Handling deviation from normality

Brice Ozenne

May 4, 2021

A question that often arise in consultations is:
"My data is not normally distributed. What should I do?"
In my views, this is a very vague question. It is a bit like going to the doctor and asking him:
"My body temperature is outside of the 36-37°C normal range. What should I do?"
In many cases statistical methods are robust (i.e. ensure approximate type 1 error control and give nearly efficient estimators) to "small" deviations from normality. So while it is a good idea to check for normality, it is not a good idea to change statistical tool just because you get a significant p-value out of a Shapiro test. If I follow the analogy, medicine can be harmful and one should not take medicine every time ones body temperature is outside the normal range. It is however a good idea to check what is going on when your temperature is high to make sure it does not hide something bad.

To ease the discussion, we will consider a simple example where we want to compare the outcome distribution between two groups. We assume to have no missing data and no measurement error, and that no external covariate is relevant. In that case, we can visualize the distribution of the outcome per group and perform the comparison "visually". We will consider four examples (Figure 1):

- Normally distributed outcomes: no problem here.

- Student distributed outcomes: symetric and unimodal distribution but with outliers.

- Gamma distributed outcomes: asymetric distribution.

- Normally distributed outcomes with ceilling effect: many observations have exactly the same value.
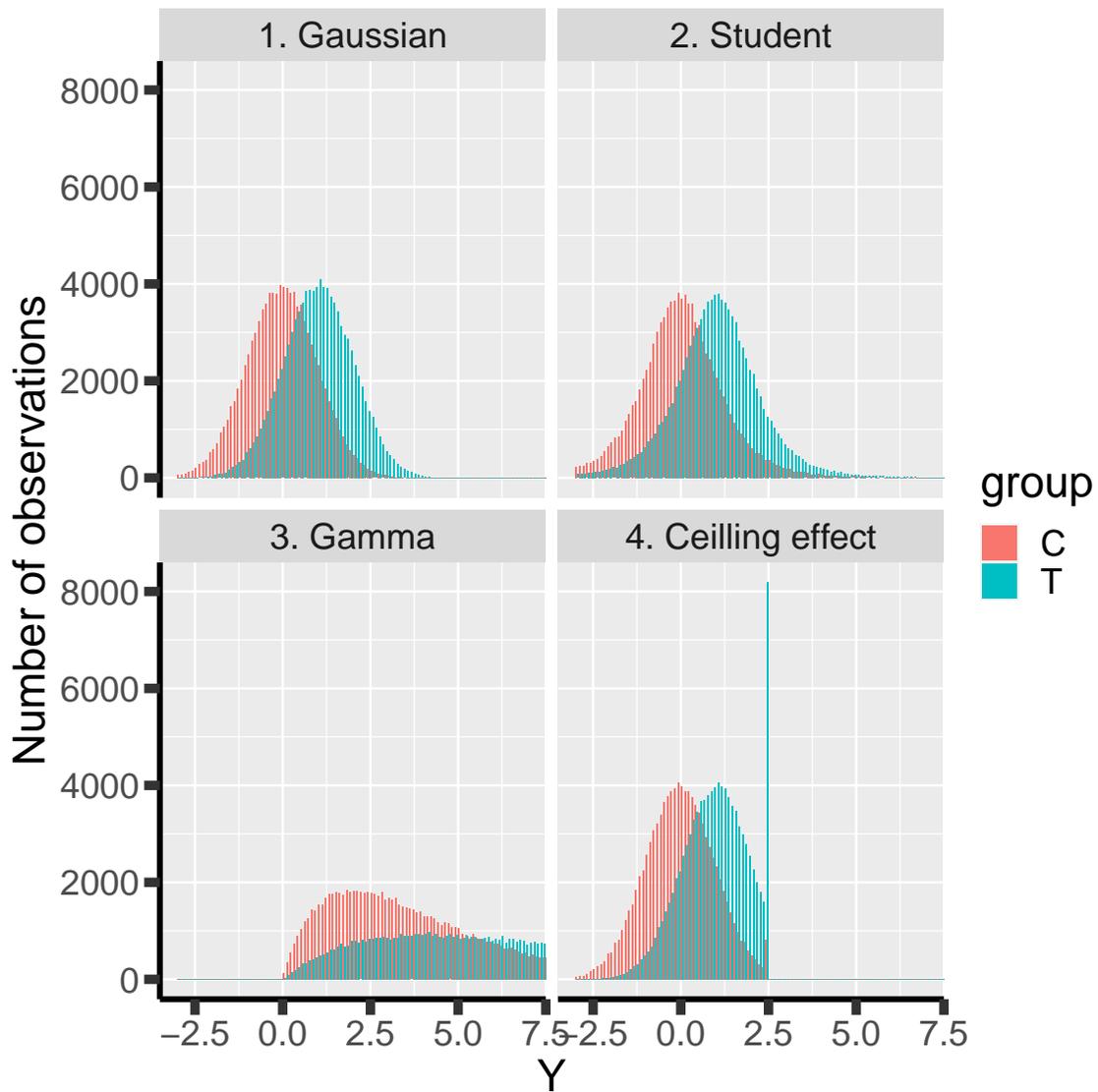
Figure 1: Example datasets.

