CPSC 4300/6300: Applied Data Science

2024

Missing data, feature selection, outliers

Dr. Aaron J. Masino

Associate Professor, School of Computing

Outline

Missing data imputation

Feature selection

Outlier detection

Missing data imputation

The problem of missing data

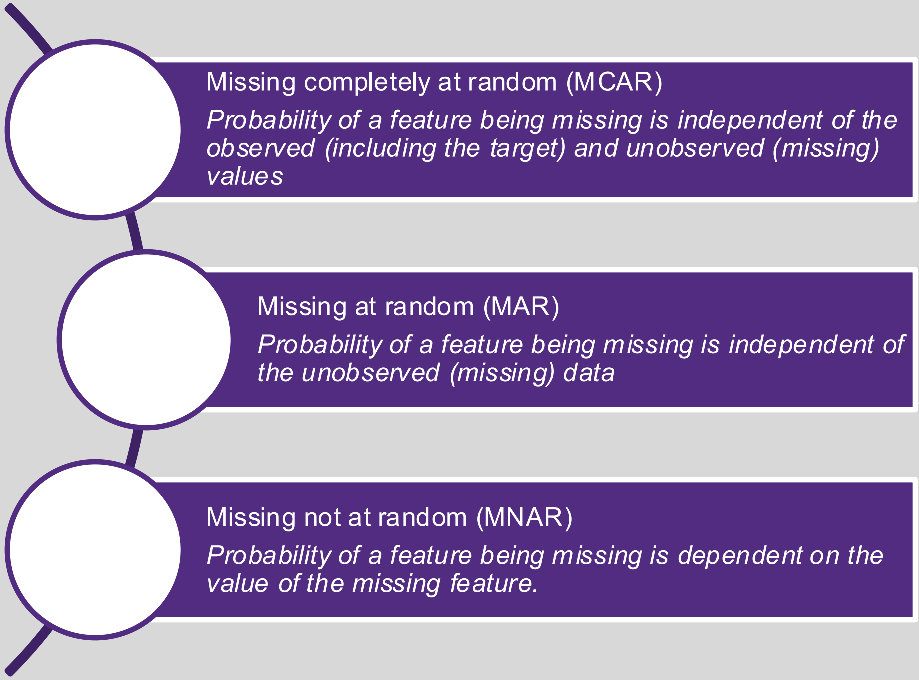
- A given sample may not have an observed value for one or more features
- Missing data can occur due to
  - Lack of compliance (e.g., patient did not complete form)
  - Measurement device failure
  - Tests not performed (frequent for retrospective analyses)
- Complete case analysis
  - Only use samples where all features are observed
  - May substantially reduce sample size
  - May introduce bias

Observed features					
Index	**Blood Type**	**SBP**	**DBP**	**Age**	**Target**
0	A-	140	90	47	1
1	B-	124	82		0
2	B-		88	31	0
3		188		67	1
4	AB-	108	76	21	1

Empty cells indicate features that are *missing* for the sample

Complete cases indicated in blue (indices 0 and 4)

Modes of missingness



Missing completely at random (MCAR)

Probability of a feature being missing is independent of the observed (including the target) and unobserved (missing) values

Missing at random (MAR)

Probability of a feature being missing is independent of the unobserved (missing) data

Missing not at random (MNAR)

Probability of a feature being missing is dependent on the value of the missing feature.

- What is the mode of my missing data?
  - MCAR - Little’s test can distinguish MCAR from not MCAR
  - No statistical test to distinguish MAR from MNAR
  - Domain experts (e.g., clinicians) may be able to assess the manner in which data is collected and advise if it is likely MAR
- Most imputation methods assume data is MCAR or MAR

