Statistics and electronics - lecture 1

The ASTA team

Contents

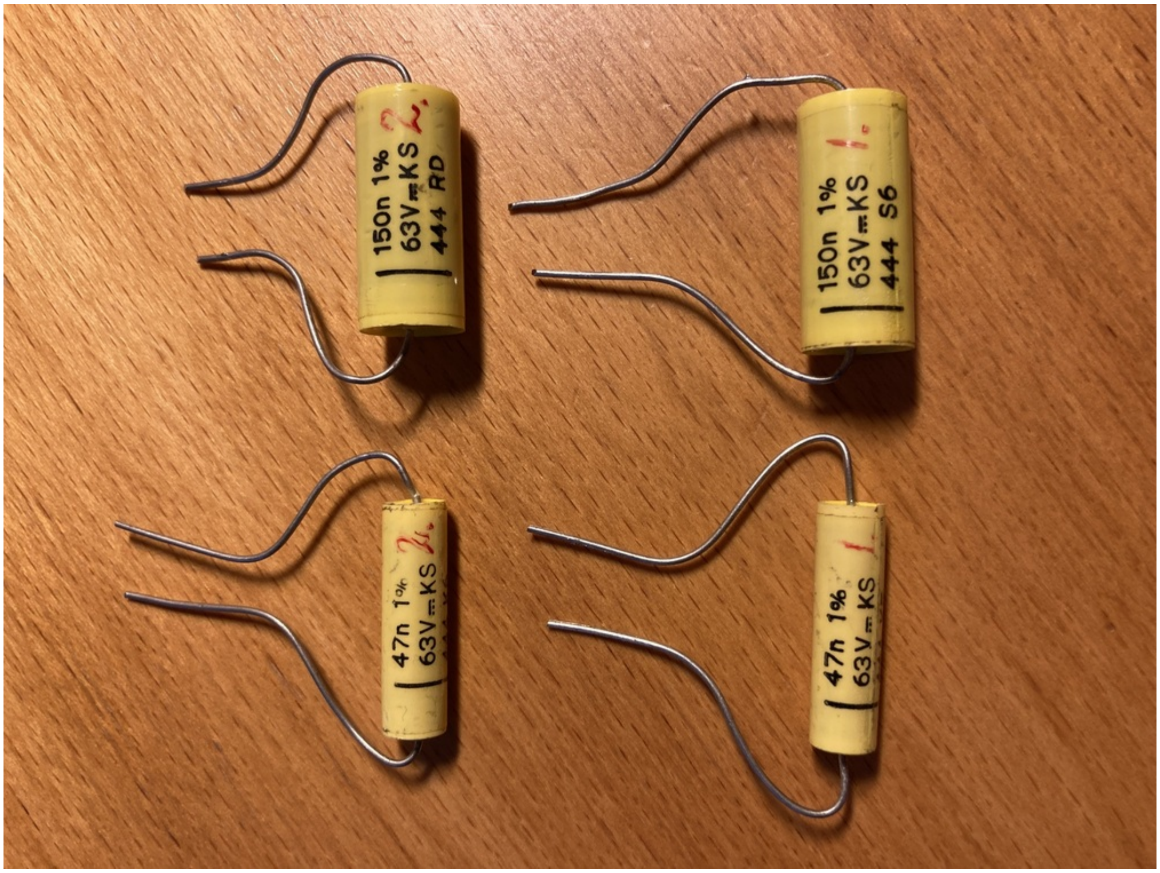
0.1 Sources of variation

Capacitors come with a nominal value for the capacitance.

- When capacitance is measured, we do not get exactly the nominal value.

We shall study 2 sources of variation:

- measurement variation due to random errors on a measuring device
- component variation due to random errors in the production process



0.2 Data from Peter Koch

Peter has done 100 independent measurements of the capacitance of each 4 of the displayed capacitors and one additional.

- Nominal values are 47, 47, 100, 150, 150 nF.
- All have a stated tolerance of 1%.

```
load(url("https://asta.math.aau.dk/datasets?file=cap_1pct.RData"))
head(capDat, 4)
```

```
##   capacity nomval   sample
## 1    45.69     47 s_1_nF47
## 2    45.71     47 s_1_nF47
## 3    45.69     47 s_1_nF47
## 4    45.71     47 s_1_nF47
```

Here we see the first 4 measurements of the first capacitor with nominal value 47nF.