⚠ if any, distributional assumptions are usually made on the resiudal terms, e.g:

$$Y = X\beta + \varepsilon \text{ where } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

and not on the outcome Y. Concretely, we don't assume that the outcome is normally distributed but that within groups (or once we remove the group effect) it is normally distributed. In example 1, the outcome is clearly not normally distributed (it is bimodal) but within groups it is normally distributed.

1 Statistical properites

When we use a statistical procedure to estimate a parameter of interest (say θ), we generally want to report an estimate of this parameter based on the data we have (denoted $\widehat{\theta}$). We generally also want to report the uncertainty around this estimate via a confidence intervals (denoted $IC_{\widehat{\theta}} = [L_{\widehat{\theta}}; U_{\widehat{\theta}}]$) or report a p-value that reflects whether a value (say θ_0) is compatible with the data at hand. We will therefore distinguish three properties for a statistical procedure:

- **validity of the estimate**: the estimate should tend to the true value as we increase the sample size (consistency) or the average estimate should tend to the true value as repeat the experiment (unbiasedness).

- **validity of the uncertainty**: the confidence interval should have proper coverage, typically they should contain the true value with probablity 95%. The p-value should have a uniform distribution under the null meaning that the type 1 error is controlled at its nominal level, typically 5%.

- **efficiency**: the procedure is optimal in the sense that it makes the best use of the data at hand.
  For instance we expect the estimates to be as precise as possible, i.e. to have the narrowest possible confidence intervals. We also expect that the type 2 error should be the smallest possible (i.e. highest power). In other terms, whenever the null hypothesis is false, the test should rejected it as often as possible.

In the following we will assume that these properties are of decreasing interest: having unbiased estimates is the most important as quantifying uncertainty and efficiency can be fixed by replicating studies and performing a meta-analysis. A non-efficient estimator can still give useful and valid results - it will "just" waste ressources but not be misleading (if used and interpreted correctly).

2 Should we worry about normality?

Most statistical procedures do not require any normality assumption to provide consistent estimates with asymptotically valid confidence intervals and p-values. For instance t-tests and linear regressions can be shown to provide consistent and asymptotically normally distributed estimates in large iid [1] samples, regardless to whether they are normally distributed as soon as their first two moments are finite. Here it is important to distringuish between the distribution of the outcome (say Y) and the distribution of the parameter of interest, often the mean of Y. Averaging "normalizes" the distribution, which is formalized in the central limit theorem, and illustrated on Figure 2:
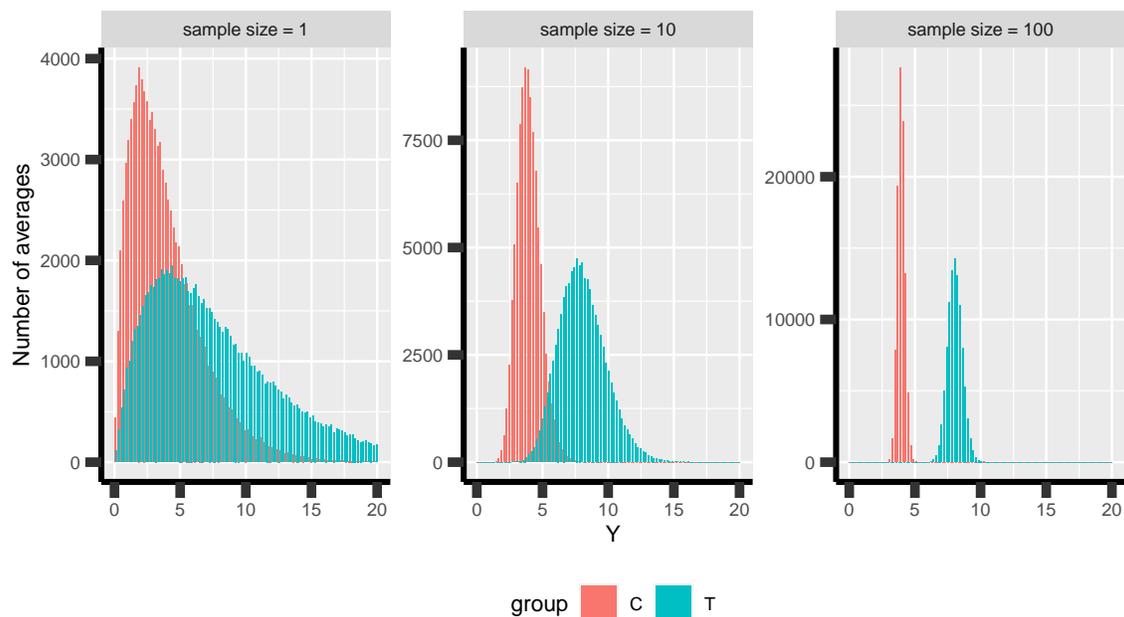

