FIT2086 Lecture 12
Revision

Daniel F. Schmidt

Faculty of Information Technology, Monash University

October 21, 2019

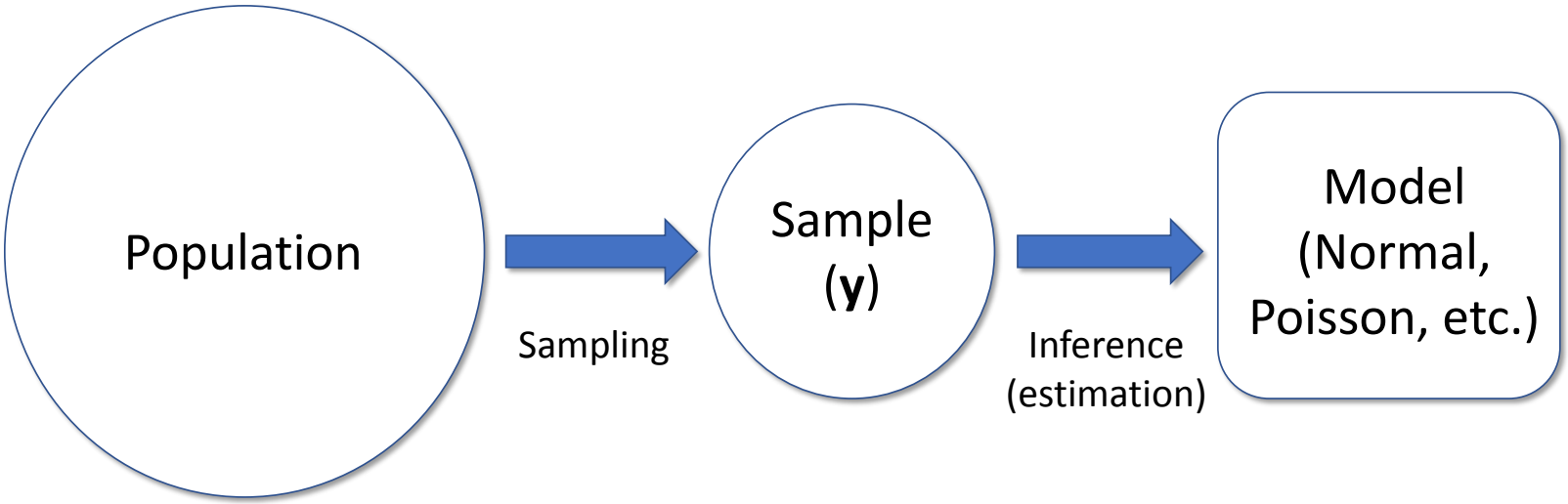
House Keeping

- Studio 12
  - A set of sample exam questions
- Assignment #3
  - Due this coming Sunday
  - Submit one PDF file and appropriate R files
- Subject Evaluation of Teaching and Units (SETU)
  - Those who have not yet done so, can you please fill out the SETU feedback
  - Greatly appreciated!

Subject Outcomes – Revision

- On completion of this unit, students should be able to:
  1. perform exploratory data analysis with descriptive statistics on given datasets; ✓
  2. construct models for inferential statistical analysis; ✓
  3. produce models for predictive statistical analysis; ✓
  4. perform fundamental random sampling, simulation and hypothesis testing for required scenarios; ✓
  5. implement a model for data analysis through programming and scripting; ✓
  6. interpret results for a variety of models. ✓

Revision from Lecture 1 (1)



- Population: A large collection of objects or items with measureable attributes
- Sample: A finite number of recordings of attributes of items from a population
- Model: A mathematical or algorithmic description of the population learned/inferred from the sample

Revision from Lecture 1 (2)

- $\mathbb{P}(X = x, Y = y)$ is joint probability of $X = x$ and $Y = y$.
  - Sum-rule (marginal probability):

$$\mathbb{P}(X = x) = \sum_{y} \mathbb{P}(X = x, Y = y)$$

  - Conditional probability

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

- Cumulative distribution function (for ordered x):

$$\mathbb{P}(X \leq x) = \sum_{x \leq x} \mathbb{P}(X = x)$$

- Also: $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$.

Revision from Lecture 2 (1)

- Let $\mathbb{P}(X = x) \equiv p(x)$; expectation and variance of $f(X)$:

$$\begin{aligned} \mathbb{E}\left[f(X)\right] &= \sum_x p(x) f(x) \\ \mathbb{V}\left[f(X)\right] &= \mathbb{E}\left[\left(X - \mathbb{E}\left[f(X)\right]\right)^2\right] \end{aligned}$$

  with integral replacing sum for continuous RVs.

- Some useful rules:
  - $\mathbb{E}\left[f(X) + g(Y)\right] = \mathbb{E}\left[f(X)\right] + \mathbb{E}\left[g(Y)\right]$
  - $\mathbb{E}\left[cf(X)\right] = c\mathbb{E}\left[f(X)\right]$
  - $\mathbb{V}\left[cf(X)\right] = c^2\mathbb{V}\left[f(X)\right]$

- If X, Y are independent RVs
  - $\mathbb{E}\left[f(X)g(Y)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right]$
  - $\mathbb{V}\left[f(X) + g(Y)\right] = \mathbb{V}\left[f(X)\right] + \mathbb{V}\left[g(Y)\right]$

Revision from Lecture 2 (2)

- Parametric distributions as models of populations
- Normal distribution; $X \in \mathbb{R}$, $X \sim N(\mu, \sigma^2)$

$$\mathbb{E}[X] = \mu, \quad \mathbb{V}[X] = \sigma^2$$

- Bernoulli distribution; $X \in \{0, 1\}$, $X \sim \text{Be}(\theta)$

$$\mathbb{E}[X] = \theta, \quad \mathbb{V}[X] = \theta(1 - \theta)$$

- Binomial distribution; $X \in \{0, 1, \ldots, n\}$, $X \sim \text{Bin}(\theta, n)$

$$\mathbb{E}[X] = n\theta, \quad \mathbb{V}[X] = n\theta(1 - \theta)$$

- Poisson distribution; $X \in \{0, 1, 2, \ldots\}$, $X \sim \text{Poi}(\lambda)$

$$\mathbb{E}[X] = \lambda, \quad \mathbb{V}[X] = \lambda$$

Revision from Lecture 3 (1)

- We looked at problem of parameter estimation
- Method of maximum likelihood

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta}\{p(\mathbf{y} \mid \theta)\}$$

- Maximum likelihood estimators for the normal

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_{\mathrm{ML}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{\mathrm{ML}})^2}$$

- Maximum likelihood estimator for Poisson

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Revision from Lecture 3 (2)

- Sampling distributions of estimators
- Bias and variance of an estimator

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta, \quad \mathrm{Var}_\theta(\hat{\theta}) = \mathbb{V}\left[\hat{\theta}\right]$$

- Mean squared error of an estimator

$$\mathrm{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \mathrm{Var}_\theta(\hat{\theta})$$

- If $Y_1, \ldots, Y_n$ have $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$ then

$$b_\mu(\bar{Y}) = 0, \quad \mathrm{Var}_\mu(\bar{Y}) = \frac{\sigma^2}{n}, \quad \mathrm{MSE}_\mu(\bar{Y}) = \frac{\sigma^2}{n}$$

  where $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$ is the sample mean
- An estimator $\hat{\theta}$ is consistent if

$$b_\theta(\hat{\theta}) \to 0, \quad \mathrm{Var}_\theta(\hat{\theta}) \to 0,$$

  as $n \to \infty$ for all θ.

