Grundkurs i statistisk teori, del 2

Paul Blomstedt

2010
Reviderad 3.5.2012

Innehåll

1 Inledning

I sannolikhetslära studerar man slumpmässiga försök av olika slag. De möjliga utfallen av ett slumpmässigt försök utgör ett utfallsrum Ω. Vidare definieras en stokastisk variabel X som avbildar utfallen till de reella talen $\mathbb{R}$. Om man känner till fördelningen för den stokastiska variabeln X kan man dra slutsatser om hurudana värden av X man sannolikt *kommer att observera* om det slumpmässiga försöket på Ω upprepas ett antal gånger. I statistisk inferens (=slutledning) studerar man samma problem omvänt. Man antar att man *har observerat* ett antal realiserade värden av X men man man känner inte fullt till dess fördelning. På basen av observationerna försöker man sedan dra slutsatser om den okända fördelningen.

Ofta kan fördelningen för X uttryckas i form av en sannolikhets- eller täthetsfunktion $f(x)$ som beror av någon *parameter* θ (som kan vara en skalär eller en vektor). Vill man göra funktionens beroende av θ explicit kan man skriva $f(x;\theta)$. Exempel på parametriska fördelningar är bl.a. *exponentialfördelningen* med parametern λ

$$f(x;\lambda) = \lambda e^{\lambda x}, \quad x \geq 0, \lambda > 0,$$

och *normalfördelningen* med parametrarna μ och σ^2

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x, \mu \in \mathbb{R}, \sigma^2 > 0.$$

I de inferensproblem som behandlas under den här kursen antas att den okända fördelningen för X egentligen är känd så när som på värdet av parametern θ. De slutsatser man drar gäller då alltså det okända värdet av θ.

I statistiska sammanhang ges utfallsrummet Ω ofta tolkningen av en *population*. Varje enskilt utfall kan enligt denna tolkning ses som en indivd i populationen. Vidare beskriver parametern θ någon egenskap av populationen med avseende på fördelningen för X. Om t.ex. X är normalfördelad med parametrarna μ och σ^2 anger väntevärdesparametern μ medelvärdet av X bland hela populationen. De observerade värdena av X som används för att dra slutsatser om θ kallas ofta *stickprov*.

1.1 Inledande exempel

1.1.1 Statistisk modell

Betrakta en stokastisk variabel X med värdemängd $\Omega_X = \{0, 1\}$. Utför vi ett slumpmässigt försök på utfallsrummet Ω antar variabeln X värdet 1 med sannolikhet p eller värdet 0 med sannolikhet $1-p$, där $p \in [0, 1]$. Som funktion av värdet $x \in \Omega_X$ kan sannolikheten uttryckas som

$$f(x; p) = \mathbb{P}(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1. \tag{1}$$

Låt oss nu anta att vi utför n oberoende försök av ovannämnda slag. Vi kan då tänka oss att vi har en vektor av n oberoende och lika fördelade (*i.i.d.*) stokastiska variabler $(X_1, \ldots, X_n)$ som resulterar i en vektor av n observerade värden $(x_1, \ldots, x_n)$. Den simultana fördelningen för de n stokastiska variablerna ges av

$$\begin{aligned} f(x_1, \ldots, x_n; p) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\ &\overset{\perp\!\!\!\perp}{=} \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}. \end{aligned} \tag{2}$$

Om vi känner till värdet på parametern p kan vi med hjälp av den simultana sannolikhetsfunktionen ovan beräkna sannolikheten för varje tänkbart resultat. I en omvänd situation kan vi tänka oss att vi inte känner till värdet på p men har observerat $(x_1, \ldots, x_n)$. Vi kan då försöka välja ett värde för p så att den simultana fördelningen $f(x_1, \ldots, x_n; p)$ enligt något kriterium på bästa möjliga sätt motsvarar den observerade vektorn. Betecknar vi den mängd av värden som p kan anta, dvs. *parameterrummet*, Ω_p, har vi en familj $\{f(x_1, \ldots, x_n; p) : p \in \Omega_p\}$ av simultana fördelningar, av vilka en antas ha givit upphov till den observerade vektorn. En sådan familj av fördelningar kallas *statistisk modell*.

1.1.2 Vita och röda bollar

Vi ska nu ge modellen ovan en konkret tolkning. Vi tänker oss att vi har en stor urna med sammanlagt N bollar av vilka K är vita och $N - K$ är röda. Bollarna i urnan representerar här ett utfallsrum eller en population.

Parametern p motsvaras av det relativa antalet vita bollar i urnan, dvs. K/N. Vi drar n bollar ur urnan med återläggning och sätter

$$X_i = \begin{cases} 1, & i\text{:te dragna bollen vit} \\ 0, & i\text{:te dragna bollen röd.} \end{cases}$$

Anmärkning. 1.1. Eftersom de n försöken antas vara oberoende dras bollarna med återläggning (utan oberoende skulle uttrycket för den simultana sannolikheten i ekvation (2) bli mera komplicerad). Är N emellertid tillräckligt stort i förhållande till n kan försöken anses vara nära på oberoende även om bollarna dras utan återläggning.

Känner vi till p kan vi enligt ekvation (2) beräkna sannolikheten för att t.ex. observera resultatet $(0, 1, 1, 0, 1)$. Uppenbart är att detta resultat har samma sannolikhet som t.ex. $(1, 0, 1, 0, 1)$ eller $(1, 1, 1, 0, 0)$ eftersom alla tre följder innehåller 3 vita bollar av totalt 5. För att bestämma sannolikheten för ett visst resultat behöver vi alltså inte använda oss av resultatet som sådant utan det räcker med en "sammanfattning" av något slag, vilket ofta är mera praktiskt. Den simultana sannolikhetsfunktionen i ekvation (2) kan som en funktion av en sådan sammanfattning skrivas om enligt

$$f(x_1, \ldots, x_n; p) = \prod_{i=1}^{n} p_i^x (1-p)^{1-x_i} = p^{k(\mathbf{x}_n)}(1-p)^{n-k(\mathbf{x}_n)}, \tag{3}$$

där $\mathbf{x}_n = (x_1, \ldots, x_n)$ och $k(\mathbf{x}_n) = \sum_{i=1}^{n} x_i$ anger antalet försök som resulterat i värdet 1. Alternativt kan vi skriva

$$f(x_1, \ldots, x_n; p) = \prod_{i=1}^{n} p_i^x (1-p)^{1-x_i} = p^{n\overline{x}}(1-p)^{n-n\overline{x}}, \tag{4}$$

där $\overline{x} = \sum_{i=1}^{n} x_i/n$ anger den relativa frekvensen av försök som resulterat i värdet 1.

Anmärkning. 1.2. Allmänt kallas en funktion $t : \mathbb{R}^n \to \mathbb{R}$ som med avseende på någon viss egenskap sammanfattar ett resultat av ett upprepat slumpmässigt försök för *statistika*. Exempel på statistikor är förutom $\sum_{i=1}^{n} x_i$ och $\overline{x}$ bl.a. $\min(x_1, \ldots, x_n)$ och $\max(x_1, \ldots, x_n)$. En statistika som m.a.p. en viss modell innehåller all relevant information kallas *tillräcklig* (jfr. t.ex. $\overline{x}$ och $\min(x_1, \ldots, x_n)$ i exemplet med bollarna). Observera att $t(X_1, \ldots, X_n)$ är en stokastisk variabel medan $t(\mathbf{x}_n) = t(x_1, \ldots, x_n)$ är ett observerat värde av den stokastiska variabeln $t(X_1, \ldots, X_n)$.

1.1.3 Skattning av den okända parametern p

Vi betraktar fortfarande urnan med bollarna. Det enda vi nu vet om innehållet är att vi har ett stort antal bollar av vilka en del är vita och en del är röda. För att få mera information om innehållet drar vi slumpmässigt n bollar ur urnan, dvs. ett stickprov. Speciellt är vi intresserade av det relativa antalet vita bollar i urnan, dvs. parametern p. Vår intuition säger oss att en bra skattning för den okända parametern är det relativa antalet vita bollar i stickprovet, vilket ges av statistikan $\tilde{p}(x_1, \ldots, x_n) = \overline{x} = \sum_{i=1}^{n} x_i/n$. Beteckningen $\tilde{p}$ används för att understryka statistikan är en skattning av p.

Vi har tidigare talat om att en parameter beskriver någon egenskap av en population med avseende på fördelningen för en stokastisk variabel X, vilket är analogt med att en statistika beskriver någon egenskap av ett stickprov. Det är därför naturligt att tänka sig att en lämpligt vald statistika utgör en bra skattning för en okänd parameter. I följande kapitel ska vi lära oss om egenskaper som gör en statistika till en ”bra” skattning samt introducera en del metoder för att ta fram skattningar i olika situationer.

2 Estimationsteori

2.1 Estimatorer och deras egenskaper

Definition. 2.1. Det observerade värdet $\tilde{\theta}_0 = \tilde{\theta}(x_1, \ldots, x_n)$ på en statistika som tagits fram för att skatta värdet på en okänd parameter θ kallas *(punkt)skattning* eller *estimat*. Den motsvarande stokastiska variabeln $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n)$ kallas *estimator*.

Eftersom sammansättningen av ett erhållet stickprov $(x_1, \ldots, x_n)$ alltid är slumpmässigt är också en skattning $\tilde{\theta}_0$ beräknat utgående från stickprovet alltid slumpmässigt. Det är därför viktigt den funktion av stickprovet som skattningen grundar sig på i genomsnitt ger värden som är nära det sanna parametervärdet θ, oberoende av sammansättningen av det observerade stickprovet. För att avgöra om en skattning är ”bra” eller ”dålig” bör vi därför studera egenskaperna hos fördelningen för den stokastiska variabeln $\tilde{\theta}(X_1, \ldots, X_n)$. I idealfall kan denna fördelning bestämmas analytiskt men ibland är man tvungen att använda sig av simulering för att få en empirisk approximation av fördelningen. Simuleringsmetoder behandlas dock inte på den här kursen.

2.1.1 Väntevärdesriktighet

Definition. 2.2. En estimator $\tilde{\theta}$ är *väntevärdesriktig* om

$$\mathbb{E}(\tilde{\theta}) = \theta$$

för alla $\theta \in \Omega_\theta$.

Den intuitiva betydelsen av väntevärdesriktighet är att estimatorn $\tilde{\theta}$ i genomsnitt varken över- eller underskattar det sanna värdet av θ. En estimator som inte är väntevärdesriktig har m.a.o. ett systematiskt fel som definieras som

$$b(\tilde{\theta}) = \mathbb{E}(\tilde{\theta}) - \theta = \mathbb{E}(\tilde{\theta} - \theta),$$

Det systematiska felet $b(\tilde{\theta})$ kallas även *bias*. Förutom ett eventuellt systematiskt fel har varje skattning ett slumpmässigt fel. Det totala felet $(\tilde{\theta} - \theta)$ är en summa av det slumpmässiga och det sytematiska felet:

$$(\tilde{\theta} - \theta) = \underbrace{[\tilde{\theta} - \mathbb{E}(\tilde{\theta})]}_{\text{slumpmässigt fel}} + \underbrace{[\mathbb{E}(\tilde{\theta}) - \theta]}_{\text{systematiskt fel=bias}} .$$

Exempel. 2.1. Antag att var och en av de stokastiska variablerna $X_1, \ldots, X_n$ följer en fördelning med väntevärdet μ (vi kräver inte oberoende). Det aritmetiska medelvärdet $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ är en väntevärdesriktig estimator för μ eftersom

$$\mathbb{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Exempel. 2.2. Låt $X_1, \ldots, X_n$ vara *i.i.d.* med gemensam varians σ^2. Om väntevärdet μ är känt får vi att

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

är en väntevärdesriktig estimator för σ^2 eftersom

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i - \mu)^2 = \sigma^2.$$

Om μ däremot är okänt är

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

en väntevärdesriktig estimator för σ^2 (övningsuppgift).

Anmärkning. 2.1. Väntevärdesriktighet är ingalunda ett absolut krav för en bra estimator. Ett *litet* systematiskt fel kan väl accepteras om estimatorn i andra avseenden kan anses ha goda egenskaper. Ibland kan väntevärdesriktighet t.o.m. vara i konflikt med andra goda egenskaper. I alla situationer är det i själva verket inte ens möjligt att konstruera en väntevärdesriktig estimator.

2.1.2 Medelkvadratfelet

För att få ett mått på hur stort det totala felet $(\tilde{\theta} - \theta)$ i genomsnitt är skulle det kännas naturligt att beräkna väntevärdet av felet. För att positiva och negativa fel emellertid inte ska väga ut varandra tar man istället väntevärdet av det kvadrerade felet (utan kvadrering får vi det systematiska felet $b(\tilde{\theta})$). Detta väntevärde kallas *medelkvadratfelet*.

