2015-04-05

STATISTIK
FÖR LÄKARSTUDENTER

Nils Karlsson

INDEX

INTRODUKTION

Här beskrivs några grundläggande begrepp som man bör känna till innan man läser vidare.

Att skriva sannolikheter

Matematiskt betecknas sannolikheten för en händelse med p. Sannolikheten för en händelse som förväntas inträffa vid varannat tillfälle kan skrivas matematiskt på något av nedanstående sätt:

$p = ½ = 50\ \% = 0{,}5$

Vid statistiska beräkningar skrivs sannolikheter som tal i intervallet 0 till 1 snarare än 0 till 100 %.

Sannolikhetslagar

Beteckningar:

$P(A)$	sannolikheten för att händelse A inträffar
$P(\bar{A})$	sannolikheten för att händelse A inte inträffar ("komplementhändelse")
$P(A \cap B)$	sannolikheten för att både händelse A och händelse B inträffar
$P(A \cup B)$	sannolikheten för att minst en av händelse A och händelse B inträffar
$P(A\|B)$	sannolikheten för att händelse A också inträffar om händelse B inträffar

Sannolikhetslagar:

$P(\bar{A}) = 1 - P(A)$	
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	additionssatsen
$P(A \cap B) = P(A) \cdot P(B\|A) = P(B) \cdot P(A\|B)$	multiplikationssatsen (beroende händelser)
$P(A\|B) = \frac{P(A \cap B)}{P(B)}$	betingad sannolikhet (beroende händelser)
$P(B\|A) = \frac{P(A\|B) \cdot P(A)}{P(B)}$	Bayes teorem (efter prästen **Thomas Bayes**)

Observera att vid oberoende händelser gäller att:

$P(A\|B) = P(A)$	
$P(B\|A) = P(B)$	
$P(A \cap B) = P(A) \cdot P(B)$	"multiplikationssatsen" (oberoende händelser)

Fakulteter

Fakulteter är särskilt användbara när man skall beräkna antalet möjliga utfall. Fakulteten för heltalet n är lika med produkten av samtliga heltal i talserien $\{1, ..., n\}$.

$$n! = 1 \cdot \ldots \cdot n$$

$$2! = 1 \cdot 2 = 2$$

$$3! = 1 \cdot 2 \cdot 3 = 6$$

Dessutom gäller att:

$$0! = 1$$

Vid n ingående element (objekt) är $n!$ olika permutationer (uppräkningar) möjliga.

Julius, Julia och deras robot Julium vill leka att de står i kö vid en statlig myndighet. Totalt är 6 olika köer möjliga: Julius-Julia-Julium, Julius-Julium-Julia, Julia-Julius-Julium, Julia-Julium-Julius, Julium-Julius-Julia samt Julium-Julia-Julius.

Antalet möjliga kombinationer när vi använder x av y tillgängliga element:

$$\binom{y}{x} = \frac{y!}{x! \cdot (y-x)!} \text{ till exempel } \binom{3}{2} = \frac{3!}{2! \cdot (3-2)!} = \frac{6}{2 \cdot 1} = 3$$ (A, B, C: A+B, A+C eller B+C)

Antalet möjliga uppräkningar när vi använder y element varav x är A och resten är B:

$$\binom{y}{x} = \frac{y!}{x! \cdot (y-x)!} \text{ till exempel } \binom{3}{2} = \frac{3!}{2! \cdot (3-2)!} = \frac{6}{2 \cdot 1} = 3$$ (AAB, ABA samt BAA)

Odds och oddskvot

Oddsen för ett utfall är lika med sannolikheten för utfallet dividerad med sannolikheten för att utfallet inte äger rum. Om 10 exponeras, 8 blir sjuka och 2 förblir friska så är oddsen för sjukdom efter exponering 4. Om 40 inte exponeras, 8 blir sjuka och 32 förblir friska så är oddsen för sjukdom efter icke-exponering 0,25.

Oddskvoten anger den relativa sannolikheten. Med föregående siffror blir oddskvoten för sjukdom lika med 4 / 0,25 = 16; exponering gör alltså att risken för sjukdom blir 16 gånger högre.

Typer av data

Kvantitativa data utgörs av mängder och mäts med siffror (till exempel ålder och vikt). Kvalitativa data utgörs istället av egenskaper (till exempel blodgrupp och kön). *Kvalitativa data kan visserligen benämnas med siffror men dessa är då att betrakta som etiketter snarare än kvantiteter.*

Kvalitativa data kan delas upp i två undertyper, nominaldata respektive ordinaldata. Nominaldata beskriver egenskaper som saknar inbördes värderingsordning och alltså inte låter sig inordnas i skalor (till exempel blodgrupp eller kön). Ordinaldata kan rangordnas men inte storleksbestämmas (till exempel betyg, där det inte går att säga att kriterierna för ett högre betyg är lika med kriterierna för ett lägre betyg multiplicerat med en viss faktor).

Kvantitativa data kan i sin tur också delas upp i två undertyper, intervalldata och kvotdata. Skillnaden mellan dem är att intervalldata kan ha negativa värden (till exempel temperatur enligt Celsiusskalan) medan kvotdata inte kan det (till exempel vikt).

Kvotdata kan omvandlas till intervalldata, intervalldata till ordinaldata och ordinaldata till nominaldata. Omvandlingar i andra riktningen är mycket vanskligare.

Kvantitativa data kan vidare sägas finnas i två varianter, diskreta data respektive kontinuerliga data. Diskreta data är begränsade till ett antal specifika värden; till exempel kan en tärning endast anta värdena {1, 2, 3, 4, 5, 6} men inte något annat värde mellan dessa. Kontinuerliga data begränsas däremot inte till ett fåtal punkter inom det studerade området; exempel på sådana data är längd, vikt och ålder som i teorin kan beskrivas med ett oändligt antal decimaler och därmed antaga ett oändligt antal möjliga värden. Kontinuerliga data måste därför beskrivas med intervaller.

Typen av data är avgörande för vilka statistiska undersökningar som är möjliga.

> *Kursbetyg är ett typiskt exempel på ordinaldata eftersom det går att rangordna dem men inte att göra en kvantitativ mätning av avståndet mellan dem (det vill säga mellan deras definitioner enligt läroplanen). Att räkna ut genomsnitt utifrån kursbetyg är således matematiskt felaktigt; äkta genomsnitt fordrar kvantitativa data. Det är också den vanligaste antagningsmetoden till högre utbildning i Sverige.*

Diagram

Venndiagram (efter logikern **John Venn**) utgörs av en karta där geometriska figurer, ofta cirklar, representerar olika händelser eller grupper. Överlappningar mellan två eller fler figurer betyder att det som figurerna representerar föreligger samtidigt (det kan till exempel vara sannolikheten för att två händelser inträffar samtidigt).

Träddiagram utgörs av en karta över möjliga, förgrenade händelsekedjor.

Lådagram (box plot) används för att visa hur urval av data är invändigt fördelade (se "deskriptiv statistik" för förklaringar av följande termer). Den centrala rutans yttre gränser utgörs av första och tredje kvartilerna medan linjen som skär rutan utgör medianen. De två utgående strecken sträcker sig 1,5 kvartilavstånd ut från lådan. Värden mer än 1,5 kvartilavstånd ut från lådan betecknas utliggare och brukar markeras i diagrammet som små cirklar. Värden mer än 3 kvartilavstånd ut från lådan betecknas extremvärden.

Stolpdiagram och **histogram** är båda exempel på stapeldiagram men det finns en viktig skillnad: stolpdiagram avser diskreta data och står separata medan histogram avser kontinuerliga data och därför saknar mellanrum mot grannstaplarna.

Sambandsdiagram (scatter plot) används för att visa hur par av variabler (till exempel personers längd och vikt) fördelar sig. De parade variablerna ger x- respektive y-koordinaterna i diagrammet. Det framgår ofta av diagrammet ifall det föreligger något tydligt samband mellan de två variablerna.

SANNOLIKHETSFÖRDELNINGAR

Sannolikhetsfördelningarna är fundamentala redskap inom statistiken, särskilt för analytisk statistik. En sannolikhetsfördelning beskriver sannolikheterna för olika utfall utifrån de valda parametrarna. Dess medelvärde/väntevärde $E(X)$ och varians $Var(X)$ kan också beräknas (se "deskriptiv statistik").

$$f(x) = formel$$

$$E(X) = formel$$

$$Var(X) = formel$$

Sannolikhetsfördelningarna finns även redovisade som referenstabeller där man kan avläsa hur stor sannolikheten är för ett visst utfall utifrån vanliga värden på parametrarna (andra parametervärden fordrar att man gör egna beräkningar eller använder ett statistiskt datorprogram).

