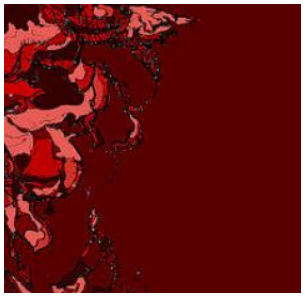
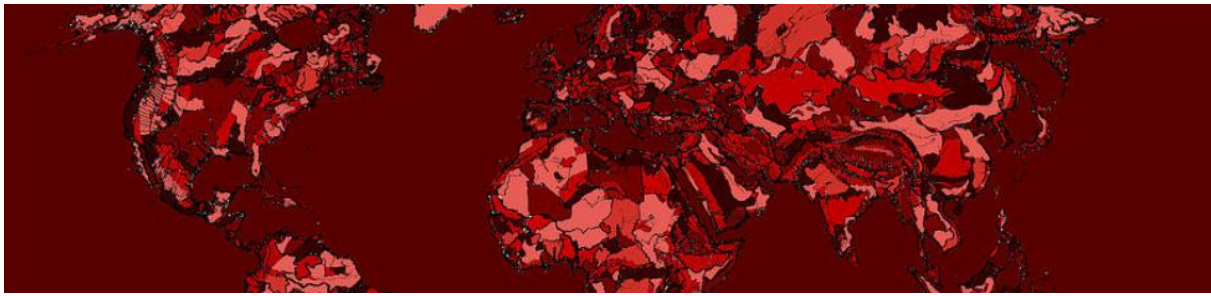
Math & Stat for MBA

Lecture Note 1

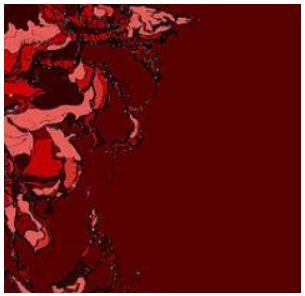
Swiss Institute of
Artificial Intelligence

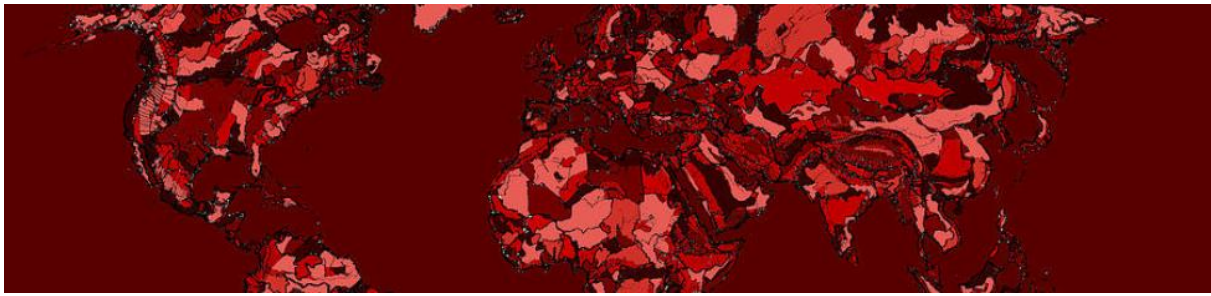




1. Basic Statistics

Swiss Institute of
Artificial Intelligence



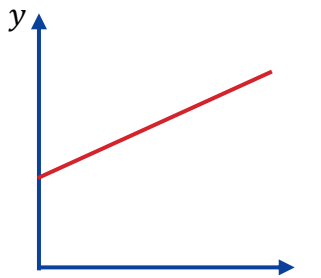


Introduction

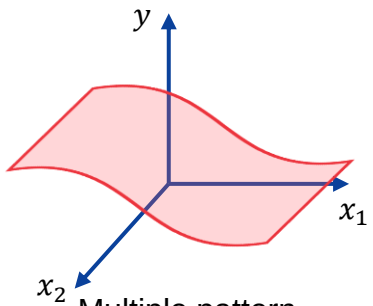
- Basic statistics

Basic statistics

- **Statistics are used everywhere**
  - Estimate the probability that incoming email is spam.
  - Pollsters can gauge the pulse of a large population on a variety of issues.



Single pattern



Multiple pattern

- **Many researchers lack a clear grasp of statistics**
  - It is a commonly accepted term that the sheer size is an important, and mostly a critical factor to determine whether the data is “BigData”.
  - The key factor of “BigData” is multiple patterns, not merely the amount of data. Only these “BigData” can provide rich information.

EXAMPLE – UNCOVERING DATA FAKERS

In 2008, a polling company called Research 2000 was hired by Daily Kos to gather approval data on top politicians.

	Favorable		Unfavorable		Undecided	
Topic	Men	Women	Men	Women	Men	Women
Obama	43	59	54	34	3	7
Pelosi	22	52	66	38	12	10
Reid	28	36	60	54	12	10
McConnell	31	17	50	70	19	13
Boehner	26	16	51	67	33	17
Cong.(D)	28	44	64	54	8	2
Cong.(R)	31	13	58	74	11	13
Party(D)	31	45	64	46	5	9
Party(R)	38	20	57	71	5	9

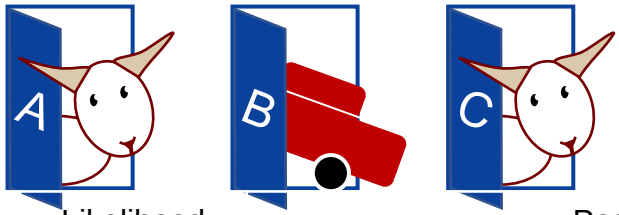
- The percentages from the men almost always had the same parity as the percentages from the women.
- The pollster not only rounded numbers but manipulated the outcome to look “clean”.
- Almost all summary data that are available in the real world are “touched”. What is required to unwrap the decoration is not a simple coding library that claims machine learning can do everything automatically.

Introduction

- Statistical thoughts and Moments

Two schools of statistical thought

- Monty Hall Problem
  - Suppose a player opened the door A, then a host opened C



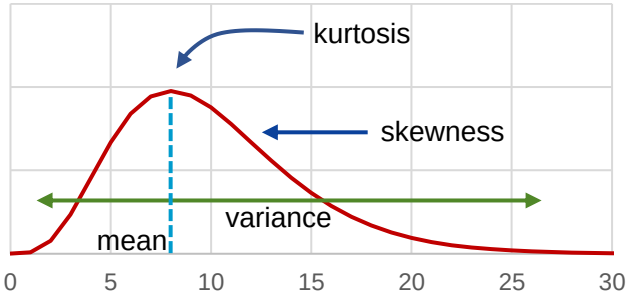
	Prior $P(H)$	Likelihood $P(D\|H)$	$P(H)\cdot P(D\|H)$	Posterior $P(H\|D)$
A	$1/3$	$1/2$	$1/6$	$1/3$
B	$1/3$	1	$1/3$	$2/3$
C	$1/3$	0	0	0

$$P(\theta|D) = \frac{P(\theta)\cdot P(D|\theta)}{P(D)} \quad \left(posterior = \frac{prior \times Likelihood}{Evidence(data)}\right)$$

- The frequentists give the same weight to all trials, but the Bayesian statisticians may give different weights to the most recent trials.
- In other words, Bayesian can be applied to cases where background states are changing overtime.

Moment

Moment is the operator to calculate average



- 1st moment : mean, median, mode
  - Expectation (1st moment operator)

$$\boldsymbol{E(x) = \sum_{all\ a} p(a)\cdot a}$$

  - Equal-weighted Mean : $p(a) = \frac{1}{n}$
  - Unequal-weighted statistics : $p(a)$ can be a probability function
- 2nd moment : variance, covariance, correlation
  - Variance (2nd moment operator)

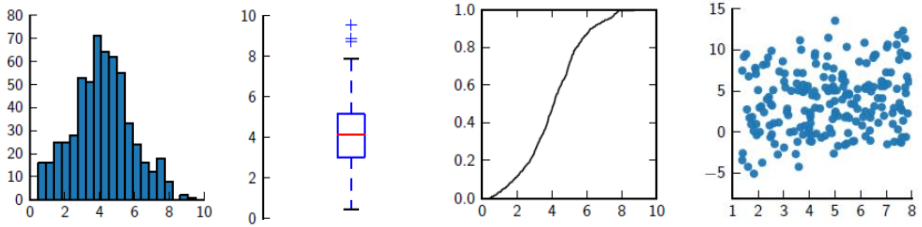
$$\boldsymbol{Var(x) = \sum_{all\ a} p(a)\cdot \left(a - E(x)\right)^2}$$

- 3rd moment : skewness
- 4th moment : kurtosis

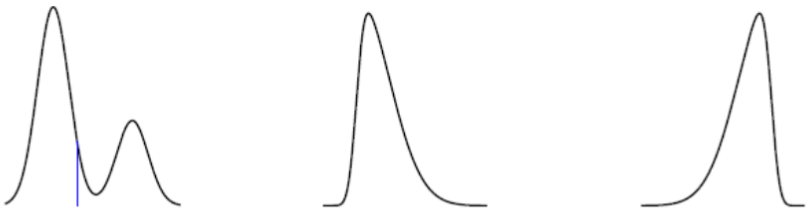
Exploratory Analysis

The different approaches for exploring data

- Plotting and visualizing data
  - This can reveal different patterns or hidden properties of the data



- Making assumptions about the data
  - It is important to check for complex effects
  - *Are the data multi-modal / skewed?*



