

Bootstrapping

per.liv@umu.se



Vad är sannolikheten att
Christer Pettersson sköt Olof
Palme?

Definition av sannolikhet

Frekventistisk statistiker



"Andel gånger som
händelsen inträffar om
processen upprepas ett
stort antal gånger"

Bayesiansk statistiker



"Degree of belief"

Statistikbegrepp

Sannolikhet för en händelse:

andel gånger som en händelse inträffar om processen upprepas många (oändligt antal) gånger.

Statistikbegrepp

P-värde vid t-test:

andel gånger som en lika stor eller ännu större gruppskillnad som den vi observerat skulle erhållas om studien skulle upprepas många gånger,

givet att det inte finns någon äkta skillnad mellan populationerna

Statistikbegrepp

Standardavvikelse:

mått på hur utspridd en variabel är kring sitt medelvärde

Statistikbegrepp

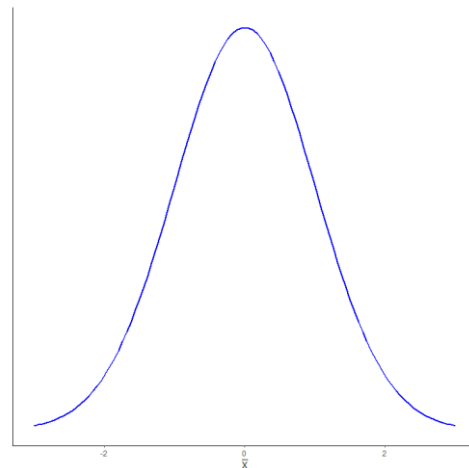
Standard error av ett medelvärde:

Standardavvikelsen för hur medelvärdena skulle fördela sig om vi skulle upprepa studien många gånger.



Upprepa många gånger

Samplingfördelning för $\bar{X}$



Standard error: standardavvikelse för medelvärkets samplingfördelning

[...om man upprepar studien många gånger]

Statistikbegrepp

95%:igt konfidensintervall av ett medelvärde:

Ett intervall som kommer innehålla populationens sanna medelvärde i 95 % av gångerna om vi upprepar studien många gånger.

Upprepa!

Upprepa!

Upprepa!

Upprepa!

Återkommande tema:

”..om vi skulle upprepa många gånger!!!!”

Upprepa!

Upprepa!

Upprepa!

Upprepa!



Problem: matematiken inte kan alltid ge oss en enkel formel för att skatta osäkerheten vid upprepning

Tex:

- Ofta när normalfördelningsantagande inte håller
- När vi ska göra komplicerade procedurer med flera källor av osäkerhet

95 %igt konfidensintervall för medelvärde

Normalfördelning:

$$\bar{X} \pm 1.96 * SEM$$

Ännu bättre med t-fördelningen:

$$\bar{X} \pm t_{0.975}(n-1) * SEM$$

**Robust mot avvikelser från icke-normalfördelning pg.a.
samplingfördelningen blir lik normalfördelning (om n stort)**

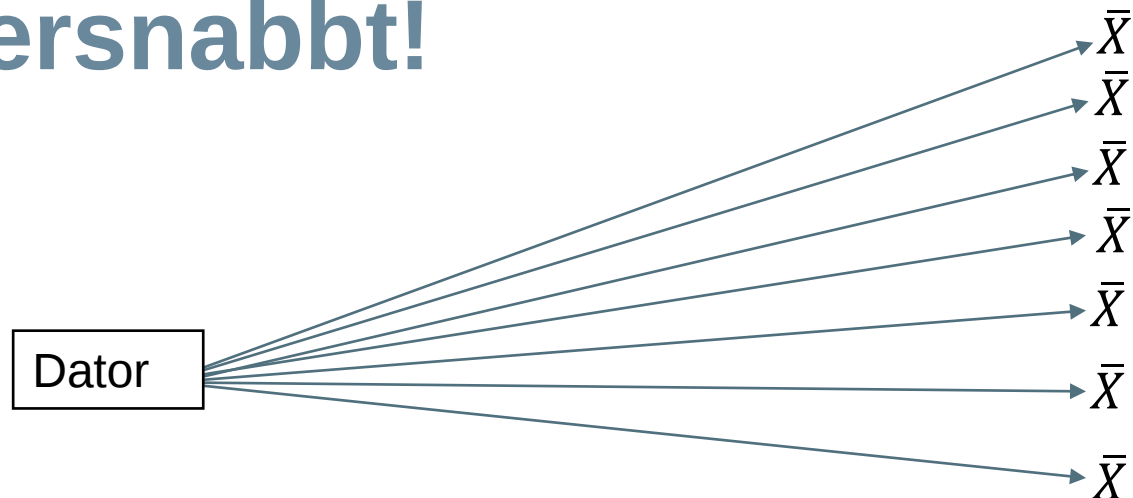


Variationskoefficienten ($= \frac{sd}{\bar{x}}$) då?

- Variation i skattning av standardavvikelsen sd
- Variation i skattning av medelvärdet $\bar{X}$
- $SE(sd)$ är mycket känslig för avvikelser från normalfördelning

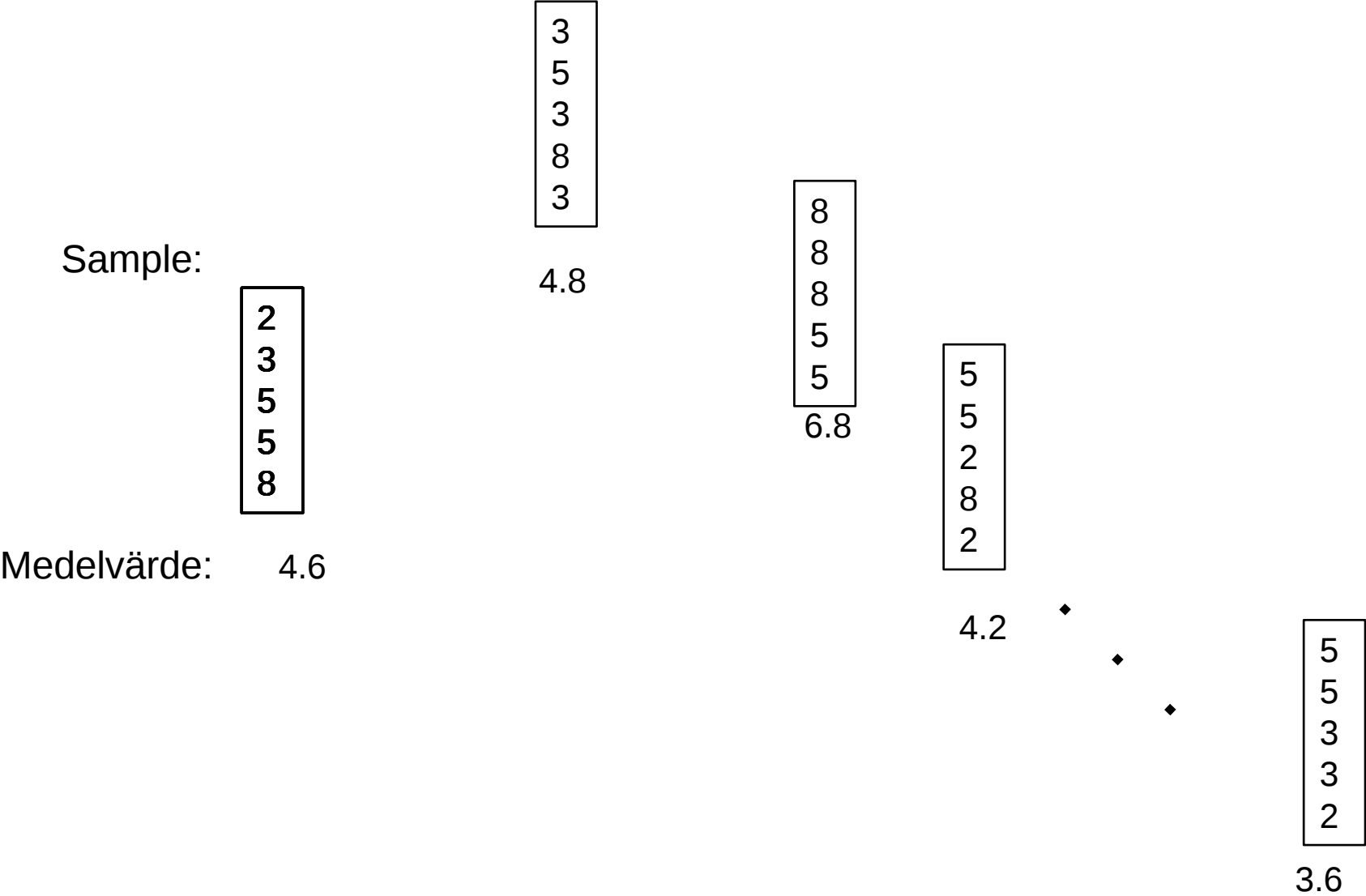
Svårt att beräkna bra konfidensintervall

