

Anvendt Statistik

Lektion 2

Sandsynlighedsregning

Sandsynlighedsfordelinger

Normalfordelingen

Stikprøvefordelinger

Sandsynlighed: Opvarmning

■ Udfald

- Resultatet af et "eksperiment" kaldes et udfald.

■ Eksempler:

- Eksperiment: Vælg en partileder / mål lysets hastighed
- Udfald: Lars / 299791 km/s

■ Hændelse

- En hændelse er en mængde af udfald.

■ Eksempler:

- Vælge en kvinde / Hastighedsmåling er ml. 299790 km/t og 299793 km/t

Sandsynlighed

■ Sandsynlighed

- Sandsynligheden for en hændelse A er andelen af gange eksperimentet resulterer i hændelsen A i det lange løb.

■ Notation

- $P(A)$ betegner sandsynligheden for hændelsen A .

■ Eksempel

- Eksperiment: Kast med en fair mønt
- $P(\text{Plat}) = 0.5$
- I det længe løb er halvdelen af møntkastene plat.

Sandsynlighed: Egenskaber og regneregler

1) $0 \leq P(A) \leq 1$

- $P(A) = 0$ - hændelsen A indtræffer aldrig.
- $P(A) = 1$ - hændelsen A indtræffer hver gang.

2) $P(\text{ikke } A) = 1 - P(A)$

- Hvis A ikke indtræffer, så må "ikke A " nødvendigvis indtræffe

3) Hvis hændelserne A og B ikke kan indtræffe samtidigt gælder:

$$P(A \text{ eller } B) = P(A) + P(B)$$

Mere om sandsynlighed

□ Betinget sandsynlighed

Hvis A og B er mulige udfald, så gælder

$$P(B \text{ givet } A) = \frac{P(A \text{ og } B)}{P(A)}$$

Hvilket kan omskrives til $P(A \text{ og } B) = P(A)P(B \text{ givet } A)$

□ Uafhængighed

To hændelser A og B er **uafhængige** hvis og kun hvis

$$P(A \text{ og } B) = P(A)P(B)$$

hvilket kan omskrives til $P(B \text{ givet } A) = P(B)$

Eksempler

- Eksperiment: Vælg en tilfældig mand.
 - Hændelse A : Den udvalgte er kortere end 170cm
 - Hændelse B : Den udvalgte er længere end 180cm
- $P(\text{højden falder ikke i intervallet 170 til 180cm}) = P(A \text{ eller } B) = P(A) + P(B)$

- Eksperiment: Vælg en voksen amerikaner
 - Hændelse A : Personen er gift $P(A) = 0.56$
 - Hændelse B : Personen er meget glad.
 - Sandsynligheden for at en gift person er meget glad er 0.40
- $P(\text{er gift og meget glad}) = P(\text{er gift}) P(\text{meget glad givet er gift}) = 0.56 * 0.40 = 0.22.$

Eksempler

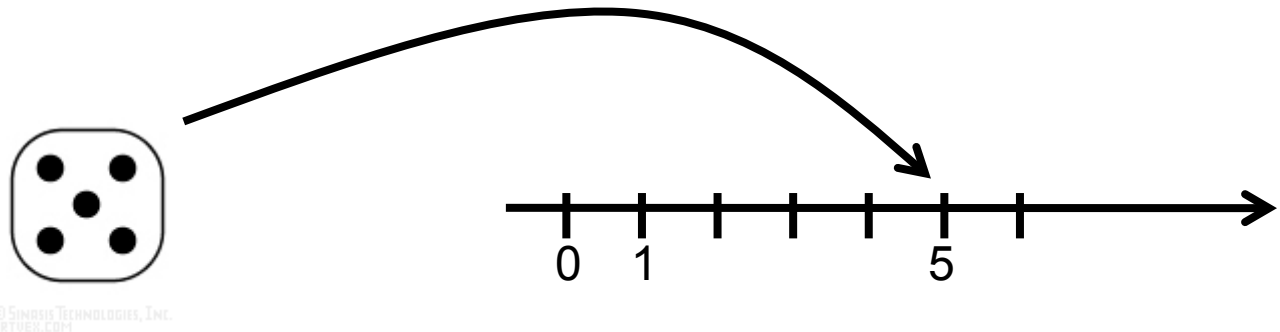
- Eksperiment: Vælg tilfældig studerende
 - A: Personen GEO studerende
 - B: Personen er en mandlig studerende
 - $P(\text{GEO givet Mandlig}) = P(\text{GEO og Mandlig})/P(\text{Mandlig})$
=

