Del I

Statistiske grundbegreber

Kapitel 1

Konkordans

Vores behandling af teoretisk statistik vil tage udgangspunkt i følgende centrale problem: Et eksperiment beskrives ved et repræsentationsrum $(\mathcal{X}, \mathbb{E})$. Vi har en ide om den sandsynlighedsmæssige mekanisme bag eksperimentet, opsummeret i et sandsynlighedsmål ν på $(\mathcal{X}, \mathbb{E})$. Vi vil ofte omtale ν som en **model** for eksperimentet.

Eksperimentet udføres, og vi observerer et punkt $x \in \mathcal{X}$. Spørgsmålet er nu om x er i overensstemmelse med hvad modellen siger, eller om vi må revurdere beskrivelsen af eksperimentet. Vi taler om **konkordans** (udledt af det latinske ord *concordantia* for enighed) hvis der er overensstemmelse, og i modsat fald om **diskordans**. Vi kan reformulere problemet med stokastiske variable:

Konkordansproblemet: *Lad* $(\Omega, \mathbb{F}, P)$ *og* $(\mathcal{X}, \mathbb{E}, \nu)$ *være sandsynlighedsfelter, og lad* $X : (\Omega, \mathbb{F}) \to (\mathcal{X}, \mathbb{E})$ *være en stokastisk variabel.*
Lad observationen $X = x$ *være givet. Er det rimeligt at hævde at* $X(P) = \nu$*, altså at* ν *er fordelingen af* X *?*

Konkordansproblemet fremstår umiddelbart som lidt af et legetøjsproblem: som et abstrakt matematisk problem, som et problem man *i virkeligheden* ikke er interesseret i at løse, hvis man er anvendt statistiker. Men det bærer i sig mange af de vanskeligheder som også kendetegner mere komplicerede statistiske problemer, og konkordansproblemet er derfor et fortræffeligt sted at skærpe sit begrebsapparat. Endnu vigtigere: vi skal se at når man skal tackle mere komplicerede problemer, så forsøger man ofte at omformulere dem, så de fremstår som simple konkordansproblemer, og

det er naturligvis afgørende om man kan gøre noget ved disse afledte konkordansproblemer.

En central erfaring med konkordansproblemet (og generelt med alle statistiske problemer) er at det ikke er muligt at besvare det stillede spørgsmål på en måde, der er immun overfor kritik. Der er **ikke** tale om et matematisk problem, men om et fortolkningsmæssigt problem, og ethvert svar involverer derfor et element af skøn. To statistikere kan udmærket skønne forskelligt. Men denne erfaring betyder ikke at alle skøn er lige gode: de principper man skønner efter kan analyseres matematisk, og egenskaber ved forskellige principper kan beskrives og vejes op mod hinanden.

1.1 Konkordansproblemer på $\mathbb{R}$

Vi vil først diskutere en række eksempler på konkordansproblemer for observationer på $\mathbb{R}$. Eksemplerne vil strække sig fra ægte anvendelser i den praktiske statistik, til det mere kunstige. Strategien vil i alle tilfælde være at sammenligne den faktisk gjorte observation x med et stort antal simulerede observationer, hvor de simulerede observationer faktisk **er** uafhængige, ν-fordelte. På den måde lader vi simulationerne give et indblik i hvordan modellen mener at observationer bør falde.

Eksempel 1.1 I 1998 blev der i Danmark født 66170 levende børn. Fordelt på køn var der 34055 drenge og 32115 piger. En simpel model for kønsfordelingen kunne basere sig på følgende postulater:

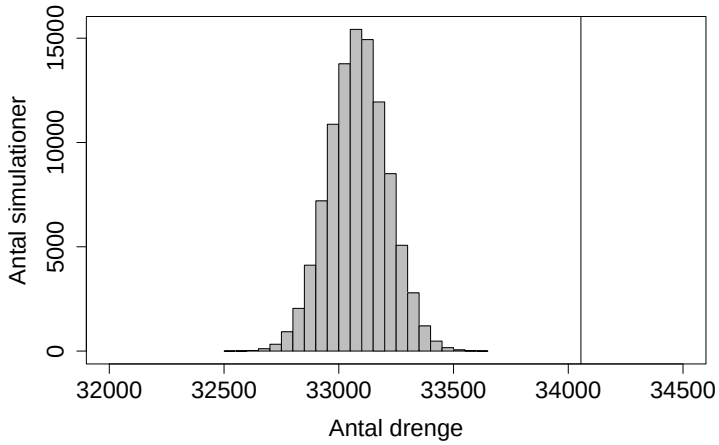
1) I hver fødsel er sandsynligheden for at få en dreng 50%.

2) Kønnet af barnet i en enkelt fødsel er uafhængigt af kønnet i de øvrige fødsler.

Man kan godt problematisere antagelse 2). For eksempel vil monozygote tvillinger have samme køn, så der er ikke uafhængighed mellem kønnet af de to børn i visse dobbeltfødsler. Men den virkeligt kritiske antagelse er den første.

Antagelse 1) og 2) leder frem til at antallet af drengebørn skulle være en observation fra en binomialfordeling med længde 66170 og succesparameter 0.5. I figur 1.1 har vi optegnet et histogram over 100.000 observationer fra en sådan binomialfordeling. Det ses at den faktisk gjorte observation på 34055 drenge overhovedet ikke svarer til de simulerede observationer.

I simulationseksperimenter skal man altid være på vagt overfor om den konklusion man drager hænger på tilfældige variationer. Men 100.000 observationer fra en Bin(66170, 0.5)-fordeling er så meget, at histogrammet er numerisk stabilt - hvis vi gentager eksperimentet, får vi et histogram frem, der er uskelneligt fra det første.

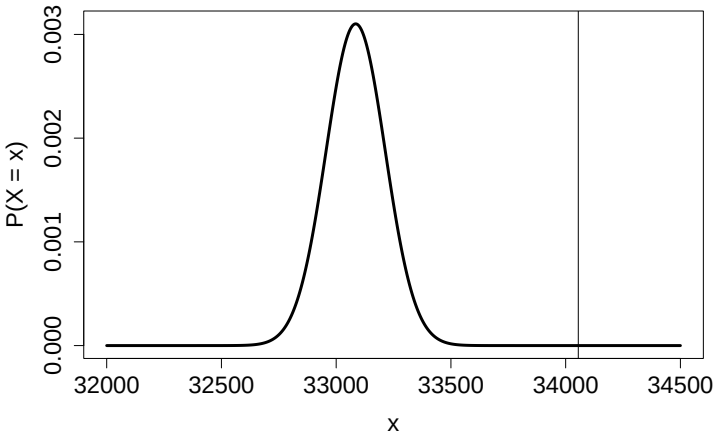


Figur 1.1: *Et histogram af 100.000 simulerede observationer fra en Bin(66170, 0.5)-fordeling. Den lodrette streg repræsenterer det faktisk observerede antal drengefødsler i Danmark i 1998.*

Intuitivt er vi på dette grundlag overbevist hinsides enhver tvivl: det passer simpelthen ikke at antallet af drengebørn født i Danmark i 1998 skulle være en observation fra en binomialfordeling med længde 66170 og succesparameter 0.5. Og dermed må mindst én af antagelserne 1) og 2) være forkert - formentlig den første.

Lad os grave lidt dybere og spørge hvorfor vi egentlig er så sikre på denne konklusion. Det er nemlig i høj grad psykologiske mekanismer der er på spil, nærmere end det er ubestridelige, logiske argumenter.

Af visuelle grunde er histogrammet i figur 1.1 baseret på en inddeling af $\mathbb{R}$ i relativt brede intervaller. Men argumentet ville fungere ligeså godt, hvis intervallerne kun indeholdt et enkelt heltal hver. Hvis vi holder observationen op mod et så fint histogram, og hvis vi lader antallet af simulationer være så stort at histogrammet er numerisk stabilt, så sammenligner vi i virkeligheden observationen med sandsynlighedsfunktionen for Bin(66170, 0.5)-fordelingen. Det gøres eksplicit i figur 1.2.



Figur 1.2: *Punktsandsynlighederne for en Bin(66170, 0.5)-fordeling. Den lodrette streg repræsenterer det faktisk observerede antal drengefødsler i Danmark i 1998. Den kraftigt optrukne "kurve" består i virkeligheden af et antal enkeltpunkter, afsat over de heltallige x-værdier.*

Hvis X er Bin(66170, 0.5)-fordelt, så er

$$P(X = 34055) = \left(\begin{array}{c} 66170 \\ 34055 \end{array} \right) \left(\frac{1}{2}\right)^{66170} \simeq 1.4 \cdot 10^{-15} .$$

Den faktisk gjorte observation har altså en sandsynlighed på stort set nul for at forekomme under modellen. Men det gælder på den anden side for alle tænkelige observationer: største punktsandsynlighed forekommer for $x = 33085$, og

$$P(X = 33085) = \left(\begin{array}{c} 66170 \\ 33085 \end{array} \right) \left(\frac{1}{2}\right)^{66170} \simeq 3.1 \cdot 10^{-3} ,$$

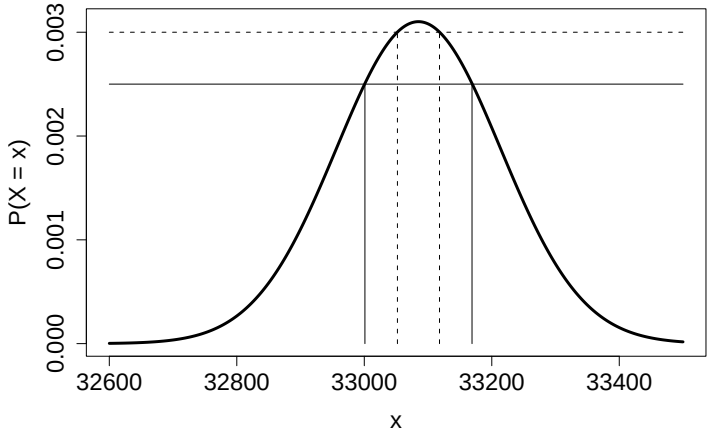
hvilket heller ikke er nogen særlig stor sandsynlighed. Der er dog mange størrelsesordner til forskel på $P(X = 34055)$ og $P(X = 33085)$, og vi konstaterer at selv om 33085 er en temmelig usandsynlig observation, så er den dog langt mere sandsynlig end 34055. Det understøttes af simulationerne, hvor 33085 således blev ramt 309 gange ud af 100.000, mens 34055 ikke blev ramt en eneste gang - største simulerede værdi var 33666. Vi drager den lære at den absolutte størrelse af punktsandsynligheder ikke betyder noget når vi skal vurdere rimeligheden/urimeligheden af en observation, men snarere *forhold* mellem punktsandsynligheder.

