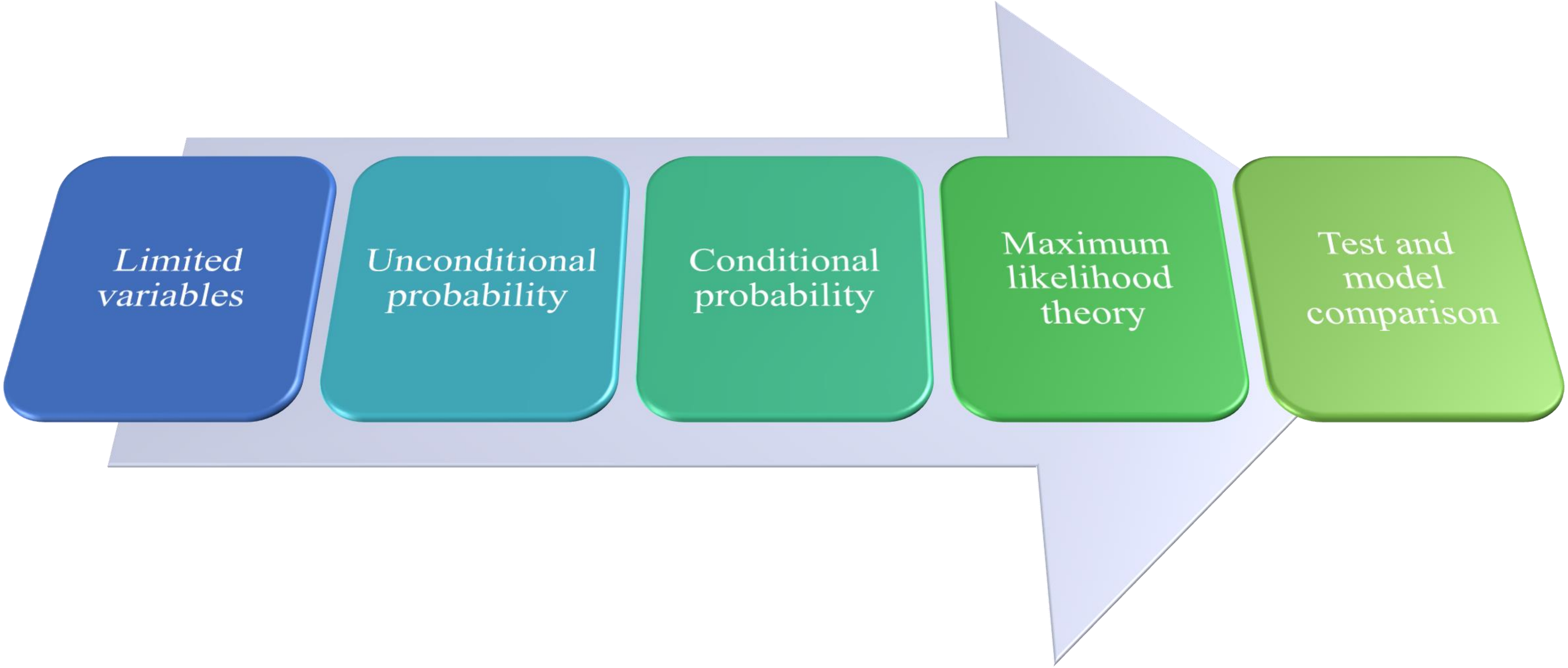
# Introduction to Econometrics

*Limited dependent variables*

ESSEC BUSINESS SCHOOL

# Outline of this course



# *Limited dependent variables*
# Introduction

# Limited variable ?

**<u>Limited dependent variable:</u>** Discrete variable takes on finite values.

**<u>Binary variables:</u> Possible values: 0/1**

<u>Any output of a dichotomous question:</u>

- Do I feel good today ? (time series)
- Did this bank collapse this year ? (cross-sectional data)
- Will this app become a killer app ? (cross-sectional data or time series)
- Is there a difference between low-cost and high-cost cheese cubes ?

**9,96 euro/kg**



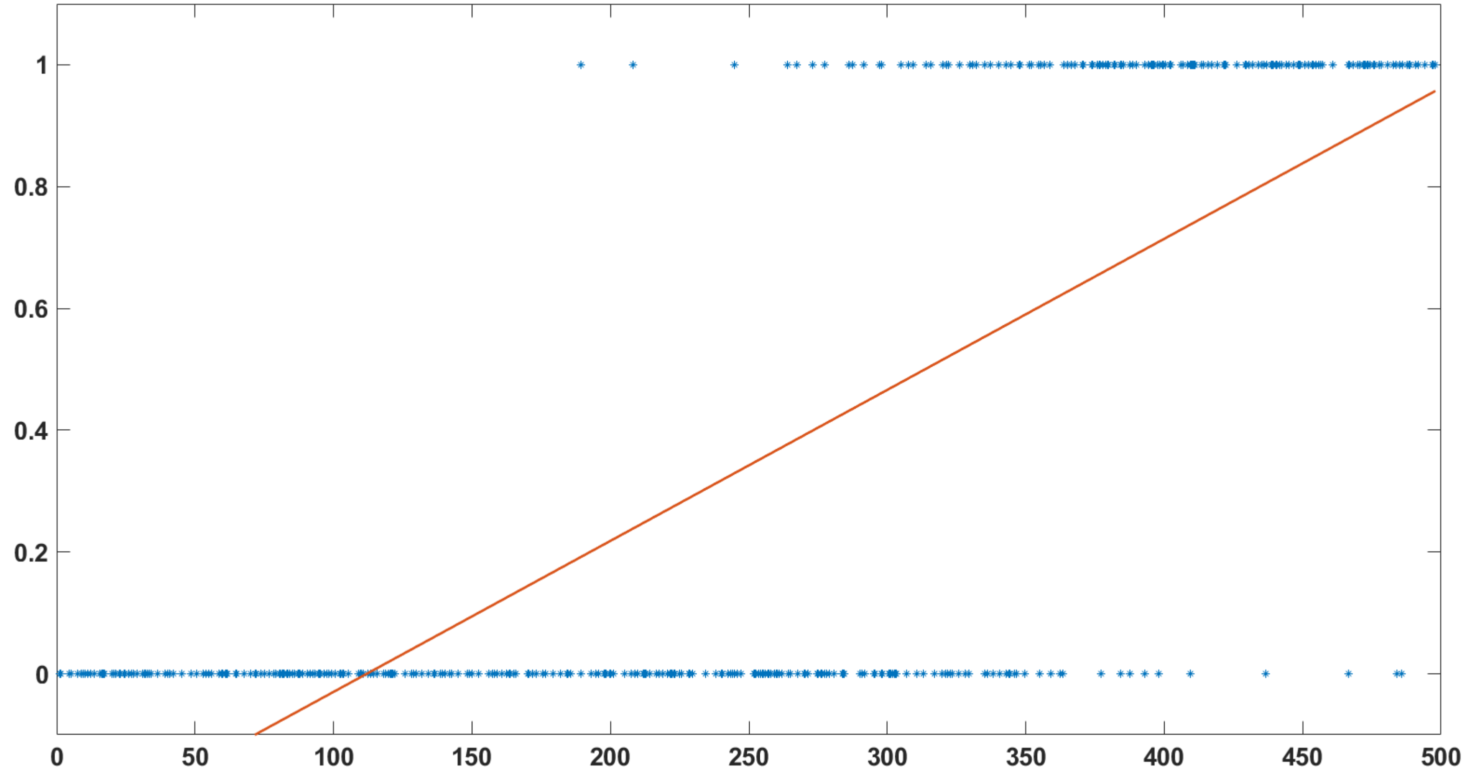
**23,27 euro/kg**

# Linear regression with limited variables

**Dependent variable:** **Paid dividend in 2019**

**Explanatory variable:** **Market capitalisation (in billions $)**

$$\mathrm{Div}_t = -0.27 + 0.0025\mathrm{cap}_t$$




**Interpretation as a probability?**

$$\mathrm{cap}_t = 50 \rightarrow \mathrm{Div}_t = -0.15$$
$$\mathrm{cap}_t = 1000 \rightarrow \mathrm{Div}_t = 2.2$$

