Statistische Methoden der Datenanalyse

Wintersemester 2012/2013

Albert-Ludwigs-Universität Freiburg

Prof. Markus Schumacher, Dr. Stan Lai

Physikalisches Institut Westbau 2 OG

E-Mail: Markus.Schumacher@physik.uni-freiburg.de

stan.lai@cern.ch

http://terascale.physik.uni-freiburg.de/lehre/ws_1213/statmethoden_ws1213

Pearsons χ^2-Test

Gegeben: N unabhängige Messwerte mit Varianzen σ_i^2 $\vec{n} = (n_1, \ldots, n_N)$

und Theorievorhersagen $\vec{\nu} = (\nu_1, \ldots, \nu_N)$:

(entweder phys. Gesetz oder WDF mit Integration über Binbreiten)

Pearsons χ^2-Teststatistik: $$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2} \text{ , where } \sigma_i^2 = V[n_i] \text{ .}$$

χ^2 = Summe der Quadrate der Abweichungen *i*-ten Messung von der *i*-ten Vorhersage, normiert auf die Varianz (Größe die in der Methode der kleinsten Quadrate minimiert wird)

Für n_i ~ aus Poissonverteilung gilt $V[n_i] = \nu_i$, also erhält man:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} .$$

Pearsons χ^2-Test

Wenn die Hypothese (Stichprobe stammt aus Theorievorhersage) wahr ist und die n_i aus Gaussverteilung mit Mittelwert ν_i und Standardabweichung σ_i, i.e., $n_i \sim \mathrm{N}(\nu_i, \sigma_i^2)$, dann folgt Pearsons χ^2-Teststatistik einer χ^2-WDF mit N Freiheitsgraden (hier mit $\chi^2 = z$):

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

Wenn the n_i aus Poissonverteilung mit $\nu_i >> 1$ (in der Praxis OK für $\nu_i > 5$) dann geht Poissonverteilung in Gaussverteilung über und daher folgt Pearsons χ^2-Teststatistik hier ebenfalls einer χ^2-WDF.

Der χ^2-Wert aus den Daten der Stichprobe ergibt dann den P-Wert gemäß:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z; N)\, dz \ .$$

The 'χ^2 pro Freiheitsgrad'

Erinnerung: für Chi-Quadrat-WDF mit *N* Freiheitsgraden für ZV z gilt

$$E[z] = N\,, \quad V[z] = 2N\,.$$

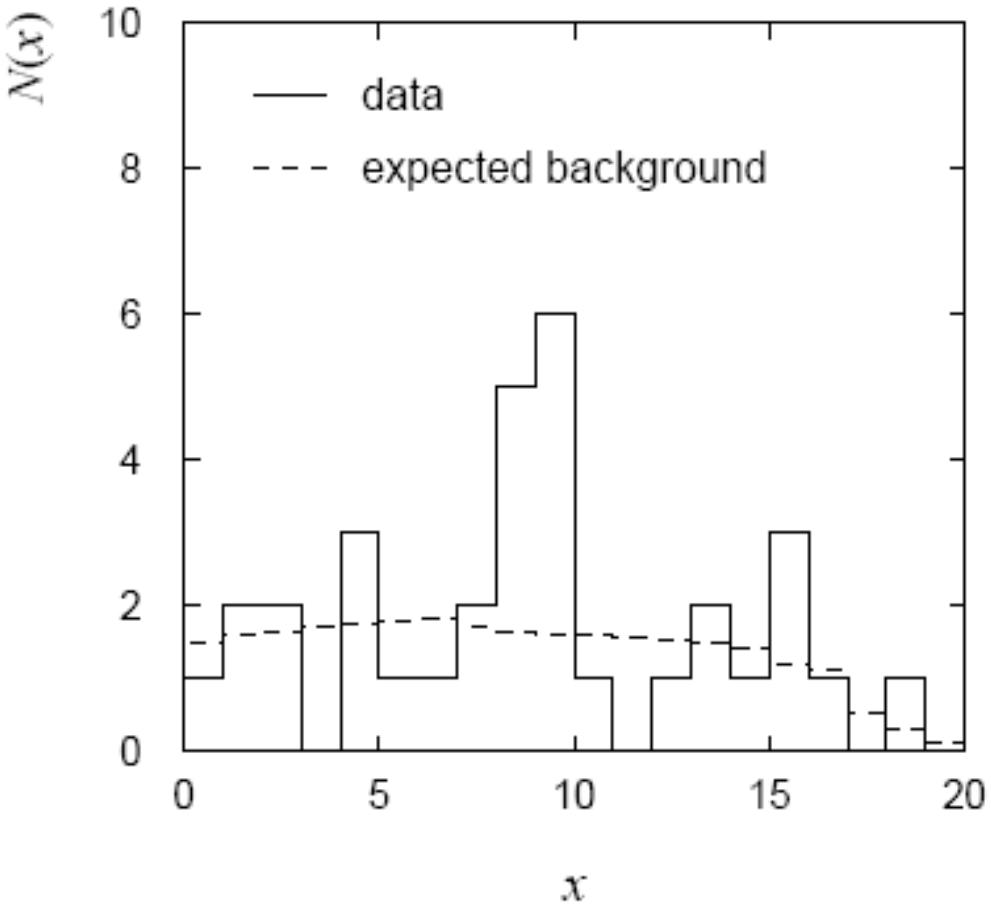
Das macht Sinn: wenn hypothetische ν_i korrekt sind, dann ist die mittlere Abweichung der n_i von ν_i gerade σ_i, also trägt jeder Term in Summe ~ 1 bei.

