ECS7024 Statistics for Artificial Intelligence and Data Science

Course Review

William Marsh

Main Themes of the Module

- Categorical and continuous variables
  - Distributions: bar charts and histograms
- Probability
  - Laws, Joint and Conditional
  - Cross-tabulation (contingency table)
- Probability distributions
  - Binomial
  - Normal: mean and variance
- Correlation
  - Scatter plots
  - Mutual information
- Sampling
- Regression
  - Continuous and logistic
  - Correlation and causes
- Hypothesis tests
  - T-test, chi-squared
  - CI and p-values
- Bootstrap
- Time series
- Bayesian statistics
  - Likelihood and prior
- Linear algebra

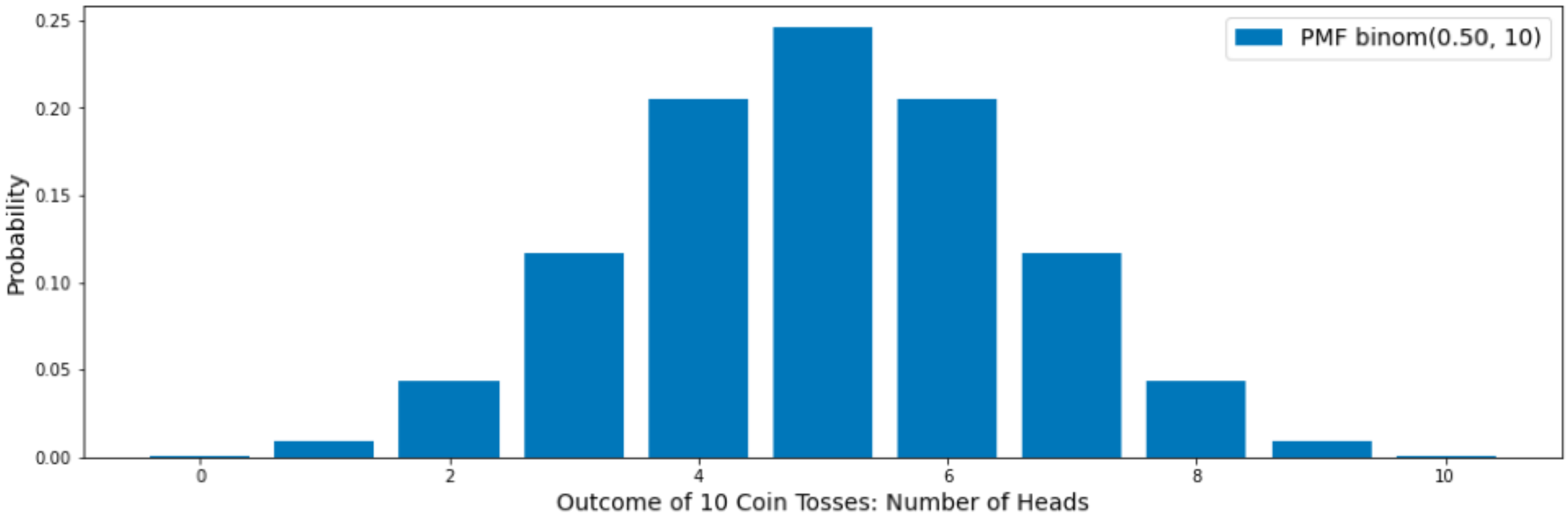
Sampling and Statistical Tests

Devices for Generating Random Events

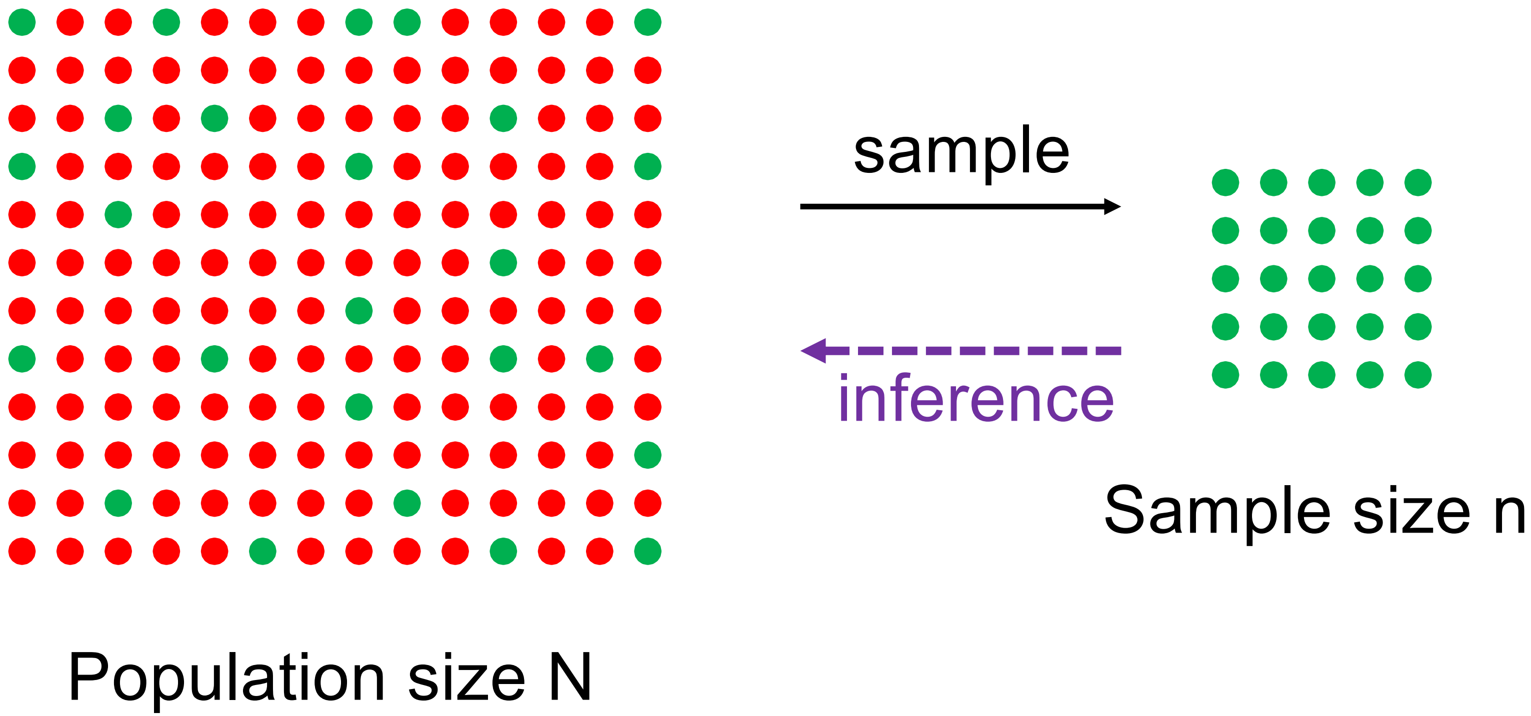
Two sequences 10 (generated with a real coin):

- H, H, T, T, H, T, T, T, H, H (5H, 5T)
- T, T, T, H, H, T, T, T, H, H (4H, 6T)