Med simulering kan vi upprepa experiment supersnabbt!



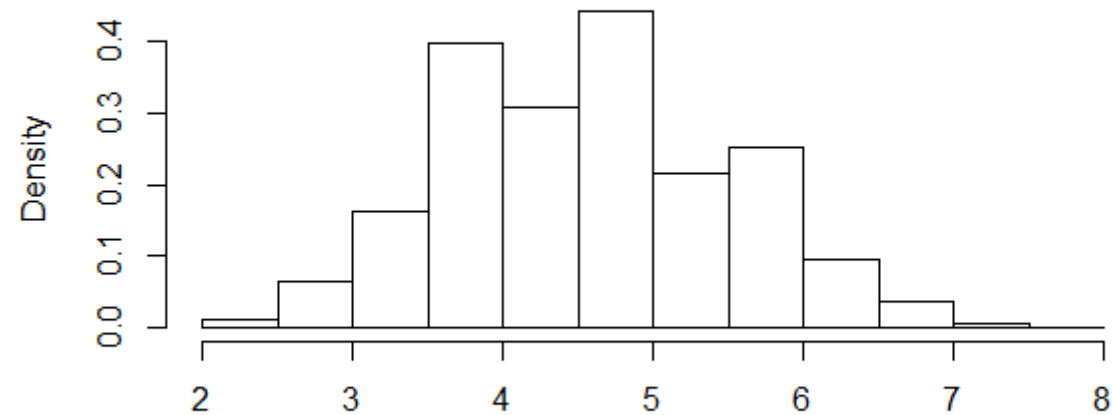
**MEN: kräver att modellen vi simulerar från överensstämmer med verkligheten
Vilken sannolikhetsfördelning ska vi simulera ifrån?**

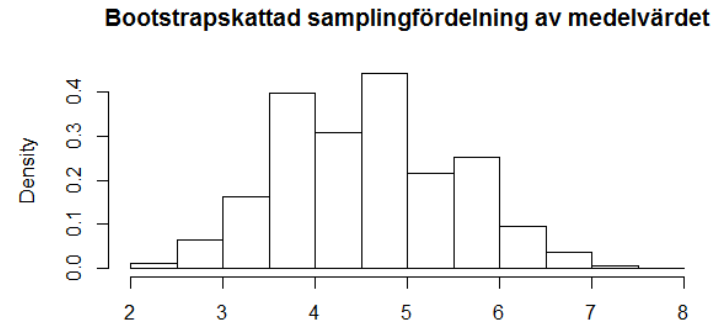
Sampling med återläggning



Samplingfördelning

Bootstrapskattad samplingfördelning av medelvärdet





Bootstrapfördelningen används sedan t.ex. till att beräkna:

- Variansrelaterade mått
 - Konfidensintervall
 - Standard errors etc...
- P-värden
- Bias
 - Validering av regressionsmodeller

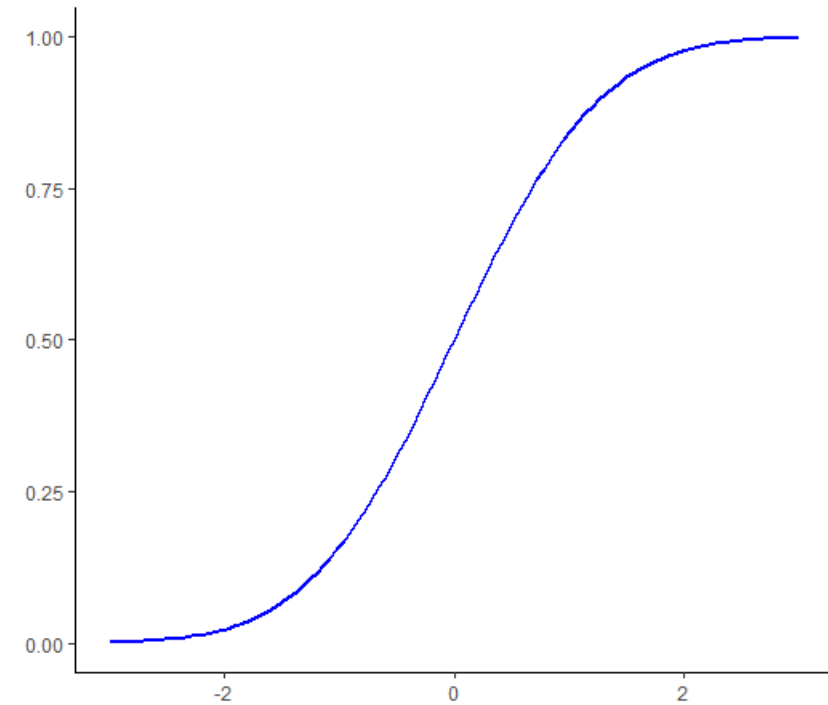
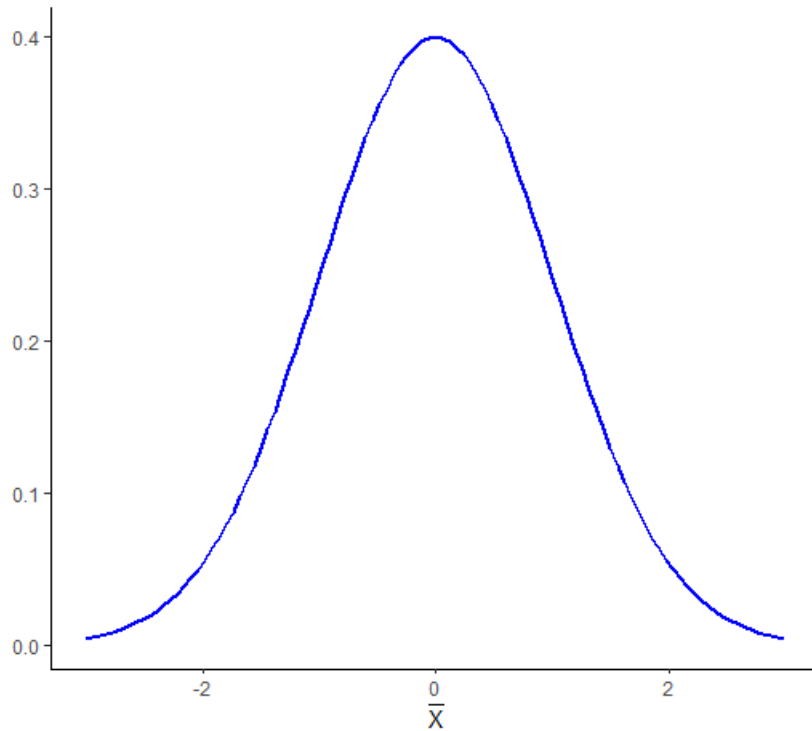
Kort historia:

- 30-tal: permutationstester utvecklas
- 50-tal: jackknife-metodiken utvecklas
- 1979: Bradley Efron (f. 1938) publicerade den första artikeln om bootstrapping¹
- Och därigenom påbörjades en revolution...

¹Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

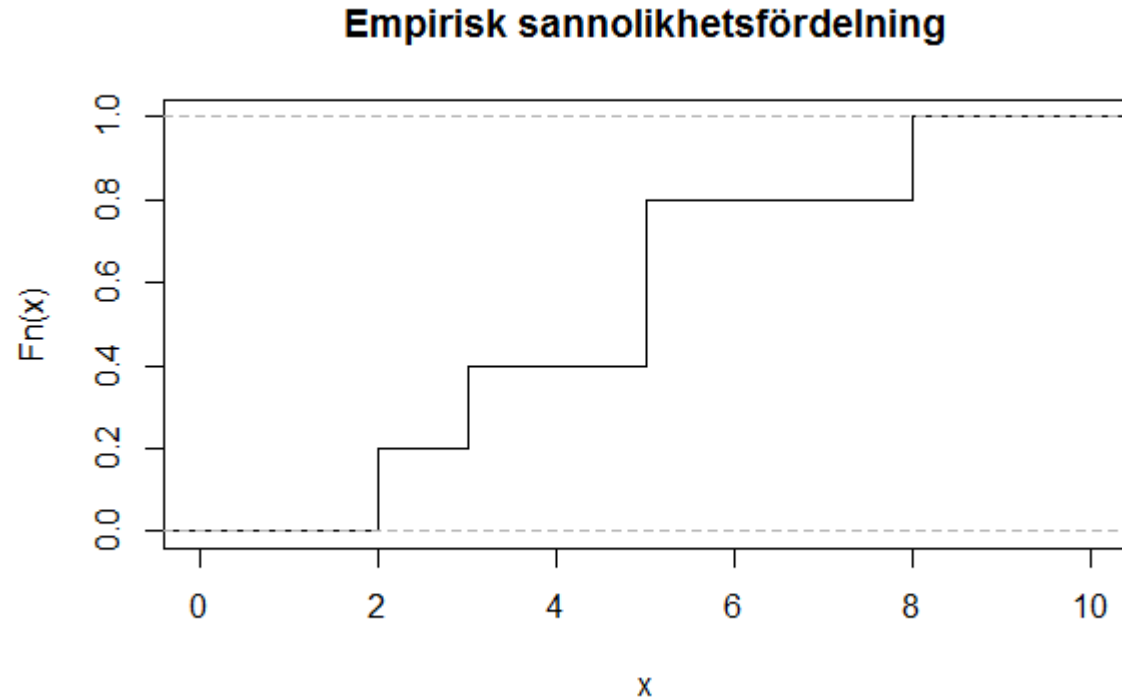
Varför fungerar bootstrapping?

Kumulativ sannolikhetsfördelning



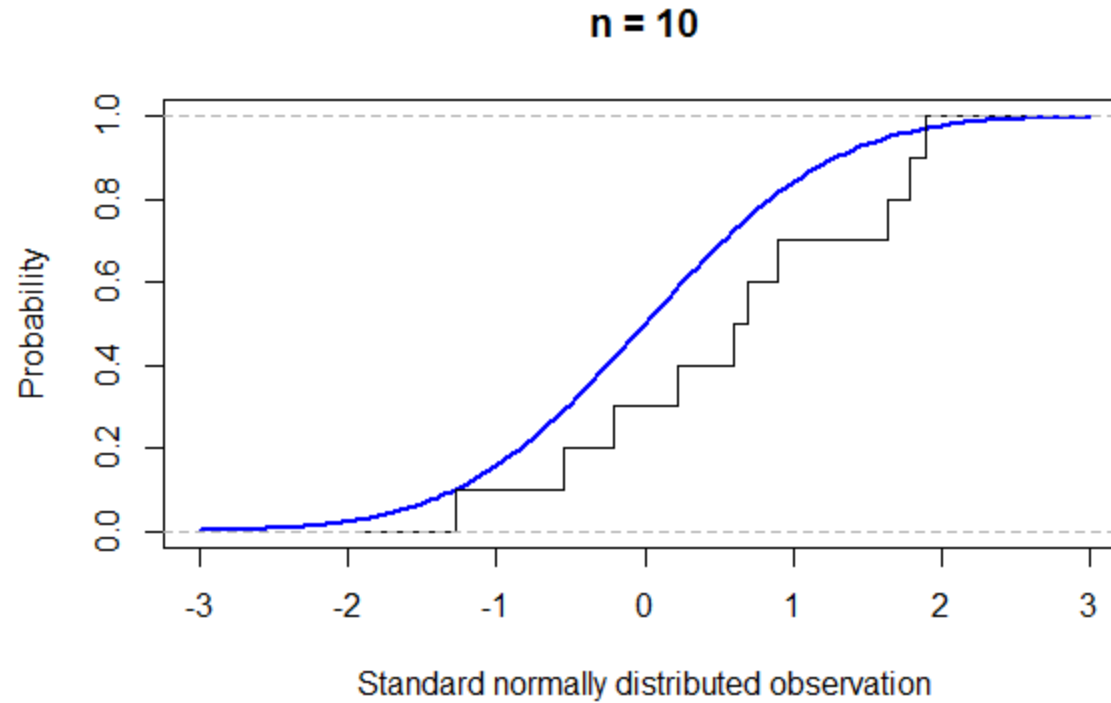
Sample:

2
3
5
5
8

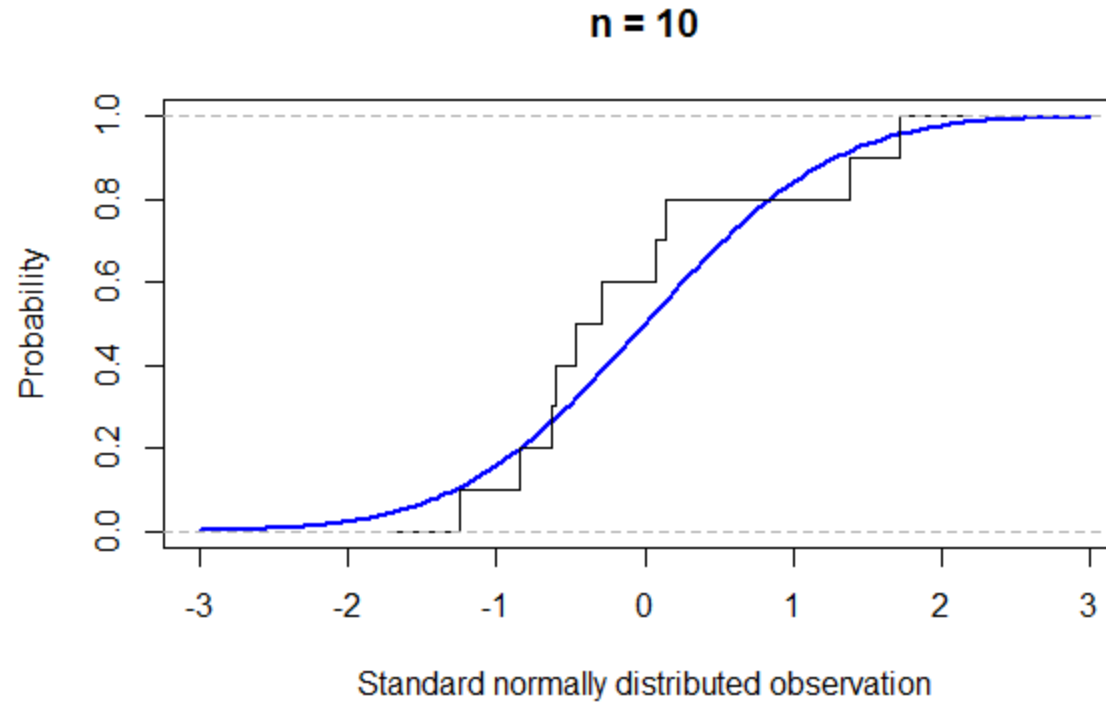


Att sampla med återläggning är att simulera från den empiriska kumulativa sannolikhetsfördelningen!

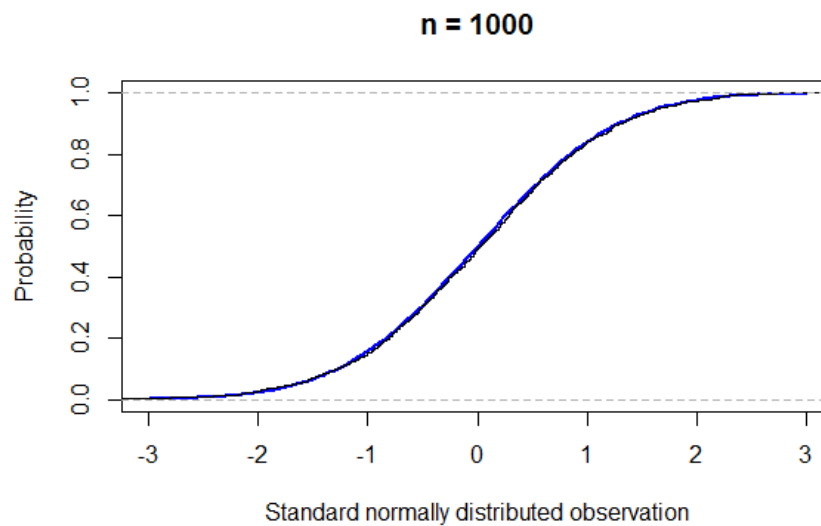
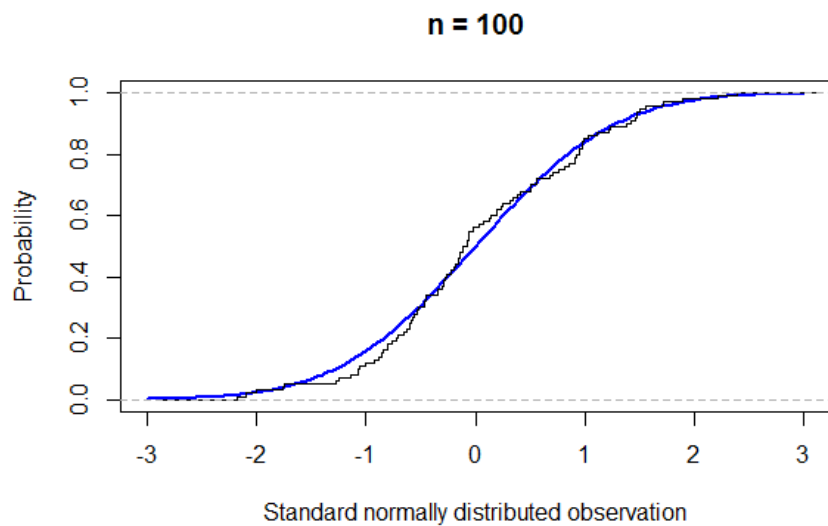
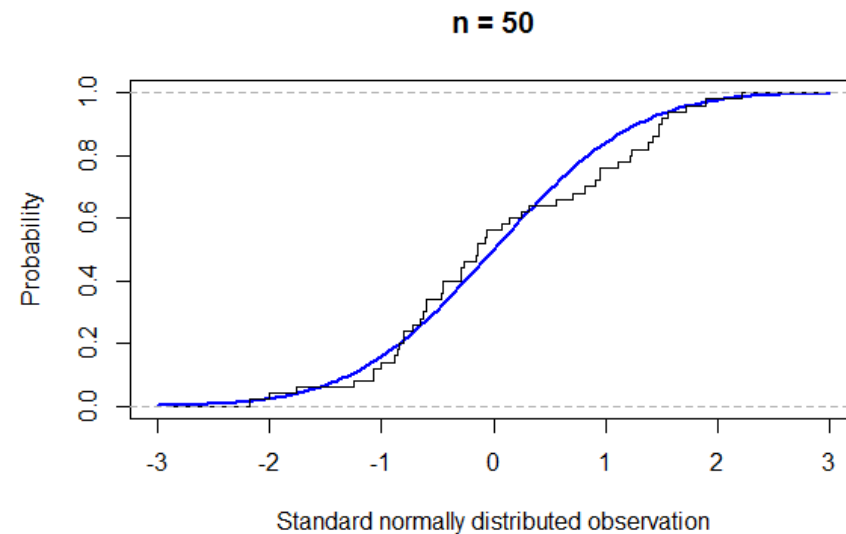
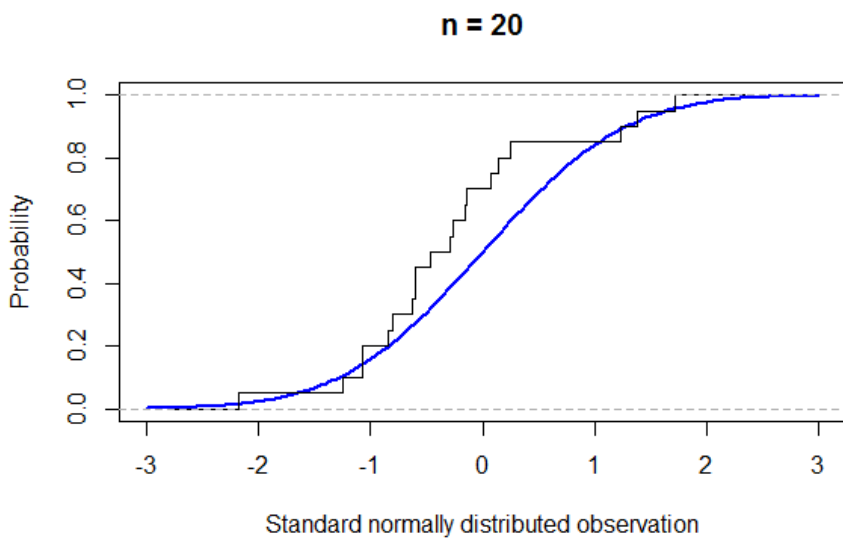
Ibland hamnar empiriska fördelningen **under** den sanna fördelningen....



