A New Look at *P*-values for RCTs

Erik van Zwet, Leiden University Medical Center

with: Simon Schwab, Andrew Gelman, Stephen Senn
Sander Greenland, Guido Imbens, Steve Goodman

NIH Collaboratory Grand Rounds
April 5, 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.

This means that the 95% confidence interval for β goes from $b - 1.96\,s$ to $b + 1.96\,s$.

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

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

z-statistic and SNR

The z-stat $z = b/s$ and the signal-to-noise ratio $SNR = \beta/s$.

$$b = \beta + N(0, s) \quad \Rightarrow \quad z = SNR + N(0, 1).$$

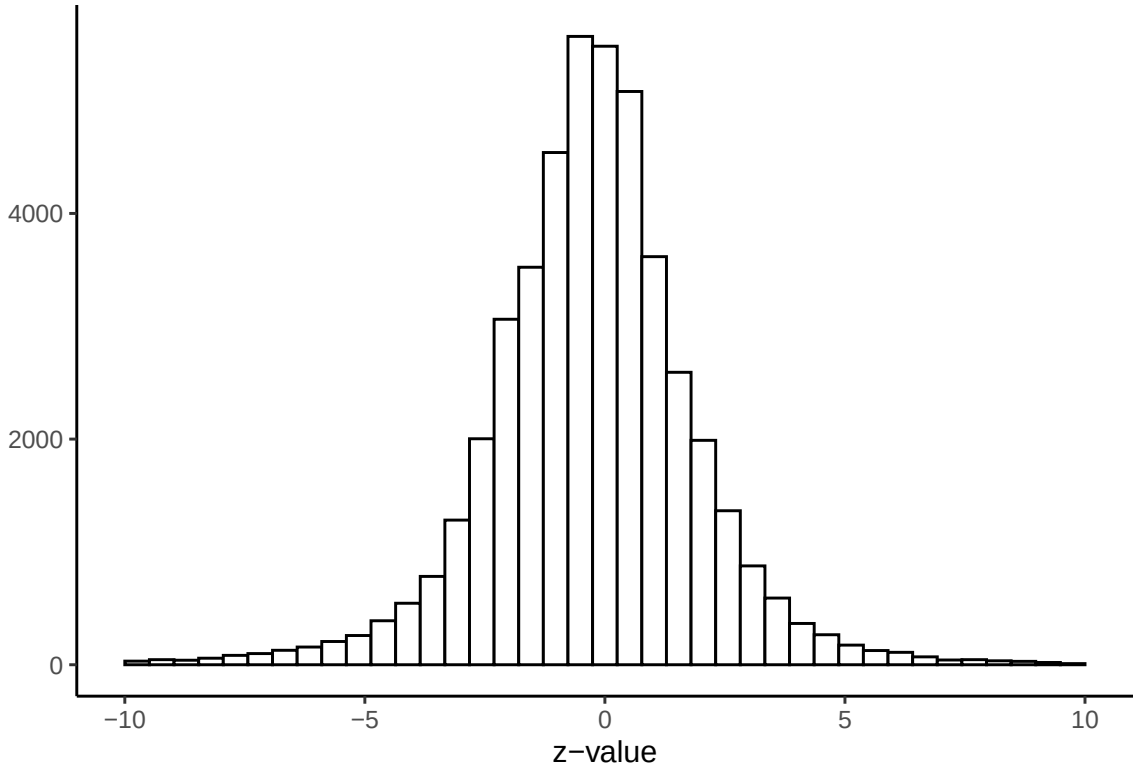
Think of the z-stat as the SNR plus standard normal "error".

Usually, we want to test $H_0 : \beta = 0$. Then:

- If $|z| > 1.96$ then the p-value is less than 0.05.
- The power depends on the SNR. For example, if $SNR = 0$ then the power is 5%, and if $SNR = 2.8$ then the power is 80%.

Cochrane Database of Systematic Reviews (CDSR)

We have about 23,000 z-stats from RCTs from the CDSR.



