Biostatistics 201a - Lecture 28 - 12/2/30

Contents

- Logistic Wrap Up
- Multiple Testing Current Event
- Course Summary + Preview of other related classes

* pages = 18

12/2/11
Lecture 28

More Logistic Regression....

Inference on β's in logistic reg:

Confidence intervals for β's:

Basic form is like MLR

$$b_j \pm Z_{\alpha/2} \cdot S_{b_j}$$

- b_j: estimate
- $Z_{\alpha/2}$: use Z instead of t to say how wide CI is.
- S_{b_j}: standard error

On the odds ratio scale the CI is obtained by exponentiation

$$\left[e^{(b_j - Z_{\alpha/2} \cdot S_{b_j})}, e^{(b_j + Z_{\alpha/2} \cdot S_{b_j})} \right]$$

i.e. take the ends of orginal CI and exponentiate.

For example w/ HIV: If you look at printout for b_7 (delivery variable) CI on log odds scale is $[-.69, -.11]$. This means the

log odds of a mom who has a C-section giving HIV to baby are between -.69 and -.11 lower than for a mom w/ natural delivery. (all else equal)

For OR we just get

$$[e^{-.69}, e^{-.11}] = [.4996, .8994]$$

Key is whole CI is below 1 (critical value for OR) - the odds of transmission are somewhere between half as high and 9/10ths as high w/ a C-section compared to natural delivery - or you can say odds are 10% - 50% lower

We can also do hypothesis tests:

H_0: $\beta_7 = 0$ - After adjusting for tx, age, etc. there is no difference in risk for those who have a C-section compared to those w/ a natural delivery

H_A: $\beta_7 \neq 0$ - even after controlling for everything else delivery type is important.

We could equally say we're testing

$OR_{C\text{-section}} = 1$ vs $\neq 1$

Test statistic (estimate, hypothesized value)

$$Z_{obs} = \frac{b_7 - 0}{s_{b_7}}$$

(s_{b_7}: standard error)

$$p\text{-value} = 2P(Z \geq |Z_{obs}|)$$

Example: $b_7 = -.4$ $s_{b_7} = .15$

$$Z_{obs} = \frac{-.4 - 0}{.15} = -2.67$$

$$p\text{-value} = 2P(Z \geq |-2.67|) \approx .004$$

Reject H_0; conclude delivery method matters

For a continuous variable we're interested in how changes in X affect odds of event:

b_j = change in log odds associated with a 1 unit change in X_j all else equal. But suppose you care about a change of Δ units in X_j rather than a change of 1? The associated change in log odds is $\Delta \cdot b_j$

Let's see what this does to odds ratios:

Person A: $X_j = x$

Person B: $X_j = x + \Delta$

otherwise the two people are the same

Person A:

$$\text{log odds} = \ln\left(\frac{p_A}{1-p_A}\right) = \beta_0 + \beta_1 x + \beta_2 X_2 + \cdots$$

Person B:

$$\text{log odds} = \ln\left(\frac{p_B}{1-p_B}\right) = \beta_0 + \beta_1 (x+\Delta) + \beta_2 X_2 + \cdots$$

matter ↑ key ↑ match

Take differences

log odds B − log odds for A

$$= \beta_1(x+\Delta) - \beta_1 x$$

$$= \beta_1 \Delta$$

$$= \ln\left(\frac{p_B}{1-p_B}\right) - \ln\left(\frac{p_A}{1-p_A}\right)$$

$$= \ln\left(OR_{B \text{ vs } A}\right)$$

Exponentiate:

$$OR_{B \text{ vs } A} = e^{\beta_1 \Delta} = \left(e^{\beta_1}\right)^{\Delta}$$

OR for a 1 unit change in X

So it turns out that a Δ change in X on the log odds scale requires raising to the power Δ on OR scale.

Example: X_4 = Age $b_4 = .1$ $e^{b_4} = 1.10$

Odds ratio 1.10 means that all else equal, every extra year of maternal age is associated with a 10% increase in the odds of transmission. What if I want to know how much two extra years of age increases odds? $\Delta = 2 \rightarrow$ on log odds scale the change is $2b_4 = 2(.1) = .2$

On OR scale I get $e^{.2} = 1.21$ (made change and then exponentiated)

Or we could say it's $(e^{b_4})^{\Delta}$

$= (1.10)^2$

$= 1.21$

Confidence intervals for change of Δ in X_j

$\Delta b_j \pm Z_{\alpha/2} \cdot \Delta \cdot s_{b_j}$ (just multiply on log odds scale)

↓ exponentiate

$[e^{(\Delta b_j - Z_{\alpha/2} \cdot \Delta \cdot s_{b_j})} \quad e^{(\Delta b_j + Z_{\alpha/2} \cdot \Delta \cdot s_{b_j})}]$

on odds ratio scale

Probabilities in Logistic Regression

Given a set of X values how do we get a predicted probability rather than odds?