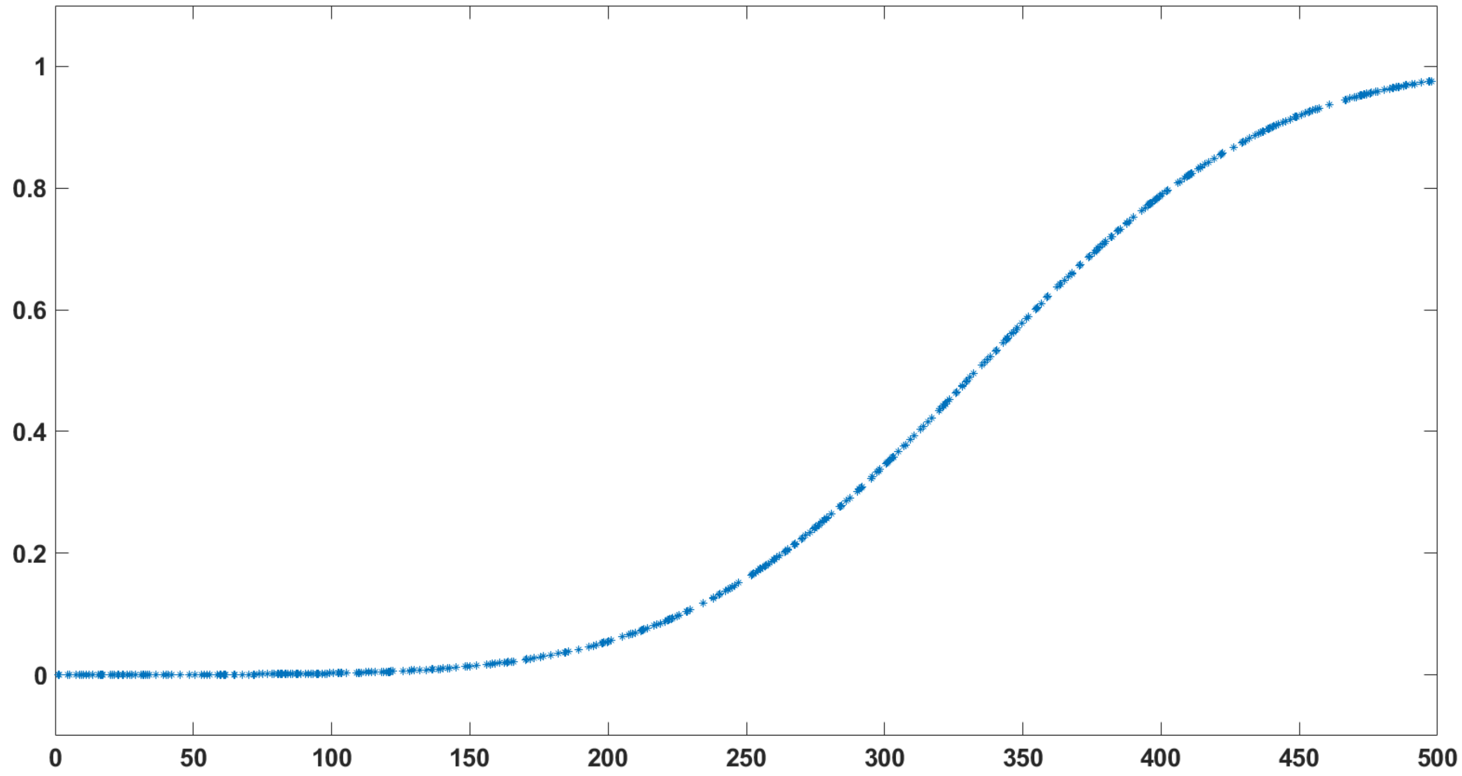
**Useful for quickly checking if a relation exists**

# Linear regression with limited variables

**Dependent variable:** **Paid dividend in 2019**

**Explanatory variable:** **Market capitalisation (in billions $)**

$$P[\mathrm{Div}_t = 1|\mathrm{cap}_t] = f(\beta_1 + \beta_2 \mathrm{cap}_t)$$




**We do not observe the probability**

**How to infer probabilities from binary data ?**

*Limited dependent variables*

**Unconditional probability**

# Model for limited variables

Are you able to distinguish between a low-cost and an expensive cheese cube ?

Dependent variable: **Distinguishing the cheese cube** → $C_t = 1 \text{ if correct}$

Unconditional probability

Average probability in the population of detecting a correct cube cheese

$$P[C_t = 1] = p$$

**How to estimate p ?**

Conditional probability

Probability given some characteristics of detecting a correct cube cheese

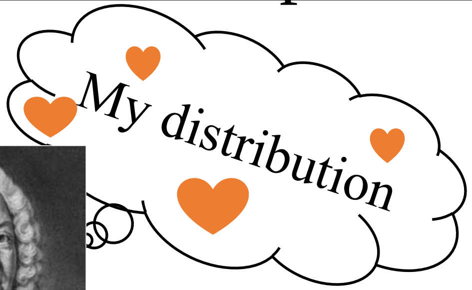
$$P[C_t = 1|x_t] = f(\beta_1 + \beta_2 x_t)$$

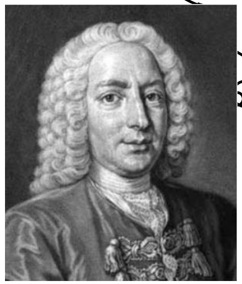
**How to estimate $\beta_1$ and $\beta_2$ ?**

# Model for limited variables

**Dependent variable:** **Distinguishing the cheese cube** → $C_t = 1 \text{ if correct}$

Unconditional probability: $P[C_t = 1] = p \longleftrightarrow P[C_t = 0] = 1 - p$



Bernoulli

**Properties:**

$$E(C_t) = p$$

$$V(C_t) = E(C_t^2) - E(C_t)^2 = p - p^2 = p(1-p)$$

**Application:** Estimation of a proportion

## Survey and polls

When individuals are humans

$C_t = 1$

$C_t = 0$


## Other proportions

When individuals are not humans

# Unconditional probability

$$P[C_t = 1] = p$$
$$N = \sum_{t=1}^{T} c_t = \# \text{ of } 1$$

**Dependent variable:** **Distinguishing the cheese cube** → $C_t = 1$ if correct

$$\text{Let us assume that } p = 0.5$$

$$T = 1$$

$P(C_1 = 0) = 0.5$ ⟷ $P(C_1 = 1) = 0.5$

$P(C_1 = 0, C_2 = 0) = 0.25$ $P(C_1 = 0, C_2 = 1) = 0.25$ $P(C_1 = 1, C_2 = 0) = 0.25$ $P(C_1 = 1, C_2 = 1) = 0.25$

**Likelihood function for two flips:**

$$P(C_1 = 0, C_2 = 1|p) = P(C_1 = 0|p)P(C_2 = 1|p) = (1-p)p = (1-p)^{\# \text{ of } 0} p^{\# \text{ of } 1}$$

**General likelihood function:**

$$P(C_1 = c_1, C_2 = c_2, \ldots, C_T = c_T|p) = \prod_{t=1}^{T} P(C_t = c_t|p) = \prod_{t=1}^{T} p^{c_t}(1-p)^{1-c_t} = p^N (1-p)^{T-N}$$

# Unconditional probability

$$P[C_t = 1] = p$$

**General likelihood function:**

$$P(C_1 = c_1, C_2 = c_2, \ldots, C_T = c_T|p) = \prod_{t=1}^{T} P(C_t = c_t|p) = \prod_{t=1}^{T} p^{c_t}(1-p)^{1-c_t} = p^{N}(1-p)^{T-N}$$

Likelihood of observing a sample: depends on *N* and *T*

Let us assume that $p = 0.5$

$T = 2$

| $N = 0$ | $N = 1$ | $N = 2$ |
|---|---|---|
| $P(C_1 = 0, C_2 = 0) = 0.25$ | $P(C_1 = 1, C_2 = 0) = 0.25$ + $P(C_1 = 0, C_2 = 1) = 0.25$ | $P(C_1 = 1, C_2 = 1) = 0.25$ |
| **Prob = 0.25** | **Prob = 0.5** | **Prob = 0.25** |

Take Away

*The likelihood of the number of successes can be expressed as*

$$f(N|T,p) = \sum_{i=1}^{\#\text{success=N}} p^N(1-p)^{T-N}$$

# Unconditional probability

$$P[C_t = 1] = p$$
$$N = \sum_{t=1}^{T} C_t = \#\text{ of } 1$$

**Additional examples**

$$p = 0.5$$

| | $N = 0$ | $N = 1$ | $N = 2$ | $N = 3$ |
|---|---|---|---|---|
| $T = 1$ | Prob = 0.5 | Prob = 0.5 | | |
| $T = 2$ | Prob = 0.25 | Prob = 0.5 | Prob = 0.25 | |
| $T = 3$ | Prob = 0.125 | Prob = 0.375 | Prob = 0.375 | Prob = 0.125 |

**Binomial distribution**

1’ to 6’

$$f(N|T,p) = \sum_{i=1}^{\#\text{success}=\text{N}} p^N (1-p)^{T-N} = \binom{T}{N} p^N (1-p)^{T-N}$$

Take Away

*The sum of Bernoulli random variables follows a Binomial distribution:*

$$C_t \sim \text{Be}(p) \text{ then } N = \sum_{t=1}^{T} C_t \sim \text{Bino}(p, T)$$

# Cheese tasting

$\boxed{P[C_t = 1] = p}$

**One trial:**

1 

2 

**Distinguishing the cheese cube** → $C_t = 1 \text{ if correct}$

**Unable to tell them apart → Random guess:**

$\boxed{p = 0.5}$

$T = 10$

| $N = 0$ | $N = 1$ | $\bullet \quad \bullet \quad \bullet$ | $N = 8$ | $N = 9$ | $N = 10$ |
|---|---|---|---|---|---|
| Prob = 0 | Prob = 0.01 | | Prob = 0.04 | Prob = 0.01 | Prob = 0 |

