AIMS VIDEO COURSES

SUPPORTING BOOKLET

PROBABILITY & STATISTICS

WITH

PROF DAVID SPIEGELHALTER

AIMS

SOUTH AFRICA

AIMS
SOUTH AFRICA

African Institute for Mathematical Sciences

6 MELROSE ROAD | MUIZENBERG | CAPE TOWN 7945 | SOUTH AFRICA
TEL: +27 (0)21 787 9320 | FAX: +27 (0)21 787 9321
EMAIL: info@aims.ac.za | WEB: www.aims.ac.za

AIMS Online Courses

The mission of the AIMS academic programme is to provide an excellent, advanced education in the mathematical sciences to talented African students in order to develop independent thinkers, researchers and problem solvers who will contribute to Africa's scientific development.

Teaching at AIMS is based on the principle of learning and understanding, rather than simply listening and writing, during classes, and on creating an atmosphere of increasing our knowledge through class discussions, through small group discussions, by formulating conjectures and assessing the evidence for them, and sometimes going down wrong paths and learning from the mistakes that led us there. The essential features of the classes at AIMS are that, in contrast to formal lecture courses, they are highly interactive, where the students engage with the lecturer throughout the class time, are encouraged to learn together in a journey of questioning and discovery, and where lecturers respond to the needs of the class rather than to a pre-determined syllabus. AIMS teaching philosophy is to promote critical and creative thinking, to experience the excitement of learning from true understanding, and to avoid rote learning directed only towards assessment.

Leading international and local experts offer the courses at AIMS, which are three weeks long (each module consisting of 30 hrs) and collectively form the coursework for a structured masters degree which also includes a research component. The advertised content is a guide, and the lecturers are encouraged, and indeed expected, to adapt daily to meet the current needs of the students.

Over the past ten years AIMS has achieved international recognition for this innovative and flexible approach. It has been the starting point for the remarkable success of our students and alumni and we all benefit from the support of many who have "witnessed the AIMS-magic and keep coming back for more."

This year we have decided to film selected courses and to make them available to a larger audience as an online facility. African universities may choose to use these courses to supplement and enhance their own postgraduate programmes. We believe this would be best achieved through engagement with AIMS. One way for this to happen, would be for AIMS to suggest or nominate a specialist tutor to spend time at the university, guiding students who follow the online programme. Where possible expert lecturers who have taught at AIMS may visit the university to give a short introduction to the course. We would welcome this interaction as well as the contribution our online courses will make to the growth of the mathematical sciences ecosystem in Africa.

Barry Green
Director & Professor of Mathematics
African Institute for Mathematical Sciences
January 2013

PROBABILITY & STATISTICS

2012

PROF DAVID SPIEGELHALTER

DAY 12

Adjusted estimates

How could we adjust for baseline imbalances?

Suppose we had a continuous outcome Y, a treatment t (=0 if 'control', =1 with 'new treatment'), and possible 'confounders' $x_1, ..., x_k$.

Confounder: something that might be correlated both with the treatment (through imbalance) and the outcome.

Build a regression model

$$\mathbb{E}[Y] = a + bt + c_1 x_1 + ... + c_k x_k$$

For *fixed* $x_1, ..., x_k$, b is the change in the expectation of Y associated with the treatment i.e.

$\mathbb{E}[Y|t = 1, \underline{x}] = a + b + c_1 x_1 + ... + c_k x_k$

$\mathbb{E}[Y|t = 0, \underline{x}] = a + c_1 x_1 + ... + c_k x_k$

and so

$\mathbb{E}[Y|t = 1, \underline{x}] - \mathbb{E}[Y|t = 0, \underline{x}] = b$

b is the 'treatment effect', *adjusted* for $x_1, ..., x_k$

Class data

```
> lm(handspan ~ gender)
Coefficients:
(Intercept)       genderM
     19.273         1.658
```

Male handspans are on average 1.66 cm wider than females

But is this just because men are taller? Need to 'adjust' for height

```
> lm(handspan ~ gender + height)
Coefficients:
(Intercept)       genderM        height
    8.06265       1.15941       0.06893
> confint(fitted)
                  2.5 %     97.5 %
genderM     -0.02769957  2.3465157
```

Adjusting for height, male handspan is still 1.16 cm wider than females. 95% confidence interval is (-0.03,2.34) so NOT significantly wider in men than women, when we 'adjust' for height.

