

Sannolikhets- och statistikteoriens huvudprinciper

Tom Britton, Stockholms Universitet

Innehåll:

1. Sannolikhets teori och slump
2. Några "klassiker"
3. Statistikens huvudidéer
 - a) Hypotestest (p-värde)
 - b) Konfidensintervall
 - c) Försöksplanering: randomisera!

Tom Britton

Syfte med föreläsningen:

Förstå huvudidéer, speciellt de som används i "all" vetenskap.

Ej att räkna med sannolikheter.

Lämplig litteratur:

Blom, Enger, Grandell, Holst, Englund: Sannolikhets teori och statistik teori med tillämpningar. *Lärobok för matematiskt skolade*

Britton o Garmo: Sannolikhetslära och statistik för lärare. *Lärobok för mindre matematiskt skolade*

Häggström: Slumpens skördar. *Strövtåg inom sannolikhets teori*

1

Tom Britton

1. Sannolikhets teori och slump

Sannolikhets teori och statistik teori = Matematisk statistik

Matematisk statistik = del av matematik som behandlar slumpmässiga fenomen

Finns slump?

Trots detta är sannolikhets teori viktigt:

bra sätt att ta hand om det man inte känner till i detalj

Sannolikhets teori började med spel och dobbel

Några förgrundsgestalter: Bernoulli (två olika), Pascal, Gauss, Kolmogorov och Cramér (svensk)

Tom Britton

Definition av sannolikhet (Kolmogorov):

Sannolikheten för en händelse A skrivs $P(A)$. $P(A)$ är en reell funktion som för att kallas sannolikhetsfunktion måste uppfylla tre krav:

- 1) $0 \leq P(A) \leq 1$
- 2) $P(\Omega) = 1$ (där Ω är mängden av alla möjliga utfall)
- 3) Om A och B är oförenliga händelser (dvs $A \cap B = \emptyset$, dvs de kan inte inträffa samtidigt) så gäller $P(A \cup B) = P(A) + P(B)$.

Hur skall man då tolka $P(A) = p$?

Några exempel:

"Sannolikheten att fastna i tullen på rutinkontroll är 0.005" (frekventistisk = vanligaste tolkning)

"Sannolikheten (eller risken) för en härdsmläta inom det närmaste decenniet är 10^{-8} " (svårare tolkning, konsekvens av vissa andra sannolikheter)

"Sannolikheten att Djurgården tar SM-guld är 90%" (subjektiv sannolikhet)

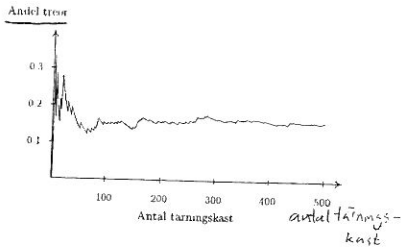
2

3

Sannolikhetens två viktigaste resultat:

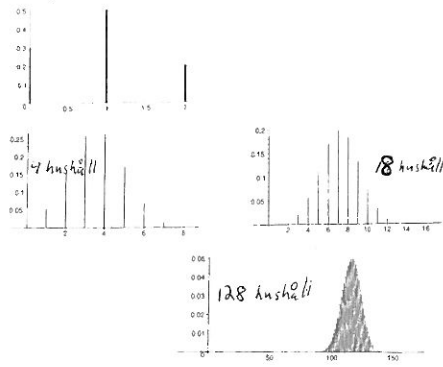
Stora talens lag: Om ett slumpexperiment med numeriskt utfall upprepas många obereonde gånger så kommer experimentens medelvärde att stabilisera sig (dvs knappast vara slumpmässigt). Värdet som det stabiliseras kring kallas experimentets väntevärde (eng. expectation)

Simulering av
500 tärningskast
 $P(3:a) = \frac{1}{6} \approx 0.167$



Centrala gränsvärdessatsen: De små avvikelserna från väntevärdet kommer vara normalfördelade. Dvs summan av många slumpstorheter är normalfördelad.

Ex. Antal bilar per hushåll:
30% har ingen bil
50% har 1 bil
20% har 2 bilar



- Efter att du pekat på en lucka kommer tävlingsledaren öppna en av de andra luckorna bakom vilket en get visar sig (Obs: detta är alltid möjligt!)

- Hon erbjuder dig därefter att byta lucka innan övriga luckor öppnas.

- Du får det som finns bakom luckan du slutligen pekar på

Fråga: Bör man byta eller behålla lucka? Vad är chansen att man vinner bilen om man byter respektive behåller sin lucka?

Svar:

Varför?

