SF1901
Föreläsning 14: Felfortplantning, medelfel, Gauss approximation, bootstrap

Jan Grandell, Gunnar Englund & Timo Koski

03.03.2016

KTH Matematik

Lärandemål:

- Felfortplantning
  - Medelfel för en punktskattning
  - Gauss' formel
- Ett exempel med minstakvadrat
- Bootstrap
  - Definition
  - Bootstrap och felfortplantning
  - Bootstrap och konfidensintervall

KTH Matematik

Felfortplantning

Felfortplantning kallas 'propagation of error' på engelska. Felfortplantning handlar om hur mätfel fortplantas till funktioner av mätvärden.
Felfortplantningen är av särskild betydelse i samband med behandlingen av toleransfrågor för funktioner av mätvärden.
Vi kommer att först diskutera Gauss' approximationsformler för fortplantning av medelfel.
Förvånande nog finns dock enkla algoritmer som ger tämligen god uppfattning om medelfelet *oavsett hur komplicerad funktionen är* (d.v.s. oavsett hur dålig/bra Gauss' approximation är eller om den alls går att använda). En sådan modern metod är **bootstrap**, som vi beskriver sist.

KTH Matematik

Felfortplantning: Exempel 1. på problemställningen

I en serie försok uppmättes svängningstiden t för en pendel samt pendelns längd l. Mätningarna kan ses som observationer av stokastiska variabler med väntevärden lika med de verkliga värdena t_o och l_o. I ett försök uppmättes $t = 2.0084$ och $l = 1.0031$. Vidare vet man av lång erfarenhet att standardavvikelserna för dessa bestämningar är 2.0084 och 0.00043. Mätningarna kan betraktas som oberoende. Efter försöket beräknas tyngdaccelerationen g enligt formeln

$$g = \frac{4\pi^2 l}{t^2}$$

Sökt: approximativa formler för väntevärde och standardavvikelse för beräknat g.
Svar: Senare

Exempel 2. på problemställningen: Medelfel för en skattning av en statistisk parameter

Vi använder variansen $V(\theta^*)$ eller, vilket i princip är samma sak, standardavvikelsen $D(\theta^*)$ som precisionsmått för en skattning θ^*. Ju mindre varians (större effektivitet), desto belåtnare är vi med skattningen. Ibland hamnar man då i en besvärlig situation: Variansen och standardavvikelsen är själva okända, emedan de beror av just den parameter som man vill skatta (och kanske av ytterligare andra okända parametrar).

KTH Matematik

Medelfel för en skattning

Om man vill få information om $D(\theta^*)$ får man försöka hitta på en skattning även av denna storhet (parameter). Konsekvensen blir att man inte får ett exakt precisionsmått utan bara ett ungefärligt. Vi betecknar denna numeriska skattning av osäkerheten med $d(\theta^*)$.

Definition

En skattning av $D(\theta^*)$ kallas medelfelet för θ^* och betecknas $d(\theta^*)$.

Detta kan nog verka förbryllande men man skall hålla isär begreppen: θ^*_{obs} är en skattning av θ och $d(\theta^*)$ är en skattning av $D(\theta^*)$.

KTH Matematik

Medelfel för en skattning: plug-in

Hur medelfelet skall väljas får avgöras från fall till fall. Man borde tillse att $d(\theta^*)$ är en konsistent skattning av $D(\theta^*)$.
Ett vanligt sätt är s.k. **plug-in**:

$$d(\theta^*) = D(\theta^*_{\text{obs}})$$

d.v.s man insätter θ^*_{obs} för θ i $D(\theta^*)$.

KTH Matematik

Fortplantning av medelfel

Felfortplantning, eller Gauss-approximation, används ibland för att approximativt beräkna medelfel för skattningar. Antag att vi har en skattning θ^*_{obs} som vi vet är approximativt väntevärdesriktig samt anser oss veta medelfelet för. Vi bildar nu en funktion $g(\theta^*_{\text{obs}})$ av skattningen. Detta kan vara aktuellt om vi är intresserade av att skatta en parameter $\psi = g(\theta)$, där en naturlig skattning är $\psi^*_{\text{obs}} = g(\theta^*_{\text{obs}})$.