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

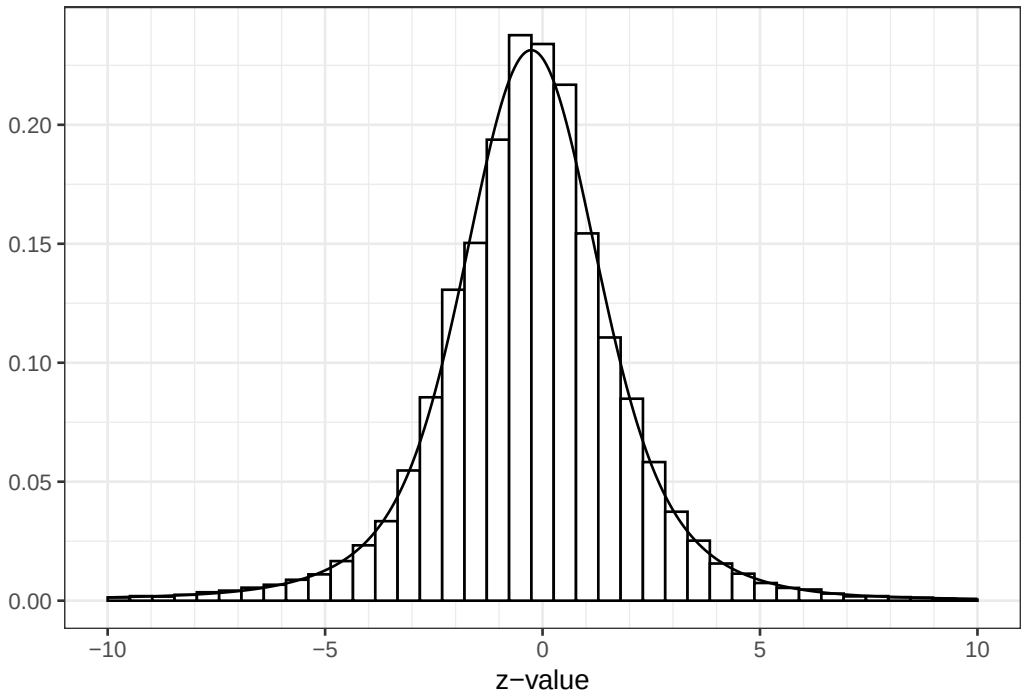
z-stat and *SNR*

We can estimate the distribution of the z-stat $z = b/s$, but also — surprisingly — of the signal-to-noise ratio $SNR = \beta/s$.

Step 1: Estimate the distribution of z directly from the observed z-stats.

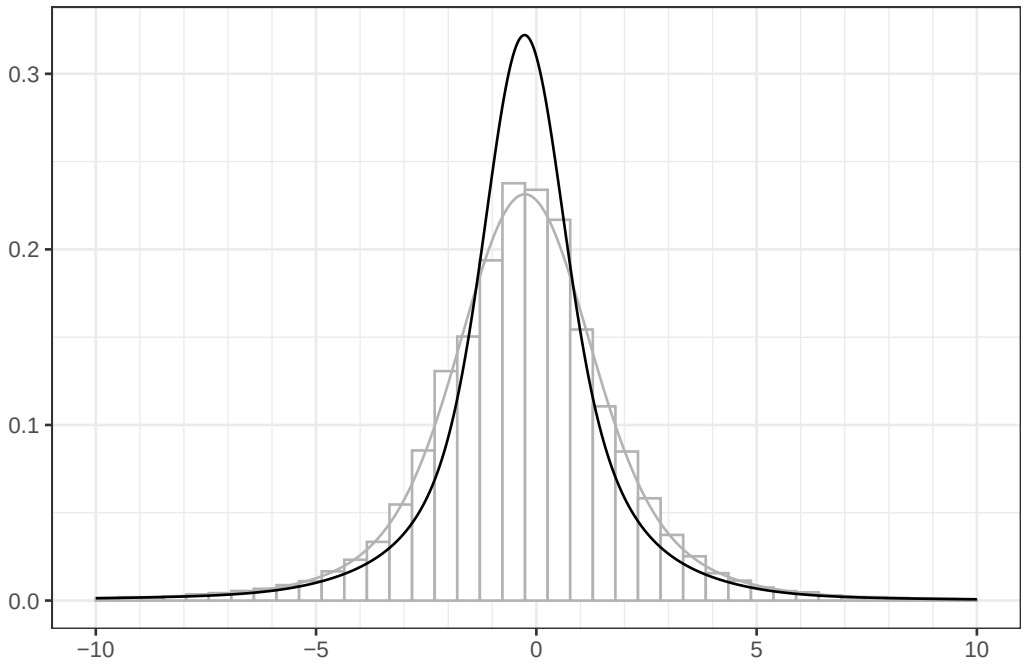
Step 2: Derive the distribution of *SNR* by "removing" the standard normal error component (i.e. denoising or deconvolution).

Step 1: Distribution of the z-stats



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

Step 2: Distribution of the *SNRs*



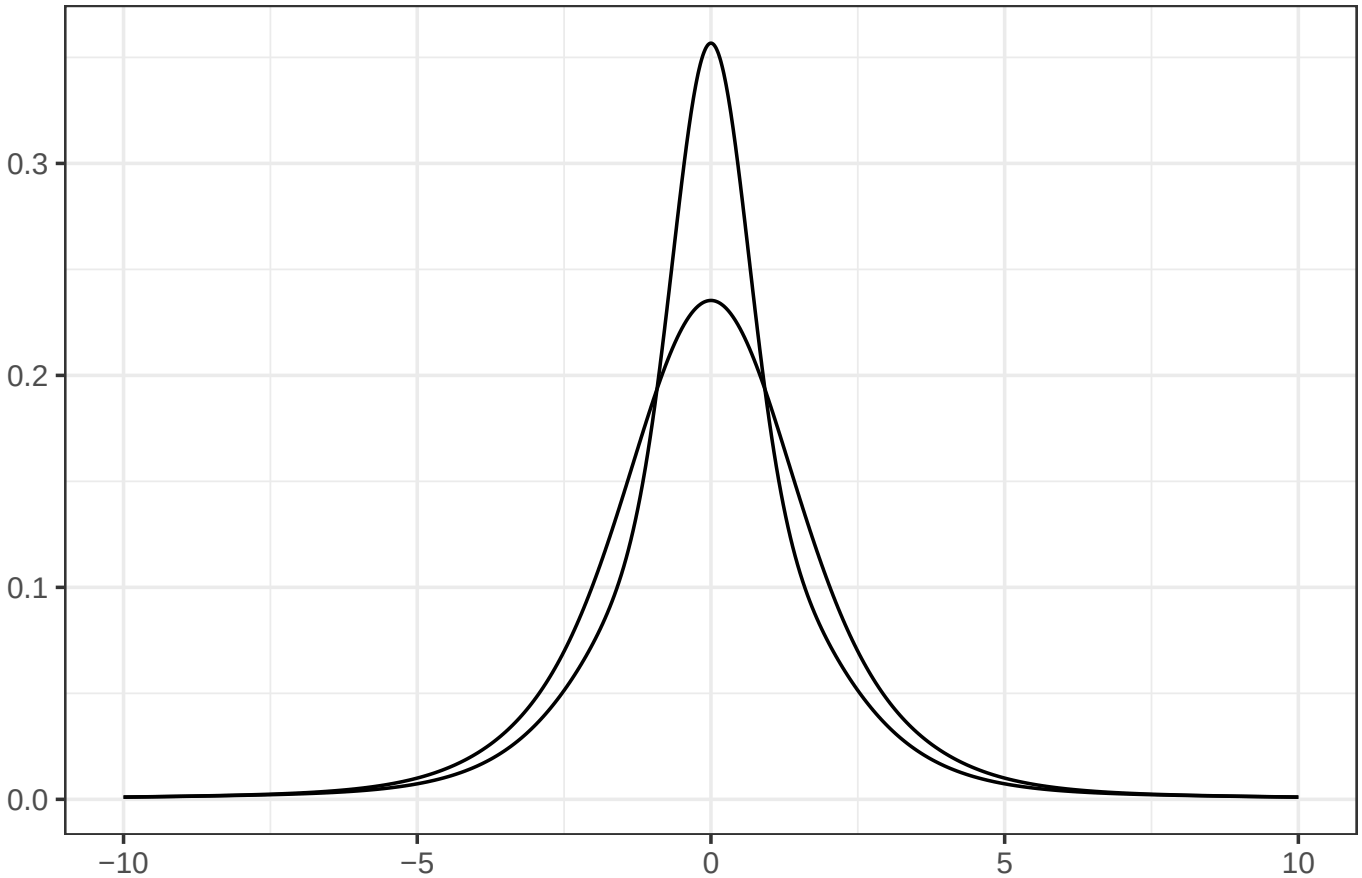
Subtract 1 from the variances of each of the mixture components.

