

TMS136

Föreläsning 7

Stickprov

- När vi pysslar med statistik handlar det ofta om att baserat på stickprovsinformation göra utlåtanden om den population stickprovet är draget ifrån
- Ett stickprov som inte ännu är observerat är en uppsättning oberoende och likafördelade s.v. $X_1, \dots, X_n$
- Ett observerat stickprov är en uppsättning tal och betecknas $x_1, \dots, x_n$
- Att beskriva ett observerat stickprov med olika mått (läge, spridning,...) kallas att göra *deskriptiv statistik*

Stickprovsmedelvärde

- Om vi observerar $x_1, \dots, x_n$ från en population väntevärde μ kan vi "skatta" μ med (det observerade) *stickprovsmedelvärdet*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Lite längre fram i kursen ska vi formalisera vad vi menar med "skatta"

Stickprovsvarians

- Om vi observerar $x_1, \dots, x_n$ från en population varians σ^2 kan vi skatta σ^2 med (den observerade) *stickprovsvariansen*

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Om vi "tar roten ur" får en skattning av (populations)standardavvikelsen σ
- Man kan undra varför det står $n - 1$ och inte n i nämnaren men längre fram ska vi se att detta ger oss en önskvärd egenskap...

Beräkningsformel

- Det är en övning att visa att

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2}{n-1}$$

Ändliga populationer

- Oftast är populationerna vi vill uttala oss om mycket större än de stickprov vi tar och då är det ingen mening att fundera på precis hur stora populationerna är
- Om vi (till äventyrs) kan observera hela populationen som här antas innehålla N objekt har vi alltså $x_1, \dots, x_N$ och då är

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad \text{och} \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Ett annat spridningsmått

- Varians är ju ett spridningsmått
- Ett annat spridningsmått som används i bland annat kvalitetsstyrning är *variationsbredd (range)*
- Om vi observerat $x_1, \dots, x_n$ är variationsbredden

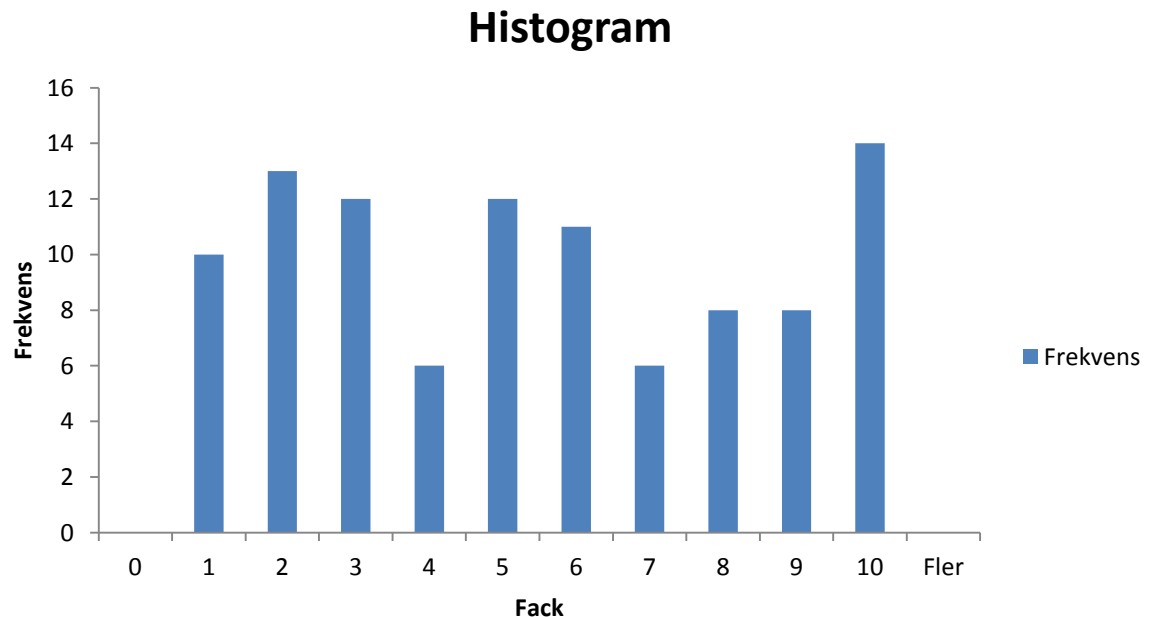
$$r = \max(x_i) - \min(x_i)$$

Histogram

- "Alla" har sett ett histogram, men vad är det för något?
- Vi observerar $x_1, \dots, x_n$
- Vi delar in x-axeln i ett antal lika breda intervall (eller fack) sådana att våra observationer ryms inom intervallen
- För varje intervall gör vi en stapel vars höjd bestäms av antalet/andelen observationer som hamnar i intervallet

