ASTA

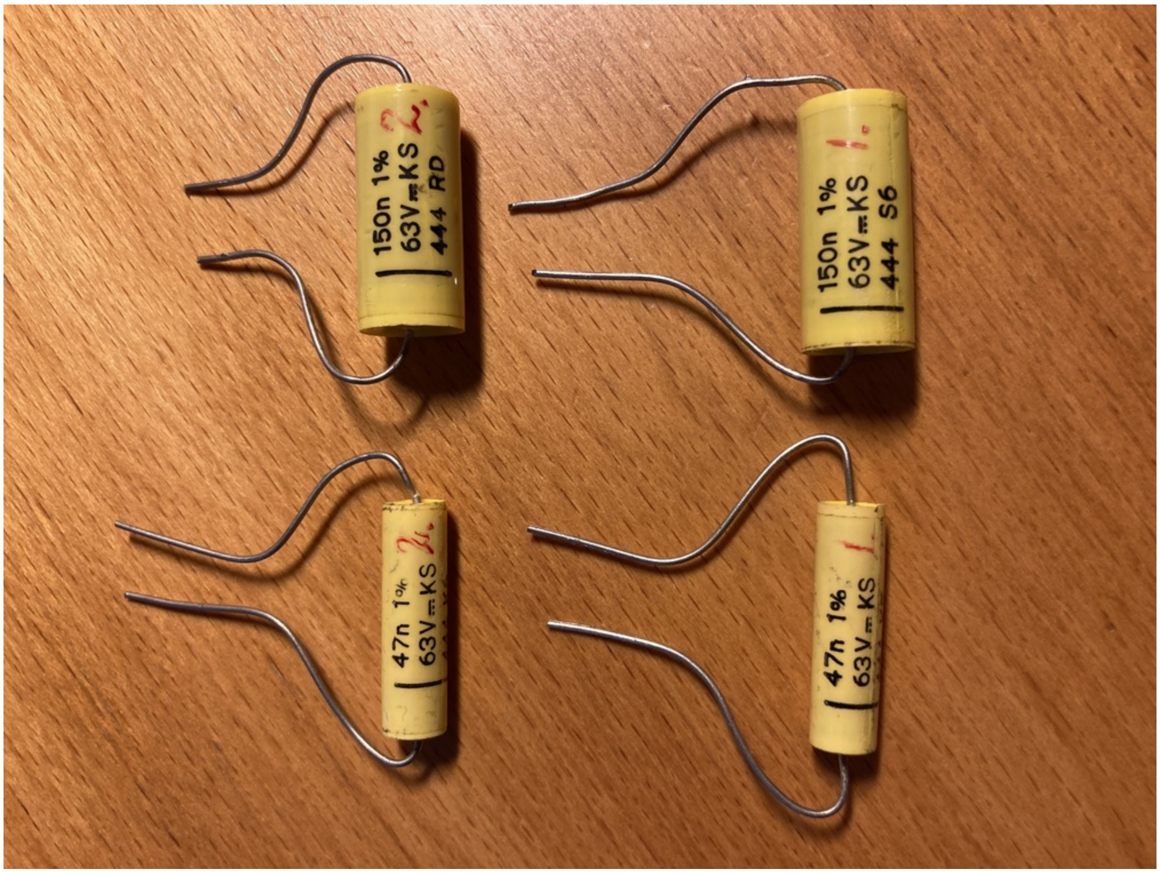
The ASTA team

Contents

0.1 Sources of variation

We shall study 2 types of variation

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



0.2 Data from Peter Koch

Peter has done 100 independent measurements of the capacity of 4 of the displayed capacitors and one additional. Nominal values are 47, 47, 100, 150, 150. All with 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 capacity measurements of the first capacitor with nominal value 47.