$P[N \geq 8] = 0.05$

**Link to statistical tests**

**Significant level of 99%** → **One mistake is allowed!**

**Significant level of 95%** → **Two mistake is allowed**

**Are we able to statistically detect the expensive cheese ?**

# Unconditional probability

$\boxed{P[C_t = 1] = p}$

**Dependent variable:** **Distinguishing the cheese cube** → $C_t = 1$ if correct

**How to estimate p ?**

$$p = 0.5,\ p = 0.25$$

| | $N = 0$ | $N = 1$ | $N = 2$ | $N = 3$ |
|---|---|---|---|---|
| $T = 1$ | Prob = 0.5<br>Prob = 0.75 | Prob = 0.5<br>Prob = 0.25 | | |
| $T = 2$ | Prob = 0.25<br>Prob = 0.562 | Prob = 0.5<br>Prob = 0.375 | Prob = 0.25<br>Prob = 0.0625 | |
| $T = 3$ | Prob = 0.125<br>Prob = 0.421 | Prob = 0.375<br>Prob = 0.421 | Prob = 0.375<br>Prob = 0.14 | Prob = 0.125<br>Prob = 0.015 |

**Binomial distribution:** $\max_p f(N|T,p) = \max_p \binom{T}{N} p^N (1-p)^{T-N}$

**Maximum likelihood estimator (MLE)**

# Unconditional probability

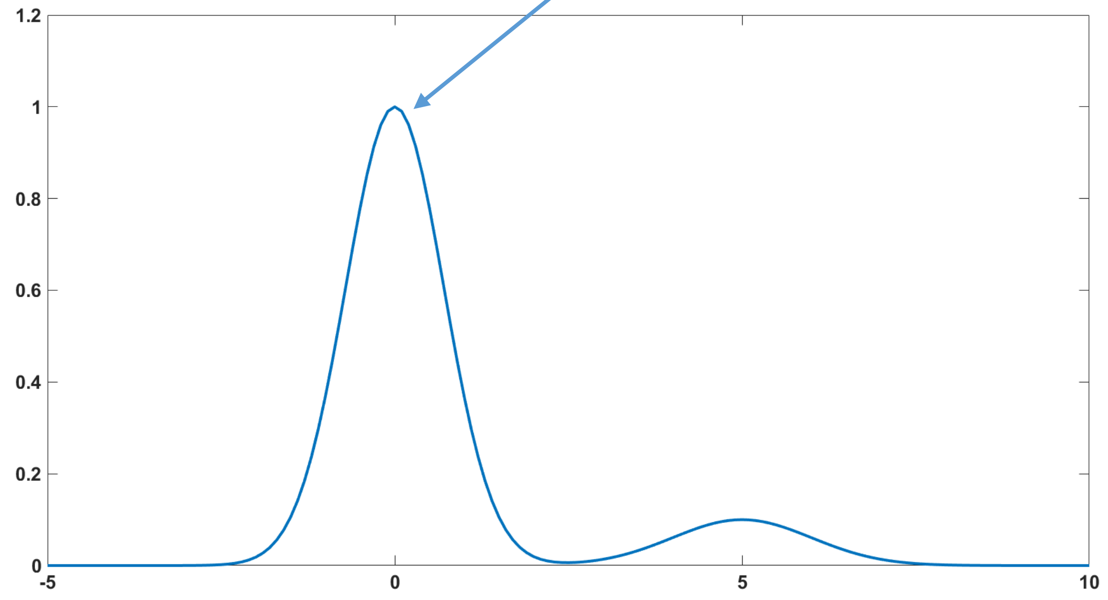
$$P[C_t = 1] = p$$
$$N = \sum_{t=1}^{T} C_t$$

**Dependent variable:** **Distinguishing the cheese cube** → $C_t = 1$ if correct

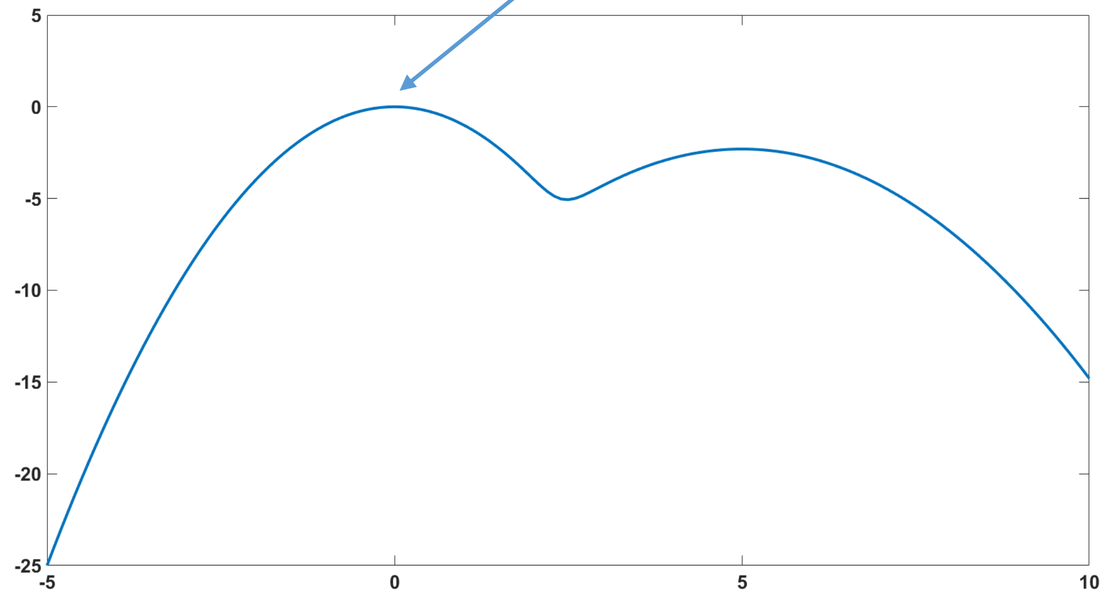
$$\max_p f(N|T,p) = \max_p \binom{T}{N} p^N (1-p)^{T-N}$$ ⇒ **Complicated function to optimize**

If $\hat{p}$ maximizes $f(N|T,p)$ then $\hat{p}$ maximizes $\ln f(N|T,p)$.

**Maximum**




**Maximum**



# Unconditional probability

$$P[C_t = 1] = p$$
$$N = \sum_{t=1}^{T} C_t$$

**Dependent variable:** **Distinguishing the cheese cube** → $C_t = 1$ if correct

$$\max_p \ln f(N|T,p) = \max_p \ln \binom{T}{N} + N\ln(p) + (T-N)\ln(1-p)$$

**Estimator of p:** $\hat{p} = \frac{N}{T}$

**Number of successes over the number of trials**

**Proof:**

$$\begin{aligned}
\frac{d\ln f(N|T,p)}{dp} &= \frac{N}{p} - \frac{(T-N)}{1-p}( \quad = 0), \\
\frac{N}{\hat{p}} &= \frac{(T-N)}{1-\hat{p}}, \\
N - N\hat{p} &= (T-N)\hat{p}, \\
\hat{p} &= \frac{N}{T}
\end{aligned}$$

# Unconditional probability

$$P[C_t = 1] = p$$
$$N = \sum_{t=1}^{T} C_t$$

**<u>Properties of the estimator ?</u>**

**<u>Note that:</u>** $\hat{p} = \frac{N}{T} = \frac{1}{T}\sum_{t=1}^{T} C_t = \bar{C}$

**Average of random variables!**

What kind of random variable is $C_t$ ?


<u>Bernoulli</u>

<u>Properties:</u>

$E(C_t) = p$

$V(C_t) = E(C_t^2) - E(C_t)^2 = p - p^2 = p(1-p)$

**<u>Assumption:</u> No dependence between the trials** → **<u>e.g. No learning curve</u>**

# Unconditional probability

$$E(C_t) = p$$
$$V(C_t) = p(1-p)$$

**Properties of the estimator:** $\hat{p} = \bar{C}$

**Unbiasedness**

$$E(\hat{p}) = \frac{1}{T}\sum_{t=1}^{T} E(C_t)$$

✓ $E(\hat{p}) = p$

**Consistency**

$$V(\hat{p}) = \frac{1}{T^2}\sum_{t=1}^{T} V(C_t)$$

✓ $V(\hat{p}) = \frac{p(1-p)}{T}$

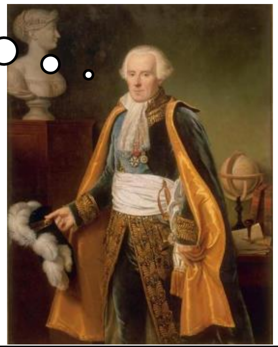
**Also: Implication of the LLN**

**Distribution of the estimator ?**

Central limit theorem: $\hat{p} = \frac{1}{T}\sum_{t=1}^{T} C_t \rightarrow N(p, \frac{p(1-p)}{T})$

Me again!

Laplace



# Statistical test when T large

$$E(C_t) = p$$
$$V(C_t) = p(1-p)$$

<u>Central limit theorem:</u> $\hat{p} = \frac{1}{T}\sum_{t=1}^{T} C_t \rightarrow N(p, \frac{p(1-p)}{T})$

<u>Hypothesis:</u> $H_0 : p = 0.4 \text{ vs } H_1 : p \neq 0.4$

<u>Under the null:</u> $\hat{p} \sim N(0.4, \frac{0.24}{T})$ $\Rightarrow$ $\frac{\hat{p}-0.4}{\sqrt{\frac{0.24}{T}}} \sim N(0,1)$

$$\boxed{\text{For large T: } Z_p \equiv \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{T}}} \sim N(0,1)}$$

**Any estimate of the statistic is a realization of a N(0,1)**

Take Away

**For large T, no need to use the binomial cumulative density function for performing statistical tests.**

# Empirical exercise

**Iphone application:** What is the proportion of Game applications ?

**What do we learn in this exercise ?**

- How to estimate the unconditional probability from a limited variable.
- How to maximize a likelihood function.
- How to test a probability.