Ibland hamnar empiriska fördelningen **över** den sanna fördelningen....

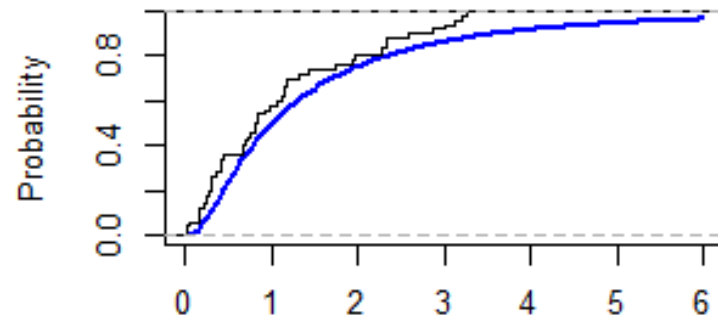


...men i genomsnitt blir det (i de flesta fall) hyfsat rätt.

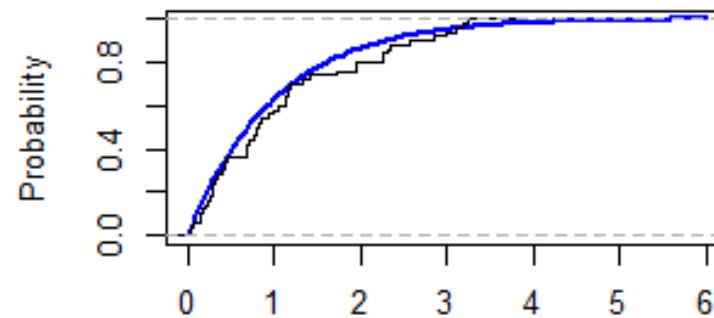


Och det fungerar (oftast) även vid skeva och "icke-normalfördelade" variabler...

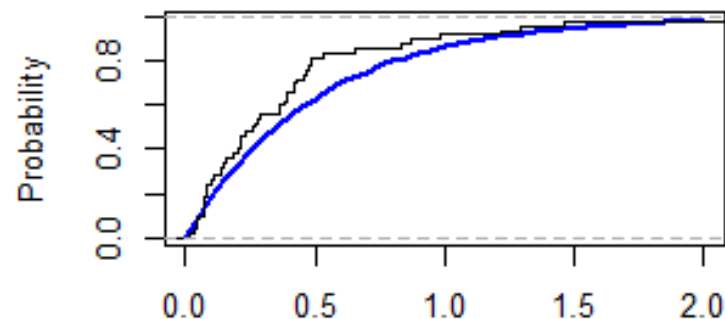
Lognormalfördelning (n = 50)



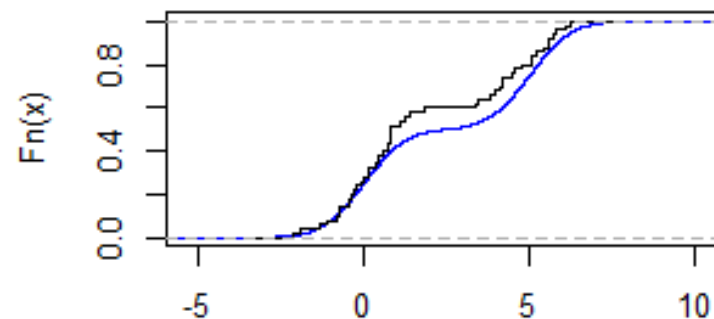
Exponentialfördelning (n = 50)



Gammafördelning (n = 50)

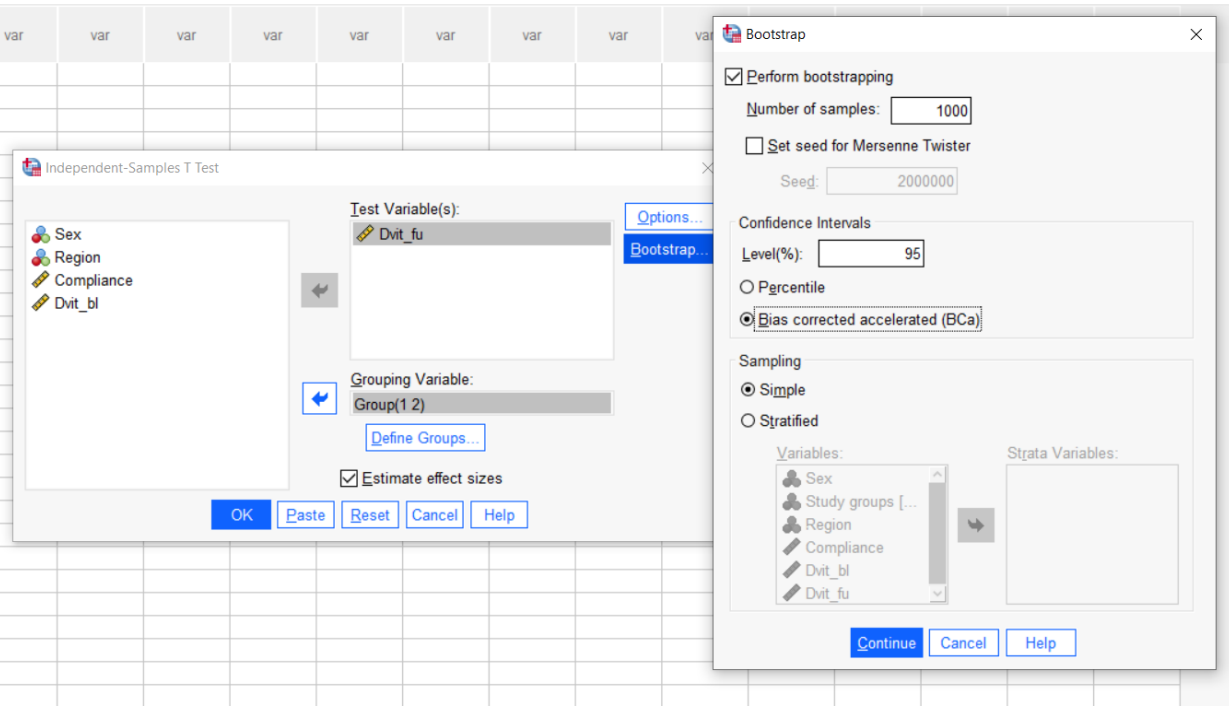


Bimodal blandfördelning (normal)



Kan alla använda bootstrapping?

