TMA4240 Statistikk H2016 [18]

Ett utvalg: estimere forventningsverdi og konfidensintervall [9.4]
Student-t fordeling [8.6]

Mette Langaas

Institutt for matematiske fag, NTNU

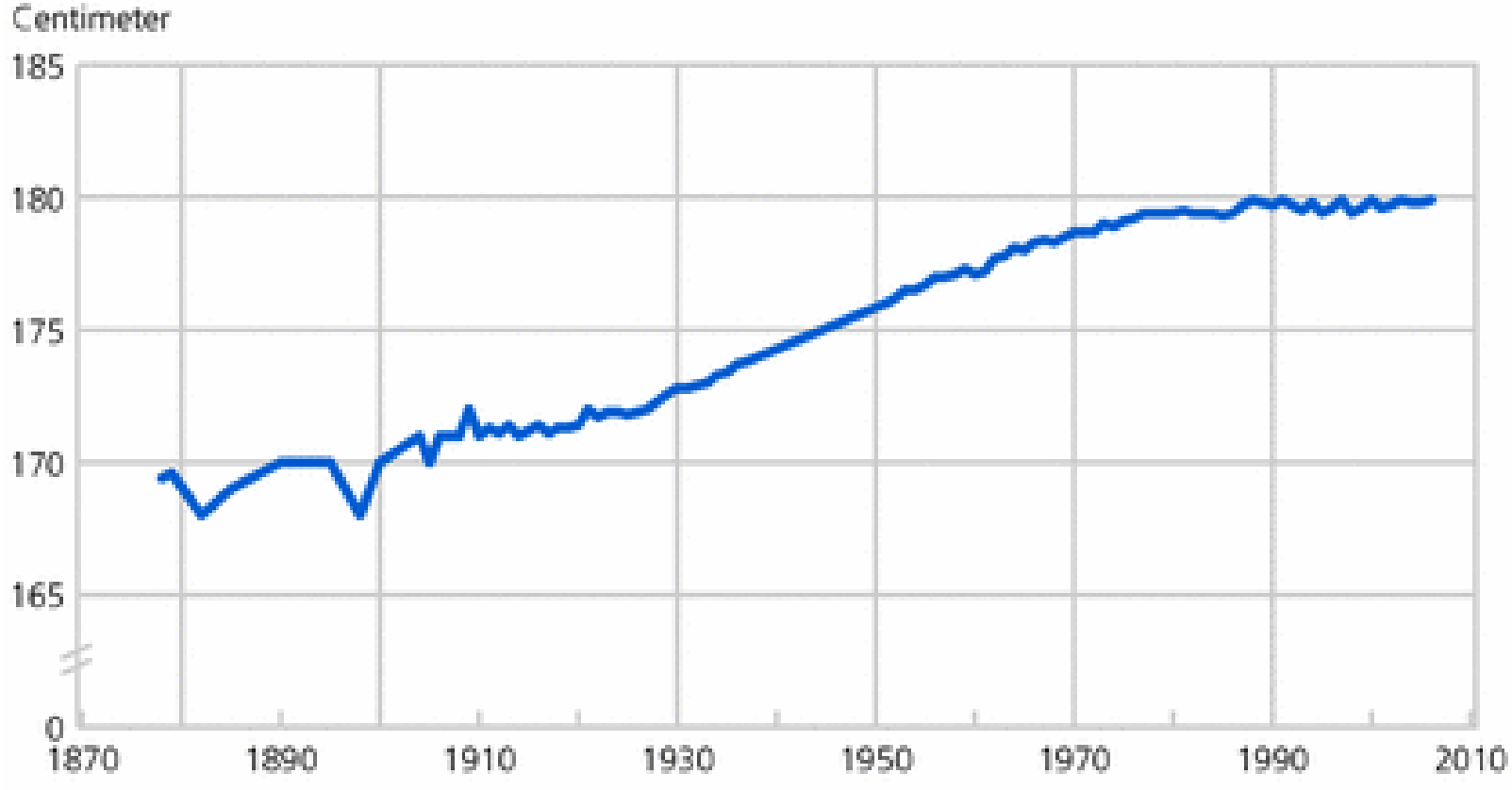
wiki.math.ntnu.no/emner/tma4240/2016h/start/

Høyde blant rekrutter

- Militære myndigheter har i århundrer målt kroppshøyde som del av de fysiske undersøkelsene under de militære mønstringene av innkalte mannskaper (sesjonene).
- Årsakene var flere, blant annet at høydedata ble brukt til identifikasjonsformål fordi dagens fotobevis var ikke tilgjengelig.
- Videre ble høydedata brukt ved produksjon av uniformer, til beregning av standard matrasjoner og til å fastsette regimentenes kampkapasitet.
- Høydemålinger av innkalte mannskaper for infanteriet i Norge ble påbudt i 1705.