Oft wird das "Chi-Quadrat pro Freiheitsgrad" χ^2/N als Mass für die Güte der Übereinstimmung angegeben. Besser und vollständiger ist es beide χ^2 und N separat anzugeben und den korrekten P-Wert auszurechnen. Betrachte zum Beispiel:

$$\chi^2 = 15,\ N = 10 \rightarrow p - \text{value} = 0.13\,,$$
$$\chi^2 = 150,\ N = 100 \rightarrow p - \text{value} = 9.0 \times 10^{-4}\,.$$

i.e. für große *N*, kann schon ein "χ^2 pro FG" geringfügig größer als 1 bereits einem sehr kleinen P-Wert entsprechen,
d.h. schlechte Übereinstimmung mit H_0

Beispiel eines χ^2-Test ohne Gaussfehler



← Dies ergibt

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

für 20 Freiheitsgrade = Bins

Nun müssen wir den P-Wert finden, aber viele Bins haben wenige oder keine Einträge. Also gilt nicht die Gaussnäherung und χ^2 folgt nicht der Chi-Quadrat-WDF.

Verwendung von MC zur Bestimmung der WDF für χ^2

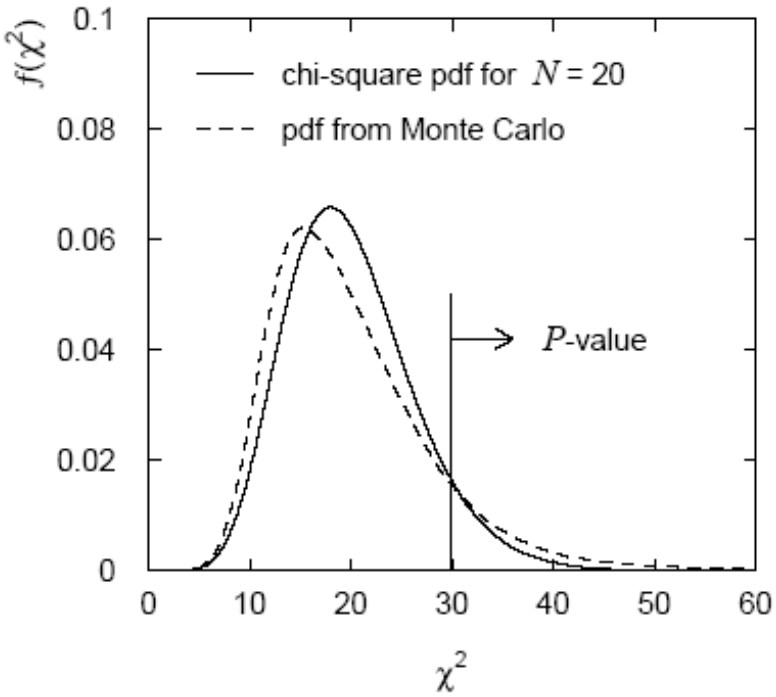
Pearsons χ^2 Teststatistik quantifiziert immer den Grad der Übereinstimmung zwischen Daten und Theorievorhersage, es ist immer ein gültige Teststatistik, aber die WDF ist für nicht gaussche Fehler nicht direkt bekannt.

Um die WDF zu bestimmen, simuliere Messungen/Stichproben mit MC-Programmen: $n_i \sim \text{Poisson}(\nu_i)\,,\ i = 1, N\,.$

Hier: 10^6 Stichproben simuliert
Bruchteil der Messungen mit $\chi^2 > 29.8$ ergibt den *P*-Wert:

$$P = 0.11$$

Wenn wir die Chi-Quadrat-WDF benutzt hätten, würden wir $P = 0.073$ erhalten.



Güte der Anpassung bei KQ-Schätzern

Der Wert von χ^2 im Minimum ist eine Mass für die Übereinstimmung der Messdaten mit der angepassten Kurve/Theorievorhersage:

$$\chi^2_{\text{min}} = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \hat{\theta}))^2}{\sigma_i^2}$$

Dieser Wert kann als Teststatistik für die Güte der Anpassung für die Theorievorhersage $\lambda(x; \theta)$ verwendet werden.

Falls die Hypothese korrekt ist, dann folgt die Teststatistik $t = \chi^2_{\text{min}}$ einer Chi-Quadrat-WDF:

$$f(t; n_{\text{d}}) = \frac{1}{2^{n_{\text{d}}/2}\Gamma(n_{\text{d}}/2)} t^{n_{\text{d}}/2-1} e^{-t/2}$$

die Anzahl der Freiheitsgrade #FG ergibt sich zu

N Messpunkte, L Parameter, M Zwangsbedingungen

keine Bins: #FG = N-L+M

Histogramm N Bins (Normierung fixiert): #FG = N-1-L+M

Verteilung der P-Werte

Der *p*-Wert ist eine Funktion der Stichprobe und daher eine Zufallsvariable, die einer gewissen Wahrscheinlichkeitsdichtefunktion folgt.

