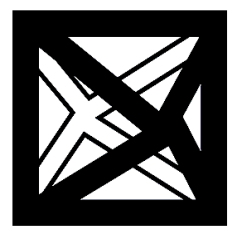
# Pure significance tests

Aldo Solari
Statistical Inference II
PhD in Economics, Statistics and Data Science
University of Milano-Bicocca



XXXVIII cycle

# Main references

- Cox and Hinkley (1976) *Theoretical Statistics.* Chapman and Hall/CRC, §3.1, §3.2, §3.3, §3.4, §6.4, §7.1, §7.2
- Cox (2006) *Principles of Statistical Inference.* Cambridge University Press, §3, §6.2.4
- Davison (2003) *Statistical Models.* Cambridge University Press. Examples 1.1, 3.11, 3.16, 7.24, 7.28

# Outline

von Bortkiewicz's data

Darwin's data

# Null hypotheses

Suppose we have data $y = (y_1, \ldots, y_n)$ and a *null hypothesis* $H_0$ concerning their distribution $F_Y(y) = \text{pr}(Y \leq y)$. It is required to examine the consistency of the data with $H_0$.

$H_0$ is said to be *simple* if it completely specifies $F_Y(y)$ and otherwise *composite*, e.g.

# Table of Contents

von Bortkiewicz's data

Darwin's data

# von Bortkiewicz's horse-kicks data

The table shows how many Prussian *Militärpersonen* died from horse-kicks in each of the 10 corps in each of the 20 successive years 1875 to 1894.

| Deaths | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Frequency | 109 | 65 | 22 | 3 | 1 |

$n = 200$ observations, 122 deaths, average $122/200 = 0.61$

i) Suppose $Y_1, \ldots, Y_n$ i.i.d. Poisson$(\theta)$. Test $H_0 : \theta = \theta_0 = 0.507$ and construct a 95% confidence interval for $\theta$.

ii) Test $H_0 : Y_1, \ldots, Y_n$ i.i.d. Poisson$(\theta)$

# Likelihood

$X = \sum_{i=1}^{n} Y_i \sim \text{Poisson}(\lambda)$ where $\lambda = n\theta$

Null hypothesis $H_0 : \lambda = \lambda_0 = n\theta_0$

Maximum likelihood estimator for $\lambda$: $\hat{\lambda} = x$

Log likelihood

$$\ell(\lambda) = \ell(\lambda; x) = \log(\lambda)x - \lambda - \log(x!)$$

with $\dfrac{\partial \ell(\lambda)}{\partial \lambda} = \dfrac{x}{\lambda} - 1$

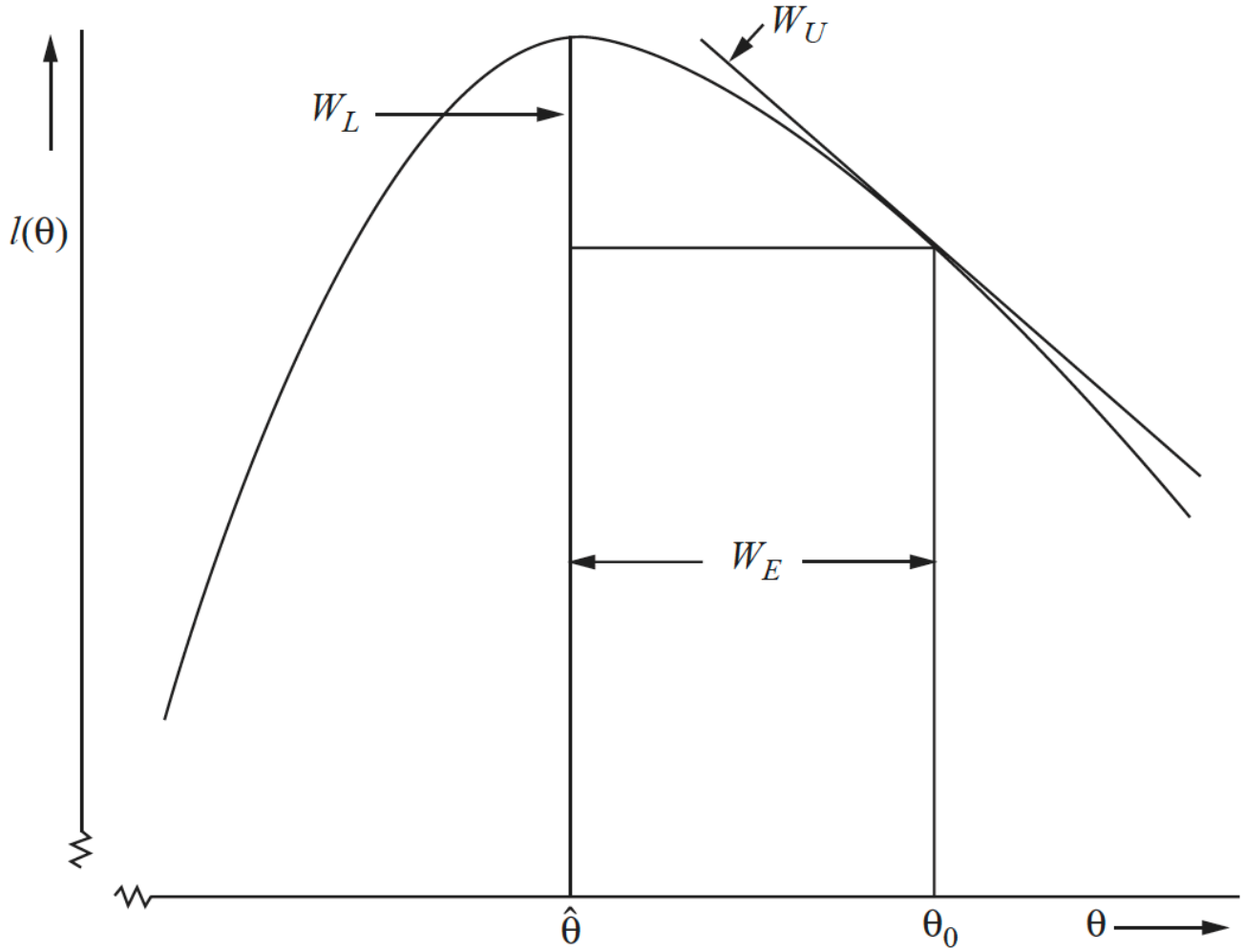
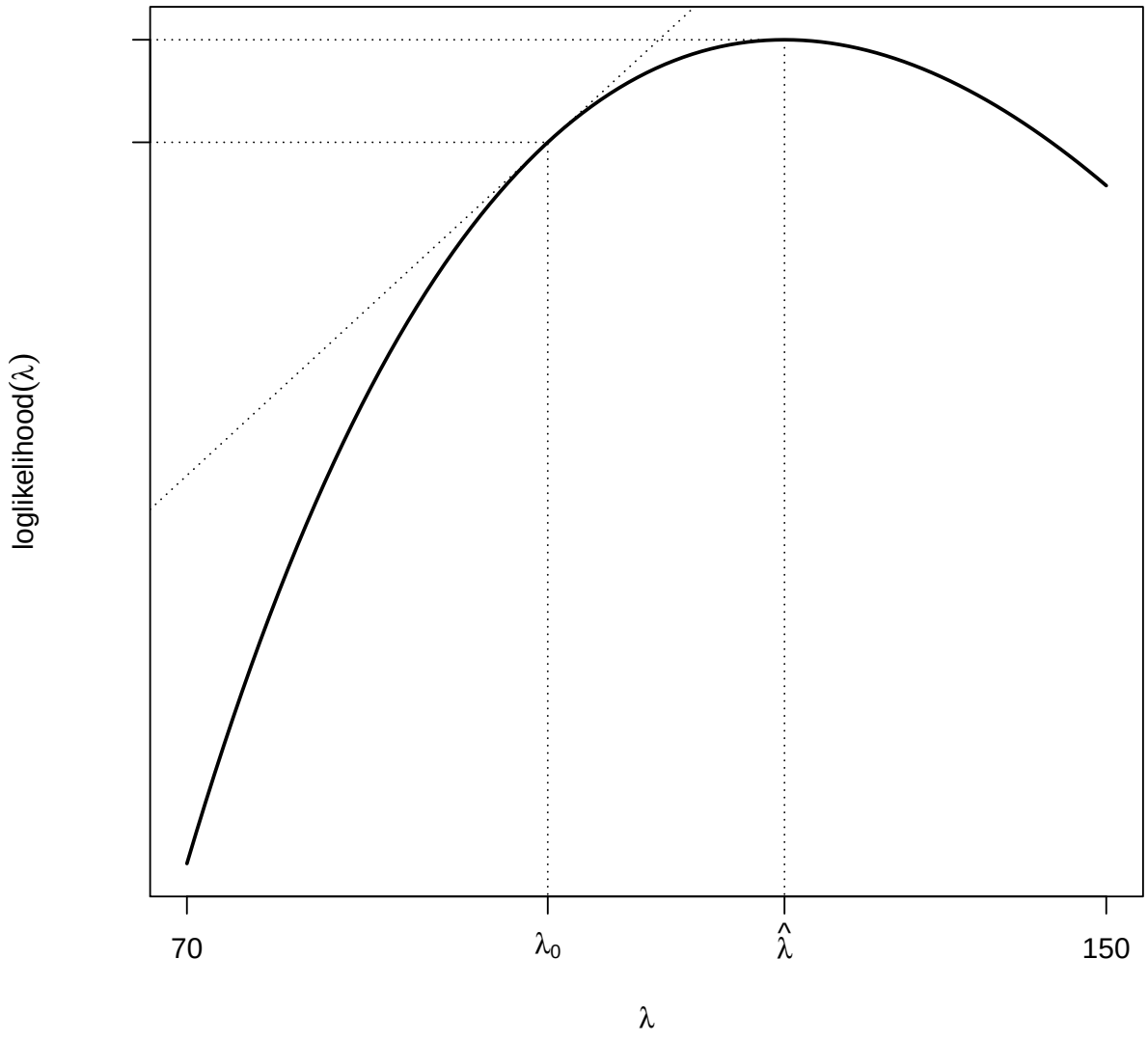


