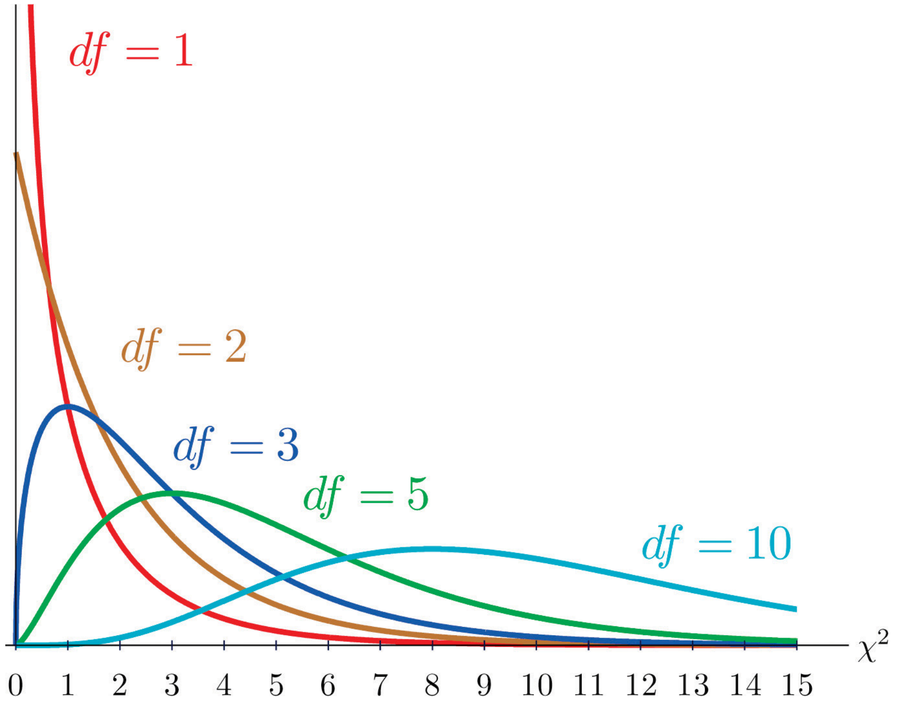
Foundational Statistics

Introduction to Frequency Analysis

F1	PS	Ps	pS	ps
PS	PPSS	PPSs	PpSS	PpSs
Ps	PPSs	PPss	PpSs	Ppss
pS	PpSS	PpSs	ppSS	ppSs
ps	PpSs	Ppss	ppSs	ppss

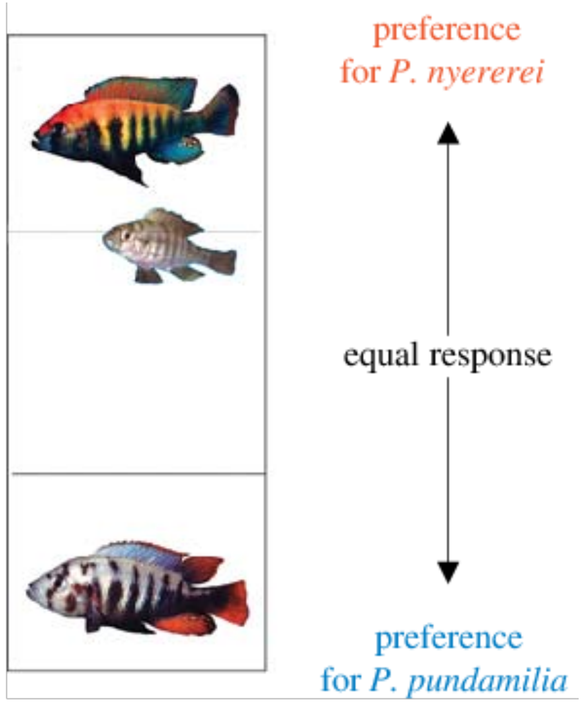
Plantes	Traitement 1	Traitement 2	$N_{k.}$
saines	21	24	44
peu infectées	10	13	23
infectées	11	10	21
très infectées	8	3	11
$N_{.l}$	50	50	100



Response variables are not always continuous, or even numeric

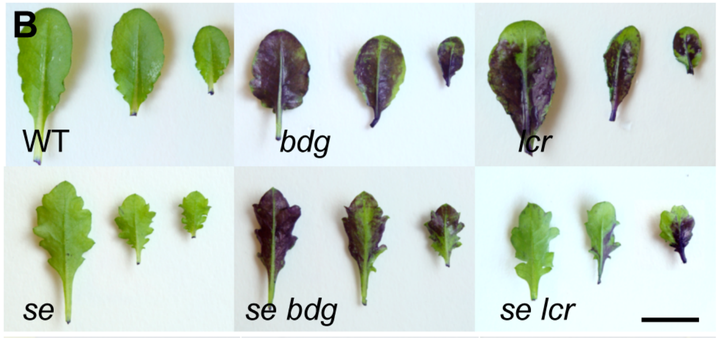
Binary responses:

- Presence / Absence
- Preference / No Preference
- Test Pass / Test Failure



>2 categorical responses:

- Discrete phenotypes or genotypes
- Ecological categories
- Scoring systems (non-ordinal or ordinal)



One approach: Count number of observations within each category to get “frequencies” -> proportions

P ( ) = 3/10 = 0.3

P ( ) = 6/8 = 0.75

Two common frequency analysis scenarios

“Goodness of fit” test: The null hypothesis is that the observed frequencies are sampled from a population with a known, inherent ratio of categories.

- Example: Phenotypic ratios expected based on a specific mechanism of genetic inheritance

“Test of independence” (aka “contingency analysis”):
The null hypothesis is that there is no association between variables (combination freqs are consistent with ind. freqs)

- Example: Associations between eye and coat color in mammals.

Goodness of fit tests: the Chi-square test

An expected count for each category is compared to the observed count, based on a specific hypothesis

Category 1:
N_1 = 22 (observed)

Category 2:
N_2 = 24 (observed)

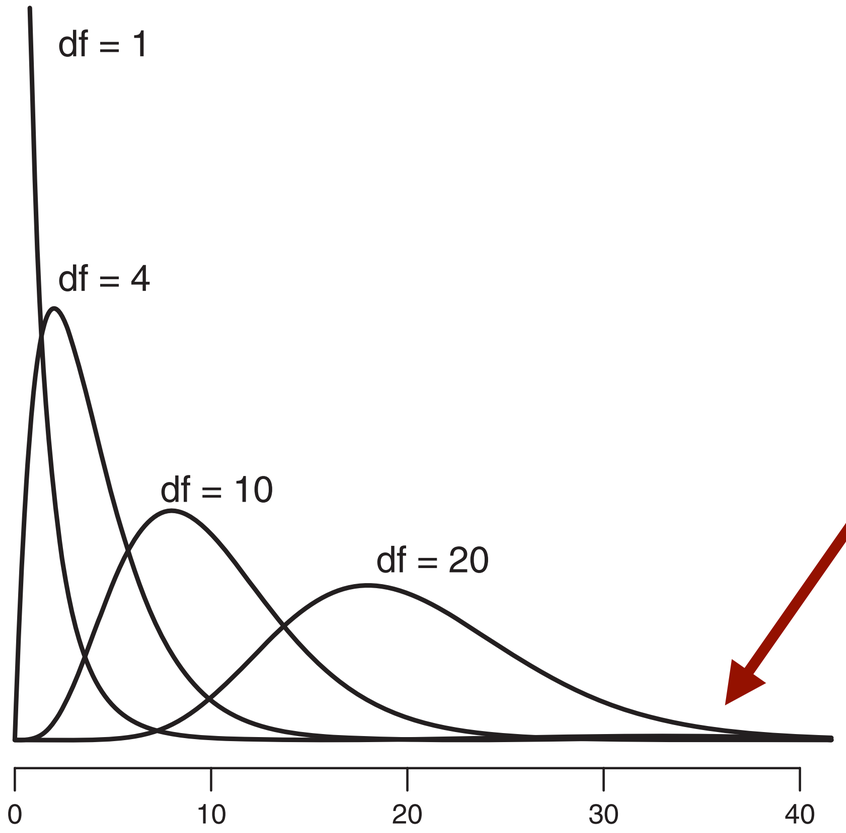
N_{total} = 46

H_0 = 1:1 ratio -> $N_{1(exp)}$ = 46*0.5 = 23
$N_{2(exp)}$ = 46*0.5 = 23

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$ Chi-square test statistic

$$\chi^2 = \frac{(22-23)^2}{23} + \frac{(24-23)^2}{23} = 0.087$$

Goodness of fit tests: the Chi-square test



$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

Large values of the test statistic will be unlikely under the null hypothesis

- The summed degree of difference between observed and expected values is represented by a chi-square statistic
- The distribution is defined by the degrees of freedom
- df for test = number of categories -1

Goodness of fit tests: Chi-square test assumptions

Independent classifications: Each observation is classified independently of all other observations, which come from a random sample.

Few categories with low expected counts: 20% or fewer of the categories should have expected counts of < 5.

- Boosting *N* can often overcome this.
- Test statistic corrections can be applied, or a randomization test is another option