Meta-analysis

Meta-analysis is analyzing previous studies in a way that tries to combine their results in a statistical way, rather than informally assessing how much they agree.

Meta-analysis is popular in biostatistics, particularly in clinical trials and in genome-wide association studies.

Part of the idea is that individual studies might be underpowered to detect results, so if we combined data from multiple studies, we could in effect increase the sample size and gain better power for testing hypotheses and more precise estimates for confidence intervals.

Meta-analysis

In the ideal case, we'd have access to the original data, and could pool the data. Suppose you have two data sets,

$$x_1, \ldots, x_n, \qquad \text{and } y_1, \ldots, y_m$$

which both measure, say decrease in systolic blood pressure after taking some medication. If you want to test the null hypothesis that there is no change in blood pressure, you could test this using either data set to get

$$t_{\text{obs},x} = \frac{\overline{x}}{s_x/\sqrt{n}}, \qquad t_{\text{obs},y} = \frac{\overline{y}}{s_y/\sqrt{m}}$$

Both t-tests would give you a p-value and you could see whether the two data sets agree on whether the medication reduces systolic blood pressure.

Meta-analysis

However, if the medication does reduce systolic blood pressure, then you are likely to get a smaller p-value if you use

$$z_i = \begin{cases} x_i & i = 1, \ldots, n \\ y_{i-n} & i = n+1, \ldots, n+m \end{cases}$$

Then do a t test as

$$t_{\text{obs},z} = \frac{\overline{z}}{s_z/\sqrt{n+m}}$$

Meta-analysis

Sometimes you'd like to combine information from different samples where you don't have access to the original data. In that case, you might only have the t-statistics or only the p-values. In particular, suppose the data from the x observations gives you a p-value of 0.08 and the data from the y observations gives you a p-value of 0.10. In this case, neither test would be significant, yet both p-values are a bit on the low side.

If you had several studies and p-values were consistent low, even though not significant, you might suspect that there really is an effect, but the studies just use samples sizes that are too low to detect the effect. Do several studies with slight evidence add up to a signficant amount of evidence for an effect?

Meta-analysis

Is there a way of combining the p-values? If p-values are uniformly distributed, then it would be a bit unlikely to get two p-values in a row that are 0.10 or less. It would be a bit like flipping a coin three times and getting heads each time, then flipping three more times, and again getting heads each time.

Fisher had an idea about using the null distribution of the p-value being uniform. Fisher combines k independent p-values using the following test statistic:

$$-2\sum_{i=1}^{k}\log(p_i) \sim \chi^2_{2k}$$

That is -2 times sum of the log p-values is χ^2 with $2k$ degrees of freedom under the null hypothesis.

Meta-analysis

The idea behind the statistic is a relationship between the uniform and exponential distributions, and then the exponential and χ^2 distributions. If p_i has an exponential distribution, then the distribution of $-\log(p_i)$ is exponential with rate 1. You can verify this using the cdf method.

$$\begin{aligned} P(-\log p_i \leq x) &= P(\log p_i \geq -x) \\ &= P(p_i \geq e^{-x}) \\ &= 1 - P(p_i < e^{-x}) \end{aligned}$$

and this is the cdf of an exponential(1). If we look at $P(-2\log p_i \leq x)$, then we get an exponential with rate $1/2$, which is equivalent to a χ^2 distribution with 2 degrees of freedom. Adding up k such independent χ^2_2 random variables results in a χ^2_{2k} random variable.

Meta-analysis

If we apply this technique to the case of the p-values of 0.07 and 0.10, we get

$$-2\sum_{i=1}^{2} p_i = -2(\log(0.08) + \log(0.10)) = 9.656627$$

We compare this to a χ^2_4. In R we get

```
> 1-pchisq(9.656627,4)
[1] 0.04662652
```

So based on Fisher's method, the combined p-value shows stronger evidence against the null hypothesis than either p-value on its own.

Meta-analysis

Is it possible that the combined evidence is weaker than the evidence from either test? Yes, for larger p-values this is possible. For example,

```
> -2*(log(.5) + log(.5))
[1] 2.772589
> 1-pchisq(2.772589,4)
[1] 0.5965735
```

It's interesting to play around with different numbers, and not entirely intuitive what will happen.

```
> -2*(log(.1) + log(.1))
[1] 9.21034
> 1-pchisq(9.21034,4)
[1] 0.05605171
> -2*(log(.05) + log(.95))
[1] 6.094051
> 1-pchisq(6.094051,4)
[1] 0.1922337
```

Meta-analysis

Fisher's method doesn't give the same value as combining the original data. Here is an example

```
> x <- rnorm(20,.1)
> y <- rnorm(30,.1)
> t.test(x)$p.value
[1] 0.8789199
> t.test(y)$p.value
[1] 0.146231
> t.test(c(x,y))$p.value
[1] 0.2455739
> -2*(log(.8789) + log(.1462))
[1] 4.103728
> 1-pchisq(4.103728,4)
[1] 0.3921498
```

In this case, the combined data gave a p-value of 0.245 and the Fisher method gave 0.392.

Meta-analysis

The above example doesn't necessarily mean that the Fisher method always gives weaker p-values than combining the data. In particular, if the data have opposite effects, then combining them could cancel out in the combined data but not in the two-sided *p*-values that are combined.

```
> x <- rnorm(20,-1)
> y <- rnorm(30,1)
> t.test(x)$p.value
[1] 0.0001414914
> t.test(y)$p.value
[1] 1.126416e-06
> t.test(c(x,y))$p.value # p-value on combined data
[1] 0.1621011
> -2*(log(.00014149) + log(1.126416e-06))
[1] 45.1195
> 1-pchisq(45.1195,4)
[1] 3.754877e-09 # Fisher p-value
```

Meta-analysis

An R function for computing the combined p-value as a function of a vector of p-values is

```
> combinep <- function(x) {
+ stat <- -2*sum(log(x))
+ return(1-pchisq(stat,2*length(x)))
+ }
> combinep(c(.08,.1))
[1] 0.04662651
> combinep(c(.08,.1,.5))
[1] 0.08705891
```

The function works for any number of input p-values.

Meta-analysis

Although Fisher came up with a method for combining p-values, the term was apparetnly first coined by Gene Glass in 1974, who said that it is the "analysis of analyses. It connotes the rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature."

Previous examples of work that could be considered meta-analysis included Pearson in 1904 studying the effects of typhoid vaccination, and a book in 1940 called "Extrasensory Perception After Sixty years", which reviewed 145 reports of ESP published over several decades.

A typical feature of meta-analysis is reviewing published journal articles that have occurred over a number of years and trying to summarize and combine the evidence from multiple sources. Although the combined p-value approach above is sort of an ideal case, in practice meta-analysis is complicated by different study designs and different populations that are used for sampling.

Meta-analysis

Although combining p-values or effect sizes is one function of a meta-analysis, it also serves to summarize the relevant literature up to that point. A nice feature of meta-analysis, especially for statisticians, is that you can write a paper and get it published in an applied journal by reviewing other people's data and without having to generate any new data yourself.

There are a number of questions of interest in a typical meta-analysis:

1. Is there evidence of an effect when combining different studies?
2. Is the effect in the same direction in different studies? (i.e., is a new drug consistently beneficial, harmful, etc.)?
3. What is the estimated effect size when combining different studies?

