BI1363 HT 2020

Test av proportioner

Oktober 2020

Adam Flöhr, BT, SLU

Binomial- och z-test för proportioner

Motsvarar *Biometri*, kap 8 (exkl. 8.6, 8.8)

I korthet

Vi vill undersöka en variabel (en egenskap) som är **binär**

Den intressanta populationsparametern är **proportionen** individer med egenskapen (***p***)

Med den skattade proportionen $\hat{p}$ kan vi genomföra hypotestest för ***p***, antingen genom ett **binomialtest** eller genom ett **z-test**

Om vi vill jämföra två grupper kan vi använda ett **z-test för två stickprov**

Binära variabler

Förekomsten av egenskapen hos en individ är en binär variabel

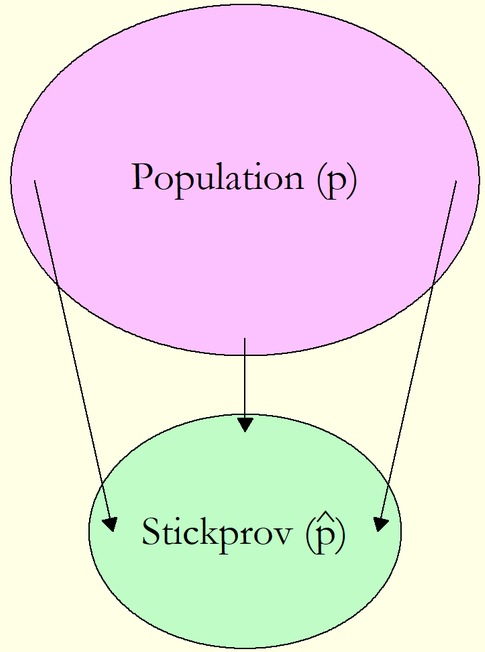
Vi vill veta hur vanlig en viss egenskap är i en population (dess proportion $\boldsymbol{p}$)

Drar ett stickprov av storlek $\boldsymbol{n}$ och tittar på antalet i stickprovet som har egenskapen (*positiva utfall*)

Den naturliga skattningen av $\boldsymbol{p}$ är

$$\hat{\boldsymbol{p}} = \frac{\boldsymbol{x}}{\boldsymbol{n}}$$

där $\boldsymbol{x}$ är antalet *positiva utfall* och $\boldsymbol{n}$ är stickprovets storlek



Fördelningen för x och $\hat{p}$

En observation ur populationen är positiv med sannolikheten p och negativ med sannolikheten $1 - p$

Antalet positiva utfall (x) bland n oberoende observationer är summan av n binära variabler

Antalet positiva utfall följer därmed en binomialfördelning med storlek n och sannolikhet p

