Statistics and electronics - lecture 2

The ASTA team

Contents

0.1 Checking for log normality

Picture of a “lot” of capacitors.

The word lot is used to identify several components produced in a single run.

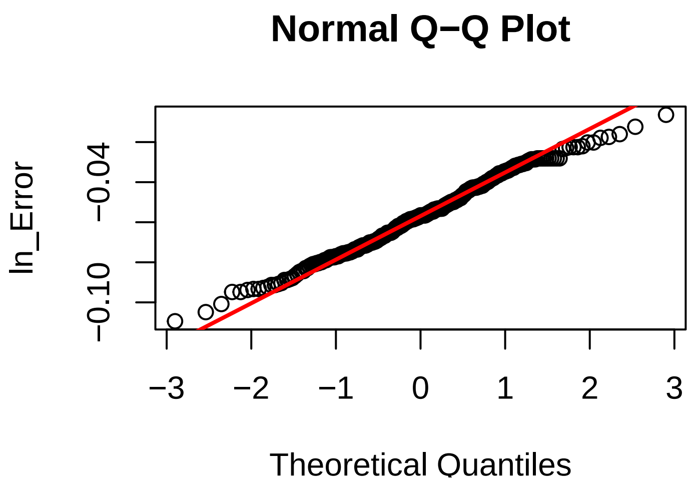
Where a run is a production series limited to a given timeinterval and fixed production parameters.

0.2 Lot variation

Peter Koch has tested 269 of the capacitors in the displayed lot.

First of all, we will check the assumption that our measurements have a log normal error.

```r
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")
```



0.3 Testing normality

The qq-plot(WMM - section 8.8) supports normality of the ln_Error.

There are several tests of normality.

Two of these are considered in WMM section 10.11:

- Gearys test
- goodness of fit

0.4 Gearys test

Consider a sample $X_1, \ldots, X_n$ and an estimate of σ - the standard deviation 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. But otherwise, it will under- or overestimate σ depending on the form of the population distribution.

0.5 Gearys test

Hence we expect that

- $U = \frac{S_1}{S_0}$ should be close to one in case of normality.

For large values of n a normal approximation yields that

- $Z = \frac{\sqrt{n}(U-1)}{0.2661}$ has a standard normal distribution **if** the sample is normal

that is, if $-2 \leq z_{obs} \leq 2$, we do not reject normality, if we test on level 5%.

```
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)/0.2261
z_obs
```

```
## [1] -1.628122
```

Hence there is no evidence of non-normality.

0.6 Goodness of fit

Is a general method for investigating whether a sample has a specific distribution.

The first example in WMM is concerned with the problem of whether a dice is balanced.

That is, all sides have probability 1/6 of showing up.

Rolling the dice 120 times we expect

- ExpectedFrequency: (20, 20, 20, 20, 20, 20)

Actually we observe

- ObservedFrequency: (20, 22, 17, 18, 19, 24)

Distance measure between observed and expected:

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

If the dice is balanced then

- X^2 has a so-called **chi-square distribution** (WMM chapter 6.7) with df=k-1=5, degrees of freedom

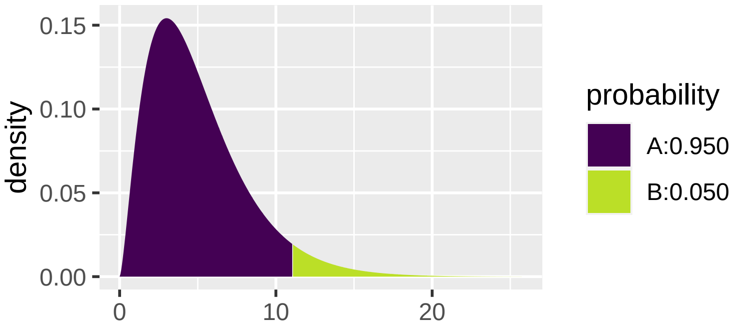
where k=6 is the number of possible outcomes.

