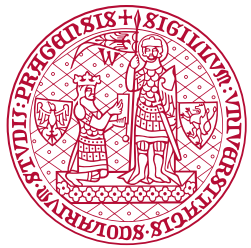
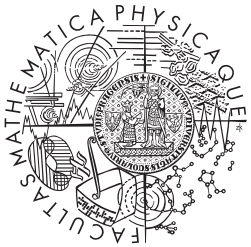
Úvod do praktické fyziky, cvičení 12
Metoda nejmenších čtevrců potřetí, test kvality fitu

Jan Matoušek

21. 12. 2020





UNIVERZITA KARLOVA
Matematicko-fyzikální
fakulta

Metoda nejmenších čtverců (cv. 10):

- Nechť mezi veličinou x a y platí vztah

$$y = \lambda(x|\boldsymbol{\theta}).$$

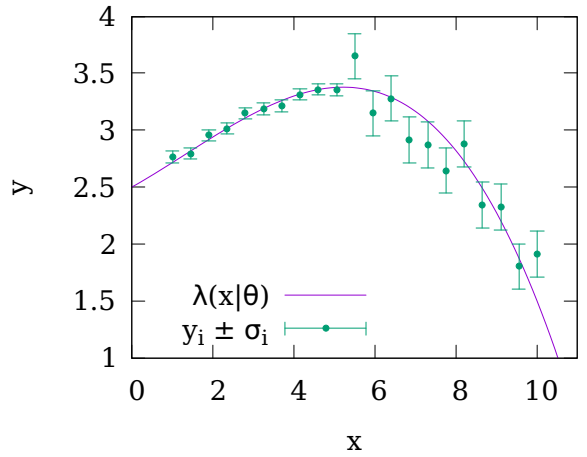
- Modelová funkce λ závisí na parametrech

$$\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_m).$$

- V bodech $x_1, x_2, ..., x_N$ jsme naměřili hodnoty $y_1, y_2, ..., y_N$ se standardními odchylkami $\sigma_1, \sigma_2, ..., \sigma_N$.
- $y_i \in N\big(\lambda(x_i|\boldsymbol{\theta}), \sigma_i\big)$
- Nejlepší odhad $\hat{\boldsymbol{\theta}}$ je ten, který minimalizuje χ^2

$$\chi^2(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} \frac{\big[y_i - \lambda(x_i|\boldsymbol{\theta})\big]^2}{\sigma_i^2}$$

(je to ten, který maximalizuje $\ln L$).



Lineární model ($\lambda(x|\boldsymbol{\theta})$ lineární v θ_i)

$$\lambda(x|\boldsymbol{\theta}) = \theta_1 f_1(x) + \theta_2 f_2(x) + ... + \theta_M f_M(x),$$

$$\frac{\partial \chi^2}{\partial \theta_i} = 0 \qquad \text{soustava lineárních rovnic.}$$

→ lze řešit analyticky (viz cvičení 10 a 11).

Nelineární model

$$\frac{\partial \chi^2}{\partial \theta_i} = 0 \qquad \text{soustava nelineárních rovnic.}$$

→ obecně nutno řešit numericky.

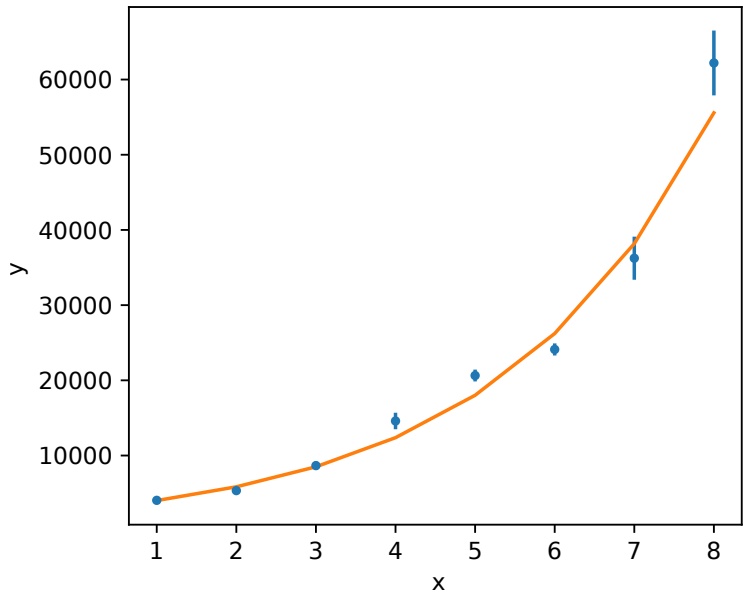
Nelineární metoda nejmenších čtverců

Příklad (`upf_cv12_covid.py`)

- Data: `upf_cv12_covid.txt` – aktivní případy COVID-19 od 24. 8. do 18. 10. [MZ ČR].
- Sloupce: číslo týdne $x_i = i$ a týdenní průměr $y_i = \overline{n}_i$, $\sigma_i^2 = \frac{1}{6}\sum_{j=1}^{7}(n_{ij} - \overline{n}_i)^2$.
- Model: exponenciální růst $\lambda(x|a,b) = ae^{bx}$.

