Box (1980) distinguishes

- **Estimation** based on posterior distribution \( f(\theta | \text{data, assumptions}) \)
Estimation vs. Criticism

Box (1980) distinguishes

– Estimation based on posterior distribution \( f(\theta \mid \text{data, assumptions}) \)
– Criticism based on prior-predictive distribution \( f(\text{data} \mid \text{assumptions}) \)

**Sampling and Bayes’ Inference in Scientific Modelling and Robustness**

*By George E. P. Box*

*University of Wisconsin–Madison*

[Read before the Royal Statistical Society at a meeting organized by the South Wales Local Group on Thursday, May 15th, 1980, the President Sir Claus Moser in the Chair]*
Estimation vs. Criticism

Box (1980) distinguishes

– Estimation based on posterior distribution $f(\theta | \text{data, assumptions})$
– Criticism based on prior-predictive distribution $f(\text{data} | \text{assumptions})$
– The assumptions include both model and prior assumptions.
Box’s Tail Probability

Box (1980) suggests to compute

\[ p_{\text{Box}} = \Pr \{ f(\text{data} \mid \text{assumptions}) < f(\text{observed data} \mid \text{assumptions}) \} \]

- Quantifies compatibility of model/prior with observed data.
- Small values of \( p_{\text{Box}} \) indicate that model or prior are discredited by the observed data.
Box’s Tail Probability

Box (1980) suggests to compute

$$p_{\text{Box}} = \Pr \{ f(\text{data} \mid \text{assumptions}) < f(\text{observed data} \mid \text{assumptions}) \}$$

- Quantifies compatibility of model/prior with observed data.
Box’s Tail Probability

Box (1980) suggests to compute

\[ p_{\text{Box}} = \Pr \{ f(\text{data} \mid \text{assumptions}) < f(\text{observed data} \mid \text{assumptions}) \} \]

- Quantifies compatibility of model/prior with observed data.
- Small values of \( p_{\text{Box}} \) indicate that model or prior are discredited by the observed data.
The $Q$-Test

- Model: Normal-normal model
- Prior: $\tau^2 = 0$
- Data: Differences between effect estimates $\hat{\theta}_i$ with standard errors $\sigma_i$

$\to f(\text{observed data} \mid \text{assumptions})$ is the $Q$-statistic

$$Q = \frac{\sum_{i<j} w_i w_j (\hat{\theta}_i - \hat{\theta}_j)^2}{\sum_{i=1}^{k} w_i}$$

where $w_i = 1/\sigma_i^2$ are “fixed-effect” weights.

$\to p_{\text{Box}}$ simplifies to the $p$-value obtained from the null distribution $Q \sim \chi^2_{k-1}$
The Generalized $Q$-Test

- Model: Normal-normal model
- Null hypothesis: $\tau^2 = \tau_0^2$
- The generalized $Q$-statistic $Q(\tau_0^2)$ now uses “random-effects” weights $w_i = 1/(\sigma_i^2 + \tau_0^2)$.
- We still have $Q(\tau_0^2) \sim \chi_{k-1}^2$ if $\tau^2 = \tau_0^2$
- Solving $Q(\tau^2) = k - 1$ for $\tau^2$ gives the Paule-Mandel estimate.
### Example: Donepezil vs. Placebo

**Donepezil vs. Placebo**  
DAD, CMCS, PDS

<table>
<thead>
<tr>
<th>Studie</th>
<th>logarithmierter RR</th>
<th>SE</th>
<th>RR (95%-KI)</th>
<th>Gewichtung</th>
<th>RR</th>
<th>95%-KI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gauthier 2002</td>
<td>-0.66</td>
<td>0.15</td>
<td></td>
<td>32.9</td>
<td>-0.66</td>
<td>[-0.95, -0.37]</td>
</tr>
<tr>
<td>Homma 2000</td>
<td>-0.35</td>
<td>0.14</td>
<td></td>
<td>34.4</td>
<td>-0.35</td>
<td>[-0.63, -0.07]</td>
</tr>
<tr>
<td>Winblad 2001</td>
<td>-0.29</td>
<td>0.15</td>
<td></td>
<td>32.7</td>
<td>-0.29</td>
<td>[-0.58, 0.00]</td>
</tr>
<tr>
<td>REM - DerSimonian-Laird</td>
<td></td>
<td></td>
<td></td>
<td>100.0</td>
<td>-0.43</td>
<td>[-0.66, -0.21]</td>
</tr>
<tr>
<td>FEM - Inverse Varianz</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.43</td>
<td>[-0.60, -0.27]</td>
</tr>
<tr>
<td>REM - Knapp-Hartung</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.43</td>
<td>[-0.93, 0.06]</td>
</tr>
<tr>
<td>Bayes - HN(0.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.43</td>
<td>[-0.64, -0.23]</td>
</tr>
</tbody>
</table>

