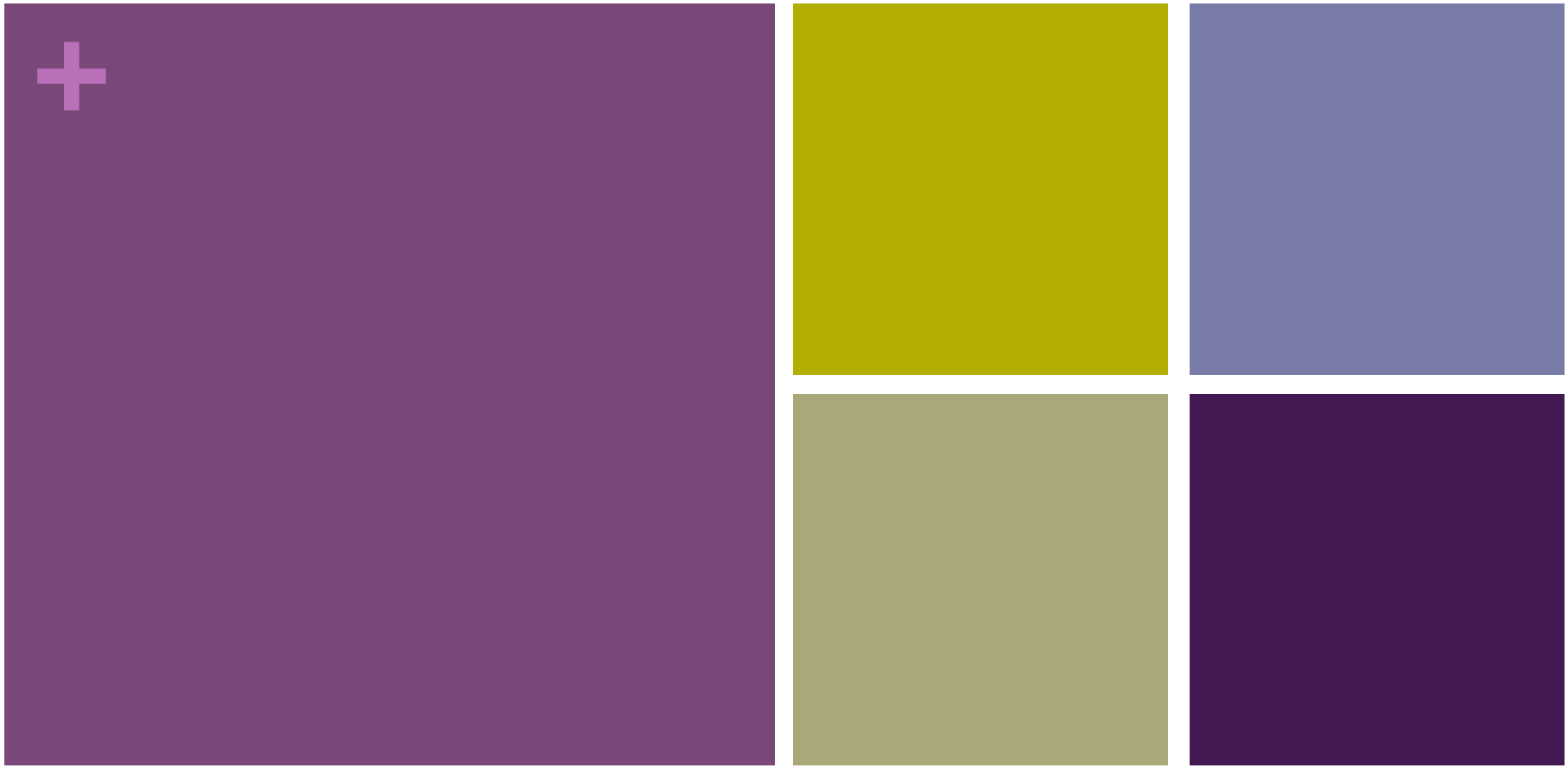


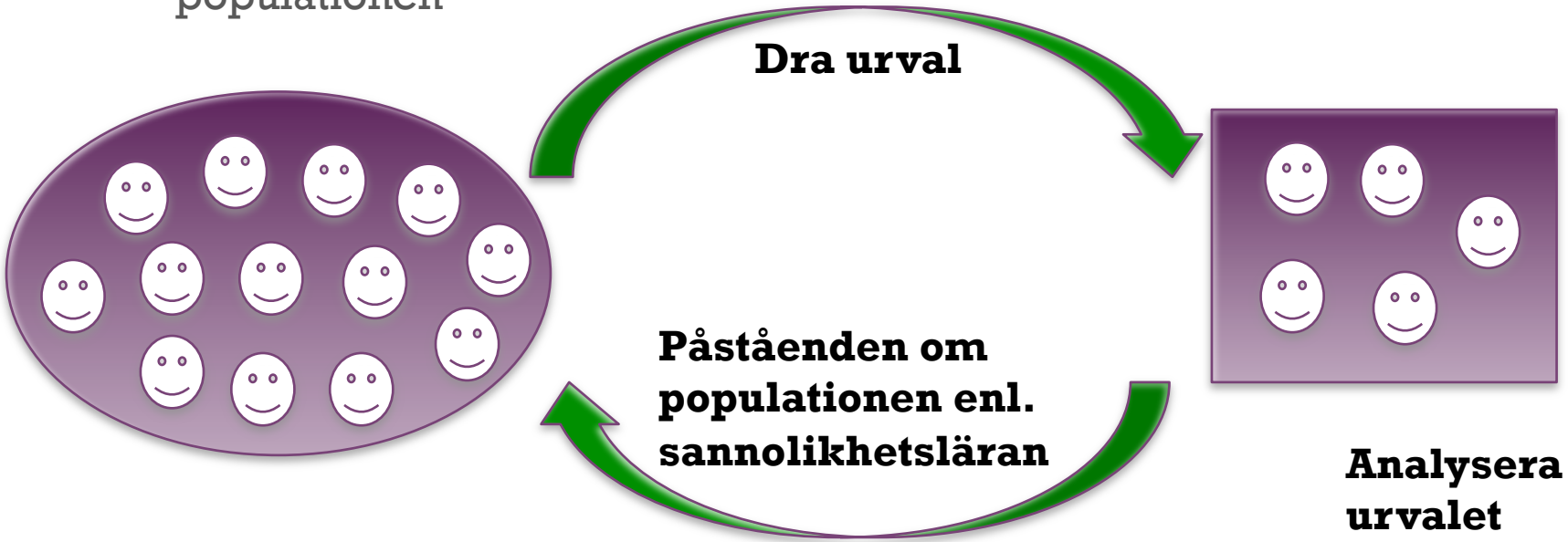
Statistiska analyser C2
Inferensstatistik

Wieland Wermke

Signifikans och Normalfördelning

Problemet med generaliseringen: inferensstatistik

- Om vi vill veta ngt. om en population, då kan vi ju fråga hela populationen.
  - men det går ju för det mesta inte!
  - därför använder man sig av inferensstatistik (slutande statistik)
  - från ett oberoende slumpmässig urval (OSU) slutar man på själva populationen



Statistisk signifikans

- När man hävdar ett samband som signifikant, påstår man att sambandet gäller också i andra urval ur samma population **(läs. ”gäller populationen” (t.ex. svenska folket))**
  - samband kan vara skillnader mellan olika grupper
  - eller kan vara korrelationer mellan två fenomen i ett urval
- Anger sannolikheten att hävda att ett samband är systematiskt (stämmer för populationen), och inte bara är slumpmässigt (ett tillfälligt resultat)
- **Signifikansen säger dock inget om hur stor effekten är, dvs. hur relevant den är (signifikans är inte relevans)!!**

Signifikansnivå och sannolikhet (p)

- Signifikansnivå
  - med en 95%/99/99.9 % vill vi vara säkra att att vår beslut inte bygger på slump, utan att ett samband existerar "verkligen" för vår population (ju högre ju bättre)
- Sannolikheten [p (eng. probability), sig., α] anger hur sannolik det är att en skillnad/samband beror på slumpen (ju mindre ju bättre)
  - Ex. Sig.=0.23 → 23% (23 av 100) sannolikhet att skillnaden är slumpmässig → resultatet är inte signifikant på en nivå av 95%
  - Sig.=0.04 → 4% sannolikhet att skillnaden är slumpmässig –> signifikant på 95%, men inte 99% nivå
  - Sig.=0.001 → 0,1% (1 av 1000) → signifikant på 99% nivå