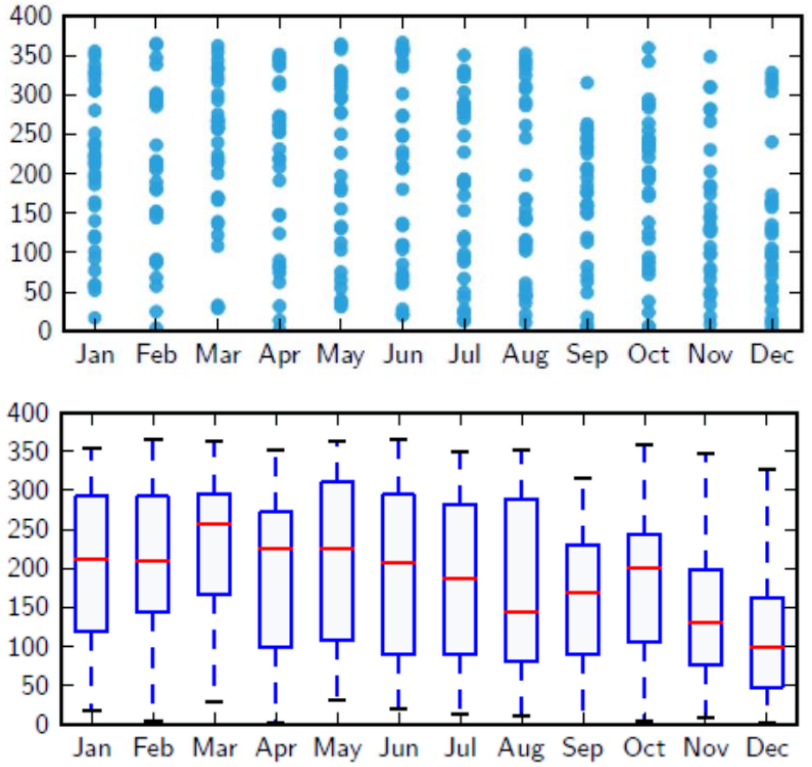
(a) A distribution with two modes. The mean is shown at the blue line.

(b) A right-skewed distribution (positive skew); the tail of the distribution extends to the right.

(c) A left-skewed distribution (negative skew); the tail of the distribution extends to the left.

EXAMPLE: VISUALIZING BIAS

- **THE VIETNAM DRAFT LOTTERY, 1970**



Can you spot a pattern from the plots above?

Why sample selection bias distort the statistical analysis?

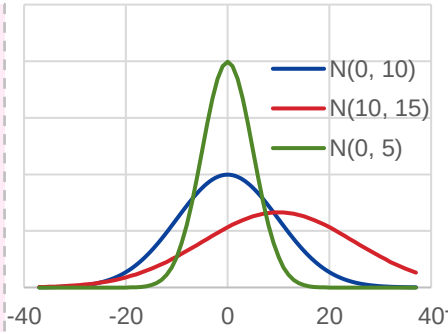
Important Distributions

- Various probabilities

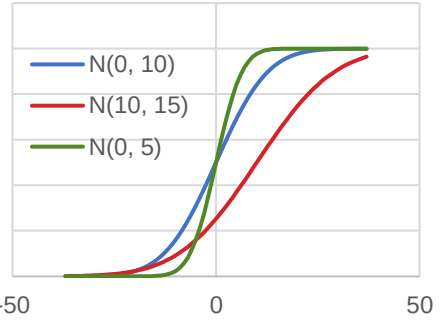
Gaussian
Normal
Student *t*

- **Normal distribution** (mean μ, variance σ^2)
  - $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $where$ $x \sim N(\mu, \sigma^2)$
  - If y $= \frac{x-\mu}{\sigma}$, y is standardized version of x. $y \sim N(0,1)$
- **Student *t* distribution** (degree of freedom $v = n - 1$)
  - Used to assessing statistical significance

PDF of normal distribution



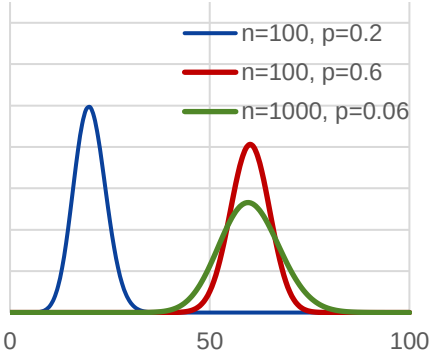
CDF of normal distribution



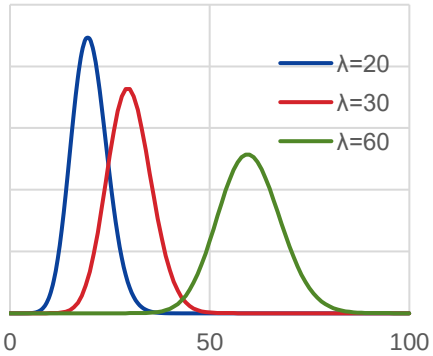
Binomial
Poisson

- **Binomial distribution**
  - sum of n independent **Bernoulli** trials (success or fail)
  - $E(x) = np$, $Var(x) = np(1-p)$
- **Poisson distribution**
  - Average number of (rare) events in a specific time period
  - $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $(x = 0,1,2 \dots)$, $E(x) = Var(x) = \lambda$

Binomial distribution



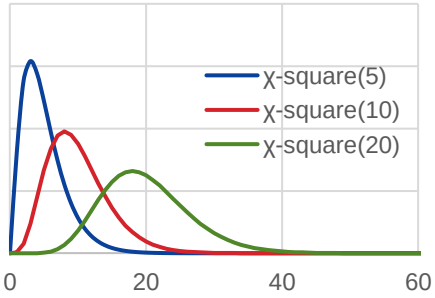
Poisson distribution



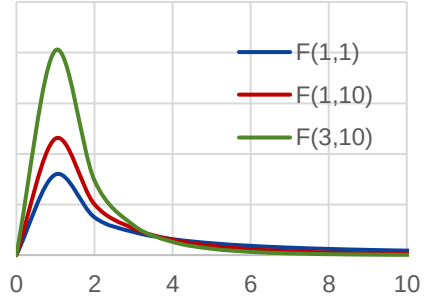
Chi-square
(χ^2)
F-dist

- **Chi-square distribution**
  - Sum of squares of n *iid*. standard normal random variables
    $\chi = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ $where$ $Z_i \sim iid.N(0,1)$
- **F-distribution**
  - Widely used in ANOVA test, F-test.

Chi-square distribution

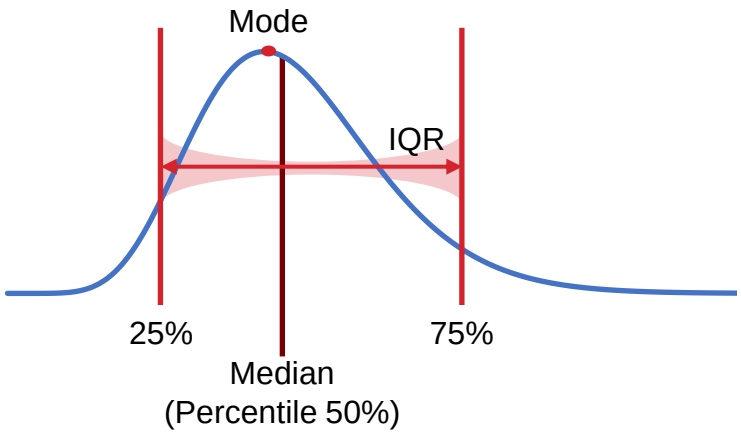


F distribution



Quantitative measures and summary statistics

Quantitative measures and summary statistics



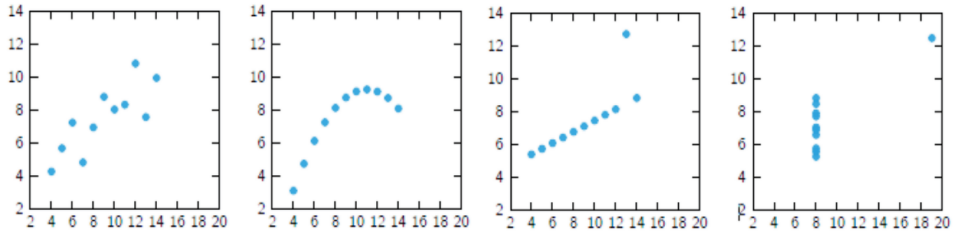
- **Useful ways of numerically summarizing data**
  - **Sample Mean** : $\bar{x} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - **Sample Variance** : $\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - **Median** : the middle value when the data are ordered
  - **Percentiles** : an extension of median to values
  - **Inter-Quartile range (IQR)** : the difference between 75th and 25th percentile
  - **Mode** : the most frequently occurring value
  - **Range** : The minimum and maximum values

EXAMPLE: ANSCOMBE'S QUARTET

- **Summary statistics may not represent true property**
  - Suppose four datasets have same properties below:

$$\bar{x} = 9, \qquad \widehat{\sigma_x^2} = 11$$
$$\bar{y} = 7.50, \qquad \widehat{\sigma_x^2} = 4.12$$
$$corr(x, y) = 0.816$$

  - However, scatterplots show very different datasets:



Source: Anscombe, F. J. (1973). Graphs in statistical analysis. *The american statistician*, *27*(1), 17-21.