Definition. 2.3. Medelkvadratfelet för en estimator $\tilde{\theta}$ definieras som

$$\mathrm{MSE}(\tilde{\theta}) = \mathbb{E}\big[(\tilde{\theta} - \theta)^2\big].$$

En alternativ formulering för medelkvadratfelet ges av följande sats.

Sats. 2.1. *Medelkvadratfelet kan uttryckas som summan*

$$\mathrm{MSE}(\tilde{\theta}) = \mathrm{Var}(\tilde{\theta}) + \big(b(\tilde{\theta})\big)^2.$$

Bevis.

$$\begin{aligned}
\mathrm{MSE}(\tilde{\theta}) &= \mathbb{E}\big[(\tilde{\theta} - \theta)^2\big] \\
&= \mathbb{E}\Big[\big((\tilde{\theta} - \mathbb{E}(\tilde{\theta})) + (\mathbb{E}(\tilde{\theta}) - \theta)\big)^2\Big] \\
&= \mathbb{E}\big[(\tilde{\theta} - \mathbb{E}(\tilde{\theta}))^2 + 2(\tilde{\theta} - \mathbb{E}(\tilde{\theta}))(\mathbb{E}(\tilde{\theta}) - \theta) + (\mathbb{E}(\tilde{\theta}) - \theta)^2\big] \\
&= \mathbb{E}\big[(\tilde{\theta} - \mathbb{E}(\tilde{\theta}))^2\big] + (\mathbb{E}(\tilde{\theta}) - \theta)^2 \\
&= \mathrm{Var}(\tilde{\theta}) + \big(b(\tilde{\theta})\big)^2.
\end{aligned}$$

□

2.1.3 Effektivitet

En följd av satsen ovan är att medelkvadratfelet hos en väntevärdesriktig estimator reduceras till $\mathrm{Var}(\tilde{\theta})$. Detta ger oss ett enkelt sätt att jämföra två väntevärdesriktiga estimatorer.

Definition. 2.4. Om två estimatorer $\tilde{\theta}_1$ och $\tilde{\theta}_2$ är väntevärdesriktiga och

$$\mathrm{Var}(\tilde{\theta}_1) \leq \mathrm{Var}(\tilde{\theta}_2)$$

för alla $\theta \in \Omega_\theta$, med sträng olikhet för något $\theta \in \Omega_\theta$, är $\tilde{\theta}_1$ *effektivare* än $\tilde{\theta}_2$.

Ibland vill man vid jämförelser ange den *relativa effektiviteten* av $\tilde{\theta}_1$ i förhållande till $\tilde{\theta}_2$, vilket definieras som

$$\frac{\mathrm{Var}(\tilde{\theta}_2)}{\mathrm{Var}(\tilde{\theta}_1)}.$$

Exempel. 2.3. Låt $X_1, \ldots, X_5$ vara *i.i.d.* och låt $X_1 \sim \mathrm{Poisson}(\lambda)$. Summan $X_1 + \cdots + X_5$ följer då en $\mathrm{Poisson}(5\lambda)$-fördelning. Vi betraktar nu två väntevärdesriktiga estimatorer

$$\begin{aligned} \tilde{\theta}_1 &= \frac{1}{5}\sum_{i=1}^{5} X_i \quad \text{och} \\ \tilde{\theta}_2 &= X_1. \end{aligned}$$

Nu är

$$\mathrm{Var}\big(\tilde{\theta}_1\big) = \frac{1}{25}\sum_{i=1}^{5}\mathrm{Var}(X_i) = \frac{1}{25}5\lambda = \frac{\lambda}{5} < \mathrm{Var}\big(\tilde{\theta}_2\big) = \mathrm{Var}(X_2) = \lambda$$

alltså är $\tilde{\theta}_1$ är effektivare än $\tilde{\theta}_2$. Den relativa effektiviteten av $\tilde{\theta}_1$ i förhållande till $\tilde{\theta}_2$ är

$$\frac{\lambda}{\lambda/5} = 5.$$

2.1.4 Konsistens

Konsistens är en asymptotisk egenskap hos en estimator. Vi studerar alltså estimatorns beteende då vi låter stickprovsstorleken n gå mot oändlighet.

Definition. 2.5. En estimator $\tilde{\theta}$ är konsistent om

$$\lim_{n\to\infty} \mathbb{P}(|\tilde{\theta} - \theta| > \epsilon) = 0$$

för alla $\theta \in \Omega_\theta$ och $\epsilon > 0$.

Denna egenskap betecknar vi $\tilde{\theta} \xrightarrow{\mathbb{P}} \theta$, dvs. $\tilde{\theta}$ konvergerar i sannolikhet mot θ (jfr. den svaga formen av stora talens lag $\overline{X} = \sum_{i=1}^{n} X_i/n \xrightarrow{\mathbb{P}} \mu$). Innebörden av definitionen ovan är att ju större vårt stickprov är desto kraftigare är fördelningen för den stokastiska variabeln $\tilde{\theta}$ koncentrerad kring θ. Eftersom definitionen som sådan kan vara besvärlig att tillämpa i praktiken ges i följande sats ett tillräckligt vilkor för konsistens.

Sats. 2.2. *Estimatorn $\tilde{\theta}$ är konsistent om följande vilkor uppfylls:*

$$\begin{aligned} &a)\ \lim_{n\to\infty} \mathbb{E}(\tilde{\theta}) = \theta \\ &b)\ \lim_{n\to\infty} \mathrm{Var}(\tilde{\theta}) = 0. \end{aligned}$$

Bevis. I beviset använder vi oss av Markovs olikhet

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}.$$

Vi utgår ifrån medelkvadratfelet $\mathrm{MSE}(\tilde{\theta}) = \mathbb{E}\big[(\tilde{\theta}-\theta)^2\big] = \mathrm{Var}(\tilde{\theta}) + \big[\mathbb{E}(\tilde{\theta})-\theta\big]^2$. Enligt satsens antagande närmar sig $\mathrm{MSE}(\tilde{\theta})$ 0 då $n \to \infty$. Låt $\epsilon > 0$ och tillämpa Markovs olikhet på $(\tilde{\theta} - \theta)^2$. Vi får då att

$$\mathbb{P}(|\tilde{\theta} - \theta| > \epsilon) = \mathbb{P}\big((\tilde{\theta} - \theta)^2 > \epsilon^2\big) \leq \frac{\mathbb{E}\big[(\tilde{\theta} - \theta)^2\big]}{\epsilon^2} \underset{n\to\infty}{\to} 0.$$

□

Exempel. 2.4. De stokastiska variablerna $X_1, \ldots, X_n$ är *i.i.d.* med väntevärde μ och varians σ^2. Medelvärdet $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ är en väntevärdesriktig estimator för μ som därtill är konsistent eftersom

$$\begin{aligned} &\text{a)}\ \lim_{n\to\infty} \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu \\ &\text{b)}\ \lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0. \end{aligned}$$

Notera att eftersom $\overline{X}$ är väntevärdesriktig gäller a) automatiskt.

2.2 Metoder för konstruktion av estimatorer

Vi har ovan diskuterat egenskaper för en bra estimator men inte sagt något om *hur* man konstruerar en estimator. Det ska vi göra i följande.

2.2.1 Maximum-likelihood-metoden

Antag att vi har ett stickprov $(x_1, \ldots, x_n)$ som det realiserade värdet av en stokastisk vektor $(X_1, \ldots, X_n)$, med simultan sannolikhets- eller täthetsfunktion $f(x_1, \ldots, x_n; \theta)$ som beror av en parameter $\theta \in \Omega_\theta$. Om vi vidare antar att $X_1, \ldots, X_n$ är *i.i.d.* får vi att

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta), \tag{5}$$

där $f(x_i; \theta)$ är den marginella sannolikhets- eller täthetsfunktionen för X_i, $i = 1, \ldots, n$.

Exempel. 2.5. Vi återgår till det inledande exemplet i kapitel 1. Vi har alltså en en urna med K vita bollar och $N - K$ röda bollar där N antas vara mycket stort. Vi betecknar $p = K/N$, $p \in [0, 1]$, där p är okänt. Vi drar nu med återläggning 5 bollar ur urnan och låter dragningarna motsvaras av de oberoende stokastiska variablerna $X_1, \ldots, X_5$ så att

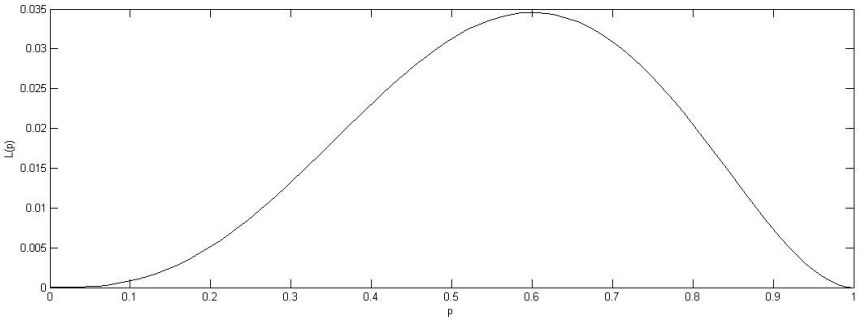
$$X_i = \begin{cases} 1, & i\text{:te dragna bollen vit} \\ 0, & i\text{:te dragna bollen röd} \end{cases}$$

för $i = 1, \ldots, 5$. För varje dragning har vi sannolikheterna $\mathbb{P}(X_i = 1) = p$ och $\mathbb{P}(X_i = 0) = 1 - p$. Variablerna X_i sägs då vara Bernoulli(p)-fördelade.

Antag nu att dragningarna resulterat i vektorn $(0, 1, 1, 0, 1)$. Vårt mål är att på basen av observationerna skatta värdet av parametern p, dvs. den relativa andelen vita bollar i urnan. Sannolikheten för den observerade vektorn är

$$\begin{aligned} \mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1) &\overset{\perp\!\!\!\perp}{=} (1-p)pp(1-p)p \\ &= p^3(1-p)^2. \end{aligned}$$

Vi frågar oss nu vilket värde för p som är det *mest trovärdiga* givet observationerna.



Figur 1: Likelihood-funktionen i exempel 2.6.

Definition. 2.6. Funktionen

$$L(\theta) = f(x_1, \ldots, x_n; \theta)$$

kallas *likelihood-funktion.*

Likelihood-funktionen (dvs. "trovärdighets"-funktionen) har samma funktionella form som den simultana sannolikhets- eller täthetsfunktionen. $L(\theta)$ uppfattas dock som en funktion av θ i och med att man fixerar $(x_1, \ldots, x_n)$ vid de observerade värdena och låter θ variera. För *i.i.d.* observationer antar likelihood-funktionen i enlighet med ekvation (5) formen

$$L(\theta) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Anmärkning. 2.2. I det diskreta fallet anger $L(\theta)$ den simultana sannolikheten för det observerade stickprovet för alla $\theta \in \Omega_\theta$. I det kontinuerliga fallet anger $L(\theta)$ den simultana tätheten för det observerade stickprovet för alla $\theta \in \Omega_\theta$.

Exempel. 2.6 (Fortsättning från föregående exempel.)**.** Vi granskar grafiskt likelihood-funktionen $L(p) = p^3(1-p)^2$, se figur 1. Vi ser att funktionen når sitt största värde vid $p = 0.6$. Analytiskt når vi samma resultat genom att derivera $L(p)$ med avseende på p och söka derivatans nollställe. Vi drar då slutsatsen att detta är det mest trovärdiga värdet för p givet observationerna.

Definition. 2.7. Det värde $\hat{\theta}_0 = \hat{\theta}(x_1, \ldots, x_n)$ som maximerar $L(\theta)$ för $\theta \in \Omega_\theta$ kallas *maximum-likelihood-skattningen*(eller *ML-skattningen*). Den motsvarande stokastiska variabeln $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ kallas *ML-estimator.*

Anmärkning. 2.3. I praktiken är det ofta lättare att maximera den logaritmerade likelihood-funktionen, dvs. *log-likelihood-funktionen*

$$l(\theta) = \log L(\theta).$$

Exempel. 2.7. Vi härleder i följande ML-estimatorn för parametern p i en Bernoulli(p)-fördelning. Sannolikhetsfunktionen för en Bernoulli-fördelad stokastisk variabel gavs i ekvation (1) i kapitel 1. Likelihood-funktionen är

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = p^k (1-p)^{n-k},$$

där $k = \sum_{i=1}^{n} x_i$, jfr. ekvation (3). Log-likelihood-funktionen blir

$$l(p) = \log L(p) = k \log p + (n-k) \log(1-p).$$

För att maximera $l(p)$ deriverar vi funktionen med avseende på p och sätter derivatan lika med noll

$$l'(p) = \frac{k}{p} - \frac{n-k}{1-p} = 0 \Leftrightarrow p = \frac{k}{n}.$$

Vi försäkrar oss ännu om att det erhållna extremvärdet är ett maximum genom att beräkna andra derivatan

$$l''(p) = -\frac{k}{p^2} - \frac{n-k}{1-p}^2$$

och sätta $p = k/n$. Vi får då att

$$l''(k/n) - \frac{n^2}{k} - \frac{n^2}{n-k} < 0$$

för alla $n > 0$ och $0 < k < n$, alltså har vi ett maximum. ML-estimatorn är slutligen $\hat{p} = \sum_{i=1}^{n} X_i/n$.

