Handling deviation from normality

Brice Ozenne

January 28, 2025

A question that often arise in consultations is:

"My data is not normally distributed. What should I do?"

I find it difficult to answer as it is vague. 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 some 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. First thing to consider is the type of variable for which the concern about normality arised:

- **outcome**: non-normality can be a threat to the interpretability of the estimate and the validity of subsquent hypothesis test. Example of non-normality are shown in Figure 3 and corresponding solution are discussed section 3.

- **exposure**: generally not a problem. One notable exception is in presence of outliers when assuming a linear exposure effect as the outliers may have an unacceptable influence on the exposure effect (see Figure 1 for example). This will be discussed in section 2.

- **covariate**: generally not a problem. One notable exception is in presence of outliers when assuming a linear exposure effect as the outliers may have an unacceptable influence on the exposure effect. Possible solutions relaxing the linearity assumption (using splines or categorizing the covariate). If the covariate is not strongly related to the outcome one can also consider excluding it from the model. It is not further discussed in this document.

Before discussing any solution, the next section what is understood as a 'good' or a 'valid' statistical procedure.

1 Statistical properties

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%.

- **robustness**: means that the validity of the estimate and uncertainty is not (dramatically) altered when the sample is altered by a fixed proportion of extreme values (say 1% of outliers).

- **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). Typically robustness can be acquired at the expense of efficiency (and interpretability).

2 Non-normal exposure

Consider the following example shown in Figure 1 (example 5.1 in Maronna et al. (2019)) evaluating the association between the contents of Western Australian rocks. Consider first a **linear relationship**:

- one point has a significant impact on the regression slope: 0.135 (p<0.001) vs 0.030 (p=0.18) when exluded. The 'red' regression line does not seems to be a reasonnable summary of the association.
- even after removal of observation 15, the exposure is still far from following a normal distribution as it seems very right-skewed. Nevertheless the 'blue' regression line seems to be a reasonnable summary of the association.
- modeling the median instead of the mean only partially solves the issue: observation 15 has still a noticeable influence on the slope (0.080 vs. 0.056). In fact in a more extreme example shown in the right panel of As shown in Figure 2 one could construct example where a small fraction of outliers would still be very influencial.

Consider instead a **non-linear** relationship:

- using thin plate regression splines seems to provide a reasonnable summary of the association with a different slope for small copper value compared to high copper content. The later slope is probably estimated with a lot of uncertainty due to only few observations with high copper content: this could be seen from the width of the confidence intervals of the regression line.

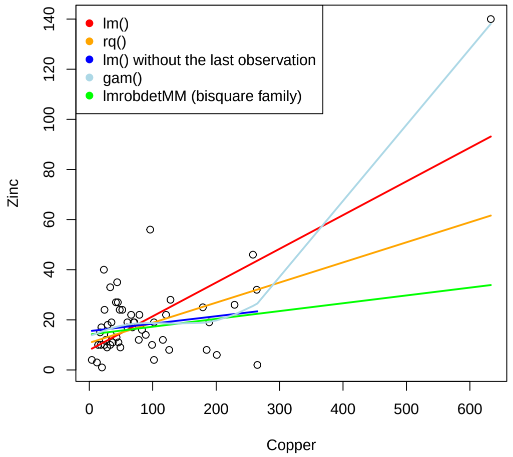


Figure 1: Real data with non-normal exposure.

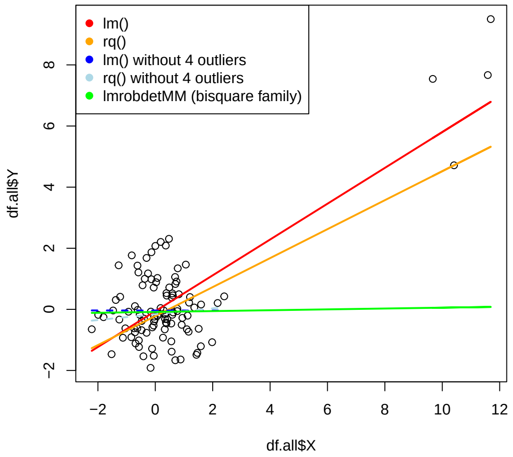


Figure 2: Simulated data with non-normal exposure.

This example illustrate several key considerations:

- the exposure does not need to be normally distributed for statistical methods to apply BUT parametric assumptions, such as linearity, make common statistical tools sensitive to extreme values.

- it can be a good idea to restrict the study of the exposure to commonly observed values. This is similar to an inclusion criteria in a clinical trial on the disease severity (e.g. not including terminally ill patients).

- some methods sometimes refered as 'robust', like median or quantile regression, have been developped to handle extreme values in the outcome not in the exposure and thus may not lead to a satisfactory solution. For instance, median regression minimizes the average of the absolute residuals and can therefore be greatly influenced by a single leverage point.