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.

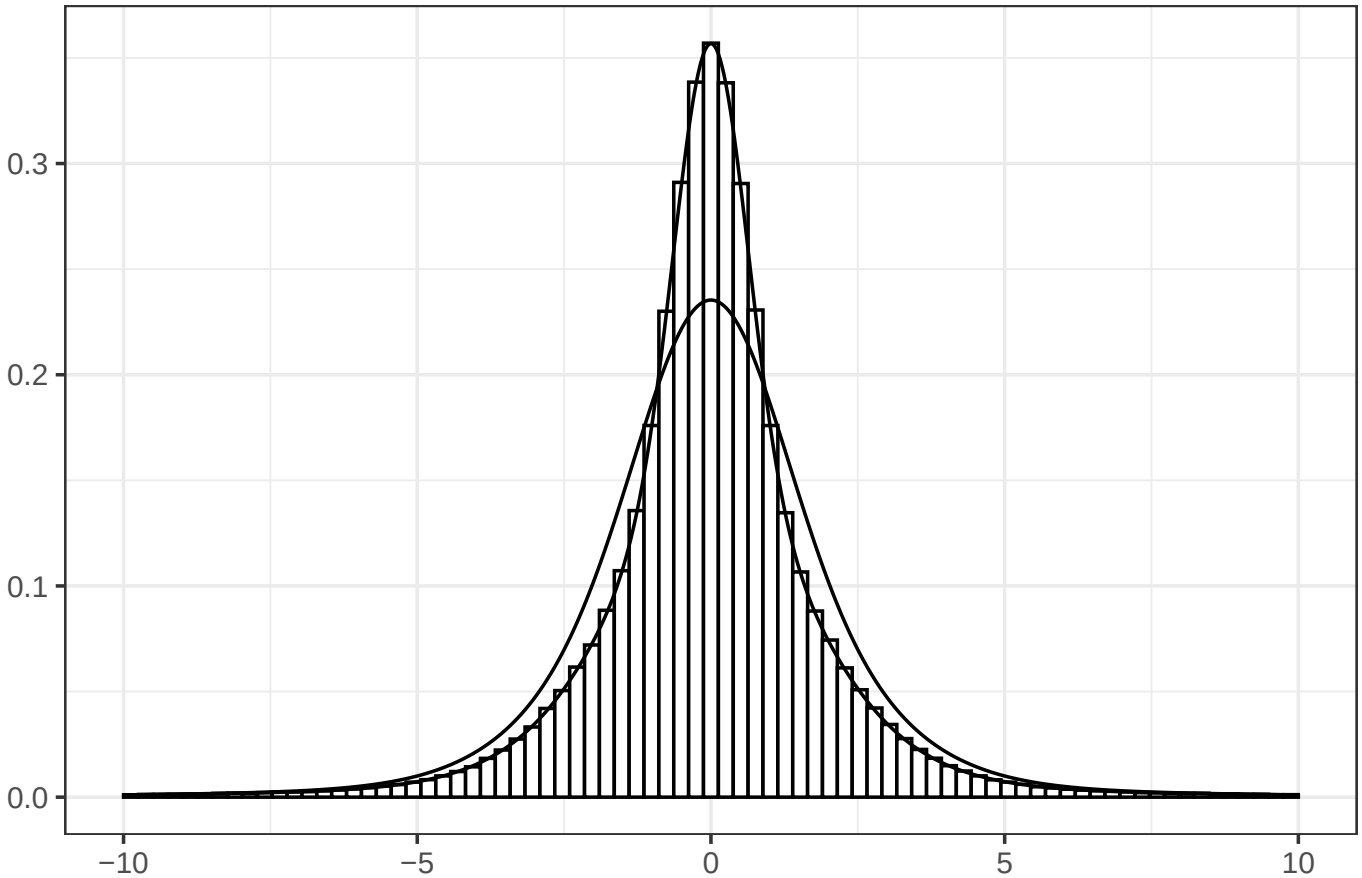
Step 1

Generate a sample a million *SNRs*.