- Mest användbar om man kan skriva egen kod
- Gå till en statistiker?
- Använda SPSS inbyggda funktioner



Independent Samples Test

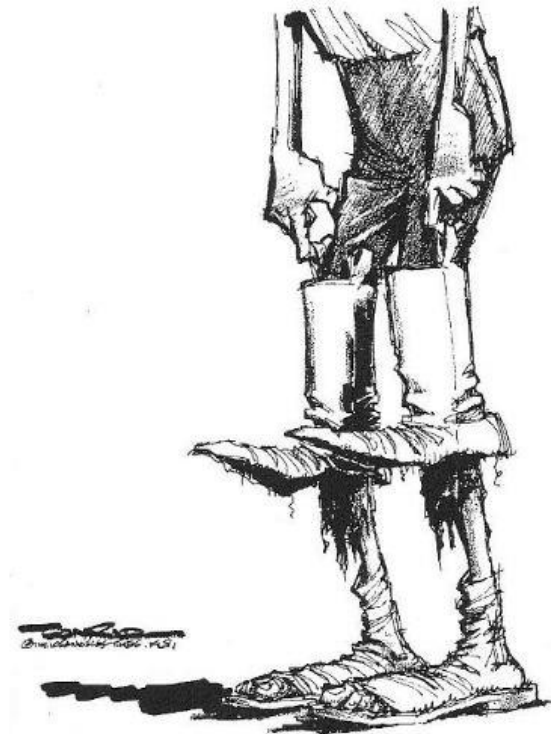
		Levene's Test for Equality of Variances		t-test for Equality of Means						95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Dvit_fu	Equal variances assumed	,244	,623	3,184	117	<,001	,002	11,18120	3,51124	4,22737	18,13503
	Equal variances not assumed			3,151	62,268	,001	,003	11,18120	3,54831	4,08883	18,27357

Bootstrap for Independent Samples Test

		Bootstrap ^a					BCa 95% Confidence Interval	
		Mean Difference	Bias	Std. Error	Sig. (2-tailed)		Lower	Upper
Dvit_fu	Equal variances assumed	11,18120	,08750	3,66216	,005		3,77761	18,63310
	Equal variances not assumed	11,18120	,08750	3,66216	,006		3,77761	18,63310

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

När fungerar inte bootstrapping?



Var försiktig om:

- Lite data
- Outliers
- Går ej att använda på extremvärden (min, max)
- Ändliga populationer
- Grupperingar/kluster/upprepade mätningar

Simuleringsexperiment:

Hur bra fungerar bootstrapping för små samplestorlekar?

- Simulera 10 000 dataset från känd fördelning med n observationer
- För varje dataset:
 - beräkna 95 %-igt konfidensintervall för medelvärdet
 - verifiera om fördelningens sanna medelvärde ligger inom konfidensintervallet
- Beräkna hur stor andel av upprepningarna som konfidensintervallet täcker in det sanna medelvärdet

De fyra olika konfidensintervallen:

- $Z: \pm 1.96*SEM$ - normalfördelningsapproximation
- $t: \pm t_{0.975}(n-1)*SEM$ - t-fördelning
- Basic bootstrap - enklaste bootstrapintervallet
- Bc_a –bootstrap - förbättrat bootstrapintervall

Simulering från normalfördelning

Resultat:

	Z	t	basic	BCa
n = 50				
n= 15				
n = 5				

De fyra olika konfidensintervallen:

- $Z: \pm 1.96*SEM$ - normalfördelningsapproximation
- $t: \pm t_{0.975}(n-1)*SEM$ - t-fördelning
- Basic bootstrap - enklaste bootstrapintervallet
- Bc_a –bootstrap - förbättrat bootstrapintervall

Simulering från normalfördelning

Resultat:

	Z	t	basic	BCa
n = 50	94,3 %	94,8 %	94,2 %	94,0 %
n= 15				
n = 5				

De fyra olika konfidensintervallen:

- $Z: \pm 1.96*SEM$ - normalfördelningsapproximation
- $t: \pm t_{0.975}(n-1)*SEM$ - t-fördelning
- Basic bootstrap - enklaste bootstrapintervallet
- Bc_a –bootstrap - förbättrat bootstrapintervall

Simulering från normalfördelning

Resultat:

	Z	t	basic	BCa
n = 50	94,3 %	94,8 %	94,2 %	94,0 %
n= 15	93,1 %	95,1 %	91,2 %	92,1 %
n = 5				

De fyra olika konfidensintervallen:

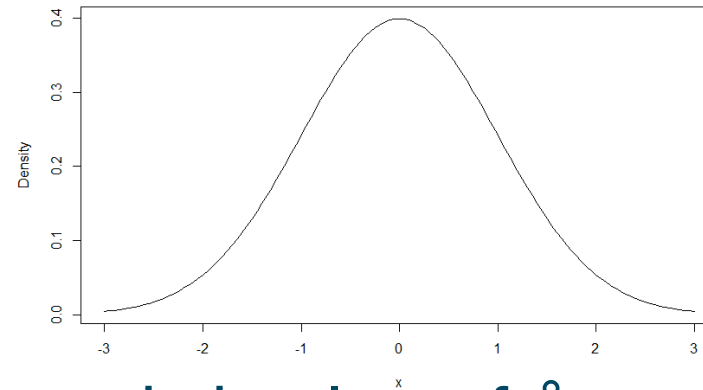
- $Z: \pm 1.96*SEM$ - normalfördelningsapproximation
- $t: \pm t_{0.975}(n-1)*SEM$ - t-fördelning
- Basic bootstrap - enklaste bootstrapintervallet
- Bc_a –bootstrap - förbättrat bootstrapintervall

Simulering från normalfördelning

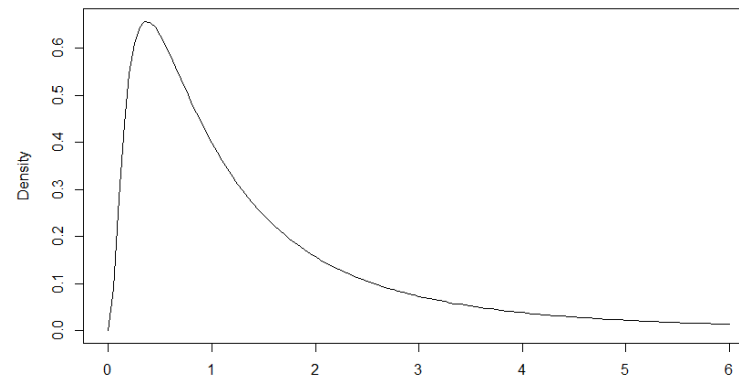
Resultat:

	Z	t	basic	BCa
n = 50	94,3 %	94,8 %	94,2 %	94,0 %
n= 15	93,1 %	95,1 %	91,2 %	92,1 %
n = 5	87,9 %	95,2 %	83,9 %	83,8 %

Men det var den "snälla" normalfördelning.