KTH Matematik

Felfortplantning

Vi undrar nu hur osäkerheten (medelfelet) i θ-skattningen "fortplantas" genom funktionen g till en osäkerhet (medelfel) i ψ-skattningen. Vad vi gör är att serieutveckla (linjärisera) funktionen g genom

$$g(x) = g(a) + (x - a)g'(a) + \text{restterm}$$

där vi anser oss kunna försumma resttermen.

$$\psi^* = g(\theta^*) \approx g(\theta) + (\theta^* - \theta)g'(\theta) + \text{restterm}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Felfortplantning: ett bra läge för att försumma resttermen

$$\psi^* = g(\theta^*) \approx g(\theta) + (\theta^* - \theta)g'(\theta) + \text{restterm}$$

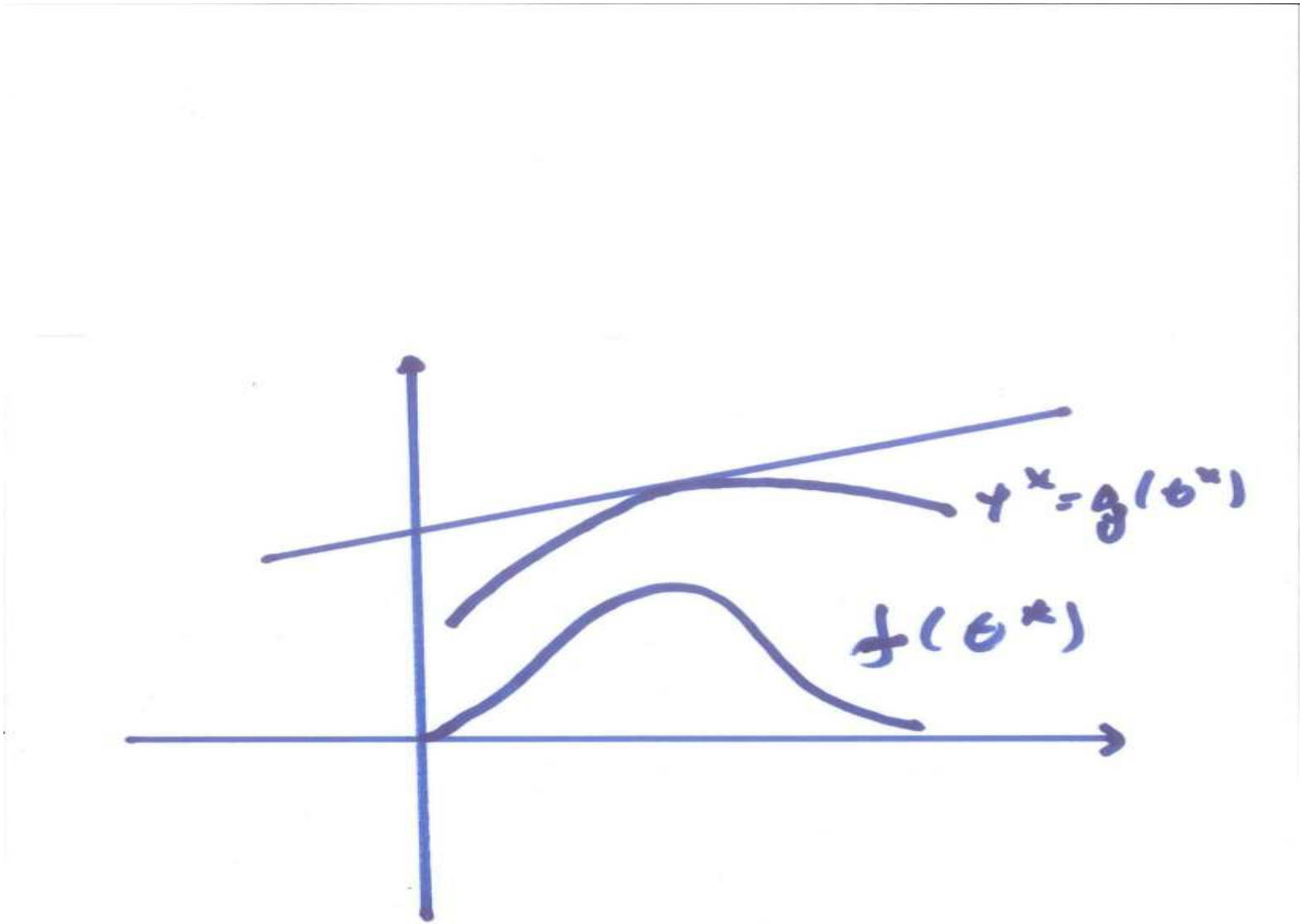
Låt I vara ett intervall, där det mesta av sannolikhetsmassan för θ^* är koncentrerad. Antag att g har kontinuerliga första och andra derivator på I. Då har resttermen det bekanta uttrycket

