1 Vote Counting Methods

Suppose there are $M$ identical studies. In each study the statistic $S$ is calculated (e.g.) ‘$t$’ or ‘$r$’.

Assume a common non-centrality parameter (or population effect) $\sigma$ in all samples and use one-sided tests. The probability of a significant result is

$$P = P(\text{significant result} \mid \sigma, n)$$

$$= \int_{C_s}^{\infty} f(s; \sigma, n) \, ds,$$

where $C_s$ is a critical value of the distribution.
Vote counting methods yield an effect if the proportion of studies with “significant” coefficients exceeds $1/3$, say. We can pick any other number (e.g. more than half).

($1/3$ arises from outcomes such as “+” and significant, “−” and significant, and “nonsignificant”) 

Let $X$ be the number of successes by this criterion.

$$X = \sum X_i$$

$$X_i \sim \begin{cases} 1 \text{ with } P \\ 0 \text{ with } 1 - P \end{cases}$$

The $X_i$ are independent. $M$ is the number of studies.
Pr(proportion of successes > \(v\)) = P\left(\frac{X}{M} > v\right)

= 1 - \sum_{i=0}^{[vM]} \binom{M}{i} P^i (1 - P)^{M-i},

where \([a]\) is the greatest integer \(\leq a\).
Assume a $t$ distribution with noncentrality parameter

$$
\delta = \frac{\mu^E - \mu^C}{\sigma},
$$

where $E$ and $C$ denote ‘experimental’ and ‘control’ respectively.

This noncentrality parameter is also called “an effect”. This produces a “$P$”. 
2 Vote Counting Methods based on L. Hedges and I. Olkin

Assume equal competence of all investigators:

1. $M$ studies, same protocol for each study

2. Independent studies (samples different and independent of each other)

3. Summarizing the same types of data and studies
We seek to test $H_0 : \Delta = 0$ where $\Delta = \mu^E - \mu^C$.

Let $C_R$ be the critical region; $S$ is a common statistic; $n$ is the sample size in each study

$$P = \Pr\{\text{rejecting the null } | \Delta, n\} = \Pr(S \in C_R | \Delta, n)$$

- Size when $\Delta = 0$ (significance)
- Power when $\Delta \neq 0$

Unbiased test: size $< \text{power for all } \Delta \neq 0$. 
2.1 Vote Counting Methods

Vote counting rule for $M$ studies. If proportion of studies significant $> \nu$, we pronounce “significant” and then proceed, e.g. $\nu = 0.6$

$$
\Pr \left( \text{Proportion of studies significant} > \nu \right) \\
= \Pr \left( \frac{\# \text{ studies sig.}}{M} > \nu \right) \\
= 1 - \sum_{i=0}^{[\nu M]} \binom{M}{i} P^i (1 - P)^{M-i},
$$

where $[\nu M] = \text{greatest integer} \leq \nu M$. 

2.2 Large Sample Analysis

Use binomial CLT: $M$ studies $Q$ yield “significant” results

\[ \frac{Q}{M} \sim \mathcal{N} \left( P, \frac{P(1-P)}{M} \right) \quad \text{as } M \to \infty. \]

$P$ is fixed.

Remember that $P$ is the power under the alternative.
Vote counting probability of an effect in studies is

\[
\Pr \left( \frac{Q}{M} > \nu \right) = \Pr \left( \frac{\frac{Q}{M} - P}{\left[ \frac{P(1 - P)}{M} \right]^{1/2}} > \frac{\nu - P}{\left[ \frac{P(1 - P)}{M} \right]^{1/2}} \right)
\]

\[\sim \mathcal{N}(0,1)\]
Two cases:

1. Suppose that \( v > P > 0 \) (cutoff is greater than \( P \)).

   \[
   \text{As } M \to \infty, \quad \frac{v - P}{P(1-P)} \to \infty, \quad \Pr \left( \frac{Q}{M} > v \right) \to 0.
   \]

   If the power of the test for each study is low, we encounter serious problems. Power of the vote counting procedure \( \to 0 \).
2. If $P > \nu$, $\Pr \left( \frac{Q}{M} > \nu \right) \to 1$.

Consider an example where $\nu = \frac{1}{3}$. Table 1 presents $1 - \Pr \left( \frac{Q}{M} > \nu \right)$ (the probability of not detecting an effect). In Figure 1, $C_0$ is our $\nu$. 
Table 1
Probability That a Standard Vote Count Fails to Detect an Effect for Various Sample, Effect, and Cluster Sizes

