Numerical Methods & Data Analysis in HEP

Mike Williams

Department of Physics

Massachusetts Institute of Technology

IDPASC Lectures
January 29 – 31, 2013

Quick Points from Lecture 1

A few quick points based on discussions with people:

- The bootstrap clearly requires that the events are independent. In the variance on the height example, this doesn't work if half the people are related.
- You can often train your BDT on data. *E.g.*, in LHCb I used large sample of pure $B_s \to D_s(KK\pi)\pi$ data to train a $D_s(KK\pi)$ "from B" BDT. I used half the data to train and half to evaluate the efficiency of any cut. I then used this to look for $B_u \to D_s(KK\pi)\phi(KK)$ (very rare). The kinematics are very similar because the system is highly boosted; thus, in this case I end up with a BDT with 100 variables that I know the efficiency of to 1%. At a lower level, you can put your hardware in a test beam and use this data to train a ANN/BDT how to identify hit clusters, *etc.*

Lecture 2

Last time we talked about using machine learning techniques to classify your data. Now that we've got the data, we want to learn something from it. Today's lecture:

- what p-values are . . . and aren't;
- 2-sample tests;
- the permutation test;
- parametric regression;
- multivariate goodness-of-fit.

Tomorrow we'll do limits, Bayesians vs Frequentists and non-parametric regression.

Notation

I've tried to avoid notation as much as possible but it'll help to define a few simple things:

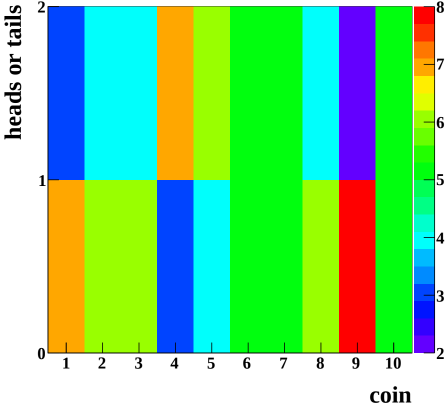
- $\vec{x}$: D-dimensional vector of all variables;
- $f(\vec{x})$: true PDF;
- $f_0(\vec{x})$: test (or fit) PDF;
- T: some test statistic that quantifies (in some way) agreement between the data and f_0 (choose smaller T to mean better agreement here . . . doesn't have to be this way in general).

You are probably familiar with at least one example of T: the χ^2 statistic.

A lot of this lecture is taken from M.Williams, *How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics*, JINST **5**, P09004 (2010).

χ^2

Say I flip 10 coins 10 times each. Are they fair?



Here the $\chi^2 = \sum (O_i - E_i)^2 / E_i = 8.4$ and $n_{\text{dof}} = 10$. Are they fair?

χ^2

- How do we know what values of χ^2 are good? Shouldn't the deviation of $\chi^2/n_{\rm dof}$ from 1 decrease as $n_{\rm dof}$ increases? Yes*.
- If I have a lot of coins and one of my coins is a heads-heads coin and the rest are fair, would looking at just the χ^2 let me know? No.

From this simple example we can see that:

- We need a quantity that tells us how well our data agrees with our hypothesis in some well-defined way.
- We need different tests for different problems. *E.g.*, are they fair is a different question than is each of them fair. The χ^2 is well suited to the former but not the latter.

*Looking forward: If $\chi^2/n_{\rm dof} = 1.1$, $p(n_{\rm dof} = 10) = 0.36$ but $p(n_{\rm dof} = 1000) = 0.01$.

p-values

If we denote the PDF of T as $g(T)$, which may depend on f_0, then the p-value is defined as:

$$p = \int_T^\infty g_{f_0}(T')dT'.$$

The p-value is the probability of finding a T-value corresponding to lesser agreement than the observed T-value if $f = f_0$. It is not the probability that $f = f_0$!