Population statistics for imputation

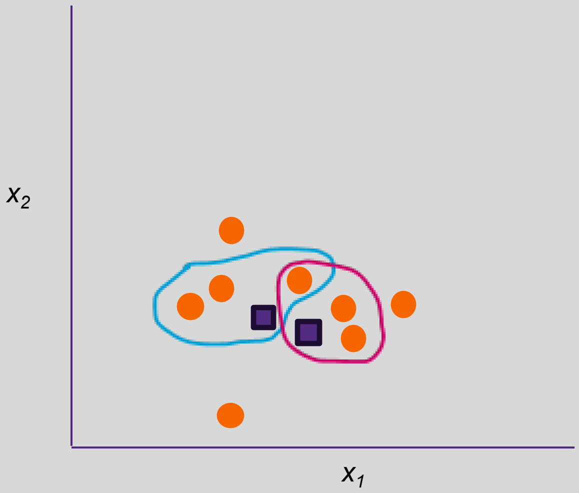
- Strategy is to replace missing values with a population statistic obtained from the samples for which the feature was observed
  - Mode – For categorical features, one can use the mode (most frequent value)
  - Mean / median – for quantitative features, one can use the mean or median
- **These methods are usually <u>not</u> a good choice**
  - They tend to introduce bias
  - Do not account for conditional relations of the missing feature with other observed features

Observed features					
Index	**Blood Type**	**SBP**	**DBP**	**Age**	**Target**
0	A-	140	90	47	1
1	B-	124	82		0
2	B-		88	31	0
3		188		67	1
4	AB-	108	76	21	1
Mode/ Mean	B-	140	84	41.5	X

Bottom row (green) indicates mode (categorical features) or mean (quantitative features). Typically, we do not impute target values (never in test set)

Nearest neighbors imputation (NNI)

- For a sample, s, with a missing value for feature p:
  - Find the k neighbors with a value for p that are closest to s based on some metric, m (e.g., Euclidean distance, Gower's distance)
  - Replace the missing value, p, for s with the mean/mode of p for the k nearest neighbors
  - Potential limitations
    - Uses only a limited subset of the data for each sample imputation
    - In high dimensional spaces (i.e., many features) all samples are far apart



Nearest Neighbors Imputation

- Samples have three features x_1, x_2, x_3
- Samples with missing feature x_3 are represented by purple squares
- Samples with observed feature x_3 are represented by orange circles
- The samples used for $k = 3$ nearest neighbors imputation for each sample are indicated by the bordered region
- Note that even though the two purple squares are closer to each other than some of the samples used for imputation, they cannot be used because feature x_3 is missing

Multivariate imputation by chained equations (MICE)

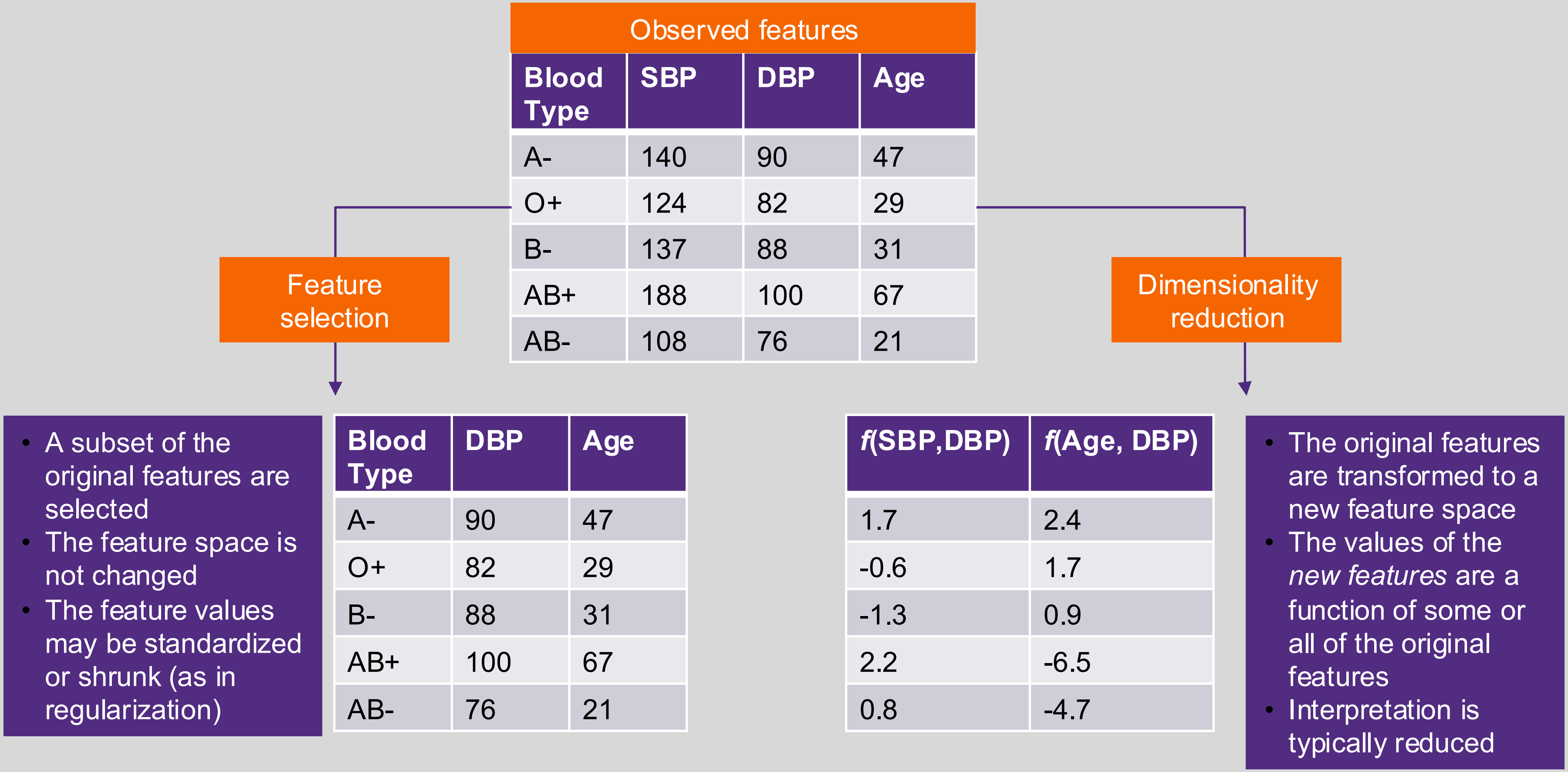
- Assume p features of which k have samples with missing values
- MICE method utilizes all the samples and all features to form estimates of the missing values
- Multiple imputation – method introduces randomness to form a distribution of imputed values for each sample
- In ML context, we typically take the first imputed value, though multiple samples can be used to evaluate model sensitivity to the imputed values

