

Forskningsmetodik 2006

lektion 2

Per Olof Hulth
hulth@physto.se

Föreläsning 2

Forskningsmetodik HT 2006

Slumpmässiga och systematiska mätfel

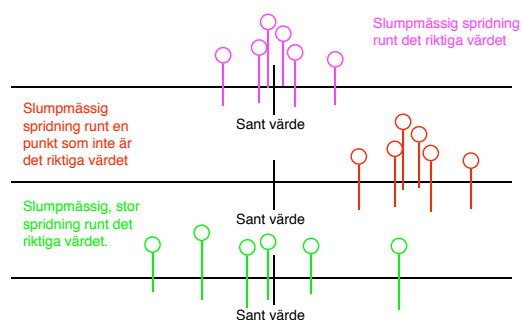
Man skiljer på två typer av fel (osäkerheter) vid mätningar:

1. Slumpmässiga fel

Positiva fel lika vanliga som negativa, små fel vanligare än stora fel. Mätningar ligger symmetriskt runt sanna värdet.

2. Systematiska fel

Ett systematiskt skift av mätningar på grund av något okänt fel, tex mätinstrumentet ej nollställt.



Slumpmässiga fel lättare att beräkna!

Föreläsning 2

Forskningsmetodik HT 2006

Ex 1 sidan 58

Medelvärde, median och typvärde

$$\text{Medelvärde} = \bar{x} = \frac{m_1 + m_2 + \dots + m_N}{N} = \frac{\text{Summan}}{\text{antalet}} = \frac{\sum_{i=1}^N m_i}{N}$$

Medelvärdet skrivs även som μ

Median delar en fördelning i mittenvärdet i en sorterad fördelning.
Tex:

22, 24, 29, 31, 55, 92 ger median = 30 (jämt antal mätpunkter)

22, 24, 29, 31, 34, 55, 92 ger median = 31 (udda antal mätpunkter)

Typvärdet är det vanligaste värdet i en fördelning

Längden på forskningsmetodikstudenter

N=23

151, 155, 156, 156, 158, 160, 164, 166, 169, 169, 170, 171, 172, 172, 174, 175, 175, 175, 178, 180, 182, 190, 193

$$\text{Medelvärde} = \bar{x} = \frac{151 + 155 + \dots + 190 + 193}{23} = \frac{3911}{23} = 170,0\text{cm}$$

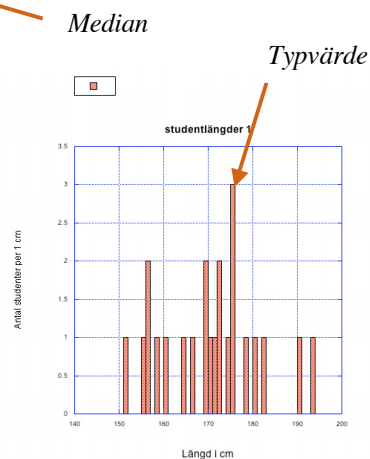
Median = 171cm

Maxvärde(typvärde) = 175cm

Medelvärde => Mean

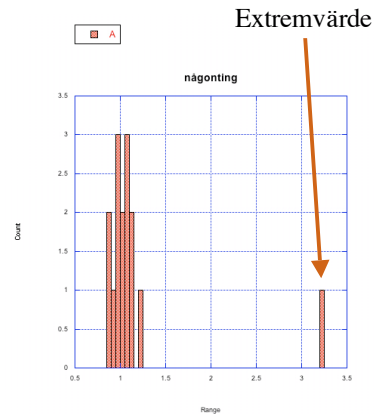
Median=> Median

Typvärde => Mode



Extremvärden, “utliggare”, “Outliers”

Extremvärden (outliers, på svenska kan man även se termen “utliggare”) kan påverka medelvärdet men påverkar ofta inte median och typvärde
Behandlingen av extremvärden är grannliga. Oftast uppstår de p g a något mätfel, men de kan också dölja den stora upptäckten.



Mätfel?

Upptäckt?

