

# Korelačná Analýza

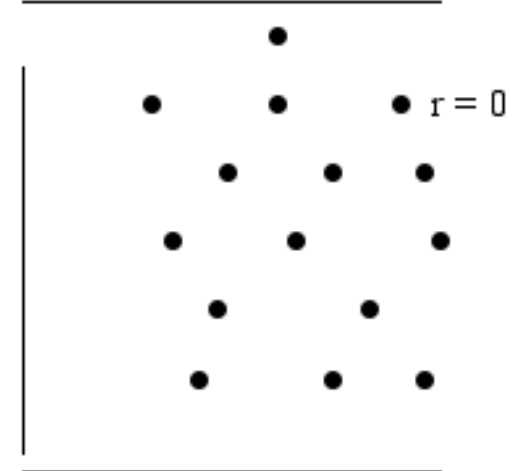
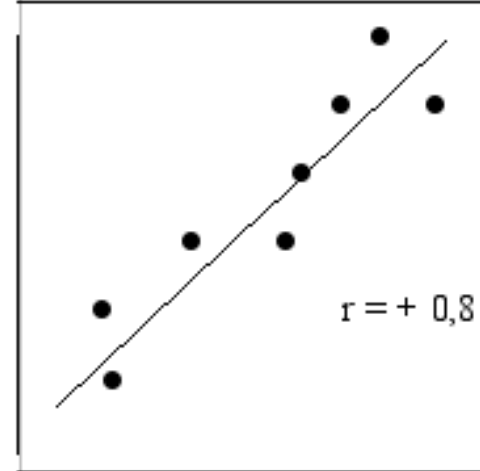