Vi ledes derfor til at opsøge de punkter der har de største punktsandsynligheder, uanset at de pågældende punktsandsynligheder egentlig ikke er særligt store. En vigtig grund til at vi føler os overbevist om at disse punkter "udpeges af modellen", er at de punkter der har relativt stor sandsynlighed under modellen "hænger sammen". For eksempel er

$$\{x \in \mathbb{N} \mid P(X = x) > 3.0 \cdot 10^{-3}\} = \{33052, 33053, \ldots, 33117, 33118\}$$

en sammenhængende sekvens af hele tal, ligesom

$$\{x \in \mathbb{N} \mid P(X = x) > 2.5 \cdot 10^{-3}\} = \{33001, 33002, \ldots, 33168, 33169\}.$$



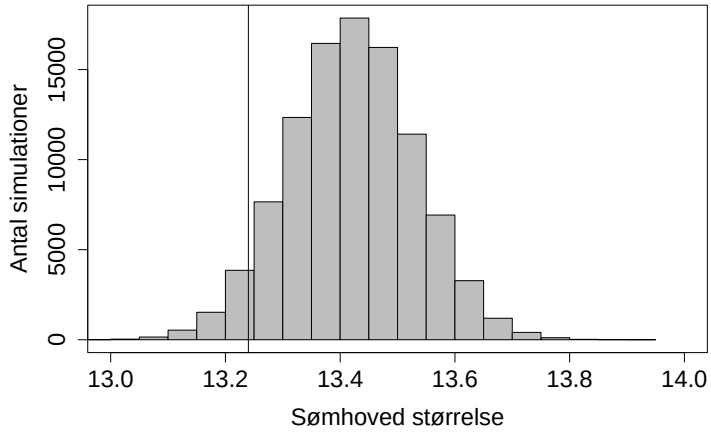
Figur 1.3: *Områder med punktsandsynligheder større end 0.0030, hhv. 0.0025 for en Bin(66170, 0.5)-fordeling.*

Dette forhold er illustreret på figur 1.3. Denne geometriske sammenhæng spiller en stor rolle for perceptionen af det oprindelige histogram, og det er den, der får os til at fornemme at observationer "bør ligge i nærheden af midtpunktet". Hvis figur 1.2 bestod af et større antal adskilte pukler, ville vores konklusion være knap så skråsikker.

∘

Eksempel 1.2 En fabrik har en maskine der producerer søm. Nøje overvågning af maskinen gennem lang tid har givet den erfaring at diameteren af et enkelt sømhoved fra et søm produceret af maskinen, kan betragtes som normalfordelt med positionsparameter 13.42 mm og skalaparameter 0.11 mm. Et søm samlet op fra fabriksgulvet viser sig at have et hoved med diameter 13.24 mm. Stammer dette søm fra fabrikkens egen maskine, eller kommer det udefra?

Set som et konkordansproblem, drejer spørgsmålet sig altså om vi med rimelighed kan betragte 13.24 som en observation fra en $\mathcal{N}(13.42, 0.11^2)$-fordeling. Vi angriber problemet ved at simulere et stort antal observationer fra en $\mathcal{N}(13.42, 0.11^2)$-fordeling, se figur 1.4.

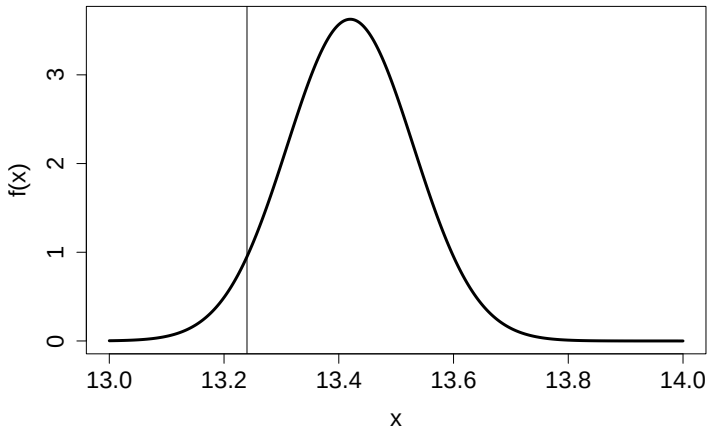


Figur 1.4: *Et histogram af 100.000 simulerede observationer fra en* $\mathcal{N}(13.42, 0.11^2)$*-fordeling. Den lodrette streg repræsenterer diameteren af det faktisk observerede sømhoved.*

Fortolkningen af figur 1.4 er knap så klar som den tilsvarende fortolkning af figur 1.1. Den faktiske observation ligger lidt yderligt i forhold til de simulerede observationer, men den falder ikke helt udenfor. En optælling viser at der trods alt er 5.046 ud af 100.000 simulerede observationer der er mindre end 13.24.

Bortset fra at den faktiske observation ikke ligger på samme måde i forhold til simulationspuklen på figur 1.1 og figur 1.4, så ligner de to tegninger jo hinanden. Men denne lighed dækker over en grundlæggende forskel: i normalfordelingseksemplet er

der ingen punkter med positiv punktsandsynlighed. Vi skal i afsnit 1.4 give et matematisk argument for følgende fortolkning: når vi i dette tilfælde holder observationen op mod et simuleret histogram, så holder vi den i virkeligheden op mod tæthedsfunktionen for $\mathcal{N}(13.42, 0.11^2)$-fordelingen. Denne sammenligning er udført direkte på figur 1.5.



Figur 1.5: *Tæthedsfunktionen for en* $\mathcal{N}(13.42, 0.11^2)$*-fordeling. Den lodrette streg repræsenterer diameteren af det faktisk observerede sømhoved.*

Som sagt er der tale om en fortolkning, eller måske nærmere en analogislutning til det diskrete tilfælde. Men har man først accepteret fortolkningen, kan man gå frem som før: Man opsøger de punkter der har størst tæthed, og fordi normalfordelingstætheden er **unimodal** ("kun én pukkel") og symmetrisk, så bliver prediktionsområdet et symmetrisk interval omkring middelværdien 13.42. Hvor bredt intervallet skal være, og om det inkluderer observationen på 13.24 vil vi senere vende tilbage til.

∘

Den grundlæggende forskel mellem eksempel 1.1 og eksempel 1.2, er at det første eksempel er diskret i sin natur, mens det andet eksempel er kontinuert. Denne forskel er ret dramatisk udfra et filosofisk synspunkt, og man kan med nogen ret spørge om denne forskel repræsenterer et reelt fænomen ved de udførte eksperimenter, eller om det er vanskeligheder vi skaber for os selv.

En usportslig reaktion på eksempel 1.2 kunne være at sige at det faktiske sømhoveds

diameter formentlig kun er målt med fire betydende cifre. Dermed er observationen "i virkeligheden" ikke kontinuert, men diskret. Ifølge argumentet har vi inddelt den reelle akse i disjunkte intervaller af længde 0.01 mm, og alt hvad vi har observeret er hvilket interval, sømhovedets ægte diameter ligger i. I dette billede er selve diameteren nok kontinuert, men vi har kun observeret en diskretiseret version. Fortsætter man ud af denne tangent, kunne man hævde at sømhovedet i virkeligheden er opbygget af atomer med en fast størrelse. Sømhovedets diameter er altså bestemt som antallet af atomer på en diagonal gange atomradius, og derfor er diameteren *principielt* diskret. Tages diskretiseringsargumentet alvorligt, er normalfordelingsmodellen under alle omstændigheder forkert.

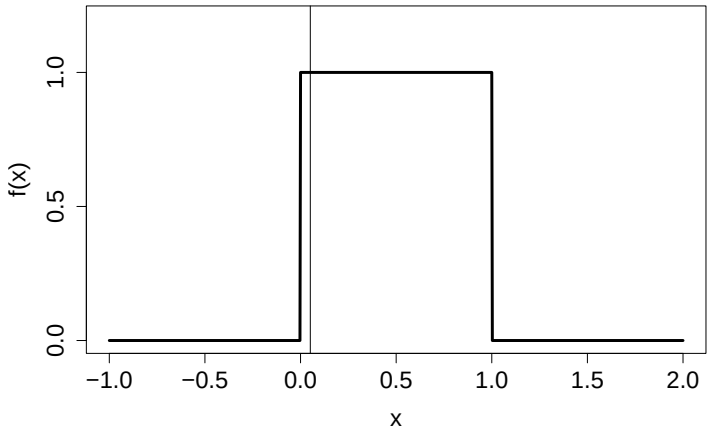
Diskretiseringsargumenterne spiller en stigende rolle i den generelle videnskabelige diskussion. Det er ikke mindst brugen af computere, der sætter emnet på dagsordenen, for tal lagres i computeren som en endelig sekvens af bits. Og i visse sammenhænge dukker denne endelighed op som et reelt fænomen - man taler da som regel om *numeriske* problemer, skønt problemstillingen er at computeraritmetik opfylder andre regler end den aritmetik vi lærte i skolen.

Der kan være meget rigtigt i disse indvendinger, men de tager alligevel fejl på et afgørende punkt: Uanset hvilken model vi bruger, så skal det ikke betragtes som den *sande* model, men som en approksimation til virkeligheden. Alle modeller gengiver virkeligheden i forvrænget form - men nogle modeller indeholder karakteristiske træk som man kan genkende fra naturen. Man bør se modeller som en slags karikaturtegning: gode karikaturer "ligner" ikke på naturalistisk facon, men man er alligevel ikke i tvivl om hvad de forestiller.

Den pointe som diskretiseringsargumenterne ser bort fra, er at kontinuerte modeller svarer meget bedre til den menneskelige intuition end de diskrete modeller. Den menneskelige hjerne giver hurtigt fortabt overfor kombinatoriske problemer, men har relativt let ved argumenter der baserer sig på kontinuitet og lineære approksimationer. Og derfor har differentialregning vist sig frugtbar igennem 400 år tværs over hele det videnskabelige felt. Man beskæftiger sig kun med diskrete modeller hvis det er tvingende nødvendigt, dvs. hvis en essentiel del af det foreliggende problem er diskret.

De to første eksempler på konkordansproblemer har været relativt simple at gå til, fordi de involverede fordelinger så tydeligt udpegede visse områder af den reelle akse som mere rimelige end andre. Men mange fordelinger har ikke denne karakter, de udpeger ikke oplagte områder, hvor observationerne bør falde:

Eksempel 1.3 Lad os sammenholde observationen $x = 0.051$ med ligefordelingen på det åbne enhedsinterval. Som i eksempel 1.2 fører en sammenligning med simulerede observationer til en sammenligning med tæthedsfunktionen for ligefordelingen, sådan som det ses i figur 1.6.



Figur 1.6: *Tæthedsfunktionen for ligefordelingen på* $(0, 1)$. *Den lodrette streg repræsenterer den postulerede observation* $x = 0.051$.

Det er umiddelbart vanskeligt at stille noget op i dette eksempel. I forhold til ligefordelingen kan den ene observation være lige så god som den anden. En observation på 0.001 er præcis lige så usandsynlig som en observation på 0.500, eller som en observation på 0.999. Hvis blot den gjorte observation ligger i enhedsintervallet, så kan den ikke opfattes som i modstrid med ligefordelingen!

∘

I eksempel 1.2 føles det ret ligetil at forholde sig til konkordansproblemet, i eksempel 1.3 er det derimod vanskeligt. Men som vi nu skal se, er det i virkeligheden samme eksempel! Og det ødelægger i nogen grad den intuitive tilgang til konkordansproblemet.

Det føles intuitivt naturligt at forlange at svaret på et konkordansproblem er **transformationsinvariant**: hvis det oprindelige problem går ud på at sammenholde observationen $x \in \mathbb{R}$ med sandsynlighedsmålet ν på $\mathbb{R}$, og hvis $t : \mathbb{R} \to \mathbb{R}$ er en strengt

voksende, kontinuert transformation, så bør man kunne svare på konkordansproblemet ved at sammenholde observationen $t(x)$ med sandsynlighedsmålet $t(\nu)$. Transformationen svarer jo blot til at man har "skiftet måleskala" i en generel forstand.

