

Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2008
The University of Texas at Austin

Unit 4: Tail Test Perils and Pitfalls: Chi-Square
Misuse, Multiple Hypotheses, Stopping Criteria

Unit 4: Tail Test Perils and Pitfalls (Summary)

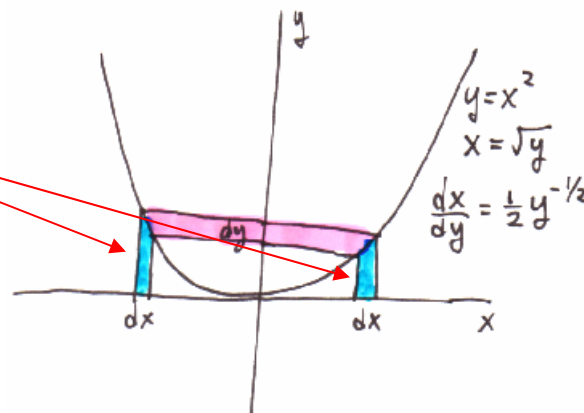
- Compute the PDF for a single term in the χ^2 sum
 - and compute its moments
 - and thus how χ^2 becomes Normal for large ν
- Fallacy of using “t-values” from Poisson statistics
 - even (especially!) in the CLT limit of large ν
 - but learn how to fix the problem by the correct Normal approximation
- How linear constraints affect χ^2
 - slice through origin of a (hyper-)spherically symmetric ball
 - therefore just reduces ν by 1
- Bonferroni corrections for multiple hypotheses
 - take $\alpha \rightarrow \alpha/N$, the most conservative possible correction
 - slavish adherence to Bonferroni is a curse on biomedical research
 - but it is better than the alternative of having a literature full of wrong results
- “Stopping rule paradoxes”
 - answer depends not only on the test, but on the stopping rule
 - sometimes causes tension between ethics and experimental design
 - Bayesian analysis is generally independent of stopping rules
- False Discovery Rate (FDR) method
 - alternative to Bonferroni, limits fraction of false discoveries
 - Benjamini-Hochberg criterion

Chi-square is so important that we need to understand its limitations, and/or how to overcome them.

You can get a statistic that is “accurately” chi-square **either** by summing (any number of) terms that are accurately squares of normal t-values, **or** by summing a large number of terms that individually have the correct mean and variance. This uses the CLT, so the exactness of chi-square is no better than its normal approximation.

Exact distribution of an individual squared t-value:

$$p_Y(y) dy = 2p_X(x) dx$$



$$\text{So, } p_Y(y) = y^{-1/2} p_X(y^{1/2}) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$$

This is (by definition) the chi-square distribution with one d.f.

Compute moments of chi-square with 1 d.f.:

```
In[31]:= py = (1 / (Sqrt[2 Pi y])) Exp[-(1 / 2) y]
```

```
Out[31]=
```

$$\frac{e^{-y/2}}{\sqrt{2\pi} \sqrt{y}}$$

```
In[32]:= Integrate[py {1, y, y^2}, {y, 0, Infinity}]
```

```
Out[32]=
```

```
{1, 1, 3}
```

```
syms y pi
py = exp(-.5*y) / sqrt(2*pi *y);
moments = int([1 y y^2]*py, 0, Inf)
moments =
[ 1, 1, 3]
```

So, $\mu = 1$, $\sigma^2 = 3 - 1 = 2$

Hence, $\text{Chisquare}(\nu) \rightarrow \text{Normal}(\nu, \sqrt{2\nu})$ as $\nu \rightarrow \infty$

If you are going to rely on the CLT and sum up lots of not-exactly-t bins, it is really important that they have the expected mean and variance.

Example: Chi-square test with small numbers of Poisson counts in some or all bins. **(People often get this wrong!)**

Recall Poisson:
$$p(n) = e^{-\mu} \frac{\mu^n}{n!}$$

```

syms mu n
poi = exp(-mu) * mu^n / gamma(n+1);
poi mean = symsum(n * poi, 0, Inf)
poi mean =
mu
poi var = simplify( symsum(n^2 *
    poi, 0, Inf) - poi mean^2 )
poi var =
mu
tmean = symsum(((n-mu)^2 / mu) *
    poi, 0, Inf)
tmean =
1
tvar = simplify( symsum(((n-mu)^2 /
    mu)^2 * poi, 0, Inf) - tmean^2)
tvar =
(2*mu+1)/mu