- Eksperiment: Kaste to terninger
 - A: Første terning er en 6'er
 - B: Anden terning er en 6'er
 - $P(\text{Slå to 6'ere}) =$

Stokastisk variabel

■ Stokastisk variabel

- Antag vi kan knytte en talværdi til hvert udfald af et eksperiment. Hvert eksperiment fører således til et tilfældigt tal.
- Dette tilfældige tal kaldes en *stokastisk variabel*.



Diskret stokastisk variabel (SV)

- En stokastisk variabel er diskret, hvis den kun kan tage adskilte værdier. Fx 0, 1, 2, 3,...
- Lad $P(k)$ betegne sandsynligheden for at den stokastiske variabel y tager værdien k .
Dvs. $P(1)$ = "sandsynligheden for y tager værdien 1".
- Der gælder
 - $0 \leq P(y) \leq 1$ for alle y .
 - $\sum_{\text{alle } y} P(y) = 1$

Eksempel

- Eksperiment: Kast to terninger
- Lad y være antallet af 6'ere i de to kast

y	$P(y)$
0	
1	
2	
Total	

Middelværdi for diskret SV

- Motivation: Vi har en stikprøve:

Udfald	0	1	2	3	4	5
Antal	1	3	60	23	12	1
Andel	1/100	3/100	60/100	23/100	12/100	1/100

- Gennemsnit

$$\begin{aligned}\bar{y} &= \frac{0 \cdot 1 + 1 \cdot 3 + 2 \cdot 60 + 3 \cdot 23 + 4 \cdot 12 + 5 \cdot 1}{100} \\ &= 0 \cdot \frac{1}{100} + 1 \cdot \frac{3}{100} + 2 \cdot \frac{60}{100} + 3 \cdot \frac{23}{100} + 4 \cdot \frac{12}{100} + 5 \cdot \frac{1}{100}\end{aligned}$$

- I det lange løb


$$\begin{aligned}\mu &= 0 \cdot P(0) + 1 \cdot P(1) + 2 \cdot P(2) + 3 \cdot P(3) + 4 \cdot P(4) + 5 \cdot P(5) \\ &= \sum_{\text{alle } y} y \cdot P(y)\end{aligned}$$

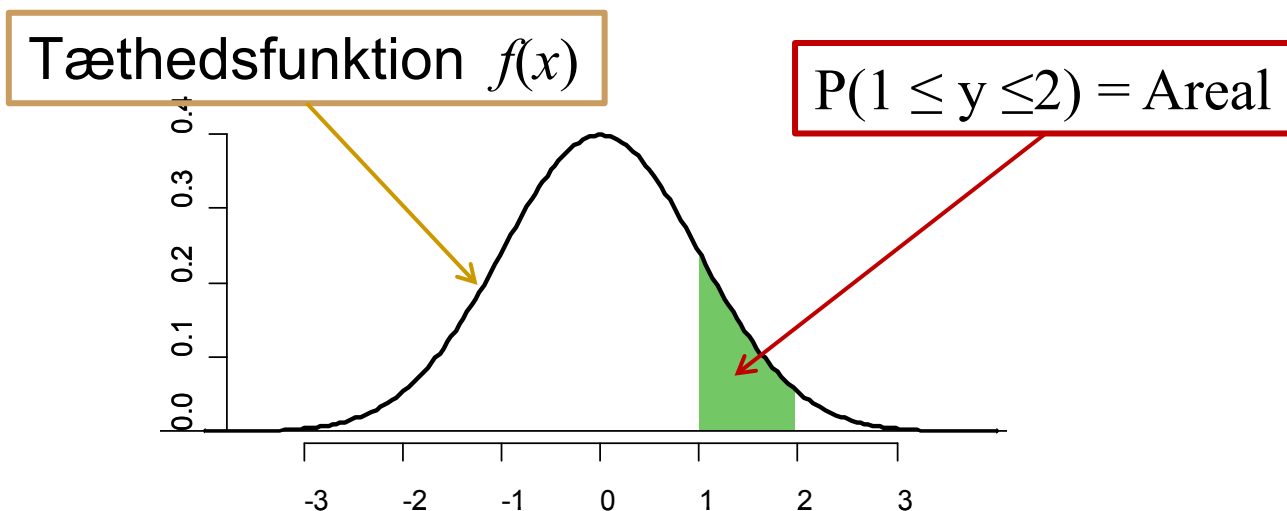
Variansen for diskret SV

- For en diskret stokastisk variabel y med middelværdi μ er variansen:

$$\begin{aligned}\sigma^2 &= \sum_{\text{alle } y} (y - \mu)^2 \cdot P(y) \\ &= (0 - \mu)^2 \cdot P(0) + (1 - \mu)^2 \cdot P(1) + \cdots + (4 - \mu)^2 \cdot P(4) + (5 - \mu)^2 \cdot P(5)\end{aligned}$$

Kontinuert stokastisk variabel

- Hvis y er en kontinuert stokastisk variabel kan den tage et kontinuum af værdier (alle værdier i et interval).
- Vi angiver **sandsynligheden** for at y falder i et interval $[a ; b]$ ved et **areal under en kurve**.



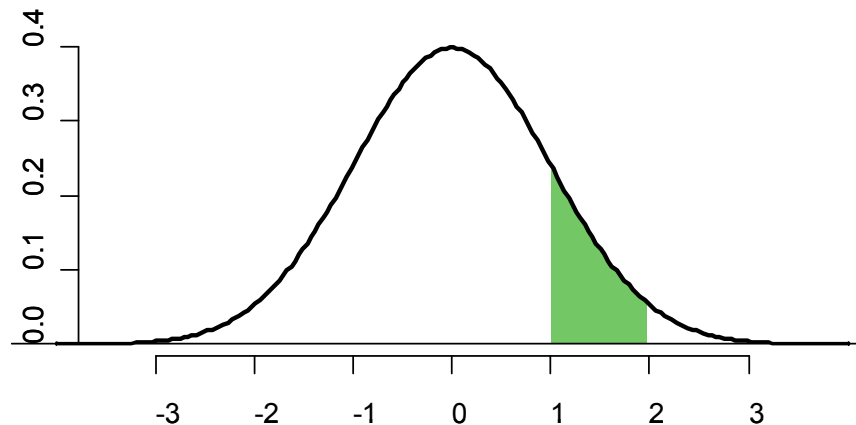
Tæthedsfunktionen

■ (Sandsynligheds)Tæthedsfunktion $f(x)$

1) $f(y) \geq 0$ for alle $y \in \mathbb{R}$

2) $\int f(y)dy = 1$ arealet under kurven f er 1

3) $P(a \leq y \leq b) = \int_a^b f(y)dy$ for $a \leq b$



Normalfordelingen

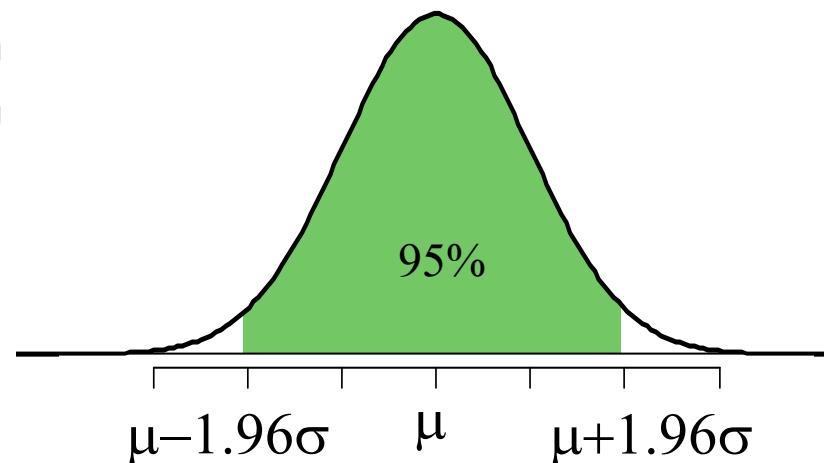
■ Normalfordelingen

- **Klokkeformet** og karakteriseret ved middelværdi μ og standardafvigelse σ .
- **Notation:** $y \sim N(\mu, \sigma^2)$ betyder at y er kontinuert stokastisk variabel, der er normalfordelt med middelværdi μ og varians σ^2 .
- Tæthedsfunktionen for normalfordelingen er