Kilde: Bore, Ragnhild Rein (2007): "Norske rekrutter har skutt i været", i På liv og død, Statistiske analyser 94, 2007, Statistisk sentralbyrå. http://www.ssb.no/a/publikasjoner/pdf/sa94/del-v-1b.pdf

Høyde blant rekrutter



[1] Gjelder menn i utskrivingsstyrken året før. 1878-1900: Tallene gjelder kun for sørlige Norge t.o.m. Nord-Trøndelag. 1878-1912: Rekrutter under 158 cm og over 185 cm er ikke med. 1913-1916: Rekrutter under 157 cm og over 186 cm er ikke med. 1938-1950: Interpolert.
Kilde: Bore 2007, Statistisk sentralbyrå.

https://www.ssb.no/helse/artikler-og-publikasjoner/vernepliktige-opp-i-vekt

Høyde blant rekrutter

108 Egenrapportert høyde og vekt for sesjonspliktige[1]

	Gjennomsnittshøyde, cm		Gjennomsnittsvekt, kg	
	Gutter	Jenter	Gutter	Jenter
1910	171,0	..	..	..
1920	171,4	..	..	..
1930	172,8	..	..	..
1937	173,8	..	..	..
1952	176,2	..	..	..
1960	177,1	..	..	..
1970	178,7	..	..	..
1980	179,4	..	..	..
1990	179,7	..	..	..
2000	179,9	..	72,8	..
2005	179,8	..	74,7	..
2010	179,5	..	75,5	..
2011 (født 1994)	179,9	167,1	73,7	62,1
2012 (født 1995)	179,9	167,1	73,9	62,5

1 Brudd i statistikken i 2011 pga. endring i Vernepliktsloven fra 1. januar 2010. Tidligere årganger har vært basert på fysiske målinger ved sesjon. Fra 2011 benyttes opplysninger fra egenrapportert nettskjema som sendes Vernepliktsverket normalt det året de fyller 17 år.

Kilde: Vernepliktsverket.

Mer informasjon: http://www.forsvaret.no/

http://www.ssb.no/a/aarbok/tab/tab-108.html

Høyde for kvinnelige studenter ved NTNU

Mål: finne ut hva gjennomsnittshøyden er for alle kvinnelige studenter ved NTNU!

Scenario 1: Jeg velger tilfeldig ut 10 kvinner nå i forelesningen og spør om høyde. Jeg finner at gjennomsnittet av de 10 høydene er 169.5 cm.

Høyde for kvinnelige studenter ved NTNU

Mål: finne ut hva gjennomsnittshøyden er for alle kvinnelige studenter ved NTNU!

Scenario 1: Jeg velger tilfeldig ut 10 kvinner nå i forelesningen og spør om høyde. Jeg finner at gjennomsnittet av de 10 høydene er 169.5 cm.

Scenario 2: Spørreundersøkelsen TMA4240: Hvor høy er du? kombinert med Kjønn="Kvinne". Det var 149 svar og gjennomsnittlig høyde var 169.5 cm.

Høyde for kvinnelige studenter ved NTNU

Mål: finne ut hva gjennomsnittshøyden er for alle kvinnelige studenter ved NTNU!

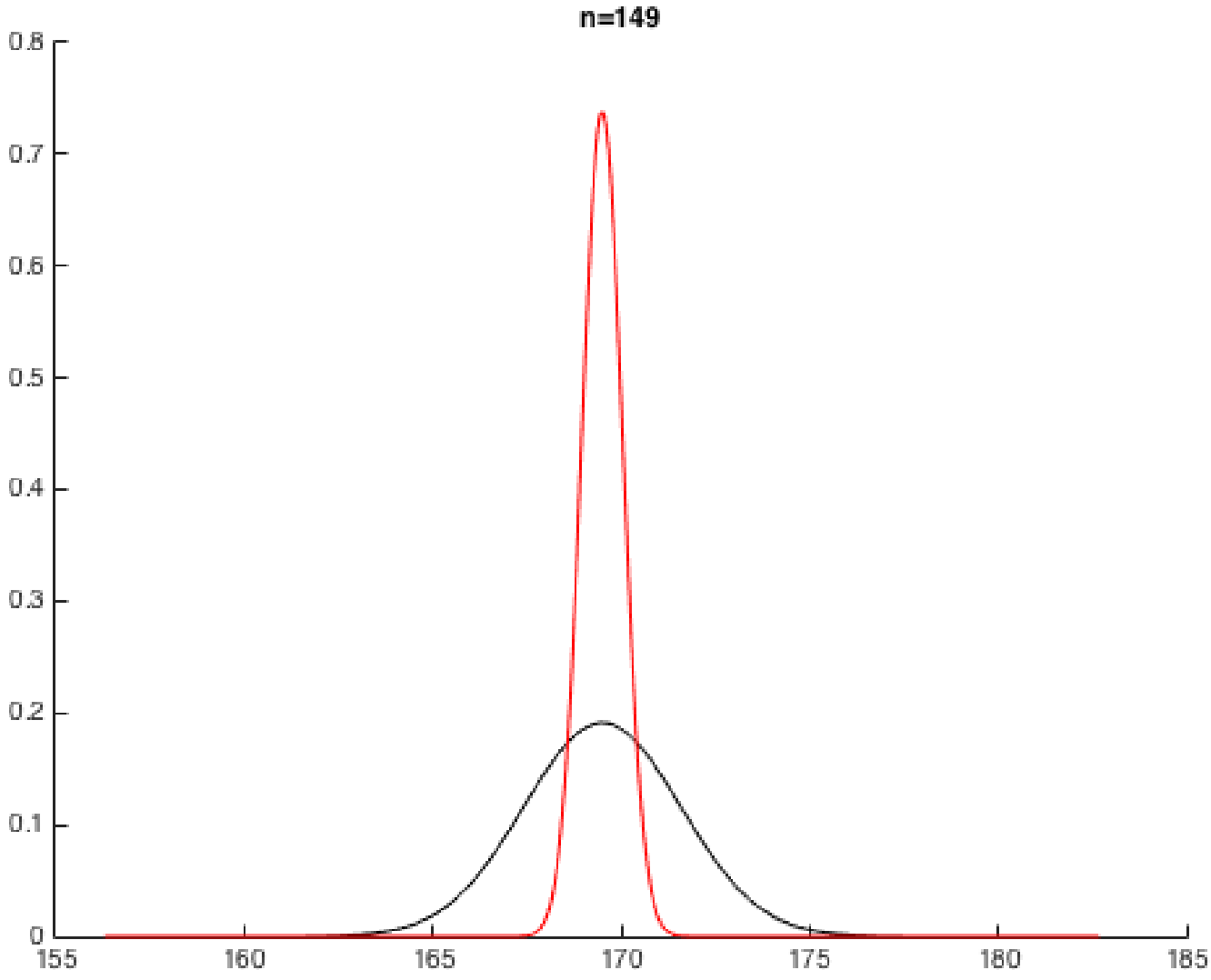
Scenario 1: Jeg velger tilfeldig ut 10 kvinner nå i forelesningen og spør om høyde. Jeg finner at gjennomsnittet av de 10 høydene er 169.5 cm.

