Statistical Methodology for Software Engineering

Hadas Lapid, PhD

Contents

Normal Distribution

- Most common distribution for the probability of a **continuous, real, random variable**

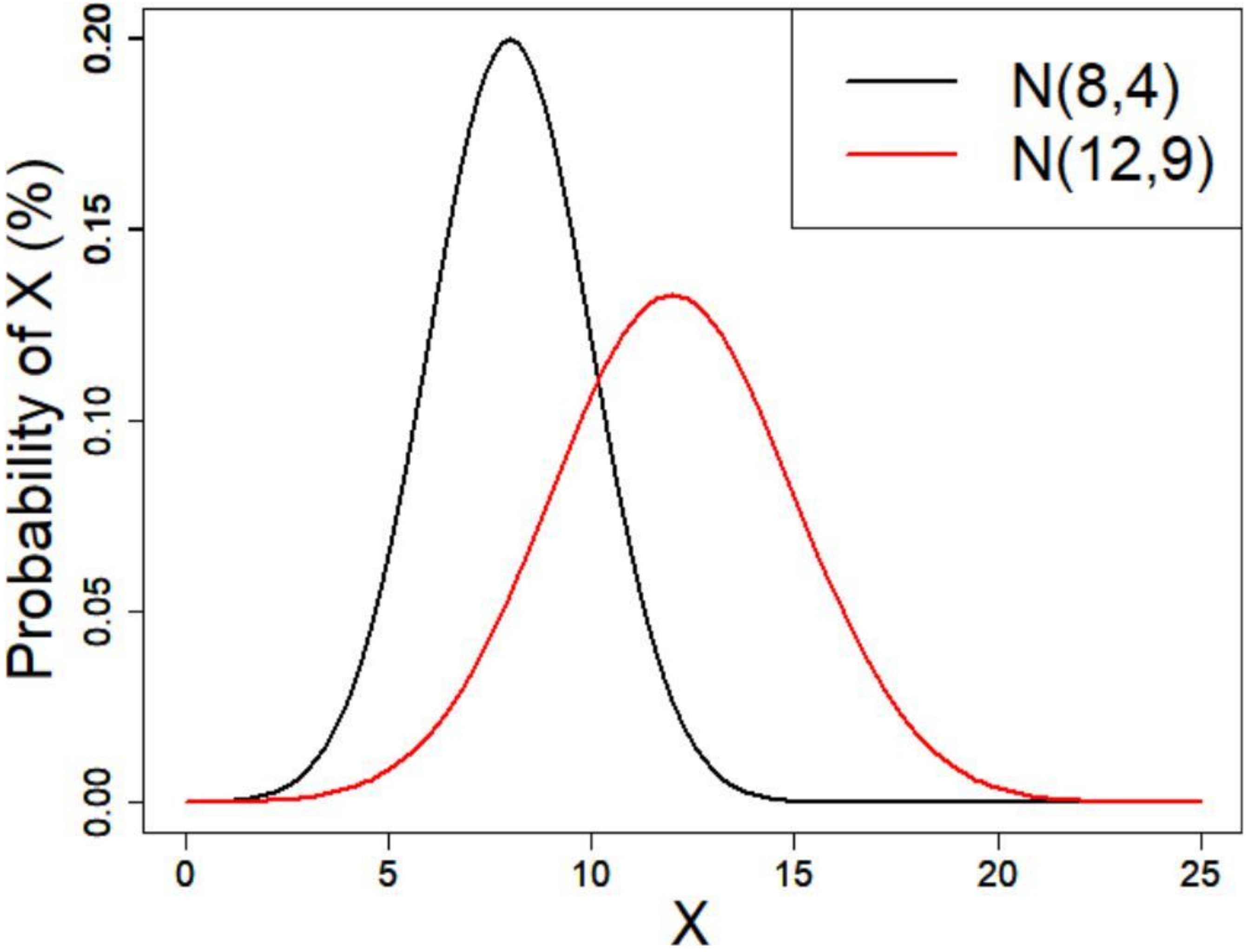
$\boldsymbol{X} \sim \boldsymbol{N}(\mu, \sigma^2)$

$\boldsymbol{\mu}$ ≡ Mean (and median)

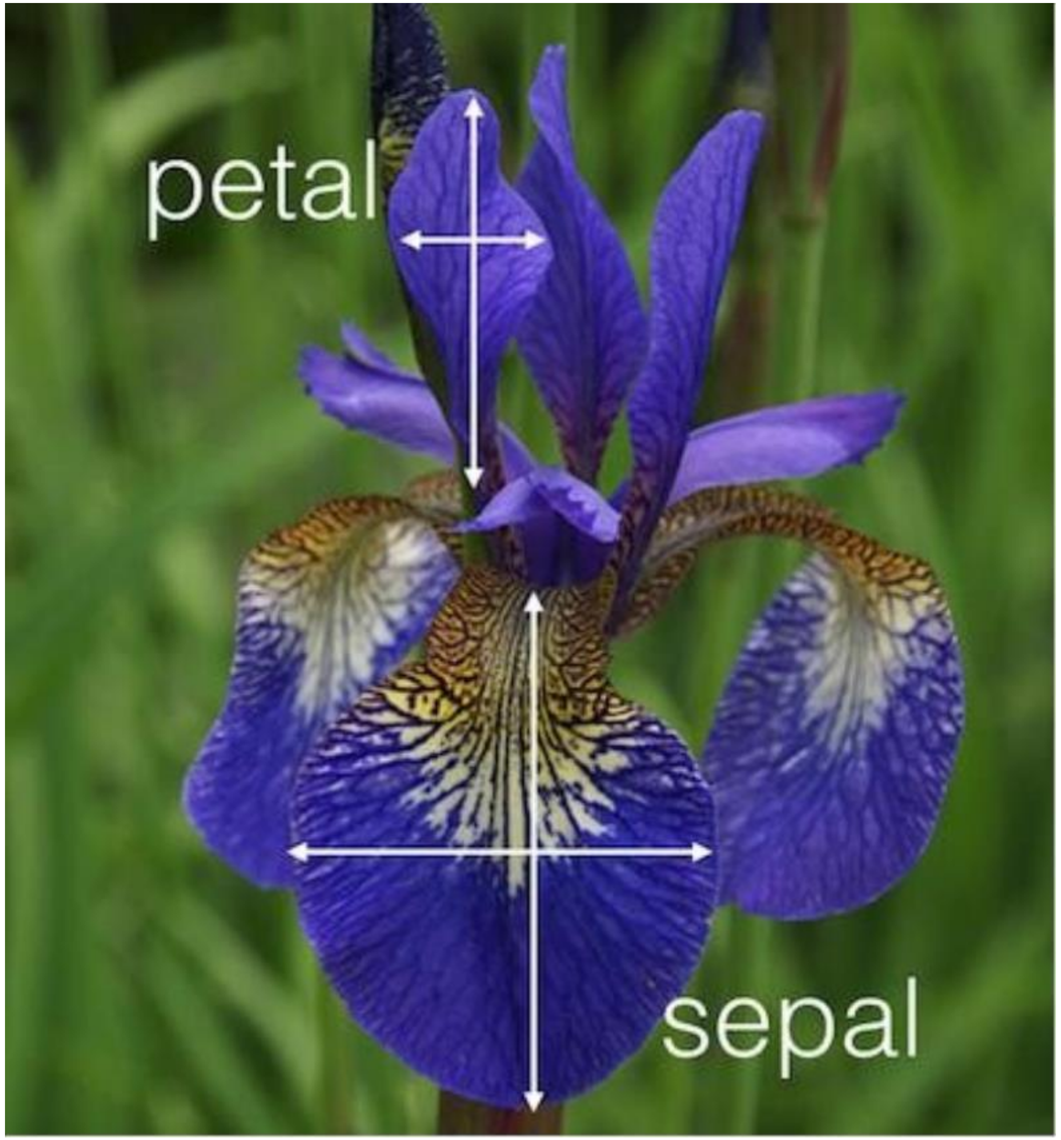
$\boldsymbol{\sigma}$ ≡ Standard deviation

