MITx:
Statistics, Computation & Applications

Statistics Refresher

Lecture 3: Multiple Hypothesis Testing

Some quotes and research findings

Giovannucci et al., Journal of the National Cancer Institute 87 (1995):

Intake of tomato sauce (p-value of 0.001), tomatoes (p-value of 0.03), and pizza (p-value of 0.05) reduce the risk of prostate cancer;

But for example tomato juice (p-value of 0.67), or cooked spinach (p-value of 0.51), and many other vegetables are not significant.

Some quotes and research findings

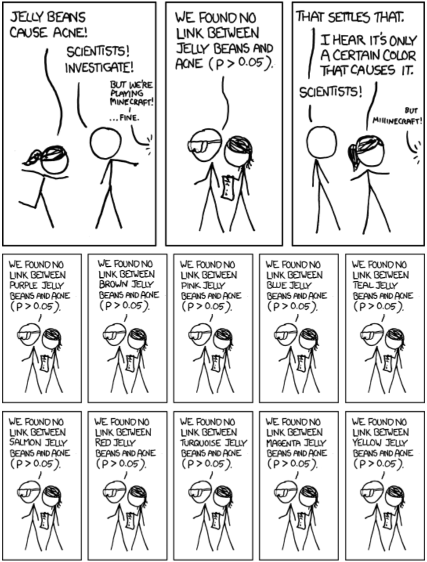
Giovannucci et al., Journal of the National Cancer Institute 87 (1995):

Intake of tomato sauce (p-value of 0.001), tomatoes (p-value of 0.03), and pizza (p-value of 0.05) reduce the risk of prostate cancer;

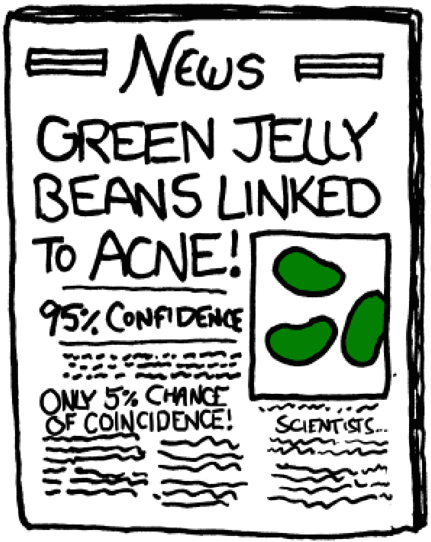
But for example tomato juice (p-value of 0.67), or cooked spinach (p-value of 0.51), and many other vegetables are not significant.

"Orange cars are less likely to have serious damages that are discovered only after the purchase."

Jelly Beans and Acne



Problematic of selective inference



http://imgs.xkcd.com/comics/significant.png

Wonder-syrup

- randomized group of 1000 people
- measure 100 variables before and after taking the syrup: weight, blood pressure, etc.
- perform a paired t-test with a significance level of 5%

Wonder-syrup

- randomized group of 1000 people
- measure 100 variables before and after taking the syrup: weight, blood pressure, etc.
- perform a paired t-test with a significance level of 5%
- $V := \#$ false significant tests: $\quad V \sim \mathrm{Binomial}(100, 0.05)$

$\Rightarrow$ in average 5 out of 100 variables show a significant effect!

Wonder-syrup

- randomized group of 1000 people
- measure 100 variables before and after taking the syrup: weight, blood pressure, etc.
- perform a paired t-test with a significance level of 5%
- $V := \#$ false significant tests: $V \sim \mathrm{Binomial}(100, 0.05)$

$\Rightarrow$ in average 5 out of 100 variables show a significant effect!

Different protection levels

Compute p-values using methods that control:

- family-wise error rate (FWER) $\leq \alpha$, where

$$\text{FWER} = \mathbb{P}(\text{at least one false significant result}) = \frac{b}{m_0}$$

$$= \mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V = 0)$$

$$= 1 - 0.95^{100} \approx 0.99$$

- false discovery rate (FDR) $\leq \alpha$, where

$$\text{FDR} = \text{expected fraction of false significant results among all significant results} = \frac{b}{n_1}$$

	declared non-sig	significant	
H_0 true	a	b	m_0
H_A true	c	d	m_1
	n_0	n_1	$\lvert M \rvert$

Bonferroni correction:

- Reject H_0 when: $m \cdot p\text{-value} \leq \alpha$
  where m is the total number of hypothesis tests performed
- Bonferroni correction implies $\mathrm{FWER} \leq \alpha$

$$
\begin{aligned}
\mathbb{P}(V \geq 1) &= \mathbb{P}(V=1) + \mathbb{P}(V=2) + \ldots + \mathbb{P}(V=m_0) \\
&\leq 0 \cdot \mathbb{P}(V=0) + 1\,\mathbb{P}(V=1) + 2\,\mathbb{P}(V=2) + \ldots + m_0\,\mathbb{P}(V=m_0) \\
&= \mathbb{E}[V] \qquad\qquad V \sim \text{Binomial}\left(m_0, \tfrac{\alpha}{m}\right) \\
&= m_0 \cdot \frac{\alpha}{m} \\
&\leq \alpha
\end{aligned}
$$

$$\mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V=0) = 1 - (1-\alpha_{ind})^{m_0} \leq 1 - (1-\alpha_{ind})^{m} \leq \alpha$$

$$\Rightarrow \alpha_{ind} = 1 - (1-\alpha)^{1/m}$$

Corrections for multiple testing

Bonferroni correction:

- Reject H_0 when: $m \cdot p\text{-value} \leq \alpha$
  where m is the total number of hypothesis tests performed
- Bonferroni correction implies $\mathrm{FWER} \leq \alpha$

Holm-Bonferroni correction:

- Sort p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(m)}$
- Reject H_0 when: $(m - i + 1)p_{(i)} \leq \alpha$ (more power than Bonferroni)
- Holm-Bonferroni correction implies $\mathrm{FWER} \leq \alpha$

$mp_{(1)} \quad (m-1)p_{(2)} \quad (m-2)p_{(3)} \quad \ldots\ldots \quad p_{(m)}$

Corrections for multiple testing

Bonferroni correction:

- Reject H_0 when: $m \cdot p\text{-value} \leq \alpha$
  where m is the total number of hypothesis tests performed
- Bonferroni correction implies $\mathrm{FWER} \leq \alpha$

Holm-Bonferroni correction:

- Sort p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(m)}$
- Reject H_0 when: $(m-i+1)p_{(i)} \leq \alpha$ (more power than Bonferroni)
- Holm-Bonferroni correction implies $\mathrm{FWER} \leq \alpha$

Benjamini-Hochberg correction:

- Sort p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(m)}$
- Reject H_0 when: $mp_{(i)}/i \leq \alpha$
- Benjamini-Hochberg correction implies $\mathrm{FDR} \leq \alpha$

Commonly accepted practice

- No correction for multiple testing when generating hypotheses (but report number of tests performed)
- $\mathrm{FDR} \leq 10\%$ in exploratory analysis or screening
  - balance between high power and low # of false significant results
- $\mathrm{FWER} \leq 5\%$ in confirmatory analysis
  - food and drug administration (FDA)

References

- Lecture by Yoav Benjamini, THE expert for multiple testing issues:

  `http://simons.berkeley.edu/talks/yoav-benjamini-2013-12-11a`