```

```

In[39]:= poi[n_] := Exp[-mu] mu^n / n!

In[48]:= poimean = Sum[n poi[n], {n, 0, Infinity}]
Out[48]=
mu

In[50]:= poivar =
Simplify[Sum[n^2 poi[n], {n, 0, Infinity}] -
    poimean^2]
Out[50]=
mu

In[51]:= tmean = Sum[ ((n - mu) ^2 / mu) poi[n], {n, 0, Infinity}]
Out[51]=
1

tvar =
Simplify[
    Sum[ ((n - mu) ^2 / mu) ^2 poi[n], {n, 0, Infinity}] -
    tmean^2]

Out[53]=
2 + 1/mu

```

So χ^2 is not Chi-square distributed! Rather, asymptotically,

$$\chi^2 \sim \text{Normal} \left(\nu, 2\nu + \sum_i n_i^{-1} \right)$$

What about zero bins? Ignore them, and don't increment ν .

“What’s the deal” on this decreasing v by the number of constraints?

$$t_i = \frac{x_i - \mu_i}{\sigma_i} \sim \text{Normal}(0, 1)$$

joint distribution on all the
t's, if they are independent

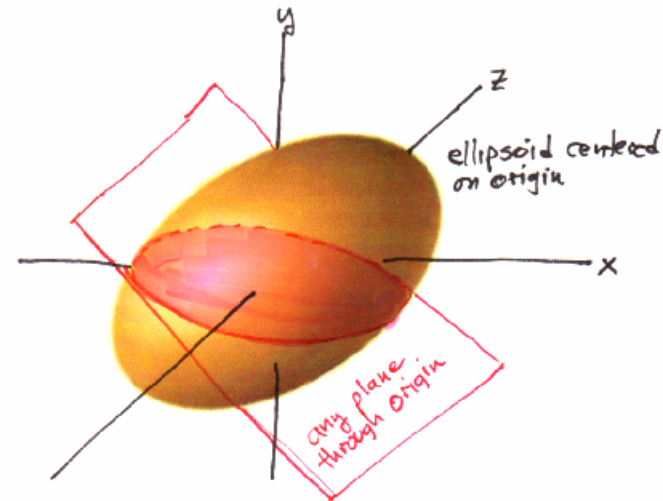
$$p(\mathbf{t}) = \prod_i p(t_i) \propto \exp\left(-\frac{1}{2} \sum_i t_i^2\right)$$

χ^2 is squared distance from origin $\sum t_i^2$

$$\text{constraint } \sum_i \alpha_i x_i = C \Rightarrow \sum_i \alpha_i \sigma_i t_i = 0,$$

$$(\text{using } \langle \sum \alpha_i x_i \rangle = \sum \alpha_i \mu_i = C)$$

Constraint is a cut. Any cut through an ellipsoid is an ellipse; any cut through a sphere is a circle.



So the distribution of distance from origin is the same as a multivariate normal “ball” in the lower number of dimensions, q.e.d.

Note: individual t 's needed to be normal. Not right if they merely have correct moments and try to use large v CLT. People often do this wrong. (Perhaps example of why resulting p-values not exactly uniform?)

(Discuss nonlinear constraints.)

So what's wrong with p-value tests?

1. Bonferroni is very conservative



Carlo Emilio Bonferroni
(1892 – 1960)

α = prob. that none of N tests will accidentally fall in their critical regions α'

$$\alpha = 1 - (1 - \alpha')^N \approx N\alpha'$$

This assumes that the N tests are all independent. That's surely not true, because, at the very least, they are looking at the same data!

The opposite limit would be to repeat the same test N times on the same data (N non-communicating graduate students open the same statistics book).

$$\alpha = \alpha'$$

The truth is always somewhere in-between.

Slavish adherence to Bonferroni is a curse on biomedical research, but it is better than the alternative of having a literature full of wrong results!

[Arnie Levine's story here.]

For large-scale screens can use False Discovery Rate (FDR) instead. (Will discuss soon.)

2. The result depends on the choice of test or (more argumentatively) what was in the mind of the experimenter

“Stopping rule” paradoxes.

Hypothesis H_0 : a coin is fair with $P(\text{heads})=0.5$

Data: in 10 flips, the first 9 are heads, then 1 tail.

Analysis Method I. Data this extreme, or more so, should occur under H_0 only

$$\frac{1 + 10 + 10 + 1}{2^{10}} = 0.0214$$

(you lose: referee wants $p < 0.01$ and tells you to get more data)



Analysis method II.