Meta-analysis

Meta-analysis is often applied in both clinical trials and GWAs. Some common tendencies include:

1. early studies often show stronger effects than later studies
2. larger sample sizes and better designs often show smaller effects
3. there can be many underpowered studies

Meta-analysis

Often, the start of a meta-analysis involves a literature review for the drug or genes of interest and their effects. Often a number of papers might show up, some of which might be excluded from the meta-analysis. Some studies might be excluded due to

1. small sample sizes
2. poor design
3. incompatible design/ different definitions with majority of studies
4. different reference populations (e.g., one study has ages 10–20, another has ages 20–60, etc.)
5. marker must not be too rare (for genetic studies)
6. study must be in English

nature genetics

om/naturegenetics

Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database

Lars Bertram[1], Matthew B McQueen[2,3], Kristina Mullin[1], Deborah Blacker[2,4] & Rudolph E Tanzi[1]

Meta-analysis: Bertram et al.

"Search strategies. To identify applicable studies, we performed a broad-range search for all publications in PubMed (National Center for Biotechnology Information; NCBI) under the keywords Alzheimer* AND (genet* OR associat*) published up until December 31, 2004, yielding nearly 19,000 articles, the first dating back to the early 1950s. Abstracts and/or titles (and full-text versions, as needed) of these papers were screened for fulfillment of AlzGene inclusion criteria. This yielded 700 papers published before 2005 eligible for inclusion in this study. Beginning January 1, 2005, we searched PubMed daily using the My NCBI e-mail subscription service for the keyword Alzheimer*. These searches were accompanied by regular screenings of tables of contents of scientific specialty journals in genetics, neurology, neuroscience and psychiatry, as well as cross-checks of references listed in individual publications. As of August 15, 2006, this search strategy led to the identification of nearly 900 different genetic association papers on Alzheimer disease eligible for inclusion in AlzGene. For the analyses presented in this manuscript, only papers published until December 1, 2005 were included (referred to as the data freeze; n = 789 papers)."

Meta-analysis

Some complications that arise include that effect sizes might be measured differently. For example, if the outcome is binary (survive, don't survive), then the results from studies might be reported as an odds ratio or a relative risk. If it is a relative risk, then you have

$$RR = \frac{P(E|A)}{P(E|B)}$$

i.e., the probability of an event given treatment A versus the probability of the event given factor B. However, if it is an odds ratio, then you have

$$OR = \frac{P(E|A)/(1 - P(E|A))}{P(E|B)/(1 - P(E|B))}$$

Here the numerator is the odds of the event given factor A and the denominator is the odds of the event given factor B.

Meta-analysis

Odds are related to probabilities by

$$O = \frac{P}{1 - P}$$

so that given the probability, you can determine the odds, and given the odds, the probability is

$$P = \frac{O}{O + 1}$$

However, given an odds ratio, we can't determine the relative risk, and vice versa.

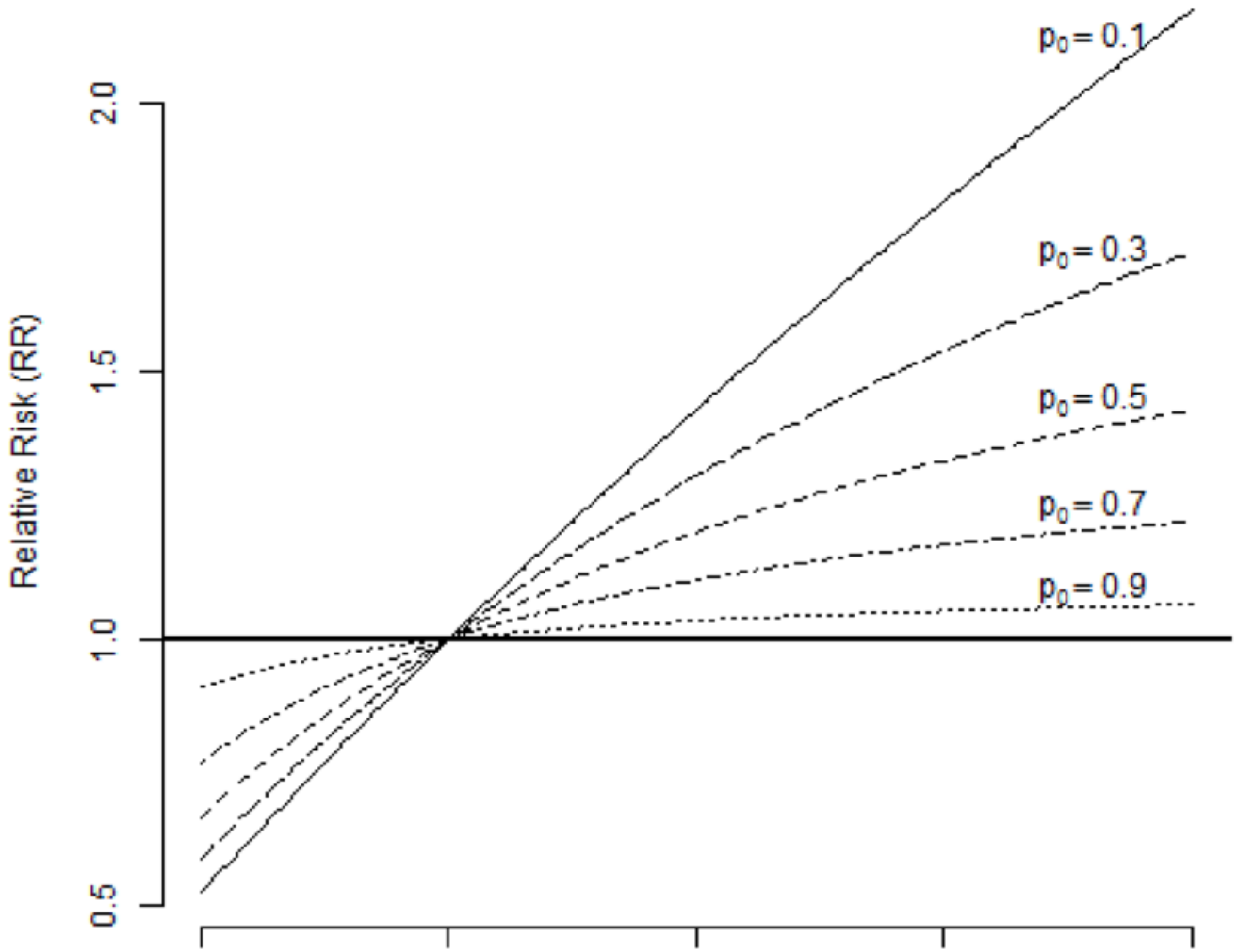
Meta-analysis

The odds ratio and relative risk are related, but conversion between one verses the other depends on the probability of the event E in the control group. If this probability is p, then the odds ratio and relative risk are related by

$$RR = OR/(1 - p + p \times OR)$$

So, if you know p for the control group, you can convert between the two. This might or might not make you exclude certain studies from a meta-analysis if you have trouble converting different effect sizes. One idea is that if you know a plausible range for p, then given an estimated OR, you can get a plausible range for RR

