Analýza dát

Úvodné sústredenie TMF

04. október 2024

Matej Badin

Prečo spracovávame exp. dáta?

- Meranie veličiny
- Hľadanie súvisu medzi meranými veličinami – korelujú?
- Overenie teoretického modelu
- Overenie hypotézy

Čím spracovávať exp. dáta?

Automatizácia zberu a analýzy dát (videá, fotografie, zvuk)

- Audacity
- ImageJ
- Tracker
- OpenCV [C++, Python, Matlab, ...]
- Textové súbory – príklad výstup dig. osciloskopu

Automatizácia a spracovanie veľkého množstva dát

- Excel
- Python [SciPy, NumPy, ...]
- R, Matlab, ...

Spracovanie chýb merania

Prečo potrebujeme vedieť určiť chybu merania?

Odmerali sme hmotnosť závažia ...

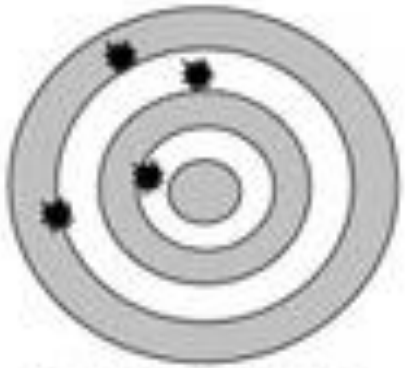
$$1.23\,\mathrm{kg}$$
$$(1.23 \pm 5)\,\mathrm{kg}$$
$$(1.23 \pm 1.1)\,\mathrm{kg}$$
$$(1.23 \pm 0.1)\,\mathrm{kg}$$
$$(1.23 \pm 0.02)\,\mathrm{kg}$$

Rôzna užitočnosť informácie

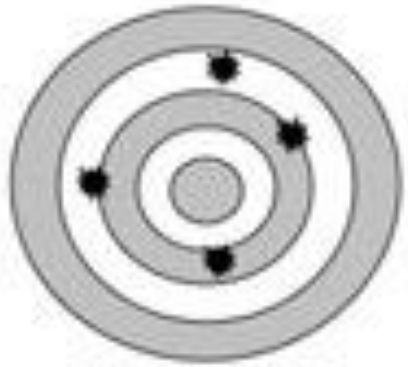
Systematické vs náhodné chyby

Accuracy – presnosť
Systematická chyba

Príklady:
- Nepresnosť prístroja
  meranie dĺžky – najmenšia jednotka
- Systematický vplyv okolia – „fúka vietor sprava“
- Nesprávna kalibrácia prístroja



Not Accurate
Not Precise



Accurate
Not Precise

Precision – opakovateľnosť
Náhodná chyba

Príklady:
- Šum
- Fluktuácie T, p
- Reakčná doba

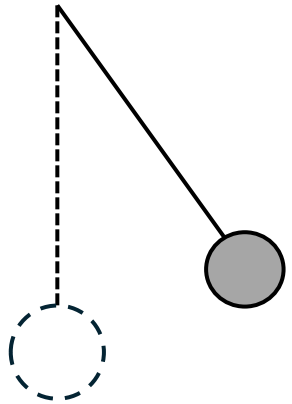
Not Accurate
Precise

Accurate
and Precise

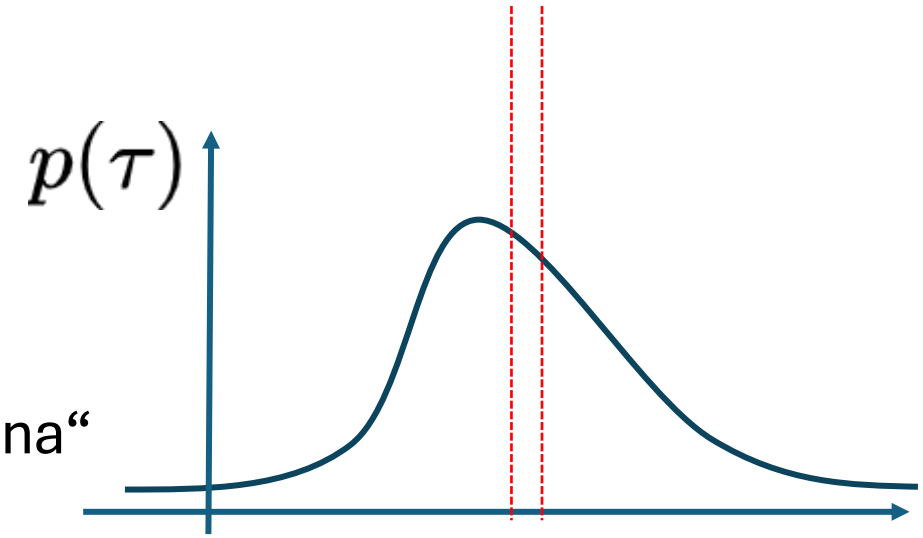
Zdroj - http://www.carolina.com/teacher-resources/Interactive/accuracy-versus-precision-beanbag-toss/

Ako **určiť** náhodnú chybu?

Príklad – chceme určiť čas kedy kyvadlo prejde rovnovážnou polohou; čas pádu, atď.
Chyba domininovaná *reakčnou dobou stlačenia stopiek*.



$$T_{\text{mer}} = T_{\text{skut}} + \tau$$

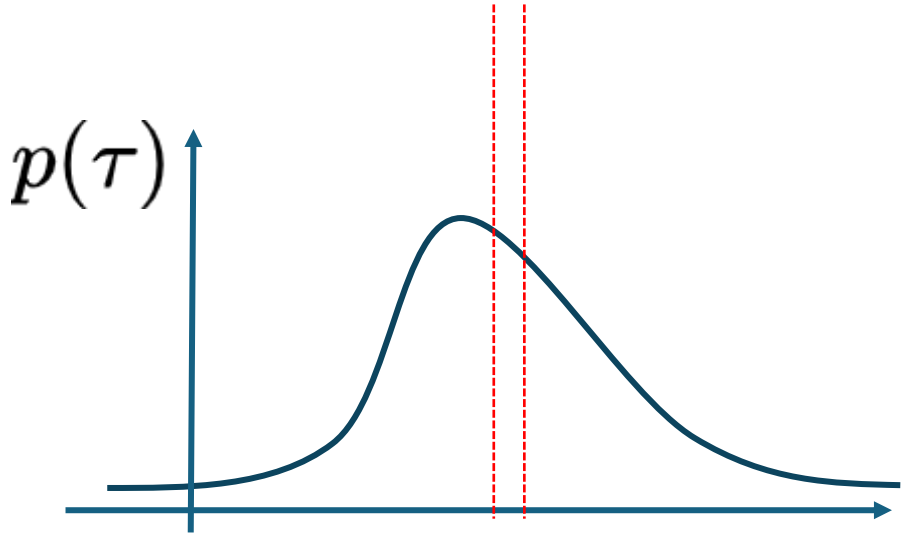


Potrebujeme zmerať reakčný čas
„**Ideálny scenár**“ – meranie viem opakovať „do nekonečna“

Zistím skutočnú pravdepodobnostnú distribúciu $p(\tau)$
→ pravdepodobnosť namerania hodnoty $\tau_i \in \{\tau_i; \tau_i + \Delta\tau\}$ je $p(\tau_i)\Delta\tau$

Ako **určiť** náhodnú chybu?