$$\text{restterm} = (\theta^* - \theta)^2 g''(z)/2,$$

där z ligger mellan θ och θ^*. Antag vidare att $|\, g''(z)\,| \leq C$ för $z \in I$. Då är $E\,|\,\text{restterm}\,| \leq CE(\theta^* - \theta)^2/2$. Om nu $g(x)$ är ungefär linjär, så är C liten. Om dessutom $V(\theta^*) = E(\theta^* - \theta)^2$ är liten, blir approximationen god.

KTH VETENSKAP OCH KONST
KTH Matematik

Felfortplantning: ett bra läge för att försumma resttermen



Felfortplantning

$$g(x) \approx g(a) + (x - a)g'(a)$$

Vad detta innebär är att vi ersätter funktionen $g(x)$ med tangenten i punkten a som (se ovan) är en hyfsad approximation åtminstone i närheten av punkten a. Om vi väljer $a = \theta$ erhåller vi

$$\psi^* = g(\theta^*) \approx g(\theta) + (\theta^* - \theta)g'(\theta).$$

KTH Matematik

Gauss-approximation

Om nu $E(\theta^*) \approx \theta$, d.v.s. att θ^*_{obs} är (åtminstone approximativt) väntevärdesriktig, och dessutom huvuddelen av sannolikhetsmassan i fördelningen för θ^* finns i det område där den linjära approximationen är god erhåller man

$$E(\psi^*) \approx g(\theta) = \psi$$

och

$$V(\psi^*) \approx (g'(\theta))^2 V(\theta^* - \theta) = (g'(\theta))^2 V(\theta^*) \approx (g'(\theta^*_{\text{obs}}))^2 V(\theta^*).$$

Gauss-approximation: linjärisering + plug-in i $g'(\theta)$.

KTH Matematik

Gauss-approximation: varning för aningslös användning

Tag

$$g(x) = x^2, \psi = \theta^2$$

och antag $E(\theta^*) \approx \theta$. Då har vi

$$E\left[(\theta^*)^2\right] = E(\psi^*) \approx g(\theta) = \theta^2$$

och

$$V(\theta^*) = E\left[(\theta^*)^2\right] - (E\left[\theta^*\right])^2 \approx \theta^2 - \theta^2 = 0$$

och

$$V(\psi^*) \approx (g'(\theta^*_{\text{obs}}))^2 V(\theta^*) \approx 0.$$

KTH Matematik

Gauss-approximation

Sats

Om θ^*_{obs} är approximativt väntevärdesriktig med medelfelet $d(\theta^*)$ så gäller att $\psi^*_{obs} = g(\theta^*_{obs})$ är approximativt väntevärdesriktig som skattning av $\psi = g(\theta)$ samt har approximativt (plug-in) medelfel $d(\psi^*) \approx |g'(\theta^*_{obs})| d(\theta^*)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Exempel: *Felfortplantning*

Låt $x_1, \ldots, x_n$ vara utfall av oberoende stokastiska variabler $X_1, \ldots, X_n$, respektive, som är *ffg*-fördelade dvs.

$$p_X(x) = p \cdot (1-p)^{(x-1)} \text{ för } x = 1, 2, \ldots$$

(a) Bestäm minsta-kvadrat-skattningen p^*_{obs} av parametern p, där $0 < p < 1$.

(b) Bestäm approximativt väntevärde $E(p^*)$ och varians $V(p^*)$ för minsta-kvadrat-skattningen p^*_{obs}.

Exempel: Felfortplantning *Minsta kvadrat*

(a) Enligt formelsamlingen är $E(X_i) = \frac{1}{p} = \mu(p)$ och därmed bildar vi minsta-kvadrat-skattningen genom att minimera funktionen

$$Q(p) = \sum_{i=1}^{n} (x_i - \mu(p))^2 .$$

och enligt tidigare fås

$$p^*_{\text{obs}} = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mu^{-1}(\overline{x}) = \frac{1}{\overline{x}},$$

där $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. SVAR: $\underline{p^*_{obs} = \frac{1}{\overline{x}}}$.

