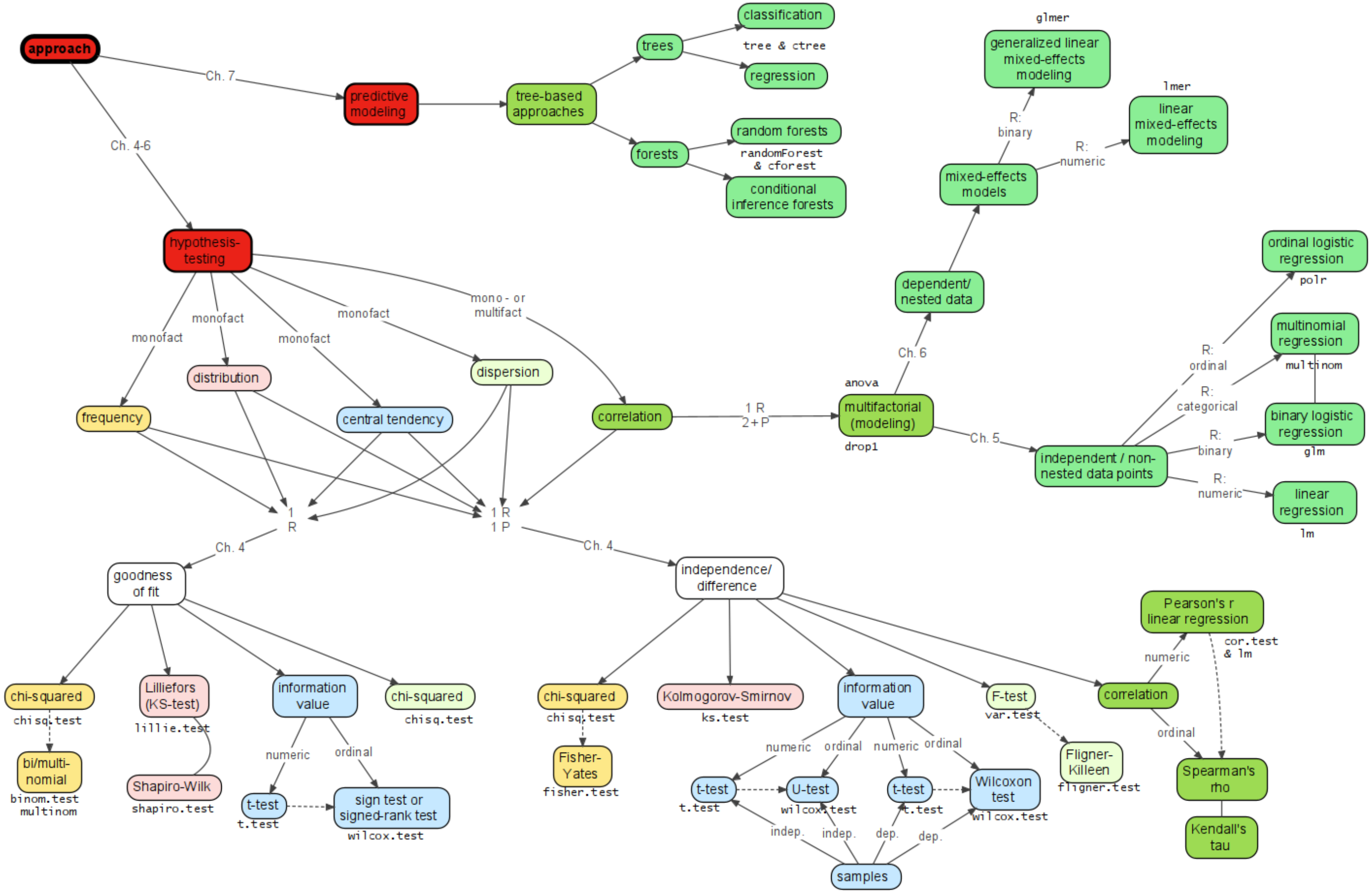
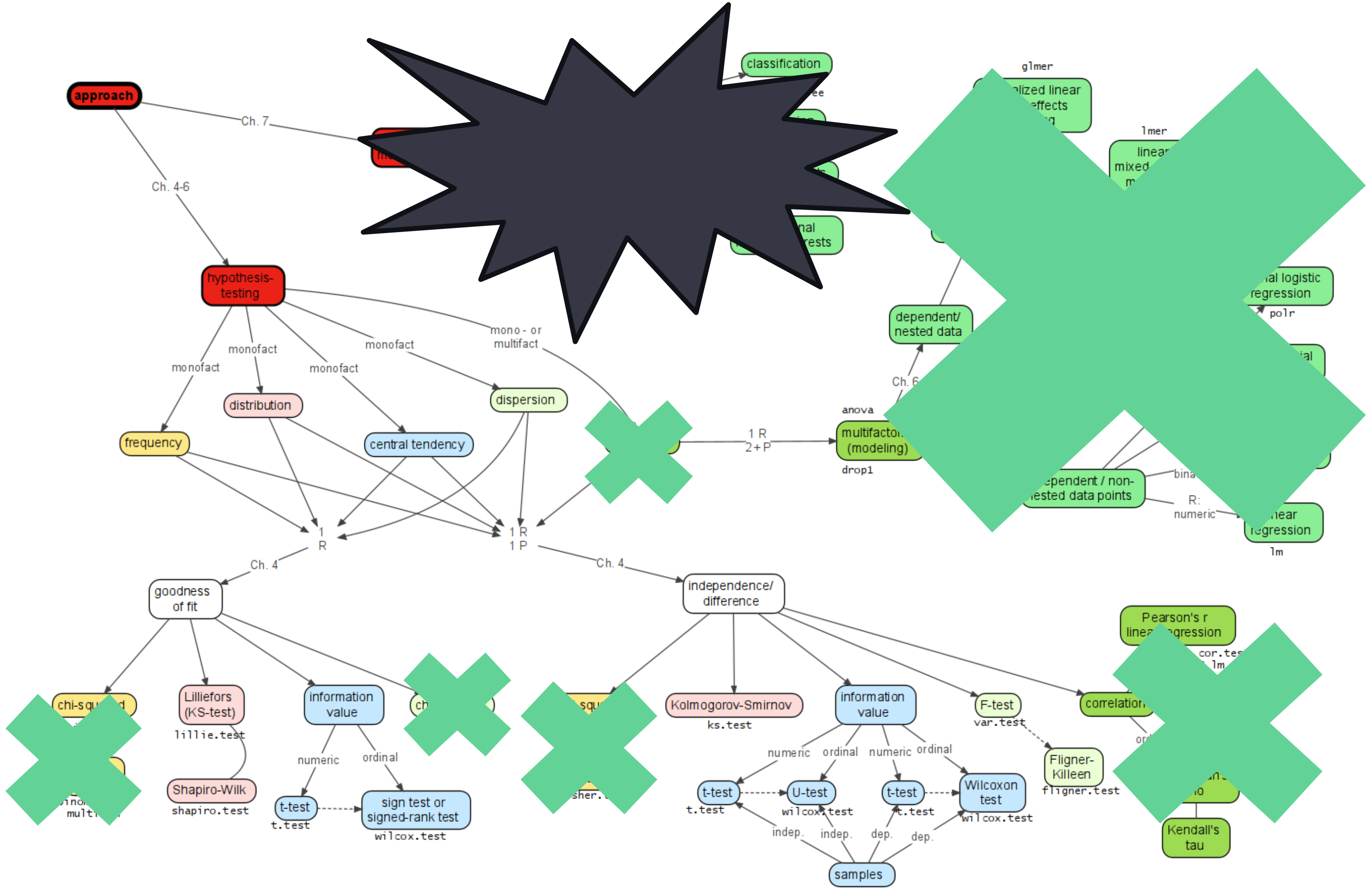