Ak poznám **skutočnú** pravdepodobnostnú distribúciu rozdelím si x-ovú os na veľa malých dielikov



- Priemerná hodnota

$$\mu(\tau) = \sum_i \tau_i p(\tau_i) \Delta\tau$$

- Disperzia

$$D(\tau) = \sum_i (\tau - \mu(\tau))^2 p(\tau_i) \Delta\tau$$

- Štandardná odchýlka

$$\sigma(\tau) = \sqrt{D(\tau)}$$

Problém – skutočnú pravdepodobnostnú distribúciu nepoznáme!

Ako **odhadnúť** náhodnú chybu?

V praxi máme iba limitovaný počet meraní N napr. 5, 10, ..., 15, ...

$$\tau_i; i \in 1, 2, \ldots, N$$

Chceme z nich čo najpresnejšie určiť meranú veličinu (v tomto prípade reakčný čas).
Teda čo najlepšie **odhadnúť** skutočný priemer $\mu(\tau)$ a skutočnú štandardnú odchýlku $\sigma(\tau)$ štatistického súboru.

„Kuchársky recept“

$$\mu(\tau) \approx \overline{\tau} = \frac{1}{N}\sum_i \tau_i$$

$$\sigma(\tau) \approx s(\tau) = \sqrt{\frac{1}{N-1}\sum_i (\tau_i - \overline{\tau})^2}$$

$$\sigma(\overline{\tau}) \approx s(\overline{\tau}) = \frac{s(\tau)}{\sqrt{N}}$$

Prečo?

Interval spoľahlivosti

Meranú veličinu potom odhadneme ako

$$\tau = \overline{\tau} \pm ts(\tau)$$

Kde ***t*** je tzv. Studentov koefficient

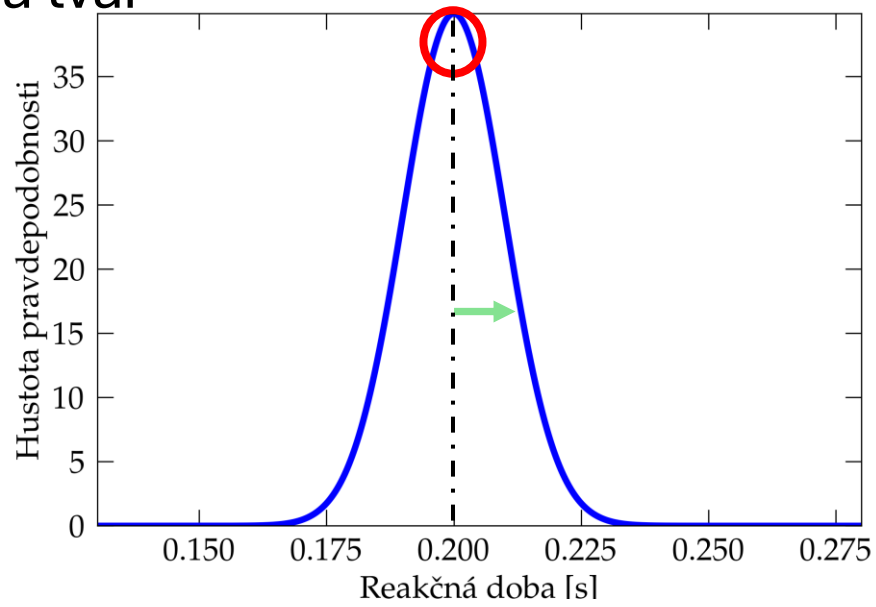
	$t(P,n)$			
n	$P = 68{,}3\%$	$P = 95{,}0\%$	$P = 99{,}0\%$	$P = 99{,}73\%$
3	1,32	4,30	9,92	19,21
4	1,20	3,18	5,84	9,22
5	1,15	2,78	4,60	6,62
6	1,11	2,57	4,03	5,51
7	1,09	2,45	3,71	4,90
8	1,09	2,37	3,50	4,53
9	1,07	2,31	3,36	4,27
10	1,06	2,26	3,25	4,09
11	1,06	2,23	3,17	3,96
12	1,05	2,20	3,11	3,85
15	1,04	2,15	2,98	3,63
20	1,03	2,08	2,86	3,45
30	1,02	2,05	2,76	3,28
50	1,01	2,01	2,68	3,16
100	1,00	1,98	2,63	3,08
∞	1,00	1,96	2,58	3,00

Viď. napr. Bohumil Vybíral – Zpracovaní dat fyz. meraní

[Nepovinný slide] Prečo sa **priemer a štatistická odchýlka takto odhaduje?**

Drvivá väčšina distribúcii chýb má normálové rozdelenie!
Potom hustota pravdepodobnosti namerať hodnotu τ má tvar

$$p(\tau) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{Exp}\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



Parametre μ a σ avšak samozrejme nepoznáme!