Meta-analysis



Meta-analysis

You can relate odds ratios and relative risks to contingency tables as follows

Treatment	Case	Control
1	a	b
2	c	d

Then we can estimate the relative risk and odds ratios as

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Meta-analysis

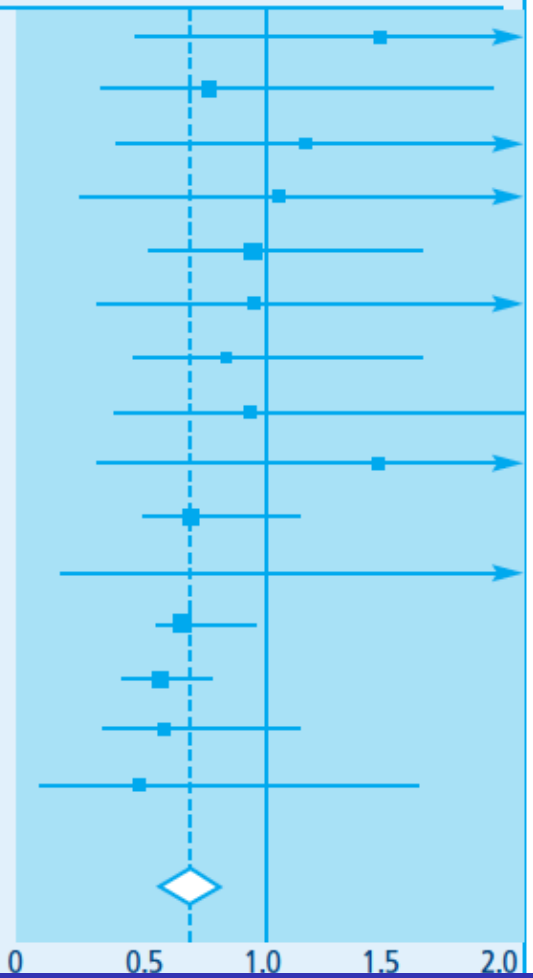
If a and b are small (number of cases is small, so the condition is rare), then $a + b \approx b$ and $c + d \approx d$. In this case $RR \approx OR$.

Meta-analysis

Often confidence intervals from different studies are plotted side-by-side, which is a useful way to summarize different studies. This is often called a *forest plot*.

Meta-analysis

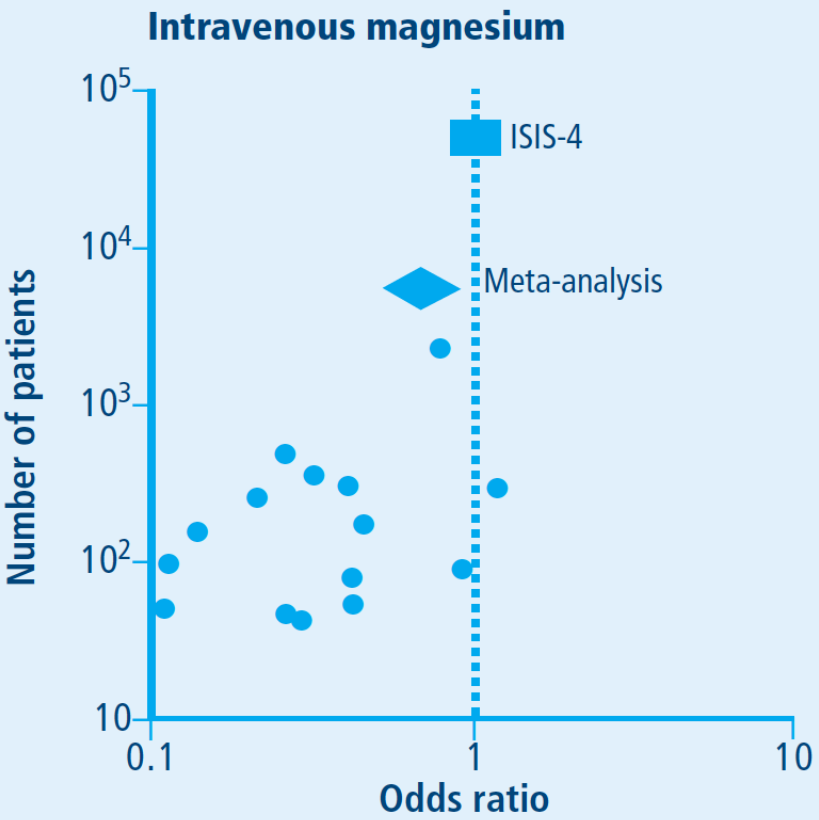
Study	No (%) of deaths/patients: Beta-blocker	Control	Beta-blocker deaths: Logrank observed – expected	Variance observed – expected
Wilcox (oxprenolol)	14/157 (8.9)	10/158 (8.9)	2.0	5.6
Norris (propanolol)	21/226 (9.3)	24/228 (9.3)	–1.4	10.2
Multicentre (propanolol)	15/100 (15.0)	12/95 (12.6)	1.2	5.8
Baber (propanolol)	28/355 (7.9)	27/365 (7.4)	0.9	12.7
Andersen (alprenolol)	61/238 (25.6)	64/242 (26.4)	–1.0	23.2
Balcon (propanolol)	14/56 (25.0)	15/58 (25.9)	–0.2	5.5
Barber (practolol)	47/221 (21.3)	53/228 (23.2)	–2.2	19.5
Wilcox (propanolol)	36/259 (13.9)	19/129 (14.7)	–0.7	10.5
CPRG (oxprenolol)	9/177 (5.1)	5/136 (3.6)	1.1	3.3
Multicentre (practolol)	102/1,533 (6.7)	127/1,520 (8.4)	–13.0	53.0
Barber (propanolol)	10/52 (19.2)	12/47 (25.5)	–1.6	4.3
BHAT (propanolol)	138/1,916 (7.2)	188/1,921 (9.8)	–24.8	74.6
Multicentre (timolol)	98/945 (10.40	152/939 (16.2)	–27.4	54.2
Hjalmarson (metoprolol)	40/698 (5.7)	62/697 (8.9)	–11.0	23.7
Wilhelmsson (alprenolol)	7/114 (6.1)	14/116 (12.1)	–3.4	4.8
Total*	640/7,047 (9.1)	784/6,879 (11.4)	–81.6	310.7

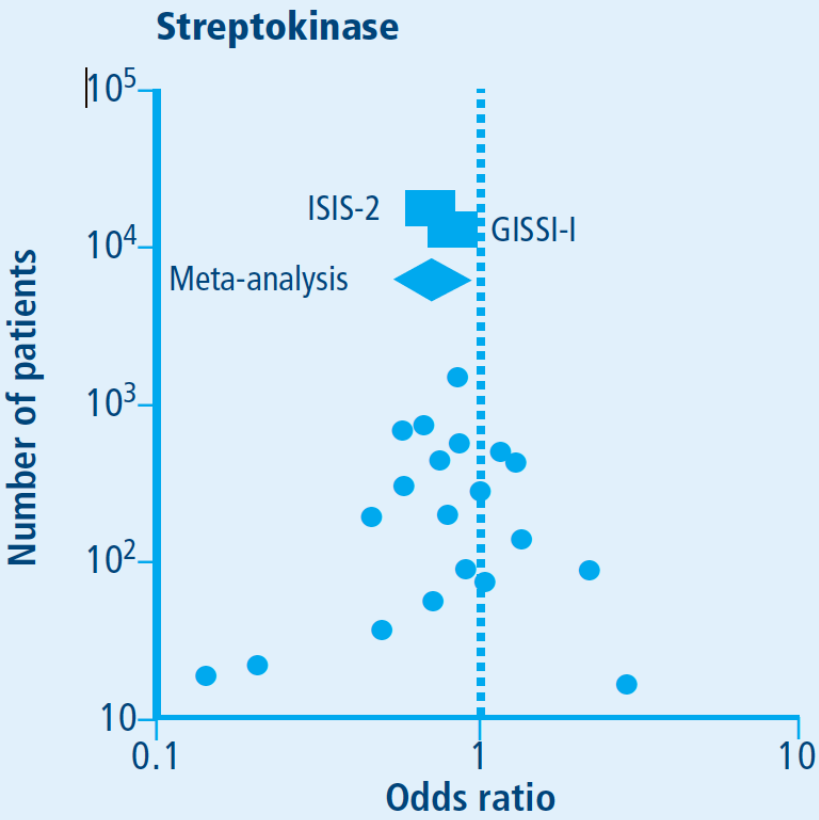


Meta-analysis

A *funnel plot* plots the effect size against the sample size. The idea is that smaller studies have more variability in effect size, so more extremes are observed. If the there is no publication bias, then both extremes should be observed, and you see a funnel-like shape. If negative findings tend not to get reported, the plot might be more asymmetrical.