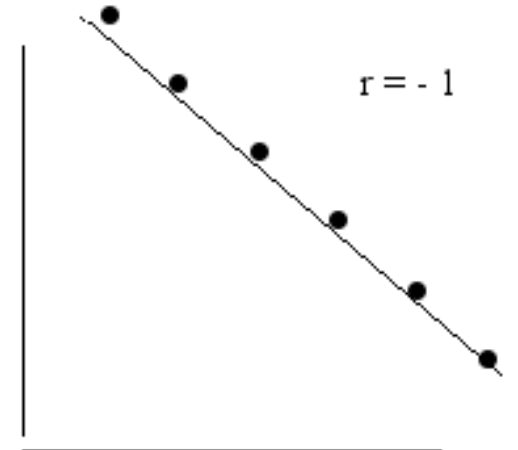
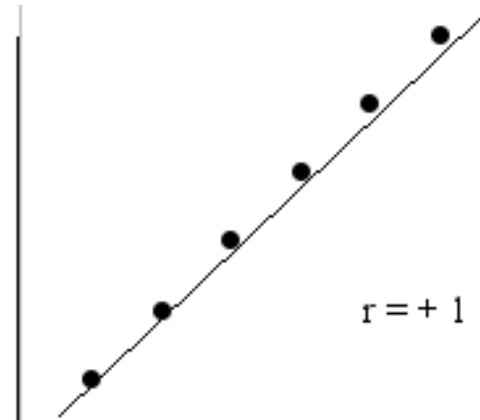
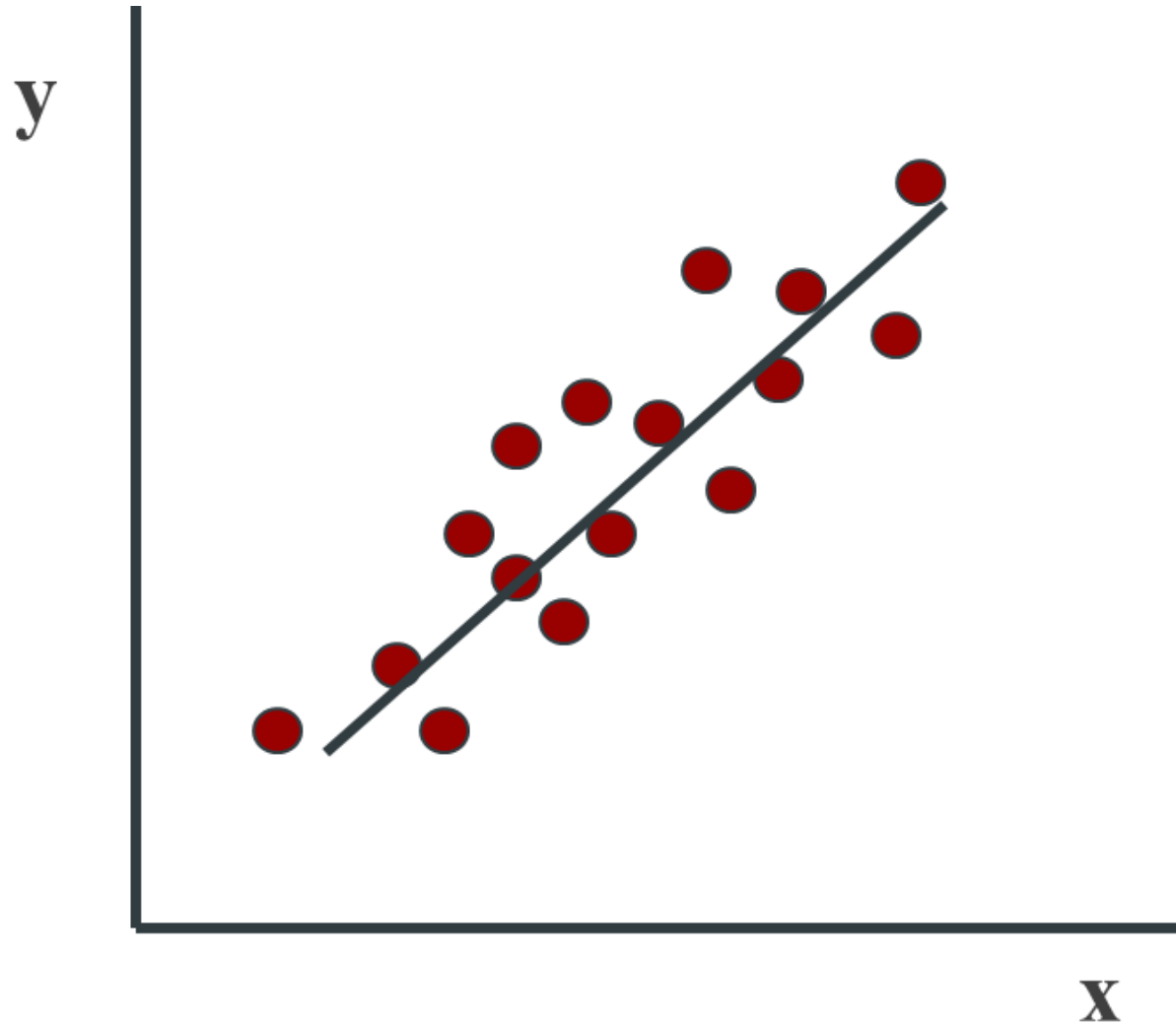


