3.3 Statistical Outlier Detection

3.3.1 Intro

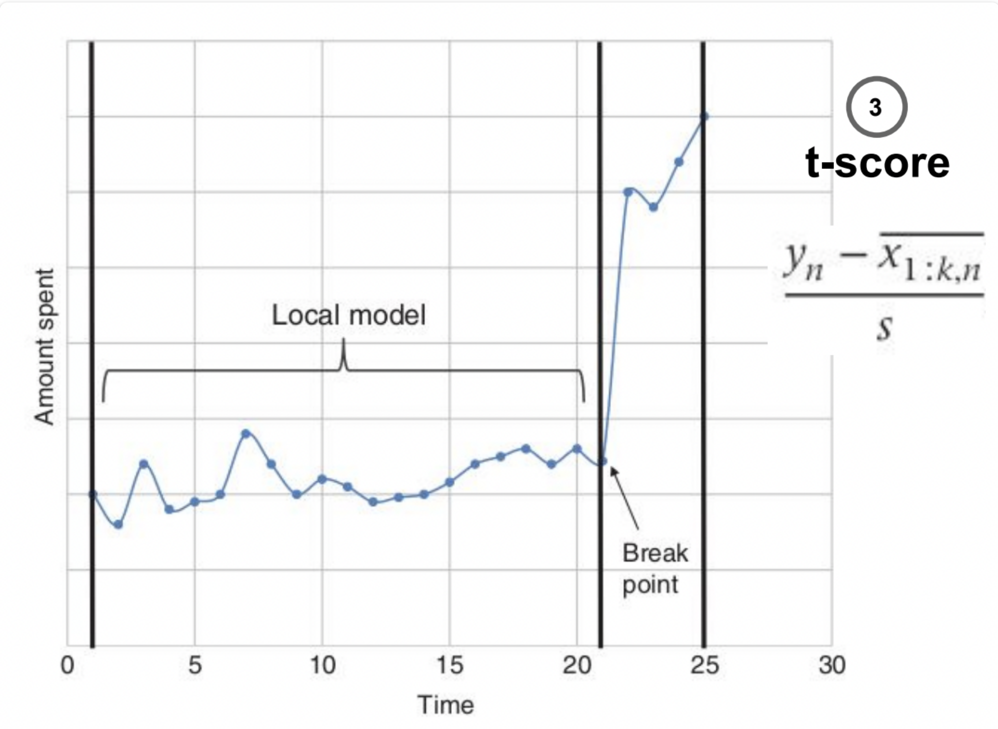
- *Z-Score*:
  - Recall: If the absolute value of the z-score is bigger than 3 can be considered as outliers.
- *Fit a distribution, or mixture of distributions*:
  - **Outliers**: observations with small values for the probability density function.
- *Break Point Analysis*.
- *Peer Group Analysis*.
- *Association Rule*.

3.3.2 Break-Point Analysis #Break-Point

#DEF **Break point indicates a sudden change in account behavior.**
We are talking about intra-account fraud detection method.

1. *Define* a fixed time window.
2. *Split* it into an "old" and "new" part.
3. *Compare* the new part with the old part.

Old part = local profile against which new observations are *compared*.



3.3.3 Peer-Group Analysis #Peer-Group

#DEF **Peer group is a group of accounts that behave similarly to the target account.**

> *When the behavior of the target account **deviates substantially** from its peers, an anomaly can be signaled.*

Peer-group analysis proceeds in two steps:

1. Peer group *identification*.
2. Anomaly *Evaluation*.

3.3.3.1 Peer-Group Analysis Steps

1. The **peer group** of a particular account **is identified**.
   - Prior *business knowledge*.
   - *Statistical* way:
     - Statistical similarity metrics (Euclidean-based metrics).
2. **Number of peers**:
   - Too *small* (too local): sensitive to noise.
   - Too *large* (too global): insensitive to local important irregularities.
3. The **behavior of the target account is contrasted with its peers**:
   - *Statistical test* (e.g., Student's t-test).
   - *Distance metric* (e.g., Mahalanobis Distance).

