Analysis for Odds and Insights: Do Uncertainty Visualisations Improve Qualityof Decisions in Visual Analysis

Abhraneel Sarma

2020-07-04 13:27:07

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(message = FALSE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rstan)
```

```
## Loading required package: StanHeaders

## rstan (Version 2.19.3, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

##
## Attaching package: 'rstan'

## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
library(brms)
```

```
## Loading required package: Rcpp
```

```
## Loading 'brms' package (version 2.13.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').
```

```
##
## Attaching package: 'brms'
```

```
## The following object is masked from 'package:rstan':
##
##     loo
```

```
## The following object is masked from 'package:stats':
##
##     ar
```

```r
library(modelr)
library(tidybayes)

options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
```

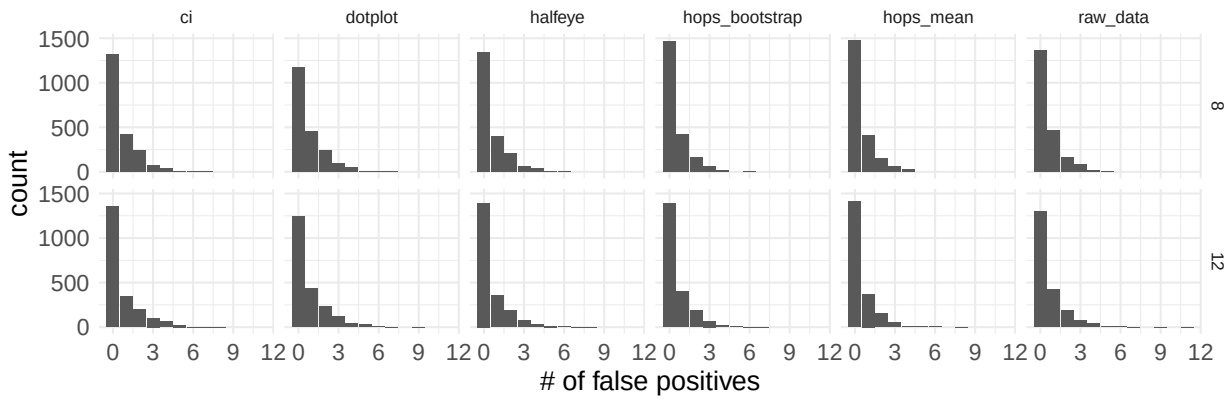
Overview

- Recruiting platform: Prolific
  - study code 29ABE0A0
  - base pay: $7
  - estimated time: 40 min
- Study website: https://mucollective.github.io/vis-insights-study/
- Database at AWS
- Exit survey at Qualtrics

Exploratory models

Number of false positives in each between and within subjects condition

Tally the false positives by each condition. The graphs below show the distribution of the number of false positives in a single trial. We can see that most people are making 2 or fewer false positives in each trial, however we do not see much differences based on the number of regions shown to participants.

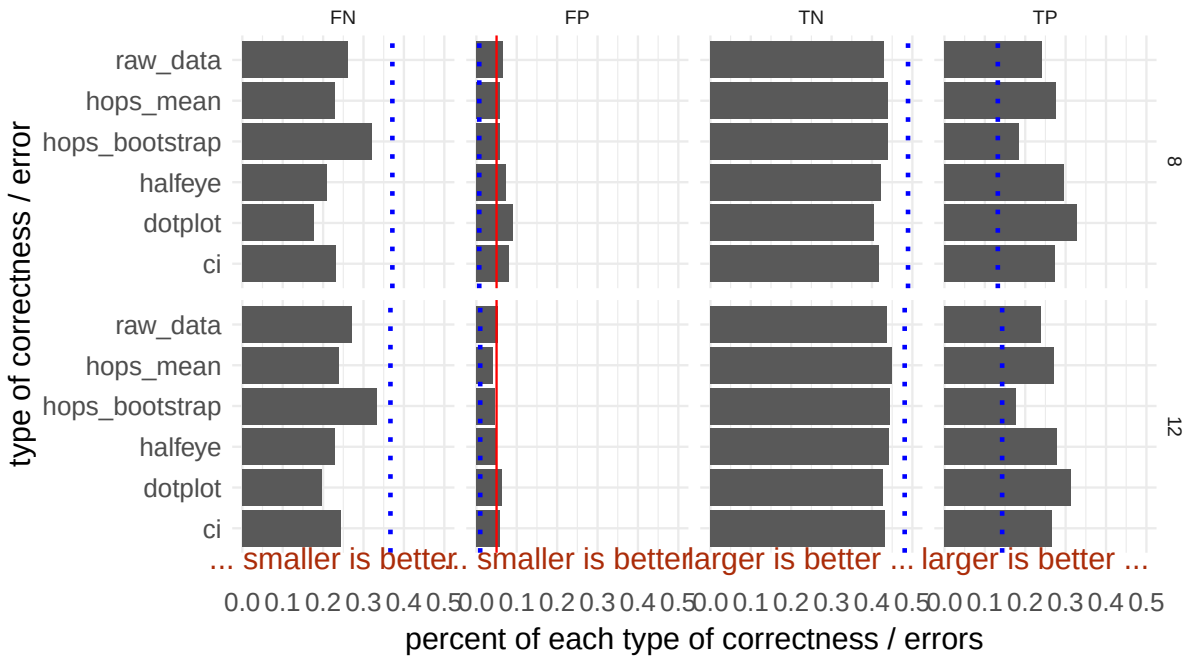


Correctness of responses in each between and within subject condition