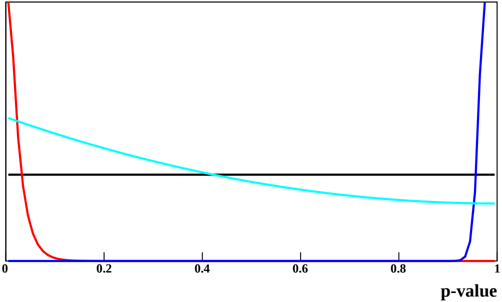
If $f = f_0$, then the p-value distribution is uniform on $(0,1)$. One can reject the hypothesis $f = f_0$ at confidence level α if $p < 1 - \alpha$; *e.g.*, the test hypothesis is rejected at 95% confidence level if $p < 0.05$.

- If $f = f_0$, $p < 0.01$ should happen in 1/100 experiments; rare but **should** happen. $p > 0.99$ should also happen 1/100. Be suspicious of these too.
- *N.b.*, one of the reasons χ^2 is so popular is that its g does not depend on f; however, this is in the limit $n \to \infty$. Everything you know about χ^2 is never strictly true on your data.

p-values

I give a physics test to 1M physics students and build g for $T =$ score. I give this test to somebody else to determine if they're a physics student.

physics students, art students, math students, professors



"Good" p-values don't mean $f = f_0$. The test may be insensitive to differences. Here, I should've asked their age (factor into T somehow) and asked less math questions (if these are alternatives I'm worried about).

2-Sample Tests

Say you collect 2 data sets: A and B. Sometimes, all you want to know is: Do A and B share the same parent PDF?

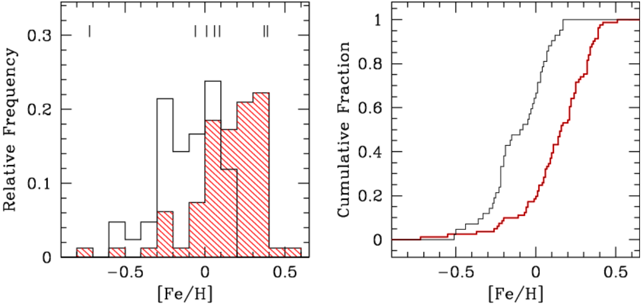
E.g.1, there may be some symmetry that relates A to B that our understanding of physics says should hold. If so, then the data should be consistent with sharing the same PDF. Simply observing the contrary would be a sign of new physics.

E.g.2, less grand: data measured from two experiments of the same process should produce the same result (or same experiment but different techniques or taken at different times, *etc*). Many systematic studies are based upon "these data should agree with those data".

How do we test this?

KS & χ^2 Tests

In 1-D, the Kolmogorov-Smirnov test is well-known and uses the max distance between the two empirical CDFs to obtain a *p*-value. It does not work in higher dimensions.



The 2-sample χ^2 test can be used in higher dimensions; however, binning often leads to quick defeat at the hands of the curse of dimensionality.

KS & χ^2 Tests

If you're in 1-D, then the KS test is a good option. For low D and large sample sizes, the χ^2 test works well. Also, there are plenty of similar tests out there: Cramer-von Mises, Anderson-Darling, Watson, . . . In many cases these are more powerful than KS.

These are both great methods and you shouldn't hesitate to use them if they apply; however, you will at some point (perhaps often) find yourself in a position where neither of these works.

So let's see what other options are out there . . .

Distance

Let's pause for a moment and define distance in a multivariate space. We'll need this a lot so I want to make sure it's clear.

One (popular) choice of a distance metric is the *normalized Euclidean distance*

$$|\vec{x}_i - \vec{x}_j|^2 = \sum_{v=1}^{D} \left(\frac{x_i^v - x_j^v}{w_v} \right)^2,$$

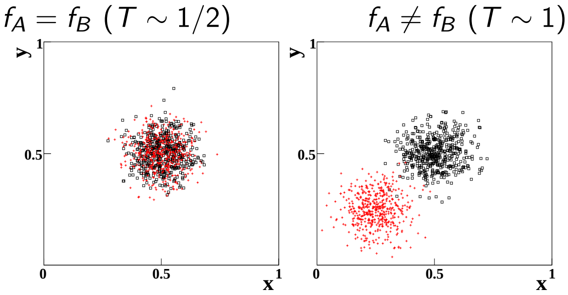
where w_v are weights for each variate. A common choice is to use the RMS for each variate. It's somewhat arbitrary, just like choosing a binning scheme is.