Step 1

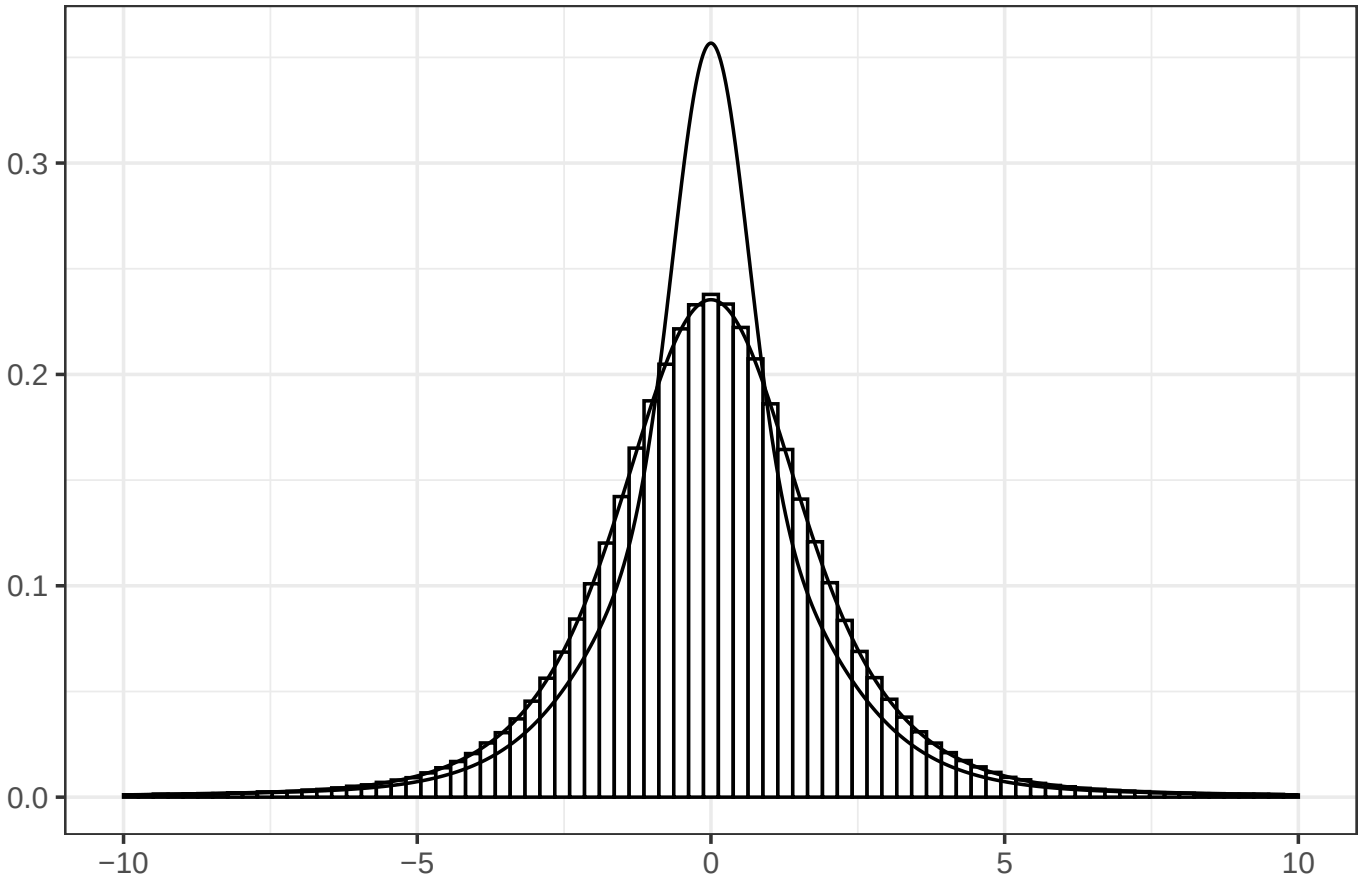
Generate a sample a million *SNRs*.



b is an unbiased estimator of β with standard error *s*. Define $z = b/s$ and SNR$=\beta/s$. Then $z = SNR + N(0, 1)$.

Step 2

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



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 *do* observe the *SNR*!

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 the synthetic CDSR is easier to understand.

