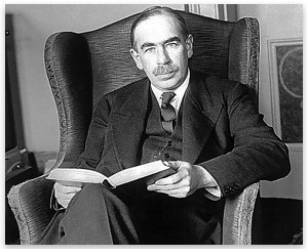
Statistics for Data Science

Introduction: Warm up & History

September 2, 2022

Daniel S. Hain
Associate Professor, PhD.
Contact: dsh@business.aau.dk

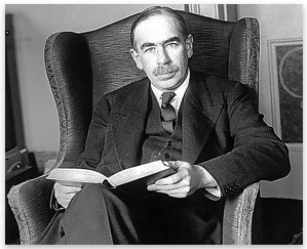
AALBORG UNIVERSITY
DENMARK



Keynes once said... (theoretically speaking)

"Consumption increases as income increases, but not as much as the increase in income"

$\rightarrow$ Meaning: The marginal propensity to consume (MPC) for an unit change in Income is greater than 0 but less than unit



Keynes once said... (theoretically speaking)

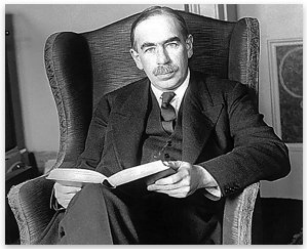
"Consumption increases as income increases, but not as much as the increase in income"

$\rightarrow$ Meaning: The marginal propensity to consume (MPC) for an unit change in Income is greater than 0 but less than unit

One could say... (Mathematically speaking)

$$Y = \alpha + \beta X \textit{ where} : 0 < \beta < 1 \rightarrow 0 < \frac{\partial Y}{\partial I} < 1 \tag{1}$$

Y = Consumption, X = Income, α = A constant, β = The marginal propensity to consume



Keynes once said... (theoretically speaking)

"Consumption increases as income increases, but not as much as the increase in income"

$\rightarrow$ Meaning: The marginal propensity to consume (MPC) for an unit change in Income is greater than 0 but less than unit

One could say... (Mathematically speaking)

$$Y = \alpha + \beta X \textit{ where} : 0 < \beta < 1 \rightarrow 0 < \frac{\partial Y}{\partial I} < 1 \quad (1)$$

Y = Consumption, X = Income, α = A constant, β = The marginal propensity to consume

An applied statistician would ask...

1. Sounds great, but is that true?
2. if yes, how big is β? $\rightarrow$ That's our job...

Agenda for the Rest of the Lecture

Reminder: Basic Statistic Measures

A Brief History of Probability Theory and Statistics
- The Origins of Probabilistic Thinking
- The Sample Space
- Pascals Triangle
- The Law of Large Numbers
- Conditional Probability & Bayesian Statistics
- The Normal Distribution
- The Error Law & Central Limit Theorem
- The Birth of (Social) Statistics

Introduction to Linear Regression Techniques
- OLS fundamentals
- OLS Assumptions

OLS metrics
- T and P Tests – Significance of Coefficients
- R2 – Model fit

Mean: Gives insight on the average value of observed variables

$$\mu \hat{=} \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2}$$

Median: Gives insight on the value of variables for the average observation

$$\eta \hat{=} M = X_{\frac{n=1}{2}} \; where: \; X_1, \ldots, X_n \rightarrow X_{min}, \ldots, X_{max} \tag{3}$$

Standard deviation: Gives insight on the amount of variation or dispersion of a set of data values.

$$sd(X) \hat{=} \sigma_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{4}$$

Covariation: Gives insight how much two random variables X and Y change together

$$cov(X, Y) \hat{=} \sigma_{XY} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n} \tag{5}$$

Correlation: Gives insight how changes in X are associated with changes in Y, and *vice versa*

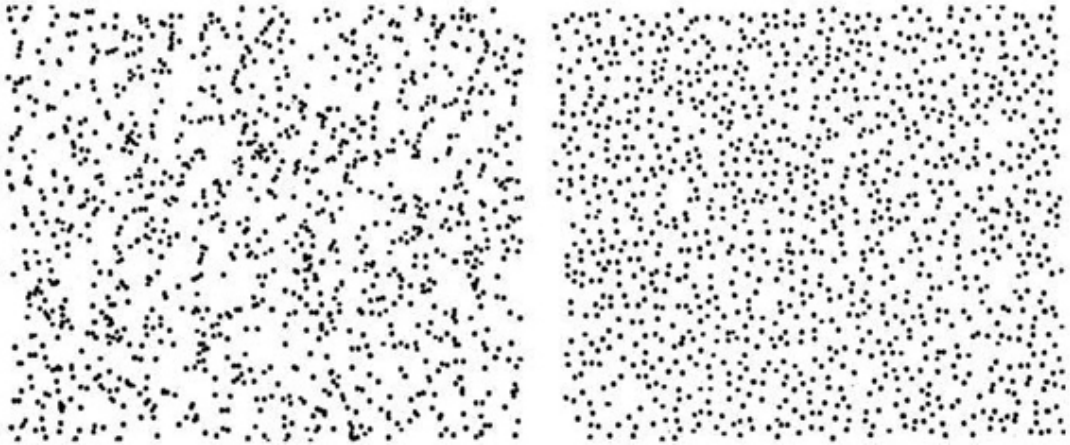
$$corr(X, Y) \hat{=} \phi_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} \tag{6}$$

$\rightarrow$ Scale free version of correlation (in units of standard deviation)

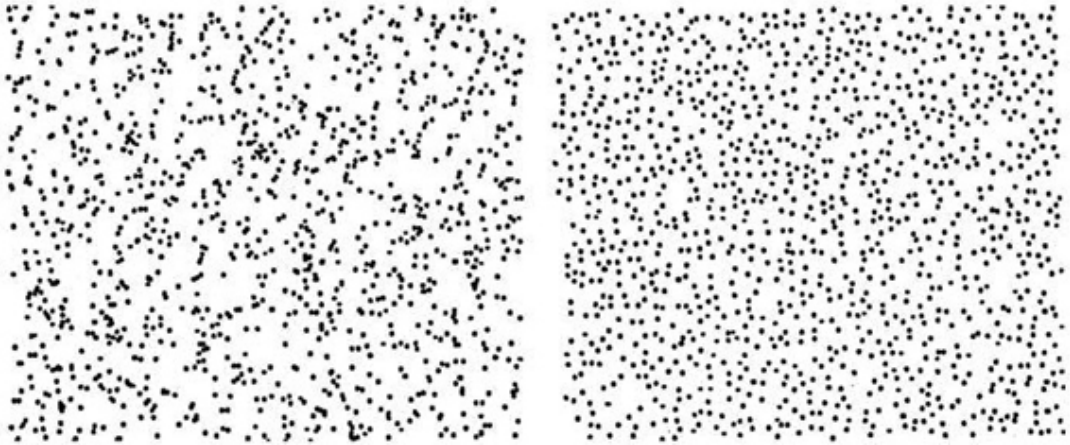
Why bother with probability

- Probability theory is the foundation on which not only econometrics but indeed statistics as a whole is built.
- Indeed, certain assumptions regarding probabilities of events and characteristics are vital for econometrics to work out...
- Main idea is that experimental (and if choosen the right sample, also observational) data usually follows some pattern of stochastic regularity wrt. the distribution of observations/events and their characteristics/variables.
- This leads to a probability distribution of events and their characteristics, which we would like to reveal.
- Why? Because this distribution helps us to develop an apply methods of statistical inference.
- Key concepts: Sample space, conditional probability, the central limit theorem, the law of large numbers, probability distributions

Our perception of randomness



Our perception of randomness



- Turns out, our brain is not really good in assessing large scale empirical data, and infer probabilities, detect pattern, causalities, ect.

Back in time: The Greeks (B.C.)

- The style of Euclid: Originating from a small set of axioms, everything is proofable. Axioms, proof, theorems, more proofs, more theorems
- Yet, no concept of probability. Why?
- Since everything that happened, happens, and will happen is given by gods, the only thing men can do is understand their will

Back in time: The Greeks (B.C.)