Från början pekar man "rätt" med slh 1/3 och "fel" med slh 2/3

"Behåll-strategi": Du vinner om du pekade rätt från början men förlorar annars

"Byt-strategi": Du förlorar om du pekade rätt från början men vinner annars

2. Några klassiska exempel

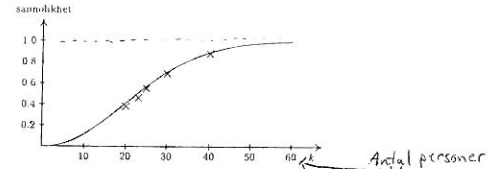
Ex 1. Födelsdagsproblemet: Vad är sannolikheten att det i en klass med 25 elever är några som fyller år samma dag?

Förenklningar: oberoende (t ex inga tvillingar), alla dagar lika sannolika, glöm skottår

Svar:

Varför så stor sannolikhet? Sannolikheten att ett enskilt par skall fylla år samma dag är ganska liten (=1/365). Men det finns många olika par ...

Sannolikheten att några har samma födelsedag



Ex 2. Bilen och getterna: En tävling. Tre luckor, en bil bakom en av luckorna och en get bakom övriga två. Tävlingsledaren vet svaret men du (=den tävlande) vet ej.

- Du får peka på en valfri lucka

Ex 3. Samlarbilder Vid köp av tuggummi får man en samlarbild. Det finns totalt 100 olika bilder

Hur många tuggummin kommer du "i medel" behöva köpa innan du har fått alla samlarbilder?

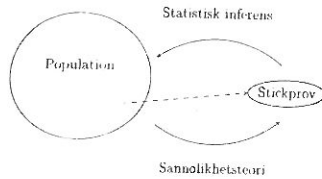
Svar:

Varför?: Majoriteten av tuggummiköp kommer göras när endast några få bilder saknas.

3. Statistikteorins huvudidéer :

SIH-teori. Uttalar sig om sannolikheten för olika utfall givet vissa modellförutsättningar

Statistikteori. Uttalar sig om vissa modellförutsättningar givet ett (eller en följd av) observerade utfall. (Ofta förutsätts vissa modellantaganden)



Ex. SIH-teori. "Om (den teoretiska) medelvikten på en svensk slaktko är 380 kg bör medelvikten hos 20 slaktkor med stor sannolikhet ligga mellan 374 kg och 386 kg."

Motsv. Ex. Statistik-teori. "Den uppmätta medelvikten bland 20 svenska slaktkor var 378 kg. Av detta drar man slutsatsen att (den teoretiska) medelvikten hos svenska slaktkor ligger mellan 372 kg och 384 kg."

Några förgrundsgestalter: Fisher, Neyman, Cramér.

8

Statistikteorins huvudprincip :

"Osannolika händelser inträffar inte"

... eller åtminstone. Om vi väljer mellan två modellantaganden och våra observerade data är osannolika för det ena antagandet men ganska sannolika för det andra, då tror vi på den andra modellen.

" $\leq 5\%$ sannolikhet = osannolikt"

Obs: Med hjälp av statistik kan man aldrig visa något med 100% säkerhet. Men man kan "tjäna" mycket om man nöjer sig med lite mindre än 100% säkerhet.

Ex. Man vill veta hur stor del av Sveriges befolkning som stöder en viss reform.

- Vill man veta den exakta andelen måste samtliga Sveriges 9 miljoner invånare tillfrågas
- Nöjer man sig att veta andelen ± 0.01 med 95% säkerhet räcker det att fråga 5000 (!) slumpmässigt utvalda

Fråga: Hur många krävs för samma precision i Kinas befolkning?

Svar:

9

a) Hypotesprövning

- Vi tvekar mellan två utsagor och skall samla in data för att avgöra vilken som verkar riktig

Ex Ökar snusning risken för cancer eller ej?

- De två utsagorna är oftast inte jämbördiga. Den ena är mer neutral och kanske den för tillfället rådande medan den andra är den vi tror på och den är mer "djärv".

Forts. Ex "Snusning ökar *inte* risken för cancer" respektive "Snusning ökar risken för cancer".

- Den neutrala kallas *nollhypotes*, H_0 , och den vi tror på kallas *alternativhypotesen*, H_A .

- Initialt utgår vi från att H_0 är sann och tittar om vårt observerade datamaterial är sannolikt eller osannolikt. I det förra fallet kvarstår H_0 (den kan inte förkastas). I det senare fallet *förkastas* H_0 till förmån för H_A .

10

Forts. Ex Scenario 1: Om 72 av 3000 snusare fick cancer medan 68 av 3000 (på andra sätt liknande) icke-snusare fick cancer kan inte H_0 förkastas: studien *kan inte* hävda att snusning "signifikant" ökar risken för cancer. Den observerade skillnaden kan uppstå av ren slump.

