2021

Basic statistical concepts for modelling

David Makowski

Key concepts

- Population
- Sample
- Estimator, estimate
- Bias and variance of an estimator
- Test
- Confidence interval
- Model

Population

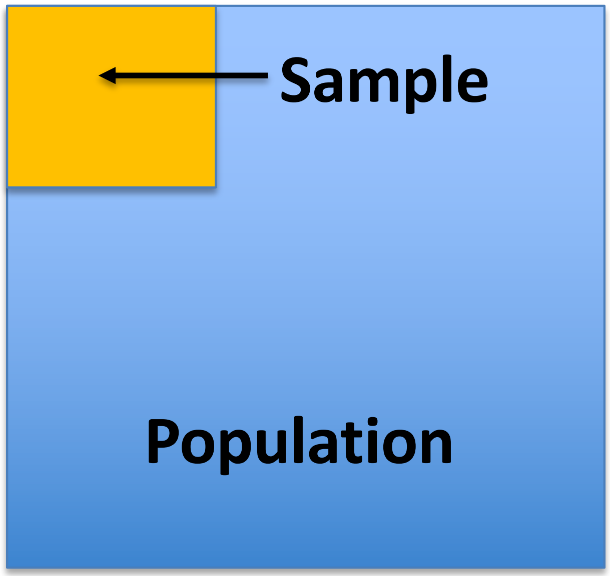
In statistics, a population is the entire pool from which a statistical sample is drawn.

A population may refer to an entire group of people, objects, events, hospital visits, or measurements.

www.investopedia.com/terms/p/population.asp

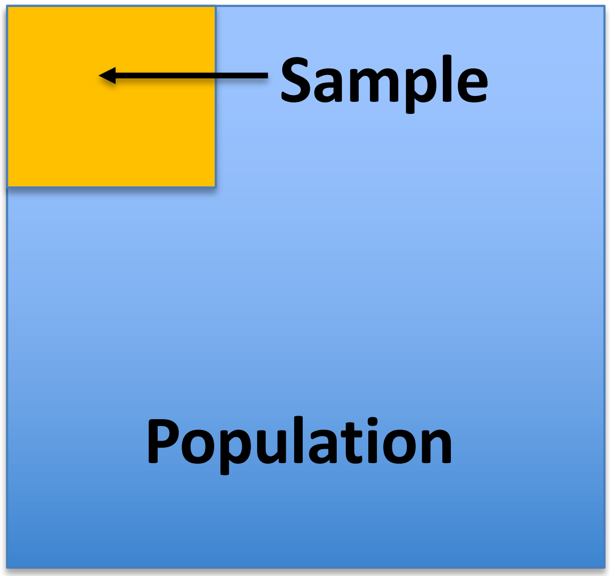
Sample

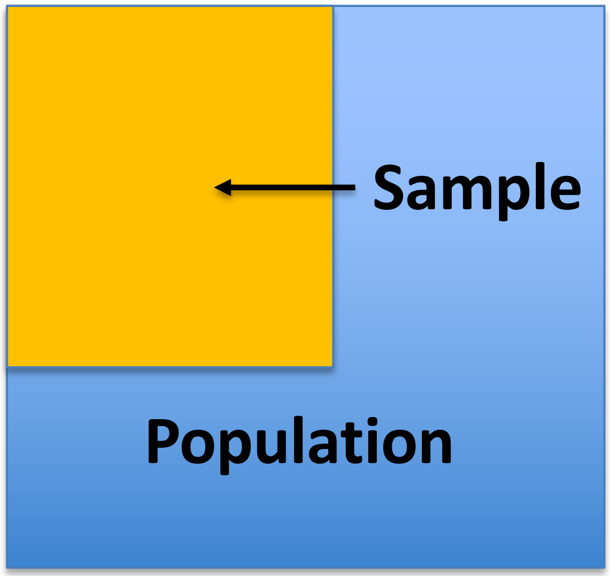
A part of a population used to estimate a characteristic of the population.



Sample

A part of a population used to estimate a characteristic of the population.





Random sample

A random sample is a sample that is chosen randomly.

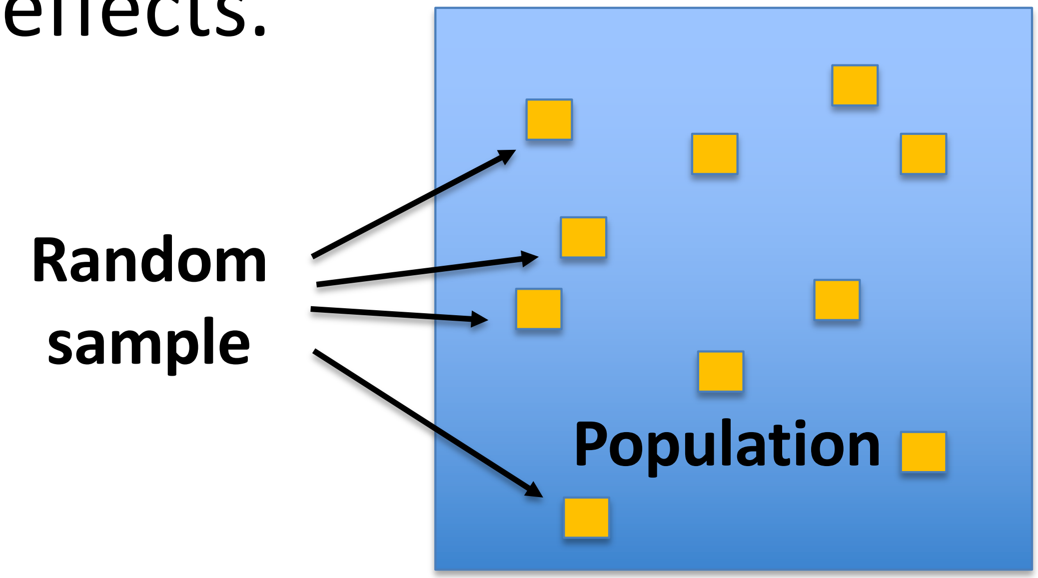
Random samples are used to avoid bias and other unwanted effects.

https://www.statisticshowto.datasciencecentral.com/simple-random-sample/

Random sample

A random sample is a sample that is chosen randomly.

Random samples are used to avoid bias and other unwanted effects.



https://www.statisticshowto.datasciencecentral.com/simple-random-sample/

Exercise

Consider the following series of numbers

1, 2, 3, 4,…,100

Generate 10 random samples of size 5 with the R function sample()

Why a random sample?

Central Limit theorem

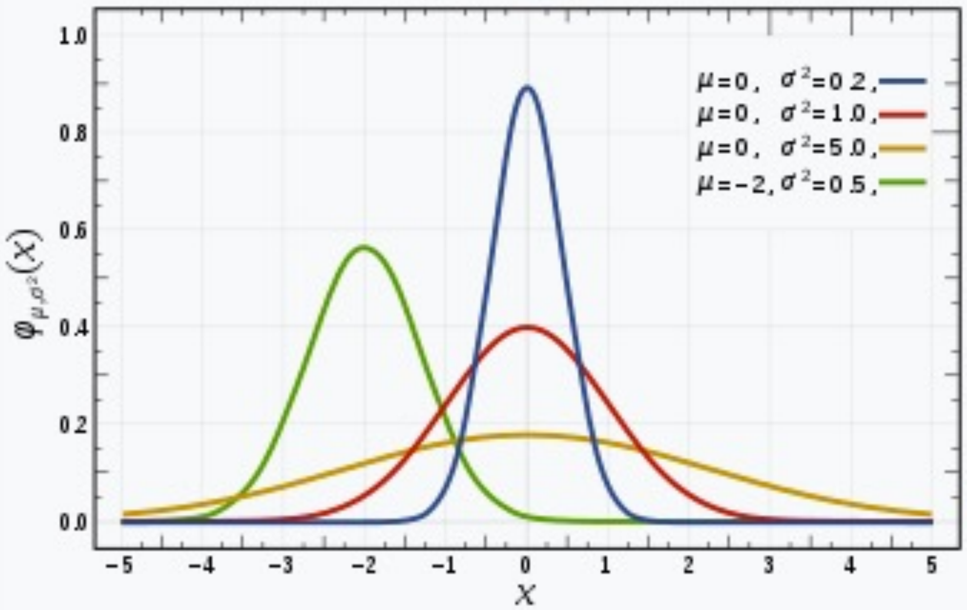
Abraham de Moivre (18th century)
Pierre Simon Marquis de Laplace (19th century)

The distributions of the average of randomly chosen observations is closely approximated by a **normal distribution**

…even if the original observations themselves are not normally distributed.

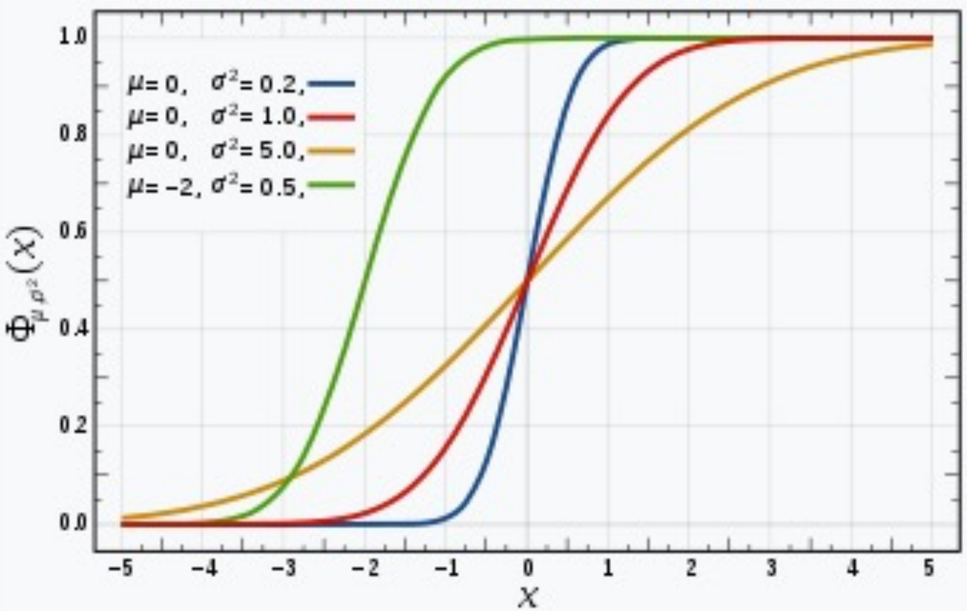
Normal Distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

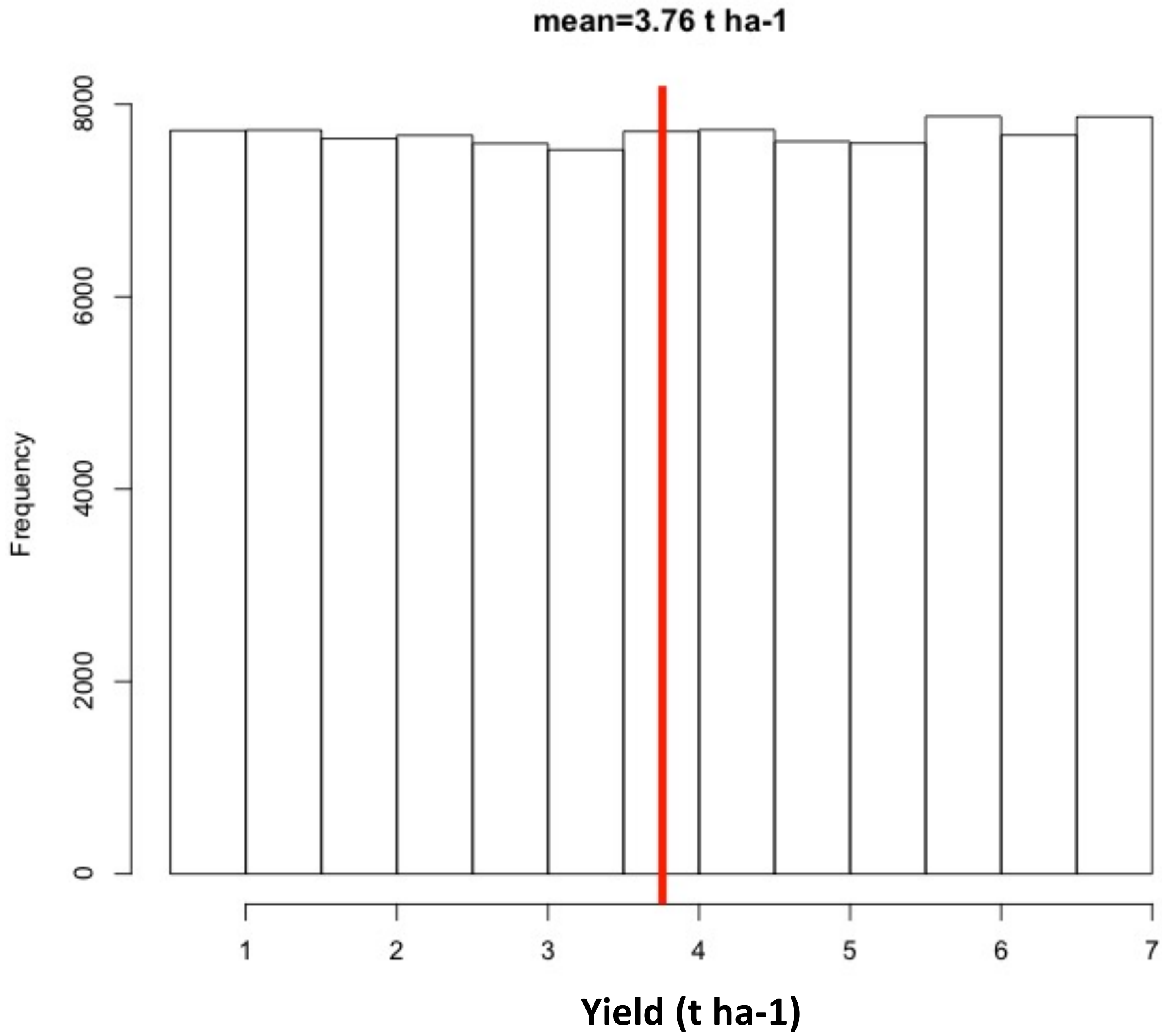
https://en.wikipedia.org/wiki/Normal_distribution