Men hvis vi forsøger at besvare konkordansproblemet på baggrund af tæthedsovervejelser, så bliver svaret **ikke** transformationsinvariant, for tætheden af billedmålet $t(\nu)$ kan se helt anderledes ud, end tætheden af ν. For eksempel, hvis F er fordelingsfunktionen for $\mathcal{N}(13.42, 0.11^2)$-fordelingen, så er $F\left(\mathcal{N}(13.42, 0.11^2)\right)$ simpelthen ligefordelingen på enhedsintervallet. Og i øvrigt er $q(13.24) = 0.051$. Så denne transformation fører det pæne eksempel 1.2 over i det ubehagelige eksempel 1.3. Og derfor står vi nu midt i suppedasen.

Hvis ikke vi skal give helt op overfor konkordansproblemet, må vi derfor acceptere at den oprindelige måleskala har en fortrinsstilling fremfor transformerede skalaer. Denne fortrinsstilling forekommer gerne matematisk sindede personer utilfredsstillende, fordi man som matematiker opfatter skalavalget som en arbitrær konvention - det er svært at forklare **hvorfor** en bestemt skala er så grundlæggende, for man kan næsten altid finde på "bagvedliggende eksperimenter", der gør at den skala man måler på må betragtes som en transformeret skala. Hvorimod man som praktisk arbejdende videnskabsmand føler det naturligt at visse skalaer er mere fundamentale end andre. Skønt de talrige slagsmål om "fortolkningen" af et eksperiment ofte handler om hvad den fundamentale skala for målingerne er.

Vi kan i et vist omfang komme ud af problemet, hvis vi indser at den oprindelige simulation i figur 1.4 primært fortæller os at vi bør være på vagt overfor meget store eller meget små observationer. Da skalaskiftet q er strengt voksende, bevarer transformationen denne ordensstruktur, og vi vil derfor mene at kritiske observationer på figur 1.6 ligger tæt ved 0 eller tæt ved 1. Præcis hvor tæt ved 0 eller 1 en observation skal ligge for at være kritisk, vil vi opstille kriterier for i afsnit 1.5.

I denne optik bruger vi den oprindelige måleskala til at afgøre i hvilken ende af skalaen de kritiske observationer ligger, men den præcise afgørelse af hvad der er kritisk og hvad der ikke er kritisk, kan godt foretages på den transformerede skala.

Bemærk at vi på denne måde i virkeligheden **ikke** har forholdt os til det konkordansproblem, der blev trukket op i eksempel 1.3. Vi har derimod sagt at hvis dette konkordansproblem er opstået ved en transformation af eksempel 1.2, så vil vi forholde os sådan og sådan til det. Man kunne i princippet forestille sig at konkordansproblemet i eksempel 1.3 var opstået på mange andre måder, og hvis det er

tilfældet, så ved vi stadig ikke hvad vi skal gøre ved det. Dette er et eksempel på hvordan man ofte inddrager **ekstra information**, inden man forsøger at besvare et konkordansproblem.

Denne diskussion har måske rystet fundamentet lidt under eksempel 1.2. Hvorfor synes vi egentlig at store og små observationer er kritiske? Ud over tæthedsargumentet, der altså ikke er så tvingende som det det kan forekomme ved første øjekast, er der endnu en psykologisk mekanisme på spil: vi kan nemlig meget nemt forestille os **alternative forklaringer** på store og små observationer, fremfor partout at forsøge at tilskrive dem den postulerede model. Hvis vi har gjort observationen $x = 10$, er det mere naturligt at forsøge at tilskrive den f.eks. en $\mathcal{N}(10, 0.11^2)$-fordeling, end partout at ville insistere på en $\mathcal{N}(13.42, 0.11^2)$-fordeling. Og der er mange andre mulige alternative forklaringer.

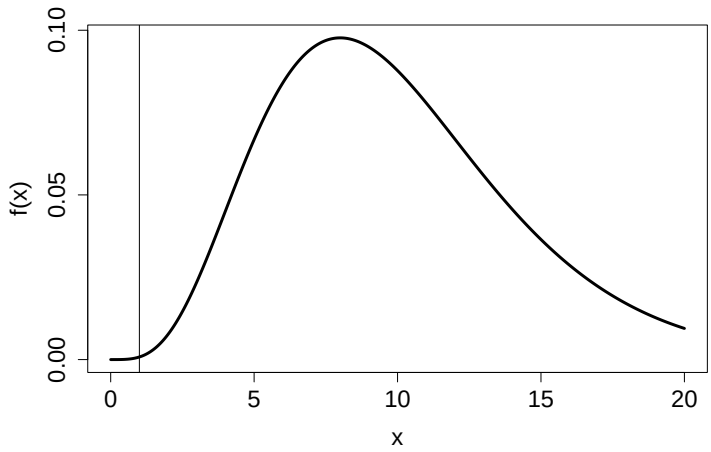
Om vi subjektivt bedømmer en observation som kritisk, har meget at gøre med hvor let vi kan forestille os alternative forklaringer. Man kan komme ud for situationer, hvor den gjorte observation forekommer ret utroværdig under modellen, men hvor man alligevel må acceptere at observationen er i konkordans med modellen, fordi man ikke kan finde på fornuftige alternative forklaringer.

Eksempel 1.4 Lad os undersøge om observationen $x = 1$ kan tænkes at stamme fra en χ^2-fordelingen med 10 frihedsgrader. Som i eksempel 1.2 fører en sammenligning med simulerede observationer til en sammenligning med tæthedsfunktionen for χ^2-fordelingen med 10 frihedsgrader, sådan som det ses i figur 1.7.

Den postulerede observation ligger meget yderligt i fordelingen, og argumenteres analogt med eksempel 1.2, vil man på denne baggrund næppe tro på at observationen stammer fra en χ^2-fordeling med 10 frihedsgrader. I forbindelse med χ^2-fordelinger, har man imidlertid ofte den ekstra information at enten passer modellen - eller også er observationen "for stor" til at pengene passer. I vores eksempel har vi en observation, der kan synes "for lille", og det vil ikke på samme måde få alarmklokkerne til at ringe. Hvis man har en fornemmelse af hvad der kan være galt, før man ser på observationen, så bør denne fornemmelse afspejle sig i den måde man vurderer tætheden på. Vi vil i eksempel 1.7 give et eksempel på en situation, hvor man har speciel grund til at være på vagt overfor store observationer i en χ^2-fordeling.

∘

Eksempel 1.5 Lad os sammenholde observationen $x = 0.1$ med arcussinus fordelingen, dvs. med B-fordelingen med formparametre $(1/2, 1/2)$. Som i eksempel 1.2



Figur 1.7: *Tæthedsfunktionen for en χ^2-fordeling med 10 frihedsgrader. Den lodrette streg repræsenterer den postulerede observation $x = 1$.*

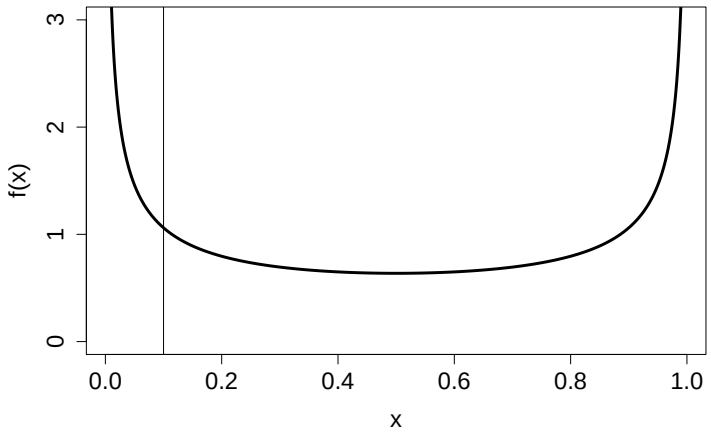
fører en sammenligning med simulerede observationer til en sammenligning med tæthedsfunktionen for arcussinus fordelingen, se figur 1.8.

Denne fordeling har ikke den unimodale karakter, som de hidtidige fordelinger har haft. Hvis vi opsøger de steder hvor tætheden er størst, får vi en forening af *to* intervaller - et interval inde omkring 0, og et andet interval ude ved 1. F.eks. er

$$\{x \in (0,1) \mid f(x) > 2\} = (0, 0.026) \cup (0.974, 1).$$

Proceduren med at opsøge de punkter der har størst tæthed, taber i intuitiv overbevisningskraft når det resulterende område ikke er sammenhængende. I dette eksempel fører argumentet til at man kun vil opfatte observationer helt inde omkring 0.5 som kritiske - og *så* kritiske er den slags observationer heller ikke, tætheden inde omkring 0.5 er ikke meget forskellig fra tætheden ude omkring den postulerede observation på 0.1.

Heldigvis er konkordansproblemet med en flerpuklet fordeling relativt sjælden - de fleste fordelinger man holder observationer op i mod vil være unimodale. Men lige præcis arcussinus fordelingen *kan* godt forekomme i praksis, f.eks. i det såkaldte ballotproblem, hvor der er afgivet et stort antal stemmer på to kandidater (*ballot* er det



Figur 1.8: *Tæthedsfunktionen for en arcussinus fordeling. Den lodrette streg repræsenterer den postulerede observation* $x = 0.1$.

engelske ord for en stemmeseddel). Selv om kandidaterne er lige populære, vil en af dem naturligvis vinde. Hvis man tæller stemmerne sekventielt, kan man registrere på hvilke tidspunkter i optællingen der er stemmelighed. Lad L være det *sidste* sådanne tidspunkt, og lad N være det samlede antal stemmer. Hvis de to kandidater vitterligt er lige populære, og hvis N er meget stor, så vil L/N kunne opfattes som en observation fra en arcussinus fordeling, ifølge den såkaldte arcussinus lov. Det intuitive indhold af arcussinus loven fremgår af figur 1.8: der er forholdsvis stor sandsynlighed for at den ene kandidat lægger sig i spidsen på et tidligt tidspunkt, og forbliver i spidsen, men der er på den anden side også stor sandsynlighed for at kandidaterne bliver ved med at skiftes til at føre lige til det sidste. Det er derimod mindre sandsynligt at en af kandidaterne midt i optællingen pludselig lægger afstand til sin konkurrent.

Hvis man i forbindelse med ballotproblemet sammenholder L/N med en arcussinus fordeling, kan det opfattes som en kontrol af at de to kandidater virkeligt er lige populære, men nok så meget af om vælgerne stemmer uafhængigt af hinanden. Hvis vælgerne reagerer på hinanden, kan man forestille sig at de på et tidspunkt begynder at stemme på den kandidat de opfatter som den sandsynlige vinder, og så kan man i princippet godt komme ud for at A og B følges ad indtil midt i optællingen, hvor A lægger afstand.

Så hvis vi tænker i baner af at kontrollere om vælgerne stemmer uafhængigt af hinanden, så er det faktisk fornuftigt nok at opfatte L/N-værdier omkring 1/2 som kritiske. Skønt det abstrakte argument om at opsøge x-værdier med store tætheder mister kraft, hvis tætheden er multimodal, så kan konkrete, kontekstbaserede argumenter altså nogen gange føre til samme konklusion.