Forts. Ex Scenario 2: Om 180 av 3000 snusare fick cancer medan 68 av 3000 (på andra sätt liknande) icke-snusare fick cancer kan H_0 förkastas: studien *kan* hävda att snusning ökar risken för cancer signifikant. Den observerade skillnaden kan "knappast" uppstå av ren slump.

En vanlig gräns för att förkasta H_0 är att sannolikheten för observerade data har mindre sannolikhet än 5% (= testets signifikansnivå).

p -värde: Sannolikheten för observerade data under H_0 kallas för testets p -värde (eng p -value). Litet p -värde betyder att vårt data är osannolikt att observera under H_0 .

11

Två potentiella fel:

Fel av typ 1: H_0 förkastas trots att den i själva verket är sann.

Detta fel begås med 5% sannolikhet (eller annan vald signifikansnivå)

Fel av typ 2: H_0 förkastas *inte* trots att den i själva verket är falsk (och H_A är sann).

Detta fel begås med en sannolikhet som beror på:

- 1) Hur mycket data som samlats in, och
- 2) Hur avvikande H_A är från H_0 .

$1 - P(\text{Fel av typ 2}) = \text{testets styrka (eng. power)}$.

12

Forts Exempel

Antag att den sanna risken för cancer i aktuell befolkningsgrupp är 0.005 för icke-snusare respektive 0.006 för snusare (vilket betyder att H_A är sann) och att studien baseras på 3000 patienter i respektive patientgrupp

Då har vi väldigt liten chans att upptäcka denna skillnad, dvs att förkasta H_0 . Testets styrka är låg

Hade vi däremot 100 000 patienter i respektive grupp är det troligt att vi kan upptäcka skillnaden och förkasta H_0 . Testets styrka är hög.

Likaså är det troligt att förkasta H_0 med endast 3000 patienter i respektive grupp om risken för cancer bland snusare är så hög som 0.03. Testets styrka är även här hög

13

OBS!! Bland många genomförda hypotestest så kommer ca 5% av testen som förkastar H_0 på 5%-nivån ha fel! Dvs i ca 5% av dessa kommer H_0 i själva verket vara sann!!

⇒ För att nya forskningsrön skall "godkännas" av forskarsamhället krävs högre signifikansnivå och flera oberoende studier som pekar i samma riktning.

Men sensationslystna tidningar kan publicera "larmrapporter" baserat på enstaka studier ...

14

b) Konfidensintervall

Bakgrund: Ibland vill man ta reda på en storhets värde snarare än att välja mellan två alternativ.

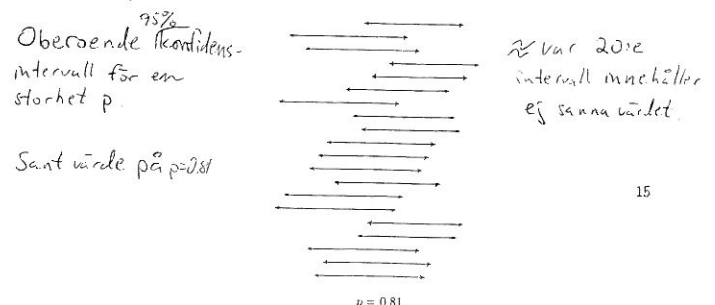
Forts. **Exempel.** Vad är risken för cancer vid snusning?

Från insamlade data kommer svaret alltid innehålla viss mått av osäkerhet.

Ett 95% *konfidensintervall* för en storhet är ett intervall som innehåller de 95% mest troliga värdena på storheten.

Ett 95% *konfidensintervall* utgörs av de värden på H_0 som inte förkastas vid ett hypotestest.

OBS!! Ett konfidensintervall är slumpmässigt! Somliga (ca 95%) innehåller det sanna värdet – resten inte.



15

c) Försöksplanering

Ofta när man planerar ett experiment får olika "enheter" olika "behandling" och man vill jämföra effekterna av de olika "behandlingarna"

Exempel:

 Patienter får ny behandling som jämförs med placebo

 Växter odlas under olika betingelser

Fishers huvudbidrag till statistikteorin:

Bestäm vilka enheter som får respektive behandling genom randomisering!!

Varför? Om man inte randomiserar riskerar man alltid att få systematiska fel. Randomisering gör att dessa jämnas ut.

Konstigt? Genom att tillföra slump ökar man vetenskapen från experiment!!

Slutord

Dagens statistikprogram är lätta att köra, men ...

Ett gott råd:

Utför och redovisa bara den statistik du förstår/behärskar
– överlåt resten åt proffs

Reklam: Statistiska Forskningsgruppen vid Stockholms Universitet (icke-vinstdrivande): www.math.su.se/matstat/sfg

Även andra statistiska konsultbyråer finns