Sampled from

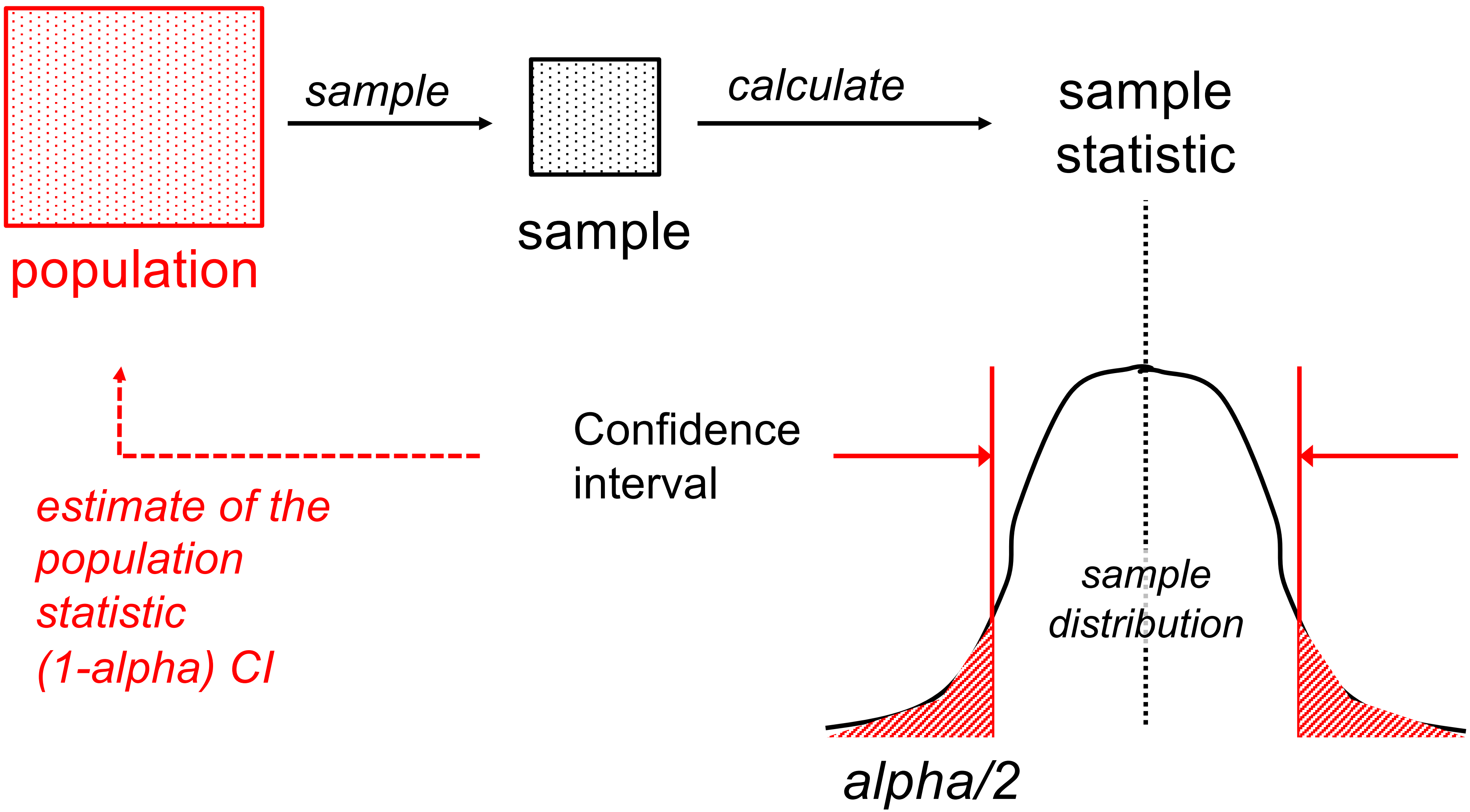


Population and Sample



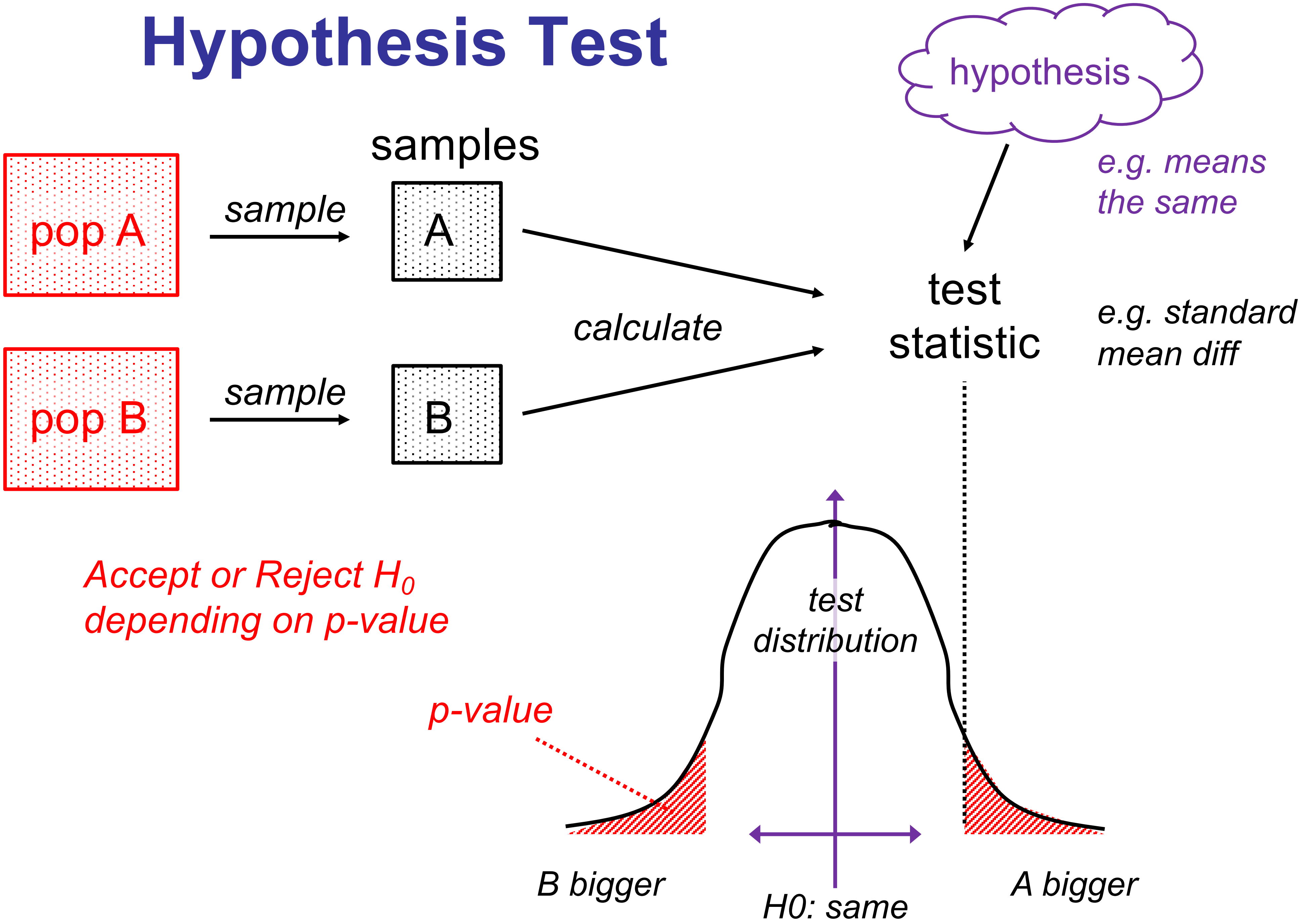
- Sample from a population
- Measure the sample (e.g. political preference)
- Statistical inference about population

Confidence Intervals



alpha is the significance threshold – choose it

Hypothesis Test



Accept or Reject H_0 depending on p-value

Some Issues

- You have to know
  - The test statistics
  - The correct distribution
  - The assumptions
- CIs and p-value can be mis-understood
  - p-value is not the probability you want
- Hypothesis testing does not consider effect size

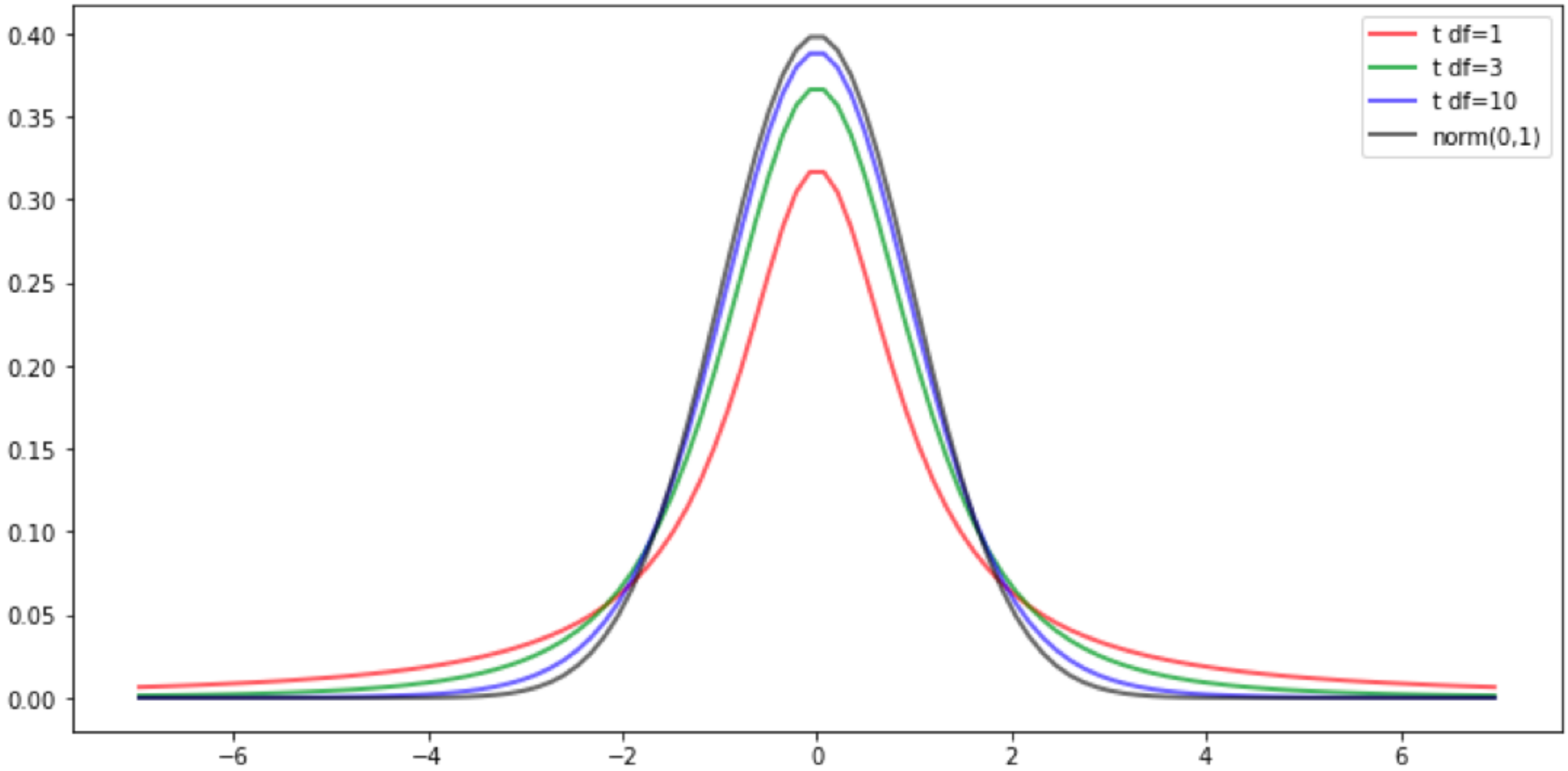
Sampling and Statistical Tests

Student's t-Distribution and Test

Sampling distribution similar to normal,
for use when variance unknown

Student's t-Distribution

- Parameter: 'degrees of freedom' df ≥ 0
  - Shift or scaled with mean and standard deviation
- Normal with 'fat tails'