- The style of Euclid: Originating from a small set of axioms, everything is proofable. Axioms, proof, theorems, more proofs, more theorems
- Yet, no concept of probability. Why?
- Since everything that happened, happens, and will happen is given by gods, the only thing men can do is understand their will

Cicero (ca.50: B.C.), Roman Statesman

"The Greeks held the geometer in the highest honor; accordingly, nothing made more brilliant progress among them than mathematicians. But we have established as the limit of this art its usefulness in measuring and counting."

- First understanding, that an event can be anticipated and predicted, even though it is the outcome of pure chance.
- First appearance of the term "probabilis"
- Yet, that's where the worlds knowledge frontier remained for ca. 1.500years.

The Renaissance: Gerolamo Cardano (ca. 1570)

Chair of medicine at the university of Pavia, published 131 books in philosophy, medicine, astronomy, physics, and one called “The Book on Games of Chance”

- Brilliant academic, and obsessive gambler.
- His insight: The sample space
- The idea: When categorizing all possible outcomes of an event, and attaching a (theoretical) probability to it, we have a measure how likely any given set of events is to occur.

The Renaissance: Gerolamo Cardano (ca. 1570)

Chair of medicine at the university of Pavia, published 131 books in philosophy, medicine, astronomy, physics, and one called “The Book on Games of Chance”

- Brilliant academic, and obsessive gambler.
- His insight: The sample space
- The idea: When categorizing all possible outcomes of an event, and attaching a (theoretical) probability to it, we have a measure how likely any given set of events is to occur.

Experiments and Events

- An experiment is a process whose outcome is not known in advance.
- Possible outcomes (or realizations) of an experiment are events.
- Set of all possible outcomes is called the *sample space* (Ω).
- Probability reflects the likelihood that an event will occur
- Cumulative probability of all events $\in \Omega = 1$

For the sake of simplicity: Pr(X=x) from now on denoted as p(x), where X represents the full range of potentially observable values, and x its observed realization.

Independence of Events

Events, X=x and Y=y are independent if: $p(x,y) = p(x) * p(y)$ or $p(x|y) = p(x)$

Marginal Probability Distribution

= The probability of event X=x, independent of the state of Y

$$p(x) = \sum_{y \in Y} p(x,y) = \sum_{y \in Y} p(x|y)p(y)$$

Joint Probability Distribution

= The probability that independent events X=x and Y=y occur together. Formally, the probability of intersection of X and Y ($X = x \cap Y = y$)

$$p(x,y) == p(x|y)\ p(y) = p(y|x)\ p(x)$$

First Example: Rolling Dices

- An experiment involves rolling a single fair die
- Each of the six faces of the die is equally likely to come up when the die is tossed
- Sample space is $\Omega\{1,2,3,4,5,6\}$
- Discrete random variable, X, takes on values 1, 2, 3, 4, 5, 6
- Since the outcome of the experiment is unknown, X is a random variable.
- The probability of event X occurring will be denoted by Pr(X).
- *Realization* of random variable is the value which actually arises (e.g. if the die is rolled, a 4 might appear).

First Example: Rolling Dices

- An experiment involves rolling a single fair die
- Each of the six faces of the die is equally likely to come up when the die is tossed
- Sample space is $\Omega\{1,2,3,4,5,6\}$
- Discrete random variable, X, takes on values 1, 2, 3, 4, 5, 6
- Since the outcome of the experiment is unknown, X is a random variable.
- The probability of event X occurring will be denoted by Pr(X).
- *Realization* of random variable is the value which actually arises (e.g. if the die is rolled, a 4 might appear).

Questions:

- What is the probability of rolling a 2?
- And what the probability of rolling a 2-4?
- Finally, the probability of rolling two times a 6 in a row.

First Example: Rolling Dices

- An experiment involves rolling a single fair die
- Each of the six faces of the die is equally likely to come up when the die is tossed
- Sample space is $\Omega\{1,2,3,4,5,6\}$
- Discrete random variable, X, takes on values 1, 2, 3, 4, 5, 6
- Since the outcome of the experiment is unknown, X is a random variable.
- The probability of event X occurring will be denoted by Pr(X).
- *Realization* of random variable is the value which actually arises (e.g. if the die is rolled, a 4 might appear).

Questions:

- What is the probability of rolling a 2?
- $p(x=1) = p(x=2) = \cdots = p(x=6) = \frac{1}{6}$
- And what the probability of rolling a 2-4?
- $p(x \in \{2,3,4\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$
- Finally, the probability of rolling two times a 6 in a row.
- $p(x=6, y=6) = \frac{1}{6} * \frac{1}{6} = \frac{1}{2}$

The "Maserati Problem"

- Suppose you are a candidate in the show "Let's make a deal."
- You can choose between 3 doors, where are behind two a goat (Zonk), and behind one a Maserati awaits you.
- You choose a door.
- To make you nervous, the moderator opens one of the remaining doors, where you find a goat.
- You get the chance to chance your pick of door, if you want.
- What to do? Stay with your pick, change, or does that not matter at all?

The "Maserati Problem"

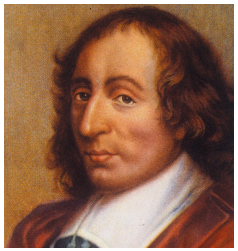
- Suppose you are a candidate in the show "Let's make a deal."
- You can choose between 3 doors, where are behind two a goat (Zonk), and behind one a Maserati awaits you.
- You choose a door.
- To make you nervous, the moderator opens one of the remaining doors, where you find a goat.
- You get the chance to chance your pick of door, if you want.
- What to do? Stay with your pick, change, or does that not matter at all?

Answer:

- For your first choice, your probability of picking the Maserati is: $\frac{1}{3}$
- If you stick to your choice, that will remain.
- With revealing one outcome, the moderator changed the sample space.
- Consequently, if you change your pick now, your probability will be: $\frac{1}{2}$
- If doing so, you can thank the moderator for an increased probability of: $\frac{1}{6}$

For further insights on probability theory, we had to wait for 150 years more, till the French revolution.

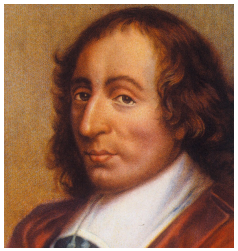
Blaise Pascal (1623 – 1662): French mathematician



- His problem: Suppose you play a game of chance with another player, where the winner takes the whole pot. The game is interrupted while some player is in the lead. So, how should the pot be split?
- Insight: If all outcomes are equally likely, summing the set of possible outcomes that lead to an final event gives you the probability of this event to occur.

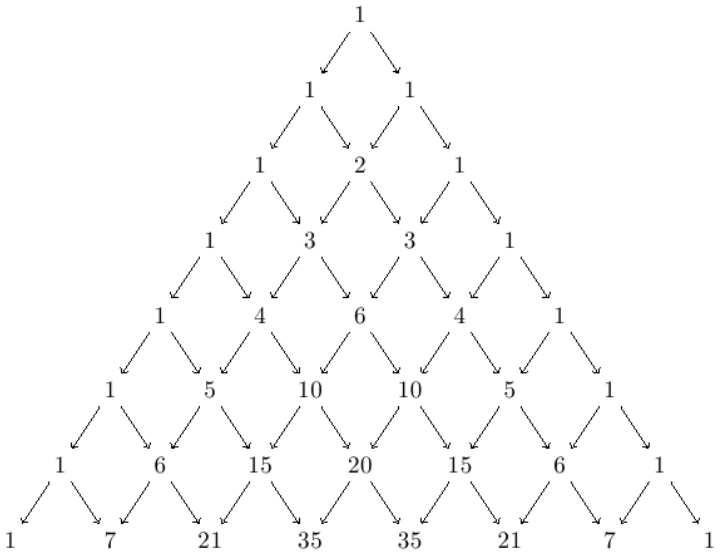
For further insights on probability theory, we had to wait for 150 years more, till the French revolution.