- Remark: The measured values are consistently below the nominal value minus the 1% tolerance: $47 - 0.47 = 46.53$.

```
table(capDat$sample)
```

```
## 
##   s_1_nF47  s_2_nF47 s_3_nF100 s_4_nF150 s_5_nF150 
##        100       100       100       100       100
```

0.3 Relative errors

- Instead of considering the raw errors

$$\text{measuredValue - nominalValue},$$

 we will consider the relative error

$$\frac{\text{measuredValue - nominalValue}}{\text{nominalValue}}.$$

- A tolerance of 0.01 means that the relative error should be within ± 0.01.

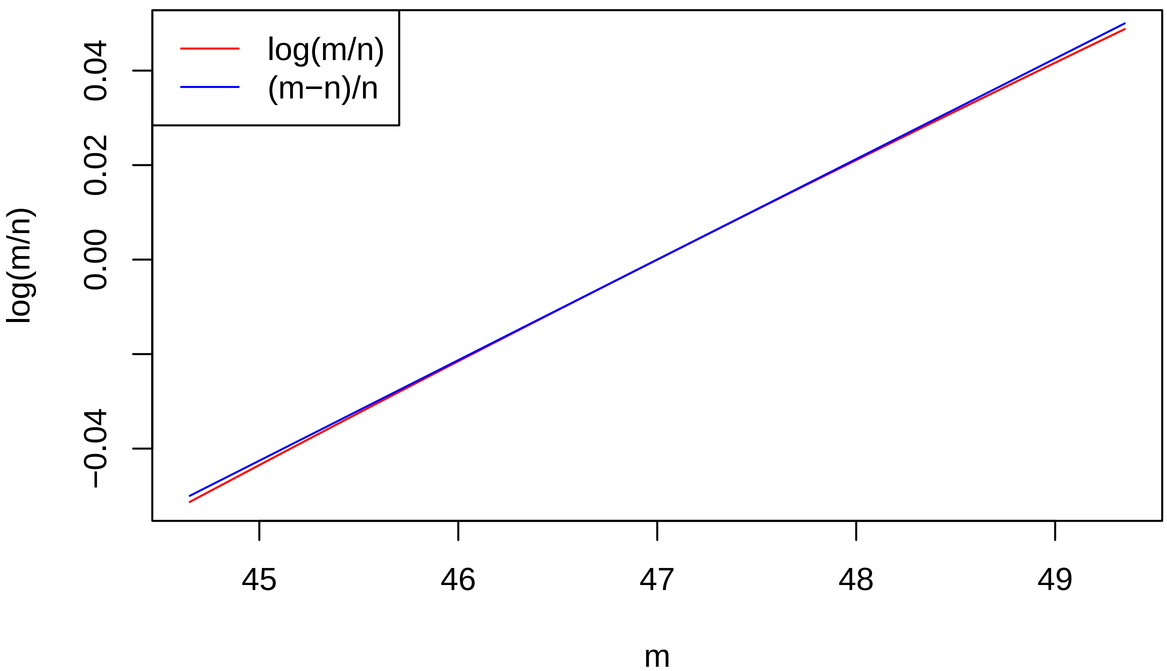
0.4 Approximation of the relative error

- Instead of looking at the relative error, we may look at the following approximation:

$$\text{lnError} = \ln\left(\frac{\text{measuredValue}}{\text{nominalValue}}\right) \approx \frac{\text{measuredValue} - \text{nominalValue}}{\text{nominalValue}}$$

- This is illustrated below with a nominal value of $n = 47$ and measured values of 47 plus/minus 5%.

```
n <- 47
m <- seq(47-5*0.01*47, 47+5*0.01*47, length.out = 100)
plot(m, log(m/n), col = "red", type = "l")
lines(m, (m - n)/n, col = "blue", type = "l")
legend("topleft", legend = c("log(m/n)", "(m-n)/n"), lty = 1, col = c("red", "blue"))
```