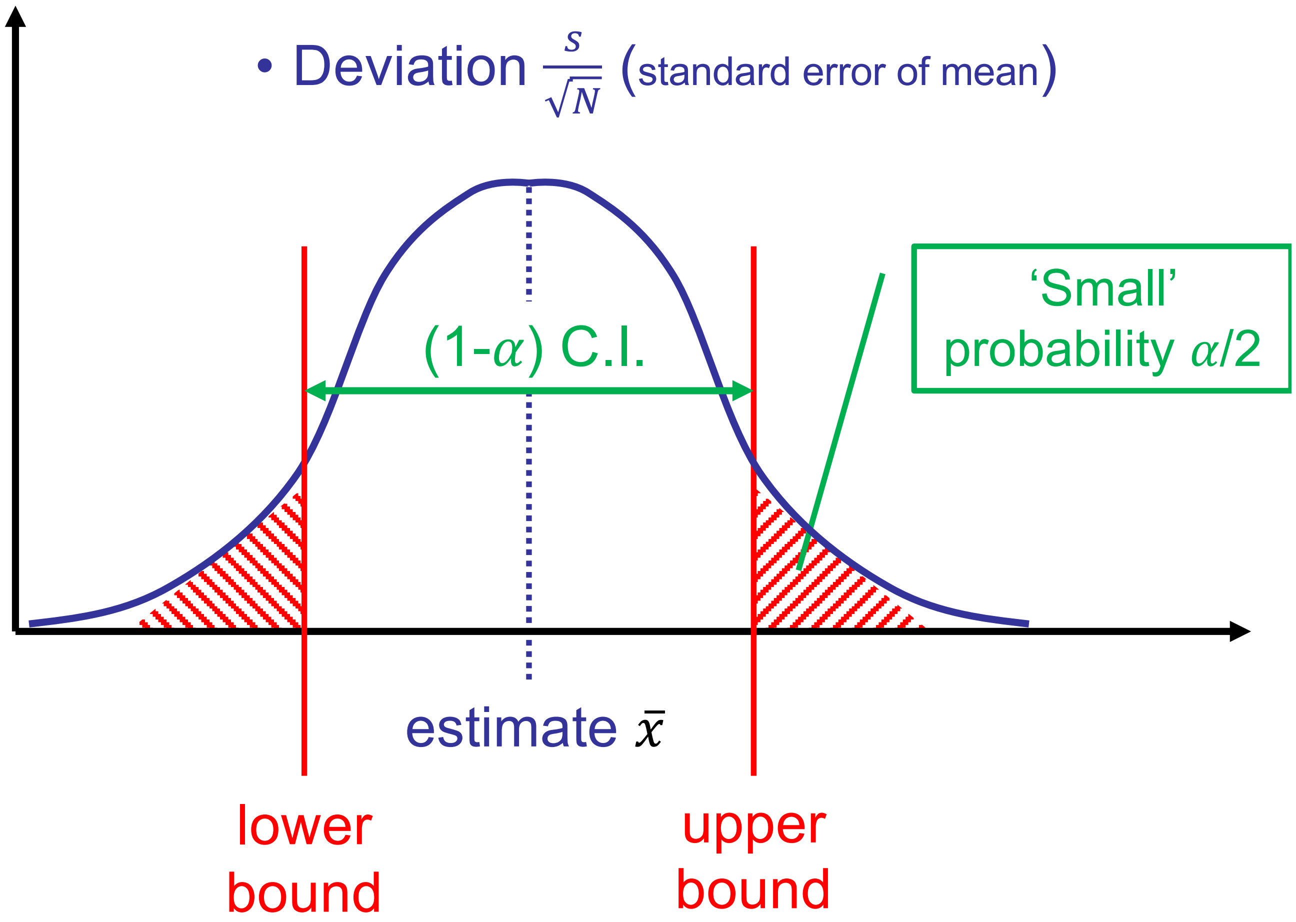


Confidence Intervals for a Mean

- Sample statistics
  - N values
  - Mean $\bar{x}$
  - Standard deviation *s*
- If 95% confidence
  - α is 2.5%
- Large sample – can use normal

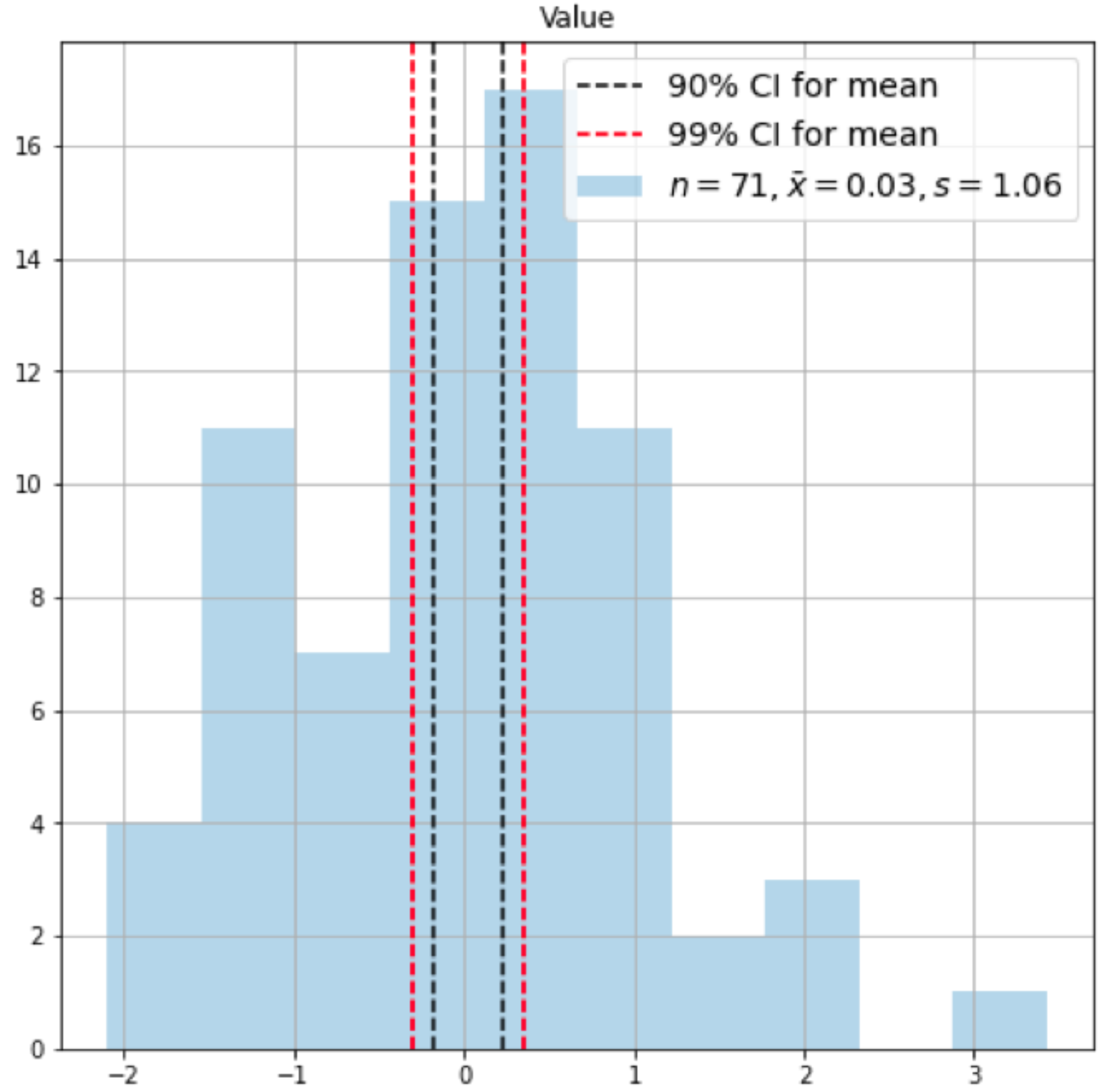
Student's t-distribution

- N-1 degrees of freedom
- Mean $\bar{x}$
- Deviation $\frac{s}{\sqrt{N}}$ (standard error of mean)



Confidence Intervals for a Mean

- Sample of data
  - From a normal
- Sample statistics
  - Mean
  - Standard deviation
- CI from
  - Student's t-distribution
  - Required p-value



Sampling and Statistical Tests

Testing Proportions in a Contingency Table

Test statistics

New distribution - χ^2

Test Statistic

- Observed

	A	B	C	D
White collar	90	60	104	95
Blue collar	30	50	51	20
No collar	30	40	45	35

- Expected
  - Assuming null hypothesis

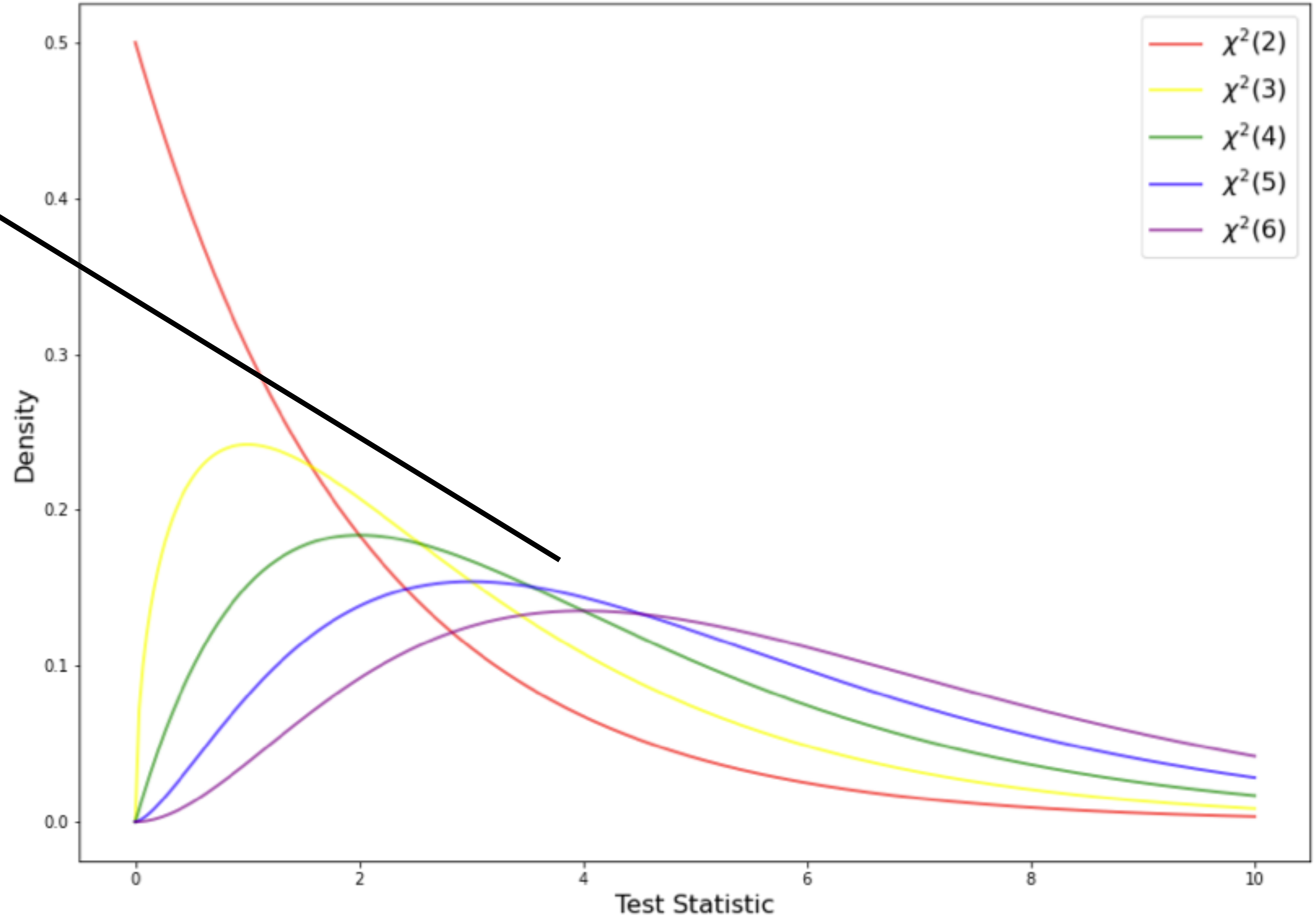
	A	B	C	D
White collar	80.5	80.5	107.4	80.5
Blue collar	34.8	34.8	46.5	34.8
No collar	34.6	34.6	46.2	34.6

$$\sum_{All\ cells} \frac{(Observed\ - Expected)^2}{Expected}$$

Chi-Squared Distribution

- Parameter: degrees of freedom
  - (number of rows – 1) * (number of columns – 1)

