Intro to basic statistics

Lilla Zöllei
Feb 5, 2009
Why’N’How
Outline

- Correlations
- *-tests
- ANOVA
Correlation

- **Covariance**: measures strength of linear link between two (numerical) random variables
  \[ \text{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] \]
  - large: strong linear link between RVs
  - small: weak OR nonlinear link between RVs
  - depends on measurement unit!!!

- **Correlation coefficient**: standardized version on covariance
  \[ \rho = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}} \]
Correlation coefficient facts

- $-1 \leq \rho \leq +1$
- $X_1 = X_2 \Rightarrow \text{Cov}(X_1, X_2) = \text{Var}(X_1) = \text{Var}(X_2)$ and $\rho(X_1, X_1) = +1$
- symmetrical: $\rho(X_1, X_2) = \rho(X_2, X_1)$
- if both have unit variances:
  \[ \text{Cov}(X_1, X_2) = \rho(X_1, X_2) \]
- distributions known through a sample (Pearson):
  \[ \rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \]
**Interpretation**

- $\rho(X_1, X_2) = \pm 1$: perfect **linear** functional relationship between $X_1$ and $X_2$
- $\rho(X_1, X_2) \approx 0$  **iff** the relationship between $X_1$ and $X_2$ is linear, their relationship is weak
  $\rightarrow$ uncorrelated variables

- $\neq$ causality
- $\neq$ independence!!!
- independence is much **stronger** than lack of correlation
Anscombe's quartets

Image: courtesy of http://wikipedia.org
Non-parametric corr. coeff.’s

- When the sample distribution is not normal the following tests are more useful:
  - Chi-square, Point biserial correlation, Spearman’s, Kendall’s, Goodman’s, Kruskal’s, …
**t-test**

- Two strong assumptions
  - samples are drawn from **normal** populations
  - for the two samples the **variances** (either known or unknown) are **identical**

- “How much can we trust the **sample mean** as a guess of the mean of the normal distribution from which the sample was drawn ?”

- Examples:
  - Single sample
  - Two samples (paired an unpaired)
One sample $t$-test with reference

$\mu_0$: reference value  
$m$: sample mean

QUESTION: "Is $m$ significantly different from $\mu_0$?"

Null hypothesis: 
$H_0 : \mu = \mu_0$. 

Image: courtesy of http://www.aiaccess.net
Assumptions about variance

- variance of the sample generating normal distribution
  - \textbf{known} \implies \textit{distribution of the standardized sample mean is the standard normal distribution} \(N(0, 1)\)
  - \textbf{unknown} \implies \textit{variance has to be estimated from the sample; but distribution of the standardized sample mean is known: (Student's) \(t\)-distribution}
t-distribution wrt $N(0,1)$

DOF = 1, 2, 3, 5, 10, 30

Image: courtesy of http://wikipedia.org
**t-test: one- or two-sided**

- **Two-sided:**
  - “Is there a significant difference between \( m \) and \( \mu_0 \)?” (in absolute!)
  - alternative hypothesis: \( H_1: \mu \neq \mu_0 \)

- **One-sided:**
  - “Is the mean of the population **larger** (smaller) than \( \mu_0 \)?”
  - alternative hypothesis \( H_1: \mu \geq \mu_0 \) \( (\mu \leq \mu_0) \)
Two paired samples $t$-test

- **QUESTION**: “Is the average shift of the observations due to the *treatment* significantly different from 0?”

  **not**: “Are the means of the two samples significantly different?”

*Image: courtesy of http://www.aiaccess.net*
Two independent samples \( t \)-test

- QUESTION: “Are the means \( m_1 \) and \( m_2 \) significantly different?”

- large difference between them rejects the null hypothesis \( H_0: \mu_1 = \mu_2 \)

Image: courtesy of http://www.aiaccess.net
Variations

• variance of the two populations
  • known  $\rightarrow$ difference between the two standardized sample means is $N(0, 1)$
  • unknown  $\rightarrow$ difference between the two standardized sample means is $t$ distributed
• can also be one- or two-sided
A more general scenario…

- What if there are more than two groups of observations?
- series of *t*-tests on every pair of groups; find at least one pair of groups where hypothesis is rejected
- **ANOVA** (global test)
REMINDER: *t*-test

- two strong assumptions
  - samples are drawn from normal populations
  - for the two samples the variances (either known or unknown) are identical
- test for normality: Kolmogorov-Smirnov test, Shapiro-Wilks test, …
- test for equality of variances: F-test, Levene’s test, Bartlett’s test…