Binomialfördelning

Binomialfördelningen för diskreta data beskriver sannolikheten för att ett utfall av två (till exempel "ja" och "nej") inträffar totalt x gånger under n försök när sannolikheten är p för just det utfallet.

$$f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

$$E(X) = n \cdot p$$

$$Var(X) = n \cdot p \cdot (1-p)$$

Binomialfördelningar skrivs förkortat:

$$Bin(n, p)$$

Om n sätts till 1 så får man Bernoullifördelningen (efter matematikern **Jakob Bernoulli**):

$$f(x) = p^x \cdot (1-p)^{1-x}$$

För fler än 2 möjliga utfall kan man istället beräkna en multinomialfördelning med k möjliga utfall:

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdot \ldots \cdot x_k!} \cdot p_1^{x_1} \cdot \ldots \cdot p_k^{x_k}$$ där x_i och p_i anger olika utfalls antal och sannolikhet

Räkneexempel:

Sannolikheten för att Barney får en "perfekt vecka" (7 av 7) om chansen per dag är 10 %:

$$f(7) = \binom{7}{7} \cdot 0{,}1^7 \cdot (1-0{,}1)^{7-7} = \frac{7!}{7! \cdot 0!} \cdot 0{,}1^7 \cdot 0{,}9^0 = 1 \cdot 0{,}1^7 \cdot 1 = 0{,}0000001$$ eller cirka 0,00001 %

Nedanför återges referenstabeller för olika värden på n och p. Observera att angivna sannolikheter utgör kumulativa sannolikheter. Sannolikheten för exakt x inträffanden fås genom att även läsa av tabellvärdet för $x - 1$ och sedan dra av det från tabellvärdet för x (alternativt genom att beräkna värdet för binomialfördelningens sannolikhetsfunktion, vilket också ger det exakta värdet).

$n = 1$

x	$p = 0{,}1$	$p = 0{,}2$	$p = 0{,}3$	$p = 0{,}4$	$p = 0{,}5$	$p = 0{,}6$	$p = 0{,}7$	$p = 0{,}8$	$p = 0{,}9$
0	0,90000	0,80000	0,70000	0,60000	0,50000	0,40000	0,30000	0,20000	0,10000
1	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

$n = 5$

x	$p = 0{,}1$	$p = 0{,}2$	$p = 0{,}3$	$p = 0{,}4$	$p = 0{,}5$	$p = 0{,}6$	$p = 0{,}7$	$p = 0{,}8$	$p = 0{,}9$
0	0,59049	0,32768	0,16807	0,07776	0,03125	0,01024	0,00243	0,00032	0,00001
1	0,91854	0,73728	0,52822	0,33696	0,18750	0,08704	0,03078	0,00672	0,00046
2	0,99144	0,94208	0,83692	0,68256	0,50000	0,31744	0,16308	0,05792	0,00856
3	0,99954	0,99328	0,96922	0,91296	0,81250	0,66304	0,47178	0,26272	0,08146
4	0,99999	0,99968	0,99757	0,98976	0,96875	0,92224	0,83193	0,67232	0,40951
5	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

$n = 6$

x	$p = 0{,}1$	$p = 0{,}2$	$p = 0{,}3$	$p = 0{,}4$	$p = 0{,}5$	$p = 0{,}6$	$p = 0{,}7$	$p = 0{,}8$	$p = 0{,}9$
0	0,53144	0,26214	0,11765	0,04666	0,01563	0,00410	0,00073	0,00006	0,00000
1	0,88574	0,65536	0,42018	0,23328	0,10938	0,04096	0,01094	0,00160	0,00006
2	0,98415	0,90112	0,74431	0,54432	0,34375	0,17920	0,07047	0,01696	0,00127
3	0,99873	0,98304	0,92953	0,82080	0,65625	0,45568	0,25569	0,09888	0,01585
4	0,99995	0,99840	0,98907	0,95904	0,89063	0,76672	0,57983	0,34464	0,11427
5	1,00000	0,99994	0,99927	0,99590	0,98438	0,95334	0,88235	0,73786	0,46856
6	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

$n = 7$

x	$p = 0{,}1$	$p = 0{,}2$	$p = 0{,}3$	$p = 0{,}4$	$p = 0{,}5$	$p = 0{,}6$	$p = 0{,}7$	$p = 0{,}8$	$p = 0{,}9$
0	0,47830	0,20972	0,08235	0,02799	0,00781	0,00164	0,00022	0,00001	0,00000
1	0,85031	0,57672	0,32942	0,15863	0,06250	0,01884	0,00379	0,00037	0,00001
2	0,97431	0,85197	0,64707	0,41990	0,22656	0,09626	0,02880	0,00467	0,00018
3	0,99727	0,96666	0,87396	0,71021	0,50000	0,28979	0,12604	0,03334	0,00273
4	0,99982	0,99533	0,97120	0,90374	0,77344	0,58010	0,35293	0,14803	0,02569
5	0,99999	0,99963	0,99621	0,98116	0,93750	0,84137	0,67058	0,42328	0,14969
6	1,00000	0,99999	0,99978	0,99836	0,99219	0,97201	0,91765	0,79028	0,52170
7	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

$n = 8$

x	$p = 0{,}1$	$p = 0{,}2$	$p = 0{,}3$	$p = 0{,}4$	$p = 0{,}5$	$p = 0{,}6$	$p = 0{,}7$	$p = 0{,}8$	$p = 0{,}9$
0	0,43047	0,16777	0,05765	0,01680	0,00391	0,00066	0,00007	0,00000	0,00000
1	0,81310	0,50332	0,25530	0,10638	0,03516	0,00852	0,00129	0,00008	0,00000
2	0,96191	0,79692	0,55177	0,31539	0,14453	0,04981	0,01129	0,00123	0,00002
3	0,99498	0,94372	0,80590	0,59409	0,36328	0,17367	0,05797	0,01041	0,00043
4	0,99957	0,98959	0,94203	0,82633	0,63672	0,40591	0,19410	0,05628	0,00502
5	0,99998	0,99877	0,98871	0,95019	0,85547	0,68461	0,44823	0,20308	0,03809
6	1,00000	0,99992	0,99871	0,99148	0,96484	0,89362	0,74470	0,49668	0,18690
7	1,00000	1,00000	0,99993	0,99934	0,99609	0,98320	0,94235	0,83223	0,56953
8	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

$n = 9$

x	$p = 0,1$	$p = 0,2$	$p = 0,3$	$p = 0,4$	$p = 0,5$	$p = 0,6$	$p = 0,7$	$p = 0,8$	$p = 0,9$
0	0,38742	0,13422	0,04035	0,01008	0,00195	0,00026	0,00002	0,00000	0,00000
1	0,77484	0,43621	0,19600	0,07054	0,01953	0,00380	0,00043	0,00002	0,00000
2	0,94703	0,73820	0,46283	0,23179	0,08984	0,02503	0,00429	0,00031	0,00000
3	0,99167	0,91436	0,72966	0,48261	0,25391	0,09935	0,02529	0,00307	0,00006
4	0,99911	0,98042	0,90119	0,73343	0,50000	0,26657	0,09881	0,01958	0,00089
5	0,99994	0,99693	0,97471	0,90065	0,74609	0,51739	0,27034	0,08564	0,00833
6	1,00000	0,99969	0,99571	0,97497	0,91016	0,76821	0,53717	0,26180	0,05297
7	1,00000	0,99998	0,99957	0,99620	0,98047	0,92946	0,80400	0,56379	0,22516
8	1,00000	1,00000	0,99998	0,99974	0,99805	0,98992	0,95965	0,86578	0,61258
9	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

$n = 10$

x	$p = 0,1$	$p = 0,2$	$p = 0,3$	$p = 0,4$	$p = 0,5$	$p = 0,6$	$p = 0,7$	$p = 0,8$	$p = 0,9$
0	0,34868	0,10737	0,02825	0,00605	0,00098	0,00010	0,00001	0,00000	0,00000
1	0,73610	0,37581	0,14931	0,04636	0,01074	0,00168	0,00014	0,00000	0,00000
2	0,92981	0,67780	0,38278	0,16729	0,05469	0,01229	0,00159	0,00008	0,00000
3	0,98720	0,87913	0,64961	0,38228	0,17188	0,05476	0,01059	0,00086	0,00001
4	0,99837	0,96721	0,84973	0,63310	0,37695	0,16624	0,04735	0,00637	0,00015
5	0,99985	0,99363	0,95265	0,83376	0,62305	0,36690	0,15027	0,03279	0,00163
6	0,99999	0,99914	0,98941	0,94524	0,82813	0,61772	0,35039	0,12087	0,01280
7	1,00000	0,99992	0,99841	0,98771	0,94531	0,83271	0,61722	0,32220	0,07019
8	1,00000	1,00000	0,99986	0,99832	0,98926	0,95364	0,85069	0,62419	0,26390
9	1,00000	1,00000	0,99999	0,99990	0,99902	0,99395	0,97175	0,89263	0,65132
10	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