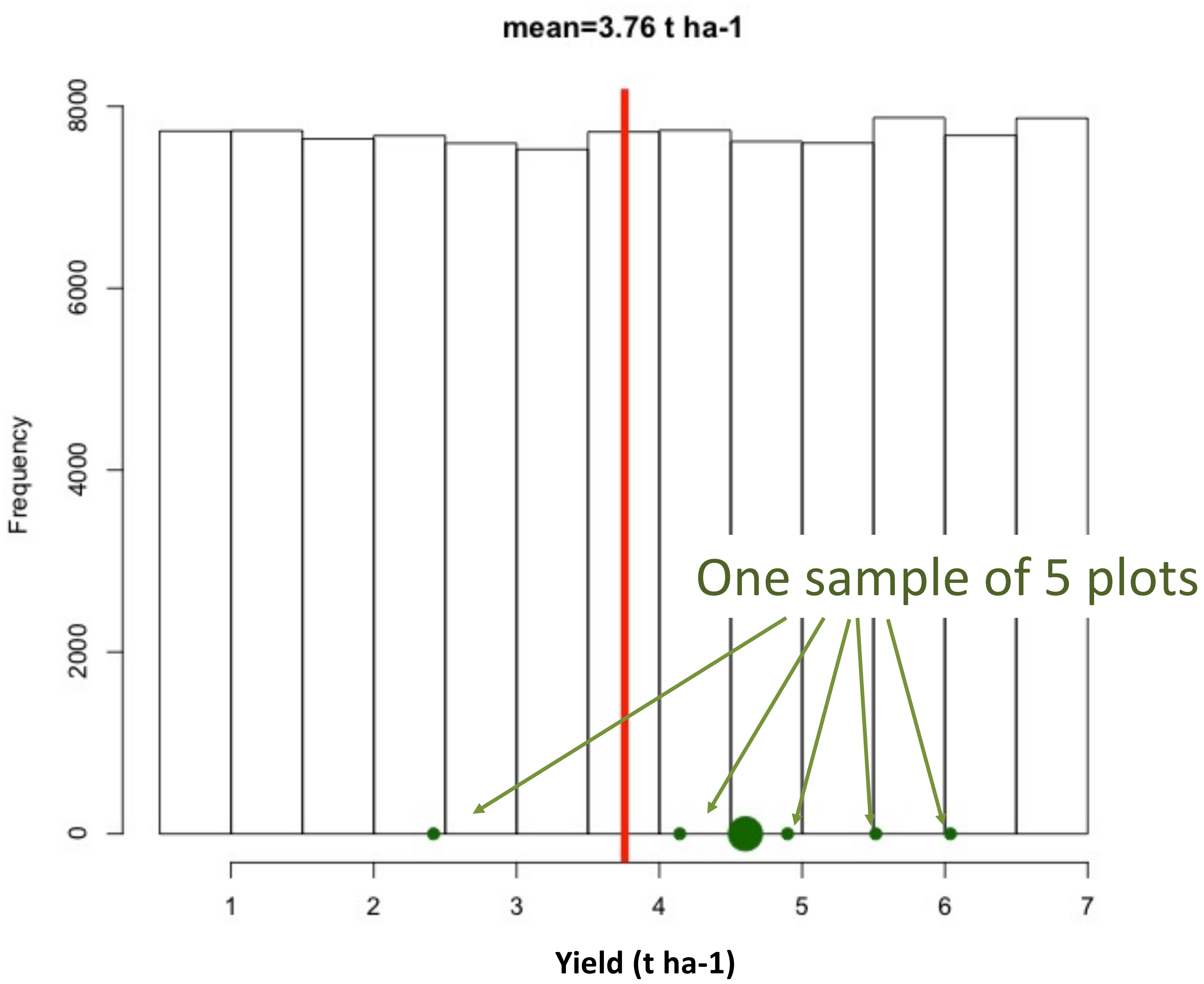
Central Limit theorem

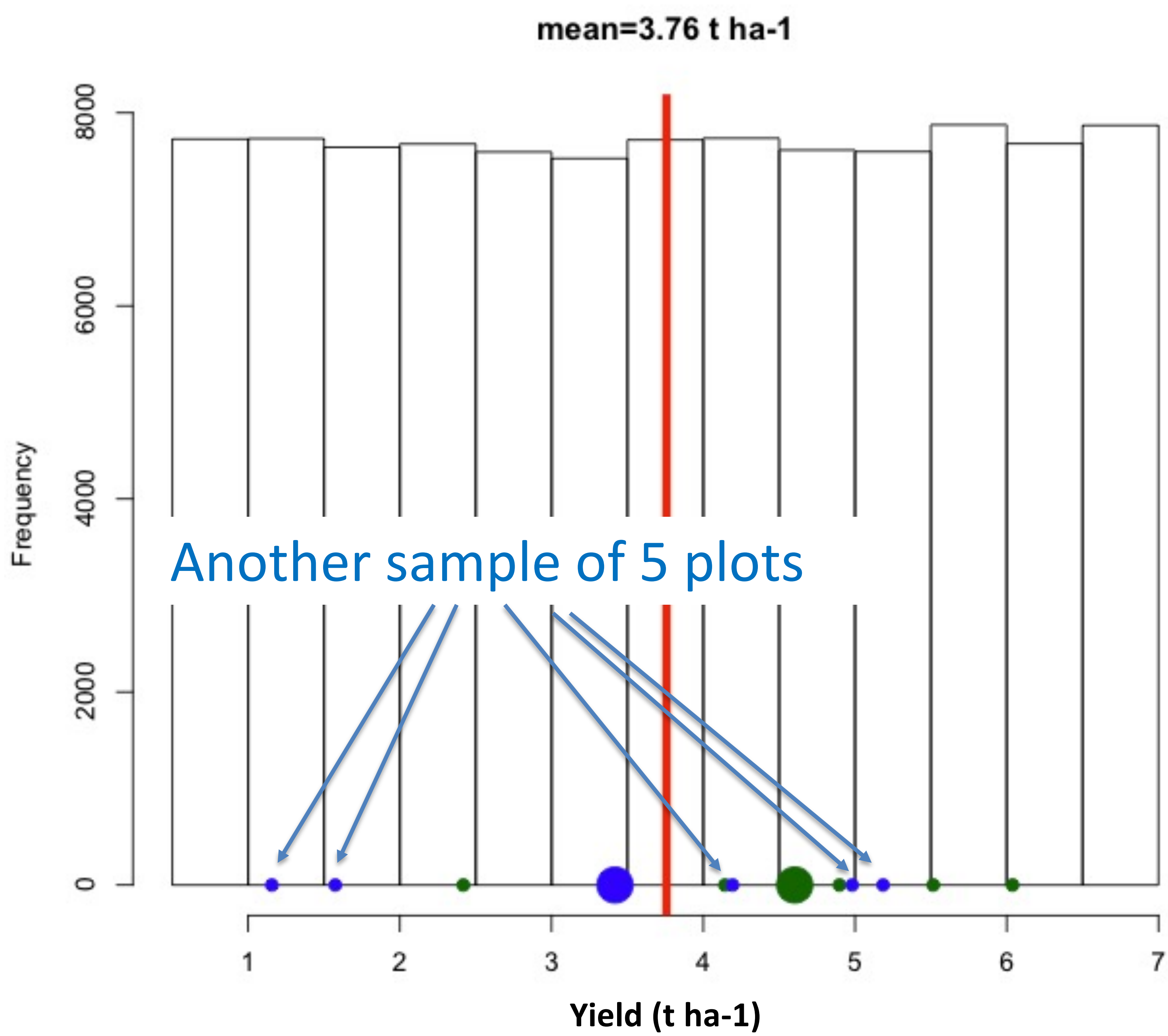
The distributions of the average of randomly chosen observations is closely approximated by a **normal distribution**

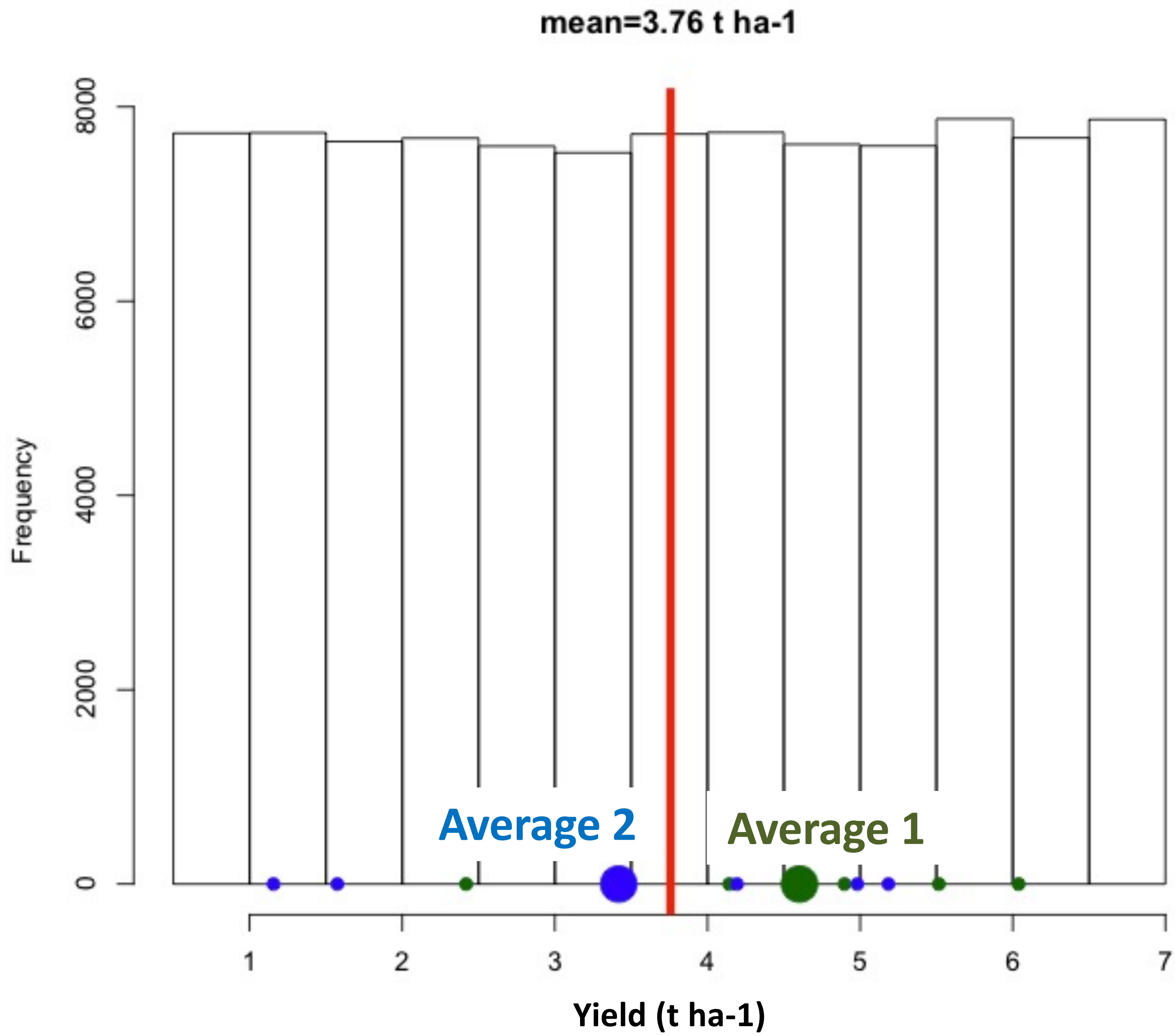
… even if the original observations themselves are not normally distributed.

