vary_distribution

Sayani Gupta

18/12/2020

Contents

1 Recap of last week 1

1.1 What worked? 1

1.2 What didn't work? 2

2 Results from this week 2

2.1 What worked? 2

2.2 To do now? 2

2.3 Structuring the simulation section of the paper? 3

3 Detail results: 3

3.1 Standard normal distribution 3

3.2 How mean and sd changes with increasing x-axis and facet levels? 3

3.3 Normal distributions with different means 3

3.4 Gamma Distribution with different locations before quantile transformation 3

3.5 Gamma Distribution with different locations after quantile transformation 3

3.6 Normalise by permuting data 3

Designs

D_{null} (No difference in distribution)
D_{var_f} (Difference in distribution only across facets)
D_{var_x} (Difference in distribution only across x-axis)
$D_{var_{all}}$ (Difference in distribution in both facets and x-axis)

1 Recap of last week

1.1 What worked?

We saw the effect on raw wpd for different choices of tuning parameter (λ) and effect of different increments (ω). We saw that raw wpd has least values for null design D_{null} and gradually increase with more deviation from the null case with maximum for $D_{var_{all}}$ (which was great). Values of λ for which D_{var_x} and D_{var_f} interact is 0.5 (nice as that's what we were expecting). Moreover, the raw wpd values increase with increasing λ for D_{var_x} and decrease with increasing λ for D_{var_f} (as expected since the measure is linearly related to λ.)

1.2 What didn't work?

Though the distribution of wpd remains the same with normal distribution with different means just by scaling the data as a prepossing step. The same doesn't happen for a Gamma distribution where the location of the distribution changes with different locations of the Gamma distribution. Box-cox transformations on the raw data did not work. It was suggested that we try out quantile transformation to bring back the quantiles of the raw data to standard normal.

2 Results from this week

2.1 What worked?

- Preprocess data: quantile transformation of the data to standard normal.

[Sort the sample $x_{(1)}, x_{(2)}, \dots, x_{(n)}$. Estimate the cumulative probabilities for each point $p_{(1)}, p_{(2)}, \dots, p_{(n)}$ such that $p_{(i)} = P(X < x_{(i)})$. Transform each observation $x_{(i)}$ into observation $y_{(i)} = \Phi^{-1}(p_{(i)})$, where Φ^{-1} is the inverse of the cumulative distribution of a standard normal distribution.]

(** All of this is being done by the R function `qqnorm`. I checked that data from different distribution (z) could be transformed to normal using `qqnorm(z)$x`. Please let me know if you think I am missing anything here**)

- check if the distribution of the raw wpd starting from a Gamma (0,5, 1) and Gamma(2, 1) distribution looks similar to the distribution of the raw wpd starting from a N(0,1) distribution. (More or less it does, yippee. (Please let me know if you think I am missing anything here Figure 4, Figure 5, Figure 8)
- how mean and sd of the distribution of raw wpd changes with increasing x-axis and facet categories - more understanding of which normalisation process to choose. (Figure 2)
- Now that quantile transformation has worked, we wanted to check which process of normalization works. Suppose $\{d_1, d_2, \dots, d_D\}$ are the JS distances from a panel

a) $wpd_{norm1} = wpd/log(D)$

b) $wpd_{norm2} = wpd - mean(wpd^*)/sd(wpd^*)$, where wpd^* 's are obtained by shuffling the data. The distribution of wpd_{norm2} is expected to have similar mean and sd for different x-axis and facet categories.

Appoach a) did not work. But Approach b) is working. (**Please let me know if you think I am missing anything from the graph**)

2.2 To do now?

- Behavior of wpd_{norm2} under different designs need to be monitored. Designs need to be constructed such that there is no **confounding factors** while deciding how distribution of wpd_{norm2} changes with different x-axis and facet categories.

- $$\mu_{j.} = \mu + j\omega$$

  (for design vary_x)

- $$\mu_{.k} = \mu + k\omega$$

  (for design vary_f)

If the designs are like the above, it would mean that mean and sd of the panels would also increase when x-axis and facet categories increase.

One way to keep mean comparable across situations would be to do the following (as discussed last with Di) Suppose we will allow a highest mean difference of M unit for each panels, which means panels with different x-axis and facet categories would also have a highest mean difference of M units.

For example, if $M = 5$ and design $= D_{var_x}$ then

$$\mu_{j.} = \mu + j\omega$$

- $\omega = 5$, for $nx = 2$
- $\omega = 5/2$, for $nx = 3$
- $\omega = 5/3$, for $nx = 4$ and so on.

Should we be bothered about it or not since we are pre-processing the data to start with?