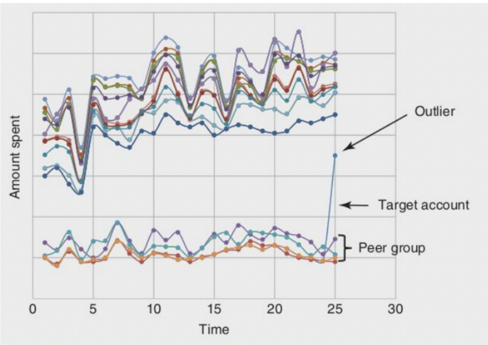
Credit Card Fraud Example:

❓ Verify whether the amount spent at time n is anomalous.

Step 1: identifying the k peers of the target account.

Step 2: Behavior comparison

t-score $\frac{y_n - \overline{x_{1:k,n}}}{s}$



3.3.4 Peer-group vs Break-point Analysis

Break-point analysis	*Peer-group analysis*
Tracks anomalies by considering **intra-account behavior**.	Tracks anomalies by considering **inter-account behavior**.

In the Christmas Period example:

Both break-point and peer-group analysis *will detect local anomalies* rather than global anomalies.

3.3.5 Association Rule Analysis #AssociationRule

💡 #IDEA **Detect frequently occurring relationships between items**

🔑 **Key input**: Transactions *database D* consisting of a *transaction identifier and a set of items I*.

An **association rule** is then an implication of the form **X ⇒ Y**,
whereby *X ⊂ I, Y ⊂ I and X ∩ Y = ∅*.

- X Rule antecedent.
- Y Rule consequent.

> ***Association rules are stochastic in nature**: statistical measures quantifying the strength of the association.*

3.3.5.1 Frequency, Support and Confidence

#DEF The **frequency** of an item set is measured by means of its **support**, *which is the percentage of total transactions in the database that contains the item set*.

$$support(x) = \frac{\#\ of\ transactions\ supporting(x)}{total\ \#\ of\ transactions}$$

#DEF **Frequent item set**: An *item set* with a *support higher than a minimum value* specified by the data scientist (e.g., 10%).

#DEF The **confidence** measures the *strength of the association* and is defined as the *conditional probability of the rule consequent, given the rule antecedent*.

$$confidence(X => Y) = P(Y|X) = \frac{support(X \cup Y)}{support(X)}$$

The data scientist has to specify a *minimum confidence* in order for an *association rule to be considered interesting*.

3.3.5.2 Insurance Fraud Example

📖 **Goal**: find *frequently occurring relationships/association rules* between the various parties involved.

Step 1: **Identify the frequent item sets**.
Item set: {insured A, police officer X, auto repair shop 1}.
Support = 3/10 -> *30%*

Claim Identifier	Parties Involved
1	insured A, police officer X, claim adjuster 1, auto repair shop 1
2	insured A, claim adjuster 2, police officer X
3	insured A, police officer Y, auto repair shop 1
4	insured A, claim adjuster 1, claim adjuster 1, police officer Y
5	insured B, claim adjuster 2, auto repair shop 2, police officer Z
6	insured A, auto repair shop 2, auto repair shop 1, police officer X
7	insured C, police officer X, auto repair shop 1
8	insured A, auto repair shop 1, police officer Z
9	insured A, auto repair shop 1, police officer X, claim adjuster 1
10	insured B, claim adjuster 3, auto repair shop 1

Step 2: **Derive Association Rules**.
Multiple association rules can be defined based on the same item set:

- If insured A and police officer X ⇒ auto repair shop 1
- If insured A and auto repair shop 1 ⇒ police officer X
- If insured A ⇒ auto repair shop 1 and police officer X

Recall: The *strength of an association rule* can be quantified by means of its **Confidence**.

⤴ "If insured A and police officer X ⇒ auto repair shop 1."

Antecedent item set: {insured A, police officer X} *occurs in 4 transactions*.

Claim Identifier	Parties Involved
1	insured A, police officer X, claim adjuster 1, auto repair shop 1
2	insured A, claim adjuster 2, police officer X
3	insured A, police officer Y, auto repair shop 1
4	insured A, claim adjuster 1, claim adjuster 1, police officer Y
5	insured B, claim adjuster 2, auto repair shop 2, police officer Z
6	insured A, auto repair shop 2, auto repair shop 1, police officer X
7	insured C, police officer X, auto repair shop 1
8	insured A, auto repair shop 1, police officer Z
9	insured A, auto repair shop 1, police officer X, claim adjuster 1
10	insured B, claim adjuster 3, auto repair shop 1

3 out of 4 transactions contain the consequent item set {auto repair shop 1}, **Confidence** = *75%*

Next chapter: Clustering