MICE with linear regression imputation model

1. For each of the k variables with missing values, fill in the missing values with random draws from the samples with values
2. For each variable, v_k, of the k variables with missing values:
   a. Perform linear regression to estimate v_k from all the other variables (using the true observations and the imputed observations)
   b. Randomly perturb the the linear regression coefficients
   c. For each sample s with missing v_k, use the perturbed coefficients to calculate a new v_k based on the feature values of s
3. Repeat step 2 for multiple cycles (typically 5-20). The imputed value for each sample from the last cycle represent the first imputed value
4. Repeat steps 1-3 m times to produce m imputed values for each missing observation

Pre-processing based feature selection

Feature selection vs. dimensionality reduction



Observed features

Blood Type	SBP	DBP	Age
A-	140	90	47
O+	124	82	29
B-	137	88	31
AB+	188	100	67
AB-	108	76	21

Feature selection

- A subset of the original features are selected
- The feature space is not changed
- The feature values may be standardized or shrunk (as in regularization)

Blood Type	DBP	Age
A-	90	47
O+	82	29
B-	88	31
AB+	100	67
AB-	76	21

Dimensionality reduction

f(SBP,DBP)	*f*(Age, DBP)
1.7	2.4
-0.6	1.7
-1.3	0.9
2.2	-6.5
0.8	-4.7

- The original features are transformed to a new feature space
- The values of the *new features* are a function of some or all of the original features
- Interpretation is typically reduced

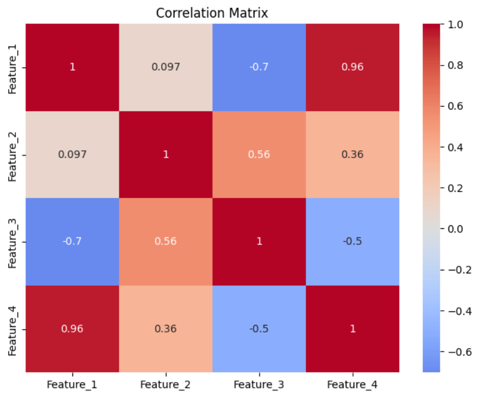
Why use feature selection?