2.3 Structuring the simulation section of the paper?

Follow the story line like this one? **Or you suggest some other ways?**

3 Detail results:

Assumptions: There is no difference in distribution between any facet or x-category since it is the null design. Process is run $nsim$ times to plot the distribution. $nsim = 200$ $\lambda = 0.67$

Questions: - How raw value of wpd changes with different nx and nfacet for different location and scale of a Normal and non-normal distribution

3.1 Standard normal distribution

3.2 How mean and sd changes with increasing x-axis and facet levels?

3.3 Normal distributions with different means

3.4 Gamma Distribution with different locations before quantile transformation

3.5 Gamma Distribution with different locations after quantile transformation

3.6 Normalise by permuting data

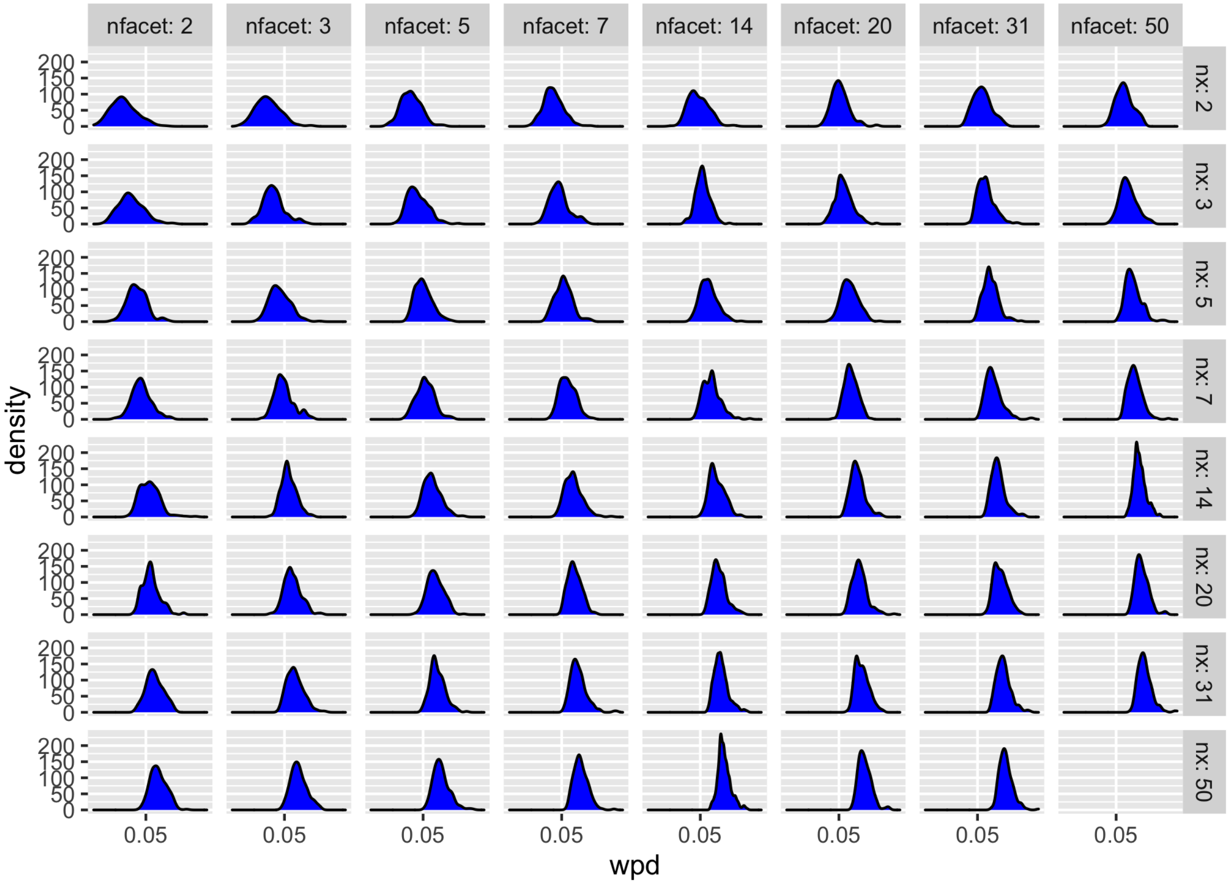

Figure 1: Density plots of raw wpd is shown for N(0,1) distribution. The densities change across different facet and x levels, which implies wpd value changes with increasing and facet levels

