The normal distribution & z-scores

→ 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 \( \mu \) (mean) and \( \sigma \) (standard deviation)
  • \( \mu \) = centre of bell curve, \( \sigma \) = 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 \( Z \): 
\[ Z := \frac{X - \mu}{\sigma} \sim N(0, 1) \]
Normal distribution & z-scores

- approximate (discrete) binomial distribution by (continuous) bell curve = normal distribution
- binomial r.v. $X \sim B(n, p) \Rightarrow$ 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 $Pr(X \geq k) \approx Pr(Y \geq L)$
- Yates’ continuity correction: $Pr(X \geq k) \approx Pr(Y \geq L - \frac{1}{2})$
- 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

Normal approximation and Yates’ correction

- Normal approximation with Yates’ correction
  \[ P(X \geq k) \]
  \[ P(Y \geq k - \frac{1}{2}) \]
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:
  - $\text{pnorm}(\delta, \text{lower}=\text{FALSE})$ and $\text{qnorm}(\alpha, \text{lower}=\text{FALSE})$
- reject $H_0$ when $|Z| \geq \delta$ or $Z^2 \geq \delta^2$ (two-sided test)
- $Z^2 \sim \chi^2_1$ has chi-squared distribution with $df = 1$
- obtain suitable thresholds from statistical tables or with R:
  - $\text{pchisq}(\delta^2, df=1, \text{lower}=\text{FALSE})$ and $\text{qchisq}(\alpha, df=1, \text{lower}=\text{FALSE})$

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**

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

<table>
<thead>
<tr>
<th>Sample #1</th>
<th>Sample #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>$O_1$</td>
</tr>
<tr>
<td>no</td>
<td>$n_1 - O_1$</td>
</tr>
</tbody>
</table>

- Corresponding expected frequencies under $H_0$:

<table>
<thead>
<tr>
<th>Sample #1</th>
<th>Sample #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>$n_1\hat{p}$</td>
</tr>
<tr>
<td>no</td>
<td>$n_1(1 - \hat{p})$</td>
</tr>
</tbody>
</table>

**Significance vs. effect size**

- 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**: $\theta := p_1/(1-p_1)$  
  
  $p_2/(1-p_2)$

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

**Effect size: difference and odds ratio**

- The `prop.test` function 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.

- The `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 $\theta$ 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 $r = p_1 / p_2$?

- 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

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)`
- `BNCS$X2 <- chisq(BNC$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$written, N.w)`
- `BNCS$rrisk <- cint$lower`
- `BNC[order(BNCS$X2, decreasing=TRUE), ]`
- `BNC[order(BNCS$rrisk, decreasing=TRUE), ]`

Application: terminology extraction

- `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, ]`
- `C$X2 <- chisq(C$f.in, C$N.in, C$f.out, C$N.out)`
- `C2$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

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

Application: collocation identification

- `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

- `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:
- and . . .

*Thank You!*