$\boldsymbol{f}(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ ≡ Distribution Function

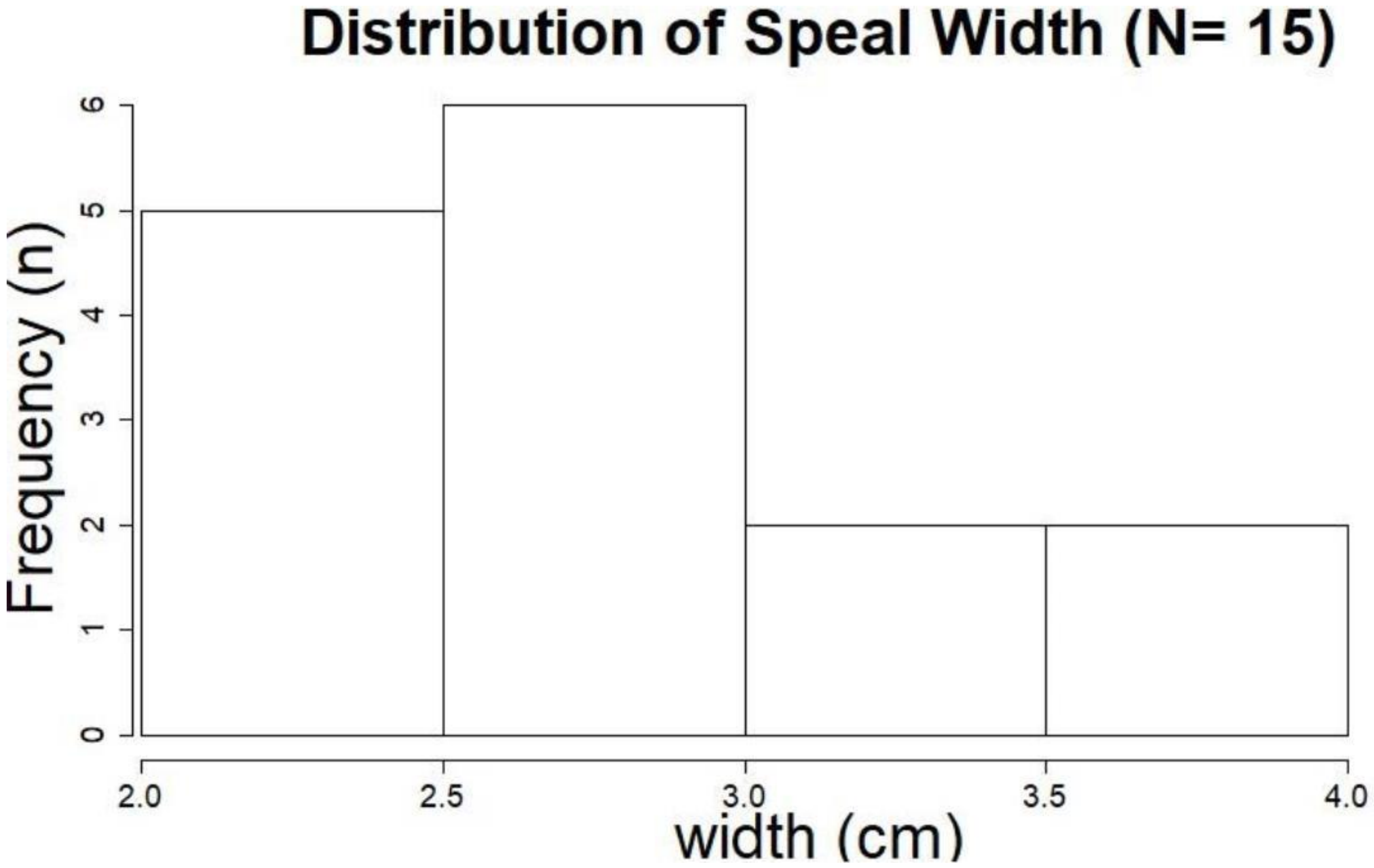
Normal Distribution

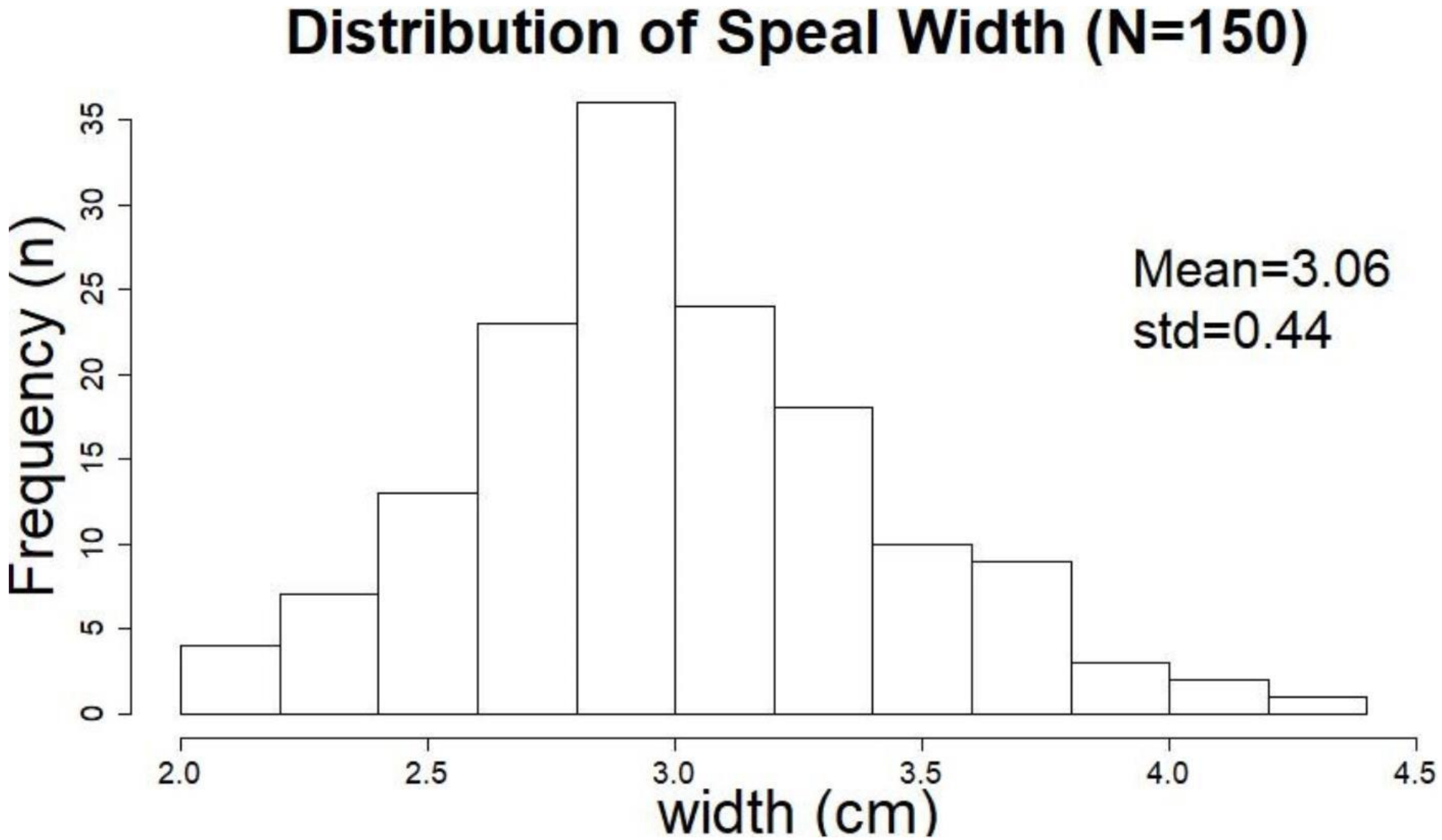


The Iris dataset



