BI1363 HT 2020
t-test för ett stickprov
Oktober 2020
Adam Flöhr, BT, SLU

t-test för ett stickprov

Motsvarar *Biometri*, kap 6

I korthet

Slumpmässig variation är en grundläggande utmaning för all kunskapsinhämtning

I en statistisk undersökning vill vi uttala oss om egenskaper i en **population** med observationer i ett **stickprov**

Konfidensintervall gör det möjligt att ringa in populationens egenskaper med en förbestämd säkerhet

Hypotestest är en metod för att dra *statistiskt säkerställda* slutsatser om populationens egenskaper

För kontinuerlig data används ofta ett **t-test**

t-test för ett stickprov testar om medelvärdet i en variabel är skilt från ett hypotesvärde

Population och stickprov

I en statistisk undersökning vill vi veta något om en egenskap (en *variabel*) i en *population*

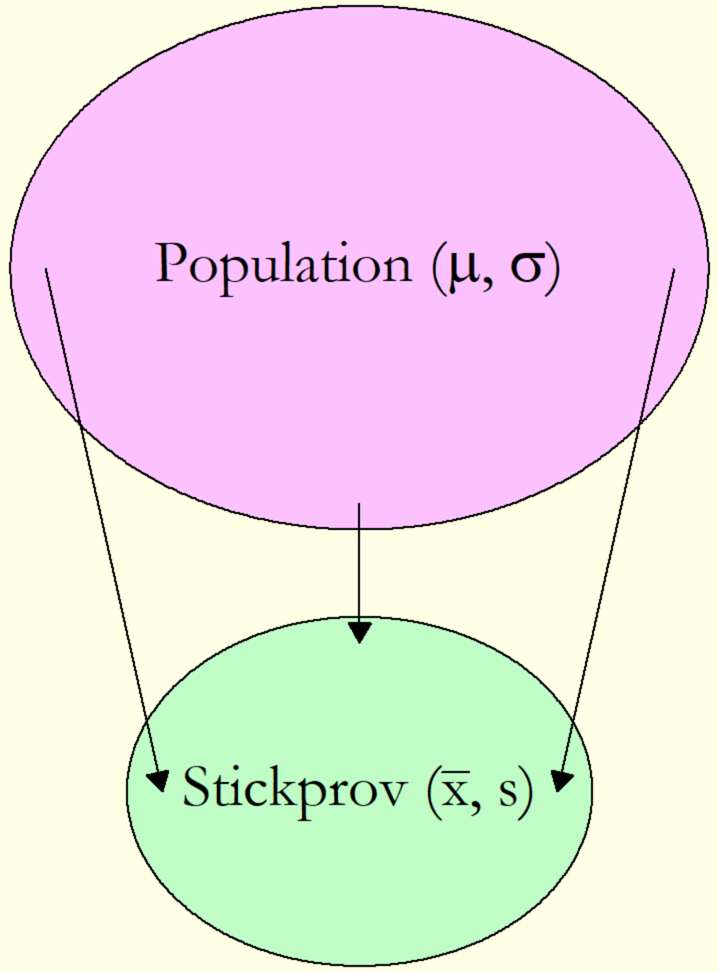
Ofta omöjligt att observera hela populationen, så vi drar ett stickprov om n observationer

Stickprovet kan användas till att *punktskatta* egenskaper som medelvärde och varians i populationen

Egenskaper i populationen betecknas med grekiska bokstäver (μ och σ) och skattningarna betecknas med romerska bokstäver ($\bar{x}$ och s)

Valet av punktskattning beror på variabeln vi undersöker

Om populationsvariabeln är normalfördelad är vi intresserade av skattade värden för medelvärde och varians (stickprovsmedelvärde och stickprovsvarians)

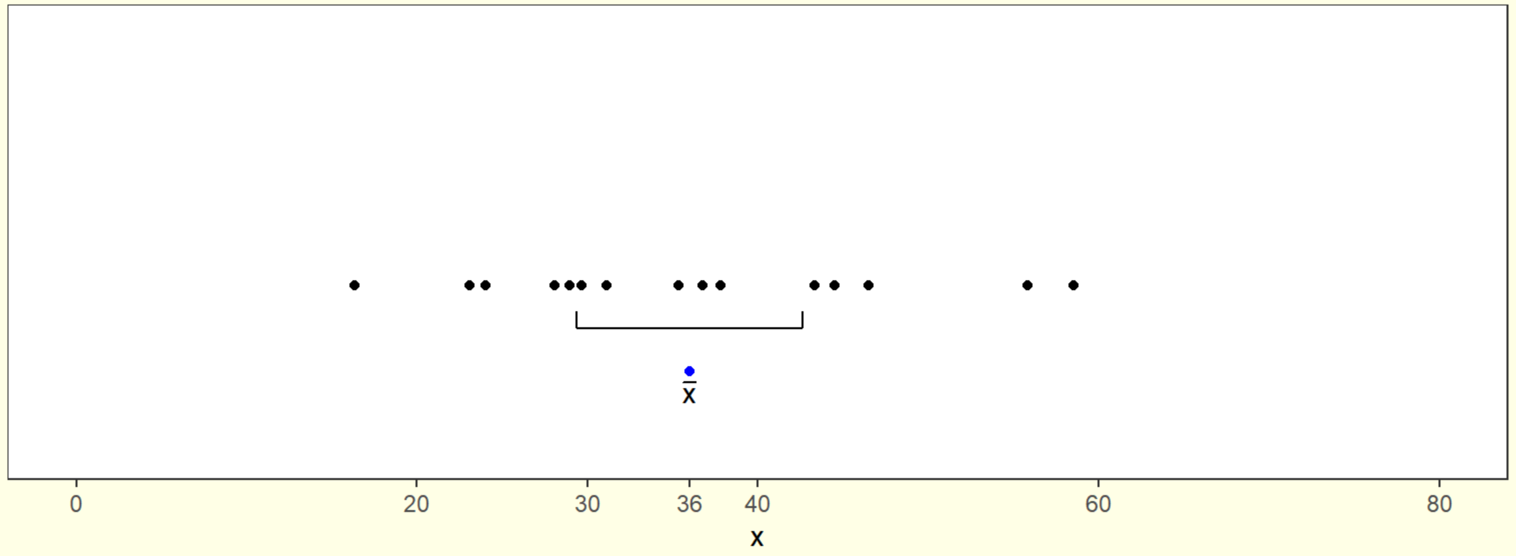


Punkskattningar är utfall av slumpvariabler

Eftersom observationerna i stickprovet är slumpmässiga kommer även punktskattningarna bero på slumpen

Vi kan använda punkskattningar för att ringa in populationsmedelvärdet μ med ett *konfidensintervall*

Ett konfidensintervall täcker μ med en bestämd konfidens, ofta 95 procent.