Ett viktigt verktyg, om vi har intervallnivå: Normalfördelning

- Grunden till undersökningar om ngn effekt är signifikant eller inte: **Normalfördelning**
- **Normalfördelning står för hela populationen, som vi inte känner!**
  - **Vi antar bara hur den måste se ut!!!**
- Det är ett **analytiskt hjälpmedel** för att kunna genomföra undersökningar (sannolikhetsläran)!
  - Hur sannolikhet är ngt (om man kastar tärningar, singlar slant)

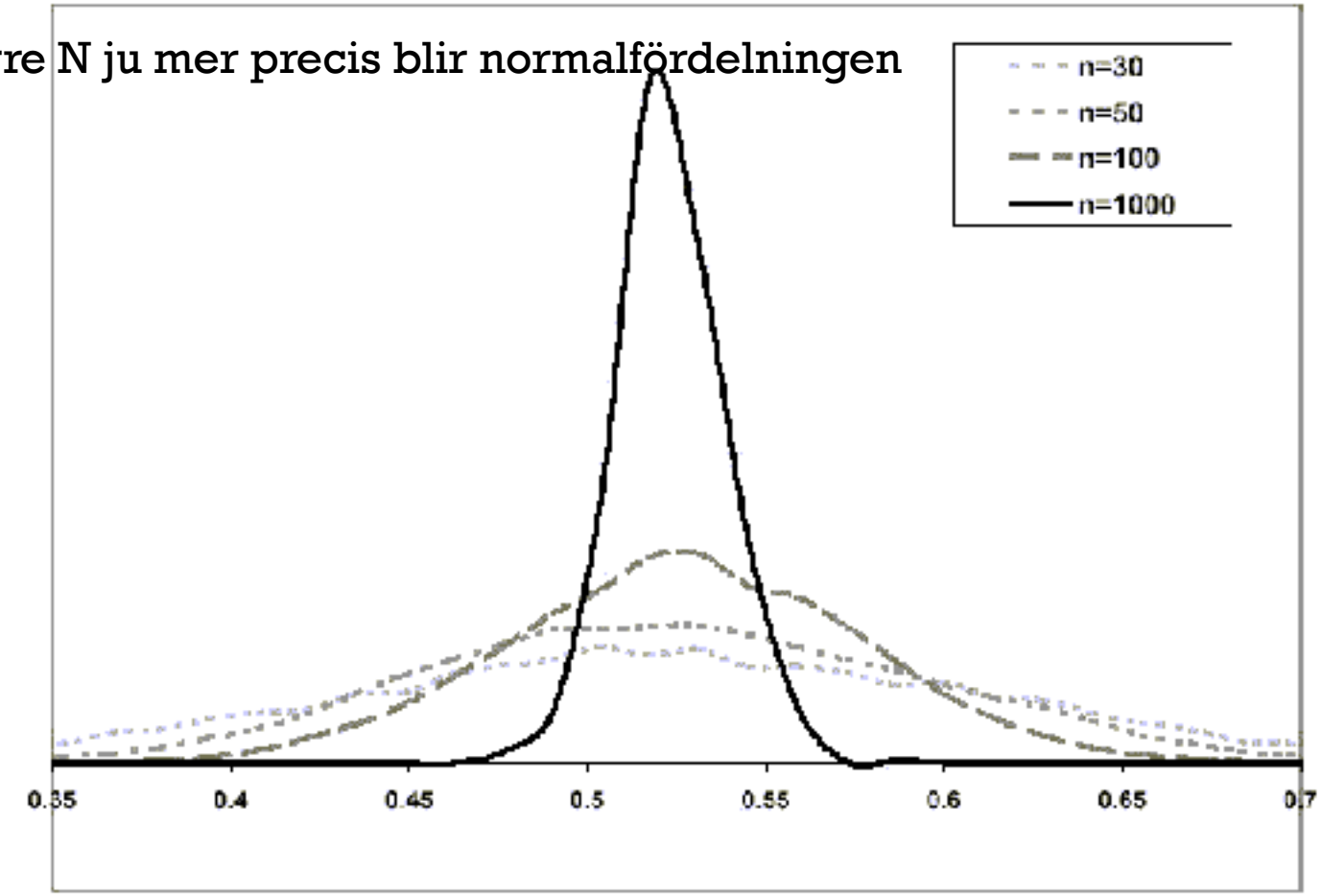
Sannolikhetlära: Centrala gränsvärdsatsen