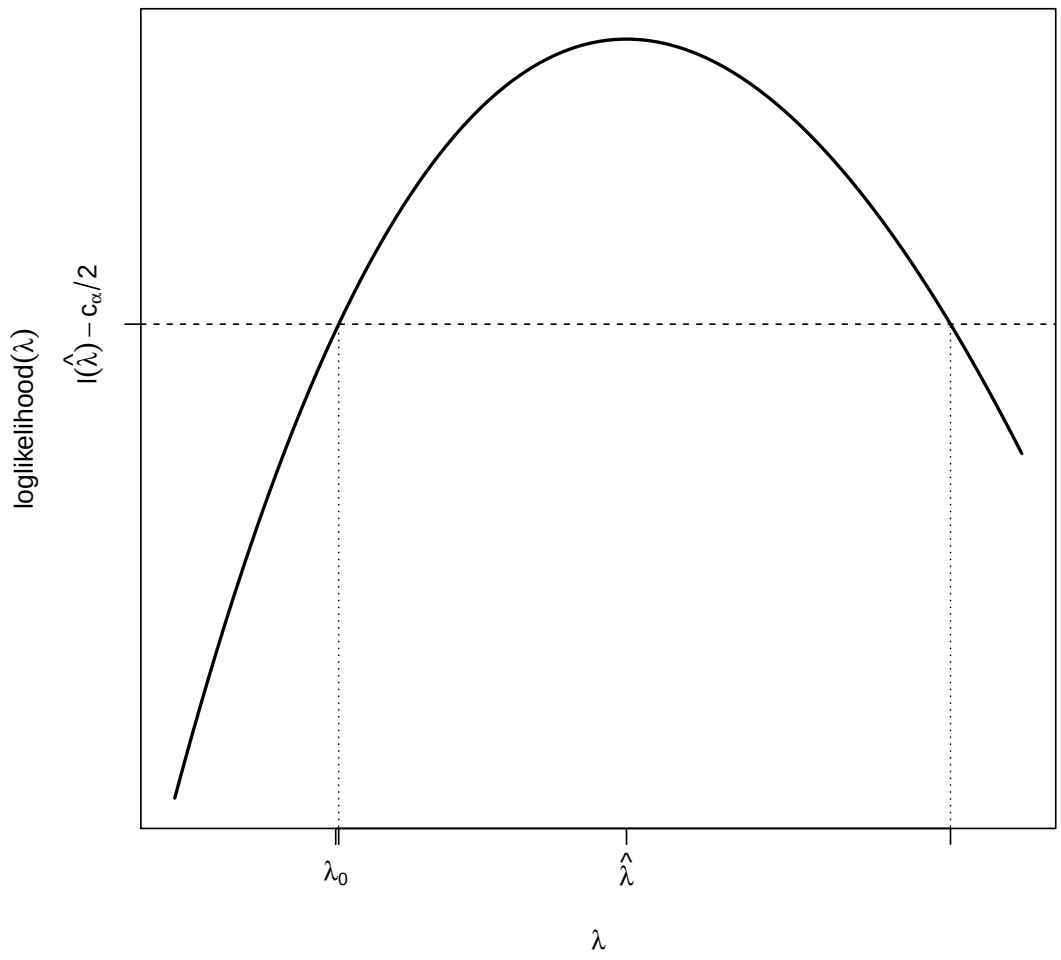

Figure 6.2. Three asymptotically equivalent ways, all based on the log likelihood function of testing null hypothesis $\theta = \theta_0$: $W_E$, horizontal distance; $W_L$ vertical distance; $W_U$ slope at null point.

Source: Cox (2006) *Principles of Statistical Inference.*


loglikelihood(λ)
70
λ0
λ̂
150
λ

$$
\begin{aligned}
T_E &= \frac{\hat{\lambda} - \lambda_0}{\sqrt{\hat{\lambda}}} \approx N(0,1) \qquad \text{(Wald)} \\
T_U &= \frac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0}} \approx N(0,1) \qquad \text{(score or Rao)} \\
T_L &= 2[\ell(\hat{\lambda}) - \ell(\lambda_0)] \approx \chi^2_1 \qquad \text{(likelihood ratio or Wilks)}
\end{aligned}
$$

| | $p$-value for $H_0 : \theta = 0.507$ | 95% confidence interval for $\theta$ |
|---|---|---|
| Wald | 0.0622 | [0.5018, 0.7182] |
| Score | 0.0408 | [0.5109, 0.7283] |
| Likelihood ratio | 0.0475 | [0.5081, 0.7247] |




Likelihood ratio $(1-\alpha)$ confidence interval

$$\{\lambda \geq 0 : \ell(\hat{\lambda}) - \ell(\lambda) \leq c_\alpha/2\}$$

where $c_\alpha$ is the $(1-\alpha)$ quantile of $\chi^2_1$

```
poisson.test(x = 122, T = 200, r = 0.507, conf.level = 0.95)

Exact Poisson test

data:  t time base: n
number of events = 122, time base =
200, p-value = 0.04672
alternative hypothesis: true event rate is not equal to 0.507
95 percent confidence interval:
 0.5065681 0.7283408
sample estimates:
event rate
      0.61
```

# $p$-value

Consider a simple $H_0$.

Let $t = t(y)$ be a function of the observations and $T = t(Y)$ be the corresponding random variable. We call $T$ a *test statistic* for testing $H_0$.

Suppose the larger the value of $t$, the stronger the evidence against $H_0$.

Then if $t_{\text{obs}} = t(y)$ is the observed value of $T$, we define

$$p_{\text{obs}} = \text{pr}_0(T \geq t_{\text{obs}}) = \text{pr}(T \geq t_{\text{obs}}; H_0)$$

the probability being evaluated under $H_0$, to be the (observed) $p$-value of the test

von Bortkiewicz's horse-kicks data:

There are many possible test statistics that might be used, for example $T = \max(Y_1, \ldots, Y_n)$ with

$$\text{pr}_\theta(T \leq t) = \Gamma(t+1, \theta)^n / \Gamma(t+1)^n$$

where $\Gamma(\cdot, \cdot)$ is the incomplete Gamma function. Given $t_{\text{obs}} = 5$ and $H_0 : \theta = \theta_0 = 0.507$, we obtain

$$p_{\text{obs}} = \text{pr}_{\theta_0}(T \geq t_{\text{obs}}) = 0.3083$$

A preference for sufficient statistics leads to $T = \sum_{i=1}^n Y_i$ with $T \sim \text{Poisson}(n\theta)$. Given $t_{\text{obs}} = 122$ and $H_0 : \theta = \theta_0 = 0.507$, we obtain

$$p_{\text{obs}} = \text{pr}_{\theta_0}(T \geq t_{\text{obs}}) = \sum_{t=122}^{\infty} \frac{0.507^t e^{-0.507}}{t!} = 0.0255$$

# $p$-value null distribution

Suppose that $T$ has null cdf $F(t) = \text{pr}_0(T \leq t)$ continuous and invertible.

