Statistics and electronics - lecture 2

The ASTA team

Contents

1 Lot variation

- Picture of a “lot” of capacitors.
- The word lot is used to identify several components produced in a single run.
  - A run is a production series limited to a given time interval and fixed production parameters.
- We expect components from the same lot to be more similar.
- Peter Koch has tested 269 of the capacitors in the displayed lot (one measurement for each).

```
Cap220=read.csv(url("https://asta.math.aau.dk/datasets?file=capacitor_lot_220_nF.txt"))[,1]
summary(Cap220)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   197.2   204.8   207.9   207.9   210.9   218.6
```

2 Testing for log normality

2.1 Log normality

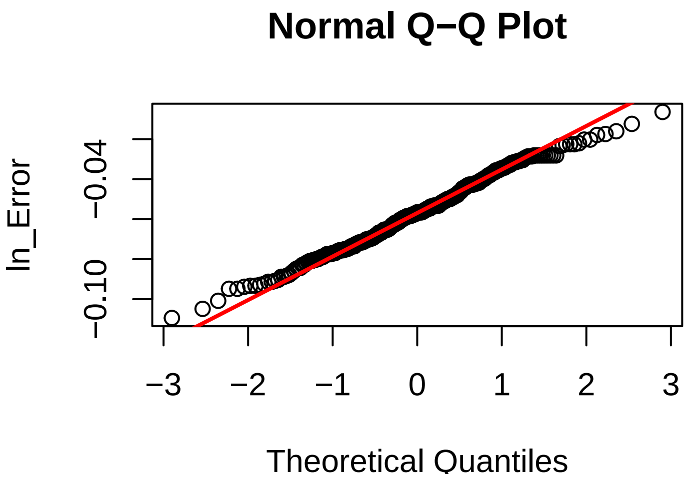
- Last time we assumed log normality of the relative measurements:

$$\ln\left(\frac{\text{measuredValue}}{\text{nominalValue}}\right) \sim \text{norm}(\mu, \sigma).$$

- The data we considered last time did not allow us check this assumtion.

- We have seen that normality can be checked with a qqplot (lecture 1.3, [WMMY] Sec. 8.8).

```
Cap220=read.csv(url("https://asta.math.aau.dk/datasets?file=capacitor_lot_220_nF.txt"))[,1]
ln_Error=log(Cap220/220)
qqnorm(ln_Error,ylab="ln_Error")
qqline(ln_Error,lwd=2,col="red")
```



- The qq-plot supports normality of `ln_Error`.

2.2 Testing normality

- One can also make a test of the null-hypothesis

$$H_0: \text{ the population has a normal distribution.}$$

- There are several tests of normality.
- Two of these are considered in [WMMY] Section 10.11:
  - Gearys test
  - goodness of fit

2.3 Gearys test

- Consider a sample $X_1, \dots, X_n$ from a population.
- We may estimate of the standard deviation σ of the population:

$$S_0 = \sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}$$

  - S_0 is *always* a good estimator of the population standard deviation σ - no matter the form of the population distribution.
- Next consider

$$S_1 = \sqrt{\frac{\pi}{2}}\sum_i |X_i - \bar{X}|/n$$

  - This is a good estimator of σ, **if** the population is normal.
  - Otherwise, it will over- or underestimate σ depending on the form of the population distribution.

2.4 Gearys test

- If the population distribution is normal, we expect that

$$U = \frac{S_1}{S_0}$$

is close to 1.

- Under the null-hypothesis,

$$Z = \frac{\sqrt{n}(U-1)}{0.2661}$$

is approximately standard normally distributed when n is large.

- That is, with a significance level of 5%, we reject the null-hypothesis if $|z_{obs}| > 1.96$.
- We can do all the computations in R.

```
mln_E=mean(ln_Error)
s1=sqrt(mean((ln_Error-mln_E)^2))
s0=sqrt(pi/2)*mean(abs(ln_Error-mln_E))
u=s1/s0
z_obs=sqrt(length(ln_Error))*(u-1)/0.2661
z_obs
```

```
## [1] -1.383383
```

- We do not reject the null-hypothesis.
- Hence there is no evidence of non-normality.