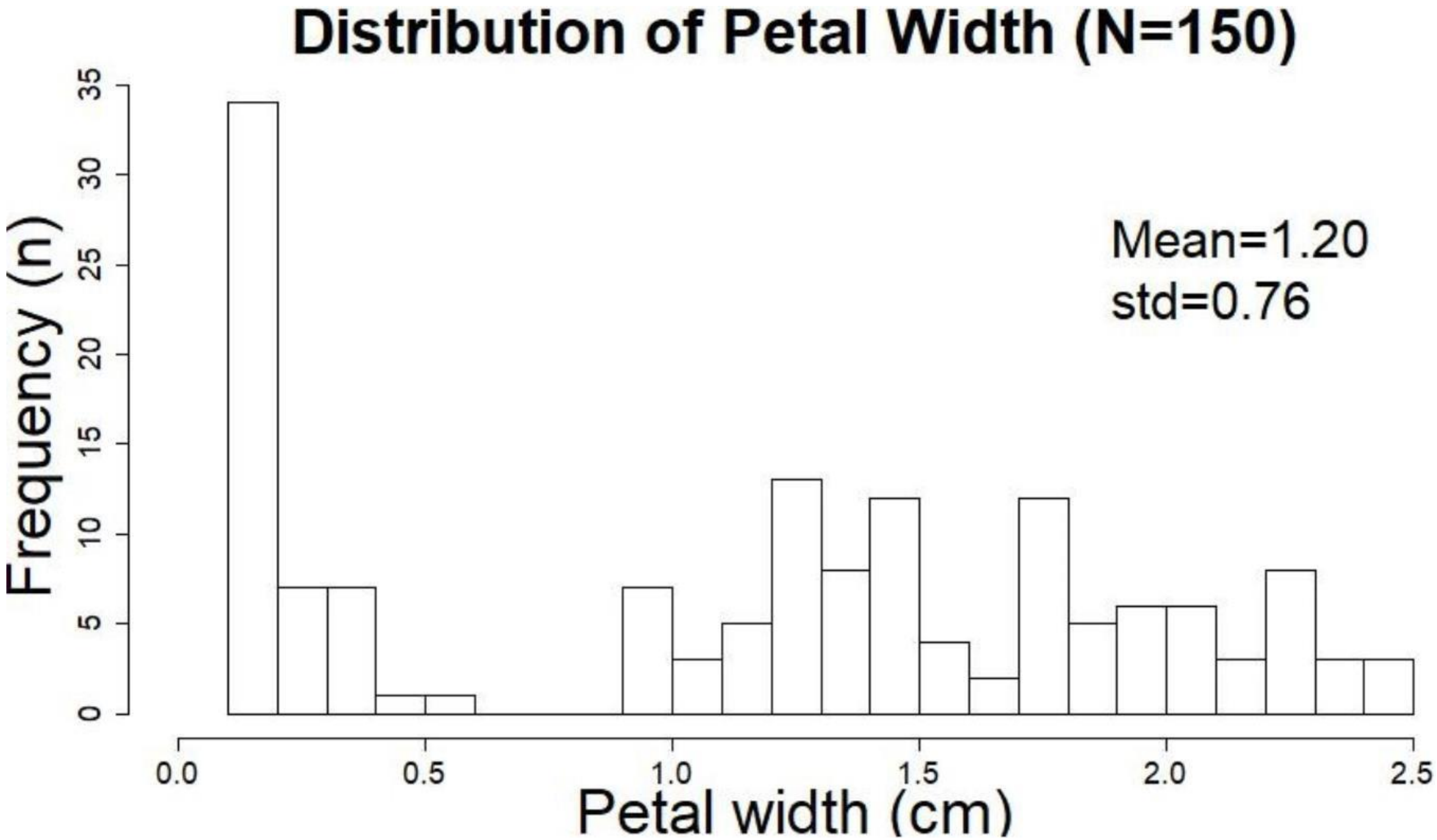
The importance of sample size




Distribution of Speal Width (N=150)
Frequency (n)
0
5
10
15
20
25
30
35
Mean=3.06
std=0.44
2.0
2.5
3.0
3.5
4.0
4.5
width (cm)