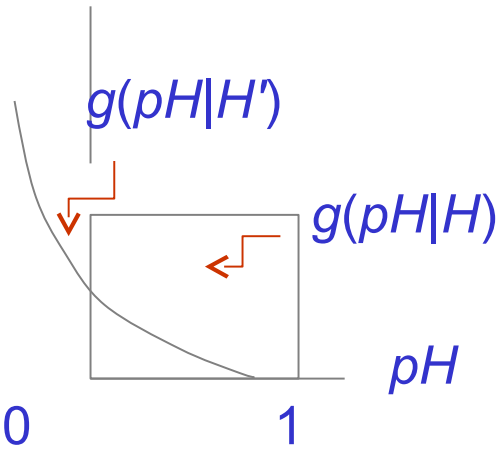
Annahme: der *P*-Wert für *H* wird aus Teststatistik $t(\boldsymbol{x})$ bestimmt gemäß

$$p_H = \int_t^\infty f(t'|H)dt'$$

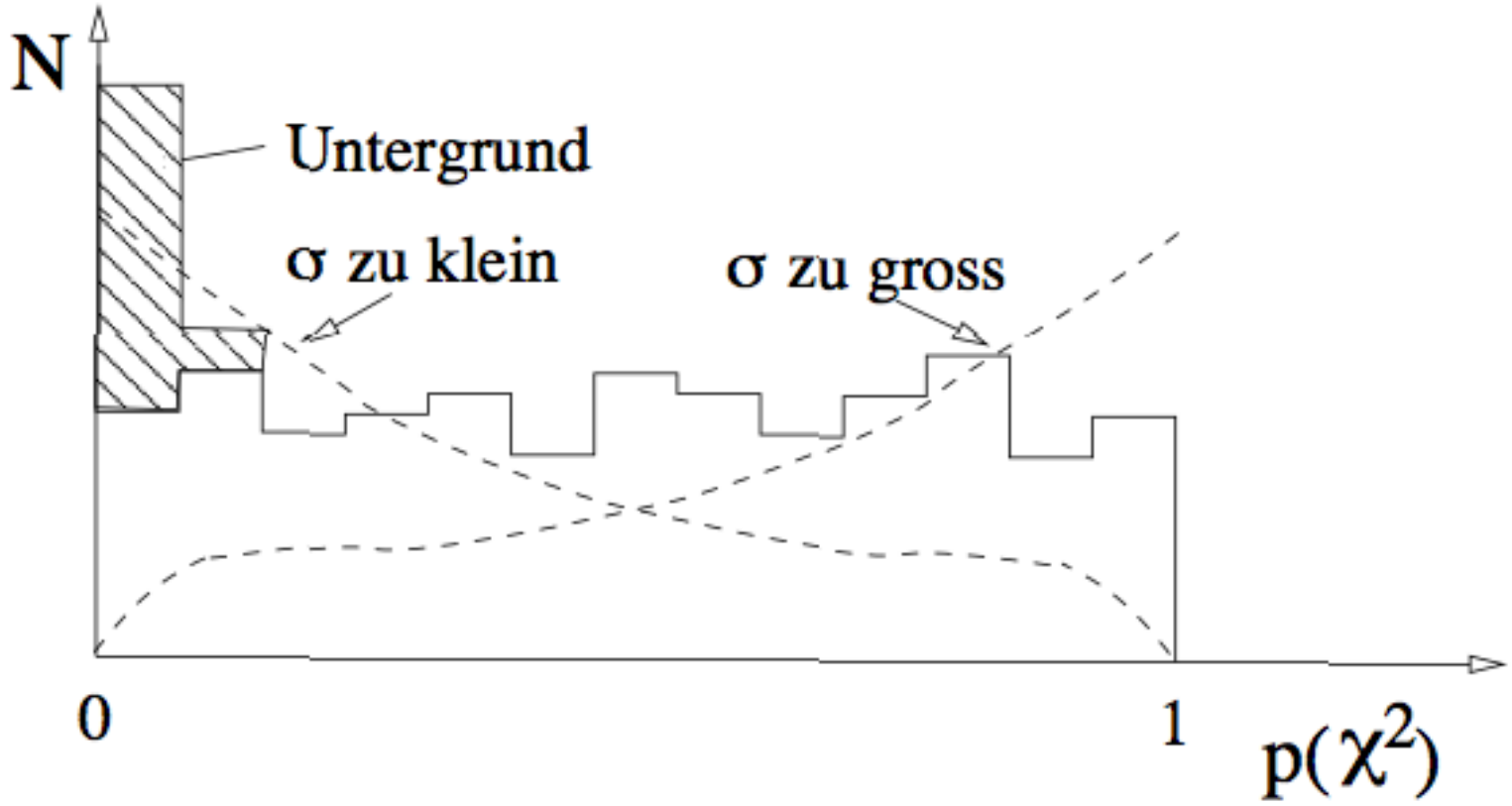
Die WDF für p_H unter der Annahme der Hypothese *H* ist

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H/\partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \le p_H \le 1)$$

i. a. gilt für kontinuierliche Daten:
unter Annahme von *H*, p_H ~ flach in [0,1]
und ist konzentriert bei Nulland für
viele Arten von Alternativhypothesen