- in case the above assumptions fail
  - Mann-Whitney test (non-parametric)
To proceed

- Select significance level $\alpha = 0.05$ or $0.01$
- $t$-test $\rightarrow p$-value

\[
\left\{ \begin{array}{l}
\text{if } (p < \alpha): \text{ reject } H_0 \\
\text{else: the data is not incompatible (at this significance level) with } H_0 \text{ (does not mean that } H_0 \text{ is true!!!)}
\end{array} \right.
\]
<table>
<thead>
<tr>
<th>SCENARIO</th>
<th>STATISTIC</th>
<th>DOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>indep. one sample with reference</td>
<td>( t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} )</td>
<td>(n-1)</td>
</tr>
<tr>
<td>indep. two samples: equal sample size, equal variance</td>
<td>( t = \frac{\bar{X_1} - \bar{X_2}}{S_{X_1X_2}\sqrt{2/n}} ) [ S_{X_1X_2} = \sqrt{\frac{S^2_{X_1} + S^2_{X_2}}{2}} ]</td>
<td>(2n-2)</td>
</tr>
<tr>
<td>indep. two samples: unequal sample size, equal variance</td>
<td>( t = \frac{\bar{X_1} - \bar{X_2}}{S_{X_1X_2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} ) [ S_{X_1X_2} = \sqrt{\frac{(n_1-1)S^2_{X_1} + (n_2-1)S^2_{X_2}}{n_1 + n_2 - 1}} ]</td>
<td>(n_1+n_2-1)</td>
</tr>
<tr>
<td>indep. two samples: unequal sample size, unequal variance (Welsch)</td>
<td>( t = \frac{\bar{X_1} - \bar{X_2}}{s_{\bar{X_1}-\bar{X_2}}} ) [ s_{\bar{X_1}-\bar{X_2}} = \sqrt{\frac{s^2_1 + s^2_2}{n_1 + n_2}} ]</td>
<td>WS</td>
</tr>
<tr>
<td>dependent (paired) two sample</td>
<td>( t = \frac{\bar{X_D} - \mu_0}{s_D/\sqrt{N}} )</td>
<td>differences are computed!</td>
</tr>
</tbody>
</table>

\( \bar{x} \): sample mean; \( \mu_0 \): population mean; \( s \): sample standard deviation; \( n \): sample size; \( \bar{X}_1 \), \( \bar{X}_2 \): sample means; \( S_{X_1X_2} \): pooled standard deviation; \( n_1 \), \( n_2 \): sample sizes; \( s_{\bar{X}_1-\bar{X}_2} \): standard error of the difference between means; \( s^2 \): sample variance; \( N \): total number of observations.
Some history...

- 1908: first $t$-distribution derivation published
- William Sealy Gosset (Student); Guinness Brewery, Dublin

- $t$-test and theory: through work of R.A. Fisher -- “Student's distribution”
ANOVA

- “ANalysis Of VAriance”
- a type of parametric statistical test
- ~ generalization of t-test for > 2 groups
- uses the F-test
- to show that there is a difference between distribution means

- also: Fisher’s ANOVA (Fisher’s F-distribution)
Example

- Assumptions about the sampling distributions: normal, identical variances (homoscedasticity), independent

Null hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$
To proceed

- Select significance level $\alpha = 0.05$ or $0.01$
- ANOVA $\rightarrow F$-value

\[
\begin{align*}
\text{if } (F < \alpha) &: \text{ reject } H_0 \\
\text{else} &: \text{ the data is not incompatible (at this significance level) with } H_0 \text{ (does not mean that } H_0 \text{ is true!!!)}
\end{align*}
\]
After ANOVA rejects $H_0$ …

- Follow-up tests to analyze the reasons why the hypothesis was rejected: planned (\textit{a priori}) vs. post hoc (\textit{a posteriori})

- \textit{post hoc} examples: Tukey’s test, Dunnett's test
Power analysis

- To determine sufficient sample size to possibly reject null hypothesis
REMININDER: ANOVA

- Strong assumptions
  - normal sampling distributions
  - homogeneous variance of the sampling distributions
- To test for them:
  - normality: Kolmogorov-Smirnov test, Shapiro-Wilk test, …
  - homoscedasticity: Bartlett test, Levene test, …
- In case the above assumptions fail
  - Kruskal-Wallace test (non-parametric)
References / useful cites

- WIKIPEDIA
- http://www.aiaccess.net/English/home.htm
  (images in the talk were obtained from above)
- http://www.physics.csbsju.edu/stats/anova.html
- http://udel.edu/~mcdonald/statanovaintro.html