Figure 2: Distribution of the estimated mean along the sample size.

This means that (almost) regardless to the input data, we will be able to estimate parameters which follows a normal distribution, i.e. for which we can quantify the uncertainty. Results from the M-estimation theory or the maximum likelihood theory can be used to show that finding parameters that minimize an error that is the lack of fit relative to individual observations lead to consistent estimates. Concretely, this means that the coverage/type 1 error control of many standard procedures such as the t-test and the linear regression will be at their nominal level in large samples, even though the normality assumption is not fullfilled (Figure 3) . . . for large enough sample sizes.

---

[1] independent and identically distributed

4

Does that mean we should not worry about normality? No:

1. we may have a valid test / consistent estimate of a meaningless parameter.

2. we may only have a small sample.

3. our estimator may not be efficient. This is usually not a problem, except when we loose so much efficiency that the estimate becomes very variable. This typically happen in presence of outliers.

4. no assuming normality complicates the understanding the group effect, as the normal distribution is one of the few distribution that can be summarized by two, easily interpretable, independent, parameters (mean and variance).

5. no assuming normality complicates other diagnostic tests, e.g. correlation is not independence and mean
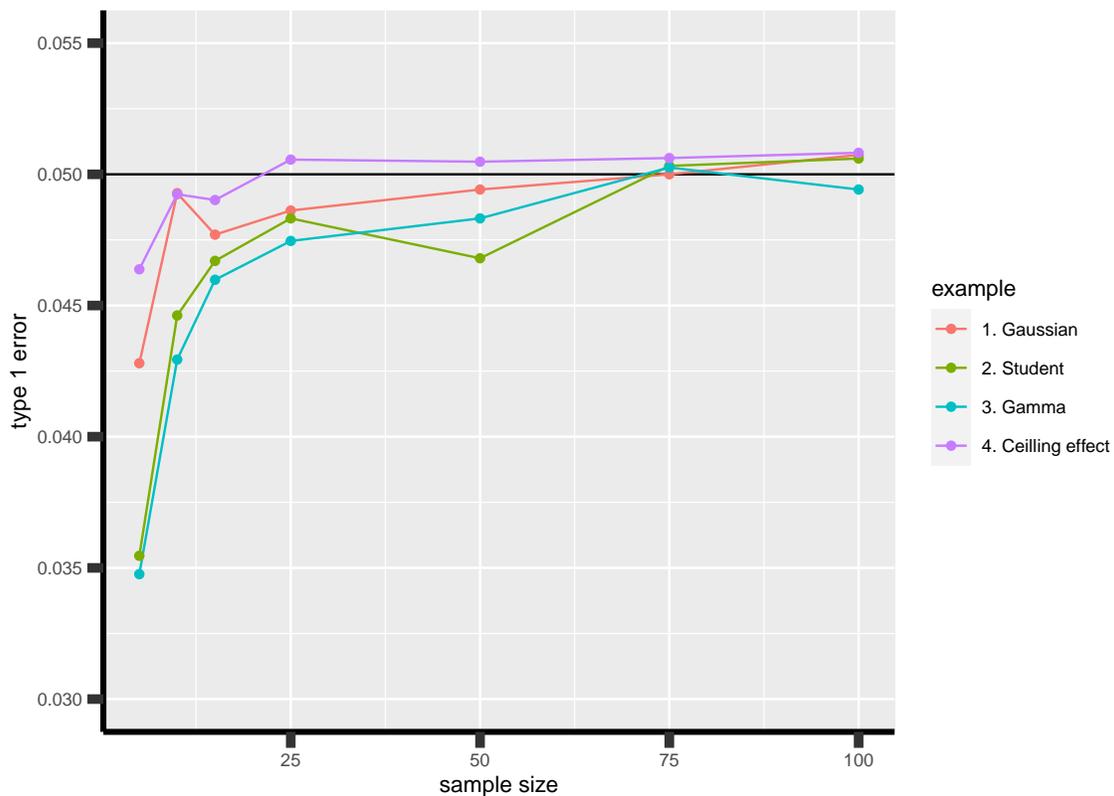
In the following we will focus on issue 1-3.