*Limited dependent variables*

**Conditional probability**

# Conditional probability

**Dependent variable:** **Distinguishing the cheese cube** → $C_t = 1 \text{ if correct}$

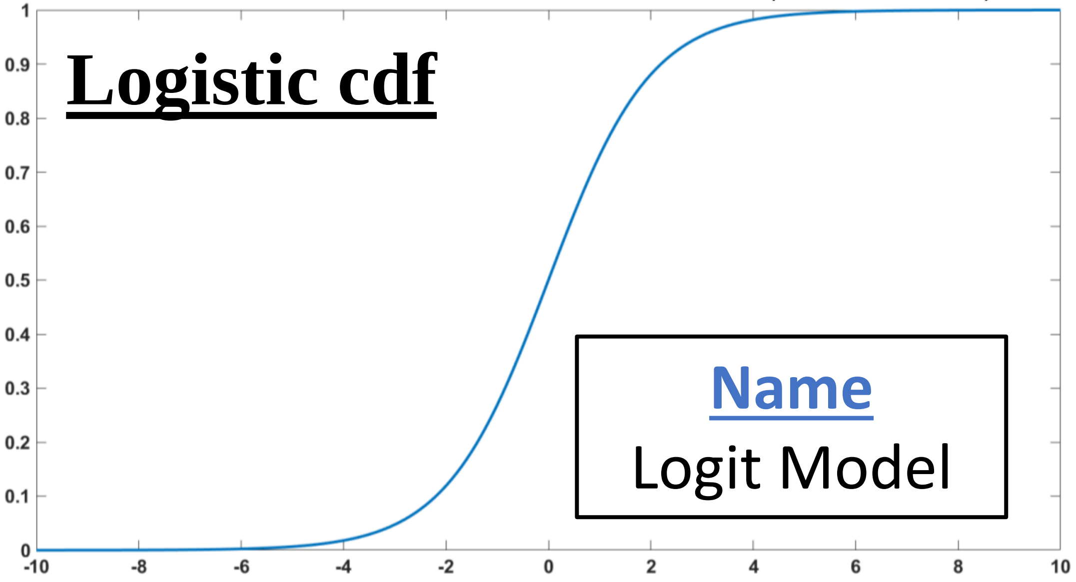
Conditional probability: $P[C_t = 1 | x_t] = f(\beta_1 + \beta_2 x_t)$

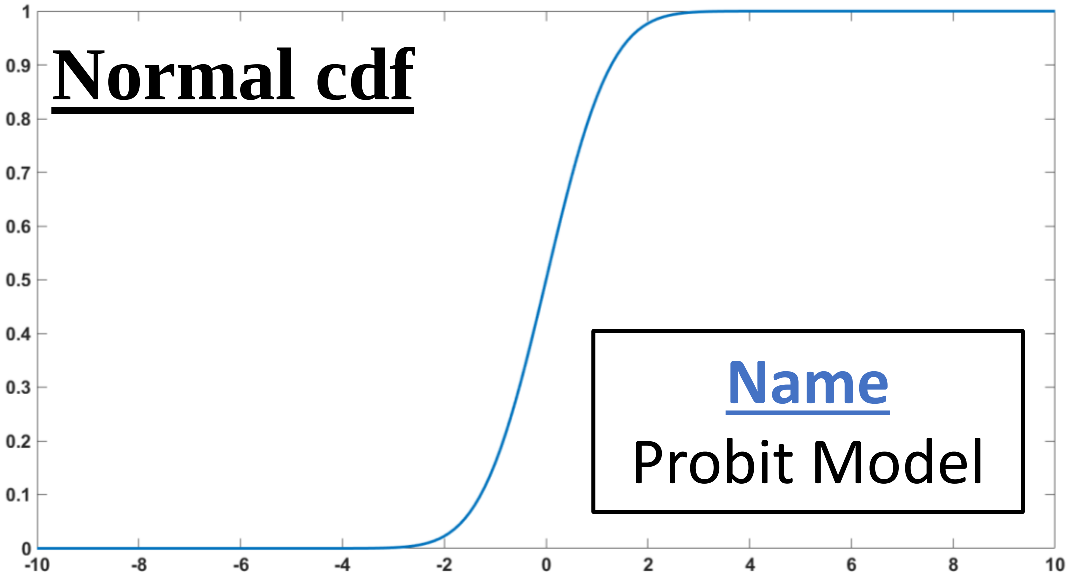
**Which function to choose ?**

**Any cumulative density function on the real support**

$$f(\beta_1 + \beta_2 x_t) = \frac{\exp(\beta_1+\beta_2 x_t)}{1+\exp(\beta_1+\beta_2 x_t)}$$

$$f(\beta_1 + \beta_2 x_t) = \Phi(\beta_1 + \beta_2 x_t) = \int_{-\infty}^{\beta_1+\beta_2 x_t} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) dz$$



# Conditional probability

**Dependent variable:** **Distinguishing the cheese cube** → $C_t = 1$ if correct

**How to estimate a conditional probability ?**

$$P[C_t = 1|x_t] = f(\beta_1 + \beta_2 x_t) = f_\beta(x_t)$$

$$T = 1$$

$$P(C_1 = 0|x_1) = 1 - f_\beta(x_1) \longleftrightarrow P(C_1 = 1|x_1) = f_\beta(x_1)$$

$$T = 2$$

$$P(C_1 = 0, C_2 = 0|x_1, x_2) = (1 - f_\beta(x_1))(1 - f_\beta(x_2)) \qquad P(C_1 = 1, C_2 = 0|x_1, x_2) = f_\beta(x_1)(1 - f_\beta(x_2))$$

$$P(C_1 = 0, C_2 = 1|x_1, x_2) = (1 - f_\beta(x_1))f_\beta(x_2) \qquad P(C_1 = 1, C_2 = 1|x_1, x_2) = f_\beta(x_1)f_\beta(x_2)$$

**General likelihood function:**

$$P(C_1 = c_1, C_2 = c_2, \ldots, C_T = c_T|x_1, \ldots, x_T) = \prod_{t=1}^{T} P(C_t = c_t|x_t) = \prod_{t=1}^{T} f_\beta(x_t)^{c_t}(1 - f_\beta(x_t))^{1-c_t}$$

Cannot be simplified with the number of successes *N* and of trials *T*

$$\prod_{t=1}^{T} f_\beta(x_t)^{c_t}(1 - f_\beta(x_t))^{1-c_t} \neq p^N(1-p)^{T-N}$$

# Conditional probability

$$\boxed{P[C_t = 1 | x_t] = f_\beta(x_t)}$$

**Maximum likelihood estimator:**

$$\max_{\beta_1,\beta_2} P(C_1 = c_1, C_2 = c_2, \ldots, C_T = c_T | x_1, \ldots, x_T) \ = \max_{\beta_1,\beta_2} \prod_{t=1}^{T} f_\beta(x_t)^{c_t} (1 - f_\beta(x_t))^{1-c_t}$$

**Complicated function to optimize**

$$\max_{\beta_1,\beta_2} \sum_{t=1}^{T} c_t \ln[f_\beta(x_t)] + (1 - c_t) \ln(1 - f_\beta(x_t))$$

**Logistic function:** $f_\beta(x_t) = \frac{\exp(\beta_1 + \beta_2 x_t)}{1 + \exp(\beta_1 + \beta_2 x_t)}$

**Normal cdf:** $f(\beta_1 + \beta_2 x_t) = \Phi(\beta_1 + \beta_2 x_t) = \int_{-\infty}^{\beta_1 + \beta_2 x_t} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) dz$

**Likelihood function cannot be maximized analytically**

# Conditional probability

6’22’’

$$P[C_t = 1|x_t] = \Phi(\beta_1 + \beta_2 x_t)$$

**Normal cdf:** $f(\beta_1 + \beta_2 x_t) = \Phi(\beta_1 + \beta_2 x_t) = \int_{-\infty}^{\beta_1+\beta_2 x_t} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})dz$

**Maximum likelihood estimator with Normal cdf:**

$$\max_{\beta_1,\beta_2} \sum_{t=1}^{T} c_t \ln[\Phi(\beta_1 + \beta_2 x_t)] + (1 - c_t) \ln(1 - \Phi(\beta_1 + \beta_2 x_t))$$

**We maximize the function using a statistical software:**

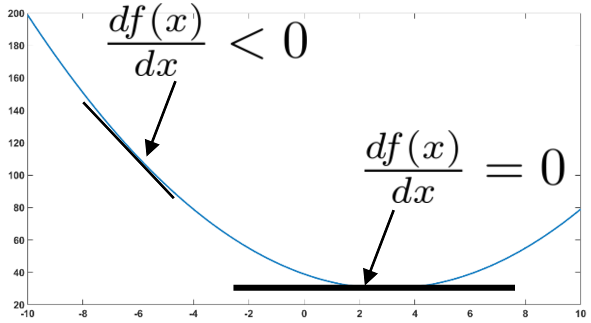
**Numerical algorithms:**

1. Steepest gradient (Based on the first derivative)
2. Newton-Raphson method (Based on the first two derivatives)
3. Nelder-Mead algorithm (derivative-free)
4. Heuristic algorithms (Particle swarm, Differential Evolution, …)