Example: ML Estimation of a Poisson (1)

- Recall the Poisson distribution with rate λ:

$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}.$$

- If $\mathbf{y} = (y_1, \ldots, y_n)$ are n integers, then the likelihood for a Poisson model is

$$\begin{aligned} p(\mathbf{y} \mid \lambda) &= \prod_{i=1}^{n} p(y_i \mid \lambda) \\ &= \left(\frac{\lambda^{y_1} \exp(-\lambda)}{y_1!}\right) \cdot \left(\frac{\lambda^{y_2} \exp(-\lambda)}{y_2!}\right) \cdots \left(\frac{\lambda^{y_n} \exp(-\lambda)}{y_n!}\right) \\ &= \frac{\lambda^{\sum_{i=1}^n y_i} \exp\left(-n\lambda\right)}{\prod_{i=1}^n y_i!} \end{aligned}$$

by $e^a e^b = e^{a+b}$ and independence of $y_1, \ldots, y_n$.

Example: ML Estimation of a Poisson (2)

- The negative log-likelihood is then

$$L(\mathbf{y} \mid \lambda) = n\lambda - \sum_{i=1}^{n} y_i \log \lambda + \sum_{i=1}^{n} \log y_i!$$

- To find the ML estimator of λ we need to minimise $L(\mathbf{y} \mid \lambda)$, or equivalently solve

$$\frac{dL(\mathbf{y} \mid \lambda)}{d\lambda} = 0,$$

for λ.

Example: ML Estimation of a Poisson (3)

- The derivative is given by

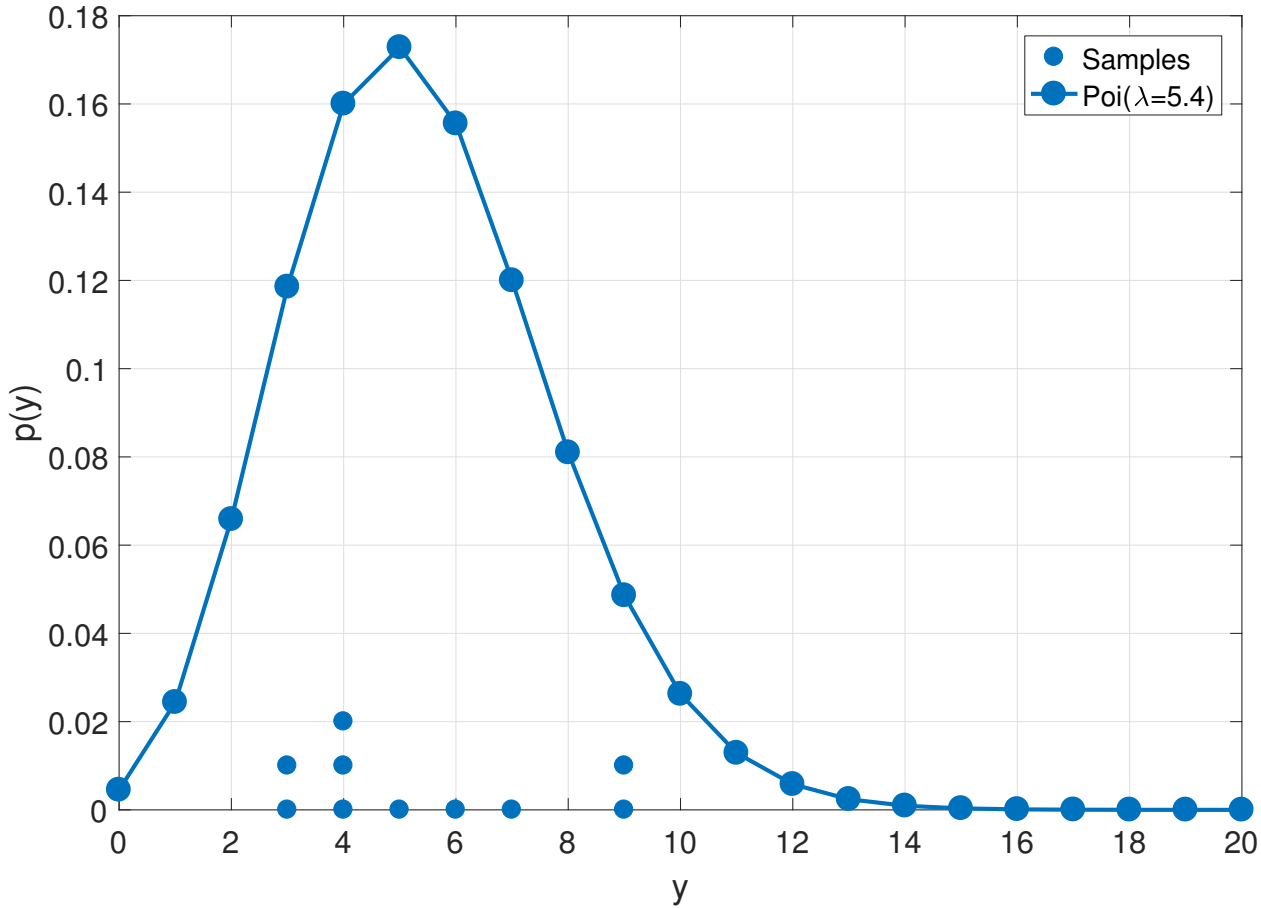
$$\frac{dL(\mathbf{y} \mid \lambda)}{d\lambda} = n - \frac{\sum_{i=1}^{n} y_i}{\lambda} \tag{1}$$

  where we use $d \log x / dx = 1/x$
- Setting (1) to zero and solving for λ yields

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- So, the ML estimator for the Poisson rate is the sample mean.
  $\Rightarrow$ this is not always the case!

Example: ML Estimation of a Poisson (4)



Data samples and the Poisson distribution fitted by maximum likelihood with $\hat{\lambda} = 5.4$. Samples were $\mathbf{y} = (7, 9, 3, 5, 3, 4, 4, 9, 4, 6)$.

Revision from Lecture 4

- Central limit theorem: if $Y_1, \ldots, Y_n$ are RVs with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$ then

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N\left(n\mu, n\sigma^2\right)$$

- Implies distribution of the sample mean $\bar{Y}$ for $Y_1, \ldots, Y_n$ with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$ satisfies

$$\bar{Y} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Confidence Intervals

- An interval estimator returns an interval of plausible values for a population parameter θ

Confidence Intervals

We use the method of confidence intervals.

We say the interval estimator $(\hat{\theta}_\alpha^-(\mathbf{y}), \hat{\theta}_\alpha^+(\mathbf{y}))$ generates a $100(1-\alpha)$-percent confidence interval, for $\alpha \in (0,1)$, if

$$\mathbb{P}\left(\theta \in (\hat{\theta}_\alpha^-(\mathbf{y}),\ \hat{\theta}_\alpha^+(\mathbf{y}))\right) = 1-\alpha,$$

where the probability is with respect to all the different samples $\mathbf{y}$ we could draw from our population.