∘

1.2 Flerdimensionale konkordansproblemer

Konkordansproblemer på $\mathbb{R}^k$ er endnu vanskeligere end på $\mathbb{R}$, fordi der opstår geometriske vanskeligheder med at beskrive hvordan "rimelighedsområder" ser ud: I en dimension kan man ikke finde på andet end intervaller, men i to dimensioner kan man forestille sig alt mellem himmel og jord: kugler, akseparallelle kasser, skævt beliggende kasser, ellipser af enhver tænkelig excentricitet og drejning, ...

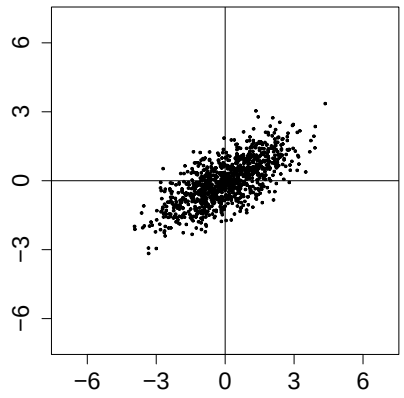
Normalt løser man problemet på den måde at man udfra modellen konstruerer en **testfunktion**, en afbildning $t : \mathbb{R}^k \to \mathbb{R}$, sådan at rimelige observationer $x \in \mathbb{R}^k$ får billede $t(x)$ et fornuftigt, genkendeligt sted på den reelle akse. Typisk vil rimelige observationer under modellen få billede inde omkring 0. Så hvis en konkret observation x har et billede $t(x)$ der ligger langt fra 0, vil man konkludere at observationen er i diskordans med modellen. Geometrisk set svarer denne procedure til at man bruger niveaukurver for testfunktionen til at afgrænse rimelighedsområder. Hvor godt dette program fungerer, er naturligvis meget afhængigt af samspillet mellem modellen ν og testfunktionen t.

Eksempel 1.6 Lad os sammenholde observationen $(-2.5, 2.5)$ med normalfordelingen på $\mathbb{R}^2$ med middelværdi $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ og varians $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. Modellen ν har således tæthed

$$f(x) = \frac{1}{2\pi} \exp\left(-\frac{1}{2} x^T \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^{-1} x\right), \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R} \tag{1.1}$$

med hensyn til det todimensionale Lebesguemål m_2. På figur 1.9 har vi simuleret 1000 observationer fra denne fordeling.

Det er tydeligt på figur 1.9 at de simulerede observationer falder inde i området omkring $(0, 0)$. En testfunktion $t : \mathbb{R}^2 \to \mathbb{R}$ der giver anledning til en fornuftig fortolk-



Figur 1.9: *1000 simulerede observationer fra den todimensionale normalfordeling (1.1).*

ning, kunne være kvadratet på den euklidiske norm,

$$t(x) = x^T x, \qquad x \in \mathbb{R}^2. \tag{1.2}$$

De fleste observationer i figur 1.9 vil få t-værdier inde omkring 0, så store t-værdier er ikke i god overensstemmelse med modellen. På figur 1.10 har vi optegnet et histogram over de 1000 observationer fra figur 1.9 transformeret med t, sammen med testfunktionen regnet ud på den postulerede observation $(-2.5, 2.5)$. Den observerede værdi af testfunktionen er så stor, at man næppe med rimelighed kan hævde at observationen stammer fra den undersøgte normalfordeling, men det er et grænsetilfælde.

Den anførte testfunktion er ikke den eneste mulige, den er faktisk ikke engang særlig god. Et andet valg med samme intuitive egenskaber kunne være den euklidiske norm i sig selv, $\sqrt{t(x)}$, men såvel $t(x)$ som $\sqrt{t(x)}$ har den defekt at de ikke tager hensyn til specielle asymmetri der tydeligt kommer til udtryk i figur 1.9. Med lidt mere viden om normalfordelinger, ville man foretrække

$$s(x) = x^T \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^{-1} x, \quad x \in \mathbb{R}^2, \tag{1.3}$$

der også kan opfattes som kvadratet på en (ikke-euklidisk) norm. Denne transformation er rettet specielt ind mod den aktuelle normalfordeling - tætheden $f(x)$ afhænger

kun af x gennem $s(x)$. Så når vi opfatter små s-værdier som konkordans, store s-værdier som diskordans, så bruger vi i virkeligheden **tætheden** for ν som målestok.

Ydermere kan man vise at $s(\nu)$ er en standardfordeling, det er en χ^2-fordelt med 2 frihedsgrader. Det har historisk været af stor betydning at vælge sine transformationer sådan at de transformerede variable har en standardfordeling, fordi man har været afhængig af tabeller over fordelingerne. I dag spiller denne pointe en mindre rolle, man kan selv generere de nødvendige tabeller på sin computer, om ikke andet kan alle fordelinger simuleres frem.

Under alle omstændigheder afslører denne testfunktion tydeligt at den postulerede observation er dybt urimelig under modellen, se midterfiguren på figur 1.10.

Lad os endelig præsentere et eksempel på en mindre heldig testfunktion. Lad $r : \mathbb{R}^2 \to \mathbb{R}$ være den lineære transformation

$$r\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 + x_2 \,. \tag{1.4}$$

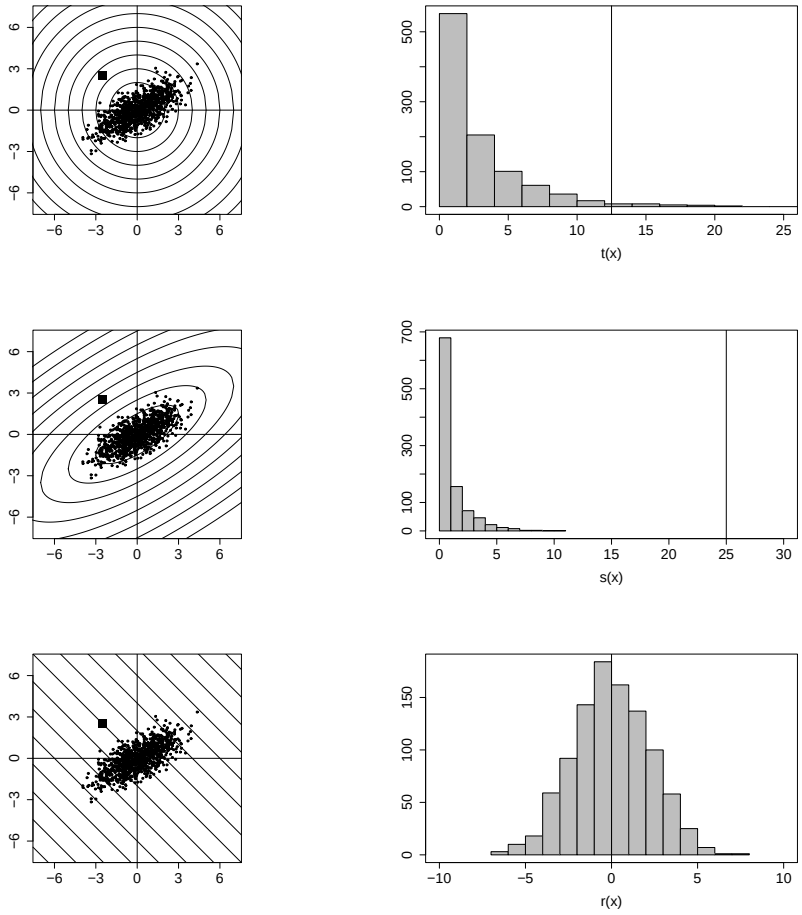
De rimelige observationer under modellen får alle r-værdier tæt ved 0 (bemærk: i modsætning til de tidligere testfunktioner, har r både positive og negative værdier), så observeres en r-værdi langt fra 0 (positiv eller negativ), vil vi igen opfatte modellen som falsificeret. Men i modsætning til tidligere kan fortolkningen ikke vendes om: der er mange x-værdier der passer dårligt med modellen, men som alligevel får en lille r-værdi. Observeres en lille r-værdi skal man være varsom med at fortolke det som om den oprindelige observation er i konkordans med modellen.

Vi siger at r har **ringe følsomhed** overfor visse afvigelser fra modellen. En nøje analyse vil vise at problemet med t er præcis det samme: t tager, modsat s, ikke hensyn til problemets asymmetri, og derfor får t nedsat følsomhed overfor visse afvigelser. Dette forhold fremgår klart af figur 1.10.

∘

Eksempel 1.7 Hvordan vil man sammenholde en observation i $\mathbb{R}^{10}$ med en model der postulerer at koordinaterne er uafhængige og hver især $\mathcal{N}(0,1)$-fordelte? Analogt med både (1.2) og (1.3) i eksempel 1.6 vil man typisk bruge testfunktionen

$$t(x_1, \ldots, x_{10}) = \sum_{i=1}^{10} {x_i}^2.$$



Figur 1.10: *Scatterplot af 1000 simulerede observationer fra den todimensionale normalfordeling (1.1) sammen med niveaukurver for testfunktionerne (1.2) øverst, (1.3) i midten og (1.4) nederst, og med histogrammer over de transformerede observationer. I alle tilfælde har vi sammenlignet med den fiktive observation (-2.5, 2.5) (det firkantede punkt på scatterplottene, den lodrette linie på histogrammerne). På scatterplottene ses det klart at den fiktive observation er urimelig under modellen. Efter transformationen ses dette forhold mere eller mindre i øverste linie, det ses meget klart i midterste linie og det ses slet ikke i nederste linie.*

Under modellen vil typiske observationer falde inde omkring origo i $\mathbb{R}^{10}$, og dermed give anledning til små t-værdier. Faktisk vil man sige at jo mindre t-værdi, jo større konkordans.

Bemærk at hvis ν er produktmålet af 10 $\mathcal{N}(0,1)$-fordelinger, så er $t(\nu)$ netop χ^2-fordelingen med 10 frihedsgrader. Og derfor står vi i den situation, der blev postuleret i eksempel 1.4, hvor man i en konkordansvurdering i forhold til en χ^2-fordeling kun er bange for meget store værdier, og ikke for værdier tæt på nul.

Men hvis man er et meget kritisk gemyt, kan man faktisk godt rejse børster ved synet af små t-værdier. Hvis $t(x_1, \ldots, x_{10})$ er lille, så må alle x_i'erne jo ligge tæt ved 0. Hvor det nok er rimeligt at de fleste x_i'er ligger tæt ved 0, så er det knap så rimeligt at de **alle** gør det - $\mathcal{N}(0,1)$-fordelingen forudsiger jo at der engang imellem kommer store observationer ved uafhængige replikationer.

Det 20. århundredes centrale statistiker, Ronald A. Fisher, skrev i 1936 en artikel (voldsomt polemisk, som alle Fishers artikler) hvor han argumenterer for at der er svindel involveret i Mendels arvelighedseksperimenter - disse eksperimenter fremstilles ellers i alle biologibøger som hjørnestenen i udviklingen af den moderne arvelighedslære, og berømmes for deres redelighed og pædagogiske opbygning. Fishers argument er meget tæt beslægtet med det just anførte, og går på at Mendels eksperimenter simpelthen passer **for** godt med teorien. Den slags argumenter har meget vanskeligt ved at overbevise praktisk arbejdende videnskabsfolk (hvis et eksperiment passer med det teoretisk forudsagte, så understøtter det teorien - ikke det modsatte!) og artiklen førte ikke til nogen varig skade i Mendels almindelige omdømme.

