# Statistical Methodology for Software Engineers

**Hadas Lapid, PhD**

# Course Syllabus

- Statistical distributions
- Hypothesis testing
- Chisquare test for goodness of fit
- Extreme value distribution and its role in sequence analysis
- Multiple hypothesis testing
- Analysis of contingency tables
- Two-sample t-test
- Analysis of variance
- Non-parametric tests
- Correlation tests
- Linear regression

# Course Requirements

- 4 HW Assignments, 20%
- Exam, 80%

# Success guaranteed if

- Review class materials before each lecture
- Practice and freely explore R software

# Resources

- All course materials on Moodle
- “An Introduction to Statistical Methods and Data Analysis” by R. Lyman Ott and Michael Longnecker (on moodle)
- Introduction to Statistics and Data Analysis _ With Exercises, Solutions and Applications in R - Heumann · Schomaker
- R Software: https://rstudio.com/products/rstudio/download/#download
- Web-textbook: http://www.statsoft.com/Textbook

# Contents Today

- **Variables**
  - types
  - Basic concepts
- **Distributions**
  - Binomial Distribution
  - Poisson Distribution
  - Normal Distribution
  - Standardized Normal Density function
- **Data example**

# Why do we need statistics?

## Typical Problems:

- Data variability
- Repeated measurements may yield different results

## Statistical Solutions:

- Provide predictions of outcome
- Draw **conclusions** about data variables (inverse problem)

## Related Subjects:

- Probability (direct problem)
- Business analytics
- Machine learning (predicting)
- Operations research (optimization)

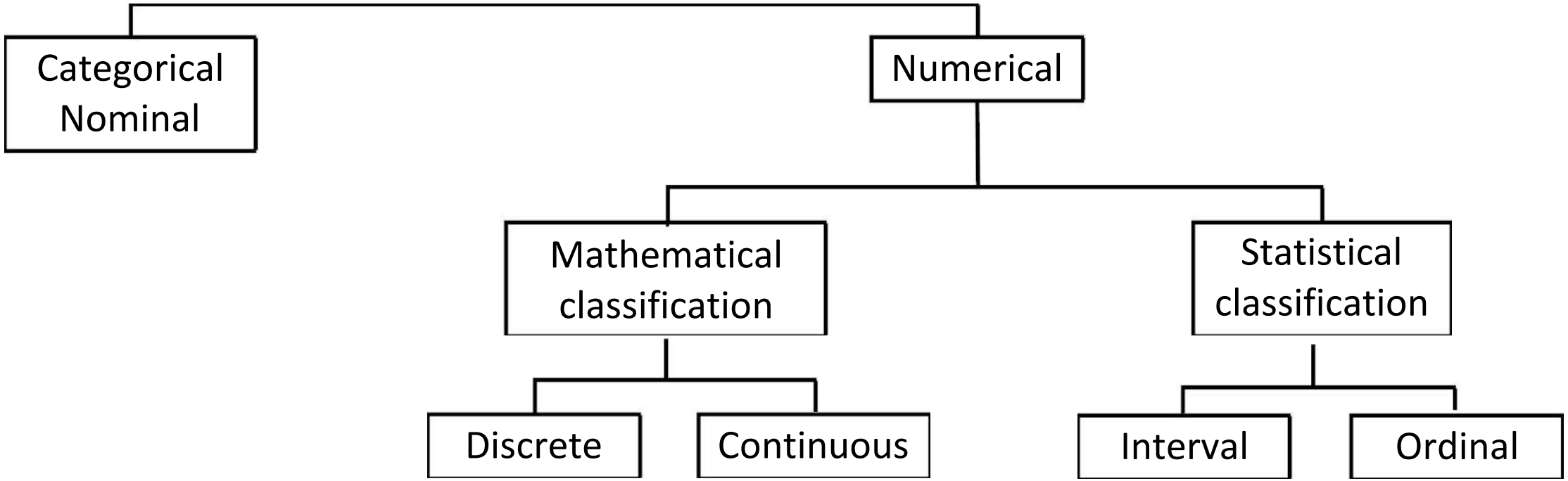
# What are variables?

- Objective/subjective measures that can be monitored, controlled or manipulated.
- Variables differ in the quantity of measurable information their scale provides

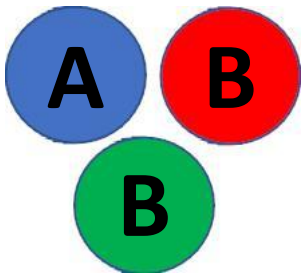
# Dependent vs. Independent variables

- Independent variables – inputs, explanatory, manipulated in the research (e.g. gender)
- Dependent variables – output, response only measured/monitored (e.g. white cell blood count)

# Variables Classification



# Nominal (Categorical) Variables

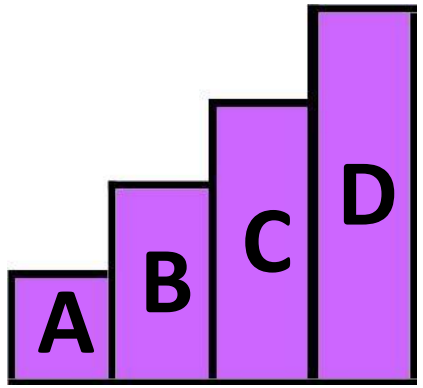



Enable qualitative classification

i.e., whether an item belongs to a distinct category

e.g., city, gender, color

# Discrete Variables

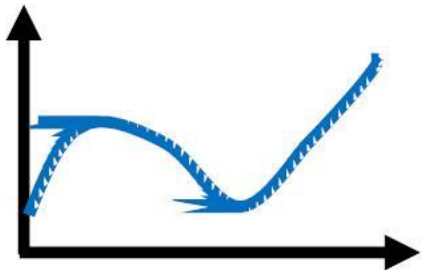



Enable order ranking

i.e., whether an item has more or less of the measured quality

e.g., Olympic rank, grades rank, disease severity level

# Continuous Variables



Quantitative measurement

i.e., how much of the measured characteristic the item displays

e.g., height, weight, temperature

# Basic Concepts

Let$(x_1, x_2, x_3 \dots x_n)$, be n samples of X measurements in some population

$\boldsymbol{\mu}$ ≡ Population mean (Expectation Value)