Class data

```
 > summary(fitted)
Residuals:
    Min      1Q  Median      3Q     Max
-3.2637 -1.3112  0.1904  1.0422  3.8701 (summary of distribution of residuals)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.06265    4.54223   1.775   0.0833 ( t = Estimate/Std error)
genderM      1.15941    0.58781   1.972   0.0553 (2-sided P value, ¿ 0.05)
height       0.06893    0.02779   2.480   0.0173 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 1.737 on 41 degrees of freedom(estimate of sigma)
Multiple R-squared: 0.2699, Adjusted R-squared: 0.2343

F-statistic: 7.578 on 2 and 41 DF,  p-value: 0.001583
```

How the eggs factor reveals your personality: Outgoing people like them poached while frying fans are hot stuff in bed

- **Scientists researched the psychology of consumer choice**
- **The preference of how to cook an egg reveals personality traits and secrets about social class and even sex drive**

By DAILY MAIL REPORTER

PUBLISHED: 00:00 GMT, 29 September 2012 | **UPDATED:** 00:55 GMT, 29 September 2012

Poached, scrambled, boiled or fried. We all have our preference for how to cook an egg.

But the choice reveals more than just our culinary tastes – it also highlights our personalities and reveals secrets about social class and even sex drive.

Scientists quizzed 1,010 adults and found that poached egg eaters are outgoing, boiled egg lovers are disorganised, fried egg fans have a high sex drive, scrambled egg aficionados are guarded and omelette eaters are self-disciplined.

Daily Mail readers were shown to prefer scrambled eggs.

The study for the British Egg Industry Council was carried out by Mindlab International, which researches the psychology of consumer choice.

It found that the average poached egg-eater is likely to be happier than most.

The egg factor: How you prefer your egg cooked reveals your personality - boiled lovers are disorganised

Boiled egg-eaters run the greatest risk of getting divorced.

Fried egg fans are usually from the skilled working class and scrambled eggs are favoured by those without children.

Andrew Joret, of the British Egg Industry Council, said: 'It's amazing to think that just by knowing someone's favourite way of eating eggs, it's possible to gauge a large amount about who they are and what they are like. But it doesn't matter how you eat your eggs – they're still nutritious, versatile and great value.'

P-values and confidence intervals - putting them into words

- $P < 0.05$ 'Evidence' against the null hypothesis (i.e. evidence that the difference is real)
- $P < 0.01$ 'Strong evidence' against the null hypothesis
- $P < 0.001$ 'Very strong evidence' against the null hypothesis
- The P-value is *not* the probability of the null hypothesis being true.
- If $P < 0.05$, you cannot say there is less than 5% chance that H_0 is true (even if everyone said this about the Higgs Boson)
- If $P > 0.05$, just say there is 'no good evidence against H_0'. Do NOT say that therefore H_0 is true.

Remember: if a 95% confidence interval does not contain the null hypothesis, this is the same as $P > 0.05$

BBC News Sport Weather Trav

NEWS SCIENCE & ENVIRONMENT

Home | UK | Africa | Asia | Europe | Latin America | Mid-East | US & Canada | Business

4 July 2012 Last updated at 07:35 GMT

27K Share

Higgs boson-like particle discovery claimed at LHC

COMMENTS (1665)

By Paul Rincon
Science editor, BBC News website, Geneva

'Dramatic'

The CMS team claimed they had seen a "bump" in their data corresponding to a particle weighing in at 125.3 gigaelectronvolts (GeV) - about 133 times heavier than the protons that lie at the heart of every atom.