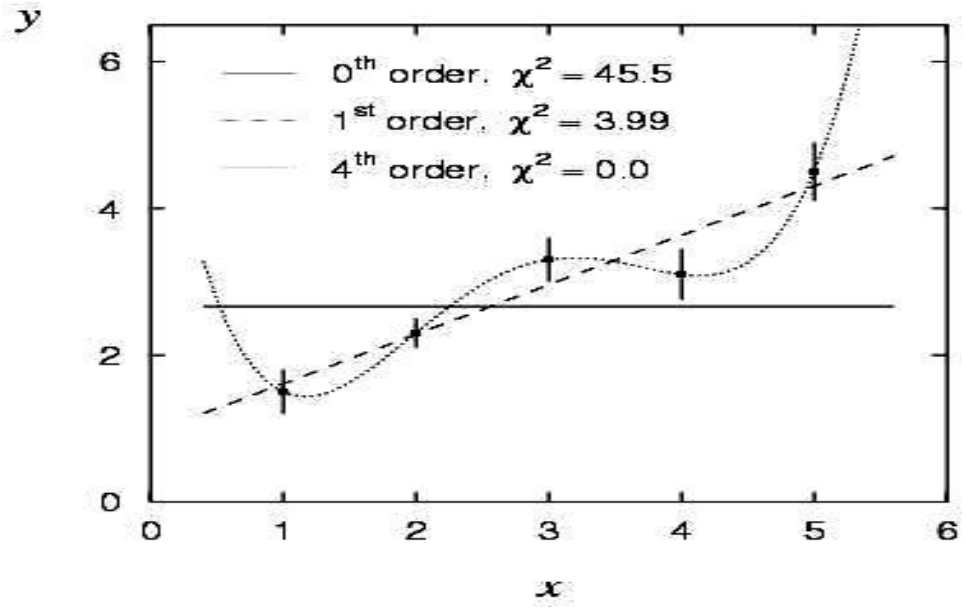
Verteilung der P-Werte bzw. Chi-Quadrat-Wkten



Hypothese korrekt, Fehlerabschätzung korrekt → flache Verteilung
Hypothese falsch → kleine P-Werte
Hypothese korrekt, Fehler zu groß → Verschiebung zu großen P-Werten
Hypothese korrekt, Fehler zu kleine → Verschiebung zu kleinen P-Werten

Beispiel der Güte der Anpassung für KQ

Anpassung eines Polynoms der Ordnung p (p+1 Parameter) an 5 Messpunkte



$$\lambda(x;\theta_0,\ldots,\theta_p) = \sum_{n=0}^{p} \theta_n x^n$$

$$p = \int_{\chi^2_{\text{min}}}^{\infty} f(t;n_{\text{d}})\,dt$$

Anpassung einer Geraden (P=2):

$$\chi^2_{\text{min}} = 3.99\,, \qquad n_{\text{d}} = 5-2 = 3\,, \qquad p = 0.263$$

Anpassung einer Konstanten (P=1):

$$\chi^2_{\text{min}} = 45.5\,, \qquad n_{\text{d}} = 5-1 = 4\,, \qquad p = 3.1\times10^{-9}$$

→ Gerade mit Steigung deutlich gegenüber Konstanten bevorzugt

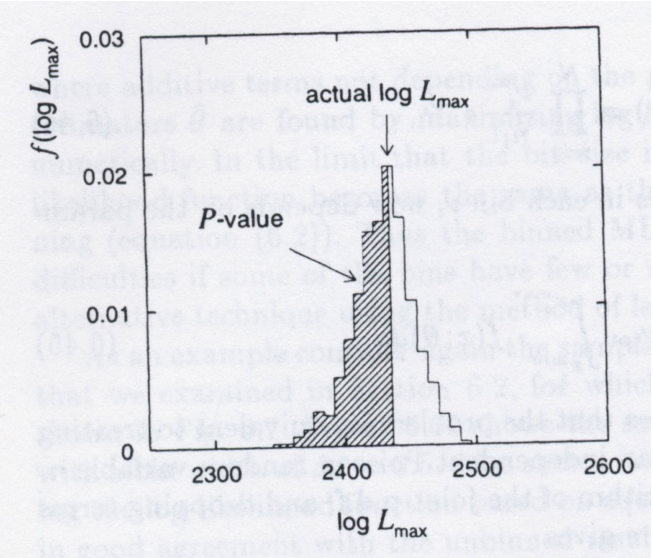
Güte der Anpassung bei ML-Schätzung

L_{max} nicht unmittelbar nutzbar, keine Aussage über WDF

a) L_{max} als Teststatistik

bestimme WDF über MC-Simulation von identischen Messungen

→ P-Wert hier:0.63



b) Histogrammierung der Messwerte

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Vergleich mit Theorievorhersage

$$\hat{\nu}_i = n_{\text{tot}} \int_{x_i^{\min}}^{x_i^{\max}} f(x;\hat{\theta})dx$$

Güte der Anpassung bei ML-Schätzung

Teststatistik gebildet aus Likelihoodfuntion des gebinnten ML-Fits

für ML-Schätzung n_{tot} fest

→ Gemeinsame WDF/Likelihood = Multinominalverteilung

für EML-Schätzung n_{tot} Poisson-ZV mit Mittelwert ν_{tot} ZV

→ Gemeinsame WDF/Likelihood = Produkt aus N Poisson-WDF

Betrachte Testatistik:

$$\lambda = \frac{L(\mathbf{n}|\boldsymbol{\nu})}{L(\mathbf{n}|\mathbf{n})} = \frac{f_{\text{joint}}(\mathbf{n};\boldsymbol{\nu})}{f_{\text{joint}}(\mathbf{n};\mathbf{n})}$$

im Nenner werden geschätze Bineinträge durch Datenwerte ersetzt

λ kann auch maximiert werden, um beste Schätzwerte zu erhalten

Güte der Anpassung bei ML-Schätzung

a) multinominal verteilte Daten (ML-Schätzung) wird Teststatistik zu

$$\lambda_M = \prod_{i=1}^{N} \left(\frac{\nu_i}{n_i} \right)^{n_i}$$

b) poissonverteilte Daten (EML-Schätzung) wird Teststatistik zu

$$\lambda_P = e^{n_{tot} - \nu_{tot}} \prod_{i=1}^{N} \left(\frac{\nu_i}{n_i} \right)^{n_i}$$

