Vooraf: waar gaat het vak statistiek eigenlijk over, wat houdt het in, wat voor soort kennis over de werkelijkheid levert het op? (Dat laatste wordt ook wel aangeduid als de epistemologische basis van de statistiek: 'epistemologie' betekent letterlijk kennisleer en gaat o.a. over het soort kennis dat aan de orde is).

Op deze vraag kunnen verschillende antwoorden gegeven worden, bijvoorbeeld 'het toepassen van de geschikte statistische procedures op de data om bepaalde conclusies te kunnen trekken of bepaalde beslissingen (bijv. wel of niet verwerpen van een nulhypothese) te kunnen nemen' of 'learning from data' (dit geldt ook voor 'machine learning').

Omdat we de werkelijkheid altijd moeten vereenvoudigen ('stileren') hebben we te maken met het modelbegrip.

Je kunt de aard en inhoud van het vak statistiek dan als volgt omschrijven:

We hebben data verkregen over een bepaald verschijnsel of 'systeem' of proces in de werkelijkheid. De data zijn een uitdrukking van dat systeem en we interpreteren dat systeem vanuit een stochastisch referentiekader. Met andere woorden, welke data er op een bepaald moment vanuit het systeem verkregen worden ligt niet vast maar wordt bepaald door toeval, wat we zullen waarnemen is daardoor tot op zekere hoogte onvoorspelbaar (data zijn niet deterministisch maar stochastisch bepaald).

Het betreffende systeem wordt gekarakteriseerd door grootheden die we parameters noemen en die specifieke waarden hebben zoals een bepaald gemiddelde, een bepaalde proportie of spreiding etc.

We benaderen het systeem met behulp van een model waarbij we onze data aan de hand van een of meer specifieke kansverdelingen relateren aan de parameters van het systeem. We zeggen dan: we gaan ervan uit dat het data-genererend proces wordt weergegeven door de betreffende kansverdeling(en). Het model is in principe wiskundig van aard maar door het toevoegen van een stochastisch element wordt het een statistisch model (zie hieronder).

Op basis van het model schatten we de waarden van de parameters (de karakteristieken van het systeem).

Omdat we die parameterwaarden alleen maar kunnen schatten (de echte parameterwaarden kennen we niet en zullen we ook nooit kennen) hebben we te maken met onzekerheid. Dit element van onzekerheid (stochastisch element) maakt het model een statistisch model (denk aan de 'error'-term in een regressievergelijking).

Het vak statistiek houdt zich in de kern bezig met het kwantificeren van die onzekerheid, dus van de onzekerheid die inherent is aan uitspraken over de parameterwaarden van het systeem.

Een voorbeeld ter illustratie.
er is een munt uit de Romeinse tijd opgegraven en we willen weten hoe zuiver die munt is

(bijvoorbeeld omdat we weten dat in die tijd belangrijke beslissingen werden genomen door zo'n munt op te gooien). We gooien de munt 20 keer en we vinden 5 keer kop.
Het systeem: de fysieke eigenschappen van de munt
Parameter: de proportie kop
Data: 5 keer kop uit 20 worpen
Kansverdeling: binomiale kansverdeling

Verderop zullen we dit vraagstuk zowel frequentistisch als Bayesiaans analyseren.

Uit het bovenstaande blijkt dat een model tenminste bestaat uit de data, een of meer parameters en een of meer kansverdelingen.
Behalve de data en een statistisch model hebben we een methode nodig die ons vertelt welke berekeningen we moeten uitvoeren, kortom, hoe we met de data en het model moeten omgaan.
Er zijn twee belangrijke paradigma's voor statistiek en dus twee verschillende methoden: de gangbare, frequentistische statistiek en de Bayesiaanse statistiek.
Bij de Bayesiaanse methode spelen een of meer 'prior'-kansverdelingen een rol (voor elke parameter een aparte 'prior').
Een belangrijk verschil tussen beide benaderingen is dat een parameter binnen de frequentistische benadering als een vaste maar onbekende grootheid, maar binnen de Bayesiaanse benadering als een kansvariabele (stochast) wordt opgevat.