Histogram exempel

- Vi har genererat 100 observationer från en kontinuerligt likformig fördelning på intervallet (0,10)



Sannolikhetsplot

- En *sannolikhetsplot* är ett grafiskt verktyg som används för att avgöra om observerade data genererats från en viss fördelning
- Många statistiska metoder förutsätter att man har normalfördelade data och för att avgöra om de data man "fått" kommer från en normalfördelning är en sannolikhetsplot ett av många användbara verktyg

Att konstruera en sannolikhetsplot

- Vi observerar $x_1, \dots, x_n$
- Vi ordnar observationerna i storleksordning från den minsta till den största $x_{(1)}, \dots, x_{(n)}$
- Sedan ”plottar” vi paren $(x_{(j)}, (j - 0.5)/n)$ på ett *sannolikhetspapper* för den föreskrivna fördelningen
- Om observationerna kommer från den föreskrivna fördelningen hamnar punkterna (någorlunda) längs en rät linje

Sannolikhetspapper

- Ett sannolikhetspapper är egentligen ett koordinatssystem där man delat in y-axeln på ett sätt som är representativt för den föreskrivna fördelningen
- Antag att vi vill kolla om observationerna kan anses normalfördelade

Sannolikhetspapper?

- Om vi kikar i standard normalfördelningstabellen i bokens appendix ser vi att $z = -0.67$ ger $\Phi(z) = 0.25$ och att $z = 0.67$ ger $\Phi(z) = 0.75$
- Som tumregel har vi att den räta linjen (längs vilken observationerna ska ligga för att anses normalfördelade) ska approximativt gå genom de punkter som delar av 25% och 75% av våra data

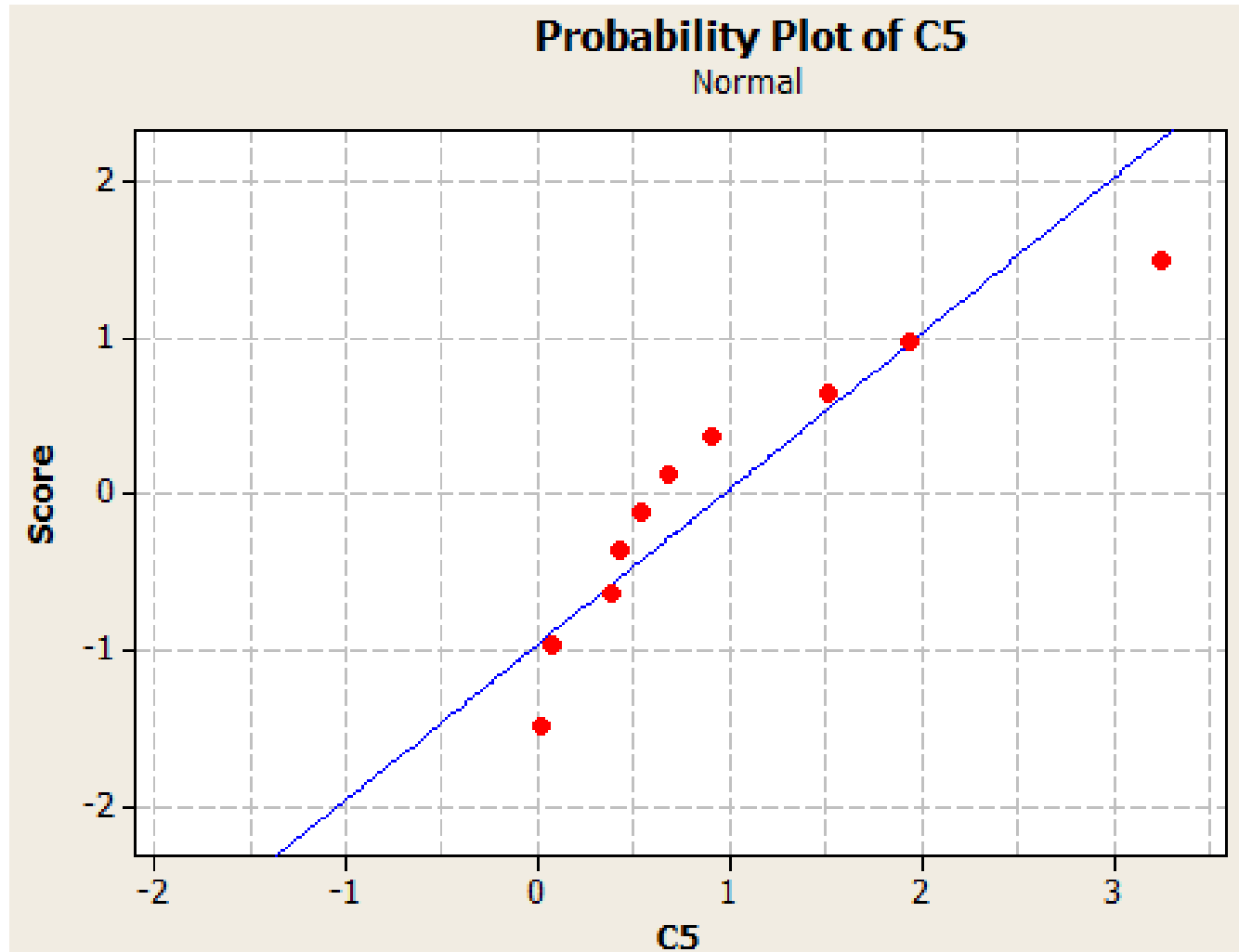
Sannolikhetspapper?