Power

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

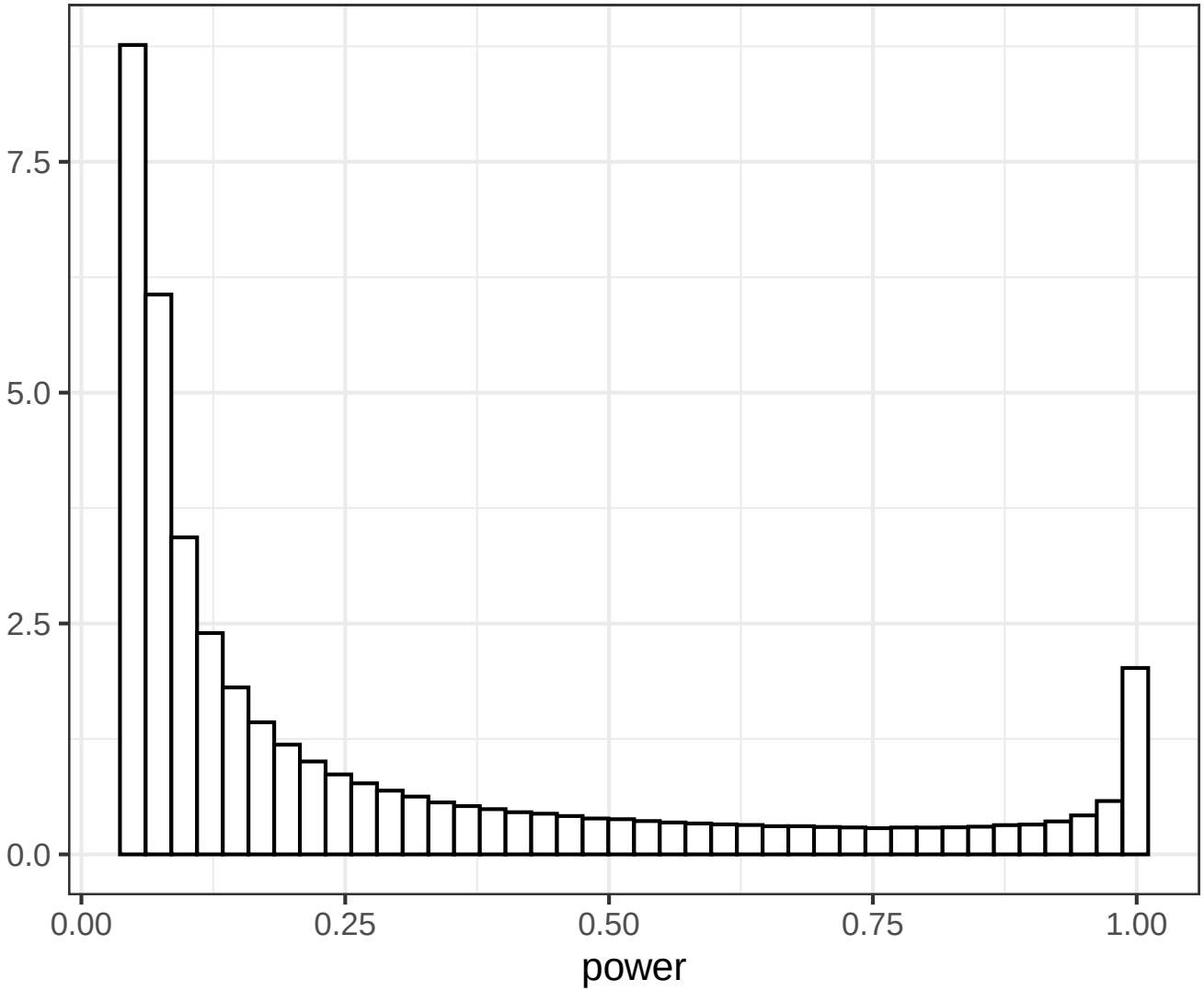
In fact, the *SNR* is larger than 2.8 in only 12% of the trials.

Let's look at the distribution of the *actual* power.

- 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 danced 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

