Intro

This is a quick simulation exploring the "error" associated with conference paper rated using the 1..5 scale with three reviewers.

We assume raters are independent and independently sampled, with baseline ratings varying around 3, and with the final ratings sampled using

$$r_{ij} \sim \mathcal{N}(\text{rev}_i + q_j, \sigma_r^2)$$

where rev_i is the baseline rating for reviewer i, q_j is the baseline rating for paper j, and σ_r^2 is the inherent variance in ratings.

We assume

$$\text{rev}_i \sim \mathcal{N}(3, \sigma_{\text{rev}}^2), \sigma_{\text{rev}} = 0.7$$

$$q_j \sim \mathcal{N}(0, \sigma_q^2), \sigma_q = 1.2$$

Overall modelling assumptions

Below, we assume (correctly) that a measurement will be "almost always" within $2.5\sigma_r$ of the "true" mean. The true mean for a particular reviewer is *that* reviewers true opionion of the paper. We also assume that different reviewers have slightly different standards. Again, we assume that a reviewer will almost always be within $2.5\sigma_{\text{rev}}$ of the average reviewer.

Conclusions

In a large group of papers, it would not be unlikely to see deviations from the "true" rating of 0.5/5-1.0/5 points, due to variation between reviewers, and due to the inherent uncertainty of ratings.

Conclusions for the PC

Whether to go against the reviewer ratings is a question of community norms and expectations more than a matter of statistical modelling. However, the modelling here indicates that it is absolutely possible for a paper rated 3.1/5 to have an underlying quality that's better than 4.0/5 and vice versa.

Of course, if you are considering this possibility, you probably have evidence for that happenning in the text

Simulation function

In this simulation, we take in a vector `q` of the "underlying quality" of each of the papers, as well as the inherent variance of reviews and the between-reviewer variance. We then simulate the review process, and return the average rating for each paper.

```
simul  <- function(N.papers, N.reviews, sigma.r, sigma.rev, q){
    rev <- rnorm(N.papers, mean = 3, sd = sigma.rev)
    r <- matrix(NA, nrow = N.reviews, ncol = N.papers)

    # review assignment
    assign <- matrix(NA, nrow = N.reviews, ncol = N.papers)
    idx  <- rep(1:N.papers, each = N.reviews)
    idx <- sample(idx)
    for(i in 1:N.papers){
        assign[,i] <- idx[((i-1)*N.reviews + 1) : (i*N.reviews)]
```

```
    }

    # review scores
    for(i in 1:N.papers){
        for(j in 1:N.reviews){
            r[j,i] <- round(rnorm(1, rev[j] + q[i], sigma.r))

        }
    }

    # clip all values in r to be between 1 and 5
    r <- pmax(r, 1)
    r <- pmin(r, 5)

    # average scores
    apply(r, 2, mean)
}

compute.rating.rmse <- function(q, avg.r, sigma.r, sigma.rev, N.papers, N.reviews, N.reps){




    # make q the same shape as avg.r (i.e., N.reps x N.papers)
    # use rep to replicate the row q N.reps times
    # make q.rep a matrix with N.reps columns and N.papers rows

    q.clipped <- pmax(q, 1)
    q.clipped <- pmin(q.clipped, 5)

    q.rep <- matrix(rep(q, N.reps), nrow = N.papers, ncol = N.reps)

    # compute SSE
    SSE <- sum((avg.r - (3 + q.rep))^2)

    # compute RMSE
    sqrt(SSE / (N.papers * N.reps))
}
```

Experiment 1: conservative estimate of variance

In this experiment, we assume that the inherent variance in reviews is 0.2^2 (i.e., the scores will almost always be within 0.5 of the "true" score for the particular reviewer), and the variance in reviewer standards is 0.2^2 (i.e., reviewers will almost always have an opinion within 0.5 of the average reviewer). We assume that the variance in paper quality is 0.8^2 (i.e., the average paper is a 3, and almost all papers are in the range $1 \ldots 5$).

```
N.papers  <- 200
N.reviews <- 3
N.reps    <- 1000
sigma.r   <- 0.2
sigma.q   <- 0.8
sigma.rev <- 0.2
```

```
sigma.r
```

```
## [1] 0.2
```

```
sigma.q
```

```
## [1] 0.8
```

```
sigma.rev
```

```
## [1] 0.2
```

```
q <- rnorm(N.papers, 0, sigma.q)
avg.r <- replicate(N.reps, simul(N.papers, N.reviews, sigma.r, sigma.rev, q))
compute.rating.rmse(q, avg.r, sigma.r, sigma.rev, N.papers, N.reviews, N.reps)
```

