Quantitative Methoden in der Molekularbiologie

12. Categorial data and non-parametric testing

WS 2015/16

Outline

a. Categorial data

b. Goodness of fit tests

c. Contingency tests

d. Non-parametric testing and quantitative data

e. Multiple testing

Categorical data

- Quantitative data: weights, lengths, counts, cycle thresholds (magnitude and order)

- Categorical data: fit into categories, e.g. color, taxonomy, phenotype, disease

Categorical data

Non-parametric testing

- Necessary if all measured data (both predictor and response) are categorical

- Counting and classifying objects in categories:
  - 1 category: goodness of fit tests (data proportional to a prior hypothesis?)
  - 2 categories: contingency tests (proportions in categories different?)

Outline

a. Categorial data

b. Goodness of fit tests

c. Contingency tests

d. Non-parametric testing and quantitative data

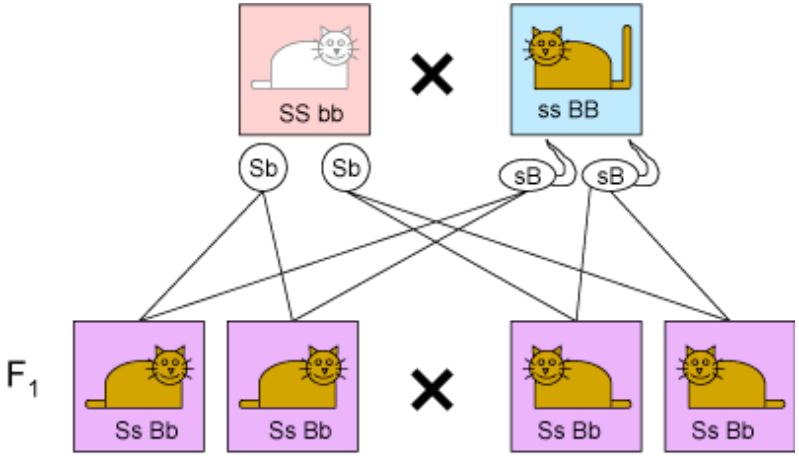
e. Multiple testing

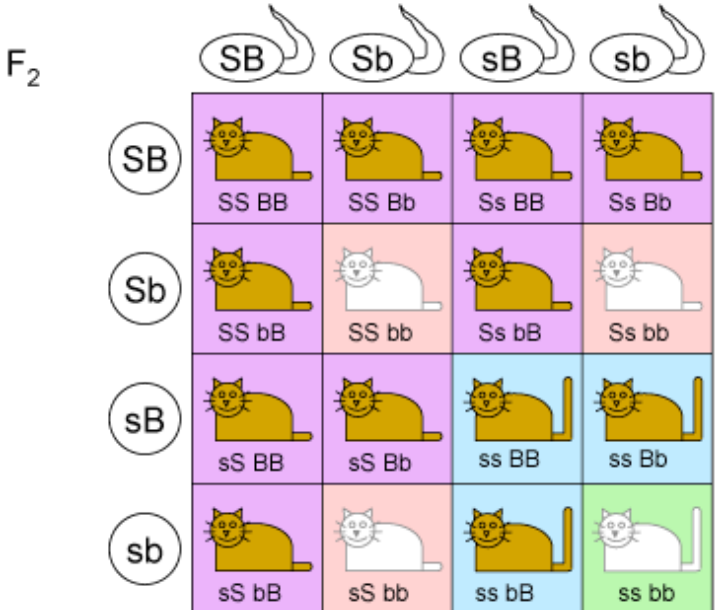
Goodness of fit tests

- Chi-squared goodness of fit test
- Calculate discrepancies between observed and expected frequencies:
  $chi^2=sum((observed-expected)^2/expected)$
- Find probability for test statistic chi^2 (Excel, R or tables, e.g. http://stattrek.com/online-calculator/chi-square.aspx)
- Example: Observed gender in family with 12 children (11 female, 1 male)
  $chi^2=(1-6)^2/6+(11-6)^2/6=8.333$
  p=0.004

Problem 1

- Mendelian rule 3
- Phenotypes in F2 generation in ratio 9:3:3:1
- Observation in 100 cats:
  - 59 brown/short
  - 20 brown/long
  - 11 white/short
  - 10 white/long





Outline

a. Categorial data

b. Goodness of fit tests

c. Contingency tests

d. Non-parametric testing and quantitative data

e. Multiple testing

Contingency tests

- Data representing two different categories
- Null hypothesis: categories are independent
- Establish contingency table:

	Variable 2: A	Variable 2: B
Variable 1: A	Count	Count
Variable 1: B	Count	Count

- Chi-squared contingency test:
  - Calculate expected frequency: row total * column total/grand total
  - Continue as in Chi^2 goodness of fit test

Problem 2

Clearance of plasmodia in blood by new malaria drug vs. reference (Chloroquine):

	Cleared	Not Cleared
Chloroquine	129	55
New drug	80	23

Problem 2

	Cleared	Not Cleared	Sum
Chloroquine	129	55	184
New drug	80	23	103
Sum	209	78	287

E = row total * column total/grand total

Expectation value for Chloroquine/Cleared:

E = 184*209/287 = 133,99

Problem 2

In R:

chisq.test(matrix(c(129,80,55,23),ncol=2,nrow=2))

By default, R uses the "pessimistic" Yates' continuity correction for discrete count values vs. continuous χ^2 distribution:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

chisq.test(matrix(c(129,80,55,23),ncol=2,nrow=2), correct=FALSE)

Outline

a. Categorial data

b. Goodness of fit tests

c. Contingency tests

d. Non-parametric testing and quantitative data

e. Multiple testing

Non-parametric tests and quantative data

- Based on:
  - ranking data
  - measures of centre and scatter (median, quantiles)
- E.g.:
  - Single Sample Tests (Wilcoxon, …)
  - Matched-Pairs Tests (Wilcoxon, …)
  - Independent sampled (Mann-Whitney, …)
- Non-parametric tests need more data than parametric tests.