- 95% confidence intervals: cover the true population parameter for 95% of possible samples we could draw from our population

CI for Normal Mean, Known Variance

- Assuming the population is normally distributed with (unknown) mean μ and (known) variance σ^2, these results yield the following 95% confidence interval for $\hat{\mu}_{\mathrm{ML}} \equiv \bar{Y}$,

$$\left(\hat{\mu}_{\mathrm{ML}} - 1.96\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

- More generally, a $100(1-\alpha)\%$ confidence interval is given by:

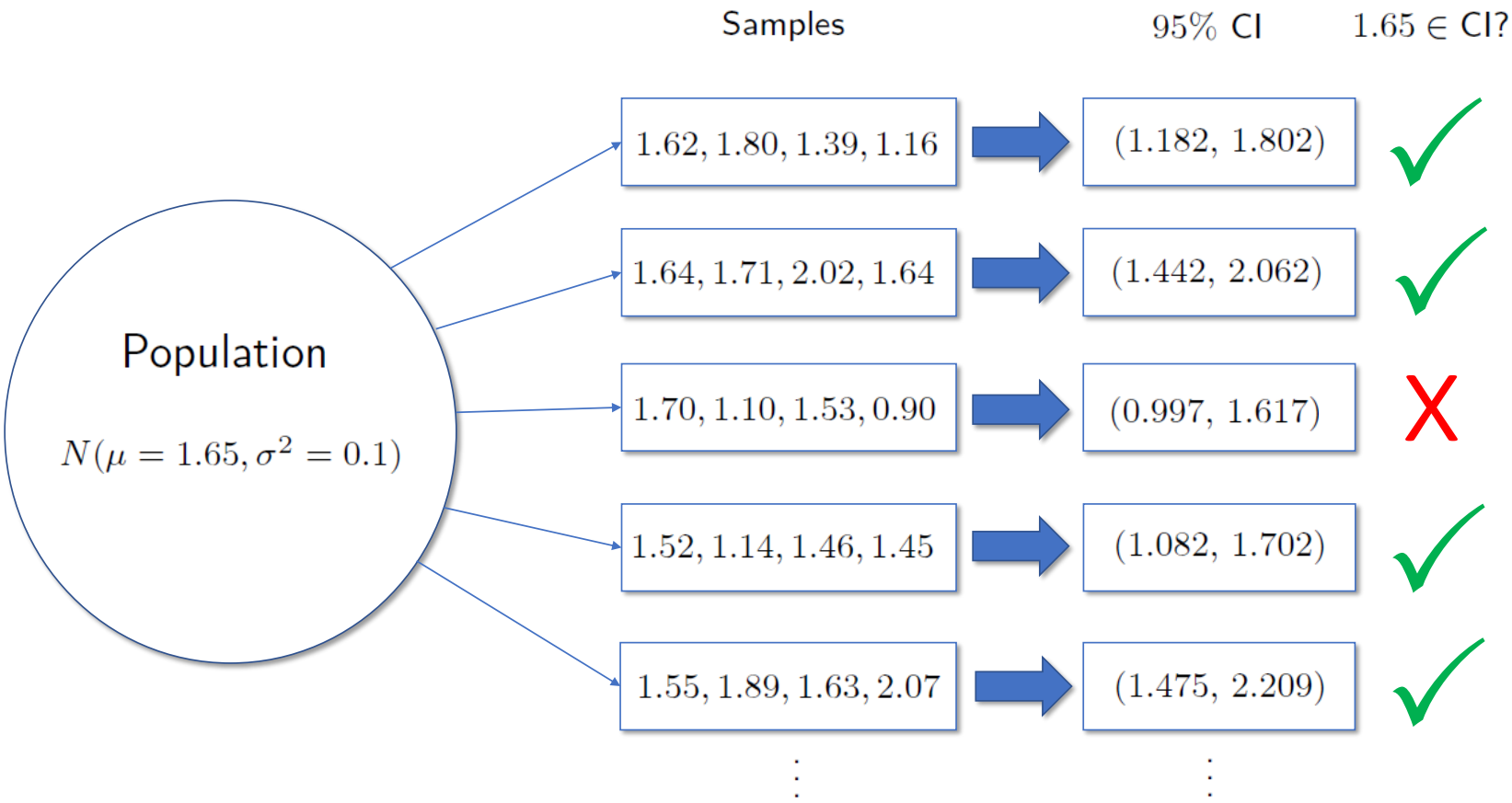
$$\left(\hat{\mu}_{\mathrm{ML}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the unit normal:

  - for $\alpha = 0.05$, $z_{0.025} = Q(p = 0.975) \approx 1,96$;
  - for $\alpha = 0.01$, $z_{0.005} = Q(p = 0.995) \approx 2.576$;
  - for general α, use $Q(p = 1-\alpha/2)$

where $Q(\cdot)$ is the quantile function for the unit normal.

CI for Normal Mean, Known Variance (2)



Cartoon showing multiple samples drawn from a $N(\mu = 1.65, \sigma^2 = 0.1)$ population, along with the 95% confidence intervals for each sample. 5% of possible samples will result in CIs that do not include $\mu = 1.65$.

Example: Normal Mean, Known Variance (1)

- Example: We have the following samples of body mass index taken people with diabetes from the Pima ethnic group

$$\mathbf{y} = (53.2, 33.6, 36.6, 42.0, 33.3, 37.8, 31.2, 43.4)$$

- Imagine we are given a value for the population variance of 43.75 which has been estimated by another, very large study of people from the Pima group.
- Task: Estimate the BMI of diabetic Pima people and construct a 95% CI
- Our best guess at the population mean BMI for Pima people with diabetes is

$$\hat{\mu}_{\mathrm{ML}} = 38.88$$

Example: Normal Mean, Known Variance (2)

- Our 95% CI is then

$$\left(38.88 - 1.96\sqrt{43.75/8},\ 38.88 + 1.96\sqrt{43.75/8}\right)$$

which is equal to

$$(34.3,\ 43.47)$$

- In words, we summarise our analysis by:

"The estimated mean BMI of people from the Pima ethnic group with diabetes (sample size $n = 8$) is $38.88\,kg/m^2$. We are 95% confident the population mean BMI for this group is between $34.3\,kg/m^2$ and $43.75\,kg/m^2$."

CI for Difference of Normal Means

- We have two samples $\mathbf{y}_A$ and $\mathbf{y}_B$
  - Wish to get confidence interval for $\mu_A - \mu_B$ (i.e., the difference in population means)
- Let us assume μ_A, μ_B, σ_A^2, σ_B^2 are all unknown
- Let $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ be unbiased estimates of the variance in sample A and B, respectively
- Then the following interval:

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}} \right)$$

is an *approximate* $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_A - \hat{\mu}_B$, with the approximation getting better for increasing n_A and n_B.