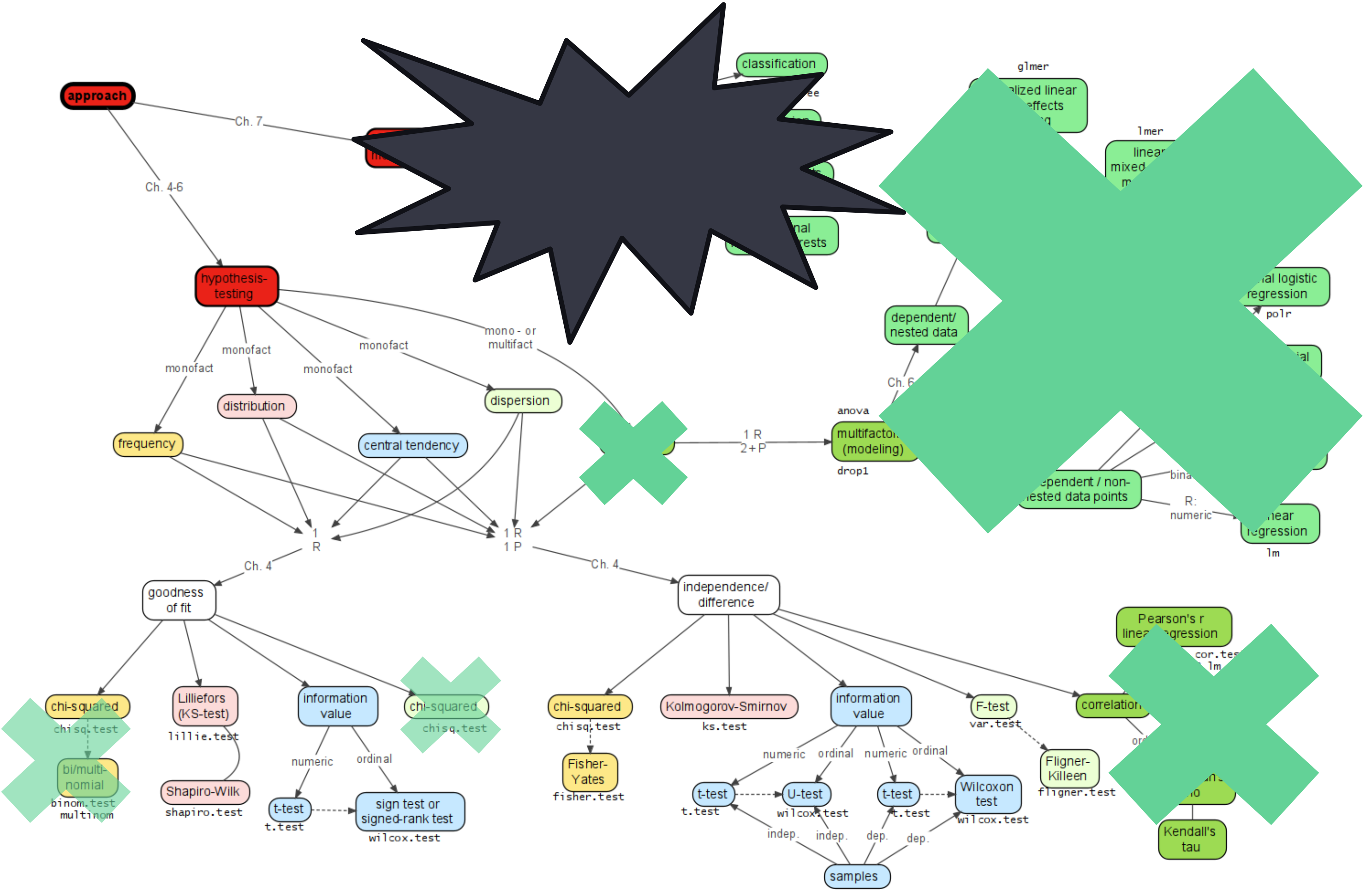
3.2 Chi-squared (X^2) test

