

# Formel- och tabellsamling för Statistik och dataanalys I, 15 hp

Kurskod: ST1101

Begrepp i **orange** typsnitt används inom del 1 av kursen.

Begrepp i **blått** typsnitt används inom del 1 och 2 av kursen.

## Deskriptiv statistik - en variabel

### Stickprovsmedelvärde

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

### Stickprovsvarians

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

### Stickprovsstandardavvikelse

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

### Standardisering - stickprov

$$z_x = \frac{x - \bar{x}}{s_x}$$

## Deskriptiv statistik - två variabler

### Stickprovskovarians

$$s_{xy} = \text{Cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

### Stickprovskorrelation

$$r_{xy} = \text{Corr}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{\text{alla data}} z_x z_y}{n - 1}$$

## Enkel linjär regression - estimation

### Skattad regressionsmodell

$$\hat{y} = b_0 + b_1 x$$

### Skattning av lutningen

$$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}$$

### Skattning av interceptet

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

### Residualvarians

$$s_e^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

### Prediktion för $x = x_i$

$$\hat{y}_i = b_0 + b_1 x_i$$

## Multipel linjär regression ( $k = 1$ för enkel)

### Skattad regressionsmodell

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

Skattningarna  $b_0, b_1, \dots, b_k$  ges av datorutskrift.

### Residualvarians

$$s_e^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k - 1}$$

### Prediktion för $x = x_i$

$$\hat{y}_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \dots + b_k x_{k,i}$$

### ANOVA-uppdelning

$$\text{SST} = \text{SSR} + \text{SSE}$$

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

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

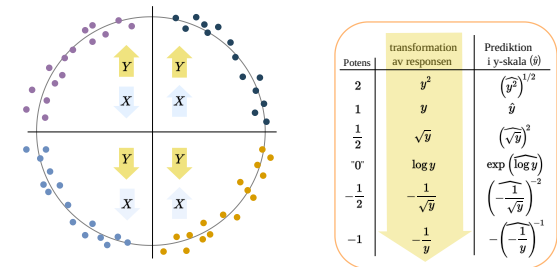
### Anpassningsmått

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$R_{\text{adj}}^2 = 1 - \frac{\text{SSE} / (n - k - 1)}{\text{SST} / (n - 1)}$$

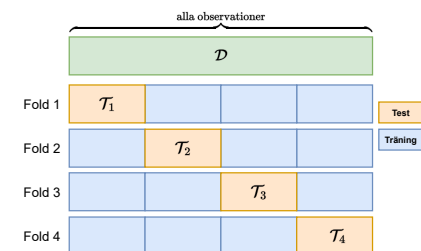
För enkel linjär regression gäller att  $R^2 = r_{xy}^2$

### Transformationer - Tukeys cirkel



### Korsvalidering

Datamaterialets observationer  $\mathcal{D} = \{1, 2, \dots, n\}$  delas upp i  $K$  delar, där varje observation är med i exakt en del.



Skattning av modellens prognosförmåga på nya data:

$$\text{SSE}_{\text{cv}} = \sum_{i \in \mathcal{I}_1} (y_i - \hat{y}_i^{(1)})^2 + \dots + \sum_{i \in \mathcal{I}_K} (y_i - \hat{y}_i^{(K)})^2$$

$$\text{RMSE}_{\text{cv}} = \sqrt{\frac{\text{SSE}_{\text{cv}}}{n}}$$

- $\mathcal{I}_k \subset \mathcal{D}$  är alla observationer som är *testdata* i fold  $k$
- $\sum_{i \in \mathcal{I}_k}$  är summan över alla testdata i fold  $k$
- $\hat{y}_i^{(k)}$  är prediktionen av  $y_i$  i fold  $k$  från en modell skattad på alla data *förutom* testdata i  $\mathcal{I}_k$ .

# Sannolikhetslära

## Additionssatsen

$$P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$$

## Multiplikationssatsen

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B})$$

## Satsen om total sannolikhet

$$P(\mathbf{A}) = \sum_{k=1}^K P(\mathbf{A}|\mathbf{B}_k)P(\mathbf{B}_k)$$

där  $\mathbf{B}_1, \dots, \mathbf{B}_K$  är en partitionering av utfallsrummet.

