A New Look at Clinical Trials

Erik van Zwet
Leiden University Medical Center

eScience Center Analytics Special Interest Group
November 28, 2024

Clinical trials

The "essence" of a clinical trial is a set of 3 numbers: (β, b, s).

- β is the unobserved, "true" effect of the treatment.
- b is a normally distributed, unbiased estimator of β with standard error s.

It's helpful to think of the estimate b as the true effect β plus a normally distributed "error":

$$b = \beta + N(0, s).$$

z-statistic and SNR

Two more quantities to consider: The z-statistic $z = b/s$ and the (unobserved!) signal-to-noise ratio $SNR = \beta/s$.

The z-statistic and the SNR have a very simple relation:

$$b = \beta + N(0, s) \quad \overset{\text{divide by } s}{\Rightarrow} \quad z = SNR + N(0, 1).$$

Think of the z-stat as the SNR plus standard normal "error".

Hypothesis testing

So, the z-statistic has the normal distribution with mean SNR and standard deviation 1.

Suppose we want to test $H_0 : \beta = 0$.

- If $|z| > 1.96$ then the p-value is less than 0.05. If $\beta = 0$ then $SNR = 0$, and we have

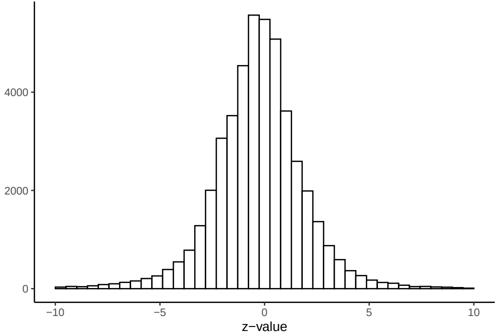
```
pnorm(-1.96,0,1) + 1 - pnorm(1.96,0,1) = 0.05
```

- The power depends on the SNR. For example, if $SNR = 2.8$ then the power is 80% because

```
pnorm(-1.96,2.8,1) + 1 - pnorm(1.96,2.8,1) = 0.8
```

Cochrane Database of Systematic Reviews (CDSR)

We have the *z*-statistics for the primary efficacy outcomes from about 23,000 randomized controlled trials (RCTs) from the CDSR.



Note: It's *not* standard normal. It would only be standard normal if all the treatments had exactly no effect.

The distribution of *z*-stats and *SNRs*

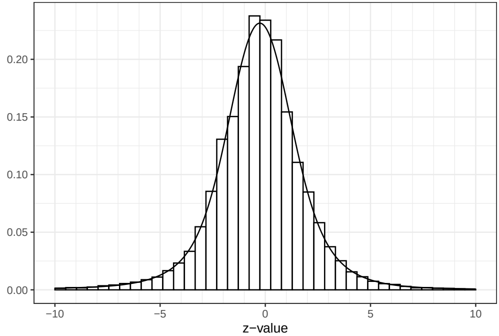
Obviously, we can estimate the distribution of the *z*-statistics across the CDSR, but also – surprisingly – of the *SNRs*.

We use the fact that the *z*-statistics are equal to the *SNRs* plus standard normal errors!

Step 1: Estimate the distribution of the *z*-statistics directly.

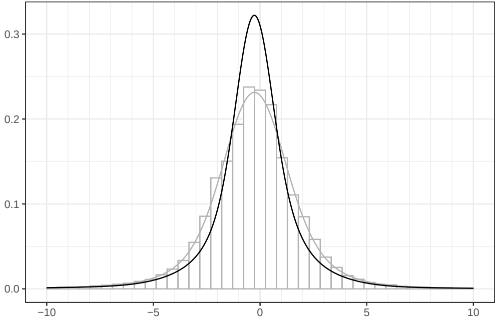
Step 2: *Derive* the distribution of the *SNRs* by removing the standard normal error component (i.e. denoising or deconvolution).

Step 1: Distribution of the z-statistics



The distribution of z is well approximated by a mixture of 4 normal components.