Revision from Lecture 5

- Hypothesis testing; test null hypothesis vs alternative

$$\begin{aligned} H_0 &: \text{null hypothesis} \\ &\text{vs} \\ H_A &: \text{alternative hypothesis} \end{aligned}$$

- A test-statistic measures how different our observed sample is from the null hypothesis
- A p-value quantifies the evidence against the null hypothesis
- A p-value is the probability of seeing a sample that results in a test statistic as extreme, or more extreme, than the one we observed, just by chance if the null was true.

Testing μ with known variance

- Assume population follows normal distribution with unknown mean and known variance σ^2; testing inequality of μ
  1. First calculate the ML estimate of the mean/sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  2. Then calculate the z-score

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\sigma/\sqrt{n})}$$

  3. Then calculate the p-value:

$$p = \begin{cases} 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases} .$$

  where $Z \sim N(0,1)$

Example: Testing if $\mu = \mu_0$ (1)

- For US women aged between 20 to 34 years of age, the population body mass index (BMI) has
  - an approximate mean of $26.8 kg/m^2$; and
  - an approximate standard deviation of $4.5 kg/m^2$.

  (Source: Center for Disease Control)
- We have BMI measured on a sample of women aged 20-34 from the Pima ethnic group, without diabetes:

$$\mathbf{y} = (46.8, 27.8, 32.5, 39.5, 32.8, 31.0, 26.2, 20.8)$$

- Using this data, can we say whether women aged 20-34 in this Pima cohort have the same average BMI as the general US population?

Example: Testing if $\mu = \mu_0$ (2)

- We want to test:
  - $H_0 : \mu = 26.8$ vs $H_A : \mu \neq 26.8$,

  μ is the population mean BMI of Pima women aged 20-34.
- The estimated mean $\hat{\mu}$ from our sample is

$$\hat{\mu} = 32.175$$

- From this we can calculate the z-score as

$$z_{\hat{\mu}} = \frac{32.175 - 26.8}{(4.5/\sqrt{8})} = 3.3784$$

- This yields a p-value of

$$\begin{aligned} 1 - \mathbb{P}(-|z_{\hat{\mu}}| < Z < |z_{\hat{\mu}}|) &= 2 * \texttt{pnorm}(-\texttt{abs}(3.3784)) \\ &= 7.29 \times 10^{-4} \end{aligned}$$

Example: Testing if $\mu = \mu_0$ (3)

- How to interpret?
- A p-value of 7.29×10^{-4} can be interpreted as follows:

  If the null was true, i.e., Pima ethnic women aged 20-34 have the same BMI as the average US woman aged 20-34, then the chance of observing a sample with as an extreme, or more extreme, difference from the null as the one that we saw would be less than $1/1371$.

- So quite unlikely to happen just by vagaries of sampling $\Rightarrow$ strong evidence against the null.

Testing difference of means, unknown variances

- We have two samples $\mathbf{y}_A$ and $\mathbf{y}_B$
  - Wish to test $\mu_A - \mu_B$ (i.e., the difference in population means)
- An approximate p-value can be computed by substituting estimates $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ into the formulae for known variance
- This give us the test statistic

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$$

  which is approximately $N(0, 1)$ for large samples.
- We can then find approximate p-values using:

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|) & \text{if } H_0 : \mu_A = \mu_B \text{ vs } H_A : \mu_A \neq \mu_B \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_A \leq \mu_B \text{ vs } H_A : \mu_A > \mu_B \\ \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_A \geq \mu_B \text{ vs } H_A : \mu_A < \mu_B \end{cases}$$

- More exact but complicated procedures exist; `t.test()` in R implements some of these

Revision from Lecture 6

- Imagine we have measured $p + 1$ variables on n individuals (people, objects, things)
- We would like to predict one of the variables using the remaining p variables

- If the variable we are predicting is categorical, we are performing classification
  - Example: predicting if someone has diabetes from medical measurements.

- If the variable we are predicting is numerical, we are performing regression
  - Example: Predicting the quality of a wine from chemical and seasonal information.

Revision from Lecure 6 (1)

- Linear regression

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$$

  - β_0 is the intercept (value of $\mathbb{E}[Y]$ when all predictors are zero)
  - β_j is a coefficient (change in $\mathbb{E}[Y]$ per unit change in $x_{j,i}$)
- Residuals (errors)

$$e_i = y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \cdots - \beta_p x_{i,p}$$

- Residual sum-of-squares

$$\mathrm{RSS}(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2$$

- Least-squares estimates linear model by finding $\beta_0, \beta_1, \ldots, \beta_p$ that minimise the RSS

Revision from Lecure 6 (2)

- R^2 measure of goodness-of-fit

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

  where TSS is the sum of squared errors for the mean model

- Sometimes our predictors are categorical variables
  - We turn them into $K - 1$ new predictors (indicator variables)

- Sometimes predictor-target relationship is nonlinear
  - Nonlinear transformations of our variables can help
  - Polynomial transformations offer general purpose nonlinear fits
  - We turn our variable into q new variables of the form:

$$x_{i,j} \;\Rightarrow\; x_{i,j},\, x_{i,j}^2,\, x_{i,j}^3,\, \ldots,\, x_{i,j}^q$$

Revision from Lecture 7 (1)

- We also have p predictor variables $X_1, \ldots, X_p$
- But now our targets are binary ($0/1$, Yes/No, etc.)
- If all predictors are also categorical we can build a classifier for Y directly using conditional probability

$$P(y \,|\, x_1, x_2, \ldots, x_p) = \frac{P(y, x_1, \ldots, x_p)}{P(x_1, \ldots, x_p)}$$

- In practice we do not know the population distribution $P(y, x_1, \ldots, x_p)$; need to estimate from sample
- Weakness: too many probabilities to estimate as p grows

Revision from Lecture 7 (2)

- A logistic regression models the conditional log-odds as

$$\log\left(\frac{\mathbb{P}(Y_i = 1 \,|\, x_{i,1}, \ldots, x_{i,p})}{\mathbb{P}(Y_i = 0 \,|\, x_{i,1}, \ldots, x_{i,p})}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \equiv \eta_i$$

- Logistic regression model of conditional probability

$$\mathbb{P}(Y_i = 1 \,|\, x_{i,1}, \ldots, x_{i,p}) = \frac{1}{1 + \exp(-\eta_i)}$$

- Estimated using maximum likelihood
- Performance measures for classification
  - Classification error
  - Sensitivity and specificity
  - Area-under-the-curve (AUC)
  - Logarithmic loss

Revision from Lecture 8 (1)

- How many predictors should we include in our linear model?
- Underfitting
  - Omitting important predictors
  - Leads to systematic error ("bias") in predicting the target
- Overfitting
  - Including spurious predictors
  - Leads our model to "learn" noise and random variation
- Methods to trade off bias and variance
  - Hypothesis testing $\beta_j = 0$ vs $\beta_j \neq 0$
    - Multiple hypothesis testing problem, Bonferroni
  - Penalized likelihood – likelihood plus complexity penalty
    - AIC, KIC, BIC, RIC
  - Cross-validation

Revision from Lecture 8 (2)