- **Centrala gränsvärdessatsen**: En urvalsfördelning, är den bara tillräckligt stor, följer normalfördelningen
  - 20-25 är det minsta antalet av observationer, där vi kan utgå från en normalfördelning (ju mer observationer ju mer tar kurvan formen)
- Normalfördelningskurvan bygger helt och hållet på två värden:
  - **Medelvärdet** (centralmått)
  - **Standardavvikelsen** (spridningsmått)

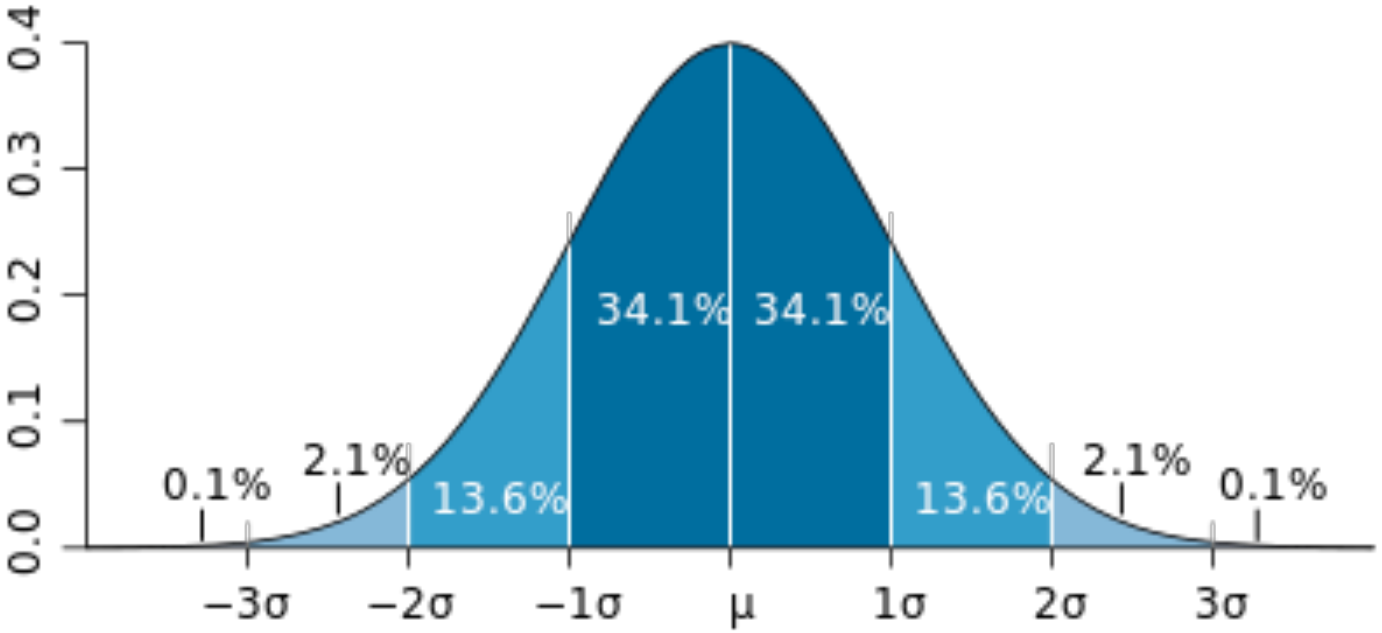
Exempel: Andel män i n stickprov

Ju närmare ni kommer en normalfördelning med era data, ju mer precis blir skattningen av resultatet!