Pravdepodobnosť namerať sadu hodnôt je (nezávislé merania)

$$p = p(\tau_1)p(\tau_2)\ldots p(\tau_N) \sim \mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_i(\tau_i-\mu)^2\right]$$

Nájdeme také $\hat{\mu}$ ako také μ, ktoré maximalizuje tento výraz (pozri v literatúre *maximum likehood estimation* [1]). Podobne aj pre σ

[1] Napr. en.wikipedia.org/wiki/Maximum_likelihood_estimation

[Nepovinný slide] Prečo sa **priemer a štatistická odchýlka sa takto odhaduje?**

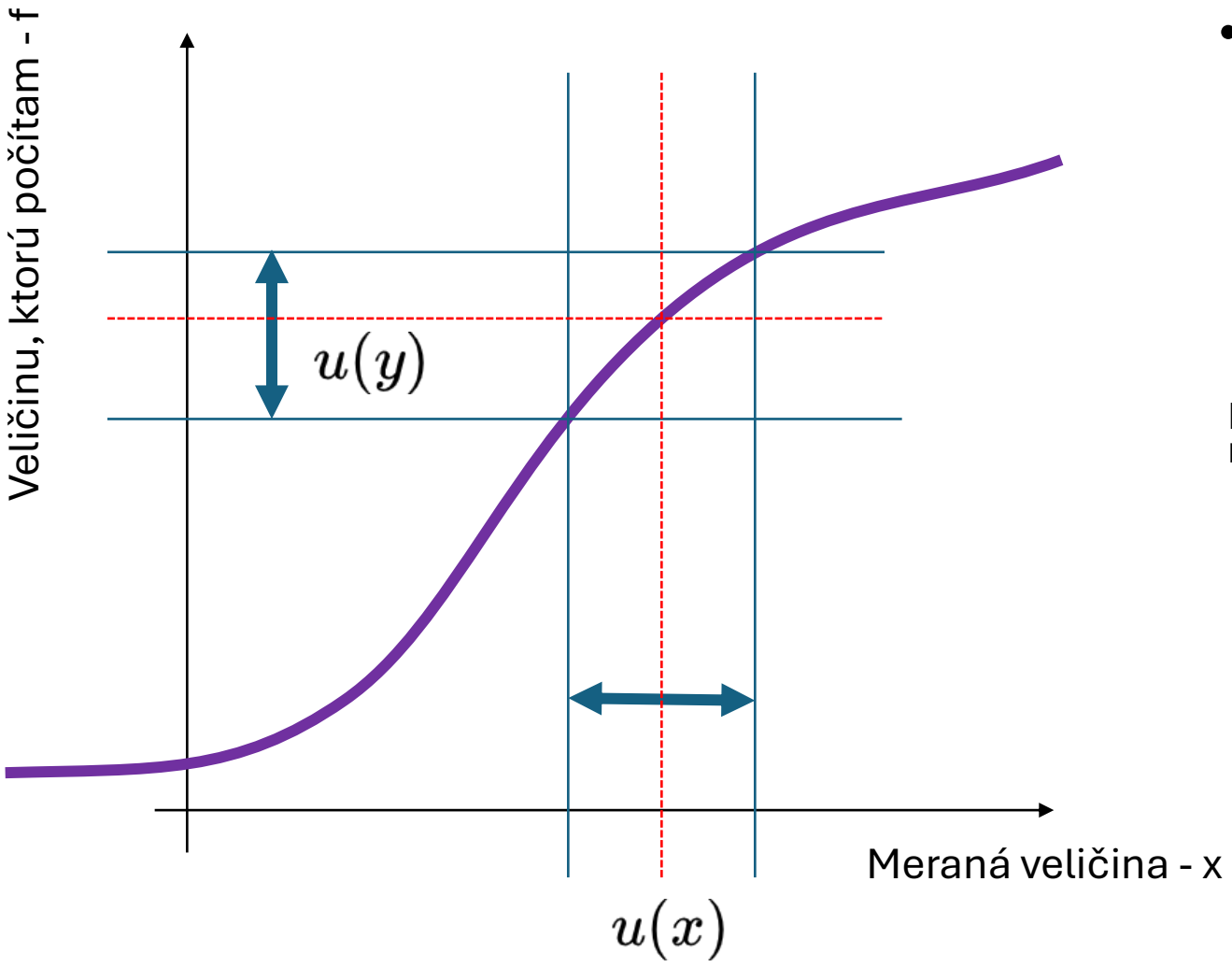
Dá sa ukázať, že takéto odhady

$$\hat{\mu} = \bar{\tau} = \frac{1}{N}\sum_i \tau_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_i (\tau_i - \bar{\tau})^2}$$

- Sú „nevychýlené“ – t.j. v priemere (pre rôzne opakovania experimentu pre fixné N) vedú k skutočným hodnotám μ a σ
- „Účinné“ – dávajú najmenšiu možnú strednú kvadratickú chybu voči skutočnému priemeru a štandardnej odchýlke pre $N \to \infty$, t.j. „nevieme vymyslieť lepší vzorček“.

Šírenie chýb



- Min vs max alebo
- Metóda linearizácie

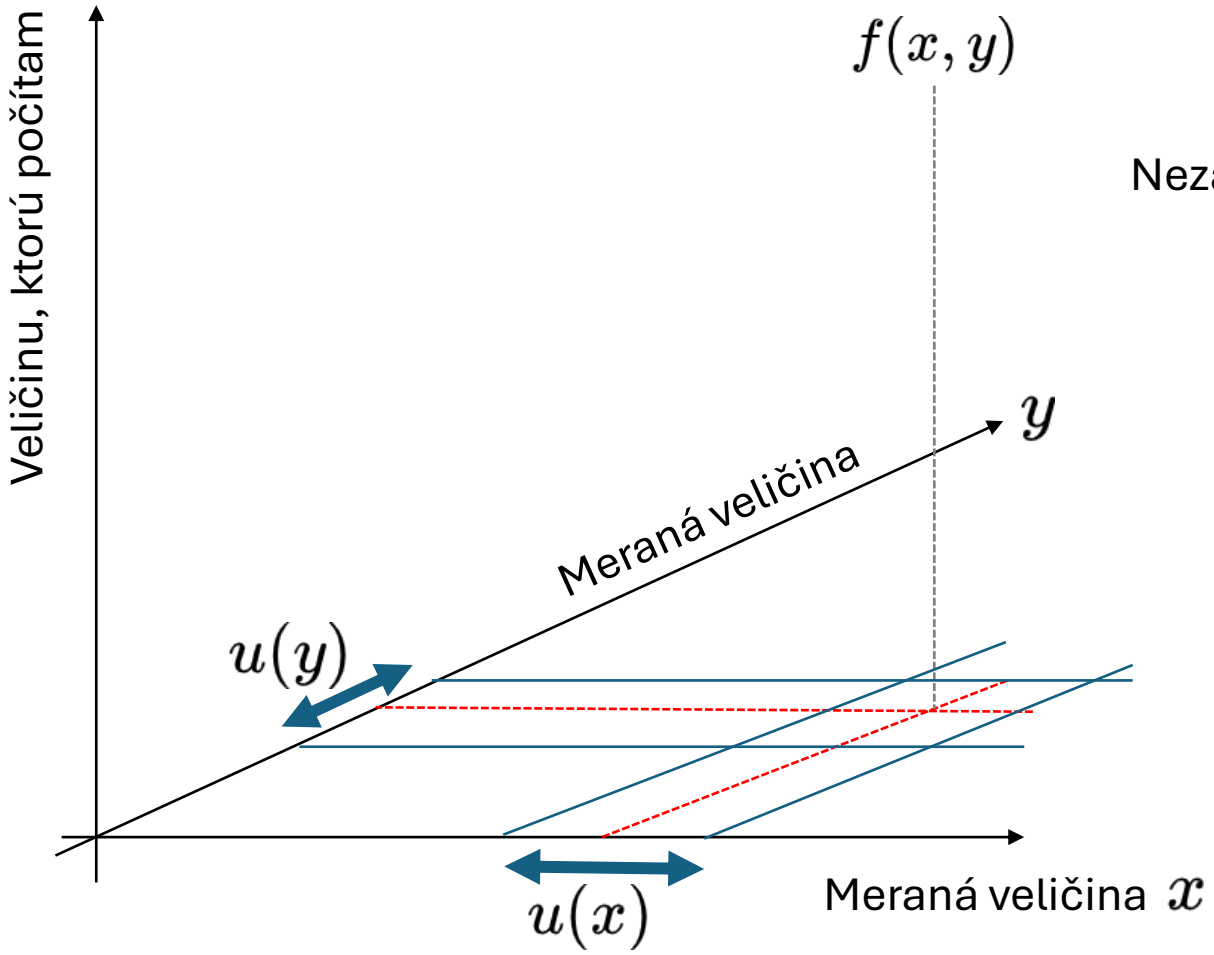
$$u(f) \approx \left|\frac{\partial f}{\partial x}\right| u(x)$$