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X_1 + \cdots + b_m X_m$$

log odds can estimated by plugging in as MLR. But we want p! We can solve for p and we get

$$p = P(\text{event}) = \frac{e^{(b_0 + b_1 X_1 + \cdots + b_m X_m)}}{1 + e^{(b_0 + b_1 X_1 + \cdots + b_m X_m)}}$$

Looks messy but not hard - just compute log odds piece, $(b_0 + b_1 X_1 + \cdots)$, exponentiate and then compute ratio

Y: Whether or not mother transmits HIV to baby (1 = Yes, 0 = No)
X1 = Indicator for whether mother is taking HAART A
X2 = Indicator for whether mother is taking HAART B
X3 = Mother's viral load (log scale)
X4 = Mother's age
X5 = Number of years the mother has been HIV positive
X6 = Number of weeks during the pregnancy the mother was on HAART
X7 = Delivery method (1 = C-section, 0 = normal delivery)

Test stat + p-value for overall model
$H_0: \beta_1 = \dots = \beta_m = 0$
H_A: not all $\beta_j = 0$

. logit HIVplus HAART_A HAART_B VLoad Age YrsHIV WksHAART Delivery

Logistic regression

Number of obs	=	300
LR chi2(7)	=	32.47
Prob > chi2	=	0.000
Pseudo R2	=	0.500

Log likelihood = -26.51722

Test of $\beta_j = 0$ vs $\beta_j \neq 0$

HIVplus	Coef = b	Std. Err. = s_b	z	P>|z| (p-value)	[95% Conf. Interval]	
HAART_A	-0.70	0.250	2.80	0.005	[-1.19,	-0.21]
HAART_B	-1.80	0.300	6.00	0.000	[-2.39,	-1.21]
VLoad	0.00001	0.0000025	4.00	0.000	[.000005,	.000015]
Age	0.10	0.050	2.00	0.046	[0.00,	0.20]
YrsHIV	0.10	0.080	1.25	0.211	[-0.06,	0.26]
WksHAART	-0.05	0.010	-5.00	0.000	[-0.07,	-0.03]
Delivery	-0.40	0.150	-2.67	0.004	[-0.69,	-0.11]
_cons	-5.00	0.500	-10.00	0.000	[-5.98,	-4.02]

CIs for β's

exponentiate — equal — equal — exponentiate

. logistic HIVplus HAART_A HAART_B VLoad Age YrsHIV WksHAART Delivery

Logistic regression

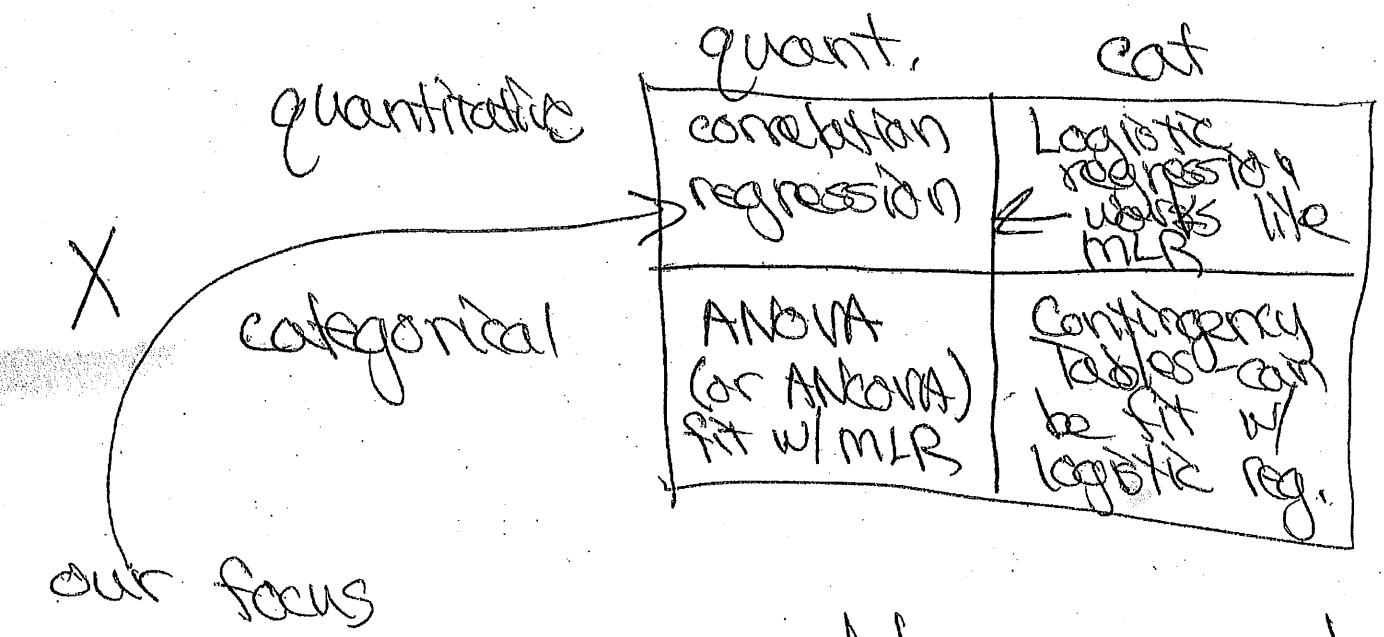
Number of obs	=	300
LR chi2(7)	=	32.47
Prob > chi2	=	0.000
Pseudo R2	=	0.500