2.2.2 Minsta-kvadrat-metoden

Låt $x_1, \ldots, x_n$ vara observerade värden av de oberoende stokastiska variablerna $X_1, \ldots, X_n$. Vi antar att

$$\mathbb{E}(X_i) = \mu_i(\theta), \quad i = 1, \ldots, n,$$

där $\mu_1, \ldots, \mu_n$ är kända funktioner av en okänd parameter $\theta \in \Omega_\theta$. För varje X_i gäller alltså att

$$X_i = \mu_i(\theta) + \epsilon_i,$$

där ϵ_i är ett slumpmässigt fel med väntevärde 0. Vi definierar till följande funktionen $Q(\theta)$

$$Q(\theta) = \sum_{i=1}^{n} \big(x_i - \mu_i(\theta)\big)^2,$$

där $x_i - \mu_i(\theta)$ är felet hos den i:te observationen om θ vore det sanna parametervärdet.

Definition. 2.8. Det värde $\tilde{\theta}_0 = \tilde{\theta}(x_1, \ldots, x_n)$ som minimerar funktionen $Q(\theta)$ för $\theta \in \Omega_\theta$ kallas *minsta-kvadrat-skattningen* (förkortas *MK-skattningen*) för θ.

Anmärkning. 2.4. Om $\mu(\theta) = \mu_1(\theta) = \cdots = \mu_n(\theta)$ får vi att

$$\begin{aligned} Q'(\theta) &= -2\mu'(\theta) \sum_{i=1}^{n} (x_i - \mu(\theta)) \\ &= -2\mu'(\theta) n(\overline{x} - \mu(\theta)), \end{aligned}$$

av vilket följer att $Q(\theta)$ minimeras då $\mu(\theta) = \overline{x} = \sum_{i=1}^{n} x_i/n$. MK-skattningen blir då

$$\tilde{\theta}_0 = \mu^{-1}(\overline{x}),$$

dvs. man ”löser ut” θ från $\mu(\theta) = \overline{x}$.

Exempel. 2.8. Lystiderna för 5 glödlampor slumpmässigt plockade ur en produktionslinje registreras. Vi antar att lystiderna X_i för var och en av lamporna följer en Exp(λ)-fördelning med $\mathbb{E}(X_i) = 1/\lambda = \theta$ för $i = 1, \ldots, 5$. De registrerade lystiderna i timmar är

$$1011, 1102, 1012, 1010, 1015.$$

För att få MK-skattningen för väntevärdet θ minimerar vi

$$Q(\theta) = \sum_{i=1}^{5} (x_i - \theta)^2.$$

Funktionen minimeras då $\theta = \overline{x}$, alltså får vi att $\tilde{\theta}_0 = 1030$h.

Exempel. 2.9. Vi har mätningar $x_1, \ldots, x_n$ av längden av sidan på en kvadrat med arean θ, där $\theta > 0$ är den okända parametern. Dessa sidmätningar är utan systematiskt fel och ses som utfall av oberoende variabler $X_1, \ldots, X_n$ med $\mathbb{E}(X_i) = \sqrt{\theta}$ och $\mathrm{Var}(X_i) = \sigma^2$. Vi bildar

$$Q(\theta) = \sum_{i=1}^{n} (x_i - \sqrt{\theta})^2,$$

som enligt ovan minimeras av $\sqrt{\theta} = \overline{x}$, varifrån vi får att MK-skattningen är $\tilde{\theta}_0 = (\overline{x})^2$.

2.2.3 Momentmetoden

Antag att de stokastiska variablerna $X_1, \ldots, X_n$ är oberoende och lika fördelade. Det första *origomomentet* α_1 (dvs. väntevärdet) för X_i, $i = 1, \ldots, n$ definieras som

$$\alpha_1 = \mathbb{E}(X_i).$$

En väntevärdesriktig estimator för α_1 är som bekant $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Mera allmänt kan visas att en väntevärdesriktig estimator för det k:te origomomentet $\alpha_k = \mathbb{E}(X_i^k)$, ifall det existerar, är det k:te *stickprovsmomentet*

$$m_k(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Antag att fördelningen för X_i beror av en d-dimensionell parameter $\theta = (\theta_1, \ldots, \theta_d)$. Då beror även momentet α_k av θ, dvs.

$$\mathbb{E}(X_i^k) = \alpha_k(\theta_1, \ldots, \theta_d).$$

Exempel. 2.10. Betrakta en $\mathrm{N}(\mu, \sigma^2)$-fördelad stokastisk variabel X. Variabelns andra origomoment är då en funktion av parametrarna μ och σ^2

$$\mathbb{E}(X^2) = \alpha_2(\mu, \sigma^2) = \mu^2 + \sigma^2.$$

Det första steget i momentmetoden är att lägga upp ett ekvationssystem

$$\begin{cases} m_1 &= \alpha_1(\theta_1, \ldots, \theta_d) \\ \quad \vdots & \\ m_d &= \alpha_d(\theta_1, \ldots, \theta_d), \end{cases}$$

där $m_k = m_k(X_1, \ldots, X_n)$ och $k = 1, \ldots, d$. Ekvationssystemet består av lika många ekvationer som det finns parametrar att skatta. Ekvationssystemet löses sedan med avseende på $(\theta_1, \ldots, \theta_d)$.

Ofta går det att finna en entydig lösning $(\tilde{\theta}_1, \ldots, \tilde{\theta}_d)$ som beror av stickprovsmomenten $m_1, \ldots, m_d$

$$\begin{cases} \tilde{\theta}_1 &= \tilde{\theta}_1(m_1, \ldots, m_d) \\ \quad \vdots & \\ \tilde{\theta}_d &= \tilde{\theta}_d(m_1, \ldots, m_d). \end{cases}$$

Vi belyser metoden genom ett exempel.

Exempel. 2.11. Betrakta de *i.i.d.* stokastiska variablerna $X_1, \ldots, X_n$, vilka alla följer $\mathrm{N}(\mu, \sigma^2)$. Det två första origomomenten för en normalfördelad variabel är $\alpha_1 = \mu$ och $\alpha_2 = \mu^2 + \sigma^2$. Vi lägger nu upp ekvationssystemet

$$\begin{cases} m_1 = \frac{1}{n}\sum_{i=1}^n X_i = & \mu = \alpha_1 \\ m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2 = & \mu^2 + \sigma^2 = \alpha_2. \end{cases}$$

Genom att lösa systemet m.a.p. μ och σ^2 får vi

$$\begin{cases} \tilde{\mu} &= \frac{1}{n}\sum_{i=1}^n X_i & = \overline{X} \\ \tilde{\sigma^2} &= \underbrace{\frac{1}{n}\sum_{i=1}^n X_i^2}_{m_2(X_1,\ldots,X_n)} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2}_{[m_1(X_1,\ldots,X_n)]^2} & = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2. \end{cases}$$

2.3 Medelfelet för en skattning

Variansen $\mathrm{Var}(\tilde{\theta})$ och därmed även standardavvikelsen $D(\tilde{\theta})$ är ett naturligt mått på precisionen för en (väntevärdesriktig) estimator $\tilde{\theta}$. Ofta beror emellertid $D(\tilde{\theta})$ av den okända parameter vi försöker skatta. M.a.o. är $D(\tilde{\theta})$ själv en okänd parameter.

Exempel. 2.12. Vi singlar en (möjligtvis obalanserad) slant n gånger. Låt $\mathbb{P}(\text{klave}) = p$, där p är okänd, och låt X beteckna antalet klavor i n försök. Vi

har då att $X \sim \text{Bin}(n, p)$. För att skatta värdet av p använder vi estimatorn $\tilde{p} = X/n$. Nu är estimatorns standarddavvikelse

$$D(\tilde{p}) = \sqrt{\text{Var}\left(\frac{X}{n}\right)} = \sqrt{\frac{1}{n^2} np(1-p)} = \sqrt{p(1-p)/n}.$$

Uttrycket beror förutom av n även av den okända parametern p. För att få ett numeriskt värde för $D(\tilde{p})$ måste vi alltså ersätta p med skattningen $\tilde{p}_0 = x/n$.

Definition. 2.9. Det skattade värdet av $D(\tilde{\theta})$ kallas *medelfelet* för $\tilde{\theta}$ och betecknas $d(\tilde{\theta})$ (ofta ser man även beteckningen $s.e.(\tilde{\theta})$).

Exempel. 2.13. Vi betraktar ett stickprov $(x_1, \ldots, x_n)$ från $\text{N}(\mu, \sigma^2)$ med okänt väntevärde μ och okänd varians σ^2. Estimatorn $\tilde{\mu} = \overline{X}$ för väntevärdet μ har standardavvikelsen

$$D(\tilde{\mu}) = \frac{\sigma}{\sqrt{n}},$$

som beror av den okända parametern σ. Vi ersätter σ med skattningen

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

och får

$$d(\tilde{\mu}) = \frac{s}{\sqrt{n}}.$$

För stora n ger detta en rimlig uppfattning om precisionen av $\tilde{\mu} = \overline{X}$.

Tyvärr finns det ingen allmän formel för hur man ska beräkna medelfelet utan detta varierar från fall till fall. Om ingen enkel lösning finns tillhanda brukar man kunna använda sig av någon datorintensiv algoritm som t.ex. *bootstrap*. Sådana algoritmer kommer vi dock inte att behandla här. Medelfelet behövs ofta då man beräknar *intervallskattningar* eller, som det även kallas, *konfidensintervall*. Intervallskattning behandlas i föjande kapitel.

3 Intervallskattning

3.1 Allmänt

Betrakta ett slumpmässigt stickprov $\mathbf{x} = (x_1, \ldots, x_n)$ från en fördelning som är beroende av en okänd parameter θ. Stickprovet uppfattas som tidigare som det observerade värdet på en stokastisk vektor $\mathbf{X} = (X_1, \ldots, X_n)$.

Definition. 3.1. Ett intervall I_θ som täcker det sanna parametervärdet θ med sannolikhet $1-\alpha$ är ett *konfidensintervall* med *konfidensnivå* $1-\alpha$, där $\alpha \in (0,1)$. Intervallets vänstra och högra ändpunkter kallas *konfidensgränser* och betecknas $a_1(\mathbf{x})$ och $a_2(\mathbf{x})$.

Anmärkning. 3.1. Konfidensgränserna $a_1(\mathbf{x})$ och $a_2(\mathbf{x})$ är observerade värden av de stokastiska variablerna $a_1(\mathbf{X})$ och $a_2(\mathbf{X})$, för vilka gäller att

$$\mathbb{P}\big(a_1(\mathbf{X}) < \theta < a_2(\mathbf{X})\big) = 1 - \alpha.$$

Talet α väljs oftast så att $1-\alpha$ motsvarar 0.9, 0.95 eller 0.99. Man talar då ofta om 90%, 95% eller 99% konfidensintervall.

Anmärkning. 3.2. Konfidensintervall kan vara förutom *tvåsidiga*, dvs.

$$I_\theta = \big(a_1(\mathbf{x}),\ a_2(\mathbf{x})\big),$$

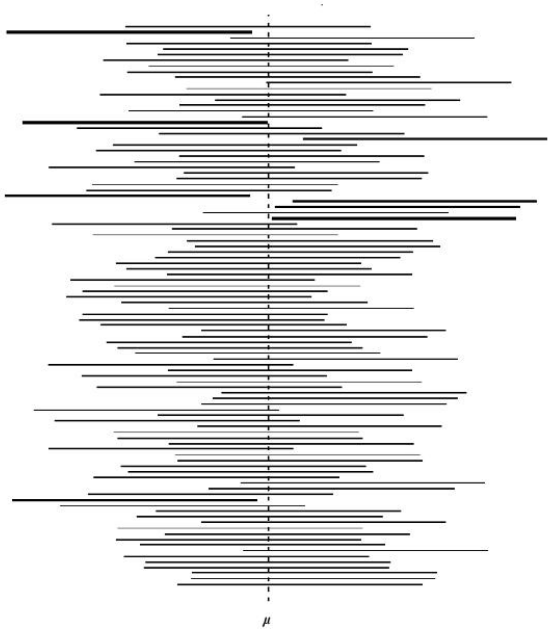
även *ensidiga*, dvs.

$$I_\theta = \big(a_1(\mathbf{x}),\ \infty\big) \ \text{ eller } \ I_\theta = \big(-\infty,\ a_2(\mathbf{x})\big).$$

Då är talen $a_1(\mathbf{x})$ resp. $a_2(\mathbf{x})$ observerade värden av de stokastiska variablerna $a_1(\mathbf{X})$ och $a_2(\mathbf{X})$ för vilka

$$\mathbb{P}\big(a_1(\mathbf{X}) < \theta\big) = 1 - \alpha \ \text{ och } \ \mathbb{P}\big(a_2(\mathbf{X}) > \theta\big) = 1 - \alpha.$$

Exempel. 3.1. Den intuitiva betydelsen av t.ex. ett 95% konfidensintervall är följande. Vi tänker oss att vi drar 100 slumpmässiga stickprov om n observationer var från en fördelning med väntevärdet μ. Utgående ifrån vart och ett av stickproven beräknar vi sedan ett 95% konfidensintervall för μ. Av samtliga intervall kommer nu i genomsnitt 95% att täcka det sanna parametervärdet μ, se figur 2. *I praktiken* brukar man ju dock endast ha tillgång till *ett* stikcprov om n observationer. Sannolikheten för att ett konfidensintervall beräknat utgående från *just detta* stickprov ska täcka μ är då 0.95.