The conclusions drawn from any test that depends on distance shouldn't depend strongly on the choice of metric.

N.b., distance-based anything doesn't work well with mixed input types (discrete and continuous).

k-Nearest-Neighbor Test

Mixing between data samples is only optimal if they share the same PDF. The kNN T is the mean fraction of like-sample NN events in the pooled sample of the two data sets.



$f_A = f_B \ (T \sim 1/2)$

$f_A \neq f_B \ (T \sim 1)$

$\langle T \rangle$ is larger for the case $f_A \neq f_B$ due to the lack of complete mixing of the two samples that occurs if their parent distributions are not the same.

- Rule of thumb: $k = 10$ tends to work well for most applications.
- See M.F. Schilling, J. Amer. Statistical Assoc. **81**, No. 395 (1986) 799-806.

Energy Test

The potential energy of a system of charged point particles interacting via potential $\psi(\Delta\vec{x})$ $(\Delta\vec{x} \equiv |\vec{x} - \vec{x}'|)$ is given by

$$\begin{aligned} T &= \tfrac{1}{2} \int \int \left(f_A(\vec{x}) - f_B(\vec{x})\right)\left(f_A(\vec{x}') - f_B(\vec{x}')\right)\psi(\Delta\vec{x}) d\vec{x} d\vec{x}' \\ &= \tfrac{1}{2} \int \int \left[f_A(\vec{x})f_A(\vec{x}') + f_B(\vec{x})f_B(\vec{x}') - 2f_A(\vec{x})f_B(\vec{x}')\right]\psi(\Delta\vec{x}) d\vec{x} d\vec{x}', \end{aligned}$$

which is minimal when $f_A = f_B$ (think about $\psi(x) = 1/x$ and recall your E&M courses; better choice here is $\psi(x) = -\log(x + \epsilon)$).

Given that we want to test this same hypothesis, this quantity seems like it may be useful. One problem: We don't know f_A or f_B!

Turns out, we don't need to . . .

B. Aslan and G. Zech, *Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding,* Nucl. Instrum. Methods **A537** (2005) 626-636.

Energy Test

T can be estimated without the need for any knowledge about the forms of f_A and f_B using the data:

$$T \approx \frac{1}{n_A(n_A-1)} \sum_{i,j>i}^{n_A} \psi(\Delta\vec{x}_{ij}) + \frac{1}{n_B(n_B-1)} \sum_{i,j>i}^{n_B} \psi(\Delta\vec{x}_{ij}) - \frac{1}{n_A n_B} \sum_{i,j}^{n_A,n_B} \psi(\Delta\vec{x}_{ij}).$$

This is simply the previous equation rewritten using the standard Monte Carlo integration approximation, along with the fact that $\int f_A(\vec{x})d\vec{x} = \int f_B(\vec{x})d\vec{x} = 1$ ($n.b.$, $\int \phi(\vec{x})f(\vec{x})\mathrm{d}\vec{x} \approx \frac{1}{n}\sum \phi(\vec{x}_i)$)

The data is sampled from the PDFs, whatever they are, so we're done! Except that we don't know what distribution T follows.

How do we get a p-value?