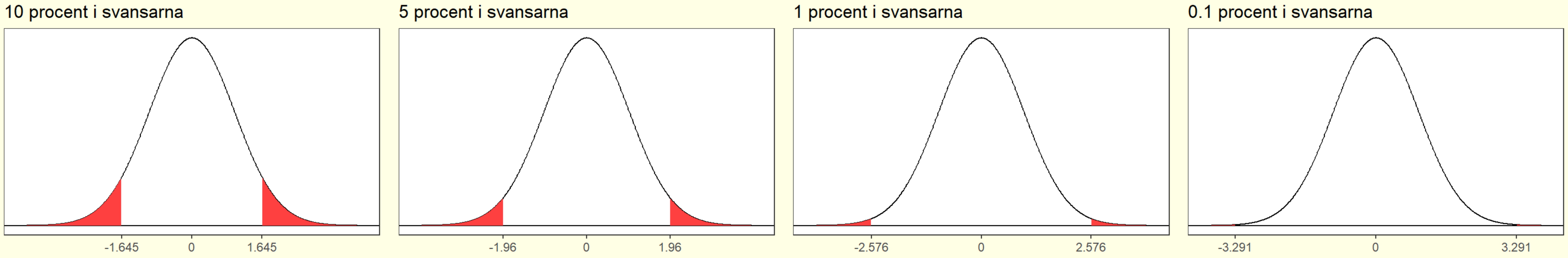


Punktskattningen av medelvärdet är $\bar{x} = 36$

Det sanna populationsmedelvärdet ligger med 95 procents konfidens i det angivna intervallet

Konfidensintervall med känd standardavvikelse

Om man har en känd populationsstandardavvikelse σ baseras konfidensintervallet på normalfördelningen



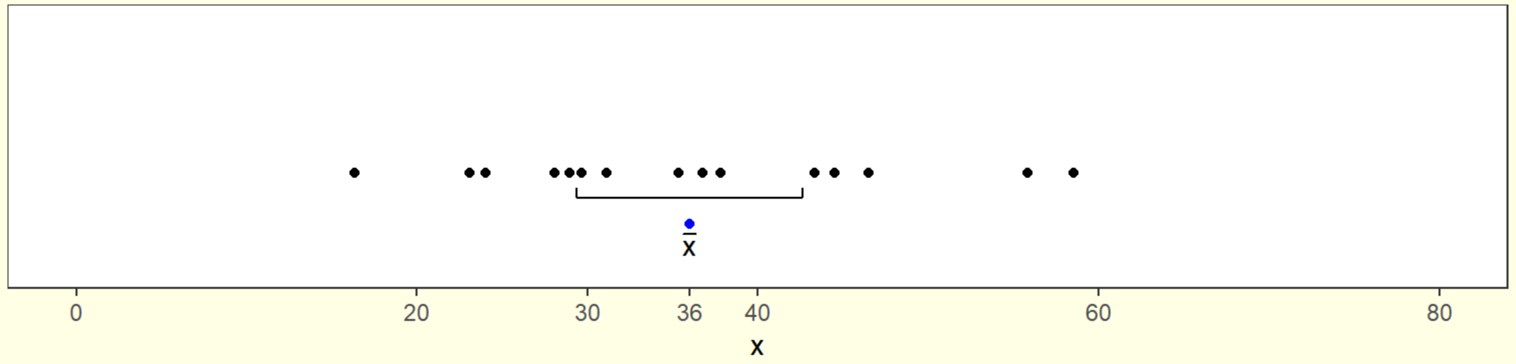
För en given svanssannolikhet α använder vi beteckningen $z_{1-\alpha/2}$ för värdet på x-axeln, t.ex $z_{(1-0.05/2)} = z_{0.975} = 1.96$ eftersom sannolikheten i svansarna bortom 1.96 är 0.05

Vi kan föra över det standardiserade intervallet till ett skattat medelvärde genom formeln

$$\bar{x} \pm z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}$$

där $\bar{x}$ är det skattade medelvärdet, $z_{(1-\alpha/2)}$ är värdet från den standardiserade normalfördelningen, σ är den kända standardavvikelsen, och n är stickprovets storlek

Konfidensintervall, känd standardavvikelsen, exempel



I datan ovan är det skattade medelvärdet $\bar{x} = 36$, den givna standardavvikelsen $\sigma = 12$, och stickprovsstorleken $n = 15$

Ett 95-procentigt konfidensintervall ges av

$$\bar{x} \pm z_{(1-0.05/2)} \frac{\sigma}{\sqrt{n}} = 36 \pm 1.96 \frac{12}{\sqrt{15}} = 36 \pm 6.07$$

vilket vi kan beräkna till $(29.93, 42.07)$

Ett 99-procentigt konfidensintervall ges av

$$\bar{x} \pm z_{(1-0.01/2)} \frac{\sigma}{\sqrt{n}} = 36 \pm 2.576 \frac{12}{\sqrt{15}} = 36 \pm 7.98$$