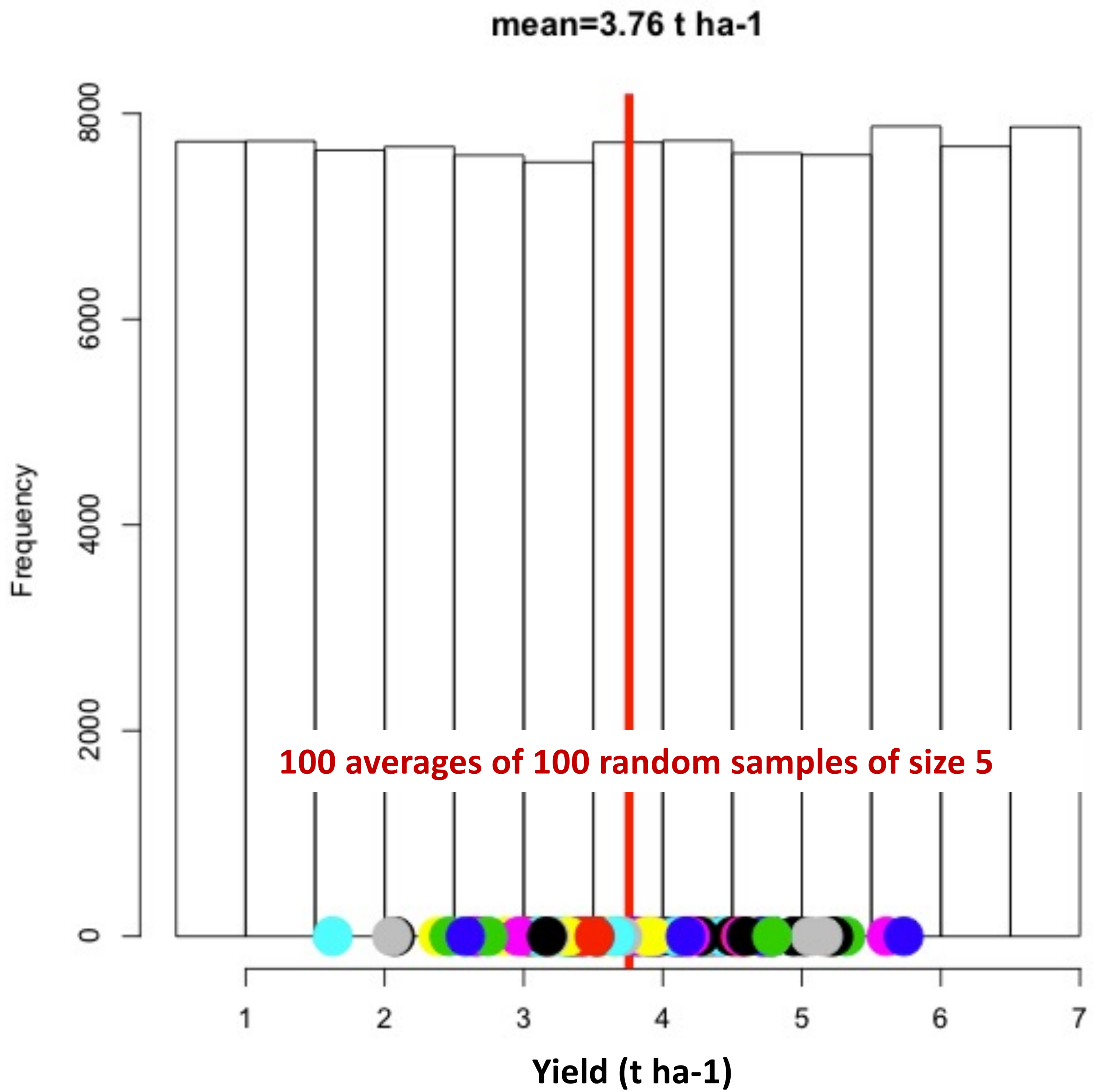
Population = 100,000 wheat plots

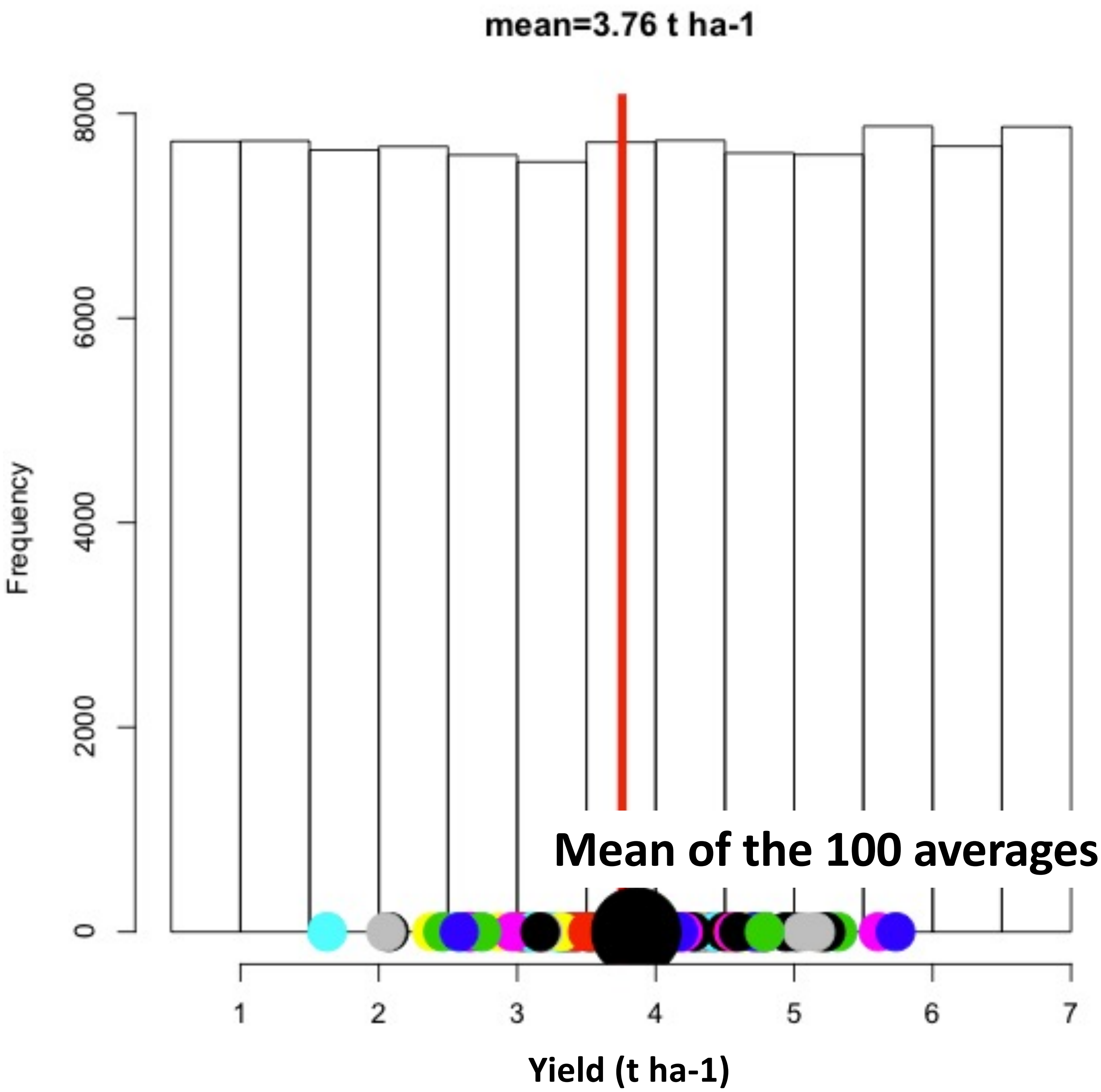


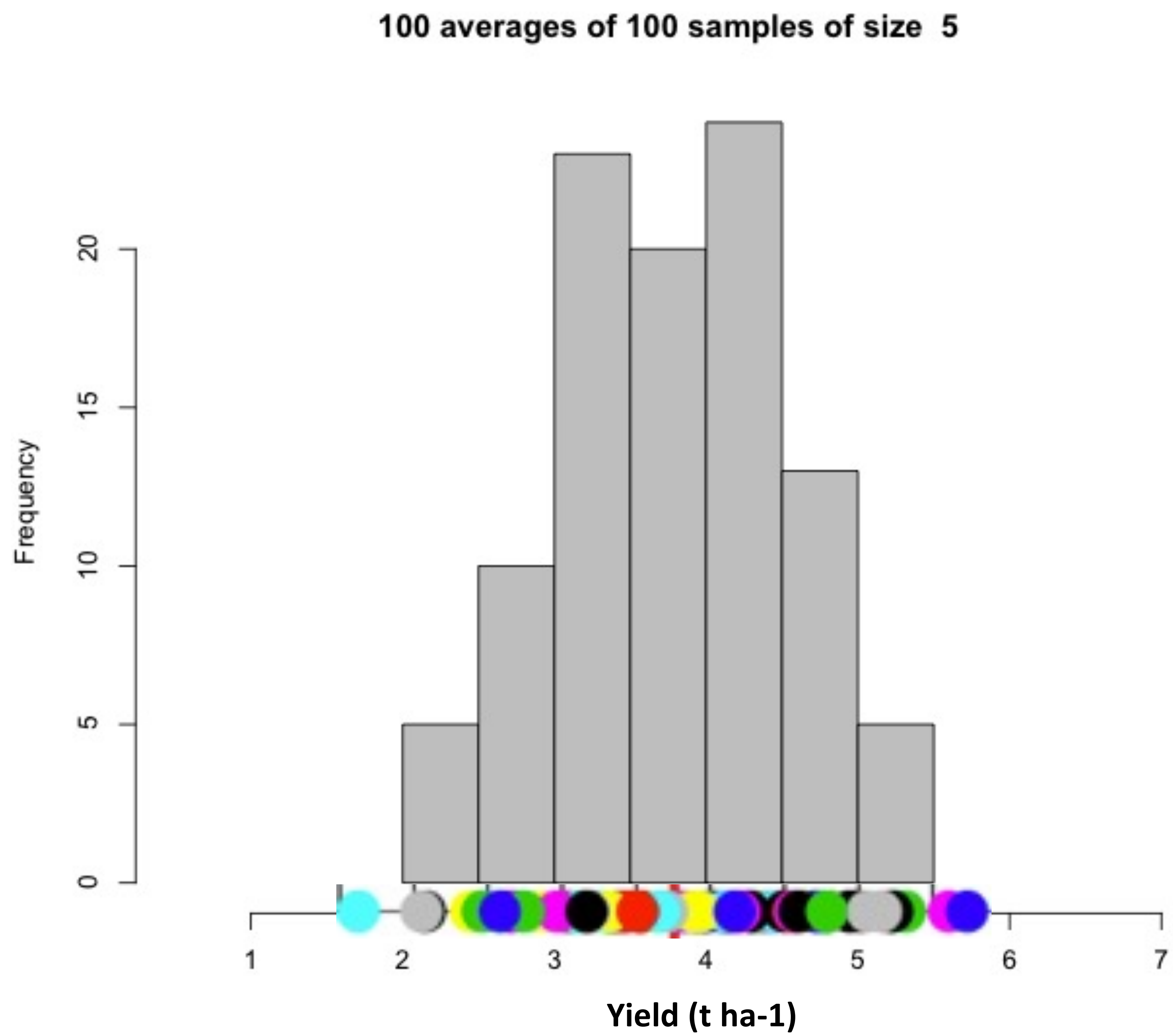

mean=3.76 t ha-1
Frequency
8000
6000
4000
2000
0
1
2
3
4
5
6
7
Yield (t ha-1)
One sample of 5 plots

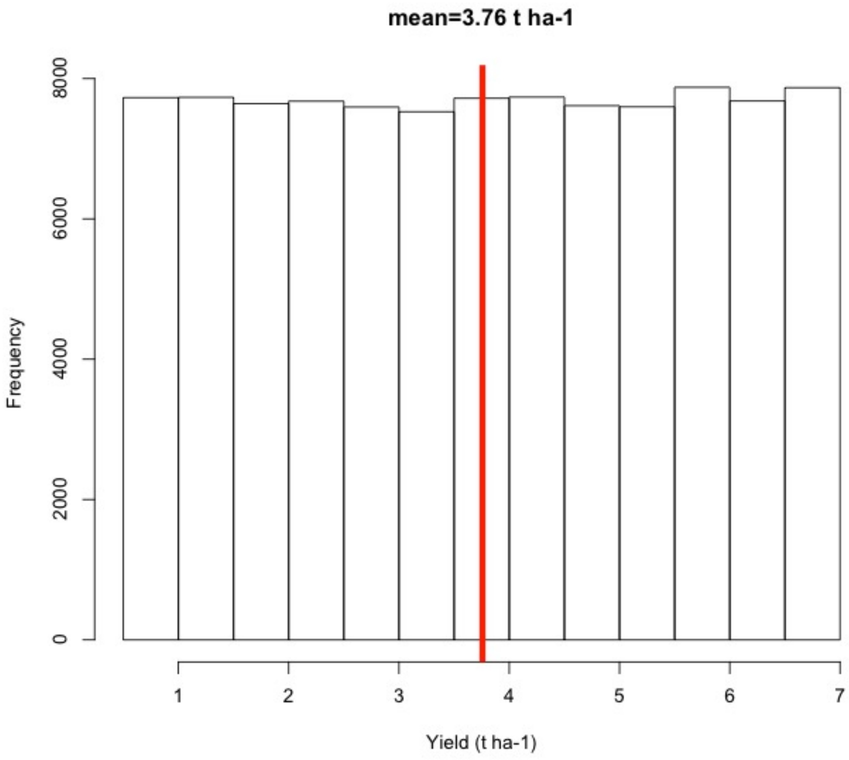

mean=3.76 t ha-1
Frequency
8000
6000
4000
2000
0
Another sample of 5 plots
1
2
3
4
5
6
7
Yield (t ha-1)


mean=3.76 t ha-1
Frequency
0
2000
4000
6000
8000
Average 2
Average 1
1
2
3
4
5
6
7
Yield (t ha-1)


mean=3.76 t ha-1
Frequency
0
2000
4000
6000
8000
1
2
3
4
5
6
7
Yield (t ha-1)
100 averages of 100 random samples of size 5

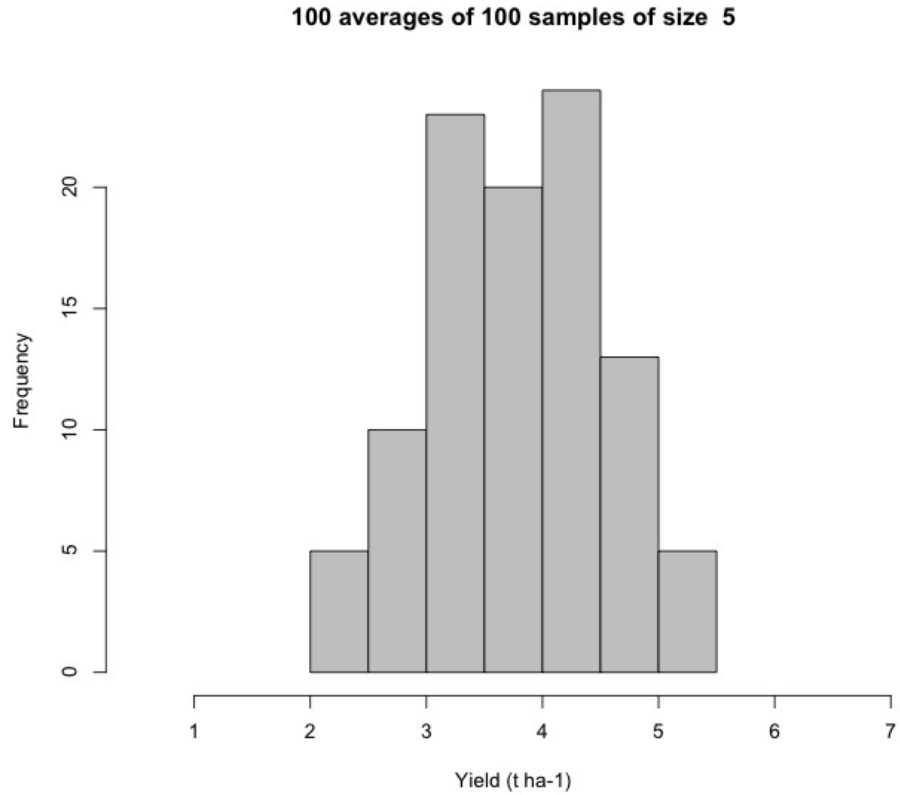

mean=3.76 t ha-1
8000
6000
4000
2000
0
Frequency
Mean of the 100 averages
1
2
3
4
5
6
7
Yield (t ha-1)

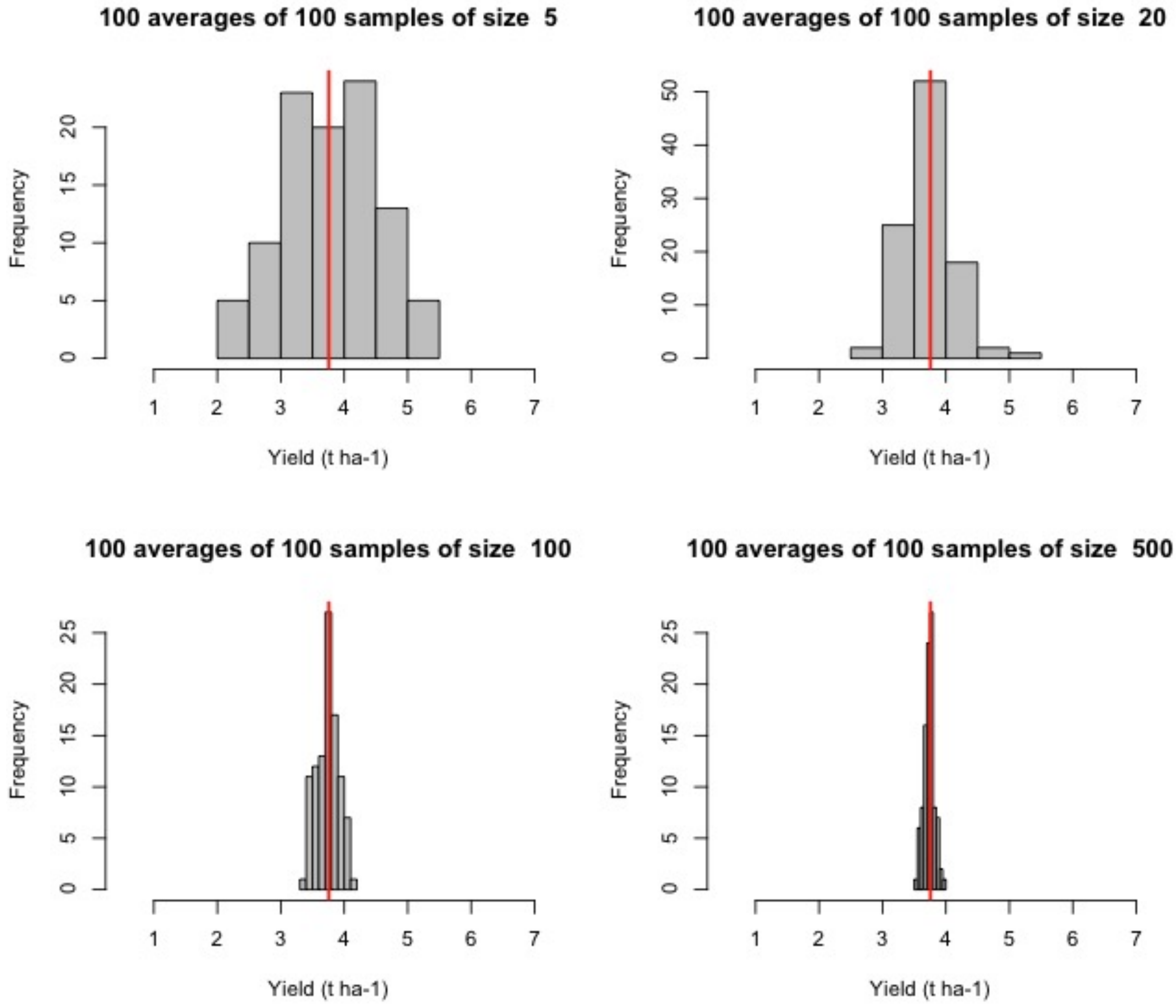

100 averages of 100 samples of size 5
Frequency
0
5
10
15
20
1
2
3
4
5
6
7
Yield (t ha-1)