Scenario 2: Spørreundersøkelsen TMA4240: Hvor høy er du? kombinert med Kjønn="Kvinne". Det var 149 svar og gjennomsnittlig høyde var 169.5 cm.

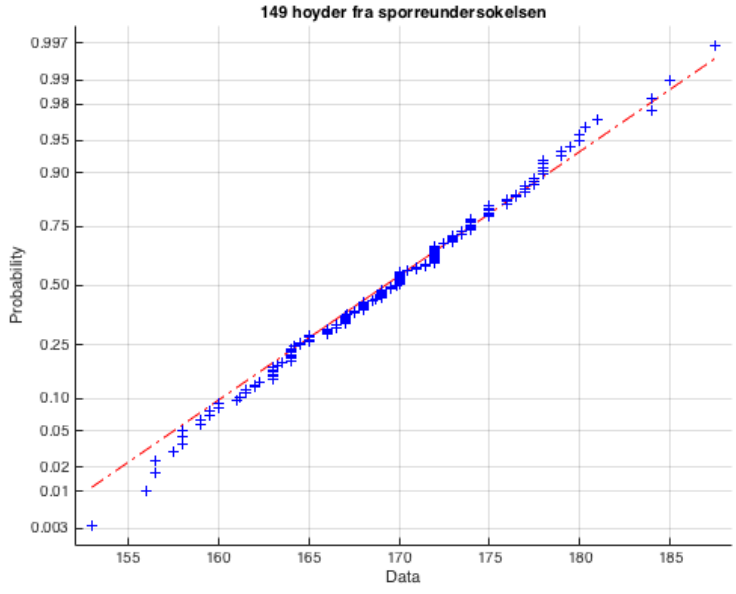
De to scenarioene gav begge et gjennomsnitt i utvalget på 169.5 cm. Dette vil vi bruke som vårt punktestimat for gjennomsnittet i populasjonen. Men, vil du stole like mye på begge disse to undersøkelsene? Bør vi i tillegg til et punktestimat komme med mer informasjon?

