**Supplement: Automated detection of over- and under-dispersion in baseline tables in randomised controlled trials**

Table of Contents

[Finding papers on *PubMed* 1](#_Toc105583189)

[Summary statistics for the PubMed central trials 2](#_Toc105583190)

[Top ten countries 2](#_Toc105583191)

[Top ten journals 3](#_Toc105583192)

[Distribution of the baseline table size from trials 3](#_Toc105583193)

[Distribution of the trial sample size 4](#_Toc105583194)

[Cumulative distribution function 4](#_Toc105583195)

[Example distributions for the simulation study results 5](#_Toc105583196)

[Distribution of p-values 6](#_Toc105583197)

[Distribution of t-statistics 6](#_Toc105583198)

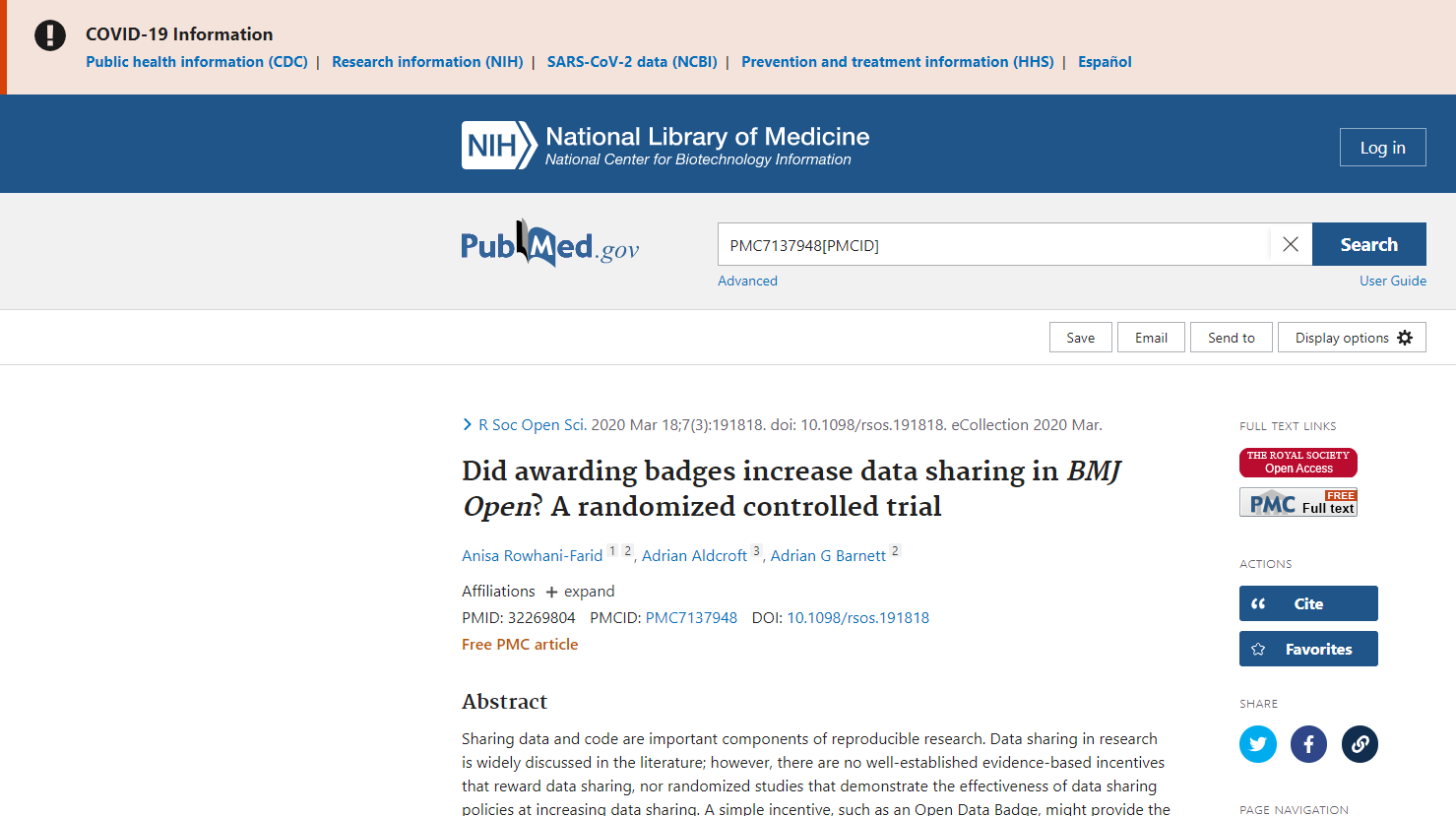
[Distribution of the precision parameter in the PubMed Central data 7](#_Toc105583199)

[Power to detect under-dispersion and baseline table size 8](#_Toc105583200)

[Reference 9](#_Toc105583201)

# Finding papers on *PubMed*

To find a paper on [*PubMed*](https://www.ncbi.nlm.nih.gov/pmc/) using the PubMed Central ID, go to the search and type the ID followed by [PMCID], for example “PMC7137948[PMCID]”. Example screen shot below.