Permutation Test

If $f_A = f_B$, then we can treat A and B as just labels. *I.e.*, for each event we recorded, there was an equal chance of it being in the A or B data set. So, any relabeling is just as likely as the what we measured.

example test with very small data sets

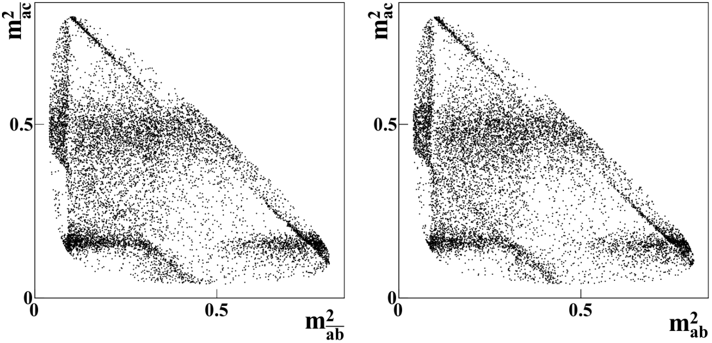
A_0, A_1, A_2	B_0, B_1, B_2	$\rightarrow T$
A_0, A_2, B_0	A_1, B_1, B_2	$\rightarrow T_0$
A_1, B_1, B_2	A_0, A_2, B_1	$\rightarrow T_1$
$\vdots$	$\vdots$	$\vdots$

Permutation test: randomly assign n_A data label A and the remaining n_B the label B; recalculate T. The p-value is fraction where $T < T_i$ is true.

This technique (Fisher 1935) works for any 2-sample test.

CP Violation

Ex) Searching for *CP* violation in the (fictitious) $X \to abc$ decay and it's charge conjugate. If *CP* is conserved, the two PDFs are the same.

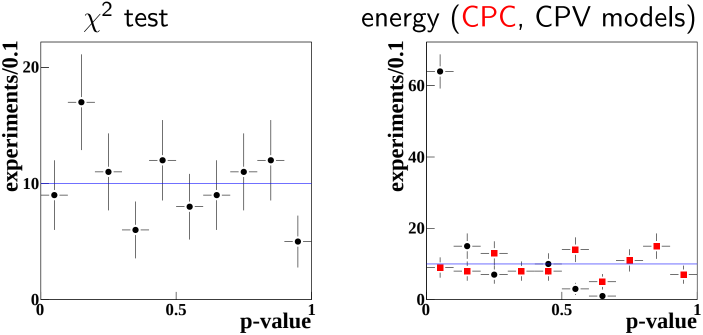


The model has a single CPV phase: $\Delta\phi_{1^-_{ac}} = 10^\circ$. Can you see the difference between the 2 data samples?

See M. Williams, *Observing CP violation in many-body decays*, Phys.Rev.D 84, 054015 (2011).

CP Violation

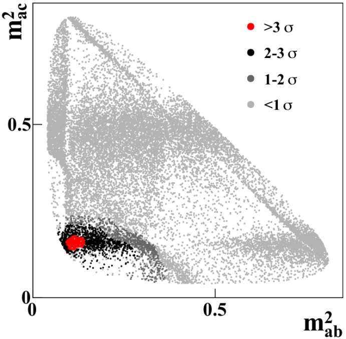
The χ^2 test is *blind* to this tiny amount of CPV; the *p*-value distribution is consistent with uniform despite the fact $f_A \neq f_B$.



The energy test can *see* even this small discrepancy between the two PDFs. It rejects $f_A = f_B$ over 50% of the time at the 95% CL.

CP Violation

Typical physicist complaint: *Sure it works great but what good is it if I don't know where the CPV is in the Dalitz plot?*



So I invented a visualization tool. It shows the pooled sample with each event in a CL band. In this example the discrepancy is where the resonance with the CPV phase interferes with another resonance (as expected).

2-Sample Tests Summary

- The KS test only works in 1-D but it is easy to use and works well.
- The χ^2 test works in any D but becomes limited by the curse of dimensionality quickly.
- The kNN test also works in any D and is easy to use but, in my experience, is only moderately more powerful than the 2-sample χ^2.
- The energy test is the most powerful 2-sample test I've ever used. It greatly outperforms the χ^2 test every time I've put them against each other. It takes a lot of CPU power to run but, in my experience, it is worth it (why spend \$100M on a detector and then balk at a few CPU days?).
- The permutation test works on any of these tests.
- If you don't care about f, why spend time trying to parametrize it?