Figure 3: Coverage of the t-test.

3 Issue 1: parameter of interest with non-normal data

By default, we generally use the mean to define our parameter of interest. In our example the difference in mean between the two groups meaning that we summarize the distribution of the outcome for each group by its mean (or expected value) and then take the difference between groups. This is somehow arbitrary, we could have used another summary statistic like the standard deviation, the median (or any other quantile), the mode, ... However it is not completely arbitrary:

- it is **convenient** to model and compute: many estimators and softwares have been developped for modeling the mean. Also this can be done in a numerically stable and efficient way.

- it is a **natural** choice if the outcome is normally distributed as the mean and the variance fully characterize the distribution so no need to model other summary statistics. In particular, for normal distributions the mean is equal to the median and the mode of the distribution.

- it is **easy to interpret** if the outcome is normally distributed as it is the average but also most likely value.

When the distribution is not normal, the last two arguments might not be true. While they approximately hold if the distribution is unimodal and symmetric, they are not valid for asymetric or bimodal distribution. For instance, the mean of a binary variable will correspond to a value that is never observed! If we look at Figure 4, we can see that the mean is far from being the most likely value (i.e. the mode). The median is slightly closer to the mode but does not really provide a satisfactory improvement.
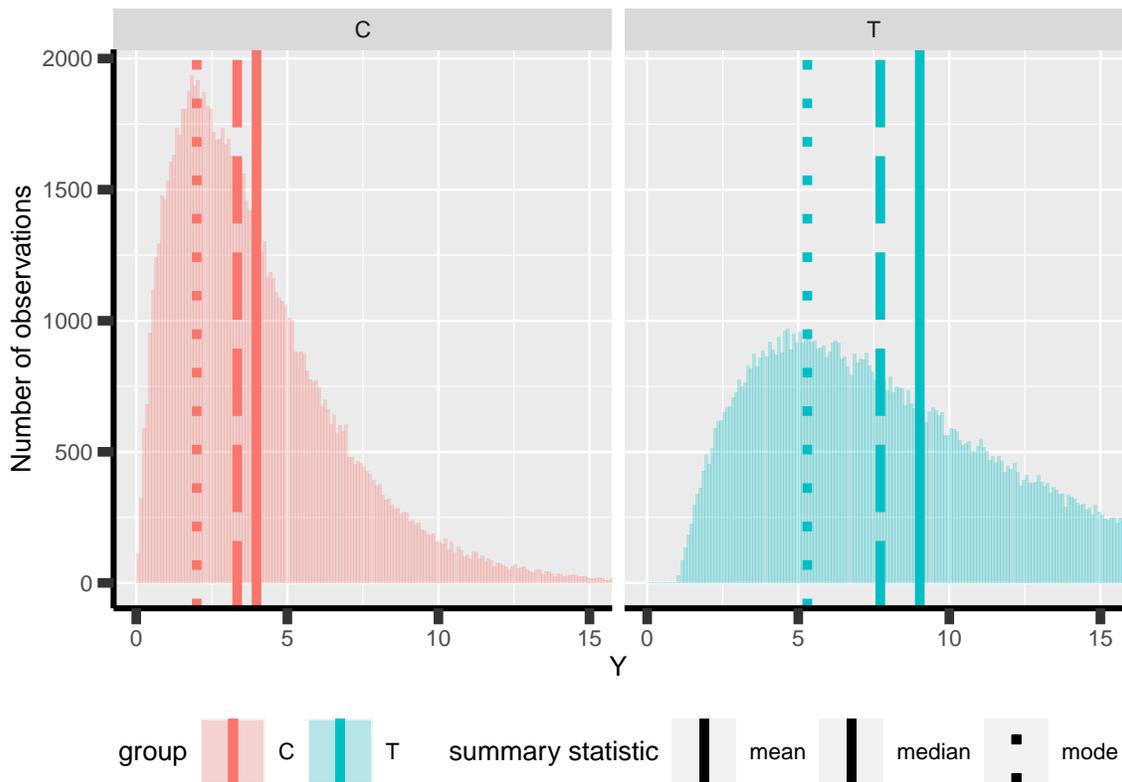
Figure 4: Mean, median, and mode two asymetric distributions.