Then $p_{\text{obs}} = 1 - F(t_{\text{obs}})$ and the corresponding random variable $P = 1 - F(T)$ has standard uniform distribution. Given $u \in [0, 1]$,

$$\begin{aligned} \text{pr}_0(P \leq u) &= \text{pr}_0(1 - F(T) \leq u) \\ &= \text{pr}_0(1 - u \leq F(T)) \\ &= \text{pr}_0(F^{-1}(1 - u) \leq T) \\ &= 1 - F(F^{-1}(1 - u)) = u \end{aligned}$$

By symmetry, the probability integral transform $U = F(T) = 1 - P$ also has a uniform distribution.

If $T$ has a discrete null distribution, $P$ is stochastically larger than uniform, i.e. $\text{pr}_0(P \leq u) \leq u$ for every $u \in [0, 1]$.

# One- and two-sided tests

Often both large and small value of the test statistic are to be regarded as evidence against $H_0$.

Calculate one-sided $p$-values

$$p^-_{\text{obs}} = \text{pr}_0(T \leq t_{\text{obs}}), \quad p^+_{\text{obs}} = \text{pr}_0(T \geq t_{\text{obs}})$$

and define

$$Q = \min(P^+, P^-)$$

as test statistic with $p$-value

$$p_{\text{obs}} = \text{pr}_0(Q \leq q_{\text{obs}})$$

In the continuous case this is $2q_{\text{obs}}$; in a discrete problem it is $q_{\text{obs}}$ plus the achievable $p$-value from the other tail of the distribution, nearest to but not exceeding $q_{\text{obs}}$

$$p_{\text{obs}} = \begin{cases} p^+_{\text{obs}} + \text{pr}_0(P^- \leq p^+_{\text{obs}}) & \text{if } p^+_{\text{obs}} \leq p^-_{\text{obs}} \\ p^-_{\text{obs}} + \text{pr}_0(P^+ \leq p^-_{\text{obs}}) & \text{if } p^+_{\text{obs}} \geq p^-_{\text{obs}} \end{cases}$$

Suppose $T \sim \text{Poisson}(\theta)$ and we want to test $H_0 : \theta = \theta_0$

$$p^+_{\text{obs}} = \text{pr}_{\theta_0}(T \geq t_{\text{obs}}) = \sum_{t=t_{\text{obs}}}^{\infty} \frac{\theta_0^t e^{-\theta_0}}{t!}$$

$$p^-_{\text{obs}} = \text{pr}_{\theta_0}(T \leq t_{\text{obs}}) = \sum_{t=0}^{t_{\text{obs}}} \frac{\theta_0^t e^{-\theta_0}}{t!}$$

For $\theta_0 = 2$ and $t_{\text{obs}} = 3$, $p_{\text{obs}} = \min(0.323, 0.857) + 0.135 = 0.458$

| $t$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $\text{pr}_{\theta_0}(T \geq t)$ | 1 | 0.865 | 0.594 | 0.323 | 0.143 | 0.053 |
| $\text{pr}_{\theta_0}(T \leq t)$ | 0.135 | 0.406 | 0.677 | 0.857 | 0.947 | 0.983 |

von Bortkiewicz's horse-kicks data:

$$p_{\text{obs}}^{+} = 0.0255$$

$$p_{\text{obs}}^{-} = 0.9795$$

$$p_{\text{obs}} = \min(0.0255, 0.9795) + 0.0212 = 0.0467$$

# Confidence sets by the inversion of a family of tests

Let $\theta$ be the parameter of interest, with $\theta \in \Theta$.

Compute $p_{\theta_0}$, a $p$-value testing $H_{\theta_0} : \theta = \theta_0$ for all $\theta_0 \in \Theta$.

Assume that $p$-values are *valid*, i.e.

$$\mathrm{pr}_\theta(P_\theta \leq u) \leq u \quad \forall\, u \in [0,1],\ \forall\, \theta \in \Theta.$$

A $(1-\alpha)$ confidence set for $\theta$ can be constructed by the inversion of the family of tests, i.e.

$$C_\alpha = \{\theta \in \Theta : p_\theta > \alpha\}$$

Then

$$\mathrm{pr}_\theta(C_\alpha \ni \theta) \geq 1 - \alpha \quad \forall\, \theta \in \Theta$$

since $\mathrm{pr}_\theta(\theta \notin C_\alpha) = \mathrm{pr}_\theta(P_\theta \leq \alpha) \leq \alpha$.

# Garwood confidence intervals for a Poisson mean

Let $T \sim \text{Poisson}(\theta)$. Garwood (1936) confidence interval $C_\alpha = [\underline{\theta}_\alpha, \bar{\theta}_\alpha]$ is based on the inversion of the family of tests

$$p_\theta = \min(1, 2\min(p_\theta^-, p_\theta^+)), \quad \theta \geq 0$$

Since $p_\theta^-$ and $p_\theta^+$ are monotonic functions of $\theta$, the limits $\underline{\theta}_\alpha$ and $\bar{\theta}_\alpha$ are the solutions to

$$\begin{aligned} p^+_{\underline{\theta}_\alpha}(t_{\text{obs}}) &= \alpha/2 \\ p^-_{\bar{\theta}_\alpha}(t_{\text{obs}}) &= \alpha/2 \end{aligned}$$

Since $p_\theta^-(t_{\text{obs}}) = 1 - G_{2(t_{\text{obs}}+1)}(2\theta)$ where $G_\nu$ is the cdf of $\chi^2$ with $\nu$ degrees of freedom

$$[\underline{\theta}_\alpha, \bar{\theta}_\alpha] = (\chi^2_{2t_{\text{obs}},\alpha/2}, \chi^2_{2(t_{\text{obs}}+1),1-\alpha/2})/2$$

# Blaker confidence intervals for a Poisson mean

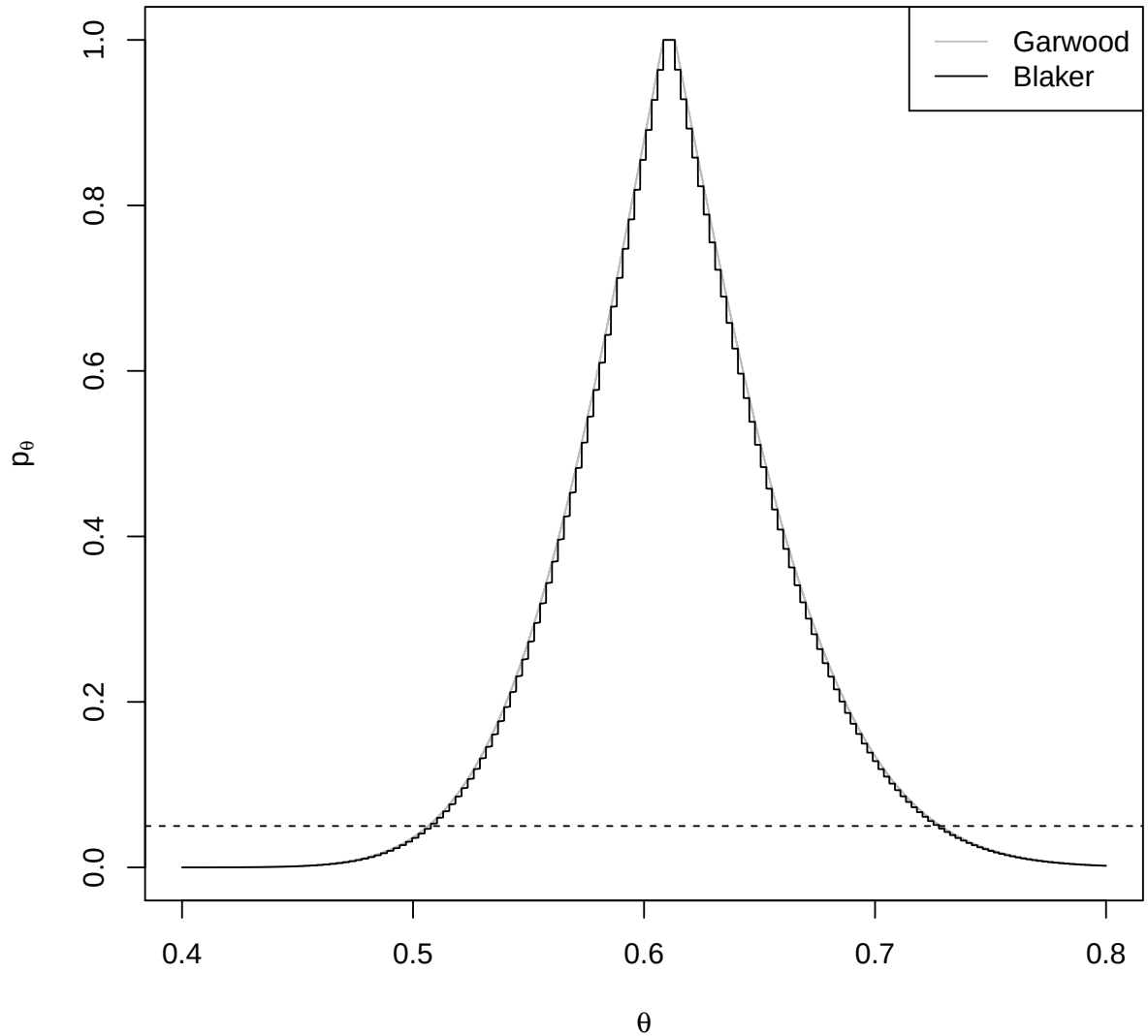
Let $T \sim \text{Poisson}(\theta)$. Blaker (2000) confidence interval $C_\alpha = [\underline{\theta}_\alpha, \bar{\theta}_\alpha]$ is based on the inversion of the family of tests