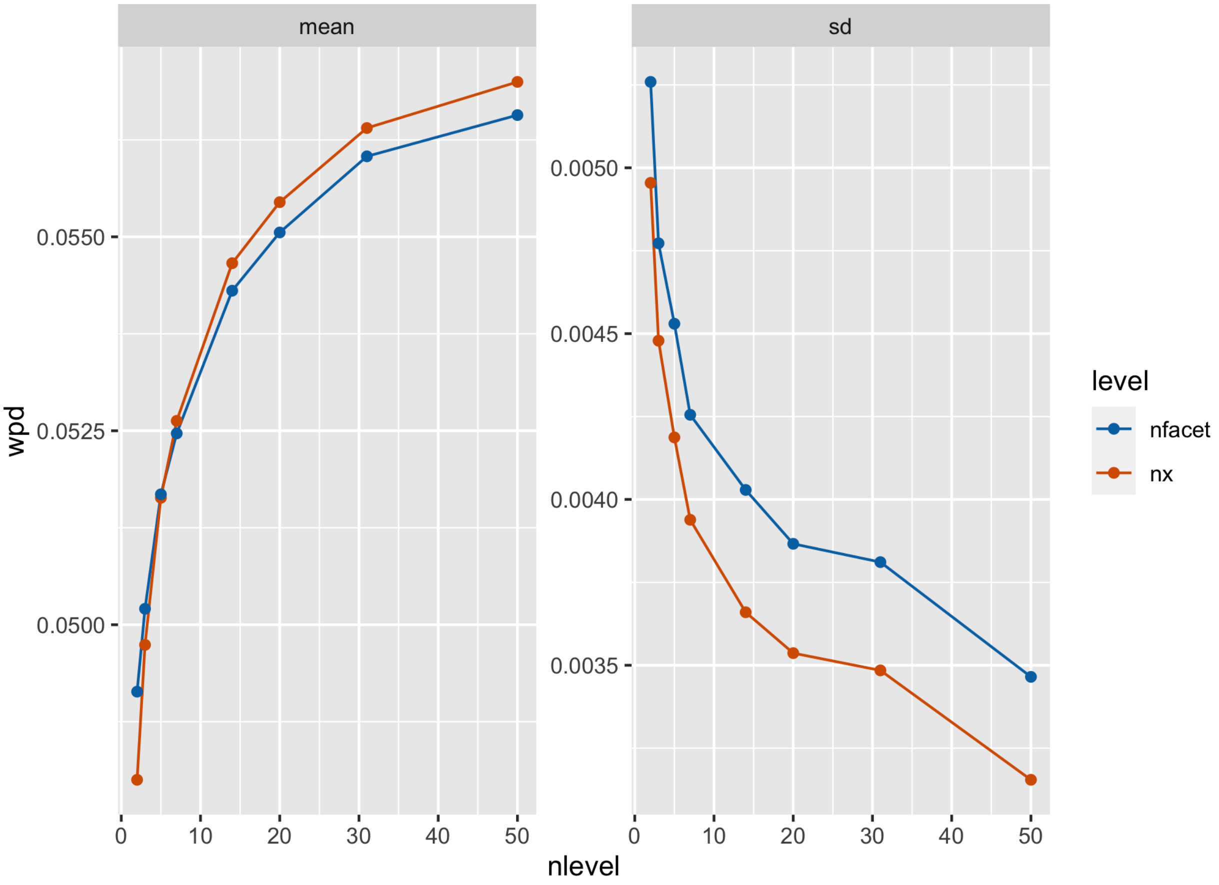

Figure 2: Movement of mean and sd for raw wpd is shown for different number of levels (nlevel) of x-axis and facets through line plots. Mean increases and sd decreases more sharply for increasing x-axis levels as compared to facet levels. It seems like both mean and standard deviation are affected more by change in the x-axis levels than the facet levels.