Ju större N ju mer precis blir normalfördelningen



Normalfördelning



Om en fördelning är normalfördelad, kan vi uttala oss med vilken chans en observation är del i själva populationen, om den har vissa egenskaper (standardavvikelse (sigma) från medelvärden (my))

Standardisering

- NE: Standardisering tar främst sikte på att underlätta kommunikation genom att skapa entydiga begrepp med definitioner och termer
- **att säkerställa *utbytbarhet och kompatibilitet genom fastläggande av mått, dimensioner, storlekar och gränssnitt***
- att åstadkomma variantbegränsning genom urval av mått, dimensioner
  (relationen blir viktig, inte det konkreta värdet)
- → **Centrering rund medelvärdet**

$$z_i = \frac{x_i - \bar{x}}{s}$$

Standardisering

- NE: Standardisering tar främst sikte på att underlätta kommu
  definit
- att säk
  fastläg
- att åst
  dimen
- → gör

Allt koncentreras omkring det sanna medelvärdet (som vi antar pga normalfördelningen)

med **95%** sannolikhet (p=.05) är värdet inom **-/+ 1,96 z värde (-/+)2sd**

med **99%** sannolikhet (p=.01) är värdet inom **-/+ 2,58 z värde**

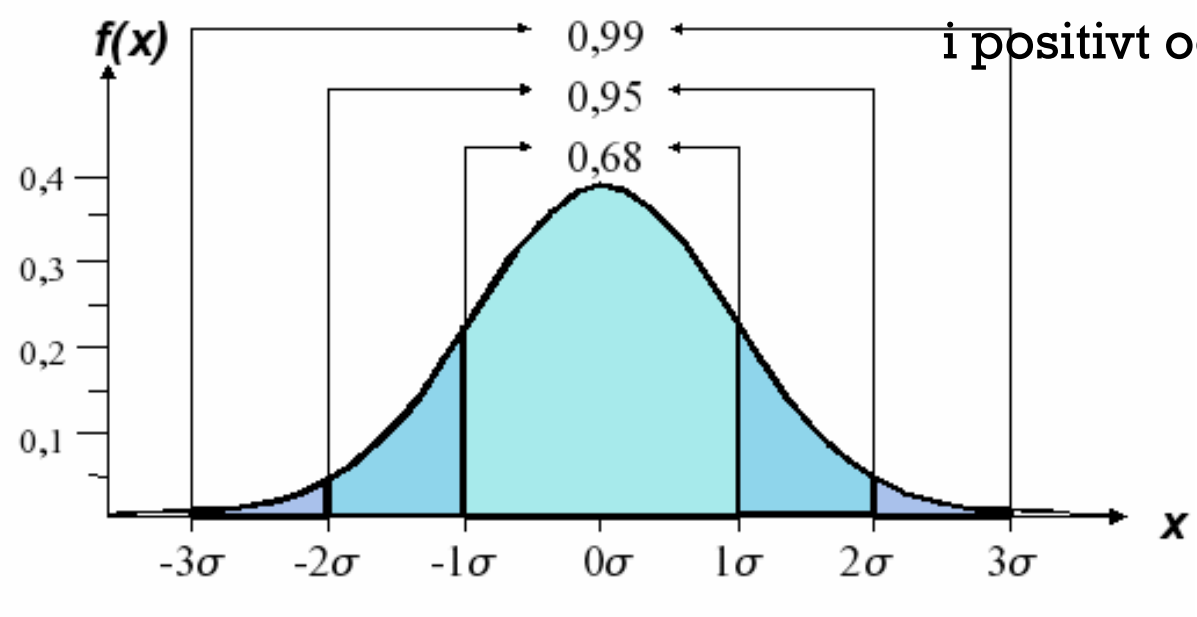
med **99,9%** sannolikhet (p=.001) är värde inom **-/+ 3,29 z värde**

Signifikanstestning

Normalfördelningen (igen)

Anger realiteten som vi tänker oss den!

Det finns medelvärde och standardiserade avvikelse i positivt och negativt håll



Stickprovsfördelning

- Varje fördelning (med tillräcklig storlek) följer normalfördelningens form (centrala gränsvärdessatsen)
- → om man drar ett antal urval för att kolla hur ofta ett fenomen dyker upp i varje urval, följer denna **stickprovsfördelning** också normalfördelningens form
  - här får vi ett **medelvärde över alla medelvärden** från n urval
  - och en **standardavvikelse** som i detta sammanhang kallas **standardfel** från n urval
  - Om urval verkligen är slumpmässiga! (Validitet/Reliabilitet)

Signifikant skillnad mellan två medelvärde

- Om skillnaden mellan två gruppers medelvärde är systematiskt och fri från slump, så kallar vi **skillnaden signifikant**

