[Date]

python-docx

[company name]

Statistic

**Deze samenvatting is gebaseerd op de slides ‘Lecture 1’ van Dennis Fok (Statistics for Data Science).**

Inhoudsopgave

# Lecture 1 — Statistics for Data Science (Volledige Samenvatting)

Erasmus School of Economics — Prof. Dennis Fok  
Periode: September – Oktober 2024

Deze samenvatting bevat een uitgebreide uitleg van alle slides uit Lecture 1, inclusief theorie, kernbegrippen, formules en Python‑voorbeelden.

# 1 Inleiding & Doel

De cursus vormt een (her)introductie tot de basis van statistiek met nadruk op kritisch denken: niet enkel **\*hoe\*** maar ook **\*waarom\*** analyses worden uitgevoerd.

Alles wordt toegepast in Python of R, met als leidraad het boek ‘Practical Statistics for Data Scientists’ (Bruce, Bruce & Gedeck, hoofdstukken 1–5).

Naast het technisch toepassen van methoden, wordt in deze cursus sterk de nadruk gelegd op het ontwikkelen van een kritische houding ten opzichte van data en statistische uitkomsten. Je leert niet alleen formules gebruiken, maar vooral ook onderliggende aannames doorzien en resultaten interpreteren binnen de context van het vraagstuk. Deze vaardigheden zijn essentieel om in de praktijk als data scientist valkuilen te vermijden en met vertrouwen beslissingen te nemen op basis van data.

# 2 Cursusstructuur en aanpak

De cursus bestaat uit zeven onderdelen:  
1. Basis van statistiek en inferentie  
2. Distributies, beschrijvende statistiek en hypothesetoetsing  
3. Toetsen van verschillen  
4. Lineaire regressie  
5. Diagnostiek en modelselectie  
6. Logistische regressie (GLM)  
7. Bayesian statistiek.  
  
Werkwijze: lezen → college → oefenen → wekelijkse opdrachten → deelopdrachten voor eindproject.

# 3 Doelen van statistiek

De drie hoofddoelen van statistiek:  
1. Data samenvatten (descriptieve statistiek)  
2. Verschillen toetsen (hypothese‑toetsing)  
3. Eigenschappen schatten van het datagenererend proces (inferentiële statistiek).  
  
Belangrijk: onzekerheid is altijd aanwezig. Resultaten berusten op aannames; wanneer die niet gelden, kunnen conclusies misleidend zijn.

Statistiek is geen trucendoos; het draait om logisch redeneren en het stellen van de juiste vragen bij elke stap van het analyseproces. Dit betekent dat je als data scientist altijd kritisch moet nagaan of de gekozen methode past bij de aard van je data en de vraagstelling. Bovendien leer je binnen deze cursus wanneer je voorzichtig moet zijn met het trekken van conclusies en hoe je valkuilen zoals bias, overfitting en misinterpretatie kunt herkennen en vermijden.

# 4 Belangrijke kernbegrippen

1. Data en variabelen:verzameling van gemeten kenmerken (features). Rol in analyse: afhankelijk (Y) of onafhankelijk (X).  
2. Steekproef en populatie:steekproef is subset van populatie; representativiteit cruciaal.  
3. Variatie en onzekerheid:toevalsvariatie, meetfouten, verschillen tussen respondenten.  
4. Modellen:vereenvoudigde weergaven van het datagenererend proces met expliciete aannames.

# 5 Typen data en variabelen

• Numeriek: continu (temperatuur) of discreet (aantal studenten)  
• Categoraal: binair (ja/nee), nominaal (kleur), ordinaal (eens/oneens)  
• Zonder duidelijke afhankelijke variabele → exploratieve statistiek In Python/R bepaalt het datatype vaak welke bewerkingen mogelijk zijn.

# 6 Steekproef vs populatie

Data kunnen afkomstig zijn uit experimenten of observaties. Controleer steeds:  
– zijn observaties onafhankelijk?  
– is er clustering (bijv. studenten binnen scholen)?  
– is er sampling bias?  
Voorbeeld van slechte steekproef: enquête op lokale markt of lage responsgraad in online survey.

# 7 Variatie, onzekerheid en significantie

Variatie is normaal. Significantie toetst of een waargenomen verschil groter is dan toevalsvariatie.  
  
\*\*Significant verschil:\*\* onwaarschijnlijk onder nulhypothese (H₀).  
\*\*Niet‑significant:\*\* verschil kan door toeval komen, maar sluit echt verschil niet uit.