Tenslotte: merk op dat
a) een statistisch model (bijv. een lineair regressiemodel) een model is dat op verschillende manieren geanalyseerd kan worden, dus op basis van verschillende methoden zoals frequentistisch of Bayesiaans. Een statistisch model op zich is dus noch frequentistisch, noch Bayesiaans. Een linear regressiemodel is een regressiemodel dat vervolgens zowel op frequentistische als op Bayesiaanse wijze geanalyseerd kan worden.

b) een gekozen model statistisch geëvalueerd moet worden, met name ten aanzien van
*'model fit' : hoe goed past het model bij de data?, denk in dit verband aan residuenanalyse en R^2-waarde bij lineaire regressie en aan de 'posterior predictive distribution' bij Bayesiaanse analyses.
**'parsimony'/complexiteit. Over het algemeen geldt: streef naar een zo eenvoudig model wanneer de verklarings-/voorspellingskracht hetzelfde is of nauwelijks meer toeneemt. R.Fisher, een van de grondleggers van de frequentistische statistiek sprak over data in termen van 'statistical currency'. Met je data kun je parameters schatten maar voor elke extra parameter die je schat lever je wel in: het aantal vrijheidsgraden neemt af en zo hou je minder data over voor andere zaken zoals het checken van de 'model fit'.
Hoe meer parameters je in je model stopt (denk bijv. aan het aantal onafhankelijke variabelen/'predictors' i.g.v. een regressiemodel), hoe beter in het algemeen je model past bij de data. Maar ook, hoe meer bronnen van onzekerheid ('error') gaan meespelen (elke parameter moet immers geschat worden) en dus hoe minder precies bijvoorbeeld je voorspellingen zullen zijn.

c) we nooit direct toegang hebben tot de werkelijkheid/een 'systeem'. Dat brengt met zich mee dat we te maken hebben met verschillende soorten fouten, 'errors'.

Zo is er de 'sampling error': omdat de data langs stochastische weg gegenereerd worden, kunnen de data per steekproef wat verschillen. Dat betekent dat we voor een specifieke dataset niet weten hoe adequaat onze data het systeem (en onze steekproefgrootheden- bijv. het steekproefgemiddelde- de corresponderende parameter) adequaat weerspiegelen. In de frequentistische statistiek komt deze 'sampling error' tot uitdrukking in de stadaardfout ('standard error'), zijnde de standaarddeviatie van de steekproevenverdeling ('sampling distribution'). De steekproevenverdeling is een kansverdeling die de mogelijke uitkomsten van een steekproefgrootheid (zoals het steekproefgemidelde) beschrijft met de bijbehorende kansen. De steekproevenverdeling is theoretisch gefundeerd in de Centrale Limiet Stelling maar kan ook empirisch - bijv. via 'bootstrapping' - afgeleid worden.
Een beperking is dat slechts voor een beperkt aantal steekproefgrootheden zo'n steekproevenverdeling langs theoretische weg afgeleid kan worden, de bootstrap kan hier uitkomst bieden (wat algemener geformuleerd: het is vaak een probleem om een geschikte kansverdeling te vinden waarmee overschrijdingskansen kunnen worden berekend).
.
In de Bayesiaanse statistiek komt de 'sampling error' tot uitdrukking in de standaarddeviatie van de 'posterior' kansverdeling. In principe zijn hier geen beperkingen aanwezig t.a.v. het aantal mogelijke steekproefgrootheden. Voor elke parameter kun je in principe een posterior kansverdeling afleiden.

Verder kun je de verkregen data beschouwen als de 'rijkdom' waarover de onderzoeker beschikt (vgl. Fisher's 'statistical currency'). Dat betekent dat de kwaliteit van de data een grote rol speelt. In dit verband hebben we te maken met de mogelijkheid van 'measurement errors'. Zo dien je aandacht te besteden aan de betrouwbaarheid van je metingen en ingeval van scores die verkregen zijn via vragenlijsten dien je na te gaan in welke mate de geobserveerde scores overeenkomen met de werkelijke scores(de score die een respondent invult en de score die echt geldt). Dit hangt onder meer samen met de deugdelijkhied van de theoretische begrippen ('constructen') die in de vragenlijst worden gebruikt: zijn de begrippen goed gedefinieerd?, zijn ze theoretische goed onderbouwd?, zijn ze adequaat geoperationaliseerd?

Daarnaast zijn er nog andere vormen van zogeheten 'non-sampling errors' zoals 'missing data errors', 'coverage errors' (populatie en/of steekproef zijn niet goed in kaart gebracht), errors vanwege het niet goed behandelen van de data enz.