# Summary statistics for the PubMed central trials

## Top ten countries

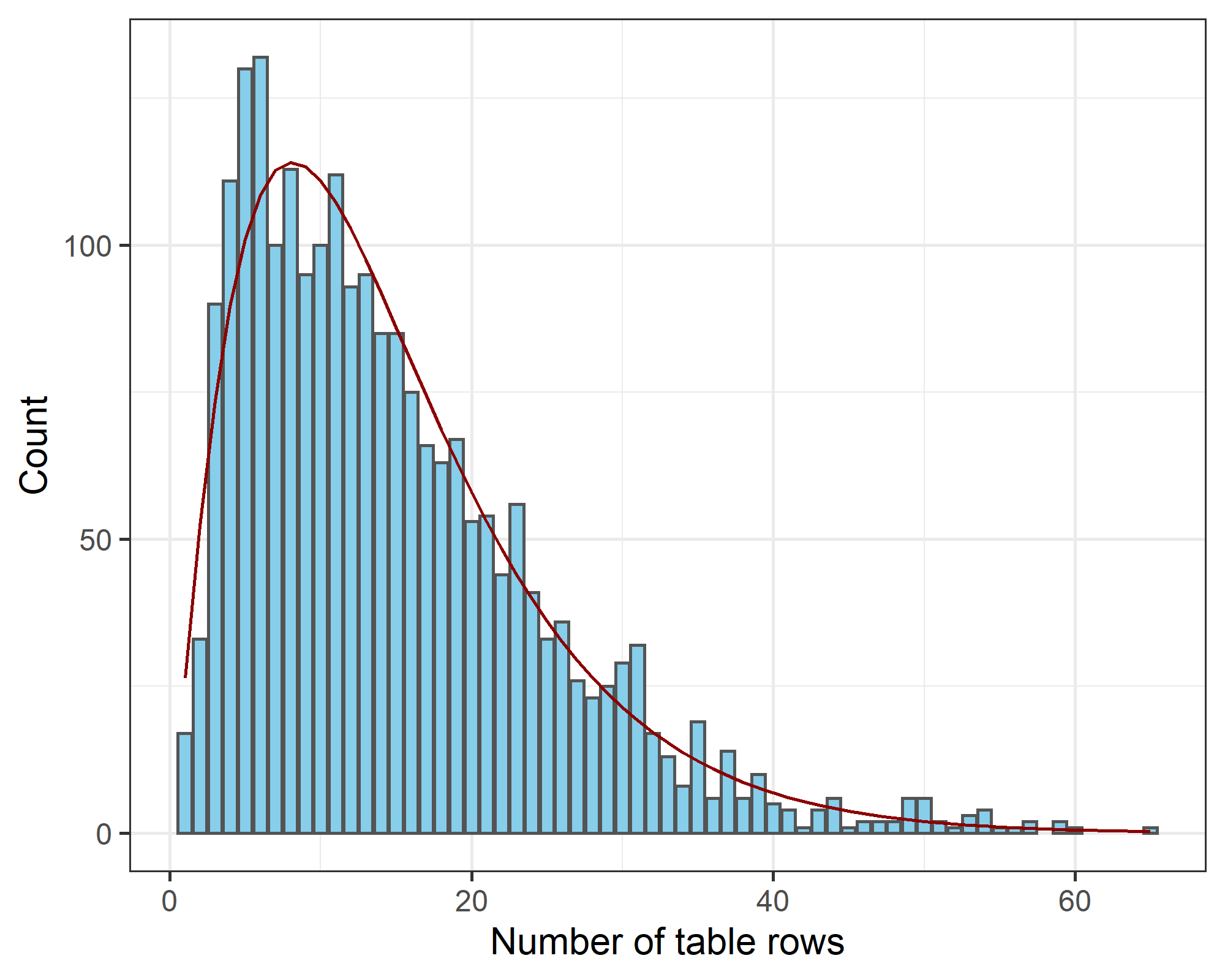
| **affiliation** | **n** |
| --- | --- |
| USA | 329 |
| China | 287 |
| Iran | 240 |
| Japan | 140 |
| South Korea | 137 |
| India | 99 |
| UK | 95 |
| Germany | 71 |
| Republic of Korea | 70 |
| Australia | 67 |

Country was based on the affiliation of the first author.