eller $(28.02, 43.98)$

Konfidensintervall med skattad standardavvikelse

Om standardavvikelsen är okänd kan vi skatta den från stickprovet

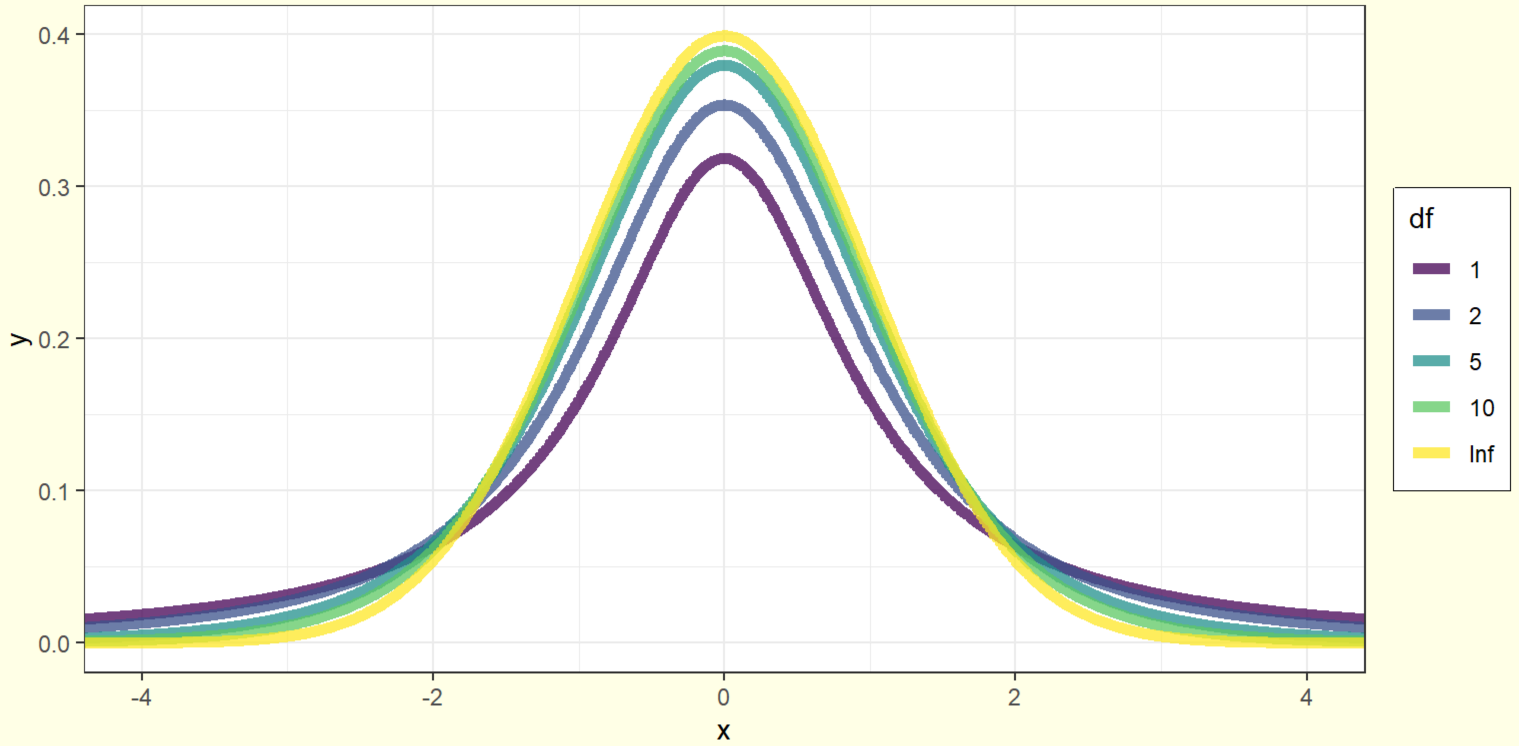
Eftersom skattningen är mindre säker är det intutivt att vi behöver ett bredare konfidensintervall

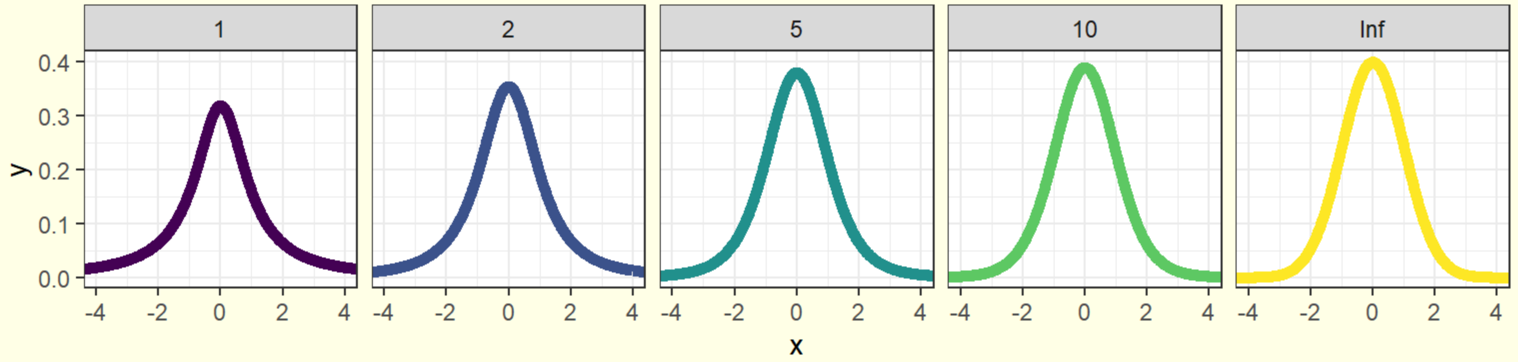
Den teoretiska konsekvensen är att vi använder en *t-fördelning* istället för en normalfördelning

t-fördelningen är en kontinuerlig fördelning med en ingående parameter - *antalet frihetsgrader* (eng. degrees of freedom, df)

t-fördelningen uppstår när man standardiserar ett medelvärde med en skattad standardavvikelse:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$





Frihetsgrader

Frihetsgrader är ett statistiskt begrepp för en datamängds storlek efter påförd begränsning

När vi skattar s använder vi serien $x_i - \bar{x}$

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

Den serien har egenskapen att den alltid summerar till noll (en begränsning)

Vid skattningen av s använder vi alltså en serie som innehåller *mindre* information än våra n observationer