- Så om vi har tio observationer drar vi linjen så att den skär $(x_{(3)}, -0.67)$ och $(x_{(8)}, 0.67)$
- För att få y-koordinaterna för punkterna som ska placeras ut standardiserar vi observationerna

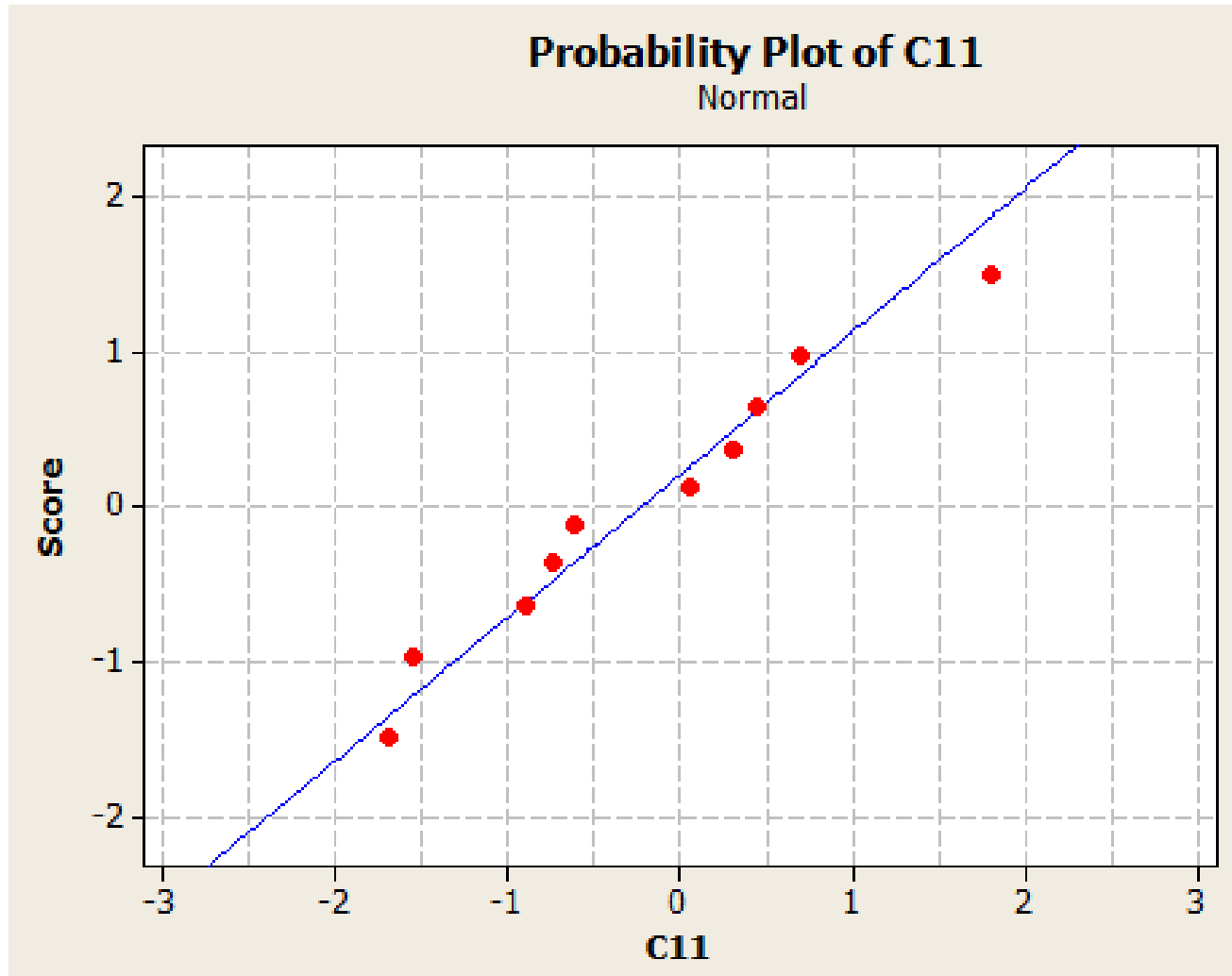
$$y_{(i)} = \frac{x_{(i)} - \bar{x}}{s}$$