100 samples of 5 yield data

100 average values




100 averages of 100 samples of size 5
100 averages of 100 samples of size 20
100 averages of 100 samples of size 100
100 averages of 100 samples of size 500
Frequency
Yield (t ha-1)

Key concepts

- **Population**
- **Sample**
- Estimator, estimate
- Bias and variance of an estimator
- Test
- Confidence interval
- Model

Key concepts

- **Population**
- **Sample**
- **Estimator, estimate**
- Bias and variance of an estimator
- Test
- Confidence interval
- Model

Estimator

A function of random variables that can be used in estimating unknown parameters of a theoretical probability distribution.

https://www.encyclopediaofmath.org/index.php/Statistical_estimator

Estimator

A rule used to calculate a quantity of interest from data

Estimator

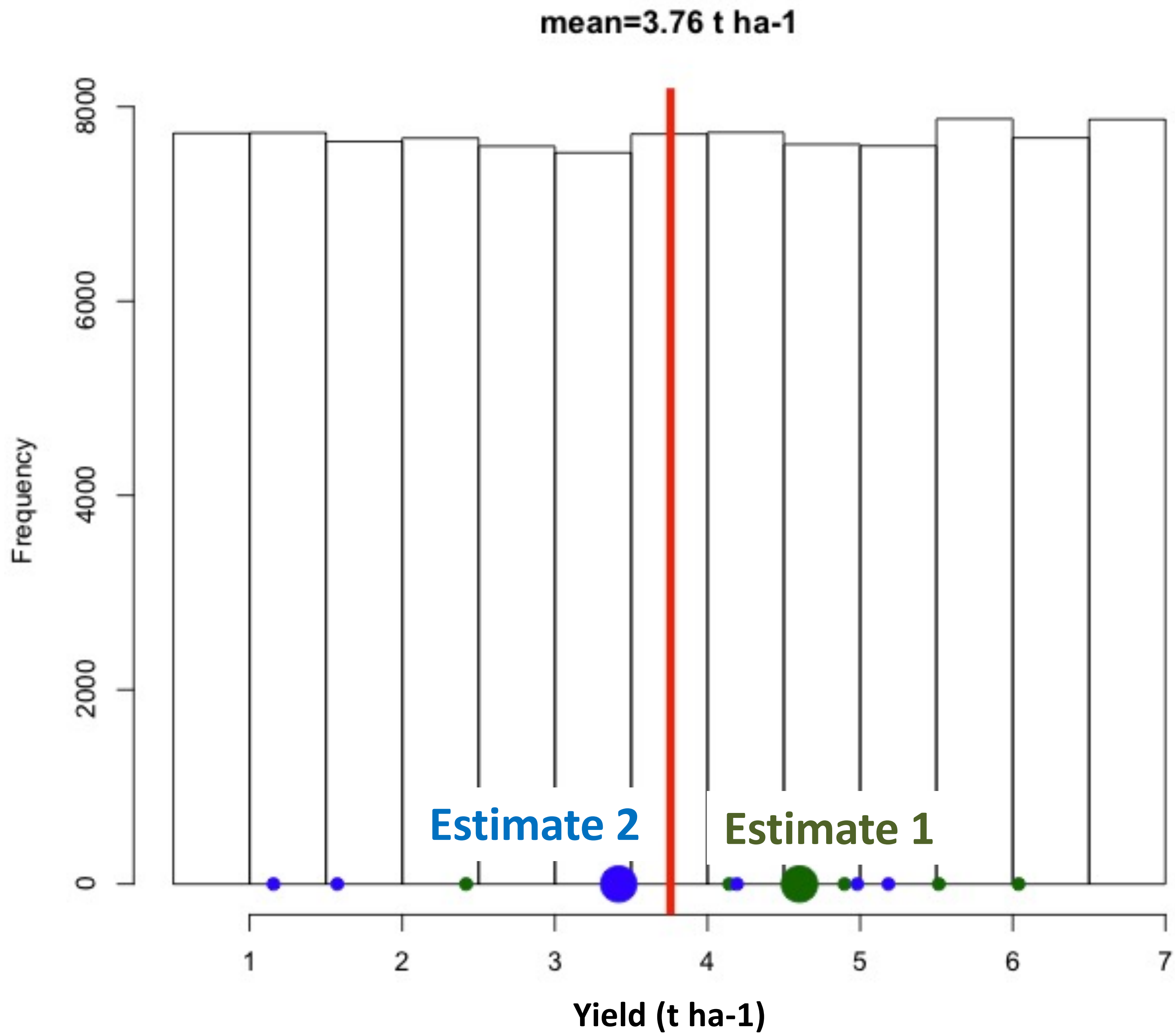
Example:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

Estimate

One value of an estimator calculated from one sample of data

$$\frac{1.1 + 2.8 + 5.8 + 6.1 + 0.8}{5}$$


mean=3.76 t ha-1
Frequency
0
2000
4000
6000
8000
Estimate 2
Estimate 1
1
2
3
4
5
6
7
Yield (t ha-1)

Bias and variance of an estimator

Bias = difference between the true value and the mean value of the estimator

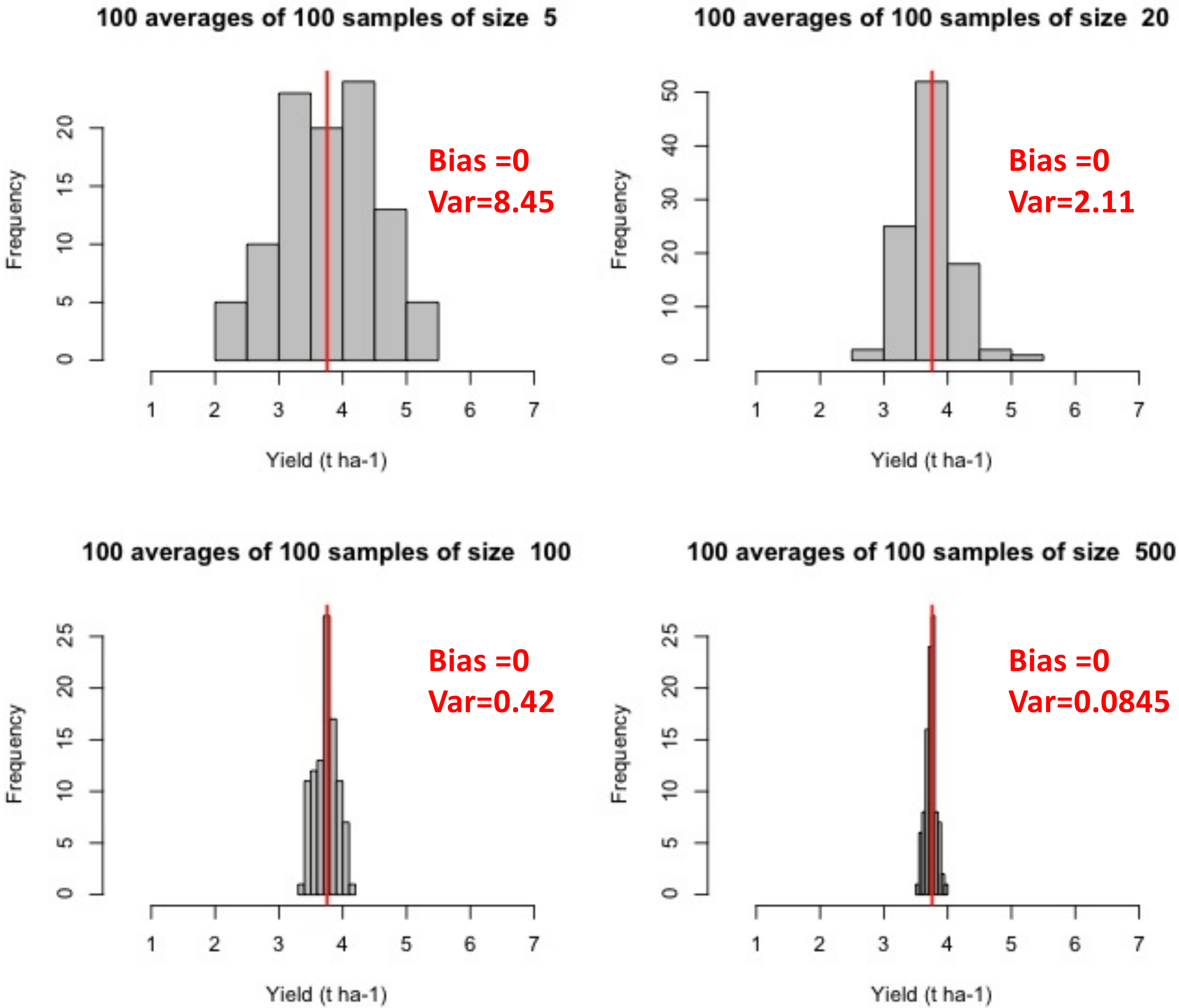
Variance = measure of the dispersion of the estimator around its mean value

Standard deviation = $\sqrt{\mathbf{Variance}}$

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

$$E(\bar{X}) = \frac{E(X_1)+E(X_2)+E(X_3)+E(X_4)+E(X_5)}{5} = \mathrm{E}(X) = 3.76$$

$$V(\bar{X}) = \frac{1}{5}V(X) = 8.45$$


100 averages of 100 samples of size 5
Bias =0
Var=8.45
Frequency
Yield (t ha-1)
100 averages of 100 samples of size 20
Bias =0
Var=2.11
Frequency
Yield (t ha-1)
100 averages of 100 samples of size 100
Bias =0
Var=0.42
Frequency
Yield (t ha-1)
100 averages of 100 samples of size 500
Bias =0
Var=0.0845
Frequency
Yield (t ha-1)

Key concepts

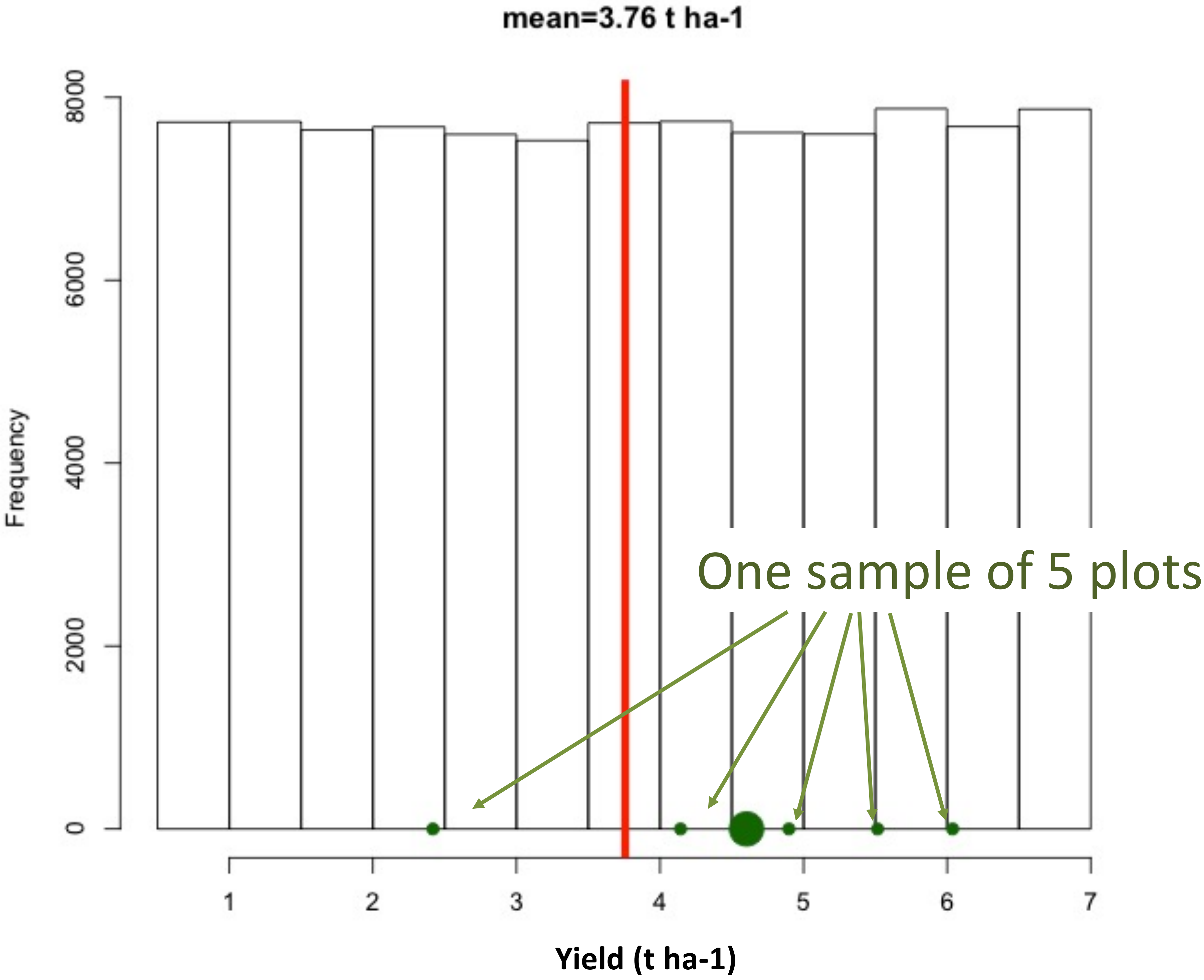
- **Population**
- **Sample**
- **Estimator, estimate**
- **Bias and variance of an estimator**
- Test
- Confidence interval
- Model

Key concepts

- **Population**
- **Sample**
- **Estimator, estimate**
- **Bias and variance of an estimator**
- **Test**
- Confidence interval
- Model

Test

Choose between two hypotheses based on a sample of observations


mean=3.76 t ha-1
8000
6000
4000
2000
0
Frequency
One sample of 5 plots
1
2
3
4
5
6
7
Yield (t ha-1)