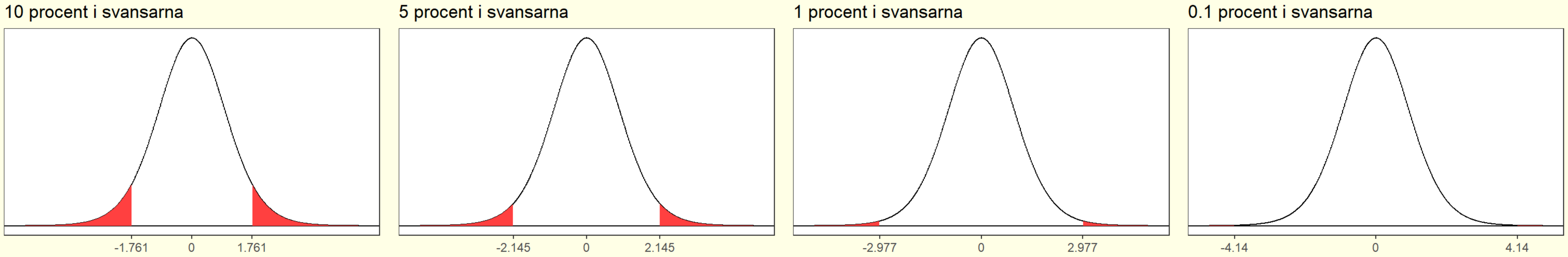
Antalet frihetsgrader ges av antalet observationer minus antalet begränsningar, så vid skattningen av s har vi $n - 1$ frihetsgrader

Om vi standardiserar med en skattad standardavvikelse får vi en t-fördelning med $n - 1$ frihetsgrader

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

Biometri, tabell 5

Vid en skattad standardavvikelse baseras konfidensintervallet på en t-fördelning med $n-1$ frihetsgrader



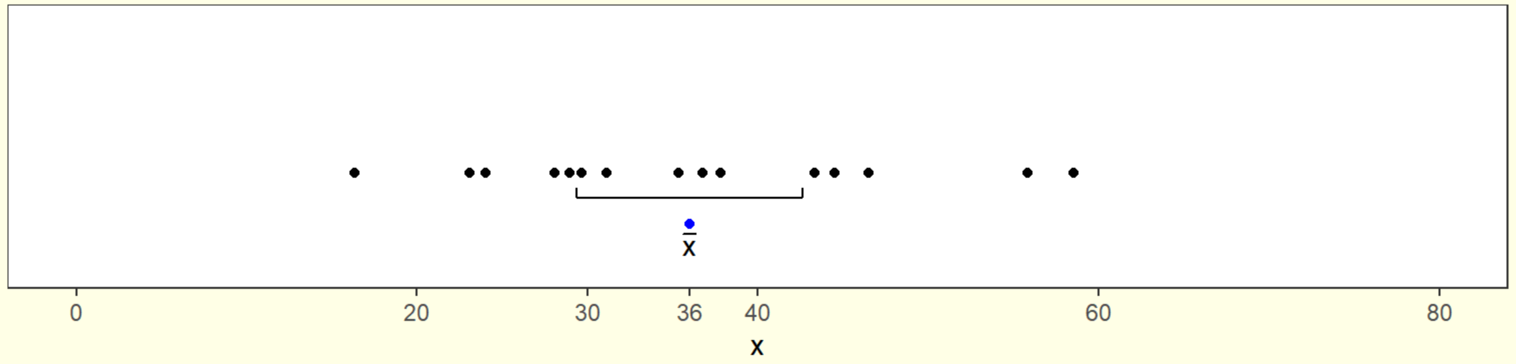
Tabellvärden för x-axeln betecknas $t_{(1-\alpha/2,df)}$ och kan hämtas från en tabell över t-fördelningen, t.ex tabell 5 i *Biometri*

	F(t)								
d.f.	0.75	0.90	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140

För ett tabellvärde motsvarande fem procent i svansarna och $n = 15$ tittar vi på $t_{(1-0.05/2,15-1)} = t_{(0.975,14)} = 2.145$

Den sista raden i tabell 5 ger värden för en t-fördelning med oändligt antal frihetsgrader, dvs en standardiserad normalfördelning

Konfidensintervall, skattad standardavvikelsen, exempel



Som tidigare är det skattade medelvärdet $\bar{x} = 36$ och stickprovsstorleken $n = 15$, men standardavvikelsen är nu skattad $s = 12$

Ett 95-procentigt konfidensintervall ges av

$$\bar{x} \pm t_{(1-0.05/2,14)} \frac{\sigma}{\sqrt{n}} = 36 \pm 2.145 \frac{12}{\sqrt{15}} = 36 \pm 6.65$$

vilket vi kan beräkna till $(29.35, 42.65)$

På grund av den skattade standardavvikelsen hämtar vi tabellvärdet från en t-fördelning och får ett bredare konfidensintervall

Termen $\frac{s}{\sqrt{n}}$ kallas *medelfelet* (eng. standard error, SE)

Tolkning av konfidensintervall

Konfidensintervallet brukar presenteras som att det täcker populationsmedelvärdet $\boldsymbol{\mu}$ med en given konfidens (ofta 95 procent)

En ny undersökning kan ge andra skattningar av $\bar{\boldsymbol{x}}$ och $\boldsymbol{s}$, och således ett annat konfidensintervall

Om vi upprepar försöket ett stort antal gånger kommer ett 95-procentigt konfidensintervall täcka populationsmedelvärdet 95 procent av gångerna

Märk särskilt att konfidensintervallet är ett intervall för medelvärdesskattningen

Intervallet ska **inte** tolkas som att 95 procent av observationerna ligger i intervallet

Slutsatser från data

Konfidensintervallet kan användas till att dra slutsatser från sin data

Värden som inte ligger i intervallet är mindre troliga att vara det sanna populationsmedelvärdet