Terug naar het voorbeeld van de munt (de analyses zijn met R uitgevoerd):
> #data: 20 keer gooien met de munt levert 5 keer kop op
> #Volgens NHST (frequentistische analyse):
> pbinom(5,20,0.5)
[1] 0.02069473
> #P-waarde = 2 x 0.02069 is ongeveer 0.042 < 0.05 -> reject H0: munt is niet zuiver

Beter (i.v.m. beschikbaarheid betrouwbaarheidsinterval) is via:
> binom.test(5,20,0.5)

Exact binomial test

data: 5 and 20
number of successes = 5, number of trials = 20, p-value = 0.04139
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.08657147 0.49104587
sample estimates:
probability of success
              0.25

Bayesiaanse data-analyse met een niet-informatieve prior (uniforme verdeling):

> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})

Attaching package: 'Bolstad'

The following object(s) are masked _by_ '.GlobalEnv':

   binodp

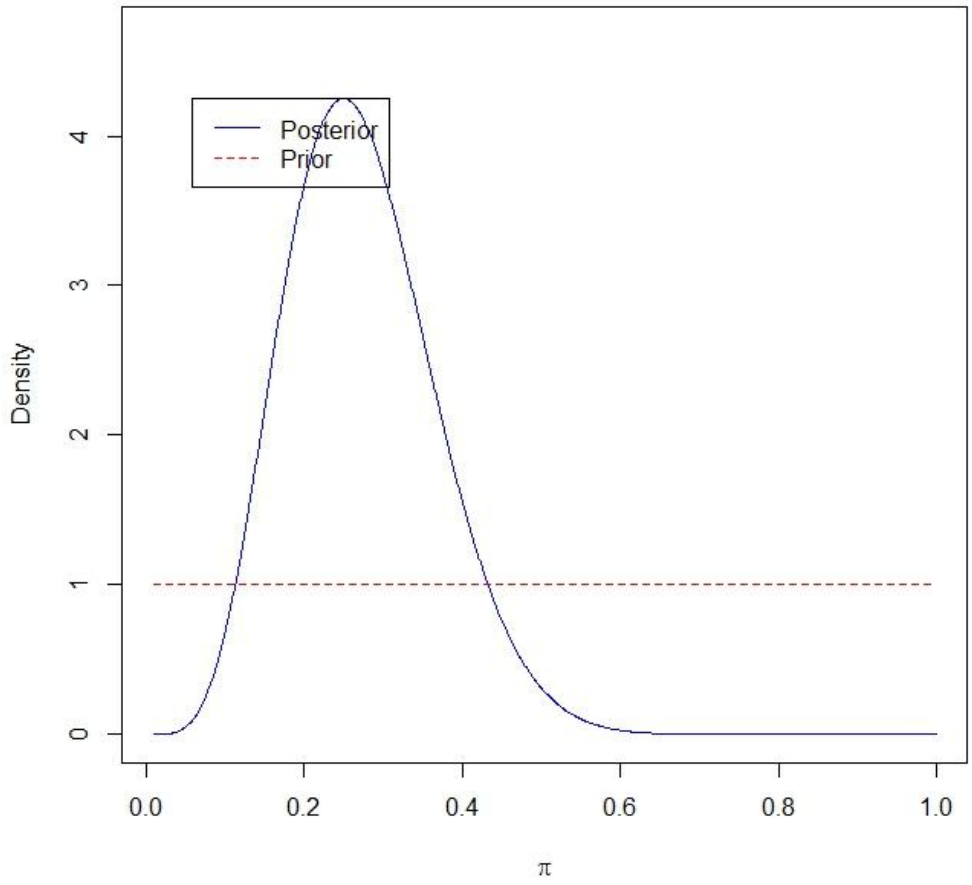
Warning message:
package 'Bolstad' was built under R version 2.15.3

> library(Bolstad)
> x=seq(0,1)
> curve(dbeta(x,1,1))
> binobp(5,20,1,1)
Posterior Mean   : 0.2727273
Posterior Variance   : 0.0086238
Posterior Std. Deviation : 0.0928643

Prob.   Quantile
------ ---------
0.005   0.0801205
0.01   0.0924654
0.025   0.1128094
0.05   0.1324482
0.5   0.2657397
0.95   0.4369763
0.975   0.4716598
0.99   0.5119804

0.995 0.5392432

95% 'credible interval': zie rode getallen



Voorbeeld van Bayesiaanse analyse (m.b.v. WinBUGS) en frequentistische analyse:

Een bioloog is geïnteresseerd in het gewicht (gemiddelde, spreiding) van een populatie slechtvalken. Er wordt een steekproef genomen van 10 slechtvalken, van elk exemplaar wordt het gewicht in grammen gemeten.
Welke uitspraak over het gewicht van de populatie kan op grond daarvan worden geformuleerd?