- solutions include relaxing parametric assumptions or using more specialized 'robust' technics, e.g. estimators based on a robust summary of the residual (e.g. median of the absolute residuals). They typically make the interpretation more challenging either because there is no more a single number describing the association or the single number is no more 'just' a mean or median difference in the population of interest.

2.1 R code

2.1.1 No covariate

Load data:

```
library(RobStatTM)
data(mineral)
mineral <- mineral[order(mineral$copper),]
head(mineral)
```

```
   copper zinc
41      4    4
48     12    3
49     14   10
36     17   15
42     18   10
26     19   17
```

Quick assessment of normality of the exposure:

```
shapiro.test(mineral$copper)
```

```
        Shapiro-Wilk normality test

data:  mineral$copper
W = 0.67314, p-value = 1.374e-09
```

Analysis:

```
library(quantreg)
library(mgcv)
library(robust)
library(RobStatTM)

set.seed(1)
e.mean <- lm(zinc ~ copper, data = mineral)
e15.mean <- lm(zinc ~ copper, data = mineral[-NROW(mineral),])
e.median <- rq(zinc ~ copper, data = mineral)
e15.median <- rq(zinc ~ copper, data = mineral[-NROW(mineral),])
e.lmRob <- lmRob(zinc ~ copper, data = mineral)
e.lmRob2 <- lmrobdetMM(zinc ~ copper, data = mineral,
                       control = lmrobdet.control(family = "bisquare"))
e.gam <- gam(zinc ~ s(copper), data = mineral)

rbind(mean.all = summary(e.mean)$coef["copper",],
      median.all = summary(e.median, se = "boot")$coef["copper",],
      rob.all = summary(e.lmRob)$coef["copper",],
      rob2.all = summary(e.lmRob2)$coef["copper",],
      mean.red = summary(e15.mean)$coef["copper",],
      median.red = summary(e15.median, se = "boot")$coef["copper",])
```

```
             Estimate Std. Error   t value     Pr(>|t|)
mean.all   0.13456951 0.01982765 6.7869632 1.181421e-08
median.all 0.08024691 0.05326053 1.5066864 1.380606e-01
rob.all    0.01471517 0.02424047 0.6070496 5.465113e-01
rob2.all   0.03118872 0.02082673 1.4975332 1.404186e-01
mean.red   0.02974749 0.02205388 1.3488551 1.834611e-01
median.red 0.05590062 0.04380294 1.2761843 2.077867e-01
```

2.1.2 With covariates

We now consider a more complex example (example 5.2 in Maronna et al. (2019)) involving multiple covariates:

```
library(robustbase)
data(wood, package = "robustbase")
head(wood)
```

```
     x1     x2    x3    x4    x5     y
1 0.573 0.1059 0.465 0.538 0.841 0.534
2 0.651 0.1356 0.527 0.545 0.887 0.535
3 0.606 0.1273 0.494 0.521 0.920 0.570
4 0.437 0.1591 0.446 0.423 0.992 0.450
5 0.547 0.1135 0.531 0.519 0.915 0.548
6 0.444 0.1628 0.429 0.411 0.984 0.431
```

We fit each model:

```
e.lm <- lm(y ~ x1 + x2 + x3 + x4 + x5, data = wood)
e.lmRob2 <- lmrobdetMM(y ~ x1 + x2 + x3 + x4 + x5, data = wood,
                       control = lmrobdet.control(family = "bisquare"))
```

And extract the partial residuals

- either adding the intercept and the contribution to the variable of interest to the residuals:

```
pres.lm <-  residuals(e.lm) + coef(e.lm)["(Intercept)"] + wood$x1 * coef(e
    .lm)["x1"]
pres.robust <- residuals(e.lmRob2) + coef(e.lmRob2)["(Intercept)"] + wood$
    x1 * coef(e.lmRob2)["x1"]
```

- or using the `residuals` method and re-centering the result (to match the original mean instead of 0):

```
Mpres.lm2 <- residuals(e.lm, type = "partial")
centerX.lm <- sum(coef(e.lm)[paste0("x",2:5)] * colMeans(wood)[paste0("x"
    ,2:5)])
pres.lm2 <- Mpres.lm2[,"x1"] + attr(Mpres.lm2, "constant") - centerX.lm
```

Both approaches give the same up to a constant:

```
range(pres.lm - pres.lm2)
```

```
[1] -1.110223e-16  1.110223e-16
```

We can then combine the partial residuals in a single data.frame:

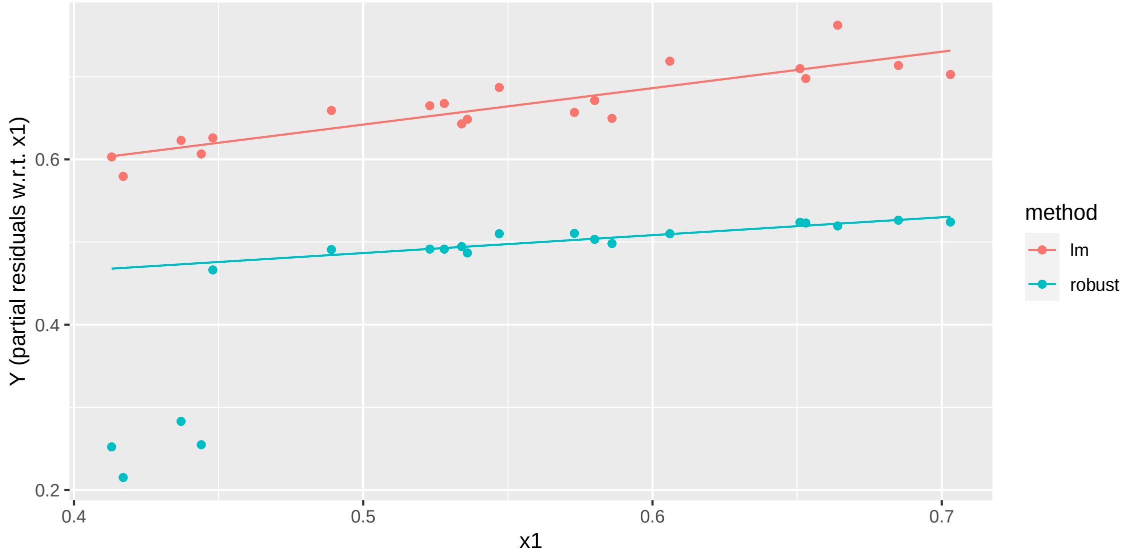
```
df.pres <- rbind(data.frame(method = "lm", x1 = wood2$x1, res = Mpres.lm),
                 data.frame(method = "robust", x1 = wood2$x1,
                            res = Mpres.robust))
```

and evaluate the model fit along x1 value, keep the other covariates at their reference level (here 0):

```
grid.data <- data.frame(x1 = seq(min(wood2$x1),max(wood2$x1),by=0.01),
                        x2 = 0, x3 = 0, x4 = 0, x5 = 0)
df.pfit <- rbind(data.frame(method = "lm", x1 = grid.data$x1,
                            fit = predict(e.lm, newdata = grid.data)),
                 data.frame(method = "robust", x1 = grid.data$x1,
                            fit = predict(e.lmRob2, newdata = grid.data))
                 )
```

to obtain the following graphical display:

```
library(ggplot2)
ggP <- ggplot(mapping = aes(x=x1))
ggP <- ggP + geom_point(data = df.pres,
                        mapping = aes(y = res, color = method))
ggP <- ggP + geom_line(data = df.pfit,
                       mapping = aes(y = fit, color = method))
ggP <- ggP + labs(x = "x1", y = "Y (partial residuals w.r.t. x1)")
ggP
```



This dataset was created to have 4 unusual observations (those with $x_1 < 0.45$). The ordinary linear regression has the better overall fit (smaller variability of the residuals) whereas the robust approach has a better fit on the subset of ’usual’ observations (smaller variability of the residuals when excluding the 4 unusual observations).

3 Non-normal outcome

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 3):

- Normally distributed outcomes: no problem here.
- Student distributed outcomes: symetric and unimodal distribution but with outliers.
- Gamma distributed outcomes: asymetric distribution. A more extreme distribution would show 'outliers'
- Normally distributed outcomes with ceilling effect: many observations have exactly the same value.

⚠ if any, distributional assumptions are usually made on the residual 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.