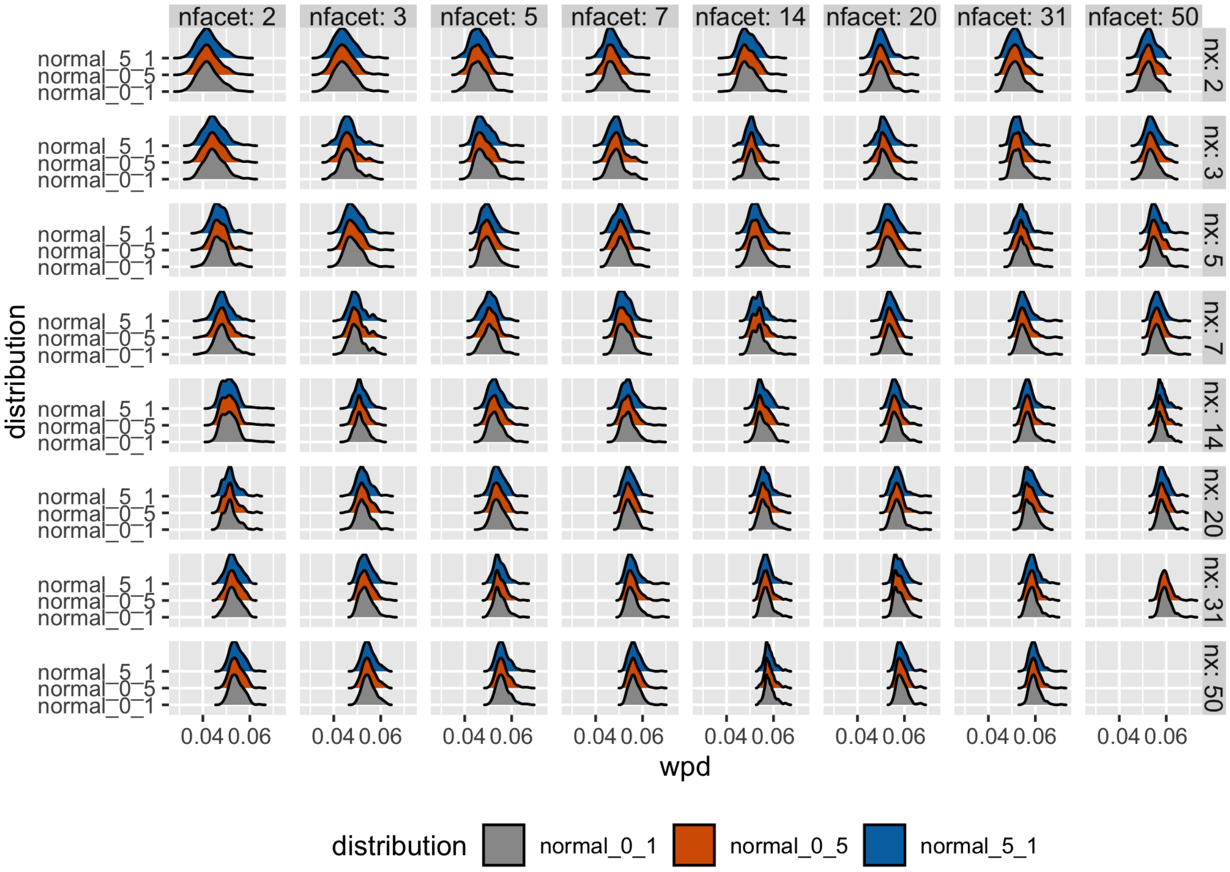

Figure 3: Ridge plots of raw wpd is shown for N(0,1), N(0,5) and N(5,1) distribution. The densities change across different facet and x levels but look same for the three normal distributions, which implies wpd value is unaffected by the change in mean value of the normal distribution