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

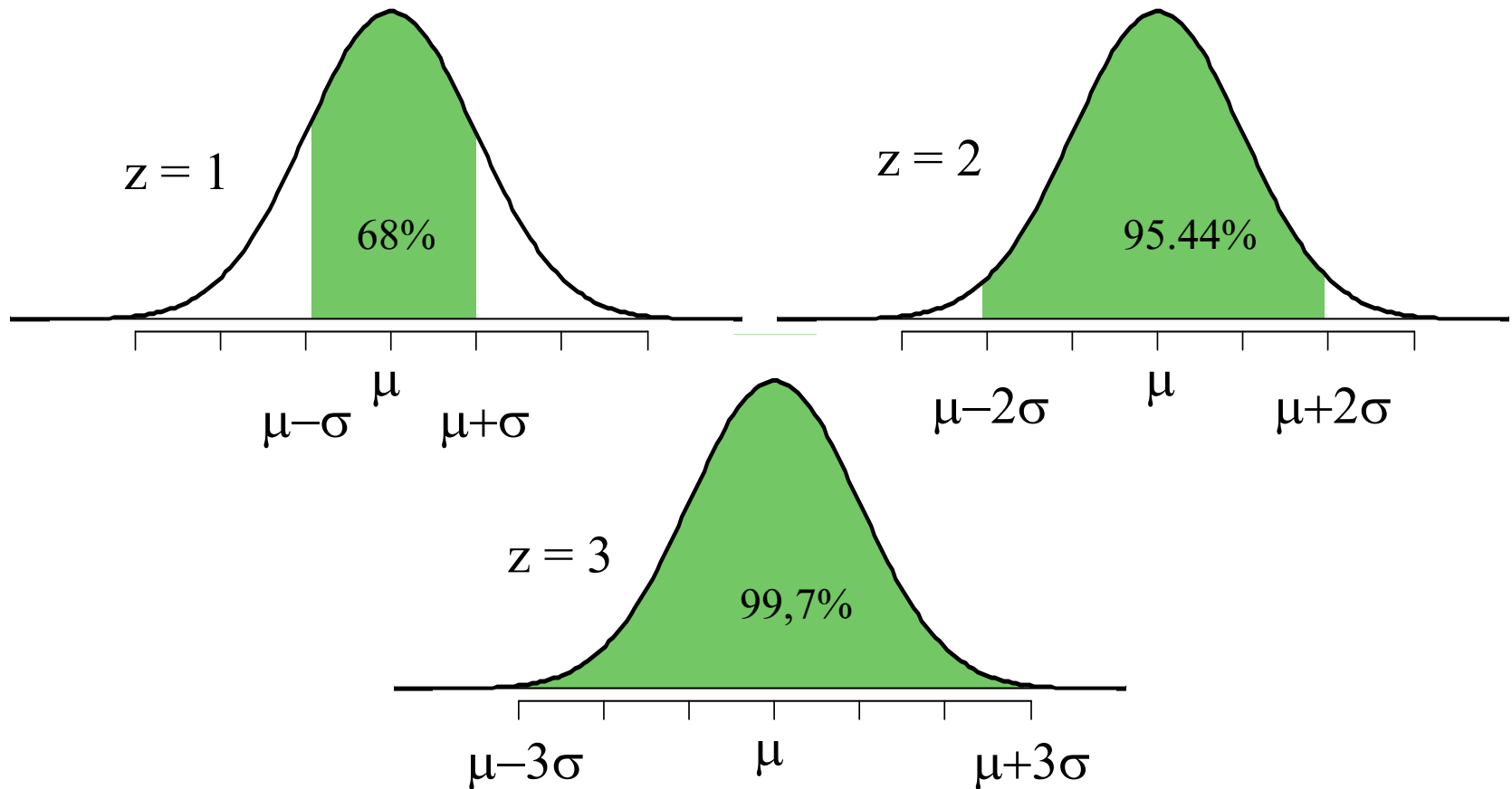
■ Egenskaber:

- **Symmetrisk** omkring μ
- $f(y) > 0$ for alle y .



