Statistical Methods Part 2

Evert & Baroni

EURAC Seminar on Statistical Methods

Part II

Stefan Evert[†] & Marco Baroni[‡]

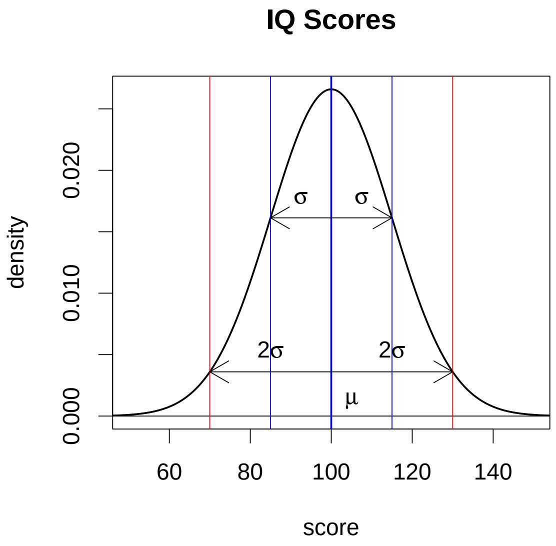
[†]Institute of Cognitive Science
Universität Osnabrück
stefan.evert@uos.de

[‡]SSLMIT / SITLEC
Università di Bologna
baroni@sslmit.unibo.it

30 September 2005

Outline

The normal distribution & z-scores



The normal distribution & z-score

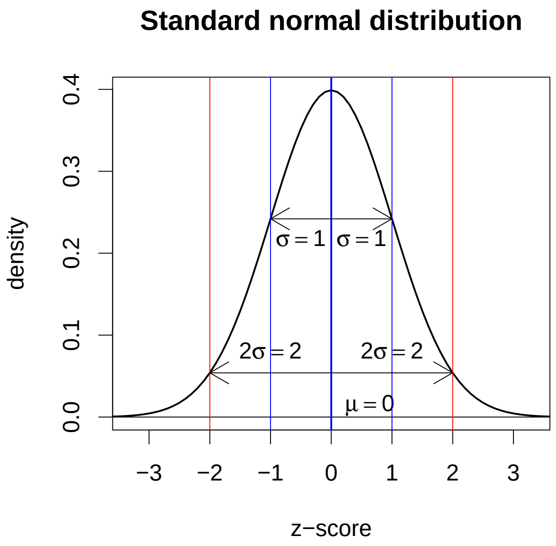
- many distributions have an (approximate) bell shape
- described by the Gaussian bell curve function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- parameters μ (**mean**) and σ (**standard deviation**)
  - μ = centre of bell curve, σ = width of bell curve
- represents distribution of a random variable $X \sim N(\mu, \sigma^2)$, called the **normal distribution** (or **Gaussian distribution**)
- **standard normal** distribution: $\mu = 0$ and $\sigma = 1$
- standardized random variable ➜ **z-score**

$$Z := \frac{X-\mu}{\sigma} \sim N(0,1)$$

Normal distribution & z-scores



UNIVERSITÄT OSNABRÜCK
Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

Normal approximation and Yates' correction

- approximate (discrete) binomial distribution by (continuous) bell curve = normal distribution
- binomial r.v. $X \sim B(n, p)$ ➜ normal r.v. $Y \sim N(\mu, \sigma^2)$
- fit parameters of bell curve to binomial distribution $B(n, p)$

$$\mu = n \cdot p$$
$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}$$

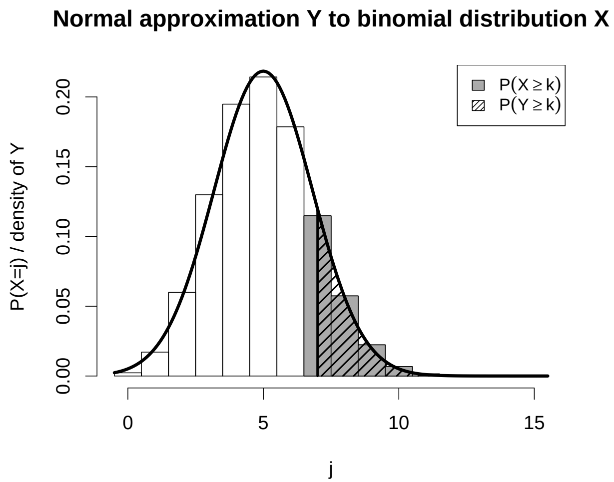
- approximate tail probabilities $\Pr(X \geq L) \approx \Pr(Y \geq L)$
- Yates' continuity correction: $\Pr(X \geq L) \approx \Pr(Y \geq L - \frac{1}{2})$
- translate to z-score:

$$Z := \frac{X - n \cdot p}{\sqrt{n \cdot p \cdot (1 - p)}} \sim N(0, 1)$$

- **R** function: `z.score(X, n, p)` in `corpora` package

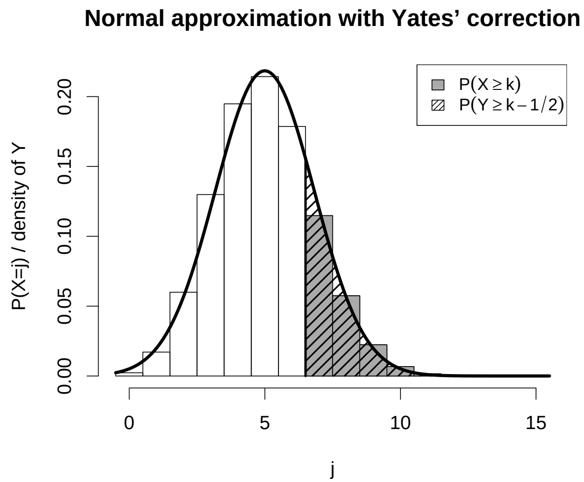
UNIVERSITÄT OSNABRÜCK
Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

Normal approximation and Yates' correction



UNIVERSITÄT OSNABRÜCK
Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

Normal approximation and Yates' correction



Inference based on the normal approximation

- use z-score Z computed from observed frequency O for hypothesis testing and parameter estimation:

$$Z := \frac{O - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$$

- when H_0 is true, $Z \sim N(0,1)$ approximately
- reject H_0 when $Z \geq \delta$ or $Z \leq -\delta$ (one-sided test)
- obtain suitable thresholds from statistical tables or with R: `pnorm(`δ`, lower=FALSE)` and `qnorm(`α`, lower=FALSE)`
- reject H_0 when $|Z| \geq \delta$ or $Z^2 \geq \delta^2$ (two-sided test)
- $Z^2 \sim \chi_1^2$ has chi-squared distribution with $\mathrm{df} = 1$
- obtain suitable thresholds from statistical tables or with R: `pchisq(`δ^2`, df=1, lower=FALSE)` and `qchisq(`α`, df=1, lower=FALSE)`

Frequency comparisons

- so far, we have assumed that we know the exact proportion of nouns in general language (➜ null hypothesis)
- but we have in fact estimated this proportion from a corpus
- ➥ need to take uncertainty of this estimate into account (particularly relevant for comparison of lexical frequencies)
- comparison of frequencies from two samples (rather than test of a pre-determined null hypothesis)
  - this is particularly suitable when the two samples have similar sizes, e.g. when comparing two technical domains (rather than technical writing with general language)

Frequency comparisons

- two **populations** with unknown proportions p_1 and p_2
- one **sample** from each population with sizes n_1 and n_2
  - samples sizes do not have to be equal
- **observed frequencies** O_1 and O_2 in the samples
- **null hypothesis** of equal proportions: $H_0 : p_1 = p_2$ (also called the null hypothesis of **homogeneity**)
- problem: we do not know the common value $p_1 = p_2$
- ➥ cannot compute the sampling distribution (under H_0)
- solution: use "pooled" estimate $\hat{p} = (O_1 + O_2) / (n_1 + n_2)$
- ➥ expected frequencies: $E_1 = n_1\hat{p}$ and $E_2 = n_2\hat{p}$
- evidence against H_0 is provided by $|O_1 - E_1|$ and $|O_2 - E_2|$