Pozor pri nelineárnych funkciách nie je pravdepodobnostné rozdelenie vo vypočítanej premenennej normálové

$$p'(f(x)) = \left|\frac{\partial f}{\partial x}\right|^{-1} p(x)$$

(Platí iba pre malé úseky, pre prosté funkcie)

Šírenie chýb - funkcie viac premenných



Príklad – meranie tiažového zrýchlenia

$$g = \frac{4\pi^2 l}{T^2}$$

Nezávislé veličiny

$$u(f) \approx \left|\frac{\partial f}{\partial x}\right| u(x) + \left|\frac{\partial f}{\partial y}\right| u(y)$$

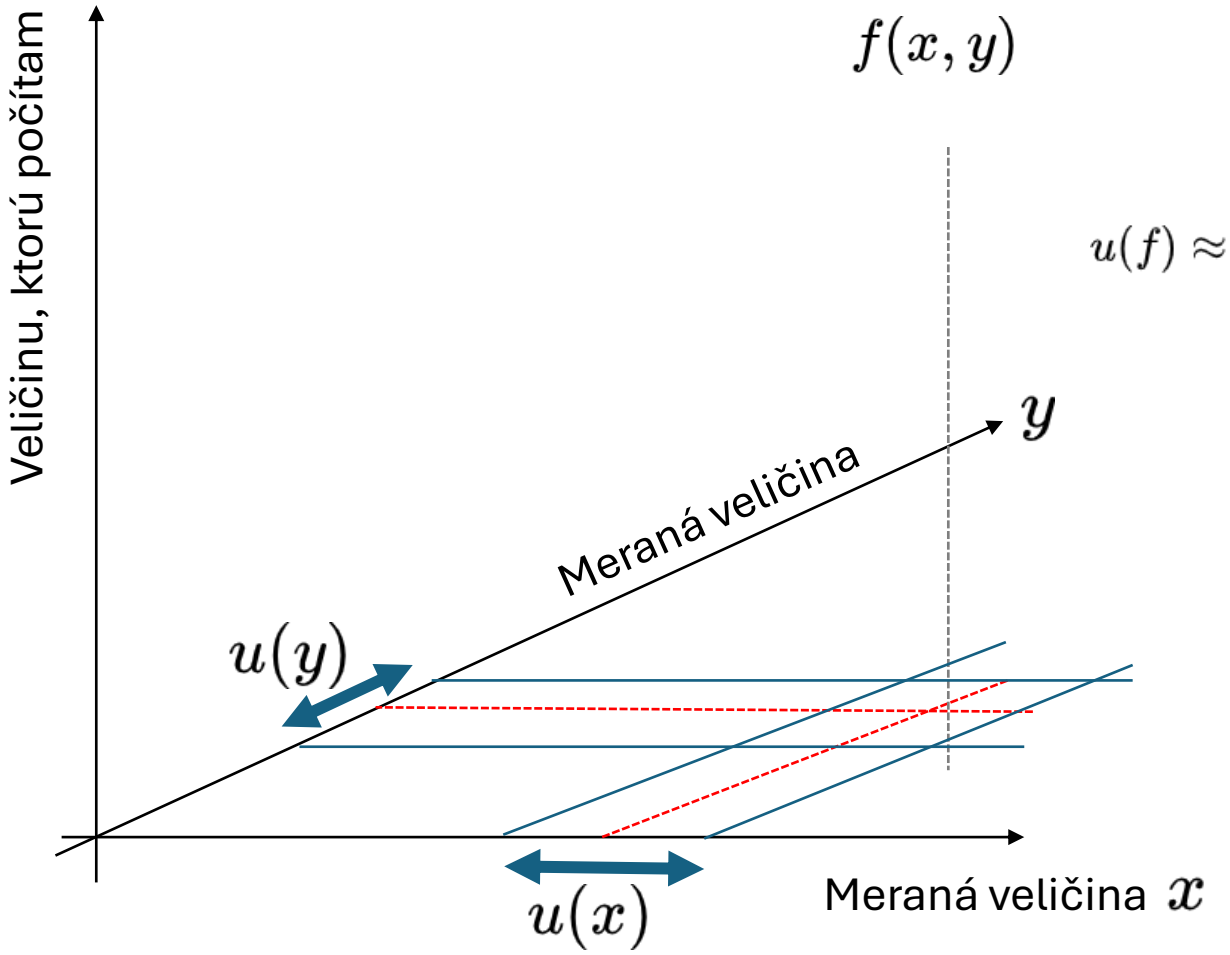
alebo

$$u^2(f) \approx \left|\frac{\partial f}{\partial x}\right|^2 u^2(x) + \left|\frac{\partial f}{\partial y}^2\right| u^2(y)$$

aj skladanie náhodnej a systematickej chyby

$$u(\tau)^2 = s(\overline{\tau})^2 + (\Delta\tau)^2$$

Šírenie chýb - funkcie viac (korelovaných) premenných



Korelované veličiny,
napr. teplomery v miestnosti, z ktorých zisťujem priemernú teplotu

$$u(f) \approx \left(\frac{\partial f}{\partial x}\right)^2 u^2(x) + \left(\frac{\partial f}{\partial y}\right)^2 u^2(y) + 2r\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right) u(x)u(y)$$