Figur 2: Hundra tvåsidiga konfidensintervall beräknade för väntevädet μ.

Anmärkning. 3.3. Ju högre konfidensnivå man väljer för ett konfidensintervall, desto längre blir intervallet och desto högre är sannolikheten för att det ska täcka det sanna parametervärdet θ. Samtidigt förlorar man dock en del information om θ. Betrakta t.ex. ett 100% konfidensintervall $I_\theta = (-\infty,\ \infty)$ för parametern θ. Vi vet med säkerhet att intervallet I_θ täcker θ men vet fortfarande inget om det sanna parametervärdet.

I resten av detta kapitel ska vi bekanta oss med några vanliga fall av konfidensintervall. Vi börjar med det absolut valigaste fallet, dvs. beräkning av konfidensintervall för väntevärdesparametern μ i en normalfördelning.

3.2 Konfidensintervall för väntevärdet i en normalfördelning

Då vi beräknar ett konfidensintervall för väntevärdet μ i en normalfördelning skiljer vi på två fall: då variansen σ^2 är känd och då σ^2 är okänd. Vi börjar med det förra fallet.

3.2.1 Variansen känd

Låt $x_1, \ldots, x_n$ vara ett oberoende slumpmässigt stickprov från $\mathrm{N}(\mu, \sigma^2)$. Vi antar vidare att variansen σ^2 är känd. Vi ska i följande två exempel stegvis härleda oss till ett 95% konfidensintervall för μ.

Exempel. 3.2. Vi börjar med att anta att vårt stickprov endast består av *ett* observerat värde x av den stokastiska variabeln $X \sim \mathrm{N}(\mu, \sigma^2)$. Standardisering av X ger oss att

$$Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

Vi söker till näst ett sådant värde $z_{\alpha/2}$ av den standardiserade normalfördelningen att $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha = 0.95$. Då vi i detta fall har $\alpha = 0.05$ blir det sökta värdet $z_{0.05/2} = 1.96$. Eftersom nu

$$\begin{aligned}\mathbb{P}\left(-1.96 < \frac{X - \mu}{\sigma} < 1.96 \right) &= \mathbb{P}(-1.96\,\sigma < X - \mu < 1.96\,\sigma) \\ &= \mathbb{P}(X - 1.96\,\sigma < \mu < X + 1.96\,\sigma) \\ &= 0.95,\end{aligned}$$

ges ett 95% konfidensintervall för μ utgående från *en* observation av

$$I_\mu = (x - 1.96\,\sigma,\ x + 1.96\,\sigma).$$

Exempel. 3.3. Låt oss nu anta att vårt stickprov istället för endast en observation består av n observationer. Vi har tidigare använt $\overline{X}$ som estimator för μ. Vi vet vidare att om $X_1, \ldots, X_n$ är *i.i.d.* och $X_1 \sim \mathrm{N}(\mu, \sigma^2)$, så följer att $\overline{X} = \sum_{i=1}^{n} X_i/n \sim \mathrm{N}(\mu, \sigma^2/n)$, vilket efter standardisering ger

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

Ett 95% konfidensintervall utgående från n observationer kan nu härledas på motsvarande sätt som i föregående exempel, dvs.

$$\begin{aligned}\mathbb{P}\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96 \right) &= \mathbb{P}\left(-1.96\,\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96\,\frac{\sigma}{\sqrt{n}} \right) \\ &= \mathbb{P}\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\,\frac{\sigma}{\sqrt{n}} \right) \\ &= 0.95.\end{aligned}$$

Konfidensintervallet för μ blir alltså slutligen

$$I_\mu = \left(\overline{x} - 1.96 \, \frac{\sigma}{\sqrt{n}}, \; \overline{x} + 1.96 \, \frac{\sigma}{\sqrt{n}} \right).$$

Vi sammanfattar exemplet i följande sats.

Sats. 3.1. *Låt $x_1, \ldots, x_n$ vara ett slumpmässigt stickprov från $N(\mu, \sigma^2)$ med μ okänt och låt $z_{\alpha/2}$ vara ett sådant värde av $Z \sim N(0,1)$ att $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. Då är*

$$I_\mu = \left(\overline{x} - z_{\alpha/2} \, \frac{\sigma}{\sqrt{n}}, \; \overline{x} + z_{\alpha/2} \, \frac{\sigma}{\sqrt{n}} \right)$$

ett tvåsidigt konfidensintervall för μ med konfidensgraden $1 - \alpha$.

Anmärkning. 3.4. Ensidiga konfidensintervall konstrueras analogt men istället för $z_{\alpha/2}$ används ett sådant värde z_α att $\mathbb{P}(Z < z_\alpha) = 1 - \alpha$ eller $\mathbb{P}(Z > -z_\alpha) = 1 - \alpha$.

Exempel. 3.4. Betrakta följande stickprov

$$x_1 = 26.62, x_2 = 27.79, x_3 = 27.79, x_4 = 27.18$$

som antas vara observerade värden från $\mathrm{N}(\mu, 0.5^2)$ med μ okänt. En skattning för μ ges av medelvärdet $\overline{x} = 27.235$. Ett konfidensintervall för μ med konfidensnivån $1 - \alpha$ ges enligt sats 3.1 av

$$\begin{aligned} I_\mu &= (27.235 - 1.96 \cdot 0.5/\sqrt{4}, \; 27.235 + 1.96 \cdot 0.5/\sqrt{4}) \\ &= (26.745, 27.725). \end{aligned}$$

3.2.2 Variansen okänd

Vi håller antagandena i övrigt som ovan med den skillnaden att variansen σ^2 är okänd. Detta innebär att även standardavvikelsen för $\overline{X}$, dvs. $\sigma/\sqrt{n}$, är okänd. Variabeln $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ vi tidigare baserade härledningen av konfidensintervall på kan alltså inte tillämpas här. Vår intuition kanske nu säger oss att vi i variabeln Z borde ersätta standardavvikelsen $\sigma/\sqrt{n}$ med medelfelet $s/\sqrt{n}$ (se avsnitt 2.3). Eftersom medelfelet emellertid endast är en skattning av standardavvikelsen, gäller oftast att $s/\sqrt{n} \neq \sigma/\sqrt{n}$. Följaktligen

kommer den nya variabeln $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ inte att följa en standardiserad normalfördelning, med den vidare påföljden att

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2}\,\frac{s}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\,\frac{s}{\sqrt{n}}\right) \neq 1-\alpha.$$

För kunna beräkna ett konfidensintervall för μ då σ^2 är okänd borde vi alltså känna till fördelningen för $\frac{\overline{X}-\mu}{s/\sqrt{n}}$. Utan bevis ges följande sats.

Sats. 3.2. *Om $X_1, \ldots, X_n$ är i.i.d. med $X_1 \sim N(\mu, \sigma^2)$ och $\overline{X} = \sum_{i=1}^{n} X_i/n$, följer den stokastiska variabeln*

$$T = \frac{\overline{X}-\mu}{s/\sqrt{n}}$$

en t-fördelning med frihetsgraderna $f = n-1$.

Anmärkning. 3.5. Precis som den standardiserade normalfördelningen är t-fördelningen "klockformad" och centrerad vid 0 med den skillnaden att dess svansar är något tjockare. Fördelningen bestäms fullt av dess frihetsgrader f. Ju större stickprovsstorleken n är (och därmed ju större f är) desto närmare $\mathrm{N}(0,1)$ är t-fördelningen.

Låt nu T vara som i satsen ovan och låt $t_{\alpha/2}(f)$ beteckna ett sådant värde av T att $\mathbb{P}\big(-t_{\alpha/2}(f) < T < t_{\alpha/2}(f)\big) = 1-\alpha$. Ett konfidensintervall med konfidensnivå $1-\alpha$ för parametern μ av en normalfördelning med okänd varians kan då härledas på liknande sätt som tidigare, dvs.

$$\begin{aligned}
\mathbb{P}\Big(-t_{\alpha/2}(f) < \tfrac{\overline{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2}(f)\Big) &= \mathbb{P}\Big(-t_{\alpha/2}(f)\,\tfrac{s}{\sqrt{n}} < \overline{X}-\mu < t_{\alpha/2}(f)\,\tfrac{s}{\sqrt{n}}\Big) \\
&= \mathbb{P}\Big(\overline{X}-t_{\alpha/2}(f)\,\tfrac{s}{\sqrt{n}} < \mu < \overline{X}+t_{\alpha/2}(f)\,\tfrac{s}{\sqrt{n}}\Big) \\
&= 1-\alpha.
\end{aligned}$$

Resultatet sammanfattas i följande sats.

Sats. 3.3. *Låt $x_1, \ldots, x_n$ vara ett slumpmässigt stickprov från $N(\mu, \sigma^2)$ med både μ och σ^2 okända och låt $t_{\alpha/2}(f)$ vara ett sådant värde av en t-fördelning med frihetsgraderna $f = n-1$ att $\mathbb{P}(-t_{\alpha/2}(f) < T < t_{\alpha/2}(f)) = 1-\alpha$. Då är*

$$I_\mu = \left(\overline{x} - t_{\alpha/2}(f)\,\frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}(f)\,\frac{s}{\sqrt{n}}\right),$$

ett tvåsidigt konfidensintervall för μ med konfidensgraden $1-\alpha$.

Exempel. 3.5. Ett stickprov om 9 produkter från en produktionslinje vägs. Vikterna i gram är

$$149, 152, 143, 147, 155, 140, 153, 148, 145.$$

Vi antar att vikterna bland hela produktionen följer en normalfördelning med okänt väntevärde och okänd varians. För att konstruera ett 99% konfidensintervall för väntevärdet μ börjar vi med att beräkna stickprovsmedelvärdet $\overline{x} = 148.0$ och stickprovsstandardavvikelsen $s = \sqrt{190/8} \approx 4.87$. Med frihetsgraderna $f = 9 - 1 = 8$ och konfidensgraden $1 - \alpha = 1 - 0.01 = 0.99$ får vi att $t_{0.01/2}(8) = 3.36$. Enligt sats 3.3 blir då 99% konfidensintervallet för μ

$$\begin{aligned} I_\mu &= \left(148.0 - 3.36 \cdot \frac{4.87}{\sqrt{9}},\ 148.0 + 3.36 \cdot \frac{4.87}{\sqrt{9}}\right) \\ &= (142.5,\ 153.5). \end{aligned}$$

3.3 Konfidensintervall för differens mellan väntevärden

Låt $X_1, \ldots, X_{n_1}$ vara oberoende och $\mathrm{N}(\mu_1, \sigma_1^2)$-fördelade. Låt vidare $Y_1, \ldots, Y_{n_1}$ vara oberoende och $\mathrm{N}(\mu_2, \sigma_2^2)$-fördelade. Vi ska nu konstruera ett konfidensintervall för $\mu_1 - \mu_2$.