## Bayes sats

$$P(\mathbf{B}_k|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{B}_k)P(\mathbf{B}_k)}{P(\mathbf{A})}$$

# Kombinatorik

**Fakultet** (eng. **factorial**) av positiva heltalet  $n$

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$$

och  $0! = 1$  per definition.

## Kombinationer och permutationer

Hur många sätt att välja $k$ element bland $n$ element?		
	med återläggning	utan återläggning
med ordning	$n^k$	${}_nP_k = \frac{n!}{(n-k)!}$
utan ordning	ej på kurs	${}_nC_k = \frac{n!}{(n-k)!k!}$

# Slumpvariablers egenskaper - en variabel

$X$  är en diskret variabel med sannolikhetsfördelning  $P(x)$ .

**Väntevärde** (eng. **expected value**)

$$\mu = E(X) = \sum_{\text{alla } x} x \cdot P(x)$$

**Varians** (eng. **variance**)

$$\sigma^2 = \text{Var}(X) = \sum_x (x - \mu)^2 \cdot P(x) = E(X^2) - \mu^2$$

**Standardavvikelse** (eng. **standard deviation**)

$$\sigma = \text{SD}(X) = \sqrt{\text{Var}(X)}$$

**Väntevärde linjärkombination** ( $c$  och  $d$  är konstanter)

$$E(c + d \cdot X) = c + d \cdot E(X)$$

**Varians linjär kombination**

$$\text{Var}(c + d \cdot X) = d^2 \cdot \text{Var}(X)$$

# Slumpvariablers egenskaper - två variabler

**Kovarians** mellan två slumpvariabler  $X$  och  $Y$

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

**Korrelation** mellan två slumpvariabler  $X$  och  $Y$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X) \cdot \text{SD}(Y)}$$

**Väntevärde för en summa av två slumpvariabler**

$$E(X + Y) = E(X) + E(Y)$$

**Väntevärde linjärkombination av två slumpvariabler**

$$E(cX + dY) = cE(X) + dE(Y)$$

**Variansen för en summa av två slumpvariabler**

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

# Medelvärde av slumpvariabler

Låt  $X_1, X_2, \dots, X_n$  vara **oberoende likafördelade** slumpvariabler med väntevärde  $\mu = E(X_i)$  och varians  $\sigma^2 = \text{Var}(X_i)$ . För stickprovmedelvärdet  $\bar{X} = \sum_{i=1}^n X_i / n$  gäller då:

**Väntevärde för medelvärdet**

$$E(\bar{X}) = \mu$$

**Varians för medelvärdet**

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

**Stora talens lag:** För alla  $\epsilon > 0$

$$P(|\bar{X} - \mu| > \epsilon) \rightarrow 0 \text{ när } n \rightarrow \infty$$

**Centrala gränsvärdesatsen (informellt):**

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ för stort } n$$

Tumregel: approximationen är tillräckligt bra om  $n \geq 30$ .