∘

1.3 Eksemplernes budskab

Vi vil nu opsummere erfaringerne fra afsnit 1.1 og 1.2, og indsætte dem i en abstrakt sammenhæng. I almindelighed er argumenterne knap så overbevisende når de gennemføres abstrakt som når de gennemføres i konkrete tilfælde. Men vi ønsker at få en generel tilgang til konkordansproblemet, en abstrakt **metode**. Så må man efterfølgende vurdere metodens egenskaber i de konkrete tilfælde. Selv hvis de abstrakte argumenter er svage, kan de jo godt lede til metoder, der fungerer godt i en bred klasse af problemer.

En grundlæggende erfaring fra eksemplerne er at vi ikke beskæftigede os med at *bekræfte* modellen - vi forsøgte at **falsificere** den. Hvis vi har afgrænset et konkordansområde, et "rimelighedsområde" for modellen, og konstaterer at den gjorte observation ligger i dette område, så har vi ikke bevist at modellen er rigtig. Forkerte modeller kan sagtens lede til forudsigelser, der tilfældigvis bekræftes eksperimentelt - især hvis de ikke siger noget særlig specifikt, og dermed leder til meget brede konkordansområder.

Hvis vi derimod konstaterer at den gjorte observation **ikke** ligger i konkordansområdet, så kan vi drage en betydningsfuld konklusion: modellen er forkert. Denne kritiske holdning gennemsiver al statistik: der er ingen modeller der er rigtige, der er ingen hypoteser der er sande - der er blot påstande vi endnu ikke har falsificeret.

I de undersøgte eksempler forsøgte vi at finde ud af hvad modellen forudsagde ved at simulere et stort antal fiktive observationer fra modellen. Det ledte os til at sammenligne den gjorte observation med *tætheden* for modellen, enten med hensyn til tællemålet eller med hensyn til et Lebesguemål. Overgangen fra simulerede observationer til tæthed er ikke en logisk nødvendighed: det er et argumentspring, eller en analogislutning. Vi vil i afsnit 1.4 give en mere præcis, matematisk begrundelse for overgangen (i hvert fald i visse modeller), men det er vigtigt at overgangen accepteres: alle vores senere overvejelser vil tage udgangspunkt i at modellens forudsigelser opsummeres i dens tæthed, på den måde at områder med (relativ) høj tæthed er "rimeligere" end områder med lav tæthed.

Der er dog mange situationer hvor denne vurdering af konkordans, baseret udelukkende på modellens "indre" egenskaber, må vejes op imod forhåndsviden om hvordan afvigelser fra modellen vil tage sig ud. Denne forhåndsviden vil ofte opstå i forbindelse med transformationer af data, se afsnit 1.2.

1.4 Modeller med tæthed

De modeller vi har set på, har haft karakteren $\nu = f \cdot \mu$, hvor μ er et grundmål på $(\mathcal{X}, \mathbb{E})$. Sammenligning med simulerede observationer under modellen har ledt frem til en sammenligning med tætheden f. Men det argument der leder fra de simulerede observationer til f er **ikke** holdbart i alle tilfælde, det er et udtryk for specielle egenskaber ved de grundmål der er benyttet, altså tællemålet på $\mathbb{Z}$ og Lebesguemålet på $\mathbb{R}^k$.

Eksempel 1.8 Ethvert sandsynlighedsmål ν på $(\mathcal{X}, \mathbb{E})$ opfylder at $\nu = 1 \cdot \nu$. Tætheden i denne fremstilling af ν, hvor ν selv opfattes som et grundmål, er altså konstant 1. Og denne tæthed kan selvfølgelig ikke bruges til at skelne rimelige observationer fra urimelige observationer.

∘

Lad os begrænse diskussionen til tilfældet $(\mathcal{X}, \mathbb{E}) = (\mathbb{R}, \mathbb{B})$. Argumentet for at benytte tætheden passerede via et **histogram**, baseret på en opdeling af den reelle akse i et antal delintervaller. Lad os som udgangspunkt benytte de disjunkte intervaller[1]

$$\left(\frac{j-1}{2^k}, \frac{j}{2^k}\right], \qquad j \in \mathbb{Z}. \tag{1.5}$$

Her er $k \in \mathbb{N}$ et fast tal, der bestemmer **finheden** af opdelingen. Lad os definere $j_k : \mathbb{R} \to \mathbb{Z}$ ved

$$j_k(x) = [2^k x], \qquad x \in \mathbb{R}.$$

Denne definition sikrer at

$$x \in \left(\frac{j_k(x)-1}{2^k}, \frac{j_k(x)}{2^k}\right] \quad \text{for alle } x \in \mathbb{R}.$$

Det **empiriske histogram** på baggrund af observationerne $x_1, \ldots, x_n$ er nu funktionen $H_{k;\, x_1,\ldots,x_n} : \mathbb{R} \to [0, \infty)$ givet ved

$$\begin{aligned} H_{k;\, x_1,\ldots,x_n}(x) &= 2^k\, \epsilon_{x_1,\ldots,x_n}\left(\left(\frac{j_k(x)-1}{2^k}, \frac{j_k(x)}{2^k}\right]\right) \\ &= \frac{2^k}{n} \sum_{i=1}^{n} 1_{((j_k(x)-1)/2^k,\ j_k(x)/2^k]}(x_i)\,, \end{aligned}$$

hvor $\epsilon_{x_1,\ldots,x_n}$ er det empiriske mål i punkterne $x_1, \ldots, x_n$. Histogramfunktionen er konstant hen over intervallerne (1.5), og normeringen sikrer at det samlede areal under grafen for $H_{k;\, x_1,\ldots,x_n}(x)$ er 1. Bemærk at "histogrammerne" i afsnit 1.1 og 1.2 **ikke** er normeret på den politisk korrekte måde.

Frekvensfortolkningen af sandsynligheder giver at med et stort antal simulerede observationer, vil det empiriske histogram stort set være identisk med et "grænsehistogram" bestemt af ν. Der gælder nemlig at

$$H_{k;\, x_1,\ldots,x_n}(x) \simeq 2^k\, \nu\left(\left(\frac{j_k(x)-1}{2^k}, \frac{j_k(x)}{2^k}\right]\right) \quad \text{for alle } \ x \in \mathbb{R}.$$

[1] Når statistiske programpakker tegner histogrammer, vil tegningen ikke være baseret på intervaller, der er fastlagt på forhånd, men på intervaller hvis antal, længde og placering afhænger af de konkrete data. Den type histogrammer er konceptuelt mere indviklede end ovenstående diskussion lader ane.

Endnu har vi ikke benyttet nogen speciel struktur af ν, argumentet er fuldstændig generelt, og man kan nu holde den faktiske observation op mod grænsehistogrammet.

Hvis ν har tæthed med hensyn til tællemålet på $\mathbb{Z}$ og vi benytter $k = 1$, så **er** grænsehistogrammet simpelthen sandsynlighedsfunktionen, eftersom der kun er ét heltal i hvert af intervallerne (1.5).

Hvis $\nu = f \cdot m$ må man argumentere via en grænseovergang: finheden af de intervaller der bestemmer histogrammerne, er jo temmelig arbitrær, og det føles naturligt at se på meget fine inddelinger, hvis man har mange observationer. Når vi således lader antallet af simulationer blive stort, vil vi også betragte fine histogrammer.

Sætning 1.9 *Lad f være en $\mathcal{M}(\mathbb{R}, \mathbb{B})$-funktion. Hvis f er integrabel mht. Lebesguemålet m, så vil*

$$2^k \int_{(j_k(x)-1)/2^k}^{j_k(x)/2^k} f(y)\, dy \to f(x) \quad \textit{for } k \to \infty \tag{1.6}$$

for m-næsten alle $x \in \mathbb{R}$.

□

Sætning 1.9 er let at vise hvis f er kontinuert, men den er sand uden anden antagelse end at f er målelig og integrabel. Sætningen kan fortolkes[2] på den måde at hvis k er stor, så har grænsehistogrammet stort set samme **form** som grafen af f. Så at holde den faktisk gjorte observation op mod et meget fint grænsehistogram, svarer til at holde observationen op mod f.

Grundmålet m spiller en stor rolle i dette argument. Et analogt resultat til sætning 1.9 kan vises for Lebesguemålet på $(\mathbb{R}^k, \mathbb{B}_k)$. Men benyttes andre mål som grundmål, bliver påstande af denne type lodret forkerte.

Hele argumentkæden, der leder fra simulerede observationer til tæthed, har den grundlæggende svaghed, at den tillægger *histogrammer* en voldsom betydning, som begrebet måske ikke helt kan bære. Måske kan de simulerede observationer opsummeres på en anden måde? Hvis man kan finde en måde, der ikke opererer med en serie af finere og finere ækvidistante inddelinger af den reelle akse, så kan man måske tildele tætheder betydning, også for andre grundmål end Lebesguemålet. Vi vil dog

[2] En præcisering kræver en form for uniform konvergens i (1.6), hvilket f.eks. kan opnås hvis f opfylder en Lipschitzbetingelse.

ikke bruge kræfter på en sådan jagt, de sandsynlighedsmål der optræder i forbindelse med analyse af den virkelige verden, har alle tæthed med hensyn til et "fornuftigt" grundmål.

1.5 Konkordansområder på fast niveau

Vi har indtil nu gennemgået en række eksempler hvor det var klart at en observation ikke passede med en givet model, og en række andre eksempler hvor det var mindre klart hvad man skulle konkludere. Og vi har argumenteret os frem til at hvis modellen har tæthed med hensyn til et fornuftigt grundmål, så bør denne tæthed spille en vigtig rolle når man forsøger at diskutere konkordans. Vi har også set at har man yderligere information om hvor kritiske observationer ligger, så bør man inddrage denne information.

Men vi har endnu ikke taget stilling til hvordan konkordansspørgsmålet *egentlig* skal besvares. Vi vil her gennemgå to forskellige strategier. Den første strategi har luret i baggrunden igennem hele diskussionen indtil nu. Den består i hårdt og brutalt at svare ja eller nej på spørgsmålet: er observationen i overensstemmelse med modellen? Man deler $\mathcal{X}$ i to, dels en **konkordansmængde** A, bestående af punkter som man opfatter som i overensstemmelse med modellen. Og dels en **kritisk mængde** A^c, bestående af alle resterende punkter, der altså opfattes som i diskordans med modellen. Egenskaberne ved en konkordansmængde A opsummeres i dens **niveau** $\alpha = \nu(A^c)$.

Dette niveau er sandsynligheden for - selv om modellen er sand - ved et uheld at gøre en observation *udenfor* konkordansmængden, og dermed konkludere at modellen nok er forkert. Risikoen for at forkaste en korrekt model for eksperimentet virker meget ubehagelig, så for en umiddelbar betragtning skal man gøre α så lille som muligt. Og det er let nok at gøre α lille: man skal blot gøre A stor. Jo mere man opfatter som i konkordans med modellen, jo sværere er det ved et uheld at komme til at konkludere at modellen er forkert.

Hvis man sætter $A = \mathcal{X}$ så er $\alpha = 0$ og man kommer aldrig til at smide modellen væk. Alle er glade - lige til man kommer i tanke om, at dette forhold også gælder hvis modellen er forkert. Uanset hvor absurd modellen er, har man i så fald forhindret sig selv i at opdage det... For at være i stand til at opdage at modellen er forkert, må man altså søge at gøre A så lille som muligt, og dermed gør man også α stor. Et godt valg af A skal afbalancere disse modstridende hensyn.