0.5 Transformation of errors

- The approximation can be justified theoretically.
- Recall the linear approximation of a function:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

- If we take

$$x_0 = 1 \tag{1}$$
$$f(x) = \ln x \tag{2}$$
$$f'(x) = 1/x, \tag{3}$$

we get

$$\ln(x) \approx \ln(x_0) + \frac{1}{x_0} \cdot (x - x_0) = x - 1.$$

- Suppose $x = m/n$. Then

$$\ln\left(\frac{m}{n}\right) \approx \frac{m}{n} - 1 = \frac{m-n}{n}$$

0.6 Transformed data

- We construct an extra `lnError` variable in the `capDat` dataset.

```
capDat = within(capDat, lnError <- log(capacity/nomval))
head(capDat, 2)
```

```
##   capacity nomval   sample     lnError
## 1    45.69     47 s_1_nF47 -0.02826815
## 2    45.71     47 s_1_nF47 -0.02783051
```

```
tail(capDat, 2)
```

```
##     capacity nomval    sample     lnError
## 499    145.7    150 s_5_nF150 -0.02908558
## 500    145.6    150 s_5_nF150 -0.02977216
```

- The resolution on Peters capacitance meter is with 1-2 decimal(s) in the 47/150 nF range, which means that only a limited number of different values(3-18) are observed for each capacitor. This means that box-plots and histograms are non-informative.

0.7 Model considerations

- Let us have a look at a summary of the data:

```
favstats(lnError~sample, data=capDat)
```

```
##      sample         min          Q1      median          Q3         max
## 1  s_1_nF47 -0.02958221 -0.02832287 -0.02804930 -0.02804930 -0.02783051
## 2  s_2_nF47 -0.02914399 -0.02783051 -0.02761176 -0.02761176 -0.02717441
## 3 s_3_nF100 -0.03521276 -0.03399638 -0.03386707 -0.03366020 -0.03334998
## 4 s_4_nF150 -0.02565975 -0.02446352 -0.02429269 -0.02429269 -0.02360987
## 5 s_5_nF150 -0.03045921 -0.02977216 -0.02908558 -0.02908558 -0.02908558
##          mean           sd   n missing
## 1 -0.02832518 0.0005062160 100       0
## 2 -0.02786346 0.0005171088 100       0
## 3 -0.03398306 0.0005057586 100       0
## 4 -0.02453879 0.0005870180 100       0
## 5 -0.02947702 0.0005543930 100       0
```

- All measurements are more than 2.3% below the nominal value.
- This must be due to a systematic error on the meter.

0.8 Sources of variation

- We now have three sources of error:
  - Systematic errors of the measurement device
  - Production errors in the individual capacitors
  - Random measurement errors
- This leads us to consider the model

$$\ln\left(\frac{\text{measuredValue}}{\text{nominalValue}}\right) = \text{systematicError} + \text{productionError} + \text{measurementError}.$$

0.9 Statistical model

- We have the model:

$$\ln\left(\frac{\text{measuredValue}}{\text{nominalValue}}\right) = \text{systematicError} + \text{productionError} + \text{measurementError}$$

- We may write the model mathematically as

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}$$

  where

  - Y_{ij} is the log error measurement (jth measurement from the ith capacitor)
  - $i = 1, \dots, k$ is the number of the capacitor
  - $j = 1, \dots, n$ is the number of the observation for that capacitor
  - $k = 5$ is the total number of capacitors
  - $n = 100$ is the number of repetitions for each capacitor
  - μ is the systematic error on the meter
  - A_i is the random production error
  - ε_{ij} is the random measurement error

- We make the following assumptions:

  - The production error A_i is normally distributed with mean 0 and variance σ_α^2,
  - The measurement error ε_{ij} is normally distributed with mean 0 and variance σ^2.

- This is called a **random effects model**, see [WMMY] Chapter 13.11.

0.10 Estimation of systematic error

- The systematic error is simply estimated by the sample mean

$$\hat{\mu} = \bar{y}_{..}$$

  - The two dots indicate that we take the average over all observations from all capacitors.

```
muhat <- mean(capDat$lnError)
muhat
```

```
## [1] -0.0288375
```

- The meter systematically reports a value, which is estimated to be 2.88% too low.