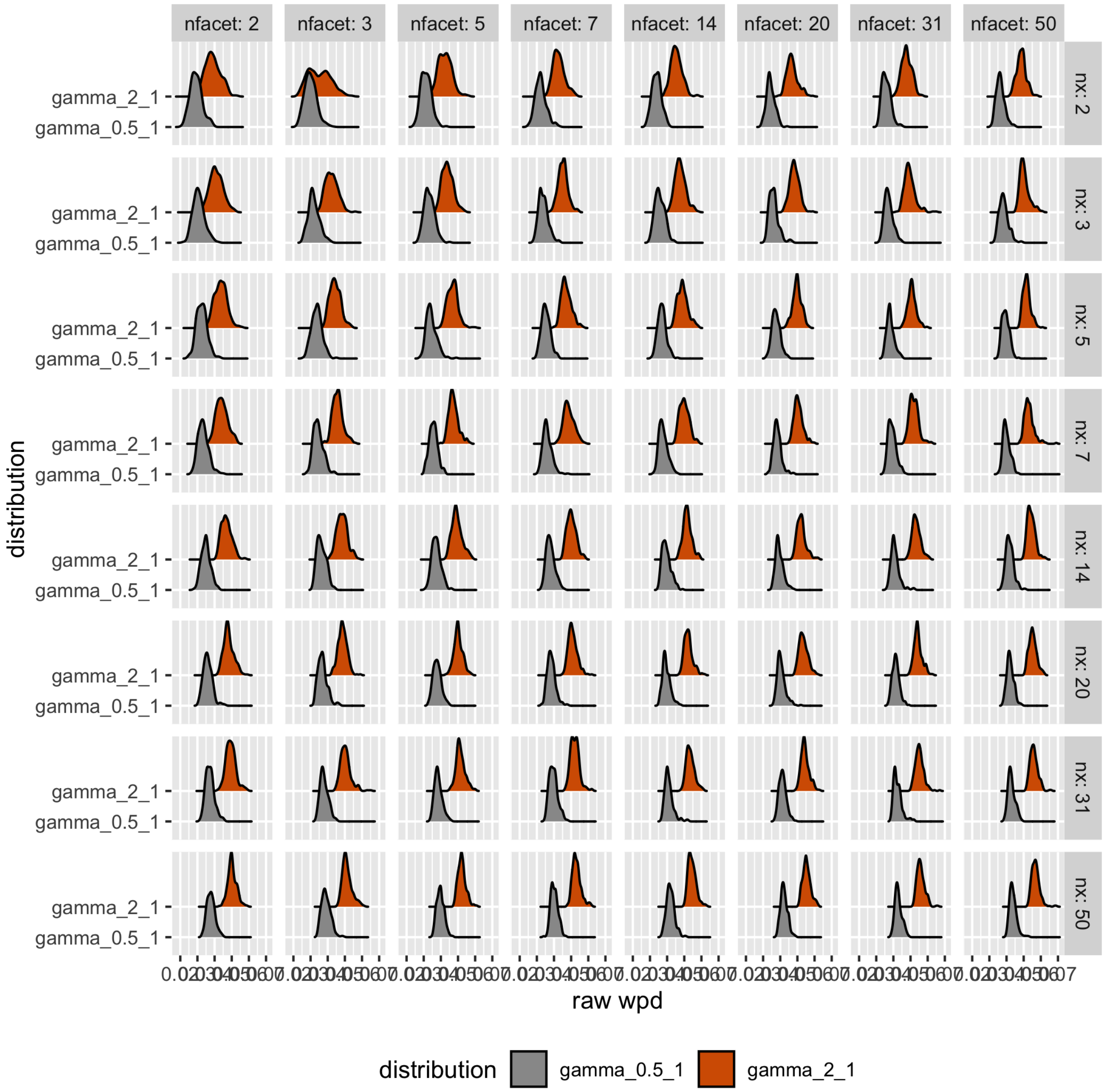

Figure 4: Ridge plots of raw wpd is shown for G(0.5,1) and G(2,1) distribution (before quantile transformation). The densities change across different facet and x levels and also for the two distributions, which implies wpd value is affected by the change in location value when distributions are not normal.

