[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.

# Lecture 2 – Statistics for Data Science

# 2.1. Distributies en basisbegrippen

Bij het werken met **statistische distributies** is het essentieel om te begrijpen **wat voor type variabele** men onderzoekt discreet of continu en om de bijbehorende **kansfuncties** correct te onderscheiden.  
Een distributie beschrijft namelijk **hoe waarschijnlijk** bepaalde waarden van een willekeurige variabele zijn. Dit inzicht vormt de basis voor bijna alle statistische analyses, van beschrijvende statistiek tot inferentiële methoden zoals hypothesetoetsing en regressieanalyse.

**Discrete variabelen**

Discrete variabelen kunnen slechts **afgebakende waarden** aannemen, zoals de uitkomst van een dobbelsteen (1–6) of het aantal kinderen in een gezin.  
Voor dergelijke variabelen worden twee fundamentele functies gebruikt:

* **Probability Mass Function (PMF)** –   
  Geeft de kans dat de variabele exact een bepaalde waarde aanneemt.
* **Cumulative Distribution Function (CDF)** –   
  Toont de kans dat de variabele kleiner dan of gelijk aan een bepaalde waarde is.

De CDF is dus de opgetelde kans tot en met een bepaalde waarde, terwijl de PMF de kans op één specifieke waardeweergeeft.  
Bijvoorbeeld: voor een eerlijke dobbelsteen is  voor , en de cumulatieve kans stijgt in stappen van .

**Continue variabelen**

Continue variabelen kunnen **elke waarde binnen een interval** aannemen, zoals lengte, gewicht of temperatuur. Omdat het aantal mogelijke waarden oneindig is, is de kans dat  precies één specifieke waarde aanneemt exact nul ().  
Daarom wordt niet met afzonderlijke kansen gewerkt, maar met **kansdichtheden**:

* **Probability Density Function (PDF)** – , zodat

De kans dat  binnen het interval  valt, is dus de oppervlakte onder de PDF tussen  en .

* **Cumulative Distribution Function (CDF)** –   
  De CDF is de integraal (oppervlakte) van de PDF en geeft de kans dat de variabele kleiner of gelijk is aan .  
  De afgeleide van de CDF levert de PDF: .

Interpretatie en betekenis

Het onderscheid tussen discrete en continue variabelen is cruciaal bij:

* het kiezen van de juiste **statistische toets** (bijv. binomiaal vs. normaal),
* het berekenen van **verwachtingswaarden en varianties**, en
* het visualiseren van data met histograms of dichtheidsplots.

Een goed begrip van kansfuncties voorkomt verkeerde interpretatie van data — bijvoorbeeld het verwarren van dichtheid (hoogte van de curve) met kans (oppervlakte onder de curve).

Vormkenmerken van verdelingen

Naast de locatie (gemiddelde) en spreiding (variantie) zijn er twee belangrijke **vormmaten**:

* **Scheefheid (Skewness):** meet de asymmetrie van de verdeling.  
  Positieve skewness → rechterstaart langer; negatieve → linkerstaart langer.
* **Kurtosis:** meet de ‘spitsheid’ of concentratie van de verdeling rond het gemiddelde.  
  Een normale verdeling heeft kurtosis ≈ 3; hogere waarden duiden op dikkere staarten.

Het analyseren van deze kenmerken helpt vast te stellen of een dataset **afwijkt van de normale verdeling** — een cruciale stap vóór het toepassen van parametrische statistische methoden.

Samenvattend:

Bij het werken met statistische distributies is het van groot belang om te begrijpen wat voor type variabele je onderzoekt — een **discrete** of een **continue** variabele — en om de bijbehorende kansfuncties correct te interpreteren. Een distributie geeft namelijk weer **hoe waarschijnlijk** bepaalde uitkomsten van een willekeurige variabele zijn. Dit vormt de basis voor het correct analyseren, interpreteren en modelleren van data.

