CS395T
Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

Lecture 8

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]."

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

This kind of analysis can be dignified by the term "meta-analysis" if you can justify your choice of priors on the basis of already published data. (Somewhat more rigorously than the above.) However, it is also 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 three Bayesian criticisms of tail tests?)

(2) 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

(3) The sanctification of certain p-values (e.g., the magic p=0.05 value) is naïve and misleading.

(on the one hand) 1 in 20 results are wrong! Imagine if we built nuclear power plants to this low a standard.

(on the other hand) the large majority of results with p=0.10 are in fact correct. These could sometimes be acted on.

Slavish adherence to p=0.05 is largely due to the young Fisher (who became arguably the greatest statistician to ever have lived).

Ronald Aylmer Fisher (1890-1962)

Fisher studied with Gossert (Student) as a young man. Gossert never approved of the p=0.05 rule, and understood as the Master Brewer that no single p-value was suitable for optimizing economic return: it depends on the relative costs of success and failure (origins of decision theory).

There is a fun article on this posted in the course forum:

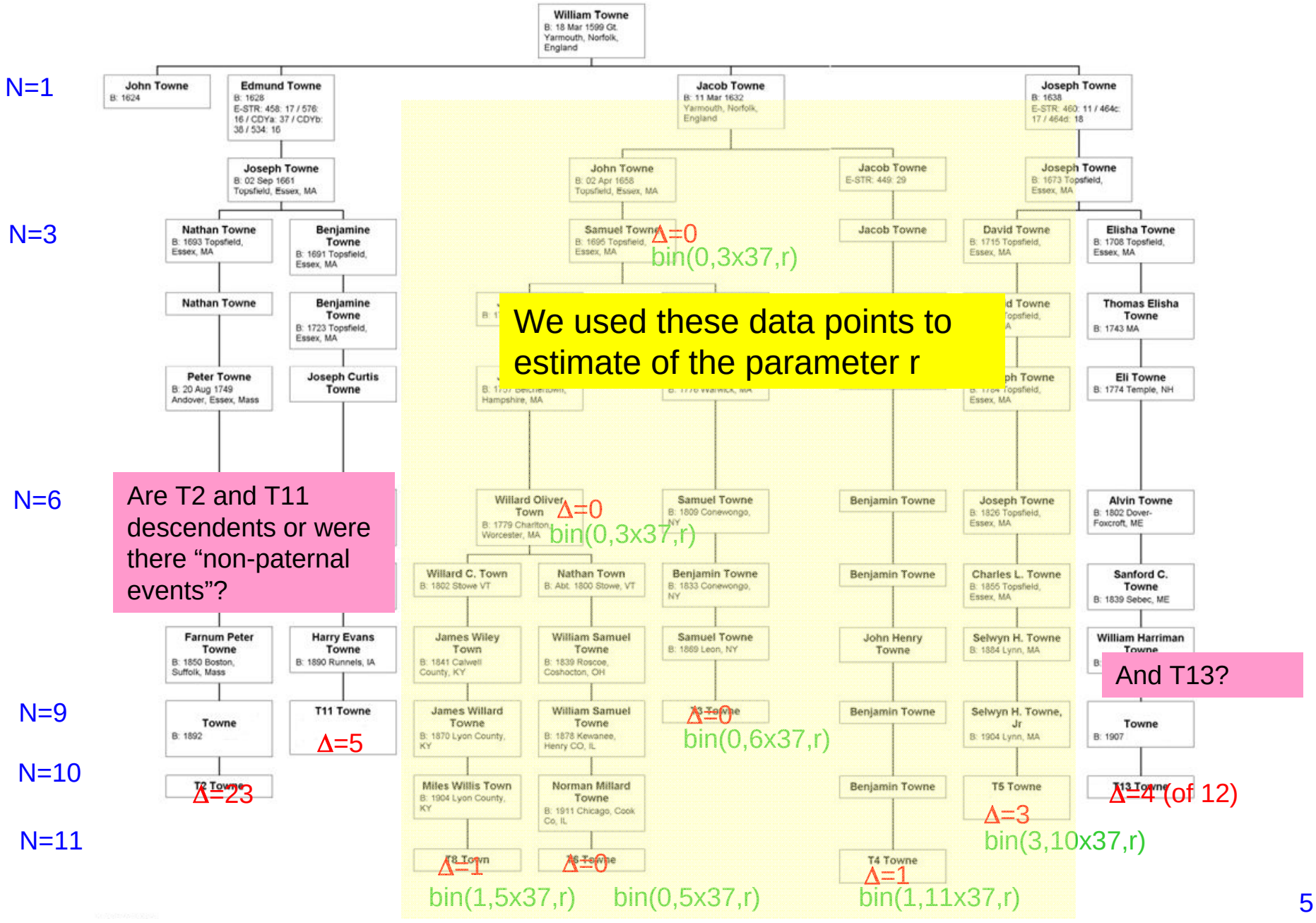
Journal of Economic Perspectives—Volume 22, Number 4—Fall 2008—Pages 199–216

Retrospectives

Guinnessometrics: The Economic Foundation of "Student's" *t*

Stephen T. Ziliak

Now that we're so adept at p-value stuff, let's go back to the Towne family.



