χ^2 Goodness-of-Fit Test

for Iris Species

Francesco Calzona

Math Modeling

χ^2 distribution

$$X \sim \chi^2_n$$

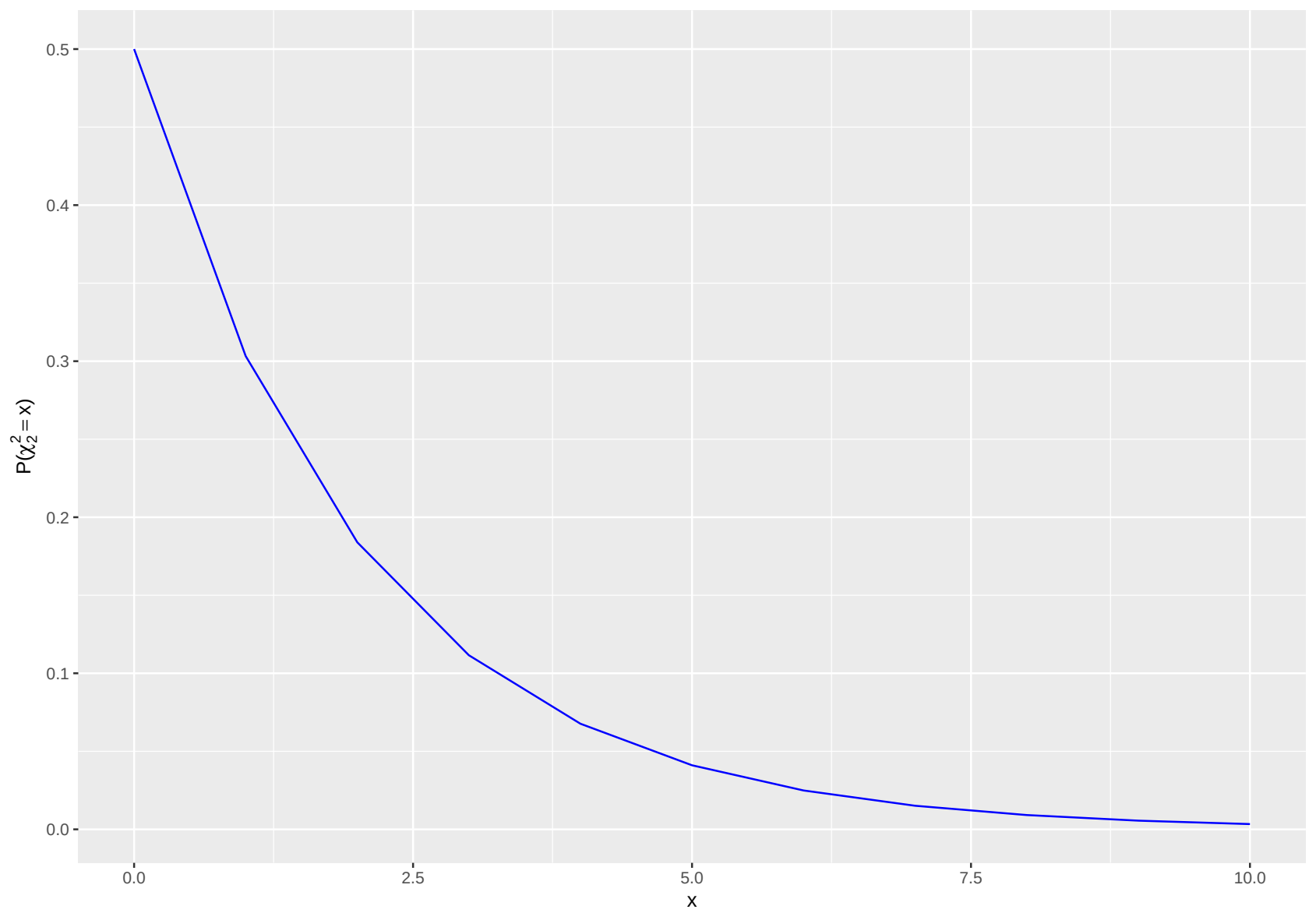
if and only if

$$X = \sum_{i=1}^{n} Z_i^2$$

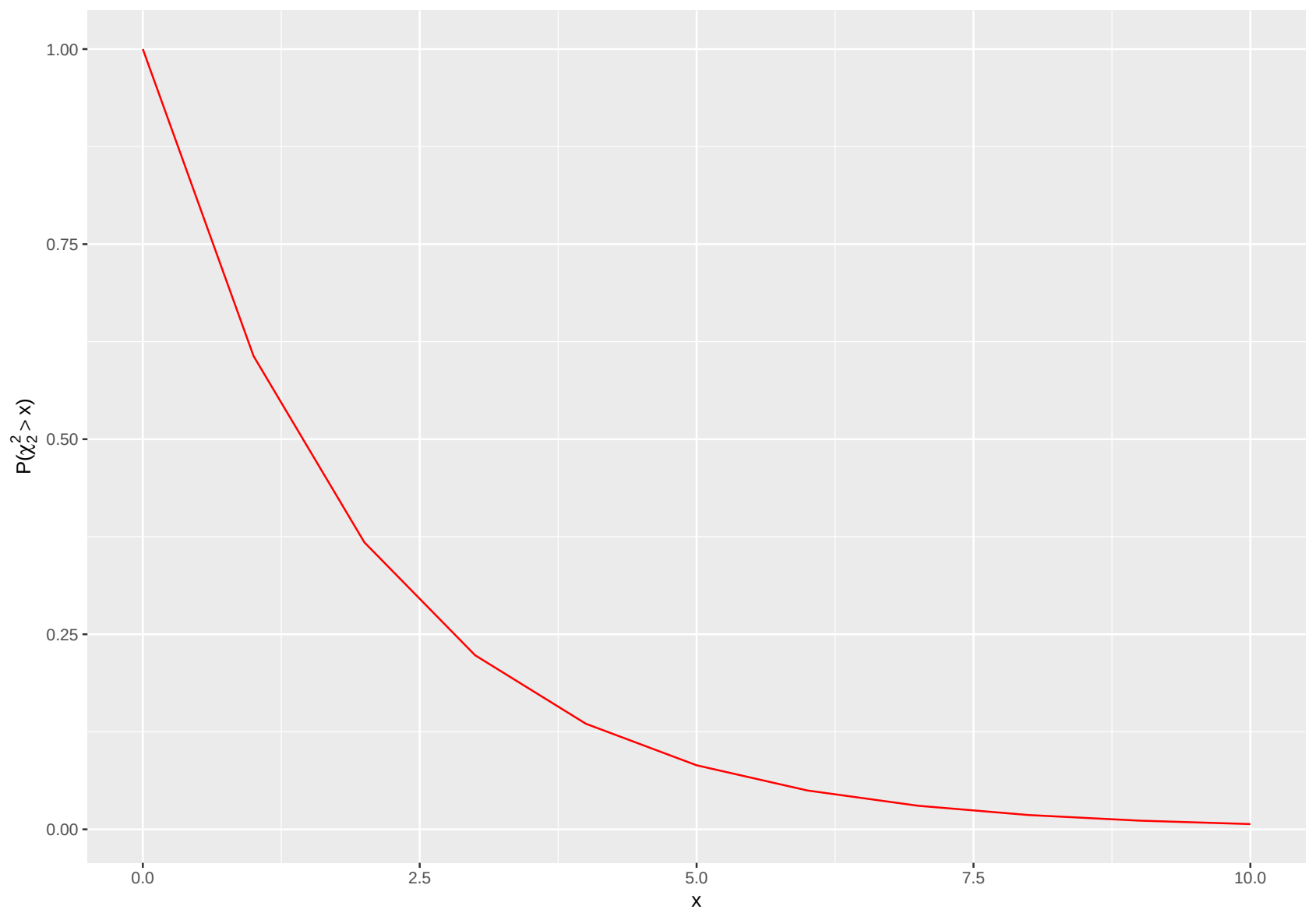
with Z_i independent and

$$Z_i \sim \mathcal{N}(0, 1)$$

χ^2_2 density function



χ^2_2 CCDF function



Approximation to the normal

Let

$$X_i \sim Pois(\mathbf{E}[X_i])$$

If $\mathbf{E}[X_i]$ is big,

$$\frac{X_i - \mathbf{E}[X_i]}{\sqrt{\mathbf{E}[X_i]}} = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Goodness-of-fit test

Assuming that our distribution of assumption H_0 is true,
how likely is it that the population assumes values **equal (or more extreme)** to those of our samples?

χ^2 Goodness-of-Fit Test

Let $C_i = \#$ counts of type i, $n = \#$ of types.