Step 2: Distribution of the *SNRs*



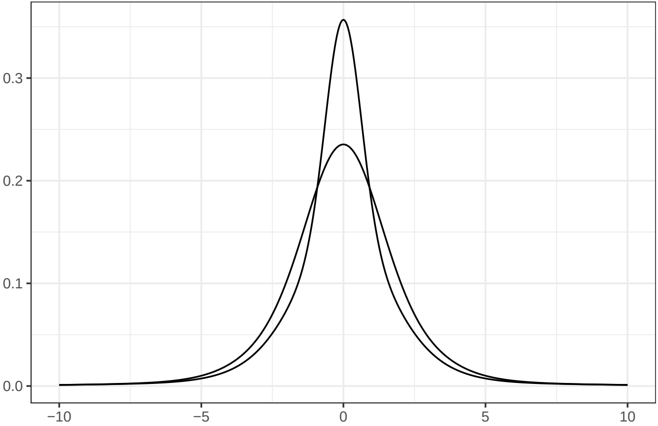
Subtract 1 from the variances of each of the mixture components.

Synthetic CDSR

We can use the estimated distributions of the z-stats and the *SNRs* to build a "synthetic" version of the CDSR with the same statistical properties as the real CDSR.

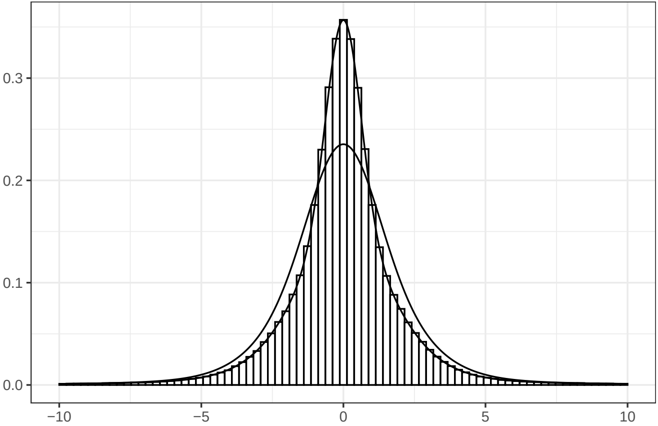
Step 1

Generate a sample of 1 million *SNRs*.



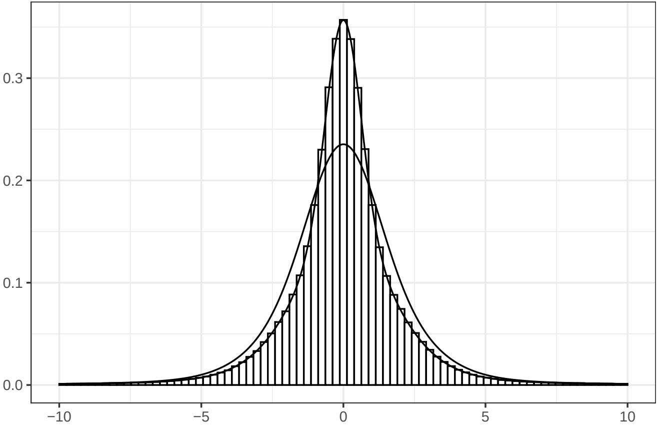
Step 1

Generate a sample of 1 million *SNRs* – done!



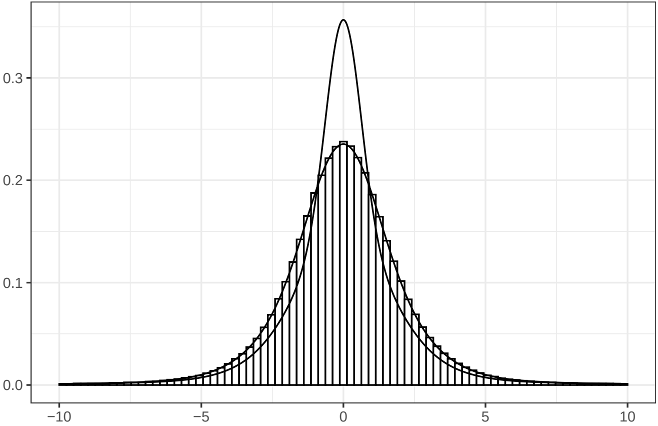
Step 2

Add normal noise: `zstat = SNR + rnorm(10^6,0,1)`



Step 2

Add normal noise: `zstat = SNR + rnorm(10^6,0,1)` – done!



Synthetic CDSR

We have now a "synthetic" version of the CDSR with a million trials — or at least their z-stats and *SNRs* — that have the same statistical properties as the real CDSR.

- In the synthetic CDSR we *observe* the *SNRs*!