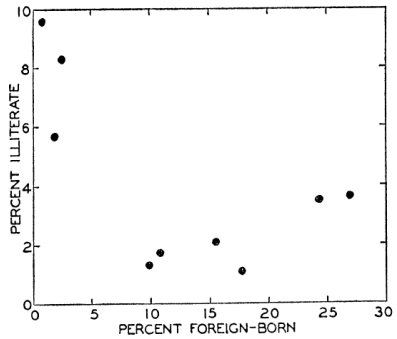
  - Same data can be drawn from the datasets only if:
    - Datasets follow identical distribution
    - Datasets have same 1st and 2nd moment
    - Datasets are scaled identically

Note: In the case of variance, we divede by $n - 1$ in the denominator and not n

Self-selection bias

EXAMPLE – WARNING OF THE DAY: ECOLOGICAL FALLACY

- **Aggregate data cannot always be used to draw conclusions about individual data**
  - In 1950, a statistician named William S. Robinson looked at each one computed the literacy rate and the fraction of immigrants.

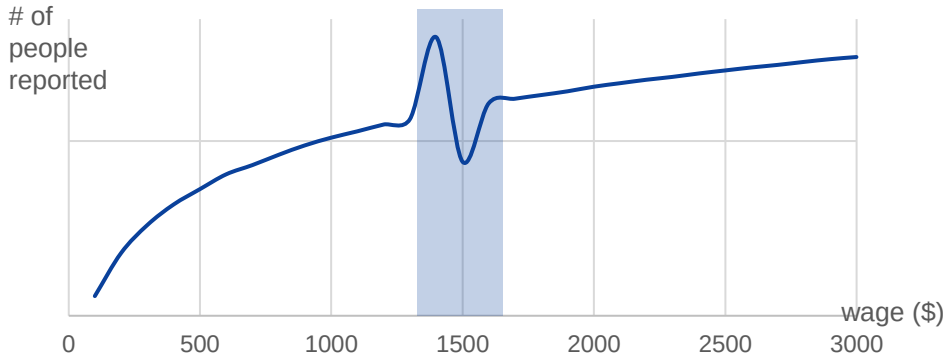


  - From the scatter plot above, we might conclude that immigrants in 1950 were more literate than non-immigrants, but in fact, the opposite was true!

	Foreign Born	Native Born	Total
Illiterate	1304	2614	3918
Literate	11913	81441	93354
Total	13217	84055	97272

  - In fact, immigrants were more likely to settle in states that already had high literacy rates.

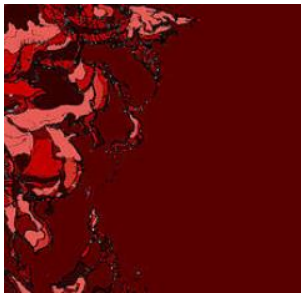
Self-selection bias

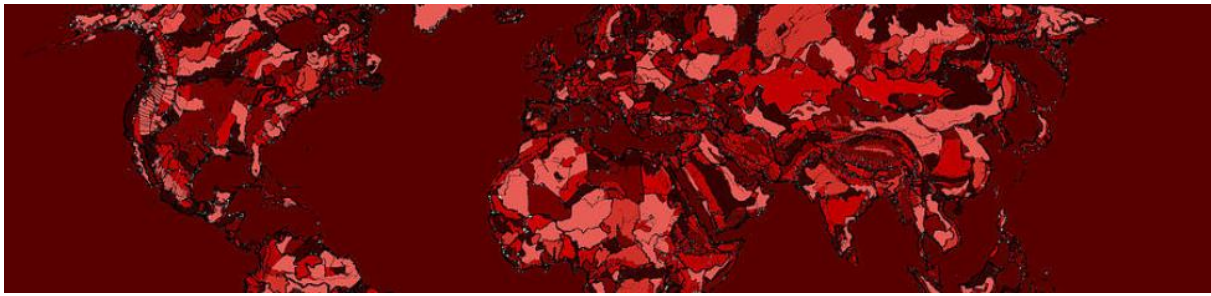


- **Self-selection bias**
  - Dependent variable that has varying level of effects to the outcome, depending on the surrounding conditions, cannot be explained by single variable.
    - e.g) Suppose an agency would like to issue stimulus check to people whose wage is under $1,500 per month.
    - People whose wage is around $1,500 would report their wage is under $1,500 to receive the check.
  - Therefore, the surrounding conditions must be controlled in advance using other variables.
  - If variables to control surrounding conditions are omitted, the model is exposed to fallacy due to unobserved or unidentified factors. **(omitted variable bias)**

2. Confidence intervals and hypothesis tests

Swiss Institute of
Artificial Intelligence





Binomial data

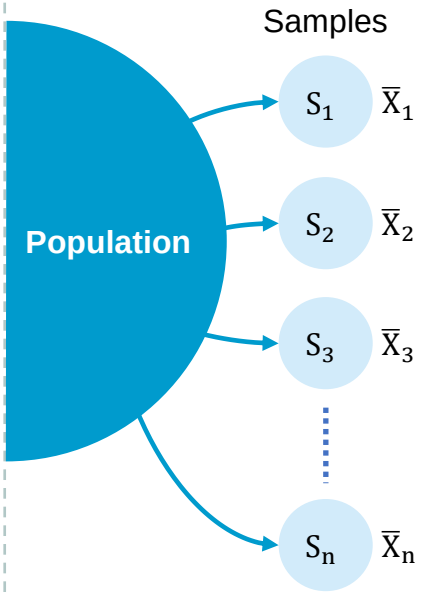
- Central Limit Theorem

Central Limit Theorem

■ **Central Limit Theorem (CLT)**

– If we collect mean of a bunch of independent random variables that all have the same distribution, the result will be approximately Gaussian

$$\text{For } \bar{X}_i \equiv \frac{X_1+X_2+\cdots+X_n}{n},\ E[X_i]=\mu,\ Var[X_i]=\sigma^2<\infty,$$
$$\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{d}N(0,\sigma^2)$$



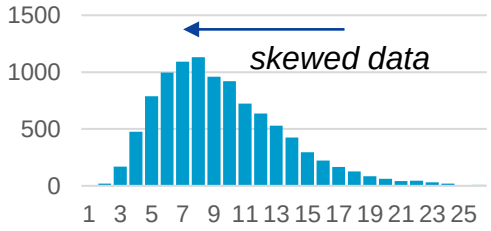
– Whichever the population distribution we have, mean of $\bar{X}_i$ repeatedly sampled from same distribution converges to Gaussian distribution

– Don't be confused with Law of Large Number (LLN) – sample average converges to the population mean when sample size $n \to \infty$

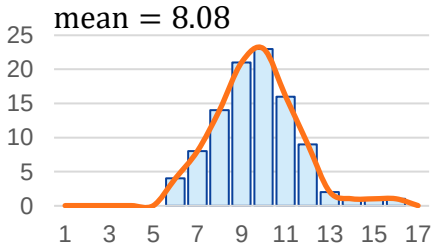
Visualize Central Limit Theorem

■ **Example of central limit theorem**

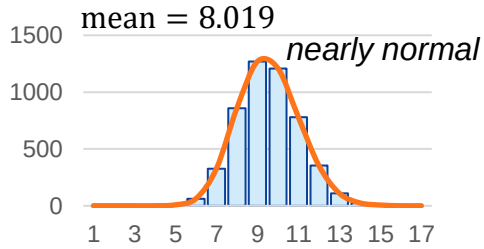
– Population data with $\mu = 8.02$



– Histogram of sampled value (sample size=30)

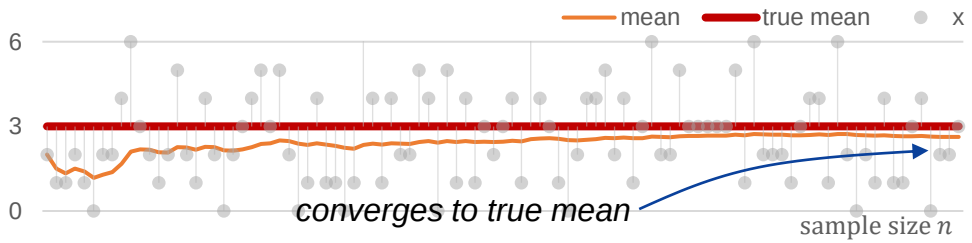


# of samples = 100



# of samples = 5,000

■ **Example of Law of Large Number**



Binomial data (Cont.)