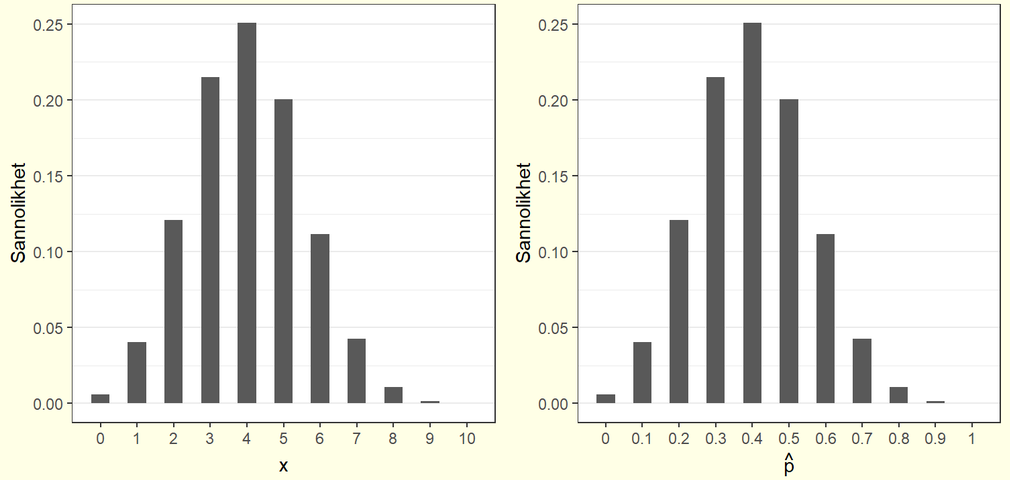
$$X \sim Bin(n, p)$$

Eftersom *antalet positiva utfall* är binomialfördelat ges fördelningen för $\hat{p}$ av en binomialfördelning skalad med n

Vi drar tio observationer ur en population där 40 procent har antikroppar mot en viss sjukdom

Antalet positiva följer en binomalfördelning med $n = 10$ och $p = 0.4$

Fördelningen för vår skattning av proportionen är en skalning av den binomialfördelningen

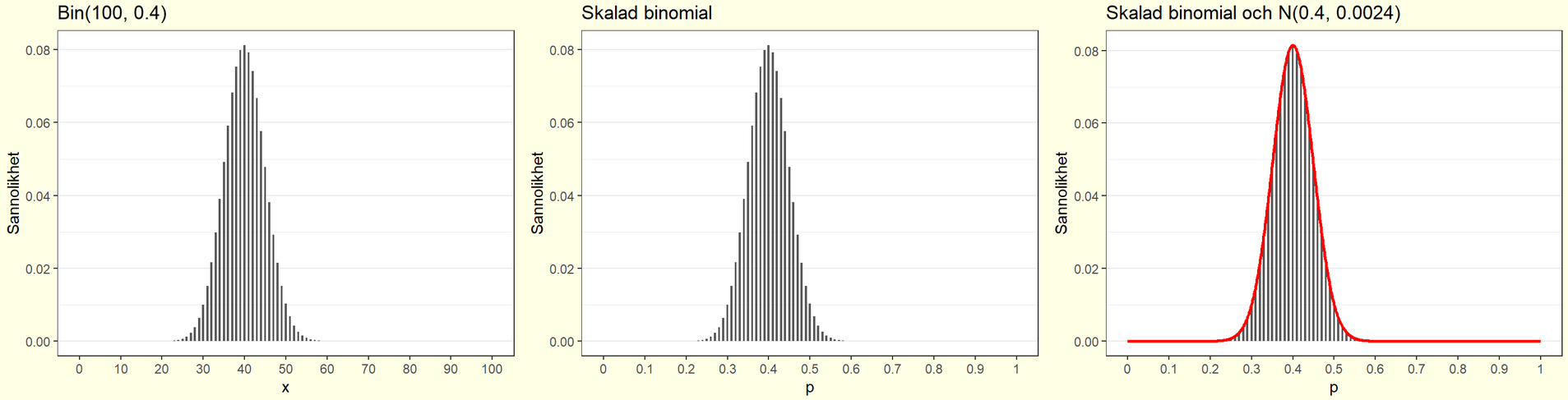


Normalapproximation av binomialfördelning

En binomialfördelning kan approximeras med en normalfördelning om n är *stort* och p *nära* 0.5

Boken *Biometri* ger tumregeln att np och $n(1-p)$ bägge ska vara större än 10

Den skalade binomialfördelningen för $\hat{p}$ kan approximeras med en normal med $\mu = p$ och $\sigma^2 = \frac{p(1-p)}{n}$



Fallet med $n = 100$ och $p = 0.4$

Vi kan antingen basera vårt test på binomialfördelningen för x ($Bin(100, 0.4)$) eller på normalfördelningen för $\hat{p}$ ($N(0.4, 0.0024)$)

Variansen i normalfördelningen ges av $\sigma^2 = \frac{p(1-p)}{n} = \frac{0.4 \cdot 0.6}{100} = 0.0024$

z-test för proportioner, ett stickprov

Ett hypotestest för att jämföra den observerade andelen $\hat{p}$ med en nollhypotes p_0