3.3.1 Varianserna kända

Vi börjar med att anta att σ_1^2 och σ_2^2 är kända. Från tidigare vet vi att

$$\overline{X} - \overline{Y} \sim \mathrm{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2),$$

av vilket följer att

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathrm{N}(0, 1).$$

Låt oss nu beteckna $D = D(\overline{X} - \overline{Y}) = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Ett konfidensintervall med konfidensnivå $1 - \alpha$ härleds då på bekant vis

$$\mathbb{P}\left(- z_{\alpha/2} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{D} < z_{\alpha/2} \right) =$$

$$\mathbb{P}\Big((\overline{X} - \overline{Y}) - z_{\alpha/2} \cdot D < \mu_1 - \mu_2 < (\overline{X} - \overline{Y}) + z_{\alpha/2} \cdot D \Big) = 1 - \alpha,$$

vilket ger oss att

$$I_{\mu_1-\mu_2} = \big((\overline{x}-\overline{y}) - z_{\alpha/2}\cdot D,\ (\overline{x}-\overline{y}) + z_{\alpha/2}\cdot D\big). \tag{6}$$

3.3.2 Varianserna lika men okända

Vi gör nu två ändringar i antagandena ovan:

1. Vi antar att $\sigma^2 = \sigma_1^2 = \sigma_2^2$.
2. Vi antar att σ^2 är okänd.

Från det första antagandet följer att

$$\overline{X} - \overline{Y} \sim \mathrm{N}\big(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)\big)$$

och efter standardisering

$$\frac{(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}} \sim \mathrm{N}(0,1).$$

Från det andra antagandet följer att $D(\overline{X}-\overline{Y}) = \sigma\sqrt{1/n_1+1/n_2}$ är okänd och måste ersättas med medelfelet $d(\overline{X}-\overline{Y})$. För att få ett uttryck för $d(\overline{X}-\overline{Y})$ börjar vi med att konstatera att i sådana fall där vi har ett stickprov var från $\mathrm{N}(\mu_1,\sigma^2)$ resp. $\mathrm{N}(\mu_2,\sigma^2)$, är en väntevärdesriktig skattning av σ^2

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)},$$

där $s_i^2,\ i=1,2$, betecknar variansen för de respektive stickproven. Medelfelet blir nu

$$d = d(\overline{X}-\overline{Y}) = s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

där

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}.$$

Konfidensintervallet

$$I_{\mu_1-\mu_2} = \big((\overline{x}-\overline{y}) - t_{\alpha/2}(f)\cdot d,\ (\overline{x}-\overline{y}) + t_{\alpha/2}(f)\cdot d\big) \tag{7}$$

med konfidensnivå $1-\alpha$ baseras på en t-fördelning med frihetsgraderna $f = (n_1-1)+(n_2-1) = n_1+n_2-2$ (jfr. konfidensintervallet i ekvation (6)).

Exempel. 3.6. Två tillverkare A och B tillverkar en produkt vars livslängd kan antas följa en normalfördelning. Den genomsnittliga lisvslängden för produkter tillverkade av A resp. B representeras då av μ_1 resp. μ_2. Varianserna antas vara lika men okända. Vi vill nu beräkna ett konfidensintervall för differensen $\mu_1 - \mu_2$. Vi drar slumpmässigt $n_1 = 80$ produkter tillverkade av A och får

$$\overline{x} = 1110 \text{ h} \quad s_1 = 105 \text{ h}$$

samt $n_2 = 100$ produkter tillverkade av B och får

$$\overline{y} = 1160 \text{ h} \quad s_2 = 130 \text{ h}.$$

Vidare är medelfelet

$$d = s\sqrt{\frac{1}{80} + \frac{1}{100}}, \quad s = \sqrt{\frac{79 \cdot 105^2 + 99 \cdot 130^2}{(80-1)+(100-1)}} \approx 119.6$$

och frihetsgraderna

$$f = 80 + 100 - 2 = 178.$$

Enligt ekvation (7) blir då t.ex. ett 99% konfidensintervall

$$\begin{aligned} I_{\mu_1-\mu_2} &= \left(-50 - t_{0.01/2}(178) \cdot 119.6\sqrt{\tfrac{1}{80}+\tfrac{1}{100}},\ -50 + t_{0.01/2}(178) \cdot 119.6\sqrt{\tfrac{1}{80}+\tfrac{1}{100}} \right) \\ &= (-50 - 2.60 \cdot 17.94,\ -50 + 2.60 \cdot 17.94) \\ &= (-96.644,\ -3.356). \end{aligned}$$

3.4 Konfidensintervall med normalapproximation

Då vi ovan för parametern θ skattade konfidensintervall av typen

$$I_\theta = \left(\hat{\theta}_0 - z_{\alpha/2} \cdot D(\hat{\theta}),\ \hat{\theta}_0 + z_{\alpha/2} \cdot D(\hat{\theta})\right)$$

eller

$$I_\theta = \left(\hat{\theta}_0 - t_{\alpha/2}(f) \cdot d(\hat{\theta}),\ \hat{\theta}_0 + t_{\alpha/2}(f) \cdot d(\hat{\theta})\right)$$

antog vi att estimatorn $\hat{\theta}$ är normalfördelad. I praktiken är ju så inte alltid fallet. I vissa fall kan emellertid ovan beskrivna metoder tillämpas approximativt om sticprovsstorleken n är tillräckligt stor.

Betrakta ett slumpmässigt stickprov från en fördelning som beror av en okänd parameter θ. Antag att vi har en estimator $\hat{\theta}$ som är *approximativt*

normalfördelad med väntevärde θ och varians $\mathrm{Var}(\hat{\theta})$. Standardavvilkelsen $D(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$ betecknar vi D. Den stokastiska variabeln $\frac{\hat{\theta}-\theta}{D}$ är då approximativt $\mathrm{N}(0,1)$-fördelad. Vi skiljer på två fall:

1. Om D inte beror av θ ger

$$I_\theta = \big(\hat{\theta}_0 - z_{\alpha/2} \cdot D(\hat{\theta}),\ \hat{\theta}_0 + z_{\alpha/2} \cdot D(\hat{\theta})\big)$$

   ett konfidensintervall för θ med approximativ konfidensnivå $1-\alpha$.

2. Om D beror av θ ger

$$I_\theta = \big(\hat{\theta}_0 - z_{\alpha/2} \cdot d(\hat{\theta}),\ \hat{\theta}_0 + z_{\alpha/2} \cdot d(\hat{\theta})\big),$$

   där $d(\hat{\theta})$ är ett lämpligt valt medelfel, ett konfidensintervall för θ med approximativ konfidensnivå $1-\alpha$.

Vi behandlar i följande ett exempel av det senare fallet.

3.4.1 Konfidensintervall för parametern p i en binomialfördelning

Låt $X \sim \mathrm{Bin}(n,p)$ där p är okänd. Vi har tidigare (t.ex. i exempel 2.12) använt oss av estimatorn $\hat{p} = X/n$ för att skatta p. Om nu n är tillräckligt stort kan man visa att X är approximativt $\mathrm{N}(np, np(1-p))$-fördelad. Likväl gäller då approximativt att

$$\hat{p} = \frac{X}{n} \sim \mathrm{N}\Big(p, \frac{p(1-p)}{n}\Big).$$

För att beräkna ett konfidensintervall för p behöver vi nu standardavvikelsen $D(\hat{p}) = \sqrt{p(1-p)/n}$. Eftersom den emellertid beror av p använder vi istället medelfelet $d(\hat{p}) = \sqrt{\hat{p}_0(1-\hat{p}_0)/n}$. Ett konfidensintervall för p med approximativ konfidensnivå $1-\alpha$ ges nu av

$$I_p = \big(\hat{p}_0 - z_{\alpha/2} \cdot d(\hat{p}),\ \hat{p}_0 + z_{\alpha/2} \cdot d(\hat{p})\big). \tag{8}$$

Exempel. 3.7. I en intervjuundersökning med $n = 250$ personer slumpmässigt valda ur en mycket stor population, visade det sig att 42 personer hade en viss åsikt H. En punktskattning för proportionen p av hela populationen med åsikt H ges av $\hat{p}_0 = 42/250 = 0.168$. Medelfelet för estimatorn $\hat{p}$ är nu

$d = \sqrt{0.168 \cdot 0.832/250} = 0.023$. Ett approximativt 95% konfidensintervall är då enligt ekvation (8)

$$\begin{aligned} I_p &= (0.168 - 1.96 \cdot 0.023,\ 0.168 - 1.96 \cdot 0.023) \\ &= (0.12,\ 0.21). \end{aligned}$$

Anmärkning. 3.6. Konfidensintervallets längd beror förutom av konfidensnivån $1 - \alpha$ även av stickprovsstorleken n. Ibland önskar man på förhand bestämma en maximal längd för ett intervall och sedan beräkna hur stort stickprovet ska vara för att denna maximala längd inte ska överskridas.

Exempel. 3.8 (Fortsättning från föregående exempel.)**.** Man önskar planera undersökningen så att man efteråt får ett ungefär 95% konfidensintervall för p med högst längden

$$I_p = (0.168 - 0.05,\ 0.168 - 0.05).$$

Man får då att

$$1.96\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}} \leq 0.05 \Leftrightarrow \left(\frac{1.96}{0.05}\right)^2 \hat{p}_0(1-\hat{p}_0) \leq n.$$

Så länge inga observationer har gjorts är $\hat{p}_0(1-\hat{p}_0)$ okänd. Alltid gäller dock att $\hat{p}_0(1-\hat{p}_0) \leq 1/4$ vilket ger att

$$\left(\frac{1.96}{0.05}\right)^2 \frac{1}{4} = 384.2 \leq n.$$

Slutsatsen är att minst 385 personer ska intervjuas för undersökningen.

4 Hypotesprövning

Exempel. 4.1. Vi är intresserade av en variabel X om vilken vi kan anta att den är (approximativt) normalfördelad i en given population. Variansen antas vara $\sigma^2 = 15^2$ och på basen av tidigare erfarenheter är man benägen att tro att väntevärdet μ ligger kring 100. Det finns dock en misstanke om att μ kan ha ökat. För att undersöka saken drar man ett stickprov om $n = 25$ enheter och räknar medelvärdet på observationerna. Resultatet blir $\overline{x} = 118.7$, vilket ju är betydligt större än 100. Frågan lyder nu om detta kan ses som ett bevis på att det faktiskt har skett en förändring i populationen?

Vår modell för populationen är nu alltså

$$X \sim \mathrm{N}(\mu, 15^2).$$

Då $n = 25$ får vi vidare att

$$\overline{X} \sim \ \mathrm{N}\big(\mu, 15^2/25\big).$$

Om ingen ingen ökning i μ skulle ha skett, skulle alltså $\mu = 100$ och

$$\overline{X} \sim \mathrm{N}(100, 15^2/25).$$

För att utvärdera hur trovärdig observationen $\overline{x} = 118.7$ är under detta antagande beräknar vi

$$\mathbb{P}\big(\overline{X} > 115\big) = 1 - \Phi\Big(\frac{115 - 100}{15/\sqrt{25}}\Big) = 0.2.87 \cdot 10^{-7},$$

där 115 är godtyckligt tal som är mindre än 118.7 Den erhållna sannolikheten är extremt liten. Vi har då två alternativa förklaringar till det observerade värdet:

1. Vi har av en slump observerat något ytterst osannolikt.
2. Vårt antagandde om att $\mu = 100$ är felaktigt.

Även om den första förklaringen är möjlig (också osannolika händelser kan ibland inträffa!) förefaller den senare förklaringen betydligt rimligare.

4.1 Hypoteser

Definition. 4.1. En *hypotes* är ett påstående gällande en fördelning. Mera specifikt gäller påståendet ofta värdet på en okänd parameter θ som fördelningen beror av.

Nollhypotesen H_0 är ett påstående vars riktighet man på basen av ett stickprov strävar efter att utvärdera. Nollhypotesen är något håller fast vid tills man blivit övertygad om att den är felaktig. Oftast formuleras även en *mothypotes* (eller *alternativ hypotes*) H_1 som ett alternativt påstående man

har större benägenhet att tro på ifall nollhypotesen verkar orimlig. Låt oss nu som tidigare beteckna parametererrummet Ω_θ. Hypoteserna

$$H_0 : \theta \in \Omega_{\theta_0} \subset \Omega_\theta$$
$$H_1 : \theta \in \Omega_{\theta_1} \subset \Omega_\theta$$

formuleras så att de är oförenliga. M.a.o. gäller att $\Omega_{\theta_0} \cap \Omega_{\theta_1} = \emptyset$ och ofta även att $\Omega_{\theta_0} \cup \Omega_{\theta_1} = \Omega_\theta$.

Definition. 4.2. Om H_0 formuleras så att

$$H_0 : \theta = \theta_0 \in \Omega_\theta,$$

dvs. att den omfattar ett enda värde av parameterrummet, kallas hypotesen *enkel*. I annat fall är den *sammansatt*.

I fortsättningen kommer vi att begränsa oss till enkla nollhypoteser.

4.2 Signifikanstest

För att testa riktigheten av en nollhypotes drar man ett stickprov $\mathbf{x} = (x_1, \ldots, x_n)$ från den fördelning hypotesen gäller. Som tidigare är stickprovet ett slumpmässigt utfall av den stokastiska vektorn $\mathbf{X} = (X_1, \ldots, X_n)$. Vidare formulerar man en *testvariabel* (eller teststatistika) $t(\mathbf{X})$, vars fördelning bestäms under antagandet att H_0 är korrekt. Till sist definierar man ett *kritiskt område* C, vilket består av alla de värden av $t(\mathbf{X})$ för vilka H_0 ter sig orimlig.