Annahme: *m* Parameter werden aus Daten geschätzt

betrachte Grenzfall großer Stichproben

Güte der Anpassung bei ML-Schätzung

Falls Hypothese über Form der theoretischen WDF korrekt ist gilt für WDF der Teststatistik λ

a) multinominal verteilte Daten (ML-Schätzung)

$$\chi^2_M = -2 \log \lambda_M = 2 \sum_{i=1}^{N} n_i \log \frac{n_i}{\hat{\nu}_i}$$

χ^2_M folgt Chi-Quadrat-WDF mit N-m-1 Freiheitsgraden

b) poissonverteilte Daten (EML-Schätzung)

$$\chi^2_P = -2 \log \lambda_P = 2 \sum_{i=1}^{N} \left(n_i \log \frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i \right)$$

χ^2_P folgt Chi-Quadrat-WDF mit N-m Freiheitsgraden

Güte der Anpassung bei ML-Schätzung

oder Anwendung des Pearsonschen χ^2-Test

a) multinominal verteilte Daten (ML-Schätzung)

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \hat{p}_i n_{\text{tot}})^2}{\hat{p}_i n_{\text{tot}}} \qquad \hat{p}_i = \hat{\nu}_i / \hat{\nu}_{\text{tot}}$$

χ^2 folgt Chi-Quadrat-WDF mit N-m-1 Freiheitsgraden für gr. Stichprobe

b) poissonverteilte Daten (EML-Schätzung)

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \hat{\nu}_i)^2}{\hat{\nu}_i}$$

χ^2 folgt Chi-Quadrat-WDF mit N-m Freiheitsgraden für gr. Stichprobe

Kolmogorov-Smirnov-Test

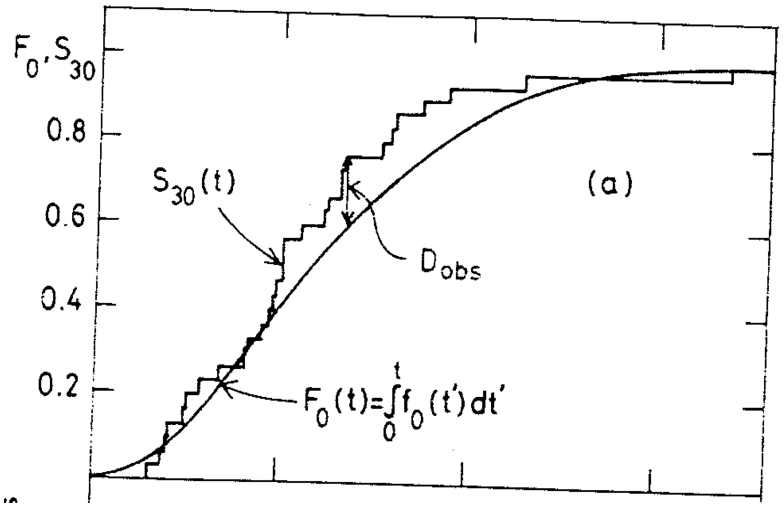
unabhängig von WDF der Grundgesamtheit
kein Binning der Daten notwendig
keine geschätzen Parameter in Theorievorhersage

gegeben: n Messwerte x_i (der Größe nach geordnet) und
Theorievorhersage f(x)

Betrachte Kummulativfunktionen für Theorievorhersage und Daten

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

$$S_n(x) = \begin{cases} 0 & x < x_1 \\ \frac{i}{n} & x_i \le x < x_{i+1} \\ 1 & x \ge x_n . \end{cases}$$