$$p_\theta = \text{pr}_\theta(Q_\theta \leq q_\theta), \quad \theta \geq 0$$

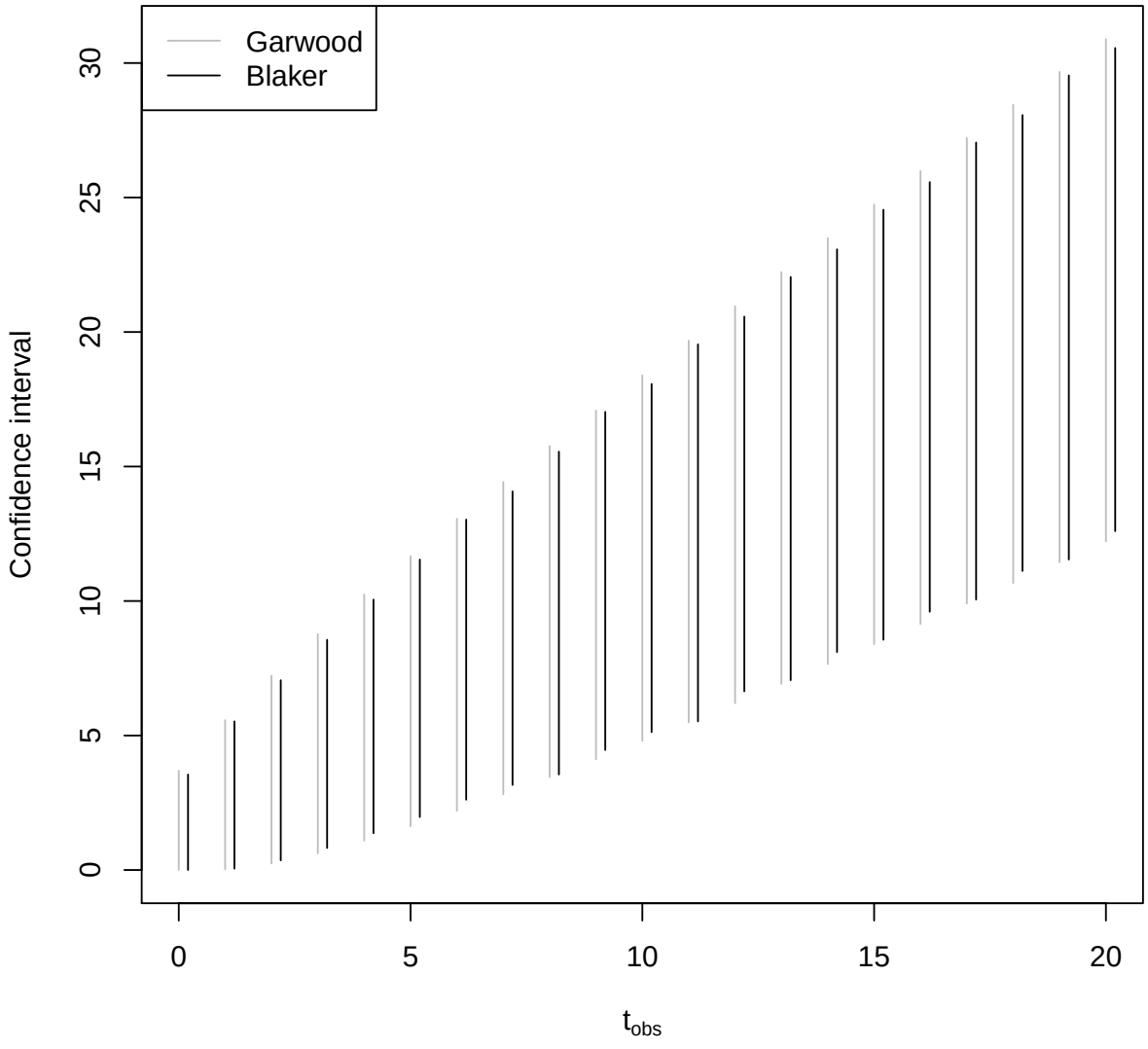
where is $Q_\theta = \min(P_\theta^-, P_\theta^+)$ and $q_\theta = \min(p_\theta^-, p_\theta^+)$

Note that Blaker's $p$-value is always smaller or equal to Garwood's $p$-value, giving shorter confidence intervals. This is because Garwood's confidence interval satisfies the stronger condition

$$\text{pr}_\theta(\underline{\theta}_\alpha > \theta) \leq \alpha/2 \quad \text{and} \quad \text{pr}_\theta(\bar{\theta}_\alpha < \theta) \leq \alpha/2$$




95% confidence intervals
$[0.5066, 0.7283]$ (Garwood) and $[0.5077, 0.7277]$ (Blaker)