Blaise Pascal (1623 – 1662): French mathematician



- His problem: Suppose you play a game of chance with another player, where the winner takes the whole pot. The game is interrupted while some player is in the lead. So, how should the pot be split?
- Insight: If all outcomes are equally likely, summing the set of possible outcomes that lead to an final event gives you the probability of this event to occur.

Pascals triangle



- A measure of possible combinations of random events leading to a certain outcome.
- Technically: Determines the coefficients which arise in binomial expansions.

Up to now (Cardano, Pascal) assumed the probability of events to be known. However, for most problems that matter, probabilities are not known a priori. What's the chance of winning a war, finding your soul-mate, or dying of cancer?

Jacob Bernoulli (1655 – 1705): Swiss mathematician

- Realization: Existing concepts of probability are not designed for situations of ignorance, where the probability of possible outcomes are theoretically defined, but practically unknown.
- His solution (in the spirit of the time): Observe outcomes. But how many, to be certain?

Up to now (Cardano, Pascal) assumed the probability of events to be known. However, for most problems that matter, probabilities are not known a priori. What's the chance of winning a war, finding your soul-mate, or dying of cancer?

Jacob Bernoulli (1655 – 1705): Swiss mathematician

- Realization: Existing concepts of probability are not designed for situations of ignorance, where the probability of possible outcomes are theoretically defined, but practically unknown.
- His solution (in the spirit of the time): Observe outcomes. But how many, to be certain?

The Law of Large Numbers (Bernoulli's "Golden Theorem")

- Concerns the way results reflect underlying probabilities when we make a large number of observations.
- Or: How likely does the outcomes reflect the true underlying probability, given the number of observations.
- If observations are picked random from the sample, then Pascal's work on the distribution of random outcomes might be helpful.
- Insight: Tolerance of error as well as tolerance of uncertainty determine the likelyhood of an outcome. Both decrease with increasing the number of observations

Formally: $Pr(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$

Thomas Bayes: 1701 – 1761, British Minister

- Catholic preacher, British minister and hobby scientist, post-mortem publication of "An Essay towards solving a Problem in the Doctrine of Chances"
- How can we infer underlying probability from observation.
- Sparked a new school of thought, where probability as an amount of epistemic confidence – the strength of beliefs, hypotheses etc. –, rather than a frequency.

Thomas Bayes: 1701 – 1761, British Minister

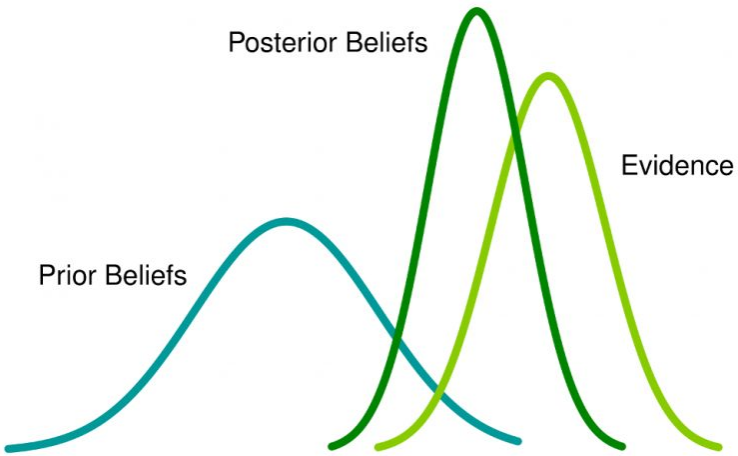
- Catholic preacher, British minister and hobby scientist, post-mortem publication of "An Essay towards solving a Problem in the Doctrine of Chances"
- How can we infer underlying probability from observation.
- Sparked a new school of thought, where probability as an amount of epistemic confidence – the strength of beliefs, hypotheses etc. –, rather than a frequency.

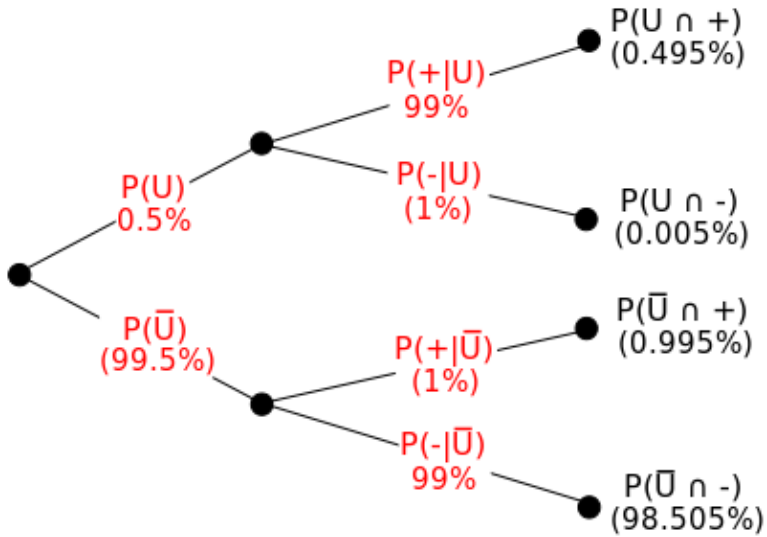
Conditional Probability

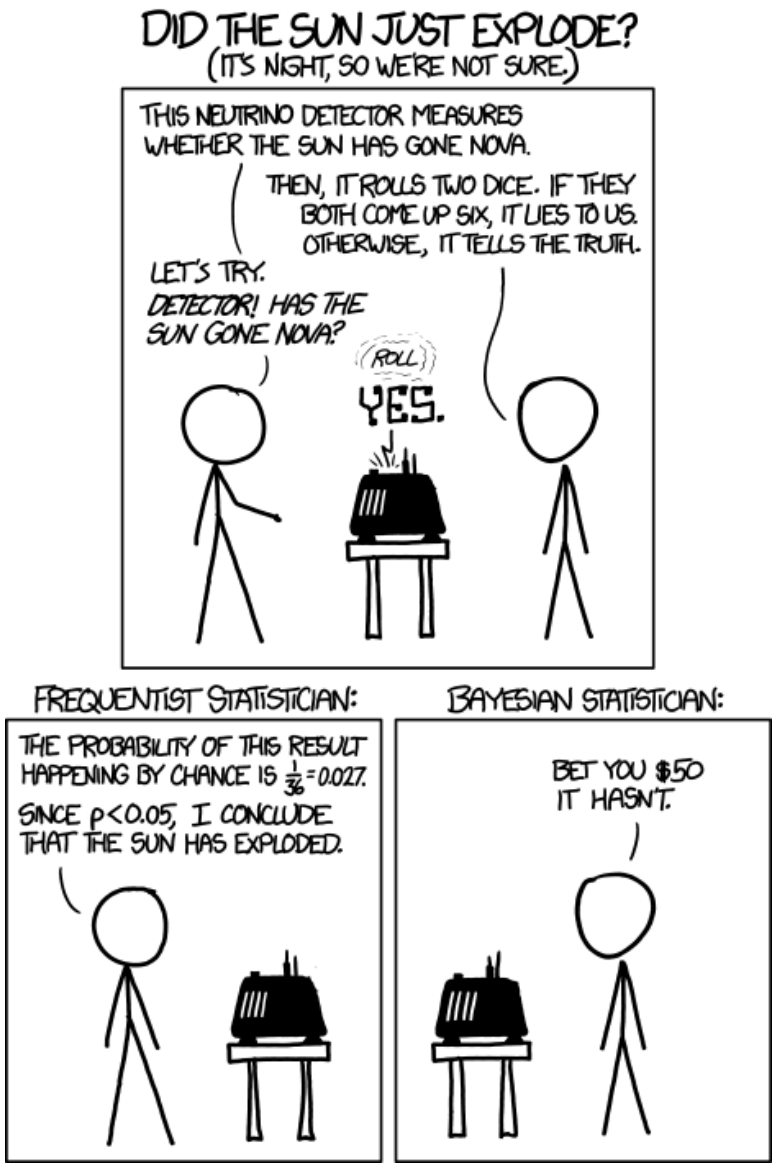
= the probability of event X occurring, given we know event Y has occurred.