Nullhypothese: $H_o: \quad S_n(x) = F_o(x)$

Teststatistik: $D_n = \max |S_n(x) - F_o(x)|$

maximale Abweichung zwischen beiden Kummulativverteilungen

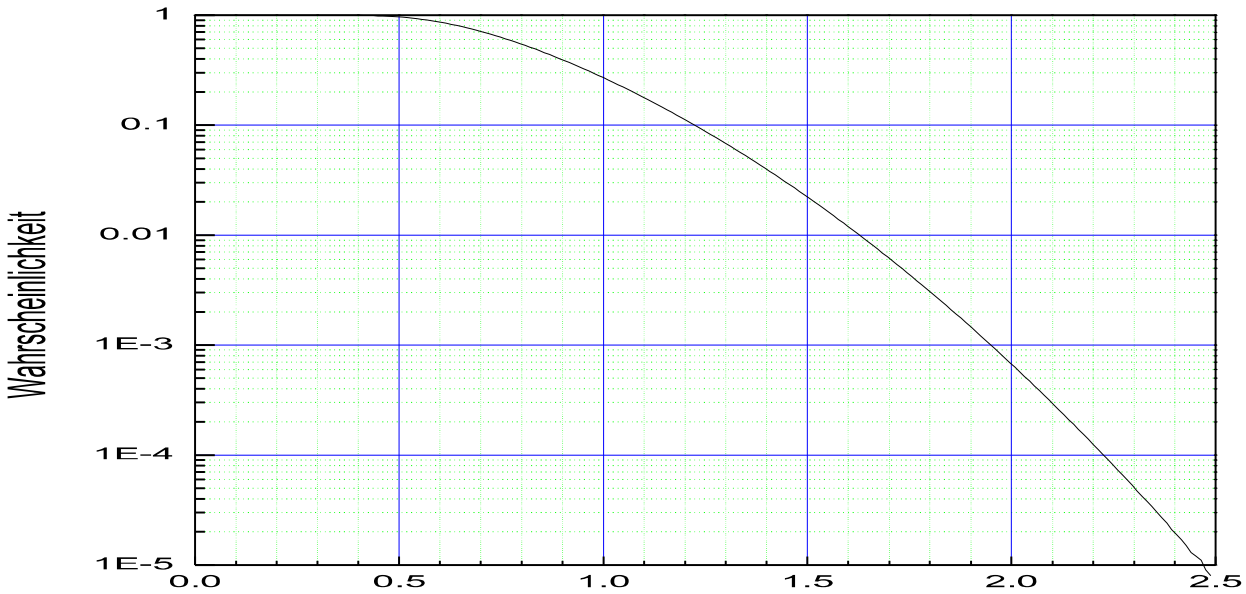
Kolmogorov-Smirnov-Test

Teststatistik D_n folgt Kolmogorov-Verteilung

Die WDF für D_n ist unabhängig von der Theorievorhersage f(x)

Für große Stichproben n gilt, dass die Kummulativ-WDF gegeben ist durch

$$\lim_{n\to\infty} P\left(D_n \leq \frac{z}{\sqrt{n}}\right) = 1 - 2\sum_{r=1}^{\infty} (-1)^{r-1} e^{-2r^2z^2}$$

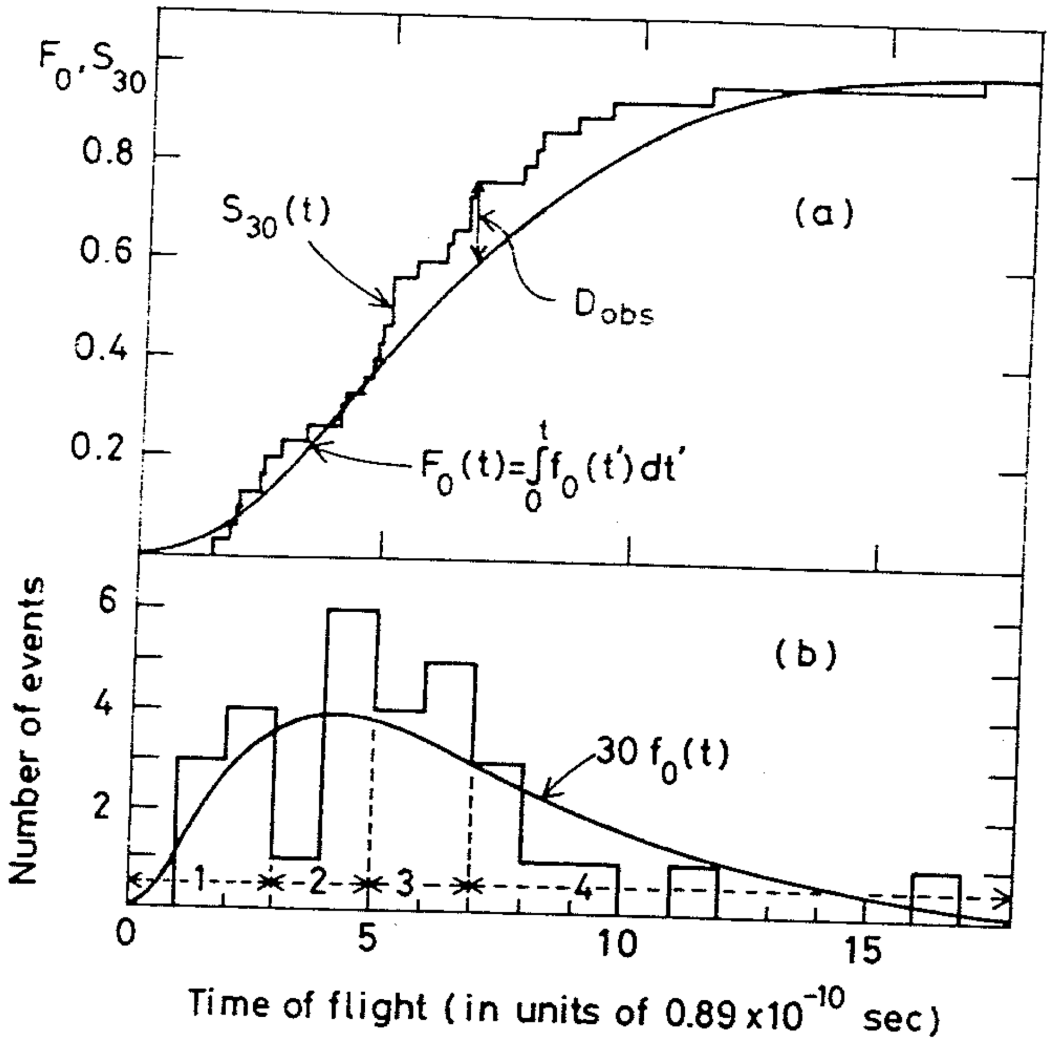


d = sqrt(n) Dn

kritische Werte im Grenzfall immer gr. als exakte Lösung

Kolmogorov-Smirnov-Test: Beispiel

30 Messungen einer Flugzeit Theorievorhersage $f_0(t)$



$D_{30} = \max |S_{30}(t) - F_o(t)| = 0.17$

P-Wert: = 0.25

sqrt(30)= 5.48
d=0.17*5.38=0.93

Vgl. mit Pearson-Test

4 Bins

(X^2_{obs}=3.0 and 3 degrees of freedom)