2.5 Goodness of fit - die example

- Goodness of fit is a general method for investigating whether a sample comes from a specific distribution.
- Before considering test for normality, we consider a simpler example (see [WMMY] Sec. 10.11).
- Suppose we roll a die. We have the null-hypothesis that the die is fair, i.e. the probabilities of the outcomes $(1, 2, 3, 4, 5, 6)$ are

$$(1/6, 1/6, 1/6, 1/6, 1/6, 1/6).$$

- Rolling the die 120 times, we expect the frequencies

$$(20, 20, 20, 20, 20, 20)$$

- Actually we observe the frequencies

$$(20, 22, 17, 18, 19, 24)$$

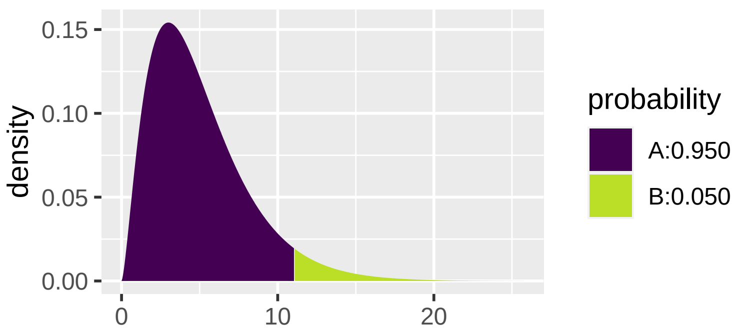
- The distance between observed and expected frequencies is measured by

$$X^2 = \sum \frac{(\text{ObservedFrequencies - ExpectedFrequencies})^2}{\text{ExpectedFrequencies}}$$

2.6 Goodness of fit - die example

- **If** the null-hypothesis is true (the die is fair), then
  - X^2 has a chi-square distribution (Lecture 1.4, [WMMY] Chapter 6.7) with df=k-1=5 degrees of freedom, where $k = 6$ is the number of possible outcomes.
  - large values of X^2 are critical for the null-hypothesis.
- For the example on the previous slide:
  - $x^2_{obs} = 1.7$

```
critical_value <- qdist("chisq", .95, df = 5)
```




```
critical_value
```