KTH Matematik

Felfortplantning

(b) Inför $g(x) = \frac{1}{x}$, för $x > 0$. Sätt $\theta = \frac{1}{p}$, d.v.s. $p = g(\theta)$, $p^* = g(\theta^*)$. Vi har $\theta^* = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Vi har

$$E(\theta^*) = E(\overline{X})$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Men och enligt formelsamlingen är $E(X_i) = \frac{1}{p}$ och således $E(\overline{X}) = \frac{1}{p} = \theta$ dvs.

$$E(\theta^*) = \theta.$$

Då ger Gauss approximation

$$E(p^*) = E(g(\theta^*)) \approx g(E(\theta^*)) = g(\theta) = p.$$

SVAR: $\underline{E(p^*) \approx p}$.

Felfortplantning

Med stöd av Gauss' approximationsformler fås nu att

$$V\left(p^*\right) = V\left(g\left(\theta^*\right)\right) \approx V\left(\theta^*\right) \cdot \left(g^{'}\left(E\left(\theta^*\right)\right)\right)^2.$$

Men eftersom vi har oberoende stokastiska variabler, så blir $V\left(\theta^*\right) = V\left(\overline{X}\right) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} V\left(X_i\right)$. Enligt formelsamlingen är $V\left(X_i\right) = \frac{1-p}{p^2}$. Detta ger $V\left(\theta^*\right) = \frac{1}{n}\frac{1-p}{p^2} (= \frac{1}{n}\theta(\theta-1))$. Eftersom $g^{'}(x) = -1/x^2$, blir $\left(g^{'}\left(E\left(\theta^*\right)\right)\right)^2 = \frac{1}{(E(\theta^*))^4} = \frac{1}{\theta^4} = p^4$.

SVAR: $\underline{V\left(p^*\right) \approx \left(1-p^*\right) \cdot \left(p^*\right)^2/n}$.

Konfidensintervall & felfortplantning

Konfidensintervall för $g(\theta)$ med approximativ konfidensgrad $1-\alpha$ (Gauss approximation används för medelfel)

$$I_{g(\theta)} = (g(\theta^*_{\text{obs}}) - \lambda_{\alpha/2}|g'(\theta^*_{\text{obs}})|d(\theta^*), g(\theta^*_{\text{obs}}) + \lambda_{\alpha/2}|g'(\theta^*_{\text{obs}})|d(\theta^*)).$$

I exemplet ovan

$$I_p = I_{1/\theta} = \frac{1}{\overline{x}} \pm \lambda_{\alpha/2}\sqrt{\frac{\left(1-\frac{1}{\overline{x}}\right)\frac{1}{\overline{x}}^2}{n}}$$

KTH Matematik

Allmänna formeln för fortplantning av varians

$X_1, X_2 \ldots, X_n$ är oberoende stokastiska variabler. Vi har en funktion $U = g\left(X_1, X_2 \ldots, X_n\right)$ som är så snäll att

$$V\left(U\right) \approx \left(\frac{\partial}{\partial x_1} g_1\right)^2 V(X_1) + \ldots + \left(\frac{\partial}{\partial x_n} g_n\right)^2 V(X_n)$$

där

$$\frac{\partial}{\partial x_i} g_i = \frac{\partial}{\partial x_i} g\left(x_1, \ldots, x_n\right) \mid_{x_1 = E[X_1], \ldots, x_n = E[X_n]}$$

Det är ingen svår sak att konstruera ett matematiskt exempel där formeln ger en rätt så dålig approximation. Men många praktiker anser att formeln fungerar fungerar oftast hyfsat väl i tillämpningar.

KTH VETENSKAP OCH KONST
KTH Matematik

Exempel 1. : fortsättning

$U = g(l, t) = \frac{4\pi^2 l}{t^2}$. Då är

$$E(U) = \frac{4\pi^2 l_o}{t_o^2}$$

och

$$\frac{\partial}{\partial l} g_1 = \frac{4\pi^2}{t_o^2}$$
$$\frac{\partial}{\partial t} g_2 = -\frac{8\pi^2 l_o}{t_o^3}.$$