Focus on the Newton-Raphson method

# Numerical optimization

**e.g. function:** $f(x) = x^2 + 4x + 2$




**Newton-Raphson:** $\hat{x} = \text{argmin}_x f(x)$

**Idea:** **At a minimum, the derivative must be zero:** $\frac{df(x)}{dx}|_{x=\hat{x}} = 0$

**Algorithm:**

1. Start from an initial point: $x_0$
2. Find a promising new point: $x_1 = x_0 + t$

**If the promising point is a minimum, its derivative is zero:** $\frac{df(x)}{dx}|_{x=x_1} = 0$

**The derivative at the promising point is approximated by a Taylor expansion:**

$$\underbrace{\frac{df(x)}{dx}|_{x=x_1}}_{=0} \approx \frac{df(x)}{dx}|_{x=x_0} + \underbrace{(x_1 - x_0)}_{=t}\frac{d^2f(x)}{dx^2}|_{x=x_0} \Rightarrow t = -\left(\frac{d^2f(x)}{dx^2}|_{x=x_0}\right)^{-1}\frac{df(x)}{dx}|_{x=x_0}$$

3. Check if the promising point is a minimum, if not come back to 1 with $x_0 = x_1$

# Numerical optimization

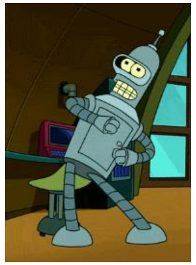
**Illustration with a function:** $f(x) = x^2 + 4x + 2$ $\Rightarrow$ $\frac{df(x)}{dx} = 2x + 4$ and $\frac{d^2f(x)}{dx^2} = 2$

**<u>Newton-Raphson algorithm:</u>**

1. Start from an initial point: $x_0 = 5$
2. Find a promising new point: $x_1 = x_0 + t$ with $t = -\left(\frac{d^2f(x)}{dx^2}\Big|_{x=x_0}\right)^{-1} \frac{df(x)}{dx}\Big|_{x=x_0}$

$\Rightarrow$ $t = -\frac{1}{2}(2x_0 + 4) = -7$ $\Rightarrow$ $x_1 = -2$

3. Check if the promising point is a minimum: $\frac{df(x)}{dx}\Big|_{x=x_1} = 2(-2) + 4 = 0$



**<u>For more complex function, it requires several iterations</u>**

**<u>Why does it work in one iteration for quadratic functions ?</u>**

The derivative is linear so the Taylor expansion is exact.

# Conditional probability

$$\boxed{P[C_t = 1|x_t] = \Phi(\beta_1 + \beta_2 x_t)}$$

## Maximum likelihood estimator with Normal cdf:

$$\max_{\beta_1,\beta_2} \sum_{t=1}^{T} c_t \ln[\Phi(\beta_1 + \beta_2 x_t)] + (1 - c_t) \ln(1 - \Phi(\beta_1 + \beta_2 x_t))$$

**Using a Numerical algorithm, we get our estimates:** $\hat{\beta}_1, \hat{\beta}_2$ ✓

## Two issues:

1. How to interpret these estimates ?

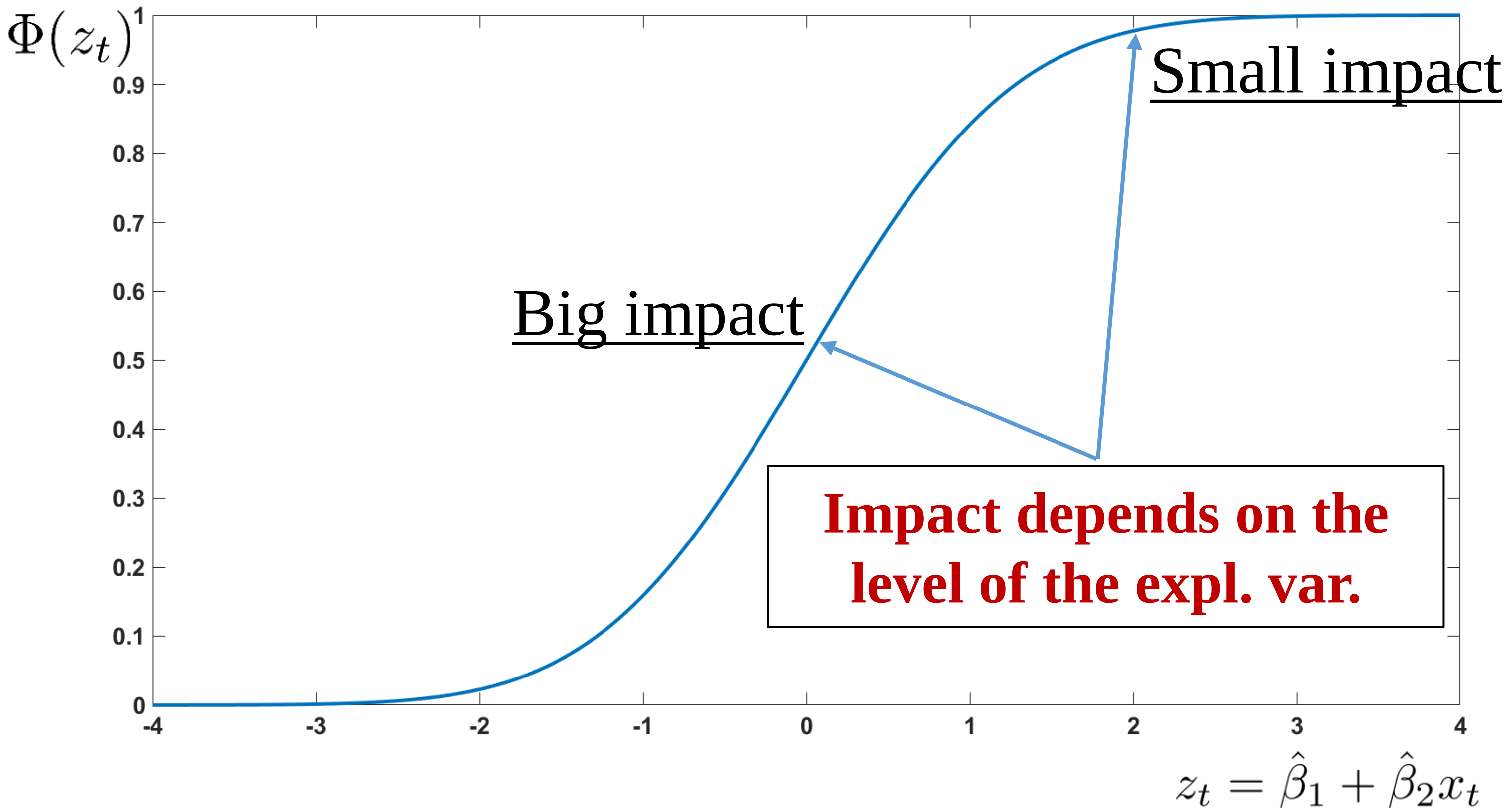
2. What are the statistical properties of our estimator ?

# Interpretation

We have our estimates: $\hat{\beta}_1, \hat{\beta}_2$

**How do I interpret them ?**

Estimates are related to the probability: $P[C_t = 1|x_t] = \Phi(\hat{\beta}_1 + \hat{\beta}_2 x_t)$




**Signs**

$\hat{\beta}_2 > 0$ : prob increases with $x_t$

$\hat{\beta}_2 < 0$ : prob decreases with $x_t$

**Exact interpretation:**

$$\frac{dP[C_t=1|x_t]}{dx_t} = \underbrace{\frac{d\Phi(\hat{\beta}_1+\hat{\beta}_2 x_t)}{d(\hat{\beta}_1+\hat{\beta}_2 x_t)}}_{\text{Density function of N(0,1)}} \hat{\beta}_2$$