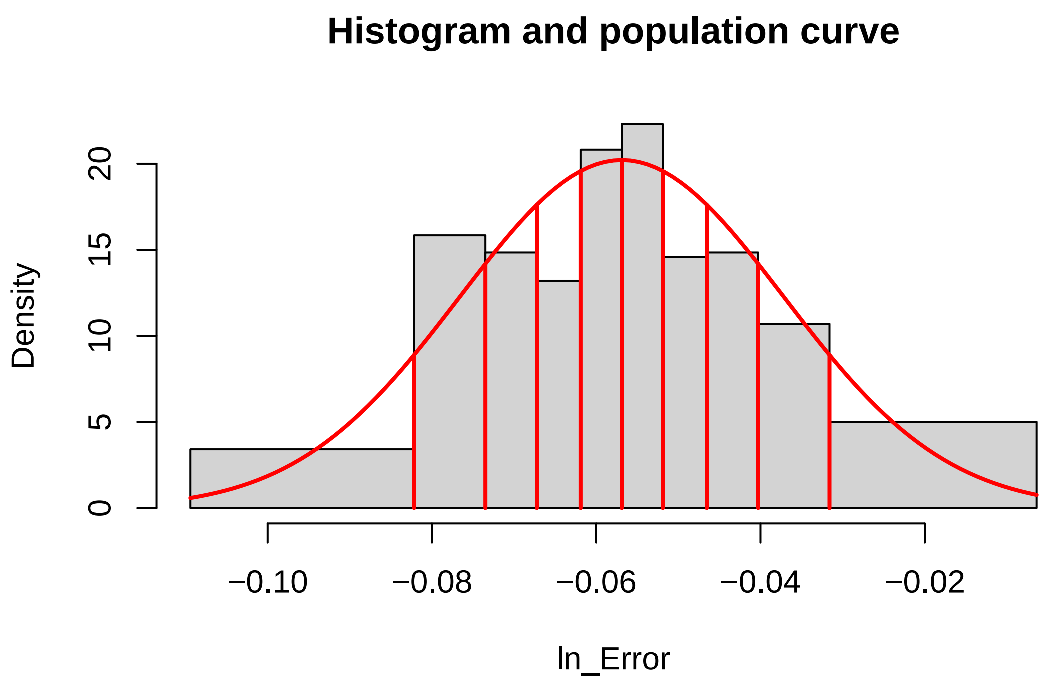
```
## [1] 11.0705
```

- At 5% significance level the critical value is 11.07, so there is no evidence against the null-hypothesis of a fair die.

2.7 Goodness of fit - normal distribution

- We assume that `ln_Error` is a sample from a normal distribution.
- We estimate its mean and standard deviation by the sample mean and sample standard deviation
- We divide the population distribution into 10 bins with equal probabilities p=10%.
  - The number of bins could be changed.
  - The bins should be so large, that the expected frequencies in each is at least 5.

```
m <- mean(ln_Error)
s <- sd(ln_Error)
breaks <- qnorm((0:10)/10, m, s)
```



- Area in each bin of the red population curve is 0.1
- As the sample size is 269 we obtain that the expected frequency is $269 * 0.1 = 26.9$ in each bin.
  - This is clearly above 5

2.8 Goodness of fit - normal distribution

- Observed frequecies:

```
observed <- table(cut(ln_Error, breaks))
names(observed) <- paste("bin", 1:10, sep = "")
observed
```

```
##  bin1  bin2  bin3  bin4  bin5  bin6  bin7  bin8  bin9 bin10
##    25    37    25    19    28    30    21    25    25    34
```

- We compute the X^2 statistic:

```
chisq_obs <- sum((observed-26.9)^2)/26.9
chisq_obs
```

```
## [1] 10.21933
```

- The degrees of freedom is the number of bins minus 3 (number of parameters + 1), i.e. df = 10-3 = 7.

2.9 Goodness of fit - normal distribution

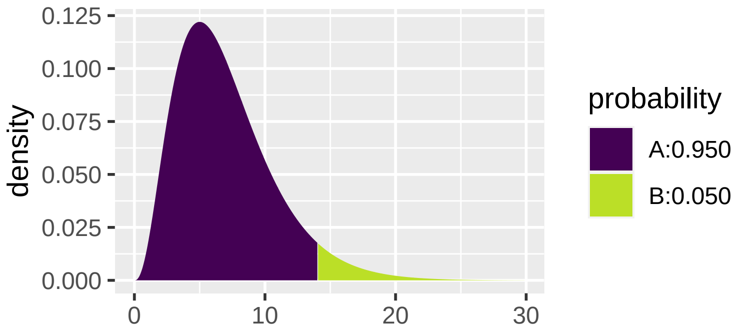
- We had computed the value of X^2

```
chisq_obs
```

```
## [1] 10.21933
```

- We find the critical value

```
critical_value <- qdist("chisq", .95, df = 7)
```




```
critical_value
```

```
## [1] 14.06714
```

- Since X^2 is smaller than the critical value, we do not reject the null-hypothesis
- We could also have used the p-value

```
p_value <- 1 - pchisq(chisq_obs, 7)
p_value
```

```
## [1] 0.1764812
```

- We do not reject normality at level 5%.

2.10 Other tests of normality

- There are many other tests of normality.
- We mention one of the most commonly used tests: Shapiro-Wilks.
- It is standard in `R`.
- We do not treat the details, but the test statistic is somewhat like a correlation for the qq-plot.
  - If the "correlation is far from 1", we reject normality.

```
shapiro.test(ln_Error)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ln_Error
## W = 0.99255, p-value = 0.1971
```

- With a p-value of 19.71%, we do not reject normality, if we test on level 5%.