M.h.a. den allmänna formeln för fortplantning av varians

$$V(U) \approx \frac{16\pi^4}{t_o^4} V(l) + \frac{64\pi^4 l_o^2}{t_o^6} V(t).$$

Med $V(t) = 2.0084^2$, $V(l) = 0.00043^2$, och två plug-in $t_o \mapsto 2.0084$ och $l \mapsto 1.0031$. fås

$$\sqrt{V(U)} = 19.635.$$

KTH Matematik

Exempel 1. : konfidensintervall

$$I_U = \frac{4\pi^2 1.0031}{2.00845^2} \pm \lambda_{\alpha/2} 19.635.$$
$$= 9.8176 \pm \lambda_{\alpha/2} 19.635.$$

KTH Matematik

Medelfel för en skattning & bootstrap

Vi har observationerna $x_1, x_2, \ldots, x_n$ som vi ser som utfall av oberoende likafördelade s.v. $X_1, \ldots, X_n$ vilkas fördelning beror av en okänd parameter $\psi = g(\theta)$. Denna skattar vi med en funktion $g\left(\theta^*_{\text{obs}}\right) = g\left(\theta^*(x_1, \ldots, x_n)\right)$. Vi vill beräkna medelfelet $d(g(\theta^*))$.
Vi beskriver nu en algoritm kallad **bootstrap**, som ger tämligen god uppfattning om medelfelet $d(g(\theta^*))$.

KTH Matematik

Bootstrap

Namnet bootstrap (på sv. stövelstropp) härstammar från uttrycket i amerikansk engelska *pull oneself up by the bootstraps*, vilket betyder att man förbättrar ens (ekonomiska/sociala o.s.v.) situation genom egna insatser utan andras hjälp. Ett uttryck med likadan betydelse är *att lyfta sig själv i håret* (jfr. baron von Münchausens äventyr).

KTH Matematik

Bootstrap

Bootstrap: Vi lottar fram ett första **fiktivt** eller **bootstrap- stickprov** ur våra mätdata genom att på måfå dra n st med återläggning från $x_1, \ldots, x_n$. Detta stickprov blir typiskt ungefär som det ursprungliga även om vissa observationer saknas och vissa finns i dubbel eller kanske tredubbel uppsättning.

KTH Matematik

Bootstrap med matlab: exempel

Vi har, till exempel,

$$x_1 = -0.2746, x_2 = -1.1730, x_3 = 1.4842, x_4 = 1.1454, x_5 = -1.6248,$$

$$x_6 = 0.9985, x_7 = 0.4571, x_8 = -1.2315, x_9 = 0.9868, x_{10} = -0.5941$$

eller

$$x = \begin{bmatrix} -0.2746 & -1.1730 & 1.4842 & 1.1454 & -1.6248 \\ 0.9985 & 0.4571 & -1.2315 & 0.9868 & -0.5941 \end{bmatrix};$$

Matlabkommandot unidrnd(10,1,10) ger tio slumptal ur den likformiga fördelningen på de tio heltalen i $1, \ldots, 10$, d.v.s $p(i) = 1/10$.

```
>> z=unidrnd(10,1,10)
z =
5 1 6 3 2 5 6 3 7 10
```

Bootstrapstickprovet ges nu av >> bssampl=x(z)

Bootstrap med matlab: exempel

Med >> bssampl=x(z)
fås bootstrap stickprovet
bssampl =
Columns 1 through 7
-1.6248 -0.2746 0.9985 1.4842 -1.1730 -1.6248 0.9985
Columns 8 through 10
1.4842 0.4571 -0.5941

KTH Matematik

Medelfel för en skattning & bootstrap

Ur detta nya stickprov beräknar vi skattningen $\theta^*_{1,\text{obs}}$ av θ. Detta upprepas nu B gånger (kanske $B = 1000$ eller $B = 10000$) och ur dessa B fiktiva stickprov får vi B skattningar $\theta^*_{1,\text{obs}}, \theta^*_{2,\text{obs}}, \ldots, \theta^*_{B,\text{obs}}$. Vi får då en medelfelsskattning genom att beräkna spridningen av dessa d.v.s.

$$d(\theta^*) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\theta^*_{i,\text{obs}} - \theta^*_{\text{medel,obs}})^2}$$

där $\theta^*_{\text{medel,obs}} = \sum_1^B \theta^*_{i,\text{obs}}/B$ är aritmetiska medelvärdet av de B skattningarna för de B fiktiva stickproven.