- All-subsets selection:
  - Try all combination of predictors to model with smallest model selection criterion score
- Forward selection algorithm:
  1. Start with the empty model;
  2. Find the predictor that reduces info criterion by most
  3. If no predictor improves model, end.
  4. Add this predictor to the model
  5. Return to Step 2
- Backwards selection is related algorithm
  - Start with the full model and remove predictors

Revision from Lecture 8 (3)

- Statistical instability;
  - Small changes in data $\Rightarrow$ big changes in model
- Ridge regression, squared penalty on coefficients

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_\lambda\right) = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \left\{ \mathrm{RSS}(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

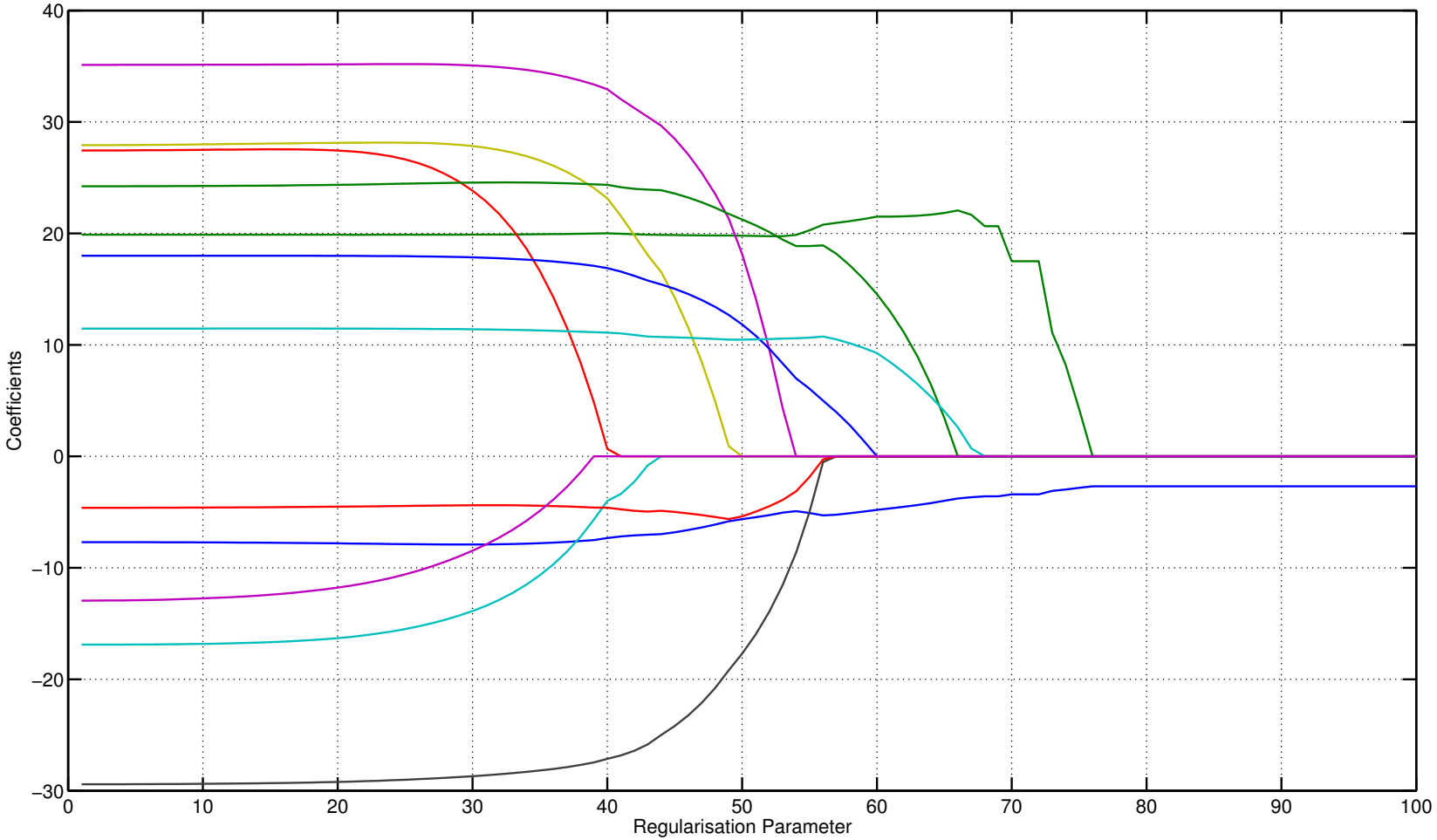
- Lasso regression, absolute penalty on coefficients

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_\lambda\right) = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \left\{ \mathrm{RSS}(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Lasso can estimate coefficients to be zero, ridge cannot
- Vary λ to get a "path" of different complexity models
  - $\lambda = 0$ most complex, $\lambda = \infty$ least complex
- Use cross validation to choose a good λ

Lasso Regression (3)

- Example lasso regression coefficient path



Revision from Lecture 9 (1)

- Machine learning methods
- Cross validation for model selection
  - Withhold data to estimate prediction error
  - K-fold CV divides data up into K equal sized groups
  - Train on $K-1$ folds, predict on the remaining fold
- Decision Trees
  - Split the data up by asking questions of the predictors
  - Number of leaves determines complexity of tree
  - Easy to interpret, flexible
- Methods for learning trees
  - Greedy growing of trees – find best split at each step
  - Backwards pruning of large tree
  - Use CV to select number of leaves in the tree

Revision from Lecture 9 (2)