Bygger på normalapproximationen av binomialfördelningen

Standardgången för hypotestest

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

z-test för proportion, ett stickprov, schema

Hypoteser

$H_0 : p = p_0$

$H_1 : p \neq p_0$

Testfunktion

$$z = \frac{\hat{p} - p_0}{\sqrt{n}\frac{p \cdot (1-p)}{n}}$$

där $\hat{p}$ skattas från stickprovet, och p_0 hämtas från nollhypotesen

Testfördelning

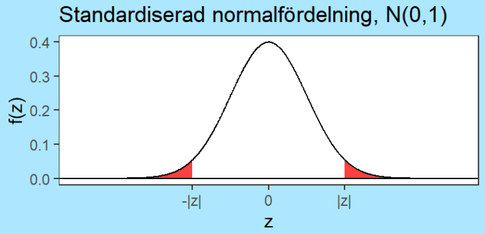
Under nollhypotesen följer z (approximativt) en standardiserad normalfördelning $N(0, 1)$

Approximationen är *giltig* om $np_0 > 10$ och $n(1 - p_0) > 10$

P-värde

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

Detta kan hämtas ur en normalfördelningstabell (*Biometri*, tabell 4)



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

z-test. Exempel

Vi undersöker mögelförekomst hos potatis och vill testa om proportionen är skild från $\mathbf{0.3}$

Vi drar ett stickprov om $\mathbf{60}$ observationer och finner $\mathbf{15}$ skadade plantor, så $\hat{\boldsymbol{p}} = \frac{\mathbf{15}}{\mathbf{60}} = \mathbf{0.25}$

Hypoteser

$$H_0 : p = 0.3 \qquad H_1 : p \neq 0.3$$

Testfunktion

$$z = \frac{\hat{p} - p_0}{\sqrt{n}\frac{p \cdot (1-p)}{n}} = \frac{0.25 - 0.3}{\sqrt{\frac{0.3 \cdot 0.7}{60}}} = -0.8452$$

Illustration av p-värdet

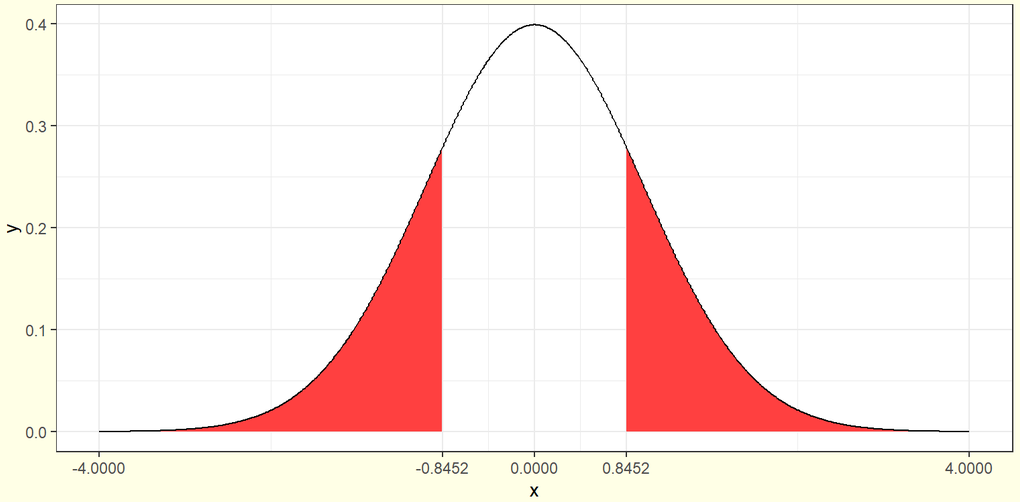
Testfunktionens värde $z = -0.8452$ är på den standardiserade skalan

Svanssannolikheten kan beräknas genom att slå i en normalfördelningstabell

Tabellen ger sannolikheten 0.8023 (för $z = 0.85$)