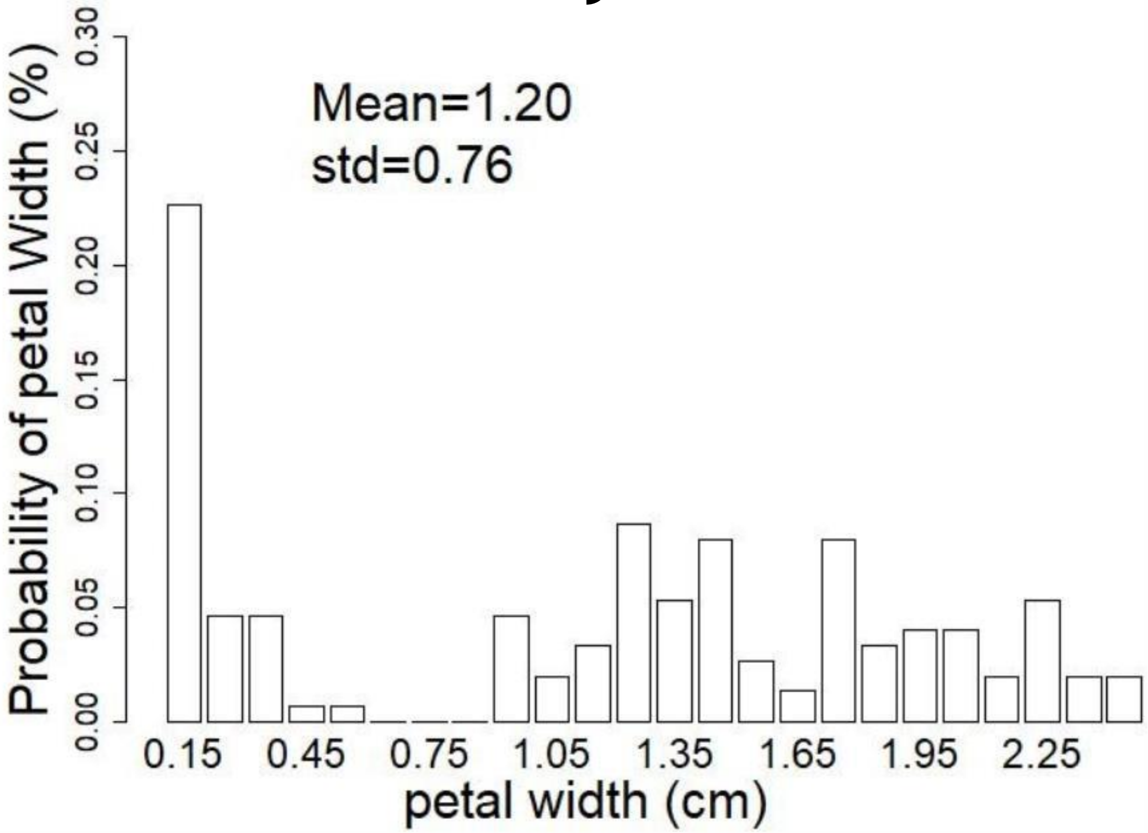
Conclusion:

NOT ALL CONTINUOUS VARIABLES ARE NORMALLY



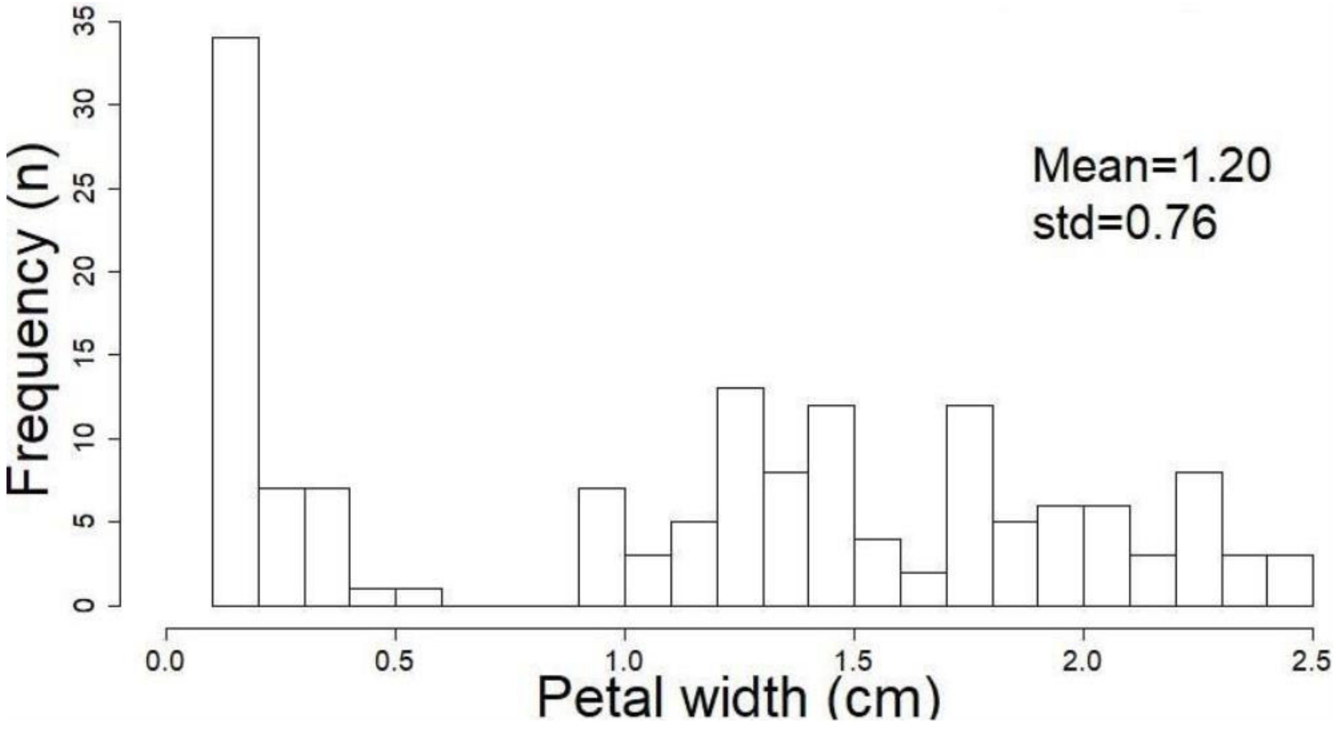
Frequency vs. Probability Distributions

Probability Distributions



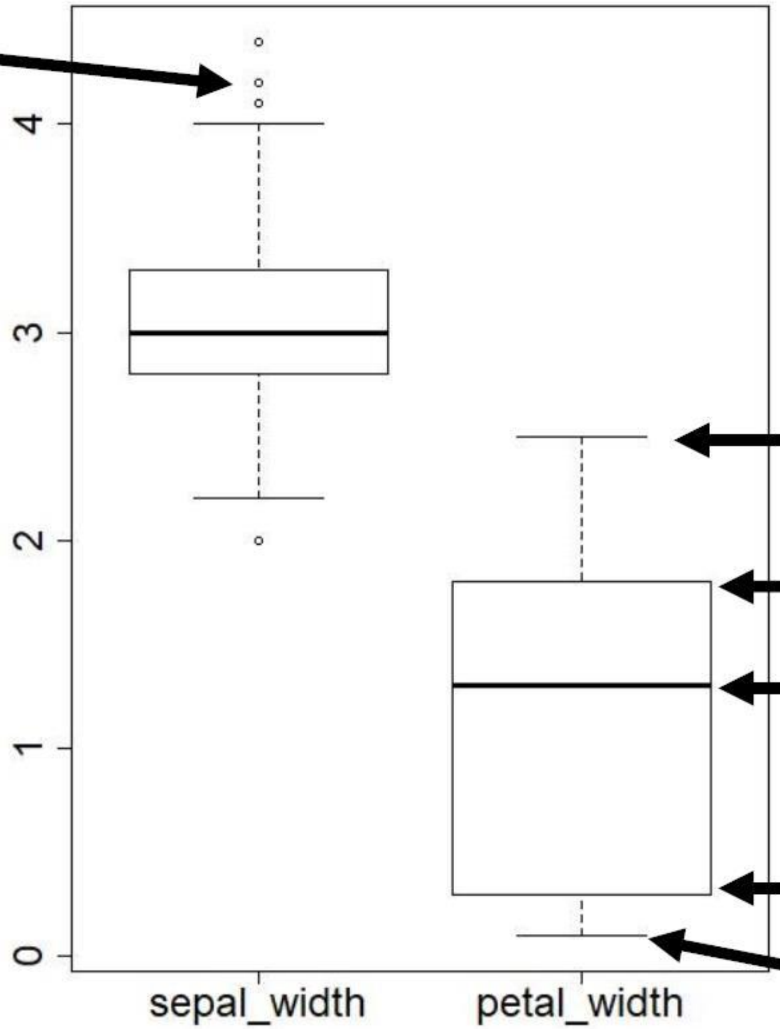
 # frequency of counts per bin

Frequency Distributions



 # counts per bin

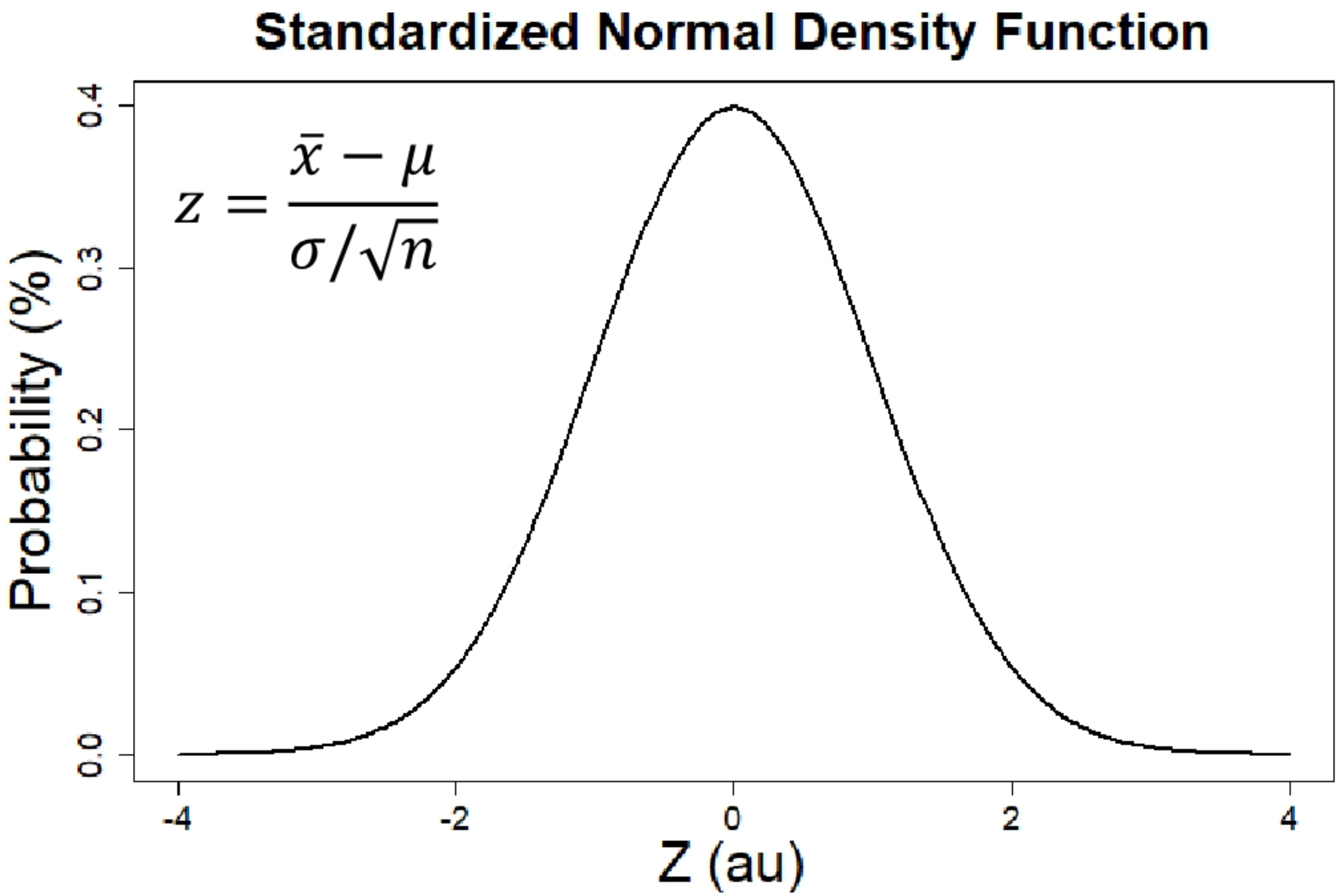
Boxplots andquantiles

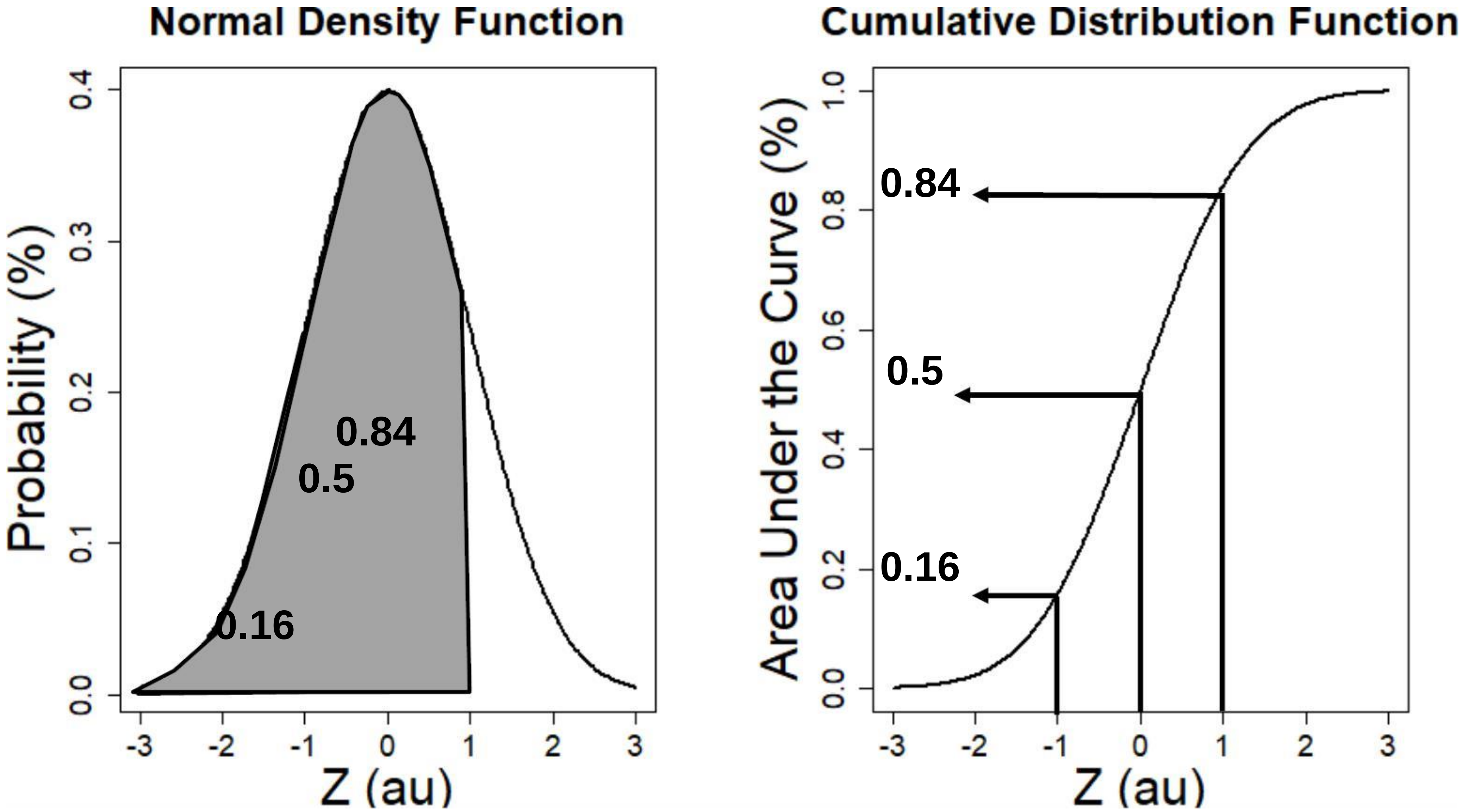



```
summary(df$sepal_width)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.000   2.800   3.000   3.057   3.300   4.400
```

The Standardized Normal Density Function

$$\boldsymbol{Z}(x|\mu = 0, \sigma = 1) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}} e^{-\frac{\boldsymbol{x}^2}{2}}$$ **Standard Distribution function**




Normal Density Function
Probability (%)
0.4