$n = 20$

x	$p = 0,1$	$p = 0,2$	$p = 0,3$	$p = 0,4$	$p = 0,5$	$p = 0,6$	$p = 0,7$	$p = 0,8$	$p = 0,9$
0	0,12158	0,01153	0,00080	0,00004	0,00000	0,00000	0,00000	0,00000	0,00000
1	0,39175	0,06918	0,00764	0,00052	0,00002	0,00000	0,00000	0,00000	0,00000
2	0,67693	0,20608	0,03548	0,00361	0,00020	0,00001	0,00000	0,00000	0,00000
3	0,86705	0,41145	0,10709	0,01596	0,00129	0,00005	0,00000	0,00000	0,00000
4	0,95683	0,62965	0,23751	0,05095	0,00591	0,00032	0,00001	0,00000	0,00000
5	0,98875	0,80421	0,41637	0,12560	0,02069	0,00161	0,00004	0,00000	0,00000
6	0,99761	0,91331	0,60801	0,25001	0,05766	0,00647	0,00026	0,00000	0,00000
7	0,99958	0,96786	0,77227	0,41589	0,13159	0,02103	0,00128	0,00002	0,00000
8	0,99994	0,99002	0,88667	0,59560	0,25172	0,05653	0,00514	0,00010	0,00000
9	0,99999	0,99741	0,95204	0,75534	0,41190	0,12752	0,01714	0,00056	0,00000
10	1,00000	0,99944	0,98286	0,87248	0,58810	0,24466	0,04796	0,00259	0,00001
11	1,00000	0,99990	0,99486	0,94347	0,74828	0,40440	0,11333	0,00998	0,00006
12	1,00000	0,99998	0,99872	0,97897	0,86841	0,58411	0,22773	0,03214	0,00042
13	1,00000	1,00000	0,99974	0,99353	0,94234	0,74999	0,39199	0,08669	0,00239
14	1,00000	1,00000	0,99996	0,99839	0,97931	0,87440	0,58363	0,19579	0,01125
15	1,00000	1,00000	0,99999	0,99968	0,99409	0,94905	0,76249	0,37035	0,04317
16	1,00000	1,00000	1,00000	0,99995	0,99871	0,98404	0,89291	0,58855	0,13295
17	1,00000	1,00000	1,00000	0,99999	0,99980	0,99639	0,96452	0,79392	0,32307
18	1,00000	1,00000	1,00000	1,00000	0,99998	0,99948	0,99236	0,93082	0,60825
19	1,00000	1,00000	1,00000	1,00000	1,00000	0,99996	0,99920	0,98847	0,87842
20	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

Poissonfördelning

Poissonfördelningen (efter matematikern **Siméon Poisson**) för diskreta data beskriver sannolikheten för att ett utfall inträffar x gånger under ett försökstillfälle när medelvärdet och variansen är λ.

$$f(x)=\frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

$$E(X)=\lambda$$

$$Var(X)=\lambda$$

Poissonfördelningar skrivs förkortat:

$$Po(\lambda)$$

Räkneexempel:

Barney bär i genomsnitt 2 kostymer per dag. Sannolikheten för en dag med 3 kostymer är:

$$f(3)=\frac{2^3 \cdot e^{-2}}{3!}=\frac{8 \cdot e^{-2}}{6}\approx 0{,}18$$ eller cirka 18 %

Nedanför återges en referenstabell för några värden på λ. Observera att de angivna sannolikheterna återigen utgör de kumulativa sannolikheterna, det vill säga summan av alla sannolikheter upp till x.

x	$\lambda=1$	$\lambda=2$	$\lambda=3$	$\lambda=4$	$\lambda=5$	$\lambda=6$	$\lambda=7$	$\lambda=8$	$\lambda=9$
0	0,36788	0,13534	0,04979	0,01832	0,00674	0,00248	0,00091	0,00034	0,00012
1	0,73576	0,40601	0,19915	0,09158	0,04043	0,01735	0,00730	0,00302	0,00123
2	0,91970	0,67668	0,42319	0,23810	0,12465	0,06197	0,02964	0,01375	0,00623
3	0,98101	0,85712	0,64723	0,43347	0,26503	0,15120	0,08177	0,04238	0,02123
4	0,99634	0,94735	0,81526	0,62884	0,44049	0,28506	0,17299	0,09963	0,05496
5	0,99941	0,98344	0,91608	0,78513	0,61596	0,44568	0,30071	0,19124	0,11569
6	0,99992	0,99547	0,96649	0,88933	0,76218	0,60630	0,44971	0,31337	0,20678
7	0,99999	0,99890	0,98810	0,94887	0,86663	0,74398	0,59871	0,45296	0,32390
8	1,00000	0,99976	0,99620	0,97864	0,93191	0,84724	0,72909	0,59255	0,45565
9	1,00000	0,99995	0,99890	0,99187	0,96817	0,91608	0,83050	0,71662	0,58741
10	1,00000	0,99999	0,99971	0,99716	0,98630	0,95738	0,90148	0,81589	0,70599
11	1,00000	1,00000	0,99993	0,99908	0,99455	0,97991	0,94665	0,88808	0,80301
12	1,00000	1,00000	0,99998	0,99973	0,99798	0,99117	0,97300	0,93620	0,87577
13	1,00000	1,00000	1,00000	0,99992	0,99930	0,99637	0,98719	0,96582	0,92615
14	1,00000	1,00000	1,00000	0,99998	0,99977	0,99860	0,99428	0,98274	0,95853
15	1,00000	1,00000	1,00000	1,00000	0,99993	0,99949	0,99759	0,99177	0,97796
16	1,00000	1,00000	1,00000	1,00000	0,99998	0,99983	0,99904	0,99628	0,98889
17	1,00000	1,00000	1,00000	1,00000	0,99999	0,99994	0,99964	0,99841	0,99468
18	1,00000	1,00000	1,00000	1,00000	1,00000	0,99998	0,99987	0,99935	0,99757
19	1,00000	1,00000	1,00000	1,00000	1,00000	0,99999	0,99996	0,99975	0,99894
20	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	0,99999	0,99991	0,99956
21	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	0,99997	0,99983
22	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	0,99999	0,99993

Normalfördelning

Gaussfördelningen (efter matematikern **Carl Gauss**), eller normalfördelningen, för kontinuerliga data är den kanske främsta sannolikhetsfördelningen. Enligt den centrala gränsvärdessatsen kommer medelvärdena av oberoende slumpmässiga urval att närma sig en normalfördelning när allt fler urval tillkommer, oavsett deras egna sannolikhetsfördelningar. Normalfördelningen är särskilt relevant för den medicinska vetenskapen eftersom många biologiska variabler är normalfördelade. Normalfördelningen är i diagram formad som en symmetrisk kulle eller "klocka" i genomskärning, med ett relativt högt mittparti och kontinuerliga svansar åt höger och vänster.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Normalfördelningar skrivs förkortat:

$$N(\mu, \sigma^2)$$

Normalfördelningens referenstabeller utgår vanligtvis från den standardiserade normalfördelningen $N(0,1)$, även känd som z-fördelningen, där $\mu = 0$ och $\sigma^2 = 1$. Genom en z-transformation kan man beräkna det Z-värde som X motsvarar: X minus medelvärdet, dividerat med standardavvikelsen.

$$Z = \frac{X - \mu}{\sigma} \text{ eller } Z = \frac{X - \bar{x}}{s}$$

Negativa värden för Z avläses som positiva, sedan subtraheras tabellvärdet från 1.

$Z = 1{,}00$ motsvarar 0,841

$Z = -1{,}00$ motsvarar då 1,000 – 0,841 = 0,159

Vid z-fördelningen motsvarar värdet på Z antalet standardavvikelser från μ!

$\mu \pm 1{,}00\sigma$ motsvarar 0,841 – 0,159 = 0,682 (68 % av normalfördelningen finns inom $1{,}00\sigma$)

$\mu \pm 1{,}96\sigma$ motsvarar 0,975 – 0,025 = 0,950 (95 % av normalfördelningen finns inom $1{,}96\sigma$)

z_y motsvarar det z-värde som vid tabellavläsning ger värdet y. Om y är mindre än 0,500 så letar man efter ett negativt z-värde och söker därför i tabellen efter värdet $1 - y$.

$z_{0{,}975} = 1{,}96$

$z_{0{,}025} = -1{,}96$

Observera att med kontinuerliga data kan man inte identifiera sannolikheter för exakta utfall! Det finns oändligt många decimaler och sannolikheten för ett exakt värde är alltså oändligt liten. Det man *kan* göra är att bedöma sannolikheten för utfall nedanför eller ovanför ett visst gränsvärde.

Räkneexempel:

Den dramatiska tiden mellan "legen" och "dary" följer en normalfördelning.

Medeltiden är 10 sekunder och standardavvikelsen är 4 s. Hur vanliga är tider över 20 s?

$$Z = \frac{20-10}{4} = \frac{10}{4} = 2{,}50$$

Tabellavläsning för 2,50 visar att tider på upp till 20 sekunder utgör 0,99379 (99,379 %).

Tider över 20 sekunder utgör då återstoden, 1 – 0,99379 = 0,00621 (0,621 %).