$$p(x|y) = \frac{p(y|x) * p(x)}{p(y)} \Rightarrow \frac{p(y|x)}{p(x|y)} = \frac{p(x)}{p(y)}$$

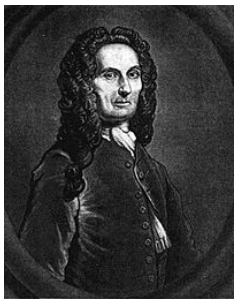
- In the Bayesian interpretation, probability measures a "degree of belief", which are not given ex ante.
- Accordingly, a probability is assigned to a hypothesis, whereas under frequentist inference, hypotheses are tested without being assigned a probability.
- Focusing on evidential probability to evaluate a hypothesis, meaning the prior probability is constantly updated to a posterior probability in the light of new, relevant data (evidence).
- Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence.
- Revival this century, where the Bayesian interpretation is embraced and utilized to develop new methods such Vector-Autogregression Models (VAR), but also inspired the machine learning community, and the idea of Bayesian updating offered new ways to model learning in general.



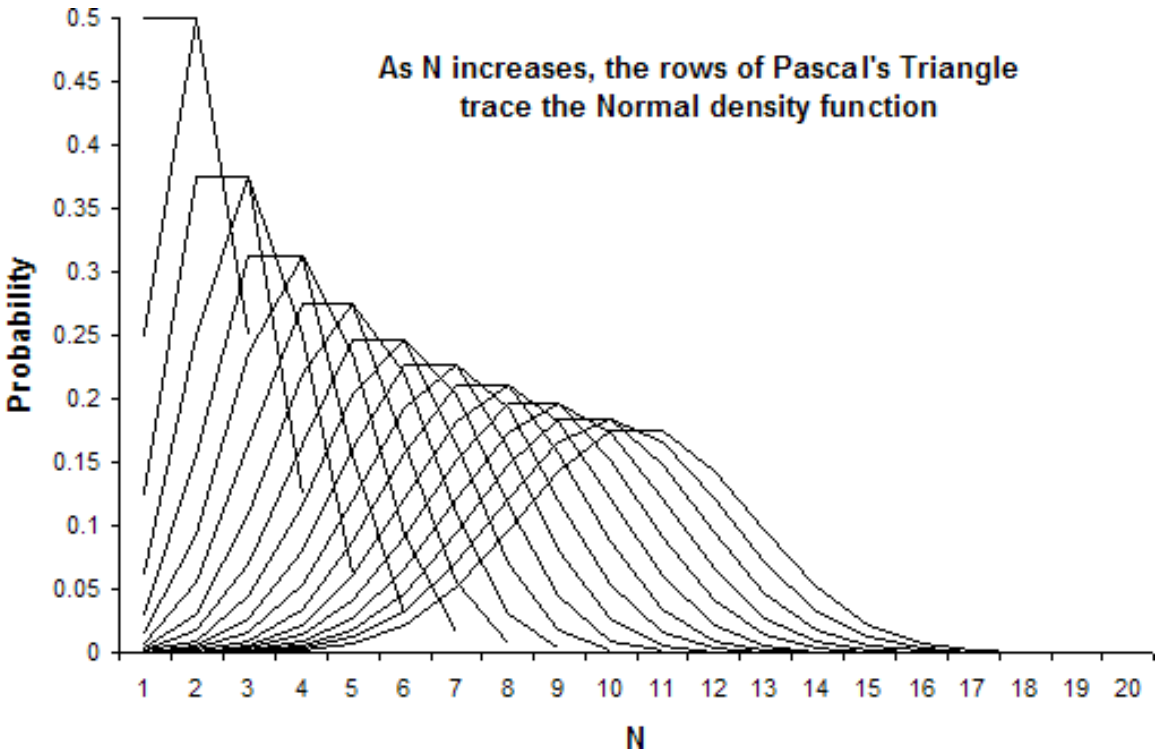


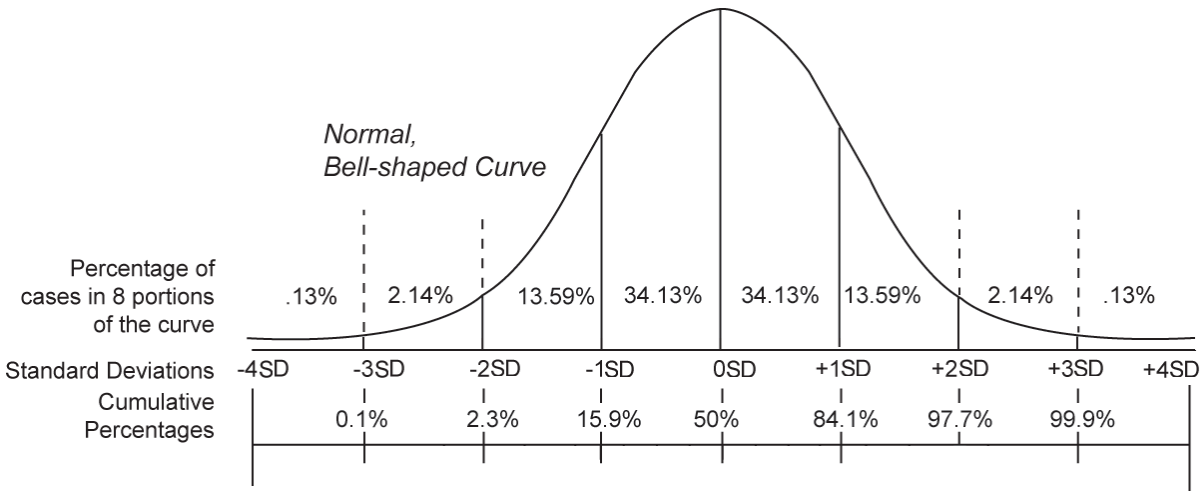


Abraham De Moivre (1667 –1754): French mathematician



- In “The doctrine of chance", development of clever method (supported by the invention of integral calculus) to solve Pascal’s triangle up to the 200th row (55 digits number)
- When you go down far enough in the Pascal triangle, and plot it, it looks like something we nowadays know as the bell curve, also known as the Normal Distribution.





The Normal Distribution

- A symetric probability density function (PDF), depicting the probability of every given event X to occur.

$$f(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}\varepsilon^{-\frac{(\chi-\mu)^2}{2\sigma^2}}} \tag{7}$$

- μ = Mean, determining the maximum of the curve
- σ = Standard deviation, determining the spread of the curve
- By definition (PDF), the entire area under the curve: $\int_{Y_{min}}^{Y_{max}} f(Y) = 1$

- Up to now, we assumed measurements of realization to be precise. But is that true? Think of wine tasting, grading of projects, counting votes by hand, ect.
- Indeed, measures are generally imperfect, especially when carried out by humans. That's one of the great contradictions in life.
- Problem: Main problem in the mid-18th century, after the formulation of Newton's law, when a main academic effort was put into celestial physics.
- The precise measurement of the movement of planets, and the calculation of their multiple gravitational pull (by hand) is complex, and prone to errors.