Nelineární fit

- Numerická minimizace χ^2, např. pomocí `scipy.optimize.curve_fit()`
- Někdy se vyplatí použít logaritmickou škálu.



Linearizace fitu

$$y_i \to \ln y_i,$$
$$\lambda(x|a,b) \to \ln\lambda(x|a,b) = \ln a + bx,$$
$$\sigma_i \to \sigma_{\ln y_i} = \frac{\partial \ln y_i}{\partial y_i}\sigma_i = \frac{\sigma_i}{y_i}.$$

- Lineární regrese, např. `numpy.polyfit()`
- Riziko: I když $y_i \in N(\lambda(x_i), \sigma_i)$, po transformaci už to platit nemusí.

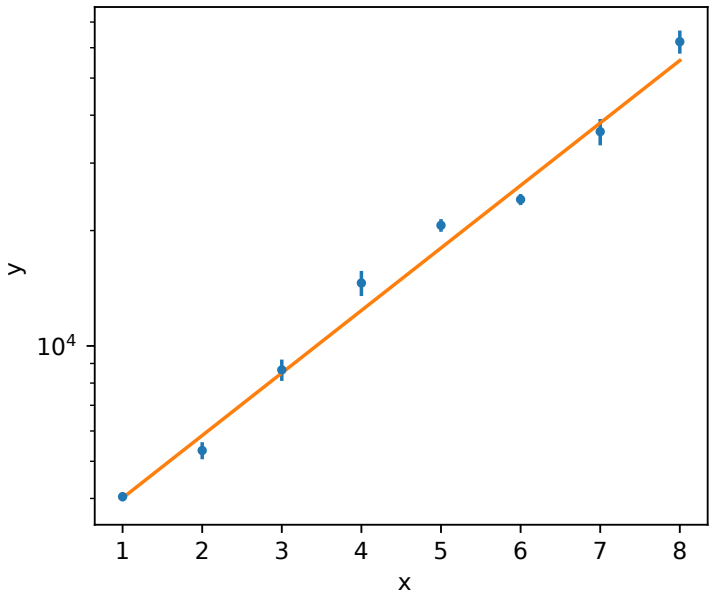
Nelineární metoda nejmenších čtverců

Příklad (`upf_cv12_covid.py`)

- Data: `upf_cv12_covid.txt` – aktivní případy COVID-19 od 24. 8. do 18. 10. [MZ ČR].
- Sloupce: číslo týdne $x_i = i$ a týdenní průměr $y_i = \overline{n}_i$, $\sigma_i^2 = \frac{1}{6}\sum_{j=1}^{7}(n_{ij} - \overline{n}_i)^2$.
- Model: exponenciální růst $\lambda(x|a,b) = ae^{bx}$.

Nelineární fit

- Numerická minimizace χ^2, např. pomocí `scipy.optimize.curve_fit()`
- Někdy se vyplatí použít logaritmickou škálu.



Linearizace fitu

$$y_i \to \ln y_i,$$
$$\lambda(x|a,b) \to \ln \lambda(x|a,b) = \ln a + bx,$$
$$\sigma_i \to \sigma_{\ln y_i} = \frac{\partial \ln y_i}{\partial y_i}\sigma_i = \frac{\sigma_i}{y_i}.$$

- Lineární regrese, např. `numpy.polyfit()`
- Riziko: I když $y_i \in N(\lambda(x_i), \sigma_i)$, po transformaci už to platit nemusí.

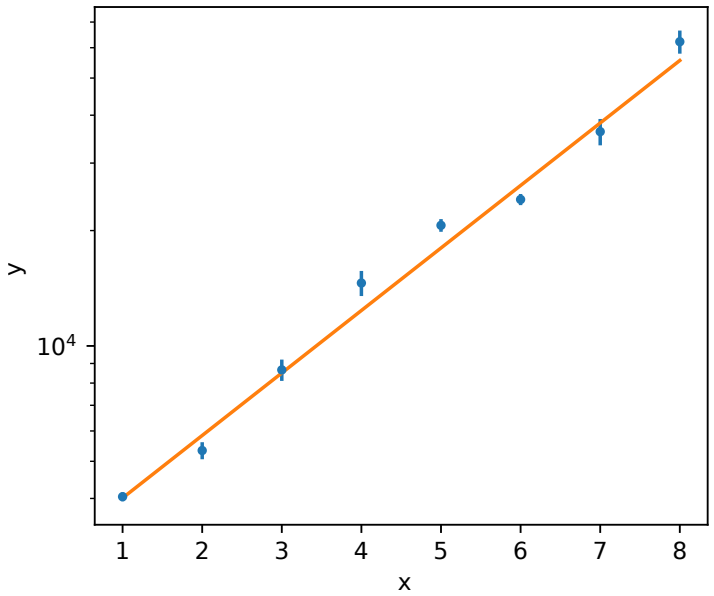
Nelineární metoda nejmenších čtverců

Příklad (`upf_cv12_covid.py`)

- Data: `upf_cv12_covid.txt` – aktivní případy COVID-19 od 24. 8. do 18. 10. [MZ ČR].
- Sloupce: číslo týdne $x_i = i$ a týdenní průměr $y_i = \overline{n}_i$, $\sigma_i^2 = \frac{1}{6}\sum_{j=1}^{7}(n_{ij} - \overline{n}_i)^2$.
- Model: exponenciální růst $\lambda(x|a,b) = ae^{bx}$.