# Statistical test

We have our estimates: $\hat{\beta}_1, \hat{\beta}_2$

**Without analytical formulas, how can we study the statistical properties ?**

**Maximum likelihood theory provides a solution**

*Limited dependent variables*

# Introduction to the maximum likelihood theory

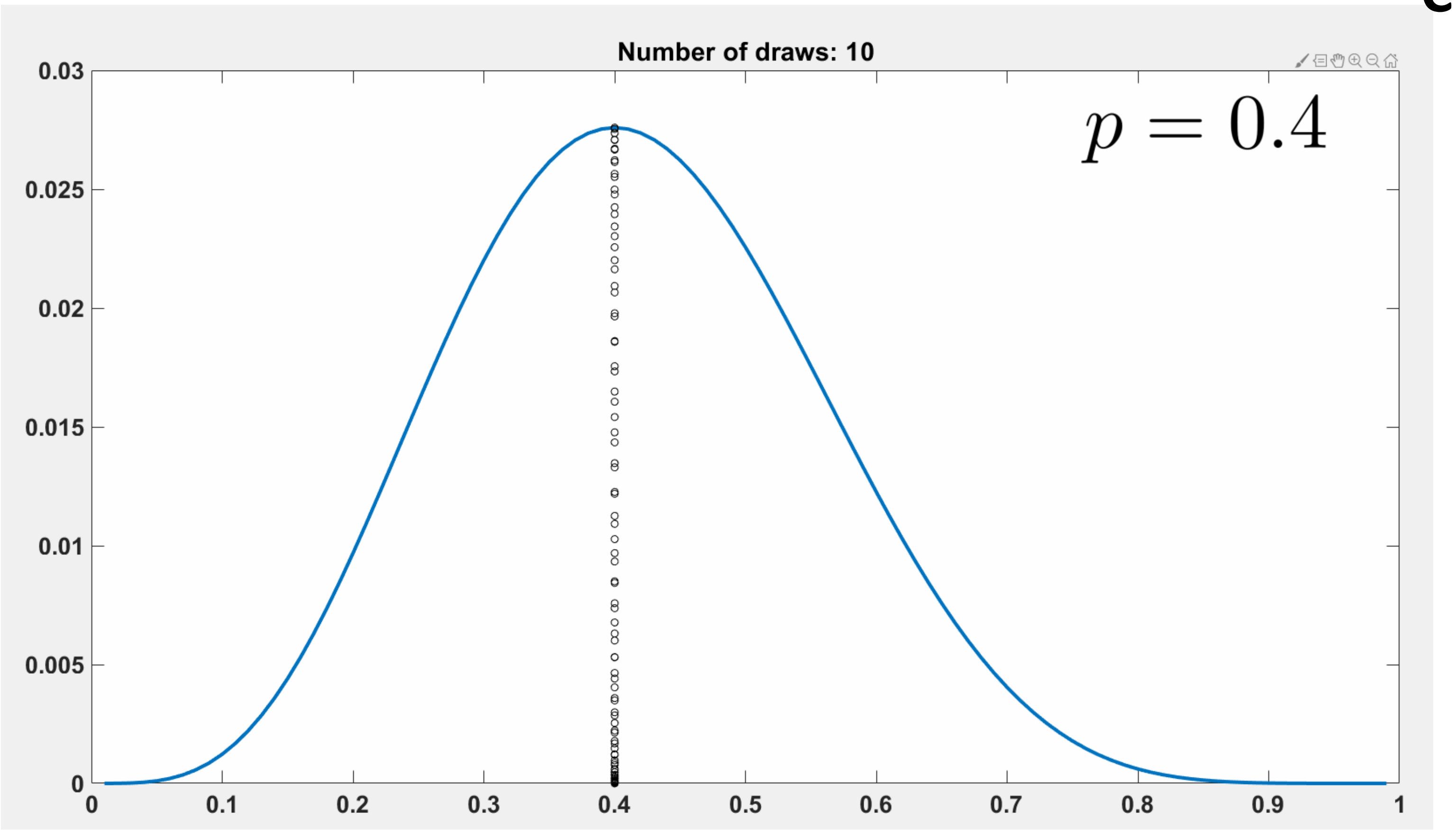
# (Relevant for empirical works)

# Maximum likelihood estimator

$$E(C_t) = p$$
$$V(C_t) = p(1-p)$$

**Maximum likelihood estimator (MLE):**

$$\hat{p} = \text{argmax}_p P(C_1, C_2, \ldots, C_T) = \text{argmax}_p \prod_{t=1}^{T} p^{C_t}(1-p)^{1-C_t}$$




**Intuition:**

**Curvature gives a sense of the variance**

$$\hat{p} \sim N(p, \sigma_p^2)$$

$$\ln f(\hat{p}|p, \sigma_p^2) = -\frac{1}{2}\ln(2\pi\sigma_p^2) - \frac{(\hat{p}-p)^2}{2\sigma_p^2}$$

$$\frac{d^2 \ln f(\hat{p}|p, \sigma_p^2)}{dp^2} = -\frac{1}{\sigma_p^2}$$

$$\boxed{\sigma_p^2 = \left(-\frac{d^2 \ln f(\hat{p}|p, \sigma_p^2)}{dp^2}\right)^{-1}}$$

# In a Nutshell: Maximum likelihood theory

**Given the log-likelihood function:**

$$\ln f(c_{1:T}|\beta = \{\beta_1, \beta_2\}) \quad = \quad \sum_{t=1}^{T} c_t \ln[\Phi(\beta_1 + \beta_2 x_t)] + (1 - c_t)\ln(1 - \Phi(\beta_1 + \beta_2 x_t))$$

1. Find the maximum likelihood estimate: $\frac{d \ln f(c_{1:T}|\beta)}{d\beta} = 0 \Rightarrow \hat{\beta} = \{\hat{\beta}_1, \hat{\beta}_2\}$

2. Compute the second derivative: $H(\beta) = \frac{d^2 \ln f(c_{1:T}|\beta)}{d\beta^2}$

3. Maximum likelihood estimator:

$$\hat{\beta}_1 \sim N(\beta_1, ((-H(\hat{p}))^{-1})_{11}) \text{ for large T.}$$

# In a Nutshell: Maximum likelihood theory

**Example: Unconditional probability**

**Slide 11: Log-likelihood function is given by**

$$\ln P(C|p) = \sum_{t=1}^{T} c_t \ln[p] + (1 - c_t) \ln(1 - p)$$

**1. Find the maximum likelihood estimate:** $\frac{d \ln P(C|p)}{dp} = 0$

$$\frac{d \ln P(C|p)}{dp} = \frac{\sum_{t=1}^{T} c_t}{p} - \frac{\sum_{t=1}^{T}(1-c_t)}{1-p} \Rightarrow \frac{\hat{p}}{1-\hat{p}} = \frac{N}{T-N} \Rightarrow \hat{p} = \frac{\sum_{t=1}^{T} c_t}{T} = \frac{N}{T}$$

$$1 - \hat{p}$$

**2. Compute the second derivative:** $H(p) = \frac{d^2 \ln P(C|p)}{dp^2}$

$$\frac{d^2 \ln P(C|p)}{dp^2} = -\frac{N}{p^2} - \frac{T-N}{(1-p)^2} \Rightarrow H(\hat{p}) = -\frac{T}{\hat{p}^2}\left[\frac{N}{T} + \frac{T-N}{T}\frac{\hat{p}^2}{(1-\hat{p})^2}\right] \Rightarrow H(\hat{p}) = -\frac{T}{\hat{p}(1-\hat{p})}$$

**3. Maximum likelihood estimator:**

$$\hat{p} \sim N(p, (-H(\hat{p}))^{-1}) \text{ for large T} \Rightarrow \hat{p} \sim N(p, \frac{\hat{p}(1-\hat{p})}{T})$$

# Empirical exercise

**Iphone application:** Does the app score help for becoming a killer app ?

## What do we learn in this exercise ?

- How to estimate the conditional probability from a limited variable.
- How to maximize a likelihood function.
- How to test the value of a parameter.

*Limited dependent variables*

# Introduction to the maximum likelihood theory (mathematical part)

# Maximum likelihood estimator

$$E(C_t) = p$$
$$V(C_t) = p(1-p)$$

Central limit theorem: $\hat{p} = \frac{1}{T}\sum_{t=1}^{T} C_t \to N(p, \frac{p(1-p)}{T})$

Maximum likelihood theory:

**Likelihood function:** $P(C|p) = \prod_{t=1}^{T} p^{C_t}(1-p)^{1-C_t}$

**Log-likelihood function:** $\ln P(C|p) = \sum_{t=1}^{T} C_t \ln p + (1-C_t)\ln(1-p)$

**First derivative:** $\frac{d\ln P(C|p)}{dp} = \sum_{t=1}^{T} \frac{C_t}{p} - \frac{(1-C_t)}{1-p}$ $\Rightarrow$ $E[\frac{C_t}{p} - \frac{(1-C_t)}{1-p}] = \frac{E(C_t)}{p} - \frac{E(1-C_t)}{1-p} = 0$

**Asymptotically, the true probability should maximize the likelihood function!**

$$\frac{d\ln P(C|p)}{dp}\Big|_{p=p} = \sum_{t=1}^{T} \frac{C_t}{p} - \frac{(1-C_t)}{1-p} \approx 0 \text{ for large T.}$$
$$\frac{d\ln P(C|p)}{dp}\Big|_{p=p} \sim N\left(0, T\ V\left(\frac{C_t}{p} - \frac{(1-C_t)}{1-p}\right)\right) \text{ for large T.}$$

# Maximum likelihood estimator

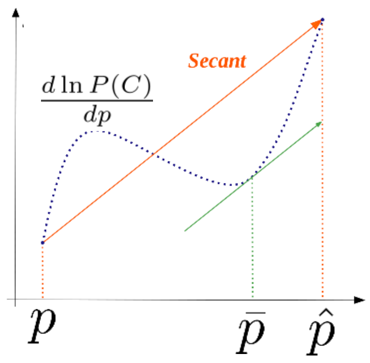
$$V\left(\frac{C_t}{p} - \frac{(1-C_t)}{1-p}\right) = E\left(\left(\frac{C_t}{p} - \frac{(1-C_t)}{1-p}\right)^2\right) = \frac{1}{p^2}p + \frac{1}{(1-p)^2}(1-p) = \frac{1}{p(1-p)}.$$

$$\boxed{\frac{d\ln P(C|p)}{dp}\Big|_{p=p} \sim N\left(0, T\ V\left(\frac{C_t}{p} - \frac{(1-C_t)}{1-p}\right)\right) \text{ for large T.}}$$