Regression

Regression involves determining the relationship between a set of dependent variables (parameters) and independent variables (data points). The name originates from one of its early uses in biology involving *regression to the mean*.

In physics, regression is (almost) always either:

- χ^2 minimization: the data are binned (in some number of dimensions) and the quantity $\chi^2 = \sum (o_c - e_c)^2/e_c$, where $o_c(e_c)$ is the number of observed(expected) counts in bin c, is minimized by varying the parameters;
- MLE: the likelihood function $\mathcal{L} = \prod f_0(\vec{x}_i)$ is maximized (really $-\log \mathcal{L}$ is minimized) by varying the parameters (sometimes the *extended* likelihood is used if $n_{\rm observed}$ is relevant).

I assume you're familiar with these techniques (stop me if you're not!).

Goodness-of-Fit Tests

A *goodness-of-fit* test describes how well a model describes a data set. GOF tests should produce *p*-value distributions that are uniform when testing the true PDF.

Need to know what alternatives we're *scared* of. Otherwise, there's no way to choose a test (remember my physics exam example?).

The χ^2 test (Pearson 1900) is the *classic* example. The χ^2 test is good but we generally expect unbinned methods to perform better (especially with sparse data). What other choices are there?

2-Sample Tests

All of the 2-sample tests we talked about earlier can be used here too. All you need to do is generate a large MC sample from the test PDF (f_0) and use that as one of the samples and the data as the other.

Of course, there will be some bias due to the fact that f_0 is determined from the data; however, provided the correlation between the fit statistic (*e.g.*, χ^2 or $\log\mathcal{L}$) and T is small, this shouldn't be an issue.

I've tested the energy test on Dalitz-plot fits where f_0 was determined from an unbinned maximum likelihood fit and found the bias in the p-value to be at the few percent level (compatible with using the asymptotic χ^2 p-values).

K-Function Test

The K-function was originally introduced by Ripley to test for 2-D uniformity and updated by Baddeley *et al.* to work in any D for any f.

Math

$$K(r) \propto \sum_{i=1}^{n} \sum_{j \neq i} \frac{I(|\vec{x}_i - \vec{x}_j| < r)}{v(i,j) f_0(\vec{x}_i) f_0(\vec{x}_j)}, \; L(r) = K(r)^{1/D},$$

where $v(i,j)$ is a volume edge-correction factor. If $f_0 = f$, then $\langle L(r) \rangle = r$. To get p use $T = (L(r) - r)_{\max}$ (similar to KS test).

Words

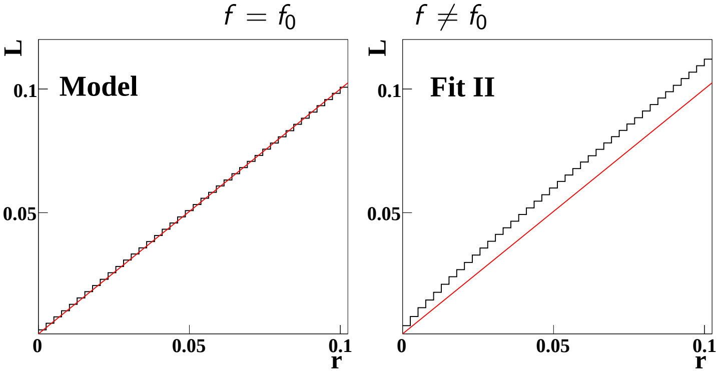
Basically, compares local event density (out to radius r) around each event to the expectated value.

B.D. Ripley, *Modelling spatial patterns*, J. Roy. Stat. Soc. B Met. **39**, No. 2 (1977) 172-212.

A.J. Baddeley, J. Møller and R. Waagepetersen, *Non- and semi-parametric estimation of interaction in inhomogeneous point patterns*, Stat. Neerl. **54**, Issue 3 (2000) 329-350.