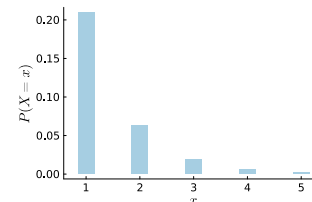
# Diskreta fördelningar

**Geometrisk fördelning:**  $X \sim \text{Geo}(p)$

$$P(X = x) = q^{x-1}p \text{ för } x = 1, 2, \dots$$

$$E(X) = \frac{1}{p}$$

$$\text{Var}(X) = \frac{q}{p^2}$$

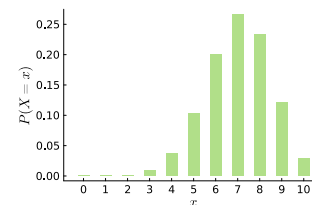


**Binomialfördelning:**  $X \sim \text{Binom}(n, p)$

$$P(X = x) = {}_nC_x \cdot p^x q^{n-x} \text{ för } x = 0, 1, 2, \dots, n$$

$$E(X) = np$$

$$\text{Var}(X) = npq$$

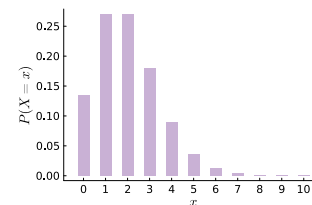


**Poissonfördelning:**  $X \sim \text{Pois}(\lambda)$

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ för } x = 0, 1, 2, \dots$$

$$E(X) = \lambda$$

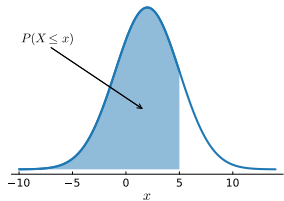
$$\text{Var}(X) = \lambda$$



## Normalfördelning och standardisering

**Normalfördelning:**  $X \sim N(\mu, \sigma)$

$$E(X) = \mu$$
$$\text{Var}(X) = \sigma^2$$

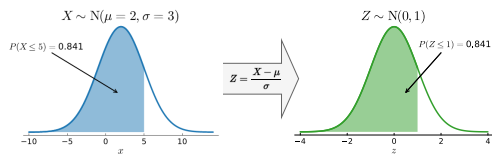


**Linjärkombination:** Om  $X \sim N(\mu, \sigma)$  och  $Y = c + d \cdot X$

$$Y \sim N(c + d \cdot \mu, |d| \cdot \sigma)$$

**Standardisering:** Om  $X \sim N(\mu, \sigma)$  så

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$



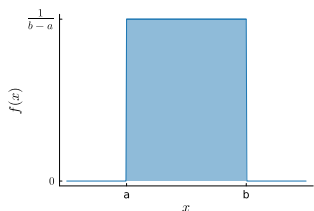
## Andra kontinuerliga fördelningar

**Likformig fördelning:**  $X \sim U(a, b)$

$$f(x) = \frac{1}{b-a} \text{ för } -a < x < b$$

$$E(X) = (a+b)/2$$

$$\text{Var}(X) = (b-a)^2/12$$



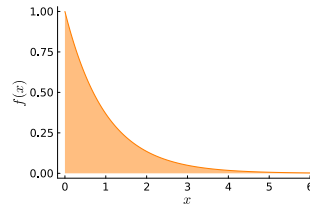
**Exponentialfördelning:**  $X \sim \text{Expon}(\lambda)$

$$f(x) = \lambda e^{-\lambda x} \text{ för } x \geq 0$$

$$P(X \leq x) = 1 - e^{-\lambda x}$$

$$E(X) = 1/\lambda$$

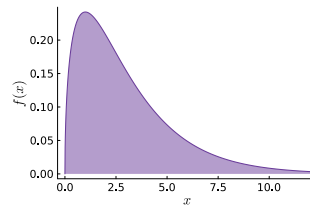
$$\text{Var}(X) = 1/\lambda^2$$



**$\chi^2$ -fördelning:**  $X \sim \text{Chi2}(\nu)$

$$E(X) = \nu$$

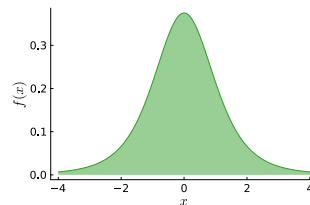
$$\text{Var}(X) = 2\nu$$



**Student t-fördelning:**  $X \sim t(\nu)$

$$E(X) = 0 \text{ om } \nu > 1$$

$$\text{Var}(X) = \frac{\nu}{\nu-2} \text{ om } \nu > 2$$



## Inferens för en population

**Konfidensintervall väntevärde** känd varians

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

**Konfidensintervall väntevärde** okänd varians

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s_x}{\sqrt{n}}$$

**Konfidensintervall andel**

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

**Teststatistika**  $H_0 : \mu = \mu_0$  känd varians

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

**Teststatistika**  $H_0 : \mu = \mu_0$  okänd varians

$$T = \frac{\bar{X} - \mu_0}{s_x / \sqrt{n}}$$

**Teststatistika proportion**  $H_0 : p = p_0$

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}}}$$

## Jämföra två oberoende populationer

**Teststatistika**  $H_0 : \mu_1 - \mu_2 = d$  kända varianser

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

**Teststatistika**  $H_0 : \mu_1 - \mu_2 = d$  okända varianser

$$T = \frac{\bar{X}_1 - \bar{X}_2 - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

## Jämföra två populationer - parade data

Data som differenser:  $D_1, D_2, \dots, D_n$ , där  $D_i = X_{1,i} - X_{2,i}$ .  
 $\bar{D} = \sum_{i=1}^n D_i / n$  och  $s_D^2$  är stickprovsvariansen för differenserna.