Korelačný koeficient

Korelácia

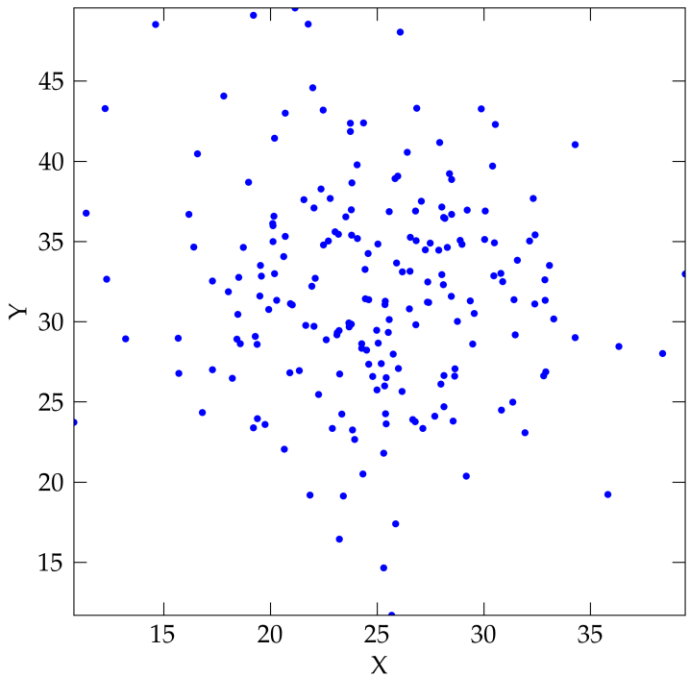
“Odmeria ako súvisí zmena jednej veličiny so zmenou druhej veličiny”