En svans motsvarar därmed $1 - 0.8023 = 0.1977$

Det tvåsidiga p-värdet ges av $2 \cdot 0.1977 = 0.3954$



Tabell 4: Normalfördelningens fördelningsfunktion

	Sista decimalen									
z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621

Konfidensintervall för proportioner, ett stickprov

Konfidensintervallet baseras på tabellvärden från den standardiserade normalfördelningen

I z-testet baseras standardfelet på nollhypotesen p_0

I konfidensintervallet är medelfelet baserat på den observerade proportionen $\hat{p}$

Formel

Det tvåsidiga konfidensintervallet ges av

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Fortsättning från tidigare exempel

$$0.25 \pm 1.96\sqrt{\frac{0.25 \cdot 0.75}{60}}$$

Detta ger intervallet $(0.140, 0.360)$

Binomialtestet

Proportioner kan också testas genom binomialfördelningen

Antalet skadade växter är $\boldsymbol{Bin(60, 0, 3)}$

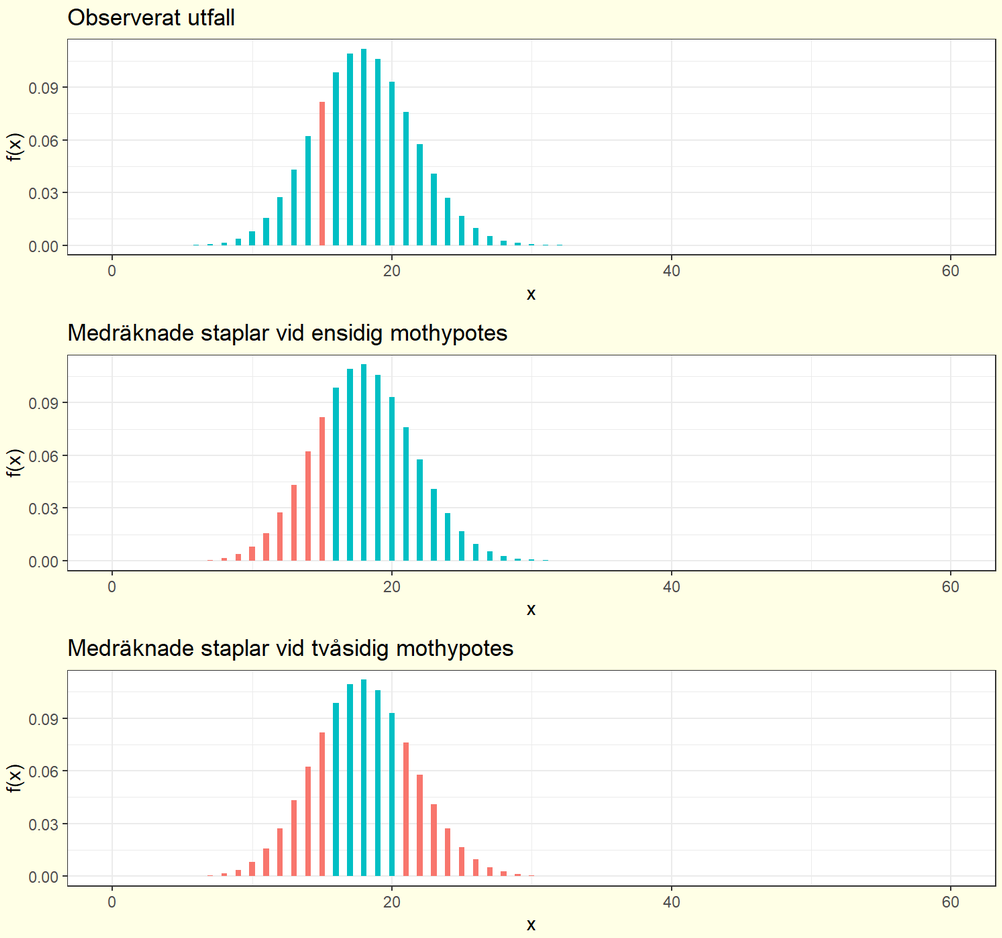
Vi observerar 15 skadade växter

Det ensidiga p-värdet (vid $\boldsymbol{H_1 : p < 0.3}$) är sannolikheten att få 15 eller färre skadade växter

