$\frac{26\ x\ 25}{40} = 16.25$
Fail	$\frac{14\ x\ 15}{40} = 5.25$	$\frac{14\ x\ 25}{40} = 8.75$

- $X^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$

$$= \frac{(7-9.75)^2}{9.75} + \frac{(19-16.25)^2}{16.25} + ... + \frac{(6-8.75)^2}{8.75}$$

$$= 3.55$$

- d.f. = (2-1) x (2-1)

  = 1
- From the table, the critical value = 3.84, observed value = 3.55
- $X^2_{ob} > X^2_{cri}$
- ∴ Fail to reject Null
- No asso

□□ □□□

irrespective to = no asso (independent)

depends on = association (not independent)

T-test VS Chi-square

T-test	About comparing numbers
Chi-square	Counts or frequency of data in different categories