Bij **discrete variabelen** gaat het om grootheden die slechts een beperkt aantal waarden kunnen aannemen, zoals het aantal keren dat een munt op “kop” valt of de uitkomst van een dobbelsteen. De verdeling wordt dan beschreven door de *probability mass function* (PMF), die de kans geeft dat de variabele een specifieke waarde aanneemt, aangeduid als . Daarnaast wordt vaak de *cumulative distribution function* (CDF) gebruikt, , die de kans weergeeft dat de uitkomst kleiner of gelijk is aan een bepaalde waarde.

Bij **continue variabelen** — zoals lengte, gewicht of temperatuur — kan de variabele oneindig veel waarden binnen een bepaald interval aannemen. Omdat de kans op één specifieke waarde precies nul is (), wordt de verdeling beschreven met een *probability density function* (PDF), aangeduid als . De kans dat een waarneming tussen twee waarden valt, wordt berekend via de oppervlakte onder de curve:

De *cumulative distribution function* is in dit geval de integraal van de PDF, en de afgeleide van de CDF is opnieuw de PDF: .

Het onderscheid tussen discrete en continue variabelen is essentieel bij het kiezen van de juiste analysemethode en het trekken van correcte conclusies uit data. Een misverstand tussen kans en kansdichtheid kan leiden tot verkeerde interpretaties van resultaten.

Naast het type kansfunctie zijn ook de **vormkenmerken van de verdeling** belangrijk. De **scheefheid (skewness)** geeft aan of de verdeling asymmetrisch is: een positieve scheefheid duidt op een langere rechterstaart, een negatieve op een langere linkerstaart. De **kurtosis** meet hoe spits of vlak de verdeling is ten opzichte van een normale verdeling: een hoge kurtosis wijst op dikkere staarten, een lage op plattere toppen.

Samen helpen deze concepten — het onderscheid tussen discrete en continue kansfuncties, en de vormkenmerken van de verdeling — bij het **begrijpen van de structuur van data** en het **toepassen van geschikte statistische modellen**. Ze vormen daarmee een onmisbare basis voor iedere datawetenschappelijke analyse.

# 2.2. Grafische illustraties

Voorbeelden van discrete en continue kansverdelingen

Een eenvoudig maar krachtig voorbeeld van een **discrete kansverdeling** is de **dobbelsteen**. Wanneer een eerlijke dobbelsteen wordt gegooid, zijn er zes mogelijke uitkomsten: 1, 2, 3, 4, 5 of 6. Elke uitkomst heeft een gelijke kans van één op zes.  
De probability mass function (PMF) wordt dan gegeven door:

De cumulative distribution function (CDF)  neemt toe in sprongen van  telkens wanneer een volgende waarde wordt bereikt. Deze trapvormige structuur is kenmerkend voor discrete verdelingen.

Bij een **continue kansverdeling** liggen de dingen anders. De variabele kan oneindig veel waarden aannemen binnen een interval, en de kans op één specifieke waarde is precies nul. De verdeling wordt dan beschreven met een probability density function (PDF), waarvan de oppervlakte onder de curve gelijk is aan 1.

Een eenvoudig voorbeeld van een continue verdeling is:

De bijbehorende cumulatieve verdelingsfunctie is:

De controle op de geldigheid van de PDF gebeurt door het berekenen van de integraal:

waaruit blijkt dat de totale kans inderdaad gelijk is aan 1.

Het verschil tussen deze twee voorbeelden maakt het onderscheid tussen discrete en continue kansverdelingen goed zichtbaar. In het discrete geval bestaat de verdeling uit afzonderlijke kanspunten, terwijl in het continue geval de kans zich over een oneindig aantal mogelijke waarden verspreidt en de oppervlakte onder de curve de kans vertegenwoordigt.

# 2.3. Python implementatie

Gebruik scipy.stats voor standaardverdelingen:  
from scipy import stats  
stats.norm.cdf(1, loc=0, scale=1) # standaardnormaal  
stats.binom.pmf(15, n=30, p=0.5) # binomiale kans  
stats.expon.ppf(0.95, scale=10) # 95e percentiel exponentieel

# 2.4. Belangrijke standaardverdelingen

Vier centrale verdelingen:  
1. Normaal: X ~ N(μ, σ²), f(x)=1/(σ√(2π)) exp(−½((x−μ)/σ)²)  
2. t-verdeling (tₙ)  
3. Chi-kwadraat (χ²ₖ)  
4. F-verdeling (F(d₁,d₂))

# 2.5. Schatting van distributies

Histogram = benadering van echte verdeling.  
Gebruik scipy.stats.ecdf(data) en fit():  
fit = stats.fit(stats.norm, data, bounds={'loc':(-4,4),'scale':(0,1)})  
fit.plot(plottype='qq')

# 2.6. Schattingsonzekerheid

We schatten μ en σ² uit steekproeven.  
Var(X̄) = σ²/n → onzekerheid daalt bij grotere n.  
Standaardfout: SE(X̄) = √(s²/n).