Medelfel för en skattning & bootstrap

Vi hade simulerat så att

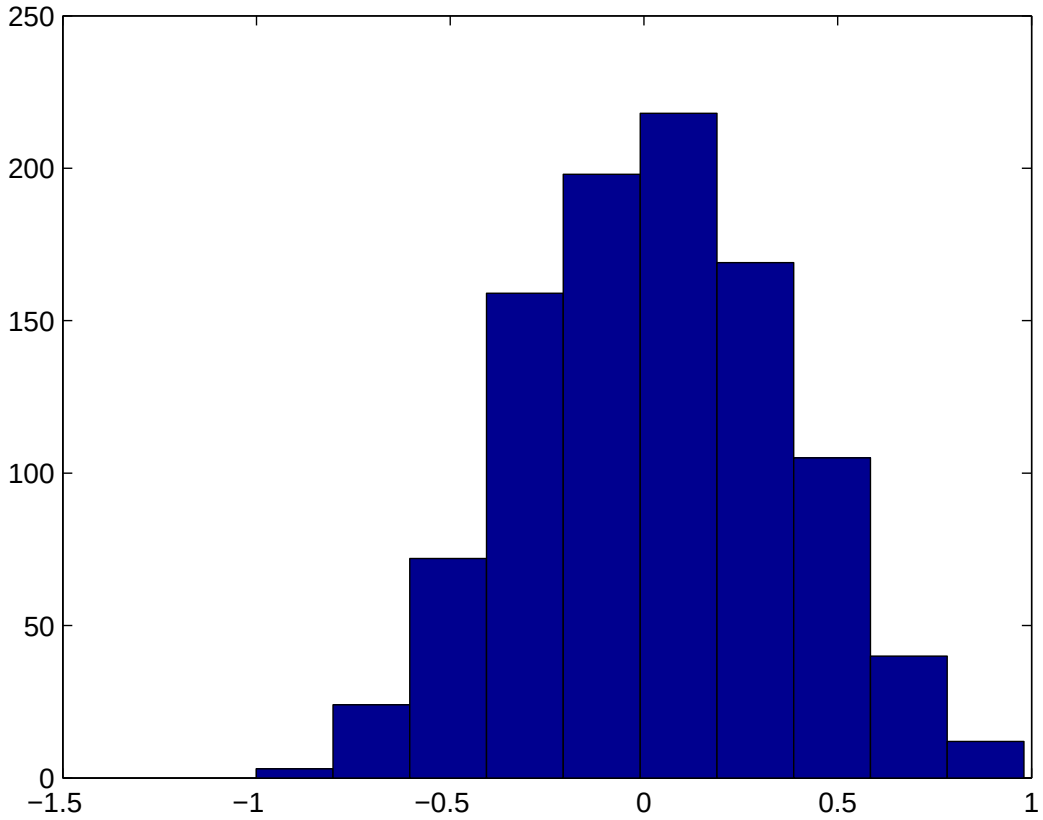
$$x = \big[\begin{array}{rrrrr} -0.2746 & -1.1730 & 1.4842 & 1.1454 & -1.6248 \\ 0.9985 & 0.4571 & -1.2315 & 0.9868 & -0.5941 \end{array} \big];$$

är tio oberoende stickprov på $X_i \in N(0,1)$. Vi har att $\theta^* = \frac{1}{10}(X_1 + \ldots + X_{10}) \in N(0, \sqrt{\frac{1}{10}})$. Om vi drar B bootstrap-stickprov från x och beräknar $\overline{x}_{i,obs}$ för $i = 1, \ldots, B$, får vi enligt receptet ovan en boostrap-skattning av $\sqrt{\frac{1}{10}} = 0.3162$. I figuren ses histogrammet för $\overline{x}_{i,obs}$ för $i = 1, \ldots, B$ med $B = 1000$, där standaravvikelsen är $d(\theta^*) = 0.339$.

KTH VETENSKAP OCH KONST
KTH Matematik

Medelfel för en skattning & bootstrap

I figuren ses histogrammet för $\overline{x}_{i,obs}$ för $i = 1, \ldots, B$ med $B = 1000$, där standaravvikelsen är 0.339.



Medelfel för en skattning & bootstrap

De fiktiva stickproven är framlottade enligt den empiriska fördelningen som lägger massan $1/n$ i vardera av de n observationerna. Denna empiriska fördelning är en skattning av den sanna fördelningen F som observationerna genererats av. Vi har därmed lyckats skapa en kopia av det ursprungliga försöket (där våra data kom från F). Vi kan då undersöka egenskaper hos den ursprungliga skattningen genom undersöka skattningen i kopian.