# Čo je to Korelácia?



- **Štatistická metrika**
- Vyjadruje **vzájomný vzťah medzi 2 premennými**
- Popisuje, do akej miery sa **zmeny jednej premennej prejavujú v zmenách druhej premennej**
- Používa sa na **zistovanie, či existuje spojitosť medzi údajmi v štatistických analýzach a strojovom učení**

# Korelační Analýza



# Ako na Meranie Korelácie

## 1. Pearsonov korelačný koeficient ( $r$ )

- Najbežnejší spôsob merania korelácie
- Určuje, ako silno a v akom smere súvisia 2 premenné
- Predpokladá, že vzťah medzi nimi je lineárny (-1 až +1)

## 2. Spearmanov korelačný koeficient ( $\rho$ )

- Keď vzťah medzi premennými nemusí byť priamo lineárny, ale stále existuje monotónna závislosť
- Keď jedna premenná rastie, druhá má tendenciu rásť alebo klesať, ale nie nevyhnutne lineárne

## 3. Kendallov tau ( $\tau$ )

- Podobný Spearmanovej korelácii, ale je založený na počte usporiadaných dvojíc
- Používa sa hlavne v prípadoch, keď máme poradové dáta (ranky) a chceme zistiť, či medzi nimi existuje vzťah
- Používa sa pri poradí hodnôt

# Pearsonov Korelačný Koeficient (r)



- **Vlastnosti:**

- Hodnota sa pohybuje v rozmedzí -1 až +1
- $r > 0$  → pozitívna korelácia (ak jedna premenná rastie, druhá tiež rastie)
- $r < 0$  → negatívna korelácia (ak jedna premenná rastie, druhá klesá)
- $r = 0$  → žiadna korelácia (medzi premennými nie je lineárny vzťah)
- Je citlivý na extrémne hodnoty (outlieri), ktoré môžu výrazne skresliť výsledok

- **Použitie:**

- Výskum v ekonómii (vzťah medzi príjmom a útratou)
- Analýza finančných trhov (závislosť medzi cenou akcií a objemom obchodovania)
- Sociálne vedy (vzťah medzi vekom a výškou príjmu)

- **Príklad:**

- Ak máme vzťah medzi počtom hodín učenia a výsledkami testu, Pearsonova korelácia nám povie, či vyšší počet hodín učenia znamená lepšie výsledky v teste.

# Pearsonov Korelačný Koeficient

## IT a Data Science



### Výkon serverov a doba odozvy

- Meranie vzťahu medzi **počtom požiadaviek za sekundu (load na serveri)** a **dobou odozvy aplikácie**
- Ak  $r \approx -0,8 \rightarrow$  viac serverových zdrojov vedie k nižšej dobe odozvy (negatívna korelácia)



### Počet chýb v kóde a skúsenosti programátora

- Analyzovanie, či viac rokov skúseností programátora znamená menej chýb v kóde
- Ak  $r \approx -0,6 \rightarrow$  skúsenejší programátori robia menej chýb.

## Marketing



### Reklamné výdavky a počet nových zákazníkov

- Skúmanie, či väčšie výdavky na online reklamu vedú k vyššiemu počtu nových zákazníkov
- Ak  $r \approx 0,7 \rightarrow$  vyššie výdavky na reklamu súvisia s vyšším prírastkom zákazníkov



### Čas strávený na webe vs. konverzný pomer

- Analýza, či zákazníci, ktorí trávajú viac času na webe, majú vyššiu pravdepodobnosť nákupu
- Ak  $r \approx 0,5 \rightarrow$  mierna pozitívna korelácia

# Pearsonov Korelačný Koeficient

## • Manažment

### ✓ Produktivita zamestnancov a dĺžka pracovného času

- Skúmanie, či dlhší pracovný čas vedie k vyššej produktivite
- Ak  $r \approx 0,2 \rightarrow$  slabá korelácia (pretože po určitej dobe výkon klesá kvôli únave)

### ✓ Výška platu a spokojnosť zamestnanca

- Skúmanie, či vyššie mzdy vedú k vyššej spokojnosti zamestnancov
- Ak  $r \approx 0,4 \rightarrow$  mierna pozitívna korelácia (nie všetko závisí od platu, dôležité sú aj benefity, pracovná kultúra)

# Spearmanov Korelačný Koeficient ( $\rho$ )



- **Vlastnosti:**

- Používa sa pre ordinálne (poradové) dáta alebo keď dáta nemajú normálne rozdelenie
- Nevyžaduje lineárny vzťah, stačí, aby bol monotónny
- Odolnejší voči outlierom ako Pearsonov koeficient
- Hodnota sa pohybuje od -1 do +1 podobne ako Pearsonov koeficient

- **Použitie:**

- Porovnávanie rebríčkov (napr. vzťah medzi rebríčkom univerzít a spokojnosťou študentov)
- Biologické merania (napr. vzťah medzi znečistením ovzdušia a výskytom respiračných chorôb)
- Analýza zákaznickej spokojnosti na stupnici od 1 do 10

- **Príklad:**

- Ak porovnáme výšku a výkon v basketbale, vzťah nemusí byť úplne lineárny, ale vyšší hráči majú tendenciu podávať lepšie výkony. Spearmanova korelácia to dokáže odhaliť.