## Top ten journals

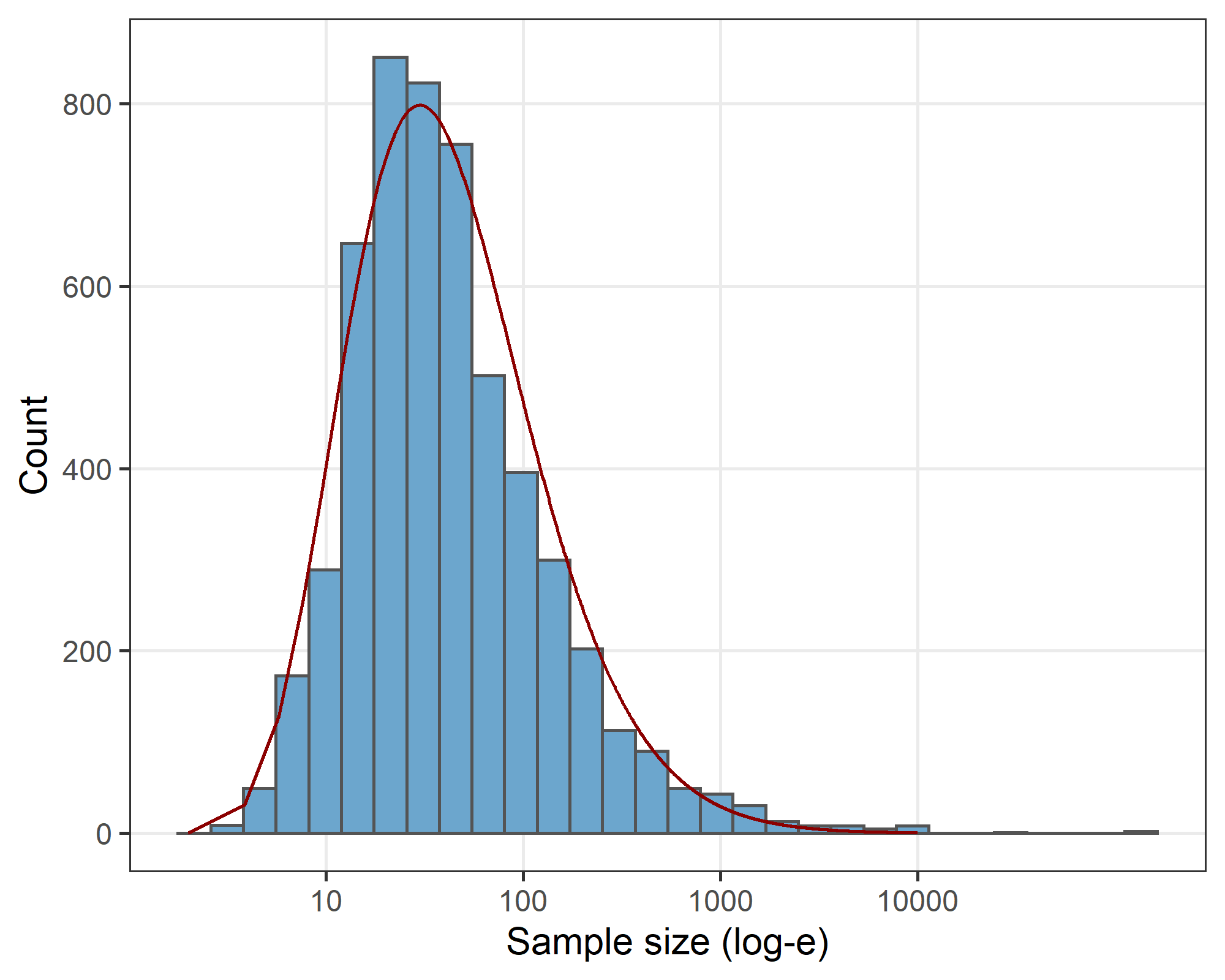
| **journal** | **n** |
| --- | --- |
| Evidence-based Complementary and Alternative Medicine : eCAM | 85 |
| Journal of Physical Therapy Science | 84 |
| Journal of Clinical Medicine | 44 |
| PLoS ONE | 32 |
| Experimental and Therapeutic Medicine | 26 |
| Anesthesiology and Pain Medicine | 25 |
| Journal of Pain Research | 25 |
| Nutrients | 25 |
| Scientific Reports | 24 |
| Frontiers in Psychology | 23 |

## Distribution of the baseline table size from trials



The mean number of rows per table was 15. The gamma distribution (red line) fitted to the row distribution has a shape parameter of 2.2 and rate of 0.15.