Vi formulerar nu följande *signifikanstest*:

$$\begin{aligned} \text{Om } t(\mathbf{x}) \in C \quad & \text{förkasta } H_0 \\ t(\mathbf{x}) \notin C \quad & \text{förkasta ej } H_0, \end{aligned}$$

där $t(\mathbf{x})$ är det observerade värdet av $t(\mathbf{X})$.

Anmärkning. 4.1. Att H_0 inte kan förkastas är inget bevis för att den är sann. Man ska därför helst inte tala om att man "godkänner H_0" (då den inte förkastas).

Exempel. 4.2. Man ser ett djur och ställer upp hypotesen H_0: Djuret är en häst. Som testvariabel tas $t(\mathbf{X})$ ="antalet ben" och som kritiskt område $C = \{0, 1, 2, 3, 5, 6, \ldots\}$. Testet lyder nu

$$\begin{aligned} \text{om } t(\mathbf{x}) < 4 \text{ eller } t(\mathbf{x}) > 4 \quad & \text{förkasta } H_0 \\ t(\mathbf{x}) = 4 \quad & \text{förkasta ej } H_0. \end{aligned}$$

Även om vi inte kan förkasta H_0 utgör detta inget bevis för att djuret vi observerat är en häst.

Anmärkning. 4.2. Det kritiska området C består ofta av intervall av typen $C = (-\infty, a)$ eller $C = (b, \infty)$, där a och b är fixa tal. Om speciellt C består av ett enda sådant intervall, sägs signifikanstestet vara *ensidigt*. Om C däremot består av både $(-\infty, a)$ och (b, ∞) där $a < b$, sägs testet vara *tvåsidigt*. Huruvida ett test är en- eller tvåsidigt bestäms av mothypotesen H_1.

Definition. 4.3. Sannolikheten

$$\mathbb{P}_{H_0}(t(\mathbf{X}) \in C) = \alpha$$

kallas för testets *signifikansnivå*. Ett värde $t(\mathbf{x})$ som faller inom det kritiska området sägs vara *signifikant* på nivån α.

I definitionen ovan har beteckningen $\mathbb{P}_{H_0}$ har använts för att understryka att sannolikheten beräknas under antagandet att H_0 är sann.

Anmärkning. 4.3. Ofta väljer man det kritiska området så att signifikansnivån α är 0.05, 0.01 eller 0.001.

Exempel. 4.3. Om vi i exempel 4.1 väljer att utföra ett ensidigt test av nollhypotesen $H_0 : \mu = 100$ mot den alternativa hypotesen $H_1 : \mu > 100$ på signifikansnivån $\alpha = 0.05$ utgörs det kriktiska området för variabeln $t(\mathbf{X}) = \overline{X}$ av intervallet $(104.93, \infty)$. Skulle vi däremot som alternativ hypotes ha $H_1 : \mu \neq 100$, dvs. ett tvåsidigt test, skulle det kritiska området utgöras av både $(-\infty, 94.12)$ och $(105.88, \infty)$.

Definition. 4.4. Den observerade signifikansnivån

$$\mathbb{P}_{H_0}(t(\mathbf{X}) \geq t(\mathbf{x})) = p$$

kallas för testets *p-värde*.

Anmärkning. 4.4. Definitionen ovan gäller ett ensidigt test av hypoteserna

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0.$$

För ett ensidigt test med mothypotesen $H_1 : \theta < \theta_0$ är p-värdet

$$\mathbb{P}_{H_0}(t(\mathbf{X}) \leq t(\mathbf{x})) = p.$$

Låt oss nu anta att $t(\mathbf{X})$ har en symmetrisk fördelning centrerad kring 0. Ett tvåsidigt test med mothypotesen $H_1 : \theta \neq \theta_0$ har då p-värdet

$$2 \cdot \mathbb{P}_{H_0}(t(\mathbf{X}) \geq |t(\mathbf{x})|) = p.$$

Exempel. 4.4. I exempel 4.1 är p-värdet för ett ensidigt test av hypoteserna $H_0 : \mu = 100$ och $H_1 : \mu > 100$

$$\mathbb{P}\big(\overline{X} \geq 118.7\big) = 2.28 \cdot 10^{-10}.$$

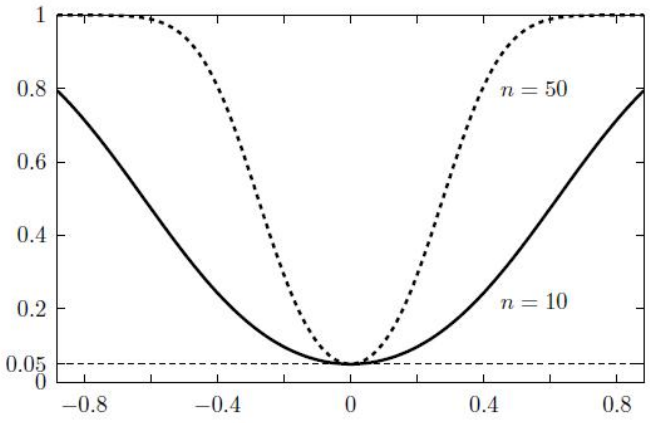
H_0 förkastas om p-värdet är "tillräckligt litet". Vad som är tillräckligt litet varierar från fall till fall men vanligt är att man jämför det erhållna värdet med en förutbestämd signifikansnivå α. Är t.ex. $p < 0.05$ förkastas H_0 på signifikasnivån $\alpha = 0.05$. M.a.o. är testresultatet signifikant på nivån 0.05.

4.3 Styrkefunktionen

Testets signifikansnivå α kan tolkas som sannolikheten att förkasta H_0 då den är sann. Man talar även om sannolikheten för ett *typ I -fel* (eller *fel av första slaget*). I hypotesprövning kan man också begå ett annat slags fel, nämligen att låta bli att förkasta H_0 då den är osann. Sannolikheten för detta fel, ett *typ II -fel* (eller *fel av andra slaget*), betecknas β. De två felen hänger ihop så att om man minskar sannolikheten för det ena kommer sannolikheten för det andra att öka. Genom att öka på stickprovsstorleken kan man dock minska på sannolikheten för båda felen samtidigt. Följande tabell ger en sammanfattning av felen:

Verklighet	Beslut	
	H_0 **förkastas**	H_0 **förkastas ej**
H_0 **sann**	Typ I -fel (α-fel)	Korrekt
H_1 **sann**	Korrekt	Typ II -fel (β-fel)

Sannolikheten för att få ett värde på testvariabeln som faller inom det kritiska området, m.a.o. att H_0 förkastas, beror uppenbarligen på vilket värde den aktuella parametern θ faktiskt har.



Figur 3: Styrkefunktionen för ett tvåsidigt test för $n = 10$ resp. $n = 50$.

Definition. 4.5. *Styrkefunktionen* är en funktion som för varje möjliga sanna värde av θ anger sannolikheten för att H_0 förkastas. Funktionen definieras som

$$\pi_\alpha(\theta) = \begin{cases} \alpha, & \theta \in \Omega_{\theta_0} \ (H_0 \text{ sann}) \\ 1-\beta, & \theta \notin \Omega_{\theta_0} \ (H_0 \text{ osann}). \end{cases}$$

För ett väl konstruerat test antar $\pi_\alpha(\theta)$ värden som är nära 1 då H_0 är osann och värden nära 0 då H_0 är sann.

Exempel. 4.5. Figur 3 visar grafen av styckefuntionen för ett tvåsidigt test av hypotesen $H_0 : \mu = 0$ mot $H_1 : \mu \neq 0$ på signifikansnivån $\alpha = 0.05$ då ett stickprov om $n = 10$ resp. $n = 50$ observationer dragits från $\mathrm{N}(\mu, 1)$.

Anmärkning. 4.5. I praktiken är man oftast intresserad av den del av funktionen för vilken H_0 är osann. (Speciellt om H_0 är enkel kan ju H_0 vara sann endast i en punkt.) Informellt anger alltså testets styrka $1 - \beta$ dess förmåga att förkasta en felaktig nollhypotes.

4.4 Test gällande väntevärdet i en normalfördelning

4.4.1 Test av $\mu = \mu_0$ då σ^2 är känd

Låt $x_1, \ldots, x_n$ vara ett stickprov från $\mathrm{N}(\mu, \sigma^2)$ med σ^2 känd. Vi vill testa den enkla hypotesen

$$H_0 : \mu = \mu_0.$$

En lämplig testvariabel är då $\overline{X} \sim \mathrm{N}(\mu_0, \sigma^2)$ som efter standardisering antar formen

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1).$$

I ett *tvåsidigt* test, dvs. ett test med mothypotes

$$H_1 : \mu \neq \mu_0$$

förkastas H_0 om $|z| > z_{\alpha/2}$. *Kritiska värden* för en del vanligt använda signifiknsnivåer är:

α	0.05	0.01	0.001
$z_{\alpha/2}$	1.96	2.54	3.29

Exempel. 4.6. Antag att en stokastisk variabel X följer en $\mathrm{N}(\mu, 2^2)$-fördelning. Man har orsak att tro att $\mu = 0$. För att utvärdera detta antagande drar man ett stickprov om $n = 10$ observationer. Stickprovsmedeltalet är $\overline{x} = 0.99$. Vi utför ett tvåsidigt test på signifikansnivån $\alpha = 0.05$ av hypoteserna

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0.$$

Testvariabeln antar värdet

$$|z| = \left|\frac{0.99 - 0}{2/\sqrt{10}}\right| \approx 1.57 < 1.96$$

alltså låter vi bli att förkasta H_0.

I ett *ensidigt* test, dvs. ett test med mothypotes

$$H_1 : \mu > \mu_0 \quad (\mu < \mu_0)$$

förkastas H_0 om $z > z_\alpha$ ($z < -z_\alpha$). Kritiska värden för en del vanligt använda signifiknsnivåer är:

α	0.05	0.01	0.001
z_α	1.64	2.33	3.09

Exempel. 4.7. Herr M påstår sig ha utvecklat en tillverkningsprocess som minskar den genomsnittliga tillverkningstiden av en viss produkt. För att utverdera hans påstående tillverkas 50 enheter med den nya metoden, vilket ger en genomsnittlig tillverkningstid på $\overline{x} = 38$ min 5 s. Testa om den nya metoden är signifikant snabare än den nya då man vet att den genomsnittliga tillverkninstiden tidigare har varit 40 min 15 s med en standardavvikelse på 1 min 30 s.

Vi formulerar följande hypoteser för ett ensidigt test på signifikansnivå $\alpha = 0.01$

$$H_0 : \mu = 40 \text{ min } 15 \text{ s}$$
$$H_1 : \mu < 40 \text{ min } 15 \text{ s}$$

Vi beräknar sedan värdet av testvariabeln

$$z = \frac{38.03 - 40.25}{1.5/\sqrt{50}} \approx -10.2 < -2.33.$$

Vi förkastar H_0 och drar slutsatsen att den nya metoden är snabbare.

4.4.2 Test av $\mu = \mu_0$ då σ^2 är okänd

Låt $x_1, \ldots, x_n$ vara ett stickprov från $\mathrm{N}(\mu, \sigma^2)$ med σ^2 okänd. Vi vill testa hypotesen

$$H_0 : \mu = \mu_0$$

En lämplig testvariabel är då

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

som följer en t-fördelning med frihetsgraderna $n - 1$. I ett *tvåsidigt* test förkastas H_0 om $|t| > t_{\alpha/2}(n-1)$. I ett *ensidigt* test förkastas H_0 om $t > t_\alpha(n-1)$ $(t < -t_\alpha(n-1))$.

Anmärkning. 4.6. Test som baserar sig på en t-fördelning brukar kallas för t-test.

Exempel. 4.8. En biltillverkare hävdar den genomsnittliga bränslekonsumptionen för en viss biltyp är 6.8 l/100 km. För att testa påståendet provkördes 10 bilar. För dessa bilar blev den genomsnittliga konsumptionen 7.0 l/100 km

och standardavvikelsen 0.2 l/100 km. Under antagandet att bränsleförbrukningen är nära på normalfördelad, använd ett ensidigt test på signifikansnivå $\alpha = 0.05$ för att testa hypoteserna

$$H_0 : \mu = 6.8 \text{ l/100 km}$$
$$H_1 : \mu > 6.8 \text{ l/100 km}$$

För ett ensidigt test på signifikansnivå $\alpha = 0.05$ baserat på en t-fördelning med $10 - 1 = 9$ frihetsgrader är det kritiska värdet $t_{0.05} = 1.83$. Då nu

$$t = \frac{7.0 - 6.8}{0.2/\sqrt{10}} \approx 3.16 > 1.83$$

förkastas H_0.