Meta-analysis





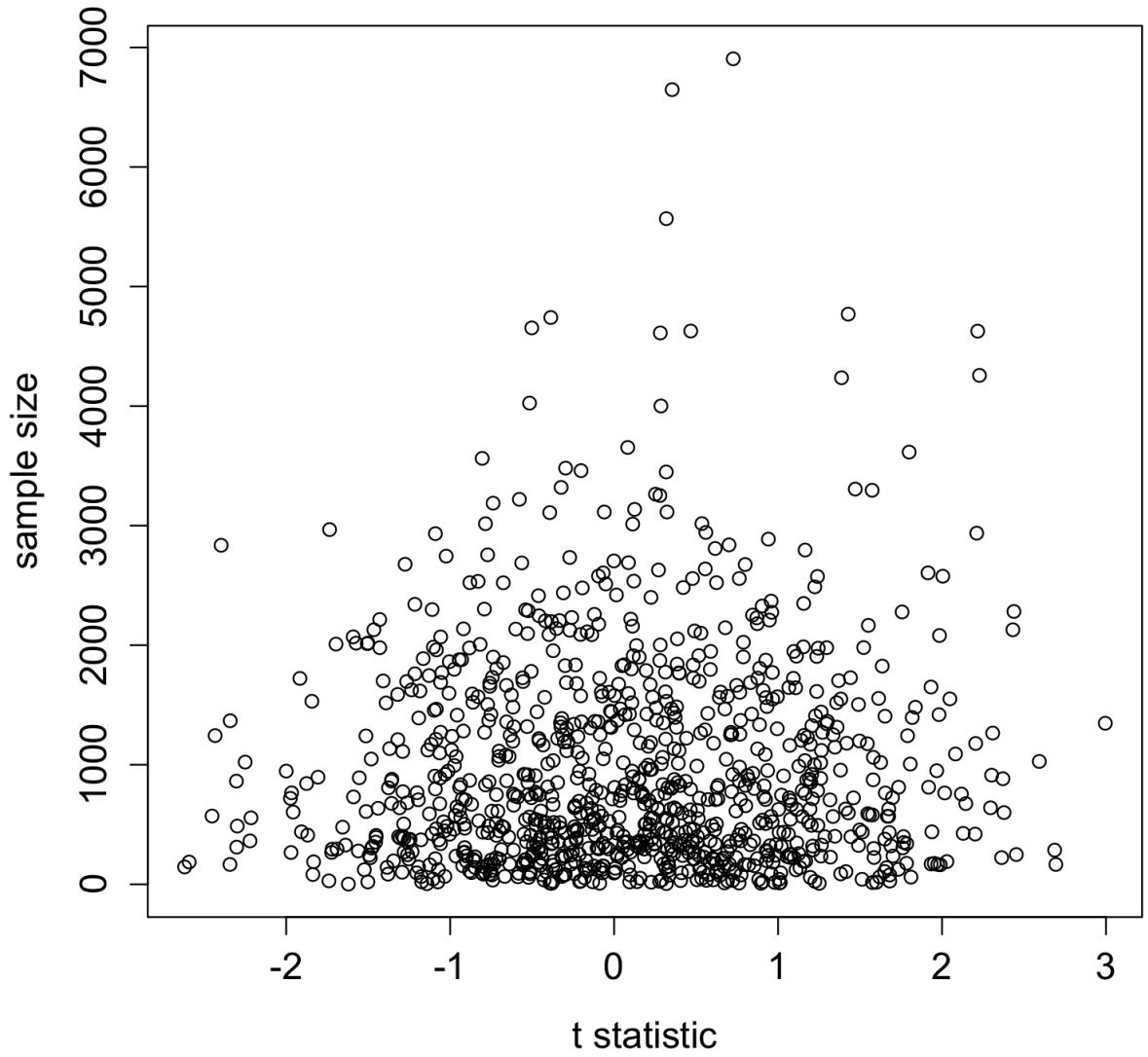
Points indicate odds ratios from small and medium sized trials, diamonds indicate combined odds ratios with 95% confidence

Meta-analysis

Here we'll do a simulation in R to verify that the funnel plot pattern is quite common.

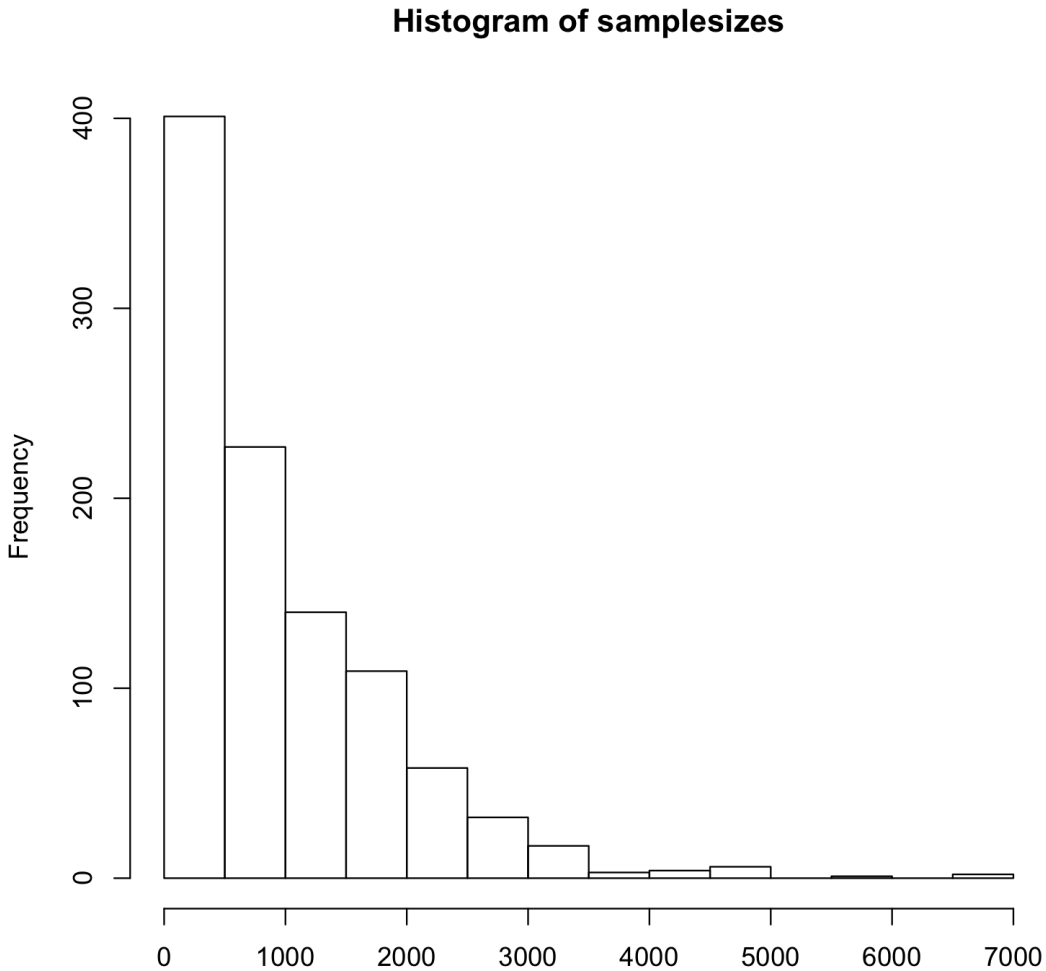
```
> samplesizes <- floor(rexp(1000,1)*1000)+2
> hist(samplesizes)
> mystat <- NULL
> for(i in 1:1000) {
+ xtemp <- rnorm(samplesizes[i])
+ mystat <- c(mystat,t.test(xtemp)$statistic)
+ }
> plot(mystat,samplesizes,xlab="t statistic",ylab="sample size
```