En relaterad metod för att dra slutsatser är *hypotestestet*

Vid ett hypotestest jämförs en specifik hypotes med observerad data

Ett resultat är *signifikant* om observationerna är osannolika givet hypotesen

Hypotestestens steg

Formell gång för att ställa observation mot hypotes

1. Hypoteser
2. Testfunktion
3. Testfördelning
4. p-värde (beräkning eller uppskattning)
5. Slutsats

HTTPS

Val av hypotestest

Typen av variabel påverkar det slumpmässiga beteendet

och man får därför skilda hypotestest för skilda variabeltyper

Under kursen tittar vi på

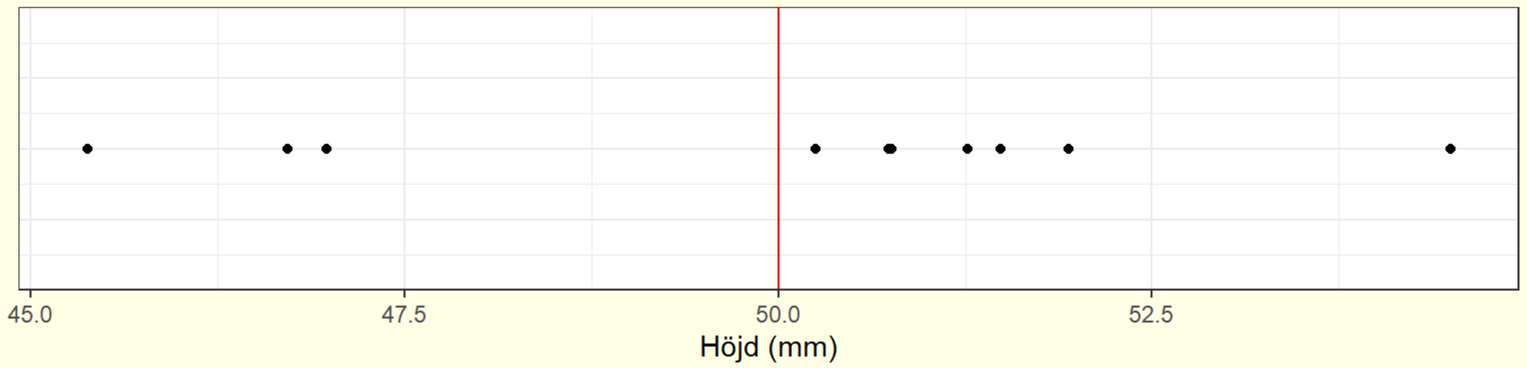
- t-test. Kontinuerlig variabel. En eller två grupper
- Binomialtest och z-test. För binär data (två möjliga utfall)
- χ^2-test. För nominella variabler
- F-test. Kontinuerlig variabel. Två eller fler grupper

t-test

t-testen är test för **kontinuerliga** variabler som följer en **normalfördelning**

t-test för ett stickprov

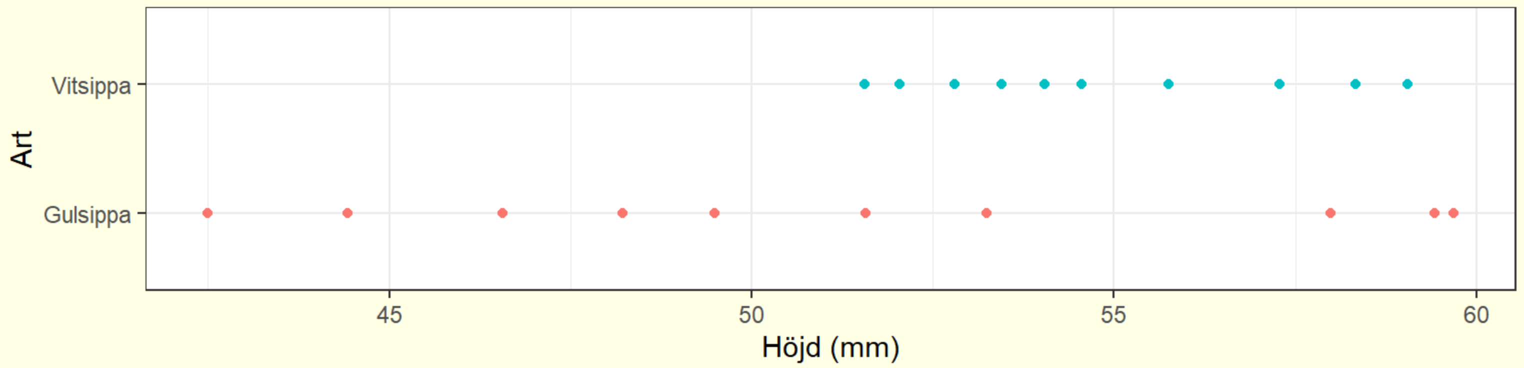
Testar om populationsmedelvärdet är skilt från en specifik nollhypotes



Är populationsmedelvärdet skilt från 50?

t-test för två stickprov

Testar om det finns någon skillnad i populationsmedelvärde mellan två populationer



Är populationsmedelvärdena skilda från varandra?