0.3
0.2
0.1
0.0
0.84
0.5
0.16
-3
-2
-1
0
1
2
3
Z (au)
Cumulative Distribution Function
Area Under the Curve (%)
1.0
0.8
0.6
0.4
0.2
0.0
0.84
0.5
0.16
-3
-2
-1
0
1
2
3
Z (au)

Why is the "Normal Distribution" important?

- The distribution of many continuous observables in nature is approximately normal
- We can describe the normal distribution in an equation

We can draw conclusions about the relations between variables with **Magnitude** and **Significance** estimates

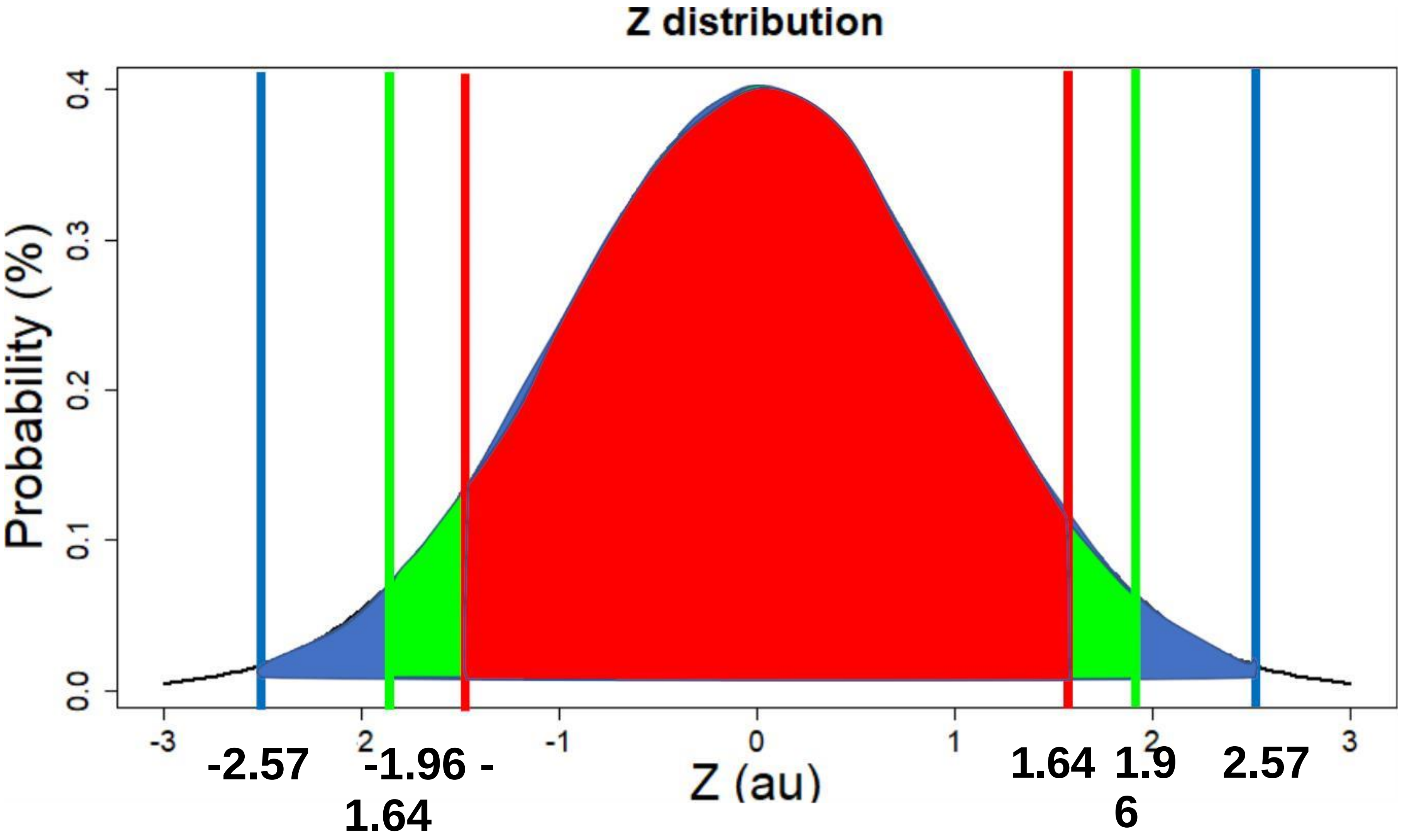
Drawing conclusions from normal distribution:
Calculating the probability of an event

- What is the probability of getting a grade of 90 or more, assuming normal distribution with mean 80 and standard deviation 10, given

  $P(x \leq 90) = 0.84$?