They claimed that by combining two data sets, they had attained a confidence level just at the "five-sigma" point - about a one-in-3.5 million chance that the signal they see would appear if there were no Higgs particle.

Is this a good interpretation? One or 2-sided?

1/pnorm(-5)

[1] 3488556

Which of you is psychic?

Guess which side the coin has come up.

20 flips, write down X or Y, and then circle it if it is right

Which of you is psychic?

Number right	1-sided P-value
13	0.132
14	0.058
15	0.021
16	0.006
17	0.001

Prob. of at least one P-value < 0.05 in 50 attempts $= 1 - 0.95^{50} = 0.92$

Prob. of at least one P-value < 0.01 in 50 attempts $= 1 - 0.99^{50} = 0.39$

Prob. of at least one P-value < 0.001 in 50 attempts $= 1 - 0.999^{50} = 0.05$

Multiple testing

If you do k independent hypothesis tests, then *even if the null hypothesis is true* -

Probability that at least one will have a P-value $P < 0.05$

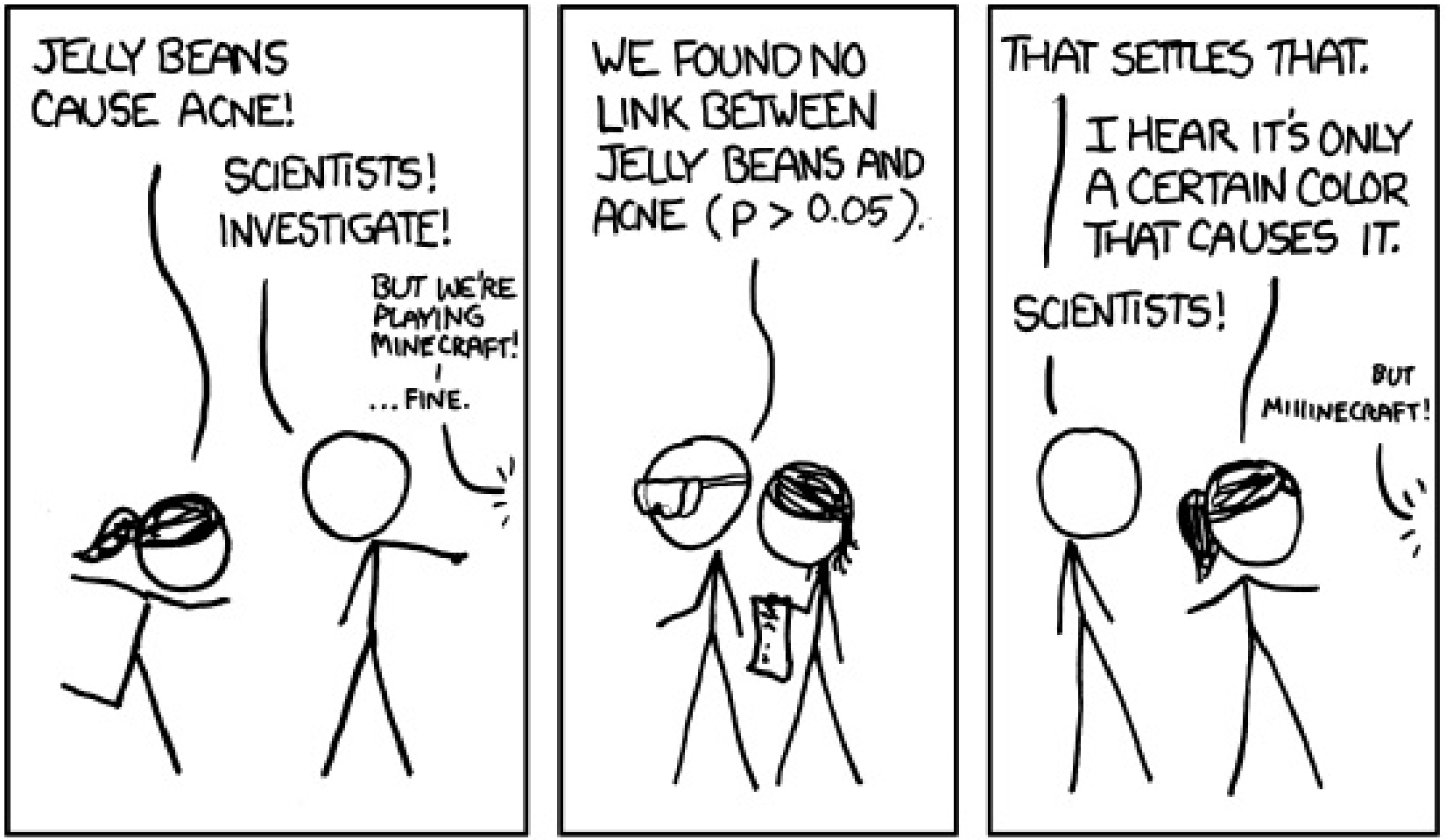
$= 1$- (Probability that they *all* will have a $P > 0.05|H_0$)