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

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

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

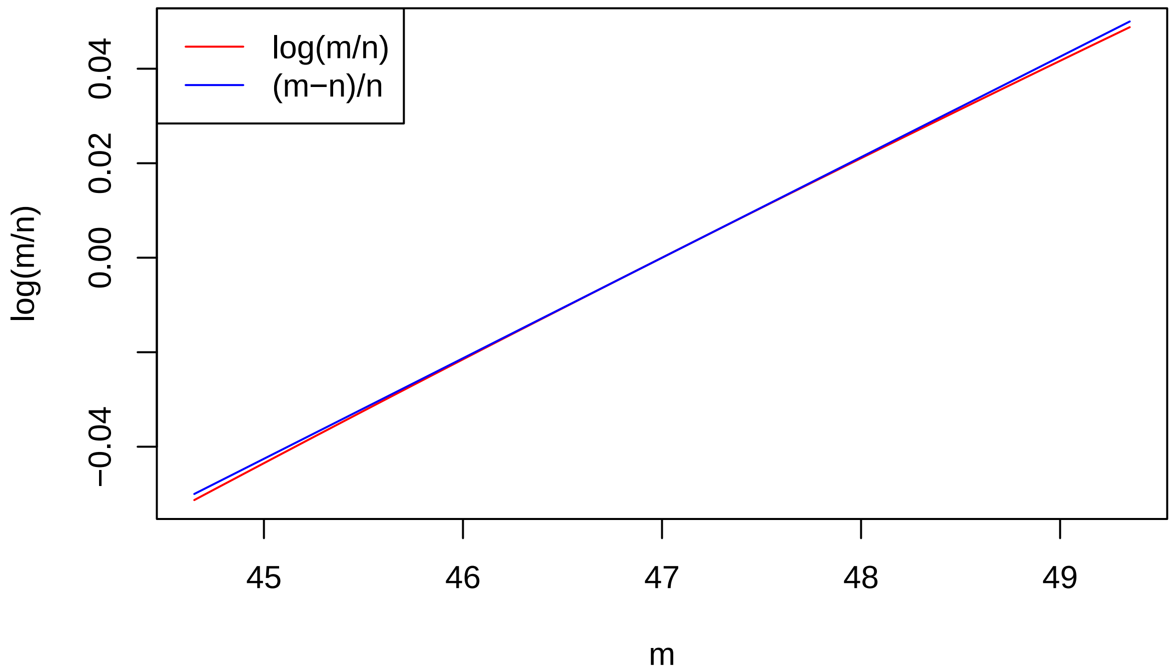
0.3 Transformation

Linearisation:

$$
\begin{aligned}
f(x) &\approx f(x_0) + f'(x_0)(x - x_0) && (1)\\
& && (2)\\
x_0 &= 1 && (3)\\
f(x) &= \log x && (4)\\
f'(x) &= 1/x && (5)\\
& && (6)\\
x &= m/n && (7)\\
\log\left(\frac{m}{n}\right) &\approx \log 1 + \frac{1}{1}\left(\frac{m}{n} - 1\right) && (8)\\
&= \frac{m - n}{n} && (9)
\end{aligned}
$$

$$\log\left(\frac{m}{n}\right) \approx \frac{m - n}{n}$$

```r
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"))
```



0.4 Transformation

$$\log\left(\frac{m}{n}\right) \approx \frac{m-n}{n}$$

Instead of the raw measurement we will consider:

lnError = ln(measuredValue/nominalValue)

Remark that by linear approximation:

lnError ≈ measuredValue/nominalValue - 1 = (measuredValue-nominalValue)/nominalValue

which is the error relative to the nominal value.

I.e.: `lnError` can be interpreted as relative error.