# Spearman Korelačný Koeficient

## IT a Data Science

### ✓ Poradie stránok vo výsledkoch vyhľadávania a organická návštevnosť

- Vyhodnocovanie, či vyššie umiestnenie stránky vo vyhľadávači (Google) vedie k vyššej návštevnosti
- Ak  $\rho \approx 0,9 \rightarrow$  silná monotónna korelácia (vyššie pozície prinášajú viac návštevníkov)

### ✓ Počet testovacích prípadov a stabilita softvéru

- Overovanie, či viac testovacích scenárov vedie k stabilnejšiemu softvéru
- Ak  $\rho \approx 0,6 \rightarrow$  stredná monotónna korelácia

## Marketing

### ✓ Pozícia produktu v rebríčku predajnosti a počet predaných kusov

- Skúmanie, či produkty na vyšších priečkach v TOP 10 rebríčkoch e-shopov majú vyšší predaj
- Ak  $\rho \approx 0,8 \rightarrow$  silná monotónna korelácia

### ✓ Poradie influencerov podľa sledovateľov a ich marketingový vplyv

- Porovnávanie počtu sledovateľov influencerov s úspešnosťou ich kampaní
- Ak  $\rho \approx 0,5 \rightarrow$  stredná korelácia (nie vždy viac followerov znamená väčší vplyv)

# Spearman Korelačný Koeficient

## Manažment

### ✓ Hodnotenie výkonnosti zamestnancov a šanca na povýšenie

- Skúmanie, či lepšie hodnotení zamestnanci majú vyššiu šancu na povýšenie
- Ak  $\rho \approx 0,7 \rightarrow$  stredne silná pozitívna korelácia

### ✓ Poradie firemných benefitov podľa dôležitosti a spokojnosť zamestnancov

- Analyzovanie, ktoré benefity sú najviac preferované zamestnancami
- Ak  $\rho \approx 0,6 \rightarrow$  mierna korelácia (nie všetky benefity ovplyvňujú spokojnosť rovnako)

# Kendallov Tau ( $\tau$ )



- **Vlastnosti:**

- Hodnota sa pohybuje medzi -1 a +1
- Vhodný na malé vzorky dát
- Menej citlivý na veľké rozdiely v hodnotách než Spearmanov koeficient
- Odolný voči nesprávne zoradeným údajom

- **Použitie:**

- Športové rebríčky (napr. či vyššie umiestnenie v kvalifikácii vedie k lepším výsledkom v hlavnom turnaji)
- Sociálne vedy (napr. poradie kandidátov v anketách verzus ich skutočné zvolenie)
- Lekársky výskum (napr. vzťah medzi hodnotením príznakov a pravdepodobnosťou diagnózy)

- **Príklad:**

- Ak máme zoznam kandidátov zoradených podľa hlasov od 2 rôznych porotcov, Kendallov tau nám povie, či majú podobné poradie (či si porotcovia hodnotenia "nerozhadzujú")

# Kendallov Tau

## IT a Data Science

### ✓ Poradie bugov podľa závažnosti a počet nahlásení

- Skúmanie, či závažnejšie chyby majú viac hlásení od používateľov
- Ak  $\tau \approx 0,75 \rightarrow$  silná zhoda medzi poradím závažnosti a počtom hlásení

### ✓ Poradie programovacích jazykov v popularite a počet nových vývojárov

- Porovnávanie rebríčkov najpopulárnejších jazykov (napr. Python, JavaScript) s nárastom počtu nových vývojárov
- Ak  $\tau \approx 0,6 \rightarrow$  mierna zhoda

## Marketing

### ✓ Poradie produktov podľa recenzií a predajnosť

- Overenie, či produkty s lepšími recenziami sa predávajú viac
- Ak  $\tau \approx 0,7 \rightarrow$  silná zhoda