Frequency comparisons: the chi-squared test

- $|O_1 - E_1|$ and $|O_2 - E_2|$ cannot be combined directly, especially when sample sizes differ substantially
- z-scores Z_1 and Z_2 are on comparable scale
- ➥ **chi-squared statistic** $X^2 := (Z_1)^2 + (Z_2)^2$
- when H_0 is true, $X^2 \sim \chi_1^2$ has a chi-squared distribution with $\mathrm{df} = 1$ (because 1 parameter has been estimated)
- Pearson's two-sided **chi-squared test** of homogeneity
- easiest **R** command is the proportions test:

`prop.test(c(`O_1`, `n_1`), c(`O_1`, `n_2`))`

Frequency comparisons: the chi-squared test

- data are often represented in a **contingency table**:

	Sample #1	Sample #2	
yes	O_1	O_2	$O := O_1 + O_2$
no	$n_1 - O_1$	$n_2 - O_2$	$n - O$
	n_1	n_2	$n := n_1 + n_2$

- corresponding **expected frequencies** under H_0:

	Sample #1	Sample #2
yes	$n_1 \hat{p}$	$n_2 \hat{p}$
no	$n_1(1-\hat{p})$	$n_2(1-\hat{p})$

Frequency comparisons: the chi-squared test

Statistical Methods
Part 2
Evert & Baroni

Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

- the `chisq.test` function expects such a contingency table
- can easily be constructed with `cont.table` function from our `corpora` package:

  `chisq.test( cont.table(`O_1`,`n_1`,`O_2`,`n_2`), ... )`

- compute X^2 and p-values directly (and for vector arguments) with our `chisq` and `chisq.pval` functions
- exact homogeneity test is **Fisher's exact test**:

  `fisher.test( cont.table(`O_1`,`n_1`,`O_2`,`n_2`), ... )`

- computes **conditional** distribution without estimate $\hat{p}$
- computationally expensive and unstable (memory/accuracy)
- `fisher.pval` is more convenient and robust (but not exact)

Significance vs. effect size

Statistical Methods
Part 2
Evert & Baroni

Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

- the p-value tells us how much **evidence** the observed data provide against H_0, but not how **different** the true parameters are from the null hypothesis $p_1 = p_2$
- for large samples, differences are almost always significant
- **significance** of evidence vs. **effect size**
- measures of effect size (= population parameters)

diff. of proportions	$\delta := p_1 - p_2$
relative risk	$r := p_1 / p_2$
odds ratio	$\theta := \dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$

- confidence intervals for effect size can be obtained with general procedure, using one of the hypothesis tests

Effect size: difference and odds ratio

Statistical Methods
Part 2
Evert & Baroni

Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

- `prop.test` returns confidence interval for the difference of proportions $p_1 - p_2$

```
result <- prop.test(c(O1, O2), c(n1, n2))
diff.lower <- result$conf.int[1]
diff.upper <- result$conf.int[2]
```

➥ not very meaningful for frequency comparisons

- `fisher.test` returns confidence interval for odds ratio

```
result <- fisher.test(cont.table(O1, n1, O2, n2))
theta.lower <- result$conf.int[1]
theta.upper <- result$conf.int[2]
```

- it is not easy to interpret θ either

➥ computationally expensive and large memory consumption

Effect size: relative risk

- most useful measure is relative risk: $r = p_1 / p_2$
- general procedure for confidence interval faces mathematical and numerical problems
- alternative approach: compute binomial confidence intervals for p_1, p_2 independently and compare them
- example: $O_1 = 99$, $n_1 = 5000$; $O_2 = 245$, $n_2 = 20000$
- 95% confidence intervals:

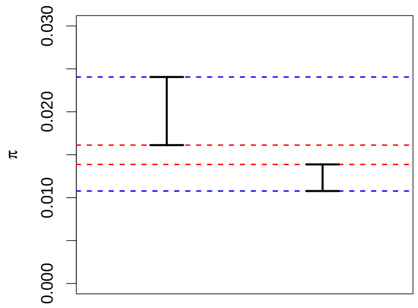
$$p_1 \in [.0161, .0241]$$
$$p_2 \in [.0108, .0139]$$

☞ what is the confidence interval for $r = p_1 / p_2$?

