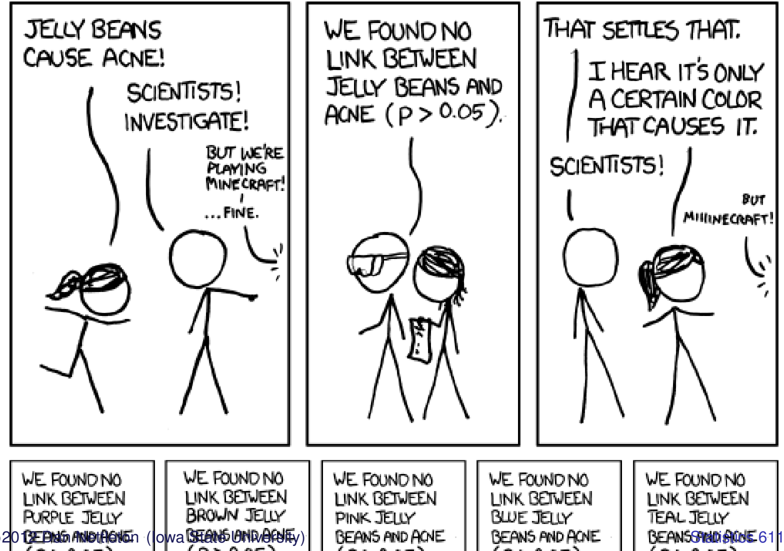
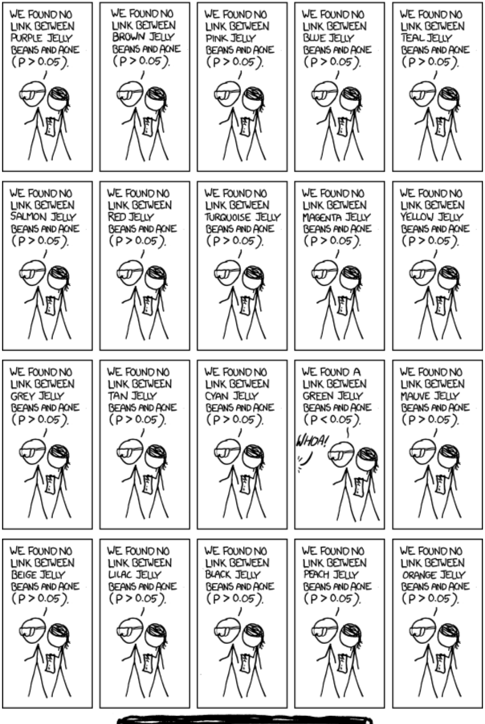
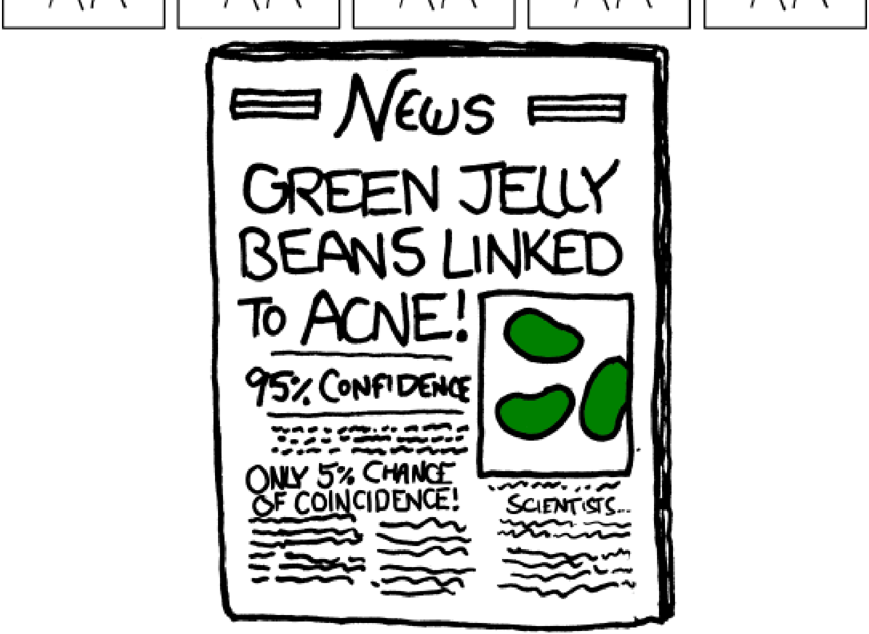
Multiple Testing

http://xkcd.com/882/




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).



NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Familywise Error Rate (FWER)

The simultaneous interval estimation methods we learned about correspond to simultaneous testing procedures that control the familywise error rate (FWER).

The familywise error rate (FWER) is the probability of one or more Type I errors when conducting a family of tests.

For example, suppose

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

and we wish to test

$$H_{0j} : \boldsymbol{c}_j' \boldsymbol{\beta} = 0 \quad \text{for} \quad j = 1, \ldots, m,$$

where each H_{0j} is a testable null hypothesis.