Garwood
Blaker
Confidence interval
0
5
10
15
20
25
30
0
5
10
15
20
$t_{obs}$

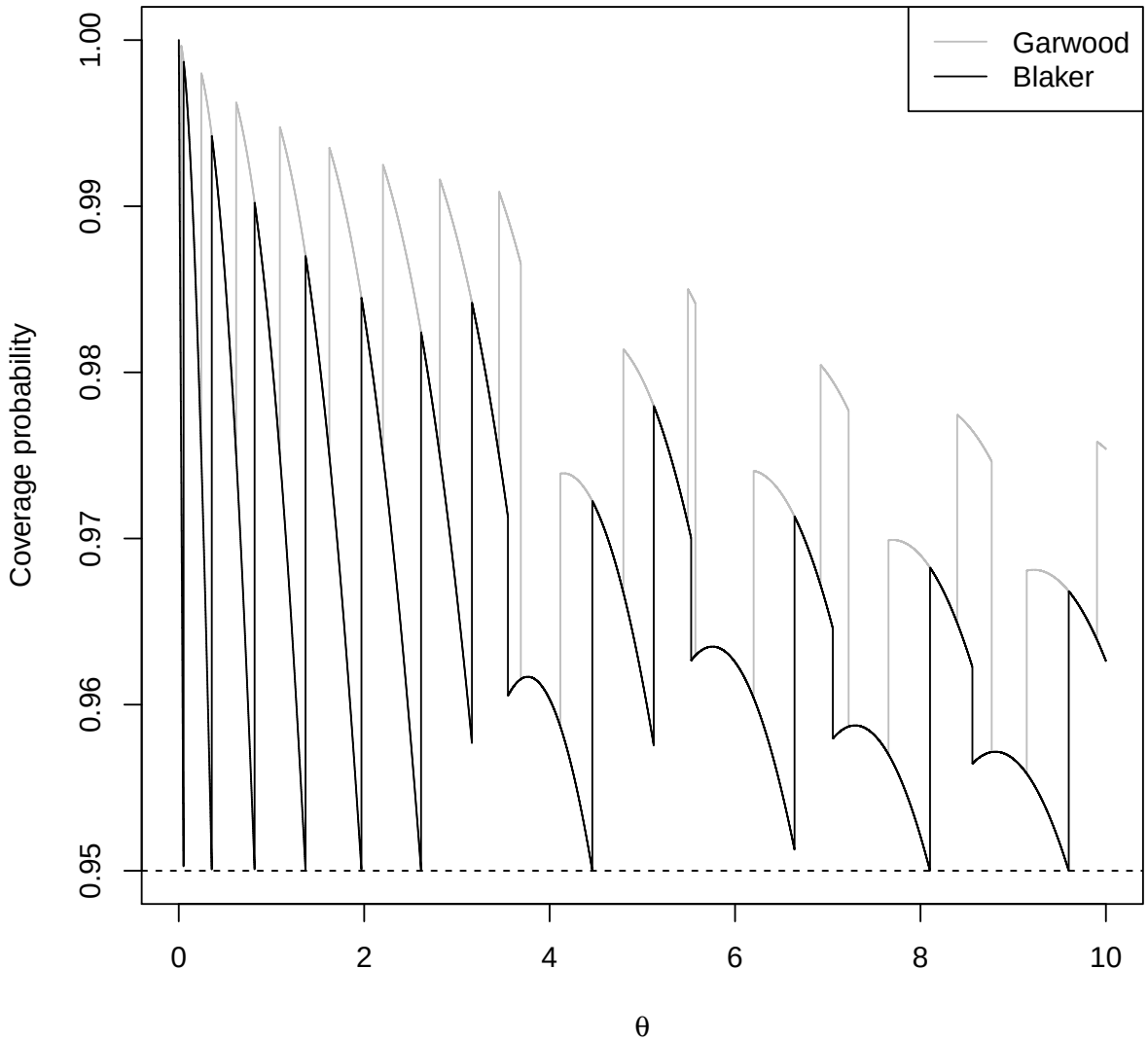

Garwood
Blaker
Coverage probability
1.00
0.99
0.98
0.97
0.96
0.95
0
2
4
6
8
10
θ

TABLE 4

| Deaths. | Frequency observed. | Expected. |
|---|---|---|
| 0 | 109 | 108·67 |
| 1 | 65 | 66·29 |
| 2 | 22 | 20·22 |
| 3 | 3 | 4·11 |
| 4 | 1 | ·63 |
| 5 | ... | ·08 |
| 6 | ... | ·01 |

Table 4 in Fisher (1925) *Statistical methods for research workers.* Oliver & Boyd

Test $H_0 : Y_1, \ldots, Y_n$ i.i.d. Poisson$(\theta)$. We have that $S = \sum_{i=1}^n Y_i$ is a sufficient statistic under $H_0$.

The *sufficiency principle* suggests to examine the conditional distribution of $Y = (Y_1, \ldots, Y_n)$ given $S = s$, i.e.

$$f_{Y|S}(y|s;\theta) = \begin{cases} \frac{s!}{\prod_{i=1}^n y_i!} \frac{1}{n^s} & \text{if } \sum_{i=1}^n y_i = s \\ 0 & \text{if } \sum_{i=1}^n y_i \neq s \end{cases}$$

which is the multinomial distribution corresponding to $s$ trials with equal probabilities $(1/n, \ldots, 1/n)$

Because the multinomial distribution is completely specified, we are testing a simple null hypothesis, but we need to choose a test statistic

The classical Pearson *goodness of fit* statistic

$$\sum_{i=1}^{n} \frac{(Y_i - \mathrm{E}_0(Y_i|S))^2}{\mathrm{E}_0(Y_i|S)}$$

with $\mathrm{E}_0(Y_i|S) = S/n = \bar{Y}$ reduces to the index of dispersion

$$T = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\bar{Y}}$$

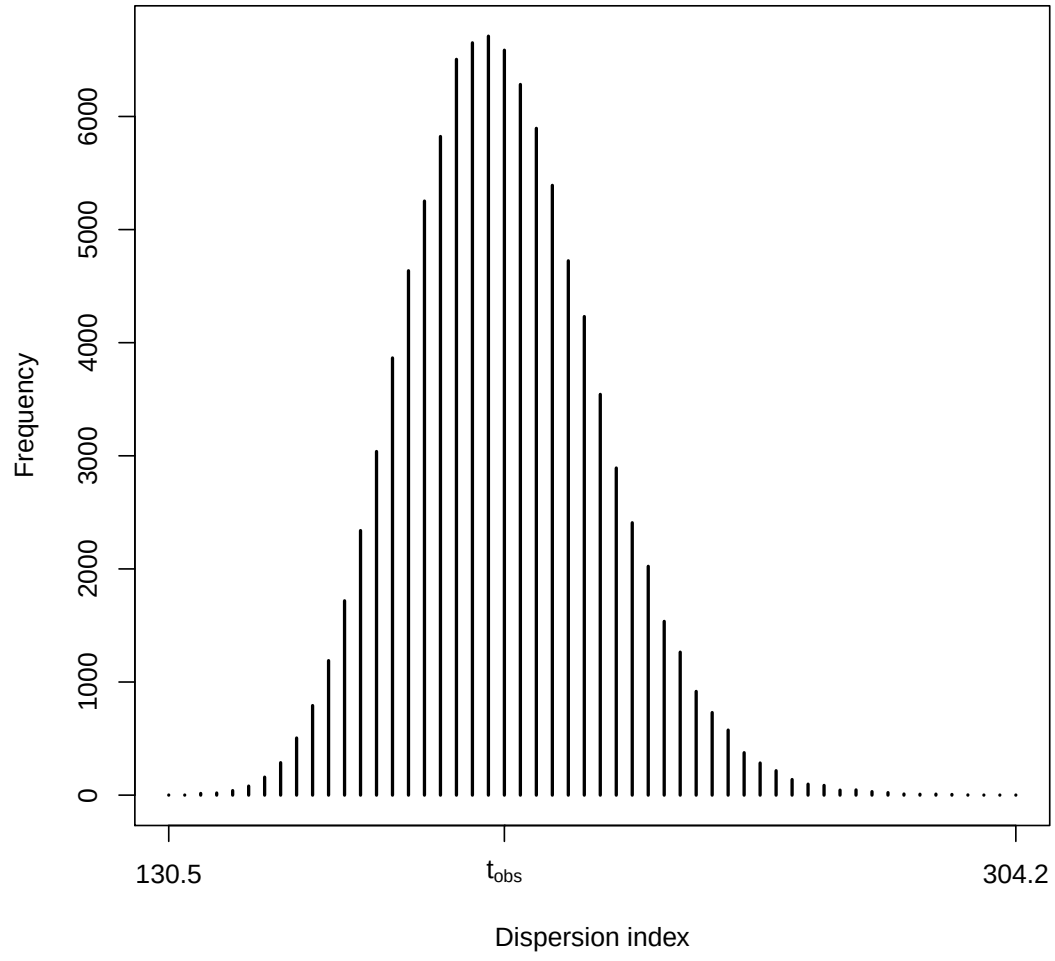
A Poisson distribution is said to be overdispersed (underdispersed) if its variance exceeds (is less than) its mean

The $\chi^2_{n-1}$ approximation for the null distribution of $T$ may be inaccurate.