Det tvåsidiga p-värdet (vid $\boldsymbol{H_1 : p \neq 0.3}$) är den summerade sannolikheten för utfall som är mindre sannolika än det observerade resultatet

Detta motsvarar alla staplar som är lägre än stapeln vid 15

Sannolikheterna kan beräknas i valfritt datorprogram, vilket ger $\boldsymbol{0.2438}$ respektive $\boldsymbol{0.4816}$



Binomialtestet, schema

Hypoteser

$H_0 : p = p_0$

$H_1 : p > p_0$ eller $H_1 : p < p_0$

Testfunktion

$x =$ **antalet positiva utfall**

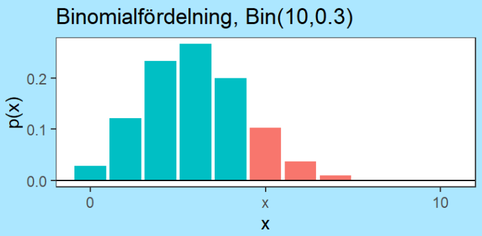
Testfördelning

Under nollhypotesen följer x en binomialfördelning med parameterar n och p_0: $Bin(n, p_0)$

P-värde

P-värdet ges av summan av staplarna över eller lika med x i testfördelningen

Detta kan hämtas ur en binomialfördelningstabell (*Biometri*, tabell 2, för $n \leq 20$)



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

Binomialtest, exempel

Vid straffavgöranden i fotboll slår lagen varannan straff

Eftersom de flesta straffar blir mål är den psykologiska bördan större på det lag som går som tvåa

I herr-VM (1998 till 2014) vann det första laget 12 av 15 straffavgöranden ($\hat{p} = 12/15 = 0.8$)

Kan vi säga att vinstsannolikheten för det lag som går först är större än 0.5?

Hypoteser

$H_0 : p_0 = 0.5$

$H_1 : p_0 > 0.5$

Testfunktion

Antalet positiva utfall är $x = 12$ av $n = 15$

Testfördelning

Under nollhypotesen kommer vår observation från en binomialfördelning med $n = 15$ och $p = 0.5$

p-värde

p-värdet beräknas som sannolikheten för det observerade utfallet eller något mer extremt

I det här fallet sannolikheten att få **12** eller fler positiva utfall

Tabell 2 ger sannolikheten för värden mindre eller lika med värdet i vänstra kolumnen

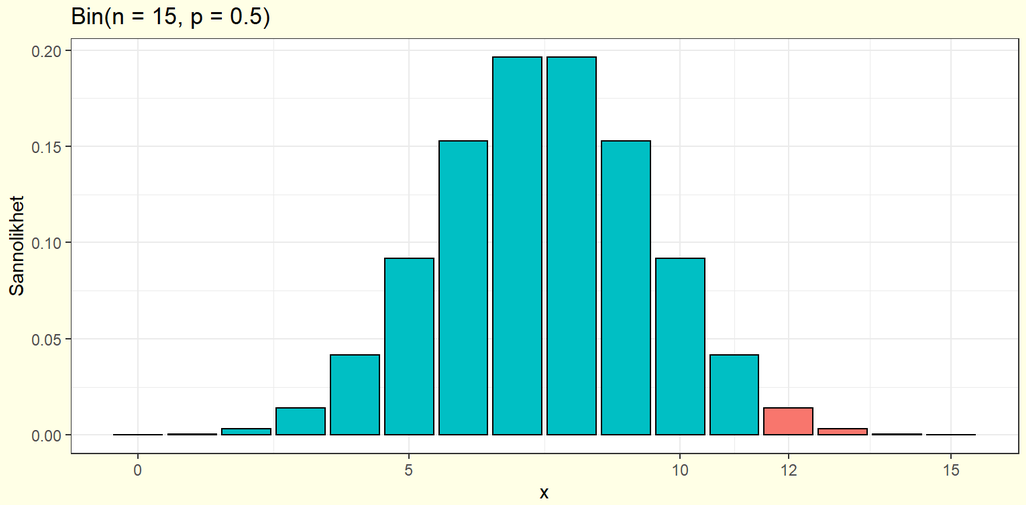
Sannolikheten att få mindre än eller lika med 11 är **0.9824**