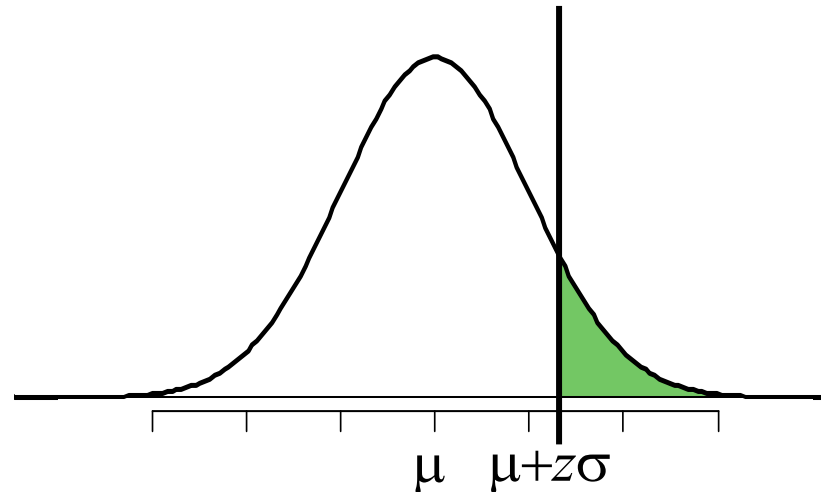
Sandsynligheder i normalfordelingen

Sandsynligheden for at fald indenfor z standardafvigelse fra middelværdien:



Sandsynligheder fra Tabel

- I Appendix A finder vi arealet af det grønne område for forskellige værdier af z .
- Antag $y \sim N(\mu, \sigma^2)$
- **Fortolkning 1:** Sandsynligheden for at y er større end $\mu + z\sigma$, for forskellige værdier af z .
- **Fortolkning 2:** Sandsynligheden for at y ligger mere end z standardafvigelser over μ .
- **Opgave:** Antag $y \sim N(\mu, \sigma^2)$. Find sandsynligheden for at y er mere end 1.26 standardafvigelser over middelværdien.



Løsning

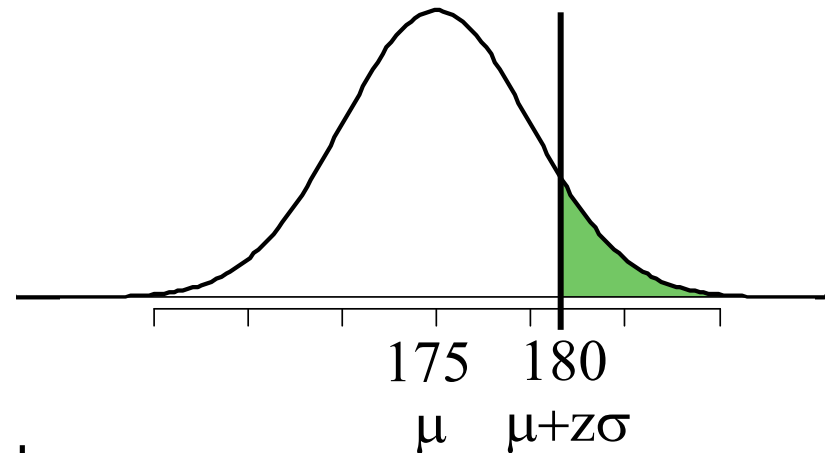
- **Opgave:** Antag $y \sim N(\mu, \sigma^2)$. Find sandsynligheden for at y er mere end 1.26 standardafvigelser over middelværdien.

z	Second Decimal Place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985

- **Svar:** $P(y \geq \mu + 1.26 \sigma) = 0.1038$
- **Bonus-spørgsmål:** $P(y \leq \mu + 1.26 \sigma) =$

Eksempel

- Antag højden blandt mænd er normalfordelt med middelværdi $\mu=175\text{cm}$ og standardafvigelse $\sigma=12\text{cm}$.
- Hvad er sandsynligheden for at en tilfældig udvalgt mand er højere end 180cm?
- Hvor mange standardafvigelse ligger 180cm over 175cm?
- $z = \text{————} \approx$
- Ifølge tabellen er sandsynligheden .
- Lad de 180cm være "y værdien", da er z-værdien $z = \frac{y - \mu}{\sigma}$
- **Bonus spørgsmål:** Hvad er sandsynligheden for at vælge en mand, der er højere end 170cm?