t-test för ett stickprov

Ett stickprov om n observationer

Skattar medelvärdet $\bar{x}$ och standardavvikelsen s

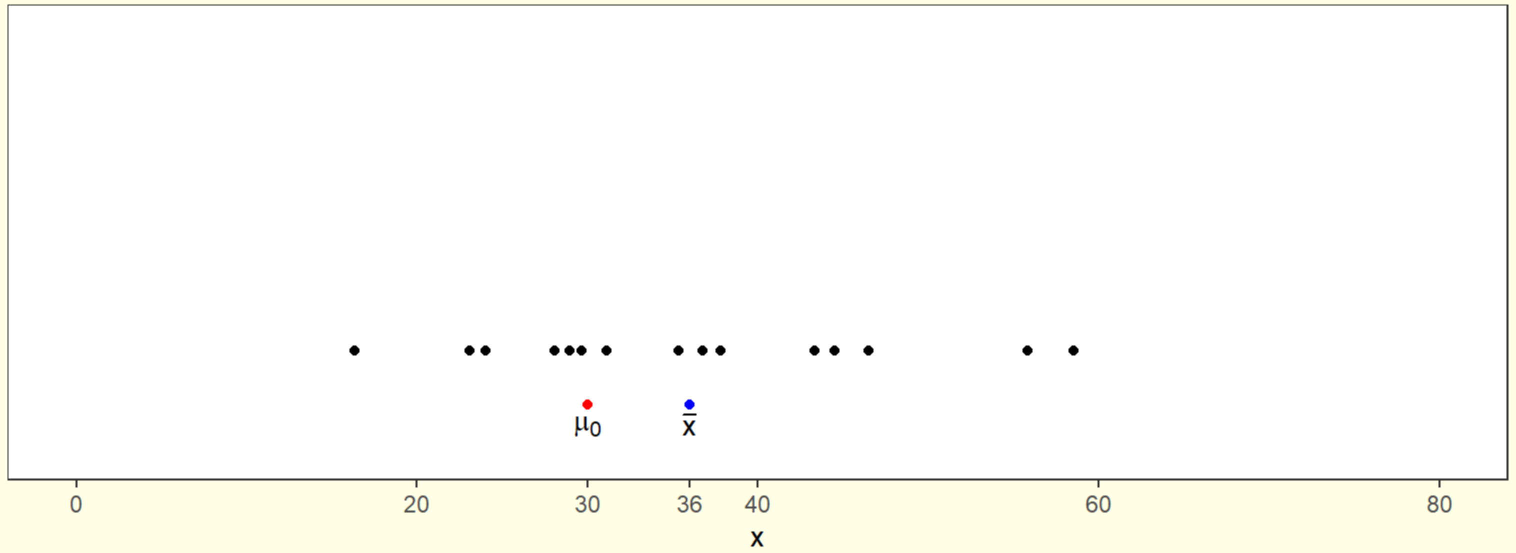
Vill ställa det mot en nollhypotes H_0

Numeriskt exempel

Man har identifierat en genotyp av pärlhyacint och vill veta om dess höjd vid blomning är skild från 30 cm

Man sätter femton exemplar och får $\bar{x} = 36$ och $s = 12$ ($s^2 = 144$)

Illustration



Vår data kommer från en fördelning med något okänt populationsmedelvärde μ

Vi har en *nollhypotes* att μ är lika med 30, $\mu_0 = \mathbf{30}$

Vad är sannolikheten för det här utfallet om medelvärdet faktiskt är 30?

Om den sannolikheten är låg, tyder det på att det sanna populationsmedelvärdet är skilt från 30

Hypoteser

Nollhypotes och alternativhypotes

$H_0 : \mu = 30$

$H_1 : \mu \neq 30$

Testfunktion

t-testets testfunktion ges av att standardisera mot nollhypotesens värde

$t = \frac{\bar{x}-\mu_0}{\sqrt{s^2/n}}$

Sätter in värden i testfunktionen

$t = \frac{36-30}{\sqrt{144/15}} = 1.936$

Testfördelning

Om nollhypotesen stämmer ska t komma från en t-fördelning med $n-1$ frihetsgrader

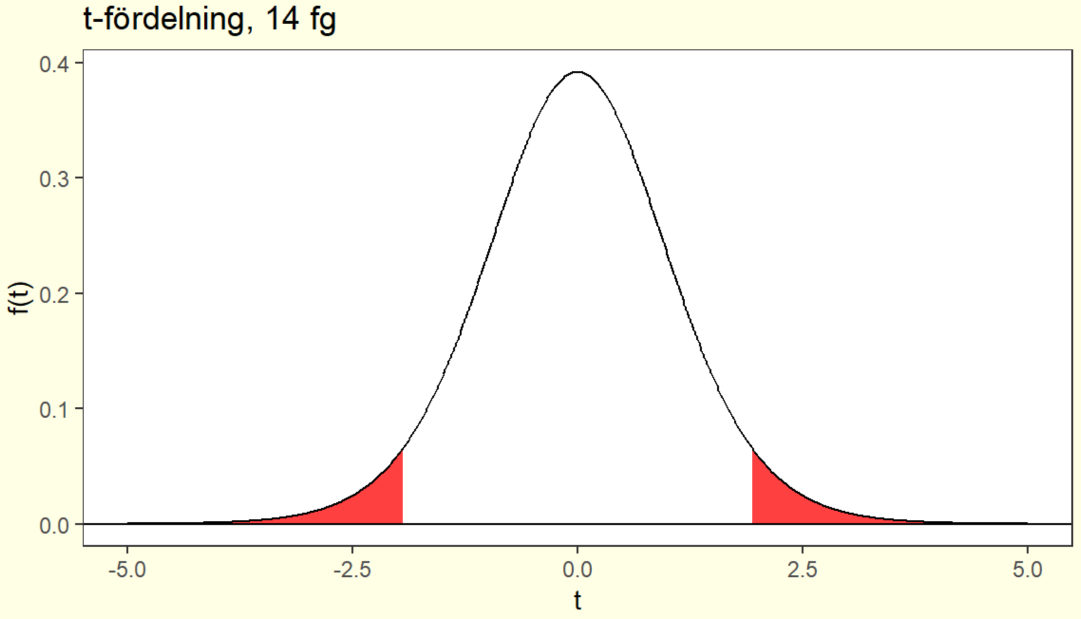
p-värde

p-värdet beräknas som ytan bortom t i en t-fördelning med $n - 1 = 14$ frihetsgrader

Eftersom alternativhypotesen är tvåsidig medtas bägge svansarna

Man kan beräkna p-värdet i valfritt datorprogram och få $p = 0.0733$

Eller titta i tabell 5, rad 14. Eftersom t ligger mellan 1.761 och 2.145 måste det motsvara ett tvåsidigt p-värde mellan 10 och 5 procent