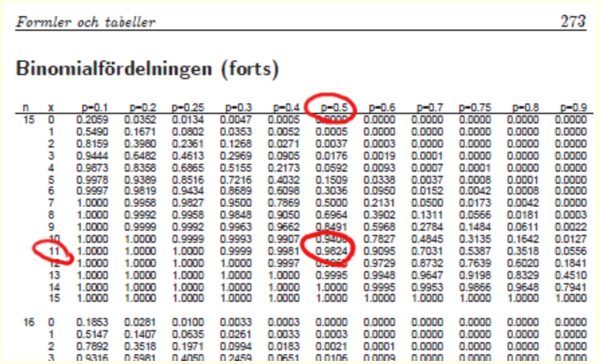
Sannolikheten att få mer eller lika med 12 är därmed
$\mathbf{1 - 0.9824 = 0.0176}$

Slutsats

Det finns en statistiskt säkerställd skillnad från **0.5**

Detta tyder på att sannolikheten att vinna är större om man går först





Formler och tabeller 273

Binomialfördelningen (forts)

n	x	p=0.1	p=0.2	p=0.25	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.75	p=0.8	p=0.9
15	0	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000	0.0000	0.0000	0.0000
	3	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000	0.0000	0.0000
	4	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0001	0.0000	0.0000
	5	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0008	0.0001	0.0000
	6	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0042	0.0008	0.0000
	7	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0173	0.0042	0.0000
	8	1.0000	0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0566	0.0181	0.0003
	9	1.0000	0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.1484	0.0611	0.0022
	10	1.0000	1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.3135	0.1642	0.0127
	11	1.0000	1.0000	1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.5387	0.3518	0.0556
	12	1.0000	1.0000	1.0000	1.0000	0.9997	0.9963	0.9729	0.8732	0.7639	0.6020	0.1841
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9948	0.9647	0.9198	0.8329	0.4510
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9866	0.9648	0.7941
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16	0	0.1853	0.0281	0.0100	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5147	0.1407	0.0635	0.0261	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.7892	0.3518	0.1971	0.0994	0.0183	0.0021	0.0001	0.0000	0.0000	0.0000	0.0000
	3	0.9316	0.5981	0.4050	0.2459	0.0651	0.0106	0.0009	0.0000	0.0000	0.0000	0.0000

Bonusuppgift. Vid VM 2018 förlorade förstalaget samtliga av fyra straffläggningar. Är resultatet fortfarande signifikant om den datan tas med?

z-test för två proportioner

Liknande generalisering som t-testet med två prov

Vi har två prover av binära data och kan uppskatta tre proportioner

$\hat{p}_1$ och $\hat{p}_2$ är proportionerna från respektive stickprov

$\hat{p}_0$ är proportionen från det sammanslagna stickprovet

z-test för proportioner, två stickprov, schema

Hypoteser

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

Testfunktion

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)(\frac{1}{n_1}+\frac{1}{n_2})}}$$

där $\hat{p}_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$

och $\hat{p}_1$ och $\hat{p}_2$ skattas

Testfördelning

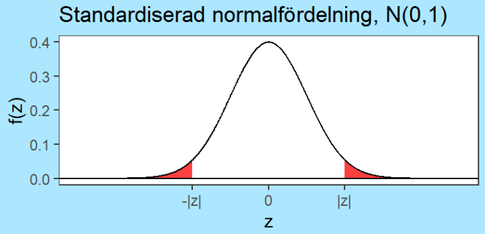
Under nollhypotesen följer z en standardiserad normalfördelning