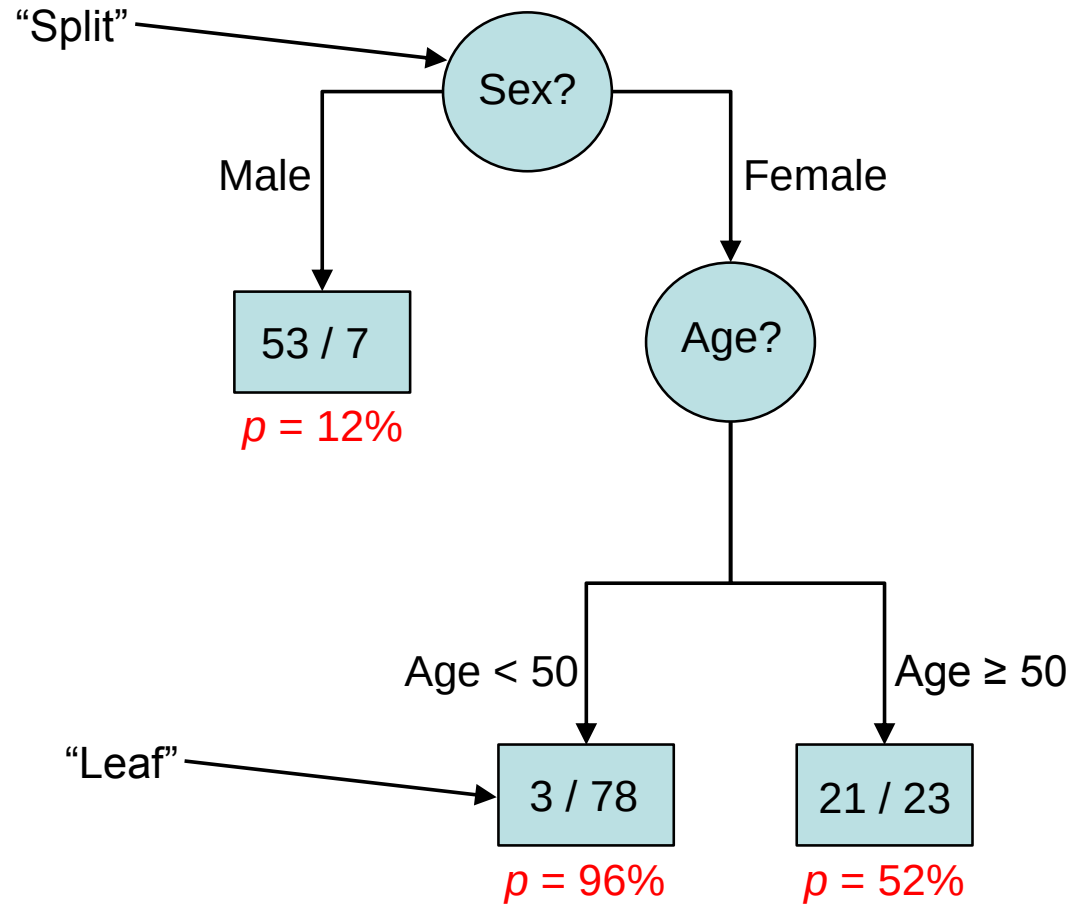
- Trees have low bias, high variance
- One solution: random forests
  - Grow many trees with guided random search
  - Aggregate predictions from the trees
  - Stable, low variance, but loses interpretability
- k-nearest neighbours (kNN) methods
  - Assume individuals similar in predictors are similar in targets
  - Find k "most similar" individuals in data to new individual
  - Use their targets to predict target for new individual
- Use CV to select k, other tuning parameters

K-fold Cross Validation

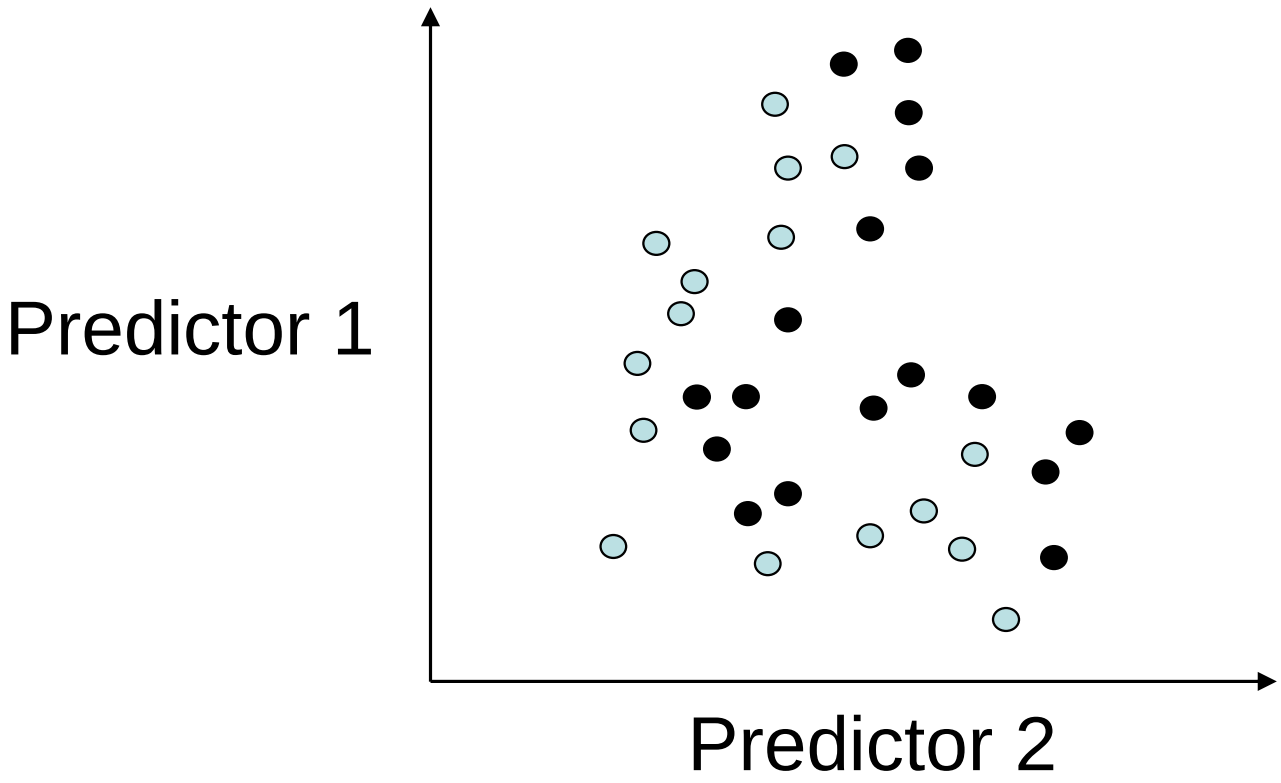
- Outer loop: try different model complexities γ
  1. For $i = 1$ to m
     1. Partition data into K equal sized, disjoint subsets
     $$\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)}, \ldots, \mathbf{y}^{(K)}$$
     2. For $k = 1$ to K
        1. Fit model $\mathcal{M}(\gamma)$ to all $\mathbf{y}^{(i)}$ except for $i = k$
        2. Use fitted model to predict onto $\mathbf{y}^{(k)}$
        3. Calculate and accumulate prediction errors
  2. Average all m accumulated prediction errors
- Once this is done, we have CV errors for each complexity γ
  $\Rightarrow$ choose the γ with the smallest estimated error
- The larger the m, the more stable the estimates (but slower)

