Calibrating the evidence in experiments with applications to meta-analysis

presented by Robert G. Staudte

Applied and Medical Statistics Session of the 2006 useR! Conference held in Vienna, Austria
15 June, 2006

Joint with Stephan Morgenthaler, EPFL
and Elena Kulinskaya, Imperial College

Calibrating the p-value

How much evidence is there in a p-value of 0.01, say, relative to 0.05?

How small must a p-value be to represent twice as much evidence against the null hypothesis as 0.05?

Calibration of the p-value

Given $X = \mu + Z$ we want to test

$\mu = 0$ against $\mu > 0$.

Observe $X = x$; then $PV(x) = \Phi(-x)$.

Under alternatives.

$PV(X) = 1 - \Phi(X)$,

where $X \sim N(\mu, 1)$.

Remarks:

- There are two p-values.
- 'Evidence' for the alternative $\mu > 0$, however it is defined, should grow at rate $\sqrt{n}$.

p	0.0005	0.001	0.01	0.05	0.1	0.2
$T(p)$	3.291	3.090	2.326	1.645	1.276	0.8416
$\frac{T(p)}{T(0.05)}$	2.000	1.879	1.414	1.000	0.779	0.511

Now suppose the experimentor makes n measurements $x_1, \ldots, x_n$ and judges the null hypothesis using the average $\bar{x}_n = (x_1 + \cdots + x_n)/n$.

The random p-value based on these n observations can be written $PV_n = 1 - \Phi(\sqrt{n}\bar{X}_n)$.

It follows that the transformed p-value $T(PV_n) = \sqrt{n}\bar{X}_n$ has an expected value $\sqrt{n}\mu$ which is proportional to the square root of the sample size.

A p-value of 0.05 should be reported as evidence 1.645 ± 1.

To test $\theta = 0$ versus $\theta > 0$, let S be a test statistic which rejects H_0 for large values of S. A measure of evidence T should satisfy:

E_1. T is monotone increasing in S;

E_2. the distribution of T is normally distributed for all values of the parameters;

E_3. the variance $\mathrm{Var}[T] = 1$ for all values of the parameters; and

E_4. the expected evidence

$$\tau = \tau(\theta) = \mathrm{E}_\theta[T]$$

is increasing in θ from $\tau(0) = 0$.

How generally applicable is the calibration scale?

For one-sample t-tests, use

$\sqrt{2\nu}\ \sinh^{-1}(\frac{t_\nu}{\sqrt{2\nu}})$

For one-sample Binomial tests, use

$2\sqrt{n}\,\{\arcsin(\sqrt{\tilde{p}}) - \arcsin(\sqrt{p_0})\}$

For Chi-squared tests with $X \sim \chi^2_\nu(\lambda)$, use $\{X - \nu/2\}^{1/2} - \nu/2^{1/2}$

Combining evidence:

Given K studies measuring possibly different effects θ_k with evidence for $\theta_k > 0$ given by

$T_k \sim N(\tau_k, 1)$,

and $\tau_k = \sqrt{n_k}\, m(\theta_k)$.

How one combines evidence in $(T_1, \ldots, T_K)$ depends on:

1. how much evidence T_Q one finds for heterogeneity of the θ_k's and

2. on the specific alternative to the joint null $\theta_1 = \ldots = \theta_K$ one wants evidence for.

The main advantage is that it is like doing meta-analysis with *known* weights.

Summary

- The evidence in the p-value is on the probit scale
- VST's will put many problems on the probit scale
- Interpreting evidence on the probit scale is simple
- Combining evidence on the probit scale is simple