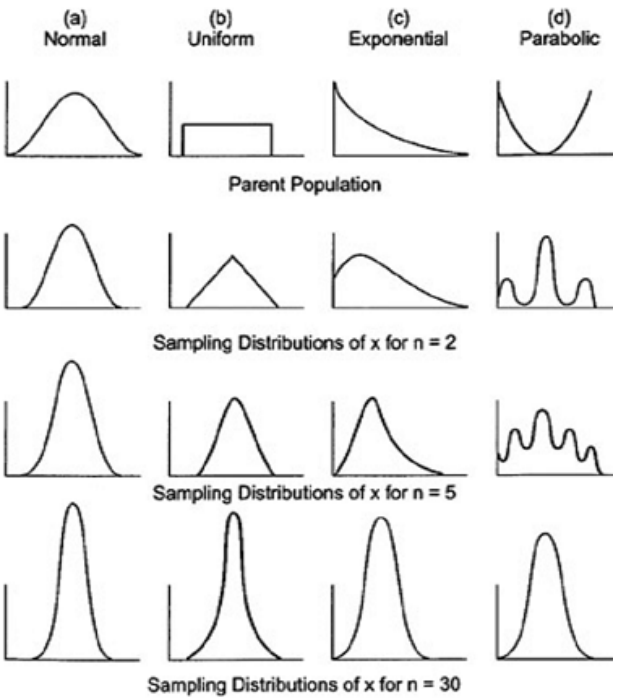
Pierre-Simon Laplace (1749 - 1827): French Universal Scholar

- Building on Gauss work on the distribution of random measurement errors in “The Theory of the Motion of Heavenly Bodies Moving around the Sun in Conic Sections”
- Final proof of the error law, widely accepted by academics.

Pierre-Simon Laplace (1749 - 1827): French Universal Scholar

- Building on Gauss work on the distribution of random measurement errors in “The Theory of the Motion of Heavenly Bodies Moving around the Sun in Conic Sections”
- Final proof of the error law, widely accepted by academics.

The Central Limit Theorem



- With increasing sample size, the sample mean is approximated by a normal distribution, regardless of the populations parent distribution.
- Final proof of the error law, widely accepted by academics.

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right). \qquad (8)$$

- By the late 17th century, now, probability theory has fully emerged.
- Yet, for the main part, it's application was limited was limited to problems in the physical sphere, where it was essential to account for measurement errors around means. This was done using the method of least squares (Gauss).
- This was about to change...

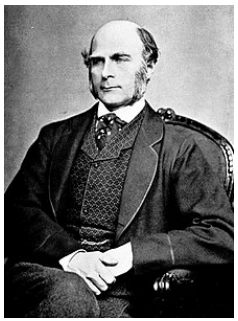
"If an intelligence, at a given instant, knew all the forces that animate nature and each constituent being; if, moreover, this intelligence were sufficiently great to submit this data to analysis, [...] nothing would be uncertain, and the future, as the past, would be present to its eyes."
Laplace, 1814 (The height of Newtonian Physics)

- This expresses a new view of the world called "determinism", where the future of individuals and societies can be foreseen just like the behavior of atoms, planets, and molecules; given the right data and method.
- And he was determined to bring the apparatus of probability theory to the course of human affairs.

Adolphe Quetelet: 1796–1874: Belgian scholar

- Obsession with measurement. He collected measures on pretty much all human related numbers, such as propensity to crime, marriage and death rates, average height, chest measurements ect.
- Whatever data he got his hands on, he found normal distributions.
- Insight: Human activity and characteristics to a large extend follows a normal distribution, making statistical inference on matters of society possible $\rightarrow$ "social physics"

Sir Francis Galton (1822 – 1911), British scholar



- Coined the term "regression" for the application Gaussian ordinary least squares techniques (OLS) to describe biological and social phenomena.
- His phenomenon of interest: That the heights of descendants of tall ancestors tend to regress down towards a normal average, coined regression toward the mean).
- Also thought-leader of eugenics (nature-vs-nurture)
- Firstly introduced survey techniques, and founded the field of psychometrics/

OLS = Ordinary Least Squares, a linear regression technique

Basic Properties

- DV: contionous,
- IV: continous, dichotonomous, categorical
- Common "allrounder"
- When to use: Explaining a phenomenon that scales and can be measured continuously

Functional form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \tag{9}$$

where:
y = DV, x_i = observed value ID_i
β_0 = Constant
β_i = Estimated effect of x_i (IV) on y (DV), slope of the linear function
ε = Error term (also denoted as u, eg. by Wooldridge)

Regressions - Example

- For instance, an econometrician could be interested in how much one's income increases for an additional year of education, *ceteris paribus*
- To answer this question, an econometrician typically collects data on education and income (and further variables, which we neglect here)

Regressions - Example

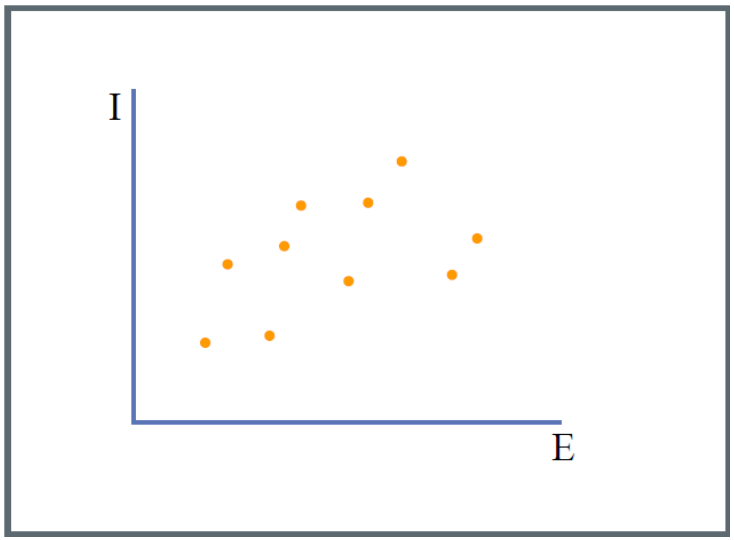
- For instance, an econometrician could be interested in how much one's income increases for an additional year of education, *ceteris paribus*
- To answer this question, an econometrician typically collects data on education and income (and further variables, which we neglect here)

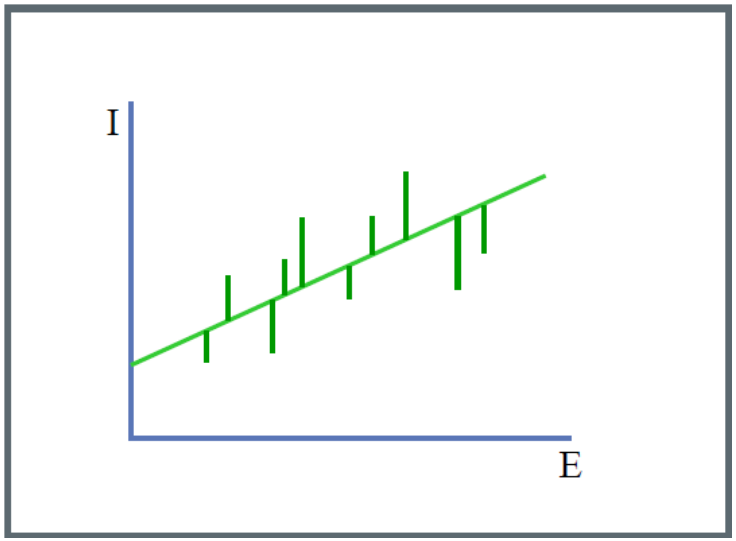
Regressions - Example

Now, we build a model, which would look somewhat like:

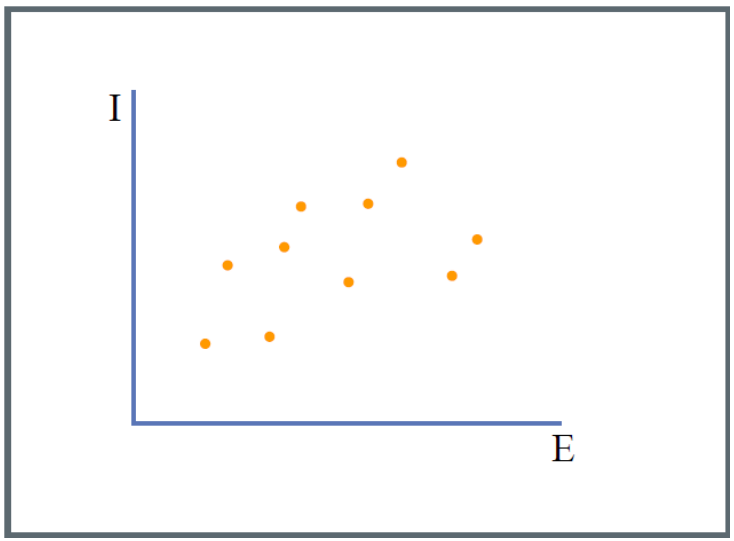
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{10}$$

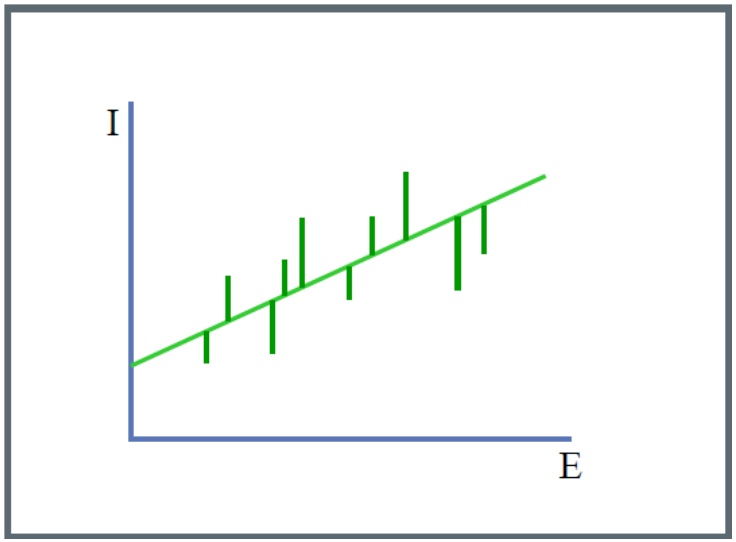
- Y = A persons annual income
- β_0 = Expected income for a person without education ($X_1 = 0$)
- X_1 = A persons years of education
- β_1 = The marginal effect of education on income
- $X_2 - X_n$ = A set of control variables, which we suspect to also matter (neglected in example)
- $\beta_2 - \beta_n$ = The marginal effect the other control variables (neglected in example)





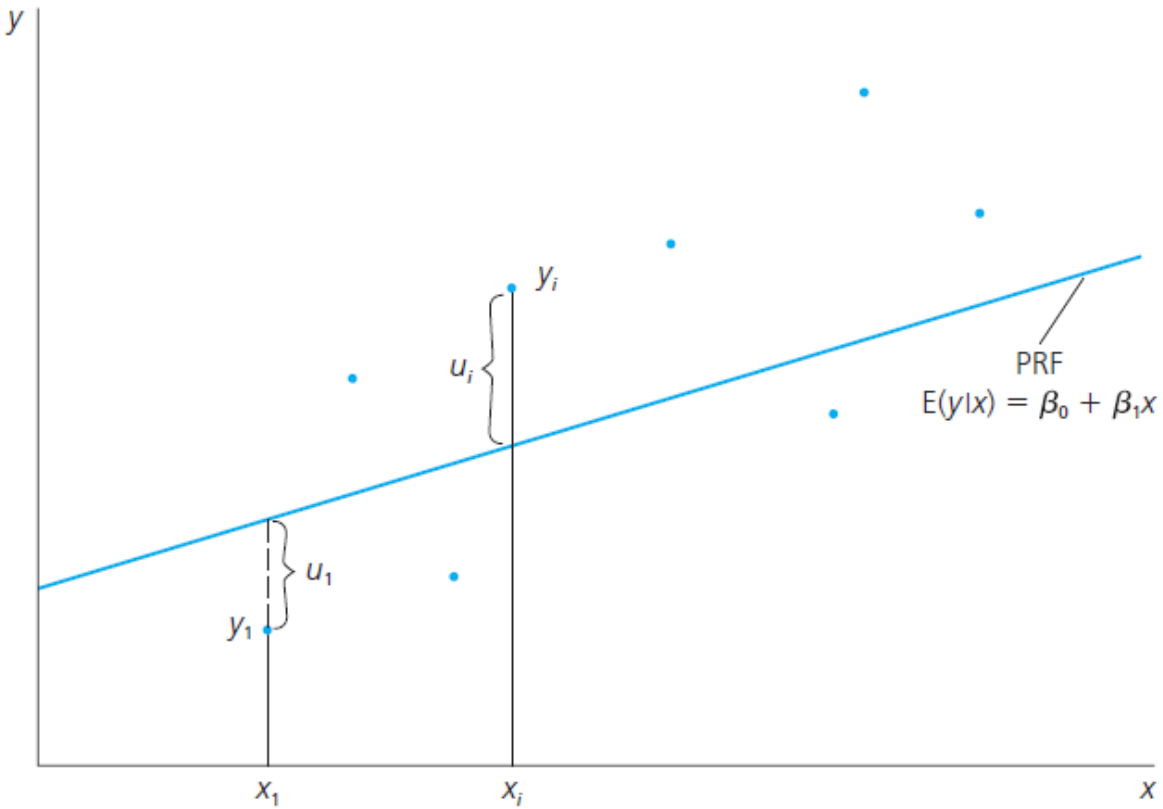
- I now collected some data on E and I. This is how it looks when I plot it.
- On first glance it shows some relationship between E and I
- We now want to estimate β_1, the marginal effect if E on I (technically equivalent to $\frac{\partial I}{\partial E}$)
- Graphically, our functional form will put a straight line through the data, with a slope of β_1





- Now our regression model estimated β_0 and β_1.
- The regressions typically fits the slope to minimize the sum of squared distances between observed datapoints and estimated function of the model ($\hat{=} min \sum_{i=1}^{n} \varepsilon_i^2$)

Illustration: A fitted Regression Line



OLS assumptions I: Unbiasedness

Under the following four assumptions, OLS is unbiased, meaning $E(\hat{\beta}) = \beta$

- **Sample Variation:** Not all x_i can have the same value
- **Random Sampling:** The x_i values must be randomly selected. In other words, there is no correlation between two different x values: $Cov(x_i, x_j) = 0$ for $\forall i \neq j$
- **Zero Conditional Mean:** The mean of the error terms, given a specific value of the independent variable x_i, is zero. $E(\varepsilon_i | X_i) = 0$.

OLS assumptions II: Efficiency

for OLS to be BLUE (Best Linear Unbiased Estimator)

- **No Heteroskedasticity:** The variance of the error terms are constant. $Var(\varepsilon | x_i) = \sigma^2$. This means that the variance of the error term ε_i does not depend on the value of x_i.
- **No Serial Correlation:** The error terms are independently distributed so that their covariance is 0. $Cov(\varepsilon_i, \varepsilon_j | x_i, x_j) = 0 \forall i \neq j$
- **Normally Distributed Errors:** Error terms normally distributed: $\varepsilon \approx N(0, \sigma^2)$

- **Sample Variation:**
  - $X_{min} \neq X_{max}, \sigma_X \neq 0$
- **Random Sampling:**
  - Easy way: Split the sample
  - Perform a Mann-Whitney-Wilcoxon test on the main variables of interest between the two samples
- **Zero Conditional Mean:**
  - Non-Zero mean not exactly testable, since it would be absorbed by the constant. What is testable: $\sigma_\varepsilon \neq 0$
  - Easy first graphical diagnosis: Plot residuals against fitted values
- **No Heteroskedasticity:**
  - Perform a Breuch Pagan test, or the more general White Test on the presence of heteroskedasticity
- **No Serial Correlation:**
  - Regress the consecutive residuals against each other and test for a significant slope.
  - If there is auto-correlation, then there should be a linear relationship between consecutive residuals.
- **Normally Distributed Errors:**
  - Easy way: graphical diagnosis.
  - Normal Probability Plot of Residuals, compares a dataset (here, your regression residuals) with a normal distribution.

