Statistik med tillämpningar i biologin

Grundläggande begrepp

Innehåll

1) Slumpvariabel
2) Utfallsrum
3) Fördelning
4) Population och stickprov
5) Inferens
6) Hypotes och Test
7) Sampling

Slumpvariabel (Stokastisk variabel)

- Resultat av ett slumpförsök (utfall) – utgången kan inte kontrolleras
- Resultatet kan inte förutspås, men vi vet (ofta) mer än ingenting:
  - Utfallsrum Ω - *alla möjliga resultat av försöket*
  - Fördelning - *vilken sannolikhet för vilket resultat*

Utfallsrum - alla möjliga utfall

diskret

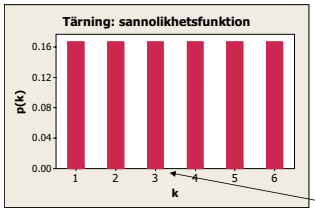
- X = resultatet av en tärningskast: Ω_X={1,2,3,4,5,6}
- Y = antalet 5:or i 100 tärningskast: Ω_Y={0,1,…100}
- Z = antalet kast tills det blir tre 5:or i råd: Ω_Z={3,4,…∞}

- X = livslängd: Ω={0, ∞} (?)
- Y = vikt av en tillfälligt utvald svensk man
- Z = omkrets av ett tillfälligt utvalt träd i en skog

kontinuerligt

Fördelning

- Med vilken sannolikhet får jag vilket resultat?
- Diskreta slumpvariabler: **sannolikhetsfunktion** (pdf[1])



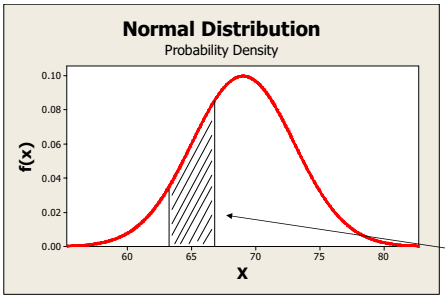
diskret

$P(X=3)=\frac{1}{6}$

[1] probability density function

Fördelning

- Med vilken sannolikhet får jag vilket resultat?
- Kontinuerliga slumpvariabler: **täthetsfunktion** (pdf)

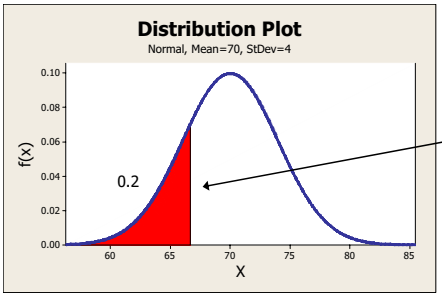


kontinuerligt

$P(63 < X \leq 67) = area$

Fördelningsfunktion (cdf[1])

- Anger sannolikheten att utfallet blir *mindre än (eller lika med)* ett visst värde.