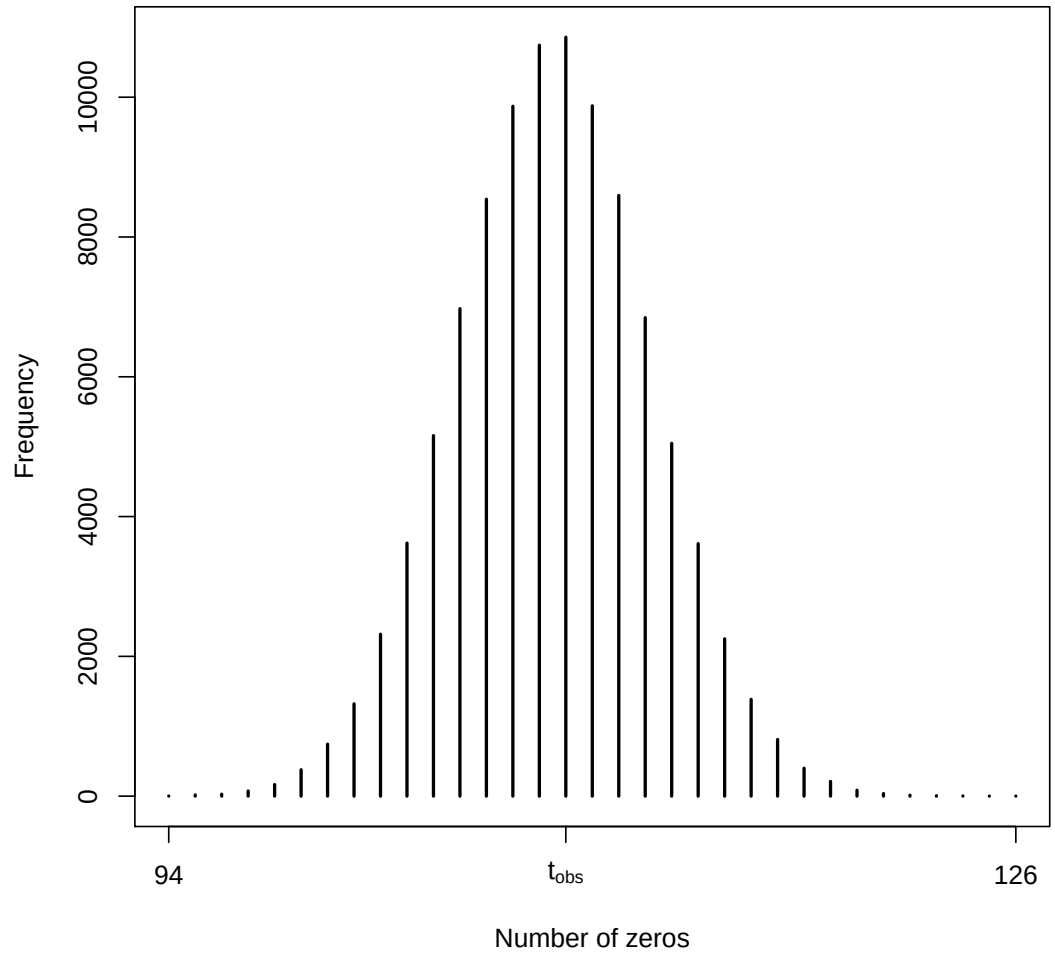
Fisher (1950) advocated for the use of the conditional null distribution of $T$. Calculations can be made by Algorithm AS 171 of Frome (1982) or by using the following Monte Carlo approximation

**Algorithm 1** Monte Carlo $p$-value

1: **for** $b = 1, \ldots, B$ **do**
2: $\quad y^{[b]} = (y_1^{[b]}, \ldots, y_n^{[b]}) \sim \text{Multinomial}(s, (1/n, \ldots, 1/n))$
3: $\quad t^{[b]} = t(y^{[b]})$
4: **end for**
5: $t_{\text{obs}} = t(y)$
6: $p^+_{\text{obs}} = \dfrac{1 + \sum_{i=1}^{n} \mathbb{1}\{t^{[b]} \geq t_{\text{obs}}\}}{B + 1}$




$$p^{+}_{\text{obs}} = 0.503, \quad \text{pr}(\chi^2_{n-1} \geq t_{\text{obs}}) = 0.48$$



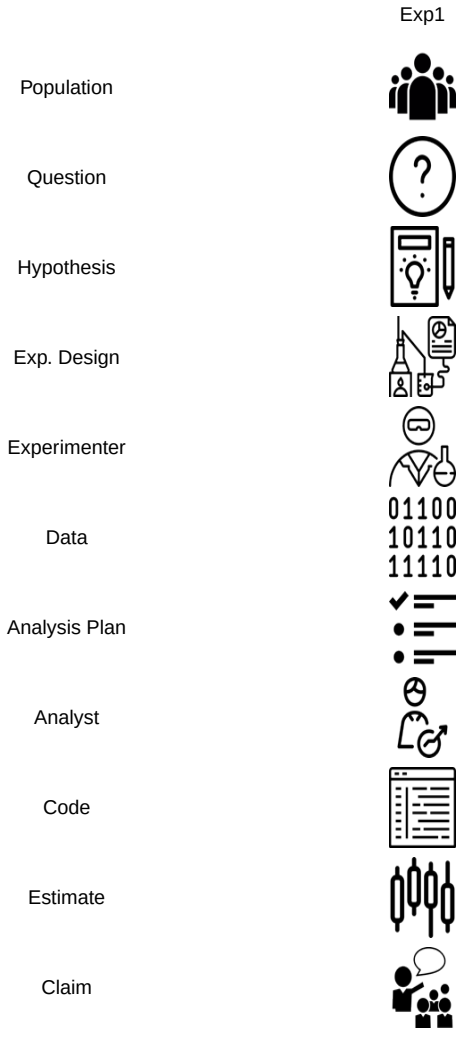

$$t_{\text{obs}} = \sum_{i=1}^{n} \mathbb{1}\{y_i = 0\}$$

# Table of Contents

von Bortkiewicz's data

Darwin's data

# Scientific experiment



# Darwin's experiment

From: Davison (2003) *Statistical Models.* Cambridge University Press.

Charles Darwin collected data over a period of years on the heights of Zea mays plants.

- *Population*: Zea mays plants
- *Hypothesis*: Height of a plant depends on the type of fertilization.
- *Experimental Design*: The plants were descended from the same parents and planted at the same time. Half of the plants were self-fertilized, and half were cross-fertilized, and the purpose of the experiment was to compare their heights (measured in eighths of an inch). To this end Darwin planted them in pairs in different pots.

*Experimental Design (cont'd)*

The focus of interest is the relation between the height of a plant and something that can be controlled by the experimenter, namely whether it is self or cross-fertilized.

This means that you can regard the height as random with a distribution that depends on the type of fertilization, which is fixed for each plant.

Note that in order to minimize differences in humidity, growing conditions, lighting, etc. Darwin had decided to plant the seeds in pairs in the same pots. The height of a plant would therefore also depend on these factors, which are not of interest, not only on the type of fertilization.

# Darwin's data

| | Pot | Cross | Self |
|---|---|---|---|
| 1 | I | 23.500 | 17.375 |
| 2 | I | 12.000 | 20.375 |
| 3 | I | 21.000 | 20.000 |
| 4 | II | 22.000 | 20.000 |
| 5 | II | 19.125 | 18.375 |
| 6 | II | 21.500 | 18.625 |
| 7 | III | 22.125 | 18.625 |
| 8 | III | 20.375 | 15.250 |
| 9 | III | 18.250 | 16.500 |
| 10 | III | 21.625 | 18.000 |
| 11 | III | 23.250 | 16.250 |
| 12 | IV | 21.000 | 18.000 |
| 13 | IV | 22.125 | 12.750 |
| 14 | IV | 23.000 | 15.500 |
| 15 | IV | 12.000 | 18.000 |

- *Analysis plan*: Define a statistical model, and be very specific about the required assumptions. Specify the null hypothesis of no fertilization effect
- *Code*: Write the code to get the results. Your code must be replicable.
- *Claim*: Briefly comment your results