- Vi placerar ut punkterna $(x_{(i)}, y_{(i)})$ i koordinatssystemet och om de ligger längs den rätta linjen är det rimligt att anta att data är normalfördelade

Så här ska det inte se ut

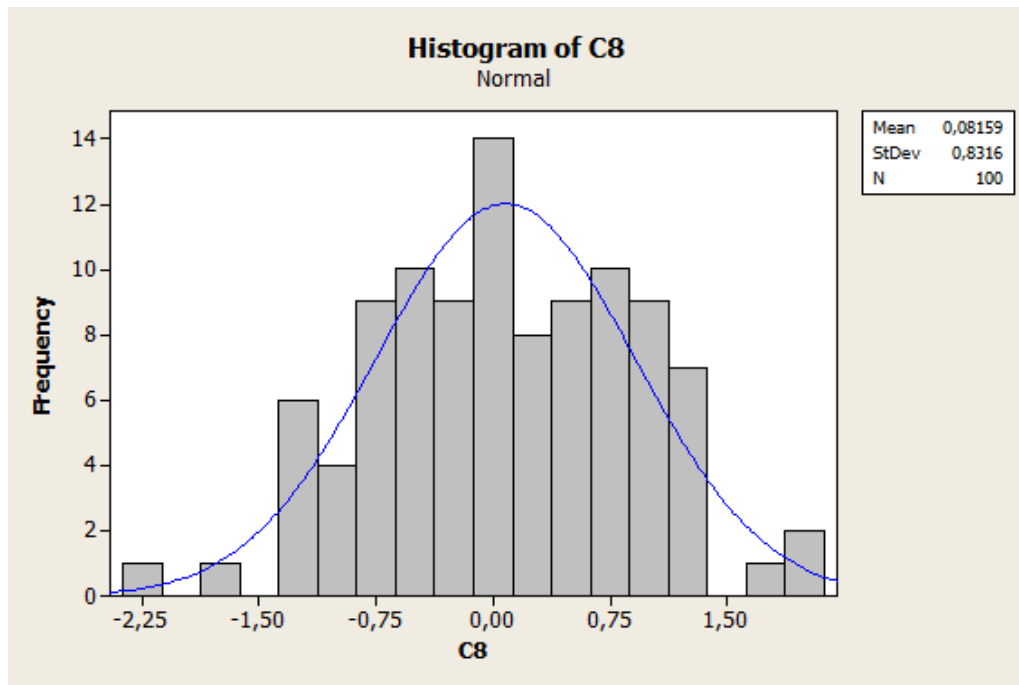


Men så här är bättre



Ett annat sätt att kolla

- För att förvissa sig om att observerade data kommer från en normalfördelning vill man se att ett histogram (med lämpligt valda intervall/fack) liknar "klockan"



Centrala gränsvärdessatsen

- *Centrala gränsvärdessatsen (CGS, Central limit theorem)* är ett remarkabelt resultat som under vissa snälla antaganden ger att stickprovsmedelvärdet är approximativt normalfördelat om stickprovet är någorlunda stort
- Detta betyder att man kan räkna sannolikheter för ett stickprovsmedelvärde utan att veta vilken fördelning stickprovet kommer ifrån