**Maximum likelihood estimator:** $\frac{d\ln P(C|p)}{dp}\Big|_{p=\hat{p}} = 0 \Rightarrow \hat{p} = \frac{1}{T}\sum_{t=1}^{T} C_t$

## Maximum likelihood theory

Mean Value Theorem: $$\underbrace{\frac{d\ln P(C|p)}{dp}\Big|_{p=\hat{p}}}_{=0} = \frac{d\ln P(C|p)}{dp}\Big|_{p=p} + (\hat{p}-p)\frac{d^2\ln P(C|p)}{dp^2}\Big|_{p=\bar{p}}$$




$$\Downarrow$$

$$(\hat{p}-p) = \left(-\frac{d^2\ln P(C|p)}{dp^2}\Big|_{p=\bar{p}}\right)^{-1}\frac{d\ln P(C|p)}{dp}\Big|_{p=p}$$

$$\Downarrow$$

$$(\hat{p}-p) \sim N\left(0, \left(\frac{-d^2\ln P(C|p)}{dp^2}\Big|_{p=p}\right)^{-1} T\ V\left(\frac{C_t}{p} - \frac{(1-C_t)}{1-p}\right)\left(\frac{-d^2\ln P(C|p)}{dp^2}\Big|_{p=p}\right)^{-1}\right) \text{ for large T.}$$

# Maximum likelihood estimator

$$V(\frac{C_t}{p} - \frac{(1-C_t)}{1-p}) = \frac{1}{p(1-p)}.$$

$$(p-\hat{p}) \sim N(0, (\frac{-d^2 \ln P(C)}{dp^2}|_{p=p})^{-1} \frac{T}{p(1-p)} (\frac{-d^2 \ln P(C)}{dp^2}|_{p=p})^{-1}) \text{ for large T.}$$

**Variance of the estimator:**

$$\begin{aligned} \frac{d^2 \ln P(C)}{dp^2} &= \frac{d}{dp}[\sum_{t=1}^{T} \frac{C_t}{p} - \frac{(1-C_t)}{1-p}], \\ &= T\frac{1}{T}\sum_{t=1}^{T} \frac{C_t - 1}{(1-p)^2} - \frac{C_t}{p^2}, \\ \frac{d^2 \ln P(C)}{dp^2}|_{p=p} &= TE(\frac{C_t - 1}{(1-p)^2} - \frac{C_t}{p^2}), \text{ for large T,} \\ &= -\frac{T}{p(1-p)} \text{ for large T.} \end{aligned}$$

$$(\hat{p}-p) \sim N(0, \frac{p(1-p)}{T}) \text{ for large T.}$$

# Maximum likelihood theory

**Maximum likelihood theory:**

**Log-likelihood function:** $\ln f(y_1, \ldots, y_T|\theta) = \sum_{t=1}^{T} \ln f(y_t|\theta)$

**Maximum likelihood estimator:** $\hat{\theta} = \text{argmax}_\theta \sum_{t=1}^{T} \ln f(y_t|\theta)$ ⇒ $\sum_{t=1}^{T} \frac{d \ln f(y_t|\theta)}{d\theta} = 0$

**Mean value Theorem:**

$$\underbrace{\frac{d \ln f(y_1, \ldots, y_T)}{d\theta}|_{\theta=\hat{\theta}}}_{=0} = \underbrace{\frac{d \ln f(y_1, \ldots, y_T)}{d\theta}|_{\theta=\theta_0}}_{\to_d N\left(0, TV\left(\frac{d \ln f(y_t|\theta)}{d\theta}\right)\right)} + (\hat{\theta} - \theta_0) \underbrace{\frac{d^2 \ln f(y_1, \ldots, y_T)}{d\theta^2}|_{\theta=\bar{\theta}}}_{\to_p TV\left(\frac{d \ln f(y_t|\theta)}{d\theta}\right)}$$

**Properties of the MLE:** $\hat{\theta} \sim N\left(\theta_0, \left(\frac{-d^2 \ln f(y_1, \ldots, y_T)}{d\theta}|_{\theta=\theta_0}\right)^{-1}\right)$ for large T.

CAN: **C**onsistent and **A**symptotically **N**ormally distributed

# Maximum likelihood theory: Example

**Linear regression with a constant:** $y_t = \beta_1 + \epsilon_t$

**OLS criterion:** $\hat{\beta}_1 = \text{argmin}_{\beta_1} \sum_{t=1}^{T} \epsilon_t^2$

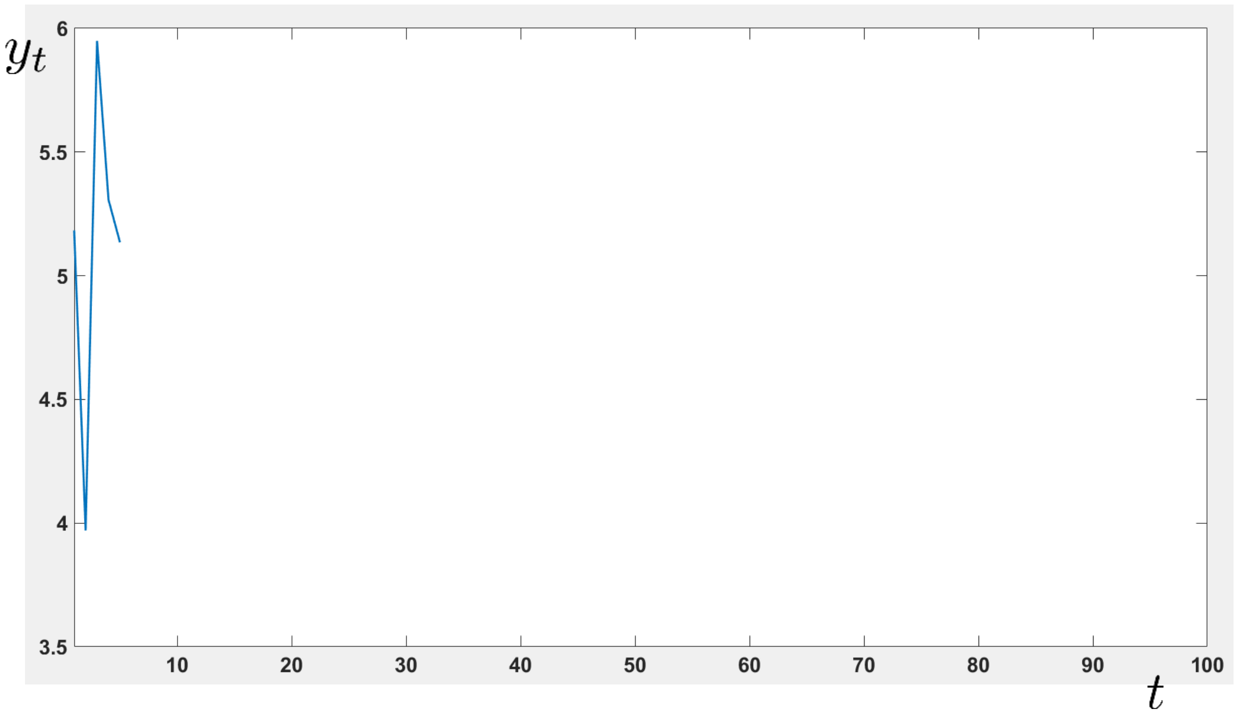
**OLS estimator:** $\hat{\beta}_1 = \frac{1}{T} \sum_{t=1}^{T} y_t$

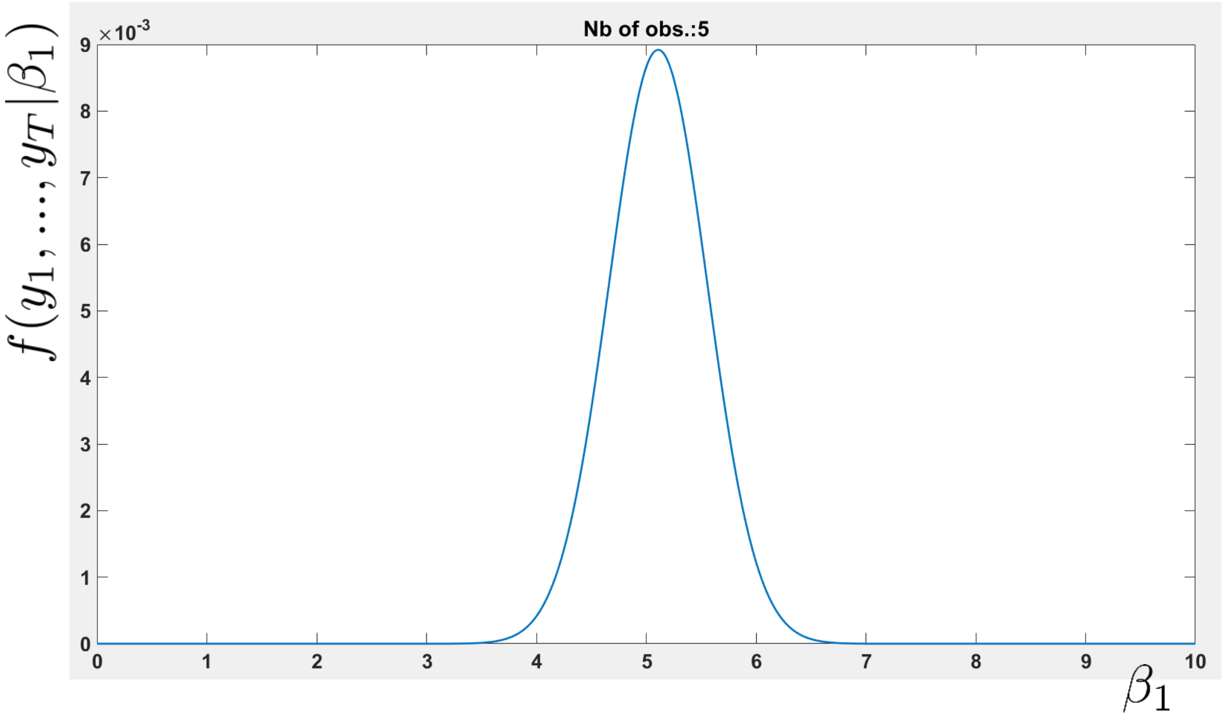
**Large sample prop:** $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{T})$