0.11 Estimation of random error

- We now try to estimate the variance σ_α^2 of the production error and the variance of the random measurement error σ^2.

- We need two types of sum of squares:

- SSA (*sum of squares between groups*) measures how much the sample means for the individual capacitors $\bar{y}_{i.}$ deviate from the total sample mean $\bar{y}_{..}$

$$SSA = n \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

- SSE (*sum of squares within groups*) measures how much the individual measurements deviate from the sample mean of the capacitor they were measured on:

$$SSE = \sum_{ij} (y_{ij} - \bar{y}_{i.})^2$$

- Intuitively, SSA is closely related to the variance of the production error σ_α^2, while SSE is closely related to the variance of the random measurement error σ^2.

0.12 Fit

- The sum of squares may be found from:

```
fit <- lm(lnError ~ sample, data = capDat)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: lnError
##            Df    Sum Sq    Mean Sq F value    Pr(>F)
## sample      4 0.0046576 0.00116440  4067.4 < 2.2e-16 ***
## Residuals 495 0.0001417 0.00000029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can extract the sum of squares as follows

```
SS <- anova(fit)$`Sum Sq`
SSA <- SS[1]
SSE <- SS[2]
SSA
```

```
## [1] 0.004657588
```

```
SSE
```

```
## [1] 0.0001417076
```

0.13 Solution

- One may show (see [WMMY] Theorem 13.4):

$$E(SSA) = (k-1)\sigma^2 + n(k-1)\sigma_\alpha^2$$
$$E(SSE) = k(n-1)\sigma^2$$

- Using the approximations

$$E(SSA) \approx SSA, \qquad E(SSE) \approx SSE$$

we obtain the estimates

$$\hat{\sigma}^2 = \frac{1}{(n-1)k} SSE = \frac{1}{99 \cdot 5} \cdot 0.0001417 = 2.86 \cdot 10^{-7}$$

$$\hat{\sigma}_\alpha^2 = \frac{1}{n(k-1)} SSA - \frac{\hat{\sigma}^2}{n} = \frac{1}{100 \cdot (5-1)} \cdot 0.0046576 - \frac{2.86 \cdot 10^{-7}}{100} = 1.16 \cdot 10^{-5}$$

0.14 Summing up

- The meter has an estimated systematic error of $\hat{\mu} = -2.88\%$.
- The estimated standard deviation of the meter is $\hat{\sigma} = \sqrt{2.86 \cdot 10^{-7}} = 0.0534\%$.
- The estimated standard deviation of the production error is $\hat{\sigma}_\alpha = \sqrt{1.16 \cdot 10^{-5}} = 0.341\%$.
- Since 99.7% (practically all) of all observations fall within $\pm 3 \cdot \sigma_\alpha$ from 0, we have that the production error falls within

$$\pm 3 \cdot 0.341\% = 1.02\%$$

  of the nominal value, which is in accordance with the tolerance of 1%.
- The total estimated variance of the log error is

$$\hat{\sigma}_\alpha^2 + \hat{\sigma}^2 = 1.16 \cdot 10^{-5} + 2.86 \cdot 10^{-7} = 1.19 \cdot 10^{-5}.$$

  - The variance is clearly dominated by the production error.
- Note that especially the estimate $\hat{\sigma}_\alpha$ is quite uncertain, since we only have measurements from 5 capacitors.

0.15 Test of no random effect

- We have the possibility of testing the hypothesis

$$H_0 : \sigma_\alpha = 0.$$