CGS

- Låt $X_1, \dots, X_n$ vara oberoende och likafördelade med $E(X_i) = \mu$ och $SD(X_i) = \sigma$
- Då gäller för stickprovsmedelvärdet

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

att

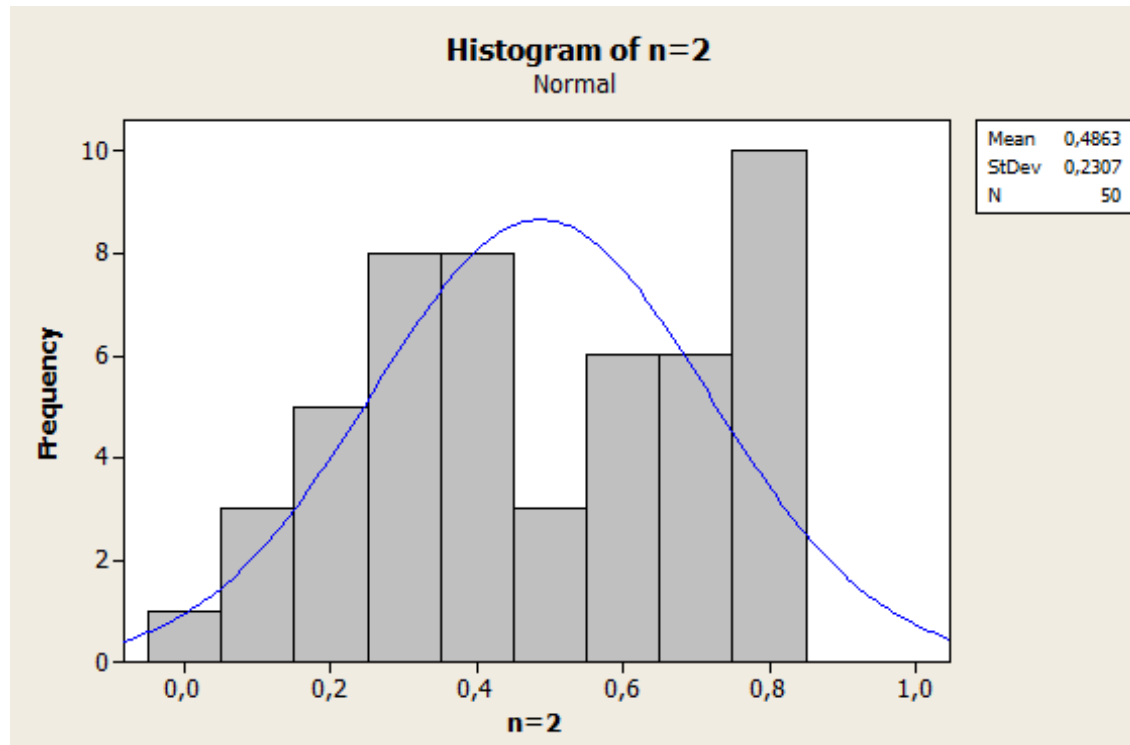
$$\lim_{n \rightarrow \infty} P \left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq x \right) = \Phi(x)$$

CGS

- På FL6 hade vi ett exempel där i räknade på en sannolikhet för medelfyllningen av läskburkar under antagande om normalfördelad påfyllning
- Med CGS räcker det att veta väntevärdet och variansen för stickprovsmedlemmarna för att kunna räkna på sannolikheter för stickprovsmedelvärdet...