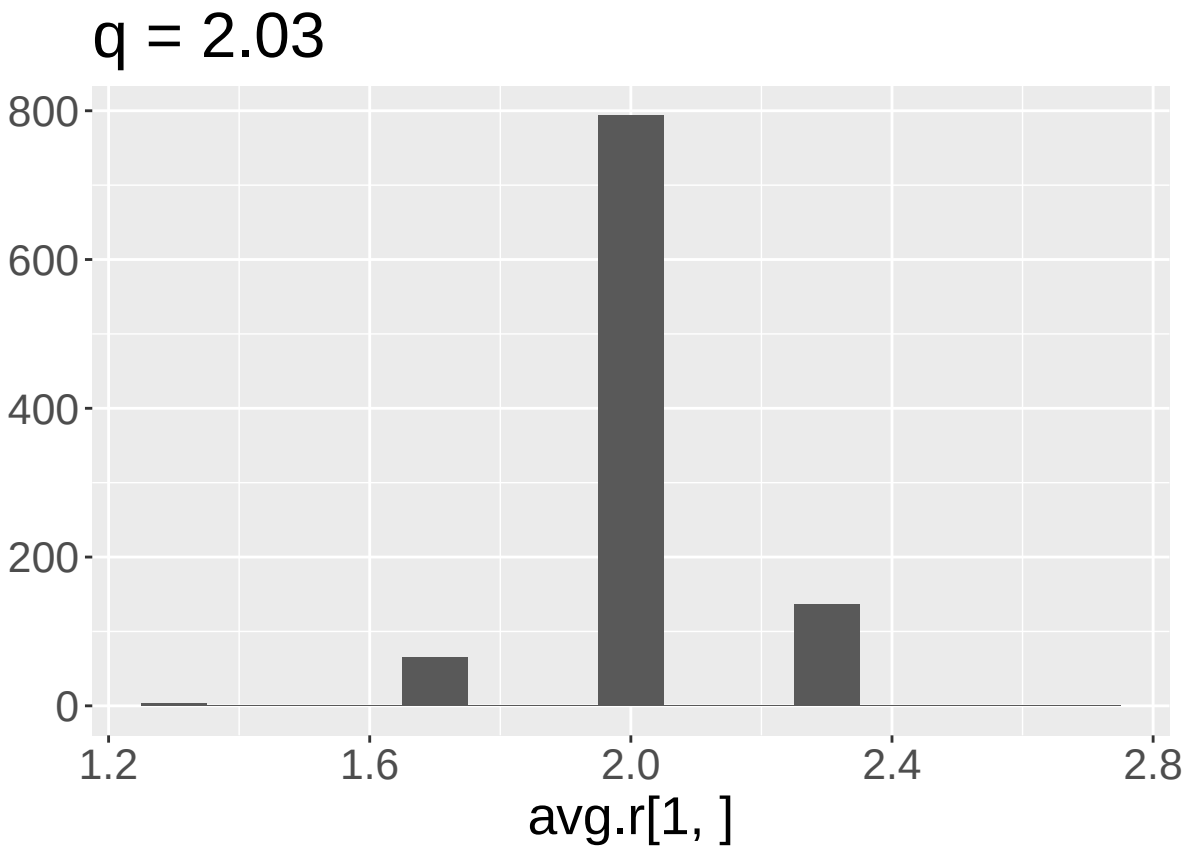
```
## [1] 0.2389256
```

This suggests that under the model assumptions, you'd expect the average rating to be within approximately 0.5 (corresponding to 2*SD) of the "true" rating.