# 8 Modellen en aannames

Een model beschrijft aannames over hoe data ontstaan. Cruciale quote: “All models are wrong, but some are useful.” (Box & Draper, 1987). Belangrijk: toets aannames (lineariteit, normaliteit, homoskedasticiteit).

# 9 Python‑implementatie (StatsModels)

Voorbeeldcode python:

|  |
| --- |
| import statsmodels.api as sm import statsmodels.formula.api as smf m = smf.ols('y ~ x1 + x2', data=df) r = m.fit() print(r.summary()) |

Beschikbare modellen: OLS, GLM, MixedLM, etc.

Met StatsModels kun je eenvoudig modellen specificeren door formules te gebruiken die vergelijkbaar zijn met de notatie in statistische literatuur. Bijvoorbeeld, een lineair regressiemodel kan worden opgezet met `**ols('Y ~ X1 + X2', data=df**)`. Vervolgens kun je de fit van het model beoordelen aan de hand van samenvattende statistieken zoals R², F-statistiek en p-waarden. Het is aan te raden om na elke analyse de aannames te evalueren via diagnostische plots, zoals **residualenplots en Q-Q plots**, zodat je zeker weet dat je interpretaties betrouwbaar zijn en niet berusten op foutieve aannames.

# 10 Workflow voor data‑analyse in Python

1. Importeren: `import pandas as pd`  
2. Data inlezen met `pd.read\_csv()`  
3. Verkennen met `df.describe()` en `df.plot()`  
4. Statistische berekeningen uitvoeren  
5. Resultaten documenteren en reproduceerbaar maken.

Een goede workflow begint vaak met het helder formuleren van je onderzoeksvraag. Daarna verzamel en structureer je de benodigde data, bijvoorbeeld door datasets te combineren of ontbrekende waarden te controleren. Vervolgens kies je geschikte analysemethoden, waarbij je steeds kritisch kijkt naar de aannames en eventuele beperkingen van je model. Tot slot is het belangrijk om je code en resultaten zó te documenteren dat anderen je stappen kunnen volgen en reproduceren – dat is echt goud waard in de wetenschap.

# 11 Beschrijvende statistiek

Gebruik grafische en numerieke samenvattingen om data te begrijpen:  
• `plot.scatter()`, `hist()`, `density()`, `boxplot()`  
Let op spreiding, vorm, correlatie en uitschieters (outliers).

# 12 entrale tendentie

Voor n waarnemingen X₁,…,Xₙ:  
• Gemiddelde: X̄ = (ΣXi)/n  
• Mediaan: 50%‑kwantiel  
• Modus: meest voorkomende waarde.  
Mediaan is robuuster dan gemiddelde bij uitschieters.

# 13 Spreidingsmaten

• Range = max − min  
• Interkwartielafstand (IQR) = Q₃ − Q₁  
• Outlier: < Q₁ − 1.5×IQR of > Q₃ + 1.5×IQR  
• Variantieschatting: s² = 1/(n−1) Σ(Xi−X̄)²  
• Standaardafwijking: s = √s²  
• Graden van vrijheid = aantal observaties − aantal geschatte parameters.

# 14 Hogere‑orde momenten

• 1e moment (verwachting): E[X] = μ  
• 2e moment (variantie): E[(X−μ)²] = σ²  
• 3e moment (scheefheid): E[(X−μ)³/σ³]  
• 4e moment (kurtosis): E[(X−μ)⁴/σ⁴]  
Voor normaalverdeling geldt: skew = 0, kurtosis = 3 (excess kurtosis = 0).

# 15 Toepassing in Python

Gebruik `.mean()`, `.median()`, `.var()`, `.std()`, `.skew()`, `.kurtosis()` op DataFrame‑kolommen voor directe berekening.

# 16 Samenvatting formules

• Gemiddelde: X̄ = (ΣXi)/n  
• Variantie: s² = (1/(n−1)) Σ(Xi−X̄)²  
• Standaardafwijking: s = √s²  
• IQR = Q₃ − Q₁  
• Outliers: < Q₁−1.5·IQR of > Q₃+1.5·IQR  
• Skewness = E[(X−μ)³/σ³]  
• Kurtosis = E[(X−μ)⁴/σ⁴]

# 17 Oefening en voorbereiding volgende week

Lees hoofdstuk 1–2 van het boek, maak opdrachten 1.1–1.4 en oefen met eigen datasets. Gebruik pandas.describe(), eenvoudige grafieken, en controleer verdelingen en relaties tussen variabelen.