Föreläsning 2

Forskningsmetodik HT 2006

Exempel 3 sidan 152

Table 4.3 Comparison of Mean, Median, and Mode

Measure	Definition	How common?	Existence	Takes every value into account?	Affected by outliers?	Advantages
Mean	$\frac{\text{sum of all values}}{\text{total number of values}}$	most familiar “average”	always exists	yes	yes	commonly understood; works well with many statistical methods
Median	middle value	common	always exists	no (aside from counting the total number of values)	no	when there are outliers, may be more representative of an “average” than the mean
Mode	most frequent value	sometimes used	may be no mode, one mode, or more than one mode	no	no	most appropriate for data at the nominal level of measurement (see Section 2.1)

Föreläsning 2

Forskningsmetodik HT 2006

Viktade medelvärden

När vissa mätvärden har större betydelse än andra använder man ett viktat medelvärde:

$$\bar{X}_{viktat} = \frac{\sum (x_i \cdot vikt_i)}{\sum_i vikt_i}$$

Viktade medelvärden används när man bestämt sig för att tillmäta vissa mätpunkter en större vikt än andra, det kan t.ex. röra sig om medelbetyg (där man kan välja att ge större vikt åt ämnen som är särskilt relevanta för en utbildning).

Ett vanligt tillfälle då man använder sig av viktade medelvärden är när man lägger samman mätresultat med olika storlekar på felen. I det läget finns det en väl definierad procedur för hur dessa skall viktas samman, så att de värden som har den lägsta osäkerheten bidrar mest till det gemensamma medelvärdet.

Föreläsning 2

Forskningsmetodik HT 2006

Exempel 7 sidan 154

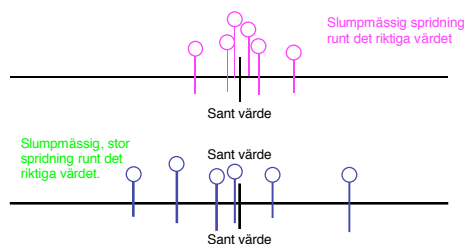
Spridningen av en fördelning

Spridningen av mätpunkterna ger information om mätningen

Fullvidd, avståndet från lägsta till högsta värdet

Kvartiler, delar upp datamängden i 4 lika stora delar

Percentiler, delar upp datamängden i 100 lika stora delar



Viktigaste måttet:

$$\text{Variansen} = \frac{\sum_i (\bar{x} - x_i)^2}{N - 1}$$

Föreläsning 2

Forskningsmetodik HT 2006

Exempel 3 sidan 163

Kvartiler



Föreläsning 2

Forskningsmetodik HT 2006

Standardavvikelsen

Standardavvikelsen är definierad som:

$$\text{standardavvikelsen}(\sigma) = \sqrt{\text{variansen}} = \sqrt{\frac{\sum_i (\bar{x} - x_i)^2}{N - 1}}$$

$\sigma(\text{sigma})$

Standardavvikelsen har samma enhet (sort) som
mätvärdena

Exempel 4 sidan 173

Föreläsning 2

Forskningsmetodik HT 2006

Tumregel för standardavvikelsen

Standardavvikelsen $\approx$ fullvidden/4

Vet vi standardavvikelsen är:

Lägsta värdet $\approx$ medelvärde - 2*standardavvikelsen

Högsta värdet $\approx$ medelvärde + 2*standardavvikelsen

Föreläsning 2

Forskningsmetodik HT 2006

Exempel 6 sidan 175

Mät längden på en parkeringsficka

Gör en mängd
mätningar och
observera
fördelningen av
mätningarna

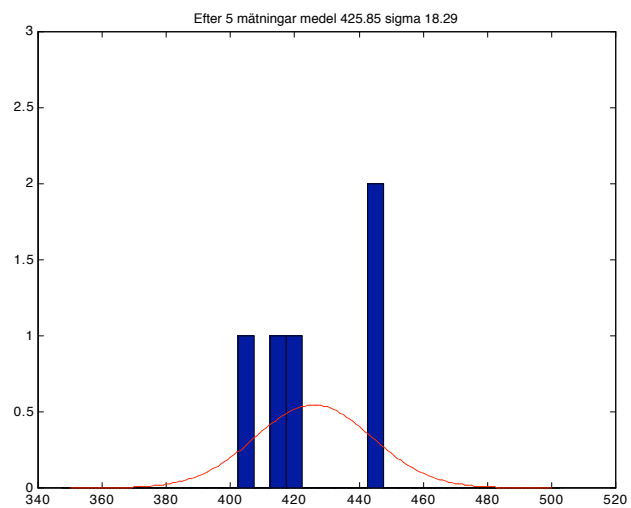
“Sann” längd
428,0 cm



← L →

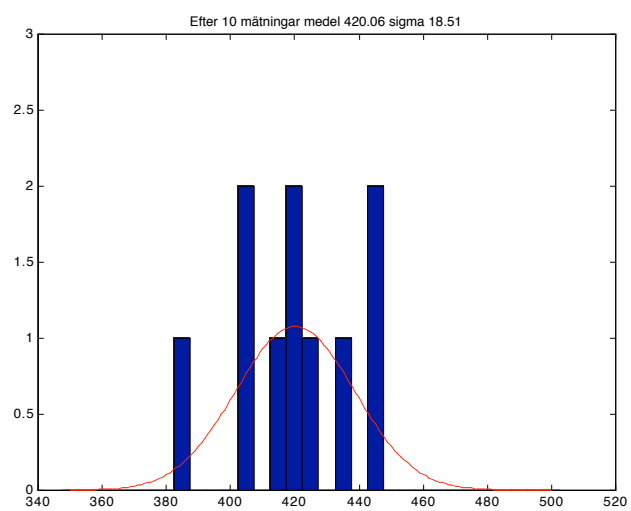
Föreläsning 2

Forskningsmetodik HT 2006



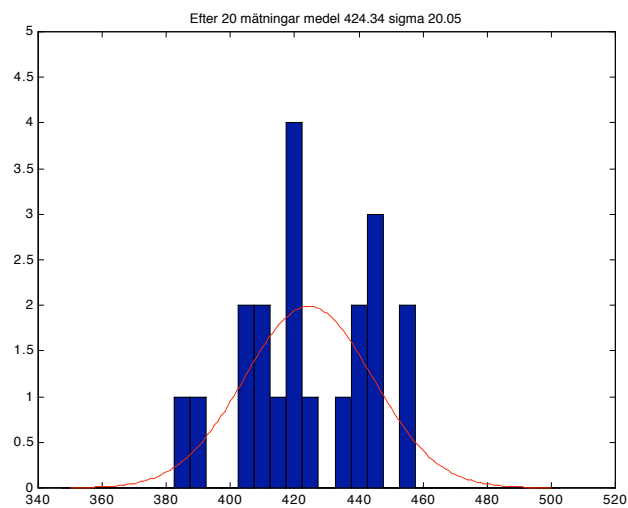
Föreläsning 2

Forskningsmetodik HT 2006



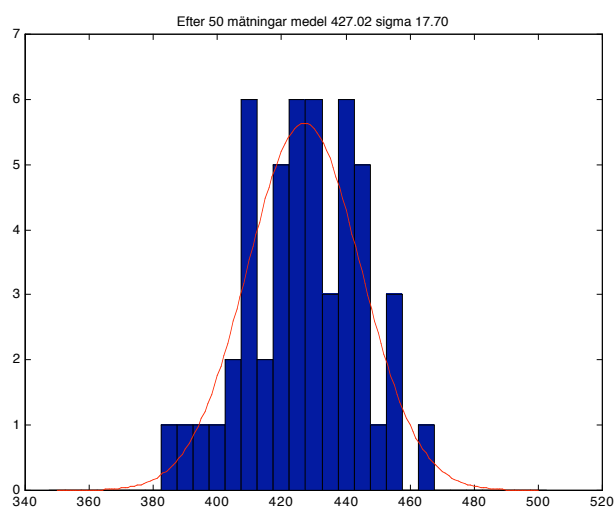
Föreläsning 2

Forskningsmetodik HT 2006



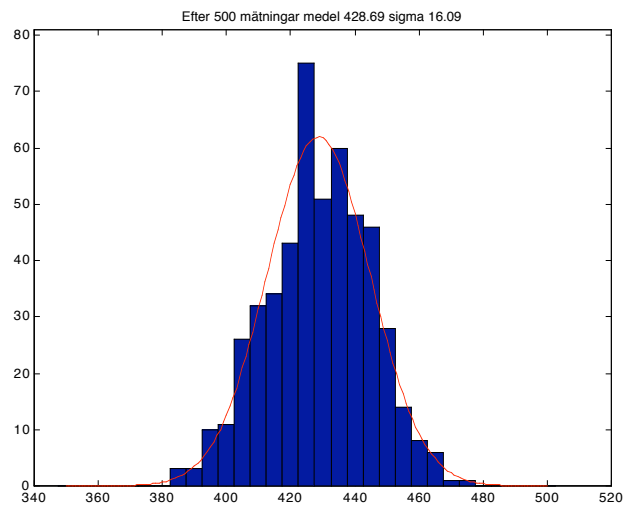
Föreläsning 2

Forskningsmetodik HT 2006



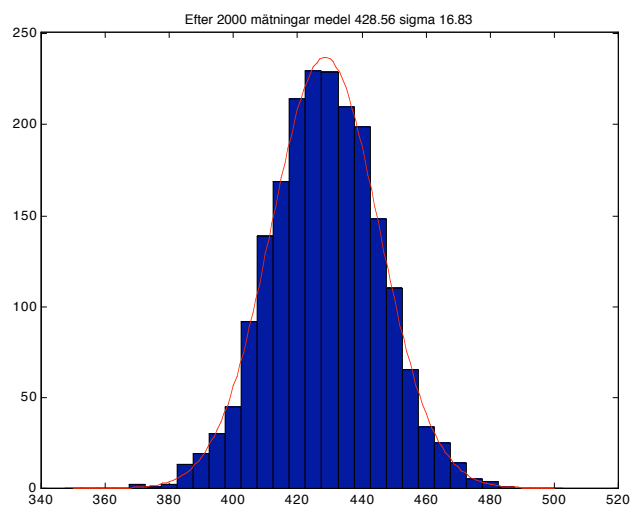
Föreläsning 2

Forskningsmetodik HT 2006