Man insisterer ofte på at α har en bestemt værdi (typisk brugte værdier er 0.05 eller 0.01), og søger så at konstruere konkordansområder A med dette niveau, under skyldig hensyntagen til hvor kritiske observationer forventes at ligge. Falder den faktisk gjorte observation udenfor A siger man at modellen er **signifikant** diskordant på niveau α. Vi vil nu angive en række konkrete konstruktioner, hvis $\mathcal{X} = \mathbb{R}$ og hvis ν er kontinuert.

1) Det tæthedsbaserede område. Hvis ν har tæthed f mht. Lebesguemålet m, og vi ikke har nogen information om hvor kritiske punkter ligger, vil vi ofte insistere på et konkordansområde af formen

$$A = \{x \in \mathbb{R} \mid f(x) > c\}$$

hvor $c > 0$ er et reelt tal, rettet ind så $\nu(A) = 1 - \alpha$.

2) Det venstrestillede område. Hvis vi ved at store værdier er kritiske, vil vi ofte insistere på et konkordansområde af formen

$$A = (-\infty, c)$$

hvor c er et reelt tal, rettet ind så $\nu(A) = 1 - \alpha$. Når det venstrestillede område bruges, ved man ofte at negative observationer ikke kan forekomme, og konkordansområdet kunne da ligeså godt vælges af formen $(0, c)$.

3) Det højrestillede område. Hvis vi ved at små værdier er kritiske, vil vi ofte insistere på et konkordansområde af formen

$$A = (c, \infty)$$

hvor c er et reelt tal, rettet ind så $\nu(A) = 1 - \alpha$. Når det højrestillede område bruges, ved man ofte at observationer udenfor enhedsintervallet ikke kan forekomme, og konkordansområdet kunne da ligeså godt vælges af formen $(c, 1)$.

4) Det symmetriske område. Hvis vi ved at numerisk store værdier er kritiske, både de negative og de positive, vil vi ofte insistere på et område af formen

$$A = (c_1, c_2)$$

hvor c_1 og c_2 er rettet ind så $\nu((-\infty, c_1]) = \frac{\alpha}{2}$ og $\nu([c_2, \infty)) = \frac{\alpha}{2}$.

Hvis ν er unimodal og symmetrisk omkring sit centrum, vil det tæthedsbaserede og det symmetriske område falde sammen. Men for asymmetriske fordelinger vil de to

områder i almindelighed være forskellige. Man vil som regel principielt foretrække det tæthedsbaserede område, men eftersom det er vanskeligere at finde ud fra de tilgængelige tabeller, lader man sig i praksis ofte nøje med det symmetriske område.

Det er vigtigt at forstå at konkordansområder af alle varianter, er udsagn om **modellen**. De kan altså udregnes før man har observeret noget som helst.

Eksempel 1.10 Lad os konstruere konkordansområder svarende til niveauet $\alpha = 0.05$ for modellen i eksempel 1.2, dvs. for $\mathcal{N}(13.42, 0.11^2)$-fordelingen. Denne fordeling er symmetrisk omkring 13.42, og tæthederne aftager når man bevæger sig væk fra denne værdi. Dermed vil det tæthedsbaserede område og det symmetriske område falde sammen. Ud fra 2.5% og 97.5% fraktilerne for en standard normalfordeling, ses dette symmetriske område let at være

$$(13.20, 13.64)$$

hvilket klart indeholder den postulerede observation på 13.24. I dette eksempel er det mindre naturligt at udregne de venstre- og højrestillede konkordansområder, men man kan naturligvis gøre det, og finder dem til at være hhv.

$$(-\infty, 13.60) \quad \text{og} \quad (13.24, \infty).$$

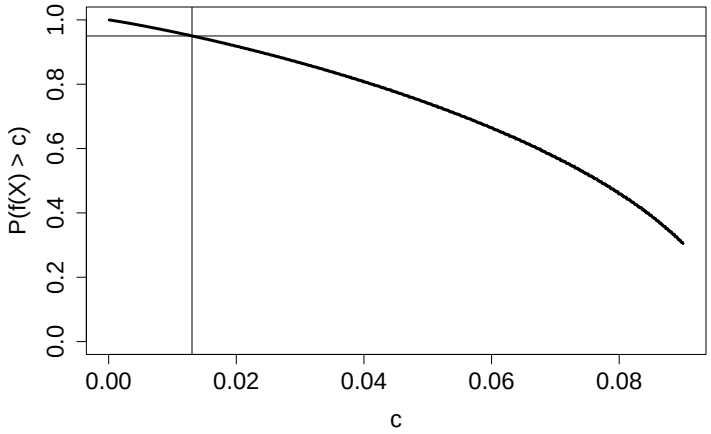
Hvis man af en eller anden grund ved at fabrikkens egen maskine er speciel derved at den laver *større* søm end verdens andre sømfremstillende maskiner, så vil fundet af et stort søm på ingen måde få en til at tvivle på at sømmet er hjemmegjort, mens fundet af et lille søm giver anledning til en sådan tvivl. Og i så tilfælde kunne det måske være relevant med et højrestillet konkordansområde (hvilket måske og måske ikke inkluderer den postulerede observation, det er lige på grænsen). Men det er svært ikke at opfatte denne argumentation som noget søgt.

∘

Eksempel 1.11 Lad os konstruere konkordansområder svarende til niveauet $\alpha = 0.05$ for modellen i eksempel 1.4, dvs. for χ^2-fordelingen med 10 frihedsgrader. Denne fordeling er asymmetrisk, så de forskellige konstruktioner falder **ikke** sammen. Det tæthedsbaserede konkordanssområde findes ved at optegne grafen for

$$c \mapsto P(f(X) > c)$$

hvor f er tætheden for χ^2-fordelingen med 10 frihedsgrader, og hvor X er en stokastisk variabel med denne fordeling, og så aflæse hvor grafen krydser niveauet 0.95.



Figur 1.11: *Grafen for $c \mapsto P(f(X) > c)$ hvor X er χ^2-fordelt med 10 frihedsgrader, og hvor f er tætheden for denne χ^2-fordeling. Den vandrette linie svarer til niveauet 0.95, mens den lodrette linie markerer hvor grafen skærer den vandrette linie.*

På figur 1.11 ser vi at dette kryds sker for $c = 0.013$, og dermed er det tæthedsbaserede konkordansområde

$$\{x \in \mathbb{R} | f(x) > 0.013\} = (2.40, 18.92).$$

Det symmetriske konkordansområde findes ud fra 2.5% og 97.5% fraktilen for χ^2-fordelingen med 10 frihedsgrader som

$$(3.25, 20.48).$$

Forskellen på disse to konkordansområder skyldes at χ^2-fordelingens skævhed. Sandsynligheden under modellen for at falde uden for det tæthedsbaserede konkordansområde er naturligvis 5%, men der er en asymmetri: sandsynligheden for at falde nedenfor er mindre end 1%, sandsynligheden for at falde ovenfor er større end 4%.

De venstre- og højrestillede områder udregnes til

$$(0, 18.31) \quad \text{og} \quad (3.94, \infty).$$

Heraf er formentlig kun det første meningsfuldt i eksemplet.

∘

Eksempel 1.12 Lad os konstruere konkordansområder svarende til niveauet $\alpha = 0.05$ for modellen i eksempel 1.6, dvs. for den todimensionale normalfordeling med middelværdi $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ og varians $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

Vi konstruerer først et område baseret på transformationen s givet i (1.3). Som anført p. 17 kan man vise, at hvis X har den angivne fordeling så er $s(X)$ χ^2-fordelt med 2 frihedsgrader, og vi må opfatte store s-værdier som kritiske. Derfor anvender vi et venstrestillet konkordansområde for $s(X)$ og på et 5% niveau får vi $I_1 = (0, 5.99)$. Fører vi dette område tilbage til de oprindelige observationers plan, får vi

$$A = s^{-1}((0, 5.99)), \tag{1.7}$$

der altså er et konkordansområde for X for et 5% niveau. Dette område er en udfyldt ellipse, som det ses på figur 1.12, hvor A er farvet sort.

Vi gentager konstruktionen, denne gang baseret på transformationen r givet i (1.4). Man kan vise at $r(X)$ er $\mathcal{N}(0, 5)$-fordelt, og numerisk store værdier (såvel positive som negative) må opfattes som kritiske. Derfor anvender vi et symmetrisk konkordansområde for $r(X)$ og på et 5% niveau får vi $I_2 = (-4.38, 4.38)$. Fører vi dette område tilbage til de oprindelige observationers plan, får vi

$$B = r^{-1}((-4.38, 4.38)), \tag{1.8}$$

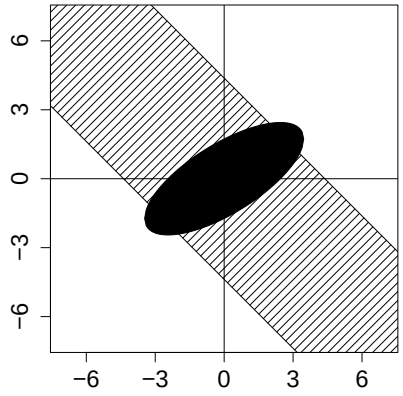
der altså er et konkordansområde for X for et 5% niveau. Dette område er en ubegrænset strimmel, som det ses på figur 1.12, hvor B er skraveret.

Sammenligner man de to variationsområder for X på figur 1.12, falder det i øjnene at de er meget forskellige, skønt de har samme sandsynlighedsindhold - vel at mærke hvis modellen er sand. Hvis modellen er forkert, vil det formentlig stadig være let at gøre en observation i B, mens det vil være vanskeligt at gøre en observation i A. Vi får bekræftet vores tidligere konklusion om at det er ineffektivt at vurdere konkordans ved hjælp af r-transformationen i denne model.

∘

Hvis ν er et sandsynlighedsmål der lever på $\mathbb{Z}$ kan man i princippet gå frem som hidtil. Men der opstår visse problemer, fordi man ikke altid for et givet α kan finde en mængde A der løser ligningen $\nu(A) = 1 - \alpha$. Idet vi ønsker at kontrollere sandsynligheden for fejlagtigt at forkaste en sand model, så er det naturligt generelt at definere f.eks. det venstrestillede konkordansområde som $A = (-\infty, c]$ hvor

$$c = \min\{x \in \mathbb{R} \mid \nu((-\infty, x]) \geq 1 - \alpha\}$$



Figur 1.12: *To konkordansområder på 5% niveau for den todimensionale normalfordeling fra eksempel 1.6. Den udfyldte mængde er givet ved (1.7), den skraverede mængde er givet ved (1.8).*

(altså så der *højst* er sandsynlighed α for at gå i vandet) og det giver nogen gange det problem at $\nu((c,\infty)) < \alpha$. Vi siger at der er forskel på områdets **nominelle** egenskaber (givet ved α) og dets **faktiske** egenskaber (givet ved $\nu((c,\infty))$. Disse forskelle er desværre et *fact of life* ved diskrete fordelinger, og et problem som man må slås med dagligt.

Eksempel 1.13 For binomialfordelingen med længde 5 og successandsynlighed 0.5 gælder der at

$$P(X = 0) = P(X = 5) = 0.031,$$

og derfor må det nominelle centrale konkordansområde på 5% være hele $\{0, 1, \ldots, 5\}$. Men det faktiske niveau for dette område er naturligvis 0%.