## Distribution of the trial sample size

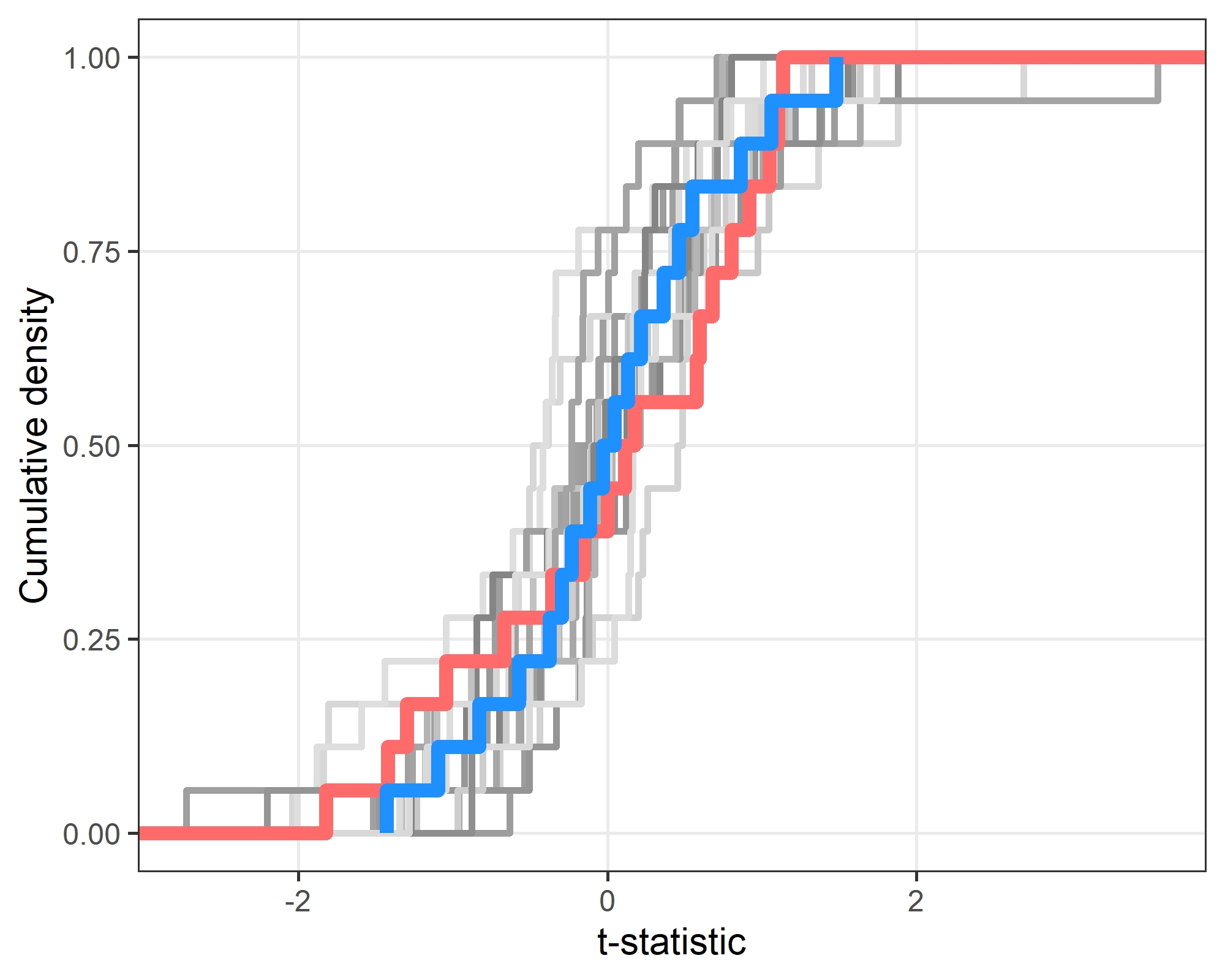


The mean trial sample size was 170. The gamma distribution (red line) fitted to the log-transformed (base e) sample size has a shape parameter of 11.2 and rate of 3.

For the simulation study we used group sample sizes and baseline table sizes that were based on the large sample size of automated extracted trials. The two figures above show the observed distributions and fitted gamma distributions. For the simulation study, we randomly generated sample sizes and table sizes from these gamma distributions.

# Cumulative distribution function

We can use the cumulative distribution function (CDF) to give a graphical comparison of the distribution of t-statistics from a trial’s baseline table. We plot the CDF of the t-statistics from the trial and compare it with data simulated under the null hypothesis of randomised groups. If the groups in the trial were correctly randomised, then the trial CDF should be similar to the CDFs from the simulated data. If the trial CDF is outside the simulated CDFs then this will indicate if the trial summary statistics are under- or over-dispersed.