$\overline{\boldsymbol{x}} = \langle \boldsymbol{x} \rangle$ ≡ Sample mean

$$\langle x \rangle = \frac{\sum_{i=1}^{n} x_i}{n}$$

# Basic Concepts

$\boldsymbol{\sigma^2}$ ≡ population variance

**S²** ≡ sample variance, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

We can assume, on average: $\boldsymbol{S^2 \approx \sigma^2}$

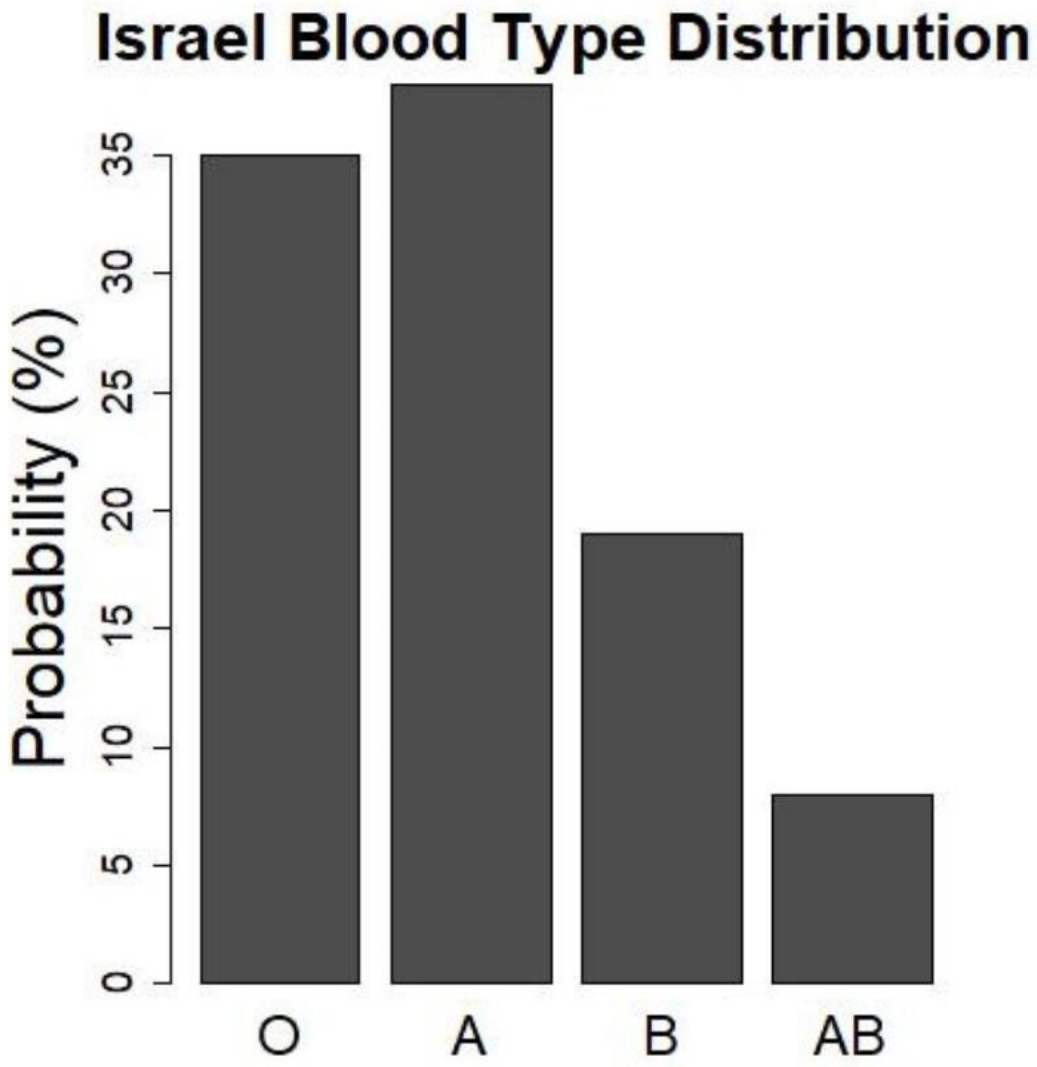
$\boldsymbol{\sigma}$ ≡ population standard deviation

**S** ≡ sample standard deviation, $S = \sqrt{S^2}$

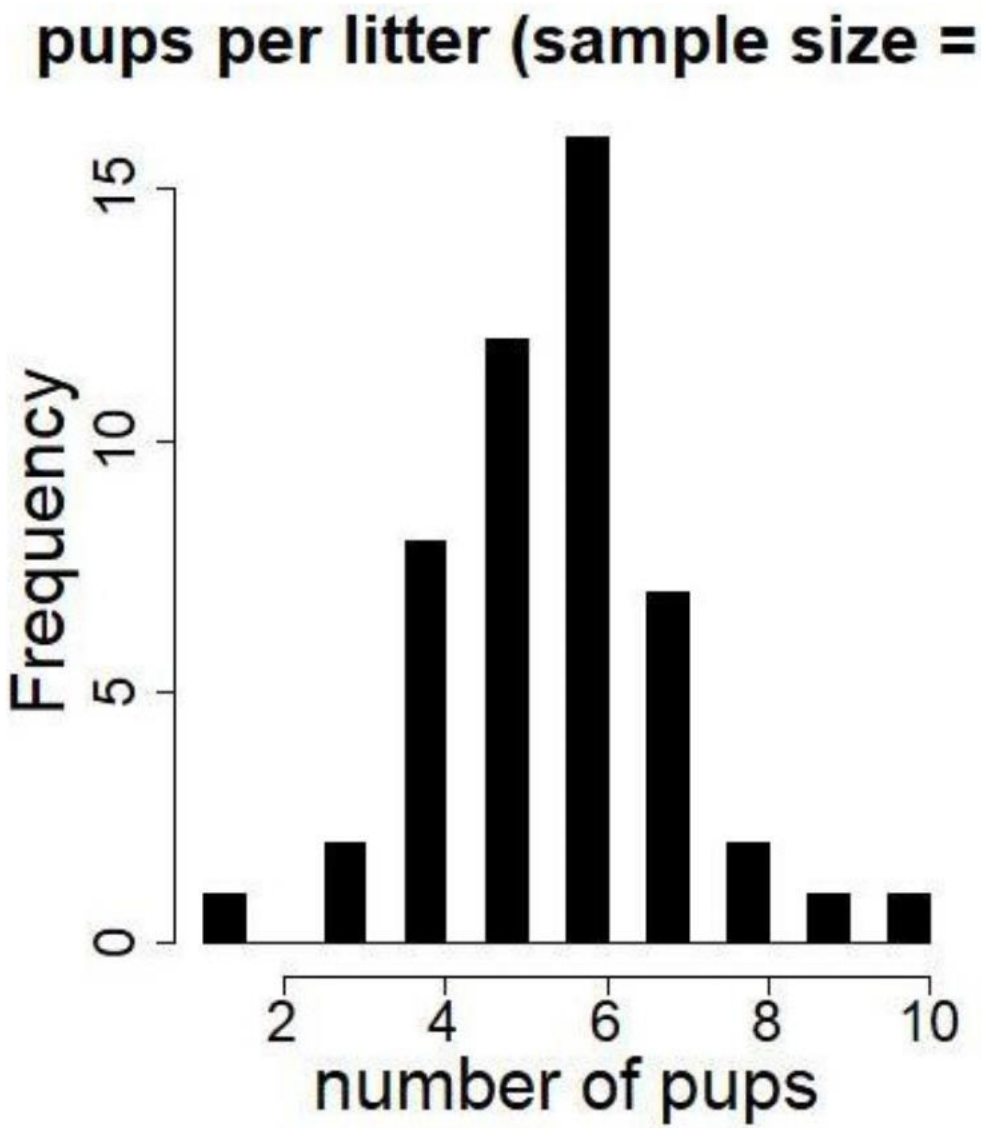
**Stderr** ≡ sample standard error of the mean,

$$s = \frac{s}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

# Categorical Variable Distribution



# Discrete Variable Distribution



# Binomial Variables

- Binomial variable is a measured outcome of an experiment, which can be either **Success (True)** or **Failure (False)**.

**ρ** = Probability of Success (roh)

**q** = **1-ρ** = Probability of Failure

# Binomial Distribution

- Binomial distribution results from **n** independent trials of a Bernoullie experiment.