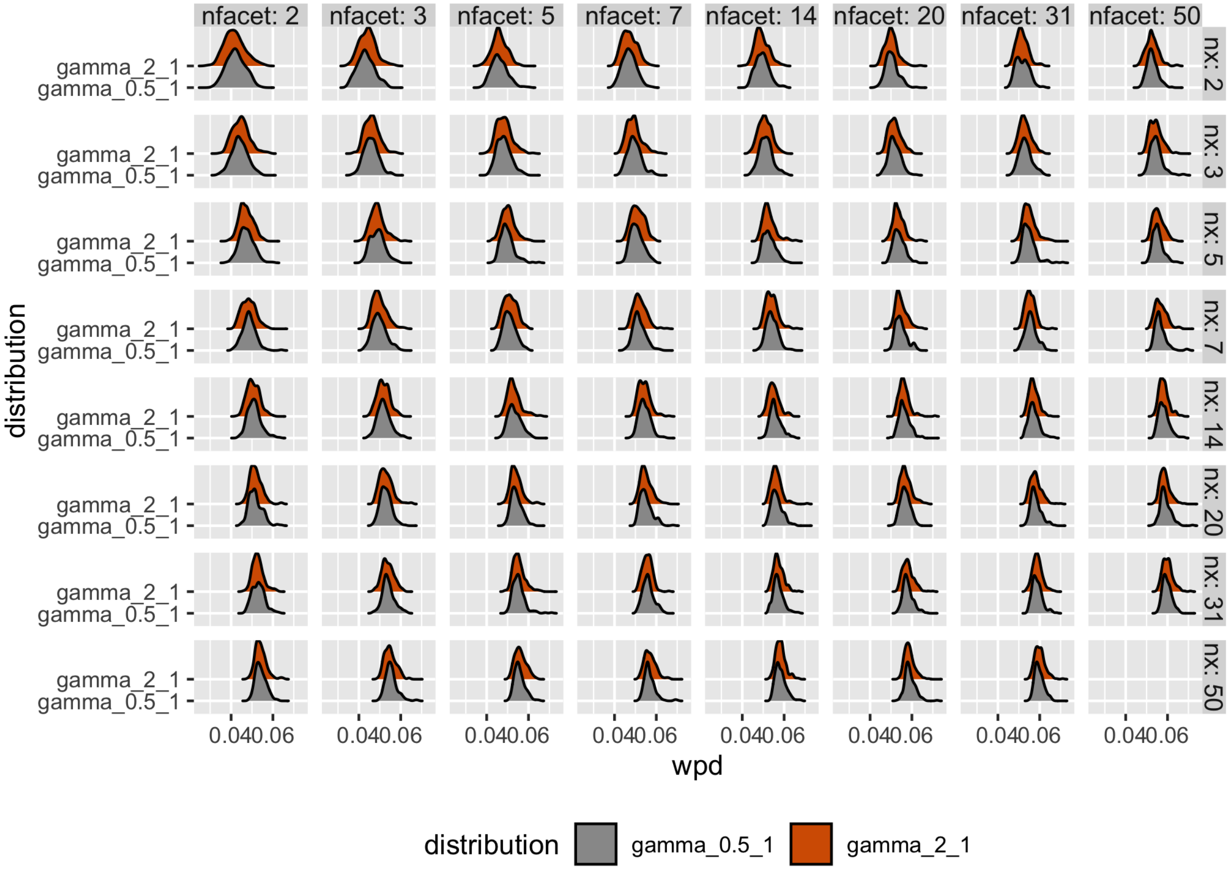

Figure 5: Ridge plots of raw wpd is shown for G(0.5,1) and G(2,1) distribution after quantile transformation looks similar and hence is unaffected by change in location.