$$r = \frac{\sum_i ((x_i - \mu_x)(y_i - \mu_y))}{\sigma_x \sigma_y}$$

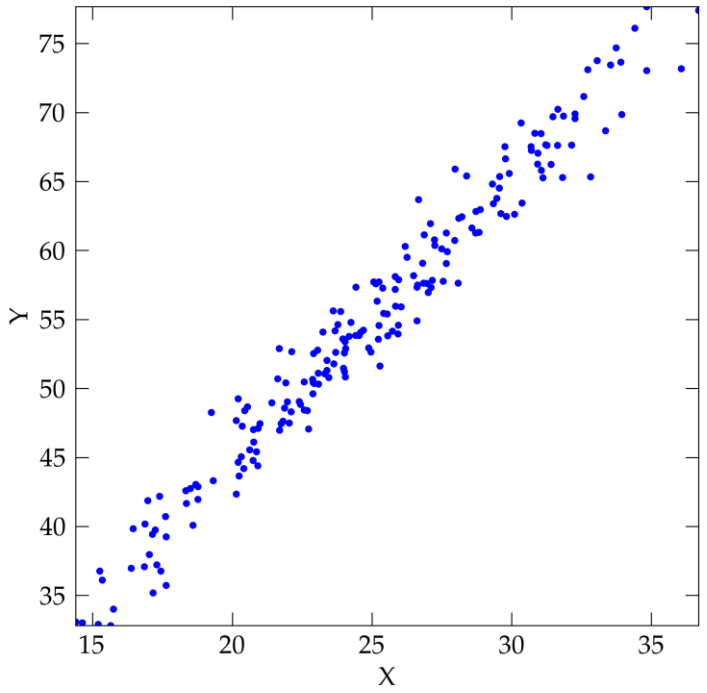
Pearsonov
korelačný koeficient

Vieme spočítať pre ľubovoľnú dvojicu premenných

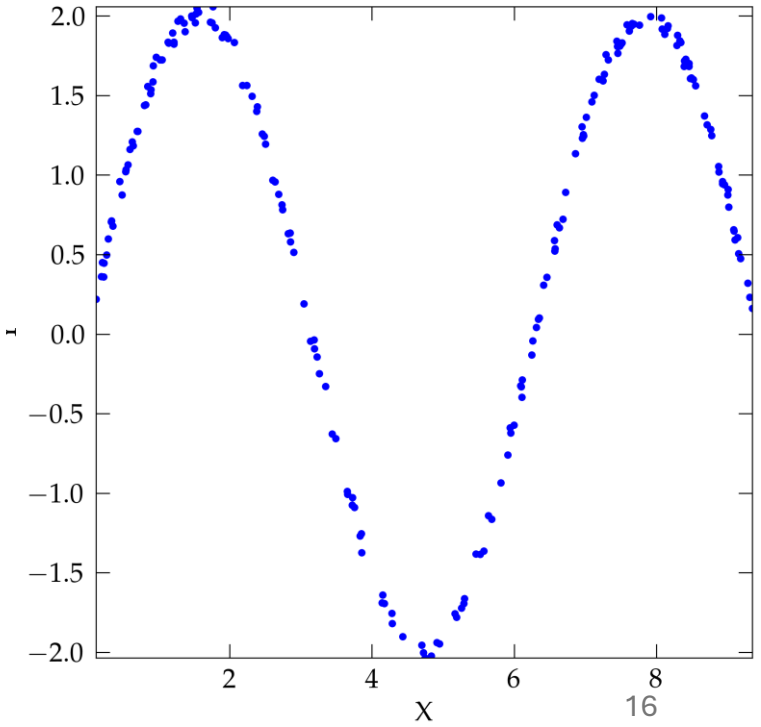
$R = -0.0826$



$R = 0.982$



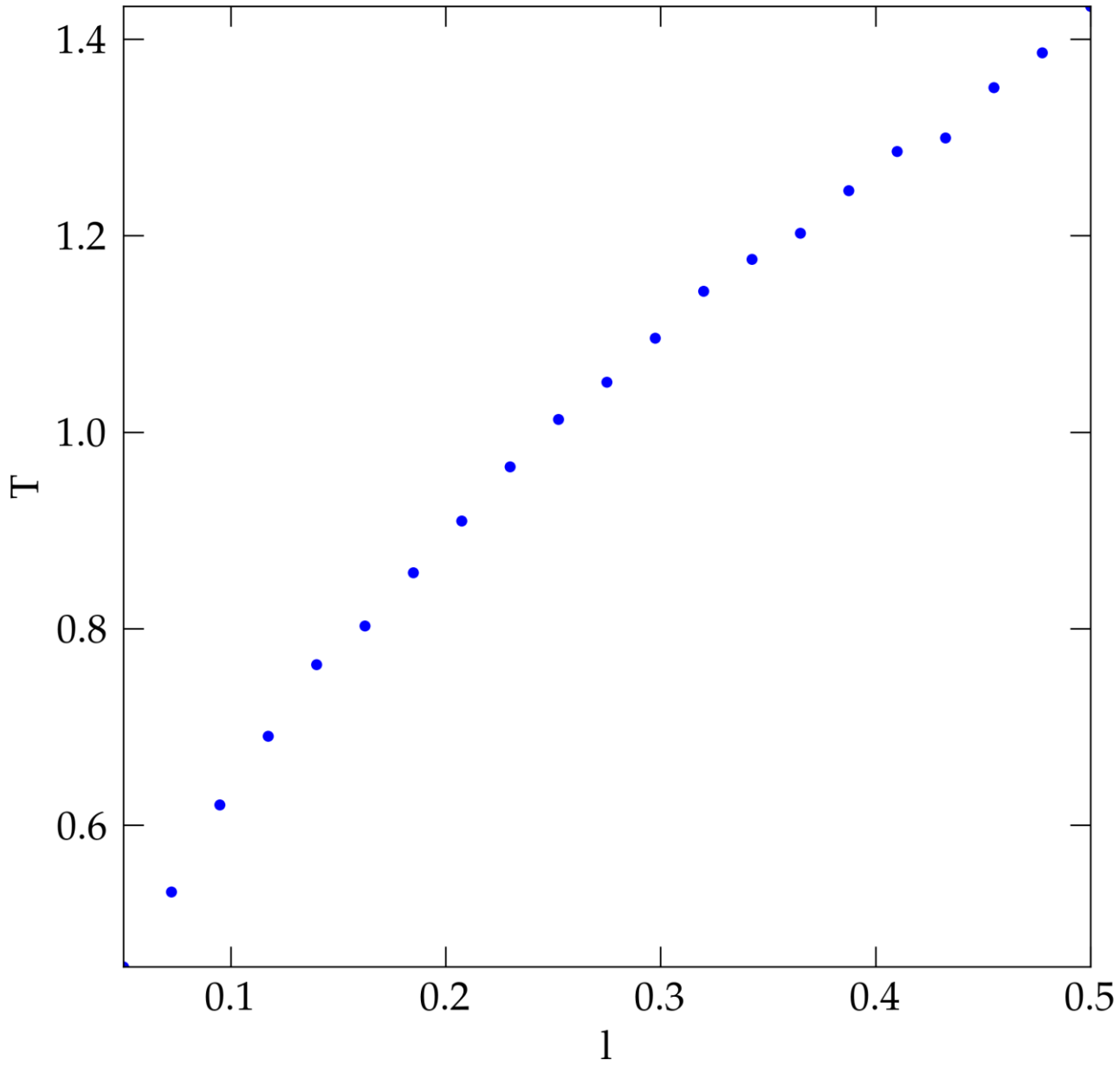
$R = -0.01176$



Porovnávanie exp dát a teoretického modelu

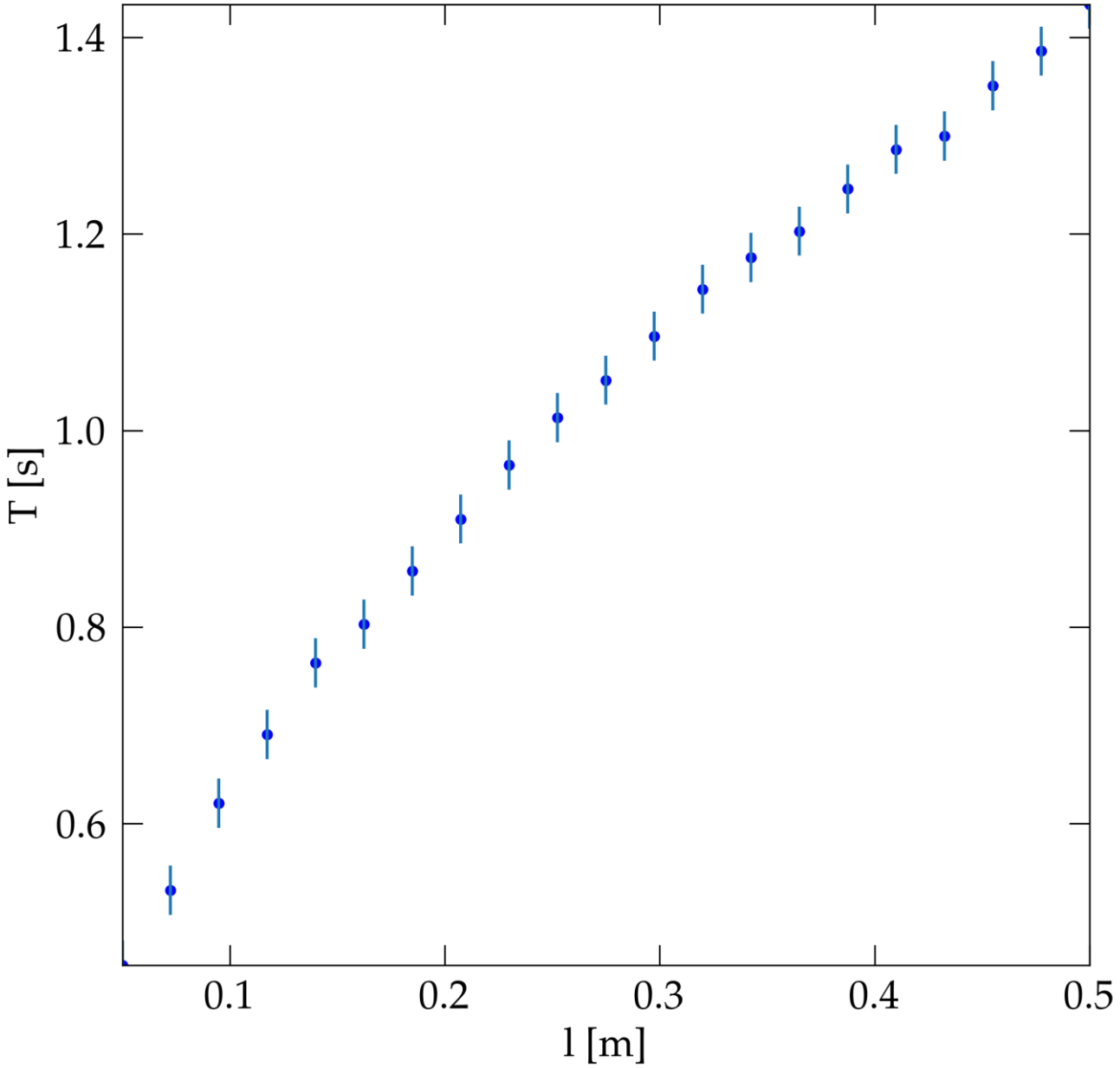
Ako na to? Vyniesť do grafu!