This will enable us to get some important insights.

PS We could have used math to get all the results that I'll show, but I think that Monte Carlo simulation is easier to understand.

Power

RCTs are designed to have 80% or 90% power for testing $H_0 : \beta = 0$ against an effect that is considered to be of minimal clinical interest, or plausible, or both.

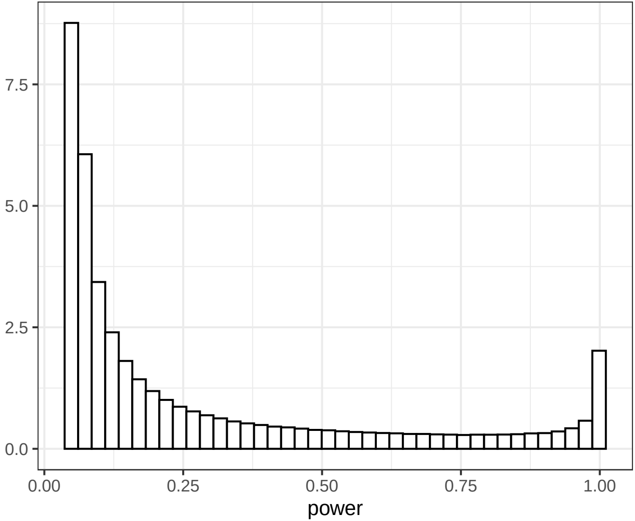
In fact, the *SNR* is larger than 2.8 in only 12% of the trials.

Let's look at the distribution of the *actual* power (i.e. the power against the true effect) in more detail.

- Take our sample of a million *SNRs*. Next,

```
power = pnorm(-1.96,SNR,1) + 1 - pnorm(1.96,SNR,1)
```

Distribution the power (median=13%, mean=29%)



Low power

The *actual* power is often very low, which won't surprise anyone who has ever been involved in a sample size calculation (which is sometimes called "the sample size samba.")

Low power has *two* consequences:

1. If $p > 0.05$ you might be discarding a useful treatment because you didn't collect enough information to show that it works.
2. If $p < 0.05$ you got very lucky. Therefore, your effect estimate is likely overestimated and replication attempts will likely fail. This is called the winner's curse.

Winner's curse

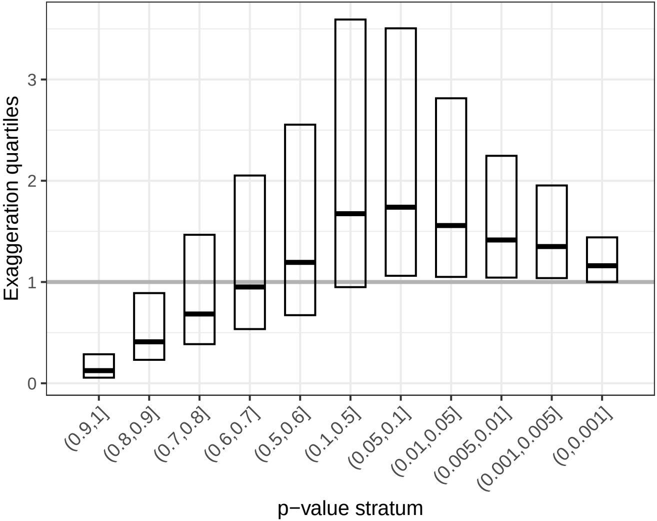
Define the exaggeration

$$\frac{|b|}{|\beta|} = \frac{|b|/s}{|\beta|/s} = \frac{|z|}{|SNR|}.$$

Take our sample of a million *SNRs*. Next,

1. `zstat = SNR + rnorm(10^6,0,1)`
2. `exaggeration = abs(zstat)/abs(SNR)`
3. `pval = 2*pnorm(-abs(zstat))`

Winner's curse (quartiles)