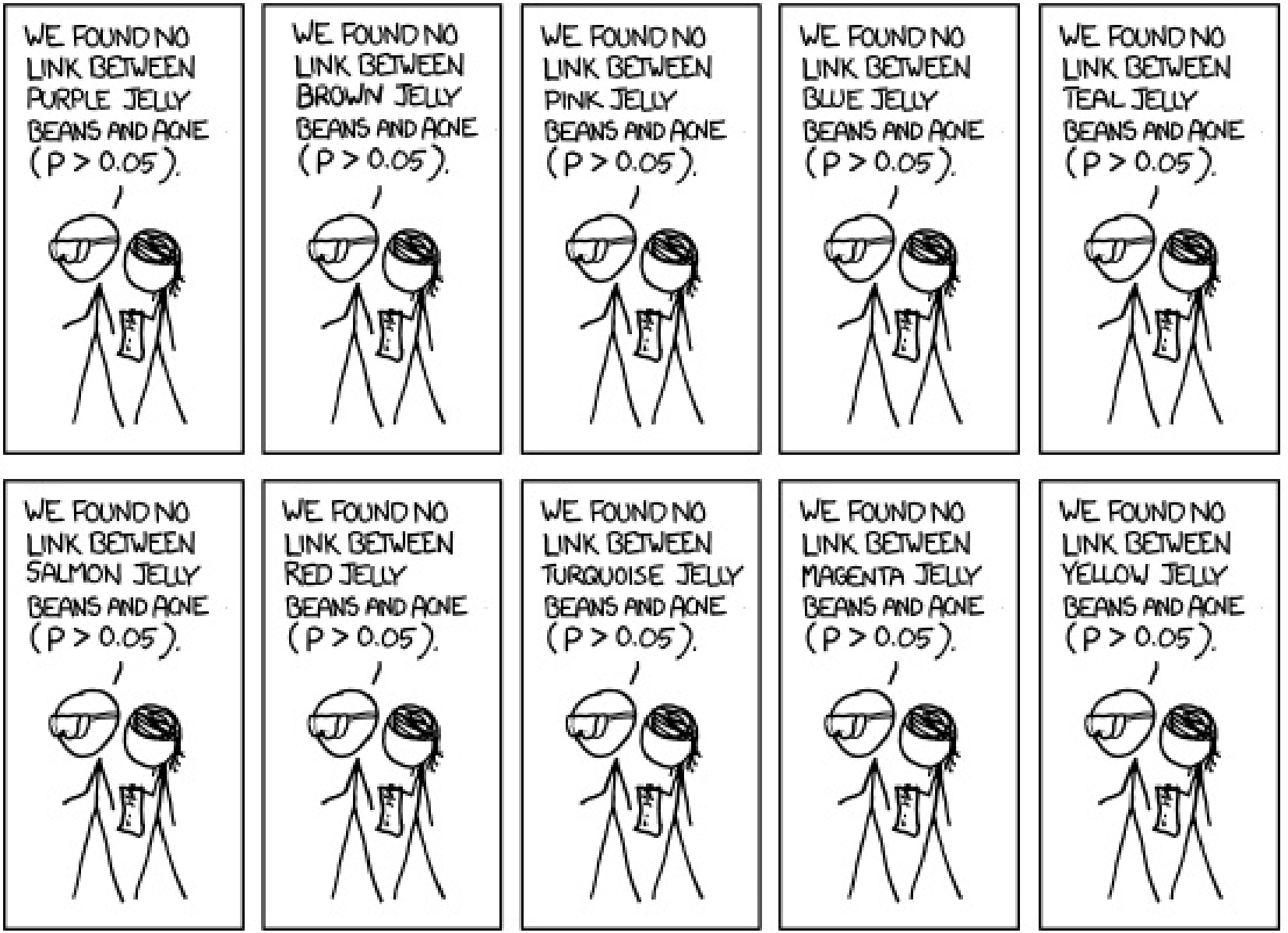
$= 1 - 0.95^k$

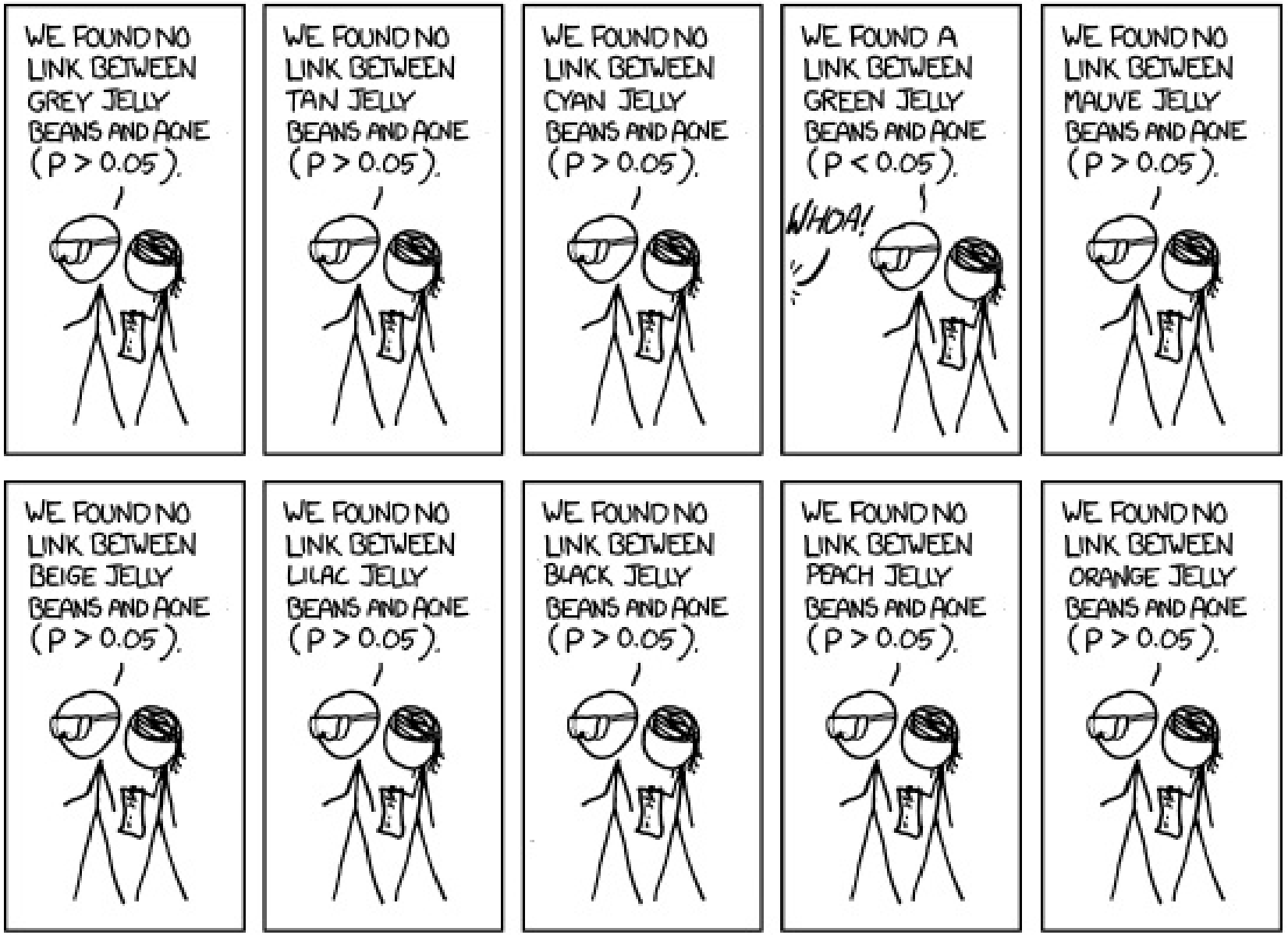
for example, if $k = 20$, then there is a $1 - 0.95^{20} = 0.64$ chance of (wrongly) rejecting H_0 in at least of the tests