$$T = 2\pi\sqrt{\frac{l}{g}}$$



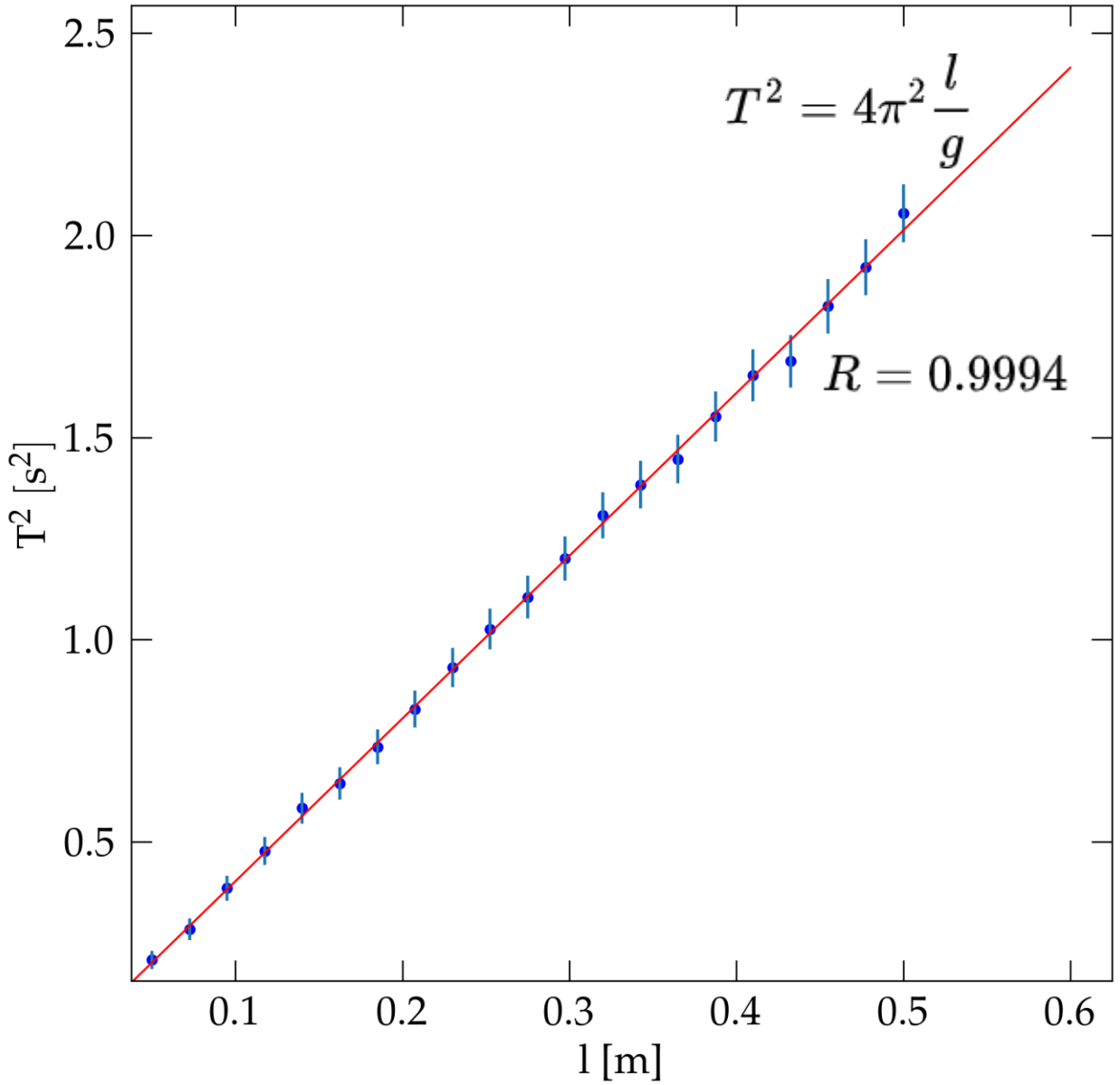
Ako na to? Vyniesť do grafu!

$$T = 2\pi\sqrt{\frac{l}{g}}$$



Ako na to? Vyniesť do grafu!