3 Sources of variation

- In lecture 4.1 we discussed 3 sources of variation:
  - systematic measurement error
  - random measurement variation
  - production variation
- Generally it is relevant to decompose the production variation in 2 components:
  - variation within lot, i.e. the variation around the lot mean

– variation between lots, i.e. the variation of the lot means.

3.1 The general model

- The completely general model would be:

$$\text{measuredValue} = \text{systematicError} + \text{lotError}$$
$$+\text{componentError} + \text{measurementError}$$

- In mathematical notation
$$Y_{k,i,j} = \mu + L_k + C_{k,i} + \varepsilon_{k,i,j}$$
  where
  - k is the number of the lot
  - i is the number of the component in lot k
  - j is the number of the measurement on component (k,i).
- The errors are assumed random and normal
  - Lot errors $L_k \sim norm(0, \sigma_l)$
  - Errors on individual component within lot $C_{k,i} \sim norm(0, \sigma_c)$
  - Measurement errors $\varepsilon_{k,i,j} \sim norm(0, \sigma_m)$

3.2 Model for our data

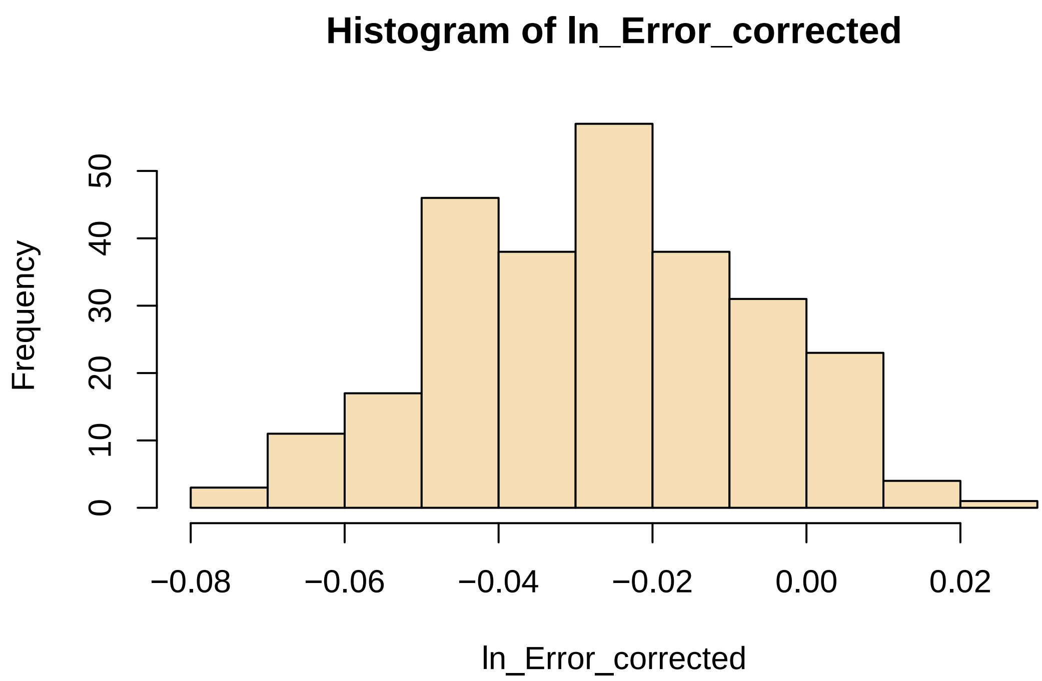
- As we have one lot only, we cannot identify the variation between lots.
  - We will consider the lot mean as fixed number μ_l
- We only have one measurement on each component
- The model for our data reduces to (since $k = 1$ and $j = 1$ we omit them from notation)
$$Y_i = \mu + \mu_l + C_i + \varepsilon_i$$
  where
  - $i = 1, \ldots, 269$ is observation number
  - μ is systematic measurement error
  - μ_l is systematic lot error
  - $C_i \sim norm(0, \sigma_c)$ is variation within lot
  - $\varepsilon_i \sim norm(0, \sigma_m)$ is measurement error