Decision Trees

- Example tree: predicting high blood pressure

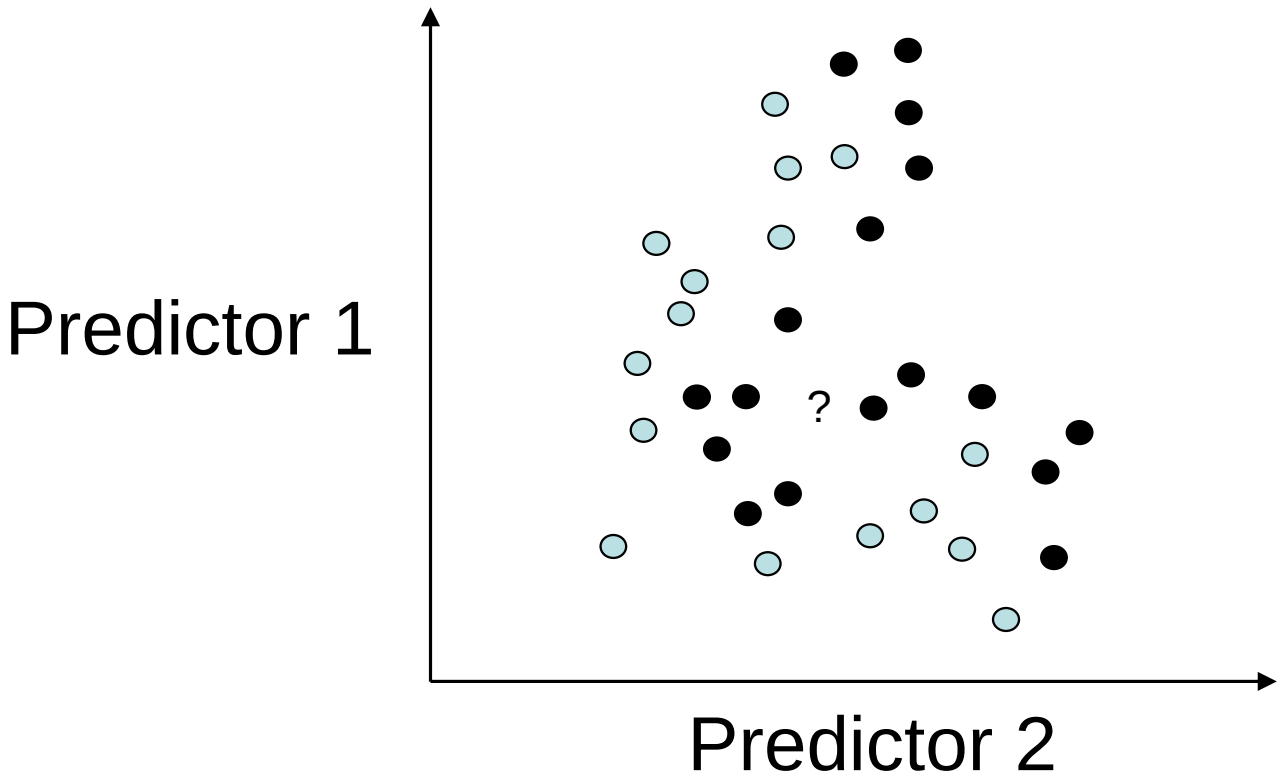


k-Nearest Neighbours Example 1 (1)



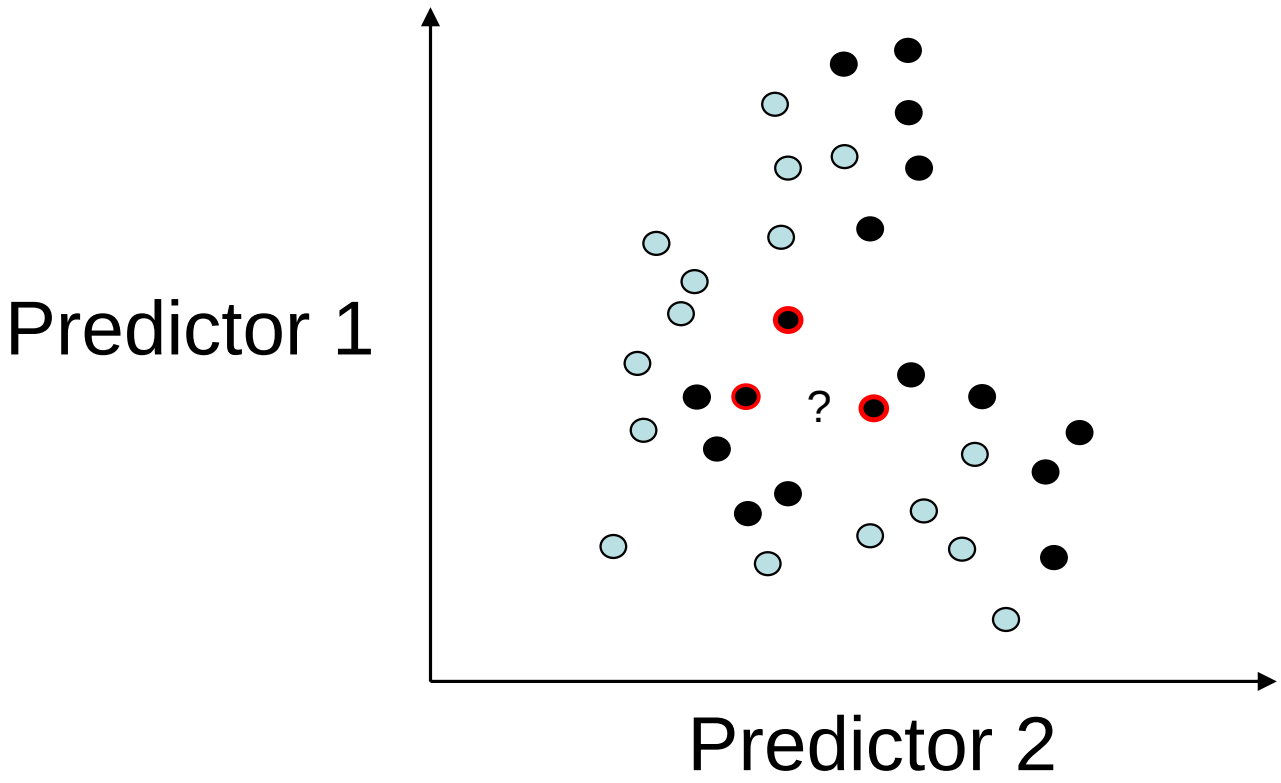
Example data set: black are individuals with disease, blue are those without

k-Nearest Neighbours Example 1 (2)



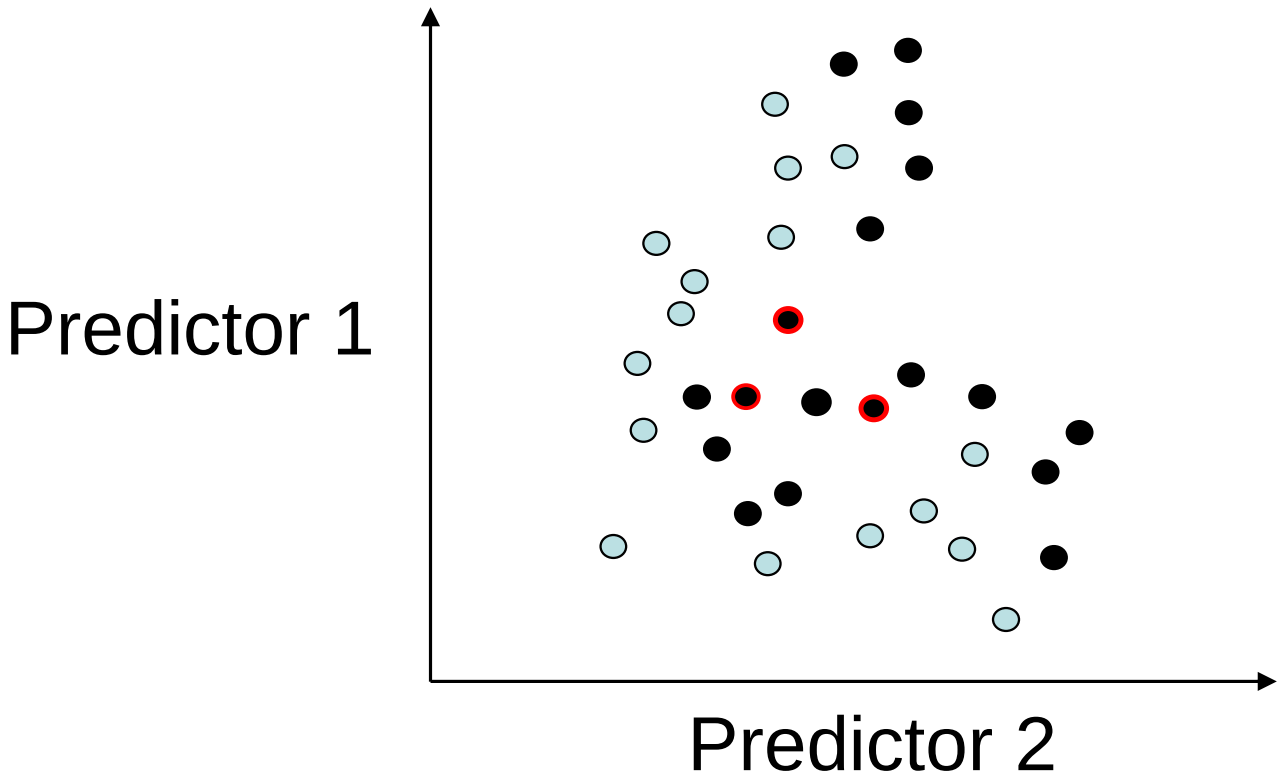
We want to predict the disease status of the individual marked with a "?" using a k nearest neighbour method with $k = 3$ neighbours