Log likelihood = -26.51722

HIVplus	OddsRatio = e^b	z	P>|z|	[95% Conf. Interval] ($e^{\text{CI for } \beta}$)	
HAART_A	0.4966	2.80	0.005	[0.3042,	0.8106]
HAART_B	0.1652	6.00	0.000	[0.0916,	0.2982]
VLoad	1.00001	4.00	0.000	[1.000005,	1.000015]
Age	1.1052	2.00	0.046	[1.0020,	1.2190]
YrsHIV	1.1052	1.25	0.211	[0.9448,	1.2928]
WksHAART	0.9512	-5.00	0.000	[0.9328,	0.9700]
Delivery	0.6703	-2.67	0.004	[0.4996,	0.8994]

Course Recap:

Focus has been on relationships among variables, specifically, using linear models. Right method to use depends (though maybe not as much as you'd think) on nature of X and Y.



In all these models we took advantage of a general framework

Basic Model (G.L.M.) & "g inverse"

$\mu_{Y|X}$ = Average value of Y for a given set of X's = $g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon)$

g^{-1}: transformation; $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$: linear combo of X's

Regular regression:

$g(\mu) = \mu$ — identity — i.e. no transformation needed

Logistic regression: $Y = \begin{cases} 1 & \text{if event happens} \\ 0 & \text{if not} \end{cases}$

$E(Y|X's) = \mu_{Y|X} = p$ = the prob. event happens

$g(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots$

In fitting these models and doing inference about parameters we've made some assumptions, especially about the distribution of points about the mean value. Under these assumptions we've been

using maximum likelihood estimates of model parameters — i.e. our guess for the population relationship the values most likely to generate our data under our assumptions.

We've looked to some extent at how to check the assumptions and to correct for them if they're wrong using transformations. When this doesn't work there are other methods:

1) Weighted analyses: So far we've treated all points equally - but may not have equal confidence in all observations or predictors. Put more weight on things you know well. Can fix problems like Constant Variance violations

2016

② Robust methods (regression and otherwise) - examples include ridge regression, principle components regression, trimmed means, ---

2016 Idea - get a similar model to classical analysis but in a way less influenced by outliers / assumption violations

③ Generalized Linear model - lots of useful link functions for other types of data - Poisson regression (count data), negative binomial regression, ordinal / multinomial logistic (>2 categories) etc. Helps if you have a known but not normal distribution

2016

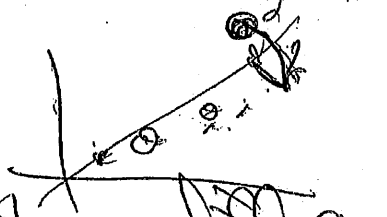
④ Non-parametric methods:

Methods that make no or very limited assumptions about underlying distributions

(Tend) to have less power than a method that assumes correct distribution)

- Rank-based methods - replace the numeric value of Y or X with its rank (from 1 to n) in the data set. Do analysis on ranks, not original values. Spearman rank correlation is just correlation on ranks. Deals well w/ outliers and non-linearities

Baskets
212, 213, 275, 276, 277



Wilcoxon rank sum test - difference of means

- Simulation-based methods - permutation tests: Shuffle group labels, recalculate test statistic, see how unusual your original was
  bootstrap: Similar idea but uses "resampling"

Let data tell you about dist'n

(5) Mixed effects / repeated measures
201, models - e.g. measurements on
411, the same subject at multiple
236 points in time or under
multiple treatment conditions
e.g. Longitudinal or crossover
designs
Problem - measurements w/in
subject tend to be correlated;
not independent.
Take this into account in
standard error calculations.