- What is the probability getting a grade between 70 and 90 under the same assumptions, given $P(x \leq 70) = 0.16$?

Confidence interval - intuition



Confidence interval for the mean (σ known)

- **Definition**:
  An interval of numbers which will include the unknown value of μ with given probability, 1-α, called the confidence interval.
- Typical confidence levels, 1-α: 0.95, 0.99, 0.9 (95%, 99%, 90%)
- Given distribute normally n samples, $x_1, x_2, \ldots x_n$,
  with known variance, σ^2, a 1-α confidence interval is given as:

$$C.I. = \bar{X} \pm Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Where $\bar{X}$ is the sample mean and $Z_{1-\frac{\alpha}{2}}$ is taken from the standardized normal distribution

Drawing conclusions from normal distribution:

Calculating the confidence interval of the population mean

Confedence level $1-\alpha/2$	$Z_{1-\alpha/2}$
0.9	1.645
0.95	1.96
0.99	2.576

- 36 cars in highway 6
- mean speed, $\bar{x} = 103kph$
- standard deviation $\sigma = 8$.

- What is the confidence interval for the mean speed for all this road's journeys?

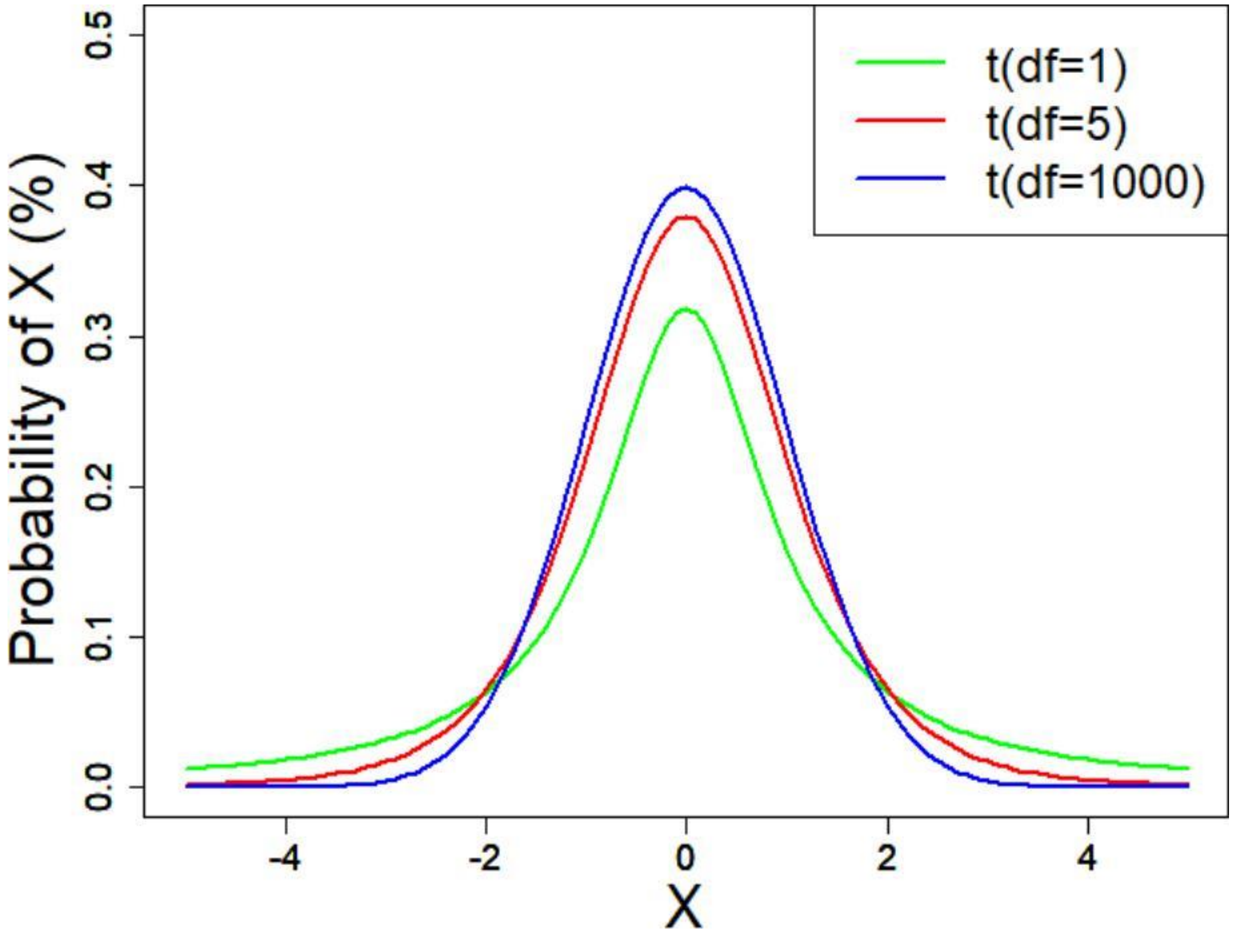
Calculating C.I. at 95%: $103 \pm 1.96 * \frac{8}{\sqrt{36}} = 103 \pm (2.6) = \ (100.4, 105.6)$

