

Statistisk översikt kurs - Föreläsning 4

Anders Fredriksson

Statistiska Institutionen
Stockholms Universitet

28 mars 2025



Stockholm
University

Föreläsning 4 - vart är vi på väg?

- **Vi börjar med två illustrationer i R:**

- Vi tittar på ett stort antal fiktiva "opinionsundersökningar" med slumpmässiga urval (punktskattning av en andel mm.)
- Vi drar slumpmässiga urval från ett dataset (en tänkt population) där två variabler är (linjärt) associerade med varandra. Vad kan vi säga, från våra urval, om hur sambandet ser ut? (regression mm.)

- **Föreläsning 5-6** tar upp bla punktskattning, konfindensintervall, inferens och hypotesprövning
- **Föreläsning 7-9** tar upp samband mellan två och flera variabler, regression, prediktion
- På vägen dit behöver vi förstå bla. slump och sannolikhet, normalfördelningen, och de stora talens lag - **Föreläsning 4**



- Slump och sannolikhet - begrepp och motivering
- Exempel med tärningskastning (många kast...)
- De stora talens lag
- Centrala gränsvärdessatsen
- Samplingfördelning
- Normalfördelning

Slump och sannolikhet - motivering

- Modellera fenomen som har en slumpkomponent.
- **Ex:** Vi väljer slumpvis 1000 röstberättigade och beräknar andelen socialdemokrater, ex. 0.40. Vilka slutsatser kan vi dra?
- Vi behöver en verktygslåda för hur vi ska hantera sådana mätningar - vilka slutsatser kan vi dra? (**inferens**)
- Här kommer sannolikhetsteori in.
- Fatta beslut i en osäker värld.
- **Ex:** Hur många skyddsmasker ska finnas i lager, hur ofta kan tunnelbanan gå utan att vi får trafikstockning?, vilken andel pensionspengar bör investeras i aktier?



Slump, slumpvariabler













- **Slump** - när ett händelseförlopp inte är deterministiskt
 - Kasta ett mynt, en tärning - vi vet inte vad resultatet kommer att bli
 - Vädret - finns ingen deterministisk modell för att förutspå exakt väder (exv., ska det snöa på fredag kl 16.00?) - det finns en slumpkomponent
 - Samhällsvetenskapliga fenomen (exv., vad bestämmer inflationen?) - vi har ekonomisk teori men det finns också en slumpkomponent
 - Om man behöver reparera sin bil eller inte - beror delvis på saker vi inte kan förutspå och inte kontrollera
- **Slumpvariabel/Stokastisk variabel** (Random/stochastic variable) - en variabel vars värden helt eller delvis bestäms av slumpen
 - Vilket värde vi får på ett tärningskast
 - Temperatur, nederbörd etc., vid olika tidpunkter
 - Inflationsutfall för olika månader
 - Årlig reparationskostnad för en bil

- **Sannolikhet** - den relativa frekvensen av ett visst utfall
- **Ex:** Hälften av slantsinglingarna leder till "krona" (sannolikhet för krona = 0.5)

- Vi utför ett **försök** ("trial"): kastar ett mynt, kastar en tärning.
- Observerar ett **utfall** ("outcome"): krona, en 6:a
- **Utfallsrummet** är alla möjliga utfall som kan inträffa (S). För en tärning: $S = \{1, 2, 3, 4, 5, 6\}$
- En **händelse** ("event") är en mängd av utfall.
- **Ex:** Två 6:or på två tärningskast, vi kan kalla detta händelsen $A = \{(6, 6)\}$.

Sannolikhet - exempel på en (annan) händelse

- Händelsen A - få summan 7 på två tärningskast
- $A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$

						
	2	3	4	5	6	7
	3	4	5	6	7	8
	4	5	6	7	8	9
	5	6	7	8	9	10
	6	7	8	9	10	11
	7	8	9	10	11	12

- Vad är sannolikheten att få en 6:a (händelsen $A = \{6\}$)?
- Att få en 6:a är ett av sex möjliga utfall
- **Matematisk (logisk) sannolikhet** Om tärningen är helt symmetrisk är varje utfall lika sannolikt

$$P(A) = \text{Sannolikhet för 1 av 6 utfall} = \frac{1}{6}$$

- **Empirisk sannolikhet:** andelen 6:or om jag kastar tärningen ett "oändligt" (mycket stort) antal gånger

$$P(A) = \frac{\text{Antal 6:or}}{\text{Antal försök}}$$

- **Subjektiv sannolikhet** Min tidigare erfarenhet av tärningskast och min uppfattning om en tärnings symmetri säger mig att min sannolikhet att få en 6:a är $1/6$.