Sammenligning – Good vs Evil

- Mr Bond har scoret 550 i en test med middelværdi $\mu = 500$ og sa $\sigma = 100$
- Dr No har scoret 30 i en test med middelværdi $\mu = 18$ og sa $\sigma = 16$.
- Hvem har den mest imponerede score?
- Løsning: Hvem ligger flest standardafvigelser over middelværdien:

□ Mr. Bond: $z = \frac{y - \mu}{\sigma} =$

□ Dr. No $z = \frac{y - \mu}{\sigma} =$

Stikprøvefordeling

- **Ide:** Bruge stikprøve-statistik til at sige noget om populationsparameter.
- **Problem:** Stikprøve-statistikken vil variere fra gang til gang – vi introducerer en vis usikkerhed i vores konklusioner.
- **Stikprøvefordelingen**
 - Stikprøvefordelingen for en statistik er den sandsynlighedsfordeling, der beskriver sandsynligheden for de mulige værdier af statistikken.

Stikprøvefordeling: Eksempel

■ **Eksempel:** Valg til guvernør

- 56.5% af 2705 vil stemme på Schwarzenegger (kilde: CNN)
- De 56.5% er *stikprøve-andelen*.
- De 56.5% er et bud på andel af populationen, der vil stemme på Schwarzenegger.

■ **Spørgsmål**

- Hvor pålidelig er denne forudsigelse (af en valgsejr)?
- Stikprøveandelen vil åbenlyst variere fra gang til gang (stikprøvefordelingen).

■ **Svar: Tankeeksperiment!**

- Antag den sande andel Schwarzenegger-stemmer er 50%. Hvor (u)sandsynlig er en stikprøve-andel på 56.5% da?

Et Simuleret Svar!

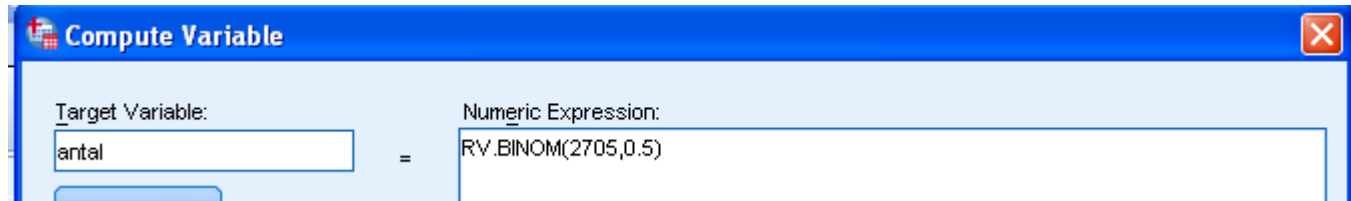
■ Forsøg

- Kast en fair mønt 2705 gang og noter andel af kroner.
- Gentag spøgen mange, mange gange...
- Hvordan fordeler andelen sig?
- Hvordan placerer 56.5% sig?

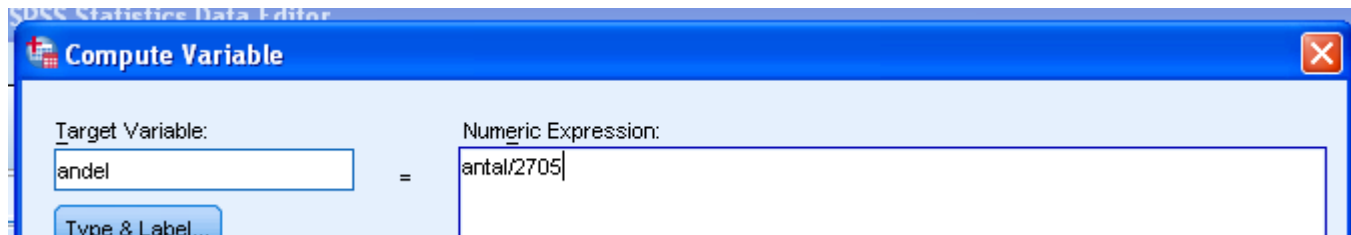
- Vi be'r SPSS om hjælp med møntkastene
-

Møntkast i SPSS

- Kør syntax-filen 1000.sps for at "narre" SPSS til at lave 1000 rækker.
- SPSS: Transform → Compute variable...
 - Simuler antal: $RV.BINOM(\text{antal kast}, \text{sandsynlighed for krone})$