# Galton's specification

Galton assumed a model where the height of a self-fertilized plant is

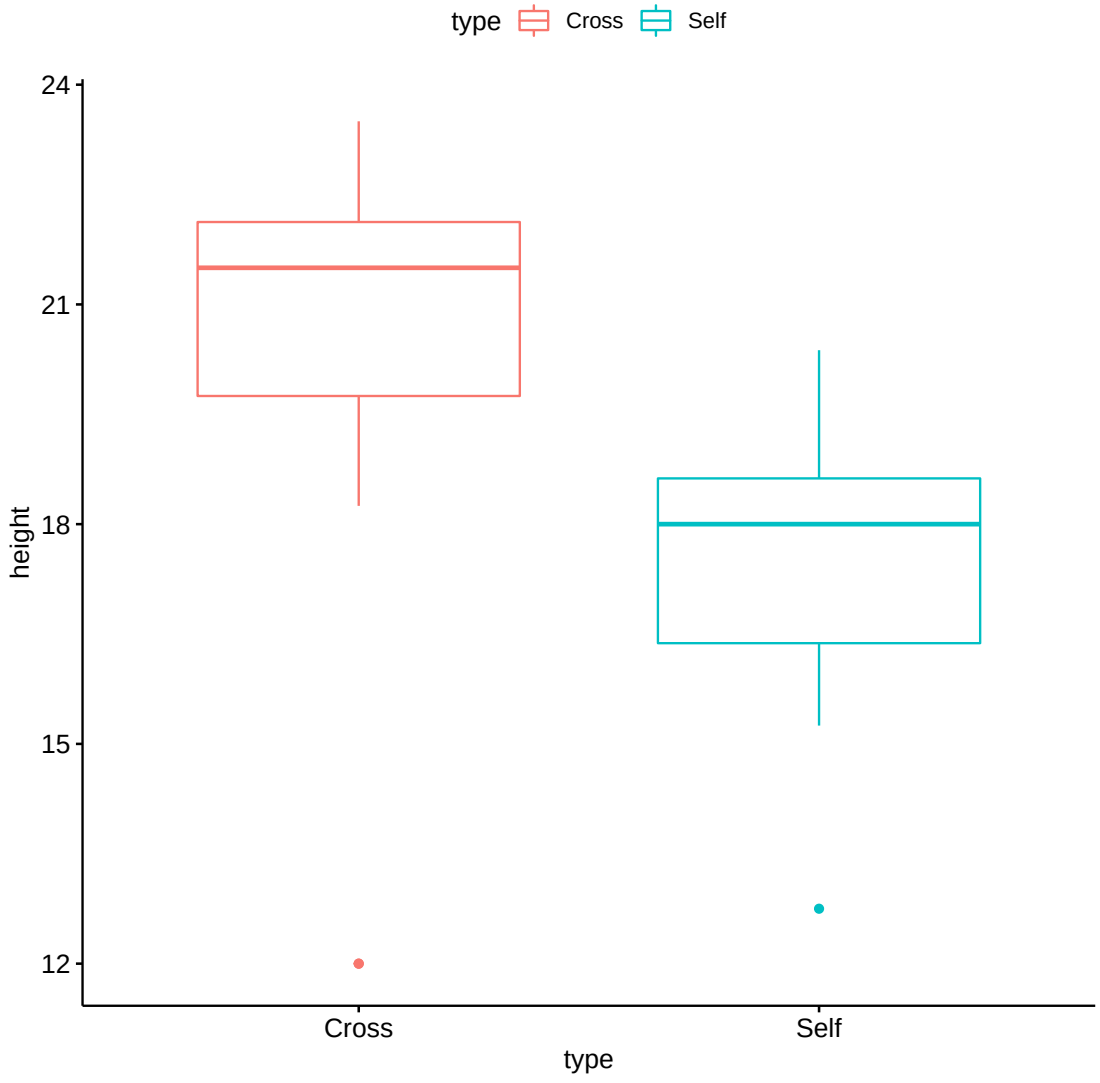
$$Y = \mu + \sigma\varepsilon$$

and of a cross-fertilized plant is

$$X = \mu + \theta + \sigma\epsilon$$

where $\mu$, $\theta$ and $\sigma$ are unknown parameters, and $\varepsilon$ and $\epsilon$ are independent random variables with mean zero and unit variance.

Observations from self-fertilized plants $Y_1, \ldots, Y_{15}$ are i.i.d. as $Y$, and observations from cross-fertilized plants $X_1, \ldots, X_{15}$ are i.i.d. as $X$.


type
Cross
Self
24
21
18
15
12
height
Cross
Self
type

If we assume that $\epsilon_i$ and $\varepsilon_i$ have a $N(0, 1)$ distribution, we can use a *two-sample t test*

```
t = 2.4371, df = 28
p-value = 0.02141

95 percent confidence interval: [0.4173, 4.816]
estimates:
mean in group Cross  mean in group Self
           20.19167             17.57500
```

# Fisher's specification

$(X_i, Y_i)$ are the heights of $i$th pair (cross-fertilized, self-fertilized)

Consider the model

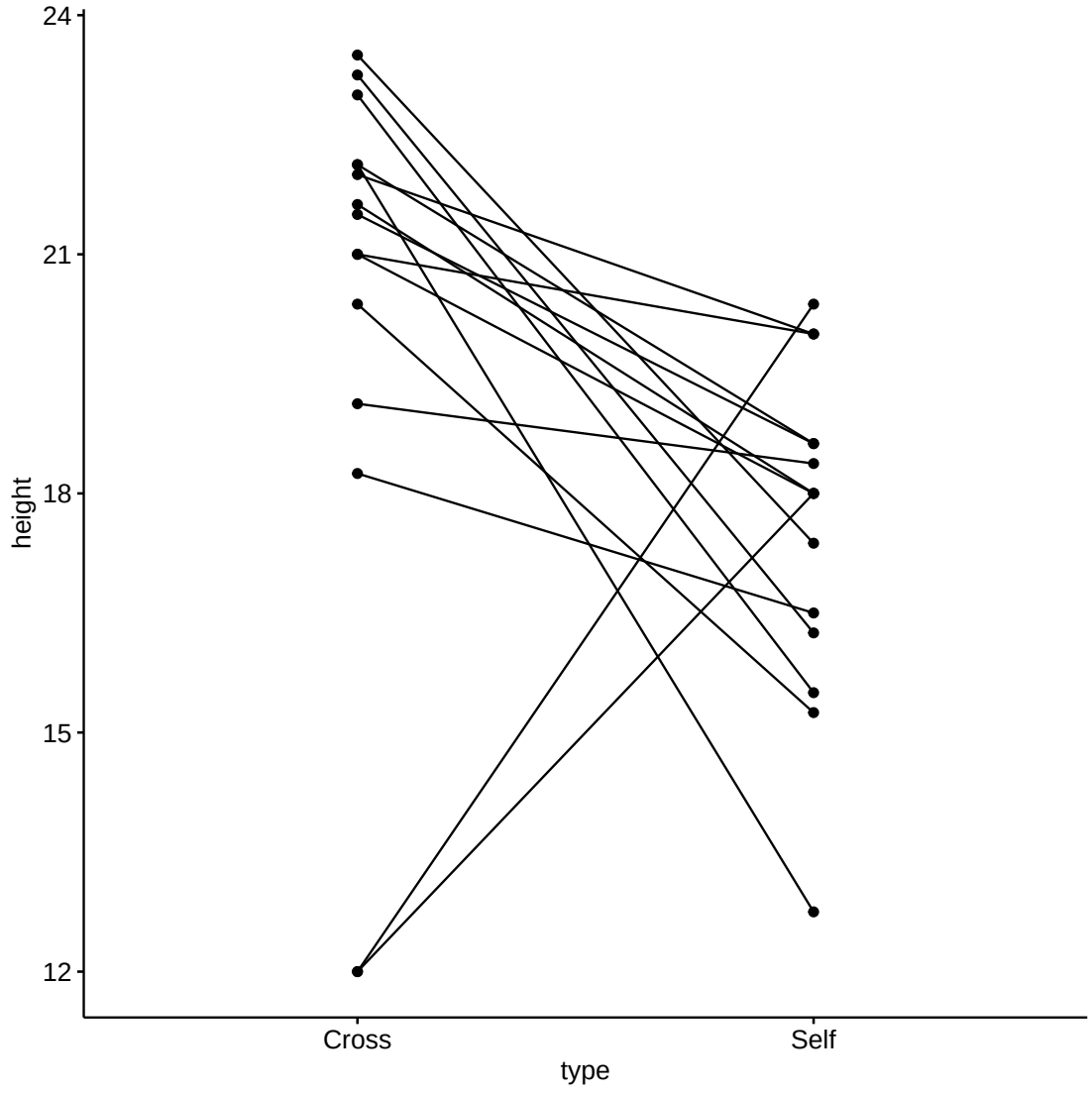
$$X_i = \mu_i + \theta + \sigma\epsilon_i \quad Y_i = \mu_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

The parameter $\mu_i$ represents the effects of the planting conditions for the $i$th pair, and $\varepsilon_i$ and $\epsilon_i$ are independent random variables with mean zero and unit variance.