K-Function Test

Example from MC (true PDF known). I found K (in 2-D Dalitz analysis) to be much more powerful than χ^2 and comparable to the energy test.



$f = f_0$ $f \neq f_0$

Can also go "low" but only if events aren't IID (common in biology).

Other Tests

There are **many** other tests on the market:

- kernel-based tests;
- distance to nearest neighbor based tests;
- *etc.*

There is no uniformly most powerful GOF test. For most applications in physics though, you can use the same test and it will be "good enough". Beware when it isn't!

For some analyses it pays to use a specialized test. This is fine . . . but be sure you know the test works. *p*-values should be uniform on test MC when you have the model right.

Confidence Intervals

After obtaining an estimator $\hat{x}$ we quote $\hat{x}^{+u}_{-l}$. If we repeat our experiment many times, 68% of the $(\hat{x} - l, \hat{x} + u)$ contain the true value of x. Notice that it is **not** the probability that the true value of x is in the interval (which is either 0 or 1 for any interval).

How do we get them? Ex.) Obtain $\hat{x}$ by maximizing the likelihood. In the limit $n \to \infty$, $\mathcal{L}$ is a parabola and $\Delta\mathcal{L} = \pm 1/2$ gives the 68% CL interval. For finite n, not true. May or may not give a good estimate for the CI.

Rule of thumb: Asymptotic stats rules tend to converge surprisingly fast. The obvious place they fail miserably though is when you measure something forbidden in the $n \to \infty$ limit (*e.g.*, less events than the bkgd expectation).

Is there a *safe* approach? Yes, toy MC (but takes a lot of CPU).

Significance

The statistical significance is the probability that a given result would occur by chance. In physics it is quoted in "units" of σ; *i.e.*, how many σ away from the mean of a Gaussian one would need to be to obtain the same *p*-value (*n.b.* $10\sigma \sim 10^{-23}$; out here nonsense lies).

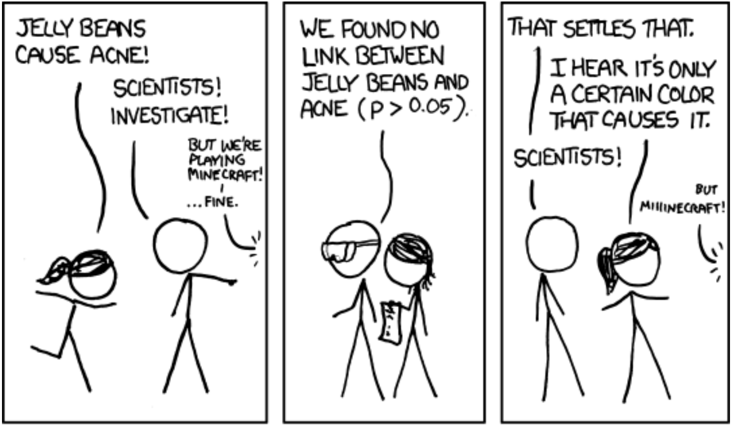
Wilks Theorem

For *nested* models $-2\Delta \log \mathcal{L}$ is asymptotically χ^2 distributed with $n_{\rm dof} = \Delta n_{\rm par}$. Physicists often screw this up:

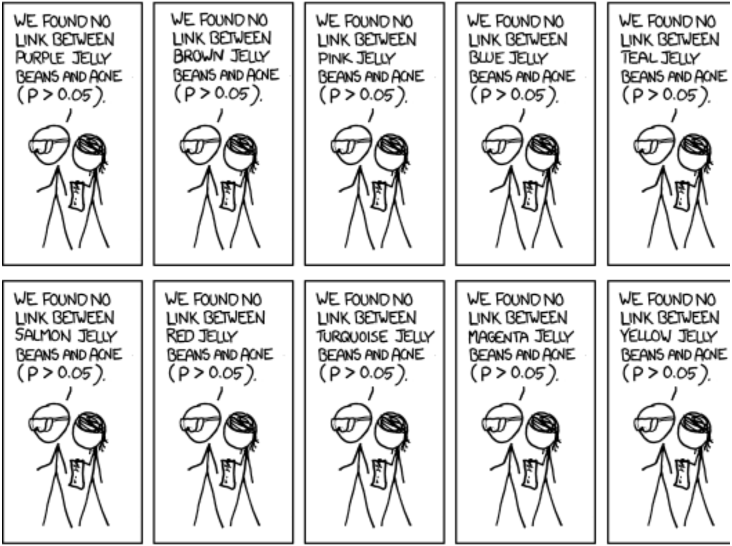
- the models **must** be nested (*e.g.*, a signal Gaussian with mean and/or width free can**not** be nested with a no signal model!);
- converting this to $n\sigma$ using $\sqrt{-2\Delta \log \mathcal{L}}$ is only (possibly) valid if $\Delta n_{\rm par} = 1$;
- this is **asymptotically** correct only.

Toy MC (again) is the way around most of these problems.

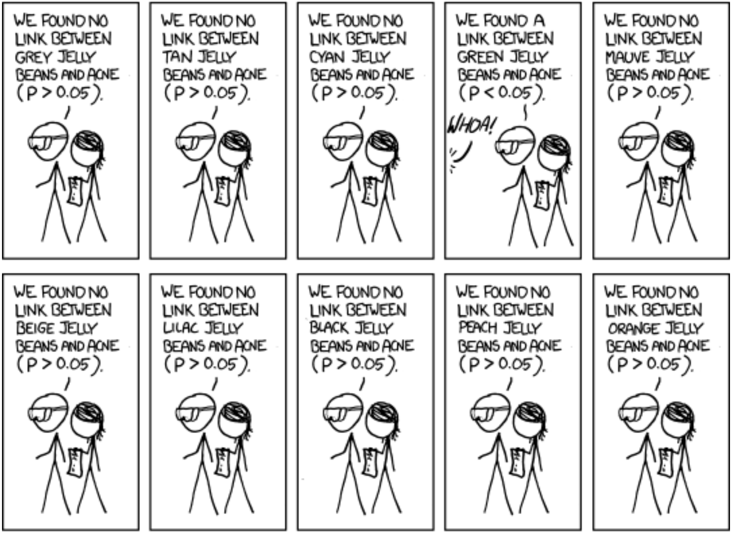
Look Elsewhere Effect


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 MIIINECRAFT!

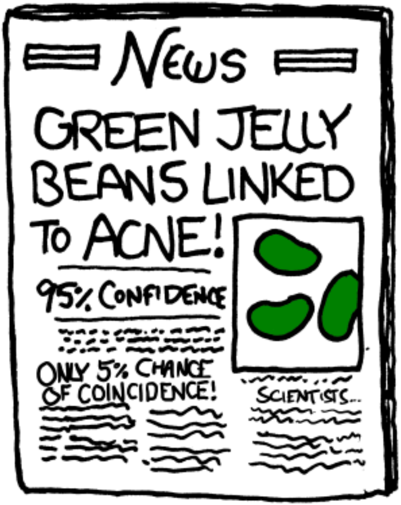
Look Elsewhere Effect


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).

Look Elsewhere Effect


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).

Look Elsewhere Effect


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

Summary

- *p*-values are not probabilities for $f = f_0$. They should be uniform if $f = f_0$. (if uncertain, check that you get uniform *p*-values).
- *p*-values from a χ^2 (or any other) limiting distribution are approximations. Can use MC or the permutation test instead (takes a lot of CPU to do if you want to go to 5σ).
- If you don't care about the PDF, don't try and fit for it. Use a 2-sample test.
- The energy test is a very powerful 2-sample/GOF test. Takes a lot of CPU but CPU is cheap compared to building/running an experiment.
- CI and significance formulas are asymptotic; often need MC.
- There is no uniformly most powerful GOF test. There are many options on the market. Make sure you get uniform *p*-values and test(s) have power against alternatives you're worried about.