---

3.3 Linear calibration

- In lecture 4.1 we developed a linear calibration to eliminate the systematic measurement error.
- To remove the systematic measurement error, we apply this calibration to our new dataset.

```
load("ab.RData")
ln_Error_corrected <- (ln_Error-ab[1])/ab[2]
hist(ln_Error_corrected, breaks = "FD", col = "wheat")
```



3.4 Model for calibrated data

- After calibration, we will assume that the systematic measurement is zero, leaving us with the model for the calibrated values:

$$Y_i = \mu_l + C_i + \varepsilon_i$$

  where
  - $i = 1, \ldots, 269$ is observation number
  - μ_l is systematic lot error
  - $C_i \sim norm(0, \sigma_c)$ is variation within lot
  - $\varepsilon_i \sim norm(0, \sigma_m)$ is measurement error
- We are this left with a normally distributed sample with
  - mean μ_l
  - variance $\sigma_c^2 + \sigma_m^2$

3.5 Estimate of parameters

- Estimate of μ_c

```
myl <- mean(ln_Error_corrected)
myl
```

```
## [1] -0.02686793
```

- That is, the systematic lot error is around -2.7%.
- Estimate of $\sigma_m^2 + \sigma_c^2$

```
var(ln_Error_corrected)
```

```
## [1] 0.0003892828
```

- That is $s_m^2 + s_c^2 = 3.9 \cdot 10^{-4}$.

- In lecture 4.1 we estimated $s_m^2 = 0.29 \cdot 10^{-6}$ and hence $s_c^2 = 3.9 \cdot 10^{-4}$

$$s_c = \sqrt{3.9 \cdot 10^{-4}} = 0.02$$

- 3 sigma limits for the corrected lot values:

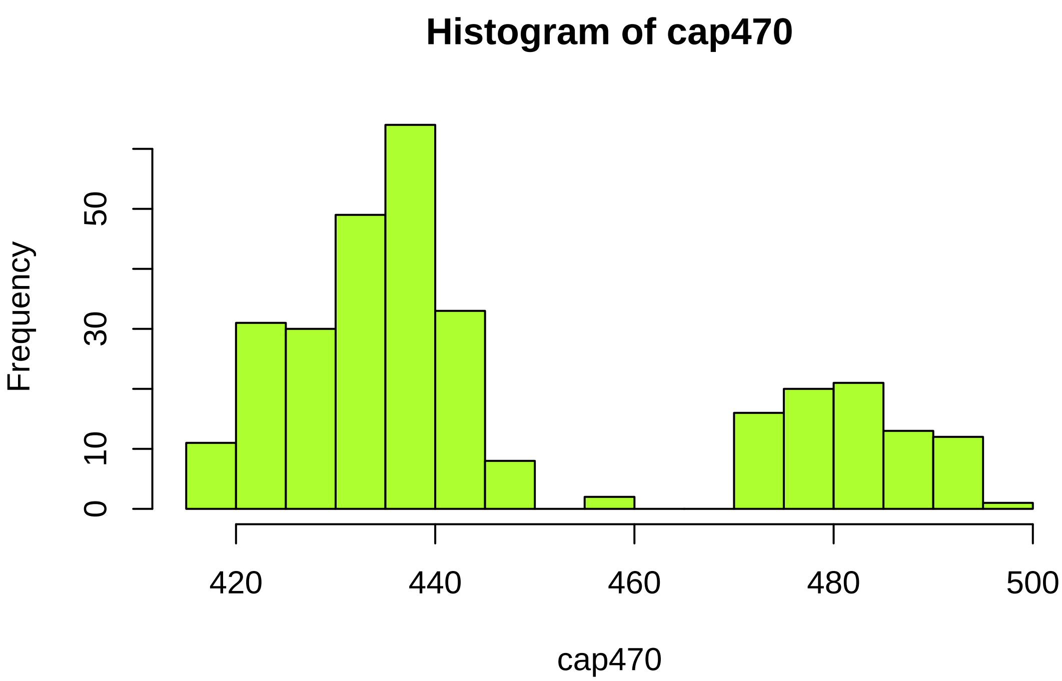
$$-2.7\% \pm 3 \cdot 2.0\% = [-8.7; 3.3]\%$$

clearly respecting the 10% tolerance.