The plot shows the trial CDF and the CDFs from 20 simulated data sets (grey lines) and the median of these simulations (blue line). The trial line (red) is similar to the simulated data. The trial is a randomised trial of high-flow oxygen with Barnett as the study statistician [1]. From the Bayesian model, the probability that the baseline table summary statistics are under- or over-dispersed are 0.112.

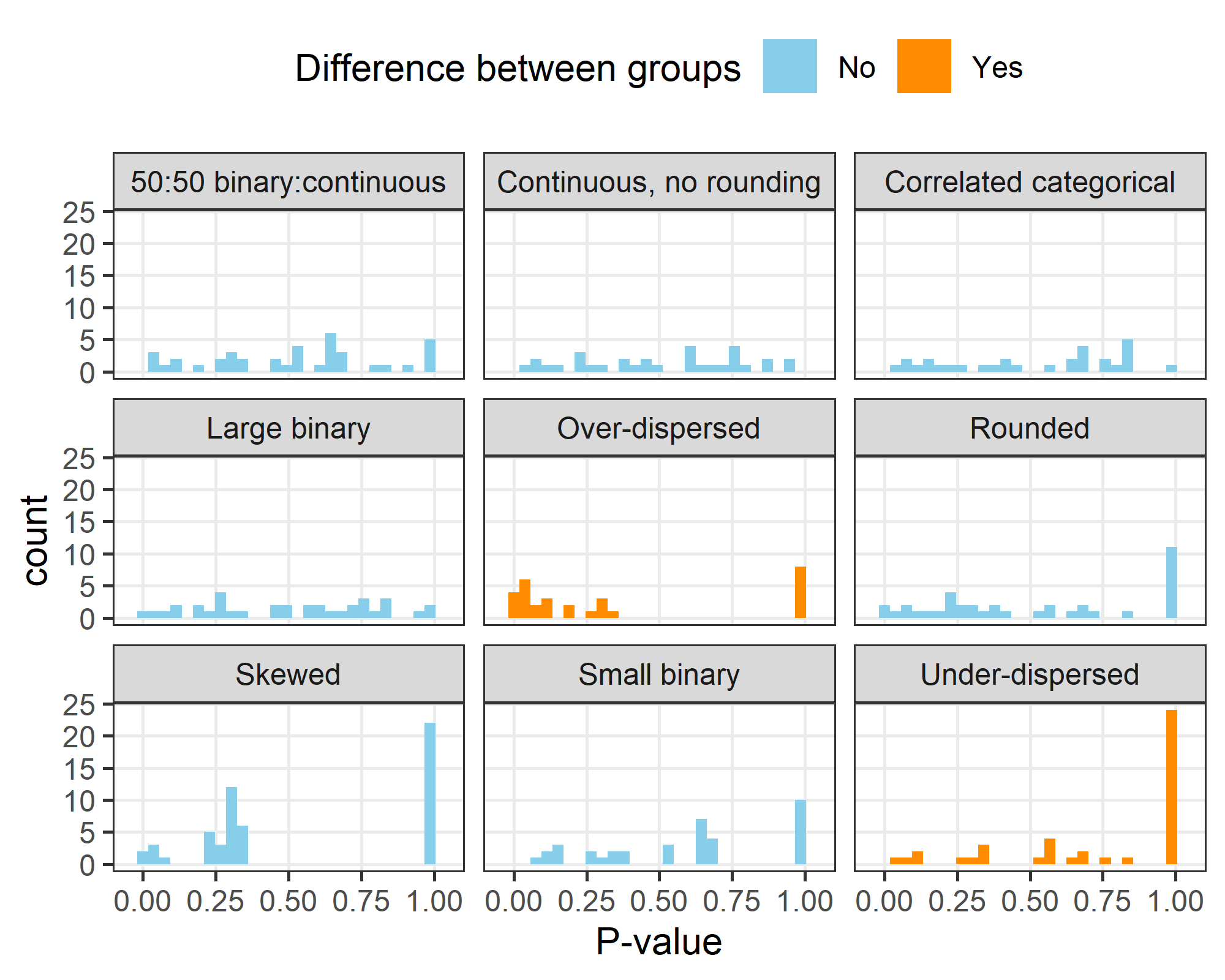
This plot and Bayesian model analysis is available for individual trials via our shiny app: <https://aushsi.shinyapps.io/baseline/>.

# Example distributions for the simulation study results

Here we show examples from the simulation study that examines the distribution of p-values and t-statistics for a range of simulated scenarios.