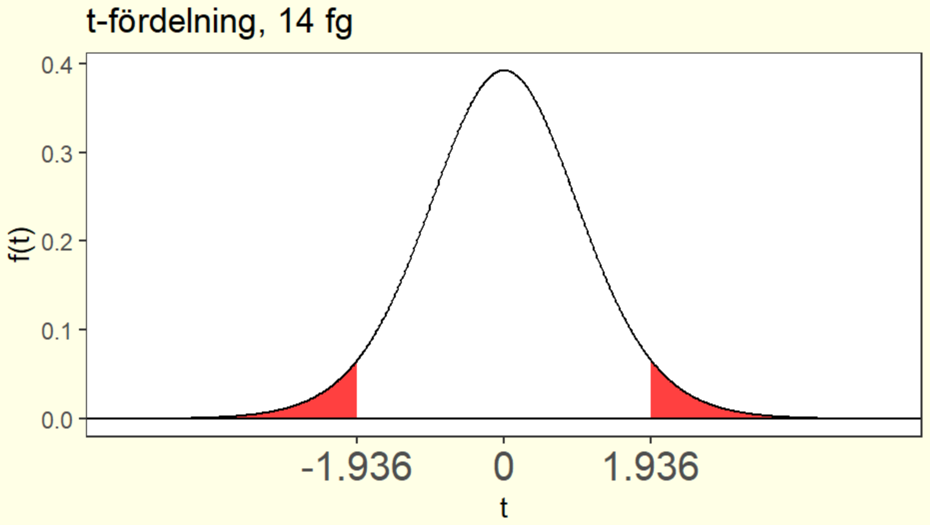


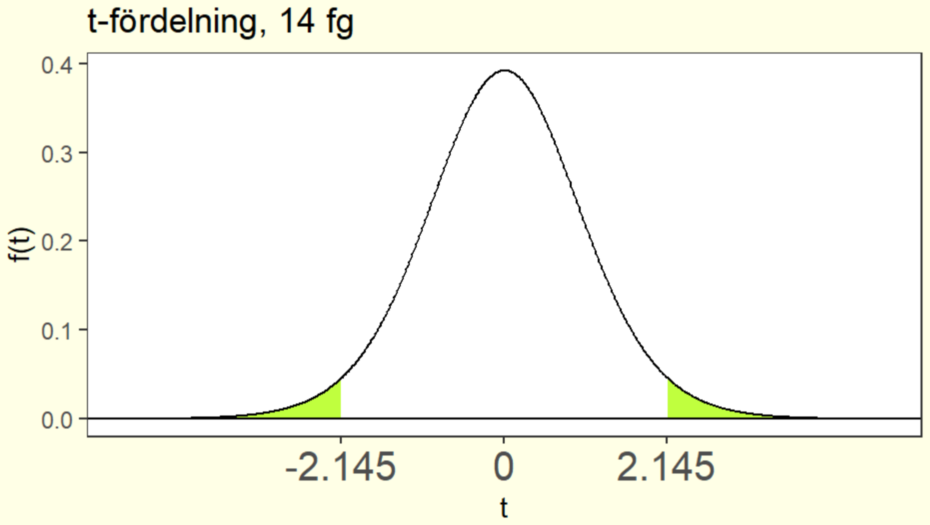
Biometri, tabell 5

Tabell 5 i *Biometri* anger t-fördelningens fördelningsfunktion

En sannolikhet anges i översta raden och ett antal frihetsgrader anges i vänstra kolumnen

Tabellens inre värde ger ett t-värde sådant att sannolikheten (*ytan under kurvan*) till vänster om det värdet ges av sannolikheten





Den gröna arean motsvarar fem procent (2.5 procent i vardera svans)

Eftersom vårt observerade t-värde 1.936 ligger närmre nollan måste *svanssannolikheten* (den röda arean) vara minst fem procent

278 *Tabell 5: t-fördelningens fördelningsfunktion*

Tabell 5: t-fördelningens fördelningsfunktion

	F(t)								
d.f.	0.75	0.90	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
∞	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.092	3.291

Exempel: För en t-fördelning med 10 frihetsgrader (d.f.) gäller det att $P(t \leq 1.812) = 0.95$. Fördelningen är symmetrisk, så för en t-fördelning med 10 frihetsgrader (d.f.) gäller det att $P(t \leq -1.812) = 0.05$.

	F(t)								
d.f.	0.75	0.90	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140

Vid 14 frihetsgrader är ytan under kurvan till vänster om 2.145 lika med 0.975

eller

$$t_{(0.975,14)} = 2.145$$

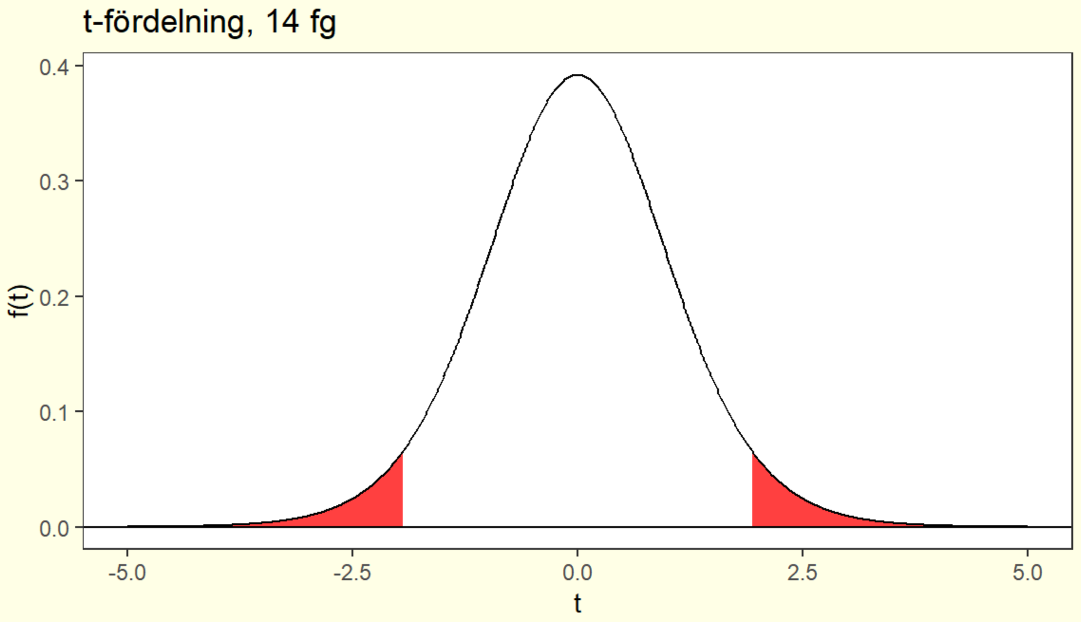
p-värde

p-värdet beräknas som ytan bortom t i en t-fördelning med $n - 1 = 14$ frihetsgrader

Eftersom alternativhypotesen är tvåsidig medtas bägge svansarna

Man kan beräkna p-värdet i valfritt datorprogram och få $p = 0.0733$

Eller titta i tabell 5, rad 14. Eftersom t ligger mellan 1.761 och 2.145 måste det motsvara ett tvåsidigt p-värde mellan 10 och 5 procent