0.7 Goodness of fit

For the actual data:

- $x^2_{obs} = 1.7$ and we need to judge whether this is higher than expected. **If** the null hypothesis is true.

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



```
critical_value
```

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

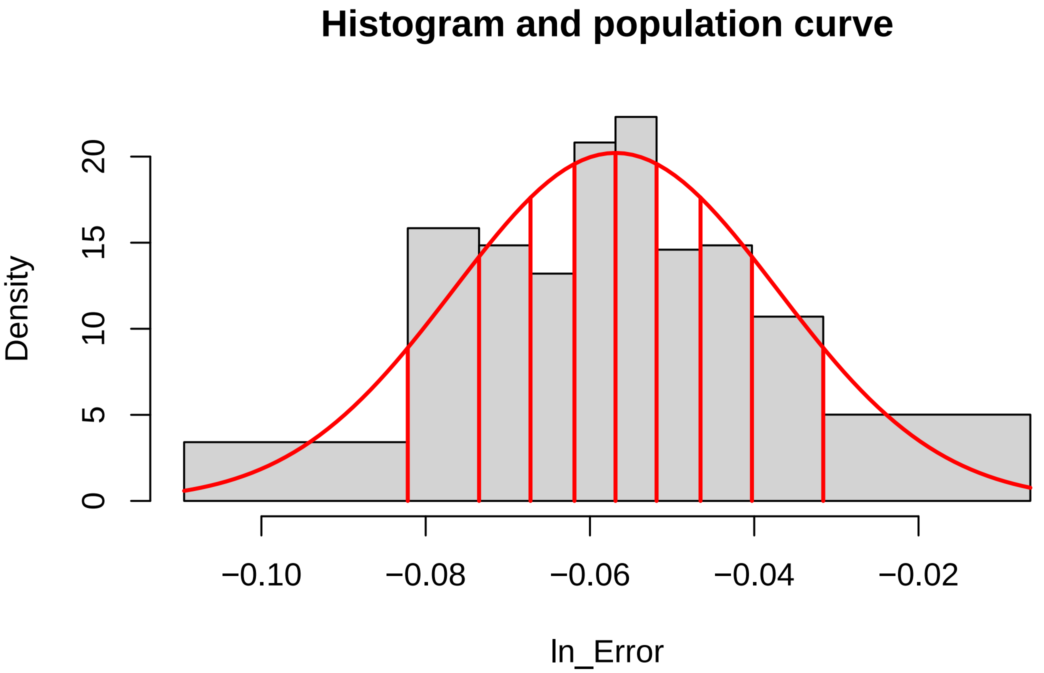
At 5% significance the critical value is 11.07, so there is no evidence of unbalancedness.

0.8 Goodness of fit - normal distribution

We assume that ln_Error is a sample from a normal distribution and divide the population distribution into 10 bins with equal probabilities p=10%.

The number of bins could be changed. It is required that the expected frequency should be 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 and as sample size is 269 we obtain

- Expected_frequency is 26.9 in each bin

0.9 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
```

X^2 statistic:

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

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

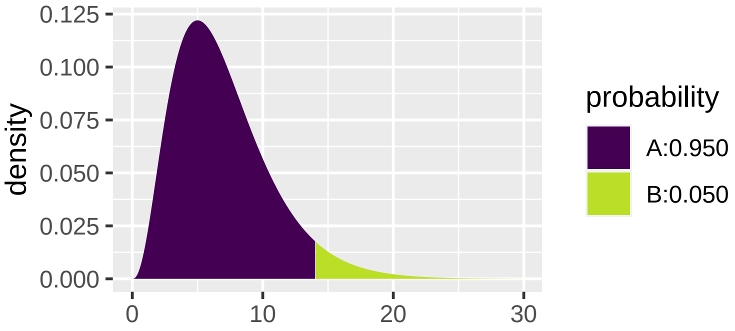
The degrees of freedom is the number of bins minus 3 (number of parameters + 1), i.e. df = 10-3 = 7.

0.10 Goodness of fit - normal distribution

```
chisq_obs
```

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

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