4 Mixture of lots

- Peter has also tested 311 capacitors with nominal value 470 nF

```
cap470 <- read.table(url("https://asta.math.aau.dk/datasets?file=capacitor_lot_470_nF2.txt"))[, 1]
hist(cap470, breaks = 15, col = "greenyellow")
```

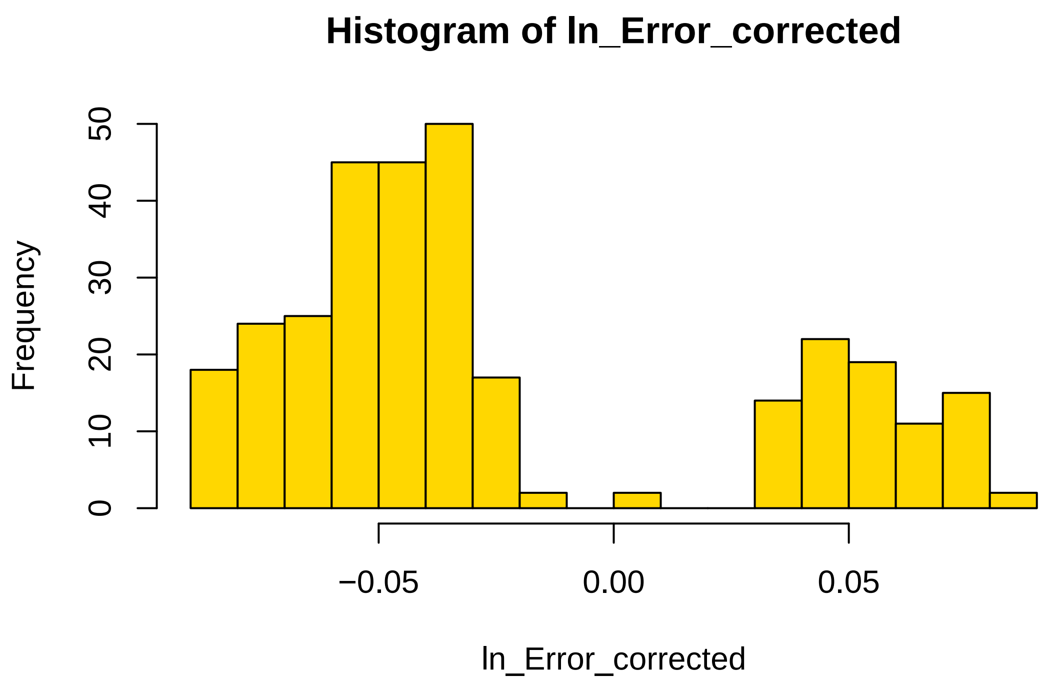


- Consulting Peter, it turned out, that his box of capacitors contained components from 2 different lots.

4.1 Transforming

- We ln-transform and calibrate:

```
ln_Error <- log(cap470/470)
ln_Error_corrected <- (ln_Error-ab[1])/ab[2]
hist(ln_Error_corrected, breaks = 15, col = "gold")
```




```
range(ln_Error_corrected)
```

```
## [1] -0.08888934  0.08323081
```

4.2 Mixture model

- We assume that the ln_Error
  - is normal with mean μ_1 if the component is from lot 1
  - is normal with mean μ_2 if the component is from lot 2
  - both distributions have variance $\sigma^2 = \sigma_m^2 + \sigma_l^2$
  - the probability of coming from lot 1 is p
- So we have 4 unknown parameters: $(\mu_1, \mu_2, \sigma, p)$.
- To estimate these, we entrust to the `R`-package `mclust`.