- Binomial distribution approximation to Poisson

Central Limit Theorem

■ **Binomial distribution approximation to Poisson**

– Binomial distribution is the sum of a bunch of independent Bernoulli random variables.

$$P[X=i] = \frac{n!}{(n-i)!\,i!}p^i(1-p)^{n-i}$$

$$= \frac{n!}{(n-i)!\,i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i}$$

$$= \underbrace{\frac{n(n-1)\cdot(n-i+1)}{n^i}}_{=1}\frac{\lambda^i}{i!}\frac{\overbrace{\left(1-\frac{\lambda}{n}\right)^n}^{=e^{-\lambda}}}{\underbrace{\left(1-\frac{\lambda}{n}\right)^i}_{=1}}$$

$$\therefore P[X=i] \approx \frac{e^{-\lambda}\lambda^i}{i!} \sim Poisson(\lambda)$$

■ **Random variable for rare events**

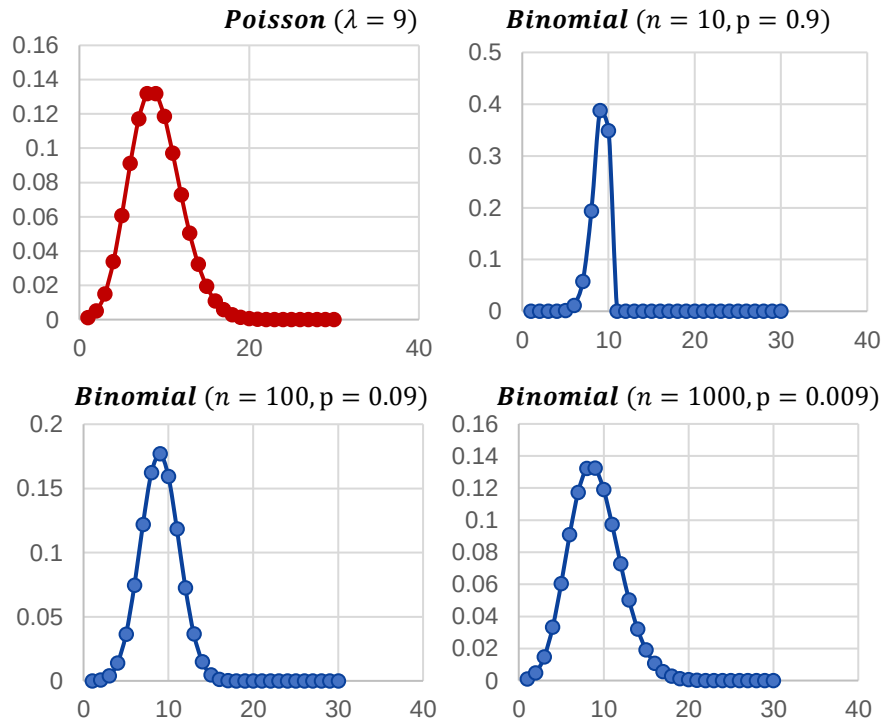
– Rare events such as error rate, a number of clicks on ads usually follow not Gaussian but Poisson distribution.

– If the data does not follow Gaussian, test statistics that are based on normal distribution cannot be used.

Visualize Binomial Distribution Approximation to Poisson

■ **Approximately converge to Poisson distribution**

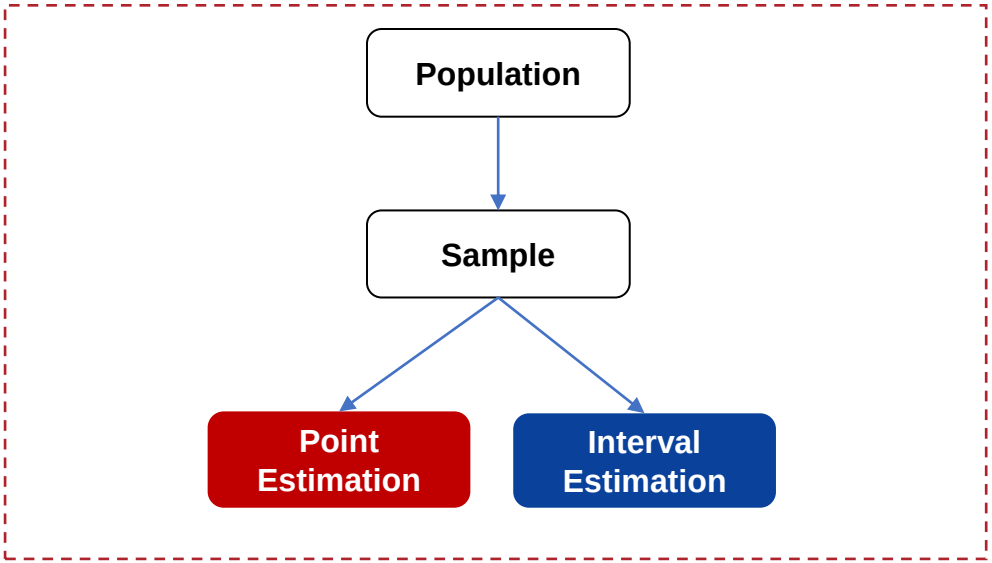
$(np = 9)$

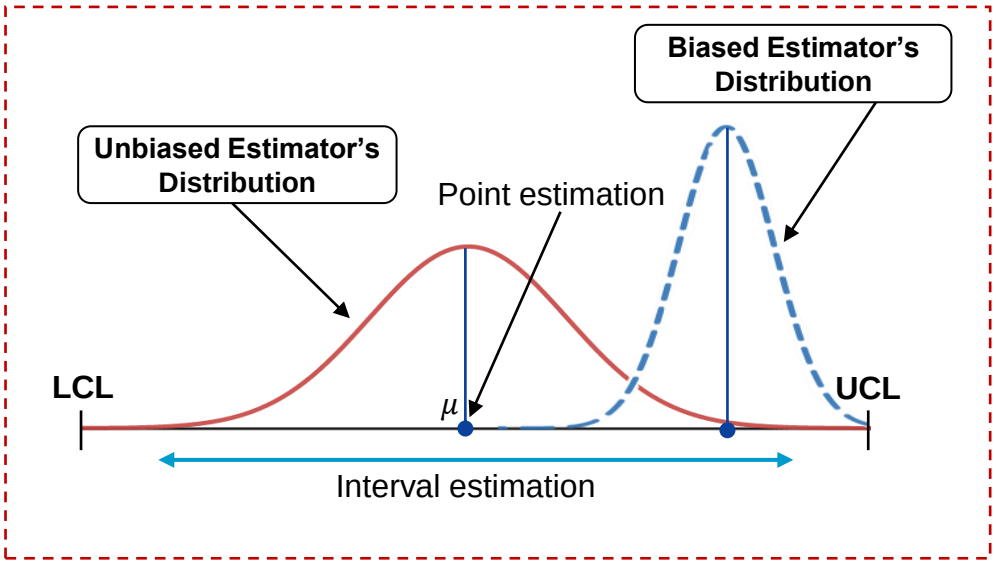


– As sample size increases, binomial distribution converges to Poisson distribution.

Methods of Estimation

- How can we inference the population parameter?





Two types of estimation

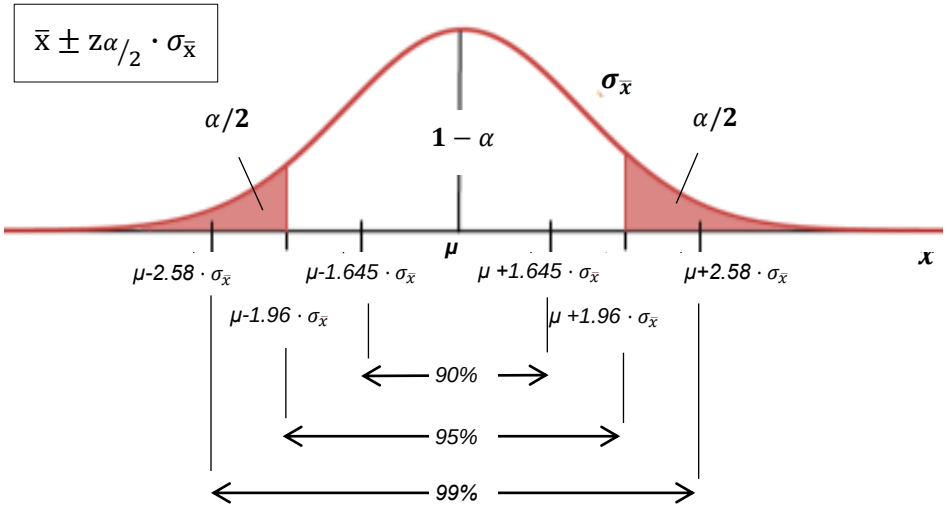
- **Point estimation**
  - It involves the use of sample data to calculate a single value of an unknown population parameter.
- **Interval estimation**
  - We use interval estimation because point estimation with a single value is difficult to represent parameters.

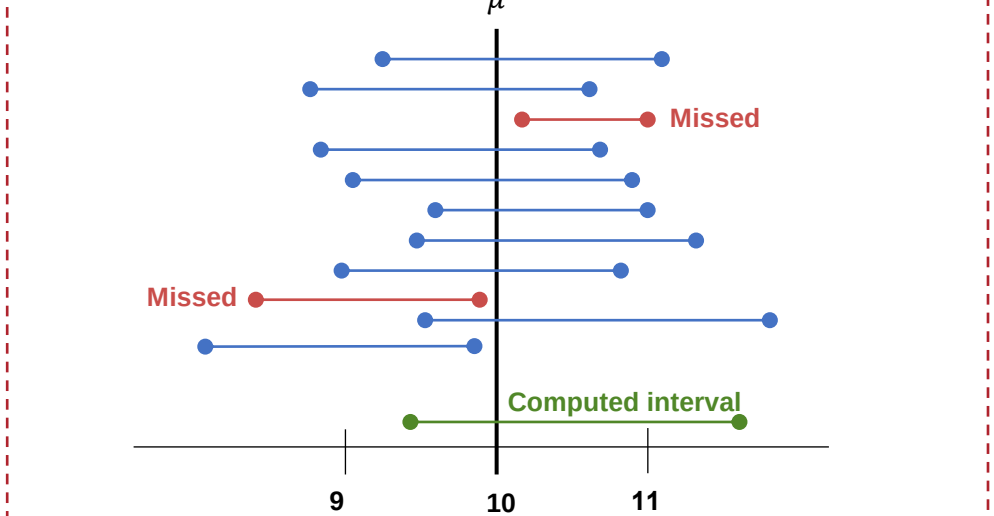
Estimate / Estimator

- **Estimate**
  - It is a specific observed numerical value used to estimate an unknown population parameter.
- **Unbiased estimator**
  - The estimator is equal to the true value within the population ($\hat{p}=p$).