- The formulas for $E(SSA)$ and $E(SSE)$ were

$$E(SSA) = (k-1)\sigma^2 + n(k-1)\sigma_\alpha^2$$
$$E(SSE) = k(n-1)\sigma^2.$$

- Under H_0, this is means that

$$\frac{1}{k-1} E(SSA) = \frac{1}{k(n-1)} E(SSE) = \sigma^2.$$

- Under H_0, the F statistic

$$F_{obs} = \frac{\frac{SSA}{k-1}}{\frac{SSE}{k(n-1)}}$$

  has an F-distribution with degrees of freedom $df_1 = k - 1$ and $df_2 = k(n-1)$.
  - Large values are critical for the null-hypothesis.
- In the capacitor dataset $F_{obs} = 4067.4$, which is highly significant (p-value close to 0).
  - Our capacitors do have some production errors.

0.16 Coefficient of variation

- Let X be a random variable with mean μ and standard deviation σ.
- If we are interested in relative variation, it is common to look at the **coefficient of variation**

$$CV(X) = \frac{\sigma}{\mu}$$

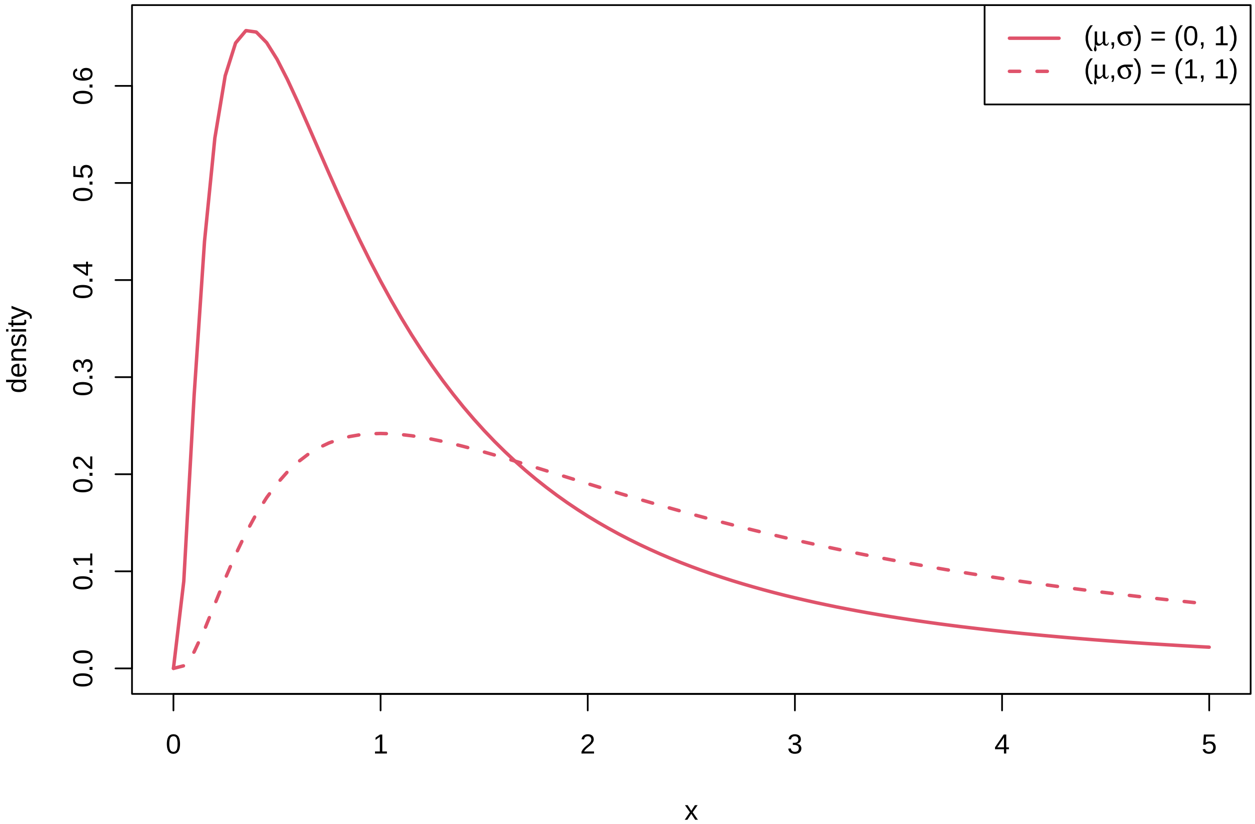
  - Standard deviation relative to the mean
  - Unit-free
- If X is normal, then 95% of our measurements are within

$$\mu \pm 2 \cdot \sigma = \mu \pm 2 \cdot \mu \cdot CV(X) = \mu(1 \pm 2 \cdot CV(X)).$$

- If e.g. $CV(X) = 0.05$, it means that 95% of all observations are within $2 \cdot 0.05 = 10\%$ of the mean.

0.17 The lognormal distribution

- In the preceeding analysis, we assumed that the log-transformed errors had a normal distribution.
- Let X be a random variable and $Y = \ln(X)$.
- We say that X has a **lognormal distribution** if Y has a normal distribution with - say - mean μ and standard deviation σ.
- Here are some plots of the density of the lognormal distribution:



0.18 Coefficient of variation for lognormal distribution

- Suppose X has a log-normal distribution, so that $Y = \ln(X)$ has a normal distribution with mean μ and standard deviation σ.
- Then the mean and variance are given by (Theorem 6.7 of [WMMY]):

$$E(X) = \exp(\mu + \sigma^2/2)$$

$$Var(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

- The coefficient of variation is then

$$CV(X) = \frac{\sqrt{Var(X)}}{E(X)} = \frac{\sqrt{\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)}}{\exp(\mu + \sigma^2/2)} = \sqrt{\exp(\sigma^2) - 1}$$

- In Peter's data we estimated the variance of the ln error to $\hat{\sigma}_\alpha^2 = 1.16 \cdot 10^{-5}$, which means that the estimated CV of the capacity measurement is

$$\widehat{CV}(X) = \sqrt{\exp\left(1.16 \cdot 10^{-5}\right) - 1} = 0.341\%.$$

0.19 Linear calibration

- In our previous analysis, we assumed, that the systematic error on the meter did not depend on nominal value.

$$\ln\left(\frac{\text{measuredValue}}{\text{nominalValue}}\right) = \text{meterError} + \text{randomError}$$

- To check this assumption consider the linear model

$$\ln(\text{measuredValue}) = \alpha + \beta \cdot \ln(\text{nominalValue}) + \varepsilon.$$

- Note that the previously considered model corresponds to $\beta = 1$.

0.20 Linear calibration fit

- We fit the linear model:

```
fit <- lm(log(capacity) ~ log(nomval), data = capDat)
summary(fit)
```

```
##
## Call:
## lm(formula = log(capacity) ~ log(nomval), data = capDat)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0064121 -0.0010784  0.0007315  0.0013879  0.0050839
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0300145  0.0011907  -25.21   <2e-16 ***
## log(nomval)  1.0002636  0.0002648 3776.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003101 on 498 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.426e+07 on 1 and 498 DF,  p-value: < 2.2e-16
```

- The slope looks close to 1.
- We may test the null-hypothesis $H_0 : \beta = 1$.

$$t_{obs} = \frac{1.0002636 - 1}{0.0002648} = 0.995.$$

This yields a p-value of around 32%.

  - It is a bit dubious to model a linear relationship with only 3 nominal values.
  - Also note that we have correlated measurements, since several measurements are made on the same capacitors.

0.21 Calibrated values

- If we stick to the linear calibration model, it is sensible to correct our measured errors according to the calibration of the meter.
- We have the model:

$$\text{measuredValue} = \alpha + \beta * \text{nominalValue}$$

- We compute the calibrated values

$$\text{calibratedValue} = (\text{measuredValue} - \alpha)/\beta$$

- We estimate the coefficients α and β and calibrate the measurements.

```
ab = coef(fit)
ab
```

```
## (Intercept) log(nomval)
## -0.03001454  1.00026359
```

```
capDat$lnError_c = (capDat$lnError - ab[1])/ab[2]
head(capDat)
```

```
##   capacity nomval   sample     lnError   lnError_c
## 1    45.69     47 s_1_nF47 -0.02826815 0.001745930
## 2    45.71     47 s_1_nF47 -0.02783051 0.002183452
## 3    45.69     47 s_1_nF47 -0.02826815 0.001745930
## 4    45.71     47 s_1_nF47 -0.02783051 0.002183452
## 5    45.70     47 s_1_nF47 -0.02804930 0.001964715
## 6    45.69     47 s_1_nF47 -0.02826815 0.001745930
```