Nelineární fit

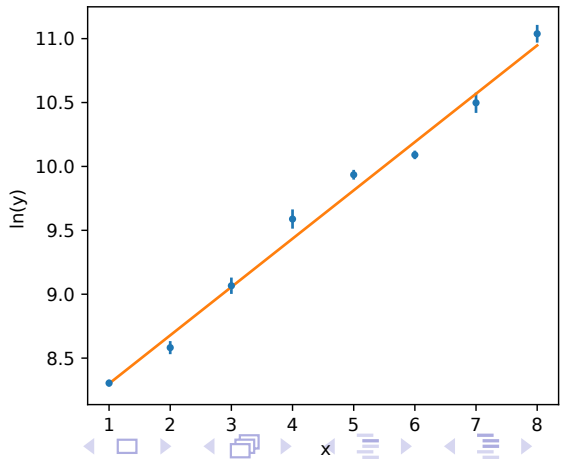
- Numerická minimizace χ^2, např. pomocí `scipy.optimize.curve_fit()`
- Někdy se vyplatí použít logaritmickou škálu.



Linearizace fitu

$$y_i \to \ln y_i,$$
$$\lambda(x|a,b) \to \ln \lambda(x|a,b) = \ln a + bx,$$
$$\sigma_i \to \sigma_{\ln y_i} = \frac{\partial \ln y_i}{\partial y_i}\sigma_i = \frac{\sigma_i}{y_i}.$$

- Lineární regrese, např. `numpy.polyfit()`
- Riziko: I když $y_i \in N(\lambda(x_i), \sigma_i)$, po transformaci už to platit nemusí.



Metoda nejmenších čtverců – chyby obou proměnných

- x i y jsou náhodné proměnné.
- Vzdálenost bodu $[x_i, y_i]$ od modelové funkce, vážená standardními odchylkami:

$$d_i^2 = \frac{(x_i - \tilde{x})^2}{\sigma_{x_i}^2} + \frac{(y_i - f(\tilde{x}))^2}{\sigma_{y_i}^2}$$