Vid avläsning av referenstabellen motsvarar olika kolumner olika värden på den andra *z*-decimalen.

z	0	1	2	3	4	5	6	7	8	9
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983

t-fördelning

Students *t*-fördelning (efter pseudonymen **Student** som användes av **William Gosset**, som vid den tiden var anställd vid **Guinness**) för kontinuerliga data används främst vid små populationer. Ju fler frihetsgrader den har, desto bättre motsvaras den av normalfördelningen.

Antalet frihetsgrader är ett mått på datavärdenas oberoende, på hur många utav dem som kan variera. Ifall det till exempel är fastslaget att v datavärden har ett visst medelvärde så räcker det med att känna till värdet på $v - 1$ datavärden för att kunna bestämma värdet på det sista av dem och det föreligger alltså $v - 1$ frihetsgrader.

$t_{(y, v)}$ motsvarar det *t*-värde som vid tabellavläsning för v frihetsgrader har sannolikhet y.

$t_{(0{,}975,\ 10)} = 2{,}228$

$t_{(0{,}025,\ 10)} = -2{,}228$

Referenstabellen för *t*-fördelningen har antalet frihetsgrader som rader och olika sannolikheter som kolumner. För negativa värden på t subtraheras den avlästa sannolikheten från 1.

v	$p = 0{,}70$	$p = 0{,}80$	$p = 0{,}90$	$p = 0{,}95$	$p = 0{,}975$	$p = 0{,}990$	$p = 0{,}995$	$p = 0{,}9990$	$p = 0{,}9995$
1	0,727	1,376	3,078	6,314	12,706	31,821	63,657	318,309	636,619
2	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,327	31,599
3	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
25	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
50	0,528	0,849	1,299	1,676	2,009	2,403	2,678	3,261	3,496
100	0,526	0,845	1,290	1,660	1,984	2,364	2,626	3,174	3,390

χ^2-fördelning

χ^2-fördelningen (chi-kvadrat-fördelningen) för kontinuerliga data används också flitigt, bland annat för att dra slutsatser om varianser och avgöra hur väl en teoretisk fördelning (en teoretisk modell, ett teoretiskt samband) matchar observationerna.

$\chi^2_{(y, v)}$ motsvarar det χ^2-värde som vid tabellavläsning för v frihetsgrader har sannolikhet y.

$\chi^2_{(0{,}975,\ 10)} = 20{,}483$

$\chi^2_{(0{,}025,\ 10)} = 3{,}247$

Referenstabellen för χ^2-fördelningen har antalet frihetsgrader som rader och olika sannolikheter som kolumner.

v	$p = 0{,}005$	$p = 0{,}010$	$p = 0{,}025$	$p = 0{,}05$	$p = 0{,}50$	$p = 0{,}95$	$p = 0{,}975$	$p = 0{,}990$	$p = 0{,}995$
1	0,000	0,000	0,001	0,004	0,455	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	1,386	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	2,366	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	3,357	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	4,351	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	5,348	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	6,346	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	7,344	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	8,343	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	9,342	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	10,341	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	11,340	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	12,340	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	13,339	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	14,339	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	15,338	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	16,338	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	17,338	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	18,338	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	19,337	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	20,337	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	21,337	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	22,337	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	23,337	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	24,337	37,652	40,646	44,314	46,928
30	13,787	14,953	16,791	18,493	29,336	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	39,335	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	49,335	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	59,335	79,082	83,298	88,379	91,952
80	51,172	53,540	57,153	60,391	79,334	101,879	106,629	112,329	116,321
100	67,328	70,065	74,222	77,929	99,334	124,342	129,561	135,807	140,169

F-fördelning

F-fördelningen (efter statistikern och biologen **Ronald Fisher** och matematikern **George Snedecor**) används bland annat inom analys och jämförelser av varianser.

$F_{(y, v1, v2)}$ motsvarar det *F*-värde som vid tabellavläsning med v_1 och v_2 frihetsgrader har sannolikhet *y*.

$F_{(0,95,\ 10,\ 10)} = 2{,}978$

Referenstabellerna för *F*-fördelningen har frihetsgraderna v_1 som kolumner respektive v_2 som rader.

$1 - p = 0{,}95$

v_2 / v_1	1	2	3	4	5	10	25	50	100
1	161,45	199,50	215,71	224,58	230,16	241,88	249,26	251,77	253,04
2	18,51	19,00	19,16	19,25	19,30	19,40	19,46	19,48	19,49
3	10,13	9,55	9,28	9,12	9,01	8,79	8,63	8,58	8,55
4	7,71	6,94	6,59	6,39	6,26	5,96	5,77	5,70	5,66
5	6,61	5,79	5,41	5,19	5,05	4,74	4,52	4,44	4,41
10	4,97	4,10	3,71	3,48	3,33	2,98	2,73	2,64	2,59
25	4,24	3,39	2,99	2,76	2,60	2,24	1,96	1,84	1,78
50	4,03	3,18	2,79	2,56	2,40	2,03	1,73	1,60	1,53
100	3,94	3,09	2,70	2,46	2,31	1,93	1,62	1,48	1,39

$1 - p = 0{,}99$

v_2 / v_1	1	2	3	4	5	10	25	50	100
1	4052,18	4999,50	5403,35	5624,58	5763,65	6055,85	6239,83	6302,52	6334,11
2	98,50	99,00	99,17	99,25	99,30	99,40	99,46	99,48	99,49
3	34,12	30,82	29,46	28,71	28,24	27,23	26,58	26,35	26,24
4	21,20	18,00	16,69	15,98	15,52	14,55	13,91	13,69	13,58
5	16,26	13,27	12,06	11,39	10,97	10,05	9,45	9,24	9,13
10	10,04	7,56	6,55	5,99	5,64	4,85	4,31	4,12	4,01
25	7,77	5,57	4,68	4,18	3,86	3,13	2,60	2,40	2,29
50	7,17	5,06	4,20	3,72	3,41	2,70	2,17	1,95	1,83
100	6,90	4,82	3,98	3,51	3,21	2,50	1,97	1,74	1,60

$1 - p = 0{,}999$

v_2 / v_1	1	2	3	4	5	10	25	50	100
1	405284,07	499999,50	540379,20	562499,58	576404,56	605620,97	624016,83	630285,38	633444,33
2	998,50	999,00	999,17	999,25	999,30	999,40	999,46	999,48	999,49
3	167,03	148,50	141,11	137,10	134,58	129,25	125,84	124,66	124,07
4	74,14	61,25	56,18	53,44	51,71	48,05	45,70	44,88	44,47
5	47,18	37,12	33,20	31,09	29,75	26,92	25,08	24,44	24,12
10	21,04	14,91	12,55	11,28	10,48	8,75	7,60	7,19	6,98
25	13,88	9,22	7,45	6,49	5,89	4,56	3,63	3,28	3,09
50	12,22	7,96	6,34	5,46	4,90	3,67	2,79	2,44	2,25
100	11,50	7,41	5,86	5,02	4,48	3,30	2,43	2,08	1,87

DESKRIPTIV STATISTIK

Deskriptiv statistik behandlar hur datamaterial kan sammanfattas och presenteras. Vilka deskriptiva mått som är lämpliga avgörs av vilken datatyp som det rör sig om.

... för alla datatyper

Typvärdet (på engelska "mode") anger det vanligast förekommande värdet i en serie. Typvärdet för talserien {1, 1, 5, 7} är 1.

Proportionen *p* avser slumpsannolikhet eller populationsandel. Motsvarigheten för stickprover är:

$\hat{p} = \frac{x}{n}$ där *n* är det totala antalet utfall och *x* är antalet utfall av den studerade typen

... för ordinala och kvantitativa data

Variationsvidden (på engelska "range") anger avståndet från det minsta värdet till det högsta värdet.

Medianvärdet, *Md*, anger det mittersta värdet vid en rangordnad uppräkning. Om antalet värden är jämnt så motsvarar medianvärdet genomsnittet av de två mittersta värdena. Med talen {1, 3, 9} så blir medianvärdet 3, med talen {1, 3, 5, 9} så blir medianvärdet 4. Medianvärdet kallas även för den andra kvartilen, *Q2*.

$0{,}50 \cdot (n+1)$ anger medianvärdets placering i uppräkningen

Första kvartilen, *Q1*, anger vilket värde som hamnar mitt emellan medianen och seriens lägre ände.
Tredje kvartilen, *Q3*, anger vilket värde som hamnar mitt emellan medianen och seriens högre ände.

$0{,}25 \cdot (n+1)$ anger vilket placeringsnummer som den första kvartilen har i uppräkningen

$0{,}75 \cdot (n+1)$ anger vilket placeringsnummer som den tredje kvartilen har i uppräkningen

Kvartilavståndet, *IQR* (på engelska "inter-quartile range"), anger skillnad i värde mellan *Q1* och *Q3*.

$IQR = Q3 - Q1$

Percentiler delar upp serien i 100 steg. *Q1*, *Q2* och *Q3* motsvaras av percentilerna #25, #50 och #75.