The $\mu_i$ could be eliminated by using the differences

$$Z_i = X_i - Y_i$$

which have mean $\theta$ and variance $2\sigma^2$


24
21
18
15
12
height
Cross
Self
type

If we assume that $\epsilon_i$ and $\varepsilon_i$ have a $N(0, 1)$ distribution, we can use a *paired t test*, or one-sample t test for the difference

$$H_0 : \theta = 0$$

```
t = 2.148, df = 14
p-value = 0.0497

95 percent confidence interval: [0.0039, 5.2294]
estimate: 2.616667
```

# Test of symmetry

Tests where the null hypothesis is formulated in terms of arbitrary distributions are called *nonparametric* or *distribution-free tests*

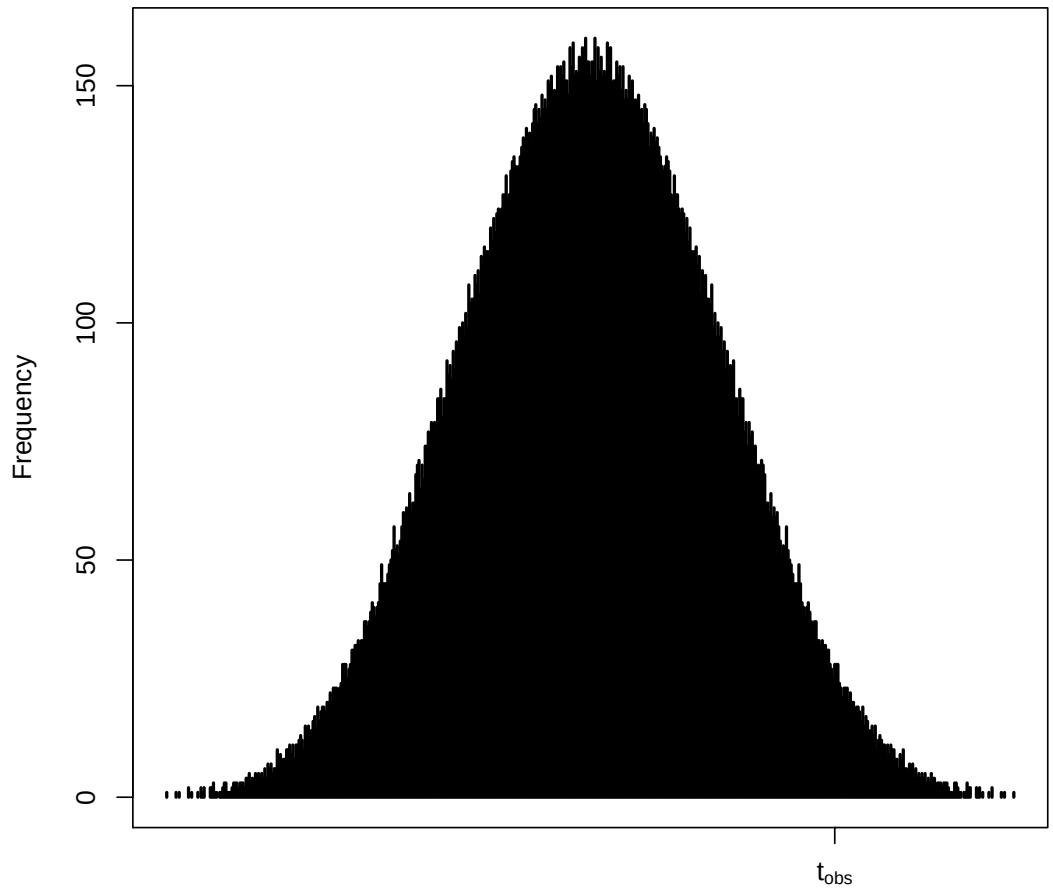
The hypothesis of no effect is then equivalent to the assumption that the distribution $F_Z$ of the difference $Z = X - Y$ is is symmetric about zero

$$H_0 : F_Z(-z) + F_Z(z) = 1$$

Under this hypothesis, all points $z$ and $-z$ have equal probability, so that the sufficient statistic is determined by the order statistics of the $|z_i|$.

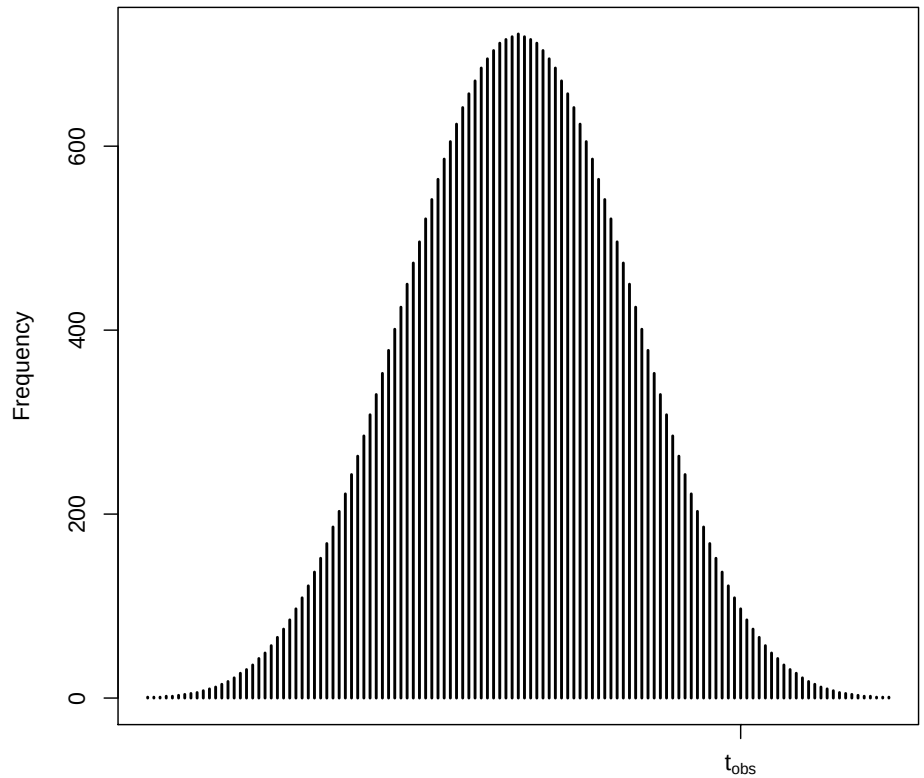
Further, conditionally on the sufficient statistic, all $2^n$ sample points $\pm z_i$ have equal probability $1/2^n$. Thus the distribution under the null hypothesis of any test statistic is in principle exactly known.

# Test of symmetry




$$t_{\text{obs}} = \sum_{i=1}^{n} z_i$$

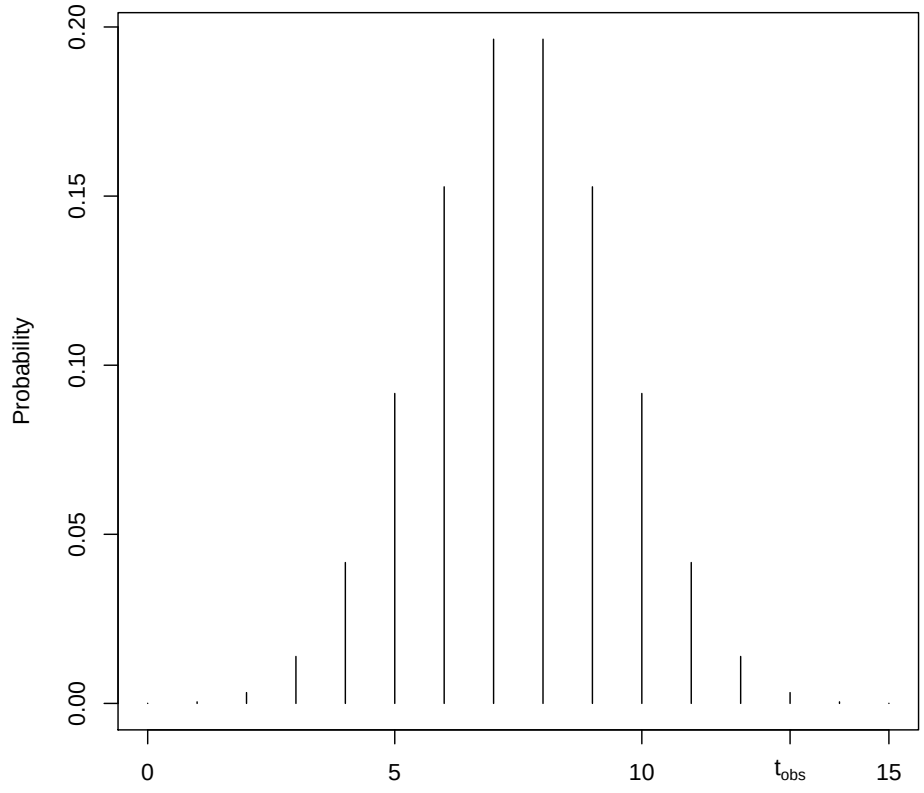
# Wilcoxon signed-rank test




$$t_{\text{obs}} = \sum_{i=1}^{n} \text{sign}(z_i) r_i$$

where $r_1, \ldots, r_n$ are the ranks of $|z_1|, \ldots, |z_n|$

# Sign test




$$H_0 : \text{pr}(Z > 0) = 0.5, \quad T = \sum_{i=1}^{n} \mathbb{1}\{Z_i > 0\} \overset{H_0}{\sim} \text{Binomial}(n, 0.5)$$