```
critical_value
```

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

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

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

We do not reject normality at level 5%.

0.11 Other tests of normality

As mentioned, there are multiple tests of normality.

We introduce one other test: 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 p-value=19.71%, we do not reject normality, if we test on level 5%.

0.12 Sources of variation

In lecture 1 we discussed

- 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.

0.13 Sources of variation

As we have one lot only, we cannot identify the variation between lots.

Our actual data are thus composed of

- systematic measurement error - call it μ_m

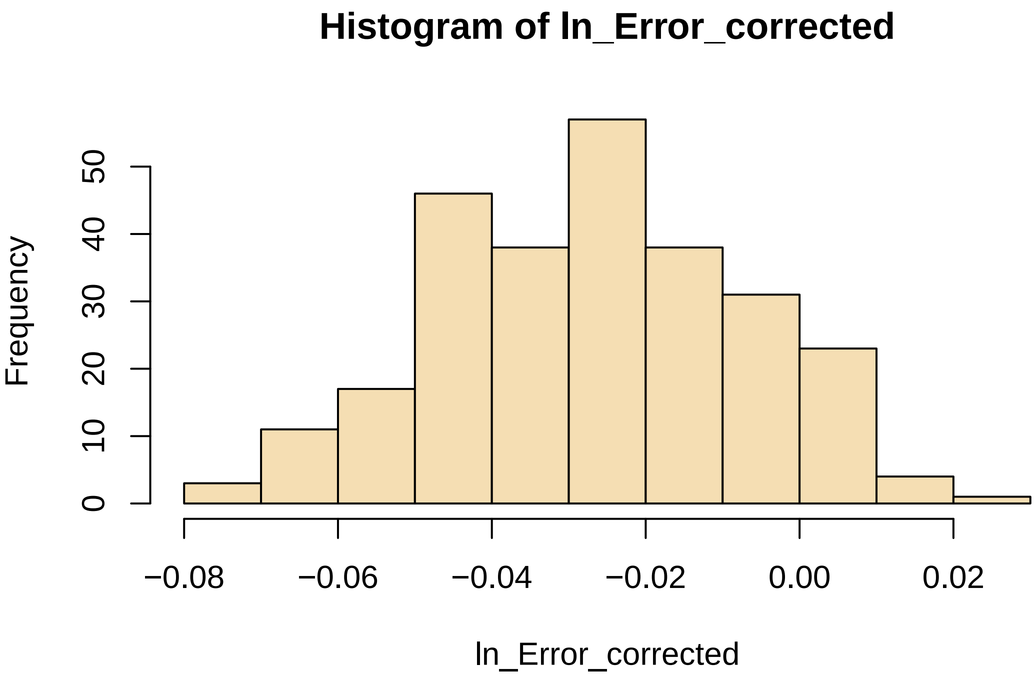
- systematic lot error - call it μ_l
- standard deviation of measurement - call it σ_m
- standard deviation within lot - call it σ_l

0.14 Linear calibration

In lecture 1 we developed a linear calibration eliminating the systematic measurement error.

Adopting this to the actual data yields

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



0.15 Sources of variation

We are now left with a sample, which has

- mean μ_l and variance $\sigma_m^2 + \sigma_l^2$

where we have assumed that the random measurement error and the random lot error are independent.

Estimate of μ_l

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

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

That is, the systematic lot error is around -2.7%.

0.16 Estimate of variances

Estimate of $\sigma_m^2 + \sigma_l^2$