- Vzdálenost bude minimální, pokud

$$\frac{\partial d_i^2}{\partial \tilde{x}} = 0$$

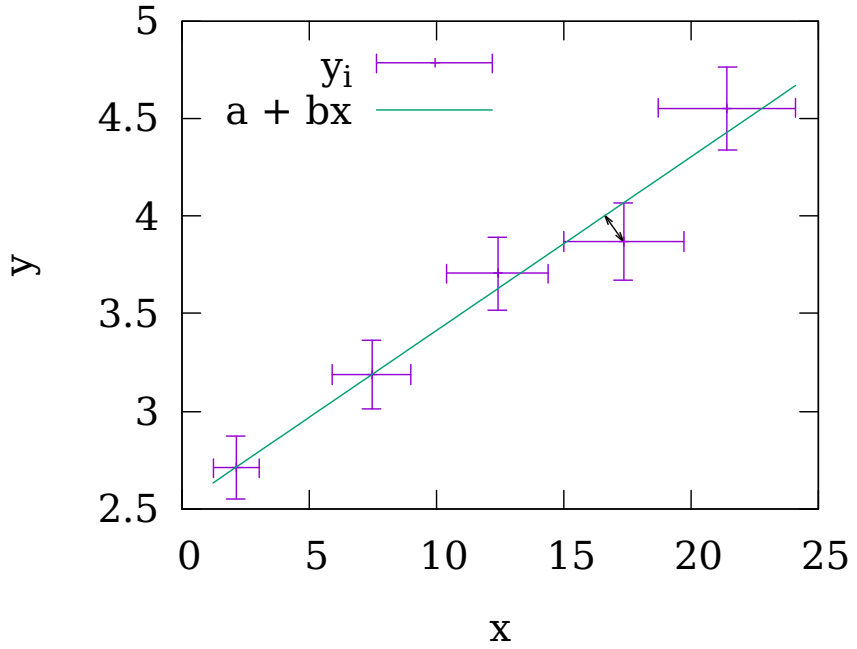
- Odtud můžeme odvodit

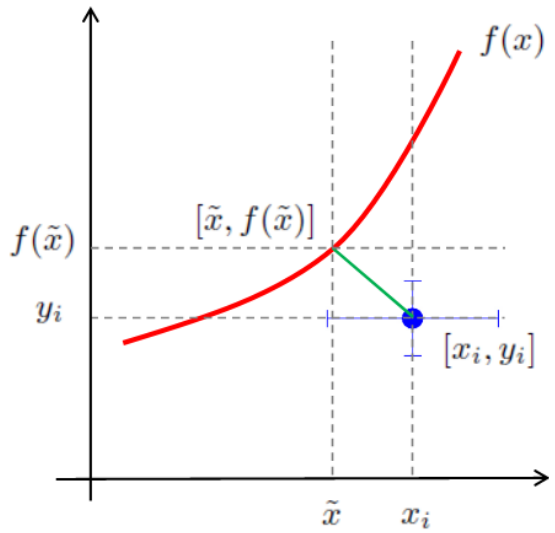
$$d_i^2 = \frac{\big(y_i - f(x_i)\big)^2}{\sigma_{y_i}^2 + \sigma_{x_i}^2 f'^2(x_i)} = \frac{\big(y_i - f(x_i)\big)^2}{\sigma_{\text{tot.}}^2}$$

- Nakonec χ^2 upravený pro chyby v y i x:

$$\chi^2 = \sum_{i=1}^{N} d_i^2.$$

- Prakticky např.:
  Root (`TGraphErrors::Fit()` automaticky),
  Gnuplot (`fit... u 1:2:3:4 xyerrors...`)





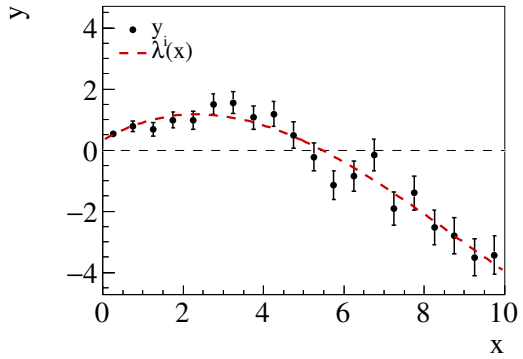
Kvalita fitu pomocí reziduí

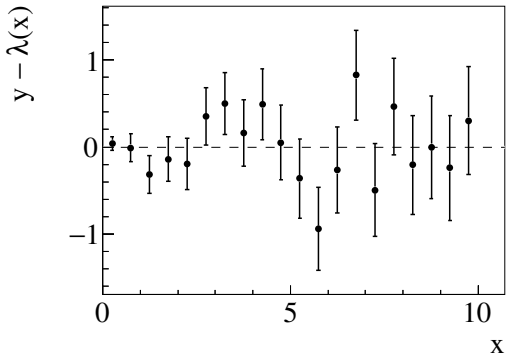
- Fitujeme body $[x_i, y_i]$ funkcí $\lambda(x|\boldsymbol{\theta})$.
- Rezidua: $r_i = y_i - \lambda(x_i|\boldsymbol{\theta})$.
- Měla by být rovnoměrně rozdělená kolem nuly.
- Hledáme systematické odchylky.
- Můžeme odhalit oblasti, kde fit nepopisuje data dobře.
- Měly by mít normální rozdělení (standardizovaná rezidua $\frac{r_i}{\sigma_{r_i}}$ normální standardní).
- Standardní odchylka reziduí:

$$\sigma_{r_i}^2 = \left(\frac{\partial r_i}{\partial y_i}\sigma_i\right)^2 + \left(\frac{\partial r_i}{\partial \lambda}\sigma_\lambda\right)^2 + 2\frac{\partial r_i}{\partial y_i}\frac{\partial r_i}{\partial \lambda}\operatorname{cov}\left(y_i, \lambda(x_i|\boldsymbol{\theta})\right)$$
$$= \sigma_i^2 + \sigma_\lambda^2 - 2\operatorname{cov}\left(y_i, \lambda(x_i|\boldsymbol{\theta})\right)$$

- $\sigma_{r_i} \approx \sigma_i$, pokud $\sigma_i \gg \sigma_\lambda$ (např. je-li bodů hodně).
- Pro vážený průměr $\lambda(x_i|\boldsymbol{\theta}) = \overline{y}$:

$$\operatorname{cov}(y_i, \overline{y}) = \sigma_{\overline{y}}^2 \qquad \rightarrow \qquad \sigma_{r_i}^2 = \sigma_i^2 - \sigma_{\overline{y}}^2.$$