If for each $j = 1, \ldots, m$, we reject $H_{0j} \iff 0 \notin I_j$ where

$$I_j = \left[\boldsymbol{c}_j'\hat{\boldsymbol{\beta}} - t_{n-r,\frac{\alpha}{2m}}\sqrt{\hat{\sigma}^2\boldsymbol{c}_j'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}_j}, \boldsymbol{c}_j'\hat{\boldsymbol{\beta}} + t_{n-r,\frac{\alpha}{2m}}\sqrt{\hat{\sigma}^2\boldsymbol{c}_j'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}_j}\right],$$

then the FWER will be bounded above by α.

Prove that this is true.

Let J_0 denote the set of indices corresponding to true null hypotheses; i.e.,

$$H_{0j} : \boldsymbol{c}_j'\boldsymbol{\beta} = 0 \text{ is true } \iff j \in J_0.$$

Then

$$\begin{aligned}
\text{FWER} &= \mathbb{P}\left(\bigcup_{j \in J_0} \{H_{0j} \text{ is rejected}\}\right) \\
&= \mathbb{P}\left(\bigcup_{j \in J_0} \{0 \notin I_j\}\right) = \mathbb{P}\left(\bigcup_{j \in J_0} \{\boldsymbol{c}_j'\boldsymbol{\beta} \notin I_j\}\right) \\
&\leq \mathbb{P}\left(\bigcup_{j=1}^{m} \{\boldsymbol{c}_j'\boldsymbol{\beta} \notin I_j\}\right) = 1 - \mathbb{P}\left(\bigcap_{j=1}^{m} \{\boldsymbol{c}_j'\boldsymbol{\beta} \in I_j\}\right) \\
&\leq 1 - (1 - \alpha) = \alpha.
\end{aligned}$$

$\square$

As another example, suppose

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, t;\ j = 1, \ldots, n,$$

where $\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{tn} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

Consider testing the family of null hypotheses

$$H_0^{(i,i^*)} : \mu_i = \mu_{i^*}, \quad 1 \leq i < i^* \leq t.$$

Suppose for each $i < i^*$ pair, we reject $H_0^{(i,i^*)}$ iff

$$0 \notin \left[\bar{y}_{i\cdot} - \bar{y}_{i^*\cdot} - \frac{\hat{\sigma}}{\sqrt{n}} R_{t,t(n-1),\alpha}, \bar{y}_{i\cdot} - \bar{y}_{i^*\cdot} + \frac{\hat{\sigma}}{\sqrt{n}} R_{t,t(n-1),\alpha}\right].$$

Then FWER = α.

Strong Control of FWER

The methods we learned about (Bonferroni, Scheffé, Tukey) correspond to multiple testing procedures that provide strong control of the FWER.

A method for testing a family of null hypotheses $H_{01}, \ldots, H_{0m}$ provides strong control of FWER at level α iff FWER $\leq \alpha$ regardless of which or how many nulls in the family are true.

Weak Control of FWER

In contrast, a method provides <u>weak control</u> of FWER at level α if $\text{FWER} \leq \alpha$ whenever all null hypotheses in the family $(H_{01}, \ldots, H_{0m})$ are true.

Strong control of FWER $\Rightarrow$ weak control of FWER.

Show by example that weak control $\not\Rightarrow$ strong control.

Suppose

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, 3;\ j = 1, \ldots, n,$$

where $\varepsilon_{11}, \ldots, \varepsilon_{3n} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

For $i = 1, 2, 3$, consider testing

$$H_{0i} : \mu_i = 0.$$

Consider the following multiple testing procedure.

Reject H_{01}, H_{02}, and H_{03} iff

$$\left| \frac{\bar{y}_{1\cdot}}{\sqrt{\hat{\sigma}^2/n}} \right| \geq t_{3(n-1),\alpha/2}.$$

If

$$\left| \frac{\bar{y}_{1\cdot}}{\sqrt{\hat{\sigma}^2/n}} \right| < t_{3(n-1),\alpha/2}$$

reject nothing.

When H_{01}, H_{02}, and H_{03} are all true,

$$\begin{aligned}
\text{FWER} &= \mathbb{P}(\cup_{i=1}^{3}\{H_{0i} \text{ rejected }\}) \\
&= \mathbb{P}\left(\left|\frac{\bar{y}_{1\cdot}}{\sqrt{\hat{\sigma}^2/n}}\right| \geq t_{3(n-1),\alpha/2}\right) \\
&= \alpha.
\end{aligned}$$

Thus, FWER is weakly controlled at level α.

Suppose H_{01} is false and H_{02} and H_{03} are true.

$$\begin{aligned} \text{FWER} &= \mathbb{P}(\cup_{i=2}^{3}\{H_{0i} \text{ rejected}\}) \\ &= \mathbb{P}\left(\left|\frac{\bar{y}_{1\cdot}}{\sqrt{\hat{\sigma}^2/n}}\right| \geq t_{3(n-1),\alpha/2}\right) \\ &> \alpha. \end{aligned}$$

- Many multiple testing procedures that provide only weak control of the FWER have been published in the statistical literature.

- Methods that provide weak control of the FWER tend to be more powerful than methods that provide strong control, but weak control is rarely sufficient.

- Thus, methods that provide strong control of FWER are preferred.