$\Rightarrow$ Applications will follow in the lab session

- While the coefficient $\hat{\beta}_i$ gives us insights on the amplitude of the correlation between x_i and y in a multivariate (*ceteris paribus*) setting, it tells us nothing about the strength of this relationship
- To interpret these coefficients and draw conclusions for real-life phenomena, we want to know much of the coefficients value can be attributed to a real empirical relationship, and how much just due to randomness

Calculation of the T Value

Let: β_0 be a known constant, $\hat{\beta}_i$ an estimator of parameter β_i, and $s.e.(\hat{\beta}_i)$ the standard error of the estimator. Then a t-statistic for this parameter is given by:

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i - \beta_0}{s.e.(\hat{\beta}_i)} \tag{11}$$

- The t-value measures the difference relative to the variation in your sample data, measured in units of standard error.
- Signal-to-noise metaphor: Signal of the effects strength versus the noise caused by the variability of the data.
- The greater the magnitude of T (positive or negative), the greater the evidence against the null hypothesis.

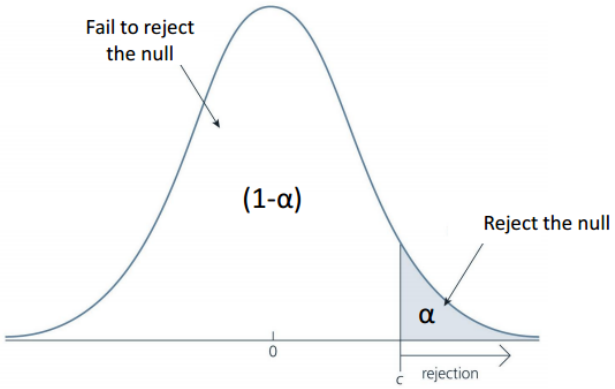
Performing a T test on a coefficient

- When performing a T-test, you usually try to find evidence of a significant difference between population means (2-sample t) or between the population mean and a hypothesized value (1-sample t).
- By default, statistical packages report t-statistic with $\beta_0 = 0$ (these t-statistics are used to test the significance of corresponding IV)
- However, when t-statistic is needed to test the hypothesis of the form $H_0 : \beta = \beta_0$, then a non-zero β_0 may be used (these t-statistics are used to evaluate the difference of means between groups).

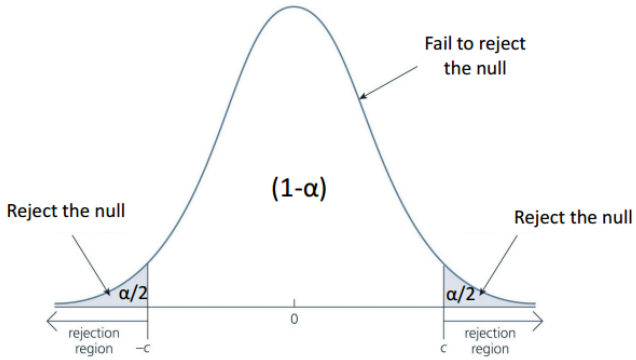
Steps

1. Set the null hypothesis (and the alternative): E.g. $H_0 : \beta_i = 0, H1 : \beta_i > 0$
2. Calculate the T-value
3. Choose a significance level α (Probability of rejecting the null if the null is actually true, committing a **Type I error**), eg.: $\alpha = 0.05$
4. Given the α level, the degrees of freedom (df = Number of variables or categories - 1), and the t-value you can look the critical value up in a standard table of significance (like back in the days...).
   $\rightarrow$ Fortunately, statistical software nowadays the significance test results and save you the trouble of looking them up in a table.

[One-Sided T test]

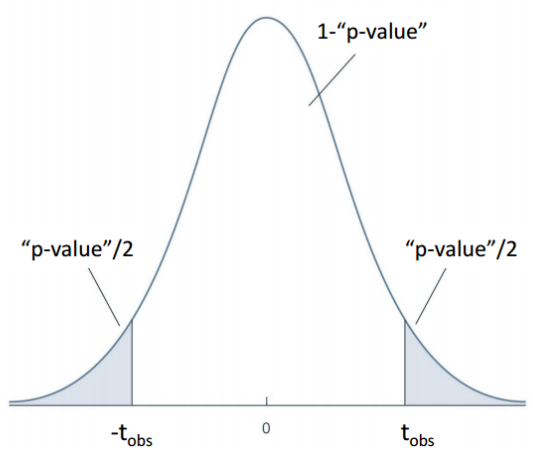


[Two-Sided T test]



One- versus Two-Sided T Test

- When using a one-sided test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction (e.g., $H_0 : \beta_j \leq 0; H_1 : \beta_j > 0$)
- When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in



The P value

- The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when H_0 is true: $Prob[|t| > |t - obs|H_0]$
- P is also described in terms of rejecting H_0 when it is actually true (Type I error). However, it is not a direct probability of this state!
- Closely connected to the T-test. Difference: While pre-choosing α in the T-test, the term P-value is used to indicate a probability that you calculate after a given study.
- Common consensus in econometrics to interpret **coefficients** above the threshold of p=0.05(*) $\rightarrow \alpha = 0.1$, 0.01(**) $\rightarrow \alpha = 0.05$, 0.001 (***) $\rightarrow \alpha = 0.01$

R^2 = Common measure of the "explanatory power" of linear models

R^2 & Adjusted R^2 Calculation

- Total sum of squares: $SST = \sum_i (y_i - \bar{y})^2$
- Explained sum of squares: $SSE = \sum_i (\hat{y} - \bar{y})^2$
- Sum of squares of residuals: $SSR = \sum_i (y_i - \hat{y})^2$
- Relationship: $SST = SSE + SSR$

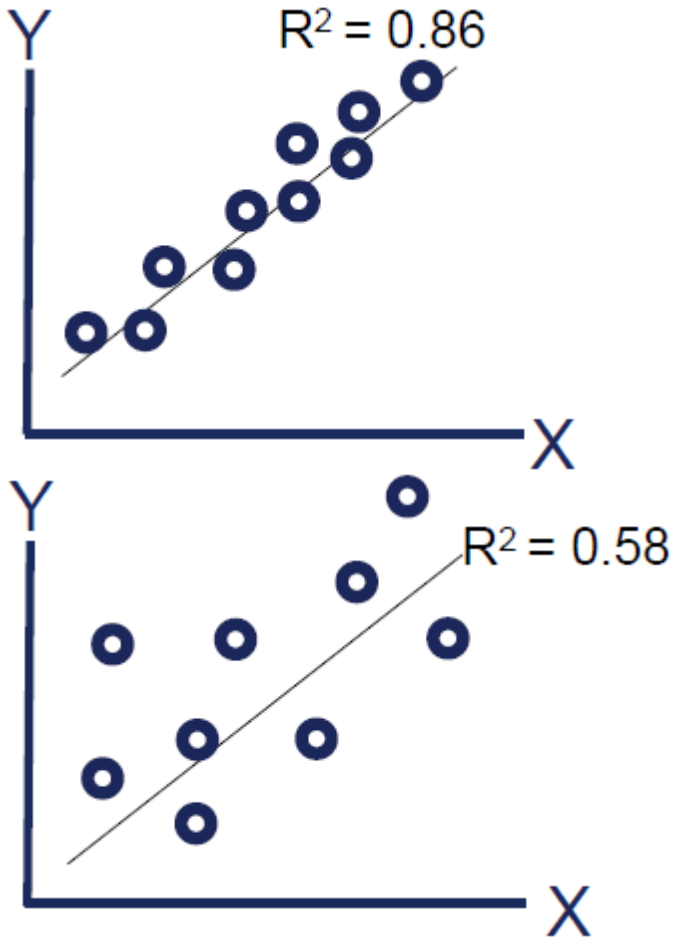
$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{12}$$

Note I: $R2$ automatically and spuriously increasing when extra explanatory variables are added to the model (given constant sample size). The adjusted R^2 ($\bar{R}^2$) penalizes extra variables, thus facilitates conscious model building.

$$\bar{R}^2 = R^2 - (1 - R^2)\frac{p}{n-k-1} = 1 - \frac{SSR/df_e}{SST/df_t} \tag{13}$$

Note II: R^2 does not indicate whether:

- The IVs are a cause of the changes in the dependent variable
- Omitted-variable bias exists
- The correct regression technique was used



Interpretation: R^2 is the ratio of the explained variation compared to the total variation; thus, it is interpreted as the fraction of the sample variation in y that is explained by x

So, what's happening here...?