1. Bayesiaans:

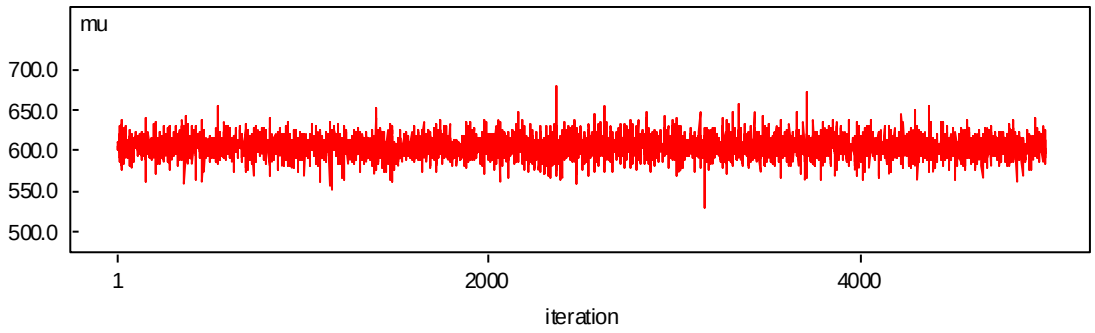
Inferring mean and standard deviation of a population distribution of peregrine falcons from a sample of observed independent data, assuming data following a Gaussian distribution.

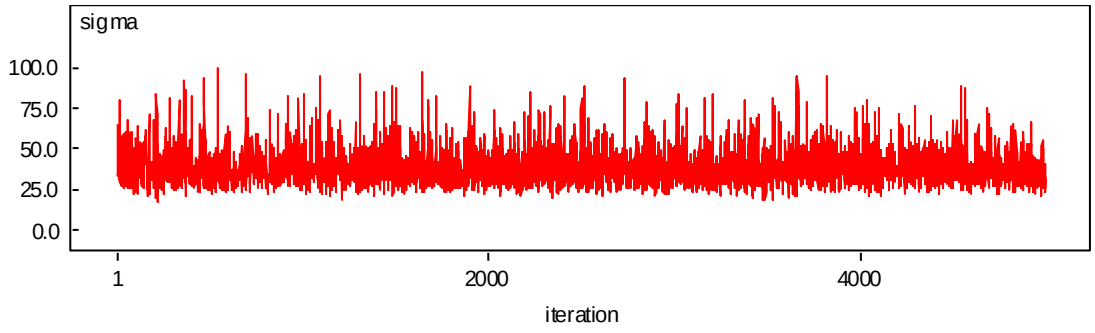
```
list(x=c(657.33,593.92,634.68,589.31,647.24,582.21,571.38,623.63,561.90,586.39),n
=10)

model{
for (i in 1:n) {
x[i]~dnorm(mu,lambda)
}

# Priors

mu~dunif(0,5000)
sigma~dunif(0,100)
lambda<- 1/pow(sigma,2)
}
```

WinBUGS OUTPUT:

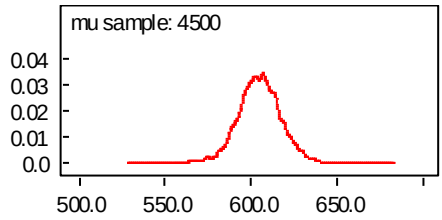
model is syntactically correct
data loaded
model compiled
initial values generated, model initialized

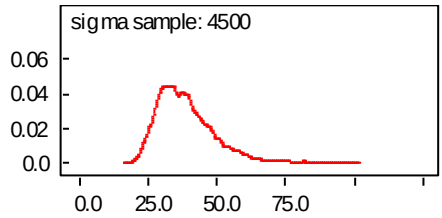
Time series





Kernel density





Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
mu	604.7	12.75	0.1991	579.5	604.6	629.9	501	4500
sigma	38.97	11.06	0.239	23.59	37.04	66.7	501	4500

2. Frequentistisch:

Theorie/Formule:

σ unknown: one-sample t test (one-sample t statistic)