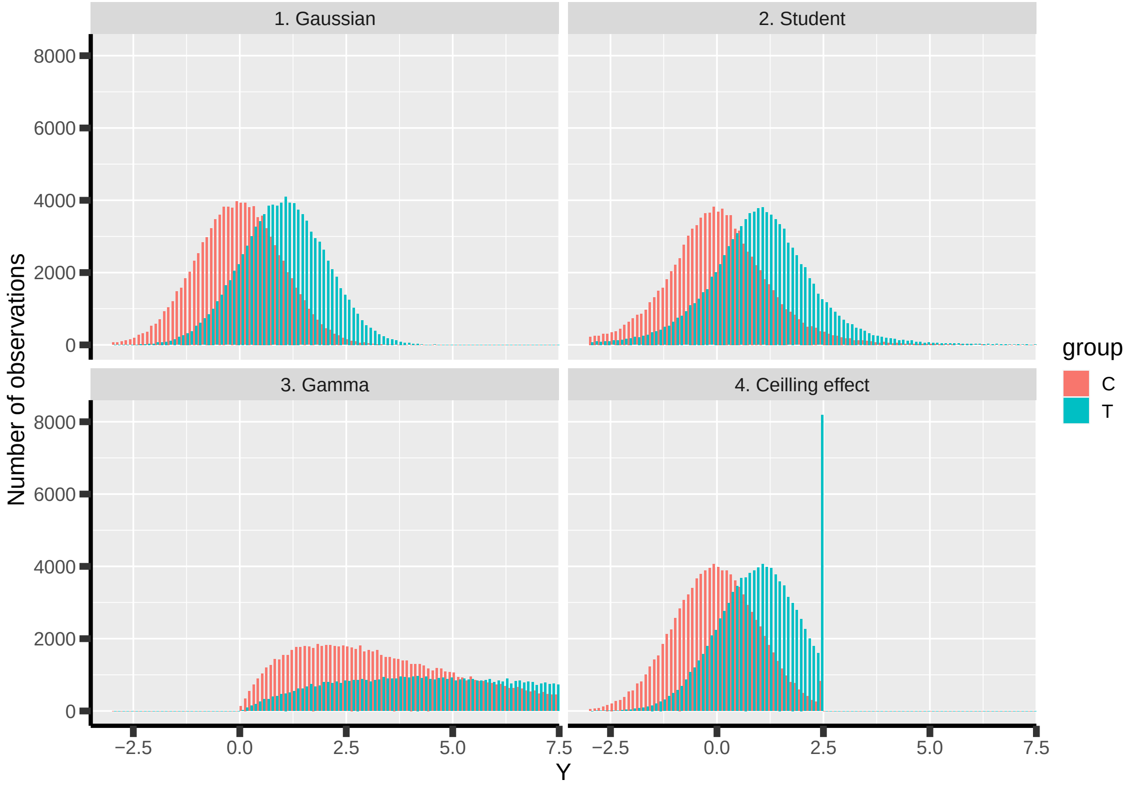


Figure 3: Example of simulated non-normal outcome.

3.1 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[2]. 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 4:

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,

[1] independent and identically distributed

[2] For some statistical tests, this requires to use robust instead of model-based standard error

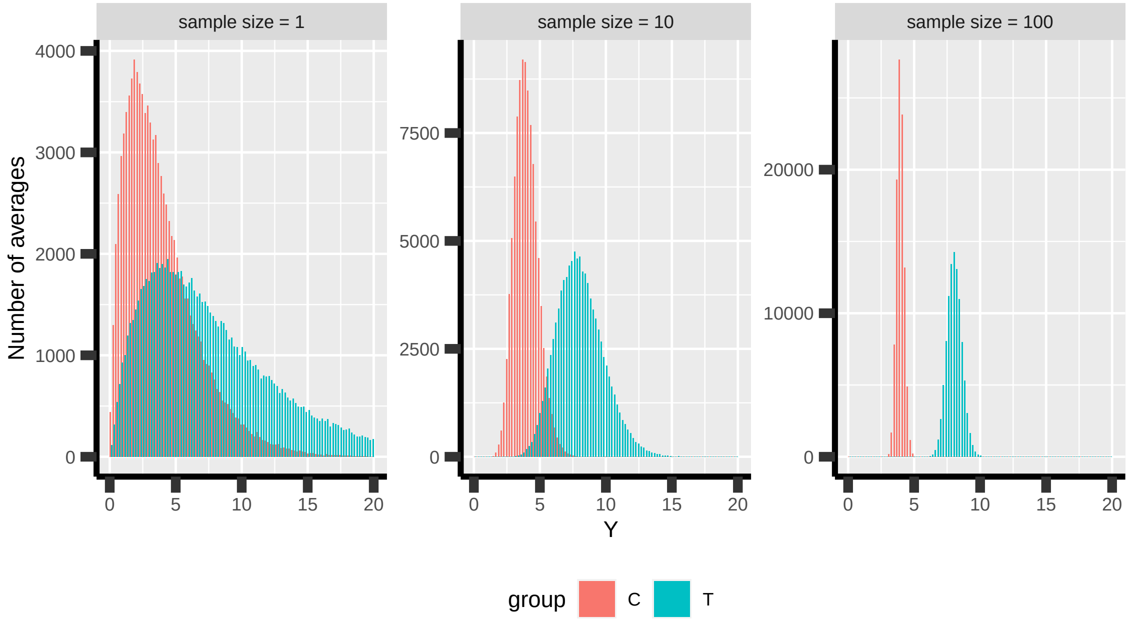


Figure 4: Distribution of the estimated mean along the sample size.

even though the normality assumption is not fullfilled (Figure 5) . . . for large enough sample sizes.

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.