Joining forces: collaboration patterns and performance of renewable energy innovators

Table 7 OLS regression on (i) non-RE innovators (PSM) and (ii). RE innovators, DV: Innovation turnover

	(1) Non-RE	(2) RE	(3) Non-RE	(4) RE	(5) Non-RE	(6) RE	(7) Non-RE	(8) RE
Firm region	6.159	−2.499	6.134	−2.213	5.443	−3.702	5.838	−3.484
	(4.174)	(4.629)	(4.073)	(4.632)	(4.166)	(4.540)	(4.091)	(4.568)
Firm legal	1.647	4.642	3.662	5.380	2.302	5.683	3.791	5.641
	(4.866)	(5.495)	(4.773)	(5.512)	(4.850)	(5.382)	(4.778)	(5.419)
Firm age_{ln}	−5.021	0.0405	−4.774	0.248	−4.445	1.306	−4.564	1.347
	(2.588)	(2.708)	(2.528)	(2.706)	(2.588)	(2.670)	(2.542)	(2.679)
Firm $empl_{ln}$	−5.160**	−1.475	−5.812***	−1.545	−4.555**	0.225	−5.517**	0.215
	(1.737)	(2.033)	(1.702)	(2.033)	(1.752)	(2.038)	(1.740)	(2.064)
coll breadth			6.198***	2.797			5.848***	0.983
			(1.707)	(1.789)			(1.759)	(1.838)
coll $breadth^2$			−0.928	−0.410			−0.904	−0.271
			(0.503)	(0.482)			(0.504)	(0.475)
R&D $intens_{ln}$					1.831*	3.604***	0.753	3.510***
					(0.875)	(0.938)	(0.913)	(1.029)
λ	−42.14**	−27.24	−35.73*	−22.72	−34.84*	−13.75	−33.26*	−13.29
	(14.31)	(16.53)	(14.09)	(16.69)	(14.66)	(16.55)	(14.41)	(16.64)
Industry controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	317	317	317	317	317	317	317	317
R-sq	0.089	0.062	0.138	0.072	0.102	0.105	0.140	0.106
Adj. *R*-sq	0.062	0.035	0.107	0.039	0.072	0.076	0.106	0.071

Bootstrapped (jackknife, 500 rep.) standard errors in parentheses
*, **, ***Indicate significance at 10, 5, and 1% levels, respectively

So, what's happening here...?

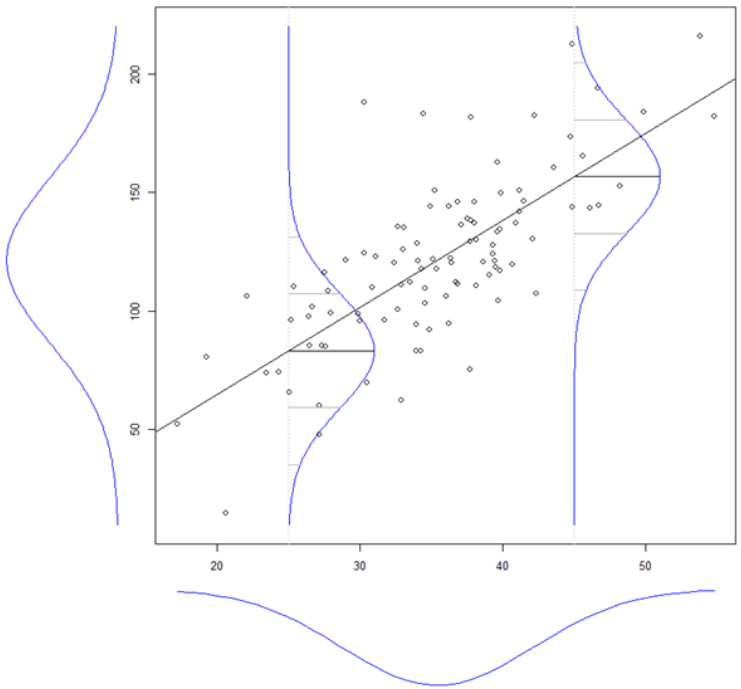
Joining forces: collaboration patterns and performance of renewable energy innovators

Table 7 OLS regression on (i) non-RE innovators (PSM) and (ii). RE innovators, DV: Innovation turnover

	(1) Non-RE	(2) RE	(3) Non-RE	(4) RE	(5) Non-RE	(6) RE	(7) Non-RE	(8) RE
Firm region	6.159	−2.499	6.134	−2.213	5.443	−3.702	5.838	−3.484
	(4.174)	(4.629)	(4.073)	(4.632)	(4.166)	(4.540)	(4.091)	(4.568)
Firm legal	1.647	4.642	3.662	5.380	2.302	5.683	3.791	5.641
	(4.866)	(5.495)	(4.773)	(5.512)	(4.850)	(5.382)	(4.778)	(5.419)
Firm age_{ln}	−5.021	0.0405	−4.774	0.248	−4.445	1.306	−4.564	1.347
	(2.588)	(2.708)	(2.528)	(2.706)	(2.588)	(2.670)	(2.542)	(2.679)
Firm $empl_{ln}$	−5.160**	−1.475	−5.812***	−1.545	−4.555**	0.225	−5.517**	0.215
	(1.737)	(2.033)	(1.702)	(2.033)	(1.752)	(2.038)	(1.740)	(2.064)
coll breadth			6.198***	2.797			5.848***	0.983
			(1.707)	(1.789)			(1.759)	(1.838)
coll $breadth^2$			−0.928	−0.410			−0.904	−0.271
			(0.503)	(0.482)			(0.504)	(0.475)
R&D $intens_{ln}$					1.831*	3.604***	0.753	3.510***
					(0.875)	(0.938)	(0.913)	(1.029)
λ	−42.14**	−27.24	−35.73*	−22.72	−34.84*	−13.75	−33.26*	−13.29
	(14.31)	(16.53)	(14.09)	(16.69)	(14.66)	(16.55)	(14.41)	(16.64)
Industry controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	317	317	317	317	317	317	317	317
R-sq	0.089	0.062	0.138	0.072	0.102	0.105	0.140	0.106
Adj. *R*-sq	0.062	0.035	0.107	0.039	0.072	0.076	0.106	0.071

Bootstrapped (jackknife, 500 rep.) standard errors in parentheses
*, **, ***Indicate significance at 10, 5, and 1% levels, respectively

- Different models with coefficient estimates
- Standard Errors in Parenthesis
- *, **, *** indicate P-Values of 0.5, 0.1, 0.01, corresponding to 10, 5, 1 percent significance level (more on that later)



- In regressions, we are interested in the impact of one or many IVs X_n on one (or in special cases, many) DV.
- At any point X_n, there exists (by assumption) a unique distribution of Y. In simple linear models, we assume this to be a normal distribution with constant variance across all X values. Later on, we will relax this assumption and allow for different and changing pdf's
- How good observed values of Y_n fit in this distribution as well as it's variance is a main determinant of the explanatory power of single variable coefficient as well as the whole model.

Thank you for your attention. Any questions?

AALBORG UNIVERSITY
DENMARK

References I