- $\boldsymbol{x}$ $\equiv$ The number of successes measured in an experiment.
- $\boldsymbol{n}$ $\equiv$ The number of trials
- Notation: $\boldsymbol{x \sim B(n, \rho)}$

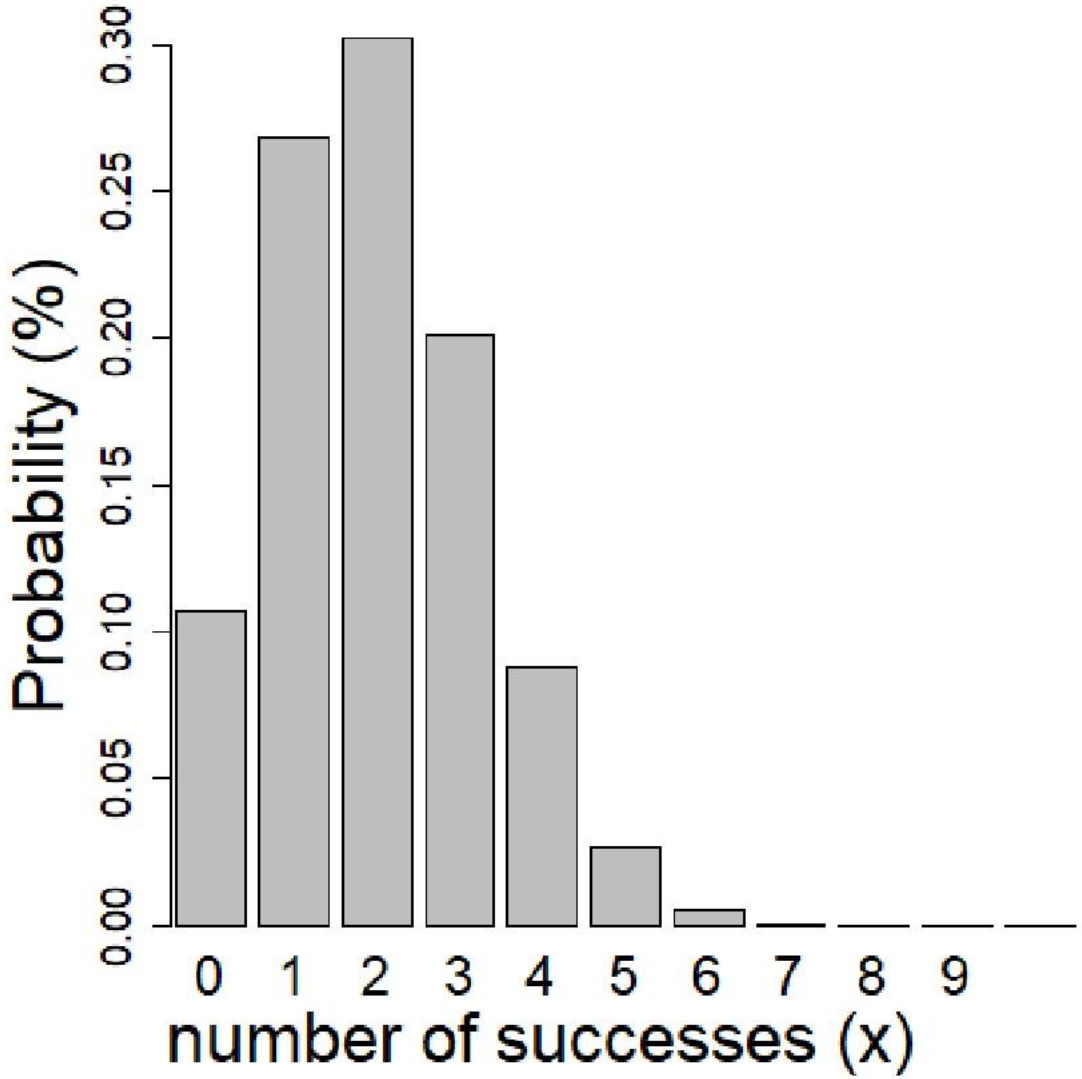
The **probability of x successes** out of n trials provided success rate ρ:

- $\boldsymbol{P(X = x) = \binom{n}{x} \rho^x q^{n-x}}$

# Binomial Distribution - Summary

- $\boldsymbol{x \sim B(n, \rho)}$
- $\boldsymbol{x}$ $\equiv$ The number of successes measured in an experiment.
- $\boldsymbol{n}$ $\equiv$ The number of trials
- $\boldsymbol{\rho}$ $\equiv$ probability of success
- $\boldsymbol{P(X = x) = \binom{n}{x} \rho^x q^{n-x} = \frac{n!}{x! \cdot (n-x)!} \rho^x q^{n-x}}$
- $\boldsymbol{\mu = np}$ Expectation Value (population mean)
- $\boldsymbol{\sigma^2 = npq}$ Variance
- $\boldsymbol{\sigma = \sqrt{npq}}$ Standard deviation

# Binomial Distribution - Example




**X = $\{x|0 < x < 10\}$**

**n = 10**

**ρ = 0.2**

**q = 0.8**

- Calculate the probability for 3 successes out of 10 $\boldsymbol{P(X = 3)}$
- Calculate the expectation value, **μ**
- Calculate the standard deviation, **σ**

# Poisson Distribution

- **Discrete probability distribution** that describes the probability of a given number of events occurring in a fixed interval of space or time

- **Examples:**

- # of calls to an emergency service per hour
- # of traffic accidents in a given street per week
- # of incoming text messages per day
- # of laser photons hitting a detector in a particular time interval

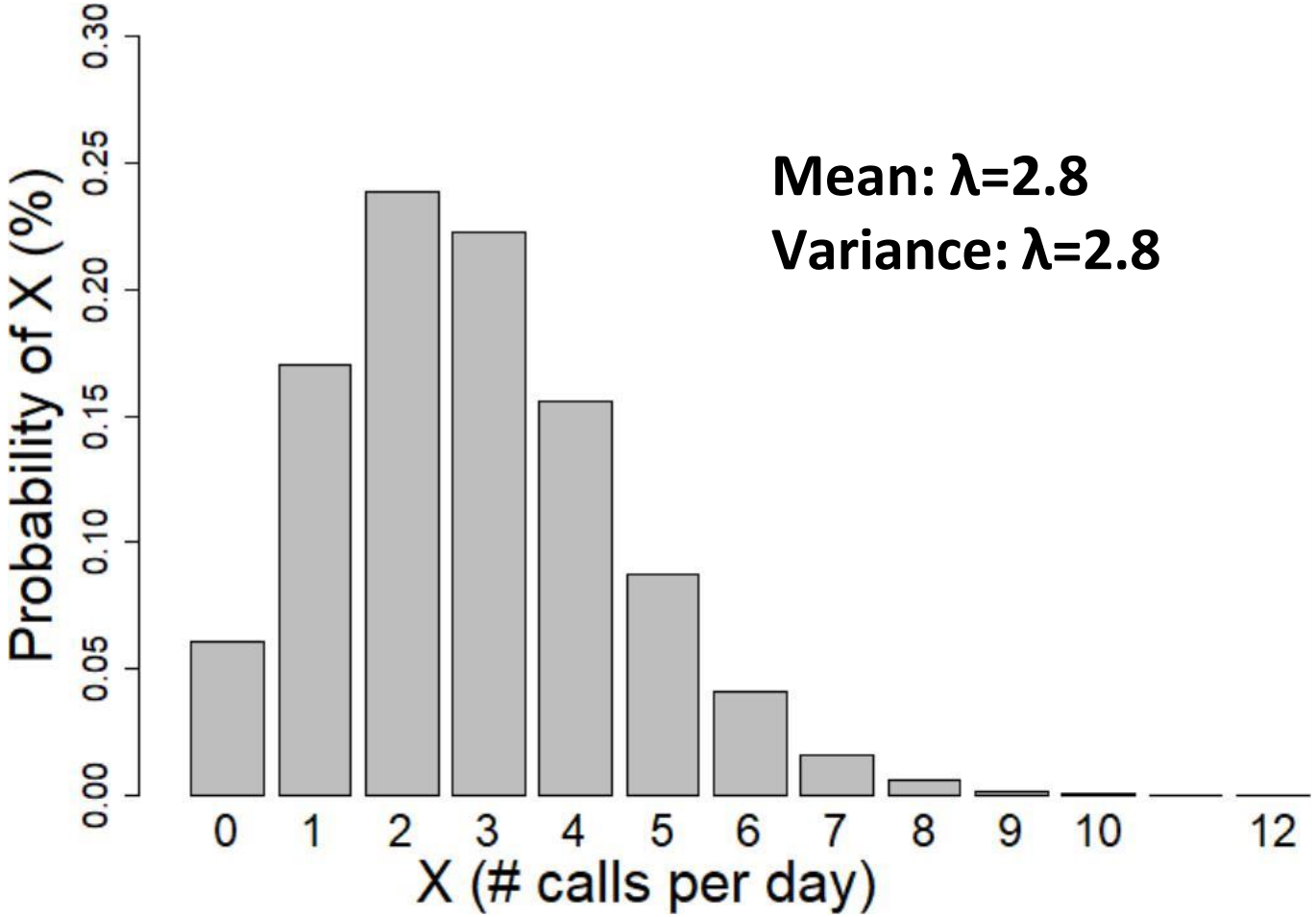
# Poisson Distribution mathematical definition