```
## # A tibble: 12 x 7
##    condition      nregions    tp    tn     fp    fn   sum
##    <chr>             <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1 ci                    8 0.273 0.417 0.0795 0.230  1.00
##  2 ci                   12 0.267 0.432 0.0588 0.243  1.00
##  3 dotplot               8 0.328 0.404 0.0913 0.176  1
##  4 dotplot              12 0.313 0.427 0.0641 0.196  1
##  5 halfeye               8 0.295 0.423 0.0736 0.209  1
##  6 halfeye              12 0.279 0.442 0.0481 0.230  1
##  7 hops_bootstrap        8 0.184 0.439 0.0580 0.319  1
##  8 hops_bootstrap       12 0.177 0.445 0.0457 0.332  1
##  9 hops_mean             8 0.276 0.439 0.0571 0.228  1
## 10 hops_mean            12 0.272 0.450 0.0407 0.238  1
## 11 raw_data              8 0.242 0.430 0.0662 0.262  1
## 12 raw_data             12 0.238 0.437 0.0536 0.272  1
```

Probability of TP/TN/FP/FN in each condition

In the following graph we show the mean values of TP, TN, FP and FN in each uncertainty visualization condition, and separated by the number of graphs that we show each participant.



Modeling

The research questions in our study are:

- RQ1: do users implicitly perform some form of multiple comparisons correction?
- RQ2: do different types of uncertainty representations help users perform multiple comparisons correction by reducing the false discovery rate (FP / (FP + TP)).

The goal of our modeling is to estimate the probability of a TP / TN / FP / FN for a given (or average trial) along with some uncertainty. Based on the results from our model, we attempt to answer our research questions.

Multiple comparisons correction We define the model and create the appropriate column in the data structure (`y`) for predicting multinomial outcome variables (`brms` requires the outcome variable to be a n × k matrix where k is the number of categories, and n is the number of responses; here # of trials × # of participants).

Model fit

Before we try to visualise the model predictions, we need to extract posterior samples from the fitted model object:

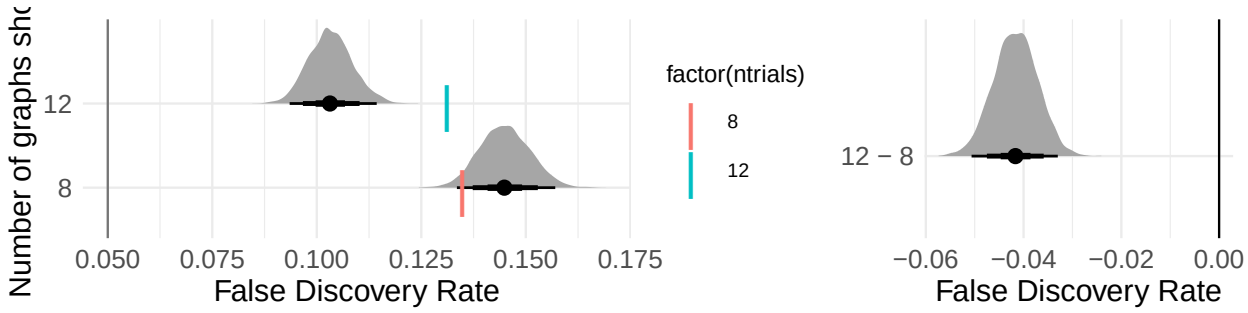
RQ 1: do users implicitly perform some form of multiple comparisons correction?

If users are not performing any form of multiple comparisons correction, then intuitively, on average, a participant in our study will have more number of False positives when presented with 12 graphs as opposed

to 8 graphs (the two within person conditions). More directly, we can compare the False Discovery rate when $nregions = 8$ vs when $nregions = 12$. If the FDR is constant or less for $nregions = 12$ compared to $nregions = 8$, then it implies that participants are performing some form of multiple comparisons correction.

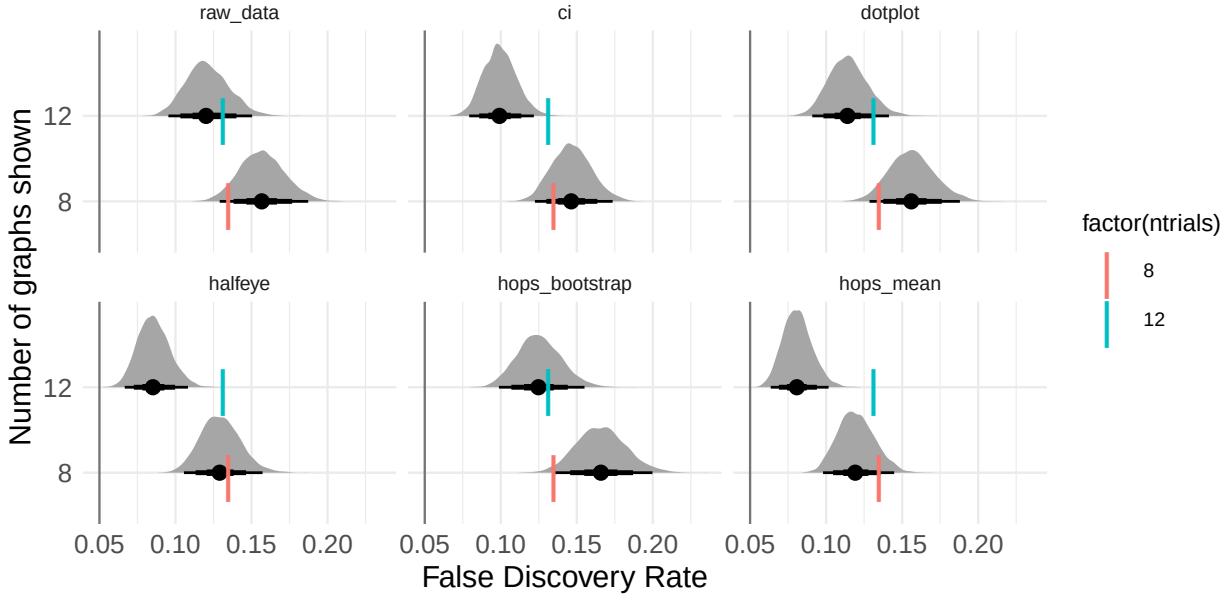