KTH Matematik

Felfortplantning & bootstrap

För varje $i = 1, \ldots, B$ (kanske $B = 1000$ eller $B = 10000$) drar vi bootstrap-stickprovet

$$(x_{i1}^*, \ldots, x_{in}^*) .$$

Vi beräknar skattningen $\psi_{i,\text{obs}} = g(\theta_{i,\text{obs}}^*)$ $i = 1, \ldots, B$, av $\psi = g(\theta)$. Vi får B skattningar $g(\theta_{1,\text{obs}}^*), g(\theta_{2,\text{obs}}^*), \ldots, g(\theta_{B,\text{obs}}^*)$.
Vi får då en medelfelsskattning $d(g(\theta^*))$ genom att beräkna spridningen av dessa d.v.s.

$$d(g(\theta^*)) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(g(\theta_{i,\text{obs}}^*) - g(\theta^*)_{\text{medel,obs}})^2}$$

där $g(\theta^*)_{\text{medel,obs}} = \sum_1^B g(\theta_{i,\text{obs}}^*)/B$ är aritmetiska medelvärdet av de B skattningarna för de B fiktiva stickproven.

KTH Matematik

Exempel fortsätts med MATLAB: Bootstrap av medelfelet hos minsta-kvadrat-skattningen p^*_{obs} av $ffg(p)$

Vi simulerar i matlab $x_1, \ldots, x_{1000}$ utfall av oberoende stokastiska variabler $X_1, \ldots, X_{1000}$, respektive, som är $ffg(0.67)$-fördelade. Detta görs genom

```
>> x=geornd(0.67,1,1000);
>> x=x+1;
```

Liksom visats ovan är minsta-kvadrat skattningen $p^*_{obs} = 1/\overline{x}$, vilket blev i detta fall

```
>> 1/mean(x)
ans =
0.6579
```

KTH Matematik

Bootstrap av medelfelet hos minsta-kvadrat-skattningen p^*_{obs} av $ffg(p)$

```
function bsstderr= stovelsffg(x,n)
phats=[];
for i=1:n
bssampl=stovelstropp(x);
phat=1/mean(bssampl);
phats=[phats phat] ;
end
bsstderr= std(phats);
```

Här anropas matlab-funktionen stovelstropp.m, som de facto framställdes redan ovan men återges i bilagan.

Exempel forts. med MATLAB: Bootstrap av medelfelet hos minsta-kvadrat-skattningen p^*_{obs} av $ffg(p)$

Vi använder funktionen stovelsffg.m (se bilagan) och får bootstrapskattningen av $d(p^*_{obs})$ med $B = 1000$ (B behöver inte vara lika med antalet ursprungliga stickprov) som

$>>$ stovelsffg(x,1000)

ans $=$

0.0120

Gauss-approximationen från ovan ger för det aktuella värdet på p^*

$$d\left(p^*\right) \approx \sqrt{\left(1-0.6579\right) 0.6579^2 / 1000} = 0.0122$$

KTH Matematik

Konfidensintervall för minsta-kvadrat-skattningen p^*_{obs} av $ffg(p)$

$$I_p = \frac{1}{\overline{x}} \pm \lambda_{\alpha/2} \sqrt{\left(1 - \frac{1}{\overline{x}}\right) \frac{1}{\overline{x}}^2 / n}$$

Med data ovan

$$I_p = 0.6579 \pm \lambda_{\alpha/2} 0.0122.$$

Detta är inte ett s.k. bootstrap-konfidensintervall för p.

KTH VETENSKAP OCH KONST
KTH Matematik

Bootstrap med matlab:

Med arbetsättet ovan måste vi skriva en ny .m fil (d.v.s. skriva om definitionen på phat) för varje ny stickprovsvariabel, vars medelfel skall undersökas.
I själva verket har Matlab en inbyggd funktion **bootstrp.m**, som tar en funktion som argument.
Mer om bootstrap kan vid behov läsas i
`http://www.math.kth.se/matstat/gru/sf2955/datortot.pdf`

KTH VETENSKAP OCH KONST
KTH Matematik

Bilaga: Bootstrap stovelstropp.m

```
function bssampl=stovelstropp(x)
m=length(x);
z=unidrnd(m,1,m);
bssampl=x(z);
```

KTH Matematik

Bilaga: Baron von Münchausen lyfter sig i håret ur ett träsk

KTH Matematik