Slutsats

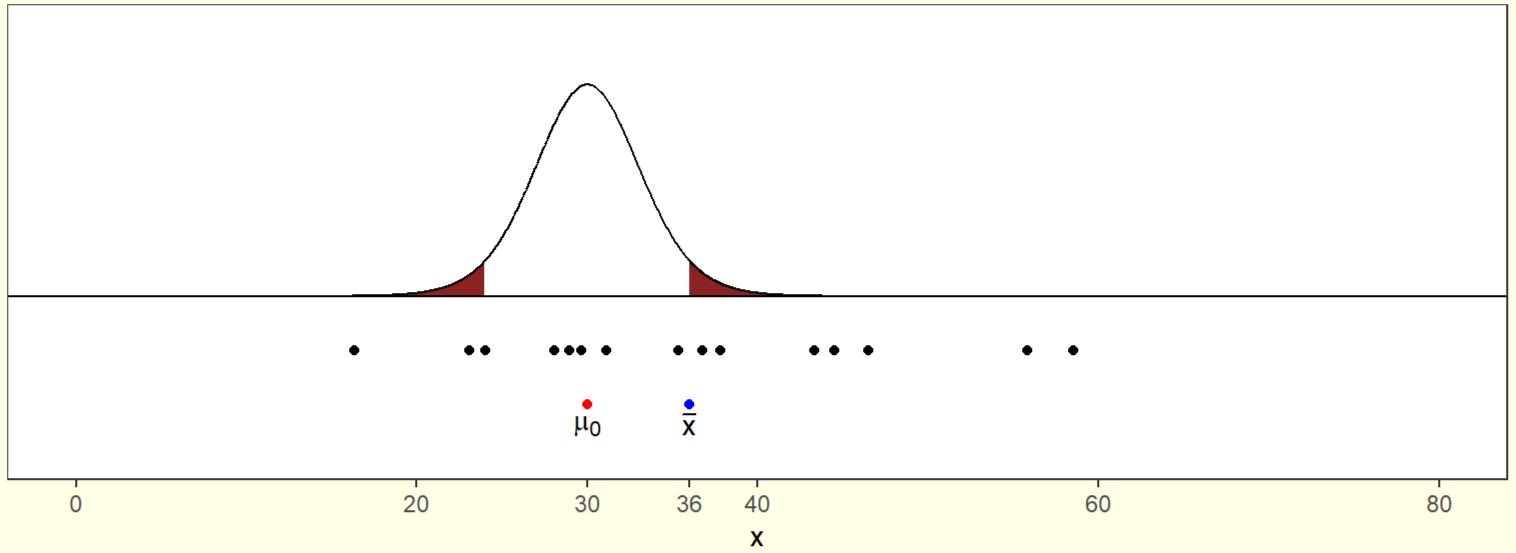
Vid slutsatsen ställs p-värdet mot en *signifikansnivå*

Signifikansnivån (α) är en förbestämd nivå från vilken man förkastar eller inte förkastar nollhypotesen

Oftast används 5, 1 eller 0.1 procent som α

Då vi har ett p-värde över fem procent kan vi *ej påvisa någon säkerställd skillnad från höjden* 30 *cm*

Illustration



Det tvåsidiga p-värdet är arean i svansen bortom $\bar{x}$, gånger två

Arean ger sannoliketen för det observerade medelvärdet om nollhypotesen stämmer

Ett lågt p-värde tyder på att nollhypotesen *inte* stämmer

Koppling mellan konfidensintervall och hypotestest

Ett konfidensintervall täcker de värden som **inte** förkastas om de används som nollhypotes i motsvarande hypotestest

Om ett visst värde ligger i ett **95**-procentigt konfidensintervall (*intervallet täcker värdet*) kommer man inte förkasta nollhypotesen att medelvärdet är det värdet på signifikansnivån **5** procent

t-test för ett stickprov, schema

Hypoteser

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

Testfunktion

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

där $\bar{x}$ och s skattas från stickprovet, och μ_0 hämtas från nollhypotesen

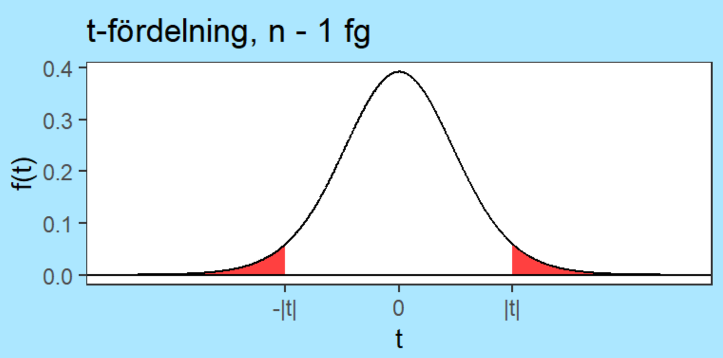
Testfördelning

Under nollhypotesen följer t en t-fördelning med $n - 1$ frihetsgrader

P-värde

P-värdet ges av arean bortom $|t|$ i testfördelningen

Vid handräkning uppskattas p-värdet genom att ställa $|t|$ mot ett tabellvärde



Svar

P-värdet ställs mot en förbestämd *signifikansnivå* (ofta 5 procent)

Vid ett lågt p-värde förkastas nollhypotesen

Vid ett högt p-värde förkastas ej nollhypotesen

Avslutande punkter

Antaganden

t-test för ett stickprov bygger på

- normalfördelning
- oberoende observationer inom stickprovet

Brister i antaganden

t-testet sägs vara *robust* och används ofta i fall där normalfördelningsantagandet inte klart gäller t.ex. för heltalsdata

Stjärnsystemet för signifikans

Symbol	**Nivå**
n.s	Ej sig
*	5%
**	1%
***	0.1%

Typ I-fel och Typ II-fel

För ett påstående finns två möjliga verkligheter (sant/falskt) och två möjliga slutsater (stämmer/stämmer inte)

	Hypotes sann	**Hypotes falsk**
Tror sann	Korrekt	Typ II-fel
Tror falsk	Typ I-fel	Korrekt

Sannolikheten för ett Typ I-fel är lika med signifikansnivån α

Slut