<table>
<thead>
<tr>
<th>m</th>
<th>n</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>1.00</td>
<td>.999</td>
<td>.998</td>
<td>.994</td>
<td>.985</td>
<td>.968</td>
<td>.935</td>
<td>.880</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>1.00</td>
<td>.998</td>
<td>.990</td>
<td>.966</td>
<td>.906</td>
<td>.987</td>
<td>.606</td>
<td>.395</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>.999</td>
<td>.995</td>
<td>.975</td>
<td>.906</td>
<td>.947</td>
<td>.502</td>
<td>.252</td>
<td>.089</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>.999</td>
<td>.991</td>
<td>.950</td>
<td>.813</td>
<td>.547</td>
<td>.254</td>
<td>.073</td>
<td>.012</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>.999</td>
<td>.986</td>
<td>.914</td>
<td>.694</td>
<td>.358</td>
<td>.105</td>
<td>.016</td>
<td>.001</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.999</td>
<td>.997</td>
<td>.989</td>
<td>.966</td>
<td>.914</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>1.00</td>
<td>1.00</td>
<td>.998</td>
<td>.988</td>
<td>.941</td>
<td>.800</td>
<td>.545</td>
<td>.265</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>1.00</td>
<td>1.00</td>
<td>.993</td>
<td>.941</td>
<td>.747</td>
<td>.400</td>
<td>.118</td>
<td>.016</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>1.00</td>
<td>.999</td>
<td>.978</td>
<td>.834</td>
<td>.463</td>
<td>.119</td>
<td>.011</td>
<td>.000</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>1.00</td>
<td>.997</td>
<td>.948</td>
<td>.672</td>
<td>.222</td>
<td>.023</td>
<td>.001</td>
<td>.000</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.998</td>
<td>.986</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.994</td>
<td>.915</td>
<td>.589</td>
<td>.174</td>
</tr>
<tr>
<td>50</td>
<td>30</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.994</td>
<td>.862</td>
<td>.363</td>
<td>.035</td>
<td>.000</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>1.00</td>
<td>1.00</td>
<td>.999</td>
<td>.942</td>
<td>.461</td>
<td>.036</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>1.00</td>
<td>1.00</td>
<td>.995</td>
<td>.773</td>
<td>.124</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

Note. Each of the m replicated studies has a common sample size n. A two-tailed t test is used to test mean differences at the .05 level of significance. An effect is detected if the proportion of positive significant results exceeds one third.
Figure 1. Effect size $\delta = (\mu^E - \mu^C)/\sigma$ below which the power of a vote count tends to zero for a collection of studies with common sample size $N$. (For each criterion $c_0$, $H_0: \delta = 0$ is rejected if the proportion of studies obtaining positive significant results exceeds $c_0$.)
3 Meta-Analysis

Theorem 1 (Pooling (Fisher)) Johnson and Kotz: Given $K$ independent studies, $U$ uniform on $(0, 1)$ implies

$$-2 \ln U \sim \chi^2(2)$$

$$\therefore -2 \ln \left( \prod_{i=1}^{K} P_i \right) \sim \chi^2(2K).$$

Reject overall if we have

$$P = -2 \sum_{i=1}^{K} \ln P_i \geq C,$$

a critical value of the $\chi^2$ distribution.
This is a strange procedure. It assigns the same weight to all studies considered.

Under Good’s reformulation, we assign weights to the $P_i$ values,

$$P_N = \prod_{i=1}^{K} P_i^{\nu_i},$$

where $-2 \ln P_N$ is a weighted average of $\chi^2$ random variables.
The question on the table is *How can we pool evidence? How can we combine tests?*

The pitfall is an assumption of homogeneity.

We also have (a) conditioning problems, and (b) problems with nuisance parameters (use invariants, etc.).
4 Scholtz’s Method for Combining $P$ Values

Most $P$ value pooling schemes for a statistic

$$T(h) = \sum_{i=1}^{k} h(P_i).$$

Scholtz’s non-decreasing function defined on $(0, 1)$. Fisher’s method:

$$h(x) = -2 \ln x$$

$$h(x) = \Phi^{-1}(x) \quad \text{(inverse normal)}$$

$$h(x) = -\ln \left( \frac{x}{1-x} \right) \quad \text{(inverse logit)}$$
The idea is to find the supremum over a properly defined set,

\[ \sup (T(h) : h), \]

a suitably restricted set with some mean restraint. \( h \) is non-increasing.

Scholtz provides an improvement on Fisher, the power of Fisher’s test is improved.
5 Bias in Picking the Most Favorable Test Statistic

(Common practice in calibration.)
Assume $K$ studies (e.g. effects of unions on relative wages). Suppose we select among the $K$ studies; take

$$Q = \min(P_i)_{i=1}^K = q_{obs}. $$

This is the lowest-rank order statistic among the $P$ values:

$$\Pr(Q \leq q_{obs} : H_0) = 1 - \Pr(P_j \geq q_{obs})^K$$

Thus, if the smallest test is 0.1, the actual size for $K = 10$ is

$$1 - (1 - 0.1)^K = 1 - (0.3487) = 65\%.$$
With 100 studies we obtain a 100% test. This is a problem with publication bias—what you see is selected to be the “most significant” outcome.
If test statistics are *not* independent, we can use the Bonferroni inequality to create

\[
\Pr \left( \bigcup_{i=1}^{I} A_i \right) \leq \sum_{i=1}^{I} \Pr(A_i).
\]

\[
\bigcap_{i=1}^{I} B_i \text{ is the complement of } \bigcup_{i=1}^{I} \bar{B}_i \in \Omega
\]
\[
\therefore \Pr \left( \bigcap_{i=1}^{I} B_i \right) = 1 - \Pr \left( \bigcup_{i=1}^{I} \bar{B}_i \right)
\]

\[
\therefore 1 - \Pr \left( \bigcap_{i=1}^{I} B_i \right) \leq \sum_{i=1}^{I} \Pr(\bar{B}_i)
\]

\[
\therefore \Pr(Q \leq q_{obs} : H_0) \leq \sum_{j} \Pr(P_j \leq q_{obs}) = K q_{obs},
\]

assuming marginals the same.