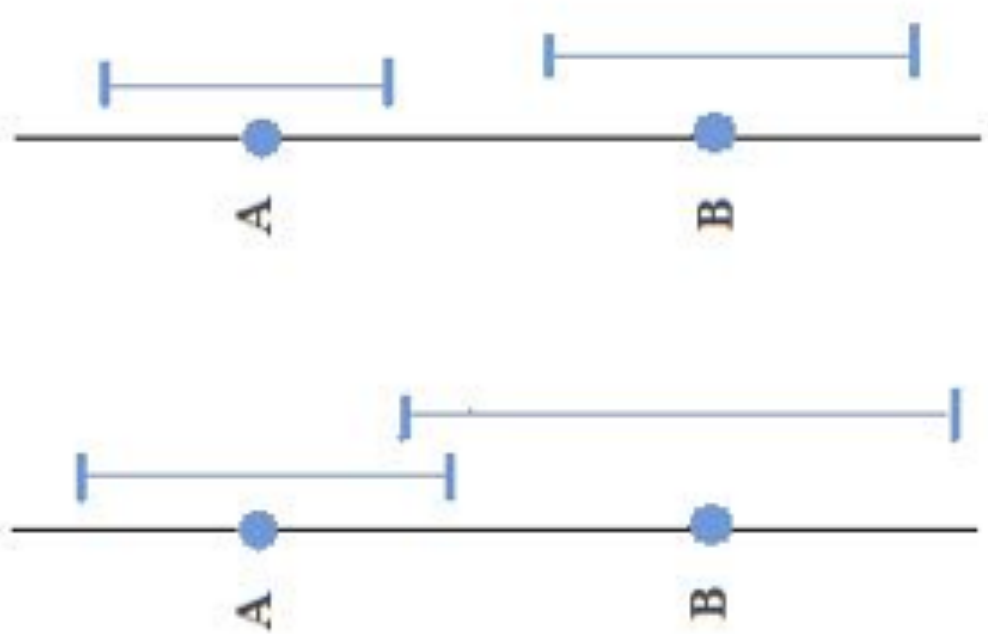
- För signifikanstestningen arbetar vi med två **hypoteser (antaganden om resultatet)**
  - **nollhypotes**: Det finns inte någon systematisk skillnad mellan två grupper
  - **mothypotes/alternativhypotes**: Det finns en systematisk skillnad mellan två grupper

Signifikanstestning

- Grundidén: vi antar en **normalfördelning**, och vi vet urvalets medelvärde/andel samt dess standardavvikelse
- Vi bestämmer själva signifikansnivån (95 %, 99 %, 99,9 %)
  - total säkerhet finns inte!
  - **båda urvalen borde vara slumpmässiga (OSU)**
- Vi räknar ut konfidensintervallen för de två grupper vi jämför
- Vi antar eller förkasta nollhypotesen, om konfidensintervallen överlappar eller inte överlappar varandra!

Signifikanstestning

- Grundidén: vi antar en **normalfördelning**, och vi vet urvalets medelvärde/andel samt dess standardavvikelse
- Vi b
  -
  -
- Vi rä
- Vi a
  konf
  vara



Signifikanstestning: nominala data: Chi två test

- Hypotestestning om det finns en signifikant skillnad mellan två underurval angående en viss egenskap (används för **data på nominalnivå**)
  - **H_0 Det finns inte någon systematisk (icke-slumpmässig) skillnad mellan de två grupperna. Skillnaden är slumpmässig**
  - **H_1 Det finns systematiska skillnader. Skillnaden är inte slumpmässig**
  - Vid en viss signifikansnivå (95 %,99 %,99.9 %)

Signifikanstestning: nominala data: Chi två test

- Man jämför i en **korstabell** det **förväntade värdet** (om värden skulle vara fördelade utan något inflyttande alls) med det **observerade värdet**
- Förväntande värdet = (radsumma * kolumnsumma)/n

Signifikanstest, Chi två test

	Kategori 1 (beroende var.)	**Kategori 2 (beroende var.)**	**Rad-summa**
Kategori 1 (oberoende var.)	*cell a* förväntat värde observerat värde	*cell b* förväntat värde observerat värde	
Kategori 2 (oberoende var.)	*cell c* förväntat värde observerat värde	*cell d* förväntat värde observerat värde	
Kolumnsumma			

Förväntat värde = (radsumma * kolumnsumma)/n

Frihetsgrader?