```
var(ln_Error_corrected)
```

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

that is $s_m^2 + s_l^2 = $ 3.9e-04

In lecture 1 we estimated $s_m^2 = $ 0.29e-06 and hence

- $s_l = $ sqrt(3.9e-04) = 2.0%.

3 sigma limits for the correct lot values:

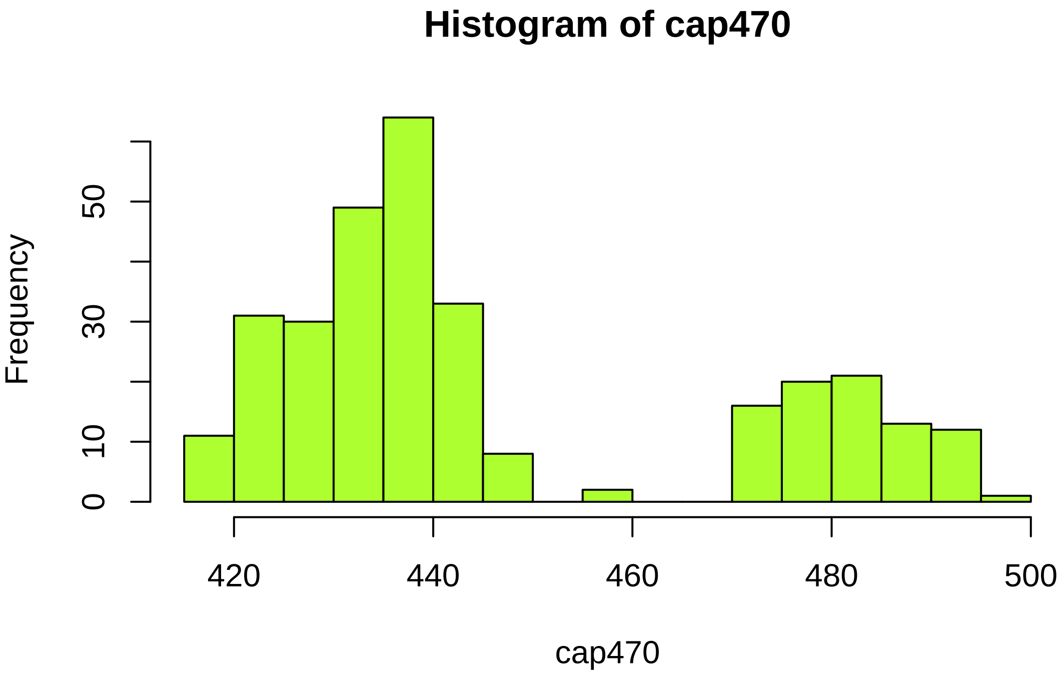
- -2,7 ± 3*2.0 = [-8.7; 3.3]%

clearly respecting the 10% tolerance.

0.17 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")
```