In the following we will discuss issues 1 to 3.

```
      0 % ~calculating
+     1 % ~11s
+     2 % ~11s
++    3 % ~11s
++    4 % ~10s
+     5 % ~10s
+     6 % ~10s
++    7 % ~10s
++    8 % ~10s
╪     9 % ~10s
╪     10% ~10s
╪╪    11% ~10s
╪╪    12% ~10s
╪     13% ~10s
╪     14% ~10s
╪╪    15% ~09s
╪╪    16% ~09s
╪     17% ~09s
╪     18% ~09s
╪╪    19% ~09s
╪╪    20% ~09s
╪     21% ~09s
╪     22% ~09s
╪╪    23% ~09s
╪╪    24% ~09s
╪     25% ~08s
╪     26% ~08s
╪╪    27% ~08s
╪╪    28% ~08s
╪     29% ~08s
╪     30% ~08s
╪╪    31% ~08s
╪╪    32% ~08s
╪     33% ~08s
╪     34% ~08s
╪╪    35% ~07s
╪╪    36% ~07s
╪     37% ~07s
╪     38% ~07s
╪╪    39% ~07s
╪╪    40% ~07s
╪     41% ~07s
╪     42% ~07s
╪╪    43% ~06s
╪╪    44% ~06s
╪     45% ~06s
╪     46% ~06s
╪╪    47% ~06s
```

```
      0 % ~calculating
┼     1 % ~11s
┼     2 % ~12s
┼┼    3 % ~12s
┼┼    4 % ~12s
┼     5 % ~11s
┼     6 % ~11s
┼┼    7 % ~11s
┼┼    8 % ~11s
╪     9 % ~11s
╪     10% ~11s
╪╪    11% ~10s
╪╪    12% ~10s
╪     13% ~10s
╪     14% ~10s
╪╪    15% ~10s
╪╪    16% ~10s
╪     17% ~10s
╪     18% ~09s
╪╪    19% ~09s
╪╪    20% ~09s
╪     21% ~09s
╪     22% ~09s
╪╪    23% ~09s
╪╪    24% ~09s
╪     25% ~09s
╪     26% ~08s
╪╪    27% ~08s
╪╪    28% ~08s
╪     29% ~08s
╪     30% ~08s
╪╪    31% ~08s
╪╪    32% ~08s
╪     33% ~08s
╪     34% ~08s
╪╪    35% ~08s
╪╪    36% ~08s
╪     37% ~08s
╪     38% ~07s
╪╪    39% ~07s
╪╪    40% ~07s
╪     41% ~07s
╪     42% ~07s
╪╪    43% ~07s
╪╪    44% ~07s
╪     45% ~07s
╪     46% ~07s
╪╪    47% ~06s
```

```
    0 % ~calculating
+   1 % ~12s
+   2 % ~12s
++  3 % ~12s
++  4 % ~12s
+   5 % ~12s
+   6 % ~12s
++  7 % ~12s
++  8 % ~11s
╪   9 % ~11s
╪   10% ~11s
╪╪  11% ~11s
╪╪  12% ~11s
╪   13% ~11s
╪   14% ~11s
╪╪  15% ~11s
╪╪  16% ~10s
╪   17% ~10s
╪   18% ~10s
╪╪  19% ~10s
╪╪  20% ~10s
╪   21% ~10s
╪   22% ~10s
╪╪  23% ~10s
╪╪  24% ~10s
╪   25% ~09s
╪   26% ~09s
╪╪  27% ~09s
╪╪  28% ~09s
╪   29% ~09s
╪   30% ~09s
╪╪  31% ~09s
╪╪  32% ~08s
╪   33% ~08s
╪   34% ~08s
╪╪  35% ~08s
╪╪  36% ~08s
╪   37% ~08s
╪   38% ~08s
╪╪  39% ~08s
╪╪  40% ~08s
╪   41% ~07s
╪   42% ~07s
╪╪  43% ~07s
╪╪  44% ~07s
╪   45% ~07s
╪   46% ~07s
╪╪  47% ~07s
```

```
      0 % ~calculating
+     1 % ~12s
+     2 % ~13s
++    3 % ~12s
++    4 % ~13s
+     5 % ~13s
+     6 % ~12s
++    7 % ~12s
++    8 % ~12s
╪     9 % ~12s
╪     10% ~11s
╪╪    11% ~11s
╪╪    12% ~11s
╪     13% ~11s
╪     14% ~11s
╪╪    15% ~11s
╪╪    16% ~11s
╪     17% ~11s
╪     18% ~10s
╪╪    19% ~10s
╪╪    20% ~10s
╪     21% ~10s