Meta-analysis

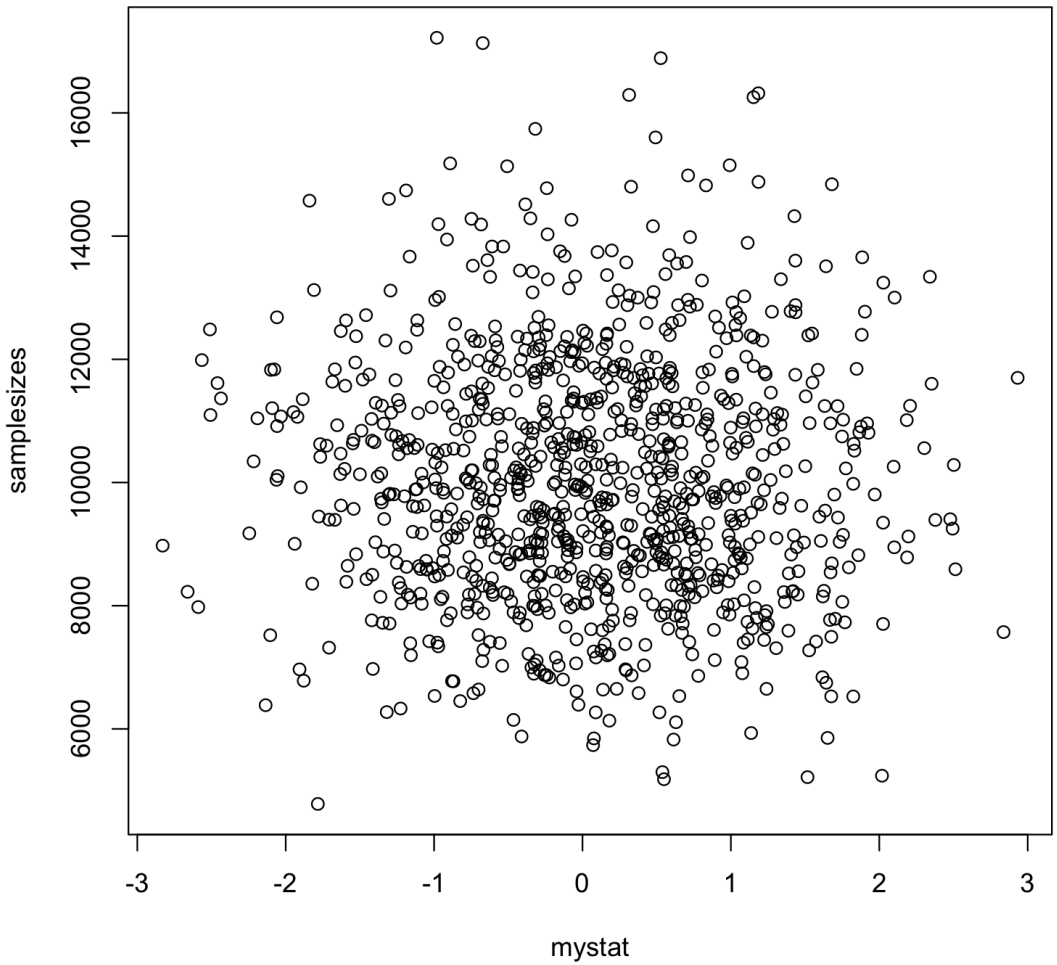


Meta-analysis

I generated skewed sample sizes for this example...



Based on the simulations I've tried, I don't get the funnel shape if the distribution of sample sizes is approximately normal:

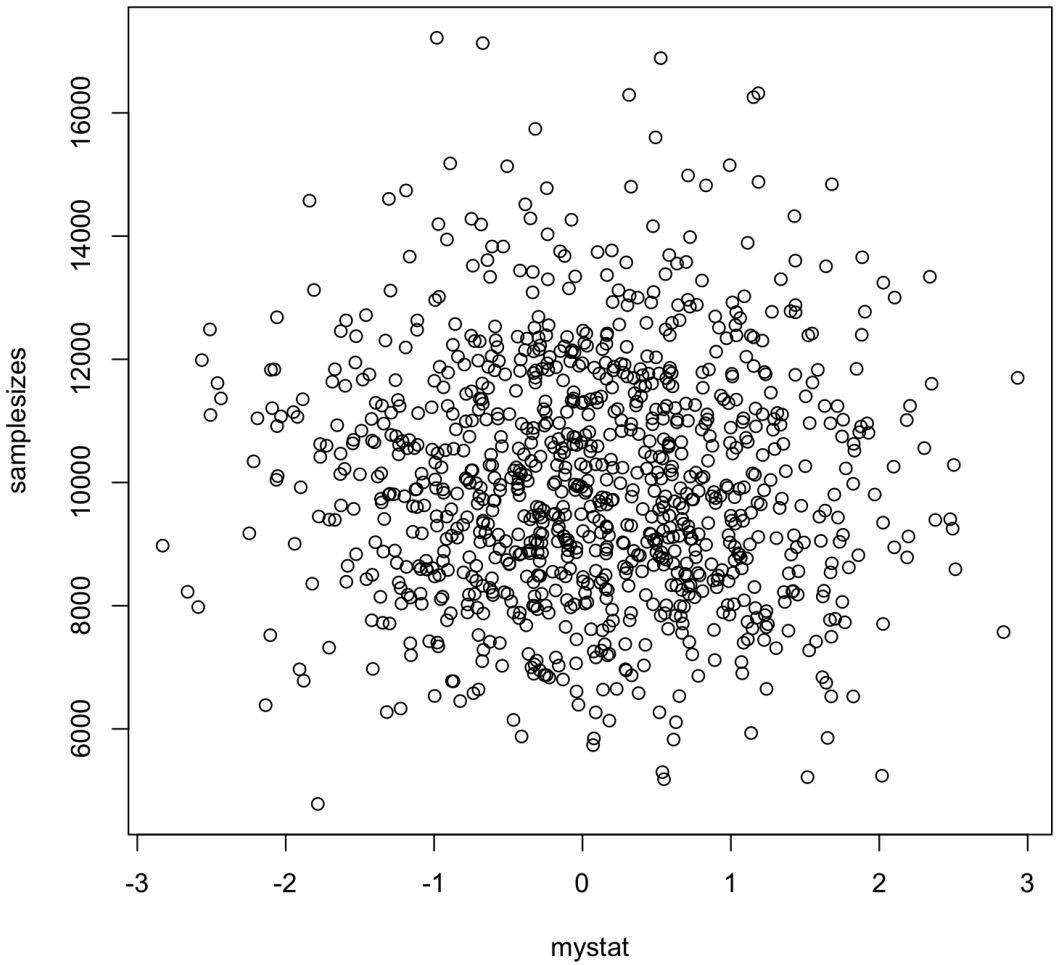


Meta-analysis

A bigger concern with funnel plots than funnel versus circular patterns is whether there is asymmetry. If only significant results are published, then you'll see some asymmetry in the plot, or gaps. Here is a simulated example where only p-values less than 0.1 are reported. This example uses normally distributed sample sizes.

```
> mystat <- NULL
> myp <- NULL
> for(i in 1:1000) {
+ xtemp <- rnorm(samplesizes[i])
+ mystat <- c(mystat,t.test(xtemp)$statistic)
+ myp <- c(myp,t.test(xtemp)$p.value)
+ }
> plot(mystat[myp <= .1],samplesizes[myp <= .1])
```