Confidence interval and interpretation

- Confidence interval and correct interpretation





Confidence interval

Interpretation

Confidence interval

- The probability that a population parameter will fall between a set of values.
- It is important to establish an appropriate confidence interval to obtain the necessary information.

Level of confidence($1-\alpha$)

- The percent of confidence intervals (from many samples) that we expect to contain the true population parameter.

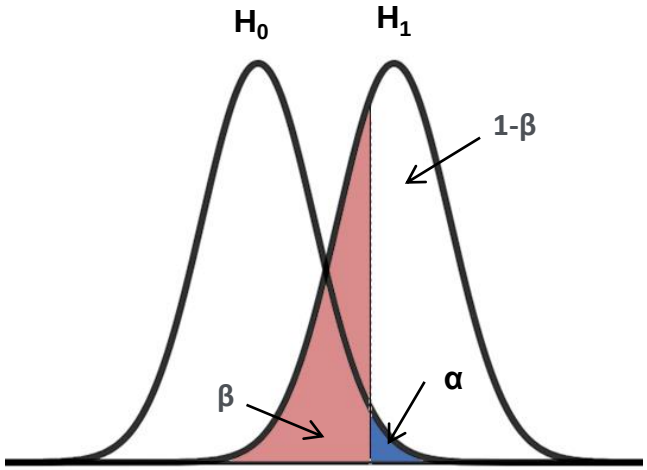
A Correct interpretation of confidence interval

- If we take samples several times and compute the interval each time, then (1-α)% of these intervals will include the actual value of μ – *We may never know which ones.*
- Different samples give different confidence intervals.

Type Ⅰ Error & Type Ⅱ Error and Statistical Power

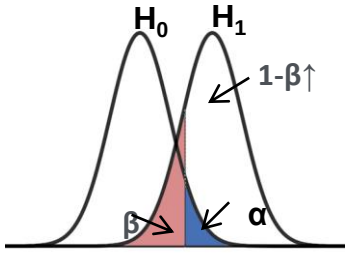
- Hypothesis Error and Increasing the statistical power

Hypothesis and Type Ⅰ Error & Type Ⅱ Error



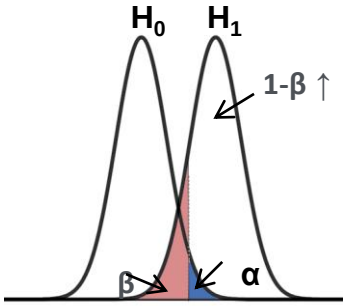
- null hypothesis(H_0) : the true proportion is, in fact p
- alternative hypothesis(H_1) : the true mean is significantly smaller than p
- Significance(α) : the probability of the study rejecting the null hypothesis, given that the null hypothesis was assumed to be true
- a threshold usually balances between Type I and Type II errors

3 Ways To Increase Statistical Power



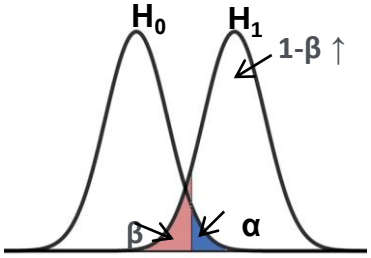
Increasing the alpha level

Alpha and beta values are offset by each other, so as alpha increases, the 1-beta value increases.



Increasing the sample size

As the number of data increases, the variance decreases.



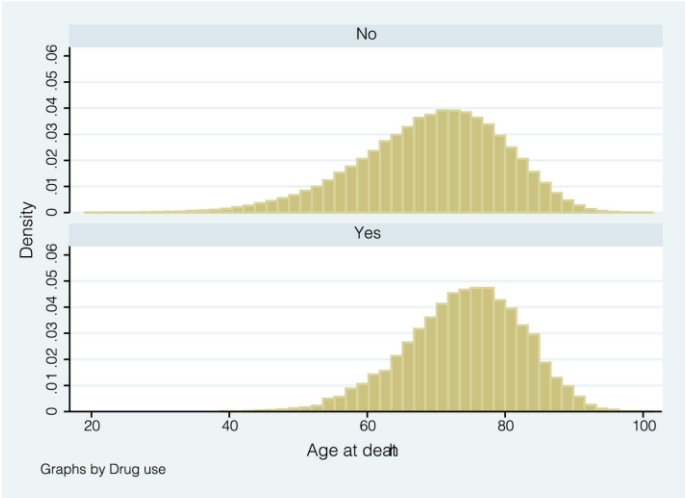
Increasing the Effect size

As the difference between the two populations increases, the statistical power improves.

hypothesis tests and Statistical power

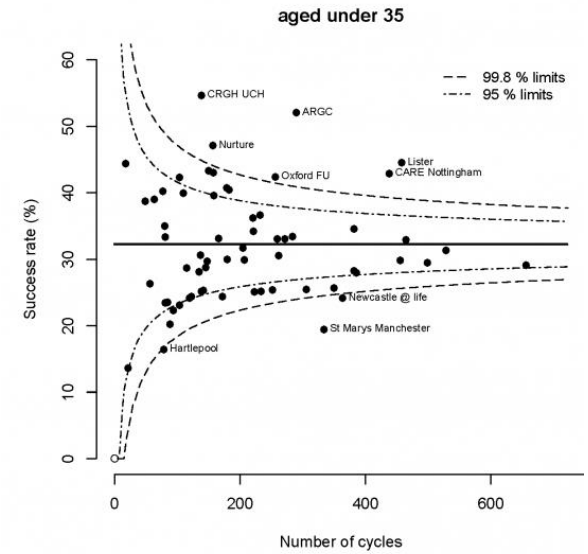
- What should we be careful about when testing hypotheses?

EXAMPLE: DRUG THERAPHY RESULTS: A WARNING ABOUT DATA COLLECTION



- **Importance of data sampling**
  - Results of a simulated drug trial measuring the effects of statin drugs on lifespan. The top guru shows the lifespan of subjects who did not receive treatment, and the bottom figure shows the lifespan of subjects who did receive it.
  - The hypothesis should not be adopted or rejected solely by changes in the mean value.
  - If the distribution function changes, so do the statistical power.
  - If the distribution function does not follow a normal distribution, we must consider various factors.

EXAMPLE: FERTILITY CLINICS

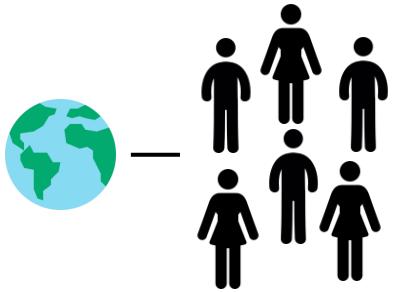


- **Data size and statistical power**
  - A funnel plot showing conception statistics from fertility clinics in the UK.
  - x-axis indicates the sample sizes, y-axis indicates the quantity of interest. The funnels (dashed lines) indicate thresholds for being significantly different from the null value of 32% (the national average).
  - large number of data does not necessarily mean that the power is good

Two-sample Test(A/B Test)

- hypothesis test with two samples

One-sample Test & Two-sample Test



Hypothesis

H_0: $\mu_1 = 0$
H_1: $\mu_1 \neq 0$

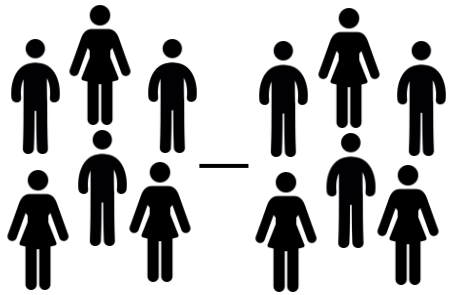
Test statistic

$$z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \quad (\sigma \text{ is known})$$

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \quad (\sigma \text{ is unknown})$$

■ **One-Sample Test**

– Is there a difference between a group and the population.



Hypothesis

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

Test statistic

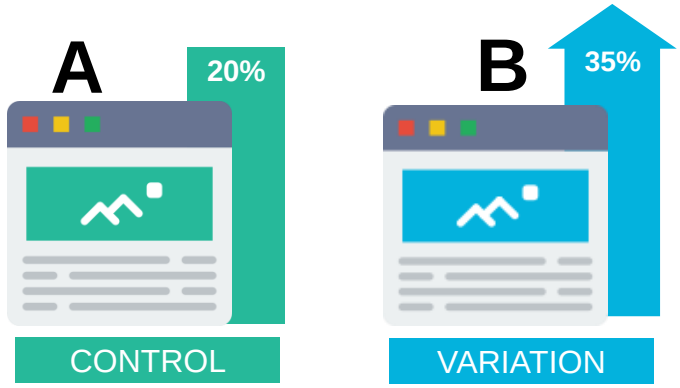
$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

$$t = \frac{\hat{\mu}^{(1)} - \hat{\mu}^{(2)}}{s_p\sqrt{(1/n_1) + (1/n_2)}}$$

■ **Two-Sample Test**

– Is there a difference between two groups.