A first example

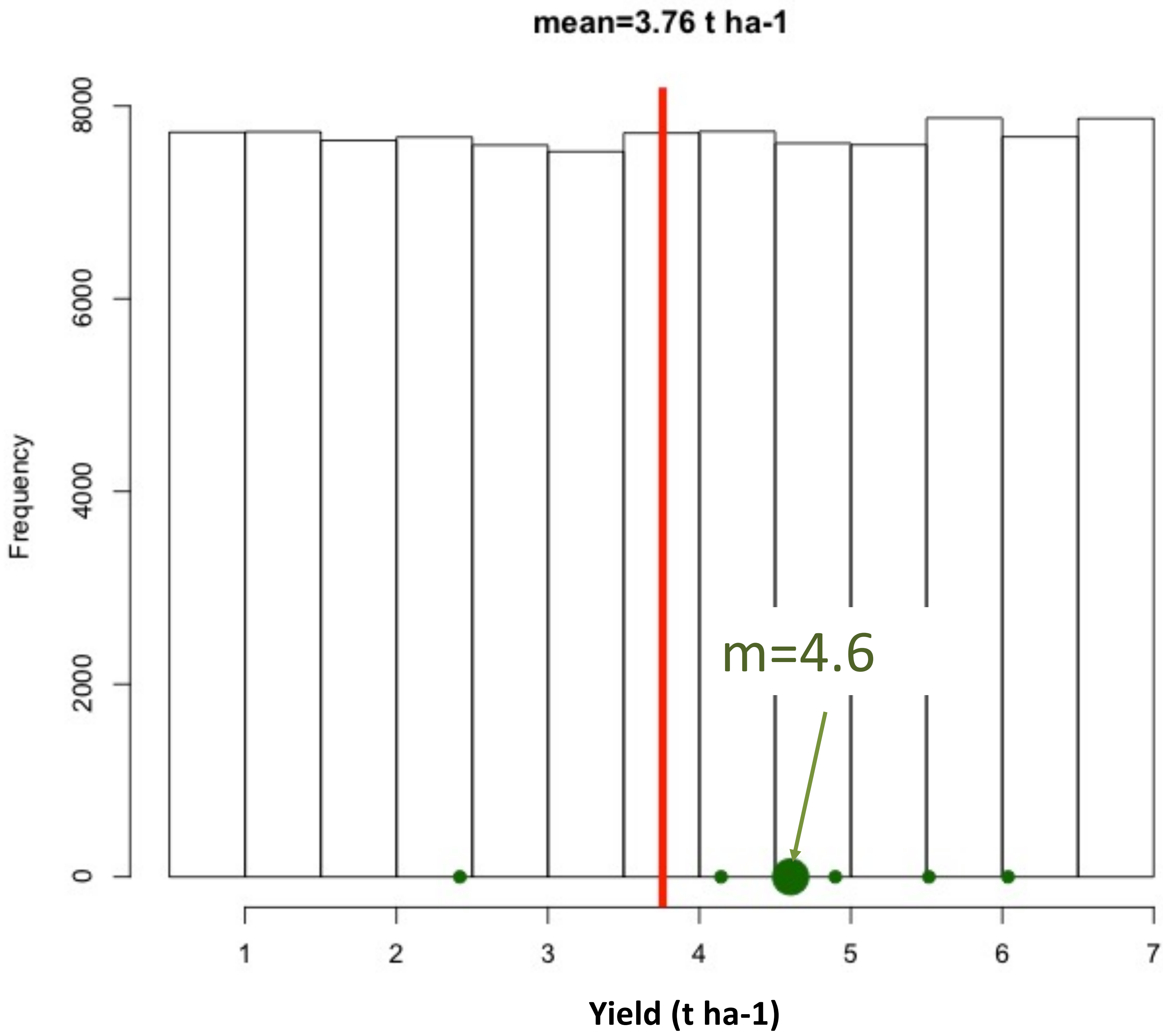
H0: true mean < 4

H1: true mean > 4

How to choose?

A naive test

If the sample mean m > 4, reject H0


mean=3.76 t ha-1
Frequency
0
2000
4000
6000
8000
1
2
3
4
5
6
7
m=4.6
Yield (t ha-1)

A naive test

If the sample mean m > 4, reject H0

H0 rejected

Error of decision

A naive test

If the sample mean m > 4, reject H0

H0 rejected

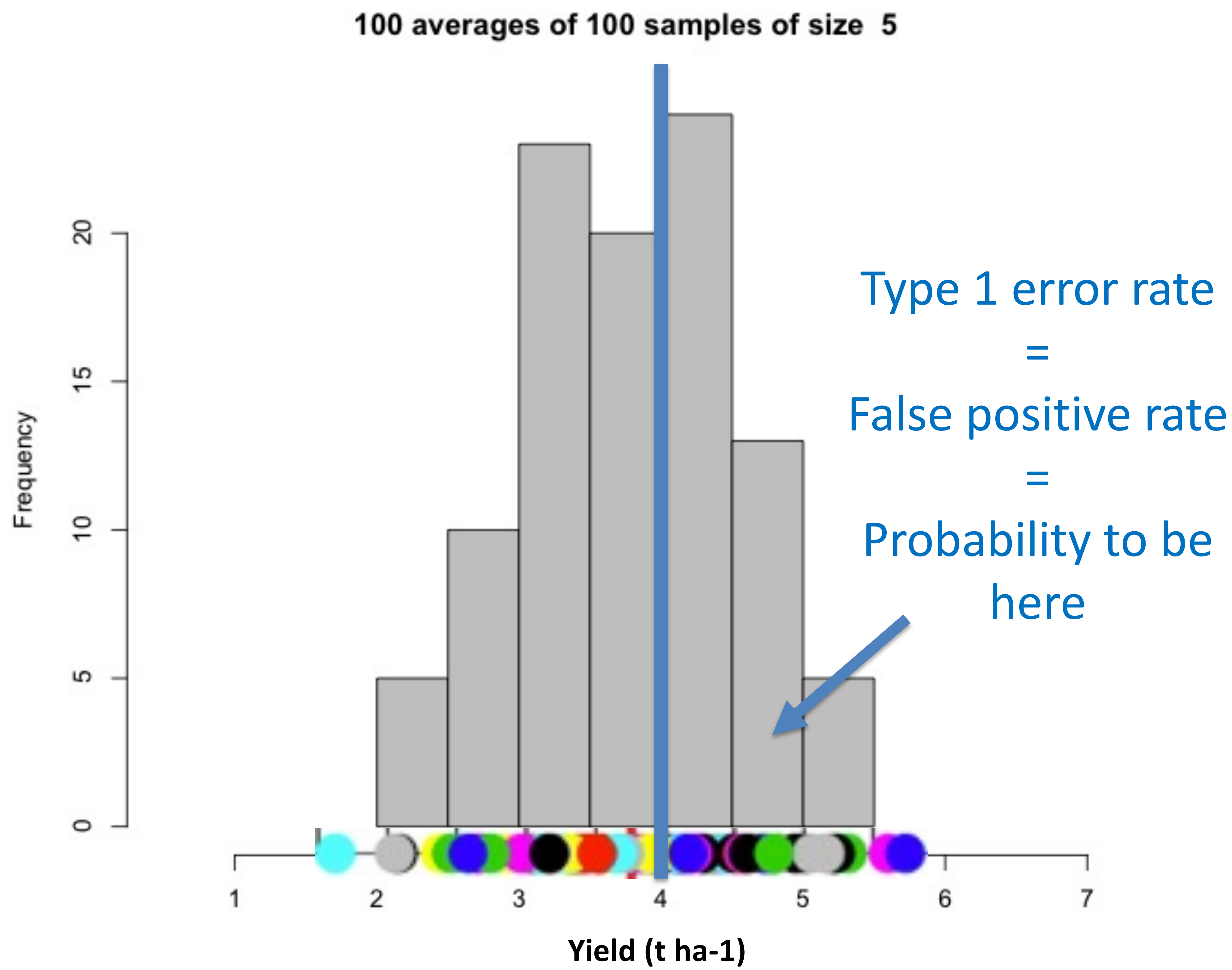
Error of decision: we reject H0 while H0 is true

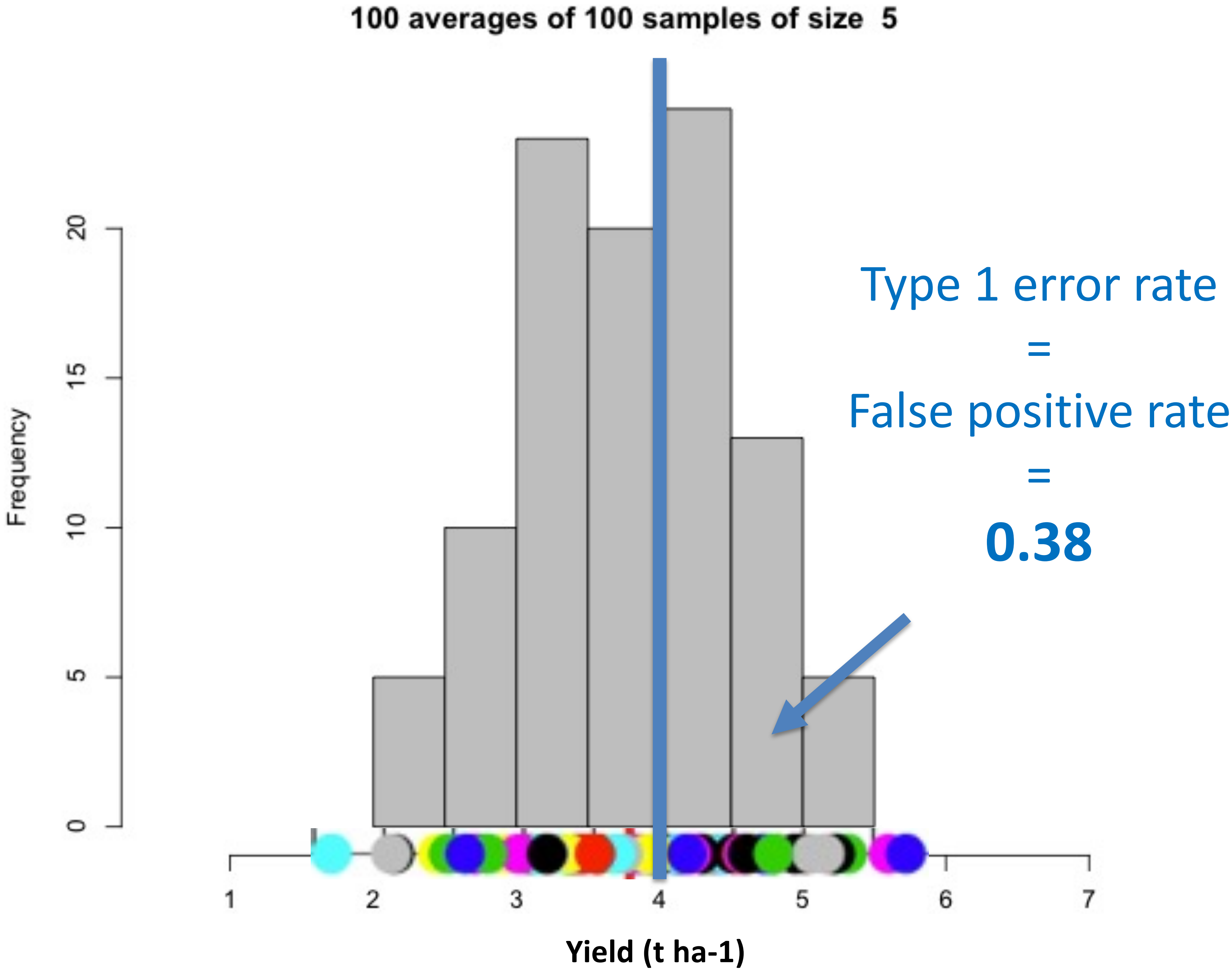
-> False positive

-> Type 1 error

What is the false positive rate of our naive test?

If the sample mean $m > 4$, reject H0


100 averages of 100 samples of size 5
Frequency
0
5
10
15
20
1
2
3
4
5
6
7
Yield (t ha-1)
Type 1 error rate
=
False positive rate
=
Probability to be
here


100 averages of 100 samples of size 5
Frequency
0
5
10
15
20
1
2
3
4
5
6
7
Yield (t ha-1)
Type 1 error rate
=
False positive rate
=
0.38

A second example

H0: true mean < 5

H1: true mean > 5

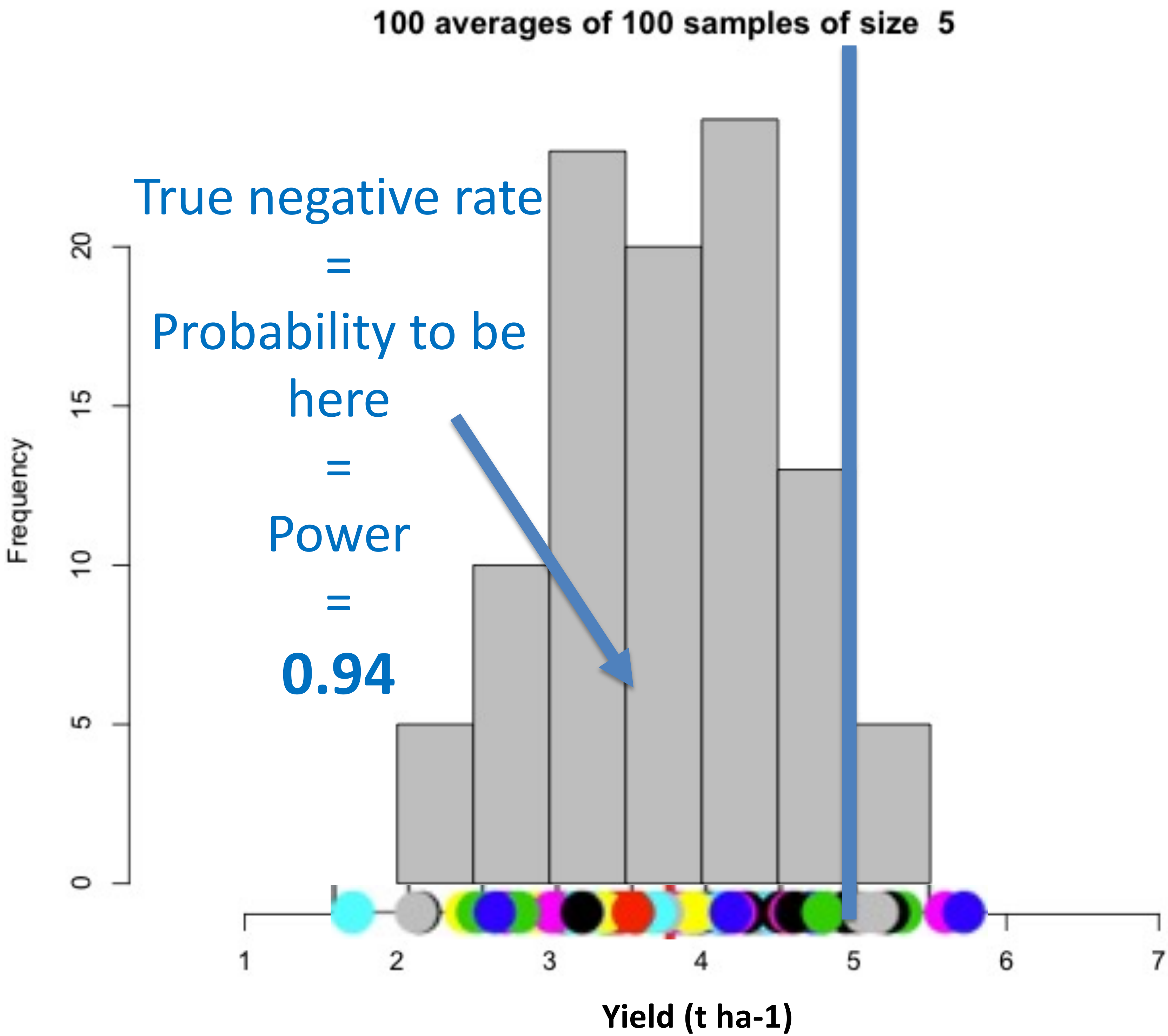
How to choose?