- Betecknar antalet av ett objekts fritt välbara rörelsemöjligheter, som är oberoende från varandra i ett system
- Om vi vet N eller medelvärdet och alla värde förutom en, så är den inte fri, de andra värde bestämmer den!
  - 1+2+3+4+x=10 x?
- Fyrfältare har **4 celler**, för att kunna definiera tabellen behövs det dock vissa parametrar, kategorier med svarsalternativ (2), antal (1), → df=1
  - df (för större tabeller (mer än två kategorier) =
    - **(antal råder – 1) * (antal kolumner – 1)**

Signifikanstest, Chi två test

- Räkna ut chi två värde :
  - $\chi^2 = \sum \frac{(Obs_{i,j} - Förv_{i,i})}{Förv_{i,i}}$ (Summa över varje tabellcell a+b +c+d)

- **Om chi värdet är olika 0, då finns det en skillnad, nu måste den ”signifikanstestas”**

- Chi två värdet kan i relation till frihetsgraderna och en innan fastlagt signifikansnivå hittas i en **chi två tabell**
  - …som visar kritiska värden, dvs. minsta värde där en skillnad är signifikant vid n frihetsgrader
  - t.ex. sannolikhet för nollhypotesen = .05 = signifikansnivå 95 %)

Chi två tabell

- jämför våra chi två värden med fördelningsvärden
- om den är ≥ värdet ur tabellen, så är värdet signifikant
- första kolumnen=frihets-graderna, översta råden=signifikans nivå

n \ γ	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	7,879	6,635	5,034	3,841	2,706	$1,58^{-3}$	$3,93^{-4}$	$9,82^{-4}$	$1,57^{-5}$	3,93
2	10,60	9,210	7,378	5,991	4,605	0,211	$0,103^{-2}$	$5,06^{-2}$	$2,01^{-2}$	1,00
3	12,84	11,34	9,348	7,815	6,251	0,584	0,352	0,216	$0,115^{-2}$	7,17
4	14,86	13,28	11,14	9,488	7,779	1,064	0,711	0,484	0,297	0,207
5	16,75	15,09	12,83	11,07	9,236	1,610	1,145	0,381	0,554	0,412
6	18,55	16,81	14,45	12,59	10,64	2,204	1,635	1,237	0,872	0,676
7	20,28	18,48	16,01	14,07	12,02	2,833	2,167	1,690	1,239	0,989
8	21,96	20,09	17,53	15,51	13,36	3,490	2,733	2,180	1,647	1,344
9	23,59	21,67	19,02	16,92	14,68	4,168	3,325	2,700	2,088	1,735
10	25,19	23,21	20,48	18,31	15,99	4,865	3,940	3,247	2,558	2,156
11	26,76	24,73	21,92	19,68	17,28	5,578	4,575	3,816	3,053	2,603
12	28,30	26,22	23,34	21,03	18,55	6,304	5,226	4,404	3,571	3,074
13	29,82	27,69	24,74	22,36	19,81	7,042	5,892	5,009	4,107	3,565
14	31,32	29,14	26,12	23,68	21,06	7,790	6,571	5,629	4,660	4,075
15	32,80	30,58	27,49	25,00	22,31	8,547	7,261	6,262	5,229	4,601
16	34,27	32,00	28,85	26,30	23,54	9,312	7,962	6,908	5,812	5,142
17	35,72	33,41	30,19	27,59	24,77	10,09	8,672	7,564	6,408	5,697
18	37,16	34,81	31,53	28,87	25,99	10,86	9,390	8,231	7,015	6,265
19	38,58	36,19	32,85	30,14	27,20	11,65	10,12	8,097	7,633	6,844
20	40,00	37,57	34,17	31,41	28,41	12,44	10,85	9,591	8,260	7,434

Exempel SPSS

- I vår material: skiljer sig studenter som är motiverat för kursen i sin temtamensangest
  - två dikotomiserade variabler:
  - *Motivation dikot.* (inte motiverat: <5, motiverat ≥ 5)
  - *Tentamensångest dikot.* (ingen ångest ≤ 2, ångest > 2)