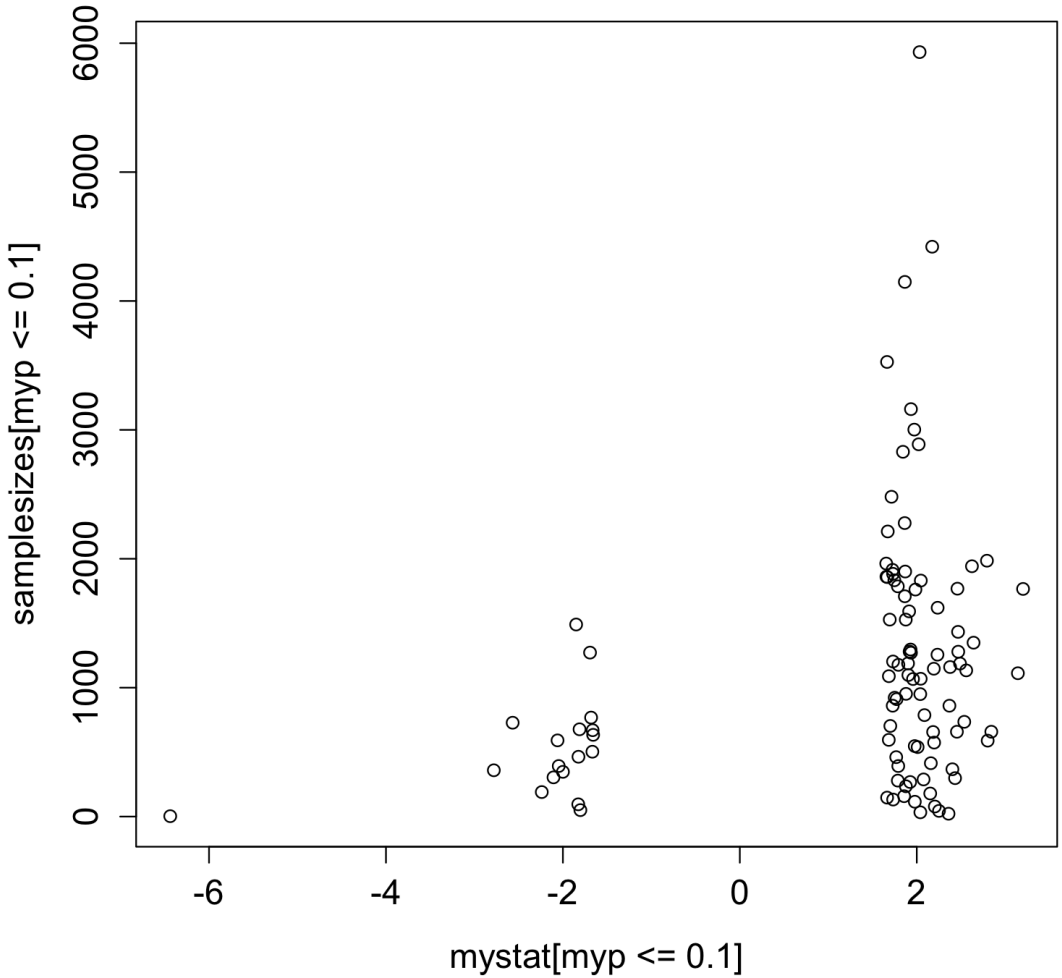
Based on the simulations I've tried, I don't get the funnel shape if the distribution of sample sizes is approximately normal:



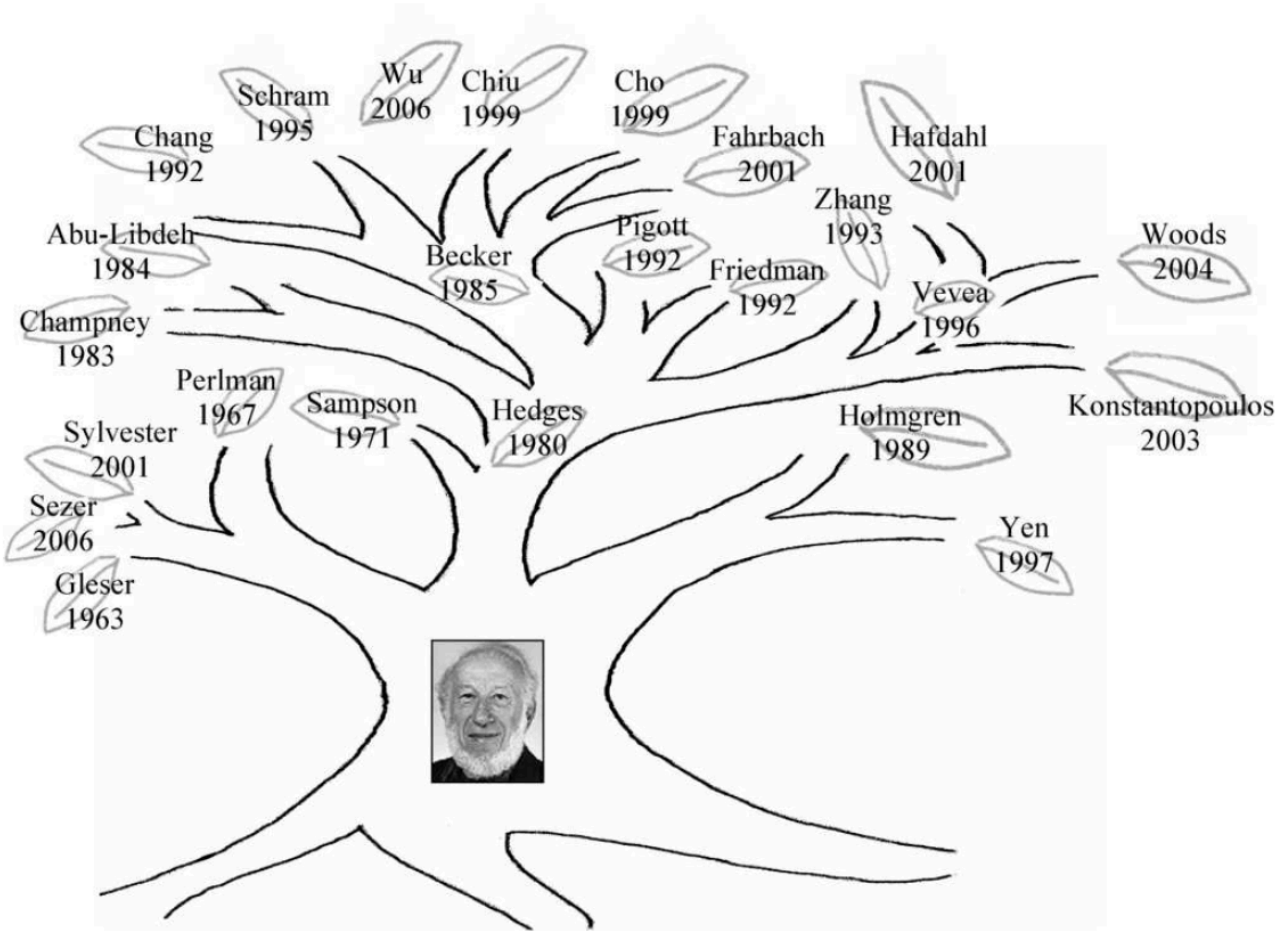
Meta-analysis

We see two clusters of points. If one-sided tests had been performed (which might be the case with χ^2 values), or if there is a genuine effect, then there would be asymmetry.

```
> mystat <- NULL
> myp <- NULL
> for(i in 1:1000) {
+ xtemp <- rnorm(samplesizes[i],.01)
+ mystat <- c(mystat,t.test(xtemp)$statistic)
+ myp <- c(myp,t.test(xtemp)$p.value)
+ }
> plot(mystat[myp <= .1],samplesizes[myp <= .1],cex.axis=1.3,c
```


samplesizes[myp <= 0.1]
0
1000
2000
3000
4000
5000
6000
-6
-4
-2
0
2
mystat[myp <= 0.1]

A particularly influential statistician in meta-analysis is Ingram Olkin (aged 90, still at Stanford). The following is an academic family tree of his PhD students and their academic descendants. *Statistical Science* 2007, Vol. 22, No. 3, 401–406 DOI: 10.1214/07-STS239



Meta-analysis

Olkin showed that vote-counting methods, which consider how many studies reject or don't reject a null hypothesis, don't work very well.