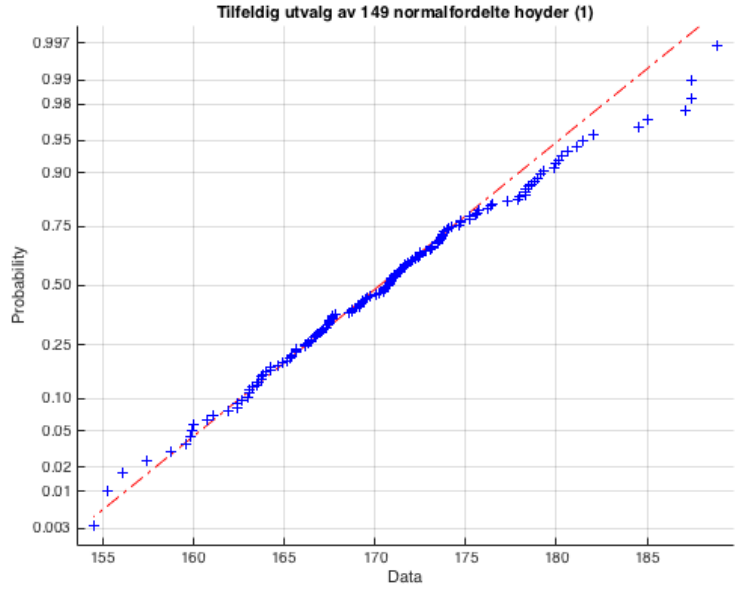
Fordeling til gjennomsnitt av 10 og 149

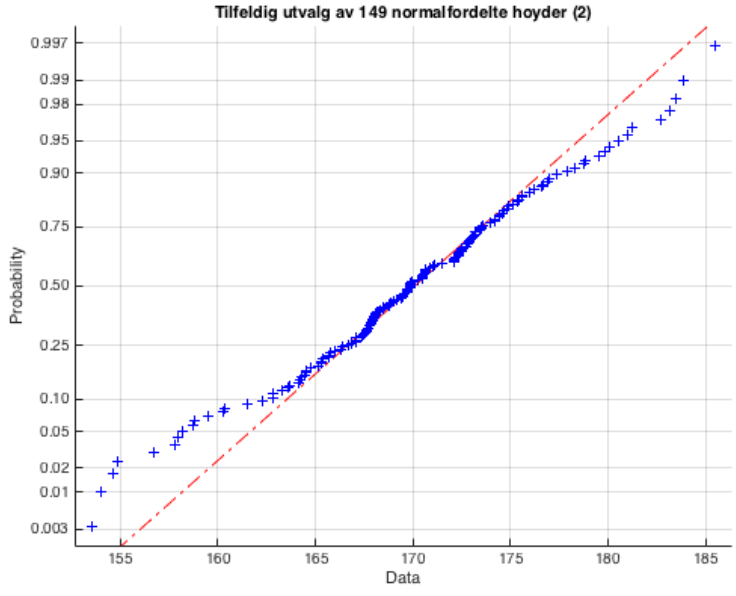
Antar høyde er normalfordelt, og observasjoner uavhengige.

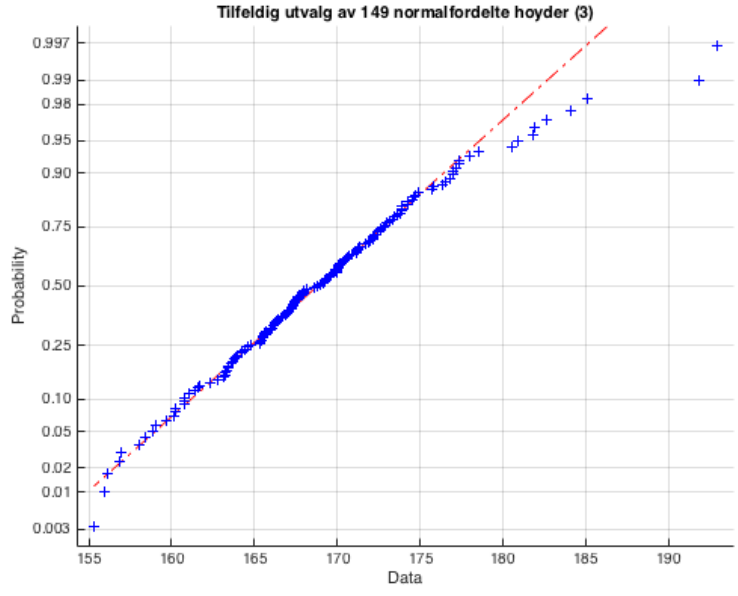


Høyde-dataene: normalfordelt? [8.8]









Fordeling til gjennomsnittet $\bar{X}$ [8.4]

TEO 8.2: Sentralgrenseteoremet La $X_1, X_2, ..., X_n$ være et tilfeldig utvalg fra en fordeling med forventning μ og varians σ^2. Da har vi at sannsynlighetsfordelingen til

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

går mot standard normalfordelingen, $n(z; 0, 1)$, når $n \to \infty$.

Merk: Det er det samme som å si at $\bar{X}$ er tilnærmet $N(\mu, \frac{\sigma^2}{n})$.

Kritiske verdier i standard normalfordelingen

$$P(Z > z_\alpha) = \alpha$$

α	z_α
.2	0.842
.15	1.036
.1	1.282
.075	1.440
.05	1.645
.04	1.751
.03	1.881
.025	1.960
.02	2.054
.01	2.326
.005	2.576
.001	3.090
.0005	3.291
.0001	3.719
.00005	3.891
.00001	4.265
.000005	4.417
.000001	4.753