P-Wert: = 0.40

Kolmogorov-WDF: kritischen Grenzen für geg.

$$P(D_n \leq d_\alpha) \simeq 1 - \alpha$$

n \ α	.20	.10	.05	.02	.01
1	.9000	.9500	.9750	.9900	.9950
2	.6838	.7764	.8419	.9000	.9293
3	.5648	.6360	.7076	.7846	.8290
4	.4927	.5652	.6239	.6889	.7342
5	.4470	.5095	.5633	.6272	.6685
6	.4104	.4680	.5193	.5774	.6166
7	.3815	.4361	.4834	.5384	.5758
8	.3583	.4096	.4543	.5065	.5418
9	.3391	.3875	.4300	.4796	.5133
10	.3226	.3687	.4093	.4566	.4889
11	.3083	.3524	.3912	.4367	.4677
12	.2958	.3382	.3754	.4192	.4491
13	.2847	.3255	.3614	.4036	.4325
14	.2748	.3142	.3489	.3897	.4176
15	.2659	.3040	.3376	.3771	.4042
16	.2578	.2947	.3273	.3657	.3920
17	.2504	.2863	.3180	.3553	.3809
18	.2436	.2785	.3094	.3457	.3706
19	.2374	.2714	.3014	.3369	.3612
20	.2316	.2647	.2941	.3287	.3524
21	.2262	.2586	.2872	.3210	.3443
22	.2212	.2528	.2809	.3139	.3367
23	.2165	.2475	.2749	.3073	.3295
24	.2121	.2424	.2693	.3010	.3229
25	.2079	.2377	.2640	.2952	.3166
26	.2040	.2332	.2591	.2896	.3106
27	.2003	.2290	.2544	.2844	.3050
28	.1968	.2250	.2499	.2794	.2997
29	.1935	.2212	.2457	.2747	.2947
30	.1903	.2176	.2417	.2702	.2899
35	.1766	.2019	.2243	.2507	.2690
40	.1655	.1891	.2101	.2349	.2521
45	.1562	.1786	.1984	.2218	.2380
50	.1484	.1696	.1884	.2107	.2260
55	.1416	.1619	.1798	.2011	.2157
60	.1357	.1551	.1723	.1927	.2067
65	.1305	.1491	.1657	.1853	.1988
70	.1259	.1438	.1598	.1786	.1917
75	.1217	.1390	.1544	.1727	.1853
80	.1179	.1347	.1496	.1673	.1795
85	.1144	.1307	.1452	.1624	.1742
90	.1113	.1271	.1412	.1579	.1694
95	.1083	.1238	.1375	.1537	.1649
100	.1056	.1207	.1340	.1499	.1608
≥ 100	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Vergleich von 2 Stichproben

gegeben: 2 Stichproben $x_1, .. x_n$ $y_1 \ldots y_m$

Nullhypothese: entstammen gleicher Grundgesamtheit,
d.h. Form der WDF ist gleich Theorievorhersage $f_0(t)$

Kolmogorov-Smirnov test

Vergleich der beiden Kummulativ funktionen $S_m(x)$ and $S_n(x)$

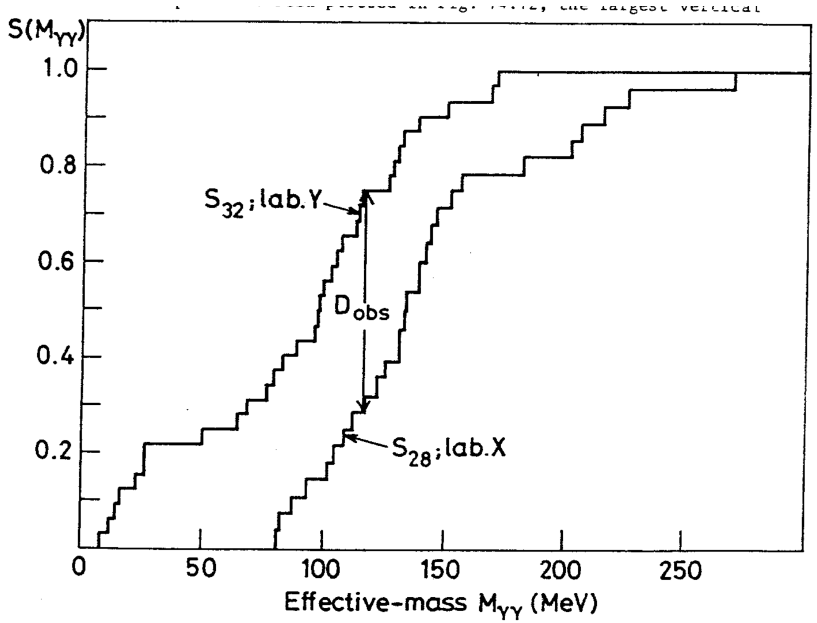
Teststatistik: $D_{mn} = \max |S_m(x) - S_n(x)|$