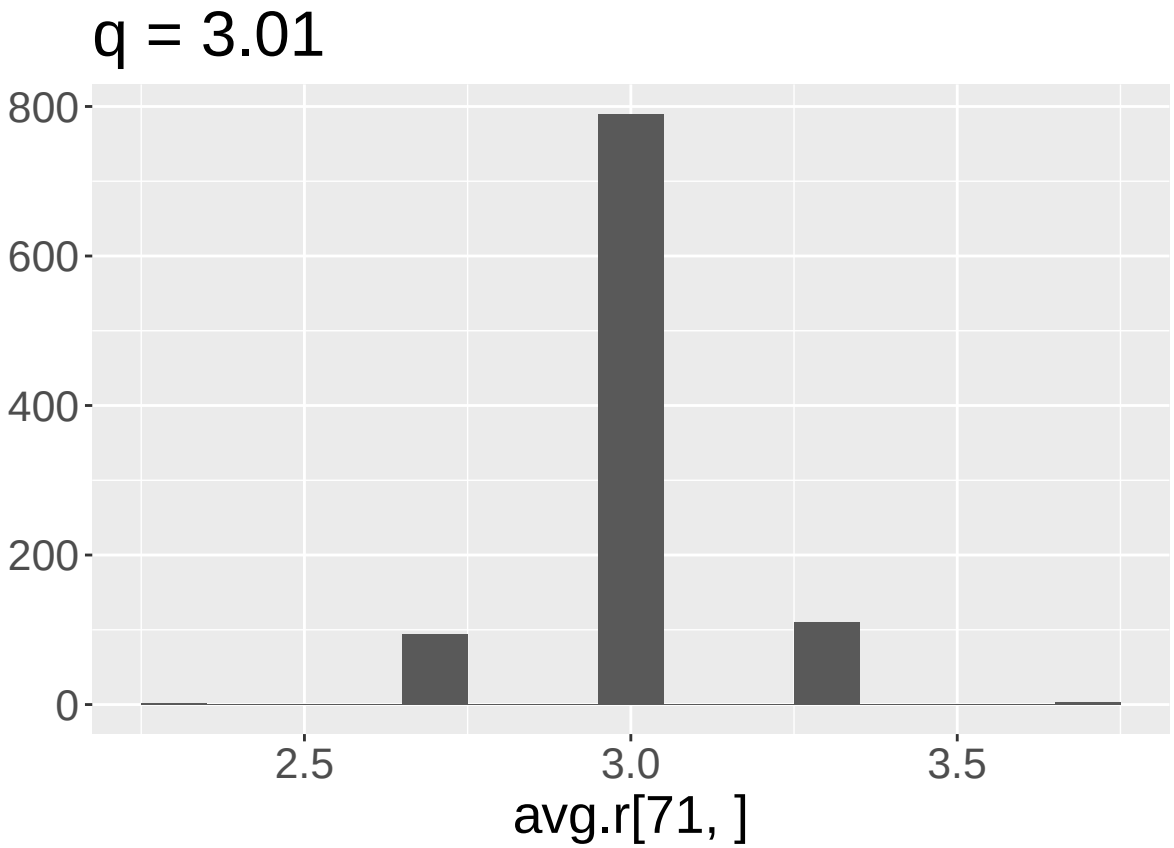
Here are examples of the distribution of the simulated average ratings:

```
qplot(avg.r[1, ], geom = "histogram", binwidth = 0.1) + ggtitle(sprintf("q = %.2f", 3+q[1])) + theme(te
```