$(x/100) \cdot (n+1)$ anger vilket placeringsnummer som percentil *x* har i uppräkningen

Observera att parametrarna nedanför använder andra tecken när de gäller hela populationen istället för stickprover; $\bar{x}$ ersätts av μ, s ersätts av σ och n ersätts av N.

Medelvärdet (på engelska "mean") är storleken på det teoretiska genomsnittliga värdet. En svaghet med medelvärden (men inte medianvärden) är att de påverkas av kraftigt avvikande värden och skeva fördelningar. Medelvärdet för ett stickprov med n mätvärden skrivs matematiskt enligt nedan:

$$\bar{x} = \frac{x_1 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variansen s^2 är ett mått på mätvärdenas avstånd till medelvärdet; större avvikelser ger större varians. Mätvärdenas avvikelser från medelvärdet kvadreras, summeras och divideras sedan med $n - 1$.

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

Standardavvikelsen s är kvadratroten ur variansen och delar enhet med mätvärdena, till exempel kg.

$$s = \sqrt{s^2}$$

Variationskoefficienten CV anger istället hur stor standardavvikelsen är jämfört med medelvärdet.

$$CV = \frac{s}{\bar{x}}$$

Ifall två oberoende urval har samma populationsvarians ($\sigma_1 = \sigma_2$) så kan den skattas som:

$$\hat{\sigma}^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}$$

Med parade mätvärden avses att två urval hämtar stickprover i par från samma (eller parade) subjekt, varefter mätvärdesparen (x och y) används för att beräkna medelskillnaden $\bar{d}$.

$$\bar{d} = \frac{d_1 + ... + d_n}{n} = \frac{\sum_{i=1}^{n} y_i - x_i}{n}$$

$$s_d^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$$

Enkel linjär regression

Ifall man har ett antal stickprovsvärden som vardera omfattar variablerna x och y så är det möjligt att beräkna vilka värden i räta linjens ekvation som bäst motsvarar tillgängliga data.

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

$$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$SS_e = SS_y - \frac{SS_{xy}^2}{SS_x}$$

$$s_e^2 = \frac{SS_e}{n-2}$$

$$b_1 = \frac{SS_{xy}}{SS_x}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Sammanfattningsvis gäller att:

$y = b_0 + b_1 \cdot x + e$ där e betecknar slumpmässiga avvikelser (i genomsnitt 0)

$$Var(b_1) = \frac{s_e^2}{SS_x}$$

$$Var(b_0) = s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)$$

Räkneexempel:

Barney blev ihop med en tjej och började snart bli allt rundare.

Efter en månad var hans vikt (x) 80 kg, efter två var den 86 kg och efter tre var den 95 kg.

Vid samtliga tillfällen var hans längd (y) 180 cm.

$$b_1 = \frac{\sum_{i=1}^{3} (x_i - 87) \cdot (y_i - 180)}{\sum_{i=1}^{3} (x_i - 87)^2} = \frac{(-7) \cdot (0) + (-1) \cdot (0) + (8) \cdot (0)}{(-7)^2 + (-1)^2 + (8)^2} = \frac{0}{114} = 0$$

$b_0 = 180 - 0 \cdot 87 = 180$ och därmed får vi att $y = 0 \cdot x + 180$

I ett xy-diagram skär denna räta linjens ekvation exakt genom alla tre datapunkterna.

Den motsvarar alltså mätvärdena alldeles perfekt. Men finns det egentligen ett samband..?

I verkligheten är många förhållanden inte strikt linjära men ibland kan det ändå vara möjligt att få ett linjärt samband att framträda genom att utföra en matematisk transformation. Det kan göras med bland annat logaritmering, tillägg av exponenter till x eller användande av x som exponent.

En ytterligare möjlighet är att flera oberoende variabler samspelar och då kan det behövas multipel regression. Detta utförs som regel med hjälp av statistiska datorprogram eftersom beräkningarna snabbt blir allt mer omfattande och komplicerade i takt med att fler oberoende variabler läggs till.

Korrelationskoefficient och förklaringsgrad

Pearsons korrelationskoefficient (efter statistikern **Karl Pearson**) är ett mått på styrkan i ett tänkt linjärt samband mellan två variabler. Värdet varierar från –1 till +1 och ju längre värdet är från 0, desto starkare är det linjära sambandet; negativa samband är också samband, fast i motsatt riktning. Korrelationskoefficienter betecknas med r för stickprover och ρ för populationer.

$$r_{xy} = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}$$

Från korrelationskoefficienten kan man sedan beräkna förklaringsgraden r^2 som anger hur stor andel av en beroende variabels värde som kan förklaras med den oberoende variabeln vid en regression.

Räkneexempel:

> Från ovan vet vi att räta linjens ekvation kan beskriva Barneys längd-viktförhållande exakt.
> Men hur starkt är egentligen det ovan nämnda "sambandet" mellan Barneys vikt och längd?
> Från uträkningen i det föregående räkneexemplet vet vi redan att kvotens täljare blir 0.
> Därmed blir $r_{xy} = 0$ och dess kvadrat $r^2 = 0$. Det finns inte någon styrka alls i "sambandet".
> Alltså: att det går att beskriva en relation med en ekvation utgör inte en korrelation!

Ifall ens mätvärden avviker betydligt från normalfördelningen, eller ifall de är av ordinaltyp, så kan man istället använda Spearmans rangkorrelationskoefficient (efter psykologen **Charles Spearman**). Först ger man mätvärdena rangordningsnummer enligt variablerna som man studerar (till exempel längd och vikt) och sedan beräknas korrelationskoefficienten utifrån dessa rangordningstal.

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(rang(x_i) - rang(y_i) \right)^2}{n \cdot (n^2 - 1)}$$

Medelvärdets standardfel

"Standard error of the mean" (SEM), på svenska "medelvärdets standardfel" eller "medelfelet", är ett mått på hur väl ett stickprovs medelvärde kan användas som uppskattning av hela populationens.

$$SEM_{\bar{x}} = s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{s^2}}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

Standardfel används vid beräkning av konfidensintervall och finns även för andra parametrar än $\bar{x}$.

Konfidensintervall

Ett konfidensintervall omfattar en viss andel av utfallen för en sannolikhetsfördelning. Den andel som ligger utanför konfidensintervallet betecknas α medan den omfattade andelen betecknas $1 - \alpha$. Vid enkelsidiga konfidensintervall är hela α samlat på samma sida i sannolikhetsfördelningen, antingen den övre (högra) eller den nedre (vänstra), medan dubbelsidiga konfidensintervall omges av två områden som vardera har storleken $\alpha/2$. En mycket vanlig omfattning på konfidensintervall är 95 % av sannolikhetsfördelningen vilket motsvaras av att α = 5 %. Till exempel innebär ett 95 % konfidensintervall för μ att det är 95 % sannolikhet för att μ ligger inom konfidensintervallet.

Konfidensintervall för μ när σ är känd:

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

$$\bar{x} - z_{(1-\alpha)} \cdot \sqrt{\frac{\sigma^2}{n}} \text{ respektive } \bar{x} + z_{(1-\alpha)} \cdot \sqrt{\frac{\sigma^2}{n}}$$

Konfidensintervall för μ när σ är okänd och urvalet är stort:

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

$$\bar{x} - z_{(1-\alpha)} \cdot \sqrt{\frac{s^2}{n}} \text{ respektive } \bar{x} + z_{(1-\alpha)} \cdot \sqrt{\frac{s^2}{n}}$$

Konfidensintervall för μ när σ är okänd och urvalet är litet och normalfördelat:

$$\bar{x} \pm t_{(1-\alpha/2, n-1)} \cdot \sqrt{\frac{s^2}{n}}$$

$$\bar{x} - t_{(1-\alpha, n-1)} \cdot \sqrt{\frac{s^2}{n}} \text{ respektive } \bar{x} + t_{(1-\alpha, n-1)} \cdot \sqrt{\frac{s^2}{n}}$$

Konfidensintervall för $\overline{D}$ (den "sanna" genomsnittliga skillnaden mellan parade data):

$$\bar{d} \pm t_{(1-\alpha/2,\,n-1)} \cdot \sqrt{\frac{s_d^2}{n}}$$

$$\bar{d} - t_{(1-\alpha,\,n-1)} \cdot \sqrt{\frac{s_d^2}{n}} \text{ respektive } \bar{d} + t_{(1-\alpha,\,n-1)} \cdot \sqrt{\frac{s_d^2}{n}}$$

Konfidensintervall för skillnaden mellan μ_A och μ_B när det kan antagas att $\sigma_A = \sigma_B$:

$$\bar{x}_A - \bar{x}_B \pm t_{(1-\alpha/2,\,n_A+n_B-2)} \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$$\bar{x}_A - \bar{x}_B - t_{(1-\alpha,\,n_A+n_B-2)} \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \text{ respektive } \bar{x}_A - \bar{x}_B + t_{(1-\alpha,\,n_A+n_B-2)} \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