$$H_0 : X_i \sim Pois(E[X_i])$$

for $i \in \{1...n\}$. The **test statistic** is

$$T_s = \sum_{i=1}^{n} \frac{(C_i - \mathbf{E}[X_i])^2}{\mathbf{E}[X_i]} \sim \chi^2_{n-1}$$

and

$$P\text{-}value = \mathbf{P}(\chi^2_{n-1} > T_s)$$

χ^2 Goodness-of-Fit Test for iris dataset

Let $C_i = \#$ counts of species i, $N = \#$ of total observations

$$H_0 : X_i \sim Pois(\frac{N}{3})$$

$$T_s = \sum_{i=1}^{3} \frac{(C_i - \frac{N}{3})^2}{\frac{N}{3}}$$

$$P\text{-}value = \mathbf{P}(\chi_2^2 > T_s)$$

R: Calculating the statistic

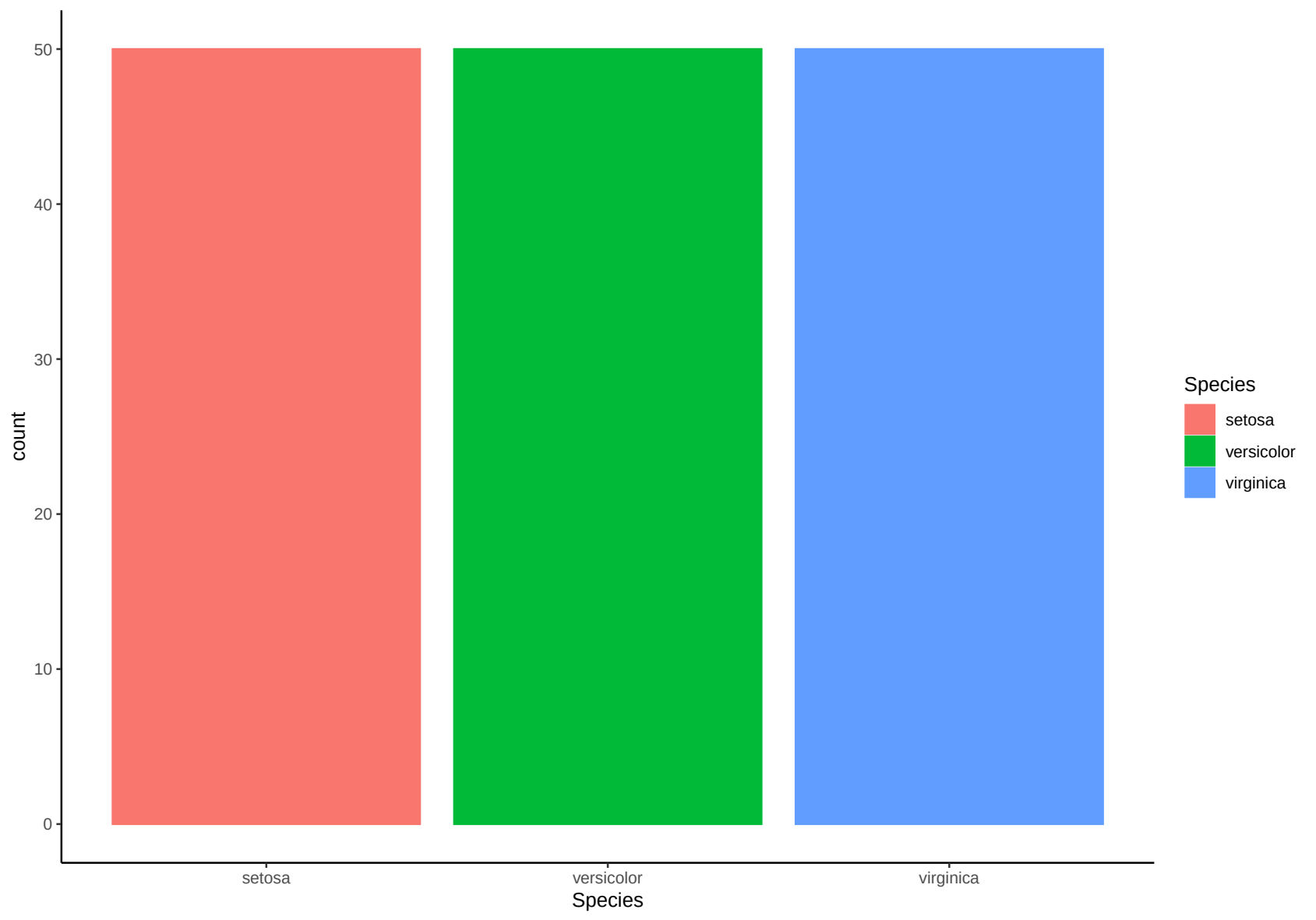
Is our hypothesis plausible?

Let's look into the dataset:

```
data(iris)
summary(iris$Species)
```

```
    setosa versicolor  virginica
        50         50         50
```

Is our hypothesis plausible?



χ^2 Goodness-of-Fit Test in R

Be careful about the `table` type!

```
chisq.test(table(iris$Species), )
```

```
    Chi-squared test for given probabilities

data:  table(iris$Species)
X-squared = 0, df = 2, p-value = 1
```

We cannot reject the hypothesis of equal frequency.

Source code

Presentation created with Quarto and Beamer.