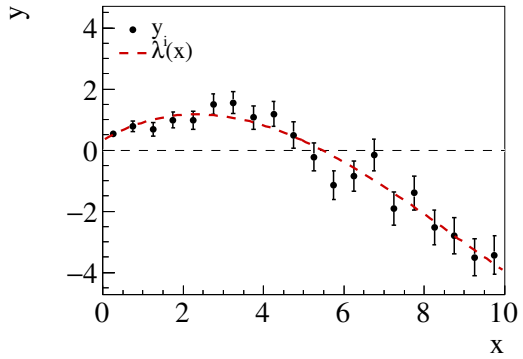
Kvalita fitu pomocí reziduí

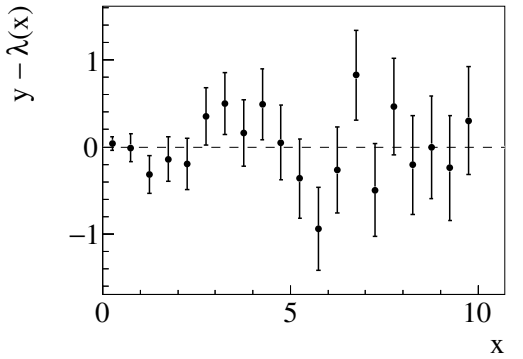
- Fitujeme body $[x_i, y_i]$ funkcí $\lambda(x|\boldsymbol{\theta})$.
- Rezidua: $r_i = y_i - \lambda(x_i|\boldsymbol{\theta})$.
- Měla by být rovnoměrně rozdělená kolem nuly.
- Hledáme systematické odchylky.
- Můžeme odhalit oblasti, kde fit nepopisuje data dobře.
- Měly by mít normální rozdělení
  (standardizovaná rezidua $\frac{r_i}{\sigma_{r_i}}$ normální standardní).
- Standardní odchylka reziduí:

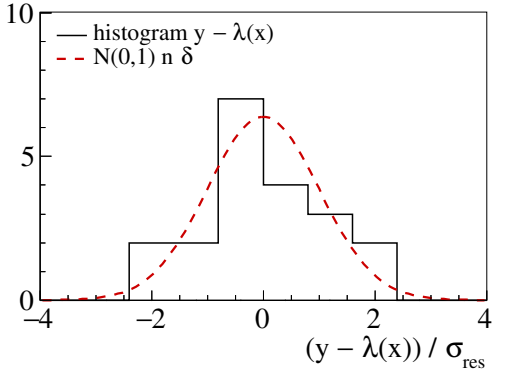
$$\begin{aligned}\sigma_{r_i}^2 &= \left(\frac{\partial r_i}{\partial y_i}\sigma_i\right)^2 + \left(\frac{\partial r_i}{\partial \lambda}\sigma_\lambda\right)^2 + 2\frac{\partial r_i}{\partial y_i}\frac{\partial r_i}{\partial \lambda}\operatorname{cov}\left(y_i, \lambda(x_i|\boldsymbol{\theta})\right) \\ &= \sigma_i^2 + \sigma_\lambda^2 - 2\operatorname{cov}\left(y_i, \lambda(x_i|\boldsymbol{\theta})\right)\end{aligned}$$

- $\sigma_{r_i} \approx \sigma_i$, pokud $\sigma_i \gg \sigma_\lambda$ (např. je-li bodů hodně).
- Pro vážený průměr $\lambda(x_i|\boldsymbol{\theta}) = \overline{y}$:

$$\operatorname{cov}(y_i, \overline{y}) = \sigma_{\overline{y}}^2 \qquad \rightarrow \qquad \sigma_{r_i}^2 = \sigma_i^2 - \sigma_{\overline{y}}^2.$$







Rozdělení χ^2

- Náhodné proměnné $z_i \in N(0,1)$.
- Náhodná proměnná $y = \sum_{i=1}^{N} z_i^2$.
- Pak $y \in \chi^2(N)$, tj. má rozdělení χ^2 s N stupni volnosti.
- Hustota pravděpodobnosti:

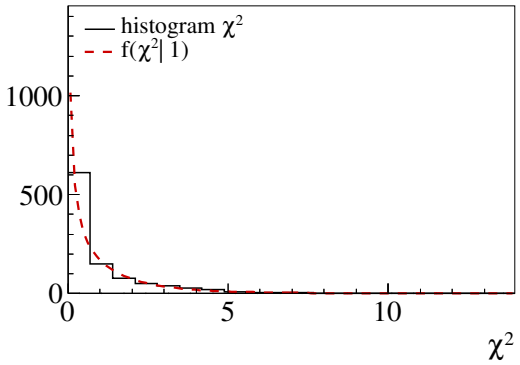
$$f_{\chi^2}(y|N) = \frac{1}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)} y^{\frac{N}{2}-1} e^{-\frac{y}{2}}, \qquad y \in [0,\infty), \quad N = 1, 2, \ldots$$