First, we need to calculate the False Positive rate without any multiple comparisons correction:

In the figure below, we see that, on average, the FDR decreases for participants when presented with more graphs. This suggests that they are likely performing some form of implicit multiple comparisons correction. The FDR for $ngraphs = 8$ is 14.5 and for $ngraphs = 12$ is 0.1.



We see that the decrease in the False Discovery Rate is, on average, a magnitude of 4 percentage points, with a 95% credible interval of [0.033, 0.051]. This implies there is almost a 30% reduction in the FDR.

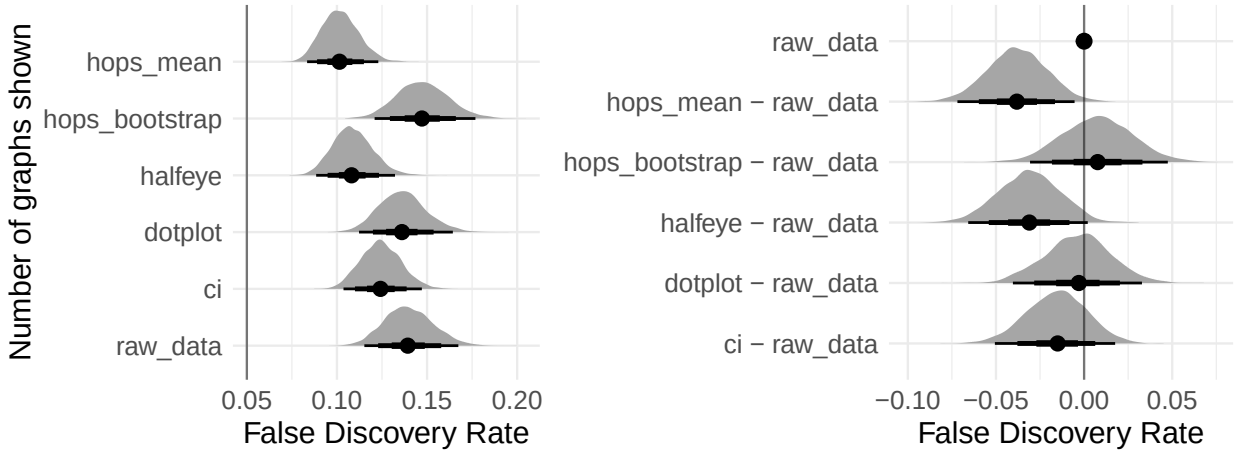
Next, since we have different visualization conditions, we inspect if this difference is persistent across all the conditions. From the following figure below, we see that the FDR decreases for participants consistently across all the uncertainty representations, suggesting that it is likely not an artifact of certain forms of visual representations. The magnitude of this decrease appears to be consistent across the different uncertainty representations as well.



Thus, our results suggest that users in our experimental set up implicitly perform some form of multiple comparisons correction. Because our experimental design incentivised participants against making False Discoveries proportionate to performing a NHST at a 95% confidence intervals, we cannot tell if participants would always behave this way. We believe that in the absence of incentives, participants may not control for False Positives in a similar manner, as suggested by the results of the study by Zgraggen et al.

RQ 2: do users implicitly perform some form of multiple comparisons correction?

To answer this question, we first look at the FDR across the different uncertainty representations, marginalised over the number of regions shown to participants. This gives us the aggregate effect over the two within-subjects conditions in the study.

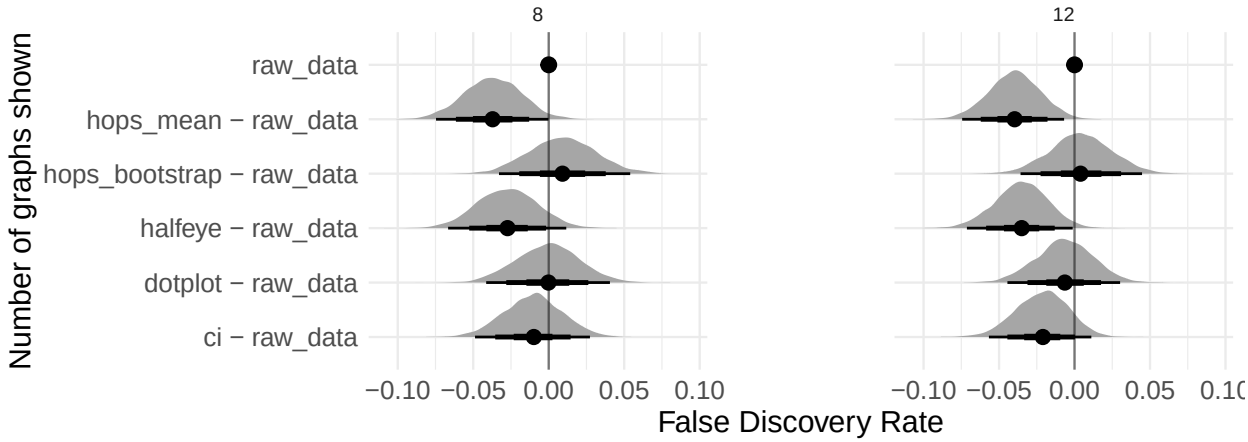


From the figure above, we can see that, on average, using uncertainty representatoions such as `Hypothetical Outcome Plot of the mean difference` and `Probability Density Function of the difference` reliably decreases the FDR, with an observed decrease of 0.04 and 0.03 percentage points respectively (95% CI: [-0.072, 0.005] and [0.065, 0.02] respectively). Some other commonly used uncertainty representations, such as Confidence Intervals appear to have a small, but unreliable effect towards decreasing (~ 1.5 percentage points, 95% CI: [-0.05, 0.02]) FDR. On the other hand, other certain other uncertainty representations such as `dotplots of the mean difference` and Hypothetical outcome plot of bootstrapped data samples' appear to have no improvement or even slightly worsen the FDR. (the exact estimates are present in the table below)