More symmetric as dof increases



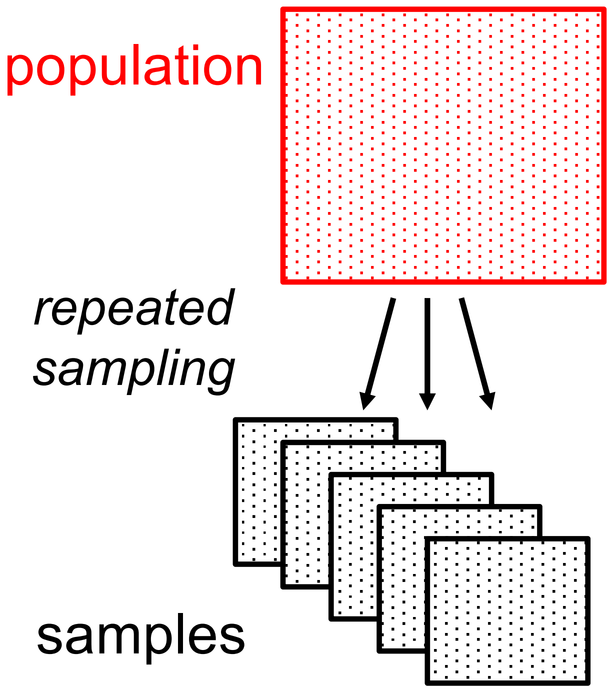
Sampling and Statistical Tests

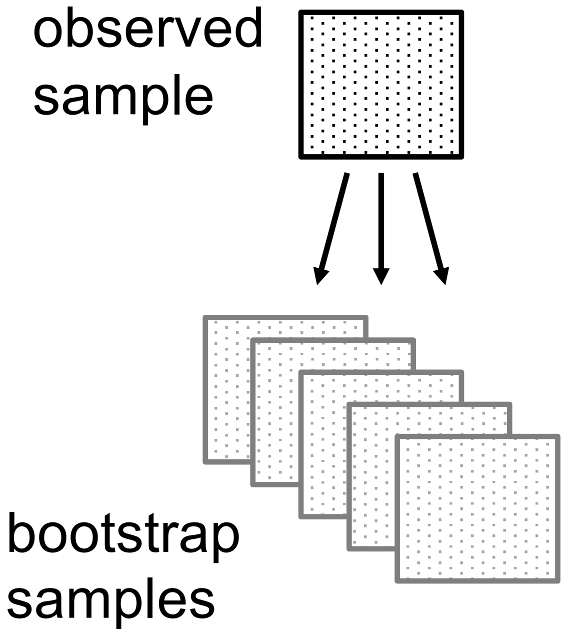
Bootstrap

CI without a Sample Distribution from Theory

Bootstrap

- In a simulation we repeated sample a known population
- In a bootstrap, we resample the sample





Sampling with Replacement

- Re-sample from the ‘observed sample’ with replacement
- Bootstrap sample
  - Same size as original
  - Some records omitted
  - Some records repeated

Related Terms

- ‘Bootstrap aggregation’ or ‘bagging’
- Resampling (with or w/o replacement)
- Permutation test

Bootstrap Steps

1. Resample from the sample
2. Calculate the statistic (e.g. mean) of interest for each new sample
3. Consider (i.e. plot) the distribution of the statistic
   – Use the quantiles to create a CI on the statistic

Understanding Statistical Tests

Using Bayes theorem to
understand p-values

Bayes Theorem

$$p(\theta \mid \text{data}) \propto p(\text{data} \mid \theta).\ p(\theta)$$

- θ is the parameter or parameters
- p(θ | data): posterior, how data changes our understanding
- p(θ) is the prior
- p(data | θ) is the likelihood; how probability of data varies with changes to θ

- Classical / frequentist statistics only has likelihood

P-Value is Not the Probability You Think It is

- Frequentist statistics only has likelihood

$$\cancel{p(\theta \mid \text{data})} \propto p(\text{data} \mid \theta).\ \cancel{p(\theta)}$$

- A p-value of 5% says:

Given the data, there is 5% probability that the null hypothesis is correct ✗ ✗

P-Value is Not the Probability You Think It is

- Frequentist statistics only has likelihood

$$\cancel{p(\theta \mid data)} \propto p(data \mid \theta).\ \cancel{p(\theta)}$$

- A p-value of 5% says:

Given the data, there is 5% probability that the null hypothesis is correct ✗ ✗

  - This is a statement about the probability of a parameter, given data
  - Only possible in Bayesian statistics

P-Value is Not the Probability You Think It is

- Frequentist statistics only has likelihood

$$\cancel{p(\theta \mid \text{data})} \propto p(\text{data} \mid \theta).\ \cancel{p(\theta)}$$

- A p-value of 5% says:

There is 5% probability of getting this data if the null hypothesis is correct ✗

P-Value is Not the Probability You Think It is

- Frequentist statistics only has likelihood

$$\cancel{p(\theta \mid \text{data})} \propto p(\text{data} \mid \theta).\ \cancel{p(\theta)}$$

- A p-value of 5% says:

There is 5% probability of getting this data if the null hypothesis is correct ✗

  - This is a statement about the probability of the data given the parameter (i.e. the hypothesis)
  - However, probability of the data is small

P-Value is Not the Probability You Think It is

- Frequentist statistics only has likelihood

$$\cancel{p(\theta \mid \text{data})} \propto p(\text{data} \mid \theta).\ \cancel{p(\theta)}$$

*There is 5% probability of getting data **this extreme** if the null hypothesis is correct*

  - Probability relates to repeating the sampling process
  - Requires a single 'distance' statistic e.g. test statistic in chi-square
  - No probabilities for model parameters or hypotheses

C/W 4

Overview

I: Paper Review

Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

- Naive / incorrect interpretation of p-values
- Short written answer
- Includes a causal diagram

2: Re-Analysis of Data

- Regression
- Bootstrap

Finding the programming difficult? Please ask for help.

What Makes Statistics Difficult?

Is Statistic Relevant to data Science

Menti code: **1351 2465**

Is Statistics Relevant to Data Science?

Reflections not answers

What is Data

- Is this data?

Another example is text

- What is it distribution?
- Can you sample from it?
- Can you visualise its distribution?

Dimensionality

- Each variable is a dimension

Variable	Meaning
Age	The person's age in years
Sex	1 = male, 0 = female
ChestPain	The chest pain experienced
RestBP	The person's resting blood pressure (mm Hg on admission to the hospital)
Chol	The person's cholesterol measurement in mg/dl
Bsugar	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
RestECG	Resting electrocardiographic measurement

- Individual located in 'N-dimensional space'
  - We looked at single variables or pairs of variables
- Challenge of high-dimensionality

Statistical Modelling

- Model: one (or some) variables determined from other
- Example: why do some students fail?
  - Statistics: what factor explain failure(in a data set)
  - ML: Can we predict failure (given a data set)?

Statistics

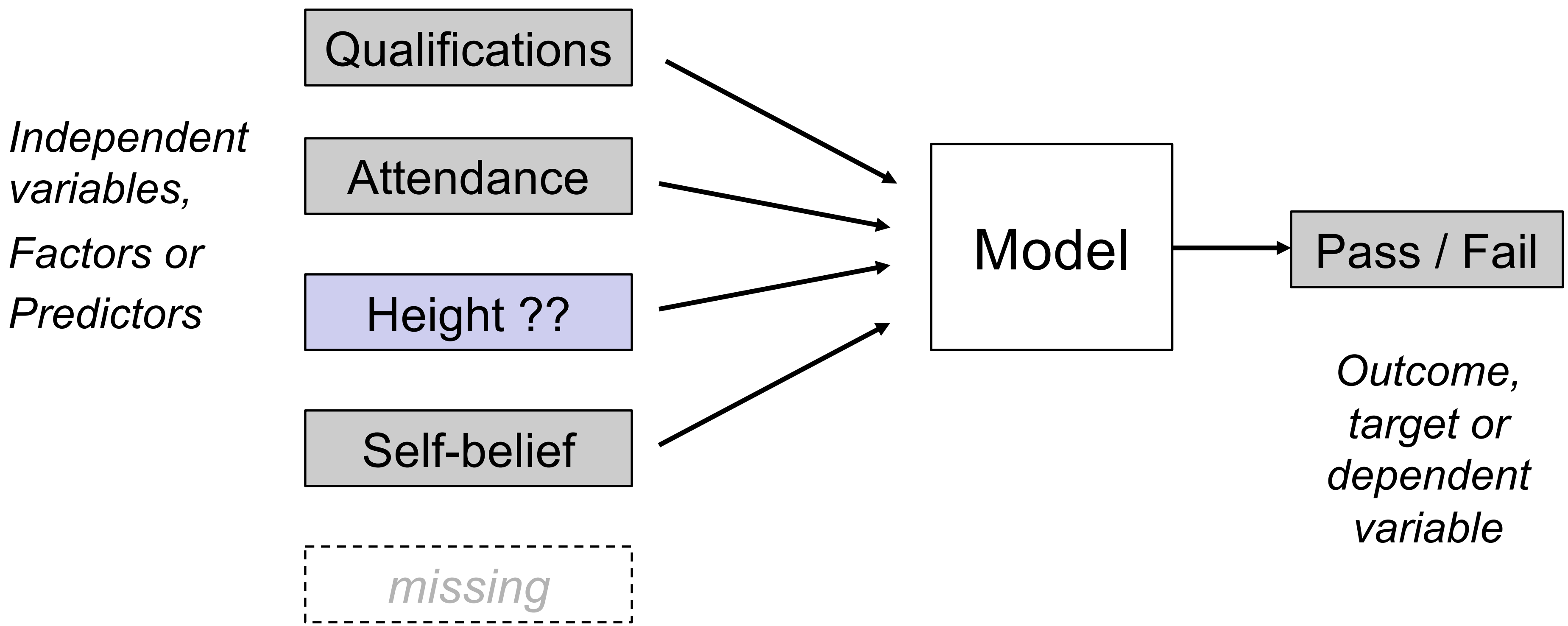
- Aim is explanation
  - Which variables?
  - Contribution of variables
- Performance: good fit
- Population

(Supervised) Machine Learning

- Aim is prediction
  - Which variables?
  - Which algorithm?
- Performance: accuracy
- Individual

Statistical Modelling & Performance

- Model: one (or some) variables determined from other
- Example: why do some students fail?
  - Statistics: what factor explain failure
  - ML: Can we predict failure?