For independence


approach
Ch. 7
Ch. 4-6
predictive modeling
tree-based approaches
trees
classification
tree & ctree
regression
forests
random forests
randomForest & cforest
conditional inference forests
hypothesis-testing
monofact
mono- or multifact
frequency
distribution
central tendency
dispersion
correlation
1 R
1 R 1 P
1 R 2 + P
multifactorial (modeling)
anova
drop1
Ch. 6
Ch. 5
dependent/ nested data
mixed-effects models
R: binary
R: numeric
generalized linear mixed-effects modeling
glmer
linear mixed-effects modeling
lmer
independent / non-nested data points
R: ordinal
R: categorical
ordinal logistic regression
polr
multinomial regression
multinom
binary logistic regression
glm
linear regression
lm
Ch. 4
goodness of fit
chi-squared
chisq.test
bi/multi-nomial
binom.test
multinom
Lilliefors (KS-test)
lillie.test
Shapiro-Wilk
shapiro.test
information value
numeric
ordinal
t-test
t.test
sign test or signed-rank test
wilcox.test
chi-squared
chisq.test
independence/ difference
chi-squared
chisq.test
Fisher-Yates
fisher.test
Kolmogorov-Smirnov
ks.test
information value
U-test
wilcox.test
Wilcoxon test
indep.
dep.
samples
F-test
var.test
Fligner-Killeen
fligner.test
correlation
Pearson's r linear regression
cor.test & lm
Spearman's rho
Kendall's tau