### ✓ Poradie sociálnych médií podľa angažovanosti a počet reklám

- Analyzovanie, či sociálne siete s vyššou angažovanosťou majú viac reklám
- Ak  $\tau \approx 0,5 \rightarrow$  mierna korelácia

# Kendallov Tau

## Manažment

### ✓ Poradie kandidátov v hodnotení HR a ich skutočné prijatie

- Skúmanie, či kandidáti odporúčaní HR majú vyššiu pravdepodobnosť prijatia do firmy
- Ak  $\tau \approx 0,8 \rightarrow$  silná korelácia




### ✓ Poradie oddelení podľa spokojnosti zamestnancov a fluktuácia

- Zisťovanie, či oddelenia s nižšou spokojnosťou majú vyššiu mieru odchodov zamestnancov.
- Ak  $\tau \approx -0,6 \rightarrow$  stredná negatívna korelácia (nižšia spokojnosť = vyššia fluktuácia)

# Prehľad Korel. Koeficientov

Typ korelácie	Kedy sa používa	Charakteristika vzťahu	Odolnosť voči outlierom
Pearsonov (r)	Lineárne vzťahy	Lineárna závislosť	Citlivý na extrémne hodnoty
Spearmanov (ρ)	Monotónne vzťahy, ordinálne dáta	Poradie hodnôt, nie priame hodnoty	Odolnejší voči outlierom
Kendallov (τ)	Porovnanie poradí	Koľko dvojíc je zoradených rovnako	Menej citlivý na odľahlé hodnoty

# Použitie Korel. Koeficientov

Korelácia	Typ vzťahu	Použitie
 <b>Pearson</b> (r)	Lineárny vzťah	Výkonnosť serverov vs. doba odozvy, reklamné výdavky vs. predaj, mzdy vs. spokojnosť zamestnancov, počet chýb v kóde vs. skúsenosti programátora, dĺžka pracovného času vs. produktivita, čas strávený na stránke vs. konverzný pomer
 <b>Spearman</b> (ρ)	Monotónny vzťah	Poradie stránok vo výsledkoch vyhľadávania vs. návštevnosť, pozícia produktu v rebríčku predajnosti vs. predajnosť, hodnotenie influencerov vs. úspešnosť kampaní, šanca na povýšenie vs. pracovný výkon, preferované benefity vs. spokojnosť zamestnancov, počet testovacích prípadov vs. stabilita softvéru
 <b>Kendall</b> (τ)	Poradie hodnôt	Poradie kandidátov HR vs. ich prijatie, ranking produktov podľa recenzií vs. predaj, firemné oddelenia podľa spokojnosti vs. fluktuácia, najpopulárnejšie programovacie jazyky vs. počet nových vývojárov, poradie bugov podľa závažnosti vs. počet nahlásení, poradie sociálnych médií podľa angažovanosti vs. počet reklám

# Chyby Korelácia



#	Chyba	Ako sa jej vyhnúť
1	Zamieňanie korelácie s kauzalitou	Overiť vzťah experimentálne alebo cez regresné modely.
2	Použitie Pearsonovej korelácie na nelineárne dáta	Vizualizovať dáta, skúsiť Spearmanovu koreláciu.
3	Ignorovanie vplyvu outlierov	Použiť robustnejšie metódy (Spearman, Kendall).
4	Práca s malou vzorkou dát	Použiť štatistické testy významnosti.
5	Korelácia spôsobená tretou premennou	Použiť multivariačnú analýzu.
6	Nesprávny typ korelácie	Spearman pre poradové dáta, Pearson pre lineárne.
7	Preceňovanie slabej korelácie	Interpretovať výsledky v kontexte.
8	Nulová korelácia neznamená žiadny vzťah	Overiť, či vzťah nie je nelineárny.
9	Použitie korelácie na kategoriálne údaje	Použiť iné štatistické testy (napr. khi-kvadrát).
10	Neoverenie štatistickej významnosti	Použiť p-hodnotu a testy významnosti.