- **Övning 1** - vi gör sannolikhetsberäkningar (SDM, kap. 12)

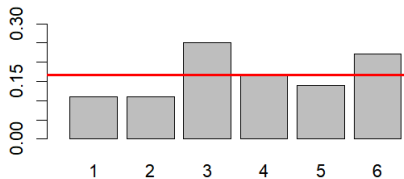
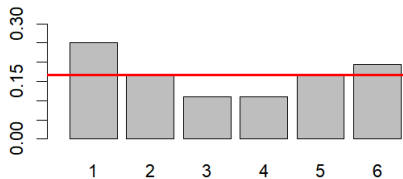
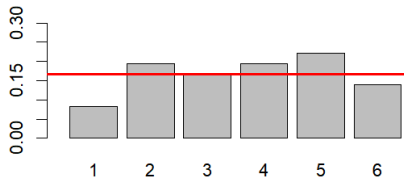
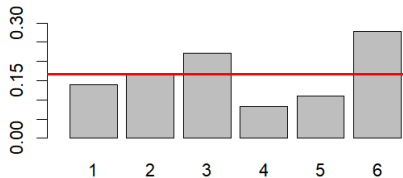


Tärningskast och slumpdragningar - exempel

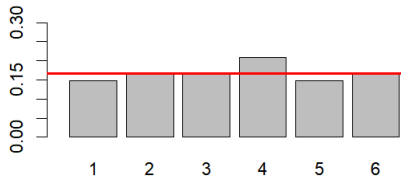
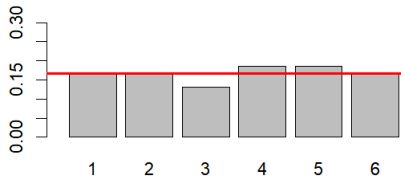
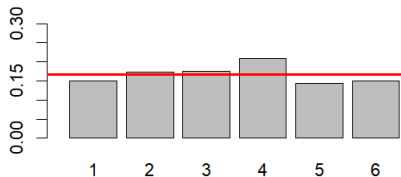
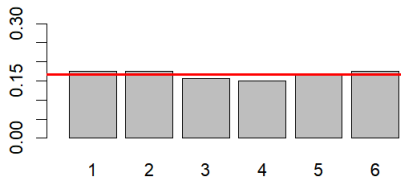
- Exempelen som följer hjälper oss (förhoppningsvis) att förstå:
 - De stora talens lag
 - Begreppet samplingfördelning
 - Centrala gränsvärdessatsen
 - Normalfördelningen



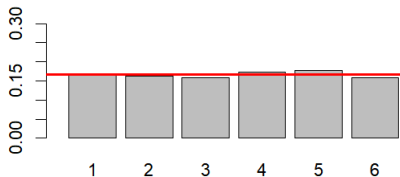
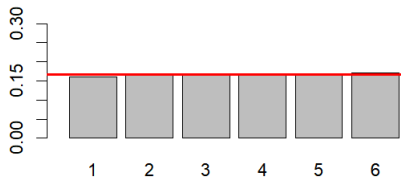
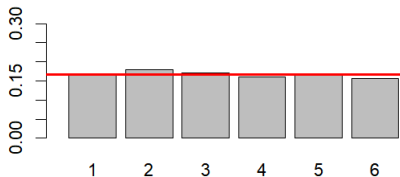
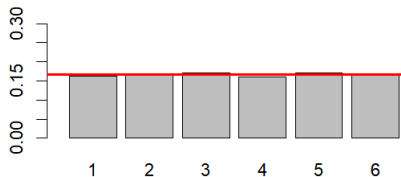
Andel 1:or, 2:or, osv. **36** tärningskast, 4 gånger (röd linje - $1/6$).