(6) Methods for missing Data
Multiple Imputation, Em algorithm
232

(7) Multivariate Methods - What
if you have multiple (related)
406 outcomes of interest? Cluster
251 analysis, factor analysis, ---

(8) Bayesian Methods - take into account prior knowledge about the system you're dealing with.

234

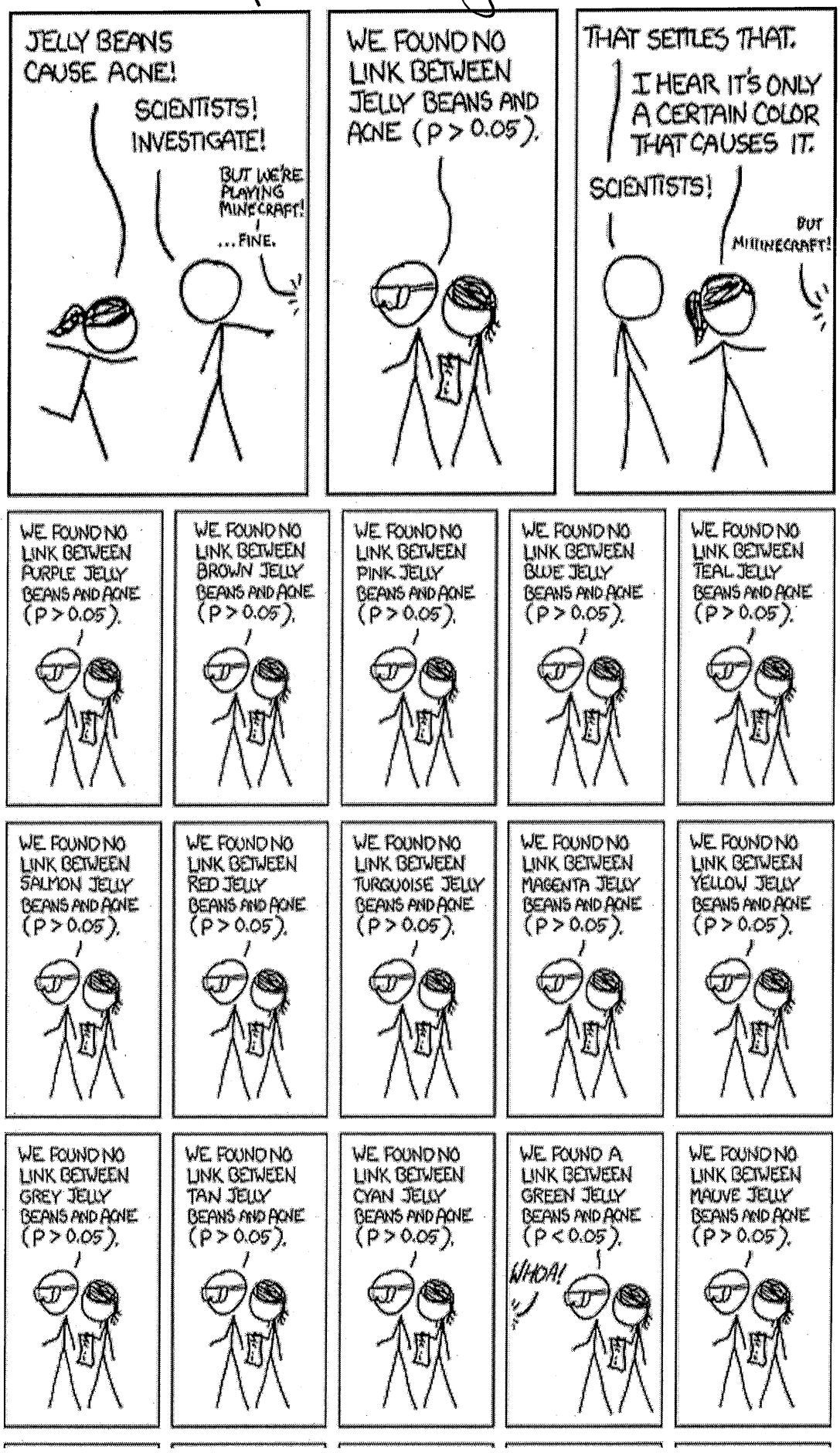
(9) Survival Analysis / Censored Data: eg. how long does someone live after being diagnosed w/cancer. Event of interest (death) may not happen during study (censoring)