Konfidensintervall för skillnaden mellan μ_A och μ_B när det kan antagas att $\sigma_A \neq \sigma_B$:

$$\bar{x}_A - \bar{x}_B \pm t_{(1-\alpha/2,\,v)} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

$$\bar{x}_A - \bar{x}_B - t_{(1-\alpha,\,v)} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \text{ respektive } \bar{x}_A - \bar{x}_B + t_{(1-\alpha,\,v)} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

där $v = \dfrac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\left(\frac{s_A^2}{n_A}\right)^2 / (n_A - 1) + \left(\frac{s_B^2}{n_B}\right)^2 / (n_B - 1)}$

Konfidensintervall för σ^2:

$$\frac{(n-1) \cdot s^2}{\chi^2_{(1-\alpha/2,\,n-1)}} < \sigma^2 < \frac{(n-1) \cdot s^2}{\chi^2_{(\alpha/2,\,n-1)}}$$

$$\frac{(n-1) \cdot s^2}{\chi^2_{(1-\alpha,\,n-1)}} < \sigma^2 \text{ respektive } \sigma^2 < \frac{(n-1) \cdot s^2}{\chi^2_{(\alpha,\,n-1)}}$$

Konfidensintervall för p när n är stort, gärna över 20:

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

$$\hat{p} - z_{(1-\alpha)} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \text{ respektive } \hat{p} + z_{(1-\alpha)} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

Konfidensintervall för skillnaden mellan p_A och p_B:

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

$$\hat{p}_A - \hat{p}_B - z_{(1-\alpha)} \cdot \sqrt{\frac{\hat{p}_A \cdot (1-\hat{p}_A)}{n} + \frac{\hat{p}_B \cdot (1-\hat{p}_B)}{n}} \text{ re. } \hat{p}_A - \hat{p}_B + z_{(1-\alpha)} \cdot \sqrt{\frac{\hat{p}_A \cdot (1-\hat{p}_A)}{n} + \frac{\hat{p}_B \cdot (1-\hat{p}_B)}{n}}$$

Konfidensintervall för parametern b_0 vid enkel linjär regression:

$$b_0 \pm t_{(1-\alpha/2,\, n-2)} \cdot \sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)}$$

$$b_0 - t_{(1-\alpha,\, n-2)} \cdot \sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)} \text{ respektive } b_0 + t_{(1-\alpha,\, n-2)} \cdot \sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)}$$

Konfidensintervall för parametern b_1 vid enkel linjär regression:

$$b_1 \pm t_{(1-\alpha/2,\, n-2)} \cdot \sqrt{\frac{s_e^2}{SS_x}}$$

$$b_1 - t_{(1-\alpha,\, n-2)} \cdot \sqrt{\frac{s_e^2}{SS_x}} \text{ respektive } b_1 + t_{(1-\alpha,\, n-2)} \cdot \sqrt{\frac{s_e^2}{SS_x}}$$

Konfidensintervall för $\bar{y}$ när x har värdet x_0 vid enkel linjär regression:

$$b_0 + b_1 \cdot x_0 \pm t_{(1-\alpha/2,\, n-2)} \cdot \sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x} \right)}$$

Prediktionsintervall för nästa enskilda värde på y när x har värdet x_0 vid enkel linjär regression:

$$b_0 + b_1 \cdot x_0 \pm t_{(1-\alpha/2,\, n-2)} \cdot \sqrt{s_e^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x} \right)}$$

ANALYTISK STATISTIK

Analytisk statistik behandlar hur datamaterial kan användas för att dra slutsatser, så kallad statistisk inferens. Vilka analytiska tester som är lämpliga styrs av dels antalet mätvärden och dessas fördelning, dels av vad det är som undersöks.

Parametriska tester kräver normalfördelade data och ett tillräckligt stort antal observationer medan icke-parametriska tester alltid kan användas. Det specifika testvalet beror därutöver även på om man gör jämförelser eller studerar samband.

Estimatorer

En estimator är en statistisk uppskattning av en parameters värde. I skrift skiljs estimatorn från parametern genom att den förses med en hatt.

$$\hat{\mu}, \hat{\sigma}, \hat{\sigma}^2, \hat{p}$$

Man kan även säga att s utgör ett stickprovs estimator av σ och $\bar{x}$ utgör ett stickprovs estimator av μ.

Estimatorer bedöms utifrån hur väl de uppnår två egenskaper, väntevärdesriktighet respektive effektivitet. Väntevärdesriktighet bedöms efter hur väl estimatorn motsvarar parametern och mäts som storleken på dess förväntade avvikelse från denna. För populationens genomsnitt μ skrivs detta:

$$Bias(\hat{\mu}) = E(\hat{\mu}) - \mu$$

En estimators effektivitet motsvaras i sin tur av dess varians.

$$Var(\hat{\mu})$$

Hur bra olika estimatorer är kan därmed bedömas som kvadraten av skillnaden mellan estimatorer och motsvarande parametrar – estimatorernas "mean squared error". För populationens medelvärde μ kan detta skrivas:

$$MSE = E(\hat{\mu} - \mu)^2 = Bias(\hat{\mu})^2 + Var(\hat{\mu})$$

Hypotesprövningar

Vid hypotesprövning ställer man upp två hypoteser, nollhypotesen H_0 och alternativhypotesen H_1. H_0 motsvarar ett skeptiskt grundantagande om att observerade fenomen bara är slumpmässiga medan H_1 motsvarar ett antagande om att normal slump inte räcker som förklaring. H_0 kan även ses som "det som ska motbevisas". Om en undersökning påvisar ett utfall som ska vara mycket ovanligt enligt H_0 så brukar det tolkas som att H_0 kan förkastas. Uttryckt som en sannolikhetsfördelning representeras ovanliga utfall av signifikansnivån α och vanliga utfall av konfidensgraden $1 - \alpha$.

En vanlig gräns för statistisk signifikans brukar vara att $\alpha = 0{,}05$ vilket även kallas för "enstjärnig signifikans". Vid tvåstjärnig signifikans är $\alpha = 0{,}01$ och vid trestjärnig signifikans är $\alpha = 0{,}001$. Ju konservativare (lägre) signifikansgräns vi väljer desto bredare blir konfidensintervallet, desto större variationer tolkas som normala och desto svårare blir det att förkasta H_0. Fördelen med att välja en konservativ signifikansnivå är att vi är säkrare på att vi inte förkastar en korrekt H_0 och accepterar en felaktig H_1 (typ 1-fel, falskt positivt resultat) medan nackdelen är att risken blir större för att vi behåller en felaktig H_0 och förkastar en korrekt H_1 (typ 2-fel, falskt negativt resultat). Det omvända gäller för en liberal (hög) signifikansgräns: större risk för typ 1-fel, mindre risk för typ 2-fel. Sannolikheten för ett typ 1-fel utgörs av α medan sannolikheten för ett typ 2-fel betecknas β.

En viktig sak att tänka på är risken för massignifikanser; ju fler försök, desto större är risken för att man accepterar någon felaktig H_1. Med $\alpha = 0{,}05$ är ungefär var 20:e signifikans falskt positiv!

Power

Sannolikheten för att acceptera en korrekt H_1 motsvarar alltså $1 - \beta$. Detta är det statistiska testets styrka, dess "power". Ju större stickprovet är, eller ju större den sökta effekten är, desto större power har testet. Under förutsättning att hela populationens standardavvikelse σ är känd och att man vet hur stor skillnad som man önskar kunna påvisa (nedan benämnt $\mu_0 - \mu_1$) vid signifikansnivån α så kan man beräkna hur stort stickprov som behövs för att uppnå en önskad nivå av power $(1 - \beta)$.

$$n > \frac{\left(z_{(1-\beta)} + z_{(1-\alpha/2)}\right)^2 \cdot \sigma^2}{\left(\mu_0 - \mu_1\right)^2}$$ för ett dubbelsidigt test

$$n > \frac{\left(z_{(1-\beta)} + z_{(1-\alpha)}\right)^2 \cdot \sigma^2}{\left(\mu_0 - \mu_1\right)^2}$$ för ett enkelsidigt test

Testfunktioner för medelvärden

Utifrån nollhypotesens antagande om populationens medelvärde, μ_0, kan man beräkna ett z-värde (eller t-värde) och avläsa sannolikheten för observationen. Är sannolikheten mindre än den valda signifikansnivån så är skillnaden statistisk signifikant och nollhypotesen kan förkastas. Glöm inte att vid dubbelsidiga tester (avvikelser uppåt och nedåt) så multipliceras avlästa sannolikheter med 2.

z-test för μ när σ är känd:

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

z-test för μ när σ är okänd och urvalet är stort:

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

t-test för μ när σ är okänd och urvalet är litet, om stickproverna är normalfördelade:

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$ där antalet frihetsgrader är $n - 1$

Testfunktion för varianser

När nollhypotesen gör ett antagande om standardavvikelsen, σ^2_0, så kan man utifrån det beräkna ett χ^2-värde och jämföra det med χ^2-värdet för den valda konfidensgraden.

$$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma_0^2}$$ där antalet frihetsgrader är $n - 1$

Testfunktion för proportioner

z-test för p utifrån nollhypotesens antagande om proportion, p_0, när n är stort (gärna över 20):

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

Testfunktioner för jämförelser av två urval

Två urval, A och B, kan jämföras för att se om det finns en statistiskt signifikant skillnad. Urvalens egenskaper avgör vilket test som är lämpligt.

t-test för $\bar{d}$:

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}}$$

t-test för skillnaden mellan μ_A och μ_B när det kan antagas att $\sigma_A = \sigma_B$:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

t-test för skillnaden mellan μ_A och μ_B när det kan antagas att $\sigma_A \neq \sigma_B$:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

där $$v = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B - 1}}$$

F-test för skillnaden mellan σ_A och σ_B:

$F = \frac{s_A^2}{s_B^2}$ vilket jämförs med gränsvärdet $F_{(1-\alpha/2,\, n_A-1,\, n_B-1)}$

z-test för skillnaden mellan p_A och p_B:

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p} \cdot (1 - \hat{p}_0) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

där $\hat{p}_0 = \frac{n_A \cdot \hat{p}_A - n_B \cdot \hat{p}_B}{n_A + n_B}$

Testfunktioner för regressionslinjer

Vid enkel linjär regression är en av variablerna beroende (y) och den andra oberoende (x).