A/B Test



$$t = \frac{\hat{\mu}^{(1)} - \hat{\mu}^{(2)}}{s_p\sqrt{(1/n_1) + (1/n_2)}} \quad \hat{\mu}^{(1)} > \hat{\mu}^{(2)} \rightarrow$$

n1	n2	t-statistic
1/500	1/500	1.5
1/100	1/900	1.3
1/900	1/100	1.7

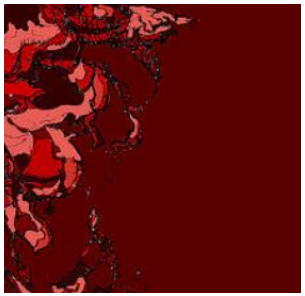
■ **A/B TEST**

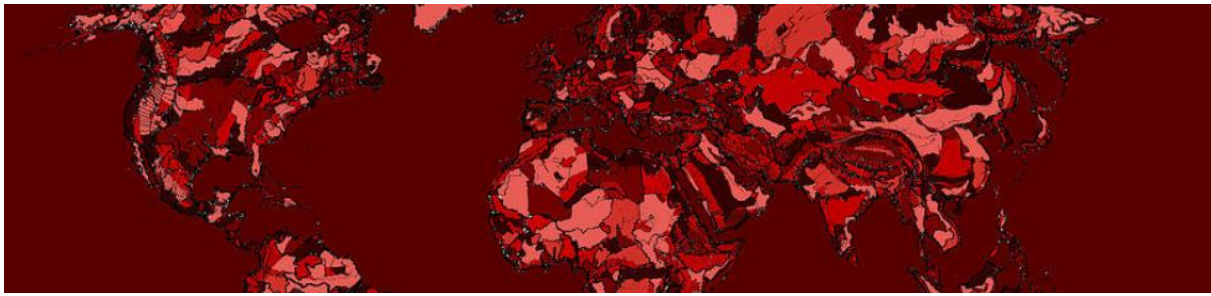
– The t statistic value varies depending on the size and weight of the data value of each sample.

– It is difficult to explain multiple variables because of the use of static data.

3. Linear regression

Swiss Institute of
Artificial Intelligence





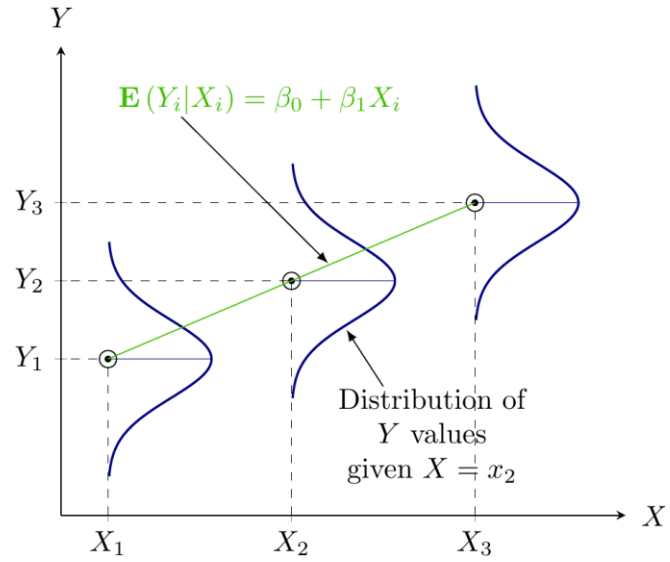
Simple linear regression

- Simple, but essential to develop intuition for multiple linear regression

Simple linear regression

Simple linear regression

- Linear regression captures the changing average of response(y) according to the change of predictors(x)
- Linear regression assumes variance of predictors can explain variance of response
- β_0 controls the height, embodying anchoring and can be removed by demeaning
  - $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow y_i - \bar{y} = \widetilde{\beta_1}(x_i - \bar{x}) + v_i$
- β_1 captures the correlation between response and predictors

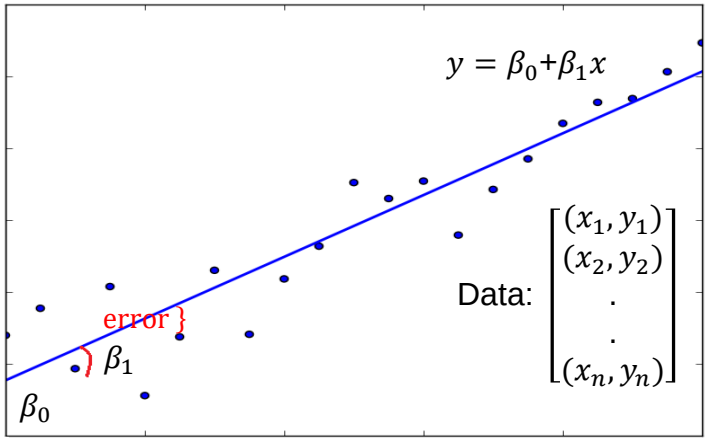


Ordinary least square estimators

Ordinary Least square estimators (scalar version)

$\boldsymbol{Goal}$: $\min_{\boldsymbol{\beta_0},\boldsymbol{\beta_1}} \sum_{i=1}^{n} [\boldsymbol{y_i} - (\boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_i})]^2$, where $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i) \sim N(0, \sigma^2)$

- There are n data points
- Assume $x_1, x_2, \ldots, x_n$ are fixed
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\boldsymbol{Cov(x,y)}}{\boldsymbol{var(x)}}$, $E(\hat{\beta}_1) = \beta_1$, $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $E(\hat{\beta}_0) = \beta_0$, $Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$
- Expectation/variance of $\hat{\beta}_0$ & $\hat{\beta}_1$ can be used to do hypothesis tests and compute confidence/prediction intervals



Tests and intervals in simple linear regression

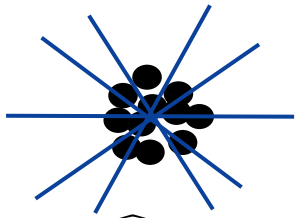
- Implement hypothesis tests and compute confidence/prediction intervals

t-statistic

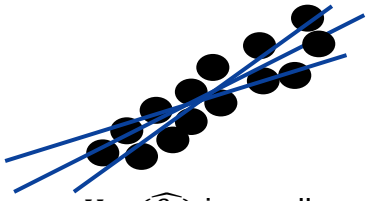
t-statistic for the slope

$$t_{\beta_1}=\frac{\widehat{\beta_1}-\beta_1}{s_{\beta_1}} \sim t(n-2)\,, \quad s_{\beta_1}=\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

- If the x-values are all really close together, or x has small swing, $Var(\hat{\beta}_1)$ becomes larger
- Increase of the number of data points n can decrease $Var(\hat{\beta}_1)$
- Larger $Var(\hat{\beta}_1)$ makes t-statistic of $\hat{\beta}_1$ lower and increases the possibility of not rejecting the null hypothesis, H_0



$Var(\widehat{\beta_1})$ is large



$Var(\widehat{\beta_1})$ is small

t-statistic for the intercept

$$t_{\beta_0}=\frac{\widehat{\beta_0}-\beta_0}{s_{\beta_0}} \sim t(n-2), \quad s_{\beta_0}=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Confidence/Prediction interval

Confidence intervals for β_0 and β_1

$$\widehat{\beta_0} \pm t_{\alpha/2}(n-2) * s_{\beta_0}, \quad \widehat{\beta_1} \pm t_{\alpha/2}(n-2) * s_{\beta_1}$$

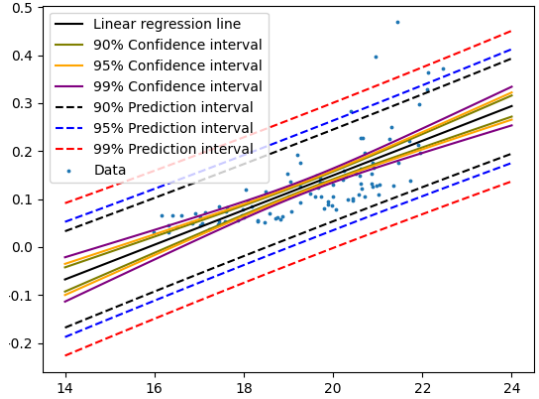
Confidence interval for the mean of out-of sample response

$$\widehat{\mu_{x^*}} \pm t_{\alpha/2}(n-2) * \hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

- μ_{x^*} is the expected value of $y(x^*)$

Prediction interval for the out-of sample response

$$\hat{y}(x^*) \pm t_{\alpha/2}(n-2) * \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$



- Prediction interval is wider than confidence interval at the same $(1-\alpha)\%$ significance level

Correlation vs Causality

- Correlation does not imply causality!