4.4.3 Test av $\mu_1 = \mu_2$ då $\sigma^2 = \sigma_1^2 = \sigma_2^2$ är okänd

Låt $x_1, \ldots, x_{n_1}$ vara ett stickprov från $\mathrm{N}(\mu_1, \sigma^2)$ och låt $y_1, \ldots, y_{n_2}$ vara ett stickprov från $\mathrm{N}(\mu_2, \sigma^2)$. Variansen σ^2 antas vara okänd. Vi vill nu testa hypotesen

$$H_0 : \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0.$$

En lämplig testvariabel är då

$$T = \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

där

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Variabeln T följer en t-fördelning med frihetsgraderna $n_1 + n_2 - 2$. I ett tvåsidigt test med mothypotes

$$H_1 : \mu_1 - \mu_2 \neq 0$$

förkastas H_0 om $|t| > t_{\alpha/2}(n_1 + n_2 - 2)$. På motsvarande förkastas H_0 i ett ensidigt test om $t > t_\alpha(n_1 + n_2 - 2)$ ($t < -t_\alpha(n_1 + n_2 - 2)$).

4.5 Test med normalapproximation

4.5.1 Binomialfördelningen: test av $p = p_0$

Låt $X \sim \text{Bin}(n, p)$, där p är okänt. Vi vill testa hypotesen

$$H_0 : p = p_0.$$

Då n är tillräckligt stort är en lämplig testvariabel

$$Z = \frac{\frac{X}{n} - p_0}{\sqrt{p_0(1-p_0)/n}},$$

som approximativt följer en standardiserad normalfördelning (jfr. avsnitt 3.4.1). De kritiska värdena är som i avsnitt 4.4.1.

Exempel. 4.9. En reklambyrå hävdar att 60% av alla personer som sett reklamen för en viss ny produkt kommer ihåg namnet på produkten 20 h efter att de sett reklamen för första gången. I en intervjuundersöking visade det sig att 156 av 300 intervjuade personer kom ihåg namnet. Vi utför ett ensidigt test på signifikansnivå $\alpha = 0.01$ för att utvädera om påståendet verkar rimligt. Hypoteserna är

$$H_0 : p = 0.60$$
$$H_1 : p < 0.60.$$

Eftersom

$$z = \frac{\frac{156}{300} - 0.60}{\sqrt{0.60 \cdot 0.40/300}} \approx -2.83 < -2.33$$

förkastas H_0.

4.6 Samband mellan hypotesprövning och intervallskattning

Det finns ett nära samband mellan ett signifikanstest på signifikansnivå α gällande en parameter θ och ett konfidensintervall för samma parameter på konfidensnivå $1 - \alpha$. Vi granskar sambandet genom följande exempel.

Exempel. 4.10. Antag att vi på basen av ett stickprov från $\mathrm{N}(\mu, \sigma^2)$ vill pröva hypotesen $H_0 : \mu = \mu_0$ mot $H_1 : \mu \neq \mu_0$ med ett tvåsidigt test på signifikansnivå α. H_0 förkastas nu om

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \quad \text{eller} \quad \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2},$$

vilket är identiskt med att H_0 förkastas om

$$\overline{x} - z_{\alpha/2} \cdot \sigma/\sqrt{n} > \mu_0 \quad \text{eller} \quad \overline{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu_0.$$

Vi känner nu igen den vänstra sidan av bägge olikheter som den undre respektive övre gränsen av ett konfidensintervall för μ på konfidensnivån $1 - \alpha$

$$I_\mu = (\overline{x} - z_{\alpha/2} \cdot \sigma/\sqrt{n}\ ,\ \overline{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n}).$$

Ett beslut gällande nollhypotesen kan alltså baseras på ett konfidensintervall: H_0 förkastas på signifikansnivån α om intervallet I_μ på motsvarande konfidensnivå $1 - \alpha$ inte täcker μ_0.

4.7 χ^2-test

χ^2-testen är en samling av test som baserar sig på en χ^2-fördelad testvariabel (se definitionen nedan). Dessa test används i situationer där man har att göra med klassindelade observationer: man registrerar vilken av k klasser en observation tillhör men behöver inte nödvändigtvis kunna ge observationerna själva något numeriskt värde. I sådana fall är man intresserad av den diskreta fördelning som *klassfrekvenserna*, dvs. antalet observationer i varje klass, bildar. I detta avsnitt ska vi ta upp två typer av χ^2-test: test av given fördelning samt oberoendetest.

Definition. 4.6. Låt $Z_1, \ldots, Z_n$ vara oberoende och låt $Z_i \sim \mathrm{N}(0, 1)$ för $i = 1, \ldots, n$. Variabeln

$$\chi^2(f) = \sum_{i=1}^{n} Z_i^2$$

följer då en χ^2-*fördelning* med $f = n$ frihetsgrader.

Anmärkning. 4.7. Man kan visa att om $X_1, \ldots, X_n$ är oberoende $\mathrm{N}(\mu, \sigma^2)$-fördelade stokastiska variabler, följer variabeln

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

en χ^2-fördelning med frihetsgraderna $f = n - 1$. Detta används t.ex. vid konstruktion av konfidensintervall för σ^2. Med liknande argument motiveras även bruket av en F-fördelad testvariabel i bl.a. variansanalys, se definition 5.1 och avsnitt 5.2.1.

4.7.1 Test av given fördelning

Låt oss anta att vi för k klasser har observerat en vektor av frekvenser $x_1, \ldots, x_k$. Vi vill nu testa hur väl den observerade fördelningen stämmer överens med en viss *förväntad fördelning* av klassfrekvenser. Låt $p_i,\ i = 1, \ldots, k$ beteckna sannolikheten i den förväntade fördelningen för att en observation ska tillhöra den i:te klassen och låt n beteckna det totala antalet observationer. Den förväntade frekvensen för klass i blir då alltså np_i. Hypoteserna är nu

H_0 : Den observerade och förväntade fördelningen lika.

H_1 : Den observerade och förväntade fördelningen olika.

Värdet på testvariabeln beräknas enligt

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}.$$

Man kan visa att då H_0 är korrekt följer testvariabeln approximativt en χ^2-fördelning med frihetsgraderna $f = k - 1$. Om den förväntade fördelningen beror av en eller flera parametrar blir frihetsgraderna $f = k - r - 1$, där r är antalet från stickprovet skattade parametrar. H_0 förkastas på signifikansnivån α om testvariabelns värde överskrider det kritiska värdet $\chi^2_\alpha(f)$.

Anmärkning. 4.8. Man brukar ha som tumregel att de förväntade frekvenserna np_i ska vara ≥ 1 och helst ≥ 5 för att testet ska fungera bra.

Anmärkning. 4.9. Vektorn av observerade frekvenser $(x_1 \ldots, x_k)$ kan uppfattas som ett utfall av en multinomialfördelad stokastisk vektor

$$(X_1, X_2, \ldots, X_k) \sim \text{ Mult}(n, p_1, p_2, \ldots, p_k).$$

Testet kan då formuleras som ett test av givna värden för de k parametrarna $p_i,\ i = 1, \ldots, k$, vilket är en generalisering av ett tvåsidigt test gällande ett givet värde för parametern p i $\text{Bin}(n, p)$, se avsnitt 4.5.1.

Exempel. 4.11. Ledningen för ett företag misstänker att en del anställda har tagit för vana att förlänga veckoslutet genom att sjukanmäla sig för fredag och/eller måndag. Saken följs upp under fyra veckors tid. Ifall ledingens misstanke är obefogad, skulle vi förvänta oss att sjukfrånvarona fördelar sig jämnt över alla arbetsdagar. Resultatet blir följande:

Veckodag	må	ti	on	to	fre
Observerade frånvaron	49	35	32	39	45
Förväntade frånvaron	40	40	40	40	40

Vi lägger upp följande hypoteser:

H_0 : Frånvarona fördelar sig jämnt över alla arbetsdagar.

H_1 : Frånvarona fördelar sig inte jämnt över alla arbetsdagar.

Värdet på testvariabeln blir nu

$$\chi^2 = \frac{(49-40)^2}{40} + \frac{(35-40)^2}{40} + \cdots + \frac{(45-40)^2}{40} = 4.9.$$

Eftersom den förväntade fördelningen är likformig behöver inga parametrar skattas. Frihetsgraderna blir således $f = 5 - 1 = 4$. Om vi utför testet på signifikansnivån $\alpha = 0.05$ jämför vi testvariabelns värde med det kritiska värdet $\chi^2_{0.05}(4) = 9.488$. Eftersom testvariabelns värde inte överskrider det kritiska värdet förkastar vi inte H_0.

4.7.2 Oberoendetest

Låt oss anta att n observationer kan placeras enligt två olika egenskaper i k respektive l klasser. Frekvenserna med avseende på de båda egenskaperna bildar nu tillsammans en tvådimensionell fördelning med de observerade cellfrekvenserna $x_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, l$.

Vi vill nu testa hur väl den observerade frekvensfördelningen stämmer överens med en förväntad fördelning under antagandet att de två egenskaperna är *oberoende*. Den simultana sannolikheten för att en observation ska tillhöra både klass i och klass j betecknar vi p_{ij}. Ifall antagandet om oberoende stämmer gäller att $p_{ij} = p_{i\cdot}p_{\cdot j}$. De marginella sannolikheterna $p_{i\cdot}$ och $p_{\cdot j}$ är okända men kan skattas enligt

$$\hat{p}_{i\cdot} = \sum_{j=1}^{l} x_{ij}/n \text{ och } \hat{p}_{\cdot j} = \sum_{i=1}^{k} x_{ij}/n.$$

Vi ställer nu upp ett test med hypoteserna

$$H_0: \text{ Egenskaperna oberoende.}$$
$$H_1: \text{ Egenskaperna ej oberoede.}$$

Testvariabeln antar värdet

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(x_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}},$$

som då H_0 är korrekt, är en observation ur en approximativ χ^2-fördelning med frihetsgraderna $f = (k-1)(l-1)$. H_0 förkastas på signifikansnivån α om $\chi^2 > \chi^2_\alpha(f)$.

Exempel. 4.12. För att undersöka effekten av ett nytt vaccin på en sjukdom ges 70 frivilliga försökspersoner vaccinet. I undersökningen ingår även en lika stor kontrollgrupp. De totalt $n = 140$ personerna följs upp under en viss tid och man erhåller följande tvådimensionella frekvensfördelning:

	Har insjuknat	Har ej insjuknat	Tot.
Har vaccinerats	20	50	70
Har ej vaccinerats	40	30	70
Tot.	60	80	140

Följande hypoteser formuleras:

$$H_0: \text{Vaccinet har ingen effekt.}$$
$$H_1: \text{Vaccinet förebygger sjukdomen.}$$

Om vaccin och sjukdom var oberoende skulle vi förvänta oss följande:

	Har insjuknat	Har ej insjuknat	
Har vaccinerats	$\frac{70\cdot 60}{140} = 30$	$\frac{70\cdot 80}{140} = 40$	70
Har ej vaccinerats	$\frac{70\cdot 60}{140} = 30$	$\frac{70\cdot 80}{140} = 40$	70
Tot.	60	80	140

På basen av de två frekvensfördelningarna räknar vi sedan värdet på testvariabeln

$$\chi^2 = \frac{(20-30)^2}{30} + \frac{(50-40)^2}{40} + \frac{(40-30)^2}{30} + \frac{(30-40)^2}{40} = 11.7.$$

Med signifikansnivån $\alpha = 0.01$ och frihetsgraderna $f = (2-1)(2-1) = 1$ förkastar vi H_0 eftersom $\chi^2 = 11.7 > \chi^2_{0.01}(1) = 6.635$.

5 Linjära modeller

Betrakta den stokastiska variabeln

$$Y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

vars väntevärde

$$\mathbb{E}(Y) = \beta_1 x_1 + \cdots + \beta_1 x_p \tag{9}$$

beror av parametrarna $\beta_1, \ldots, \beta_p$ och de givna storheterna $x_1, \ldots, x_p$. Termen ϵ är ett slumpmässigt fel om vilket antas att

$$\begin{aligned} \mathbb{E}(\epsilon) &= 0 \\ \mathrm{Var}(\epsilon) &= \sigma^2 \\ \epsilon &\sim \mathrm{N}(0, \sigma^2). \end{aligned} \tag{10}$$

Av antagandena (9) och (10) följer att

$$Y \sim \mathrm{N}(\beta_1 x_1 + \cdots + \beta_1 x_p, \sigma^2).$$

Ett stickprov om n observationer av Y uppfattas som ett utfall av den stokastiska vektorn

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

där $Y_1, \ldots, Y_n$ antas vara oberoende med lika varians. Ekvationen ovan kan med matrisnotation kompaktare skrivs som

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon. \tag{11}$$

Den modell som specificerats i ekvation (11) och för vilken antagandena (9) och (10) gäller kallas för *linjär modell*. Linjära modeller innefattar som specialfall ett stort antal olika statistiska modeller, bl.a. enkel- och multipel regression, kovariansanalys samt en- och flervägs variansanalys. Vi ska i följande bekanta oss med två av dessa.