“I forgot to tell you,” says the experimenter, “my protocol was to flip until a tail and record $N (=9)$, the number of heads.”

$$\text{Under } H_0 \quad p(N) = 2^{-(N+1)}$$

$$p(\geq N) = 2^{-(N+1)} \left(1 + \frac{1}{2} + \frac{1}{4} + \dots\right) = 2^{-N}$$

$$P(\geq 9) = 2^{-9} = 0.00195$$

(Nature hold the presses!)

Stopping rule effects are a serious methodological issue in biomedical research, where for ethical reasons stopping criteria may depend on outcomes in complicated and unpredictable ways, or be ad hoc after the experiment starts (and rightly so – see next slide!)



April 8, 2006

British Rethinking Rules After Ill-Fated Drug Trial

By [ELISABETH ROSENTHAL](#),
International Herald Tribune

In February, when Rob O. saw the text message from Parexel International pop up on his cellphone in London — "healthy males needed for a drug trial" for £2,000, about \$3,500 — it seemed like a harmless opportunity to make some much-needed cash. Parexel, based in Waltham, Mass., contracts with drug makers to test new medicines.

Just weeks later, the previously healthy 31-year-old was in intensive care at London's Northwick Park Hospital — wires running directly into his heart and arteries, on dialysis, his immune system, liver, kidneys and lungs all failing — the victim of a drug trial gone disastrously bad.

One of six healthy young men to receive TGN1412, a novel type of immune stimulant that had never before been tried in humans, Rob O. took part in a study that is sending shock waves through the research world and causing regulators to rethink procedures for testing certain powerful new drugs.

Although tests of TGN1412 in monkeys showed no significant trouble, all six human subjects nearly died. One is still hospitalized and the others, though discharged, still have impaired immune systems, their future health uncertain.

On Wednesday, after releasing its interim report on the trial as well as previously confidential scientific documents that were part of the application for a trial permit, the British government announced it was convening an international panel of experts to "consider what necessary changes to clinical trials may be required" for such novel compounds.

The outcome "could potentially affect clinical trials regulation worldwide," the announcement said. In statements this week, both Parexel and **the drug's manufacturer, TeGenero, emphasized that they had complied with all regulatory requirements and conducted the trial according to the approved protocol.** But they declined to answer questions e-mailed to them about the specifics of the science involved.

"The companies have worked according to strict standards applicable for such type of studies," said Kristin Kaufmann, a spokeswoman for TeGenero.



What would be a Bayesian approach?

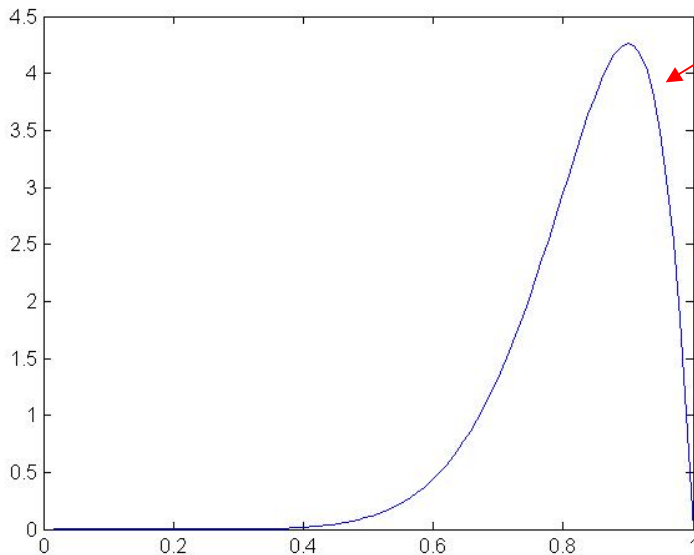
H_p is the hypothesis that prob = p .

$P(H_p)$ is its probability.

$$P(H_p|\text{data}) \propto P(\text{data}|H_p)P(H_p) \propto p^9(1-p)$$

$$P(H_p|\text{data}) = \frac{p^9(1-p)}{\int_0^1 p^9(1-p)dp}$$

```
xx = linspace(0, 1);  
plot(xx, betapdf(xx, 10, 2), '-');
```



The curve is the answer.
We might, however, summarize it in various ways:

```
y1 = betapdf(.5, 10, 2)  
y1 =  
    0.1074  
y2 = betapdf(.9, 10, 2)  
y2 =  
    4.2616  
quad(@(x)betapdf(x, 10, 2), 0, .5)  
ans =  
    0.0059
```



For an example in which we might use a more sophisticated prior, suppose the data is **10 heads in a row**.

“Hmm. When people make me watch them flip coins, 95% of the time it’s a (nearly) fair coin [A], 4% of the time it’s a double-headed [B] or double-tailed coin [C], and 1% of the time something else weird is happening [D].”