0.5 Transformed data

```
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 2/1 decimal(s) in the 47/150 nF range, which means that only a limited number of different values(3-8) are observed. Meaning that box- or histogram-plots are noninformative.

0.6 Model considerations

- The measurements are more than 2.7% below the nominal value. This must be due to a systematic error on the meter.

In this case we have as earlier mentioned two further sources of error:

- ln(measuredValue / nominalValue) = systematicError + productionError + measurementError

0.7 Statistical model

$$\ln(\text{measuredValue / nominalValue}) = \text{systematicError} + \text{productionError} + \text{measurementError}$$

We formulate the model:

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

where

- Y_{ij} is the log error measurement
- μ is the systematic error on the meter
- A_i is the random production error
- ε_{ij} is the random measurement error
- $i = 1, 2, 3, 4, k = 5$ is the number of the 5 samples
- $j = 1, \ldots, n = 100$ is the number of the observation in each sample

0.8 Assumptions

This is the model treated in WMM chapter 13.11, where it is assumed that

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

0.9 Estimation of systematic error

The systematic error is simply estimated by the mean

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

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

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

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

0.10 Estimation of random error

Notation from WMM chapter 13.3:

- $SSA = n\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$ (related to production error)
- $SSE = \sum_{ij} (y_{ij} - \bar{y}_{i.})^2$ (related to measurement error)

Theorem 13.4 states:

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

0.11 Fit

```
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
```

where we read

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

- SSA $= 0.00466$ and SSE $= 0.000142$

0.12 Solution

Solving the equations

- `SSA = E(SSA)` and `SSE = E(SSE)`

yields

- $\hat{\sigma}_\alpha^2 = \frac{1}{n}(\frac{SSA}{k-1} - \hat{\sigma}^2) = 11.64 \times 10^{-6}$
- $\hat{\sigma}^2 = \frac{SSE}{k(n-1)} = 0.29 \times 10^{-6}$

0.13 Summing up

- the meter has an estimated systematic error of -2.88%
- the estimated standard error of the meter is $\sqrt{0.29 \times 10^{-6}} = 0.054\%$