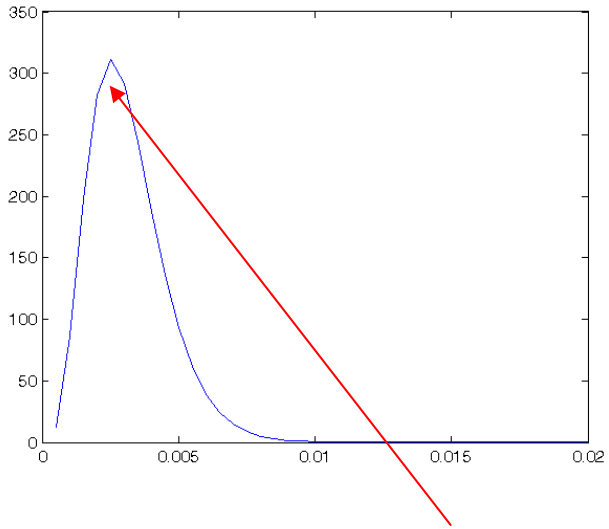
If we really knew *r*, then a p-value (tail) test on T2, T11, and T13 would be straightforward,

$$p_{\mathrm{tail},11} = \sum_{k=5}^{37} \mathrm{bin}(k, 9 \times 37, r)$$

notice how the "neglect backmutation" assumption makes this slightly dodgy

The problem is we have only Bayesian (uncertain) knowledge about *r*

$$P(r|\mathrm{data}) = \mathrm{bin}(0, 3 \times 37, r)\,\mathrm{bin}(0, 3 \times 37, r)\,\mathrm{bin}(1, 5 \times 37, r)\,\mathrm{bin}(0, 5 \times 37, r)$$
$$\times\, \mathrm{bin}(0, 6 \times 37, r)\,\mathrm{bin}(1, 11 \times 37, r)\,\mathrm{bin}(3, 10 \times 37, r)/r$$

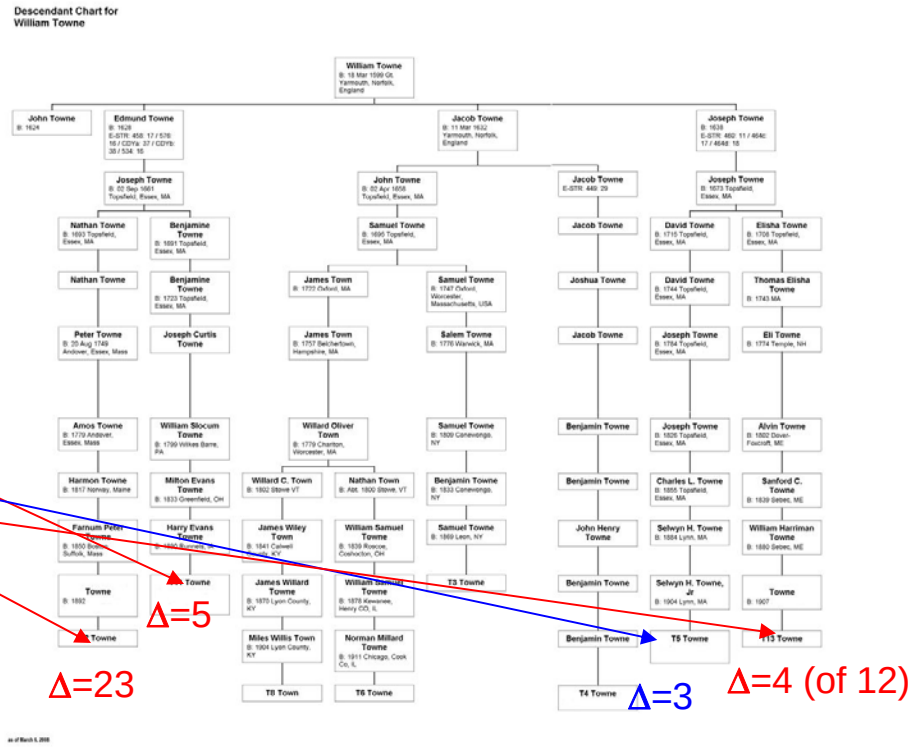


A common frequentist practice is to use the maximum likelihood estimate of r. **This is just wrong** (except asymptotically if the distribution of r were very narrow) because T11's extreme tail probabilities will be dominated by the extreme (but possible) values of r.