Konfidensintervall for μ med σ kjent

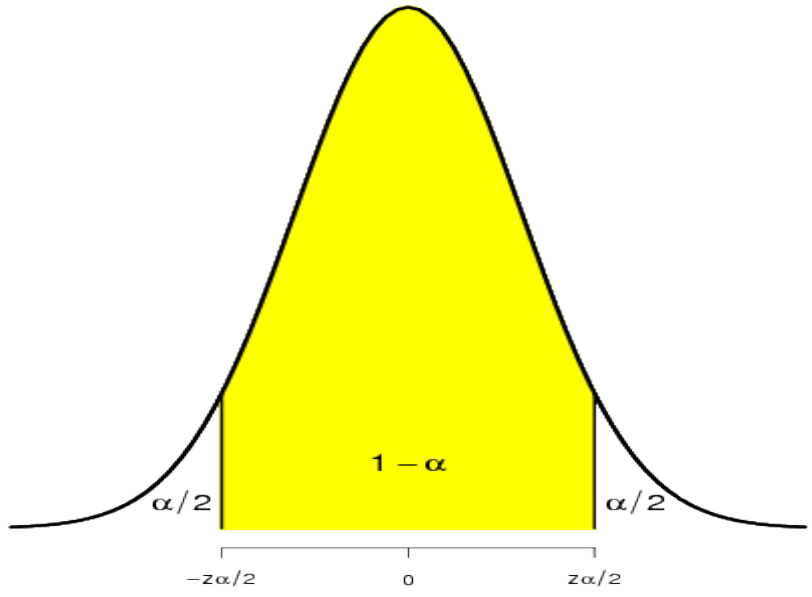
- Hvis $\bar{x}$ er gjennomsnittet av et tilfeldig utvalg av størrelse n fra en populasjon med kjent varians σ^2, så er et (1-α)100% konfidensintervall for μ

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

hvor $z_{\frac{\alpha}{2}}$ er verdien i standard normalfordelingen som har areal $\frac{\alpha}{2}$ til høyre, dvs. $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Konfidensintervall for μ med σ kjent

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$



3 spørsmål

På utdelte hvite lapp svarer du på følgende tre spørsmål - og leverer så lappen i boks på vei ut av auditoriet

1. Hvor mange prosent av denne forelesningen forstod du?
2. Er det 90% konfidensintervall bredere eller smalere enn et 95% konfidensintervall?
3. Er det noe TMA4240-teamet – inkludert foreleser - kan gjøre slik at du kan øke din **læring** i emnet? Hva?

T og t-fordeling

COR: La $X_1, X_2, ..., X_n$ være uavhengige stokastiske variabler som alle er normalfordelte med samme forventning μ og samme standardavvik σ. La

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{og} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Da er den stokastiske variablen

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

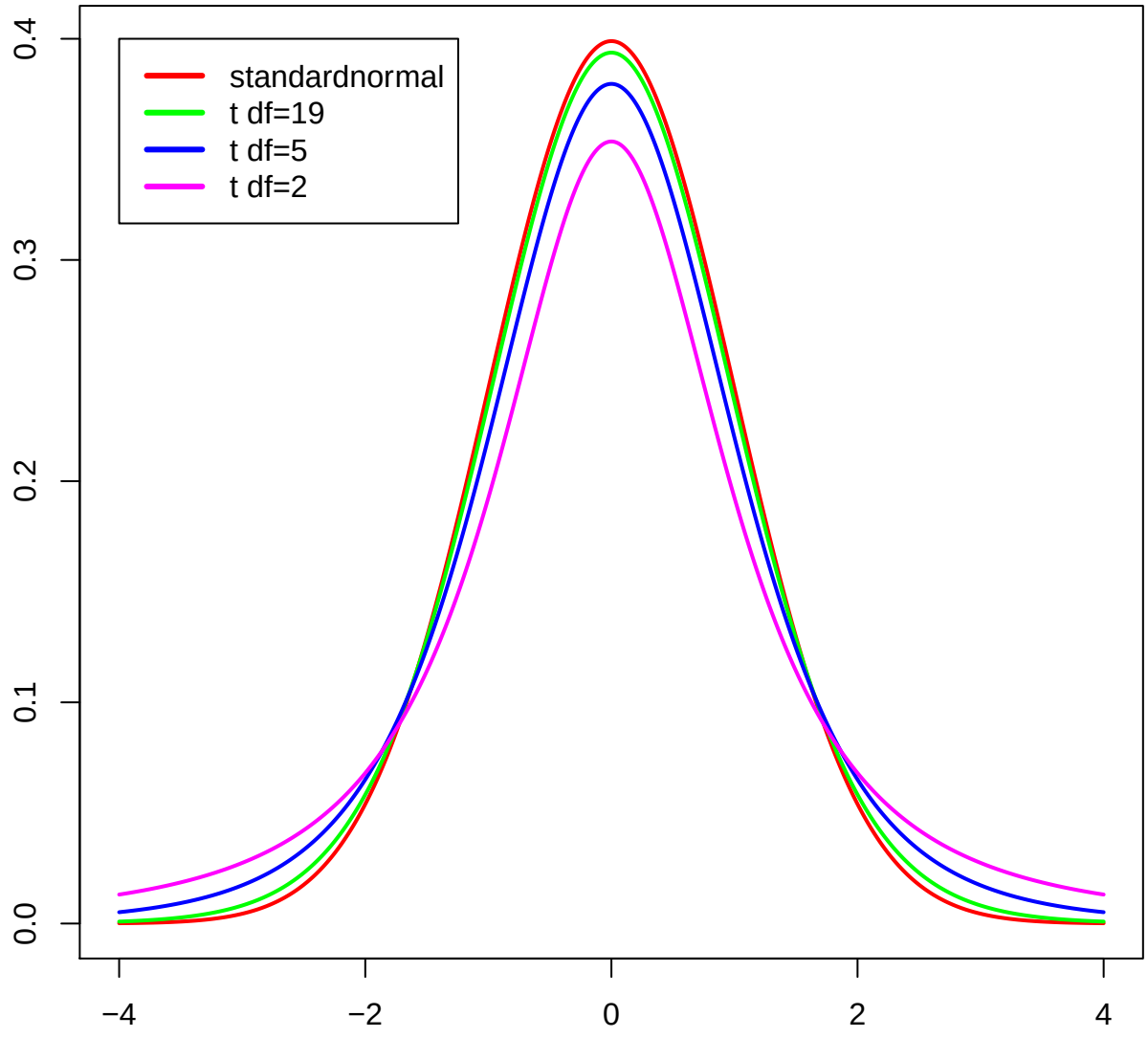
t-fordelt med $\nu = (n-1)$ frihetsgrader.