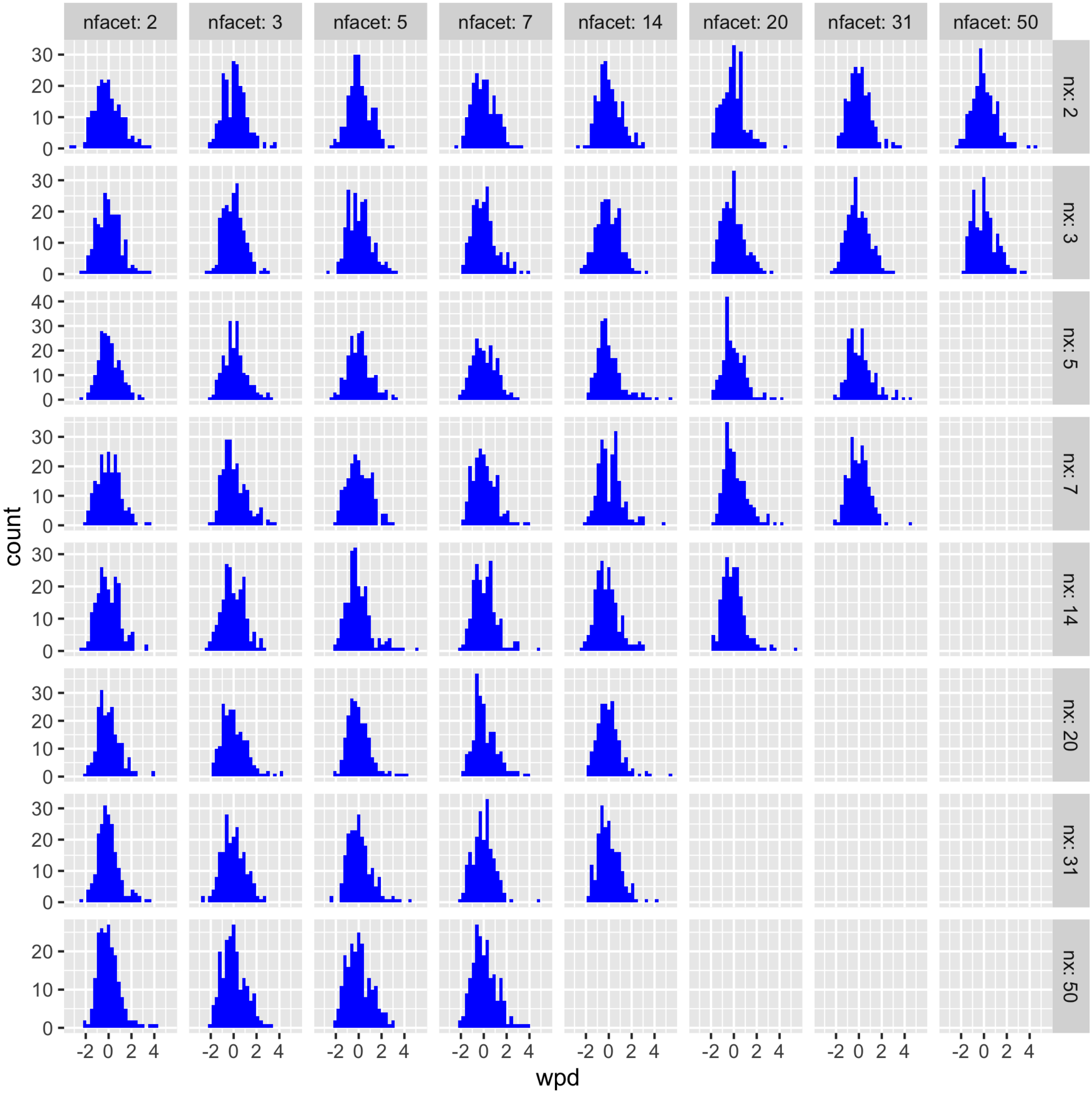

Figure 6: Density plots of normalised wpd is shown for Gamma(2,1) distribution. The mean and sd of the distribution looks almost similar for different facet and x-axis categories

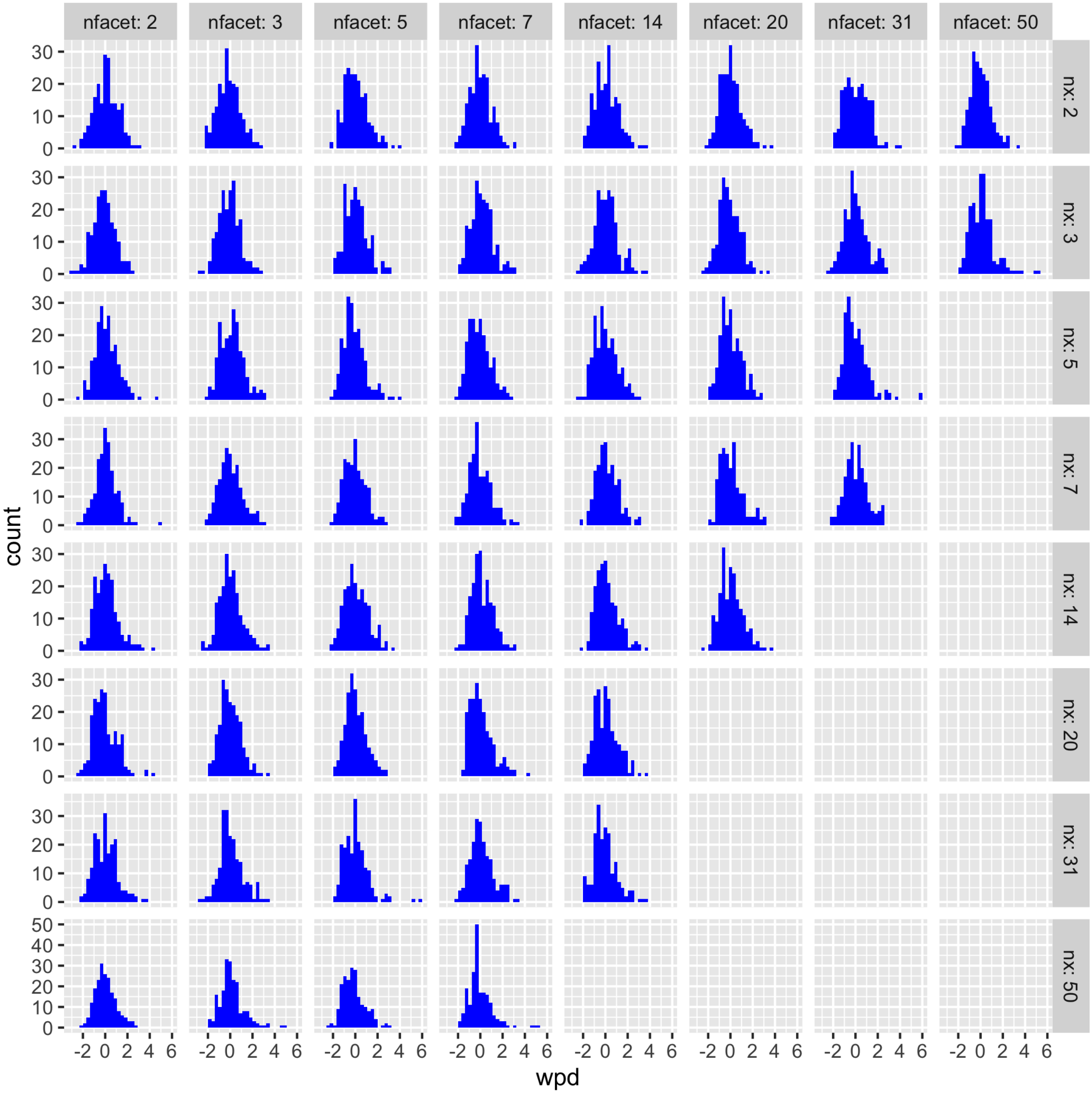

Figure 7: Density plots of normalised wpd is shown for Gamma(0.5,1) distribution, which is more skewed. The mean and sd of the distribution looks almost similar for different facet and x-axis categories