- ***H_0: Båda grupper skiljer sig inte, H_1:Båda grupper skiljer sig***

Motivation dikotomisiert * Tentamensångest dikotomiserat Crosstabulation

Count

		Tentamensångest dikotomiserat		Total
		ingen/lite tentamensångest	tentamensångest/ingen tentamensångest	
Motivation dikotomisiert	inte/lite motiverat	4	14	18
	mer/mycket motiverat	8	26	34
Total		12	40	52

Exempel SPSS

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.011[a]	1	.915		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.011	1	.915		
Fisher's Exact Test				1.000	.601
Linear-by-Linear Association	.011	1	.916		
N of Valid Cases	52				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.15.

b. Computed only for a 2x2 table

Chi två värde

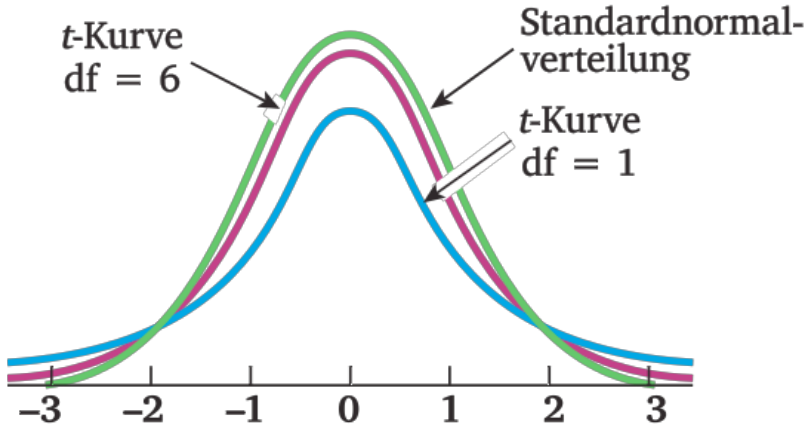
frihetsgrader
alltid 1 i en fyrfältare

Signifikans (p)

p/sig = .915: Chansen att skillnaden inte är signifikant/slumpartad = 91,5% procent, Nollhypotesen gäller

T test

- Undersöka hypoteser, där medelvärde är känt , men varians inte (→ används för intervallskalnivåer)
- Bygger på t fördelningen, som är också en fördelningskurva, dock fungerar bra för mindre urval (=>30), med växande n motsvarar t en normalfördelningen



T tests användning

- En grupps T test
  - Vi har undersöka om en grupps medelvärde skiljer sig signifikant från ett medelvärde man anger
- Oberoende T test
  - Skiljer sig medelvärdet av två grupper sig signifikant, som undersöktes samtidigt (män/kvinnor)
- Beroende T test
  - Skiljer sig medelvärdet av två grupper som är beroende av varandra i datamaterialet (mammor/söner eller, medelvärde i en longitudinell studie)
    - *kommer inte vara med idag*

Oberoende T test

- Skiljer sig tjänsteman barn och arbetarbarn signifikant ang. sin ångest att fråga om statistikhjälp?
  - $H_0: m_{tjänstemän} = m_{arbetar}$ $H_1: m_{tjänstemän} \neq m_{arbetar}$
- **Analyze→Compare Means→Independent sampels T Test**
  - Define groups efter kategorisiffror
    - Group 1: 0, Group 2: 1

Group Statistics

	Familjbakgrund	N	Mean	Std. Deviation	Std. Error Mean
Fråga om Hjälp Ångest	Arbetarfamilj	28	1.8980	.73036	.13802
	Tjänstmanfamilj	22	1.7662	.77941	.16617

Oberoende T test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Fråga om Hjälp Ångest	Equal variances assumed	.114	.737	.615	48	.542	.13173	.21431	-.29916	.56262
	Equal variances not assumed			.610	43.769	.545	.13173	.21602	-.30369	.56714

skillnad i spridningen

t värde

frihetsgrader

Signifikans (p) inte riktad

Medelvärdesskillnad mellan grupperna

Konfidensintervall av skillnaden plus se

p/sig = .542 Chansen att skillnaden inte är signifikant/slumpartad = 54 procent, Nollhypotesen gäller