W. S. Gosset alias Student

Historisk: Student-t fordelingen

- W.S. Gosset (1876-1937) was employed by the Guinness Brewing Company of Dublin.
- Sample sizes available for experimentation in brewing were necessarily small, and Gosset knew that a correct way of dealing with small samples was needed.
- He consulted Karl Pearson (1857-1936) of Universiy College in London about the problem. Pearson told him the current state of knowledge was unsatisfactory.
- The following year Gosset undertook a course of study under Pearson. An outcome of his study was the publication in 1908 of Gosset's paper on "The Probable Error of a Mean," which introduced a form of what later became known as Student's t-distribution.
- Gosset's paper was published under the pseudonym "Student."
- The modern form of Student's t-distribution was derived by R.A. Fisher and first published in 1925.

t-fordelingen



DEF: t-fordeling

TEO 8.5: La Z være en standard normalfordelt stokastisk variabel og V være en kjikvadrat-fordelt stokastisk variabel med ν frihetsgrader. Hvis Z og V er uavhengige, er fordelingen til den stokastiske variablen T

$$T = \frac{Z}{\sqrt{V/\nu}}$$

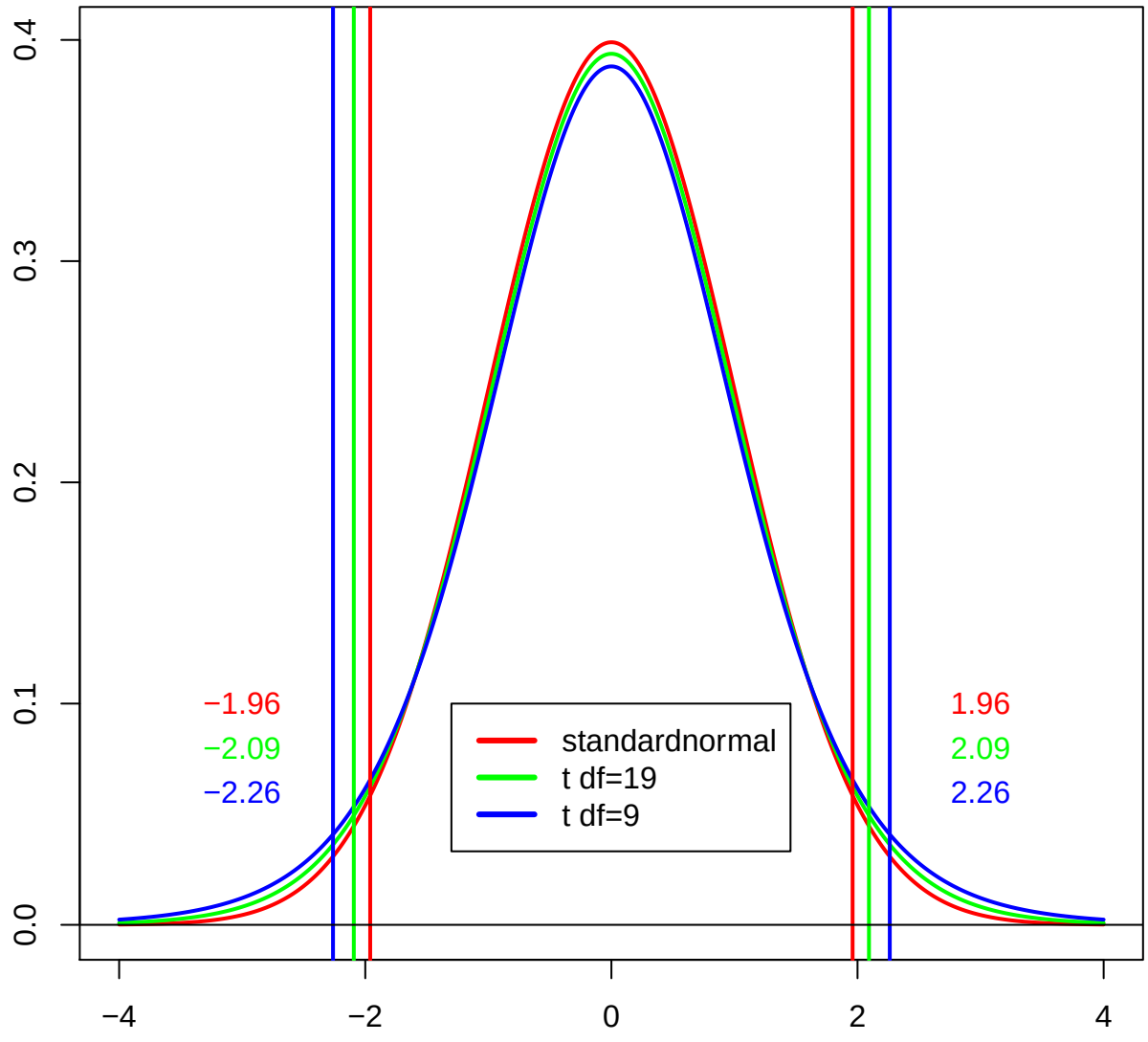
gitt ved sannsynlighetstettheten

$$h(t) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}}(1+\frac{t^2}{\nu})^{-(\nu+1)/2}$$

for $-\infty < t < \infty$. Denne fordelingen har navnet (Student) $t-$fordelingen med ν frihetsgrader.

- $\mathrm{E}(T) = 0$ hvis $\nu \geq 2$.
- $\mathrm{Var}(T) = \frac{\nu}{\nu-2}$ hvis $\nu \geq 3$.

Kvantiler N og t: $\alpha/2 = 0.025$



Kritiske verdier i t-fordelingen

$$P(T > t_{\alpha,\nu}) = \alpha$$

$\nu \backslash \alpha$	.150	.100	.075	.050	.025	.010	.005	.001	.0005
1	1.963	3.078	4.165	6.314	12.706	31.821	63.657	318.309	636.619
2	1.386	1.886	2.282	2.920	4.303	6.965	9.925	22.327	31.599
3	1.250	1.638	1.924	2.353	3.182	4.541	5.841	10.215	12.924
4	1.190	1.533	1.778	2.132	2.776	3.747	4.604	7.173	8.610