- Udregn andel: $\text{andel} = \text{antal} / 2705$

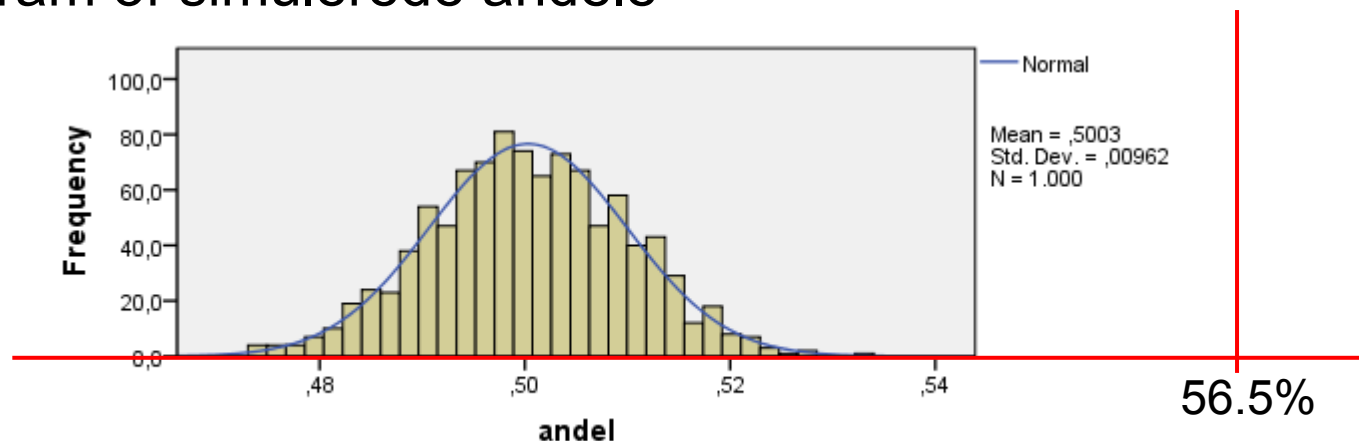


Stikprøvefordelingen af andele

■ Resultat i SPSS

	dummy	antal	andel	var	var	var
1	1,00	1383,00	,51			
2	1,00	1383,00	,51			
3	1,00	1381,00	,51			

■ Histogram of simulerede andele



- Var der i virkeligheden dødt løb, så er en stikprøveandel på 56.5% meget usandsynlig.
- Dvs. vi kan være ret sikre på at Mr S. vinder.

Stikprøvefordeling for $\bar{y}$

- Stikprøve-gennemsnittet $\bar{y}$ er en (stokastisk) variabel, da y variere (tilfældigt) fra gang til gang. Faktisk vil $\bar{y}$ varierer omkring populations-gennemsnittet μ .
- Antag vi har en stikprøve af størrelse n fra en population med middelværdi μ og s.a. σ .
- Stikprøvefordelingen for $\bar{y}$ har da
 - middelværdi μ
 - standardafvigelse $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ (betegnes standardfejlen)

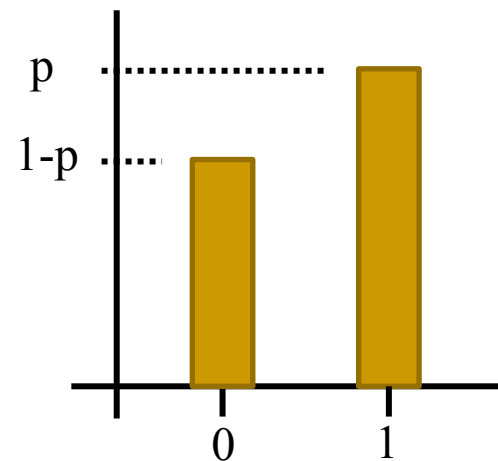
Stemmer

- Lad variablen y betegne om vil stemme på Mr. S eller ej.

- Antag $y = 1$ (Stemme på Schwrarzenegger)
 $y = 0$ (Stemme på ham den anden)

- Antag $P(1) = p$ ($0 \leq p \leq 1$)
 $P(0) = 1 - p$

- Da gælder $\mu = p$ og $\sigma = \sqrt{p(1-p)}$



- Stemmeandelen er et gennemsnit af mange y 'er.
- **Bemærk:** m er populationsandelen af stemmer på Mr. S.