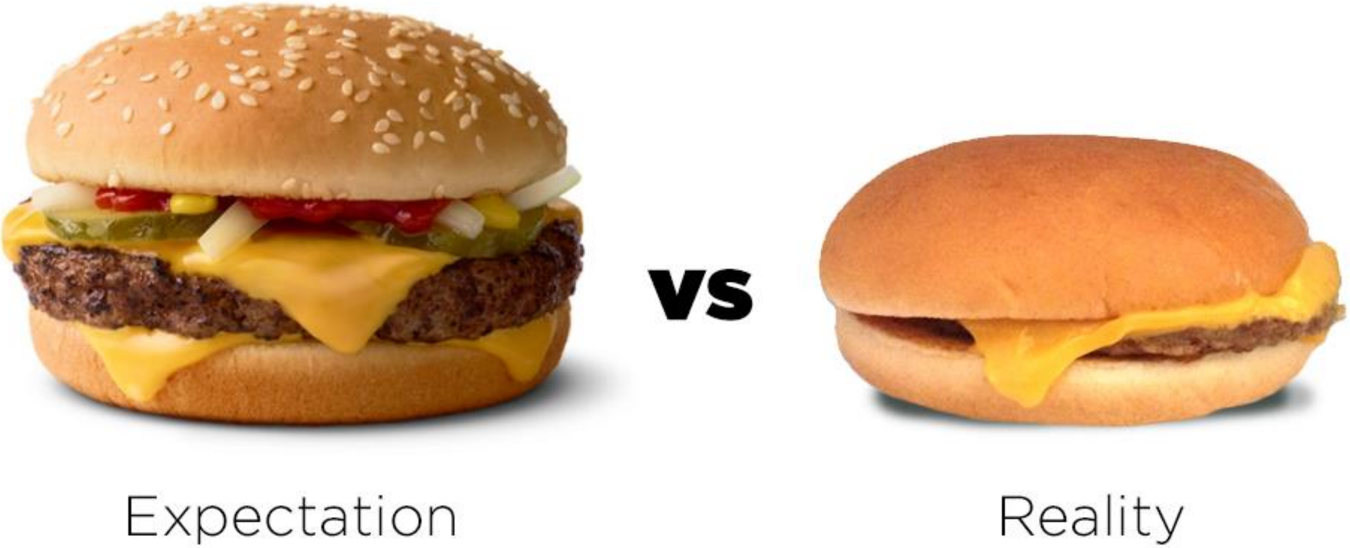
approach
Ch. 7
Ch. 4-6
classification
glmer
lmer
hypothesis-testing
monofact
monofact
monofact
monofact
mono- or multifact
distribution
dispersion
frequency
central tendency
1 R
2+P
anova
drop1
dependent/ nested data
Ch. 6
polr
R: numeric
lm
1 R
1 R 1 P
Ch. 4
Ch. 4
goodness of fit
independence/ difference
Lilliefors (KS-test)
lillie.test
Shapiro-Wilk
shapiro.test
information value
numeric
ordinal
t-test
t.test
sign test or signed-rank test
wilcox.test
Kolmogorov-Smirnov
ks.test
information value
numeric
ordinal
numeric
ordinal
t-test
t.test
U-test
wilcox.test
t-test
t.test
Wilcoxon test
wilcox.test
indep.
indep.
dep.
dep.
samples
F-test
var.test
Fligner-Killeen
fligner.test
correlation
Kendall's tau