$H_0 : \mu = \mu_0$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad d_f = n - 1$$

R-command:

t.test(data, mu= .... (μ_0), conf.level=....)

```
> gewicht=c(657.33,593.92,634.68,589.31,647.24,582.21,571.38,623.63,561.90,586.39)
> t.test(gewicht,mu=400)

        One Sample t-test
```

```
data:  gewicht
t = 19.4554, df = 9, p-value = 1.158e-08
alternative hypothesis: true mean is not equal to 400
95 percent confidence interval:
 580.9862 628.6118
sample estimates:
mean of x
  604.799
```

Frequentistische (conventionele, ook wel klassieke) versus Bayesiaanse statistiek

Klassiek (frequentistisch):
vraag: hoe waarschijnlijk zijn de verkregen data vergeleken met een aangenomen nulhypothese-waarde? (in de context van vele herhalingen van het experiment/onderzoek),

$$p(data|H_0) = ?$$

specifieker:

vanuit de frequentistische benadering wordt de 'likelihood' geschat dat dergelijke data gevonden zullen worden onder de aanname dat de nulhypothese waar is, ze is gebaseerd op de verwachte frequentie dat de data zoals die welke we gevonden hebben (en, in het geval van significantietoetsing, nog extremere data) zouden voorkomen indien we dezelfde procedure van dataverzameling en -analyse vele malen zouden herhalen.
Frequentistische methoden focussen dus op de frequentie van de verkregen data, in het kader van hypothetische replicaties van de betreffende steekproef.

Bayesiaans:
vraag: gegeven de data en eventueel eerder verkregen informatie, wat is de kans op een bepaalde stand van zaken in de werkelijkheid?

$$p(H|data) = ?$$

ad 1)
Een voorbeeld van een onderzoekssituatie waarin de Bayesiaanse benadering de voorkeur verdient:

Een ecoloog wil weten of in een vijver in stad A. een bepaalde kikkersoort leeft. Gedurende haar eerste bezoek aan de vijver probeert ze 20 minuten lang sporen en geluiden van deze

kikkers op te pikken, echter zonder resultaat.

Nu is er uit eerder onderzoek het een en ander bekend over deze kikkersoort:
- wanneer de kikker er leeft, wordt die slechts in 80% van de gevallen ook daadwerkelijk waargenomen of gehoord
- het voorkomen van deze kikkersoort in een vijver hangt o.m. af van het type vegetatie en de omvang van de vijver en ook van de locatie (gemeten aan de hand van de asfaltdichtheid van de directe omgeving).

De vraag die beantwoord moet worden is welke (kans)uitspraak de ecoloog kan formuleren omtrent het wel of niet aanwezig zijn van deze kikkersoort, gegeven het feit dat ze geen kikker heeft waargenomen.

Klassieke benadering:

Nulhypothese: er zijn geen kikkers
P-waarde:
$$p(data|H_0) = p(geen-kikkers-waargenomen|H_0 : er-zijn-geen-kikkers) = 1$$
Conclusie: nulhypothese niet verwerpen

Nulhypothese: er zijn wel kikkers
P-waarde:
$$p(data|H_0) = p(geen-kikkers-waargenomen|H_0 : er-zijn-wel-kikkers) = 0.20$$
Conclusie: nulhypothese niet verwerpen

Bayesiaanse benadering:

$$p(aanw|niet-gezien) = \frac{p(aanw)\times p(niet-gezien|aanw)}{\{p(aanw)\times p(niet-gezien|aanw)\}+\{p(niet-aanw)\times p(niet-gezien|niet-aanw)\}} =$$
$$\frac{0.5\times 0.2}{(0.5\times 0.2)+(0.5\times 1)} = \frac{0.1}{0.6} = \frac{1}{6}$$

BOOMDIAGRAM!!

Stel nu dat op grond van bovengenoemde factoren (type vegetatie enz.) de prior kans dat er wel kikkers in de vijver leven gelijk is aan 0.75 ('high-quality habitat'), dan krijgen we:

$$p(aanw|niet-gezien)=\frac{p(aanw)\times p(niet-gezien|aanw)}{\{p(aanw)\times p(niet-gezien|aanw)\}+\{p(niet-aanw)\times p(niet-gezien|niet-aanw)\}}=$$

$$\frac{0.75\times 0.2}{(0.75\times 0.2)+(0.25\times 1)}=\frac{0.15}{0.40}=\frac{3}{8}$$