- $\boldsymbol{x} \sim \boldsymbol{P}(\boldsymbol{\lambda}, \boldsymbol{\lambda})$
- **λ ≡ Poisson Variable**
- $\boldsymbol{\mu} = \lambda$ , population mean
- $\boldsymbol{\sigma}^{\mathbf{2}} = \lambda$ , variance
- $\boldsymbol{P}_{\boldsymbol{i} \geq \mathbf{0}} = \dfrac{\boldsymbol{e}^{-\boldsymbol{\lambda}} \boldsymbol{\lambda}^{\boldsymbol{i}}}{\boldsymbol{i}!}$ probability of i occurrences per unit

- When the number of trials (n) goes to infinity, and the expected number of successes ($p_x$) is small then

$$F_{binomial}(x; n, p_x) = F_{Poisson}(x; \lambda = np_x)$$

# Poisson Distribution with Poisson variable λ=2.8



# Normal Distribution

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

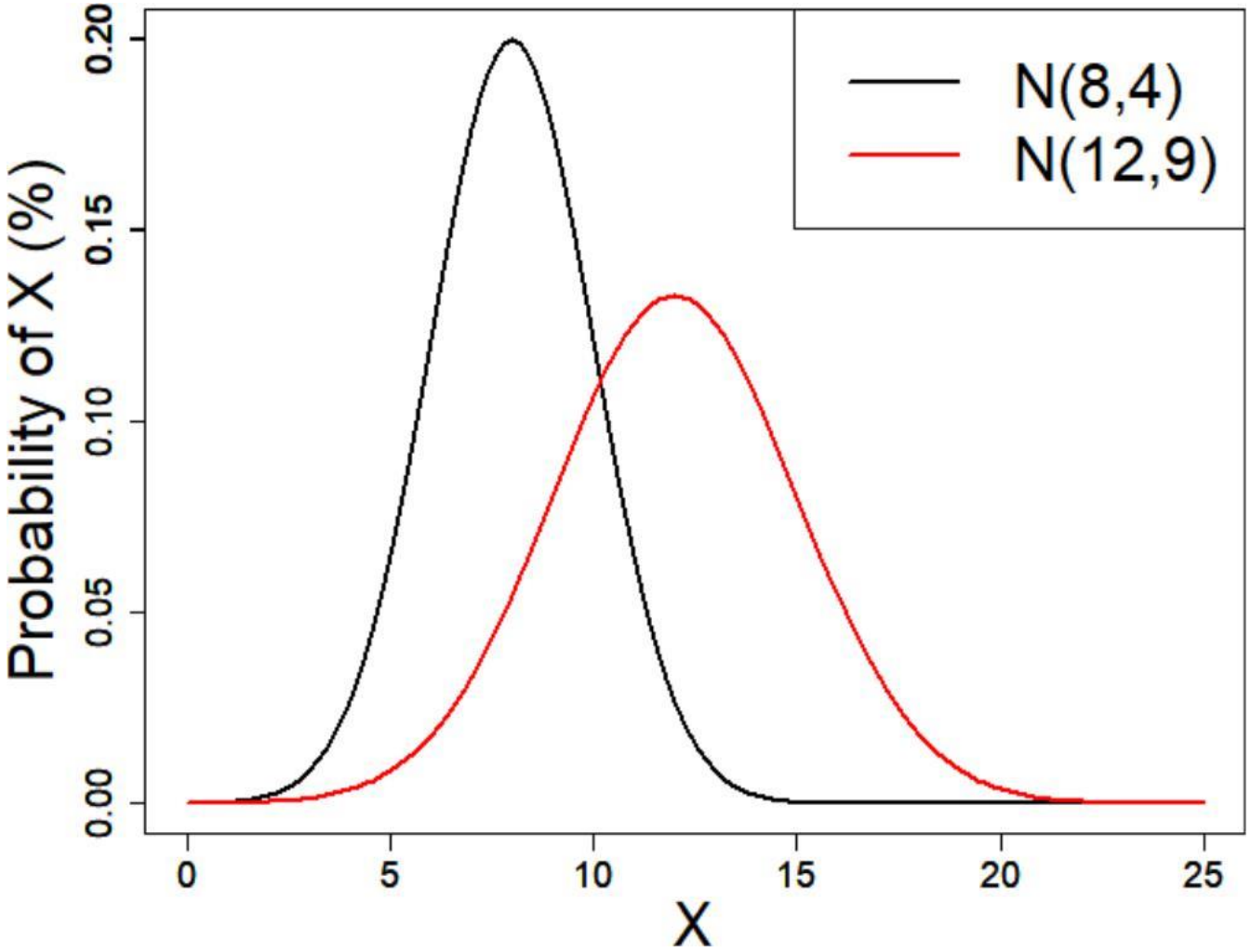
$$\boldsymbol{X \sim N\left(\mu, \sigma^2\right)}$$

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

$\boldsymbol{\sigma}$ $\equiv$ Standard deviation

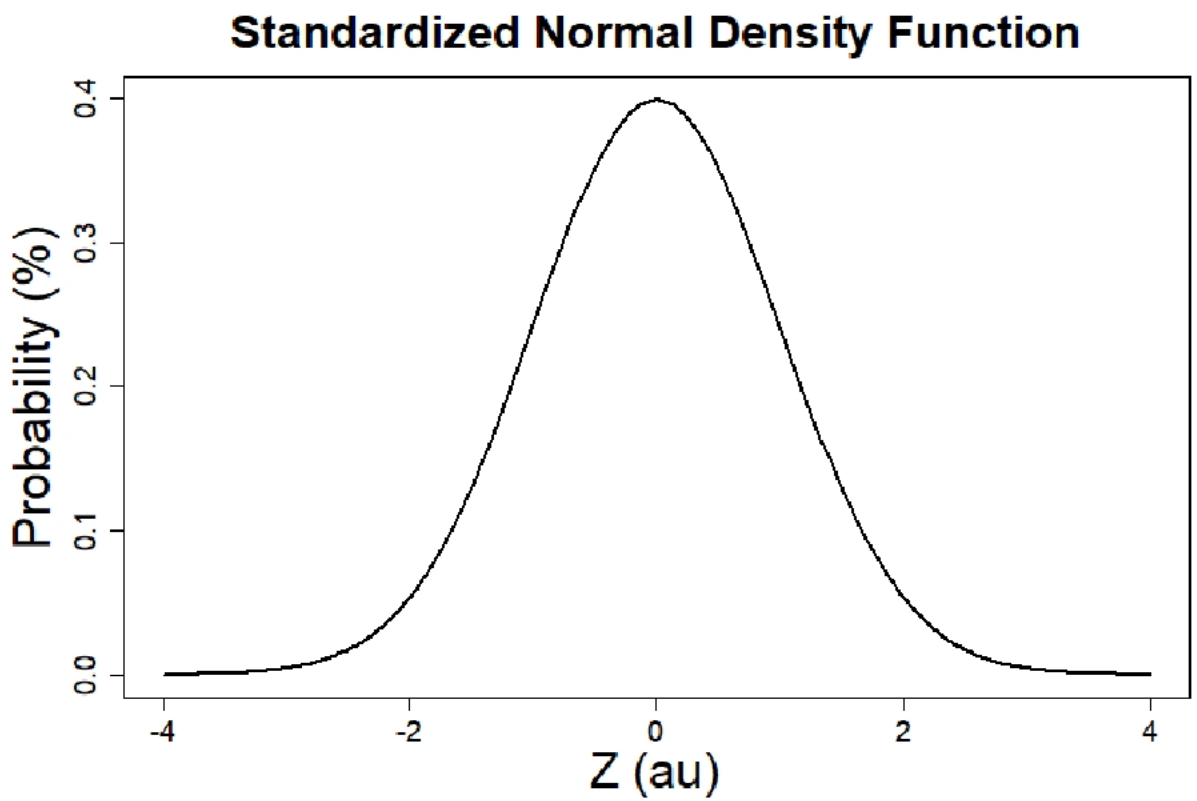
$$\boldsymbol{f}(x|\mu,\sigma) = \frac{\mathbf{1}}{\boldsymbol{\sigma}\sqrt{\mathbf{2}\boldsymbol{\pi}}} e^{-\frac{\mathbf{1}}{\mathbf{2}}\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)^{\mathbf{2}}} \equiv \text{Distribution function}$$