**Teststatistika parade data**  $H_0 : \mu_1 - \mu_2 = d$  okänd varians

$$T = \frac{\bar{D} - d_0}{s_D / \sqrt{n}}$$

## Enkel linjär regression - inferens

### Populationsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{och} \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon)$$

### Standardfel för $b_1$

$$s_{b_1} = \frac{s_e}{\sqrt{(n-1)s_x}}$$

### Konfidensintervall för $b_1$

$$b_1 \pm t_{\alpha/2, n-2} \cdot s_{b_1}$$

### Teststatistika för $H_0 : \beta_1 = \beta_1^{(0)}$

$$T = \frac{b_1 - \beta_1^{(0)}}{s_{b_1}}$$

### Prediktionsintervall vid $x_*$

$$\hat{y}_* \pm t_{\alpha/2, n-2} \cdot \sqrt{\frac{s_e^2}{n} + s_{b_1}^2 (x_* - \bar{x})^2 + s_e^2}$$

---

## Multipel linjär regression - inferens

### Populationsmodell

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i} + \varepsilon_i \quad \text{och} \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon)$$

### Standardfel för $b_j$

Komplicerad formel, ges av datorutskift.

### Konfidensintervall för $b_j$

$$b_j \pm t_{\alpha/2, n-k-1} \cdot s_{b_j}$$

### Teststatistika för $H_0 : \beta_j = \beta_j^{(0)}$

$$T = \frac{b_j - \beta_j^{(0)}}{s_{b_j}}$$

---

## Chi2-test

### Teststatistika för $\chi^2$ -test

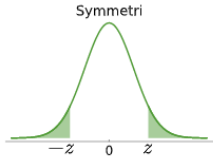
$$\chi^2 = \sum_{\text{alla celler}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

### Frihetsgrader för goodness-of-fit:

$$\nu = N - 1 \quad \text{där } N \text{ är antalet kategorier/bins}$$

# Normalfördelning

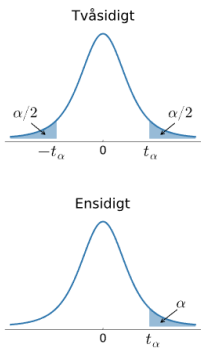
Tabellen ger sannolikheten  $\Phi(z) = P(Z \leq z)$  för olika  $z$  där  $Z$  är standardnormal,  $Z \sim N(0, 1)$ . Sannolikheter i den vänstra svansen fås genom symmetri:  $P(Z \leq -z) = 1 - P(Z \leq z)$ .



Andra decimalen i  $z$

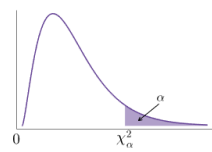
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

# Student-*t* fördelning



Konfidensnivå:	80%	90%	95%	98%	99%
Tvåsidig sannolikhet:	0.200	0.100	0.050	0.020	0.010
Ensidig sannolikhet:	0.100	0.050	0.025	0.010	0.005
df					
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
32	1.309	1.694	2.037	2.449	2.738
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
75	1.293	1.665	1.992	2.377	2.643
100	1.290	1.660	1.984	2.364	2.626
120	1.289	1.658	1.980	2.358	2.617
140	1.288	1.656	1.977	2.353	2.611
180	1.286	1.653	1.973	2.347	2.603
250	1.285	1.651	1.969	2.341	2.596
400	1.284	1.649	1.966	2.336	2.588
1000	1.282	1.646	1.962	2.330	2.581
oändligt	1.282	1.645	1.960	2.326	2.576

$\chi^2$ -fördelning



Sannolikhet i höger svans:	0.100	0.050	0.025	0.010	0.005
df					
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.195	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672
40	51.805	55.758	59.342	63.691	66.766
50	63.167	67.505	71.420	76.154	79.490
60	74.397	79.082	83.298	88.379	91.952
70	85.527	90.531	95.023	100.425	104.215
80	96.578	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169