Tumregel: $n_1\hat{p}_0, n_2\hat{p}_0, n_1(1-\hat{p}_0)$ och $n_2(1-\hat{p}_0)$ bör alla vara större än **10**

P-värde

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

Detta kan hämtas ur en normalfördelningstabell (*Biometri*, tabell 4)



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

Två stickprov, exempel

Träd i södra och norra Sverige kontrolleras för en viss sjukdom

$\mathbf{50}$ av $\mathbf{100}$ är sjuka i söder och $\mathbf{52}$ av $\mathbf{200}$ är sjuka i norr

$$\hat{p}_1 = \frac{50}{100} = 0.50 \qquad \hat{p}_2 = \frac{52}{200} = 0.26$$

Vi vill testa om regionerna har skilda proportioner sjuka träd

Den sammanslagna proportionen är $\hat{p}_0 = \frac{50+52}{100+200} = 0.34$

Hypoteser

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

Testfunktion

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.50 - 0.26}{\sqrt{0.34(1-0.34)\left(\frac{1}{100} + \frac{1}{200}\right)}} = 4.137$$

Testfördelning

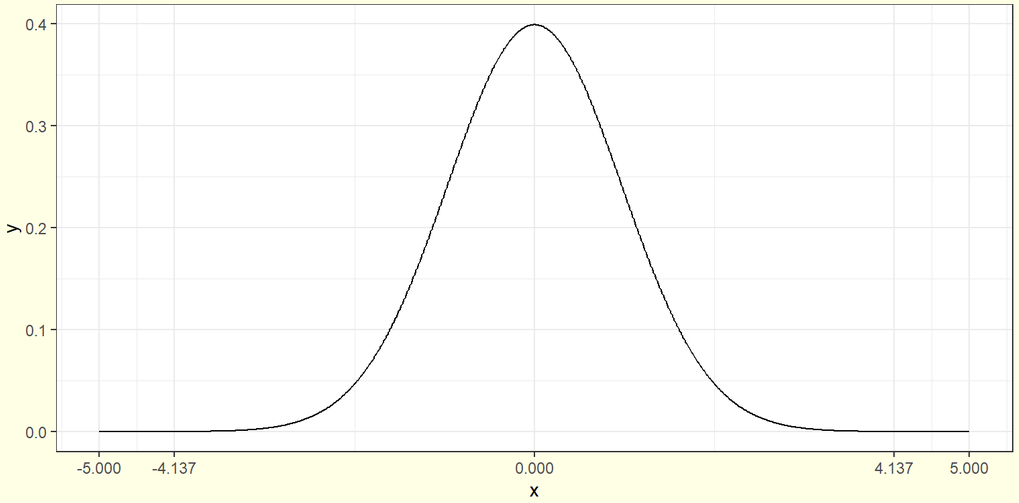
Under nollhypotesen följer z en standardiserad normalfördelning $N(0,1)$

P-värde

p-värdet ges av arean under kruvan bortom $z = 4.137$

I det här fallet är den arean mycket liten

Normalfördelningstabellen (*Biometri*, tabell 4) ger att p-värdet i alla fall är mindre än **0.001**



Slutsats

Ett lågt p-värde ger att vi förkastar nollhypotesen

Det finns en statistiskt säkerställd skillnad mellan proportionerna sjuka träd

Två stickprov, Konfidensintervall

Konfidensintervallet för skillnaden mellan proportioner ges av

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Ett 95-procentigt konfidensintervall för exemplet

$$0.5 - 0.26 \pm 1.96\sqrt{0.34 \cdot 0.66(0.01 + 0.005)} = 0.24 \pm 0.11$$

eller $[0.13, 0.35]$

Antagandet

Binomaltestet och z-testerna bygger på oberoende observationer

z-testerna bygger dessutom på en giltig normalapproximation

Slut