Föreläsning 2

Forskningsmetodik HT 2006



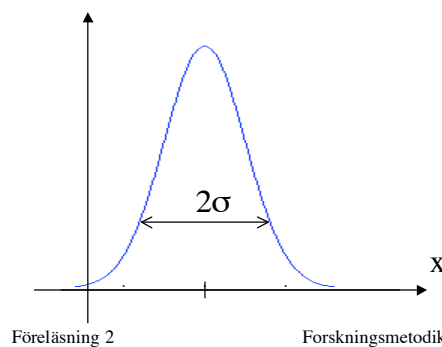
Föreläsning 2

Forskningsmetodik HT 2006

Centrala Gränsvärdesatssen

Om vi summerar ett stort antal slumpmässigt fördelade tal, så kommer den asymptotiska fördelningen för summan att gå mot en *normalfördelning*

Detta gäller oberoende av hur fördelningen ser ut för de termer som ingår i summan!!



Normalfördelningen:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Forskningsmetodik HT 2006

Centrala Gränsvärdesatssen

Medelvärden av fördelningen av medelvärden närmar sig det sanna medelvärdet, μ , för stora N

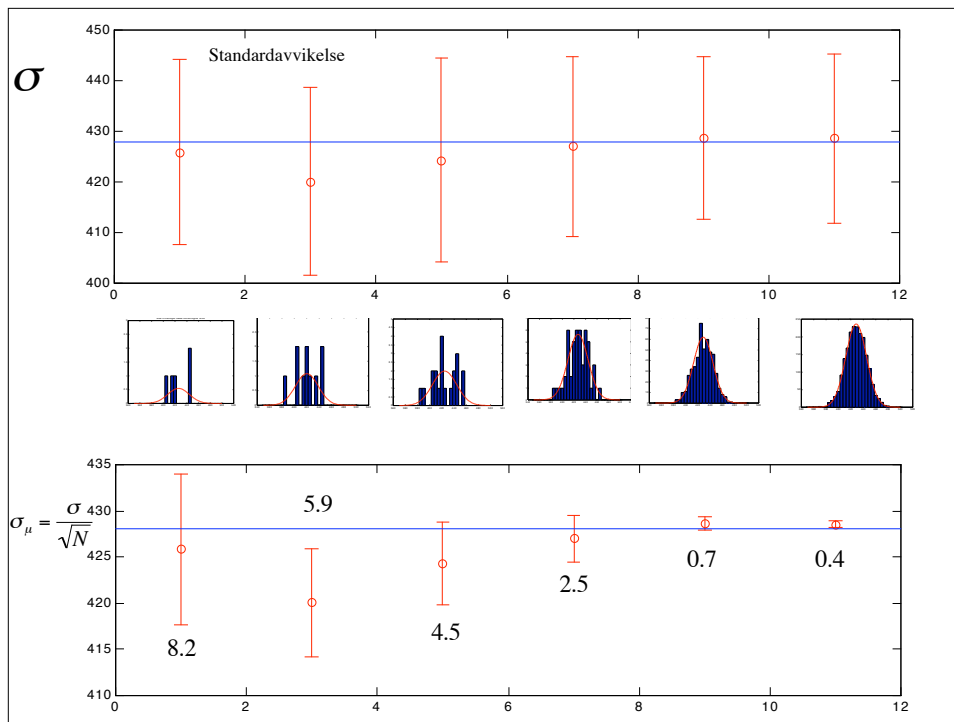
Standardavvikelsen för fördelningen av medelvärden går mot

$$\sigma_{\mu} = \frac{\sigma}{\sqrt{N}}$$

“De stora talens välsignelse”