5	1.156	1.476	1.699	2.015	2.571	3.365	4.032	5.893	6.869
6	1.134	1.440	1.650	1.943	2.447	3.143	3.707	5.208	5.959
7	1.119	1.415	1.617	1.895	2.365	2.998	3.499	4.785	5.408
8	1.108	1.397	1.592	1.860	2.306	2.896	3.355	4.501	5.041
9	1.100	1.383	1.574	1.833	2.262	2.821	3.250	4.297	4.781
10	1.093	1.372	1.559	1.812	2.228	2.764	3.169	4.144	4.587
11	1.088	1.363	1.548	1.796	2.201	2.718	3.106	4.025	4.437
12	1.083	1.356	1.538	1.782	2.179	2.681	3.055	3.930	4.318
13	1.079	1.350	1.530	1.771	2.160	2.650	3.012	3.852	4.221
14	1.076	1.345	1.523	1.761	2.145	2.624	2.977	3.787	4.140
15	1.074	1.341	1.517	1.753	2.131	2.602	2.947	3.733	4.073
16	1.071	1.337	1.512	1.746	2.120	2.583	2.921	3.686	4.015
17	1.069	1.333	1.508	1.740	2.110	2.567	2.898	3.646	3.965
18	1.067	1.330	1.504	1.734	2.101	2.552	2.878	3.610	3.922
19	1.066	1.328	1.500	1.729	2.093	2.539	2.861	3.579	3.883
20	1.064	1.325	1.497	1.725	2.086	2.528	2.845	3.552	3.850
21	1.063	1.323	1.494	1.721	2.080	2.518	2.831	3.527	3.819
22	1.061	1.321	1.492	1.717	2.074	2.508	2.819	3.505	3.792
23	1.060	1.319	1.489	1.714	2.069	2.500	2.807	3.485	3.768
24	1.059	1.318	1.487	1.711	2.064	2.492	2.797	3.467	3.745
25	1.058	1.316	1.485	1.708	2.060	2.485	2.787	3.450	3.725
26	1.058	1.315	1.483	1.706	2.056	2.479	2.779	3.435	3.707
27	1.057	1.314	1.482	1.703	2.052	2.473	2.771	3.421	3.690
28	1.056	1.313	1.480	1.701	2.048	2.467	2.763	3.408	3.674
29	1.055	1.311	1.479	1.699	2.045	2.462	2.756	3.396	3.659
30	1.055	1.310	1.477	1.697	2.042	2.457	2.750	3.385	3.646
35	1.052	1.306	1.472	1.690	2.030	2.438	2.724	3.340	3.591
40	1.050	1.303	1.468	1.684	2.021	2.423	2.704	3.307	3.551
50	1.047	1.299	1.462	1.676	2.009	2.403	2.678	3.261	3.496
60	1.045	1.296	1.458	1.671	2.000	2.390	2.660	3.232	3.460
80	1.043	1.292	1.453	1.664	1.990	2.374	2.639	3.195	3.416
100	1.042	1.290	1.451	1.660	1.984	2.364	2.626	3.174	3.390
120	1.041	1.289	1.449	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.036	1.282	1.440	1.645	1.960	2.326	2.576	3.090	3.291