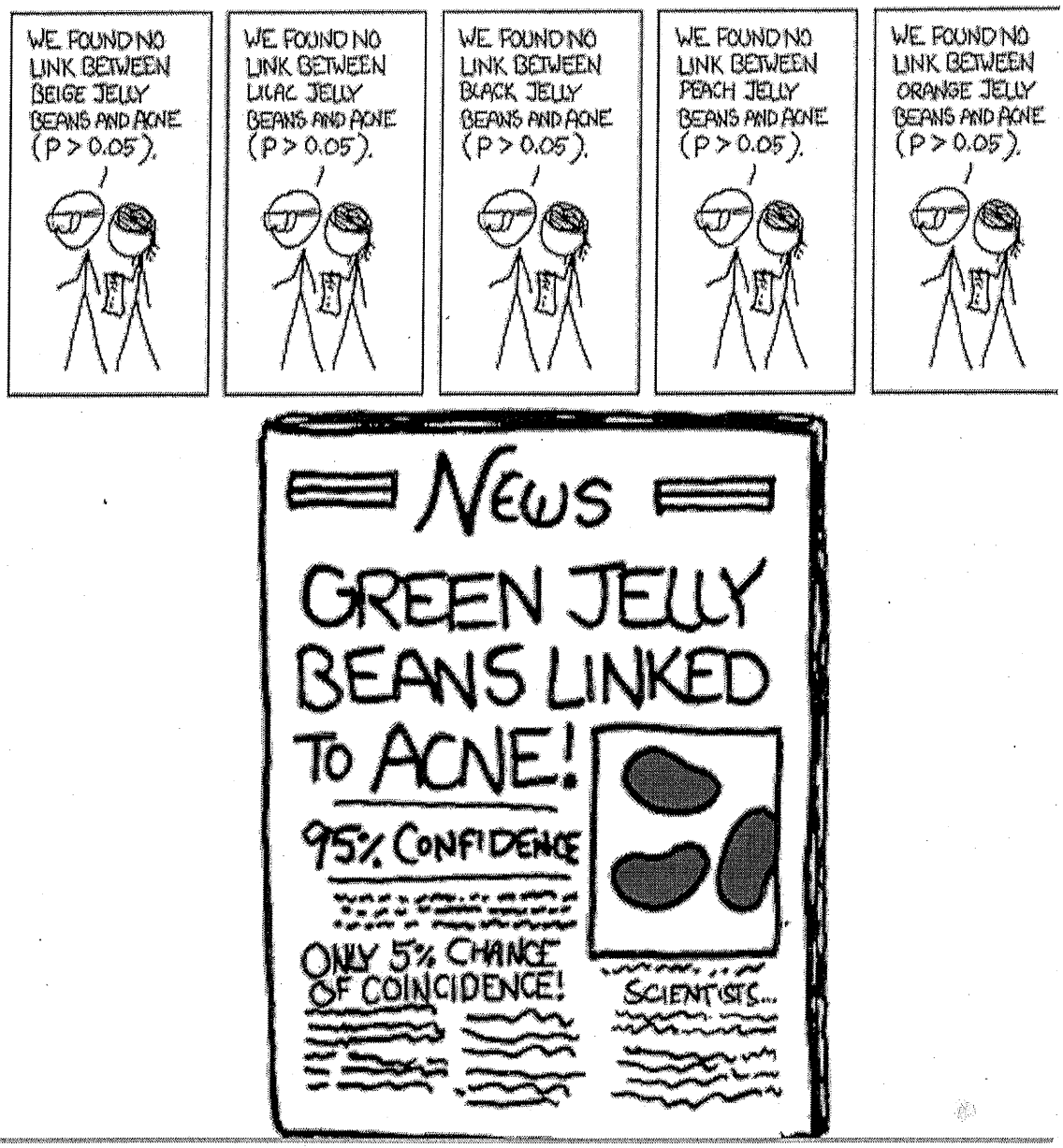
215

(10) Meta Analysis - combine results from multiple studies?

?

Multiple Testing Current Event




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...