For each of the seven simulation types, we selected a simulated result based on the mean and standard deviation, selecting the simulation that was closest to the overall mean and standard deviation using the Euclidean distance. This was an attempt to show a typical simulation. We limited the comparison to simulations with at least 30 rows in the baseline table.

## Distribution of p-values

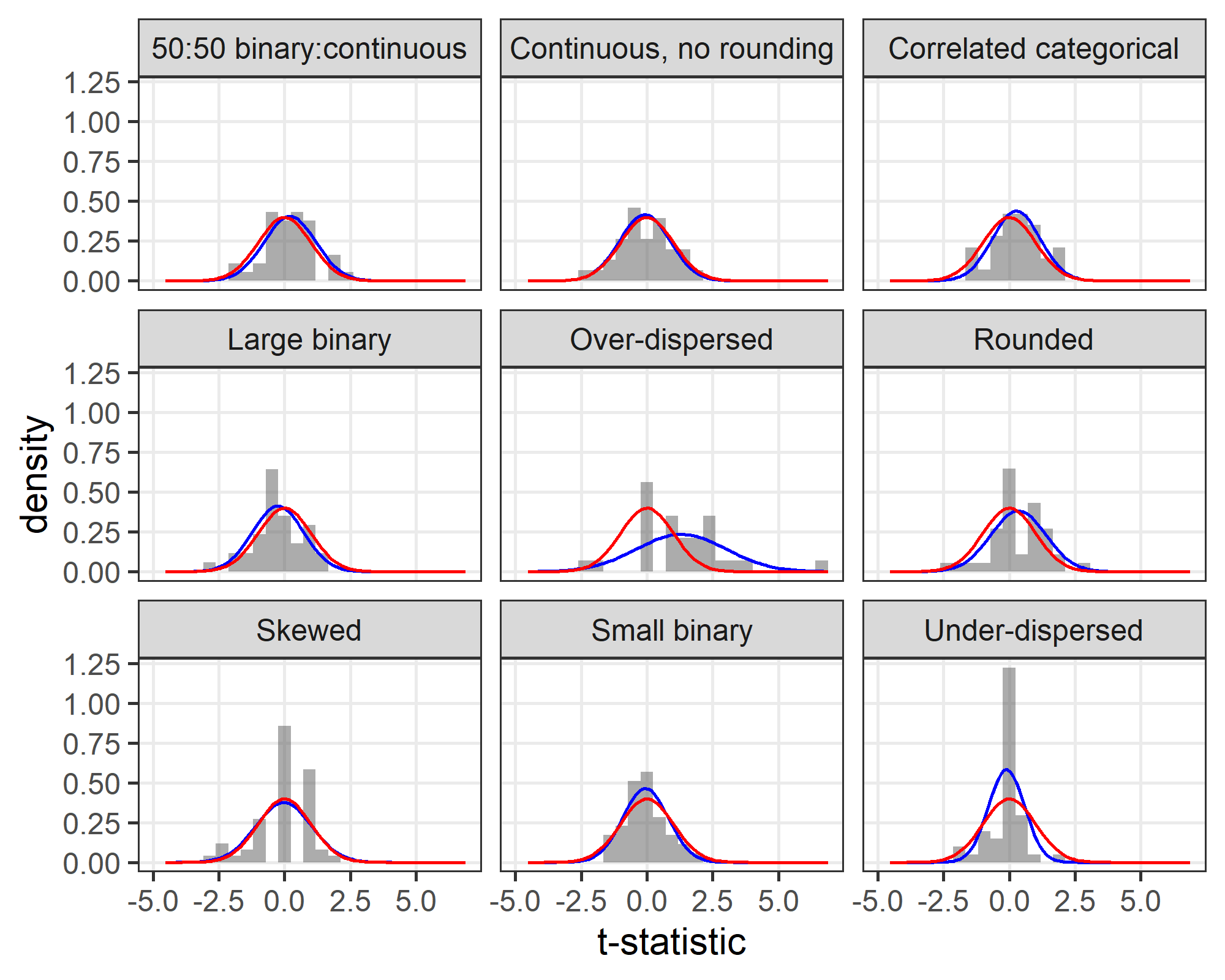


The example shows how skewed data can create a distribution of p-values that are not uniform, despite there being no difference between the randomised groups. There are also some spikes in the small binary simulation creating a non-uniform distribution. The rounded data has a spike at 1 as the rounding blurs differences between groups, making them appear similar.

The p-value distribution is clearly not uniform for the over-dispersed data with many small p-values, and shows a slight bump towards larger p-values for the under-dispersed data.