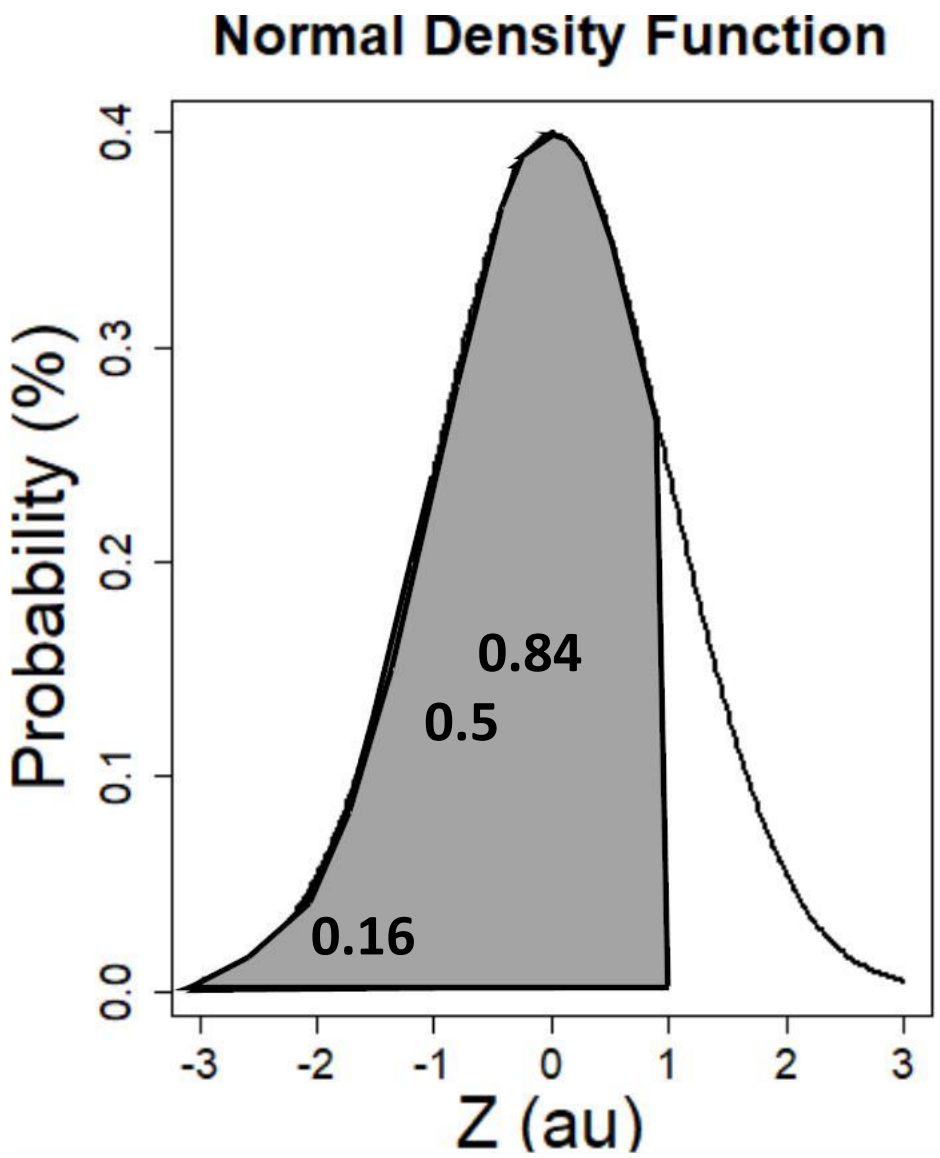
# NormalDistribution

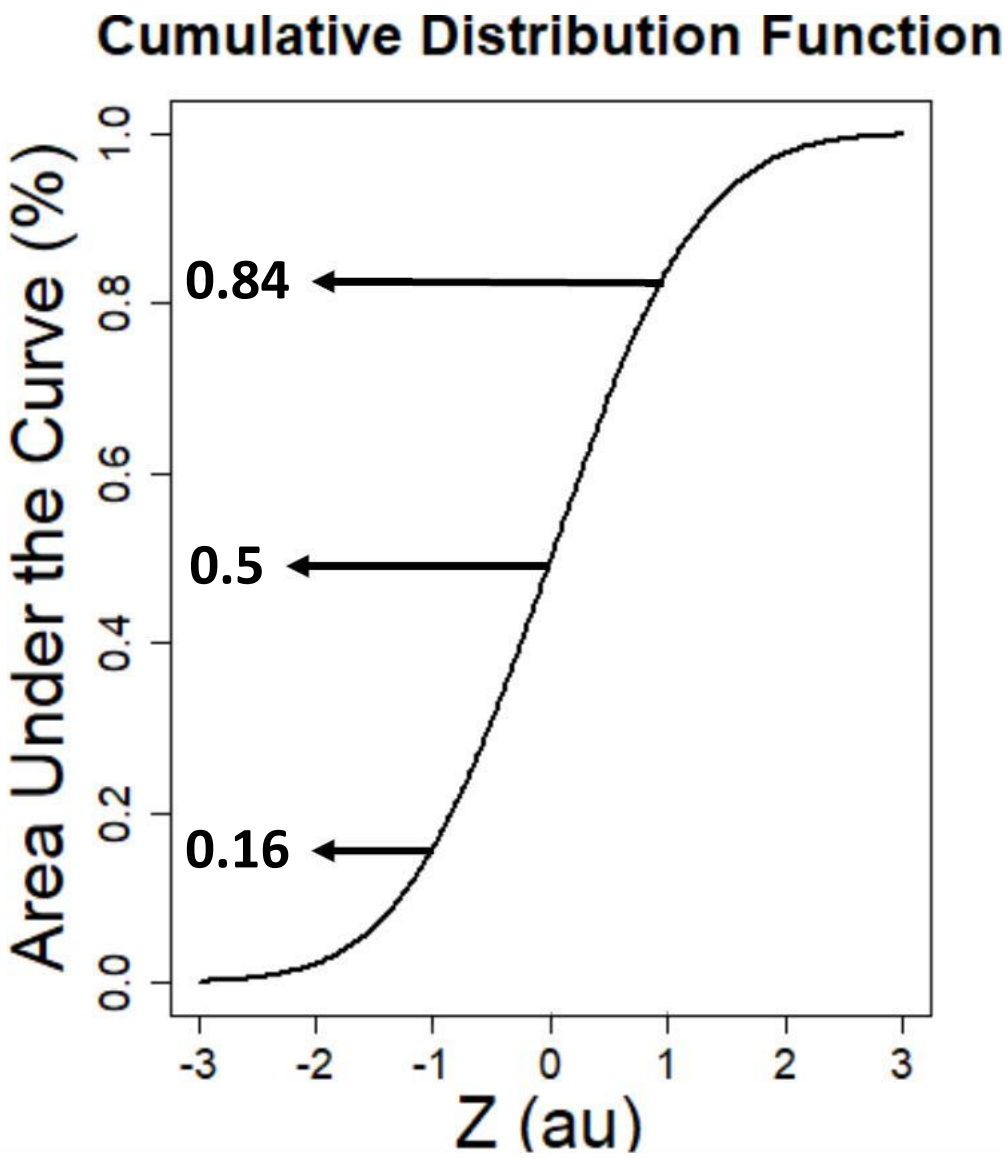


# The Standardized Normal Density Function

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




Normal Density Function
Probability (%)
0.0
0.1
0.2
0.3
0.4
0.84
0.5
0.16
-3
-2
-1
0
1
2
3
Z (au)