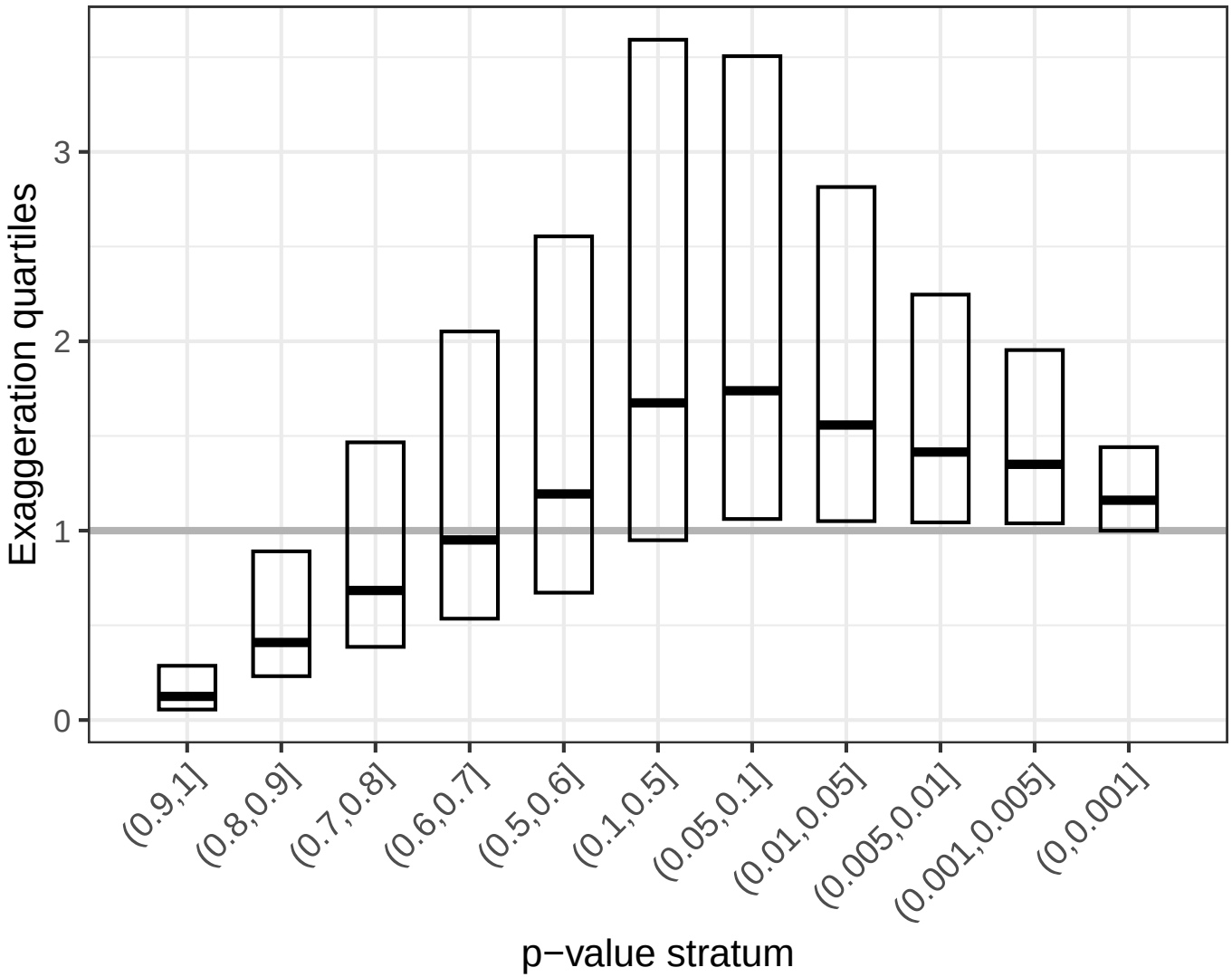
Define the exaggeration

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

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)