```
## # A tibble: 5 x 7
##   condition                 fp_rate  .lower   .upper .width .point .interval
##   <chr>                       <dbl>   <dbl>    <dbl>  <dbl> <chr>  <chr>
## 1 ci - raw_data            -0.0150  -0.0506  0.0176    0.95 median qi
## 2 dotplot - raw_data       -0.00300 -0.0404  0.0328    0.95 median qi
## 3 halfeye - raw_data       -0.0311  -0.0657  0.00208   0.95 median qi
## 4 hops_bootstrap - raw_data  0.00770 -0.0307  0.0475    0.95 median qi
## 5 hops_mean - raw_data     -0.0382  -0.0719 -0.00541   0.95 median qi
```

Are these differences consistent across the number of regions shown?

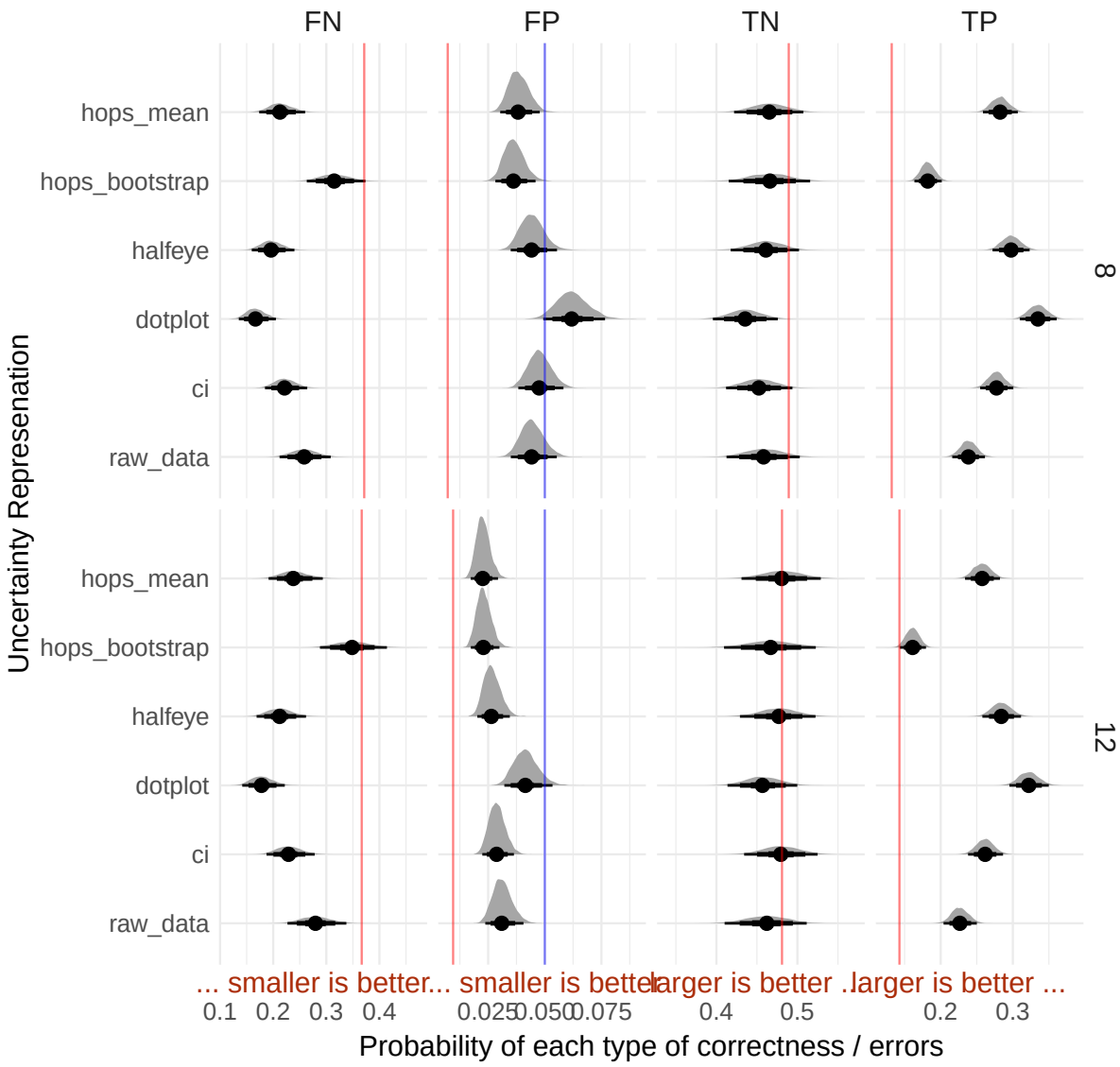
Based on the plots below, we find that the differences are fairly consistent across the within-subjects manipulation.



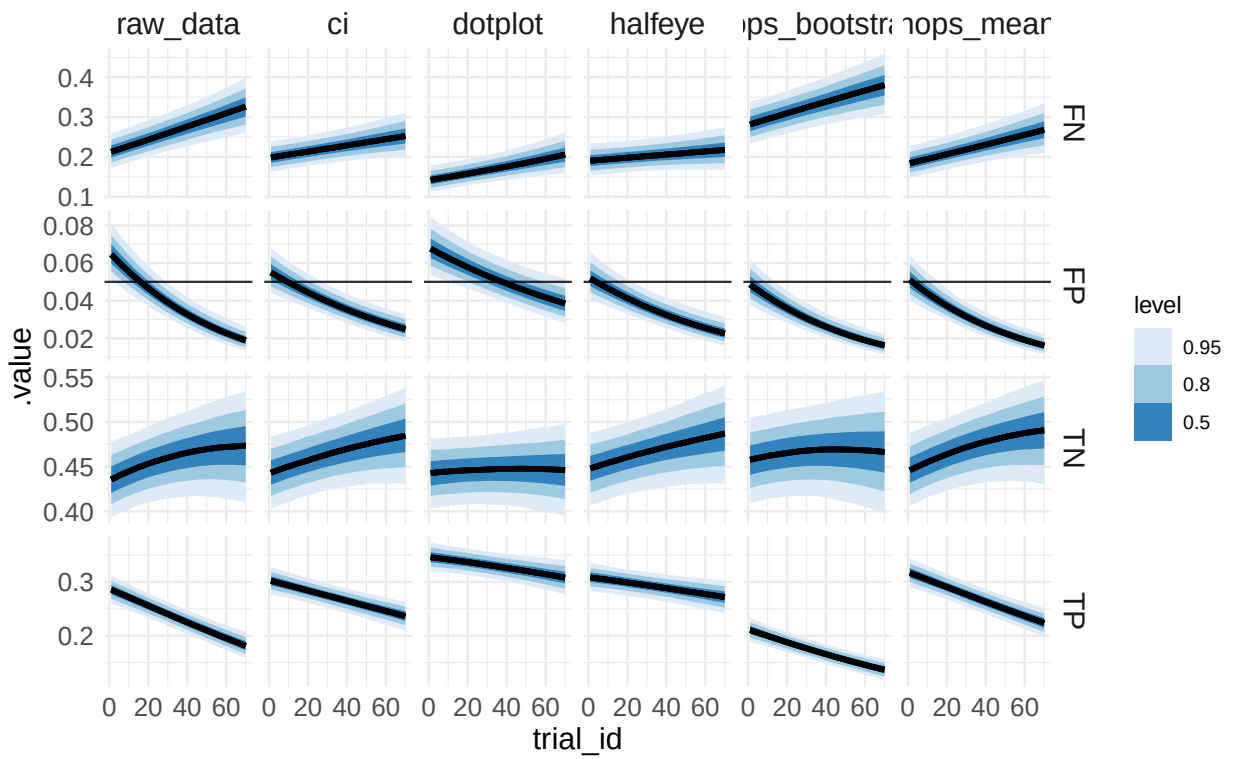
Exploratory analysis

We first take a look at the probability of TP/TN/FP/FN in each uncertainty representation condition, marginalised over *nregions*. Interestingly, we find that there isn't a lot of difference in the probability of an average user making a FP on an average trial, including some of the uncertainty representations with better FDR; except for the dotplot condition, all the other uncertainty representations appear comparable to the baseline (`raw data`) in terms of probability of False Positives in an average trial. The improvement in FDR usually arises from the analysts being able to correctly identify