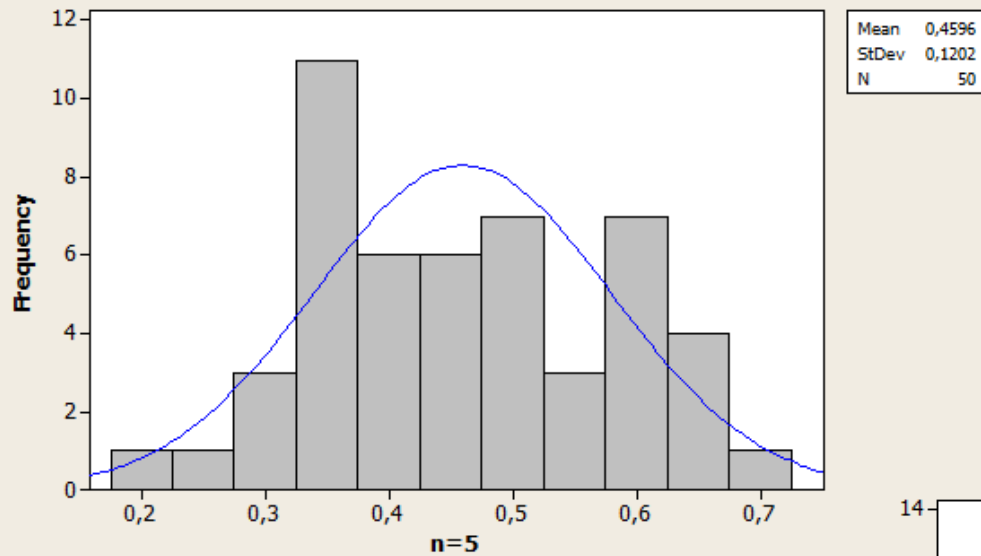
CGS

- Histogrammen nedan visar (i varje diagram) 50 observationer av stickprovsmedelvärden baserade på 2, 5, 10, 20 och 50 observationer av kontinuerligt likformigt fördelade (på (0,1)) S.V.