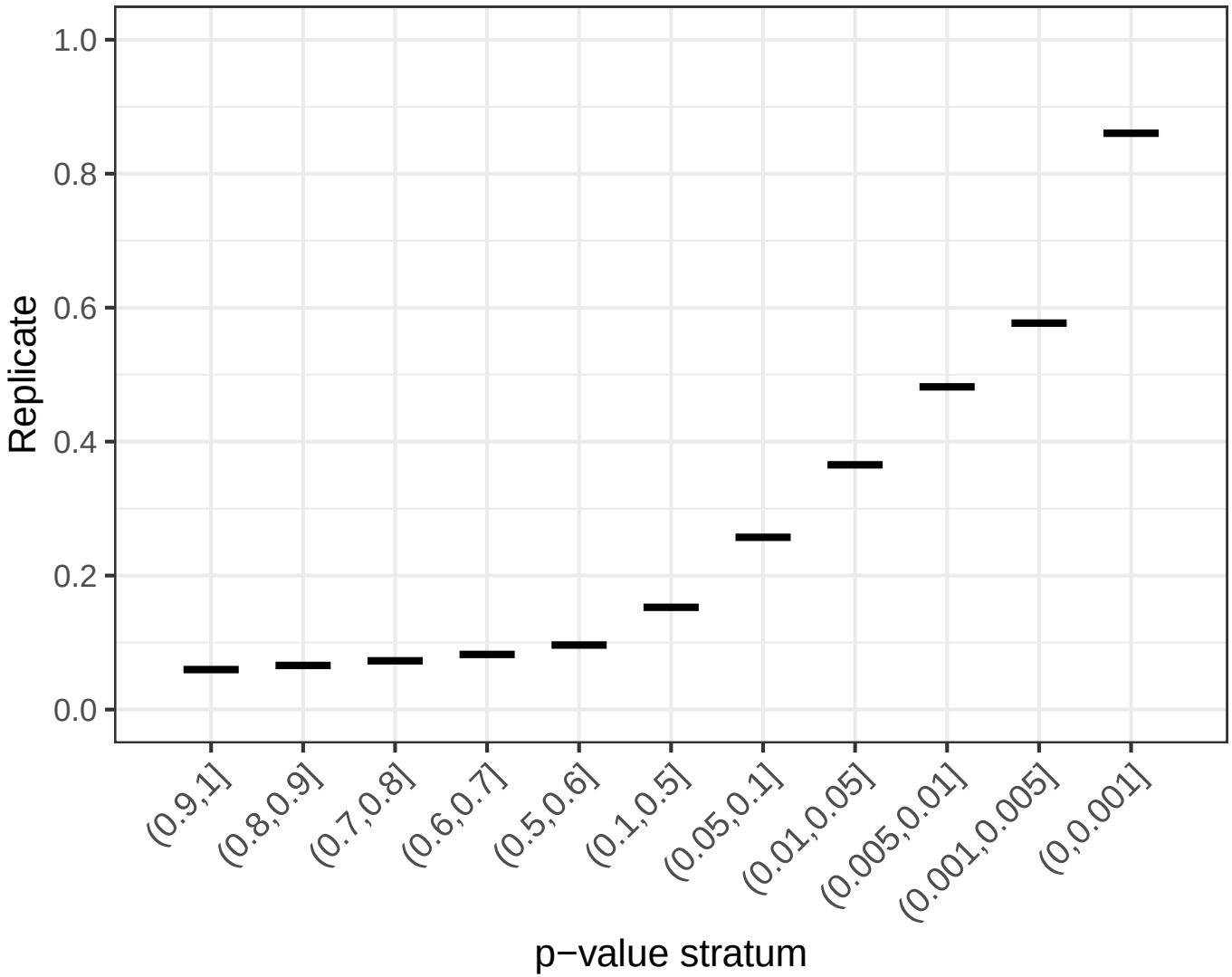
Replication probability

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

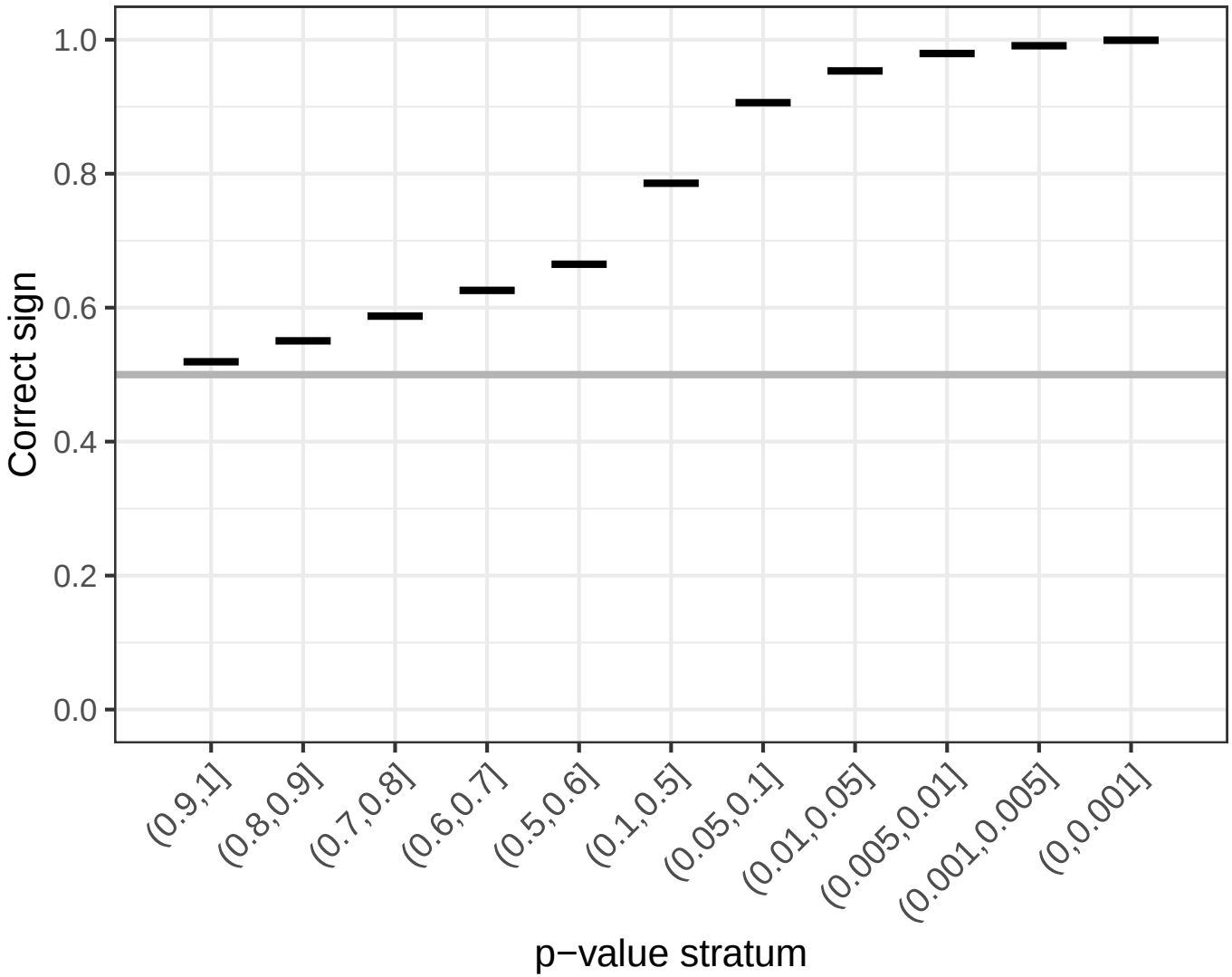
Note that

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

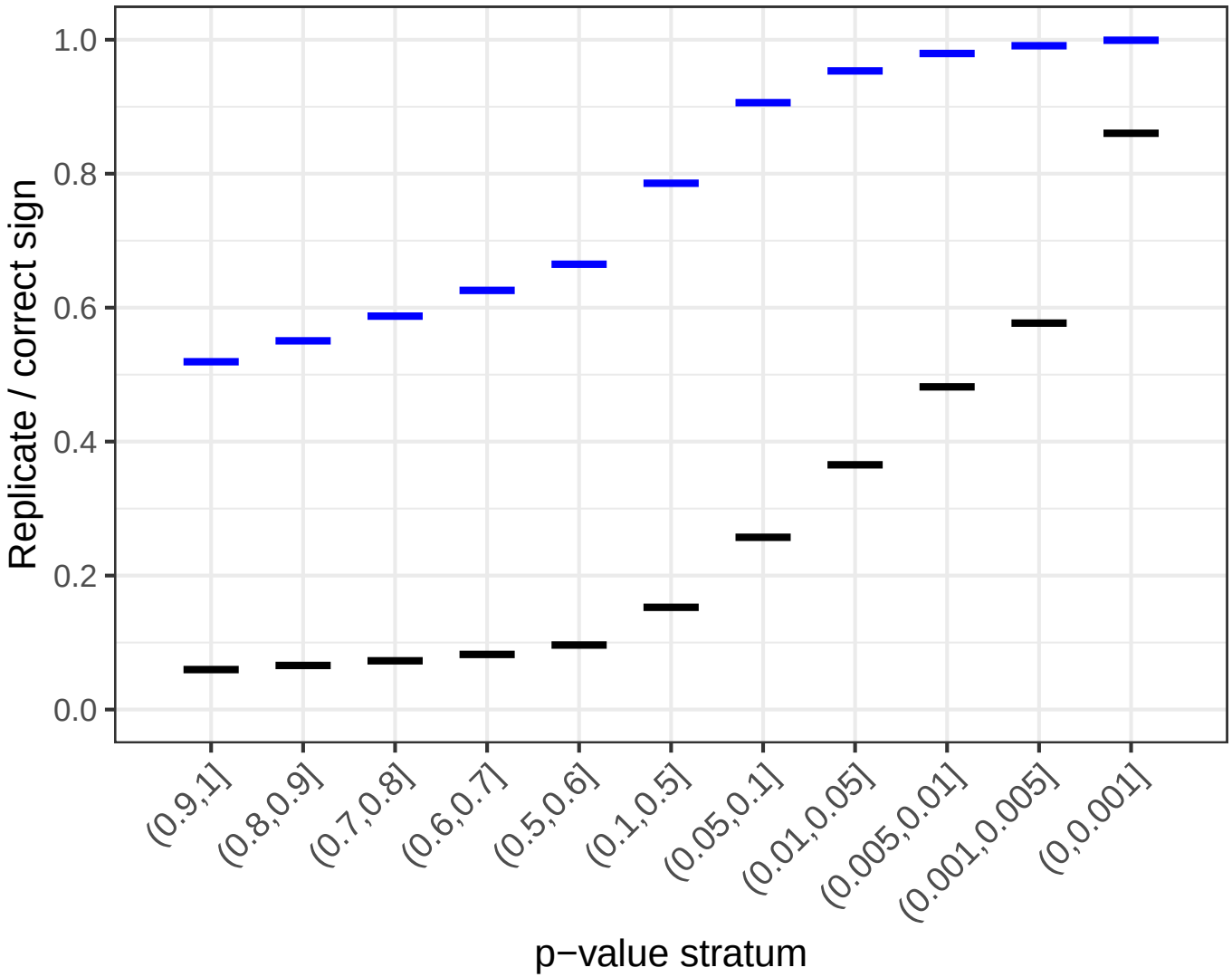
Take our sample of a million *SNRs*. Next,