- the estimated standard error of the production is $\sqrt{11.64 \times 10^{-6}} = 0.34\%$. So the 3-sigma limit is 1.02%, which is in accordance with the tolerance of 1%. It should be noted that the estimate is insecure, as it is based on 4 degrees of freedom only.

The estimated variance on log error

- $0.29 \times 10^{-6} + 11.64 \times 10^{-6} = 11.93 \times 10^{-6}$

is clearly dominated by the production error.

0.14 Test of no random effect

We have the possibility of testing the hypothesis

- H_0: $\sigma_\alpha = 0$

This is equivalent to

- $E(SSA/(k-1)) = E(SSE/k/(n-1)) = \sigma^2$

Under H_0 the statistic

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

has an F-distribution with degrees of freedom $(k-1, k(n-1))$

In the actual case $f_{obs} = 4067.4$, which is highly significant (p-value=0).

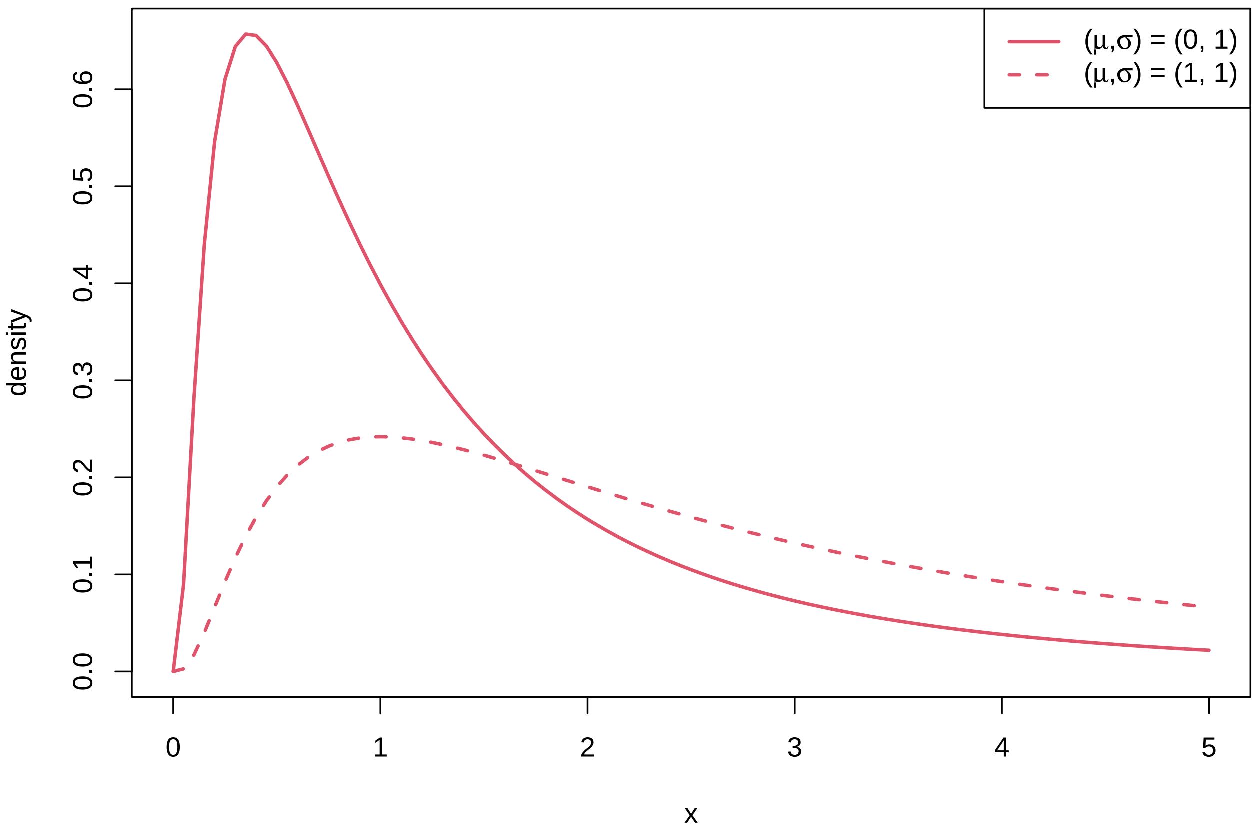
0.15 Lognormal variation

In the preceeding we assumed normal errors after a log transformation.

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.

Density plots:



0.16 Moments of lognormal

If $Y = ln(X)$ has a normal distribution with mean μ and standard deviation σ, then Theorem 6.7 of WMM states:

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

If we are interested in relative variation, it is common to look at the coefficient of variation

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

if e.g. CV=0.05 then 95% of our measurements are within

- $\mu \pm 2\sigma = \mu \pm 2 * 0.05\mu = \mu(1 \pm 0.1)$

i.e. most observations are within 10% of the mean.

0.17 CV of Lognormal

If $Y = ln(X)$ has a normal distribution with mean μ and standard deviation σ, we calculate CV for X as

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

In Peter's data we estimated the variance of the log error to 11.64×10^{-6}, which means that the estimated CV of the capacity measurement is

- $\text{CV} = \sqrt{\exp\left(11.64 \times 10^{-6}\right) - 1} = 0.34\%$.

i.e., **if** we correct for the systematic error of the meter, then our measurements are extremely precise.

0.18 Linear calibration

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

To check this assumption consider the model

- $Y = \ln(\text{measuredValue})$ is a linear model of $x = \ln(\text{nominalValue})$
- $Y = \alpha + \beta x + \varepsilon$

where we have previously assumed slope(β) equal to 1.

0.19 Linear calibration fit