approach
Ch. 7
Ch. 4-6
hypothesis-testing
monofact
mono- or multifact
frequency
distribution
central tendency
dispersion
1 R
1 R 1 P
Ch. 4
goodness of fit
chi-squared
chisq.test
bi/multi-nomial
binom.test
multinom
Lilliefors (KS-test)
lillie.test
Shapiro-Wilk
shapiro.test
information value
numeric
ordinal
t-test
t.test
sign test or signed-rank test
wilcox.test
independence/ difference
Fisher-Yates
fisher.test
Kolmogorov-Smirnov
ks.test
U-test
Wilcoxon test
indep.
dep.
samples
F-test
var.test
Fligner-Killeen
fligner.test
correlation
Kendall's tau
classification
1 R 2+ P
multifactorial (modeling)
anova
drop1
Ch. 6
dependent/ nested data
glmer
lmer
polr
R: numeric
lm
cor.test

When to use a chi-squared test for independence?

- To study the **correlation** between **(two) categorical variables**
- **NOT** to study **causality:**
  - No dependent vs independent variable
  - Not a directional test

Essence: expectation vs reality



Essence: expectation vs observation

Chi-squared test compares **expected** and **observed** frequencies, and calculates the difference

- **Observed frequencies =** counts from your experiment / corpus / …
- **Expected frequencies =** expected counts
  (mostly counts that follow from H_0,
  sometimes theoretically motivated distribution,
  results from a well-known paper …)

Example

Let’s say you have a group of 215 voters, and you want to study the **correlation** between the political **candidate** they vote for and the **newspaper** they read

Example

Let’s say you have a group of 215 voters, and you want to study the **correlation** between the political **candidate** they vote for and the **newspaper** they read

Observed frequencies = results from e.g. your survey:

	Candidate A	Candidate B	**Total**
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- **Expected frequencies** based on H_0: even distribution

  ≠ all cells identical (215/4 = 53.75)

  = ratios of cells are equal to each other and the marginal totals

Each cell = (rowtotal*coltotal)/grandtotal

	Candidate A	Candidate B	Total
Newspaper 1			**100**
Newspaper 2			**115**
Total	**125**	**90**	**215**

Example

- **Expected frequencies** based on H_0: even distribution

  ≠ all cells identical (215/4 = 53.75)

  = ratios of cells are equal to each other and the marginal totals

Each cell = (rowtotal*coltotal)/grandtotal

	Candidate A	Candidate B	Total
Newspaper 1	(100*125)/215	(100*90)/215	**100**
Newspaper 2	(115*125)/215	(115*90)/215	**115**
Total	**125**	**90**	**215**

Example

- **Expected frequencies** based on H_0: even distribution

  ≠ all cells identical (215/4 = 53.75)

  = ratios of cells are equal to each other and the marginal totals

Each cell = (rowtotal*coltotal)/grandtotal

	Candidate A	Candidate B	Total
Newspaper 1	58	42	**100**
Newspaper 2	67	48	**115**
Total	**125**	**90**	**215**

Example

Observation vs expectation:

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

	Candidate A	Candidate B	Total
Newspaper 1	58	42	**100**
Newspaper 2	67	48	**115**
Total	**125**	**90**	**215**

Calculation of X^2-value

$$\sum_{i=1}^{n} \frac{(observed\ - expected)^2}{expected}$$

Calculation of X^2-value

$$\sum_{i=1}^{n} \frac{(observed - expected)^2}{expected}$$

Calculation of X^2-value

$$\sum_{i=1}^{n} \frac{(observed - expected)^2}{expected}$$

→ 0 when all observed freqs correspond to expected freqs

H_0: X^2 = 0

Calculation of X^2-value

$$\sum_{i=1}^{n} \frac{(observed - expected)^2}{expected}$$

→ 0 when all observed freqs correspond to expected freqs

H_0: $X^2 = 0$

→ increases as differences between observed and expected freqs increase

H_1: $X^2 > 0$

Example

70	30
55	60

58	42
67	48

Let’s calculate X^2:

$$\sum_{i=1}^{n} \frac{(observed\ - expected)^2}{expected}$$

Example

70	30
55	60

58	42
67	48

Let's calculate X^2:

$$\sum_{i=1}^{n} \frac{(observed\ - expected)^2}{expected}$$

$$= \frac{(70-58)^2}{58} +$$

Example

70	30
55	60

58	42
67	48

Let's calculate X^2:

$$\sum_{i=1}^{n} \frac{(observed\ - expected)^2}{expected}$$

$$= \frac{(70-58)^2}{58} + \frac{(30-42)^2}{42} +$$

Example

70	30
55	60

58	42
67	48

Let's calculate X^2:

$$\sum_{i=1}^{n} \frac{(observed\ - expected)^2}{expected}$$

$$= \frac{(70-58)^2}{58} + \frac{(30-42)^2}{42} + \frac{(55-67)^2}{67} +$$

Example

70	30
55	60

58	42
67	48

Let's calculate X^2:

$$\sum_{i=1}^{n} \frac{(observed - expected)^2}{expected}$$

$$= \frac{(70-58)^2}{58} + \frac{(30-42)^2}{42} + \frac{(55-67)^2}{67} + \frac{(60-48)^2}{48}$$

$= 11.06$

Significance?

- X^2-value alone cannot indicate whether differences between observed and expected freqs are <u>significant</u>

- Look X^2-value up in X^2-table
  - Depending on degrees of freedom (= maximum number of logically independent values)
  - Depending on desired significance level

  → **Value must be >= tabulated value for significance**

Example

- X^2-value = 11.06
- Degrees of freedom: **(nr_rows – 1)*(nr_cols – 1)**

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- X^2-value = 11.06
- Degrees of freedom: **(nr_rows – 1)*(nr_cols – 1)** = 1

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- X^2-value = 11.06
- Degrees of freedom = 1
- Level of significance: $p <= 0.05$ (convention humanities)

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

→ Minimal (critical) X^2-value to reach significance = 3.841

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	$\boxed{3.841}$	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

Example

X^2-value = 11.06 > critical value of 3.841

→ **significant correlation between candidate and newspaper**

The observed freqs are significantly different from the expected freqs if H_0 would be true (i.e. no correlation between newspaper and political candidate), with $p <= 0.05$

Chi-squared test in Python

- No need to look up table
- Built-in function for indepedence: scipi.stats.chi2_contingency()
  - ⇔ goodness-of-fit: scipi.stats.chisquare()

Chi-squared value depends on sample size

The same (proportional) distribution in a **larger** dataset results in:

- a **larger** chi-squared value
- a **smaller** p-value

Example

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

$X^2 = 11.06$

Example

	Candidate A	Candidate B	Total
Newspaper 1	70*10	30*10	**100***10
Newspaper 2	55*10	60*10	**115***10
Total	**125***10	**90***10	**215***10

$X^2 = 11.06*10 = 110.6$

Chi-squared value: depends on sample size

- Issue because:
  - Just because a sample is larger, does not mean that the relation of the values to each other has changed (e.g. become stronger) too

- **Therefore: do not report a chi-squared test without measure of effect size!**
  - = Measure of the size of your effect
  - = how strong is the association?
  - Unaffected by sample size
  - E.g. odds ratio, Cramer's V

Hypothesis testing: workflow

Test statistic

- Chi Square, (rank) correlation, regression…

→

P-value

- < 0,05 → significant association

→

(Measure of effect size)

→

Reject/accept H_0/H_1

Measure of effect size: odds ratio

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Odds for reader newspaper1 to vote for candidate A:

Odds for reader newspaper2 to vote for candidate A:

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Odds for reader newspaper1 to vote for candidate A: 70/30

Odds for reader newspaper2 to vote for candidate A: 55/60

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Odds for reader newspaper1 to vote for candidate A: 70/30

Odds for reader newspaper2 to vote for candidate A: 55/60

Odds ratio: odds_newspaper1 / odds_newspaper2

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Odds for reader newspaper1 to vote for candidate A: 70/30

Odds for reader newspaper2 to vote for candidate A: 55/60

Odds ratio: (70/30) / (55/60) = 2.55 (range 1 to +inf)

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

Odds for reader newspaper1 to vote for candidate A: 70/30

Odds for reader newspaper2 to vote for candidate A: 55/60

- Odds ratio: $\frac{odds1}{odds2}$

Odds ratio: (70/30) / (55/60) = 2.55 (range 1 to +inf)

OR odds_newspaper2 / odds_newspaper1

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Odds for reader newspaper1 to vote for candidate A: 70/30