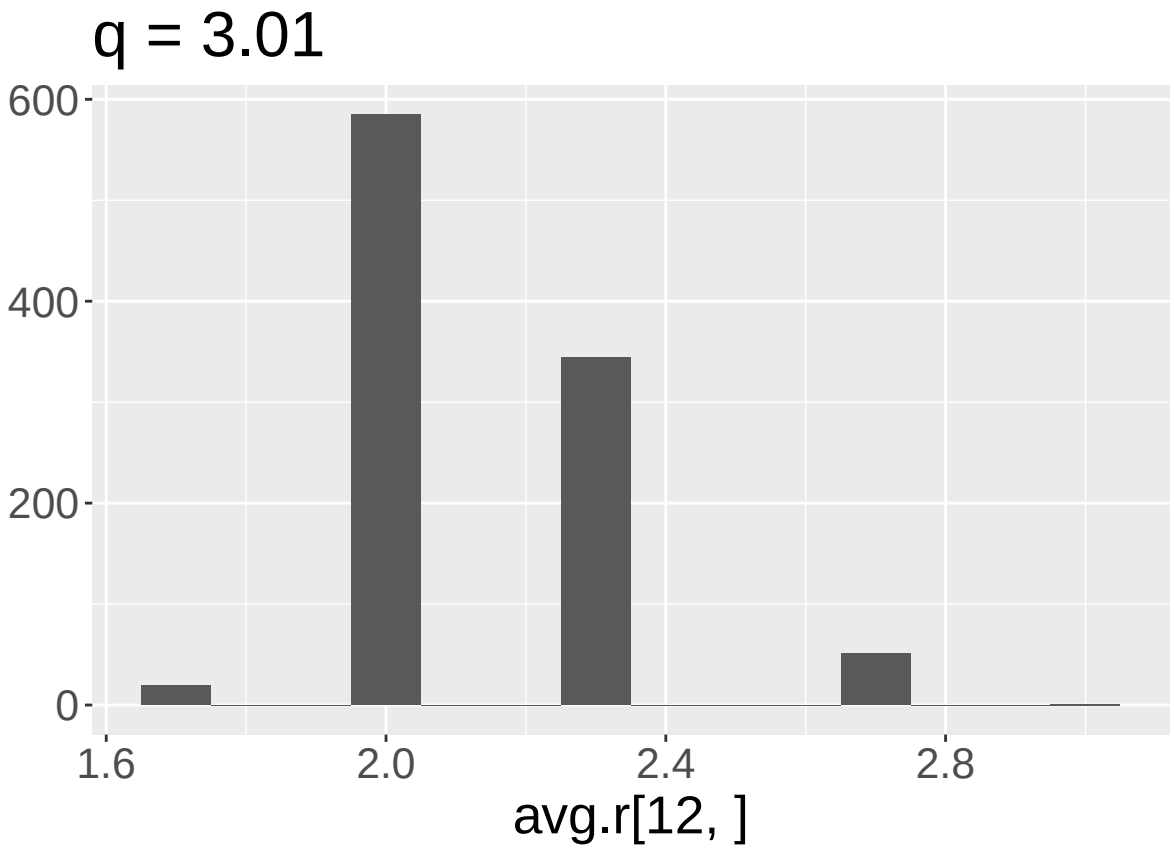
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
qplot(avg.r[71, ], geom = "histogram", binwidth = 0.1) + ggtitle(sprintf("q = %.2f", 3+q[71])) + theme(
```




```
qplot(avg.r[12, ], geom = "histogram", binwidth = 0.1) + ggtitle(sprintf("q = %.2f", 3+q[71])) + theme(
```



Experiment 2: less conservative estimate of variance

Here, we'll increase the standard deviations by a factor of 1.5. This means that reviewers can have an underlying difference of opinion of up to 0.75/5 points from the average reviewer, and the reviewer ratings can differ by up to 0.75/5 points from the average reviewer's rating.

```
N.papers  <- 200
N.reviews <- 3
N.reps    <- 1000
sigma.r   <- 0.4
sigma.q   <- 0.8
sigma.rev <- 0.4


sigma.r
```

```
## [1] 0.4
```

```
sigma.q
```

```
## [1] 0.8
```

```
sigma.rev
```

```
## [1] 0.4
```

```
q <- rnorm(N.papers, 0, sigma.q)
avg.r <- replicate(N.reps, simul(N.papers, N.reviews, sigma.r, sigma.rev, q))
compute.rating.rmse(q, avg.r, sigma.r, sigma.rev, N.papers, N.reviews, N.reps)
```

```
## [1] 0.361028
```

Under less conservative assumptions, the $2.5RMSE$ is 0.9, meaning a rating could be "off" by 0.9/5 points.