Modelling: Prediction and Explanation

Linear Regression

Equation of a Line (1 Independent Variable)

- Two parameters
  - Intercept: Y when X = 0
  - Slope: increase in Y when X increases by 1

$$Y = 2.5 * X + 3$$

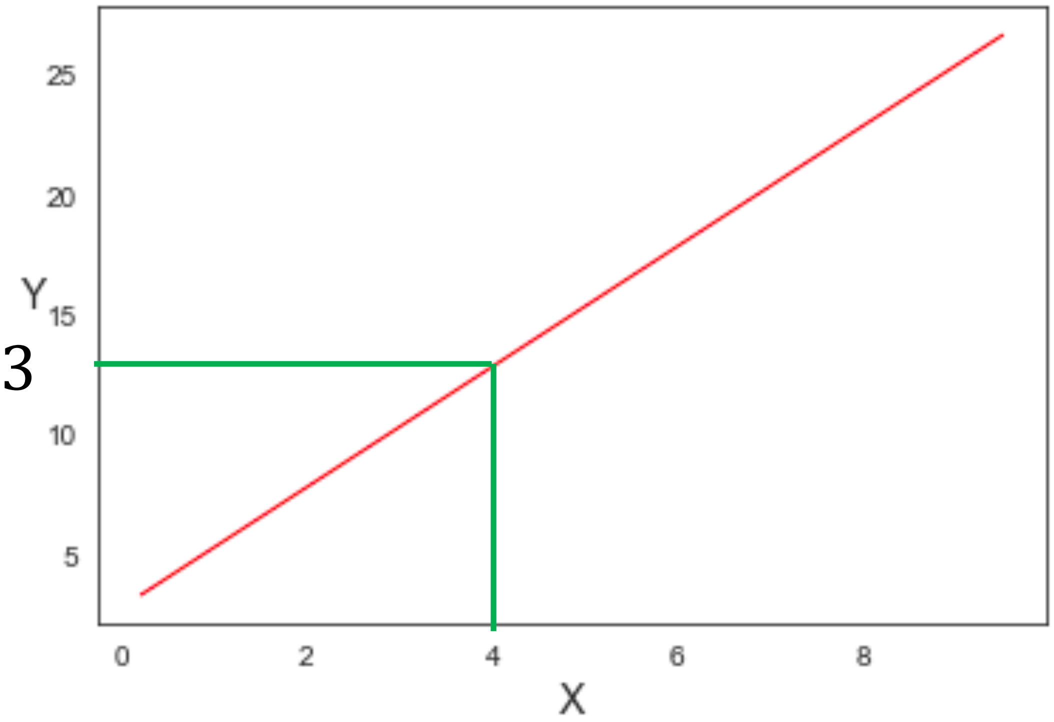
Dependent

Slope

Independent

Intercept

$$Y = (2.5 * 4) + 3 = 13$$



Linear Regression Assumptions

- Can have multiple independent variables (predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Each independent variable X_i
  - Adds or subtracts to Y in independently of other X_j
  - Has it's own 'coefficient' β_i
  - Linear: the same change in X_i gives same change in Y

- Cannot be true if X_i and X_j are correlated

Explaining using Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Each β shows the importance of its predictor
- β can be +ve or –ve
- If β very ‘small’ then predictor not important
  - Size is relative to other predictors
  - Standardise range of Xs

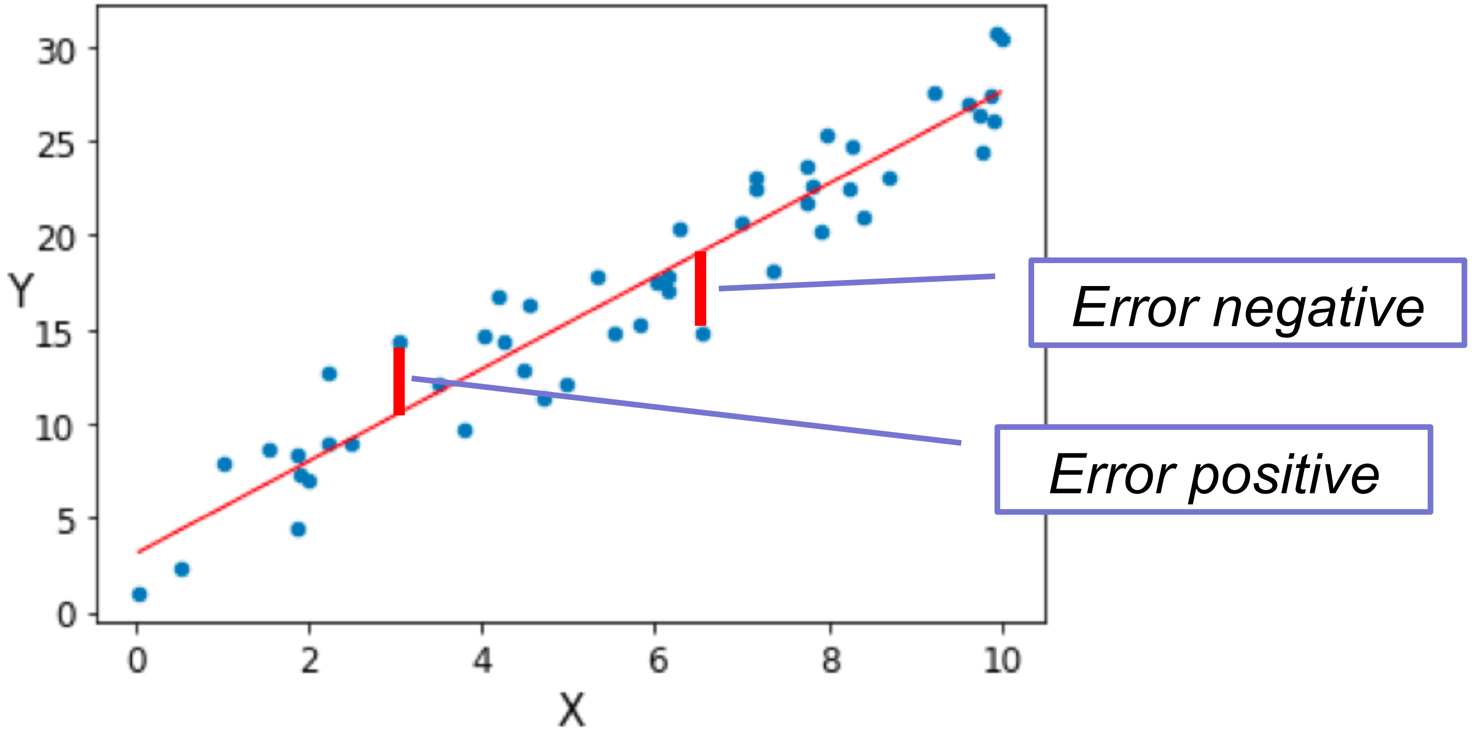
- *What about missing predictors?*

Regression Line for Data Points

- Points are not exactly on a line

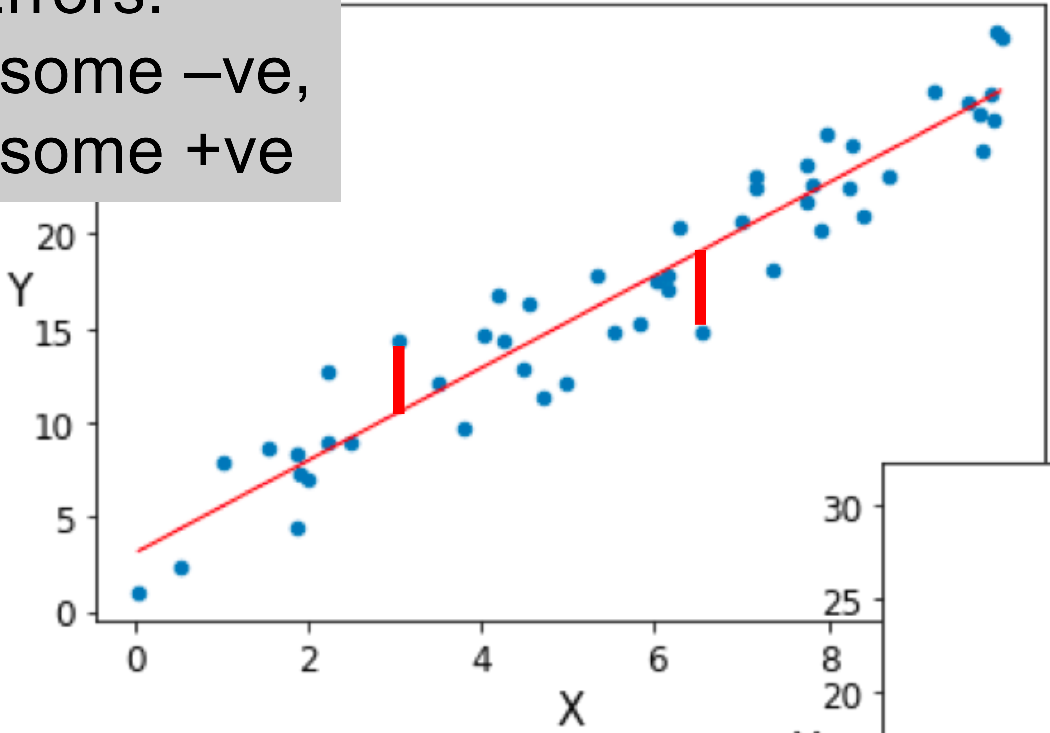
$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