Case A:	$0.95 \times (0.5)^{10} = 0.00093$	0.043
Case B	$0.02 \times 1^{10} = 0.02$	0.915
Case C	$0.02 \times 0^{10} = 0$	0.000
Case D	$0.01 \times \int_0^1 p^{10} dp = 0.00091$	0.042

This kind of analysis is not usually publishable, unless you can justify your choice of prior on the basis of already published data. (In such a case it is dignified by the term “meta-analysis”.) However, it is a good way to live your life, especially if you are a person who likes to make bets!



(Can you remember that we were listing things bad about p-value tests?)

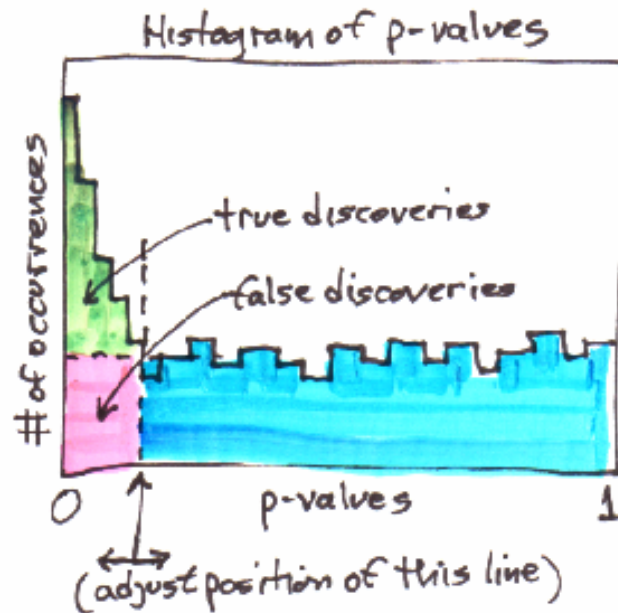
3. Not suitable for comparing hypotheses quantitatively. Best you can do is rule one out, leaving the other viable. Ratio of p-values is not anything meaningful!

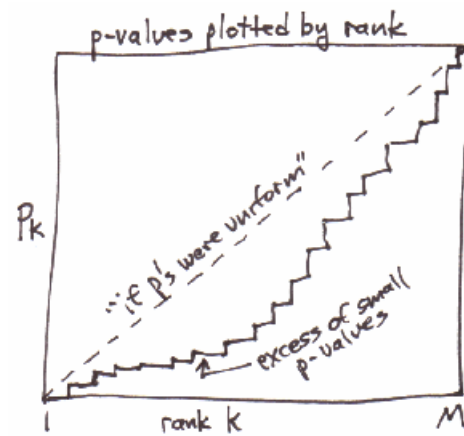
you should go learn about Likelihood Ratio tests, but I personally think that Bayes odds ratio is easier to compute and easier to interpret

False Discovery Rate (Benjamini & Hochberg)

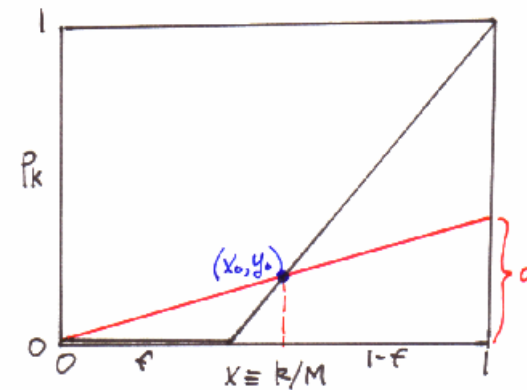
This is often a good alternative to Bonferroni, when the latter is too conservative.

- You have a lot of p-values
 - e.g., one per drug for 1000 drugs
 - or, one per gene for 10000 genes
- They are not uniform
 - there is an excess at small values
 - so some must be “causal”
- How do you set p to control α , the fraction of discovery calls that are false?
 - say, $\alpha = 5\%$





idealized version:



Prescription: call as discoveries all $p_j \leq \frac{j}{M} \alpha$

Proof:

$$\alpha x_0 = \frac{x_0 - f}{1 - f} \Rightarrow x_0 = \frac{f}{1 - \alpha(1 - f)}$$

$$\Rightarrow \text{FDR} = \frac{x_0 - f}{x_0} = \alpha(1 - f) < \alpha$$

(There are fancier proofs for the nonidealized version.)

OK, enough p-values for now.
Let's get back to random number generators.