True Positives more accurately where we see large differences.The probability of False Negative also varies substantially across the different conditions, with `HOPS bootstrap` performing worse and all the other uncertainty representations performing better than the baseline condition (with dotplot performing the best).There is also variation in the probability of making a False Negative across the conditions, but little or no difference between the probability of making a True Negative.

Because it is difficult to compare the rates of TP / TN / FP / FN across conditions, we will use metrics developed in ML such as F-scores and Matthews Correlation Coefficient to obtain a composite score. Another way of comparing the different conditions would be to use the payout which serves as the incentive for the different participants.

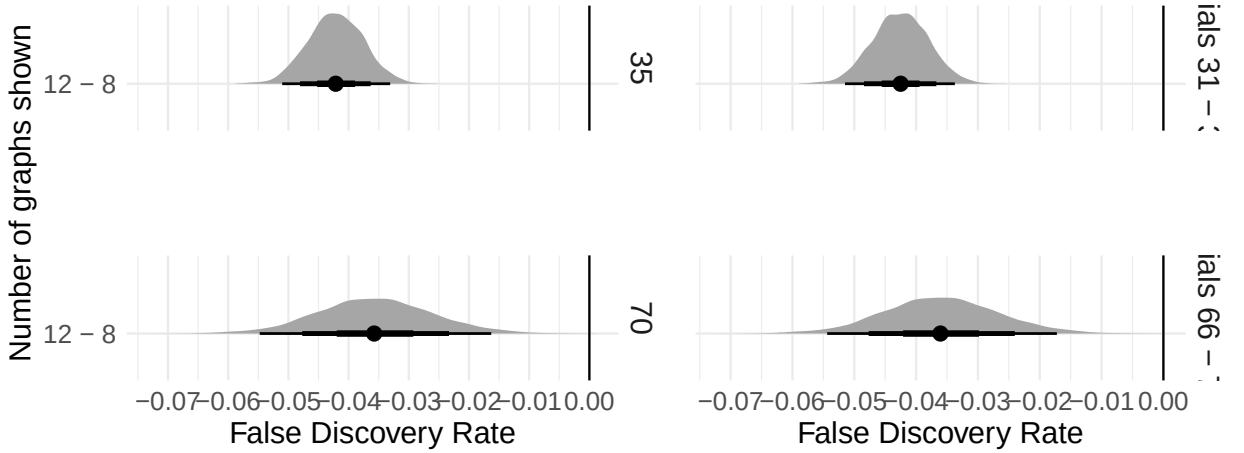


Learning effects? Before we compare the composite scores, we look at the potential learning effects in our primary research questions. In repeated measures experimental designs such as this, where we also provide participants feedback (in the first 5 trials in each block), we might expect to see some learning effect (or at least variation in the responses over the course of the trials). In the figure below, we plot the change in the probability of TP/TN/FP/FN in each condition.