4.3 Fitting a mixture

- We fit the model

```
library(mclust)
fit <- Mclust(ln_Error_corrected, 2 , "E")# 2 clusters; "E"qual variances
pr <- fit$parameters$pro[1]
pr
```

```
## [1] 0.728314
```

- The chance of coming from lot 1 is around 73%.

```
means <- fit$parameters$mean
means
```

```
##            1            2
## -0.05174452  0.05406515
```

- The mean in lot 1 is around -5.2%
- The mean in lot 2 is around 5.4%

```
sigma <- sqrt(fit$parameters$variance$sigmasq)
sigma
```

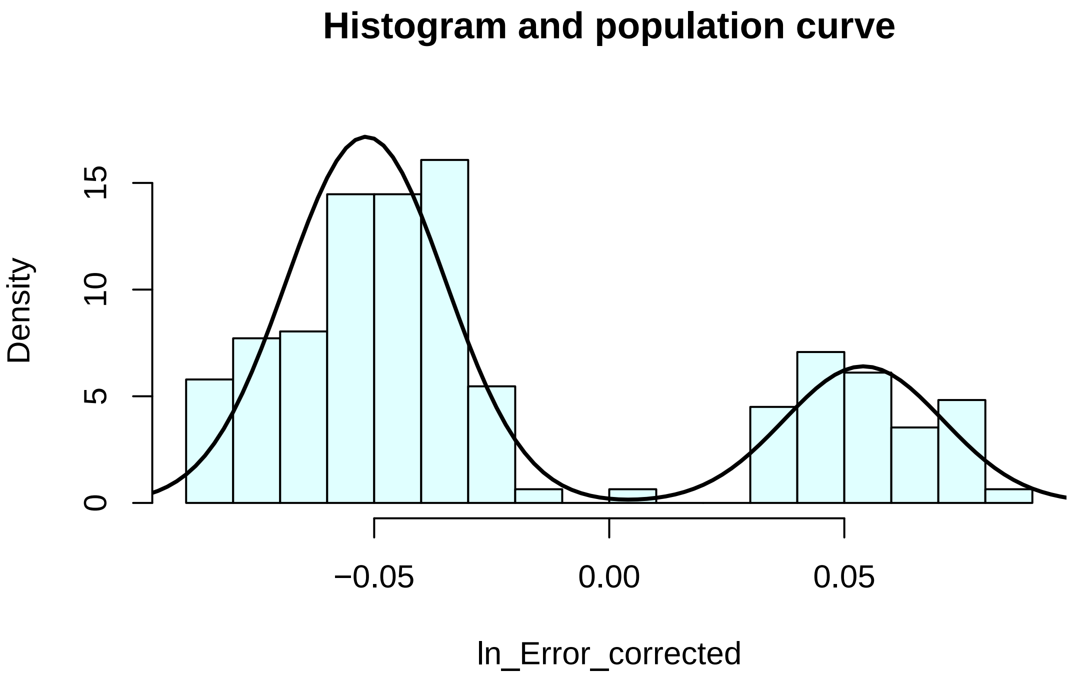
```
## [1] 0.01692654
```

- σ is around 1.7%

4.4 Comparing model and data

- We compare the histogram with the fitted normal curves.

```
hist(ln_Error_corrected,breaks=15,col="lightcyan",probability = TRUE,ylim=c(0,18),main="Histogram and po
curve(pr*dnorm(x,means[1],sigma)+(1-pr)*dnorm(x,means[2],sigma),-.1,.1,add=TRUE,lwd=2)
```



4.5 Concluding remarks

- Estimate of σ was 1.7%. In relation to the 220 nF lot we estimated 2.0%, which is comparable.
  - 3 sigma limits for the correct lot 1 values:

$$-5.2\% \pm 3 * 1.7\% = [-10.3; -0.1]\%$$

  - 3 sigma limits for the correct lot 2 values:

$$5.4\% \pm 3 * 1.7\% = [0.3; 10.5]\%$$

- The lots do not completely respect the tolerance of 10%. However, in the sample the minimum is -8.9% and the maximum 8.3%.

- The difference in lot means is $5.4\% - (-5,2)\% = 10.6$.
- This indicates that the variation between lots is much greater than the variation within lots.
- This is also clearly illustrated by the histogram/density plots.