UNIVERSITÄT OSNABRÜCK
Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

Effect size: relative risk

- $p_1 \in [.0161, .0241]$, $p_2 \in [.0108, .0139]$



➥ $r \in \left[\dfrac{.0161}{.0139}, \dfrac{.0241}{.0108}\right] = [1.16, 2.23]$

UNIVERSITÄT OSNABRÜCK
Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

Effect size: relative risk

- conservative estimate for relative risk:
  - estimate p_1 and p_2 individually
  - calculate range of possible values for r from the two binomial confidence intervals
  - need to adjust confidence level for individual estimates
- implemented by `rel.risk.cint` function (`corpora`)

  `rel.risk.cint(`O_1`,` n_1`,` O_2`,` n_2`, ... )`

  - also accepts vectors as arguments
  - one-sided and two-sided confidence intervals
  - individual estimates based on binomial or z-score test

UNIVERSITÄT OSNABRÜCK
Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

Application: terminology extraction

- goal: identify **domain-specific** words
  - i.e. words that are used more often in special language (from a particular domain) than in general language
- frequency comparison: domain corpus vs. reference corpus
- use hypothesis test to find out whether there is significant evidence for a difference in usage (proportions)
- estimate effect size to find out how large the difference is
- **ranking** of term candidates:
  - by p-values (from one-sided or two-sided test)
  - by relative risk (or other measure of effect size)

Application: terminology extraction


```
library(corpora)
data(BNCcomparison); BNC <- BNCcomparison
N.s <- sum(BNC$spoken); N.w <- sum(BNC$written)
BNC$X2 <- chisq(BNC$spoken, N.s, BNC$written, N.w)
signif <-
   chisq.pval(BNC$spoken, N.s, BNC$written, N.w) < .01
BNC[!signif, ]
cint <-
   rel.risk.cint(BNC$spoken, N.s, BNC$written, N.w)
BNC$rrisk <- cint$lower
BNC[order(BNC$X2, decreasing=TRUE), ]
BNC[order(BNC$rrisk, decreasing=TRUE), ]
```

Application: collocation identification

Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

- collocation identification as frequency comparison
- given a **node**, say the verb *put*, divide corpus into
  1. tokens within the **cotext** of the node instances (e.g. spans of 5 tokens to each side of the node)
  2. tokens outside the cotext of the node instances
- any word within 1. is a potential **collocate**
- compare frequencies of word in 1. and 2.
- different **measures of collocativity**, mostly used for ranking
- ➥ learn more at http://www.collocations.de/

Application: collocation identification

Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

```
library(corpora)
data(BNCInChargeOf); C <- BNCInChargeOf
signif <-
   chisq.pval(C$f.in, C$N.in, C$f.out, C$N.out) < .01
sum(signif)
C[!signif, ]
C2 <- C[ , c("collocate", "f.in", "f.out")]
C2$X2 <- chisq(C$f.in, C$N.in, C$f.out, C$N.out)
cint <-
   rel.risk.cint(C$f.in, C$N.in, C$f.out, C$N.out)
C2$rrisk <- cint$lower
```

Application: collocation identification

Statistical Methods Part 2
Evert & Baroni
Exact & asymptotic inference
Frequency comparisons
Significance vs. effect size
Application to terminology extraction
Application to collocation identification
Last words

```
rank.X2 <- order(C2$X2, decreasing=TRUE)
rank.rrisk <- order(C2$rrisk, decreasing=TRUE)
C2$rank.X2 <- rank(-C2$X2, ties="random")
C2$rank.rrisk <- rank(-C2$rrisk, ties="random")
C2[rank.X2[1:30], ]
C2[rank.rrisk[1:30], ]
```

Our last words . . .

- **representative** and **balanced** samples
- the **unit of sampling** and the randomness assumption
- methods for data on **interval scale**
  - z-scores ➜ t-scores and the t-test
  - chi-squared test ➜ F-test
- **non-parametric** methods (without normal approximation)
- **correlation** techniques & **linear models**
- **lexical statistics** (Zipf's law, word freq. distributions)
- Eric Weisstein's World of Mathematics: `http://mathworld.wolfram.com/`
- DeGroot, M. H. and Schervish, M. J. (2002). *Probability and Statistics*. Addison Wesley, Boston, 3rd edition.
- and . . .

Thank You!