- Improved model performance
  - Reduces overfitting (variance) particularly when the ratio $^{n}/_{p} < 10$ where n is the number of training samples and p is the number of features
  - Better test set performance
- Greater model interpretability
  - Linear models – all feature selection methods reduce the number of features in the model with non-zero coefficients
  - Non-linear models – interpretability improvement depends on the feature selection method

Unsupervised feature selection

- Performed as a data preprocessing step
- Uses the training data only to filter features
- Variance threshold – removes variables with low variance
  1. Select a threshold, α
  2. For each feature x in X, compute sample variance $var(x)$
  3. If $var(x) < \alpha$ remove x from the feature set X
- Collinearity threshold – removes one variable from a pair of highly correlated variables
  1. Select a threshold, γ (typically set 0.9 or 0.95)
  2. For each feature pair, (x_i, x_j) compute $cor(x_i, x_j)$
  3. If $cor(x_i, x_j)$> γ, remove x_i from the feature set X

Use variance thresholding with caution. It can lead to removing features that are important, particularly in the presence of class imbalance (classification) or skewed distributions (regression).



Correlation Matrix

	Feature_1	Feature_2	Feature_3	Feature_4
Feature_1	1	0.097	-0.7	0.96
Feature_2	0.097	1	0.56	0.36
Feature_3	-0.7	0.56	1	-0.5
Feature_4	0.96	0.36	-0.5	1

Filter methods for feature selection

- Performed as a data preprocessing step
- Using the training data only
- Perform a statistical test to determine association between individual features and target
- Select the K (count or percent) features with strongest statistical association

Which statistical comparison to use?

Input Feature Type	Target Type	Methods
Quantitative	Quantitative (regression)	Pearson, Spearman, Mutual Information
Qualitative	Quantitative (regression)	ANOVA, Mutual Information
Mixed	Quantitative (regression)	Mutual Information
Quantitative	Qualitative (classification)	ANOVA, Mutual Information
Qualitative	Qualitative (classification)	Chi-Squared, Mutual Information
Mixed	Qualitative (classification)	Mutual Information

ANOVA : tests linear association
Mutual Information : non-parametric, captures any association

Outlier detection

Motivation for Outlier Analysis

Fraud Detection (Credit card, telecommunications, criminal activity in e-Commerce)

Customized Marketing (high/low income buying habits)

Medical Treatments (unusual responses to various drugs)

Analysis of performance statistics (professional athletes)

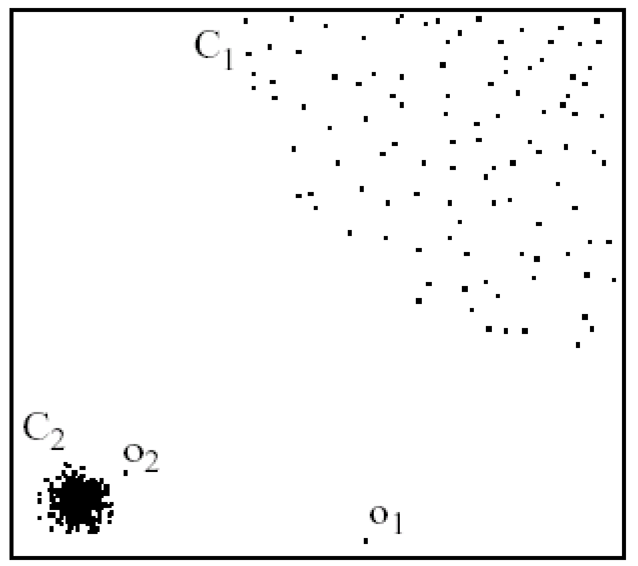
Weather Prediction

Financial Applications (loan approval, stock tracking)