# 2.7. Betrouwbaarheidsintervallen

Als σ² bekend: X̄ ± 1.96√(σ²/n)  
Als σ² onbekend: X̄ ± t₀.₉₇₅,ₙ₋₁ √(s²/n)  
Voorbeeld (huisprijsdata): [4967.998, 5332.533].

# 2.8. Central Limit Theorem (CLT)

### Wat als het N aantal waarnemingen klein is?

Wanneer het aantal waarnemingen (n) klein is, kunnen we niet zomaar aannemen dat het steekproefgemiddelde normaal verdeeld is. De **Centrale Limietstelling (CLT)** stelt weliswaar dat de verdeling van steekproefgemiddelden bij toenemende steekproefgrootte steeds beter de normale verdeling benadert, maar deze eigenschap geldt slechts **ongeveer**en **pas bij voldoende grote steekproeven**.

Bij kleine steekproeven (bijvoorbeeld n < 30) — zeker wanneer de onderliggende data afwijkt van een normale verdeling — kan dit leiden tot **onbetrouwbare of misleidende conclusies**. De gebruikelijke betrouwbaarheidsintervallen en hypothesetoetsen die zijn gebaseerd op de normale of t-verdeling, gaan dan mogelijk uit van verkeerde aannames. Hierdoor kunnen de berekende betrouwbaarheden of p-waarden afwijken van hun werkelijke waarden.

In de praktijk betekent dit dat het geschatte **betrouwbaarheidsinterval** niet altijd de bedoelde dekking heeft: een zogenaamd 95%-interval kan in werkelijkheid bijvoorbeeld maar 90% of juist 98% van de tijd het echte populatiegemiddelde bevatten. Ook kunnen statistische toetsen **te streng of te soepel** zijn, wat de kans op verkeerde beslissingen (type-I- of type-II-fouten) vergroot.

Wanneer de steekproefgrootte beperkt is, is het daarom belangrijk om **de verdeling van de data zorgvuldig te onderzoeken**. Als de verdeling duidelijk afwijkt van normaal, moet men extra voorzichtig zijn met het toepassen van standaard parametische methoden.

Een oplossing is het gebruik van **niet-parametrische methoden**, die geen specifieke veronderstellingen doen over de verdelingsvorm. Voorbeelden hiervan zijn de **Wilcoxon-toets** (als alternatief voor de t-toets) en de **Mann–Whitney-U-toets** voor onafhankelijke steekproeven. Deze toetsen baseren zich op de rangorde van de data in plaats van op de ruwe waarden, waardoor ze robuuster zijn bij niet-normale verdelingen.

Een andere veelgebruikte aanpak is de **bootstrap-techniek**. Hierbij wordt herhaaldelijk met teruglegging uit de steekproef getrokken om een empirische verdeling van de schatter (zoals het gemiddelde) op te bouwen. Op basis van deze herhalingen kunnen betrouwbaarheidsintervallen worden berekend zonder aan te nemen dat de data normaal verdeeld is. Dit maakt bootstrap-methoden bijzonder geschikt voor situaties met kleine steekproeven of onbekende verdelingen.

Samenvattend is het bij kleine steekproeven verstandig om:

* niet blind te vertrouwen op de Centrale Limietstelling,
* eerst de verdeling van de data te analyseren, en
* indien nodig alternatieve methoden te gebruiken die minder afhankelijk zijn van de aanname van normaliteit.

Voor **grote steekproeven** geldt wél dat:

en dus dat het steekproefgemiddelde zelfs bij niet-normale data **ongeveer normaal verdeeld** zal zijn.  
Maar zolang n klein blijft, zijn **de aannames van de CLT te zwak** om volledig op te vertrouwen, en moet men overschakelen op aangepaste of niet-parametrische methoden.