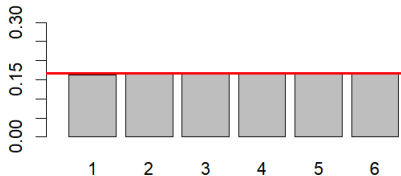
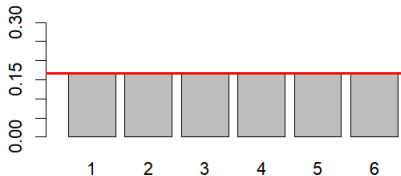
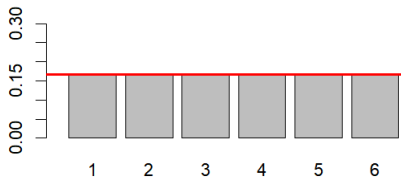
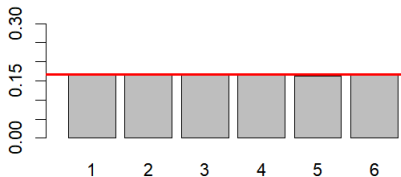
Andel 1:or, 2:or, osv. **360** tärningskast, 4 gånger (röd linje - $1/6$).



Andel 1:or, 2:or, osv. **3600** tärningskast, 4 gånger (röd linje - $1/6$).



Andel 1:or, 2:or, osv. **36000** tärningskast, 4 gånger (röd linje - $1/6$).

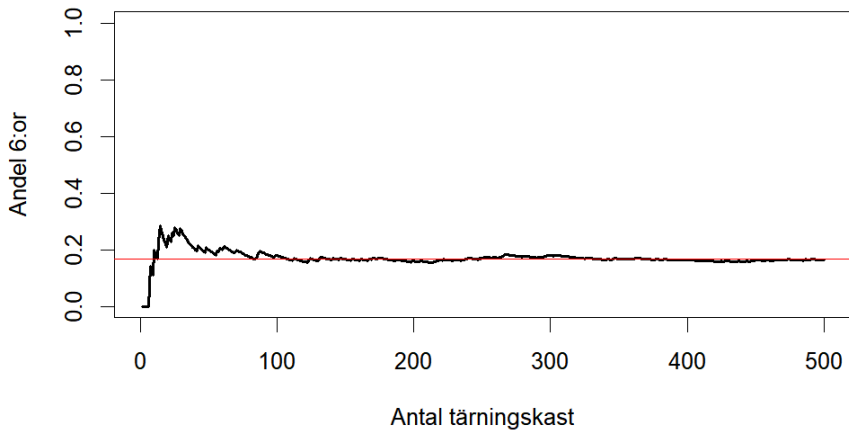


Stora talens lag ("Law of large numbers")

- **Stora talens lag:** När vi gör tillräckligt många försök, med **oberoende** händelser, konvergerar andelen "lyckade" försök (exv: en 6:a på en tärning) till ett bestämt tal (sannolikheten)
- **Oberoende** händelser (tärningar) - vad som kommer upp på tärningen i ett visst tärningskast beror inte på vad som kommer upp i andra kast
- **Oberoende** händelser (generellt) - händelserna A och B är oberoende om vetskapen att B har inträffat inte påverkar sannolikheten för A. Och vice versa.



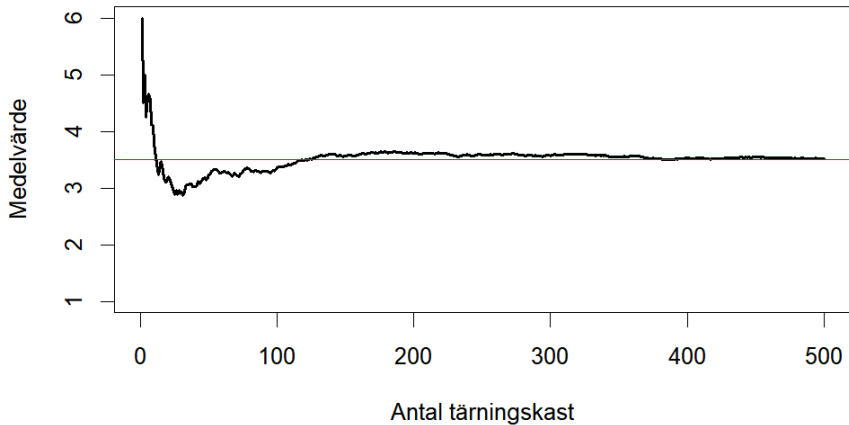
Utveckling av andel 6:or, 500 tärningskast