k-Nearest Neighbours Example 1 (3)



We find the closest $k = 3$ individuals

k-Nearest Neighbours Example 1 (4)



They all have the disease, so we predict that our new individual will also have the disease

Revision from Lecture 10 (1)

- Unsupervised Learning
- n individuals, q attributes
  - No "target"; instead, discover structure inherent in data
- Clustering
  - Model the population as K distinct sub-populations
  - Learn both K and the subpopulation parameters
- k-means clustering algorithm
  - Allocate individuals to closest clusters
  - Re-estimate cluster centres
  - Iterate until convergence

Revision from Lecture 10 (2)

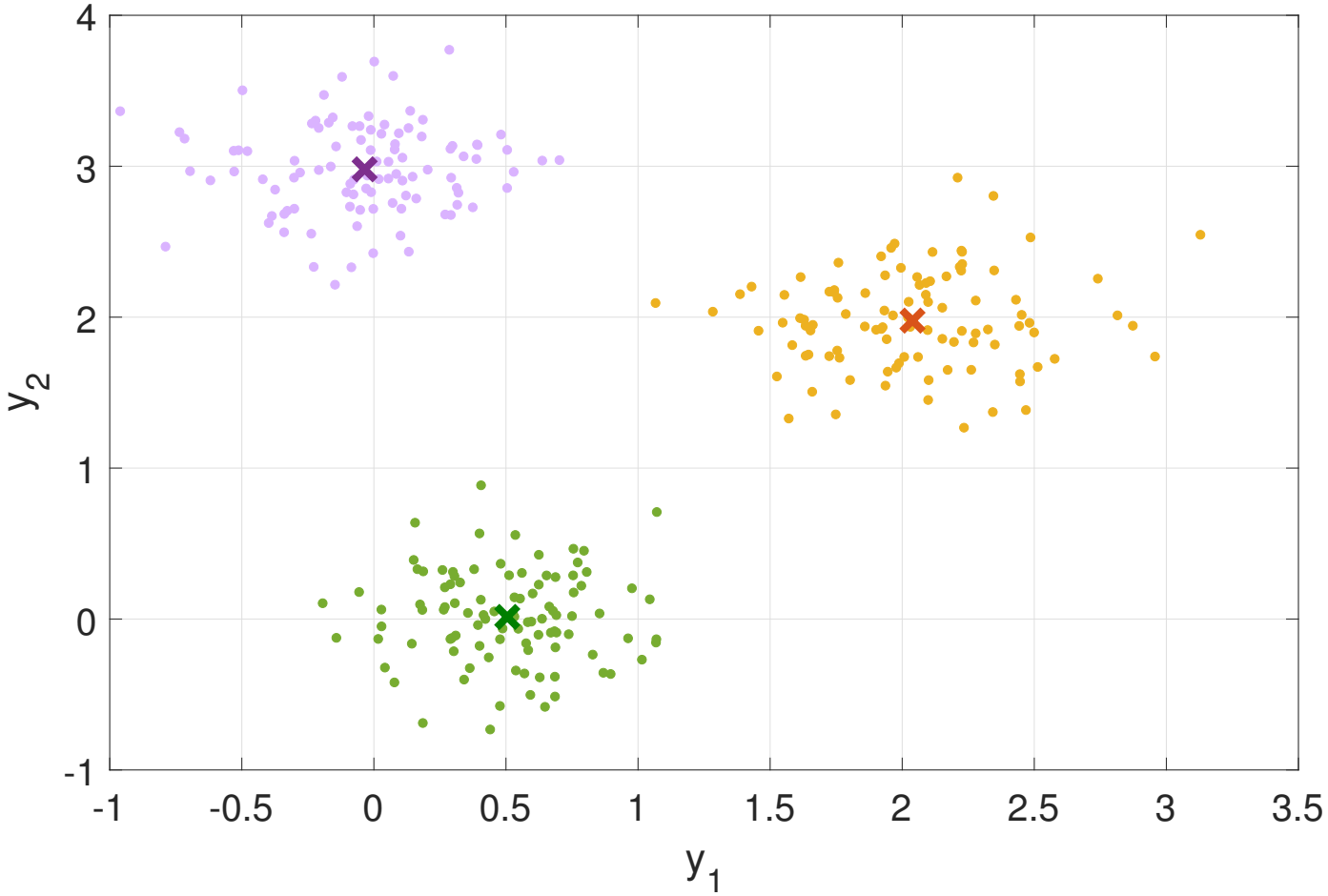
- Mixture modelling
  - Models data as a mixture of probability distributions (subpopulations)

$$p(y_{i,j}) = \sum_{k=1}^{K} \alpha_k p(y_{i,j} \mid \boldsymbol{\theta}_{k,j})$$

  - Each subpopulation characterised by distributions
  - "Soft" assignment of individuals to classes
  - Intrinsic classification
- Matrix completion
  - Missing entries in our data matrix
  - Try to estimate the values of the missing data
  - Applications: imputation, recommender systems
  - Mixture modelling approach, k-NN approaches

K-means Clustering

- Find the centroids that minimise within-cluster sum-of-squares



Beyond FIT2086 ...

- I hope you have enjoyed your first taste of data science, and learned a lot
- I have put some links up in week 12 to books/book chapters you can read if you want to learn more about data science
- FIT3154 Advanced data analysis is a good follow up if you enjoyed this
  - The Bayesian approach to statistical inference (probably the fastest growing area of data science)
  - Explore advanced models for data analysis/prediction, and what links different models together
  - Understand limits on learning
  - Learn *how* and *why* different models perform well (why does a neural network outperform a polynomial, and when will it fail to do so?)
- Also FIT3181 Deep Learning

Good luck!

- Good luck in the exam!