```
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 is more than close to 1. But is actually extremely significantly different from 1 (tvalue=3776.74 »» 3).

Clearly, it is a bit dubious to assume a linear relationship, as we only have 3 nominal values.

0.20 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:

- $$\text{measuredError} = \alpha + \beta * \text{correctError}$$
- $$\text{correctError} = (\text{measuredError} - \alpha)/\beta$$

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

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

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

0.21 Calibrated data

```
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
```

The calibrated data now shows that the production error on component s_1_nF47 is in the vicinity of 0.2%. Well below the tolerance 1%.

0.22 Checking for log normality

Picture of a “lot” of capacitors.

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

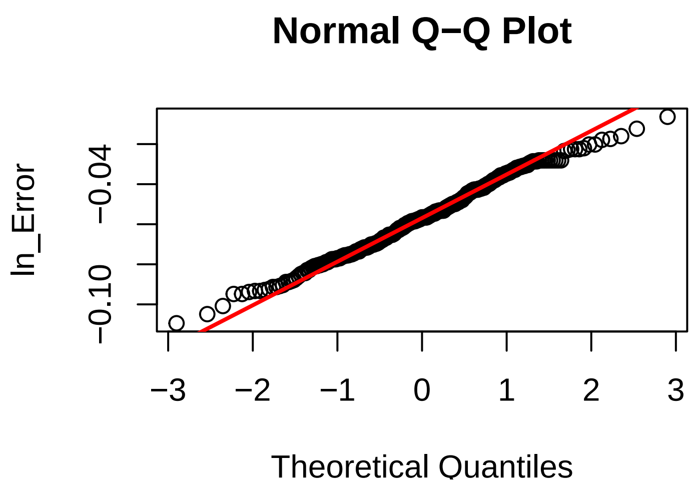
Where a run is a production series limited to a given timeinterval and fixed production parameters.

0.23 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.

```
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")
```



0.24 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

0.25 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.

0.26 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.

0.27 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

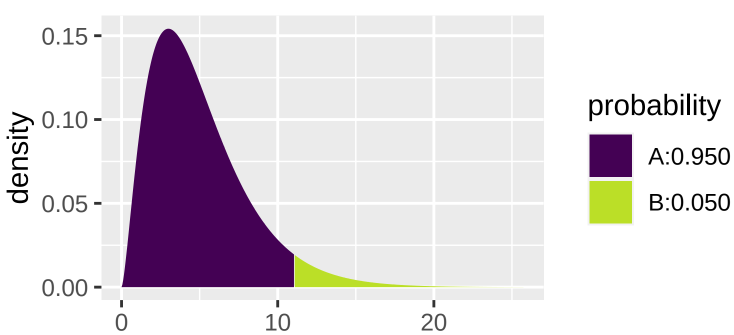
where k=6 is the number of possible outcomes.

0.28 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
```

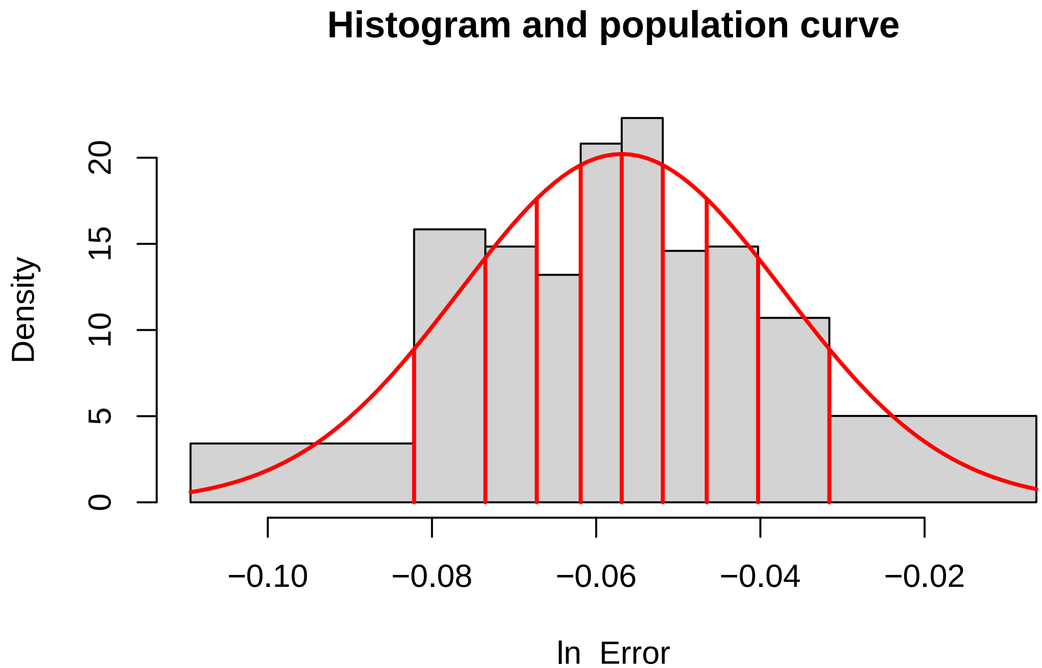
At 5% significance the critical value is 11.07, so there is no evidence of unbalancedness.

0.29 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

0.30 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
```