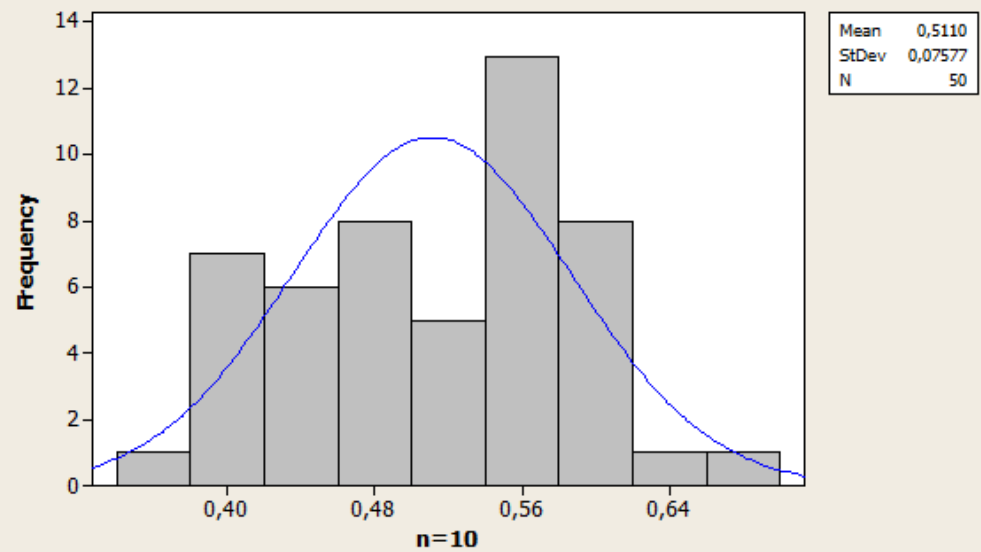


CGS

Histogram of n=5
Normal

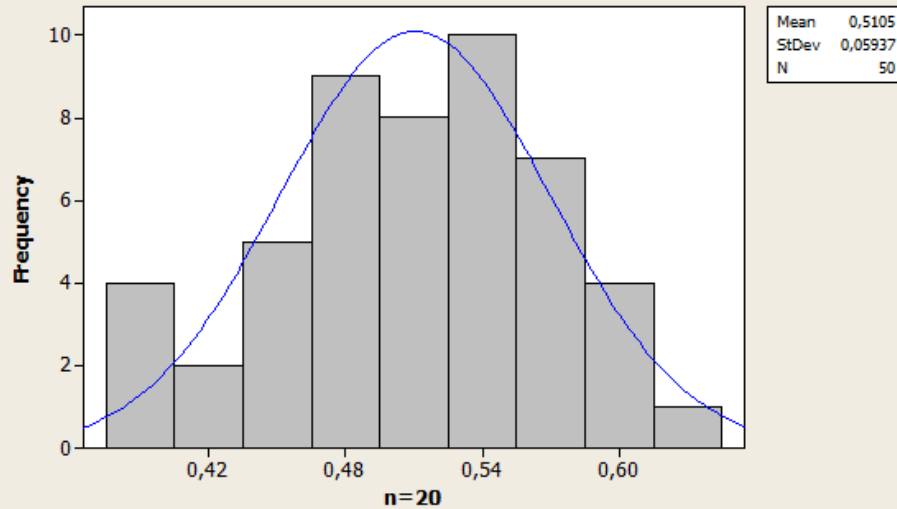


Histogram of n=10
Normal

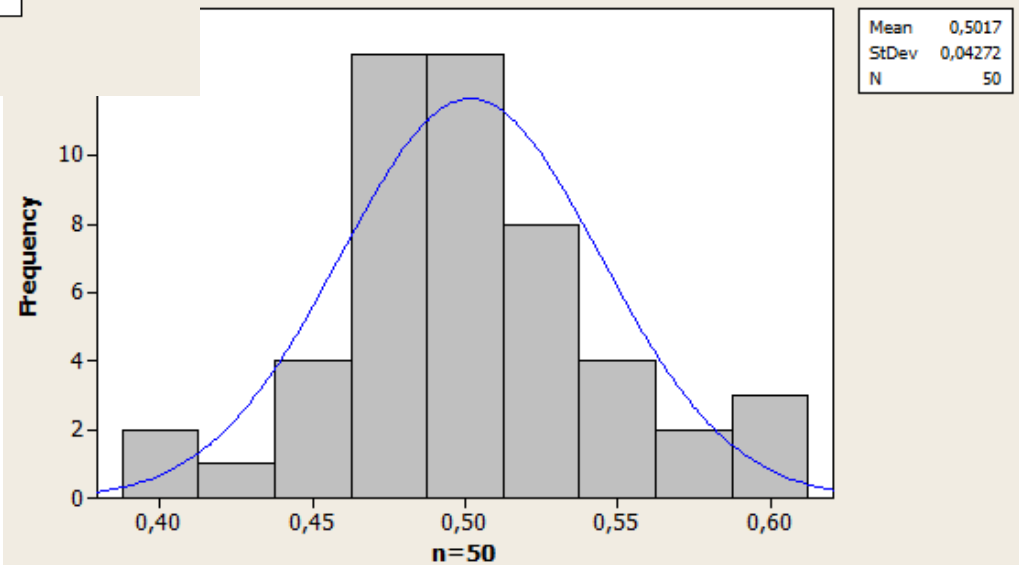


CGS

Histogram of n=20
Normal



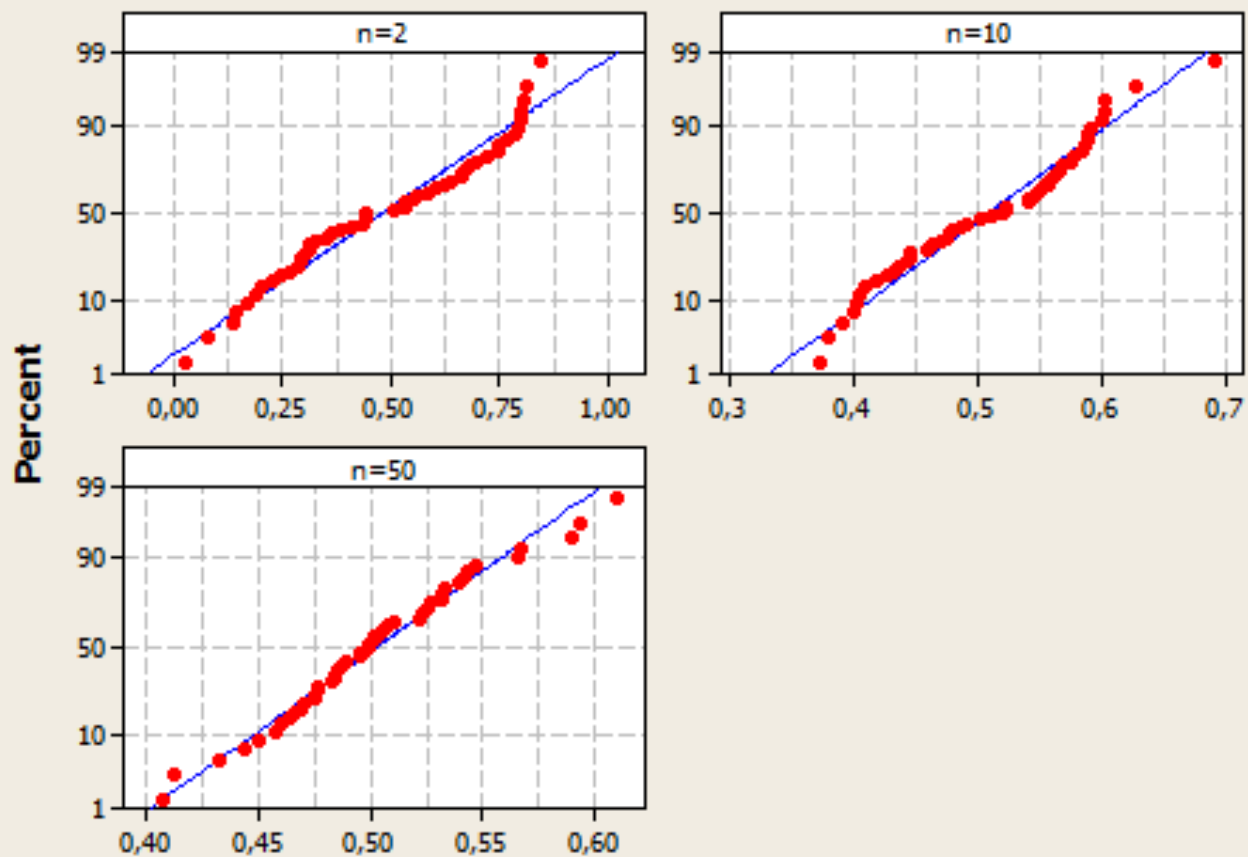
Histogram of n=50
Normal



CGS

Probability Plot of $n=2$; $n=10$; $n=50$

Normal



Följder av CGS

- Vi har tidigare sett att man kan approximera binomialfördelning med normalfördelning om parametern n i binomialfördelningen är tillräckligt stor
- Att detta funkar är en följd av CGS eftersom vi kan betrakta en binomialfördelad s.v. som en summa av oberoende och likafördelade "noll-ett-värda" Bernoulli-variabler...

Följder av CGS

- Vi har också sett att man kan approximera Poissonfördelning med normalfördelning om parametern λ i Poissonfördelningen är tillräckligt stor
- Att detta funkar är en följd av CGS eftersom vi kan betrakta en Poissonfördelad s.v. med ett stort λ som en summa av oberoende Poissonfördelade alla med samma "lilla" λ

Följder av CGS

- Vi formulerade ju ”slutklämmen” i CGS som

$$\lim_{n \rightarrow \infty} P \left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq x \right) = \Phi(x)$$