╪     22% ~10s
╪╪    23% ~10s
╪╪    24% ~10s
╪     25% ~09s
╪     26% ~09s
╪╪    27% ~09s
╪╪    28% ~09s
╪     29% ~09s
╪     30% ~09s
╪╪    31% ~09s
╪╪    32% ~08s
╪     33% ~08s
╪     34% ~08s
╪╪    35% ~08s
╪╪    36% ~08s
╪     37% ~08s
╪     38% ~08s
╪╪    39% ~08s
╪╪    40% ~07s
╪     41% ~07s
╪     42% ~07s
╪╪    43% ~07s
╪╪    44% ~07s
╪     45% ~07s
╪     46% ~07s
╪╪    47% ~07s
```

```
     0 % ~calculating
+    1 % ~12s
+    2 % ~12s
++   3 % ~12s
++   4 % ~12s
+    5 % ~12s
+    6 % ~12s
++   7 % ~12s
++   8 % ~12s
ǂ    9 % ~18s
ǂ    10% ~23s
ǂǂ   11% ~27s
ǂǂ   12% ~30s
ǂ    13% ~32s
ǂ    14% ~34s
ǂǂ   15% ~36s
ǂǂ   16% ~37s
ǂ    17% ~38s
ǂ    18% ~39s
ǂǂ   19% ~40s
ǂǂ   20% ~40s
ǂ    21% ~41s
ǂ    22% ~41s
ǂǂ   23% ~41s
ǂǂ   24% ~42s
ǂ    25% ~42s
ǂ    26% ~41s
ǂǂ   27% ~41s
ǂǂ   28% ~41s
ǂ    29% ~41s
ǂ    30% ~41s
ǂǂ   31% ~41s
ǂǂ   32% ~41s
ǂ    33% ~41s
ǂ    34% ~41s
ǂǂ   35% ~40s
ǂǂ   36% ~40s
ǂ    37% ~39s
ǂ    38% ~37s
ǂǂ   39% ~36s
ǂǂ   40% ~35s
ǂ    41% ~33s
ǂ    42% ~32s
ǂǂ   43% ~31s
ǂǂ   44% ~30s
ǂ    45% ~29s
ǂ    46% ~28s
ǂǂ   47% ~27s
```

```
     0 % ~calculating
+    1 % ~12s
+    2 % ~12s
++   3 % ~13s
++   4 % ~12s
+    5 % ~12s
+    6 % ~12s
++   7 % ~12s
++   8 % ~12s
╪    9 % ~12s
╪    10% ~12s
╪╪   11% ~11s
╪╪   12% ~11s
╪    13% ~11s
╪    14% ~11s
╪╪   15% ~11s
╪╪   16% ~11s
╪    17% ~11s
╪    18% ~10s
╪╪   19% ~10s
╪╪   20% ~10s
╪    21% ~10s
╪    22% ~10s
╪╪   23% ~10s
╪╪   24% ~10s
╪    25% ~10s
╪    26% ~10s
╪╪   27% ~10s
╪╪   28% ~10s
╪    29% ~10s
╪    30% ~10s
╪╪   31% ~10s
╪╪   32% ~10s
╪    33% ~09s
╪    34% ~09s
╪╪   35% ~09s
╪╪   36% ~09s
╪    37% ~09s
╪    38% ~09s
╪╪   39% ~09s
╪╪   40% ~09s
╪    41% ~09s
╪    42% ~08s
╪╪   43% ~08s
╪╪   44% ~08s
╪    45% ~08s
╪    46% ~08s
╪╪   47% ~08s
```

```
      0 % ~calculating
+     1 % ~12s
+     2 % ~12s
++    3 % ~12s
++    4 % ~12s
+     5 % ~12s
+     6 % ~12s
++    7 % ~12s
++    8 % ~12s
╪     9 % ~11s
╪     10% ~11s
╪╪    11% ~11s
╪╪    12% ~11s
╪     13% ~11s
╪     14% ~11s
╪╪    15% ~11s
╪╪    16% ~11s
╪     17% ~11s
╪     18% ~10s
╪╪    19% ~10s
╪╪    20% ~10s
╪     21% ~10s
╪     22% ~10s
╪╪    23% ~10s
╪╪    24% ~10s
╪     25% ~10s
╪     26% ~09s
╪╪    27% ~09s
╪╪    28% ~09s
╪     29% ~09s
╪     30% ~09s
╪╪    31% ~09s
╪╪    32% ~09s
╪     33% ~09s
╪     34% ~09s
╪╪    35% ~09s
╪╪    36% ~09s
╪     37% ~08s
╪     38% ~08s
╪╪    39% ~08s
╪╪    40% ~08s
╪     41% ~08s
╪     42% ~08s
╪╪    43% ~08s
╪╪    44% ~08s
╪     45% ~08s
╪     46% ~08s
╪╪    47% ~07s
```