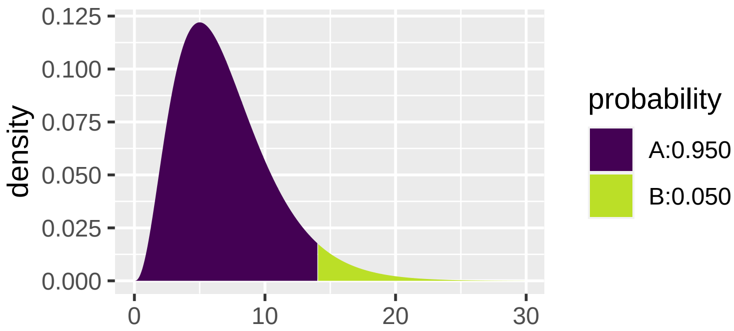
The degrees of freedom is the number of bins minus 3 (number of parameters + 1), i.e. df = 10-3 = 7.

0.31 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%.

0.32 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%.

0.33 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.

0.34 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

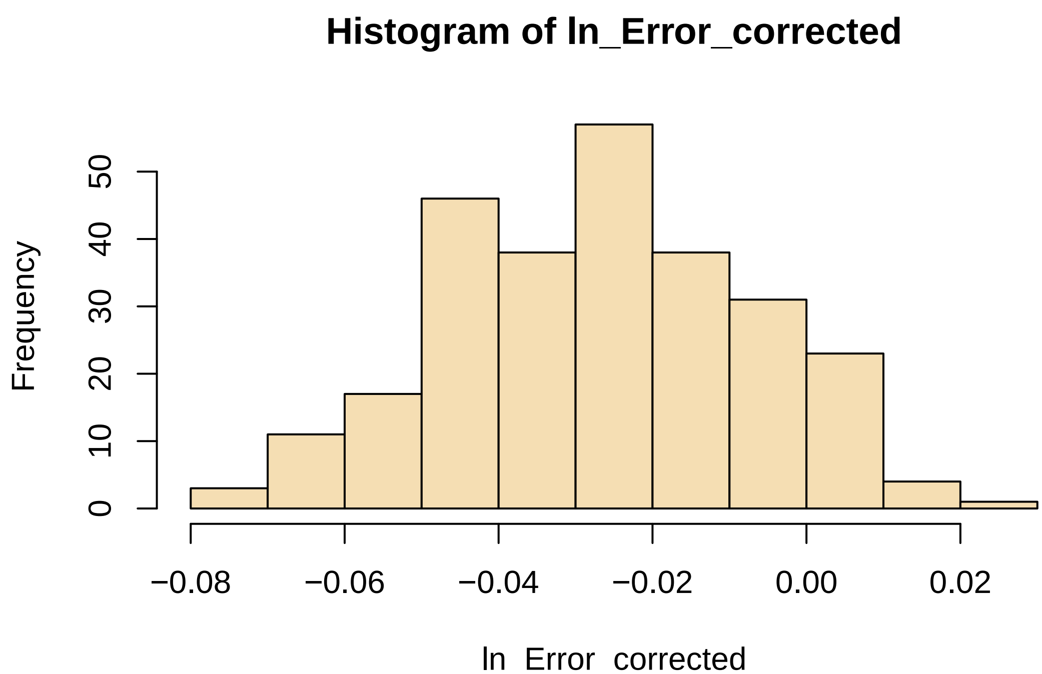
- systematic lot error - call it μ_l
- standard deviation of measurement - call it σ_m
- standard deviation within lot - call it σ_l

0.35 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")
```



0.36 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%.

0.37 Estimate of variances

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