Correlation

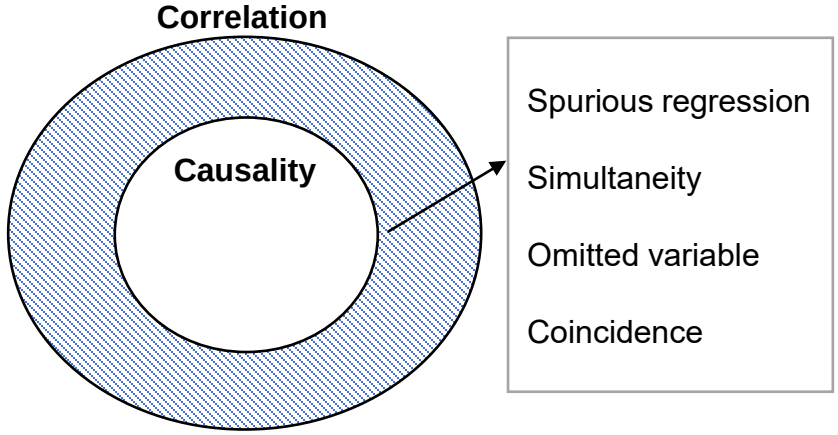
Correlation

- Linear regression just shows the correlation of response(y) & predictors(x)
- $\hat{\beta}_1 = \frac{Cov(x,y)}{var(x)}$ captures the relationship even though we can measure the statistic such as changing average
- However, linear regression tells us the *correlation*, not the causality
- Just choosing predictors which have high correlation with response is not reasonable

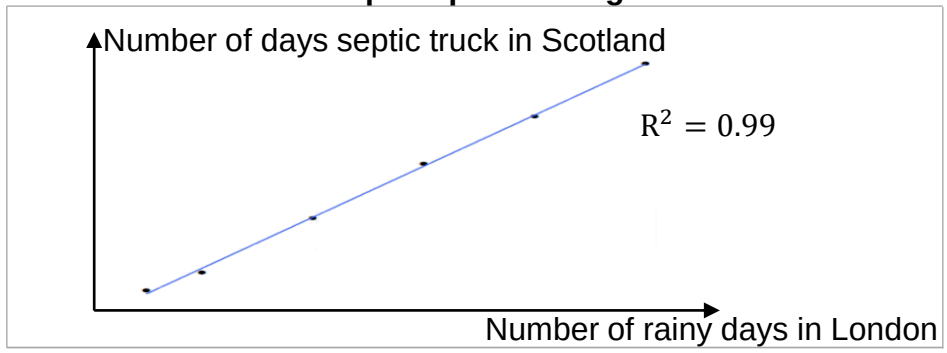
Causality

Causality

- Causality is applied to cases where action A causes outcome B
- Correlation can be mistranslated if spurious regression, simultaneity, omitted variable, coincidence exist
  – Example: number of days septic truck emerges in Scotland and number of rainy days in London
- We cannot prove that the causality exists and humankind should judge whether the causality exists
- However, whether the causality does not exist can be proved by Granger causality



Example- spurious regression



- Emerge of septic truck and rainy days are everyday events

Interpolation vs extrapolation

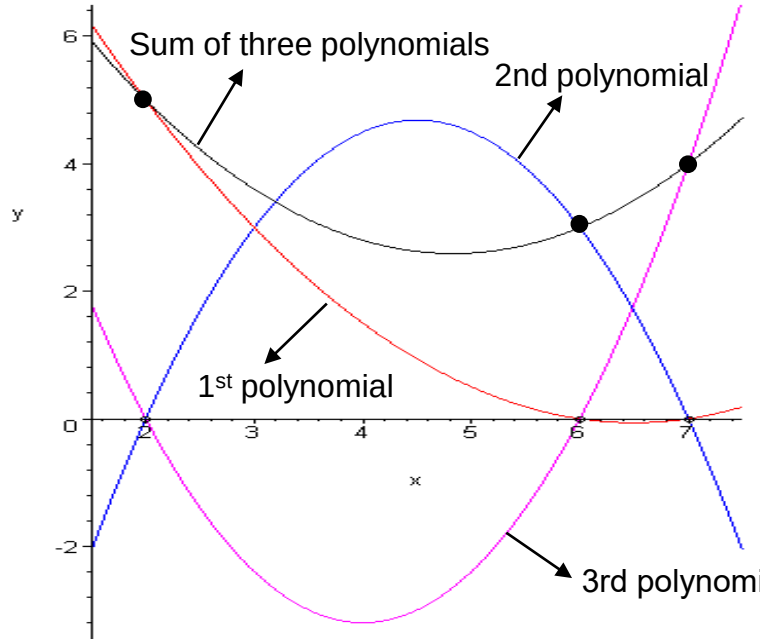
- How to make non-linearity?

Creation of non-linearity

Non-linearity

- The most common way to create non-linearity is to use polynomials, where $(n-1)$th order polynomial that passes through n data points is solved
- $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1}$
  - $x = (x_1, x_2, \ldots, x_n)^T$ is known, while a_0, a_1,…a_{n-1} are unknown
- $$\begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & \ldots & x_2^{n-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n-1} & x_{n-1}^2 & \ldots & \ldots & x_{n-1}^{n-1} \\ 1 & x_n & x_n^2 & \ldots & \ldots & x_n^{n-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ . \\ . \\ . \\ a_{n-1} \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ . \\ . \\ . \\ f(x_n) \end{bmatrix}$$
  - Polynomial interpolation problem can be shown as above
- $f_1(x) = \frac{x-x_1}{x_0-x_1} f(x_0) + \frac{x-x_0}{x_1-x_0} f(x_1)$ for the first-order version
- $f_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$ for the second-order version (shown right)

Example- creation of second-order polynomial



- Goal: create black line which connects three black points using polynomials
- The black line is the sum of three terms, 1st polynomial(red line), 2nd polynomial(blue line) and 3rd polynomial(pink line)
- As the number of terms added increases, higher order polynomials emerge and total error becomes smaller

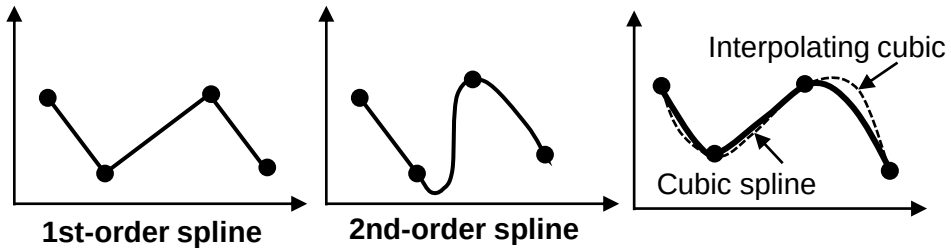
Interpolation vs extrapolation

- Examples of interpolation/extrapolation

Interpolation / Extrapolation

Interpolation

- Prediction of y for a value of x which is within the interval of points that are observed in the original data
  – Example: estimation of interest rate
- We should reasonably determine to use linear/non-linear model
  – Example: Estimating the orbit of Halley's Comet
- Spline interpolation fits low-degree polynomials to small subsets of the values instead of fitting a single, high-order polynomial

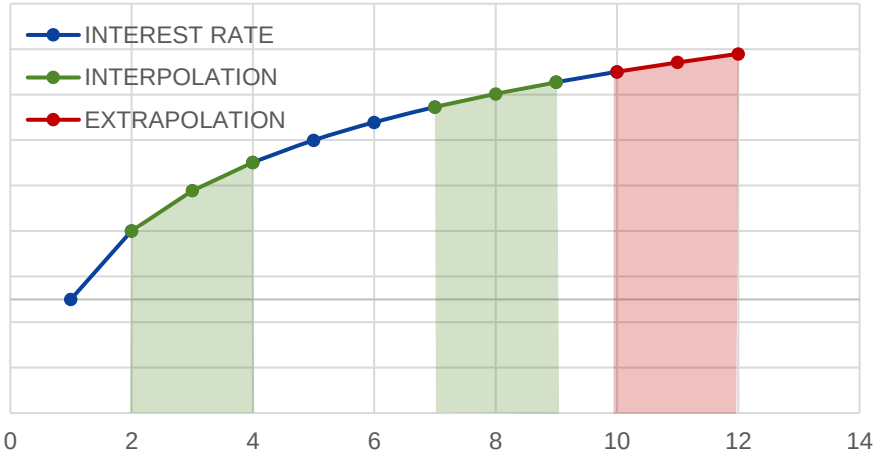


Extrapolation

- Prediction of y for a value of x which is outside the range of values that are observed in the original data
  – Example: 12 years interest rate with the same data above

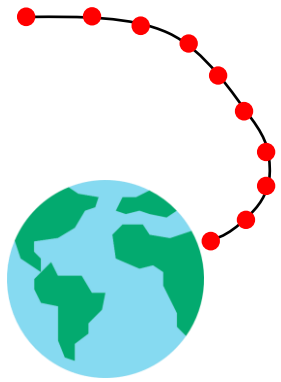
Example- inter/extrapolation of interest rate, Halley's Comet

Estimation of interest rate



Orbit of Halley's Comet

- Fitting with linear model would not be a reasonable choice to compute the orbit of Halley's Comet
- Comet tends to move while following a curve, not a linear line
- We should think about which function to use in order to make non-linear fit



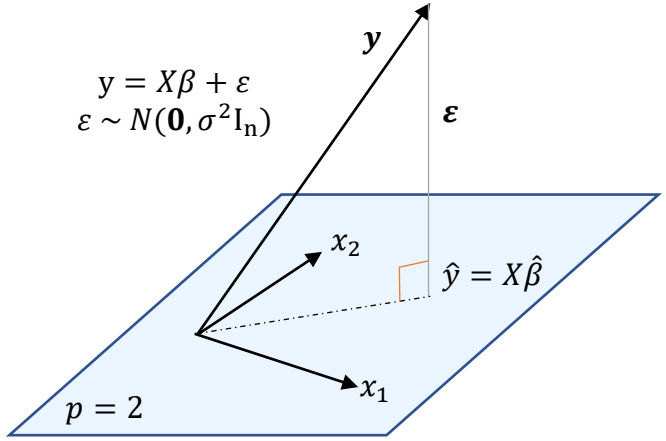
Multiple linear regression

- What if more than 2 predictor variables exist?