An approach developed by Olkin is to let T_{ij} be the effect of treatment j in study i. Then the collection of variables forms a matrix

$$T = \begin{pmatrix} T_{11} & \cdots & T_{iJ} \\ T_{21} & \cdots & T_{2J} \\ \vdots & & \vdots \\ T_{I1} & \cdots & T_{IJ} \end{pmatrix}$$

The matrix T is then a random matrix which can be studied using techniques from multivariate statistics.

Two basic types of meta-analysis assume that either that the true effect of a given treatment is the same in every study (fixed effect), or that the true effect of a given treatment in a study is random (random effect).

Approximate Bayesian Computation

Approximate Bayesian Computation is a type of Bayesian inference. In Bayesian inference, the likelihood of the data is used, but inferences combine information from the data and prior beliefs about the distribution of the data. The prior puts a probability distribution on parameters, meaning that parameters in the model are considered random variables.

Treating parameters as random variables is something that separates Bayesian statistics from frequentist statistics.

Combining the prior and the likelihood results in what is called a posterior distribution, which is a distribution describing the parameters. The idea is that if you have beliefs about plausible valies for the parameters, then these beliefs can be updated when you see new evidence (data).

Approximate Bayesian Computation

The relationship between the prior, the posterior, and the likelihood are often written as

$$p(\theta|\text{data}) \propto p(\theta)L(\theta|\text{data})$$

In other words, the posterior distribution of θ given the data is proportional to the likelihood times the prior. This comes from Bayes rule:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}$$

The denominator is constant with respect to θ, and that is why the proportionality symbol is often used. The value of the denominator is whatever makes the right hand side integrate to 1. However, since θ is not fixed, but is random, the right hand side must be integrated with respect to θ. This is typically not tractable, especially when θ is multi-dimensional.

Approximate Bayesian Computation

An additional complication is that the probability of the data is usually not known, but can be computed as

$$\int p(\text{data}|\theta)p(\theta)d\theta$$

but usually this quantity is difficult to compute.

Markov chain Monte Carlo gives a way to use simulation to approximate the posterior distribution. It generates a sequence: $\theta^{(1)}, \ldots, \theta^{(M)}$, and this sequence can be interpreted as simulated values from the posterior distribution.

These simulated values in theory converge to the posterior distribution. However, convergence can be very slow, requiring millions, or even billions, of iterations. Convergence can also be difficult to check, and for some problems can take days, weeks, or months of computer time. This has especially been a problem in population genetics where data consists of large amounts of DNA sequences and likelihoods are difficult to compute.

Approximate Bayesian Computation

Approximate Bayesian Computation is an alternative to MCMC as a way to approximate the posterior distribution. The basic idea is to use the prior distribution of θ to simulate values of θ directly. For a given simulated value of $\theta^{(i)}$, a data set can be simulated. Summary statistics from the simulation are used to compare to the same summary statistics obtained from the original data.

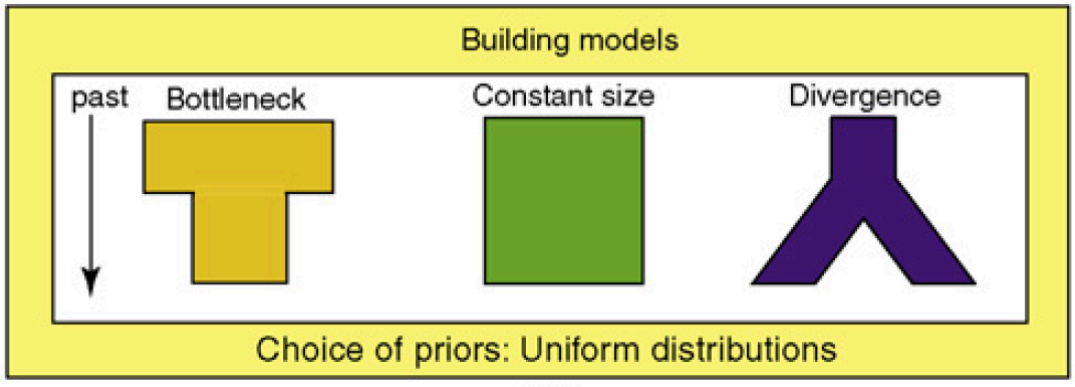
If the summary statistics in the simulated data set match the summary statistics of the original data to within a given level of tolerance, then the simulated θ value is "accepted". The accepted values of θ,

$$\theta^{(j_1)}, \ldots, \theta^{(j_M)}$$

are then used as an approximation to the posterior distirbution of θ. For example, if θ is one dimensional, then you can approximate $P(\theta > a)$ using the proportion of $\theta^{(j)}$ values (the proportion of accepted simulated θs) that are greater than a.

Approximate Bayesian Computation

Model selection can also be done using ABC. A popular applcation is in population genetics, where the goal is to understand past demographic scenarios. For example, current patterns in diversity for some population could have been caused perhaps by past population size changes, constant population size, or a case where populations have diverged.