## Distribution of t-statistics

The plot below shows the t-distributions for the same seven examples from the p-value plot above. The red line is a standard normal distribution and the blue line is a normal distribution fitted to the observed density.

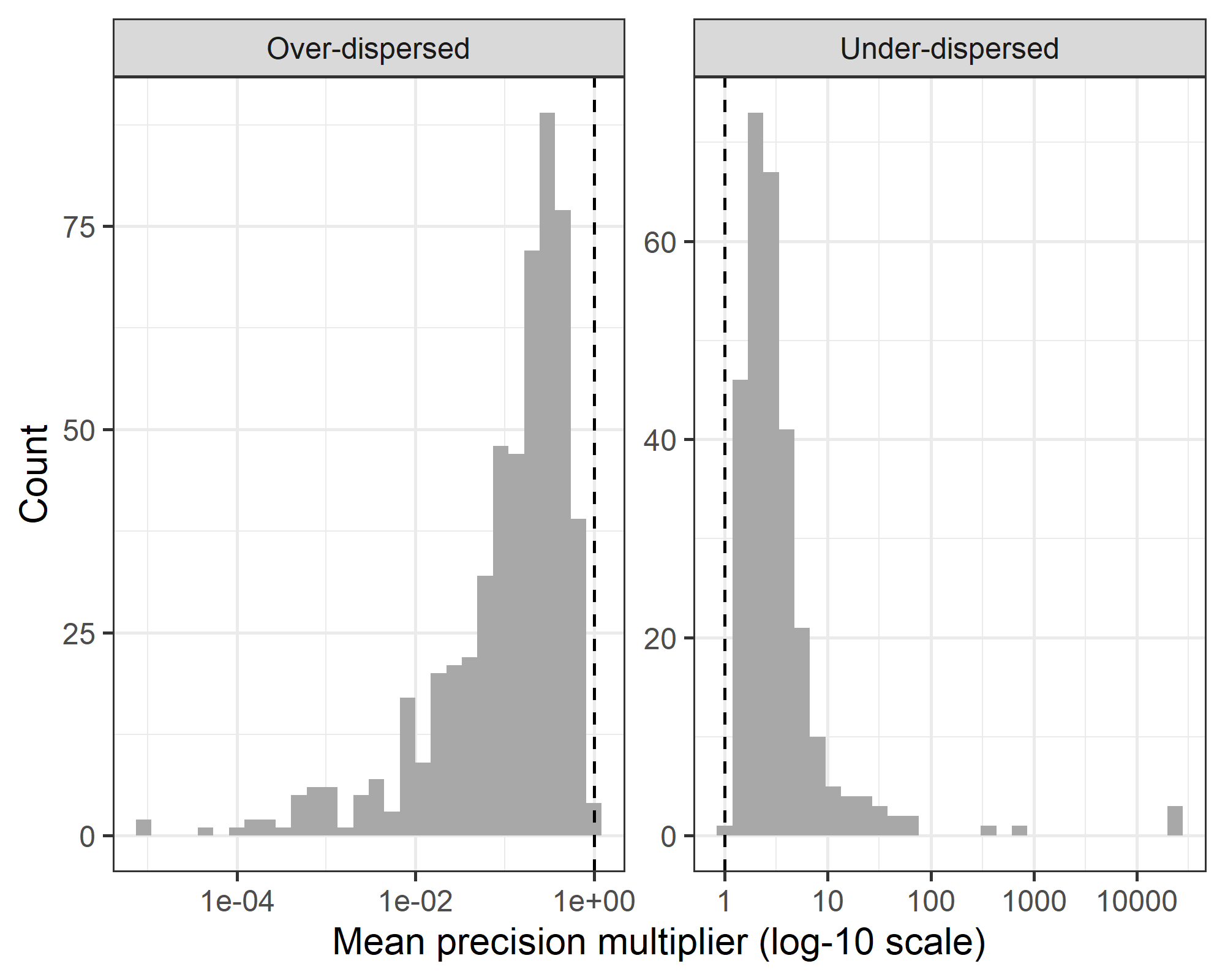


The skewed data has clustered t-statistics rather than a smooth distribution, however the fitted normal distribution is close to the expected standard normal distribution (zero mean and standard deviation of 1). Hence, this simulation with skewed data is not flagged by the Bayesian model.

The over-dispersed distribution is clearly different to a standard normal. The under-dispersed distribution is slightly narrower than the standard normal.