# 2.9. Samenvattend overzicht

**Stap 1: Probability Mass Function (PMF)**

**Definitie**  
De PMF beschrijft de kans dat een **discrete variabele** een specifieke waarde aanneemt:

Voorbeeld: bij een dobbelsteen .

**Belangrijke conclusie**  
De PMF vormt de basis van kansrekening bij discrete data.  
🡺 **Zonder PMF geen inzicht in de totale kansverdeling.**  
🡺Dit inzicht is nodig om over te stappen naar continue variabelen (PDF’s) en integralen als vervanging van sommen.

**Stap 2: Probability Density Function (PDF)**

**Definitie**  
Voor **continue variabelen** wordt de kansdichtheid beschreven door een PDF:

De kans op één exacte waarde is nul (), maar de kans over een interval wordt bepaald door de oppervlakte onder de curve.

**Belangrijke conclusie**  
PDF’s maken het mogelijk om kansen te berekenen bij continue data.  
🡺 **De focus verschuift van afzonderlijke waarden naar oppervlakten (integralen).**  
🡺 Deze gedachte is essentieel voor de stap naar de cumulatieve verdeling (CDF).

**Stap 3: Cumulative Distribution Function (CDF)**

**Definitie**  
De CDF geeft de opgetelde kans tot en met een bepaalde waarde:

en haar afgeleide is de PDF: .

**Belangrijke conclusie**  
De CDF biedt een volledig overzicht van de verdeling en vormt de schakel tussen kansdichtheid en realisaties in data.  
🡺 **Met F(x) kunnen we kwantielen, percentielen en overschrijdingskansen berekenen.**  
🡺 Dit begrip is cruciaal om straks betrouwbaarheidsintervallen te kunnen opstellen.

**Stap 4: Standaardfout (SE)**

**Definitie**  
De standaardfout meet de variatie van een schatter (zoals het steekproefgemiddelde):

**Belangrijke conclusie**  
De SE geeft aan **hoe precies** een steekproef het populatiegemiddelde benadert.  
🡺**Kleinere SE = grotere betrouwbaarheid van de schatting.**  
🡺 Deze maat vormt de bouwsteen voor het berekenen van betrouwbaarheidsintervallen.

**Stap 5: Betrouwbaarheidsintervallen (Confidence Intervals, CI)**

**Definitie**  
Op basis van de standaardfout kunnen we een interval schatten waarbinnen het populatiegemiddelde met een bepaalde waarschijnlijkheid ligt.

* Als σ bekend:
* Als σ onbekend:

**Belangrijke conclusie**  
CI’s vertalen de variantie van een schatting naar een interpreteerbaar interval.  
🡺**Ze vormen de brug tussen beschrijvende en inferentiële statistiek.**  
🡺Zonder CI’s weten we niet hoe betrouwbaar onze schatting van μ werkelijk is.

**Stap 6: Centrale Limietstelling (CLT)**

**Definitie**  
De CLT stelt dat voor grote steekproeven:

Zelfs als de onderliggende data niet normaal verdeeld is, wordt het steekproefgemiddelde bij grote n ongeveer normaal verdeeld.

**Belangrijke conclusie**  
De CLT rechtvaardigt het gebruik van normale en t-verdelingen bij inferentie.  
🡺**Bij grote n mogen we normaalmodellen gebruiken, zelfs bij niet-normale data.**  
🡺Bij kleine n moeten we echter alternatieven (zoals bootstrap of niet-parametrische methoden) overwegen.

**Eindconclusie**

Deze opeenvolgende stappen vormen samen het fundament van statistische inferentie:

1. PMF/PDF → beschrijven hoe kansen verdeeld zijn.
2. CDF → begrijpen hoe cumulatieve kansen zich opbouwen.
3. SE → kwantificeren hoeveel onzekerheid een schatting bevat.
4. CI → vertalen die onzekerheid naar interpreteerbare grenzen.
5. CLT → rechtvaardigen waarom deze aanpak wiskundig geldig is bij grote steekproeven.