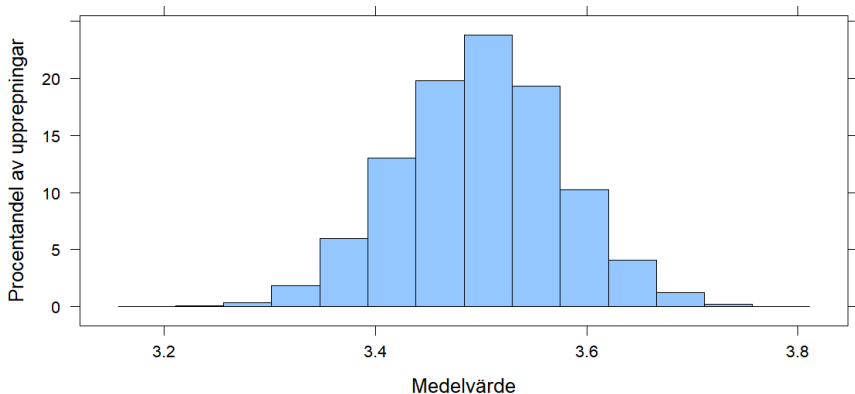
Samplingfördelning ("Sampling distribution")

- Om vi kastar en tärning ett visst antal gånger, exv. 500, och repeterar förfarandet väldigt många gånger
- Vi skulle exempelvis kunna få
 - 80 var av 1-5, och 100 6:or
 - 100 1:or, 80 var av 2-6
 - 84 var av 1-5, och 80 6:or
 - 99 var av 1-5, och fem 6:or (extremt osannolikt, men teoretiskt möjligt)
 - en 1:a och 499 6:or (ännu mer osannolikt, men teoretiskt möjligt)
 - det finns en enorm mängd möjliga kombinationer...
- Fördelningen av medelvärdena, från alla möjliga kombinationer, kallas **samplingfördelningen**
- I praktiken kan vi inte ta fram alla kombinationer, istället slumpar vi ett stort antal gånger, exv. 10000 gånger, 500 tärningskast

Utveckling av medelvärdet, 500 tärningskast



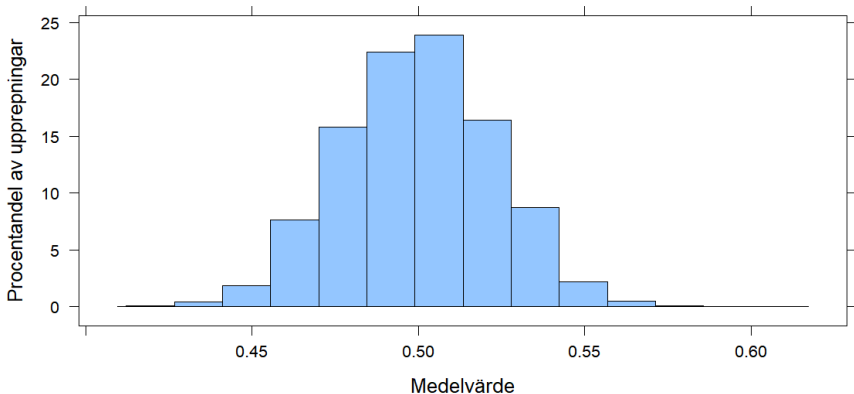
Fördelning av medelvärde vid 10000 upprepningar av 500 tärningskast



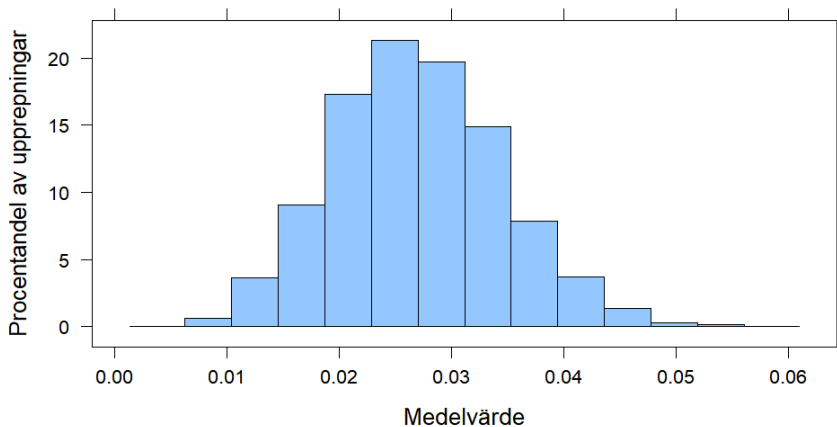
Andra exempel på samplingfördelningar

- Slantsingling
- Andel "Huset vinner" i roulette (nummer 0-36, 37 nummer, huset vinner på nollan)
- Fler exempel kan ges - tid mellan olika individers ankomst till ett myndighetskontor, medelantal gäster på restaurang, antal tillverkningsfel per dag, andel socialdemokrater i opinionsundersökning, osv. (under antagande om oberoende)