Vad händer om vi simulerar från en skev fördelning, t.ex. lognormalfördelning?



De fyra olika konfidensintervallen:

- $Z: \pm 1.96*SEM$ - normalfördelningsapproximation
- $t: \pm t_{0.975}(n-1)*SEM$ - t-fördelning
- Basic bootstrap - enklaste bootstrapintervallet
- Bc_a –bootstrap - förbättrat bootstrapintervall

Simulering från lognormalfördelning

Resultat:

	Z	t	basic	BCa
n = 50	90.6 %	90,9 %	88,5 %	91,3 %
n = 15				
n = 5				

De fyra olika konfidensintervallen:

- $Z: \pm 1.96*SEM$ - normalfördelningsapproximation
- $t: \pm t_{0.975}(n-1)*SEM$ - t-fördelning
- Basic bootstrap - enklaste bootstrapintervallet
- Bc_a –bootstrap - förbättrat bootstrapintervall

Simulering från lognormalfördelning

Resultat:

	Z	t	basic	BCa
n = 50	90.6 %	90,9 %	88,5 %	91,3 %
n = 15	84,5 %	86,0 %	81,6 %	85,9 %
n = 5				

De fyra olika konfidensintervallen:

- $Z: \pm 1.96*SEM$ - normalfördelningsapproximation
- $t: \pm t_{0.975}(n-1)*SEM$ - t-fördelning
- Basic bootstrap - enklaste bootstrapintervallet
- Bc_a –bootstrap - förbättrat bootstrapintervall

Simulering från lognormalfördelning

Resultat:

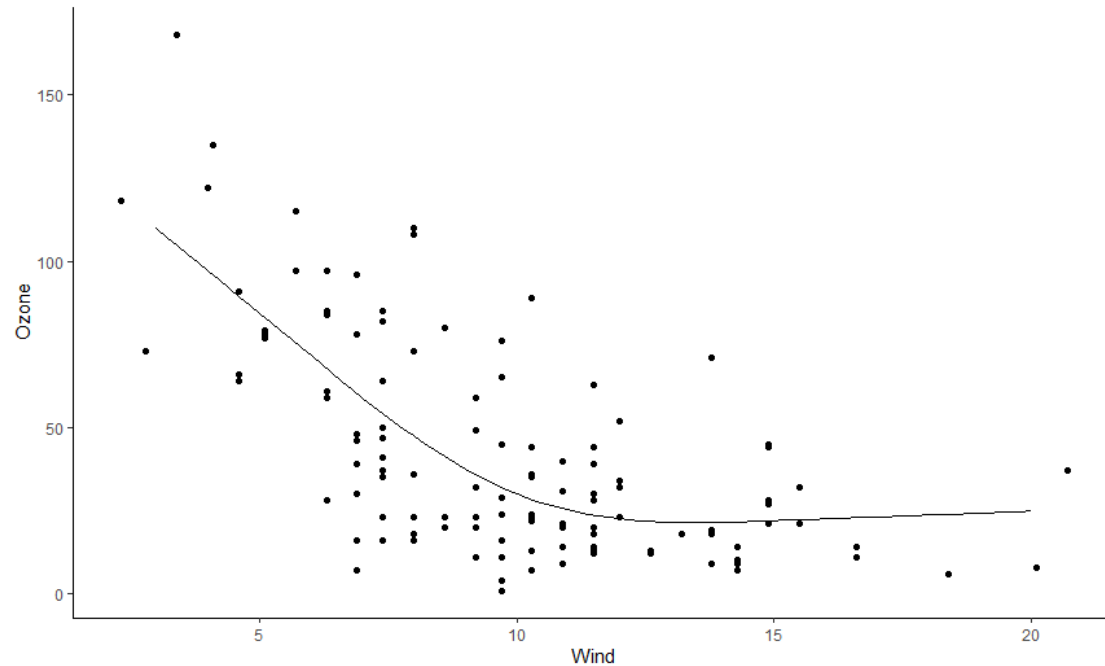
	Z	t	basic	BCa
n = 50	90.6 %	90,9 %	88,5 %	91,3 %
n = 15	84,5 %	86,0 %	81,6 %	85,9 %
n = 5	74.4 %	82.2 %	69.2 %	73.9 %

Slutsatser från vår minisimuleringsstudie

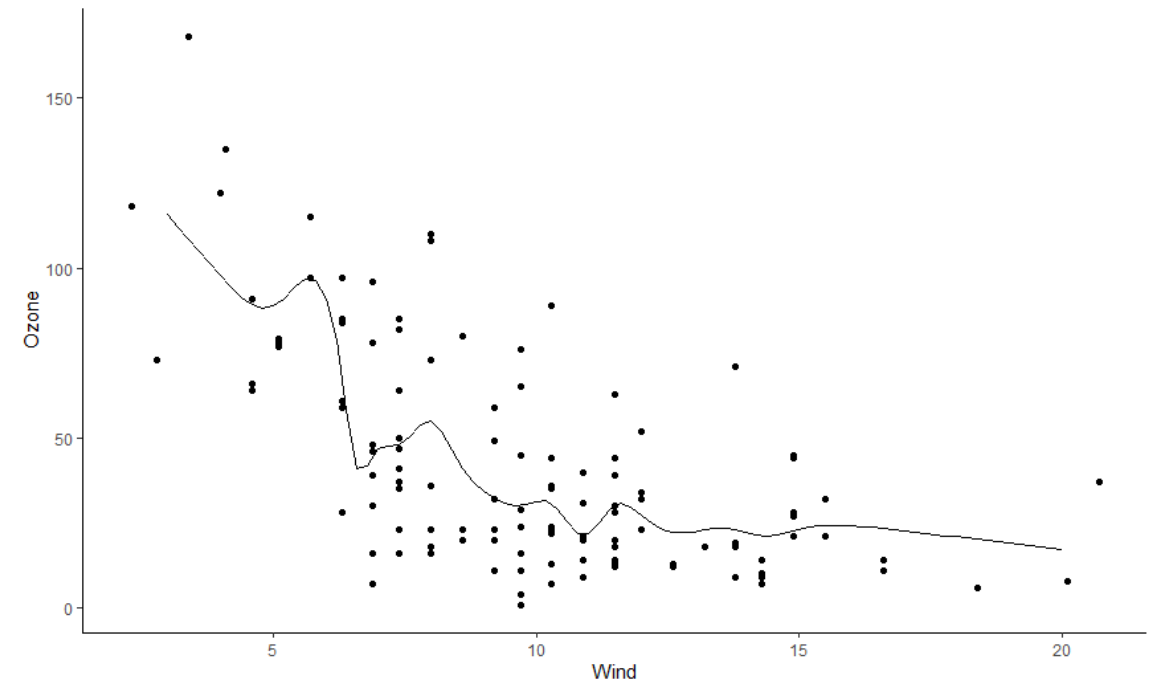
- Bootstrapping fungerar bra om stickprovsstorleken är stor
- Om samplet är "litet" (under 20), fungerar den sämre
- Högre samplestorlek krävs om bootstrapfördelningen är skev
- För beräkning av konfidensintervall för medelvärde tycks t-fördelningen vara minst lika bra som bootstrapping även vid icke-normalfördelning