This indicates that there might be an effect of learning that which might affect our interpretation of the results. Specifically, if the effect of learning is more significant than any of the effects of the different conditions that we manipulate, we might not expect to see any differences in the results when we compare the final trial. Another common practice is to compare the results of the final 5-10 trials (we take 5 here). We choose these points specifically because the within-subjects manipulation is performed in two blocks of 35 trials each giving us reasonable end points.

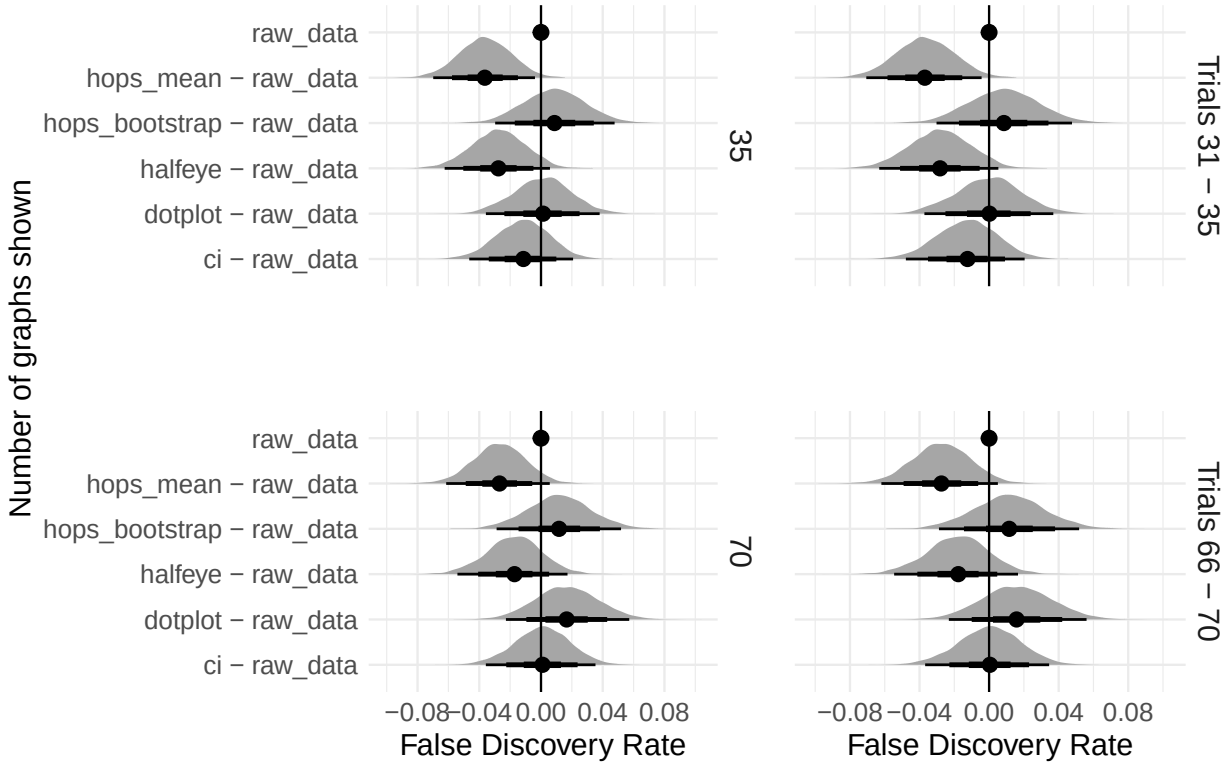
From the figure below, we see that the our participants still perform some form of implicit multiple comparisons correction, and this persists even after accounting for the potential effect of learning.



Next we test if there were any effects of learning on the differences between the uncertainty representations. In other words, does the effect of learning dominate over the effect of the uncertainty representation?

From the figure below, we can see that the effect persists for the unertainty representations which reduces FDR (`probability density plot` and `HOPs` of the mean difference), although the magnitude of the mean effect is smaller by around 1 percentage point. This indicates that certain forms of uncertainty representations

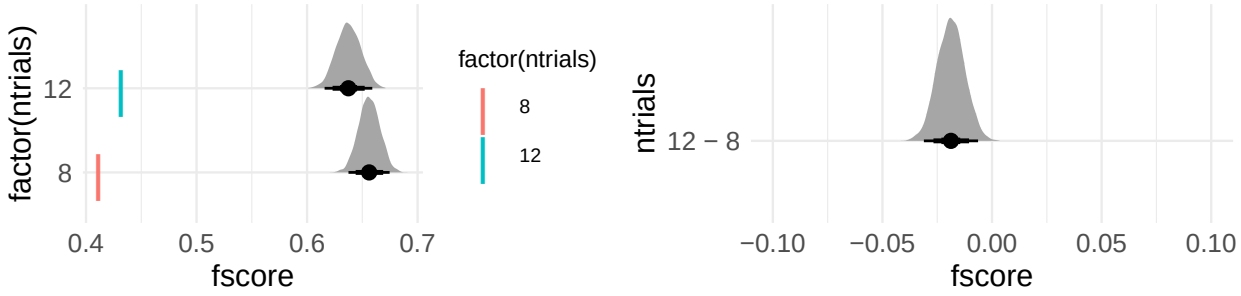
are reliably better at reducing the FDR.