Multiple linear regression

Multiple linear regression

- Multiple linear regression is an extension of simple linear regression
- $\mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{i1} & x_{i2} & \dots & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_i \\ \dots \\ \beta_p \end{bmatrix}, \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_i \\ \dots \\ \varepsilon_n \end{bmatrix}$
- In matrix perspective, multiple linear regression is trying to find the the projection of response($\hat{y}$) into the vector space of predictors(X) such that $\hat{y}$ is orthogonal to ε



OLS/ Rank deficiency

Ordinary Least square estimators (matrix version)

$$\boldsymbol{Goal}: \ \boldsymbol{\min_{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|^{\boldsymbol{2}}, \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathrm{I_n})$$

- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-\boldsymbol{1}}\boldsymbol{X'y}, \quad E(\hat{\beta}) = \beta, \quad Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

Rank deficiency

- Inputting same data due to the lack of data can cause rank deficiency
- In this case, X has not full rank and OLS estimator cannot be obtained
- The number of equations become less than the number of variables

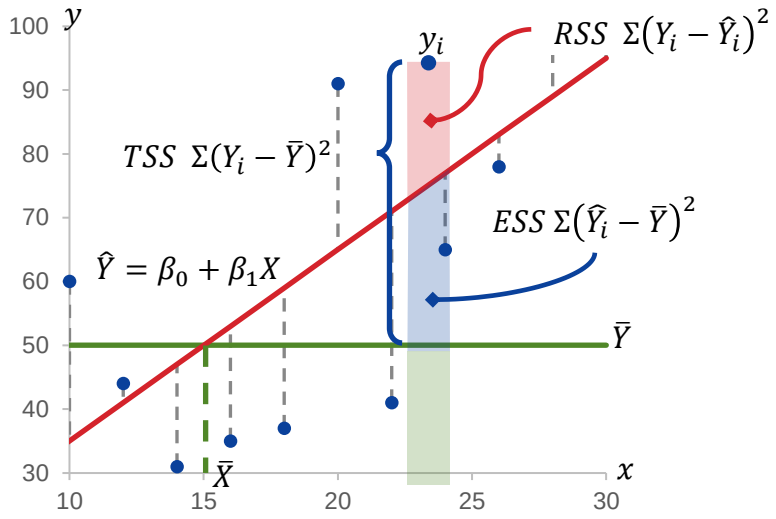
$$\mathbf{X} = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 3 & 5 \\ 1 & 3 & 5 \\ 2 & 6 & 3 \\ 2 & 7 & 6 \end{bmatrix}$$

$\text{rank}(\mathbf{X}) = 3$

Analysis of variance and model evaluation

- Usage of 2nd moment to test performance of the model

Analysis of variance



Analysis of variance

- $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$
  - $\hat{y}_i - \bar{y}$ is the difference explained by model, $y_i - \hat{y}_i$ is residual
- $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $(TSS = ESS + RSS)$
  - Now, we can focus on variance, which is 2nd moment
- Why 2nd moment is important?
  - When one data explains other data, the key is swing (variance)
  - Generally, data with large swing tend to explain the other data well

Model evaluation

Coefficient of determination (R^2)

- Coefficient of determination is interpreted as the fraction of variability in the data explained by model

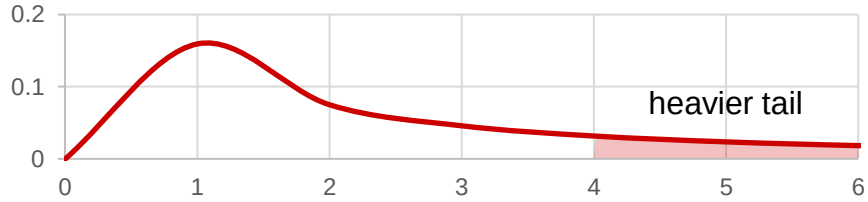
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}, \qquad 0 \le R^2 \le 1$$

F-test

- F-test use variance to measure performance of the model

$$F = \frac{ESS/p}{RSS/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)} \sim F(p, n-p-1)$$

- High ESS and low RSS means the model can fit well to data
- Note that $\frac{ESS}{\sigma^2} \sim \chi^2(p)$ and $\frac{RSS}{\sigma^2} \sim \chi^2(n-p-1)$
- As the explanatory power gets better (R^2 becomes larger), F-statistic becomes higher
  - Since F-distribution has one-sided thick tail than the normal/student *t* distribution, F-test can increase the power of outliers to the maximum



Appendix

- Estimation of beta

Find the beta

$y_t = \beta_1 + \beta_2 x_t + \epsilon_t, t = 1, \ldots, T.$
The transpose of a matrix X, denoted by X'

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix},\ i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},\ x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{bmatrix},\ X = \begin{bmatrix} i & x \end{bmatrix},\ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},\ \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}$$

$$Y = X\beta + \epsilon$$
$$X'Y = X'(X\beta + \epsilon) = X'X\beta + X'\epsilon$$
$$(X'X)^{-1}X'Y = (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'\epsilon$$
$$(X'X)^{-1}(X'X) = I,\ E((X'X)^{-1}X'\epsilon) = 0$$
$$\beta = (X'X)^{-1}X'Y$$

$$X'X = \begin{bmatrix} i' \\ x^T \end{bmatrix} \begin{bmatrix} i & x \end{bmatrix} = \begin{bmatrix} i'i & i'x \\ x'i & x'x \end{bmatrix} = \begin{bmatrix} T & \sum_t x_t \\ \sum_t x_t & \sum_t x_t^2 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} \sum_t x_t^2 & -\sum_t x_t \\ -\sum_t x_t & T \end{bmatrix} / \triangle$$

where $\triangle = |X'X| =$
$$= T\sum_t x_t^2 - (\sum_t x_t)^2 = T\sum_t x_t^2 - T^2(\tfrac{1}{T}\sum_t x_t)^2$$
$$= T\sum_t x_t^2 - T^2\bar{x}^2 = T(\sum_t x_t^2 - T\bar{x}^2) = T(\sum_t x_t^2 - 2T\bar{x}^2 + \bar{x}^2)$$
$$= T\sum_t (x_t - \bar{x})^2.$$

Find the beta

$$X'Y = \begin{bmatrix} i' \\ x' \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} \sum_t y_t \\ \sum_t x_t y_t \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \tfrac{1}{\triangle} \begin{bmatrix} \sum_t x_t^2 & -\sum_t x_t \\ -\sum_t x_t & T \end{bmatrix} \begin{bmatrix} \sum_t y_t \\ \sum_t x_t y_t \end{bmatrix} = \tfrac{1}{\triangle} \begin{bmatrix} \sum_t x_t^2 \sum_t y_t - \sum_t x_t \sum_t x_t y_t \\ -\sum_t x_t \sum_t y_t + T\sum_t x_t y_t \end{bmatrix}$$

$$\beta_1 = \frac{\sum_t x_t^2 \sum_t y_t - \sum_t x_t \sum_t x_t y_t}{T\sum_t (x_t - \bar{x})^2}$$
$$= \frac{\bar{y}\sum_t x_t^2 - \bar{x}\sum_t x_t y_t}{\sum_t (x_t - \bar{x})^2}$$
$$= \frac{\bar{y}\sum_t (x_t - x_t^2) + T\bar{x}^2\bar{y} - \bar{x}\sum_t x_t y_t}{\sum_t (x_t - \bar{x})^2}$$
$$= \bar{y} - \frac{\bar{x}\sum_t (x_t - \bar{x}) y_t}{\sum_t (x_t - \bar{x})^2}$$

$$\beta_2 = \frac{-\sum_t x_t \sum_t y_t + T\sum_t x_t y_t}{T\sum_t (x_t - \bar{x})^2}$$
$$= \frac{-\bar{x}\sum_t y_t + \sum_t x_t y_t}{\sum_t (x_t - \bar{x})^2}$$
$$= \frac{\sum_t (x_t - \bar{x}) y_t}{\sum_t (x_t - \bar{x})^2}$$

Appendix

- Estimation of variance of beta

Find the Var(beta)

Derive the beta variance-covariance matrix.

$$
\begin{aligned}
Cov(\hat{\beta}) &= E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] && \text{(by definition)} \\
&= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] && (\because E(\hat{\beta}) = \beta) \\
&= E[(X'X)^{-1}(X')e)((X'X)^{-1}(X')e)'] && (\because \hat{\beta} = \beta + (X'X)^{-1}(X')e) \\
&= E[(X'X)^{-1}(X')e)(e'X)(X'X)^{-1})] && (\because (AB)' = B'A', (X'X)^{-1\prime} = (X'X)^{-1}) \\
&= E[(X'X)^{-1}(X')(ee')X(X'X)^{-1})] \\
&= (X'X)^{-1}(X')E[ee']X(X'X)^{-1} \\
&= (X'X)^{-1}(X')\sigma^2 I_n X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}
$$

$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

$$
Cov(\hat{\beta}) = \begin{bmatrix}
Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & \dots & Cov(\hat{\beta}_1, \hat{\beta}_k) \\
Cov(\hat{\beta}_2, \hat{\beta}_1) & Var(\hat{\beta}_2) & \dots & Cov(\hat{\beta}_2, \hat{\beta}_k) \\
\vdots & \vdots & \vdots & \vdots \\
Cov(\hat{\beta}_k, \hat{\beta}_1) & Cov(\hat{\beta}_k, \hat{\beta}_2) & \dots & Var(\hat{\beta}_k)
\end{bmatrix}
$$

$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Find the Var(beta)

Express $(X'X)^{-1}$ as matrix $[a_{ii}]$,

$$
Cov(\hat{\beta}) = \sigma^2 \begin{bmatrix}
a_{11} & a_{12} & \dots & a_{1i} \\
a_{21} & a_{22} & \dots & a_{2i} \\
\vdots & \vdots & \vdots & \vdots \\
a_{i1} & a_{i2} & \dots & a_{ii}
\end{bmatrix}
$$

Eventually, we knew the expected value and covariance of the random variable beta. Therefore, we can express the distribution of the probability variable beta as follows.

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2_{ii})$$