Odds for reader newspaper2 to vote for candidate A: 55/60

Odds ratio: (70/30) / (55/60) = 2.55 (range 1 to +inf)

OR inverse: (55/60) / (70/30) = 0.39 (range 0 to 1)

	Candidate A	Candidate B	Total
Newspaper 1	70	30	**100**
Newspaper 2	55	60	**115**
Total	**125**	**90**	**215**

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Odds for reader newspaper1 to vote for candidate A: 70/30

Odds for reader newspaper2 to vote for candidate A: 55/60

Odds ratio: (70/30) / (55/60) = 2.55 (range 1 to +inf)

OR inverse: (55/60) / (70/30) = 0.39 (range 0 to 1)

→ Both scores are equivalent!!
But many people find the first one easier to interpret

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

- Odds ratio: $\frac{odds1}{odds2}$

Odds for reader newspaper1 to vote for candidate A: 70/30

Odds for reader newspaper2 to vote for candidate A: 55/60

Odds ratio: (70/30) / (55/60) = 2.55 (range 1 to +inf)

OR inverse: (55/60) / (70/30) = 0.39 (range 0 to 1)

- Interpretation: The odds of voting for candidate A are 2.55 times higher for readers of newspaper1 than for readers of newspaper2

Example

- Odds = $\frac{probability(event)}{1-probability(event)}$

Odds for reader newspaper1 to vote for candidate A: 70**0**/30**0**

Odds for reader newspaper2 to vote for candidate A: 55**0**/60**0**

- Odds ratio: $\frac{odds1}{odds2}$

Odds ratio: (70**0**/30**0**) / (55**0**/60**0**) = 2.55 (range 1 to +inf)

OR inverse: (55**0**/60**0**) / (70**0**/30**0**) = 0.39 (range 0 to 1)

- Odds ratio does not depend on sample size!

Measure of effect size: Cramer's V

= chi-squared value normalized for sample size

$$\sqrt{\frac{X^2}{n * (\min(nrows, ncols) - 1)}}$$

Example

$$\sqrt{\frac{X^2}{n * (\min(nrows, ncols) - 1)}}$$

Example

$$\sqrt{\frac{11.06}{215 * (\min(2,2) - 1)}}$$

= 0.23
(ranges 0 to 1)

Cramers V	Interpretation association
0,00- 0,10	Negligible
0,11 – 0,30	Weak
0,31 – 0,50	Moderate
0,51 – 0,80	Relatively strong
0,81– 0,99	Strong
1	Very strong

Example

$$\sqrt{\frac{11.06 * \mathbf{10}}{2150\mathbf{0} * (\min(2,2) - 1)}}$$

= 0.23
(ranges 0 to 1)

→ **Not affected by sample size!**

Measures of effect size

- Necessary to get an idea of the **size** of your effect
  - = strength association
- Does not change your chi-squared test or p-value (significance)
  - = coincidence or not

→ <u>additional</u> information to report along with your chi-squared test

Final things to keep in mind…

- ALWAYS calculate chi-squared value on the **raw** frequencies, not percentages
  → Why?

Final things to keep in mind…

- ALWAYS calculate chi-squared value on the **raw** frequencies, not percentages

  → because actual counts / actual sample size matters!

Final things to keep in mind…

- ALWAYS calculate chi-squared value on the **raw** frequencies, not percentages

  → because actual counts / actual sample size matters!

  - larger samples: larger chi-squared value

  - too small counts (<u>expected</u> cell count < 5): other test needed

  (e.g. Fisher's Exact Test)

Final things to keep in mind…

- ALWAYS calculate chi-squared value on the **raw** frequencies, not percentages
- Chi-squared test is only for **independent** datapoints

Final things to keep in mind…

- ALWAYS calculate chi-squared value on the **raw** frequencies, not percentages

- Chi-squared test is only for **independent** datapoints

  → For dependent data / repeated measurements:
  see session on mixed effects regression

Workflow chi-squared

1. Research question with dependent variable y and independent variable x
2. Contingency table with row and column totals
3. Formulate H0 & H1 about association x and y
4. Check assumptions: Is data independent? Are sample size and expected cell counts sufficiently large?
5. Calculate $\chi 2$, p-value & measure of effect size
6. Make conclusions about significance and interpretation $\chi 2$ and measure of effect size
7. Accept or discard H0