One “modern” way to proceed is to integrate the p-value over the posterior probability of all estimated quantities. This is called the “**posterior predictive p-value**” and is an example of a set of methods loosely called “**empirical Bayes**”.

$$p_{\mathrm{tail},11} = \int_0^\infty \sum_{k=5}^{37} \mathrm{bin}(k, 9\times 37, r)\, P(r|\mathrm{data})dr \Big/ \int_0^\infty P(r|\mathrm{data})dr$$

```
t11tail =
    0.0104
t2tail =
  1.0036e-013
t5tail =
    0.1288
t13tail =
    0.0013
```



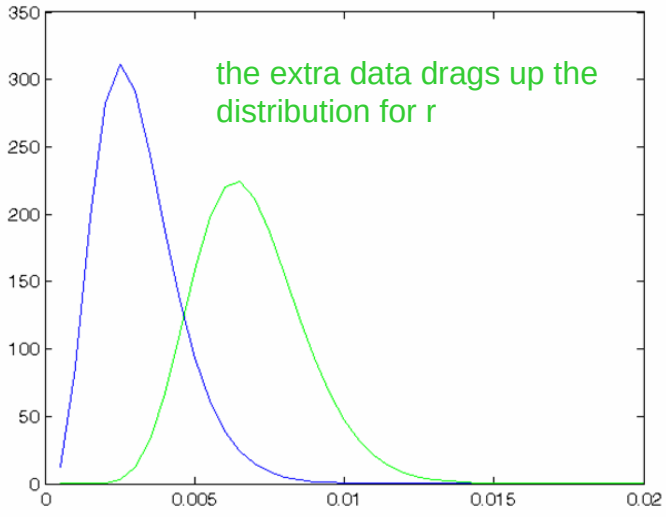
So the three questionables are all unlikely to be descendents.

This would be a satisfactory end to the Towne story, except that we tainted the data by tail trimming. While T2 is hopeless, what if we had included T11 and T13?

$$P(r|\text{data}) \propto P_{\text{previous}}(r|\text{data}) \times \text{bin}(5, 9 \times 37, r)\, \text{bin}(4, 10 \times 12, r)$$

```
t11tail =
    0.0953
t2tail =
  3.2348e-011
t13tail =
    0.0122
```

So suddenly there is hope for T11.
T2 and T13 still strongly ruled out.



This is an actual methodological problem with "posterior predictive p-value". Data is being used twice: once to get the posterior, then again to test itself. Often you can get away with this (e.g., try posterior both with and without questionable data). But in this example T11 is left ambiguous.

This is when we need real 

(We'll return to the Towne family one more time, later.)

Let's talk about multiple hypothesis testing.

The "Bonferroni correction" is widely used.

Carlo Emilio Bonferroni
(1892 – 1960)

It is very conservative, hence usually not the most powerful test.

α = prob. that <u>one or more</u> 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 rarely true.

The opposite limit would be to repeat the <u>same test</u> 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.

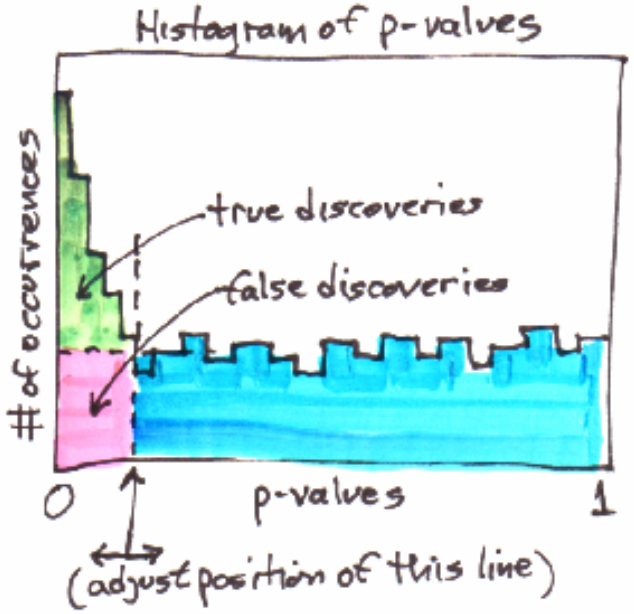
Slavish adherence to Bonferroni is a curse on biomedical research, but it is better than the alternative of having a literature full of wrong results!