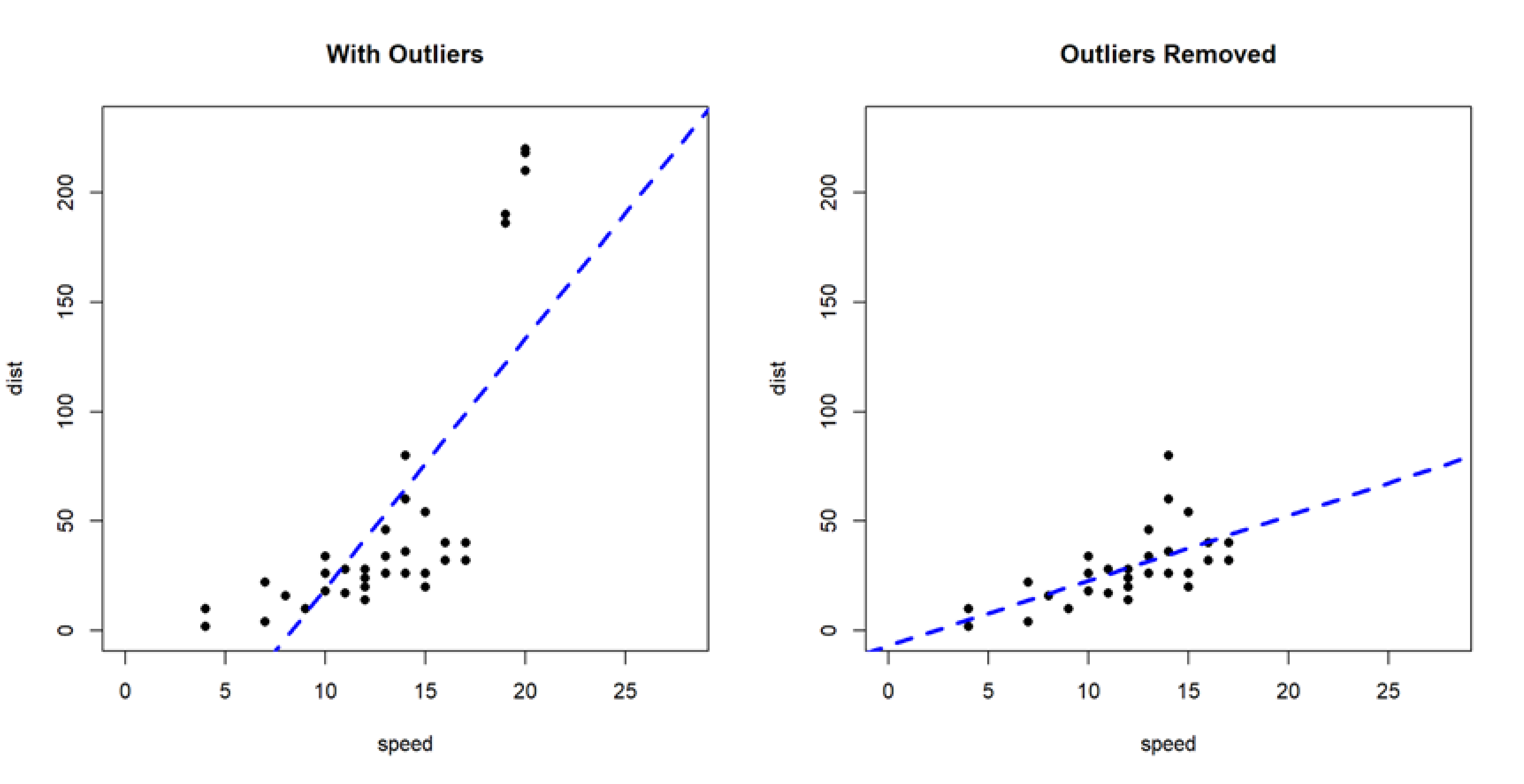
“One person’s noise could be another person’s signal.”

What is an outlier?

- Observations inconsistent with rest of the dataset – Global Outlier
- Special outliers – Local Outlier
  - Observations inconsistent with their neighborhoods
  - A local instability or discontinuity



Why are outliers important?