Fördelning av medelvärde vid 10000 upprepningar av 500 slantsinglingar



Fördelning av andelen "huset vinner" vid 10000 upprepningar av 500 roulettespel (europeisk variant)



Centrala gränsvärdessatsen, normalfördelningen

- Graferna ser väldigt lika ut (om än på olika skala)
- Fördelningarna liknar alla **normalfördelningen**
- Detta beror på **Centrala Gränsvärdessatsen**:

Samplingfördelningen av medelvärden av **oberoende försök (observationer)**, från vilken fördelning som helst, går mot en normalfördelning (om antal försök är tillräckligt stort, och utfallen är ändliga tal.)

- Det spelar mao. ingen roll vilken fördelning data i sig har, så länge observationerna är oberoende (och ändliga) kommer medelvärdesfördelningen från olika slumpmässiga urval att gå mot en normalfördelning
- Exempelfördelningar på tavlan



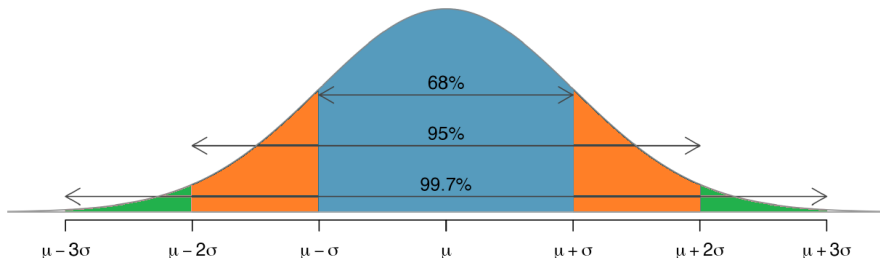
Normalfördelningen ("normal distribution")

- En sannolikhetsfördelning som är symmetrisk kring medelvärdet.
- Definieras av två parametrar:
 - Medelvärde (μ): Det centrala värdet.
 - Standardavvikelse (σ): Mått på spridning.
- Symmetrisk, klockformad kurva.
- Medelvärde, median och typvärde är lika.
- Vi kommer träffa på normalfördelningen igen på föreläsning 5 och 6, bland annat när vi ska dra slutsatser från statistiska undersökningar (inferens)
- Gå igenom i kapitel 13



Mer om normalfördelningen

- Följer 68-95-99.7-regeln:
 - 68% av värdena ligger inom 1σ från medelvärdet.
 - 95% inom 2σ .
 - 99.7% inom 3σ .



Standardisering av Normalfördelningen

- Av olika skäl är det praktiskt att "standardisera" normalfördelningen
 - Vi vill kunna jämföra fördelningar, där exempelvis ett dataset har variabler i SEK och ett annat i USD
 - Oberoende av skala på vårt data vill vi ta reda hur sannolikt ett visst värde är, för detta används standardiserade tabeller/R-kommandon
 - Vi arbetar därför med en fördelning centrerad kring 0 och med standarvavvikelse 1 (**standardnormalform**)
- Omvandla till så kallade Z-värden.

$$Z = \frac{X - \mu}{\sigma}$$



Exempel på standardisering

- Givet: $x = 75, \mu = 70, \sigma = 5$

- Beräkna z-värde:

$$z = \frac{75 - 70}{5} = 1$$

- Tolkning:

- Värdet 75 är 1 standardavvikelse över medelvärdet.
- Enligt en Z-tabell (föreläsning 5, 6) är sannolikheten för att $X \leq 75$ lika med 0.8413 (eller 84.13%).

Denna version av dokumentet: 2025-03-28

Materialet i Statistisk översikt kurs har tagits fram av Ulf Högnäs och Anders Fredriksson, med inspiration och ibland direkt användande av material från andra kurser och personer, bland annat kurserna Statistik och dataanalys 1-3, med material av Michael Carlson, Ellinor Fackle Fornius, Jessica Franzén, Oskar Gustafsson, Oscar Oelrich, Mona Sfaxi, Karl Sigfrid, Mattias Villani, med flera.