Comparison of the different using composite metrics (F-scores, MCC and Payout) *F-scores* are a common metric used in ML to get a composite score for the performance of an algorithm and takes into account the number of True Positives, False Positives and False Negatives. It is given by F-score = $\frac{2precision}{precision+recall} = \frac{2TP}{2TP+FN+FP}$. We can use this to compare performance across the different uncertainty representation conditions.

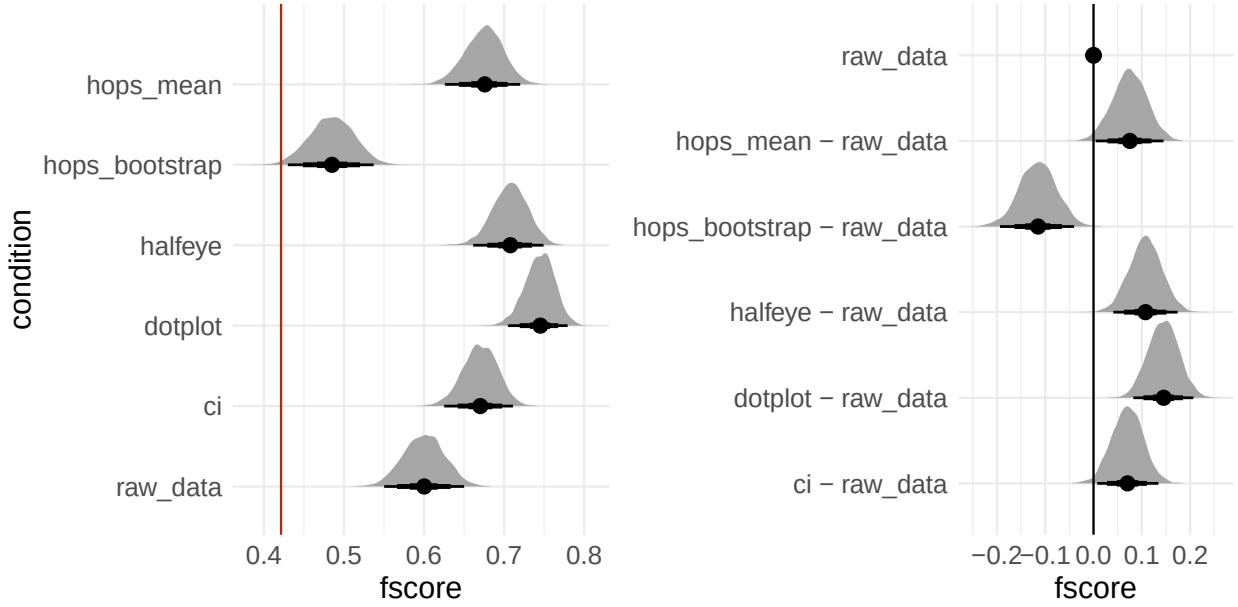
In this analysis, we would compare the F-scores of an average user to those of the BH procedure (which we consider optimal for this task).

First we compare the difference in F-scores when we manipulate *nregions* i.e. the number of graphs presented to participants (8 or 12). We see that F-scores actually decrease when participants were presented with more graphs (which might be because the decrease in FDR might also entail a decrease in the number of True Positives).



Next, comparing the difference in F-scores between different uncertainty representations, we see that all the uncertainty representations, except HOPs of bootstrapped data samples reliably increase F-scores. In other words, the accuracy increases when these uncertainty representations are used. Interestingly, the

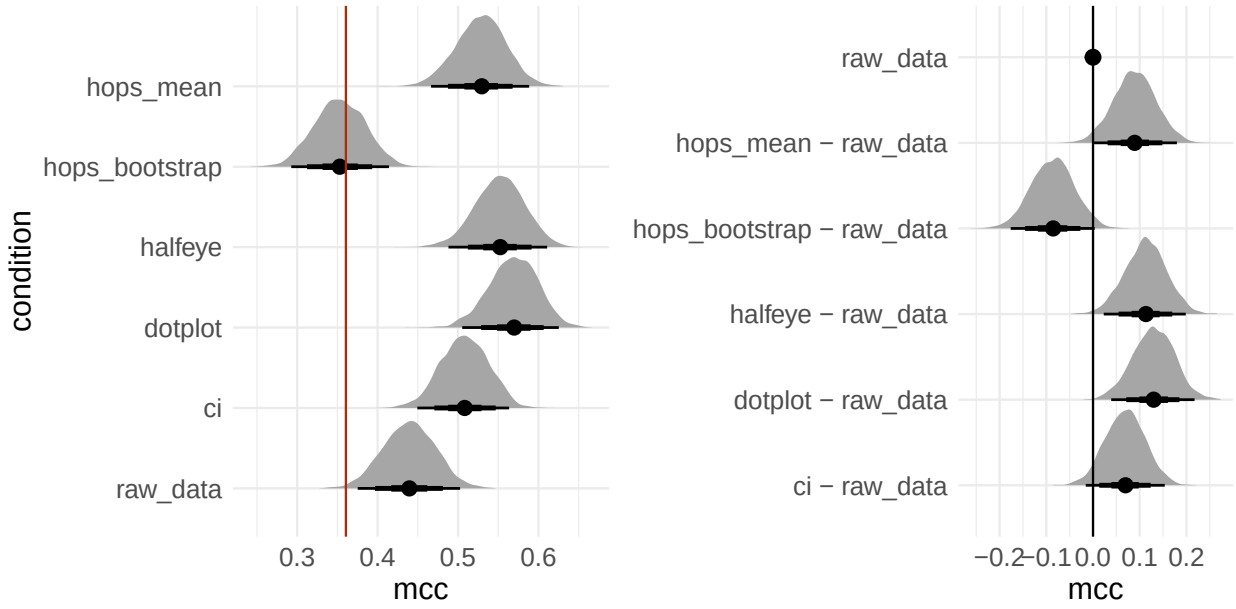
dotplot results in the highest accuracy (by almost 15 percentage points, 95% CI: [0.08, 0.2]) compared to the Baseline. Other uncertainty representations such as `probability density plots`, `HOPS` and `95% confidence intervals` of the mean difference also improves the F-scores.