error

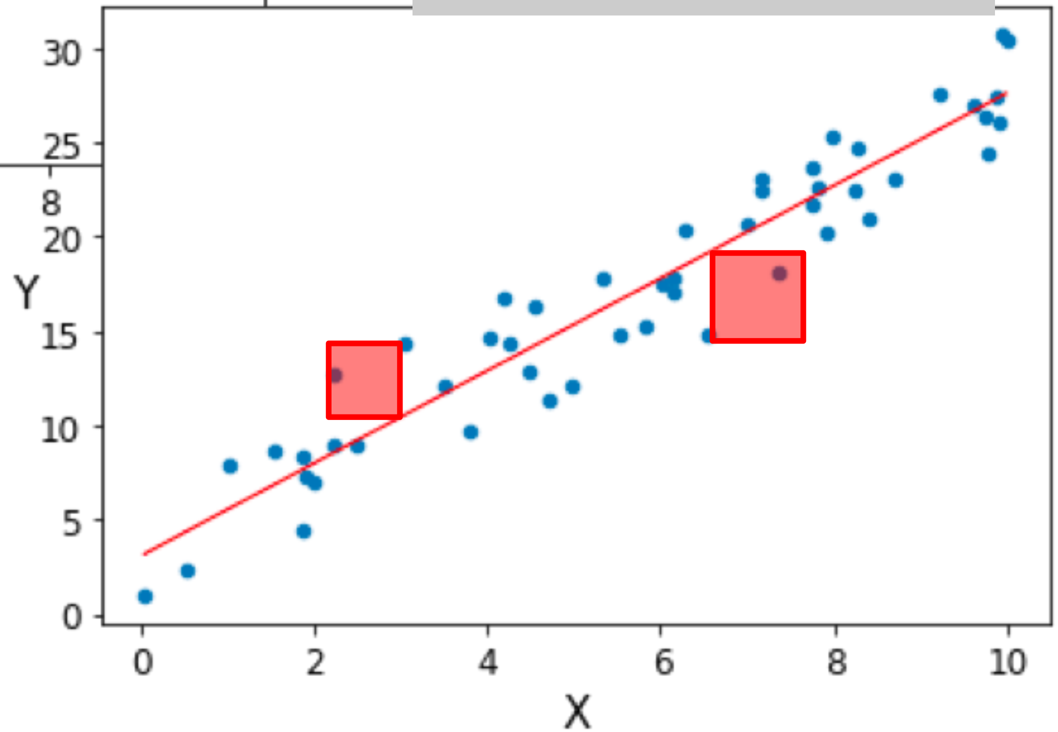


Minimise Residual Sum of Squares

Errors:
some –ve,
some +ve



Errors squared:
All +ve
Min => balance



Warning: remove outliers

Residuals (Errors)

Prediction – if the point on the line

$$\widehat{y_i}_i = \beta_0 + \beta_1 x_{1i}$$

Actual – off the line by an error

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

Residuals
(errors)

$$e_i = y_i - \widehat{y_i}_i$$

‘Best’ Fit and Distribution of Errors

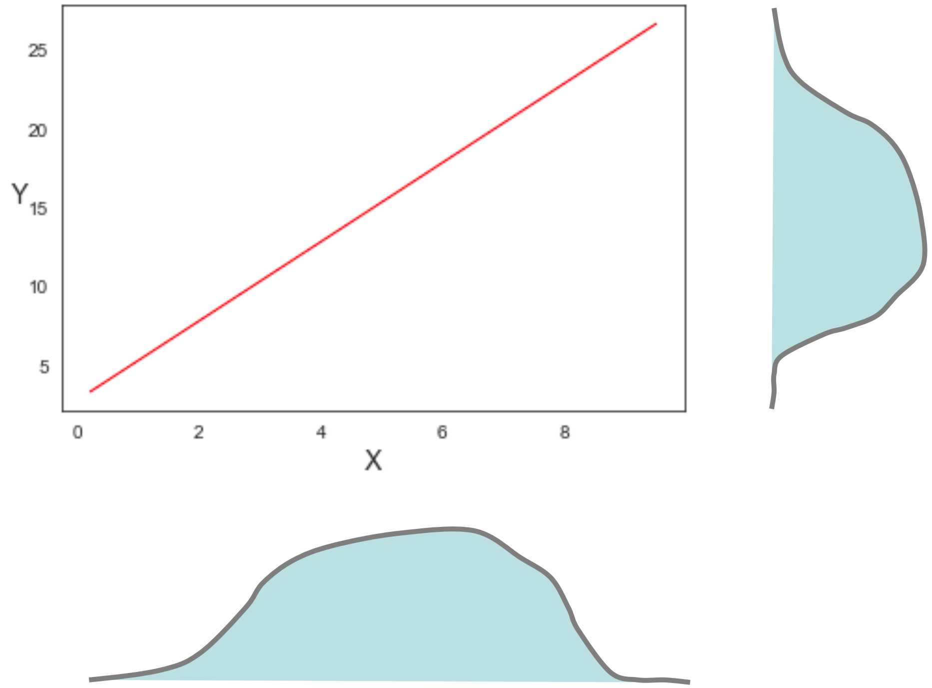
- Theory assumes that distribution of residuals (errors) is normal

- You can check this
  – Plot the distribution
  – QQplot for normality

- If distribution of residuals skewed, then the parameters may not be ‘best’

Goodness of Fit: R^2

- R^2 is popular: *coefficient of determination*
- Range 0 to 1
- Proportion of the variance of Y that is predictable from X
  - Rest of the variance due to errors
  - i.e. missing predictors



Goodness of Fit: R^2

- Proportion of the variance Y that is predictable from X

$$R^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- Perfect prediction the $R^2 = 1$
- If we always predict $\bar{y}$ then $R^2 = 0$

- *Note: this is not the most general definition, but it applies in linear regression*

Goodness of Fit: RMSE

- RMSE: root mean squared error
- RMSE = $\sqrt{\frac{1}{N}\sum_i {e_i}^2} = \sqrt{\frac{1}{N}\sum_i \left(y_i - \widehat{y_i}_i\right)^2}$
  - Instead of N, sometimes N – p – 1 (for p predictors) as number of degrees of freedom

- More common in ML
  - Accuracy of predictor for continuous variable

Issues for Regression

- Issue 1: Enough Data?
  - Each has β to be estimated from data
  - A statistical model can be too complex for the data
  - Most statistical models have more parameters

- Issue 2: co-linearity
  - Remember assumption: predictor independent
  - Always check correlation of predictors

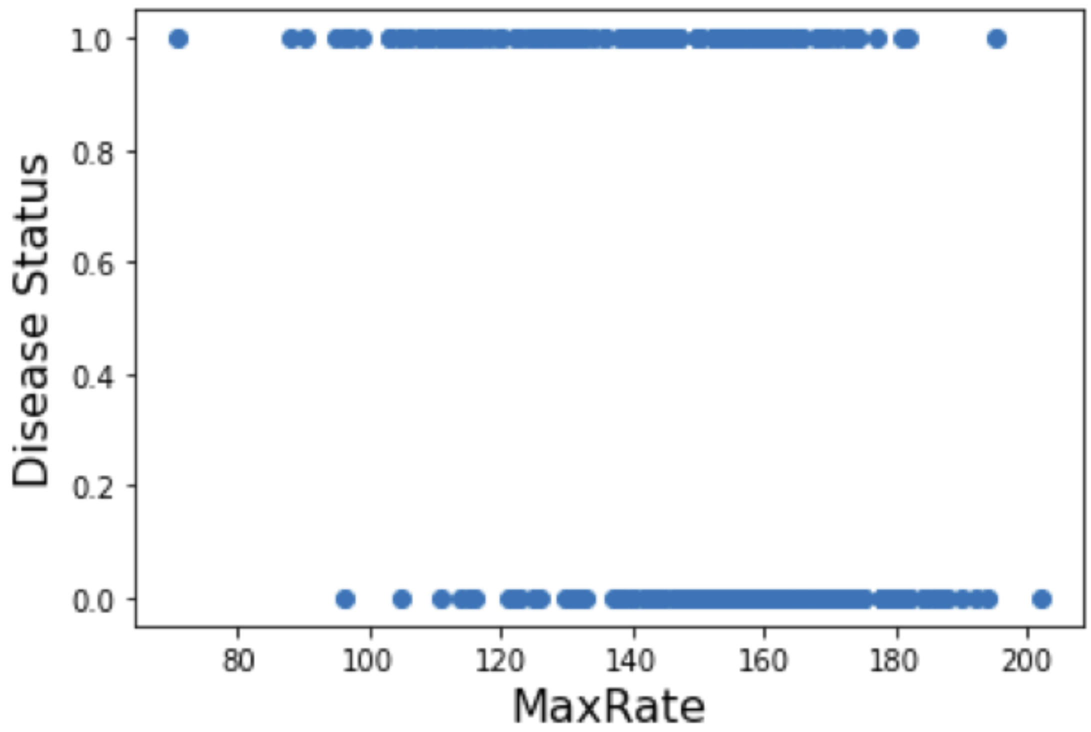
- Stepwise regression
  - Algorithm for choosing best set of predictors

Modelling: Prediction and Explanation

Logistic Regression

Problem: Regression with Binary Target

- How to use a linear regression for target with 2 values?