Heterogenität: $Q=3.70$, $df=2$, $p=0.157$, $I^2=45.9\%$

Gesamteffekt (REM - DerSimonian-Laird): Z-Score=-3.78, $p<0.001$, Tau=0.135
Example: Donepezil vs. Placebo
Checking the Heterogeneity Prior

- Suppose we now have a half-normal prior $f(\tau)$ with $E(\tau) = \tau_0$

![Graph of HN(0.25) prior]

$\rightarrow p_{Box}$ is now based on

$$\tilde{Q} = \int Q(\tau^2)f(\tau^2)d\tau^2$$

- The $\chi^2$-distribution still holds because $\tau^2$ is a pivot for $Q(\tau^2)$

$\rightarrow p_{Box}$ can be easily calculated through Monte Carlo simulation.
Example: Donepezil vs. Placebo

![Graph showing the comparison between Donepezil and Placebo with Q statistic and p-value.]

- Fixed
- Half-normal prior
Type-I Error Assessment

\[ k = 3, \tau_0 \sim f(\tau) \]

Normal distribution
t(4) distribution

\[ \text{Density} \]

\[ \begin{array}{ccc}
0.0 & 0.4 & 0.8 \\
10\% & & \\
\end{array} \]

\[ \begin{array}{ccc}
0.0 & 0.4 & 0.8 \\
12\% & & \\
\end{array} \]
Power Assessment

\[ k = 3, \text{ fixed } \tau_0 = 0.2 = E(\tau) \]

Normal distribution

\[ \text{Density} \]

\[ 0.0 \quad 0.4 \quad 0.8 \]

\[ \text{p-value} \]

\[ 0.0 \quad 0.5 \quad 1.0 \quad 1.5 \quad 2.0 \quad 2.5 \quad 3.0 \quad 3.5 \]

\[ 10\% \]

\[ \text{t(4) distribution} \]

\[ \text{Density} \]

\[ 0.0 \quad 0.4 \quad 0.8 \]

\[ \text{p-value} \]

\[ 0.0 \quad 0.5 \quad 1.0 \quad 1.5 \quad 2.0 \quad 2.5 \quad 3.0 \quad 3.5 \]

\[ 12\% \]
Power Assessment

\[ k = 3, \text{ fixed } \tau_0 = 0.4 = 2 \text{E}(\tau) \]

**Normal distribution**

**t(4) distribution**
Power Assessment

\( k = 3, \text{ fixed } \tau_0 = 0.8 = 4E(\tau) \)

**Normal distribution**

- **Density**
  - 3.5
  - 3.0
  - 2.5
  - 2.0
  - 1.5
  - 1.0
  - 0.5
  - 0.0
  
- **p-value**
  - 0.0
  - 0.4
  - 0.8

- **25% Normal distribution**

**t(4) distribution**

- **Density**
  - 3.5
  - 3.0
  - 2.5
  - 2.0
  - 1.5
  - 1.0
  - 0.5
  - 0.0

- **p-value**
  - 0.0
  - 0.4
  - 0.8

- **20% t(4) distribution**
Aggregated Power Assessment for 10 Studies

$k = 3$, fixed $\tau_0 = 0.8 = 4E(\tau)$

**Normal distribution**

**t(4) distribution**

<table>
<thead>
<tr>
<th>p-value</th>
<th>Normal distribution</th>
<th>t(4) distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>67%</td>
<td>51%</td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

- \( Q \)-Test can be generalized to check heterogeneity prior and other model assumptions.
- Has low power for meta-analyses with very small \( k \).
- Power can be increased by summation of \( Q \)-statistic across meta-analyses.
Summary and Discussion

- Q-Test can be generalized to check heterogeneity prior and other model assumptions.
- Has low power for meta-analyses with very small $k$.
- Power can be increased by summation of $Q$-statistic across meta-analyses.

- Heterogeneity prior is not the only critical assumption
- Normality assumption may also be wrong due to publication bias etc.
References