TREE-1239; No. of Pages 9
ARTICLE IN PRESS

Review

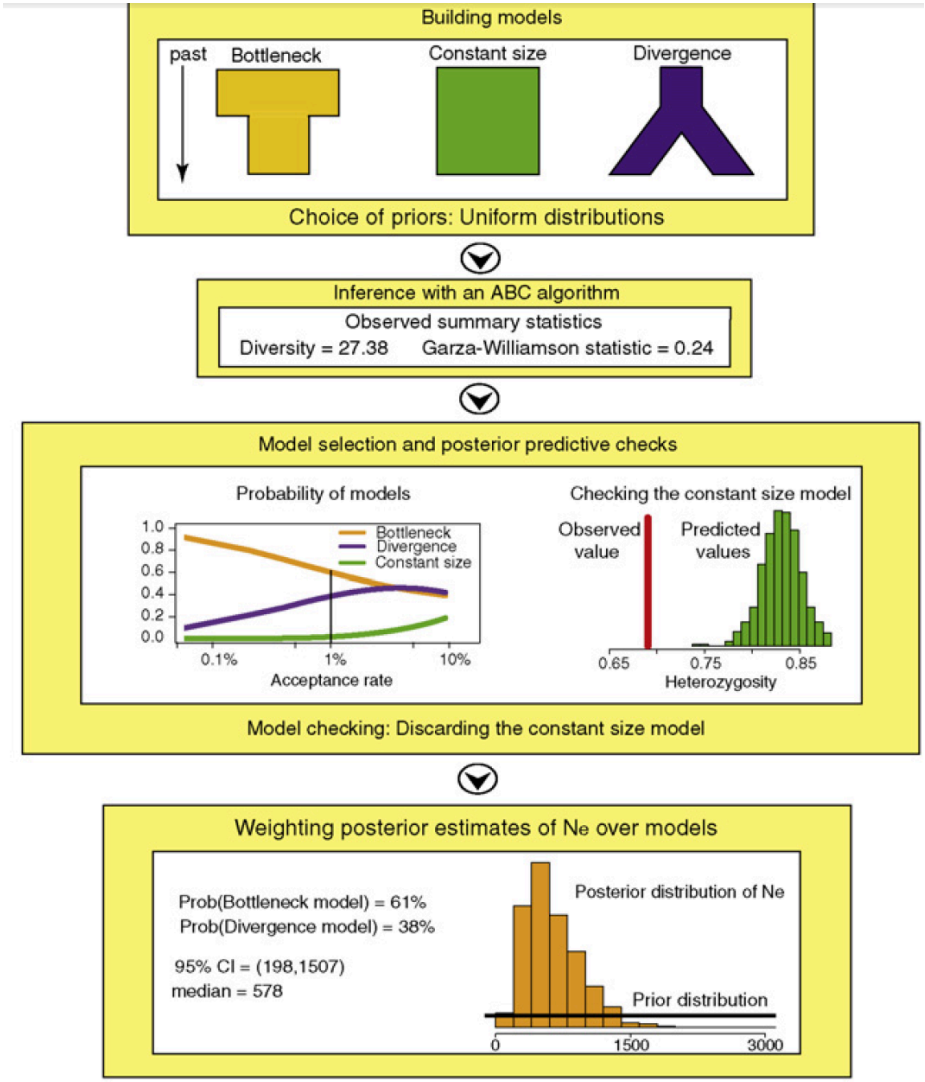
Cell PRESS

Approximate Bayesian Computation (ABC) in practice

Katalin Csilléry[1], Michael G.B. Blum[1], Oscar E. Gaggiotti[2] and Olivier François[1]

[1] Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité, Centre National de la Recherche Scientifique UMR5525, Université Joseph Fourier, 38706 La Tronche, France
[2] Laboratoire d'Ecologie Alpine, Centre National de la Recherche Scientifique UMR5553, Université Joseph Fourier, 38041 Grenoble, France

Approximate Bayesian Computation



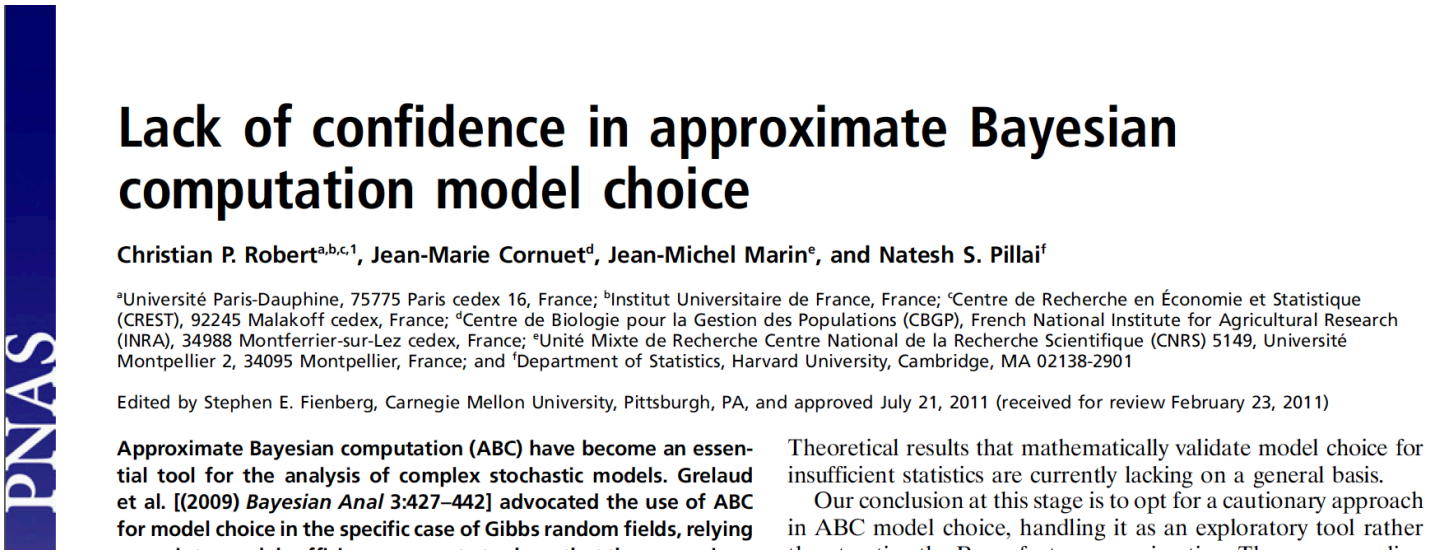
Approximate Bayesian Computation

One issue with ABC is that the summary statistics need to be sufficient statistics in order for the inferences to be valid. If the summary statistics are not sufficient statistics, then inferences can be misleading. In population genetics, it often isn't clear which statistics are sufficient and which are not, so proving that something is a sufficient statistic can be an important contribution. Sometimes, researchers proceed with ABC without knowing whether their statistics are sufficient or not.

For the demography example, what is of interest are the three models as well as the population sizes. If you use a summary statistic such as the proportion of people who are homozygous, it isn't obvious that different combinations of models and population sizes will necessarily yield the different distributions of the summary statistic. It must be the case that a summary statistic will have a different distribution for unique parameter combinations that you are estimating.

Approximate Bayesian Computation

"without further justification, ABC methods cannot be trusted to discriminate between two competing models when based on insufficient summary statistics."



PNAS

Lack of confidence in approximate Bayesian computation model choice

Christian P. Robert[a,b,c,1], Jean-Marie Cornuet[d], Jean-Michel Marin[e], and Natesh S. Pillai[f]

[a]Université Paris-Dauphine, 75775 Paris cedex 16, France; [b]Institut Universitaire de France, France; [c]Centre de Recherche en Économie et Statistique (CREST), 92245 Malakoff cedex, France; [d]Centre de Biologie pour la Gestion des Populations (CBGP), French National Institute for Agricultural Research (INRA), 34988 Montferrier-sur-Lez cedex, France; [e]Unité Mixte de Recherche Centre National de la Recherche Scientifique (CNRS) 5149, Université Montpellier 2, 34095 Montpellier, France; and [f]Department of Statistics, Harvard University, Cambridge, MA 02138-2901

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved July 21, 2011 (received for review February 23, 2011)

Approximate Bayesian computation (ABC) have become an essential tool for the analysis of complex stochastic models. Grelaud et al. [(2009) *Bayesian Anal* 3:427–442] advocated the use of ABC for model choice in the specific case of Gibbs random fields, relying

Theoretical results that mathematically validate model choice for insufficient statistics are currently lacking on a general basis.

Our conclusion at this stage is to opt for a cautionary approach in ABC model choice, handling it as an exploratory tool rather

Approximate Bayesian Computation

The sort of trick in model selection is to replace some family of distributions with a larger family that includes different models. For example, suppose you are not sure whether data is exponential or χ^2. Then instead of having two families of distributions:F_λ (for the exponential) G_k (for the χ^2, we have a larger family of distributions, $\mathcal{F}_{\theta,m}$ If $m = 1$, then the distribution is exponential with rate θ. If $m = 2$, then the distribution is χ^2 with mean θ. Things can get weirder if the different models have different numbers of parameters.

A point made in the Robert et al. PNAS article is that if T is a sufficient statistic for both distributions F and G, it doesn't follow that it is a sufficient statistic for $\mathcal{F}$.