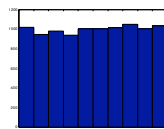
Föreläsning 2

Forskningsmetodik HT 2006



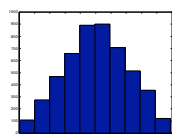
Exempel:

Summera tal
slumpvis fördelade
mellan 0 och 1

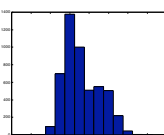


Ursprungliga fördelningen

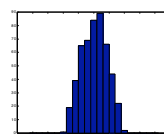
Se figur 5.27
sidan 212!



Summan av två tal



Summan av fem tal

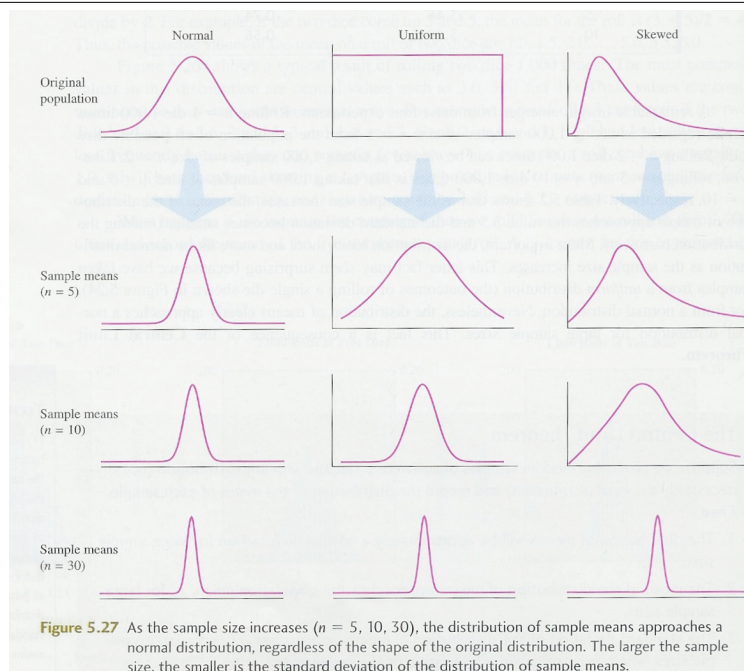


Summan av tjugo tal

När kan man förvänta sig en normalfördelning?

1. De flesta värdena skall ligga nära medelvärdet.
2. Positiva avvikelser är lika vanliga som negativa.
3. Stora avvikelser är mindre sannolika än små.
4. Den totala avvikelsen för ett mätvärde är summan av en mängd slumpmässiga bidrag.

(Strikt matematiskt kommer punkt 4, tack vare centrala gränsvärdessatsen leda till de övrig punkterna)



Föreläsning 2

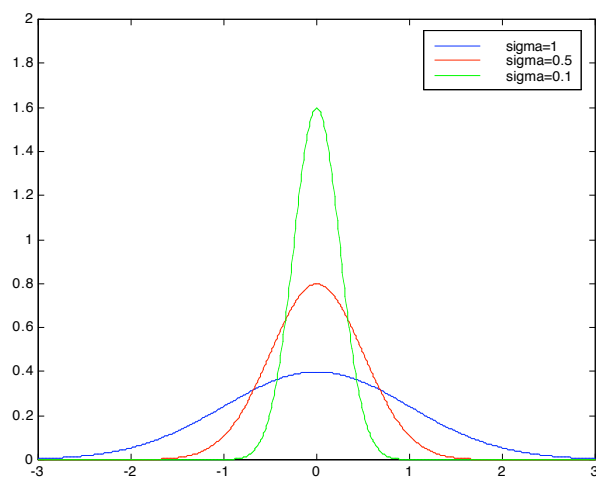
Forskningsmetodik HT 2006

Normalfördelningen

Maximum vid $x=\mu$

Symmetrisk runt $x= \mu$

Ytan under kurvan
"normaliserad" till 1



Föreläsning 2

Forskningsmetodik HT 2006

Normalfördelningen

För att förstå hur centralvärdessatsen sats kan hjälpa oss så kan vi betrakta resultatet av en mätning som beroende av det sanna värdet till vilket har adderats slumpvisa bidrag från ett stort antal okända (och en del kända) felkällor. Det totala felet i mätningen är då summan av alla dessa bidrag.