**ML: Additional assumption:** $\epsilon_t \sim N(0, \sigma^2)$ (independent)

**Model fully characterized:** $y_t | \beta_1 \sim N(\beta_1, \sigma^2)$

**ML criterion - Likelihood function:** $f(y_1, \ldots, y_T | \beta_1) = \prod_{t=1}^{T} f(y_t | \beta_1) = \prod_{t=1}^{T} \frac{\exp\left(-\frac{(y_t - \beta_1)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$



# Maximum likelihood theory: Example

**OLS criterion:** $\hat{\beta}_1 = \text{argmin}_{\beta_1} \sum_{t=1}^{T} \epsilon_t^2$

**OLS estimator:** $\hat{\beta}_1 = \frac{1}{T} \sum_{t=1}^{T} y_t$

**Large sample prop:** $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{T})$

**Linear regression with a constant:** $y_t = \beta_1 + \epsilon_t$

**ML criterion - Likelihood function:**

$$f(y_1, ..., y_T|\beta_1) = \prod_{t=1}^{T} f(y_t|\beta_1) = \prod_{t=1}^{T} \frac{\exp\left(-\frac{(y_t-\beta_1)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$$

**ML estimator :** $\hat{\beta}_1 = \text{argmax}_{\beta_1} \ln f(y_1, ..., y_T|\beta_1) = \text{argmax}_{\beta_1} \sum_{t=1}^{T} \left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_t-\beta_1)^2}{2\sigma^2}\right)$

$\Rightarrow$ $\hat{\beta}_1 = \text{argmax}_{\beta_1} - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \beta_1)^2$ $\Rightarrow$ $\hat{\beta}_1 = \frac{1}{T} \sum_{t=1}^{T} y_t$

Same estimator as the OLS estimator

**Large sample prop:** $\hat{\beta}_1 \sim N\left(\beta_1, \left(\frac{-d^2 \ln f(y_1,...,y_T|\beta_1)}{d\beta_1^2}\right)^{-1}\right) = N(\beta_1, \frac{\sigma^2}{T})$

$\Rightarrow$

$$\frac{d^2 \ln f(y_1, ..., y_T|\beta_1)}{d\beta_1^2} = \frac{d^2}{d\beta_1^2}\left[\sum_{t=1}^{T} \left( -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_t - \beta_1)^2}{2\sigma^2}\right)\right],$$

$$= -\sum_{t=1}^{T} \frac{1}{\sigma^2} \quad \left(= -\frac{T}{\sigma^2}\right)$$

# Summary

## 1. Fully Characterized your model:

**Linear regression:**

$$y_t | x_t \sim N(\beta_1 + \beta_2 x_t, \sigma^2)$$

**Probit model:**

$$P[C_t = 1 | x_t] = \Phi(\beta_1 + \beta_2 x_t)$$

## 2. Write down the likelihood function:

$$\prod_{t=1}^{T} \frac{\exp\left(-\frac{(y_t - \beta_1 - \beta_2 x_t)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$$

$$\prod_{t=1}^{T} \Phi(\beta_1 + \beta_2 x_t)^{c_t} (1 - \Phi(\beta_1 + \beta_2 x_t))^{1-c_t}$$

## 3. Compute the log-likelihood function:

$$\sum_{t=1}^{T} \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_t - \beta_1 - \beta_2 x_t)^2}{2\sigma^2} \right))$$

$$\sum_{t=1}^{T} c_t \ln[\Phi(\beta_1 + \beta_2 x_t)] + (1 - c_t) \ln(1 - \Phi(\beta_1 + \beta_2 x_t))$$

## 4. Maximize the log-likelihood function with respect to the parameters

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x},\ \hat{\beta}_2 = \frac{\sum_{t=1}^{T} (x_t - \bar{x})(y - \bar{y})}{\sum_{t=1}^{T} (x_t - \bar{x})^2},\ \hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{\epsilon}_t^2}{T}$$

No analytical formula

## 5. Compute the Hessian at the maximum likelihood estimate:

$$H(\hat{\theta}) = \frac{d^2 \ln f(y_1, \ldots, y_T)}{d\theta d\theta}$$

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\sigma}^2 \end{pmatrix} \sim N\left( \begin{pmatrix} \beta_1 \\ \beta_2 \\ \sigma^2 \end{pmatrix}, (-H(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2))^{-1} \right)$$

**Large T**

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N\left( \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, (-H(\hat{\beta}_1, \hat{\beta}_2))^{-1} \right)$$

## *Limited dependent variables*

# Model comparison

# Likelihood ratio test

$$m = \# \text{ of restrictions} \\ = K_{\mathrm{U}} - K_{\mathrm{R}}$$

## How to compare nested models ?

**Likelihood ratio test (equivalent of the Fisher test in a ML context).**

**Full model (or unrestricted model):**

$$P[C_t = 1 | x_t] = \Phi(\beta_1 + \beta_2 x_t)$$

$$\ln f(c_1, ..., c_T | \hat{\beta}, x_{1:T}) =$$
$$\sum_{t=1}^{T} c_t \log[\Phi(\hat{\beta}_1 + \hat{\beta}_2 x_t)] + (1 - c_t) \log(1 - \Phi(\hat{\beta}_1 + \hat{\beta}_2 x_t))$$

**Nested model (or restricted model):**

$$H_0 : \text{Restrictions on the model hold}$$

$$P[C_t = 1 | x_t] = \Phi(\beta_1) \Rightarrow H_0 : \beta_2 = 0$$

$$\ln f(c_1, ..., c_T | \tilde{\beta}_1) =$$
$$\sum_{t=1}^{T} c_t \log[\Phi(\tilde{\beta}_1)] + (1 - c_t) \log(1 - \Phi(\tilde{\beta}_1))$$

**Likelihood ratio test**

$$2[\ln f(c_1, ..., c_T | \hat{\beta}, x_{1:T}) - \ln f(c_1, ..., c_T | \tilde{\beta}_1)] \sim \chi^2(m) \text{ under } H_0$$

# Information Criteria

## In general, how to compare models with the same dependent variable ?

**Information criteria have been proposed for general comparisons.**

Advantage: Simple and intuitive

Drawback: Multiple criteria exist and can lead to select different models.

**Example:**

- Multiple models to compare: Model 1 to Model 10.
- Compute the information criterion for each model.
- The best model exhibits the highest value of the information criterion.

**Information criteria (IC):** $\text{Likelihood function evaluated at the MLE - } \underbrace{\text{penalty}}_{\text{Differs from one IC to another}}$

**Two popular information criteria:** AIC and BIC

# Information Criteria

- Akaïke Information criterion (**AIC**) :

$$AIC = \ln f(c_1, ..., c_T | \hat{\beta}, x_{1:T}) - \underbrace{K}_{\text{penalty}_{AIC}} \quad \text{where } K \text{ denotes \# parameters}$$

- Bayesian Information criterion (**BIC**) :

$$BIC = \ln f(c_1, ..., c_T | \hat{\beta}, x_{1:T}) - \underbrace{\frac{K}{2} \log T}_{\text{penalty}_{BIC}}$$

**AIC penalty is smaller than the BIC penalty**

**BIC leads to select smaller models than AIC**

# AIC and BIC justifications

- Akaïke Information criterion (**AIC**) : $AIC = \ln f(c_1, ..., c_T | \hat{\beta}, x_{1:T}) - K$

**Asymptotic justification:**

- Selected model minimizes the Kullback-Leibler divergence (with respect to the true model)

No claim that the selected model is the good one.

The selected model should be good at predicting the data

- Bayesian Information criterion (**BIC**) : $BIC = \ln f(c_1, ..., c_T | \hat{\beta}, x_{1:T}) - \frac{K}{2} \log T$

**Asymptotic justification:**

- BIC is proportional to the marginal likelihood of the Bayesian model with uninformative priors.

Selected model is the true model!

The true model must be among the models in competition.

$\boxed{\text{BIC is favoured in financial research}}$