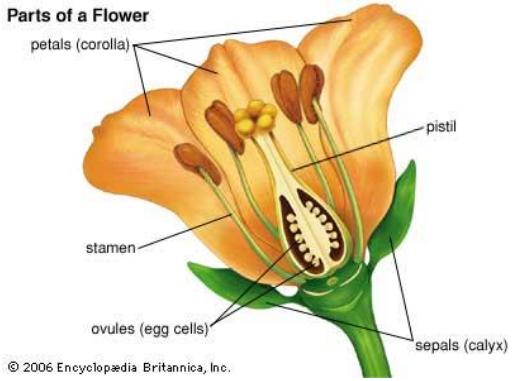

Cumulative Distribution Function
Area Under the Curve (%)
0.0
0.2
0.4
0.6
0.8
1.0
0.84
0.5
0.16
-3
-2
-1
0
1
2
3
Z (au)

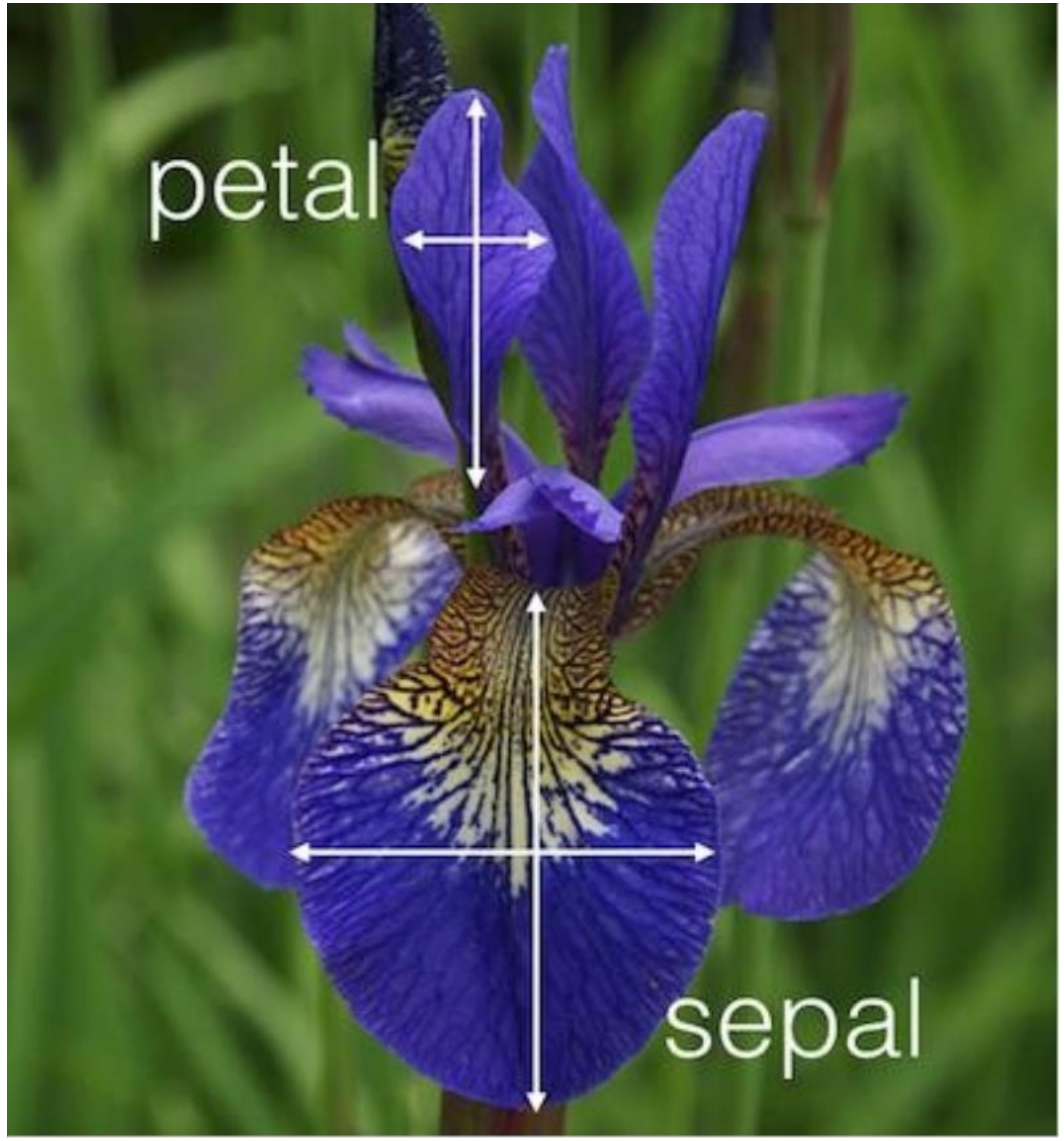
# Why is the "Normal Distribution" important?

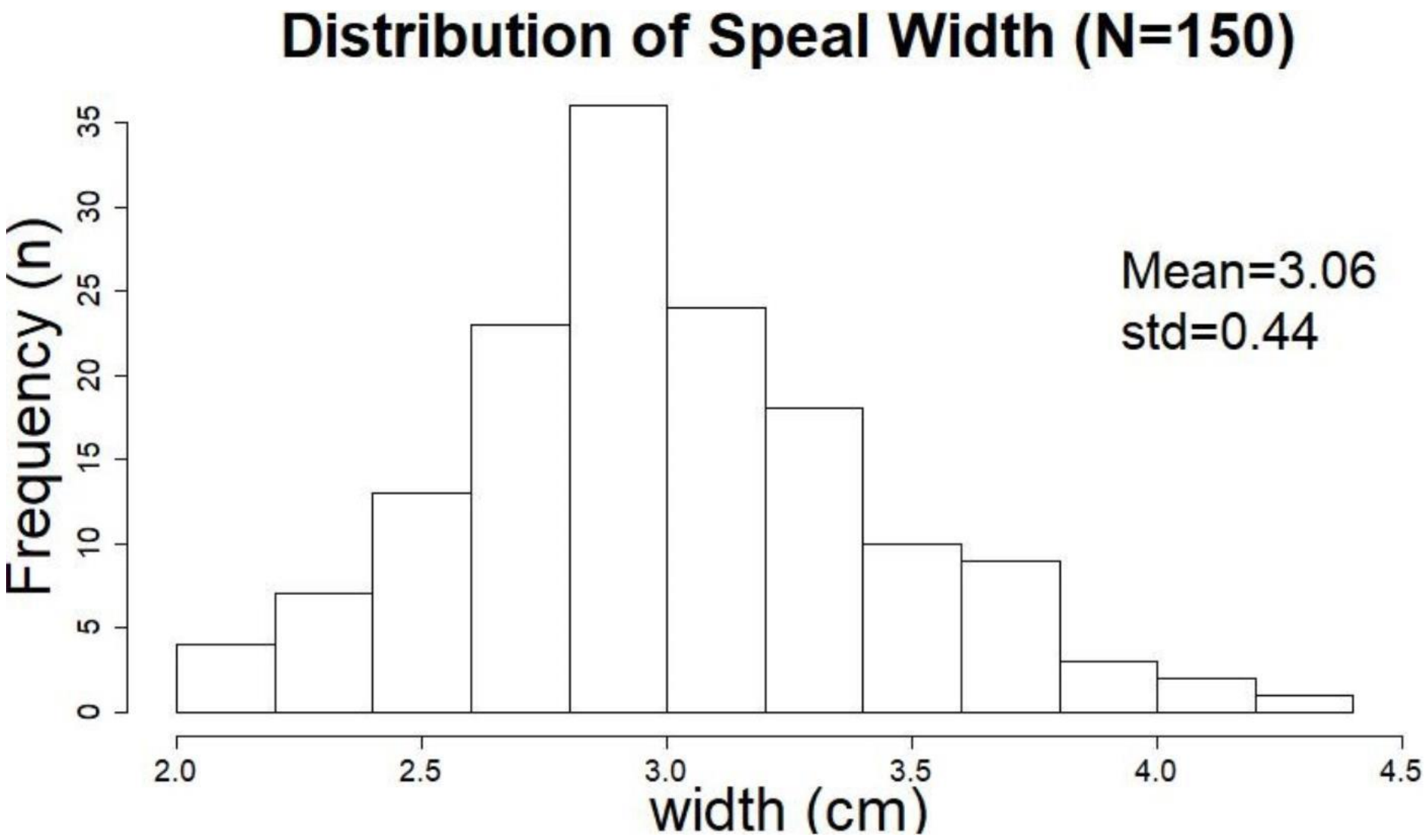
- The distribution of many continuous observables in nature is approximately normal
- We can describe the normal distribution in an equation