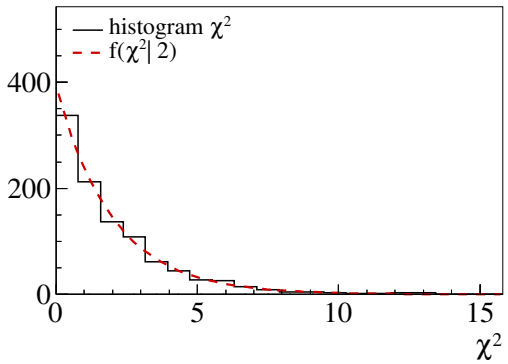
- Gamma funkce: definovaná integrálem

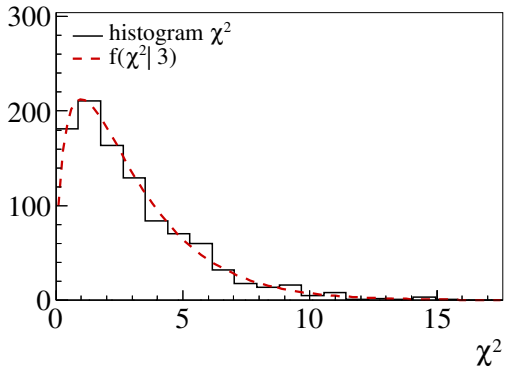
$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t.$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \qquad \Gamma(x+1) = x\Gamma(x), \qquad \Gamma(n) = (n-1)!$$

- Momenty rozdělení: $E[y] = N$, $V[y] = 2N$.







Rozdělení χ^2

- Náhodné proměnné $z_i \in N(0,1)$.
- Náhodná proměnná $y = \sum_{i=1}^{N} z_i^2$.
- Pak $y \in \chi^2(N)$, tj. má rozdělení χ^2 s N stupni volnosti.
- Hustota pravděpodobnosti:

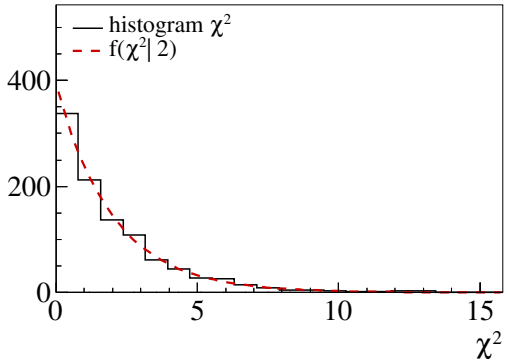
$$f_{\chi^2}(y|N) = \frac{1}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)} y^{\frac{N}{2}-1} e^{-\frac{y}{2}}, \qquad y \in [0,\infty), \quad N = 1, 2, \ldots$$

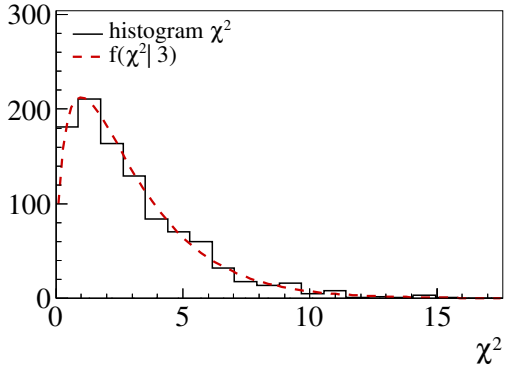
- Gamma funkce: definovaná integrálem

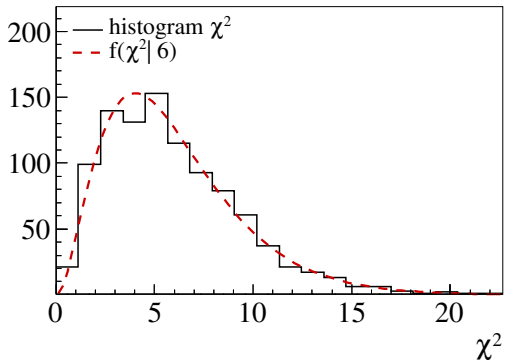
$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t.$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \qquad \Gamma(x+1) = x\Gamma(x), \qquad \Gamma(n) = (n-1)!$$

- Momenty rozdělení: $E[y] = N$, $V[y] = 2N$.







Rozdělení χ^2

- Náhodné proměnné $z_i \in N(0,1)$.
- Náhodná proměnná $y = \sum_{i=1}^{N} z_i^2$.
- Pak $y \in \chi^2(N)$, tj. má rozdělení χ^2 s N stupni volnosti.
- Hustota pravděpodobnosti:

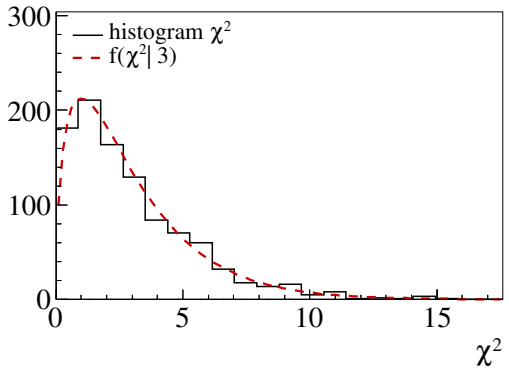
$$f_{\chi^2}(y|N) = \frac{1}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)} y^{\frac{N}{2}-1} e^{-\frac{y}{2}}, \qquad y \in [0,\infty), \quad N = 1, 2, \ldots$$