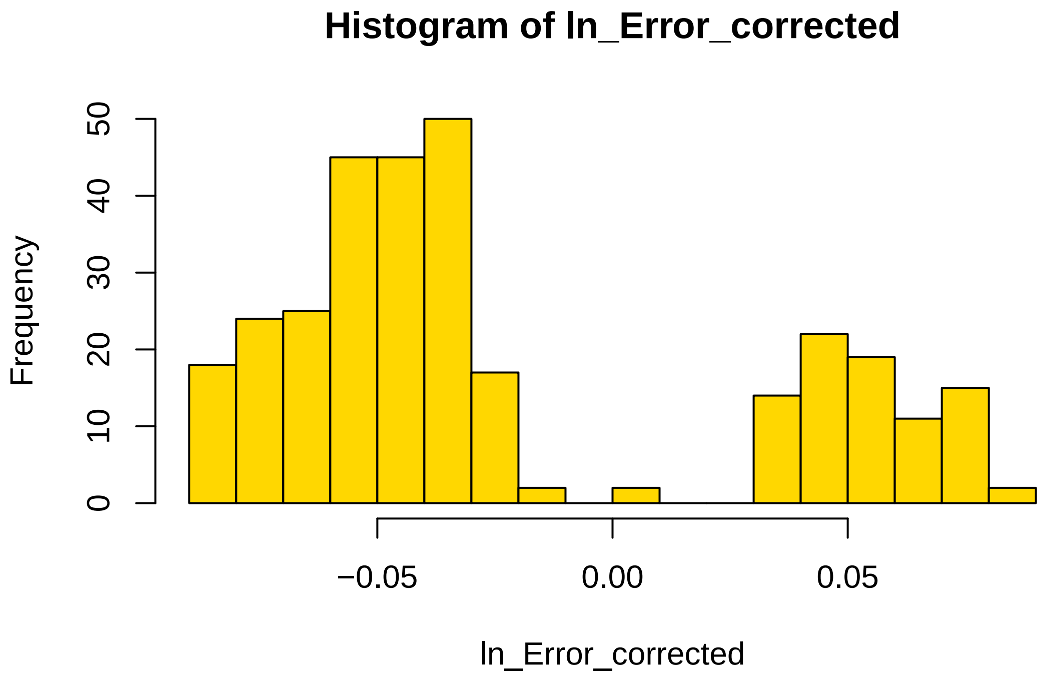


Consulting Peter, it turned out, that his box of capacitors contained components from 2 different lots.

0.18 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")
```



```r
range(ln_Error_corrected)
```

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

0.19 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: $(\mu_1, \mu_2, \sigma, p)$.

How to estimate these, we entrust to the `R`-package `mclust`.

0.20 Fitting a mixture

```r
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 lot1 is around 73%.

```r
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%

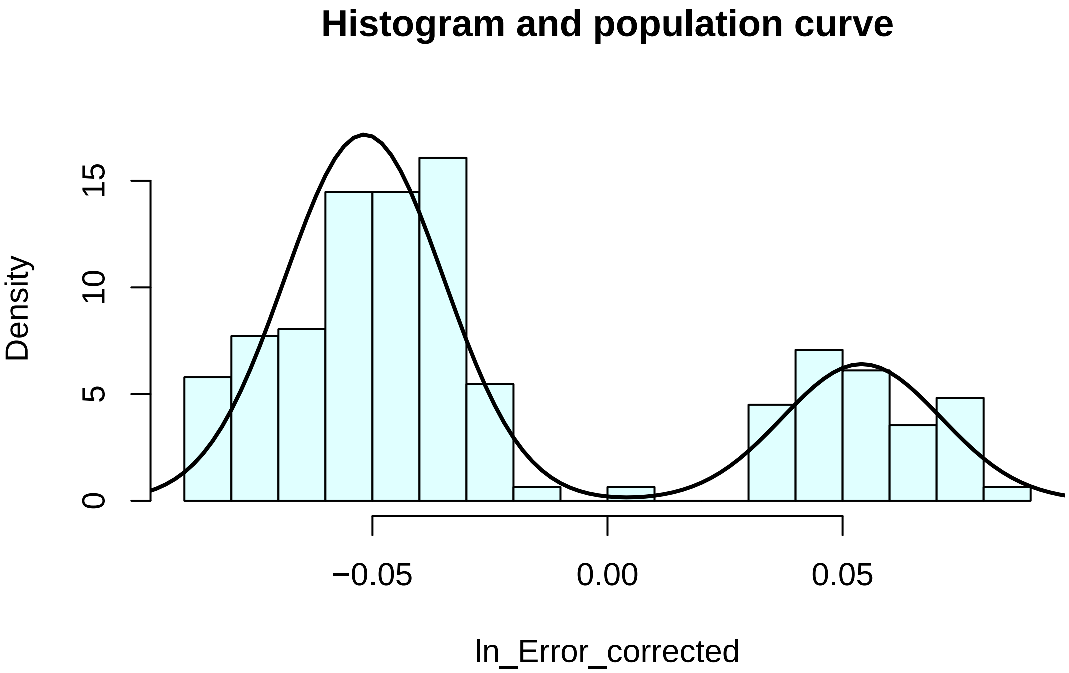
```
sigma <- sqrt(fit$parameters$variance$sigmasq)
sigma
```

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

- σ is around 1.7%

0.21 Comparing model and data

```
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)
```



0.22 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 ± 3*1.7=[-10.3;-0.1]%
- 3 sigma limits for the correct lot 2 values: 5.4 ± 3*1.7=[0.3;10.5]%

do not completely respect the tolerance 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.

Which is also clearly illustrated by the histogram/density plots.