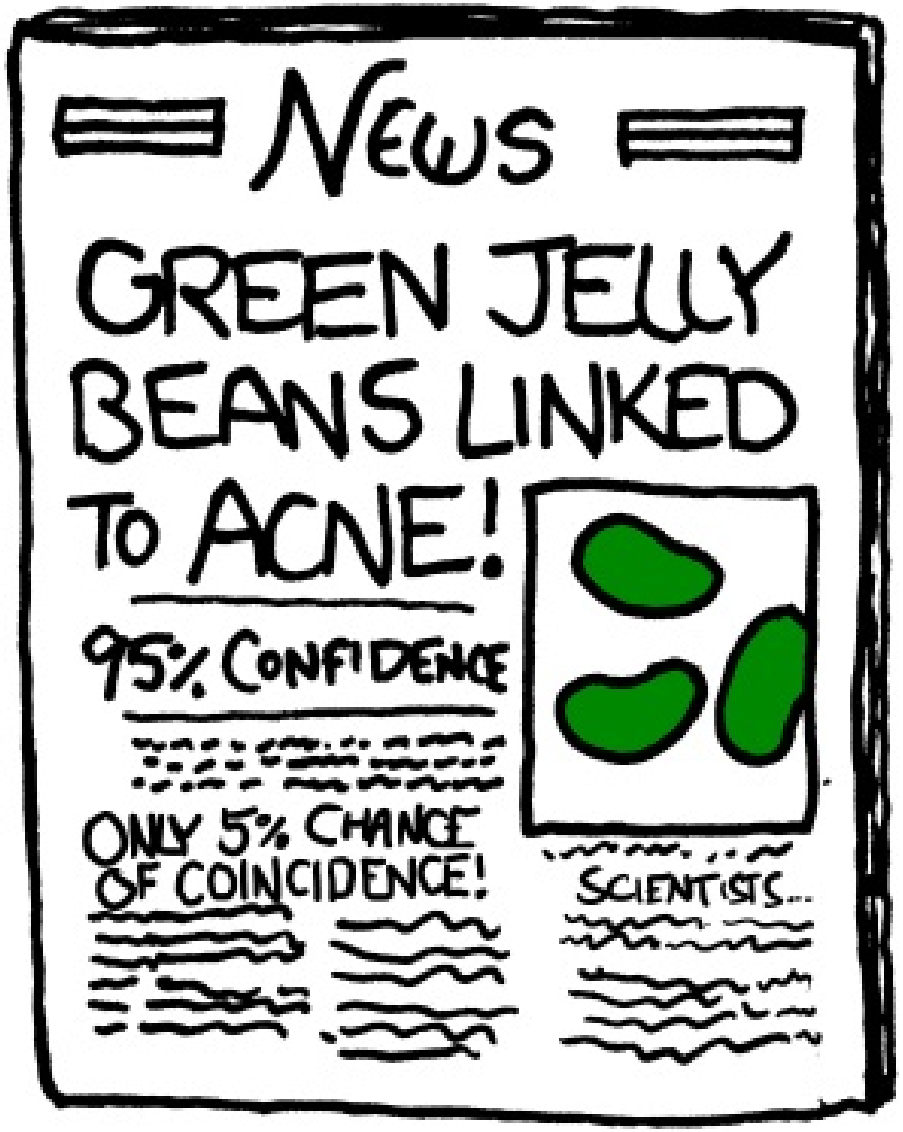
So if we do enough testing we are bound to reject H_0