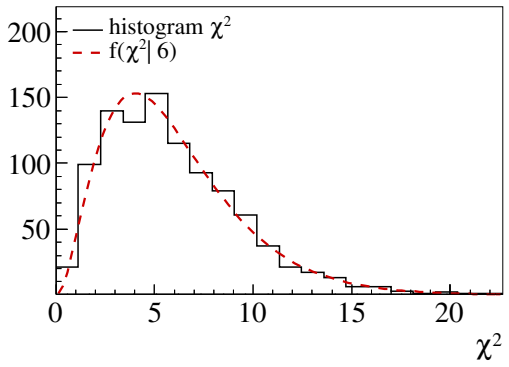
- Gamma funkce: definovaná integrálem

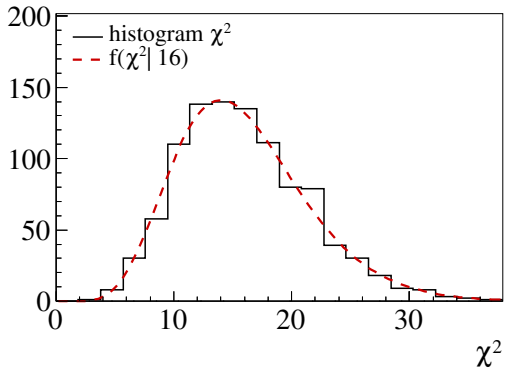
$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t.$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \qquad \Gamma(x+1) = x\Gamma(x), \qquad \Gamma(n) = (n-1)!$$

- Momenty rozdělení: $E[y] = N$, $V[y] = 2N$.







Rozdělení χ^2

- Náhodné proměnné $z_i \in N(0,1)$.
- Náhodná proměnná $y = \sum_{i=1}^{N} z_i^2$.
- Pak $y \in \chi^2(N)$, tj. má rozdělení χ^2 s N stupni volnosti.
- Hustota pravděpodobnosti:

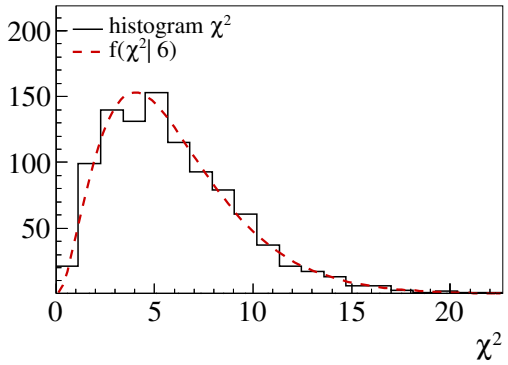
$$f_{\chi^2}(y|N) = \frac{1}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)} y^{\frac{N}{2}-1} e^{-\frac{y}{2}}, \qquad y \in [0,\infty), \quad N = 1, 2, \ldots$$

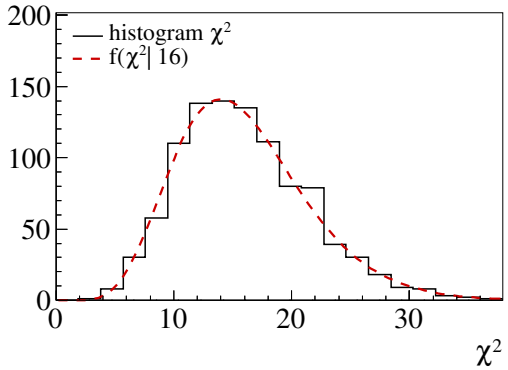
- Gamma funkce: definovaná integrálem

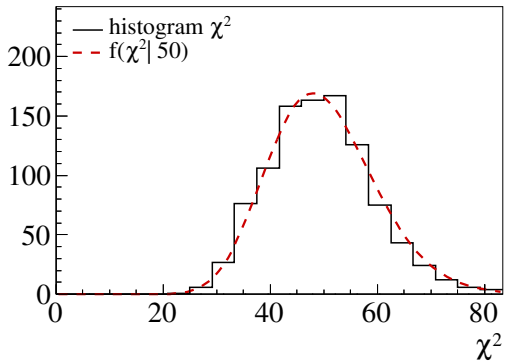
$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t.$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \qquad \Gamma(x+1) = x\Gamma(x), \qquad \Gamma(n) = (n-1)!$$

- Momenty rozdělení: $E[y] = N$, $V[y] = 2N$.