Example

Human liver cells are treated with a new drug or an old one, in order to assess side effects. The cell growth is visually scored on a scale from 1 (poor) to 10 (perfect). Does the new drug have less side effects as the old one?

Replicate	New drug	Old drug
1	9	5
2	8	6
3	6	4
4	8	6
5	7	5
6	6	7
7	8	6

Assumptions

- The observations must be random
- The observations must be independent from each other

- Samples are assumed to come from populations having the same shape.

Mann-Whitney U test

In R:

```
wilcox.test(c(9,8,6,8,7,6,8),c(5,6,4,6,5,7,6))
```

Outline

a. Categorial data

b. Goodness of fit tests

c. Contingency tests

d. Non-parametric testing and quantitative data

e. Multiple testing

Multiple testing

- Necessary if multiple hypotheses are tested on the same data
- Typical for genetic data (associations genotype-phenotype) or expression data (enrichment of functions in differentially expressed genes)
- Many methods available -> different stringencies

Bonferroni correction: multiply p-values and number of tests

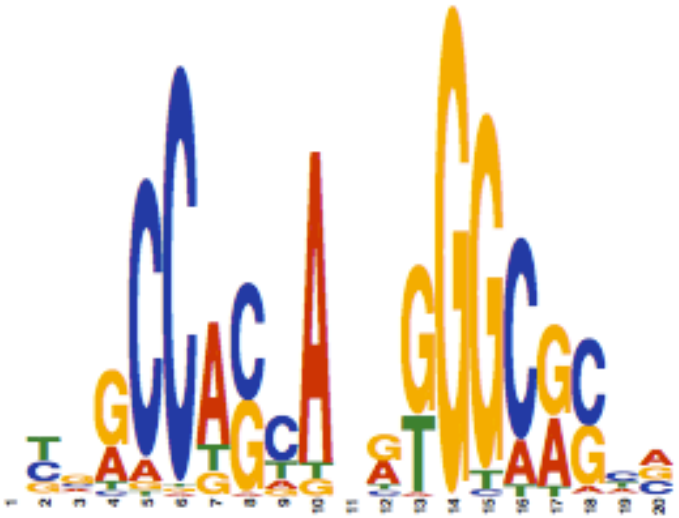
Phenotype/Disease	Rank	Uncorrected p-Value (chi^2)	Multiplicator	Corrected p-Value (chi^2)
Obesity/weight	1	0.003	10	0.03
Heart failure	2	0.007	10	0.07
Alzheimer	3	0.01	10	0.1
Diabetes Type 2	4	0.05	10	0.5
Colon cancer	5	0.09	10	0.9
Breast cancer	6	0.1	10	1
Prostate cancer	7	0.2	10	2
Phenylketonuria	8	0.5	10	5
Leukemia	9	0.7	10	7
Osteoporosis	10	0.9	10	9

FDR correction: multiply p-values and number of tests divided by rank

Phenotype/Disease	Rank	Uncorrected p-Value (chi^2)	Multiplicator	Corrected p-Value (chi^2)
Obesity/weight	1	0.003	10	0.03
Heart failure	2	0.007	5	0.035
Alzheimer	3	0.01	3.33	0.033
Diabetes Type 2	4	0.05	2.5	0.125
Colon cancer	5	0.09	2	0.18
Breast cancer	6	0.1	1.67	0.167
Prostate cancer	7	0.2	1.43	0.286
Phenylketonuria	8	0.5	1.25	0.625
Leukemia	9	0.7	1.11	0.778
Osteoporosis	10	0.9	1	0.9

Example: transcription factor binding site (see primer by William Noble)

a



b

Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	–	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	–	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	CCCGCCAGCAGAGGGAGCCG	24.71
13406383	+	CTAGCCACCAGGTGGCGGTG	24.71
18613020	+	CCCGCCAGCAGAGGGAGCCG	24.71
31980801	+	ACGCCCAGCAGGGGGCGCCG	24.71
32909754	–	TGGCTCCCCCTGGCGGCCGG	24.71
25683654	+	TCGGCCACTAGGGGGCACTA	24.58
31116990	–	GGCCGCCACCTTGTGGCCAG	24.58
29615421	–	CTCTGCCCTCTGGTGGCTGC	24.46
6024389	+	GTTGCCACCAGAGGGCACTA	24.46
26610753	–	CACTGCCCTCTGCTGGCCCA	24.34
26912791	–	GGGCGCCACCTGGCGGTCAC	24.34
20446267	+	CTGCCCACCAGGGGGCAGCG	24.22
21872506	–	TGGCGCCACCTGGCGGCAGC	24.22