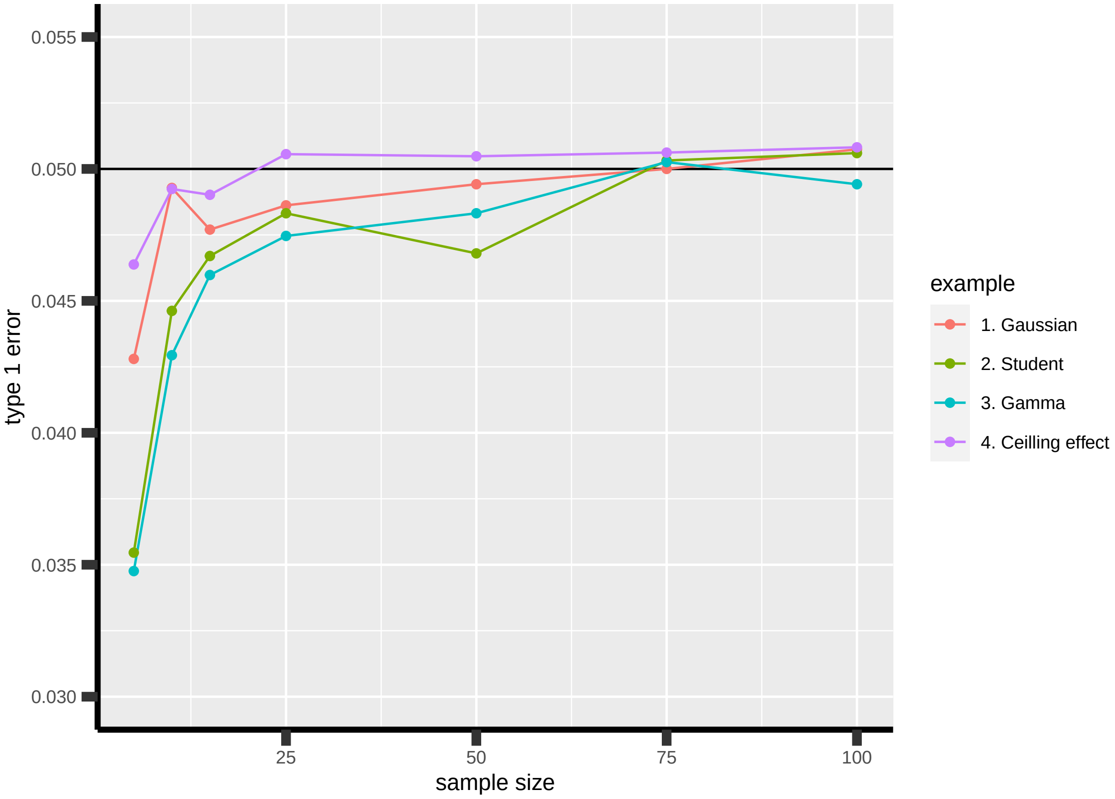


Figure 5: Coverage of the t-test.

Note: not assuming normality complicates the understanding the group effect: the normal distribution is one of the few distribution that can be summarized by two, easily interpretable, independent, parameters (mean and variance).

3.2 Issue 1: parameter of interest

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 (also refered to as '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 6, we can see that the mean is not the most likely value (i.e. the mode). The median is slightly closer to the mode but does not really provide a satisfactory improvement.

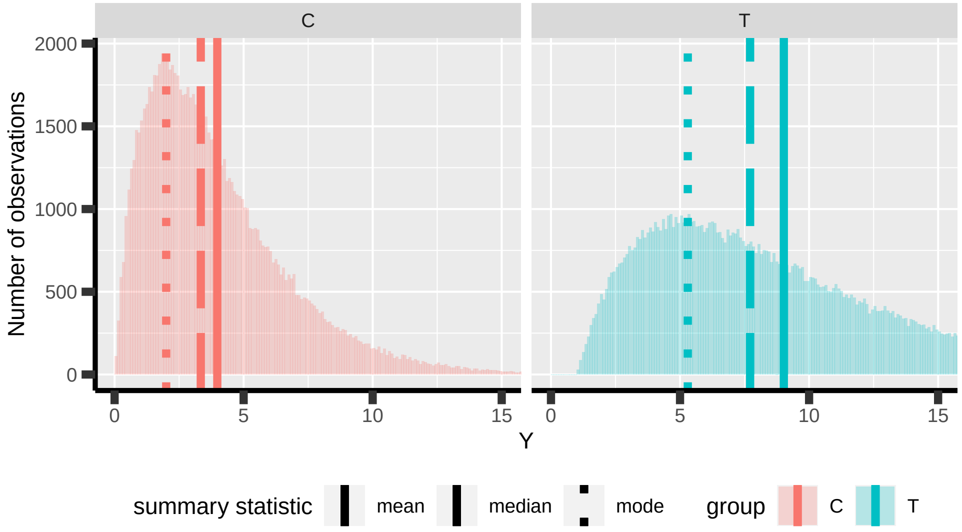


Figure 6: Mean, median, and mode two asymetric distributions.