5.1 Enkel regression

Låt $(x_1, y_1), \ldots, (x_n, y_n)$ vara ett stickprov av parvisa observationer av vilka $x_1, \ldots, x_n$ är givna storheter (som eventuellt kan vara utfall av stokastiska variabler) och $y_1, \ldots, y_n$ är utfall av de stokastiska variablerna

$$Y_i \sim \ \mathrm{N}(\beta_1 + \beta_2 x_i, \sigma^2),\ i = 1, \ldots, n,$$

dvs.

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

där $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. Matrisen $\mathbf{X}$ i den linjära modellen (11) antar nu alltså formen

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

och vektorn β formen

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

5.1.1 Punktskattning

Målsättningen är att finna ekvationen för en sådan linje

$$y = \beta_1 + \beta_2 x \tag{12}$$

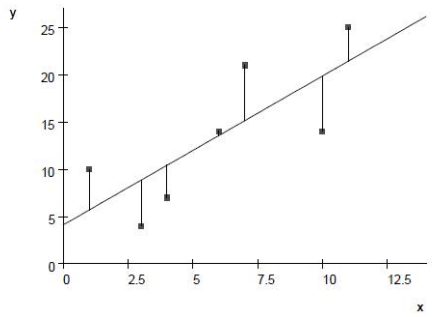
som på bästa möjliga sätt beskriver det *genomsnittliga sambandet* mellan observationerna y_i och x_i. Vi ska alltså skatta värdena på de okända parametrarna β_1 och β_2, m.a.o. skärningspunkten med y-akseln och riktningskoefficienten. I regressionsmodeller skattas parametrarna ofta med minstakvadrat-metoden (se avsnitt 2.2.2).

Anmärkning. 5.1. Man kan visa att då felen ϵ_i är normalfördelade överensstämmer MK-skattninarna med motsvarande ML-skattninar.

MK-skattningarna erhålls genom att minimera

$$Q(\beta_1, \beta_2) = \sum_{i=1}^{n} \big(y_i - \underbrace{(\beta_1 + \beta_2 x_i)}_{\mu_i(\beta_1,\beta_2)}\big)^2,$$

se figur 4. Sätter vi de partiella derivatorna



Figur 4: Skattningarna $\hat{\beta}_1$ och $\hat{\beta}_2$ är de värden för vilka summan av de kvadrerade lodräta avstånden mellan observationerna (y_i, x_i) och linjen $\beta_1 + \beta_2 x$ minimeras.

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i) \quad \text{och} \quad \frac{\partial Q}{\partial \beta_2} = -2\sum_{i=1}^{n} x_i(y_i - \beta_1 - \beta_2 x_i)$$

lika med 0 och löser för β_2 och β_1 får vi

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = r_{xy}\frac{s_y}{s_x} \tag{13}$$

och

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}, \tag{14}$$

där s_x och s_y i ekvation (13) är skattade standardavvikelser för de respektive variablerna och

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

är stickprovskorrelationen mellan variablerna. Genom insättning av skattningarna (14) och (13) i ekvation (12) får vi den skattade *regressionslinjen*

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x.$$

Den skattade *regressionsparametern* $\hat{\beta}_2$ anger effekten av den *förklarande variabeln* x på den *beroende variabeln* y, dvs. hur mycket y ökar då x ökar med en enhet.

Anmärkning. 5.2. Vi gör i detta avsnitt i beteckning ingen skillnad på skattningarna $\hat{\beta}_1$ och $\hat{\beta}_2$ och de estimatorer som skattningarna är observationer av.

5.1.2 Konfidensintervall

Vi är nu intressedrade av parametern β_2. Man kan visa att standardavvikelsen för $\hat{\beta}_2$ är

$$D(\hat{\beta}_2) = \sigma / \sqrt{\textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Då variansen σ^2 är okänd ges för den en väntevärdesriktig skattning av

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} \big(y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i)\big)^2$$

och vi får således medelfelet

$$d(\hat{\beta}_2) = s / \sqrt{\textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Man kan vidare visa att skattningen $\hat{\beta}_2$, se ekvation (13), kan uttryckas som en linjär kombination av $y_1, \ldots, y_n$, vilka är observationer av normalfördelade stokastiska variabler $Y_1, \ldots, Y_n$. Således är även $\hat{\beta}_2$ normalfördelad och vi kan därmed enligt samma principer som i avsnitt 3.2 bestämma ett konfidensintervall för $\hat{\beta}_2$, dvs.

$$I_{\beta_2} = \big(\hat{\beta}_2 - z_{\alpha/2} \cdot D(\hat{\beta}_2)\ ,\ \hat{\beta}_2 + z_{\alpha/2} \cdot D(\hat{\beta}_2)\big)$$

eller då σ^2 är okänd

$$I_{\beta_2} = \big(\hat{\beta}_2 - t_{\alpha/2}(n-2) \cdot d(\hat{\beta}_2)\ ,\ \hat{\beta}_2 + t_{\alpha/2}(n-2) \cdot d(\hat{\beta}_2)\big).$$

5.1.3 Hypotesprövning

Ofta vill man testa huruvida regressionsparametern β_2 skiljer sig signifikant från 0. Logiken är då den att om $\beta_2 = 0$ är väntevärdet för Y konstant

för alla värden av x, alltså finns inget linjärt samband mellan variablerna. Hypoteserna i ett sådant test är

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0.$$

Vi antar nu att σ^2 är okänd. Testet baserar sig då på en t-fördelning med frihetsgraderna $f = n-2$. Värdet på testvariabeln får man genom att dividera skattningen med medelfelet, dvs.

$$t = \frac{\hat{\beta}_2}{d(\hat{\beta}_2)}.$$

H_0 förkastas som bekant om $|t| > t_{\alpha/2}(n-2)$.

I multipel regression där antalet förklarande variabler är $(p-1) \geq 2$ kan man utöver ovannämnda test av enskilda parametrar även utföra ett allmänt test av hypoteserna

$$\begin{aligned} &H_0 : \beta_2 = \cdots = \beta_p = 0 \\ &H_1 : \beta_j \neq 0 \text{ för något index } j = 2, \ldots, p. \end{aligned} \tag{15}$$

Ett liknande test ska vi nedan bekanta oss med i form av variansanalys. Test av denna typ baserar sig på en F-fördelad testvariabel.

Definition. 5.1. Antag att de stokastiska variablerna $X_1, \ldots, X_m, Z_1, \ldots, Z_n$ är oberoende och $\mathrm{N}(0,1)$-fördelade. Variabeln

$$F(m,n) = \frac{\frac{1}{m}(X_1^2 + \cdots + X_m^2)}{\frac{1}{n}(Z_1^2 + \cdots + Z_n^2)} = \frac{\frac{1}{m}\chi^2(m)}{\frac{1}{n}\chi^2(n)}$$

följer då en F-fördelning med frihetsgraderna (m,n).

5.2 Envägs variansanalys

Låt oss nu anta att vi har p stickprov om storlekarna $n_1, \ldots, n_p$, där $\sum_{j=1}^{p} n_j = n$. Stickproven uppfattas som utfall av de stokastiska variablerna

$$Y_{ji} \sim \ \mathrm{N}(\mu_j, \sigma^2),\ i = 1, \ldots, n_j,\ j = 1, \ldots, p,$$

dvs.

$$Y_{ji} = \mu_j + \epsilon_{ji}, \tag{16}$$

där $\epsilon_{ji} \sim \mathrm{N}(0, \sigma^2)$.

Anmärkning. 5.3. Den linjära modellen (11) har i detta fall

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_p} \end{bmatrix}$$

och

$$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix},$$

där $\mathbf{1}_{n_j}$ betecknar en kolumnvektor $(1, 1, \ldots, 1)^{\mathrm{T}}$ av dimension n_j.

5.2.1 Test av likhet mellan väntevärdena

Vi frågar oss nu huruvida väntevärdena $\mu_1, \ldots, \mu_p$ för de p populationerna skiljer sig från varandra. Vi vill alltså testa riktigheten av nollhypotsen

$$H_0 : \mu_1 = \cdots = \mu_p.$$

Vi har redan i avsnitt 4.4.3 bekantat oss med ett t-test för likhet mellan väntevärden av $p = 2$ normalfördelade populationer med lika varians. *Variansanalys* (förkortas ofta ANOVA från *analysis of variance*) är en generalisering av detta test för $p \geq 2$ väntevärden. Namnet har sitt ursprung i att den F-fördelade testvariabeln som testet baserar sig på kan tolkas som en kvot av variansskattningar bildade på två olika sätt.

Ekvation (16) kan alternativt formuleras som

$$Y_{ji} = \mu + \gamma_j + \epsilon_{ji}$$

så att μ är det gemensamma väntevärdet för samtliga p populationer och $\gamma_j = \mu_j - \mu$ anger hur mycket väntevärdet för den j:te populationen skiljer sig från μ. Formuleringen ger oss hypoteserna

$$\begin{aligned} &H_0 : \gamma_1 = \cdots = \gamma_p = 0 \\ &H_1 : \gamma_j \neq 0 \text{ för något index } j = 1, \ldots, p, \end{aligned}$$

jfr. hypoteserna (15) för det allmänna testet i linjär regression.

Det gemensamma väntevärdet μ skattas med

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{p} y_{ji} = \frac{1}{n}\sum_{j=1}^{p} n_j \overline{y}_j.$$

Vidare skattas den gemensamma variansen σ^2 med

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n_j}\sum_{j=1}^{p}(y_{ji} - \overline{y})^2 = \frac{SS_T}{n-1}.$$

Kvadratsumman SS_T i täljaren som reflekterar den sammanlagda *totala* variationen bland samtliga stickprov kan uppdelas enligt

$$SS_T = SS_W + SS_B,$$

där

$$SS_W = \sum_{i=1}^{n_j}\sum_{j=1}^{p}(y_{ji} - \overline{y}_j)^2 = \sum_{j=1}^{p}(n_j - 1)s_j^2$$

står för den sammanlagda variationen *inom* (*within*) de enskilda stickproven och

$$SS_B = \sum_{i=1}^{n_j}\sum_{j=1}^{p}(\overline{y}_j - \overline{y})^2 = \sum_{j=1}^{p} n_j(\overline{y}_j - \overline{y})^2$$

står för den sammanlagda variationen *mellan* (*between*) stickproven. Testvariabeln

$$F = \frac{SS_B/(p-1)}{SS_W/(n-p)}$$

är nu en kvot av variansskattningar beräknade utgående från kvadratsummorna SS_B och SS_W. Logiken är att om variationen mellan stickproven är stor jämfört med variationen inom stickproven har man skäl att ifrågasätta antagandet om att stickproven härstammar från populationer med lika väntevärde. Man kan visa att om H_0 är korrekt följer testvariabeln en F-fördelning med frihetsgraderna $[(p-1),(n-p)]$, se definition 5.1 och anmärkning 4.7. H_0 förkastas då på signifikansnivån α om $F > F_\alpha[(p-1),(n-p)]$.

Exempel. 5.1. Vi drar ett stickprov var från fyra normalt fördelade populationer $\mathrm{N}(\mu_j, \sigma^2),\ j = 1, \ldots, 4$. Av tabellen nedan framgår storleken, medelvärdet och variansen för de respektive stickproven:

n_j	7	5	6	6
$\overline{y}_j$	2.4	3.3	3.4	2.6
s_j^2	0.25	0.34	0.10	0.05

Vi vill nu utföra ett test av nollhypotesen

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

på signifikansnivån $\alpha = 0.05$. Vi gör följande uträkningar:

$$\begin{aligned}
n &= 7 + 5 + 6 + 6 = 24, \\
\overline{y} &= \frac{1}{24}(7 \cdot 2.4 + 5 \cdot 3.3 + 6 \cdot 3.4 + 6 \cdot 2.6) = 2.9, \\
SS_W &= \sum_{j=1}^{p} (n_j - 1) s_j^2 \\
&= (7-1) \cdot 0.25 + (5-1) \cdot 0.34 + (6-1) \cdot 0.10 + (6-1) \cdot 0.05 = 3.6, \\
SS_B &= \sum_{j=1}^{p} n_j (\overline{y}_j - \overline{y})^2 \\
&= 7 \cdot (2.4 - 2.9)^2 + 5 \cdot (3.3 - 2.9)^2 + 6 \cdot (3.4 - 2.9)^2 + 6 \cdot (2.6 - 2.9)^2 = 4.6.
\end{aligned}$$

Nu är

$$F = \frac{4.6/(4-1)}{3.6/(24-4)} = \frac{1.53}{0.18} = 8.66,$$

vilket är större än $F_{0.05}(3, 20) = 3.10$, alltså förkastas H_0. Beslutet kunde vi alternativt ha baserat på att p-värdet $\mathbb{P}(F(3, 20) \geq 8.66) = 0.0007$ är så pass litet att H_0 verkar orimlig.