Logistic Regression: Key Ideas

1. Predict a probability
   – Advantage: it's a number; use it to choose class
   – Problem: range 0 to 1
   – $p = f(\beta_0 + \beta_1 . x_1 + \beta_2 . x_2)$ – choose a suitable *f()*

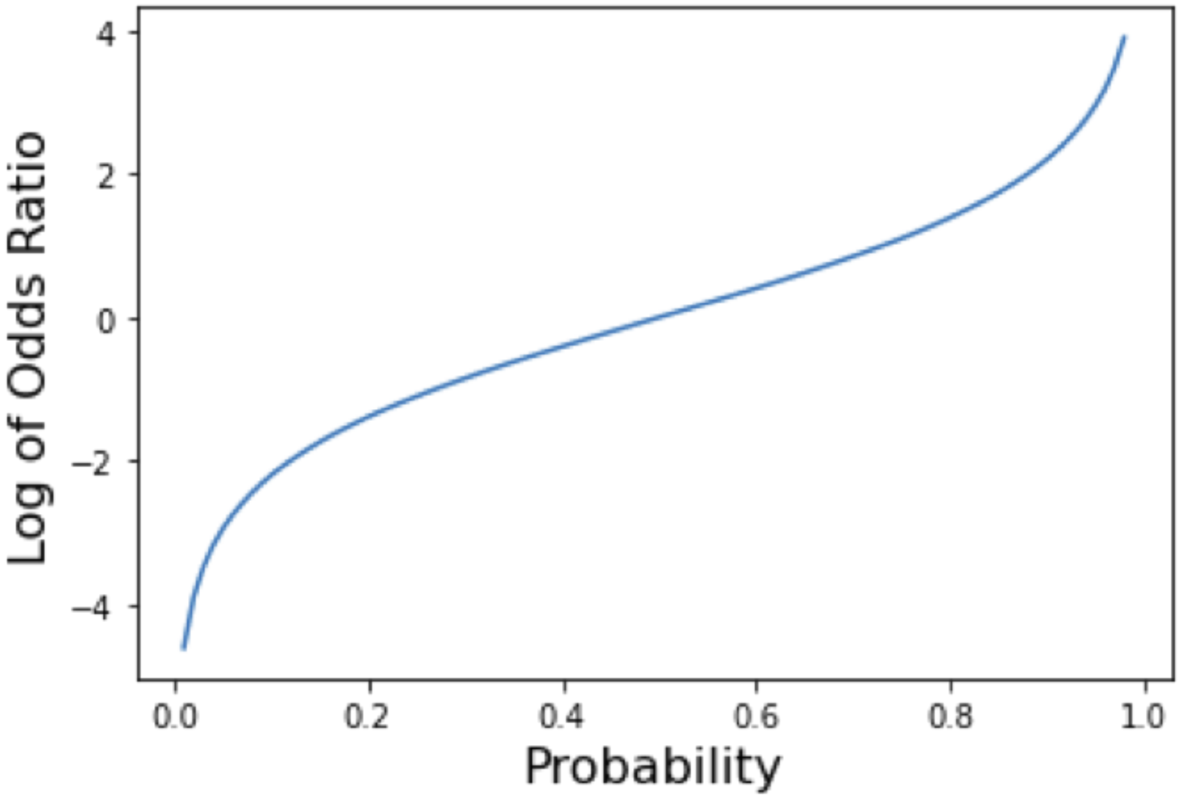
2. Predict odds p(Y=true) / p(Y=false)
   – Advantage: range is 0 upwards
   – Problem: not linear; cannot be negative

3. Predict the log of the odds
   – Solution: range over -∞ to +∞

Logit: Log Odds

- Maps probability *p* to range -∞ to +∞

Log of odds ratio: $logit(p(x)) = \ln(\frac{p(x)}{1-p(x)})$



Not the only possible conversion

Getting the Probability & Class

- Logit regression
  - Linear regression on log odds
  - $logit(p(x)) = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2$
- Odds
  - Reverse the log: $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1.x_1+\beta_2.x_2}$
- Probability
  - Reverse the odds: $p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1.x_1+\beta_2.x_2)}}$

- Class: y is True if p > 50% (*a possible threshold*)

Modelling: Prediction and Explanation

Accuracy, Confusion Matrix and AUC

Applies to any binary classifier

Confusion Matrix

- Compare actual and predicted

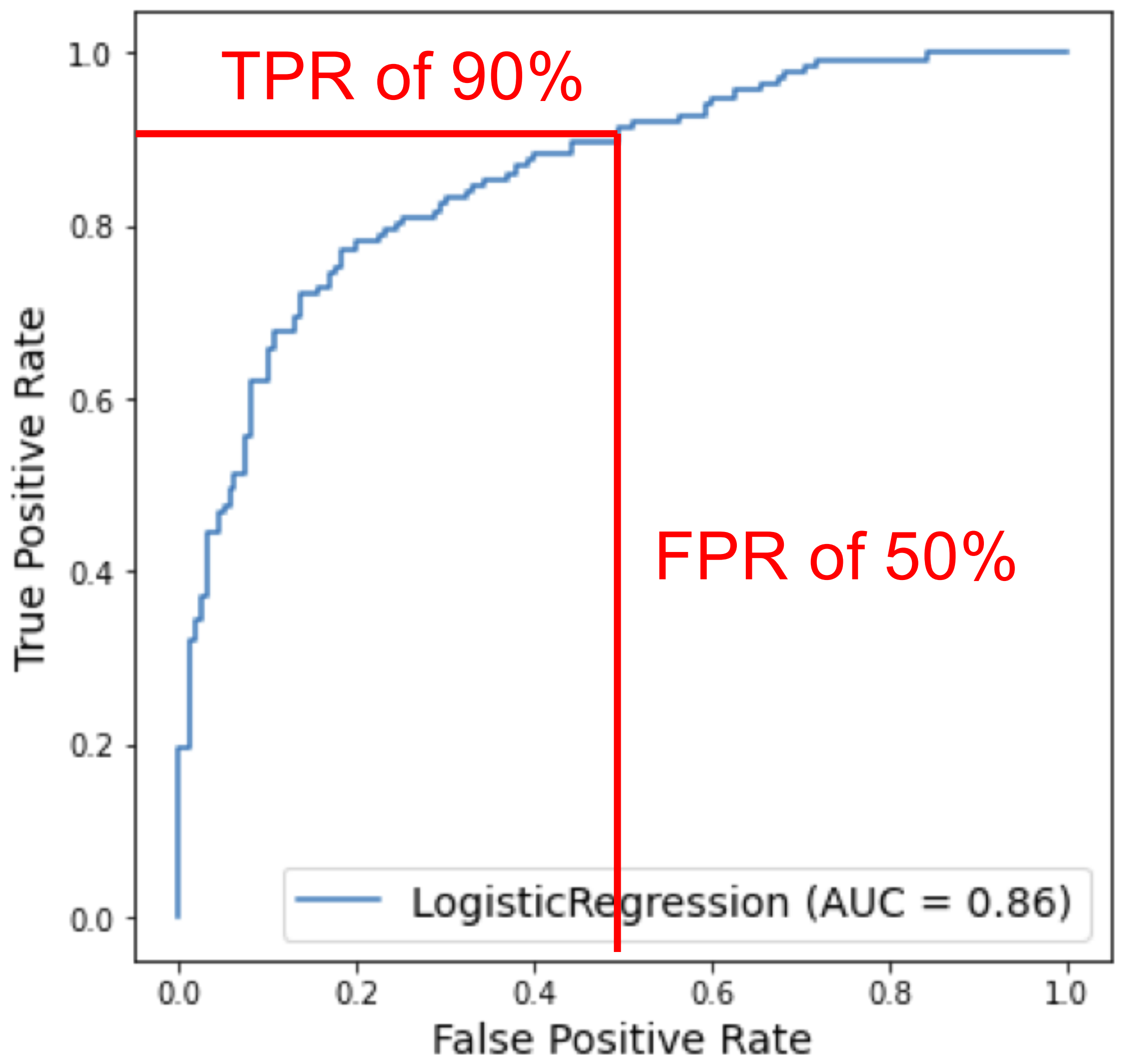
		Predicted Disease Status	
		Positive	**Negative**
True Disease Status	**Positive**	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

- Classification depends on probability threshold
- Are both types of error equal?

ROC: Sensitivity v Specificity

- Y axis
  - TPR (Sensitivity)
- X axis
  - FPR (1 – Specificity)

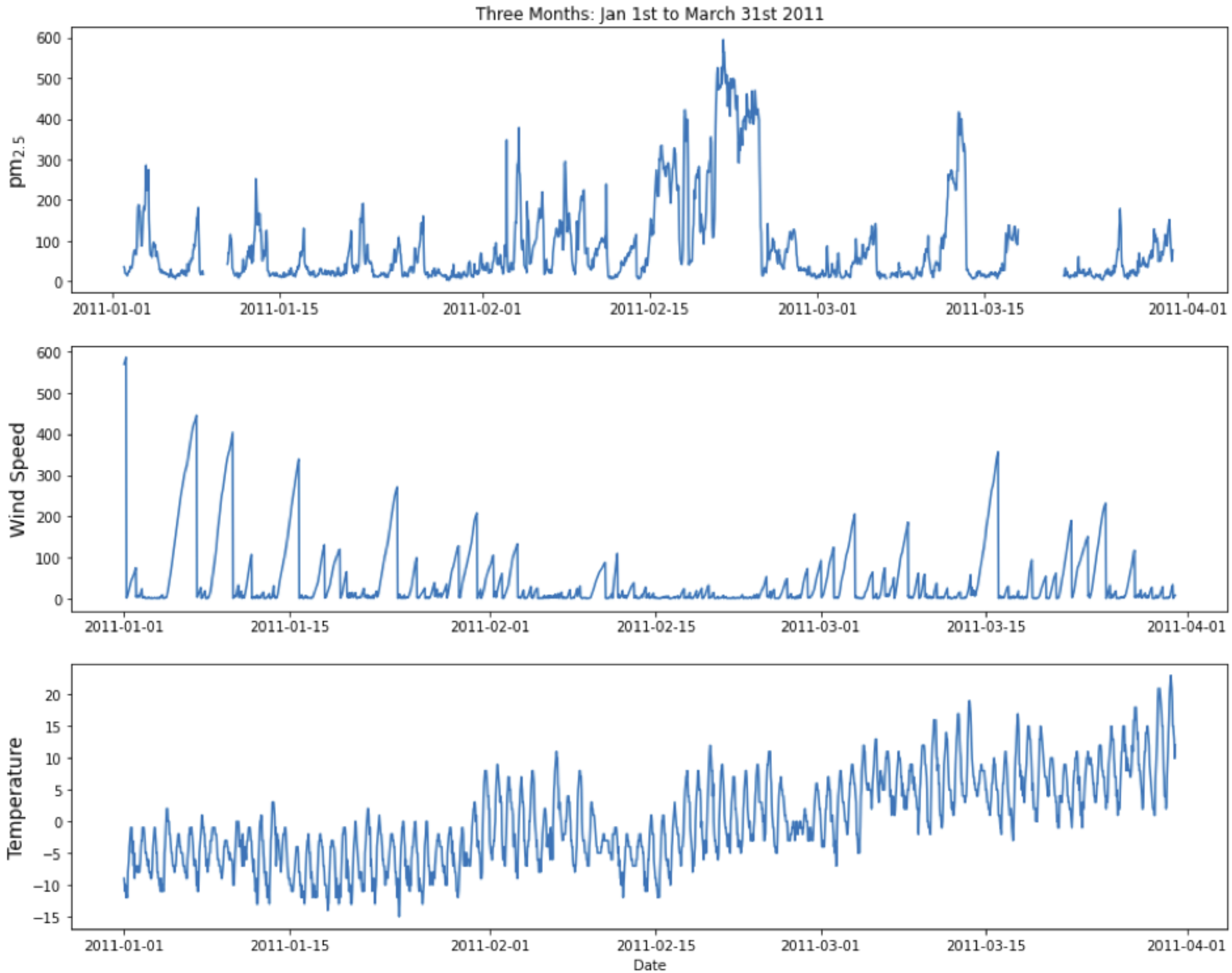
- Curve
  - Possible operating points
  - Given by threshold
- AUC: measure of performance