```
var(ln_Error_corrected)
```

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

that is $s_m^2 + s_l^2 = 3.9\text{e-}04$

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

- $s_l = \text{sqrt}(3.9\text{e-}04) = 2.0\%$.

3 sigma limits for the correct lot values:

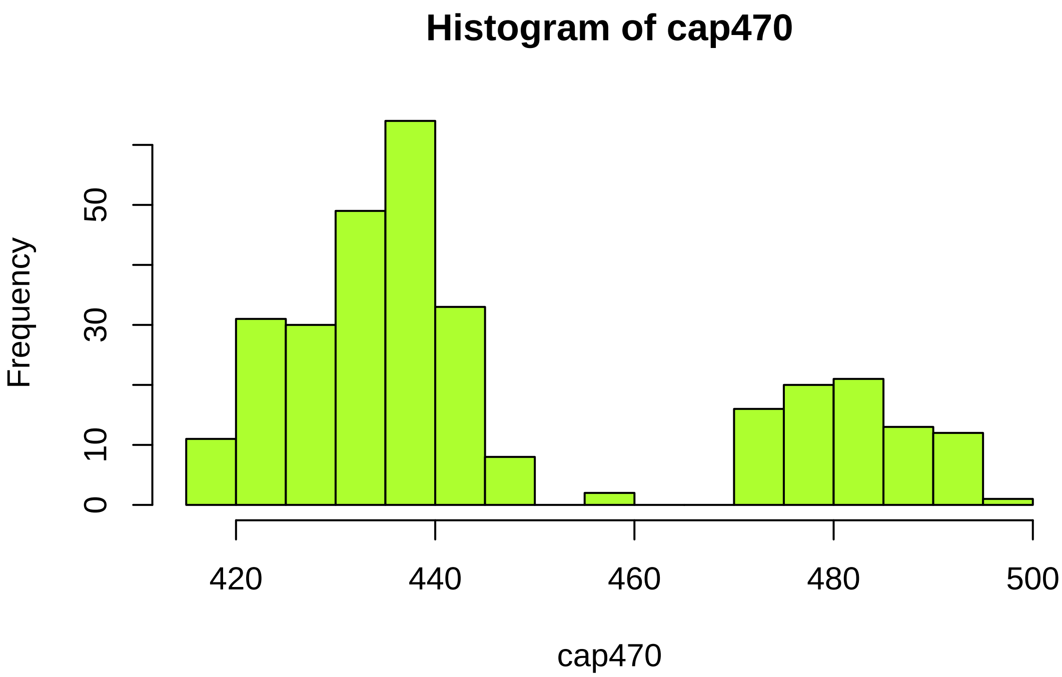
- -2,7 $\pm$ 3*2.0 = [-8.7; 3.3]%

clearly respecting the 10% tolerance.

0.38 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")
```