- Detta betyder ju även att $\bar{X}$ är approximativt normalfördelad med väntevärde μ och standardavvikelse $\sigma / \sqrt{n}$
- Det betyder också att $\sum_{i=1}^n X_i$ är approximativt normalfördelad med väntevärde $n\mu$ och standardavvikelse $\sqrt{n}\sigma$

Exempel

- När man tillverkar en viss typ av spik är väntevärdet för massan hos en spik 30g och standardavvikelsen är 1g
- Vad är sannolikheten att 200 spikar tillsammans inte väger mer än 5990g om vi antar att spikarnas massor är oberoende av varandra?

Exempel forts

- Låt X_i för $i = 1, \dots, 200$ vara massan av en spik. Vi har att $\mu = E(X_i) = 30$ och $\sigma = SD(X_i) = 1$
- Massan för 200 spik är $\sum_{i=1}^{200} X_i$ och det följer att $E\left(\sum_{i=1}^{200} X_i\right) = 200 \cdot 30 = 6000$ och att $SD\left(\sum_{i=1}^{200} X_i\right) = \sqrt{200} \cdot 1 = \sqrt{200}$

Exempel forts

- Vi har då att

$$\begin{aligned} & P\left(\sum_{i=1}^{200} X_i \leq 5990\right) \\ &= P\left(\frac{\sum_{i=1}^{200} X_i - 6000}{\sqrt{200}} \leq \frac{5990 - 6000}{\sqrt{200}}\right) \\ &\approx \Phi(-0.71) \approx 0.24 \end{aligned}$$

Viktigt

- Massan för 200 spik kan **INTE** skrivas $200X_i$!
- $SD(200X_i) = 200$ men $SD(\sum_{i=1}^{200} X_i) = \sqrt{200}$
så $200X_i$ har inte (ens) samma fördelning som $\sum_{i=1}^{200} X_i \dots$