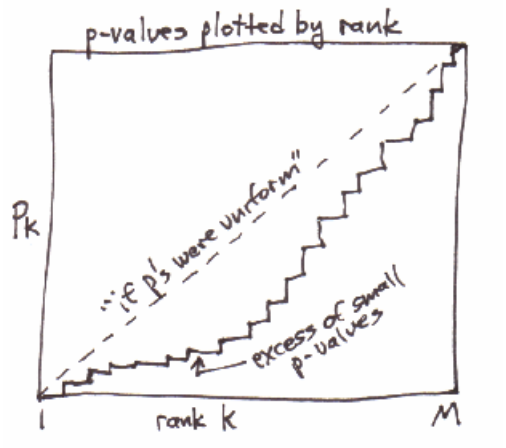
For large-scale screens can use False Discovery Rate (FDR) instead.

False Discovery Rate (Benjamini & Hochberg)

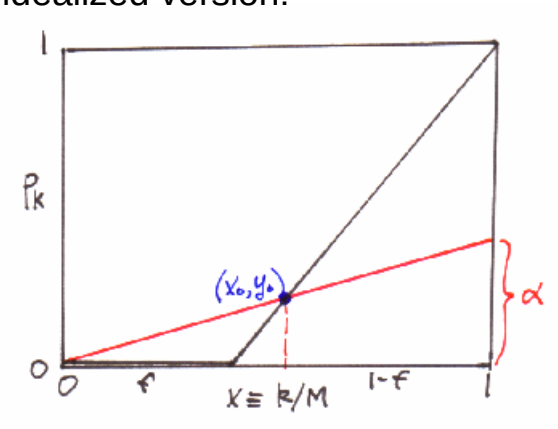
This is often a good alternative to Bonferroni, when the latter is too conservative.

- 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, α = 5%





idealized version:



Prescription: call as discoveries all $p_k < \frac{k}{M}\alpha$

Proof:

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

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

(There are fancier proofs for the nonidealized version.)

OK, enough p-values for now.

Keep in mind that we are still "closet Bayesians", however…

Bayesians have much less difficulty with multiple hypotheses in the happy case that they are EME.

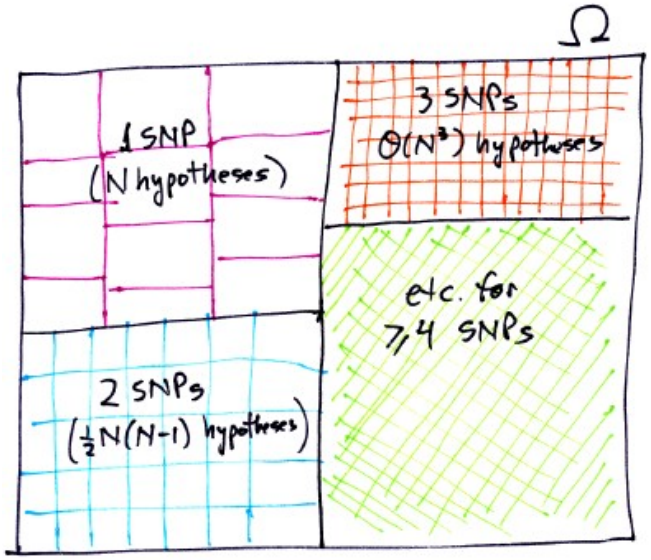
Example: We have a model where one, or a combination of, single nucleotide polymorphisms (SNPs) causes a particular kind of cancer.

We genotype patients and controls for N SNPs, each with 2 alleles.

So there are 2^N measurable hypotheses, and under each we can compute $P(\text{data}|H_i)$

p-value with Bonferroni would make a statistically significant finding impossible!

We are saved by the prior $P(H_i)$ which must have $\sum P(H_i) = 1$



Quite typically, our prior for models with "one main factor" (here, one SNP) will be larger than with "two main factors" (2 SNPs) and so on.

Now, do the "Bayes thing" and see if the evidence factor increases any individual model to high posterior probability.

Look, Ma, no multiple hypothesis correction!

(Let me explain where we're going here…)

- Building up prerequisites to do a fairly sophisticated treatment of model fitting
  - Bayes parameter estimation ✓
  - p-value tail tests ✓
  - really understand multivariate normal and covariance
  - really understand chi-square
- Then, we get to appreciate the actual model fitting stuff
  - fitted parameters
  - their uncertainty expressed in several different ways
  - goodness-of-fit
- And it will in turn be a nice "platform" for learning some other things
  - bootstrap resampling