Test kvality fitu pomocí χ^2

- Nezávislá měření $y_1, y_2, ..., y_N$.
- $y_i \in N\big(\lambda(x_i|\boldsymbol{\theta}), \sigma_i\big)$
- λ modelová funkce s parametry $\theta_1, \theta_2, ...\theta_M$.
- Potom

$$\chi^2(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} \frac{\big[y_i - \lambda(x_i|\boldsymbol{\theta})\big]^2}{\sigma_i^2}$$

  má rozdělení $\chi^2(N - M)$.
- Počet stupňů volnosti je $N - M$, tedy počet bodů mínus počet fitovaných parametrů.

$$E[\chi^2] = N - M, \qquad V[\chi^2] = 2(N - M).$$

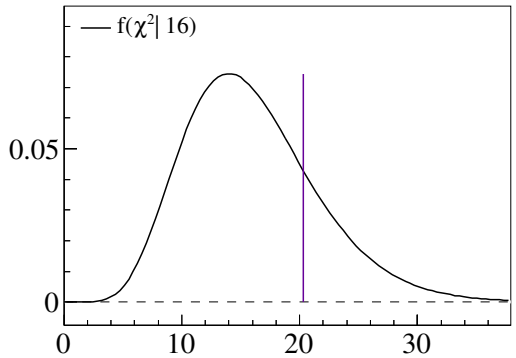
- χ^2 na počet stupňů volnosti

$$E\left[\frac{\chi^2}{N - M}\right] = 1.$$

- Pro jeden fit spočítáme jeho χ_0^2 a jeho p-hodnotu:

$$P(y \geq \chi_0^2) = \int_{\chi_0^2}^{\infty} f_{\chi^2}(y|N - M)\mathrm{d}y = 1 - F_{\chi^2}(\chi_0^2|N - M)$$

- Je-li $P\big(y \geq \chi_0^2\big) < \alpha$, zamítneme nulovou hypotézu, tj. že data lze popsat funkcí $\lambda(x, \boldsymbol{\theta})$.
- α je hladina signifikance, obvykle 0.05 nebo 0.01.



χ_0^2 spočítaný z jedné sady nafitovaných dat, porovnaný s hustotou pravděpodobnosti χ^2. P-hodnota odpovídá integrálu od vertikální čáry doprava.

Test kvality fitu pomocí χ^2

- Nezávislá měření $y_1, y_2, ..., y_N$.
- $y_i \in N\big(\lambda(x_i|\boldsymbol{\theta}), \sigma_i\big)$
- λ modelová funkce s parametry $\theta_1, \theta_2, ...\theta_M$.
- Potom

$$\chi^2(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} \frac{\big[y_i - \lambda(x_i|\boldsymbol{\theta})\big]^2}{\sigma_i^2}$$

  má rozdělení $\chi^2(N - M)$.
- Počet stupňů volnosti je $N - M$, tedy počet bodů mínus počet fitovaných parametrů.

$$E[\chi^2] = N - M, \qquad V[\chi^2] = 2(N - M).$$

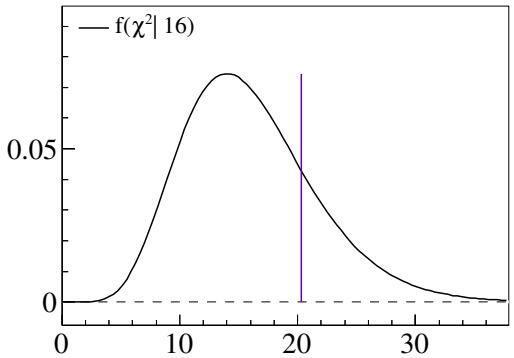
- χ^2 na počet stupňů volnosti

$$E\left[\frac{\chi^2}{N - M}\right] = 1.$$

- Pro jeden fit spočítáme jeho χ_0^2 a jeho p-hodnotu:

$$P\big(y \geq \chi_0^2\big) = \int_{\chi_0^2}^{\infty} f_{\chi^2}(y|N - M)\mathrm{d}y = 1 - F_{\chi^2}(\chi_0^2|N - M)$$

- Je-li $P\big(y \geq \chi_0^2\big) < \alpha$, zamítneme nulovou hypotézu, tj. že data lze popsat funkcí $\lambda(x, \boldsymbol{\theta})$.
- α je hladina signifikance, obvykle 0.05 nebo 0.01.



χ_0^2 spočítaný z jedné sady nafitovaných dat, porovnaný s hustotou pravděpodobnosti χ^2. P-hodnota odpovídá integrálu od vertikální čáry doprava.

Test kvality fitu pomocí χ^2

- Programy pro zpracování dat poskytují F_{χ^2} nebo přímo p-hodnotu:
  - gnuplot automaticky vypisuje p-hodnotu při fitu.
  - $F_{\chi^2,r}$ = Excel: `CHISQ.DIST(x, r, true)`, Root: `ROOT::Math::chisquared_cdf(x, r)`, Python: `stats.chi2.cdf(x, r)`...
- Lze zjistit i jaký χ^2 odpovídá určité hladině signifikance:
  - $\chi^2 = F^{-1}(P, r)$ = Excel: `CHISQ.INV(P,r)`, Root: `ROOT::Math::chisquared_quantile(P,r)`...
- Tabulka p-hodnot pro počet stupňů volnosti 1–10:

Počet stupňů volnosti	χ^2								hladina signifikance $\alpha = 5\,\%$ ↓	hladina signifikance $\alpha = 1\,\%$ ↓	
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P[y≥ χ^2]	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

- Pro počet stupňů volnosti > 10 konverguje k $N(k, \sqrt{2k})$.