Två exempel på andra tillämpningar

Fiffig tillämpning 1: validering av prediktionsmodell



Överanpassning



Överanpassad modell

- Visst mått på överanpassning har vi jämt
- Ger bias i mått på hur bra modellen är att prediktera
 - Linjär regression: R^2 ,
 - Binära utfall: AUC

Om vi använder modellen på nya data kommer vi inte få lika bra prediktioner som modellens R^2 /AUC lovar

Validering

- Ett vanlig sätt att validera en prediktionsmodell är split-sample validering
 - Dela data i två delar 70% / 30%
 - Anpassa modellen på 70% av data
 - Testa hur bra modellen är att prediktera dom övriga 30%
- Nackdel: vi utnyttjar bara 70% av data när vi skattar modellen
- Alternativ:
 - Bootstrapping för att skatta hur mycket för optimistisk våra R^2 /AUC är¹

¹Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4), 361-387.

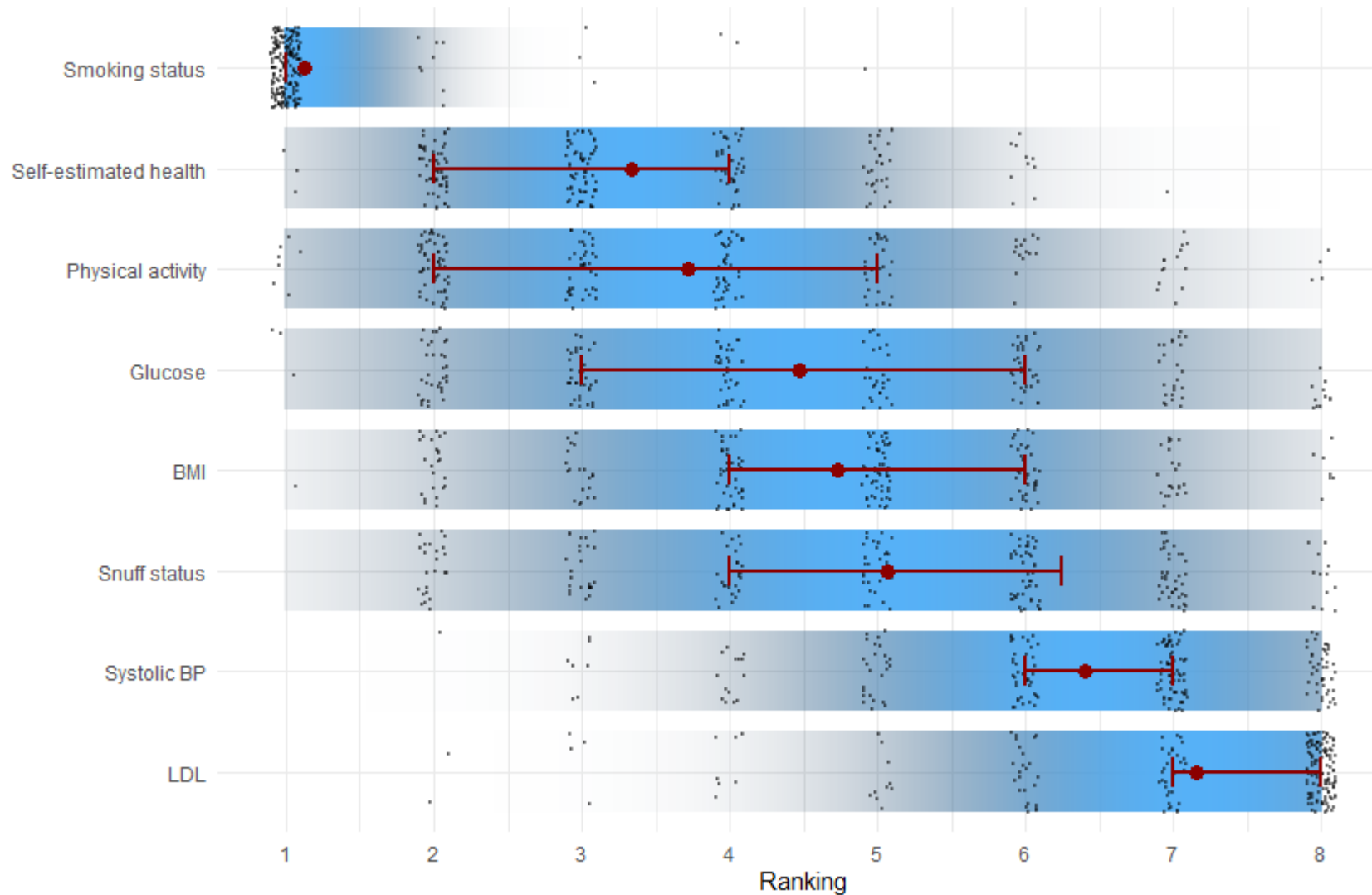
Fiffig tillämpning 2: stabilitet av rangordning

Problem: Vi vill veta vilken faktor som är viktigast för att man ska åldras med god hälsa.

Metod: utvärdera prediktiv förmåga av hälsosamt åldrande hos några friskfaktorer och rangordna

Men om vi upprepade vår datainsamling, skulle samma variabler vara viktigast?

Hur stabil är vår rangordning egentligen?



Varje liten svart punkt = en rangordning för en riskfaktor.

Röda punkterna = medelrangordningen över de 200 simuleringarna för varje riskfaktor

Röda error bars = kvartilerna av rangordningarna för varje riskfaktor



Frågor?

<https://www.canvas.umu.se/courses/2600/>

Nästa seminarium:

18/1 kl 12-13 Nyfiken på interaktioner

Interaktioner är ett vanligt begrepp inom statistik och epidemiologi. Begreppet har dock olika betydelse beroende på sammanhang. I den här föreläsningen ger vi en introduktion till interaktioner och förklarar skillnaden mellan interaktioner inom statistiken och begreppen biological/causal-interactions inom epidemiologin. Vi berör begrepp som additiv och multiplikativ interaktion; effektmodifikation; Relative excess risk due to Interaction (RERI); och Attributable proportion due to interaction (AP) och förklarar dess användningsområden.

Gabriel Granåsen, Region Västerbotten/ Institutionen för folkhälsa och klinisk medicin, Umeå universitet

Programmet för vårterminen kommer strax



Kontakta oss gärna!

Kontaktuppgifter till lokalt Kliniskt forskningscentrum hittar du på www.forumnorr.se
eller kontakta
forumnorr@regionvasterbotten.se