Causes of Outliers

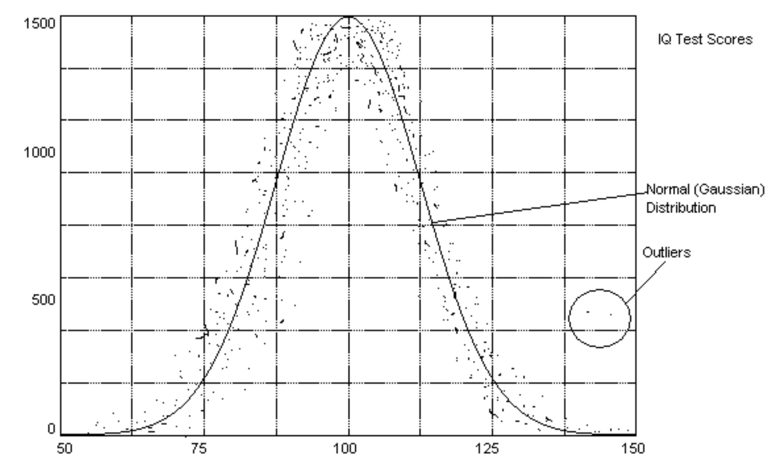
Poor data quality / contamination

Low quality measurements, malfunctioning equipment, manual error

Correct but exceptional data

Statistical-Based Outlier Detection (Distribution-based)

- Assumptions:
  - Knowledge of data (distribution, mean, variance)
- Statistical discordancy test
  - Data is assumed to be part of a working hypothesis (working hypothesis)
  - Each data object in the dataset is compared to the working hypothesis and is either accepted in the working hypothesis or rejected as discordant into an alternative hypothesis (outliers)



Working Hypothesis: $H : o_i \in F$, where $i = 1,2,\ldots,n.$

Discordancy Test: is o_i in F within standard deviation $= 15$

Alternative Hypothesis:

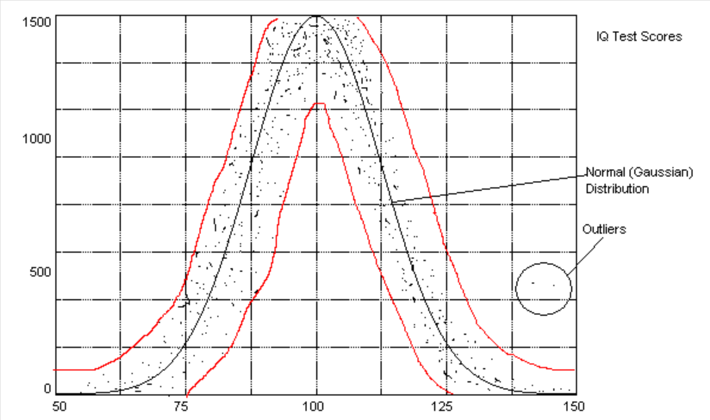
-Inherent Distribution: $\overline{H} : o_i \in G$, where $i = 1,2,\ldots,n.$

-Mixture Distribution: $\overline{H} : o_i \in (1-\lambda)F + \lambda G$, where $i = 1,2,\ldots,n.$

-Slippage Distibution: $\overline{H} : o_i \in (1-\lambda)F + \lambda F'$, where $i = 1,2,\ldots,n.$

Statistical-Based Outlier Detection (Distribution-based)

- Assumptions:
  - Knowledge of data (distribution, mean, variance)
- Statistical discordancy test
  - Data is assumed to be part of a working hypothesis (working hypothesis)
  - Each data object in the dataset is compared to the working hypothesis and is either accepted in the working hypothesis or rejected as discordant into an alternative hypothesis (outliers)



Working Hypothesis: $H: o_i \in F$, where $i = 1,2,\ldots,n.$

Discordancy Test: is o_i in F within standard deviation $= 15$

Alternative Hypothesis:

-Inherent Distribution: $\overline{H}: o_i \in G$, where $i = 1,2,\ldots,n.$

-Mixture Distribution: $\overline{H}: o_i \in (1-\lambda)F + \lambda G$, where $i = 1,2,\ldots,n.$