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

# The Iris dataset

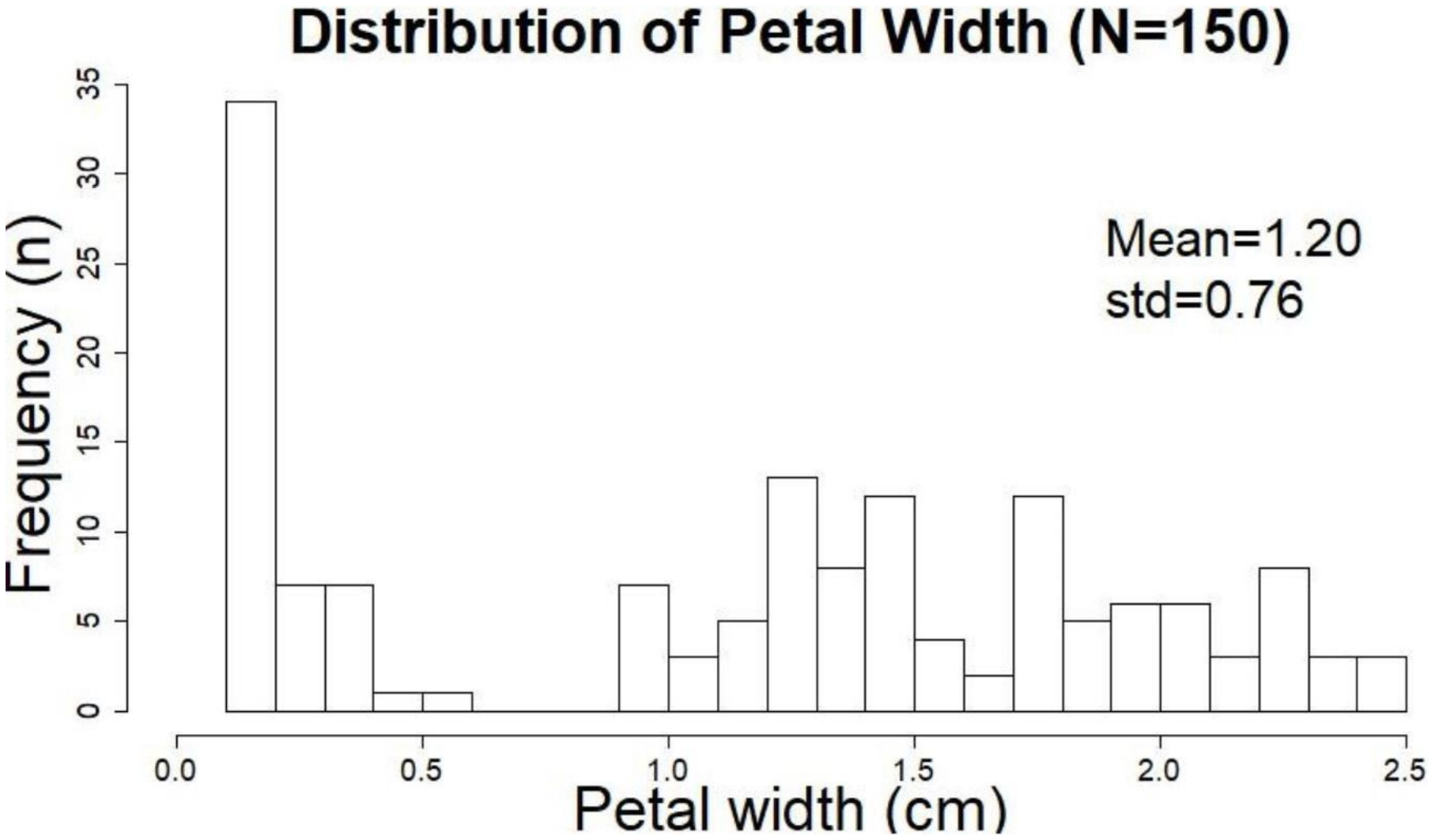



Distribution of Speal Width (N=150)
Frequency (n)
0
5
10
15
20
25
30
35
Mean=3.06
std=0.44
2.0
2.5
3.0
3.5
4.0
4.5
width (cm)

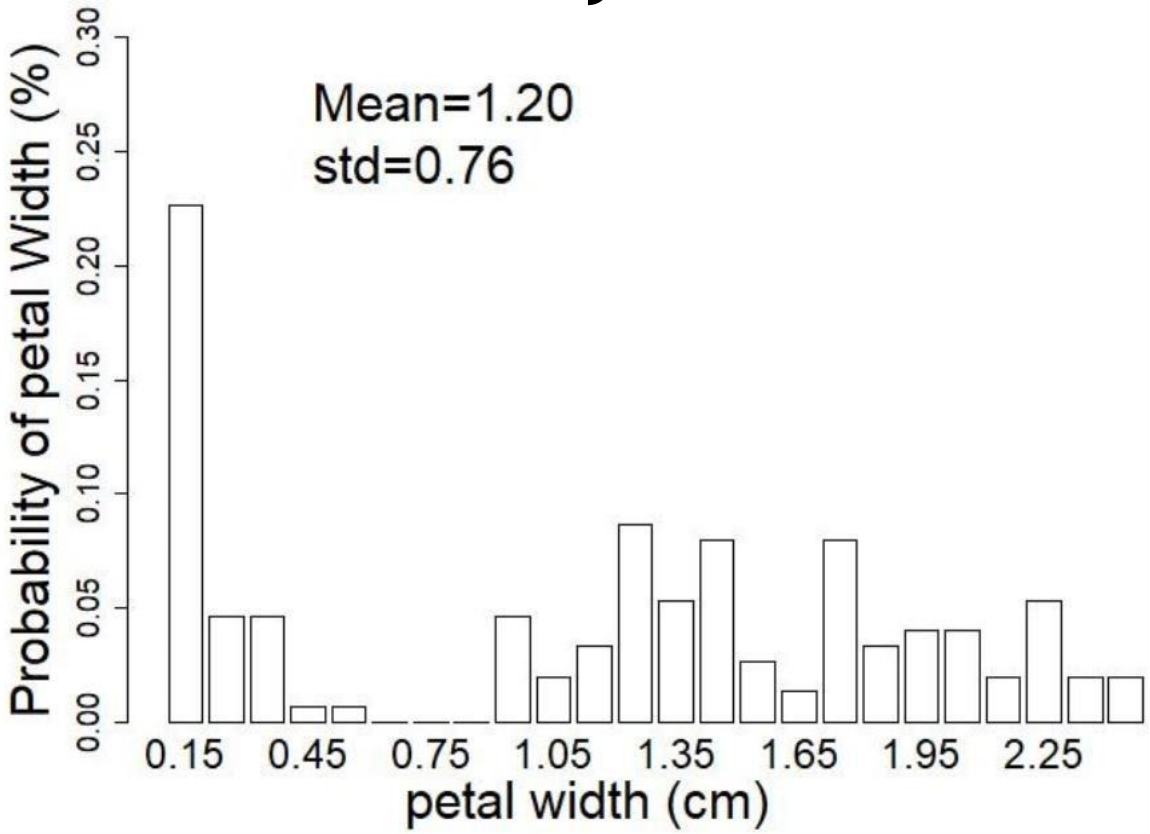
# Conclusion:

NOT ALL CONTINUOUS VARIABLES ARE NORMALLY DISTRIBUTTED



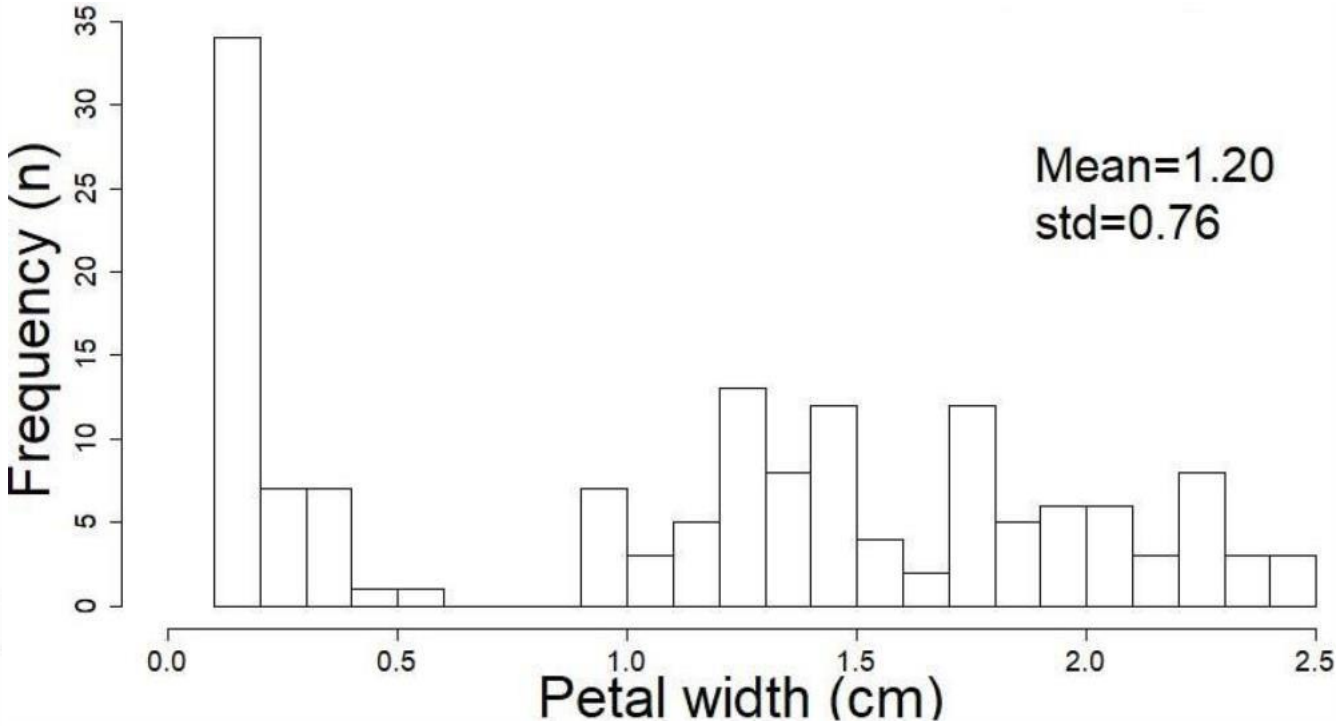
# Frequency vs. Probability Distributions

## Probability Distributions



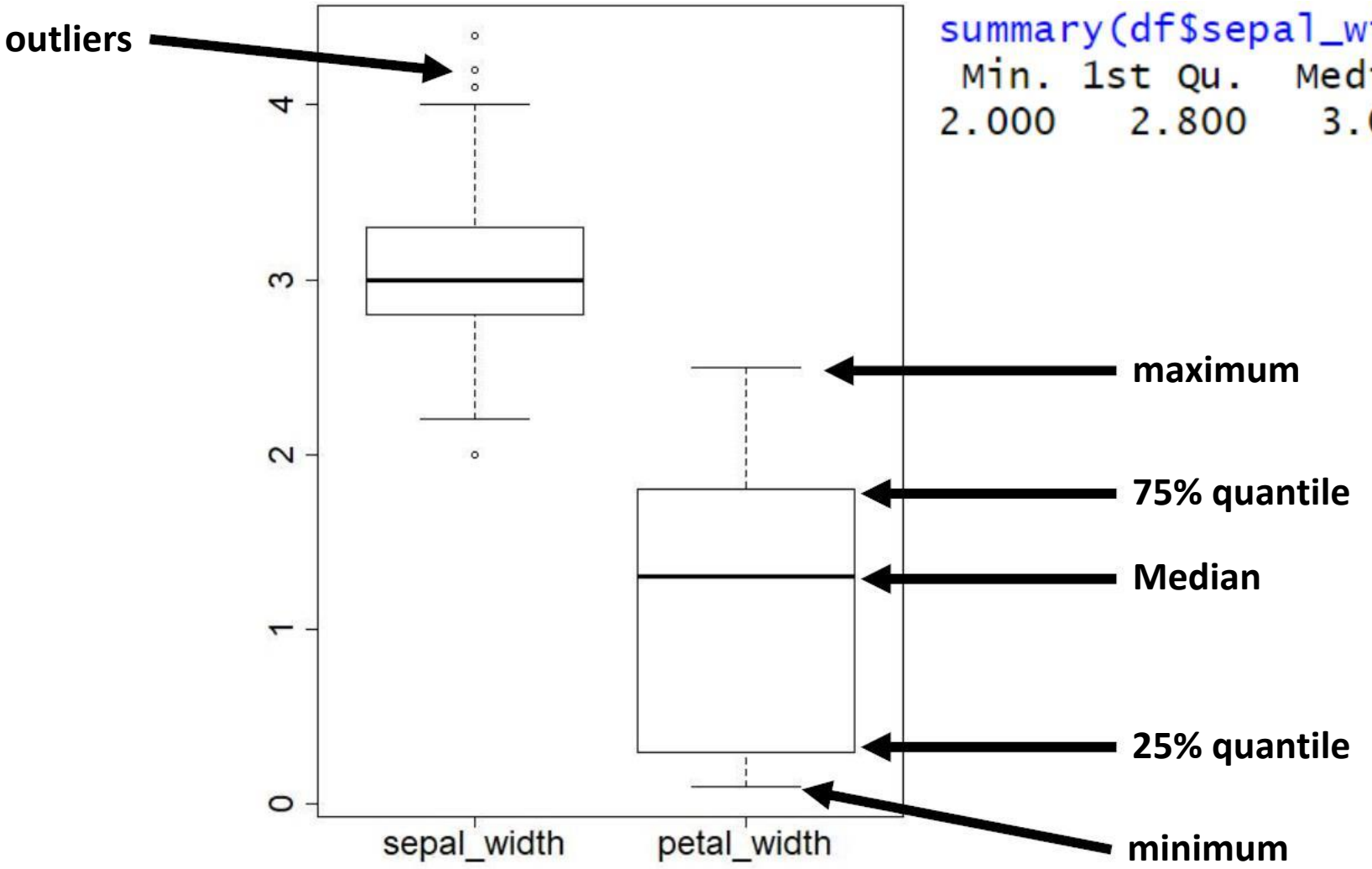

*# frequency of counts per bin*

## Frequency Distributions




*# counts per bin*

# Boxplots and quantiles




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