Für gr. Stichproben:

$$\lim_{m,n\to\infty} P\left(D_{mn} \le z\sqrt{\frac{1}{m}+\frac{1}{n}}\right) = 1 - 2\sum_{r=1}^{\infty}(-1)^{r-1} e^{-2r^2z^2}$$

Rückführung auf kritischen Wert für
1 Stichprobentest (n) d_α of D_n

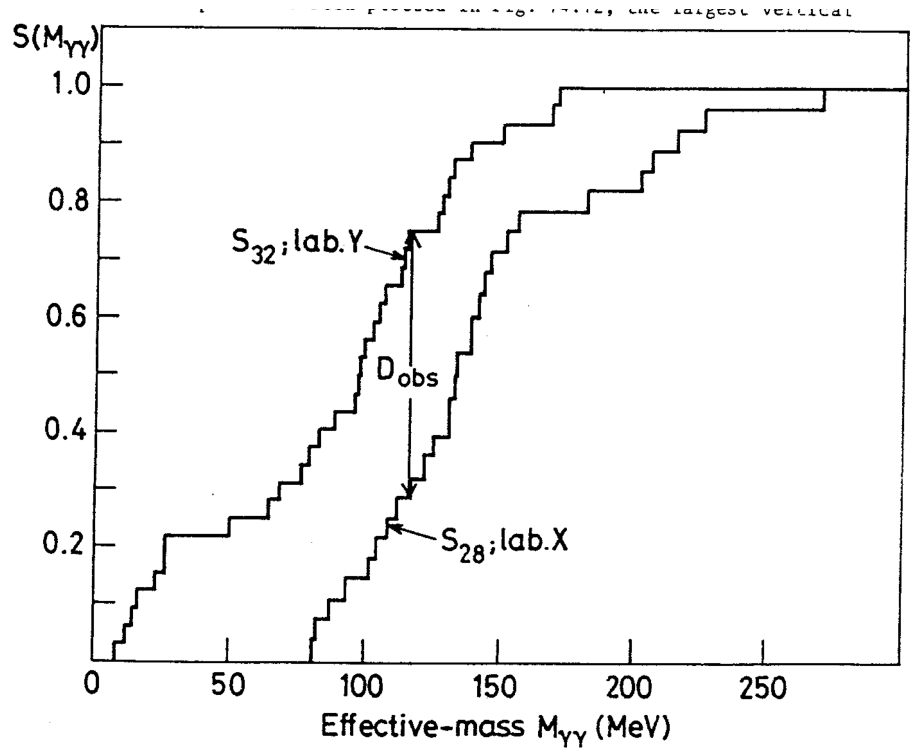
$$D_\alpha = d_\alpha \sqrt{1+\frac{n}{m}}$$



Vergleich von 2 Stichproben: Kolmogorov-Bsp.

gegeben: 2 Stichproben $x_1, .. x_{28}$ und $x_1 \ldots x_{32}$

$$D_{obs} = \max\left|S_{32}(M_{\gamma\gamma}) - S_{28}(M_{\gamma\gamma})\right| = \frac{24}{32} - \frac{8}{28} = 0.46$$



1 Sample-Test für n=28
kritischer Wert d für α=0.05

$$d_{.05} = 0.2499$$

Übertragung auf 2 Stichprobentest: $D_{.05} = 0.2499 \cdot \sqrt{1 + \frac{28}{32}} = 0.34$

→ Nullhypothese verwerfen wenn Signifikanzniveau auf 5% fixiert war

Vergleich von Histogrammen: Pearson χ^2-Test

gegeben: J Histogramme mit jeweils I Bins

→ J unabhängige Multionomialverteilungen

im j-ten Histogramm seien die Binwahrscheinlichkeiten gegeben durch:

$$p_{1j}, p_{2j}, \ldots, p_{Ij}; \qquad \sum_{i=1}^{I} p_{ij} = 1$$

die Einträge im i-ten Bin und die Gesamtanzahl im j-ten Histogramm seien:

$$n_{1j}, n_{2j}, \ldots, n_{Ij}; \qquad \sum_{i=1}^{I} n_{ij} = n_{\cdot j}$$

die Gesamtanzahl aller Bineinträge über alle J Histogramme sei:

$$\sum_{j=1}^{J} n_{\cdot j} = \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij} = n$$

Vergleich von Histogrammen: Pearson χ^2-Test

Nullhypothese: die Wkt für Bin i in allen J Histogramme ist gleich

$$H_o: \quad p_{i1} = p_{i2} = \cdots = p_{iJ} \qquad i = 1,2,\cdots,I$$

Für die unbekannten p_i ergibt sich aus der ML-Schätzung:

$$\hat{p}_{i\cdot} = \sum_{j=1}^{J} n_{ij}/n \qquad i = 1,2,\cdots,I$$

Diese Schätzwerte für p_i erfüllen Normierungsbedingung: $\sum_{i=1}^{I} \hat{p}_{i\cdot} = 1$

d.h. nur i-1 sind unabhängige Schätzwerte

Teststatistik: $$x^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - \hat{p}_{i\cdot} n_{\cdot j})^2}{\hat{p}_{i\cdot} n_{\cdot j}}$$

folgt im Grenzfall großer Bineinträge und bei Korrektheit der Nullhypothese einer Chi2-WDF mit (I-1)(J-1)=IJ+1-I-J Freiheitsgraden
(IJ-J unabhängige Beobachtungen und I-1 geschätze Parameter)