A naive test

If the sample mean m > 5, reject H0

As here m=4.6, we accept H0

No error of decision here: True negative


100 averages of 100 samples of size 5
True negative rate
=
Probability to be
here
=
Power
=
0.94
Frequency
0
5
10
15
20
1
2
3
4
5
6
7
Yield (t ha-1)

Two types of error

- Type 1: Reject H0 while H0 true
  - False positive rate
  - Alpha risk

- Type 2: Accept H0 while H0 wrong
  - False negative rate
  - Beta risk
  - Equal to 1-Power

Two types of error

A good test is a test with

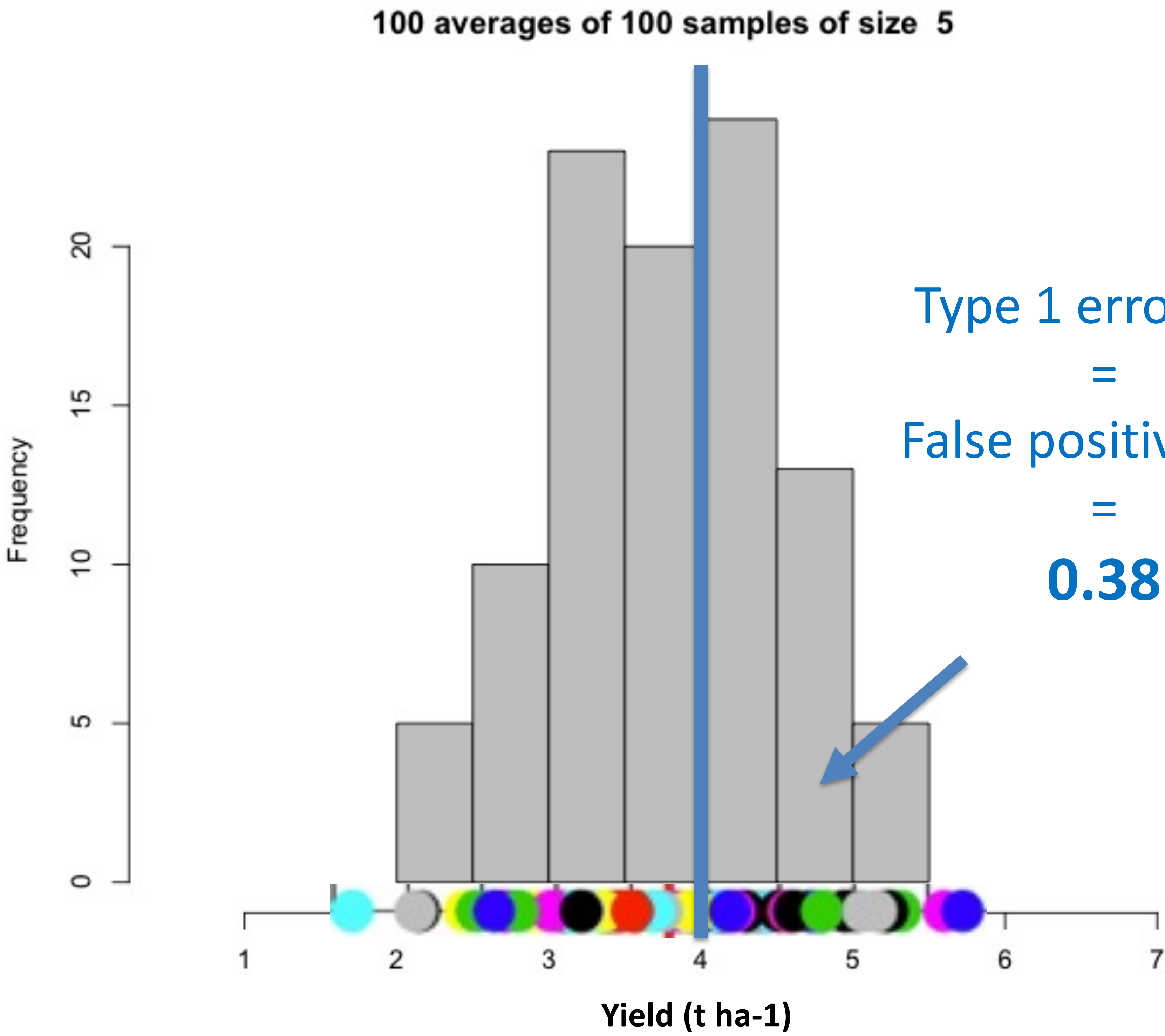
- A small type 1 error rate
- A small type 2 error rate (i.e., a high power)

First example

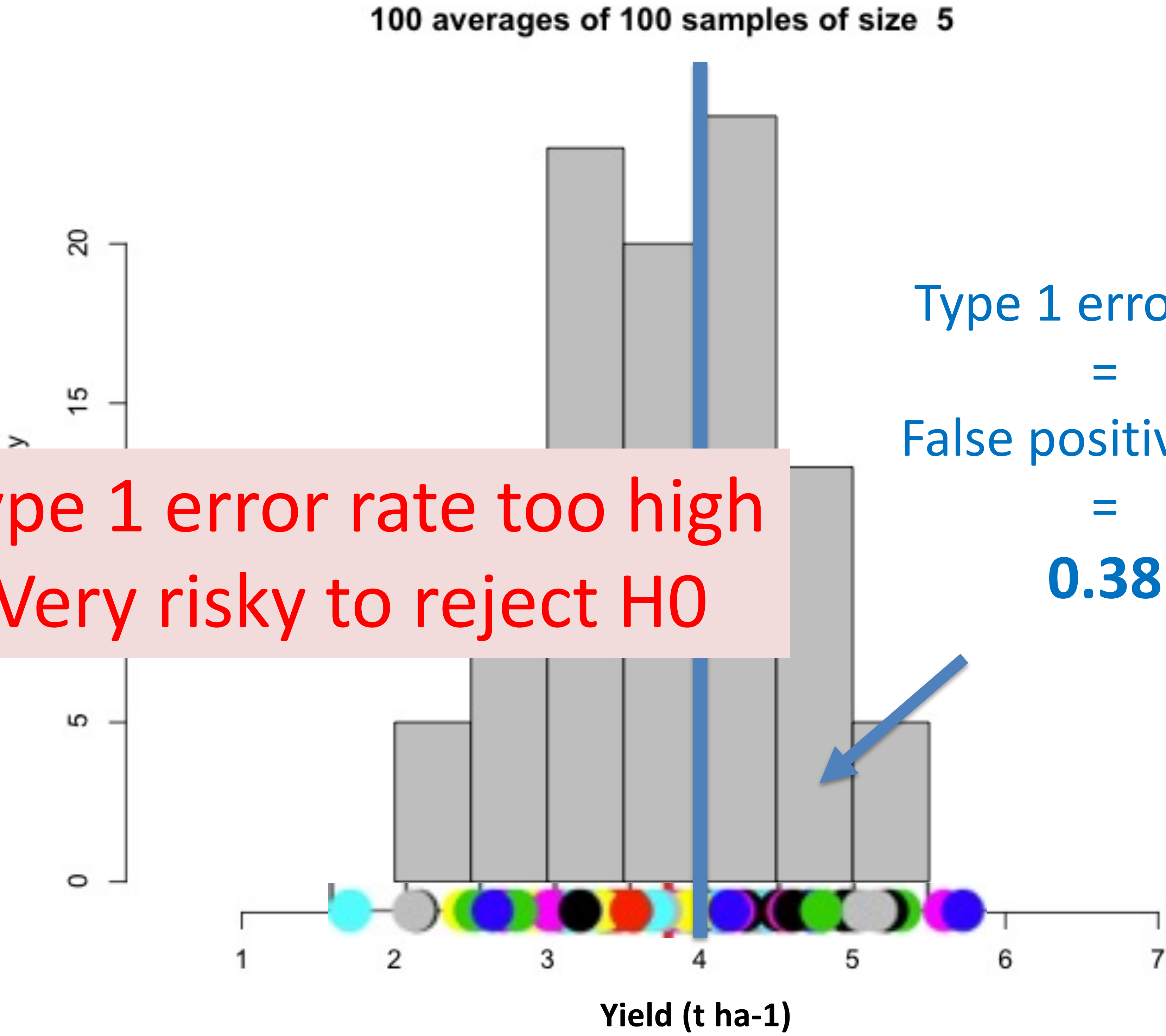
H0: true mean < 4

H1: true mean > 4

How to choose?



Type 1 error rate
=
False positive rate
=
0.38



Type 1 error rate too high
Very risky to reject H0

Type 1 error rate
=
False positive rate
=
0.38

First example

H0: true mean < 4

H1: true mean > 4

How to choose?

A better test

Define $T = (m-4)/s$

m = sample mean

s = standard error

T measures how far the value of *m* is from 4

If T is large enough, we reject H0

A better test

Define $T = (m-4)/s$

m = sample mean

s = standard error

A better test

Define $T = (m-4)/s$

m = sample mean

s = standard error

= standard deviation/sqrt(sample size)

A better test

Define T = (m-4)/s

m = sample mean

s = standard error

= standard deviation/sqrt(sample size)

$$m = \frac{X1 + X2 + X3 + X4 + X5}{5}$$

$$s = \sqrt{\frac{1}{5}\frac{(X1-m)^2+(X2-m)^2+(X3-m)^2+(X4-m)^2+(X5-m)^2}{5-1}}$$

A better test

Define T = (m-4)/s

m = sample mean

s = standard error

= standard deviation/sqrt(sample size)

$$m = \frac{X1 + X2 + X3 + X4 + X5}{5}$$

$$s= \sqrt{\frac{1}{5}\frac{(X1-m)^2+(X2-m)^2+(X3-m)^2+(X4-m)^2+(X5-m)^2}{5-1}}$$

If T> K, reject H0

How to choose K?

- Set a max acceptable value for the type 1 error rate
    Ex: 0.05 i.e., 5%

- Choose K in order to stay below this value according to some probability distribution, here, the *student distribution*

t test

Define $T = (m-4)/s$

m = sample mean
s= standard error

If $T >$ 95% quantile of a student distribution, reject H0

t test with R

```
> Y=c(4.9,4.15,6.3,2.4,5.5)
> Y
[1] 4.90 4.15 6.30 2.40 5.50
> t.test(x=Y,mu=4,alternative="greater")

	One Sample t-test

data:  Y
t = 0.9788, df = 4, p-value = 0.1915
alternative hypothesis: true mean is greater than 4
95 percent confidence interval:
 3.234287      Inf
sample estimates:
mean of x
     4.65
```

t test with R

```
> Y=c(4.9,4.15,6.3,2.4,5.5)
> Y
[1] 4.90 4.15 6.30 2.40 5.50
> t.test(x=Y,mu=4,alternative="greater")

	One Sample t-test

data:  Y
t = 0.9788, df = 4, p-value = 0.1915
alternative hypothesis: true mean is greater than 4
95 percent confidence interval:
 3.234287      Inf
sample estimates:
mean of x
     4.65
```

(mean(Y)-4)/(sd(Y)/sqrt(5))

mean(Y)

t test with R

```
> Y=c(4.9,4.15,6.3,2.4,5.5)
> Y
[1] 4.90 4.15 6.30 2.40 5.50
> t.test(x=Y,mu=4,alternative="greater")

	One Sample t-test

data:  Y
t = 0.9788, df = 4, p-value = 0.1915
alternative hypothesis: true mean is greater than 4
95 percent confidence interval:
 3.234287      Inf
sample estimates:
mean of x
     4.65
```

Type 1 error rate

p value >5%

Too risky to reject H0

Exercise

Five yield data: 1.2, 4.2, 5.0, 5.2, 1.6

H0: True mean <1 t ha-1

H1: True mean > 1 t ha-1

Use a t test to test this hypothesis

Exercise

Five yield data: 1.2, 4.2, 5.0, 5.2, 1.6

H0: True mean <2 t ha-1

H1: True mean > 2 t ha-1

Use a t test to test this hypothesis

Exercise

Five yield data: 1.2, 4.2, 5.0, 5.2, 1.6

H0: True mean >6 t ha-1

H1: True mean <6 t ha-1

Use a t test to test this hypothesis

Key concepts

- **Population**
- **Sample**
- **Estimator, estimate**
- **Bias and variance of an estimator**
- **Test**
- Confidence interval
- Model

Key concepts

- **Population**
- **Sample**
- **Estimator, estimate**
- **Bias and variance of an estimator**
- **Test**
- **Confidence interval**
- Model

Confidence interval

Range of values that contains the true value with a certain probability

95% confidence interval of a mean

IC95 = [L, U]

P(L< True mean< U) = 0.95

L and U are calculated from the sample of data

Example

```
> Y
[1] 4.90 4.15 6.30 2.40 5.50
> t.test(Y,conf.level=0.95)

    One Sample t-test

data:  Y
t = 7.0022, df = 4, p-value = 0.00219
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.806223 6.493777
sample estimates:
mean of x
     4.65
```

L U

Exercise

Five yield data: 1.2, 4.2, 5.0, 5.2, 1.6

Ten yield data: 1.2, 4.2, 5.0, 5.2, 1.6, 2.8, 3.4, 6.1, 4.1, 3.2

Calculate

- 95% confidence interval with 5 and 10 data
- 90% confidence interval with 5 and 10 data

Key concepts

- **Population**
- **Sample**
- **Estimator, estimate**
- **Bias and variance of an estimator**
- **Test**
- **Confidence interval**
- Model

Key concepts

- **Population**
- **Sample**
- **Estimator, estimate**
- **Bias and variance of an estimator**
- **Test**
- **Confidence interval**
- **Model**

What is a statistical model?

- A particular type of mathematical model

- A model including measurable components

… and unmeasurable components

- Some of the model components are defined as random variables

What is a statistical model?

Parameters

$$Y = f\left(X, \theta, \varepsilon\right)$$

Outputs

Inputs

Residuals

What is a **linear** statistical model?