In such a case, it can be a good idea to define a new parameter of interest. One could for instance apply a transformation that normalizes the distribution (e.g. log-transformation, see Figure 5), estimate the mean of the transformed data (here 1.1 vs 1.8), and compare them across groups (here 0.7). In the case of a log-transformation, the back-transformed difference and get a nice interpretation has a nice interpration: it is a multiplicative effect (exp(0.7)=2, i.e. the mean in the treatment is twice larger than in the control group). So, instead of studying an additive group effect (on the mean), the parameter of interest is a multiplicative group effect (on the mean).
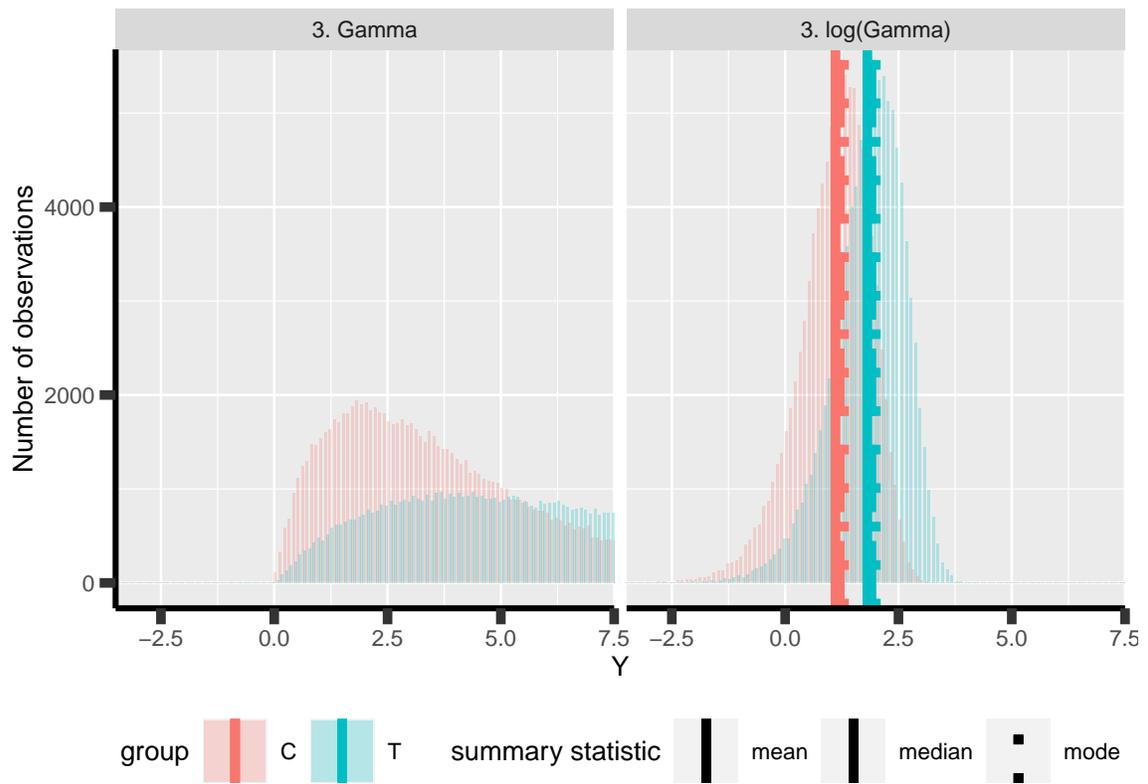
Figure 5: Mean, median, and mode on the log-transformed data

TODO: Mann-Whitney parameter (e.g. if one cannot find a good transformation), Zero inflated model (example 4)

4 Issue 2: handling small samples with non-normal data

Bias correction
    Permutation tests
    Robust standard error

5   Issue 3: handling outliers

No good solution here. Check the influence of each observation. Median more robust than mean.

6   References