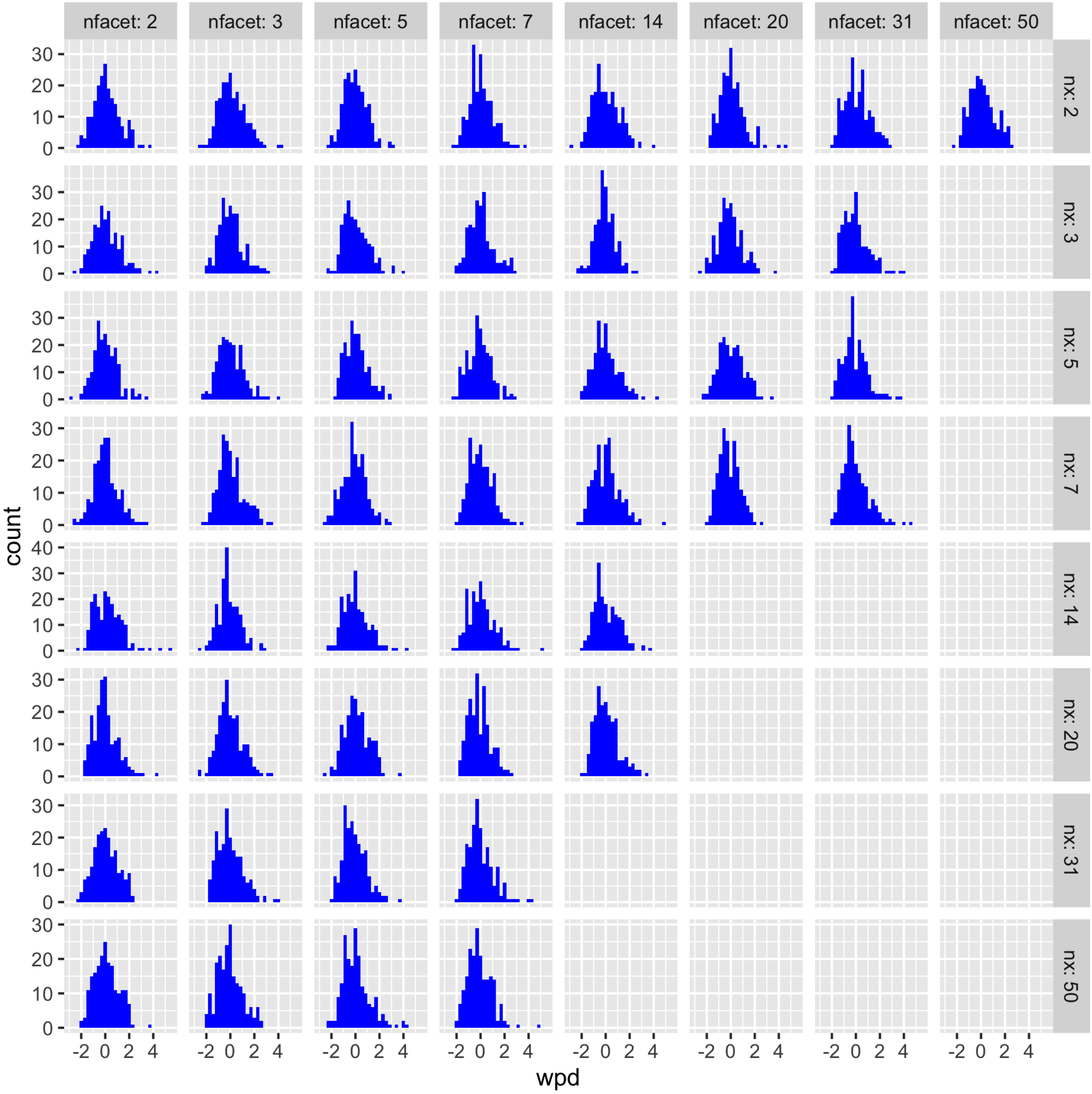

Figure 8: Density plots of normalised wpd is shown for Normal(0,1) distribution. The mean and sd of the distribution looks almost similar for different facet and x-axis categories.