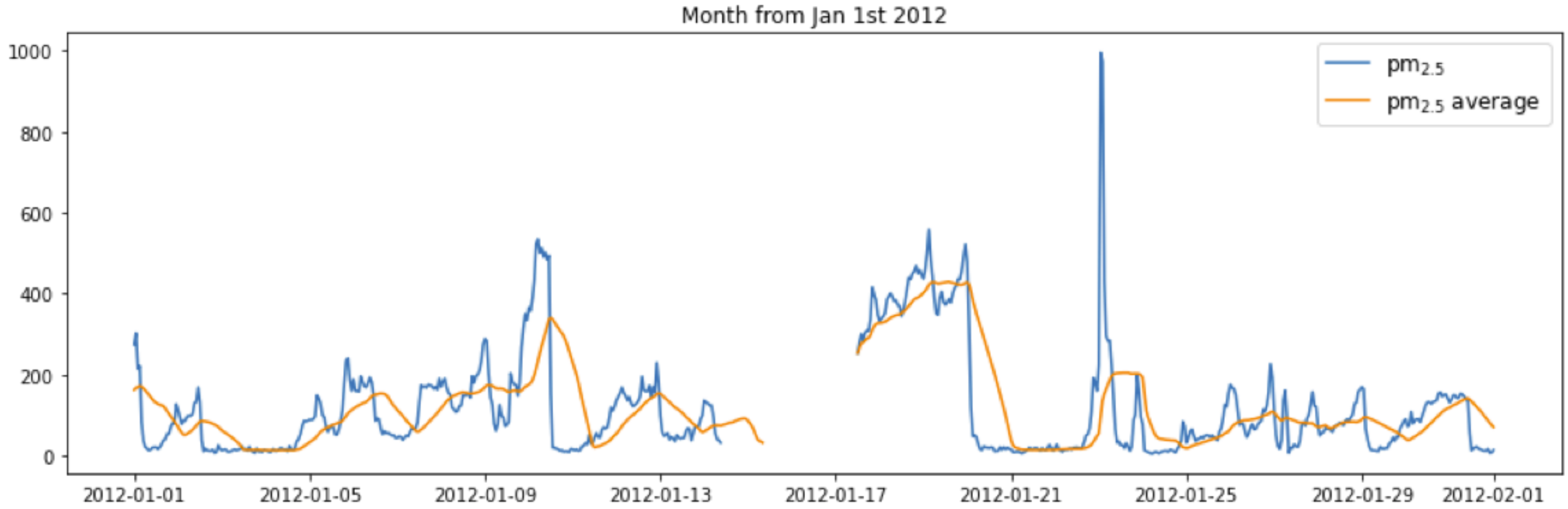
Time Series Analysis

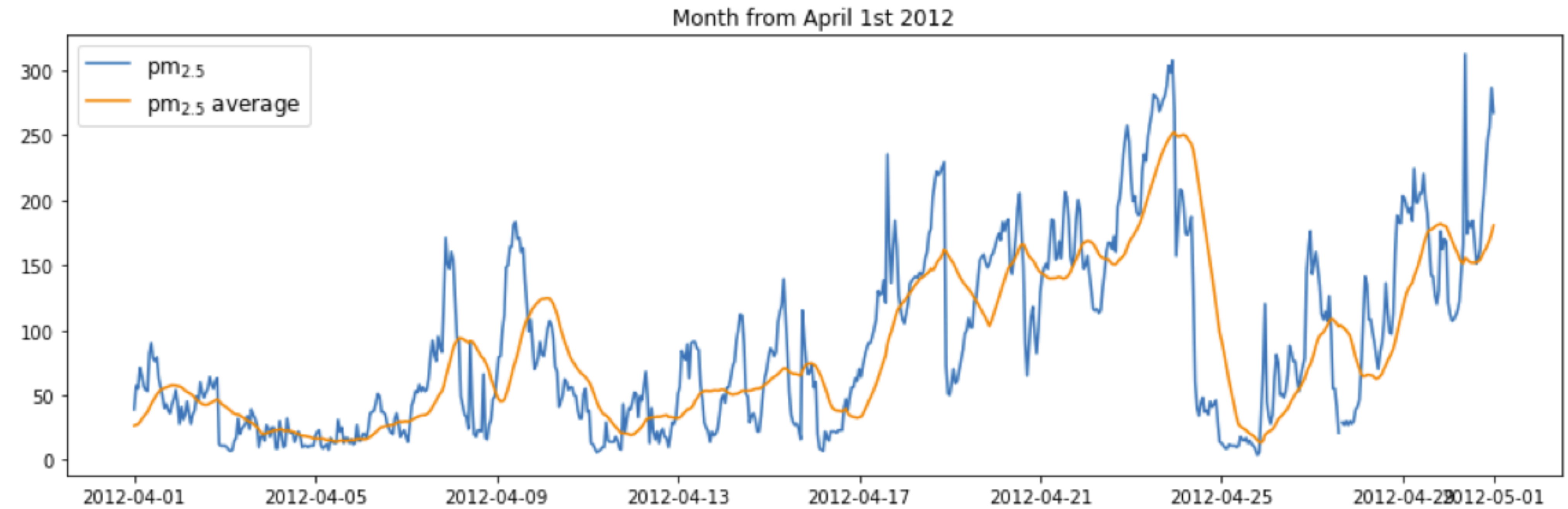
Meanings of Time

Selecting a Range: 3 Months



Rolling Average: Daily





Modelling Time Series Data

- Combined elements of
  - Trend
  - Auto-regression
  - Multiple time series

Trend: Time as a Predictor

- Regression of the form

$$y_t = \beta_0 + \beta_1 t + \epsilon$$

- Combined with other terms

Auto-Regression: Earlier Values Predict

- Use an earlier value (lagged value) to predict later value

$$y_t = \beta_0 + \beta_1 y_{t-T} + \epsilon$$

- Used for prediction

Regression with Multiple Time Series

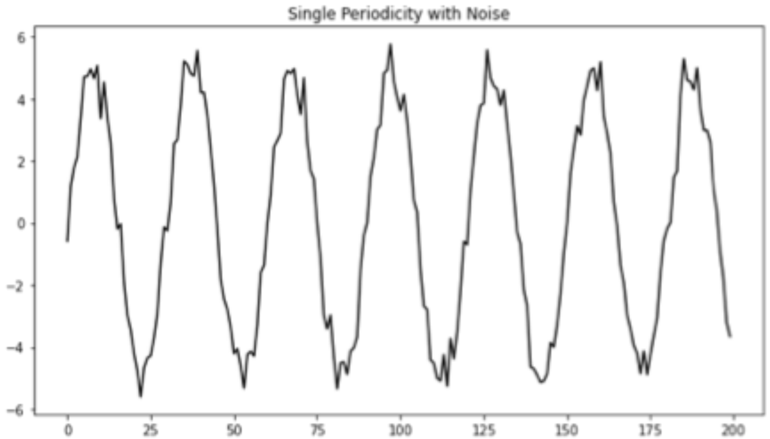
- Values of some time series predict another time series

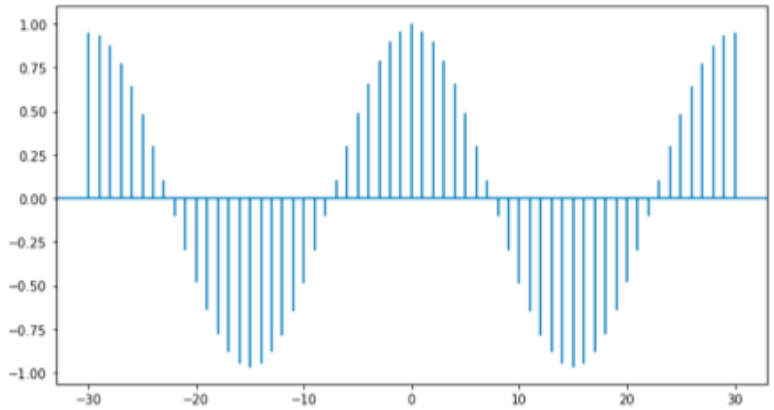
$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \epsilon$$

- Without lag, not a prediction
- Used to understand causal patterns

Auto Correlation

- Generalise correlation
  - Correlation as a function of lag





- Does not ‘see’ other periodicity
  - ‘Partial auto-correlation’
  - Frequency decomposition

Summary

Is Statistics Relevant to Data Science/ML?

- Disciplines developed separately
  - Slowly converging
- Two aims: both relevant
  - Prediction
  - Explanation
- Sampling and uncertainty
  - Increasing depends to ‘explain’ prediction
  - Understand the performance of models
- Model building
  - All statistics involves model building
  - Statistical validity versus validity of model
  - ML offers new models

Summary

- I am not a statistician!
- Trying to present statistics in a way that is relevant to data science
  - Work in progress

- Classical statistical tests not very relevant
- Bootstrap brilliant
- Bayesian thinking increasingly important
- Linear models remain surprisingly relevant