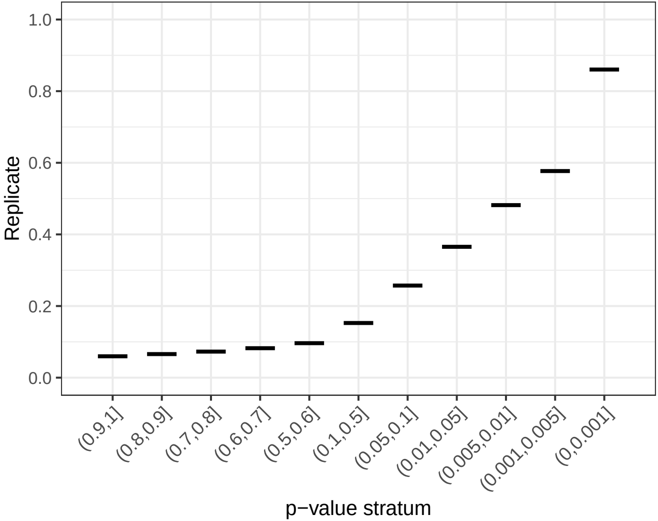
Replication probability (predictive power)

The probability of a significant result when a study with a particular *p*-value would be repeated exactly.

Take our sample of a million *SNRs*. Next,

1. `zstat = SNR + rnorm(10^6,0,1)`
2. `pval = 2*pnorm(-abs(zstat))`
3. `zrepl = SNR + rnorm(10^6,0,1)`
4. `prepl = 2*pnorm(-abs(zrepl))`

Replication probability



Sign agreement

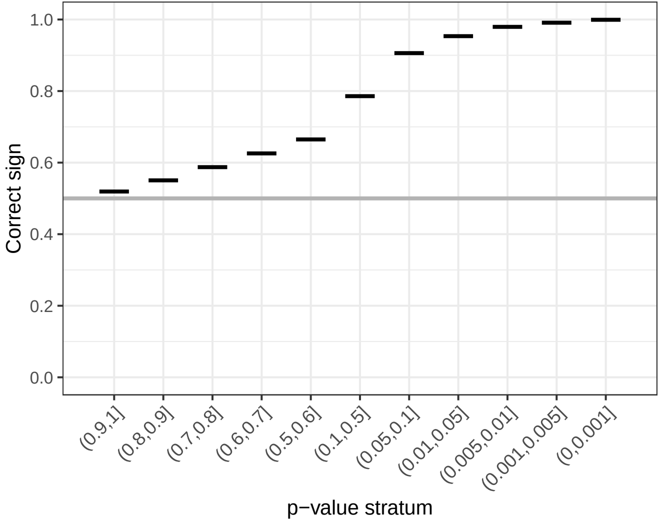
Note that

$$\beta \times b > 0 \;\Leftrightarrow\; SNR \times z > 0.$$

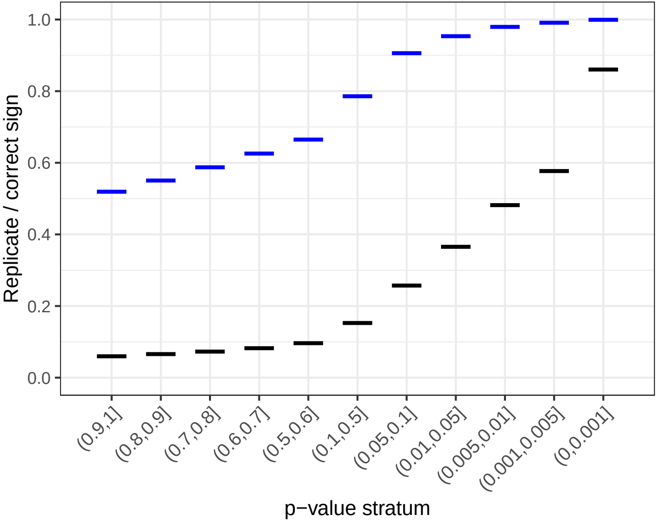
Take our sample of a million *SNRs*. Next,

1. `zstat = SNR + rnorm(10^6,0,1)`
2. `agree = (SNR * z > 0)`
3. `pval = 2*pnorm(-abs(zstat))`

Sign agreement



Mind the gap!