∘

Eksempel 1.14 Lad os konstruere konkordanssområde svarende til niveauet $\alpha = 0.05$ for modellen i eksempel 1.1, dvs. for binomialfordelingen med længde 66170 og succesparameter 0.5. Idet denne fordeling er symmetrisk omkring 33085, og idet sandsynlighedsfunktionen aftager når vi bevæger os væk fra denne værdi, vil det

tæthedsbaserede og det symmetriske konkordanssområde være identiske. Vi finder at

$$P(X \leq 32832) = 0.0248, \quad P(X \leq 32833) = 0.0253$$

og derfor er det centrale konkordansområde med et nominelt niveau på 5%

$$\{32833 \ldots, 33337\}.$$

Vi ser endvidere at det faktiske niveau for dette område er givet ud fra

$$P(32833 \leq X \leq 33337) = 0.9504.$$

I dette tilfælde er der ikke stor forskel på det faktiske niveau 4.96% og det nominelle niveau på 5%. Det skyldes at punktsandsynlighederne er meget små.

∘

1.6 Konkordans via p-værdier

I en mere raffineret tilgang til konkordansproblemet svarer man hverken ja eller nej til spørgsmålet om hvorvidt observationen passer med modellen. Man tager i stedet udgangspunkt i at enhver observation i et vist omfang modsiger enhver model, og forsøger at **kvantificere** i hvor høj grad den konkrete observation modsiger den givne model. Den normale måde det sker på, er at man udregner sandsynligheden (under modellen) for en observation der modsiger modellen i *mindst samme grad* som den faktisk gjorte observation. Det udregnede tal kaldes observationens **signifikanssandsynlighed** eller blot dens p**-værdi**.

For at kunne bruge denne ide i praksis, må man kunne afgøre hvilken af to observationer x og x' der modsiger modellen mest. Vi giver her et antal muligheder, der kan bruges hvis $\mathcal{X} = \mathbb{R}$.

1) Den tæthedsbaserede p-værdi. Hvis ν har tæthed f mht. Lebesguemålet m, og vi ikke har nogen information om hvor kritiske punkter ligger, vil vi opfatte x som i bedre overensstemmelse med modellen end x' hvis $f(x) \geq f(x')$. For observationen x udregnes derfor

$$P(f(X) \leq f(x))$$

altså sandsynligheden for at se en mindre tæthed end tætheden i det faktisk observerede punkt.

2) Den venstrestillede p-værdi. Hvis vi ved at store værdier er kritiske, vil vi opfatte x som i bedre overensstemmelse med modellen end x' hvis $x \leq x'$. For observationen x udregnes derfor

$$P(X \geq x).$$

Hvis F er fordelingsfunktionen for ν, udregner vi altså $1 - F(x - 0)$.

3) Den højrestillede p-værdi. Hvis vi ved at små værdier er kritiske, vil vi opfatte x som i bedre overensstemmelse med modellen end x' hvis $x \geq x'$. For observationen x udregnes derfor

$$P(X \leq x).$$

Hvis F er fordelingsfunktionen for ν, udregner vi simpelthen $F(x)$.

4) Den symmetriske p-værdi. Her udregnes

$$2 \cdot \min\{P(X \leq x), P(X \geq x)\}.$$

Ideen bag den symmetriske p-værdi er at den hale af fordelingen, som den gjorte observation befinder sig i, får lov at tælle dobbelt. Hvis F er fordelingsfunktionen for ν, udregner vi altså

$$2 \cdot \min\{F(x), 1 - F(x - 0)\}.$$

Hvis $\nu = f \cdot m$ og hvis tætheden f er unimodal og symmetrisk omkring sit toppunkt, så falder den tæthedsbaserede og den symmetriske p-værdi sammen. For asymmetriske fordelinger bliver de to konstruktioner derimod forskellige.

Retningslinierne for hvilken p-værdi man bør udregne, ligner de tilsvarende retningslinier for konstruktion af konkordansområder på fast niveau: i fravær af yderligere information vil man foretrække den tæthedsbaserede p-værdi - omend den kan være vanskelig at regne ud, og praktiske forhold derfor kan tale for en symmetrisk p-værdi. Har man yderligere information, der gør at man er specielt bange for store (eller små) observationer, så udregner man naturligvis en venstrestillet (eller højrestillet) p-værdi.

Sammenhængen mellem p-værdier og konkordansområder på fast niveau er følgende: man kan efter en vis procedure (det være sig tæthedsbaseret, venstrestillet, højrestillet eller symmetrisk) konstruere konkordansområder svarende til samtlige niveauer $\alpha \in (0, 1)$. Lad $A(\alpha)$ være konkordansområdet svarende til niveau α.

Disse konkordansområder ligger inde i hinanden, sådan at små α'er giver store $A(\alpha)$-mængder. En konkret x-værdi vil ligge i $A(\alpha)$ hvis blot α er tilstrækkeligt tæt ved nul. Hvis p-værdien for x (udregnet efter den analoge procedure) er lig p, så gælder der at

$$p = \sup\{\alpha \in (0,1) \mid x \in A(\alpha)\}.$$

Vi ser altså at p er det største niveau hvorpå vi ville opfatte x som i konkordans med modellen.

De fleste statistikere er enige om at p-værdien er væsentlig mere informativ end den blotte afrapportering af om modellen accepteres eller forkastes på et givet niveau. Man skal dog være opmærksom på at kun **små** p-værdier er numerisk meningsfulde: der er ingen grund til at hidse sig op over at en vis p-værdi er 90%, mens en anden kun er 50% - begge dele betyder at observationen ikke modsiger modellen nævneværdigt. Man støder ofte på følgende tommelfingerregel:

p-værdi	Fortolkning
0% - 1%	Stærk evidens mod modellen
1% - 5%	Moderat evidens mod modellen
5% - 10%	Svag evidens mod modellen
10% - 100%	Ingen evidens mod modellen

Eksempel 1.15 Vi vil finde p-værdier svarende til observationen og modellen i eksempel 1.2, dvs. for observationen 13.24 i forhold til $\mathcal{N}(13.42, 0.11^2)$-fordelingen. Da fordelingen er symmetrisk omkring sit centrum, og da tæthederne aftager når man bevæger sig væk fra denne værdi, vil de to centrale p-værdier (den tæthedsbaserede og den symmetriske) være ens. Vi standardiserer observationen til

$$\frac{x - 13.42}{0.11} = -1.64$$

hvilket er 0.051-fraktil i en standard normalfordeling. Og dermed er den centrale p-værdi af den gjorte observation 0.101. De venstre- og højrestillede p-værdier udregnes tilsvarende til 0.949 hhv. 0.051. Som tidligere nævnt, skal man være ganske fantasifuld for at finde de venstre- og højrestillede p-værdier relevante i dette eksempel.

∘

Eksempel 1.16 Vi vil finde p-værdier svarende til observationen og modellen i eksempel 1.4, altså for observationen $x = 1$ i forhold til χ^2-fordelingen med 10 friheds-

grader. Hvis f betegner tætheden for den relevante χ^2-fordeling, ser vi at

$$f(1) = f(27.519) = 0.00079,$$

og dermed er den tæthedsbaserede p-værdi lig

$$1 - P(1 < X < 27.519) = 0.0023.$$

Den symmetriske p-værdi findes til

$$2 \cdot \min\{P(X \leq 1), P(X \geq 1)\} = 2P(X \leq 1) = 0.0003$$

hvilket er faktor 10 mindre end den tæthedsbaserede p-værdi, skønt de begge er enige om at der er kraftig evidens mod modellen. Den venstrestillede p-værdi udregnes til 0.9998, hvilket fører til den stik modsatte konklusion: observationen er i fin overensstemmelse med modellen.

Hvis man skal udføre en konkordansundersøgelse af en observation i forhold til en χ^2-fordeling, vil man næsten altid anse store værdier for kritiske, og derfor er den venstrestillede p-værdi formentlig den mest relevante i dette eksempel, sammen med dens budskab om konkordans.

∘

Eksempel 1.17 Vi vil finde p-værdier for den fiktive observation (-2.5, 2.5) i forhold til modellen i eksempel 1.6, dvs. for den todimensionale normalfordeling med middelværdi $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ og varians $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

Vi udregner først en p-værdi, baseret på transformationen s givet i (1.3). Som anført p. 17 kan man vise, at hvis X har den angivne fordeling så er $s(X)$ χ^2-fordelt med 2 frihedsgrader, og vi må opfatte store s-værdier som kritiske. Den fiktive observation får en s-værdi på 31.25, hvilket er 0.9999998-fraktil i χ^2-fordelingen med to frihedsgrader. Dermed er den venstrestillede p-værdi $2 \cdot 10^{-7}$. Dette er en ekstremt kraftig evidens mod at den oprindelige observation skulle stamme fra den angivne todimensionale normalfordeling.

Vi kan også udregne en p-værdi, baseret på transformationen r givet i (1.4). Man kan vise at $r(X)$ er $\mathcal{N}(0, 5)$-fordelt, og numerisk store værdier (såvel positive som negative) må opfattes som kritiske. Derfor udregner vi en central p-værdi. Den fiktive observation får en r-værdi på 0, hvilket er 0.5-fraktil i $\mathcal{N}(0, 5)$-fordelingen. Dermed er den centrale p-værdi lig 1. Denne måde at betragte observation og model på, leder

ikke til nogen iøjnefaldende diskordans, tværtimod. Vi får altså repeteret at det er ineffektivt at vurdere konkordans ved hjælp af r-transformationen i denne model.

∘

Ved udregning af p-værdier, er der ikke specielle problemer med diskrete fordelinger - der er intet skel mellem nominelle og faktiske p-værdier. Men problemerne kan snige sig ind ad bagvejen i forbindelse med *fortolkningen* af p-værdierne, især hvis man holder sig firkantet til grænserne 1%, 5% og 10% som skel mellem kvalitativt forskellige konklusioner.

Eksempel 1.18 Vi vil finde p-værdier for observationen 34055 i forhold til Bin(66170, 0.5)-fordelingen, sådan som problemet fremgår af eksempel 1.1. Man kan vise at hvis X er Bin(66170, 0.5)-fordelt, så er

$$P(X \geq 34055) = 2.38 \cdot 10^{-14}.$$

Idet både små og store værdier er kritiske, vil vi finde en central p-værdi, og den udregnes til at være $4.75 \cdot 10^{-14}$, hvilket er hysterisk stærk evidens mod modellen.

∘

1.7 Approksimative konkordansproblemer

Ofte støder man ind i en variant af konkordansproblemet, hvor man godt **ved** at modellen er forkert, men hvor man har en forestilling om at den i en eller anden forstand er meget tæt på at være rigtig. Vi kan give problemet følgende formulering:

Det approksimative konkordansproblem: *Lad* $(\Omega, \mathbb{F}, P)$ *og* $(\mathcal{X}, \mathbb{E}, \nu)$ *være sandsynlighedsfelter, og lad* $X : (\Omega, \mathbb{F}) \to (\mathcal{X}, \mathbb{E})$ *være en stokastisk variabel.*
Lad observationen $X = x$ *være givet. Er det rimeligt at hævde at* $X(P) \approx \nu$*?*

Så længe vi ikke forklarer hvad vi mener med at to fordelinger er "approksimativt ens", er dette spørgsmål ud fra en matematisk synsvinkel endnu mere upræcist end det oprindelige konkordansproblem. Vi kunne i princippet godt præcisere. Vi kunne indføre et afstandsbegreb mellem fordelinger, f.eks. kunne vi lade afstanden mellem to fordelinger ν_1 og ν_2 på $(\mathbb{R}, \mathbb{B})$ være bestemt som

$$d(\nu_1, \nu_2) = \sup_{x \in \mathbb{R}} |F_1(x) - F_2(x)| \tag{1.9}$$

hvor F_1 og F_2 er de to tilhørende fordelingsfunktioner. Dette afstandsbegreb er vel hårdt for mange formål, men der findes en række blødere metrikker man kunne bruge i stedet. Det approksimative konkordansproblem kan nu formuleres på denne måde: hvis $X = x$, kan man så med rimelighed hævde at $d(X(P), \nu) < \epsilon$ for et givet ϵ?