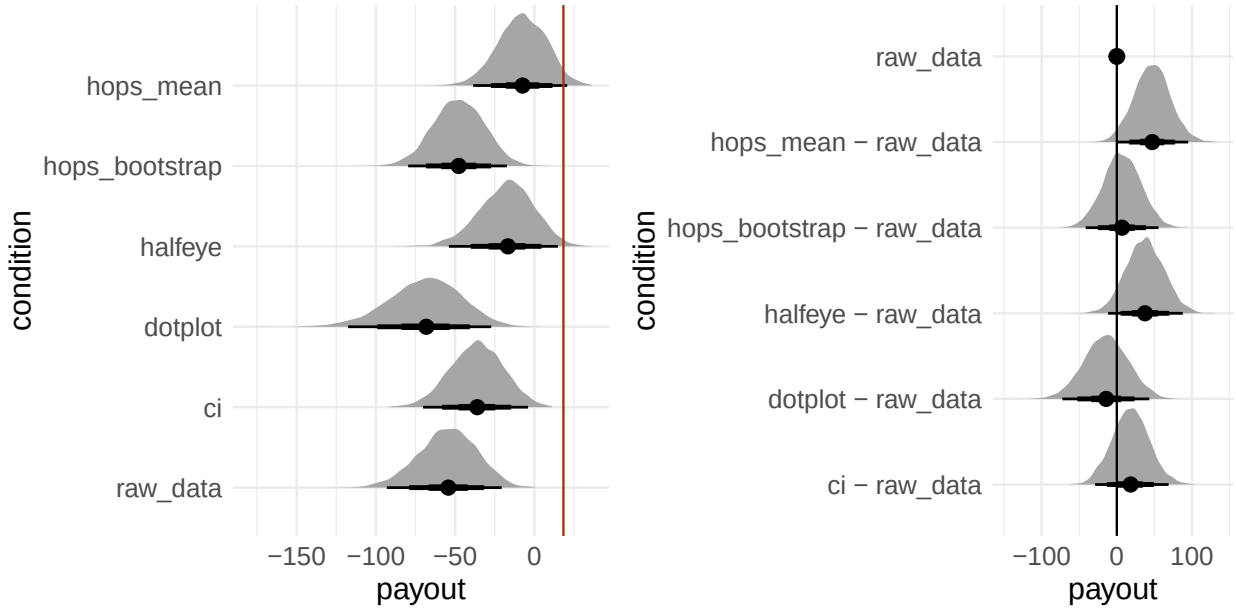
The following table summarise the difference in F-scores of the different conditions compared to the baseline.

```
## # A tibble: 5 x 7
##   condition                fscore   .lower  .upper .width .point .interval
##   <chr>                     <dbl>    <dbl>   <dbl>  <dbl> <chr>  <chr>
## 1 ci - raw_data            0.0704  0.00753  0.134    0.95 median qi
## 2 dotplot - raw_data       0.145   0.0826   0.207    0.95 median qi
## 3 halfeye - raw_data       0.108   0.0410   0.174    0.95 median qi
## 4 hops_bootstrap - raw_data -0.115  -0.194   -0.0403   0.95 median qi
## 5 hops_mean - raw_data      0.0753  0.00431  0.145    0.95 median qi
```

Matthews Correlation Coefficient: One common drawback of the F-score is that it does not take into account False Negatives. The Matthew's Correlation Coefficient is a proposed measure to address this limitation, and is calculated as $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.



In our study, we incentivise participants using a payout scheme. It might be the case that participants are optimising for the incentives provided. Hence we compare the average payout across the different conditions. Based on this measure, we see that participants in the `probability density plots` and `HOPS` of the mean difference, on average, have higher payouts.



Power analysis