# Distribution of the precision parameter in the PubMed Central data

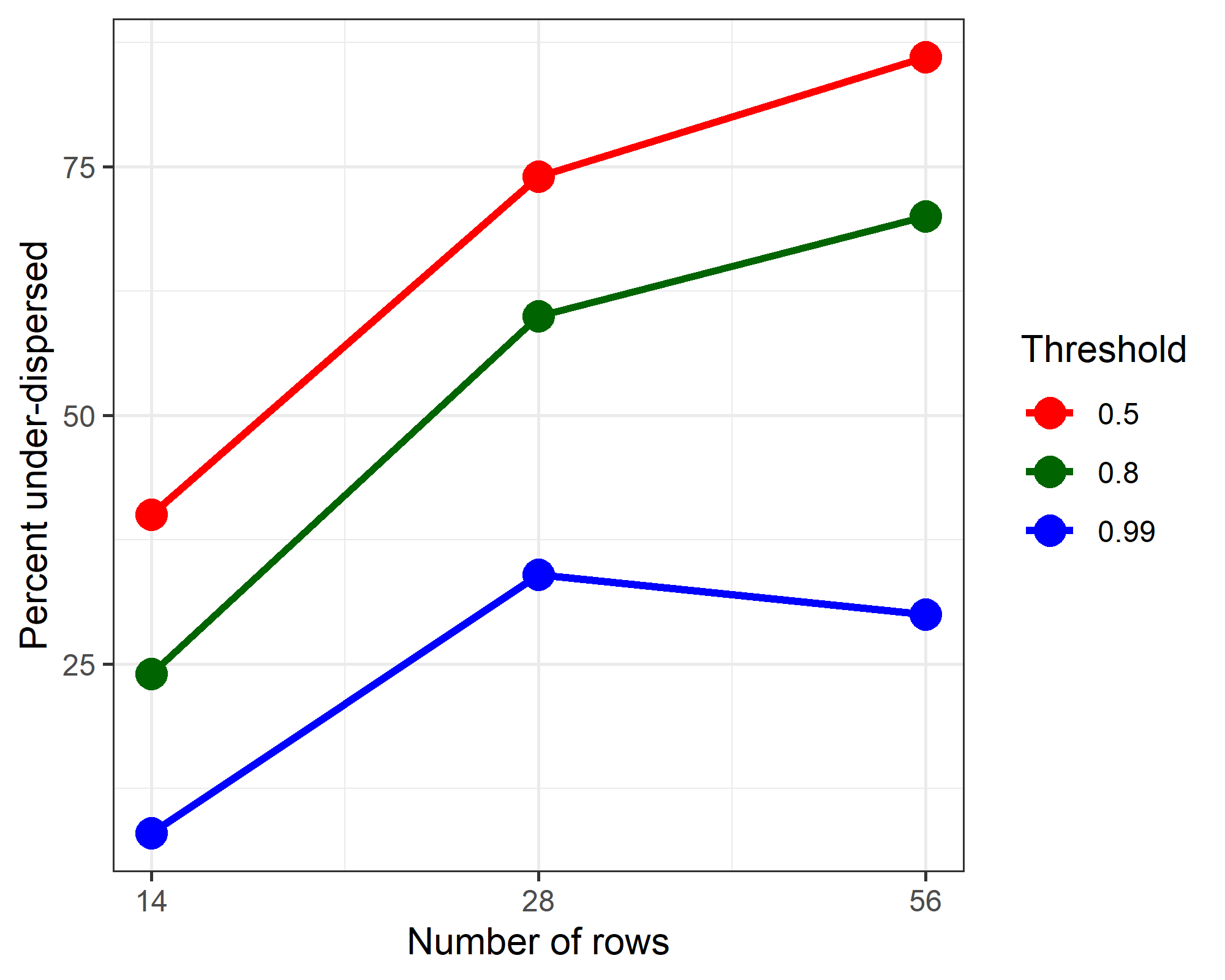
The plot below shows the distribution of the precision multiplier () for the trials extracted from PubMed Central. We plot the estimated mean from the Bayesian model. The plot is for trials where the median probability for the precision was non-zero. The panels are split by those under-dispersed (precision less than 1) and over-dispersed (precision greater than 1).



The precision axis in the plot was truncated at 1,000,000 as there was 1 outlier that made it difficult to visualise the distribution.

# Power to detect under-dispersion and baseline table size

The graph below shows a result from an additional simulation where we increased the number of rows in the baseline table and examined the power of the Bayesian model to detect under-dispersion.



# Reference

1 Corley A, Bull T, Spooner AJ, *et al.* Direct extubation onto high-flow nasal cannulae post-cardiac surgery versus standard treatment in patients with a BMI 30: A randomised controlled trial. *Intensive Care Medicine* 2015;**41**:887–94. doi:[10.1007/s00134-015-3765-6](https://doi.org/10.1007/s00134-015-3765-6)