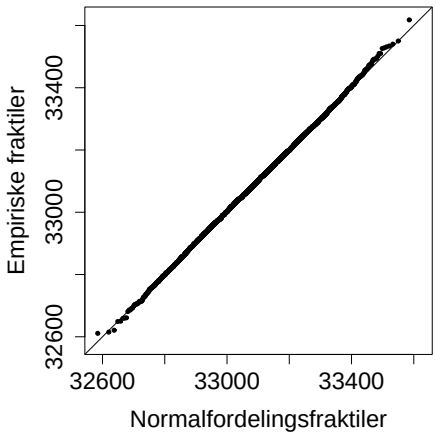
Det er en vigtig pointe at vi ikke ønsker at gå ind på disse præciseringer. Vi vil tværtimod insistere på at opfatte det approksimative spørgsmål som upræcist, som et spørgsmål der kræver fortolkning og som kræver indsigt i de videnskabelige problemer, der giver anledning til at konkordansspørgsmålet rejses.

En lang række statistiske problemer har i deres formulering indbygget en *størrelse* N af det udførte eksperiment. Dette N repræsenterer som regel antallet af udførte deleksperimenter, antallet af undersøgte individer eller lignende. En étdimensional opsummering af hele det udførte eksperiment vil da ofte have en fordeling som er approksimativt kendt, og hvor forskellene mellem den faktiske fordeling og den approksimative fordeling forsvinder for $N \to \infty$. Vi kalder metoder der baserer sig på denne form for approksimation for **asymptotiske metoder**.

Eksempel 1.19 I eksempel 1.1 diskuterede vi en binomialfordeling med længde N og successandsynlighed 0.5. Pointen er at N var meget stor, og det er velkendt at en sådan binomialfordeling er tæt beslægtet med en normalfordeling med samme middelværdi og varians. Altså $\text{Bin}(N, 0.5) \approx \mathcal{N}\left(N/2, N/4\right)$.

Som illustration af dette forhold har vi i figur 1.13 optegnet et QQ-plot af 10.000 simulerede observationer fra en Bin(66170, 0.5)-fordeling mod $\mathcal{N}(33085, 16, 542, 5)$. Som det ses er overenstemmelsen endog meget god, ingen ville protestere hvis vi hævdede at observationerne faktisk stammede fra den pågældende normalfordeling. Og overensstemmelsen bliver endnu bedre i grænsen $N \to \infty$.

Vi fristes derfor til at holde den faktisk gjorte observation $x = 34055$ op - eller at **evaluere** den - mod den approksimerende normalfordeling fremfor mod den eksakte binomialfordeling. Denne approksimative evaluering er let at udføre - alt hvad man har brug for, er en tabel over fraktilerne i en standard normalfordeling. Hvorimod den eksakte evaluering kræver adgang til en Bin(66170, 0.5)-fordeling, hvad man formentlig kun har via en computer. Evalueres x i den approksimative normalfordeling fås en **approksimativ** p**-værdi** på $4.64 \cdot 10^{-14}$ - ikke helt det samme resultat som i eksempel 1.18, men meget tæt på.



Figur 1.13: *Et QQ-plot af 10.000 simulerede observationer fra en Bin(66170, 0.5)-fordeling mod* $\mathcal{N}(33085, 16, 542, 5)$*-fordelingen.*

I praksis ville man være tilbøjelig til at udregne

$$Y = \frac{(X - N/2)^2}{N/4}.$$

Hvis X var eksakt $\mathcal{N}(N/2, N/4)$-fordelt, så ville Y være eksakt χ^2-fordelt med 1 frihedsgrad. Hvis X er Bin$(N, 0.5)$-fordelt, og dermed approksimativt normalfordelt, så er Y approksimativt χ^2-fordelt med 1 frihedsgrad, og approksimationen er god, når N er stor. Idet X-værdier langt fra $N/2$ er kritiske for binomialfordelingsmodellen, må store Y-værdier opfattes som kritiske. I det konkrete eksempel udregner vi

$$y = \frac{(34055 - 33085)^2}{16,542.5} = 56.88,$$

hvilket giver en venstrestillet approksimativ p-værdi på $4.64 \cdot 10^{-14}$ som før. ∘

Eksempel 1.20 I eksempel 1.5 blev der refereret til den såkaldte arcussinus lov. Den omhandlede en størrelse L/N der er diskret i sin natur - både L og N er heltal - og

derfor ikke kan være eksakt arcussinus fordelt. Men approksimationen af arcussinus fordelingen til den eksakte fordeling er god når N er stor.

∘

Det fundamentale asymptotiske resultat er den centrale grænsesætning (*CLT - central limit theorem*) der i sin simpleste form siger følgende:

Sætning 1.21 (CLT) *Hvis $X_1, X_2, \ldots, X_N$ er uafhængige, identisk fordelte stokastiske variable med*

$$EX_1 = 0, \quad EX_1{}^2 = 1,$$

og hvis vi sætter $Y = \frac{1}{N}\sum_{i=1}^{N} X_i$, så er

$$Y(P) \approx \mathcal{N}(0, \frac{1}{N}).$$

□

Det er dette resultat, der ligger bag approksimationen af binomialfordelinger med normalfordelinger i eksempel 1.19. Der findes mange udvidelser til andre situationer: med afhængige variable og/eller med variable der ikke er identisk fordelte. Der findes endvidere et væld af afledte resultater - f.eks. χ^2-approksimationen i eksempel 1.19. Der findes også præcise, kvantitative varianter. Her præsenterer vi en, der gør brug af afstandsbegrebet fra (1.9):

Sætning 1.22 (Berry-Esseens CLT) *Hvis $X_1, X_2, \ldots, X_N$ er uafhængige, identisk fordelte stokastiske variable med*

$$EX_1 = 0, \quad EX_1{}^2 = 1, \quad E|X_1|^3 < \infty$$

og hvis vi sætter $Y = \frac{1}{N}\sum_{i=1}^{N} X_i$, så er

$$d\left(Y(P),\ \mathcal{N}(0, \frac{1}{N})\right) \leq \frac{3\,E|X_1|^3}{\sqrt{N}}.$$

□

Disse approksimative resultater er utroligt vigtige for statistisk metodik, både historisk set og aktuelt. Historisk har et af de centrale problemer været hvordan man skulle klare sig med et begrænset antal tabeller over fordelinger til sin rådighed. Det kan være vanskeligt i dag at forstå **hvor** afgørende denne praktiske vanskelighed har været. Men det er klart at de asymptotiske metoder i høj grad har været svaret på problemet - vi så i eksempel 1.19 hvordan en enkelt tabel over normalfordelingen kunne erstatte tabeller over mange, mange binomialfordelinger. En bonusgevinst er at de fordelinger, der kan optræde som grænsefordelinger, ofte hører med til standardudstyret og de er derfor tabelleret i forvejen.

Det kan virke som om disse praktiske vanskeligheder er blevet uaktuelle med fremkomsten af computere. Men det er et optisk bedrag. I princippet kan man i dag finde enhver fordeling eksakt, hvis man virkelig vil, men der er ofte et betydeligt numerisk arbejde forbundet med at beskrive de eksakte fordelinger, et arbejde der ikke altid står mål med udbyttet af de mange, mange konkordansproblemer man vurderer i løbet af et videnskabeligt projekt. Her har man nærmere brug for et skud fra hoften, et skøn over konkordansen. Selv om moderne statistikpakker har et stort arsenal af indbyggede fordelinger, så har de ingenlunde alle de fordelinger man støder på i praksis. I den forstand har computerne kun givet os et større antal tabeller end vi havde før, men vi har stadig ikke nok... Derfor er vi i praksis stadig afhængige af de asymptotiske metoder.

Men måske endnu vigtigere end den praktiske betydning, er den rolle de asymptotiske metoder spiller for den teoretiske forståelse af statistiske teknikker. En lang række resultater bliver intuitivt klare hvis de diskuteres i grænsen $N \to \infty$, og beviser i matematisk statistik drejer sig ofte om at gøre rede for at de intuitive asymptotiske resultater i passende forstand er opfyldt for $N < \infty$.

I eksempel 1.19 modstod vi ikke fristelsen til at evaluere den faktisk gjorte observation mod den approksimerende fordeling, og på den måde udregnede vi approksimative p-værdier. Vi kunne tilsvarende have fundet approksimative konkordansområder på niveau α ved at finde et eksakt konkordansområde for den approksimerende fordeling.

Denne tilgang til det approksimative konkordansproblem skaber et nyt skel mellem en metodes nominelle og faktiske egenskaber. De nominelle egenskaber er de egenskaber metoderne ville have, hvis de brugte approksimationer var rigtige identiteter - de faktiske egenskaber er dem de i virkeligheden har. Hvis A er et konkordansområde på niveau α for den approksimerende fordeling ν, så har det nominelt niveau α for den

konkordansundersøgelse vi er i gang med - men det har faktisk niveau $1 - P(X \in A)$. Hvis approksimationen $X(P) \approx \nu$ er dårlig, kan der være betydelig forskel på det nominelle og det faktiske niveau. Tilsvarende kan der være store forskelle mellem den faktiske og den approksimative p-værdi for en konkret observation.

Hvis de asymptotiske metoder skal være nyttige, må man sørge for at disse forskelle ikke bliver for store. Man skal kun stole på de asymptotiske resultater hvis der faktisk er et stort antal gentagelser involveret i ens eksperiment. Hvis man i forbindelse med en CLT-approksimation finder en approksimativ p-værdi på 5%, og gerne vil være sikker på at den faktiske p-værdi er under 10%, så kræver det ifølge Berry-Esseens sætning adskillige hundreder observationer, lidt afhængigt af hvor tunge halerne er. Nu er formelle betingelser, der skal sikre gode approksimationer, som regel meget strengere end hvad der i praksis er nødvendigt. Men moralen er at det kræver en sikker hånd, ledt af erfaring, at bruge asymptotiske metoder uden at begå fejlslutninger.

1.8 Opgaver

OPGAVE 1.1. Lad X være t-fordelt med 3 frihedsgrader. Find et centralt konkordansområde for X, både på niveau 5% og på niveau 1%.

Lad $x = 3.4$. Udregn den centrale p-værdi for x i forhold til fordelingen af X. Kan vi betragte x som en realisation af den stokastiske variabel X?

OPGAVE 1.2. Lad X være Γ-fordelt med formparameter 5 og skalaparameter 3. Find et venstrestillet konkordansområde for X, både på niveau 5% og på niveau 1%.

Lad $x = 27.0$. Udregn den venstrestillede p-værdi for x i forhold til fordelingen af X. Kan vi betragte x som en realisation af den stokastiske variabel X?