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

Residualen ε_i utgör den slumpmässiga och normalfördelade skillnaden mellan regressionslinjen och det uppmätta mätvärdet vid x_i (se "linjär regression" för skattningar av β_1 och β_0 samt definitioner av SS_x och S_e^2).

t-test för β_0:

$$t = \frac{b_0 - \beta_{0,0}}{\sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)}}$$ där antalet frihetsgrader är $n - 2$

t-test för β_1:

$$t = \frac{b_1 - \beta_{1,0}}{\sqrt{\frac{s_e^2}{SS_x}}}$$ där antalet frihetsgrader är $n - 2$

(där $\beta_{1,0} = 0$ kan användas för ett tvåsidigt test av om det alls finns något linjärt samband)

Testfunktioner för korrelationskoefficienter

t-test för ρ_{XY} (populationens Pearson-korrelation mellan X och Y):

$$t = r_{xy} \cdot \sqrt{\frac{(n-2)}{(1-r^2)}}$$ där antalet frihetsgrader är $n - 2$

t-test för ρ_S (populationens Spearman-rangkorrelation mellan X och Y):

$$t = r_s \cdot \sqrt{\frac{(n-2)}{(1-r^2)}}$$ där antalet frihetsgrader är $n - 2$

Testfunktioner för variansanalys (ANOVA)

Variansanalys används för att jämföra tre eller fler populationers medelvärden. Variansanalys brukar göras med statistikprogram eftersom beräkningsvolymen kan bli stor och detta stycke avgränsas därför till några enkla exempel som bör ge en fingervisning om vad ANOVA kan användas till.

Vid ensidig variansanalys finns det a dataserier med n mätvärden vardera. Det totala medelvärdet betecknas μ, dataseriernas medelvärdens avvikelser från detta betecknas a_i och enskilda mätvärden betecknas y_{ij}. Enskilda mätvärdens avvikelser från dataseriernas medelvärden utgör residualen ε_{ij}.

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Kvadratsumman för datamängdens totala variation betecknas SS_T. Den kan anses bestå av dels variationen mellan dataserierna, SS_A, och dels variationen inom dataserierna, SS_E.

$$SS_T = SS_A + SS_E$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 \quad \text{där antalet frihetsgrader är } n \cdot a - 1$$

$$SS_A = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_i - \bar{y})^2 \quad \text{där antalet frihetsgrader är } a - 1$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 \quad \text{där antalet frihetsgrader är } n \cdot a - a$$

F-test för dataseriernas medelvärden (H_0 är att alla medelvärden är lika):

$$F = \frac{\left(\frac{SS_A}{a-1}\right)}{\left(\frac{SS_E}{n \cdot a - a}\right)} \quad \text{där antalet frihetsgrader är } a - 1 \text{ respektive } n \cdot a - a$$

En variant av ovanstående är att mätvärdena kan vara indelade i n block, där varje dataserie har ett mätvärde i varje block och de olika blocken motsvarar olika omständigheter (inte olikt parade data).

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

$$SS_T = SS_A + SS_{block} + SS_E$$

$$SS_{block} = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_j - \bar{y})^2 \quad \text{där antalet frihetsgrader är } n - 1$$

$$SS_E = SS_T - SS_A - SS_{block} \quad \text{där antalet frihetsgrader är } (a-1)\cdot(n-1)$$

$$F = \frac{\left(\frac{SS_A}{a-1}\right)}{\left(\frac{SS_E}{(a-1)\cdot(n-1)}\right)} \quad \text{där antalet frihetsgrader är } a - 1 \text{ respektive } (a-1)\cdot(n-1)$$

Testfunktioner för sannolikhetsfördelningar av icke-kvantitativa data

Om man har en nollhypotes om sannolikheterna för olika alternativa utfall, till exempel hur många barn som föds med olika ögonfärger i en familj, så kan det testas huruvida observationerna stöder nollhypotesen. I formlerna nedanför betecknar p_i sannolikheten och o_i antalet observationer för kategori (möjligt utfall) nummer i (av c). Det totala antalet observationer är n. Pålitliga resultat fordrar att $n \cdot p_i$ är minst 5 för varje kategori; om inte så är ett alternativ att slå ihop kategorier.

χ^2-test för i förväg bestämda sannolikheter:

$$\chi^2 = \sum_{i=1}^{c} \frac{(o_i - n \cdot p_i)^2}{n \cdot p_i}$$ där antalet frihetsgrader är $c - 1$

χ^2-test för att en viss typ av sannolikhetsfördelning föreligger (p_i beräknas från urvalet):

$$\chi^2 = \sum_{i=1}^{c} \frac{(o_i - n \cdot p_i)^2}{n \cdot p_i}$$ där antalet frihetsgrader är $c - 1 -$ [antal skattade parametrar bakom p_i]

Det är även möjligt att bedöma ifall olika stickprover följer samma fördelning eller skiljer sig åt. I formlerna nedanför betecknar r stickprover (rader i tabell) och c typer av utfall (kolumner i tabell). För olika kombinationer av dessa (rutor i tabell) betecknas observationer o_{ij} och sannolikheter p_{ij}.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - n \cdot \hat{p}_{ij})^2}{n \cdot \hat{p}_{ij}}$$ där antalet frihetsgrader är $(r - 1) \cdot (c - 1)$

där $\hat{p}_{ij} = \frac{\sum_{i=1}^{r} o_{ij}}{n} \cdot \frac{\sum_{j=1}^{c} o_{ij}}{n}$ eller $\hat{p}_{ij} = \frac{[\text{summa rad } i] \cdot [\text{summa kolumn } j]}{n^2}$ för kombinationen ij

Räkneexempel:

Barney har många strategier för att nå sina mål, ofta genom att låtsas vara något han inte är. Ibland låtsas han vara astronaut, ibland tidsresenär, ibland yogainstruktör, ibland blind. Som nollhypotes antar vi att dessa fyra är lika vanliga; $p_1 = p_2 = p_3 = p_4 = 1{,}00 / 4 = 0{,}25$. Efter 20 observationer blev fördelningen {10, 2, 6, 2} för fredagar.

$$\chi^2 = \frac{(20 - 20 \cdot 0{,}25)^2}{20 \cdot 0{,}25} + \frac{(2 - 20 \cdot 0{,}25)^2}{20 \cdot 0{,}25} + \frac{(6 - 20 \cdot 0{,}25)^2}{20 \cdot 0{,}25} + \frac{(2 - 20 \cdot 0{,}25)^2}{20 \cdot 0{,}25}$$

$$\chi^2 = \frac{(20-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(2-5)^2}{5} = \frac{(15)^2 + (-3)^2 + (1)^2 + (-3)^2}{5} = \frac{244}{5} = 48{,}8$$

Avläsning av χ^2-fördelningen för $20 - 1 = 19$ frihetsgrader och $p = 0{,}995$ ger 38,6. Observationerna skiljer sig alltså mycket signifikant från nollhypotesen, som förkastas. Vill man gå längre så kan man jämföra om olika veckodagar har olika strategifördelningar.

Icke-parametriska testfunktioner

Icke-parametriska testfunktioner lämpar sig för ordinaldata och icke normalfördelade kvantitativa data. De fordrar dock ett större antal observationer för att uppnå en viss signifikansnivå.

För att testa om ett enkelt eller parat stickprov har en median som avviker signifikant från nollhypotesens Md_0 kan man använda teckentestet. Teckentestet anses vara det första statistiska testet och användes år 1710 av läkaren **John Arbuthnot** i en studie där han visade att slumpen inte räckte som förklaring till att det föddes fler pojkar än flickor.
Sannolikheten för att en median ska överskridas av m utav n värden ges av binomialfördelningen för n med $p = 0{,}50$ (vid dubbelsidigt test multipliceras avläst sannolikhet med 2). Om avläst sannolikhet är lägre än vald signifikansnivå så förkastas H_0.