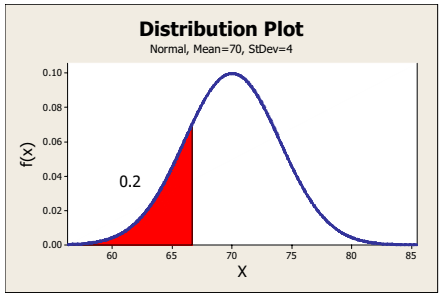


$$P(X \le 67) = area = F_X(67)$$

[1] cumulative distribution function

Fördelningsfunktion

- argument = Ω, funktionsvärden = arean



$67 \rightarrow 0.2 \quad F(67) = 0.2$

$70 \rightarrow 0.5 \quad F(70) = 0.5$

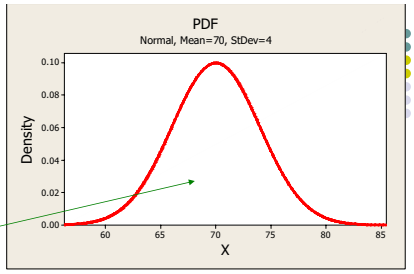
$73 \rightarrow 0.8 \quad F(73) = 0.8$

$F(-\infty) = 0$

$F(\infty) = 1 \quad hela\ arean = 1$

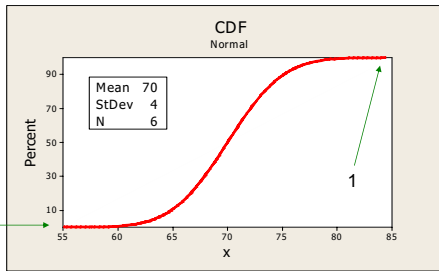
Täthetsfunktion

arean = 1



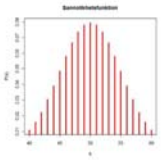
Fördelnings-funktion

0



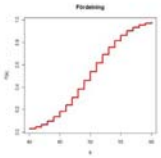
Diskreta slumpvariabler

$$F_X(x) = P(X \le x) = \sum_{k \le x} p_X(k)$$

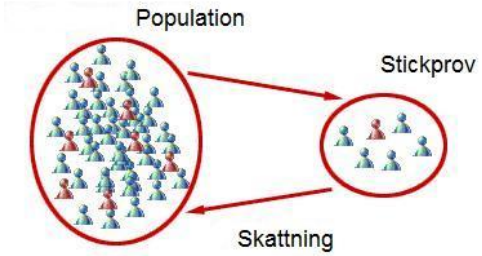


Kontinuerliga slumpvariabler

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt$$



Population och stickprov



Inferens

- dra slutsatser om en population utifrån ett stickprov:
  - Vikten av 100 st. sill → vikten av sill (i hela östersjön)
  - Längden av 50 tonåringar → längden av tonåringar i Sverige
  - Medikament: blodtryck hos 30 pers → blodtryck hos alla människor
- uttalanden om egenskaper för en hel population utifrån beräknade värden på ett urval av denna population
- inte möjlig att uttala sig med fullständig säkerhet om en population → *konfidens, signifikans*

Inferens

Inferens statistik är metoder för att dra slutsatser om en population utifrån ett urval. Inferens statistik kallas också för induktiv statistik eller induktion. Inferens statistik innebär att vi uttalar oss om egenskaper för en hel population utifrån beräknade värden på ett urval av denna population. Inferens statistik innebär att vi inte med fullständig säkerhet kan uttala oss om en population, vi kan dock göra väl underbyggda uttalanden om populationen. Att göra statistiska undersökningar på hela populationen blir oftast alldeles för dyrt då en population kan bestå av flera miljoner objekt (observationer). Inferens statistik är en viktig vetenskap för att kunna göra väl underbyggda uttalanden om en population till en rimlig kostnad.

Hypoteser och test

- En hypotes angående (en parameter av) en population ställs upp: p=½
- Hypotesen testas:

1. Ta ett stickprov från denna population:

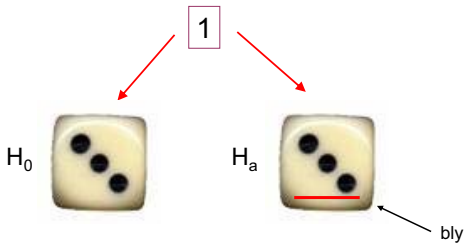
2. Beräkna om stickprovet ”går ihop” med hypotesen (är stickprovet ”sannolikt” när man antar att hypotesen stämmer?):

p=0.5 ⟶ P(8 x framsidan) = 0.00391

3. Förkasta hypotesen om stickprovet ”är osannolikt” → alternativa hypotesen gäller

Hypotes

- H_0: en tärning är inte manipulerad



När förkastas en hypotes ?

H_0: tärningen är inte manipulerad

Stickprov: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
→ inte omöjligt, men mycket osannolikt (intuition)
→ H_0 förkastas: H_a: tärningen *är* manipulerad

Stickprov: 1 1 2 1 1 6 5 1 1 4 1 1 1 4 2 1 1 3 1 1 1 5 5 6 1
→ P = 0,000001, mycket osannolikt *(Binomialfördelning)*
→ H_0 förkastas också

Stickprov: 1 3 2 1 3 2 5 2 1 4 2 1 3 4 1 6 1 3 5 2 1 5 5 6 4
→ P = 0,064, inte så osannolikt *(Binomialfördelning)*
→ H_0 förkastas inte – resultatet är inte **signifikant**
→ vi utgår ifrån att tärningen är inte manipulerad

Exempel

- Hela populationen (genomsnitt):
  - 2.9 av 100000 personer/år insjuknar i leukemi
- En stad nära ett kärnkraftverk:
  - 4.1 per 100000 personer insjuknar årligen i leukemi
- H_0: kärnkraftverket har *inte* något inflytande på insjuknandet
  - Hur sannolikt är en sådan avvikning från genomsnittet?
  - Är avvikningen mycket osannolikt (kan inte förklaras med slumpen) → förkasta H_0
- Vi måste kunna beräkna denna sannolikhet !

Insamling av data (sampling)

- Storlek av stickprovet (sample size)
  - Beror på sprindningen i populationen
- Två stickprov: $n_1 = n_2$
- Regression: förmånlig fördelning av mätpunkterna
- Parade observationer vs. oberoende grupper

Experimental design

- Completely randomized
- Ramdomized block design
- Latin square design