Replicate / correct sign
1.0
0.8
0.6
0.4
0.2
0.0
(0.9,1]
(0.8,0.9]
(0.7,0.8]
(0.6,0.7]
(0.5,0.6]
(0.1,0.5]
(0.05,0.1]
(0.01,0.05]
(0.005,0.01]
(0.001,0.005]
(0,0.001]
p−value stratum

Take home

Many trials of low power against the true effect. This has *two* consequences:

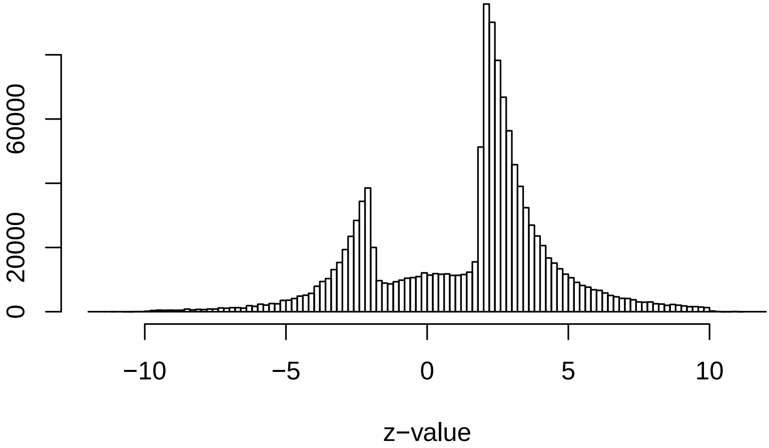
1. If $p > 0.05$ you might be discarding a useful treatment because you didn't collect enough information to show that it works.
2. If $p < 0.05$ you got very lucky. Therefore, your effect estimate is likely overestimated and replication attempts will likely fail. This is called the winner's curse.

A potential solution is the give the effect estimate a "haircut" (shrinkage estimation).

Three thorny issues

1. Exchangeability.
2. Coining.
3. Publication bias.

A million *z*-values from Medline (Barnett and Wren, 2019)



Further reading

1. with Andrew Gelman: A proposal for informative default priors scaled by the standard error of estimates (2022) in *The American Statistician*
2. with Simon Schwab and Stephen Senn: The statistical properties of RCTs and a proposal for shrinkage (2021) in *Statistics in Medicine*
3. with Simon Schwab and Sander Greenland: Addressing exaggeration of effects from single RCTs (2022) in *Significance*
4. with Steven Goodman: How large should the next study be? Predictive power and sample size requirements for replication studies (2022) in *Statistics in Medicine*
5. with Lu Tian and Robert Tibshirani: Evaluating a shrinkage estimator for the treatment effect in clinical trials. (2023) in *Statistics in Medicine*
6. with Andrew Gelman, Sander Greenland, Guido Imbens, Simon Schwab and Steven Goodman: A new look at p values for randomized clinical trials. (2024) in *NEJM Evidence*

Coverage

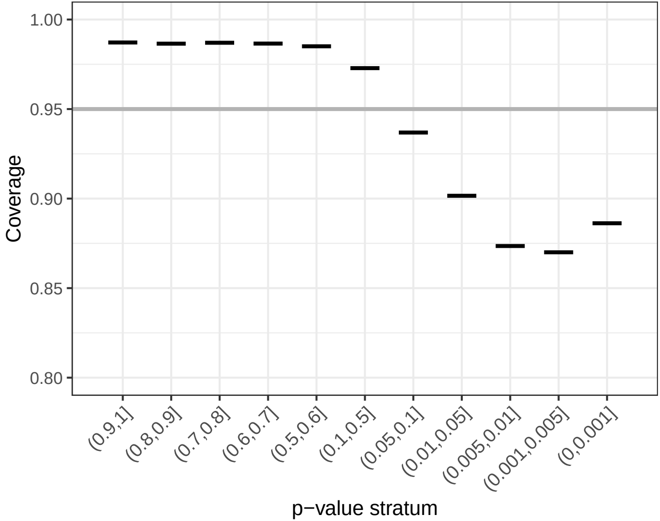
Note that

$$b - 1.96s < \beta < b + 1.96s \;\Leftrightarrow\; z - 1.96 < SNR < z + 1.96$$

Take our sample of a million *SNRs*. Next,

1. `zstat = SNR + rnorm(10^6,0,1)`
2. `cover = (SNR > z - 1.96) & (SNR < z + 1.96)`
3. `pval = 2*pnorm(-abs(zstat))`

Coverage


Coverage
1.00
0.95
0.90
0.85
0.80
(0.9,1]
(0.8,0.9]
(0.7,0.8]
(0.6,0.7]
(0.5,0.6]
(0.1,0.5]
(0.05,0.1]
(0.01,0.05]
(0.005,0.01]
(0.001,0.005]
(0,0.001]
p−value stratum