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

Three thorny issues

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

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

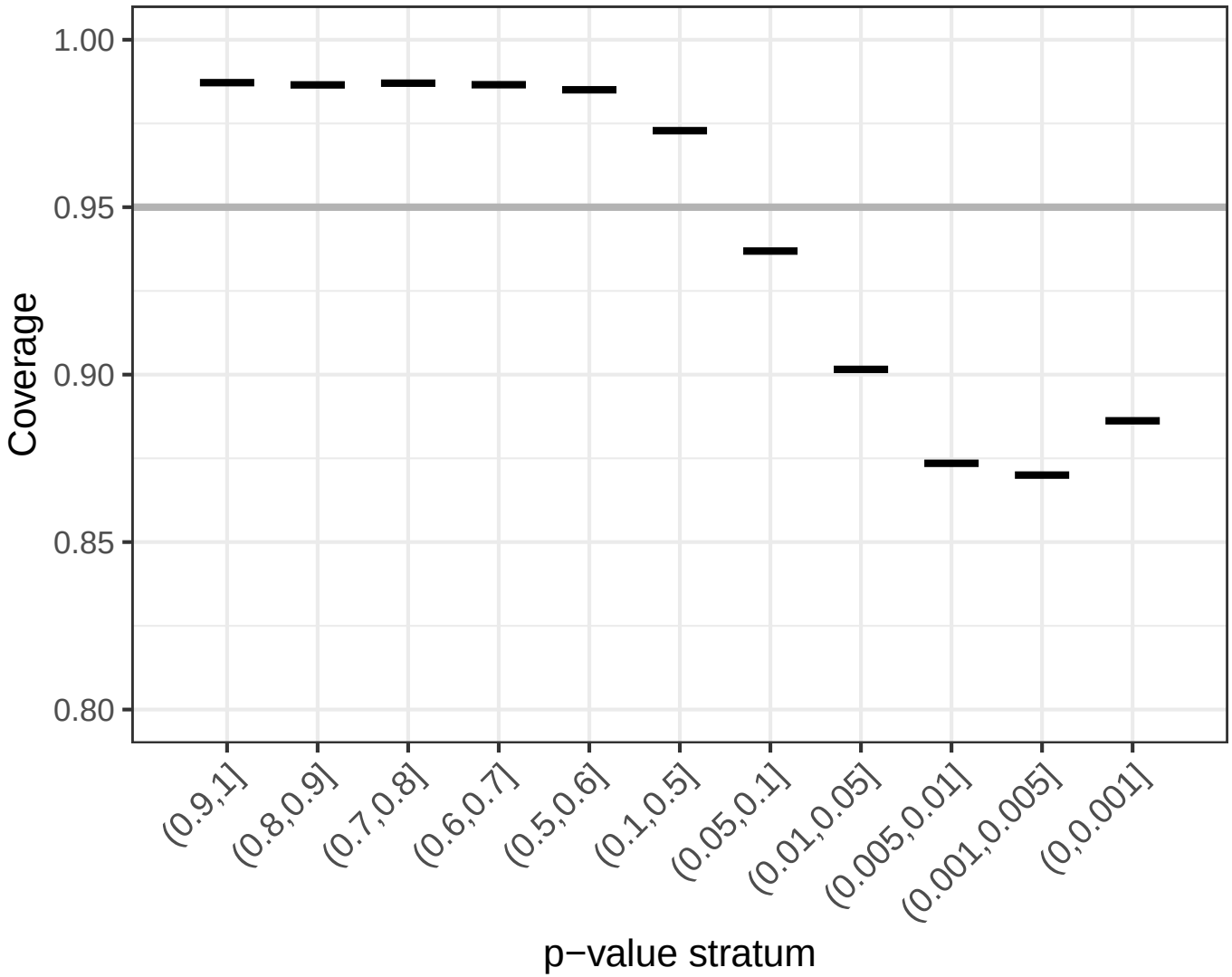
Note that

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

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


1.00
0.95
0.90
0.85
0.80
Coverage
(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