Räkneexempel:

> IQ-medianen bland Barneys tillfälliga kvinnliga bekantskaper påstås vara högst 95.
> Barney accepterar utmaningen och under en månad genomför han postcoitala IQ-tester.
> Resultatet av stickprovet blev totalt 8 mätvärden {85, 90, 95, 100, 100, 105, 110, 120}.
> Totalt 5 av 8 värden överskrider Md_0 vilket har sannolikheten 1 – 0,85547 = 0,14453.
> En sannolikhet på 14 % räcker inte för att avfärda nollhypotesen.

När man ska testa om det föreligger en signifikant skillnadsmedian för två parade populationer så kan man använda Wilcoxons teckenrangtest. Först beräknas varje datapars inbördes skillnader.

$$d_i = y_i - x_i \quad \text{för parserien } \{1, ..., n\}$$

Sedan rangordnas d_i utifrån absolutvärdenas storlek (oberoende av om d_i är positiva eller negativa). Slutligen räknar man ut rangsummorna $R_{positiv}$ som är rangsumman för alla d_i som har positiva värden och $R_{negativ}$ som är rangsumman för alla d_i som har negativa värden. Den mindre rangsumman blir vår teststatistika W. Signifikant avvikelse från H_0 föreligger när W är mindre än det kritiska värdet för n med den valda signifikansnivån α (se tabellerna på nästa sida). Om n är större än 25 så kan man dock överväga att använda en normalapproximation istället för tabellavläsning.

$$z = \frac{W - \frac{n \cdot (n+1)}{4}}{\sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{24}}}$$

Kritiska värden för Wilcoxons teckenrangtest. Rad väljs utifrån n och kolumn utifrån önskat α.

	Enkelsidiga				Dubbelsidiga				
n	$\alpha = 0{,}050$	$\alpha = 0{,}025$	$\alpha = 0{,}010$	$\alpha = 0{,}005$	n	$\alpha = 0{,}100$	$\alpha = 0{,}050$	$\alpha = 0{,}020$	$\alpha = 0{,}010$
6	2				6	2			
7	3	2			7	3	2		
8	5	3	1		8	5	3	1	
9	8	5	3	1	9	8	5	3	1
10	10	8	5	3	10	10	8	5	3
11	13	10	7	5	11	13	10	7	5
12	17	13	9	7	12	17	13	9	7
13	21	17	12	9	13	21	17	12	9
14	25	21	15	12	14	25	21	15	12
15	30	25	19	15	15	30	25	19	15
16	35	29	23	19	16	35	29	23	19
17	41	34	27	23	17	41	34	27	23
18	47	40	32	27	18	47	40	32	27
19	53	46	37	32	19	53	46	37	32
20	60	52	43	37	20	60	52	43	37
21	67	58	49	42	21	67	58	49	42
22	75	65	55	48	22	75	65	55	48
23	83	73	62	54	23	83	73	62	54
24	91	81	69	61	24	91	81	69	61
25	100	89	76	68	25	100	89	76	68
26	110	98	84	75	26	110	98	84	75
27	119	107	92	83	27	119	107	92	83
28	130	116	101	91	28	130	116	101	91
29	140	126	110	100	29	140	126	110	100
30	151	137	120	109	30	151	137	120	109
31	163	147	130	118	31	163	147	130	118
32	175	159	140	128	32	175	159	140	128
33	187	170	151	138	33	187	170	151	138
34	200	182	162	148	34	200	182	162	148
35	213	195	173	159	35	213	195	173	159
36	227	208	185	171	36	227	208	185	171
37	241	221	198	182	37	241	221	198	182
38	256	235	211	194	38	256	235	211	194
39	271	249	224	207	39	271	249	224	207
40	286	264	238	220	40	286	264	238	220
41	302	279	252	233	41	302	279	252	233
42	319	294	266	247	42	319	294	266	247
43	336	310	280	261	43	336	310	280	261
44	353	327	296	276	44	353	327	296	276
45	371	343	312	291	45	371	343	312	291
46	389	361	328	307	46	389	361	328	307
47	407	378	345	322	47	407	378	345	322
48	426	396	362	339	48	426	396	362	339
49	446	415	379	355	49	446	415	379	355
50	466	434	397	373	50	466	434	397	373

För att testa om två grupper har samma median kan man använda Mann-Whitney-Wilcoxons *U*-test (efter matematikern **Henry Mann**, studenten **Donald Whitney** och statistikern **Frank Wilcoxon**). Först rangordnas alla mätvärden. Lika stora mätvärden får sitt rangintervalls median som rangtal.

Exempel: urvalen {a, d, e} och {b, c, d, f} får rangtalen {1, 4.5, 6} och {2, 3, 4.5, 7}.

Sedan summeras urvalens rangtal till rangsummorna R_1 och R_2.

Exempel: serierna {1, 4.5, 6} och {2, 3, 4.5, 7} får rangsummorna $R_1 = 11{,}5$ och $R_2 = 16{,}5$.

Slutligen beräknas *U*-värden utifrån de respektive rangsummorna.

$$U_1 = n_1 \cdot n_2 + (n_1 \cdot (n_1 + 1))/2 - R_1$$

$$U_2 = n_1 \cdot n_2 + (n_2 \cdot (n_2 + 1))/2 - R_2$$

Om antingen n_1 eller n_2 är 10 eller lägre så jämförs det mindre av *U*-värdena med tabellvärden över kritiska gränser (se tabellerna på nästa sida). Om *U*-värdet är lägre än tabellvärdet för n_1 och n_2 så förkastas H_0. Om både n_1 och n_2 är större än 10 så används en normalapproximation istället för tabellavläsning.

$$z = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}}$$ där U är U_1 eller U_2 (båda ger samma resultat, ty $n_1 \cdot n_2 = U_1 + U_2$)

Kruskal-Wallis test (efter matematikern **William Kruskal** och ekonomen **Wilson Wallis**) motsvarar Mann-Whitney-Wilcoxons *U*-test men för *k* populationer. Det kan liknas vid envägs-ANOVA för ordinaldata eller icke normalfördelade kvantitativa data.

Först rangordnas alla mätvärden, från minst till störst och oberoende av urvalsgruppstillhörighet, och sedan beräknas rangsummorna för de olika urvalsgrupperna. Urvalsgrupperna n_i behöver inte vara av samma storlek men om varje n_i är större än 5 så kan man beräkna χ^2-teststatistikan *K*.

$$K = \frac{12}{n \cdot (n+1)} \cdot \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3 \cdot (n+1)$$ där antalet frihetsgrader är $k - 1$

Vid statistisk signifikans skiljer sig urvalsgrupperna.

Kritiska värden för Mann-Whitney-Wilcoxons *U*-test. Rad och kolumn väljs utifrån n_1 respektive n_2.

$\alpha = 0{,}050$ dubbelsidigt (0,025 enkelsidigt)

n_2 / n_1	2	3	4	5	6	7	8	9	10
2									
3					1	1	2	2	3
4				1	2	3	4	4	5
5			1	2	3	5	6	7	8
6		1	2	3	5	6	8	10	11
7		1	3	5	6	8	10	12	14
8		2	4	6	8	10	13	15	17
9		2	4	7	10	12	15	17	21
10		3	5	8	11	14	17	20	23
11		3	6	9	13	16	19	23	26
12	1	4	7	11	14	18	22	26	29
13	1	4	8	12	16	20	24	28	33
14	1	5	9	13	17	22	26	31	36
15	1	5	10	14	19	24	29	34	39
16	1	6	11	15	21	26	31	37	42
17	2	6	11	17	22	28	34	39	45
18	2	7	12	18	24	30	36	42	48
19	2	7	13	19	25	32	38	45	52
20	2	8	14	20	27	34	41	48	55

$\alpha = 0{,}010$ dubbelsidigt (0,005 enkelsidigt)

n_2 / n_1	2	3	4	5	6	7	8	9	10
2									
3									
4							1	1	2
5						1	1	2	3	4
6					1	2	3	4	5	6
7					1	3	4	6	7	9
8				1	2	4	6	7	9	11
9				1	3	5	7	9	11	13
10				2	4	6	9	11	13	16
11				2	5	7	10	13	16	18
12			1	3	6	9	12	15	18	21
13			1	3	7	10	13	17	20	24
14			1	4	7	11	15	18	22	26
15			2	5	8	12	16	20	24	29
16			2	5	9	13	18	22	27	31
17			2	6	10	15	19	24	29	34
18			2	6	11	16	21	26	31	37
19			3	7	12	17	22	28	33	39
20			3	8	13	18	24	30	36	42

ADDENDA ET CORRIGENDA

Tillägg, korrigeringar och andra förändringar som har gjorts sedan originalversionen.

2015-04-05

Beskrivningen av centrala gränsvärdessatsen har förbättrats.