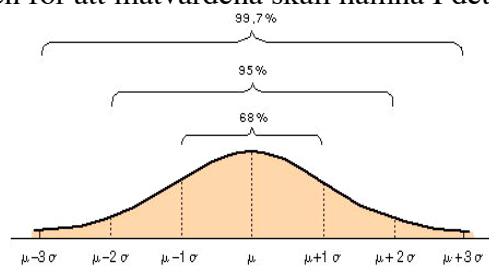
Enligt centrala gränsvärdessatsen är då denna summa (= det totala mätfelet) normalfördelat.

Resultatet av ett stort antal mätningar kommer då att spridas runt det sanna värdet, och fördelningen av mätningarna runt detta kommer att ha en form som ges av normalfördelningen.

Föreläsning 2

Forskningsmetodik HT 2006

Eftersom normalfördelningen är “normaliserad” - så kan man tolka ytan under kurvan mellan två x-värden som den relativa frekvensen för att mätvärdena skall hamna i detta intervall.



The “68-95-99,7 regel”

68% av datapunkterna faller inom 1 standardavvikelse från medelvärdet

95% av datapunkterna faller inom 2 standardavvikelse från medelvärdet

99,7% av datapunkterna faller inom 3 standardavvikelse från medelvärdet

Föreläsning 2

Forskningsmetodik HT 2006

Exempel 1 sid 199, ex 3 sid 201

Ett standardiserat mått på avvikelsen ges av (mätvärdet - medelvärde) / standardavvikelsen. I boken kallas detta mått "standard score" eller "z-score". Vanligare i naturvetenskap är kanske termen "pull"

$$z = \frac{\text{mätvärde} - \text{medelvärde}}{\text{standardavvikelsen}}$$

z är antalet standardavvikelser som mätvärdet avviker från medelvärde

Exempel 1 sidan 213

Föreläsning 2

Forskningsmetodik HT 2006

Table 5.1 Standard Scores and Percentiles for a Normal Distribution (cumulative values from the left)

Standard score	%	Standard score	%	Standard score	%	Standard score	%
-3.5	0.02	-1.0	15.87	0.0	50.00	1.1	86.43
-3.0	0.13	-0.95	17.11	0.05	51.99	1.2	88.49
-2.9	0.19	-0.90	18.41	0.10	53.98	1.3	90.32
-2.8	0.26	-0.85	19.77	0.15	55.96	1.4	91.92
-2.7	0.35	-0.80	21.19	0.20	57.93	1.5	93.32
-2.6	0.47	-0.75	22.66	0.25	59.87	1.6	94.52
-2.5	0.62	-0.70	24.20	0.30	61.79	1.7	95.54
-2.4	0.82	-0.65	25.78	0.35	63.68	1.8	96.41
-2.3	1.07	-0.60	27.43	0.40	65.54	1.9	97.13
-2.2	1.39	-0.55	29.12	0.45	67.36	2.0	97.72
-2.1	1.79	-0.50	30.85	0.50	69.15	2.1	98.21
-2.0	2.28	-0.45	32.64	0.55	70.88	2.2	98.61
-1.9	2.87	-0.40	34.46	0.60	72.57	2.3	98.93
-1.8	3.59	-0.35	36.32	0.65	74.22	2.4	99.18
-1.7	4.46	-0.30	38.21	0.70	75.80	2.5	99.38
-1.6	5.48	-0.25	40.13	0.75	77.34	2.6	99.53
-1.5	6.68	-0.20	42.07	0.80	78.81	2.7	99.65
-1.4	8.08	-0.15	44.04	0.85	80.23	2.8	99.74
-1.3	9.68	-0.10	46.02	0.90	81.59	2.9	99.81
-1.2	11.51	-0.05	48.01	0.95	82.89	3.0	99.87
-1.1	13.57	0.0	50.00	1.0	84.13	3.5	99.98

NOTE: The % column gives the percentage of values in the distribution less than the corresponding standard score.