Conclusion: μ is found in the range of 100.4 and 105.6 in 95% confidence

At home, try the same for μ at 90% and at 99% confidence

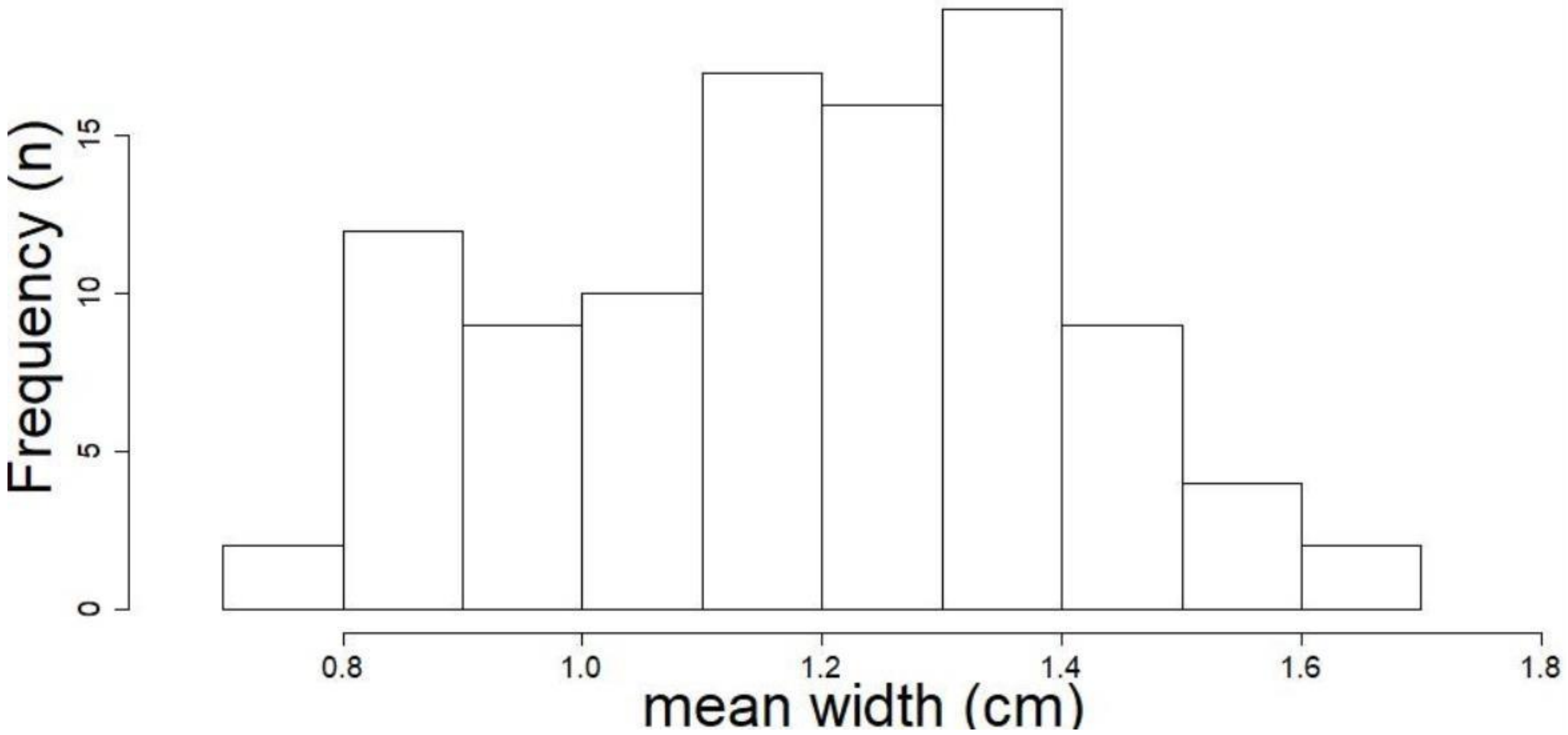
t(df)
Standardized normal density function without known σ

For n samples, and $\upsilon = n - 1$ degrees of freedom, we can assume that σ ~ s



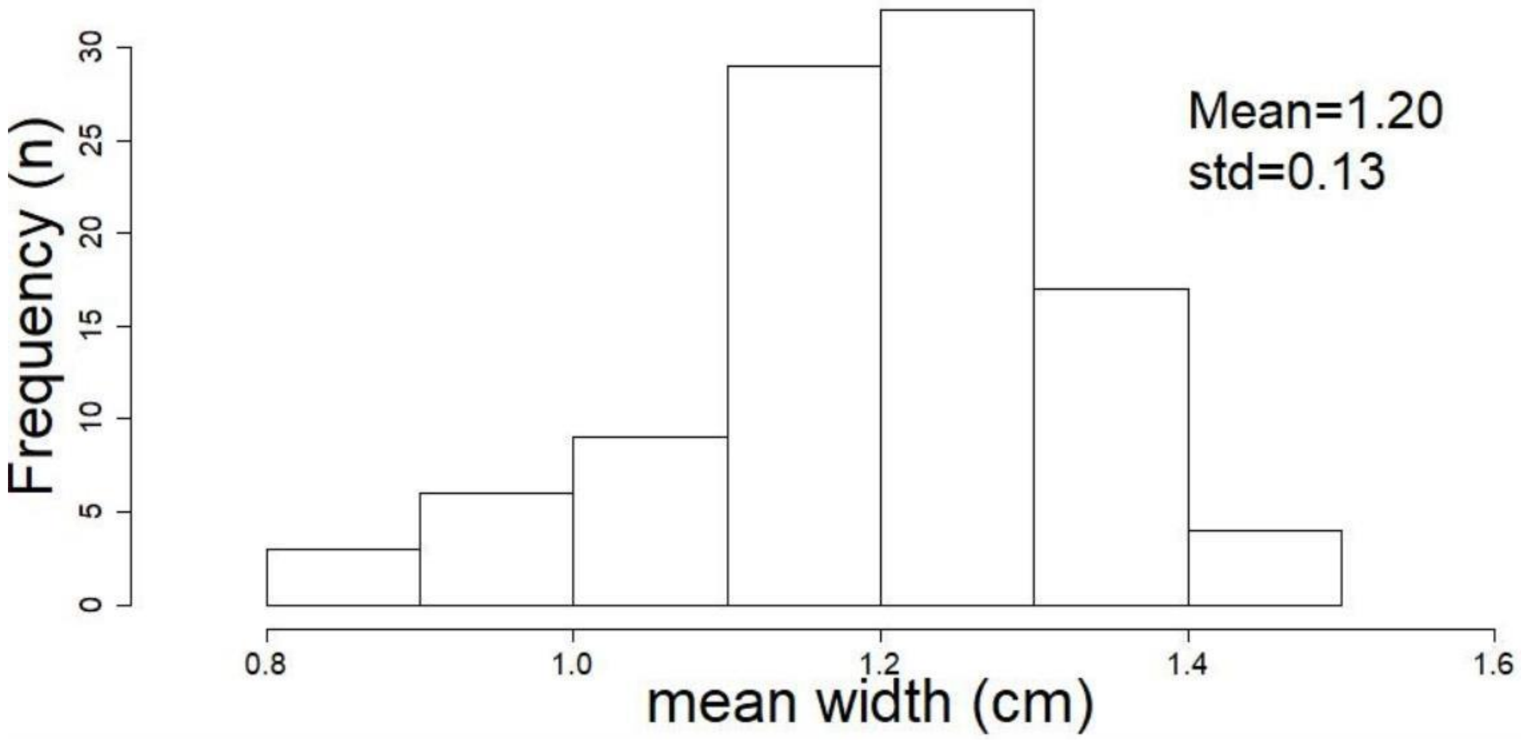
Central limit theorem - intuition

Distribution of the means of 100 random samples of n=10 petal widths



Central limit theorem - intuition

Distribution of the means of 100 random samples of n=30 petal widths



Central limit theorem – informal definition

- The **probability distribution of the mean** of repeated, independent, large enough, random samples of a variable **will always be normal**, even if the variable itself does not distribute normally.
- Large enough ≈ 30
- $E(\bar{X}) \approx \mu$
- $S(\bar{X}) \approx \frac{\sigma}{\sqrt{n}}$
- $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Hypothesis Testing
Basic terminology

- Unknown population parameter (e.g., μ)
- Null hypothesis: H_0
- Alternative hypothesis: H_1
- One-sided / two sided alternative
- Test statistic
- Statistical test
- Rejection region, R
- Probability of Type I error, σ
- Probability of Type II error, β
- Power
- p value