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 7), 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 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). Technically this requires additional assumptions, such as homoschedasticity, that are not discussed here.

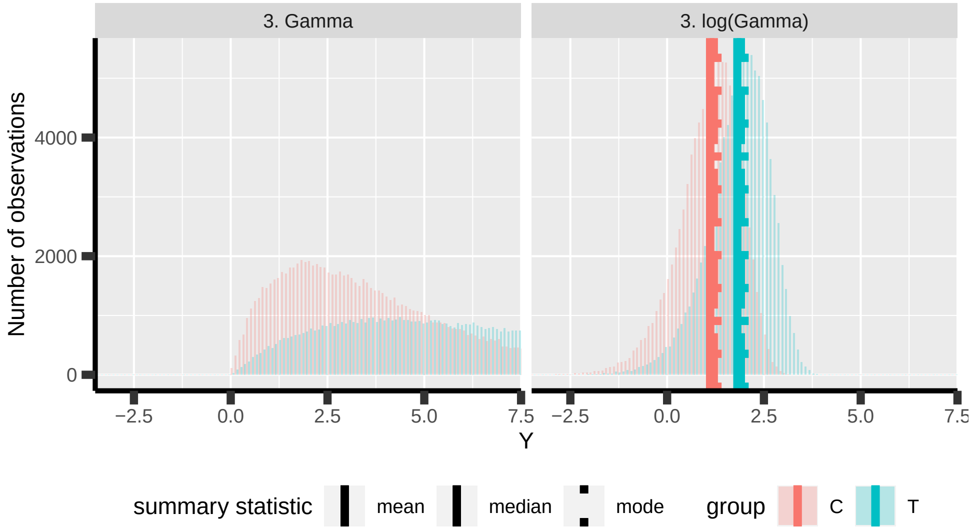


Figure 7: Mean, median, and mode on the log-transformed data

There are other possible parameter of interest, e.g.:

- The **Mann-Whitney parameter** $\mathbb{P}\left[X \geq Y\right] + \frac{1}{2}\mathbb{P}\left[X = Y\right]$: this is the probability that a randomly chosen individual from the active group has a larger value than a randomly individual from the control group.

  → it is closely related to the Wilcoxon-Mann-Whitney test and the AUC

  → not (completely) straightforward causal interpretation (Fay et al., 2018).

  → implementation: see the function `wmwTest` from the asht package
  ⚠ in presence of heteroschedasticity (variance that differs between groups) one should use another tool (see the BuyseTest package)

- One could dichotomize the outcome to **compare the probability of a high outcome value** between the two groups. This can be relevant in presence of a important ceiling effect.

  → implementation: see the function `uncondExact2x2` from the exact2x2 package for comparing proportions.

3.3 Issue 2: handling small samples

In small samples, traditional methods will not provide a very accurate type 1 error control or coverage as illustrated in Figure 5.

- **permutation methods** can be used to obtain exact type 1 error control under exchangeability. Exchangeability is violated when testing a mean difference between the groups while there is a difference in variance. In such a case studentized permutation should be used instead (Chung and Romano, 2016).
→ this will produce valid p-values but no confidence intervals

- **bootstrap resampling methods** can be used to reduce the coverage error error in small samples. This includes studentized non-parametric bootstrap where the bootstrap test statistic is used to estimate the quantiles used in the confidence intervals (instead +/- 1.96) or bias-corrected and accelerated (BCa) bootstrap interval (see the boot package).
⚠ Not all bootstrap methods have good sample properties, e.g. the 'standard' non-parametric bootstrap using the quantiles of the boostrap distribution of the parameter of interest does not have very attractive small sample properties.

There are also analytic correction for improving the small sample properties but there typically are specific to a statistical model/test and are not discussed here.

3.4 Issue 3: handling outliers

Most of the statistics will quantify some kind of average difference between groups. One observation with a very large value may have large influence on this average. If that is a concern, rank-based statistics (e.g. median, Mann-Whitney parameter, probability of a high-value) may be seen as more fair statistics in the sense that all observations have the same weight on the summary statistic.

⚠ Artificially reducing the outcome value (e.g. to be at most the mean plus 2 standard deviation) is generally a bad idea: it will induce a downward bias in the estimated mean and can lead to inflated type 1 error (if the probability of a large value is group dependent).

4 References

Chung, E. and Romano, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference*, 168:97–105.

Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., and Gabriel, E. E. (2018). Causal estimands and confidence intervals associated with wilcoxon-mann-whitney tests in randomized experiments. *Statistics in Medicine*, 37(20):2923–2937.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.