Konfidensintervall for μ med σ ukjent

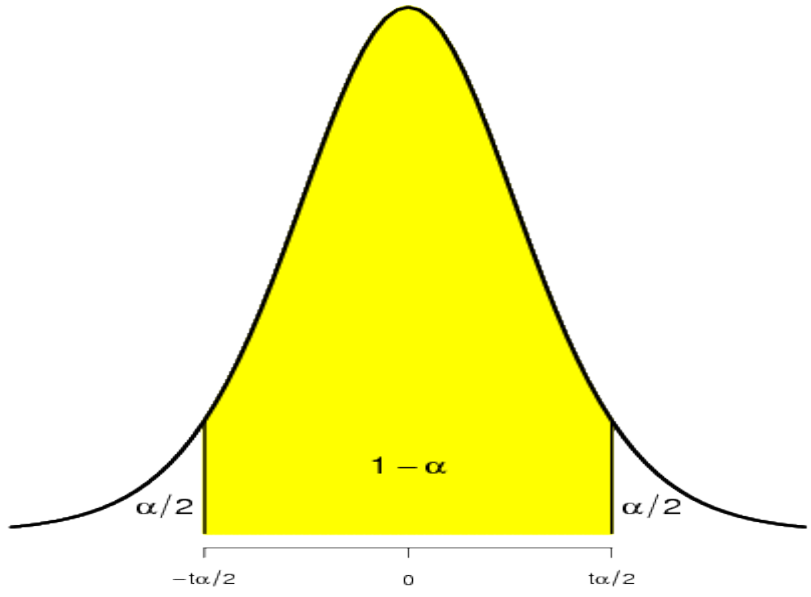
- Hvis $\bar{x}$ er gjennomsnittet og s er estimert standardavvik av et tilfeldig utvalg av størrelse n fra en populasjon med ukjent varians σ^2, så er et (1-α)100% konfidensintervall for μ

$$\bar{x} - t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}}$$

  hvor $t_{\frac{\alpha}{2},(n-1)}$ er verdien i t-fordelingen med $n-1$ frihetsgrader som har areal $\frac{\alpha}{2}$ til høyre, dvs. $P(T > t_{\frac{\alpha}{2},(n-1)}) = \frac{\alpha}{2}$.

Konfidensintervall for μ med σ ukjent

$$\bar{x} - t_{\frac{\alpha}{2},(n-1)}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2},(n-1)}\frac{s}{\sqrt{n}}$$



Fordelingen til S^2

- Resultat: $V = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - \bar{Z}^2$ er kjikvadrat-fordelt med $\nu = n - 1$ frihetsgrader. Fordi:

  i) $X_1, ..., X_n$ u.i.f. normal, $\mathrm{E}(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$.

  ii) $Z_i = \frac{X_i - \mu}{\sigma}$ er standard normalfordelt, og $\bar{Z} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ er standard normalfordelt.

  iii) $Z_i^2 = \left(\frac{X_i - \mu}{\sigma}\right)^2$ er kjikvadrat-fordelt med 1 frihetsgrad.

  $\bar{Z}^2 = \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$ er kjikvadrat-fordelt med 1 frihetsgrad.

  iv) $\sum_{i=1}^{n} Z_i^2$ er kjikvadratfordelt med n frihetsgrader.

  v) $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$, og dermed $V = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - \bar{Z}^2$

  vi) $\sum_{i=1}^{n} Z_i^2$ og $\bar{Z}^2$ er uavhengige.

Læringsmål

- Punktestimering: forstå at et punktestimat kun gir informasjon om et "beste gjett" og ikke sier noe om presisjonen i estimatet.
- Være med de ulike stegene i utledningen av et konfidensintervall.
- Forstå, ved å trekke repeterte utvalg, hva et $(1-\alpha)\cdot 100\%$ konfidensintervall betyr.
- Forstå hvilke størrelser som påvirker bredden til et konfidensintervall for forventningsverdien (utvalgsstørrelsen, konfidensnivå, kjent eller ukjent varians, variansen til estimatoren som benyttes).
- Forstå at en t-fordeling har tyngre haler enn en normalfordeling, og hva "tyngre" haler betyr.
- Forstå hvorfor en normalfordeling må erstattes med en t-fordeling når vi ser på en standardisert størrelse basert på gjennomsnittet av normalfordelte variabler der variansen ikke er kjent.