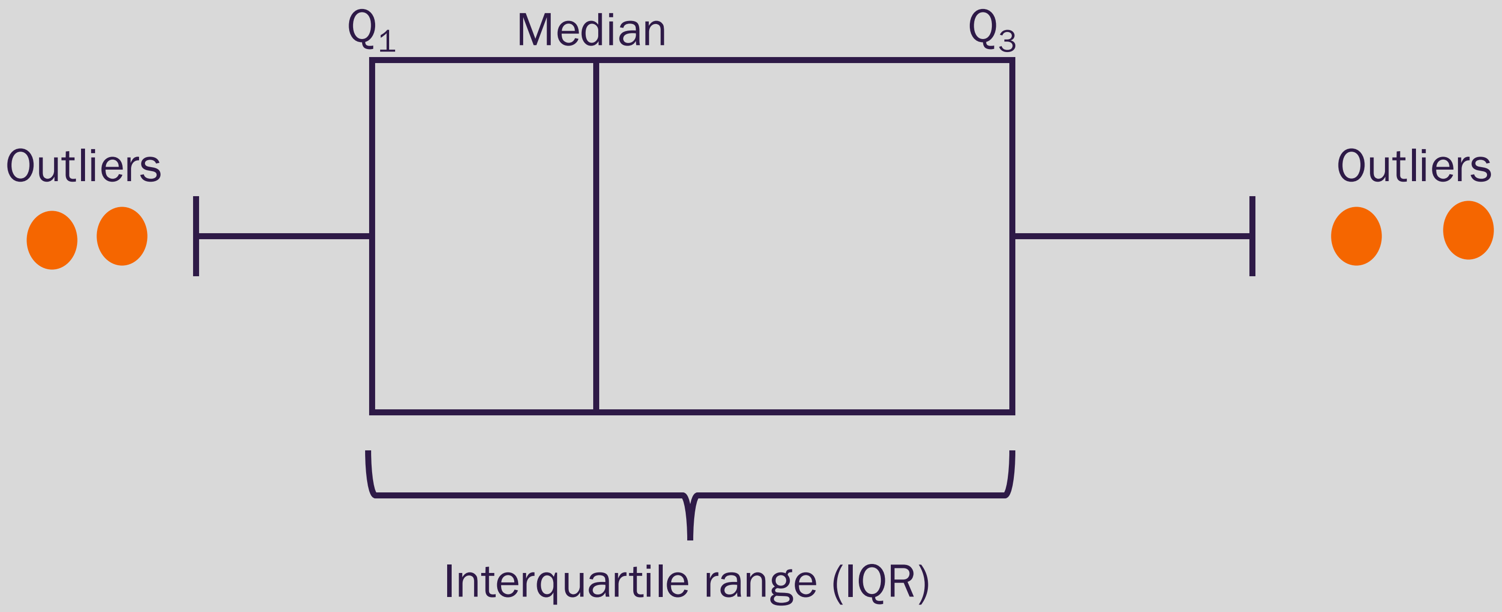
-Slippage Distibution: $\overline{H}: o_i \in (1-\lambda)F + \lambda F'$, where $i = 1,2,\ldots,n.$

Which values are statistical outliers?

Establish a boundary based on the IQR

- Lower boundary: $Q_1 - 1.5IQR$
- Upper boundary: $Q_3 + 1.5IQR$

Values outside the boundary are considered outliers



Statistical-Based Outlier Detection

Strengths

- Most outlier research has been done in this area, many data distributions are known

Weakness

- Almost all of the statistical models are univariate (only handle one attribute) and those that are multivariate only efficiently handle k<4
- All models assume the distribution is known –this is not always the case
- Outlier detection is completely subjective to the distribution used

Deviation-Based Outlier Detection

Simulate a mechanism familiar to human being: after seeing a series of similar data, an element disturbing the series is considered an exception

We'll consider Sequential Exception Techniques though there are other methods

Sequential Exception

- Select subsets of data I_j (j=1,2,…,n) from the dataset I
- Compare the dissimilarity of I and ($I-I_j$)

$$D = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

- Find minimum subset I_j that most reduces dissimilarity
- Smoothing factor

$$SF(I_j) = c(I - I_j) \times [D(I) - D(I - I_j)]$$

  - D is a dissimilarity function
  - C is a cardinality function, for example, the number of elements in the dataset

Example

Let the data set I be the set of integer values {1,4,4,4}

I_j	$I- I_j$	$C(I- I_j)$	$D(I- I_j)$	$SF(I_j)$
{}	{1,4,4,4}	4	1.69	0.00
{4}	{1,4,4}	3	2.00	-0.93
{4,4}	{1,4}	2	2.25	-1.12
{4,4,4}	{1}	1	0.00	1.69
{1}	{4,4,4}	3	0.00	5.07
{1,4}	{4,4}	2	0.00	3.38
{1,4,4}	{4}	1	0.00	1.69

Note, when I_j = {}, D(I) = D(I-I_j) = 1.69, SF(I_j)=0

When I_j={1}, SF(Ij) has the maximum value, so {1} is the outlier set

Questions?