Stemmer – fortsat

- For populationen har vi altså
 - at $\mu = p$ og $\sigma = \sqrt{p(1-p)}$
 - Dvs. y har middelværdi μ og standardfejl $\sigma_{\bar{y}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$
- Jf. tommelfingerregel, er vil 95% af alle stikprøveandele ligge i intervallet $\mu \pm 2\sigma_{\bar{y}}$
- Antager vi $p=0.50$ (dødt løb) har vi: $p \pm 2 \cdot \sqrt{p(1-p)} / \sqrt{n} = 0.50 \pm 2 \cdot 0.5 / \sqrt{2705}$
- Bemærk: Intervallet bliver kortere, hvis vi øger n .

Central grænseværdisætning (CLT)

■ Central grænseværdisætning

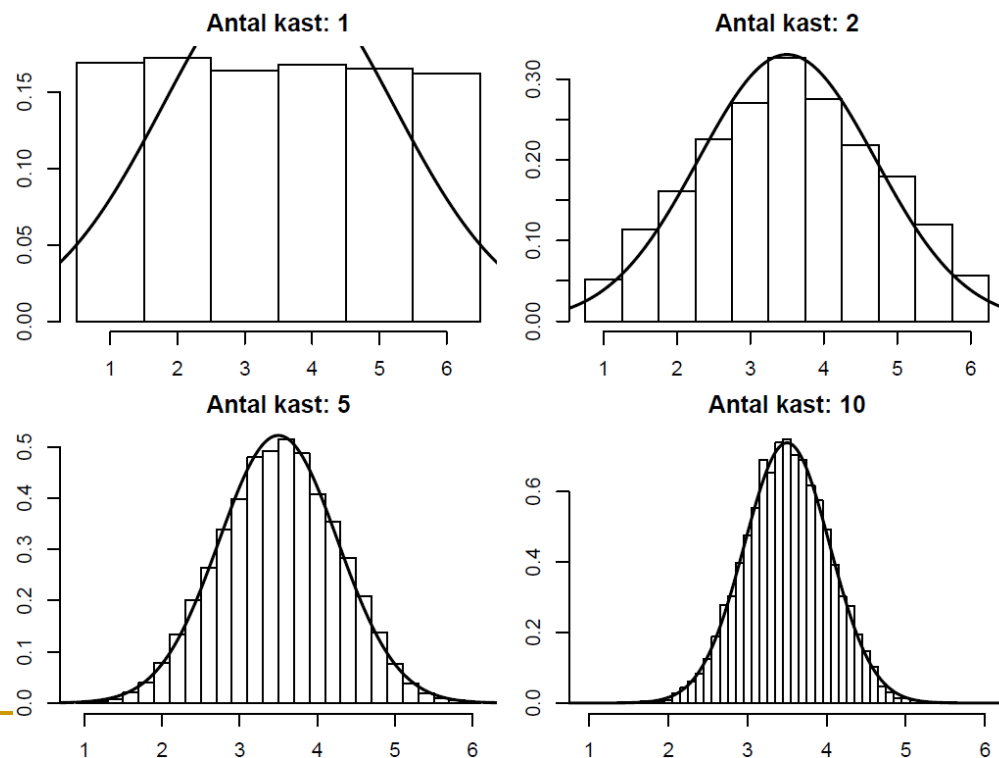
- For en tilfældig stikprøve med en tilstrækkelig **stor stikprøvestørrelse** n , vil stikprøvefordelingen af stikprøvegennemsnittet $\bar{y}$ være ca. **normalfordelt**.

■ Eksempel:

Det **gennemsnitlige antal øjne** i k kast med en terning.

Til høje: $k = 1, 2, 5, 10$

Bemærk: Allerede med $k = 10$ kast er gennemsnittet meget lig en normalfordeling.



Eksempel

- En stikprøveandel er (ca.) normalfordelt, hvis stikprøvestørrelsen er stor og populationsandelen ikke er for tæt på 0 eller 1.
- Stikprøve andelen er normalfordel med middelværdi p og standardafvigelse $\sqrt{p(1-p)/n}$.
- Hvis der er dødt løb ($p = 0.50$), hvad er da sandsynligheden for at se en stikprøveandel på 0.565 eller større ved en stikprøve på 2705?

$$z = \frac{y - \mu}{\sigma} = \frac{0.565 - 0.5}{\sqrt{0.5(1-0.5)/2705}} =$$