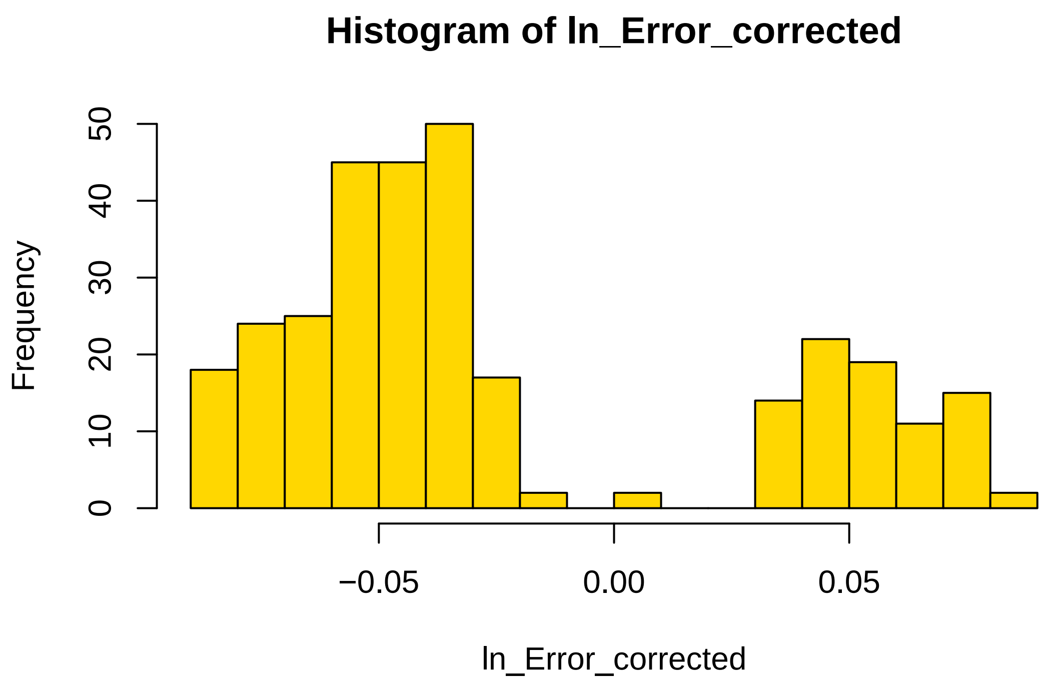


Consulting Peter, it turned out, that his box of capacitors contained components from 2 different lots.

0.39 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
```

0.40 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`.

0.41 Fitting a mixture

```
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%.

```
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%

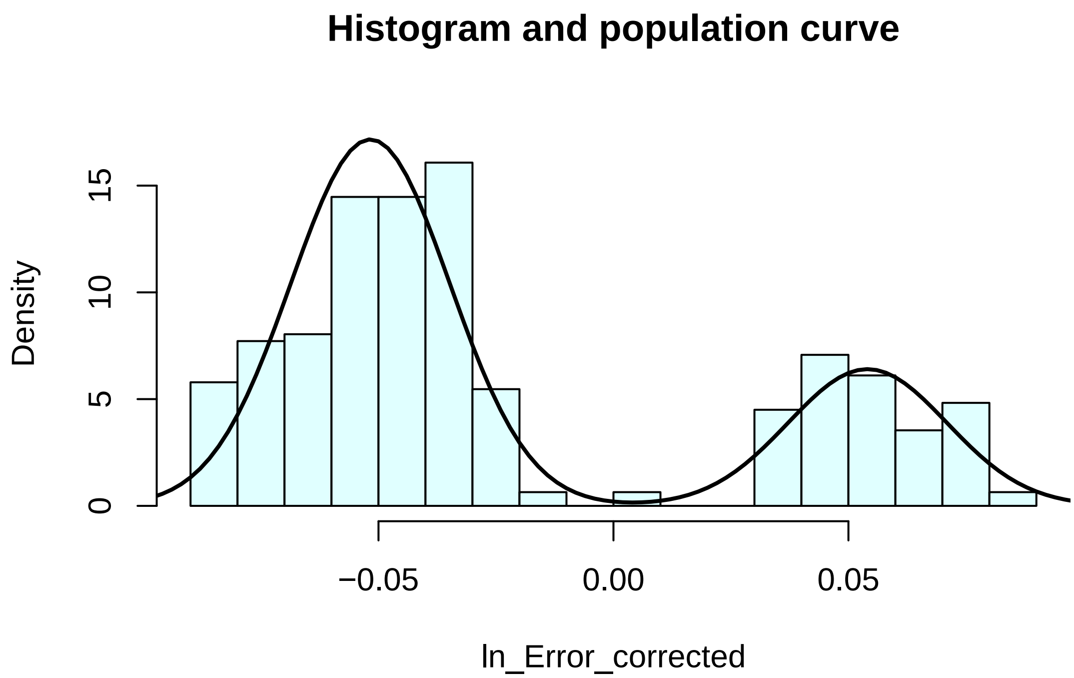
```
sigma <- sqrt(fit$parameters$variance$sigmasq)
sigma
```

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

- σ is around 1.7%

0.42 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)
```



0.43 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]%

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.