Parameters

$$Y = X\theta + \varepsilon$$

Outputs

Inputs

Residuals

$$
\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_P \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}
$$

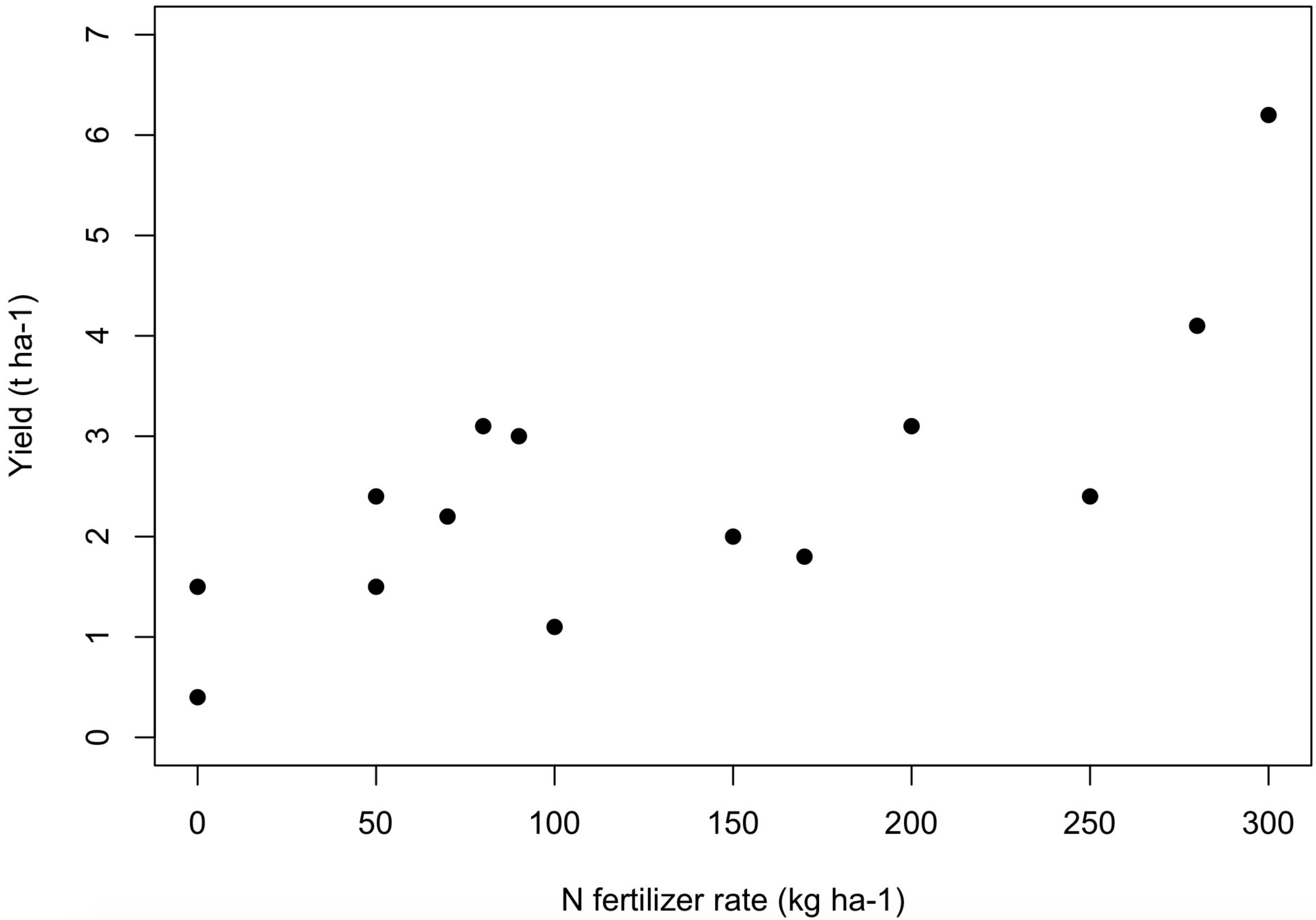
$$
y_2 = x_{21}\theta_1 + x_{22}\theta_2 + \dots + x_{2P}\theta_P + \varepsilon_2
$$

Applications

- Test whether an output (*Y)* is related to one or several inputs (*X*)
  → Statistical test

- Quantify effect of input X on output Y
  → Estimation and confidence interval

- Predict *Y* as a function of *X*
  → Prediction

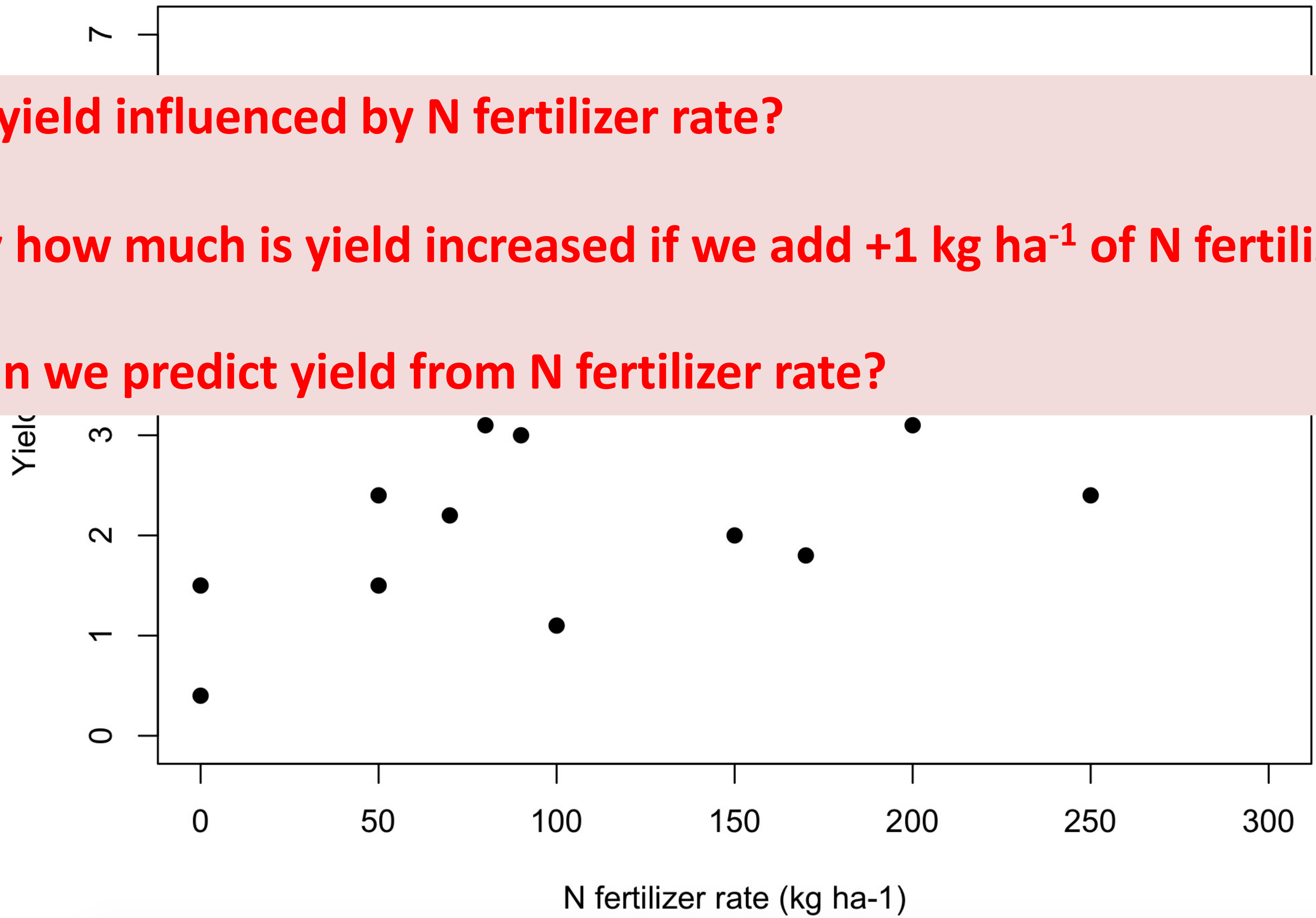


Yield (t ha-1)
0
1
2
3
4
5
6
7
0
50
100
150
200
250
300
N fertilizer rate (kg ha-1)

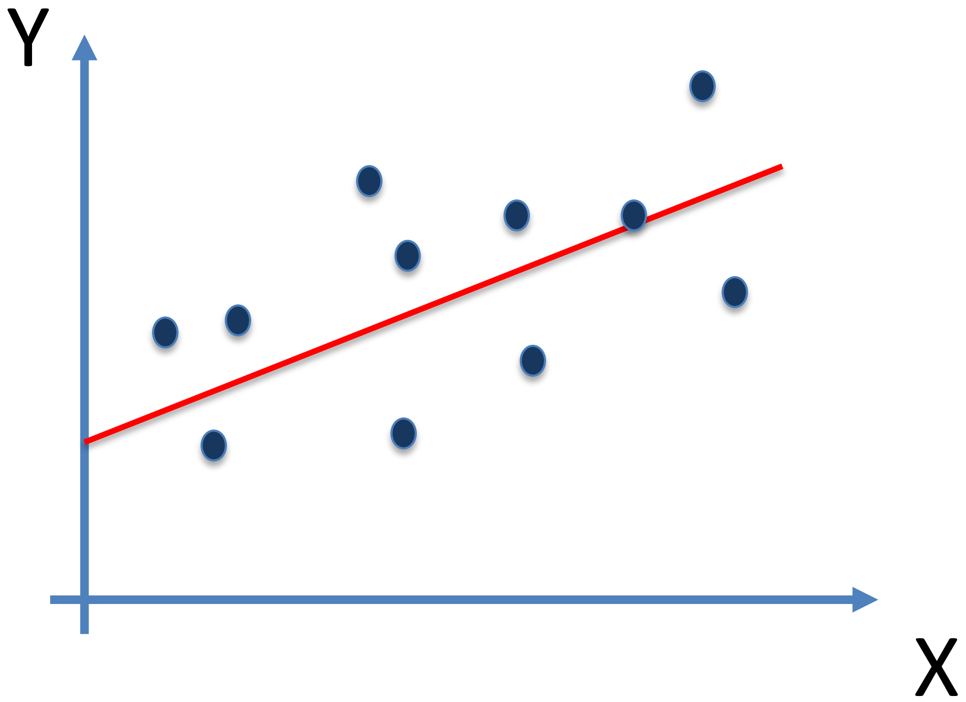
Is yield influenced by N fertilizer rate?

By how much is yield increased if we add +1 kg ha^{-1} of N fertilizer ?

Can we predict yield from N fertilizer rate?



$$Y = \alpha + \beta X + \varepsilon$$



Estimation

Use estimators to compute parameter values from a sample of data

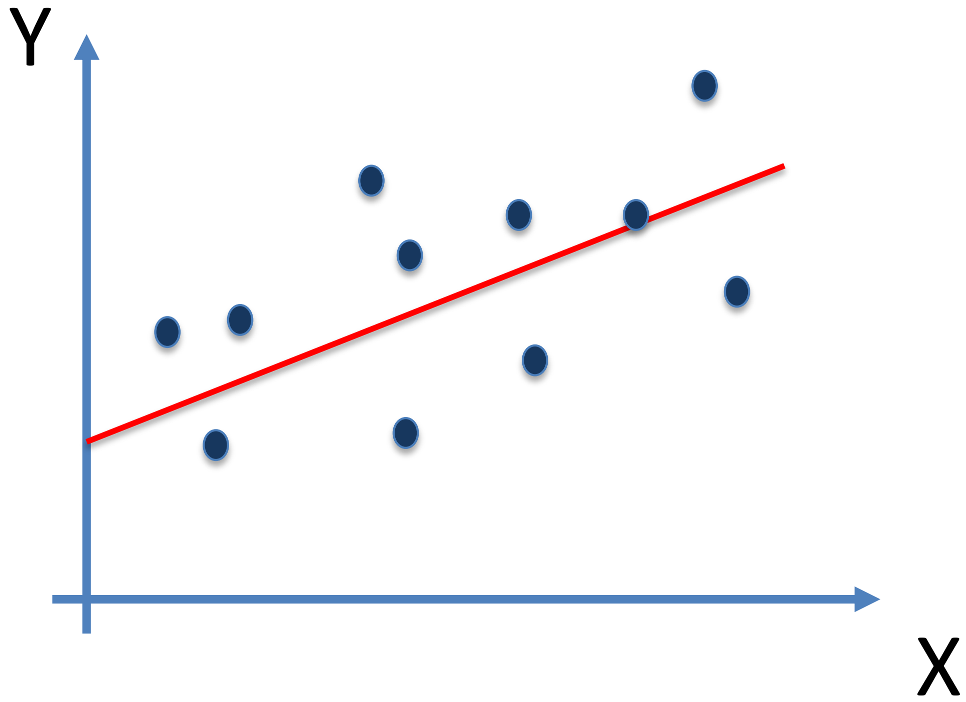
Classic estimators: Ordinary least squares

- Unbiaised
- With small variances (under some assumptions)

Ordinary least squares

Estimate the parameters by minimizing

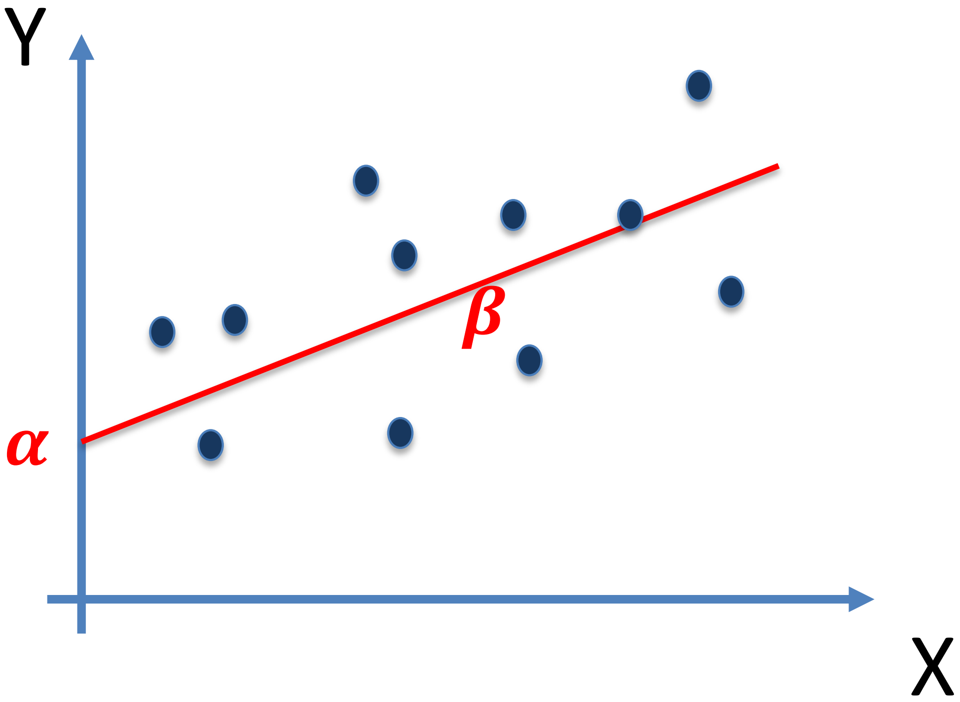
$$OLS = \sum_{i=1}^{N}\left[y_i - \left(\alpha + \beta x_i\right)\right]^2$$