This is the *multiple testing* problem


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


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...

What do people do about multiple testing problem?

- Adjust the P-values for multiple testing: e.g. the 'Bonferroni' adjustment multiplies P-values by the number of tests done.
- Use 99% confidence intervals if calculating many intervals.
- If there have been lots of tests, demand a VERY low P-value (e.g. $P < 0.000001$)
- Demand replication of the 'discovery' by other teams
- Try to control the 'False Discovery Rate' (out of things found to be significant, what proportion are 'false discoveries'?)

The Higgs Boson team did almost all of these!

Stats practical

Adapt commands from the previous .R files for the following.

1. Read in the data from the 'UN / Europe' experiment - it is in `UN-data.csv`.
2. Plot all the estimates as a histogram
3. Plot the estimates for each group X (prompt.X=10) and Y (prompt.X=60) as a histogram [see class.data.R for how this was done using the "lattice" package]
4. type the command `prompt.X==10` and see that it produces a logical array with TRUE/FALSE according to then value of prompt.X
5. Find the sample mean and variance for each group. You could select the two groups by using `prompt.X==10` in square brackets after `estimate` - this picks out the elements of `estimate` corresponding to TRUE

   ```
   x=estimate[prompt.X==10]   # select guesses of people given the prompt X=10
   y=estimate[prompt.X==60]    #  select guesses of people given the prompt X=60
   ```

6. Assuming each group represents a random sample from the population, estimate the standard error for each sample mean.
7. Find an approximate 95% confidence interval for the difference in population means, using the formula given previously

   $$\overline{X} - \overline{Y} \pm 1.96\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}.$$

   Do this calculation in R
8. Test the hypothesis that there is no difference between the groups (2-sided). You can use the `t.test` function. Compare the 95% confidence interval they give with the one you calculated above.
9. Put your conclusions into words!

Incidentally, out of 193 Member States of the UN, 48 (25%) are in Europe

`http://www.blatantworld.com/feature/europe/united_nations_member_states.html`

How does this compare with the opinions?

Do you think the 'wisdom of crowds' worked in this case?

Stats assignment

Data in `sac.indsex.dat` concern 706 unrelated South African Coloured TB cases and controls (cPHENO), age, sex and individual genetic ancestry proportion from Yoruba (cYRI), Khoesan(cSAN), European (cCEU), East-Asian (cCHB) and Indian (cGIH).

The question is to investigate the relationship between TB status (cPHENO), gender and genetic ancestry

1. Read in the data from the file `sac.indsex.dat` - this is not a CSV file so use `dat=read.table(....` instead of `dat=read.csv`
2. Check the content of the data-file with `summary` and listing it by typing `dat`. Check you understand what each column means (ignore the NULL column)
3. `attach` the file as in previous examples, so you can just use the variable names
4. Look at each variable using a suitable plot
5. Look at correlations between the variables using suitable plots. You might want to look at QuickR on scatter plots to see how to produce a nice table of scatterplots between many different variables - there are a number of ways of doing this
6. How are the different genetic ancestry proportions related to each other?
7. Check how people with TB (cPHENO=1) differ from those without TB (cPHENO=0) using appropriate tests and plots. Are they younger, and do they have different genetic ancestry?
8. You might use the `lm(cPHENO ~ ....)` function to see how multiple variables predict TB status. See slides in stats lecture 5.
9. Produce a short report, with graphs, summarising your conclusions