Ordinary least squares

Estimate the parameters by minimizing

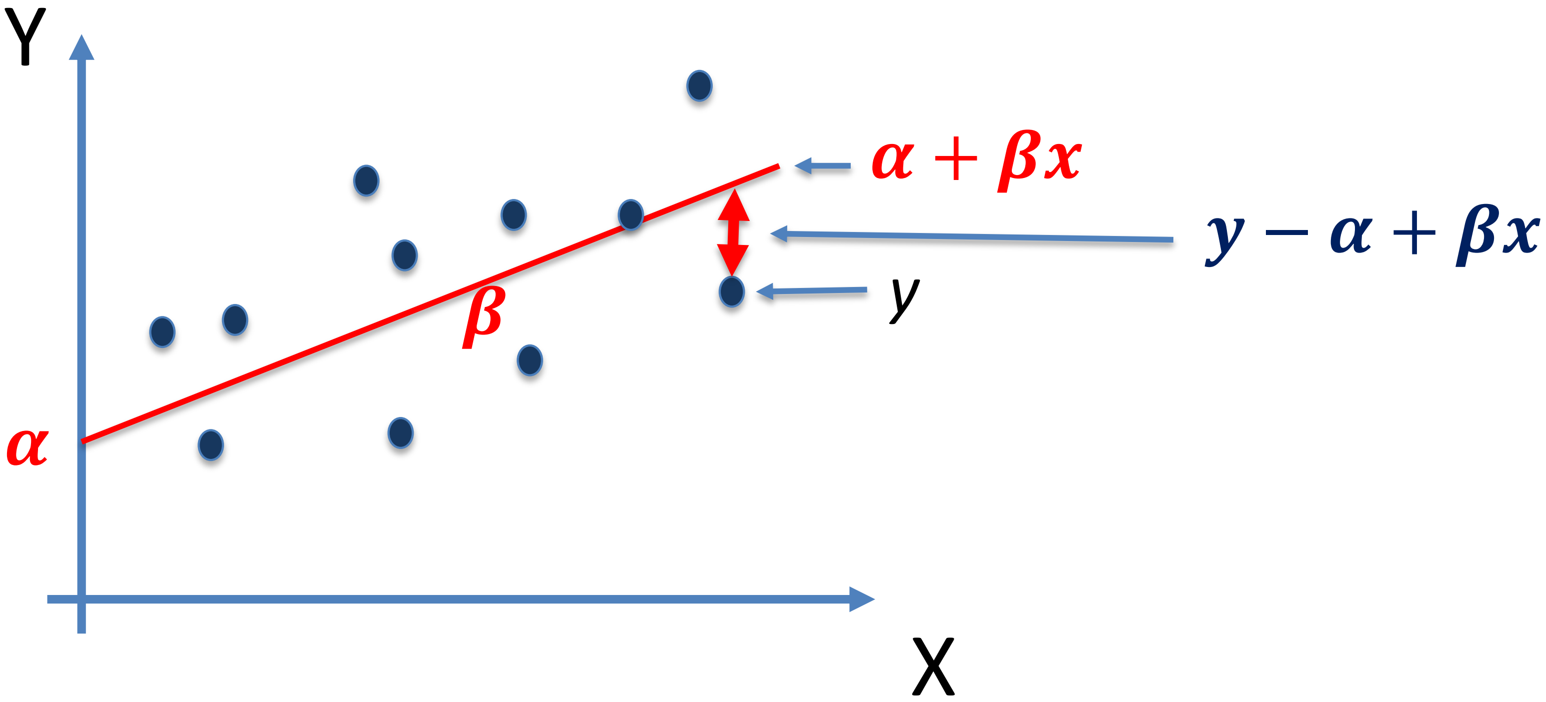
$$OLS = \sum_{i=1}^{N} \left[y_i - \left(\alpha + \beta x_i \right) \right]^2$$



Ordinary least squares

Estimate the parameters by minimizing

$$OLS = \sum_{i=1}^{N}\left[y_i - \left(\alpha + \beta x_i\right)\right]^2$$



Function « lm() » de R

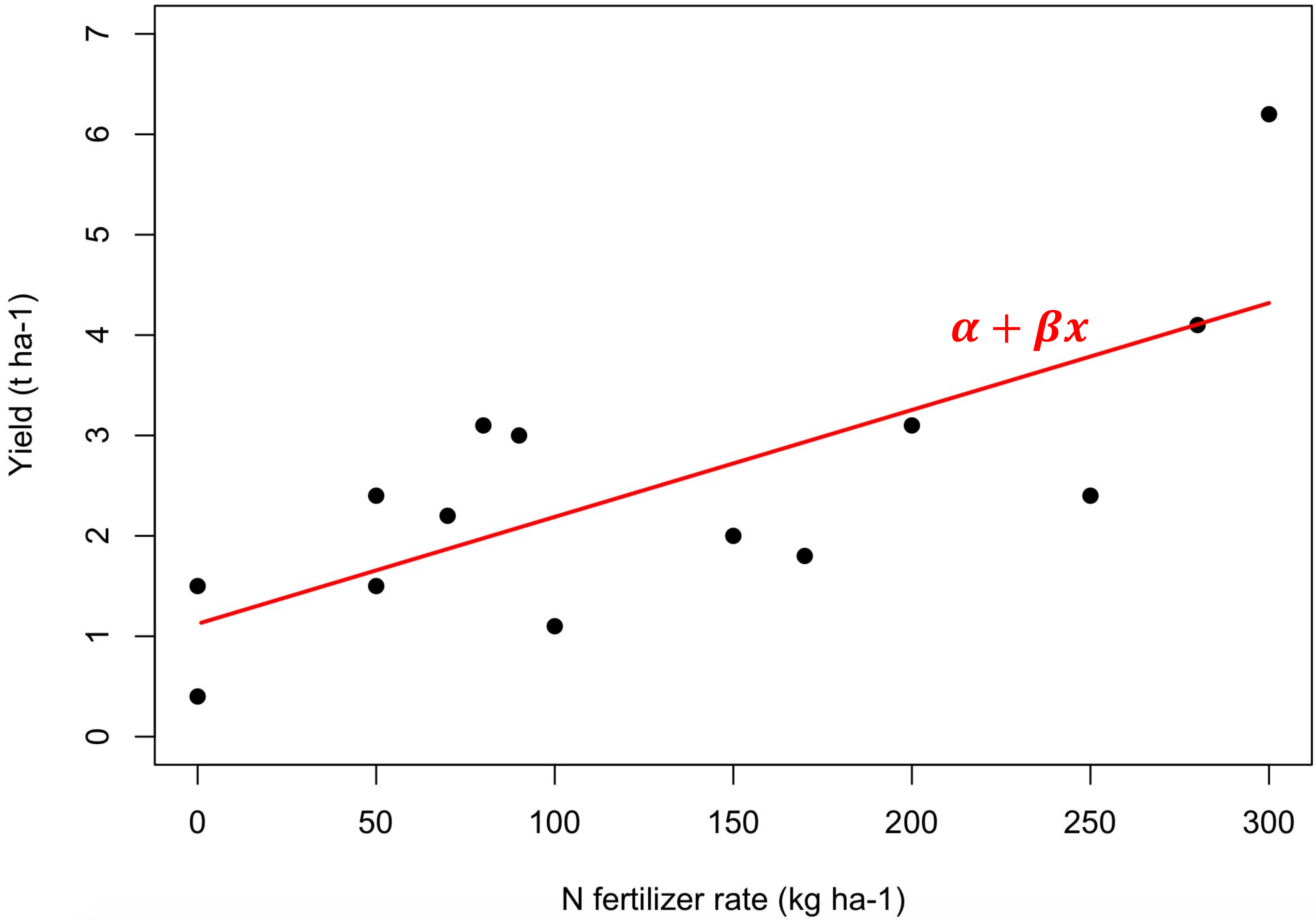
```
Dose<-c(0,250,100,50,70,170,300,50,80,90,0,280,200,150)
Obs<-c(1.5,2.4,1.1,1.5,2.2,1.8,6.2,2.4,3.1,3.0,0.4,4.1,3.1,2)

plot(Dose,Obs, xlab="N fertilizer rate (kg ha-1)", ylab="Yield (t ha-1)", ylim=c(0,7),pch=19)

Mod<-lm(Obs~Dose)
summary(Mod)
```

α (Intercept)

β Dose

```
> summary(Mod)

Call:
lm(formula = Obs ~ Dose)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.38665 -0.72333 -0.08014  0.65167  1.88080 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.123915   0.445230   2.524  0.02670 * 
Dose        0.010651   0.002789   3.818  0.00245 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9973 on 12 degrees of freedom
Multiple R-squared:  0.5485,	Adjusted R-squared:  0.5109 
F-statistic: 14.58 on 1 and 12 DF,  p-value: 0.002446
```

```r
plot(Dose,Obs, xlab="N fertilizer rate (kg ha-1)", ylab="Yield (t ha-1)", ylim=c(0,7),pch=19)

Mod<-lm(Obs~Dose)
summary(Mod)

D<-1:300

pred<-coef(Mod)[1]+coef(Mod)[2]*D

lines(D,pred,col="red",lwd=2)
```



Yield (t ha-1)
0
1
2
3
4
5
6
7
0
50
100
150
200
250
300
N fertilizer rate (kg ha-1)
$\alpha + \beta x$

Test on the effect of N fertilizer

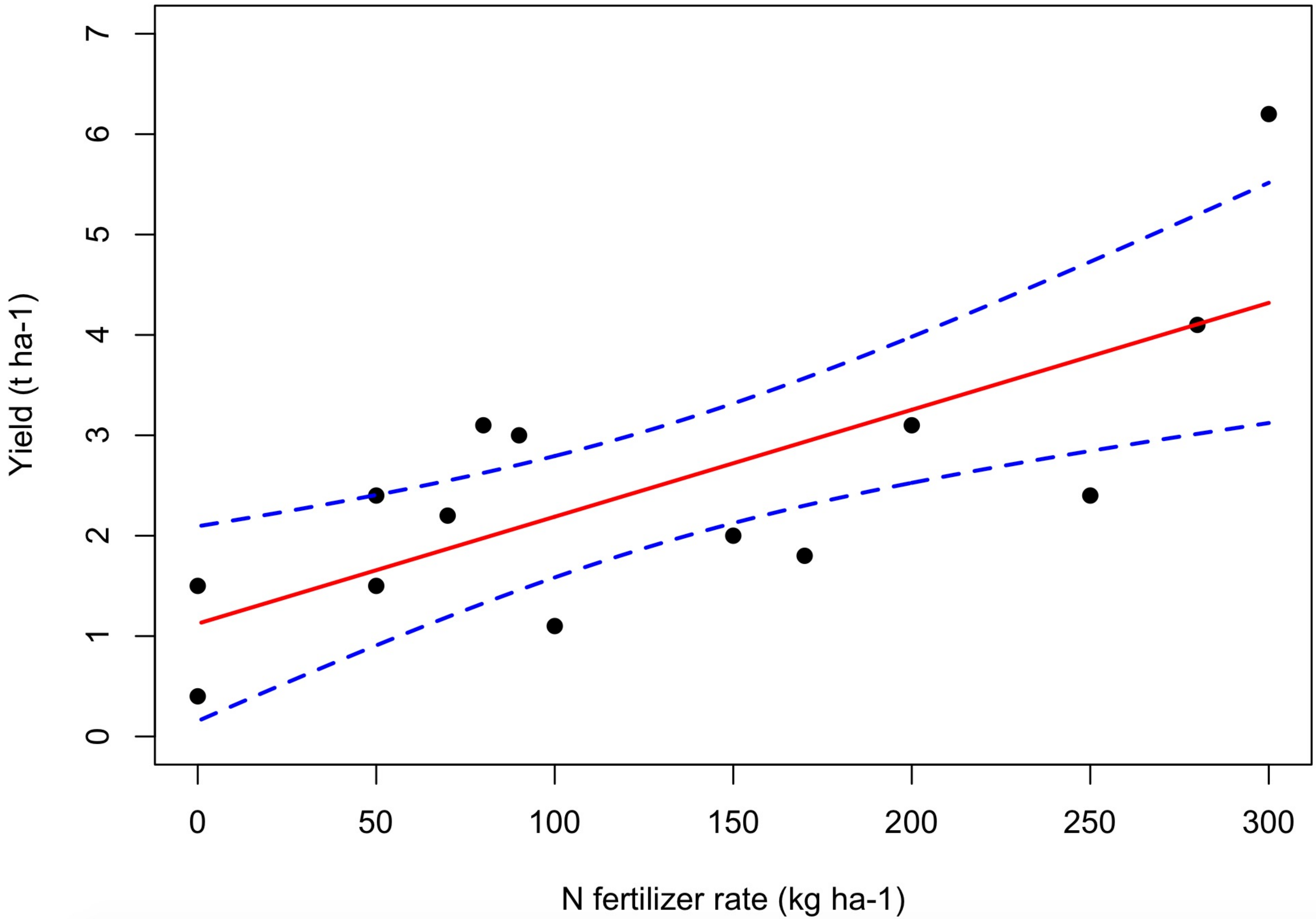
H_0 : « $\beta = 0$ » against H_1 : « $\beta \neq 0$ »

Test on the effect of N fertilizer

H_0 : « $\beta = 0$ » against H_1 : « $\beta \neq 0$ »

```
> summary(Mod)

Call:
lm(formula = Obs ~ Dose)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.38665 -0.72333 -0.08014  0.65167  1.88080 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.123915   0.445230   2.524  0.02670 * 
Dose        0.010651   0.002789   3.818  0.00245 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9973 on 12 degrees of freedom
Multiple R-squared:  0.5485,	Adjusted R-squared:  0.5109 
F-statistic: 14.58 on 1 and 12 DF,  p-value: 0.002446
```

Confidence intervals



```r
Dose<-c(0,250,100,50,70,170,300,50,80,90,0,280,200,150)
Obs<-c(1.5,2.4,1.1,1.5,2.2,1.8,6.2,2.4,3.1,3.0,0.4,4.1,3.1,2)

plot(Dose,Obs, xlab="N fertilizer rate (kg ha-1)", ylab="Yield (t ha-1)", ylim=c(0,7),pch=19)

Mod<-lm(Obs~Dose)
summary(Mod)

D<-1:300

pred<-coef(Mod)[1]+coef(Mod)[2]*D

lines(D,pred,col="red",lwd=2)

predIC<-predict(Mod,newdata=data.frame(Dose=D),interval="confidence",level=0.95)

predIC

lines(D,predIC[,2],lty=2,lwd=2, col="blue")
lines(D,predIC[,3],lty=2,lwd=2, col="blue")
```

Model evaluation

```
> summary(Mod)

Call:
lm(formula = Obs ~ Dose)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.38665 -0.72333 -0.08014  0.65167  1.88080 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.123915   0.445230   2.524  0.02670 * 
Dose        0.010651   0.002789   3.818  0.00245 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9973 on 12 degrees of freedom
Multiple R-squared: 0.5485,	Adjusted R-squared:  0.5109 
F-statistic: 14.58 on 1 and 12 DF,  p-value: 0.002446
```

Efficiency

RMSE

Conclusion
Main steps for the development of a model

- Definition of inputs X and outputs Y
- Definition of equations f
- Estimation of parameters θ
- Tests and model assessment
- Practical use