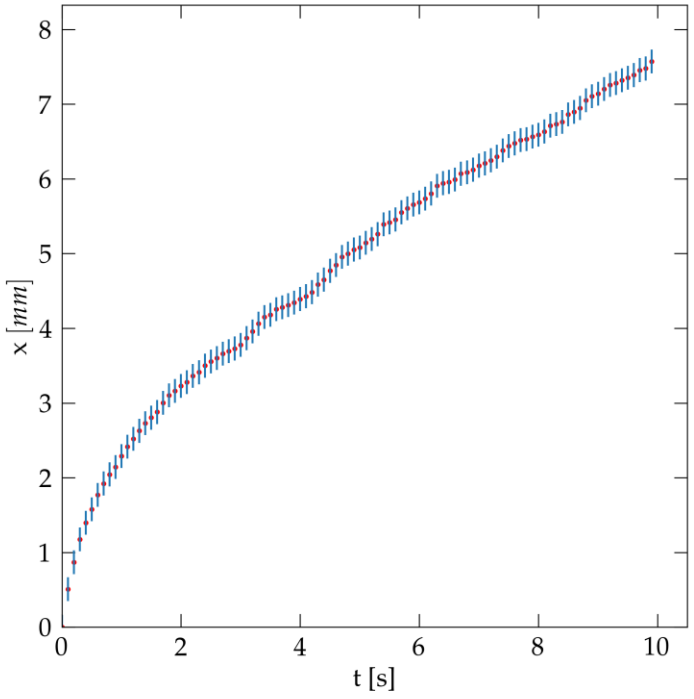
$$T = 2\pi\sqrt{\frac{l}{g}}$$

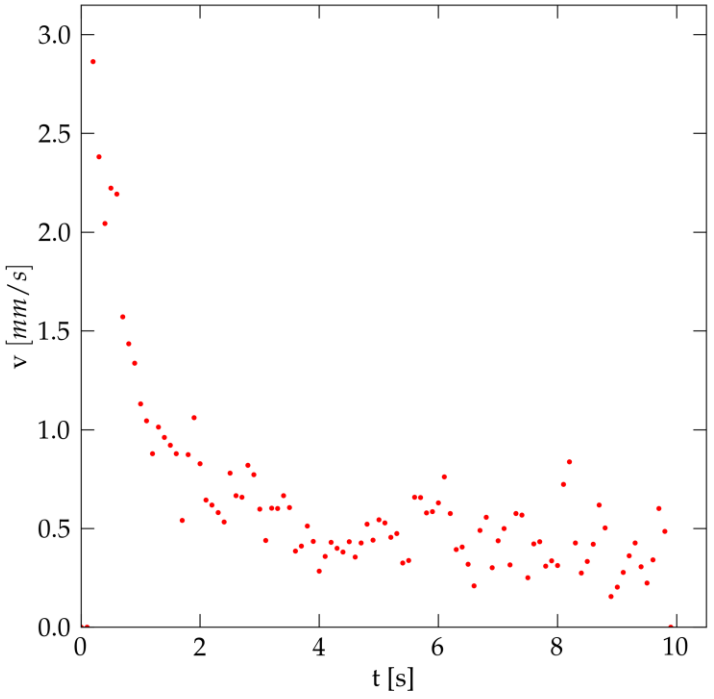


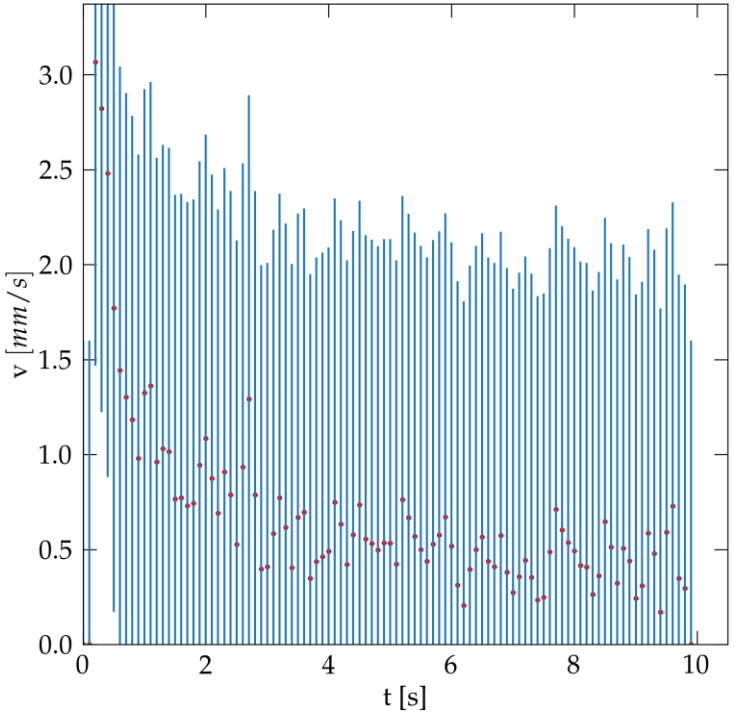
Čo by sa dalo urobiť ešte lepšie?

Porovnávať polohu alebo rýchlosť, ... ?

$$F = -A - Cv^2$$







Zisťovanie parametrov teoretického modelu z exp. dát

Metóda najmenších štvorcov

Častokrát chceme z nameranej závislosti x_i, y_i
Určiť parametre modelu, majúc na pamäti nejaký konkrétny model
napr. $y(x) = ax + b$

Chceme teda z nameraných dvojíc x_i, y_i
Zistiť optimálnu sadu parametrov a, b

Ak je model správny, tak $y_i = ax_i + b + \varepsilon_i$

Distribúciu chýb nepoznáme, ale predpokladáme, že je
z normálneho rozdelenia $\varepsilon \sim N(0, \sigma^2)$

Optimálne parametre minimalizujúce $L = \sum_i (y_i - (ax_i + b)))^2$

Metóda najmenších štvorcov

Optimálne parametre minimalizujúce $L = \sum_i (y_i - (ax_i + b)))^2$

$$\frac{\partial L}{\partial a} \stackrel{!}{=} 0; \frac{\partial L}{\partial b} \stackrel{!}{=} 0$$

Sústava lineárnych rovníc pre a, b

$$a = \frac{\sum_i x_i y_i - \frac{1}{N} \sum_i x_i \sum_i y_i}{\sum_i x_i^2 - \frac{1}{N} \left(\sum_i x_i^2\right)^2}$$

$$b = \frac{1}{N} \sum y_i - a \frac{1}{N} \sum x_i$$

Prečo najmenšie štvorce odchýlok?

Predpokladáme, normálne rozdelenie chýb v y-vej premennej

$$p(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{Exp}\left[-\frac{\epsilon^2}{2\sigma^2}\right]$$

$$\epsilon = y_{\mathrm{namerane}} - y_{\mathrm{model}}$$

Spomeňme si na metódu *maximum likelihood*
Pravdepodobnosť namerať N-ticu dát

$$p = p(\epsilon_1)p(\epsilon_2)\ldots p(\epsilon_N)\Delta\epsilon_1\ldots\Delta\epsilon_N \sim \mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_i (y_i - (ax_i + b_i))^2\right]$$

Hľadáme taký model, ktorý túto pravdepodobnosť maximalizuje

Ak majú vstupné dáta rôznu chybu?

Predpokladáme, normálne rozdelenie chýb v y-vej premennej

$$p(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{Exp}\left[-\frac{\epsilon^2}{2\sigma^2}\right]$$

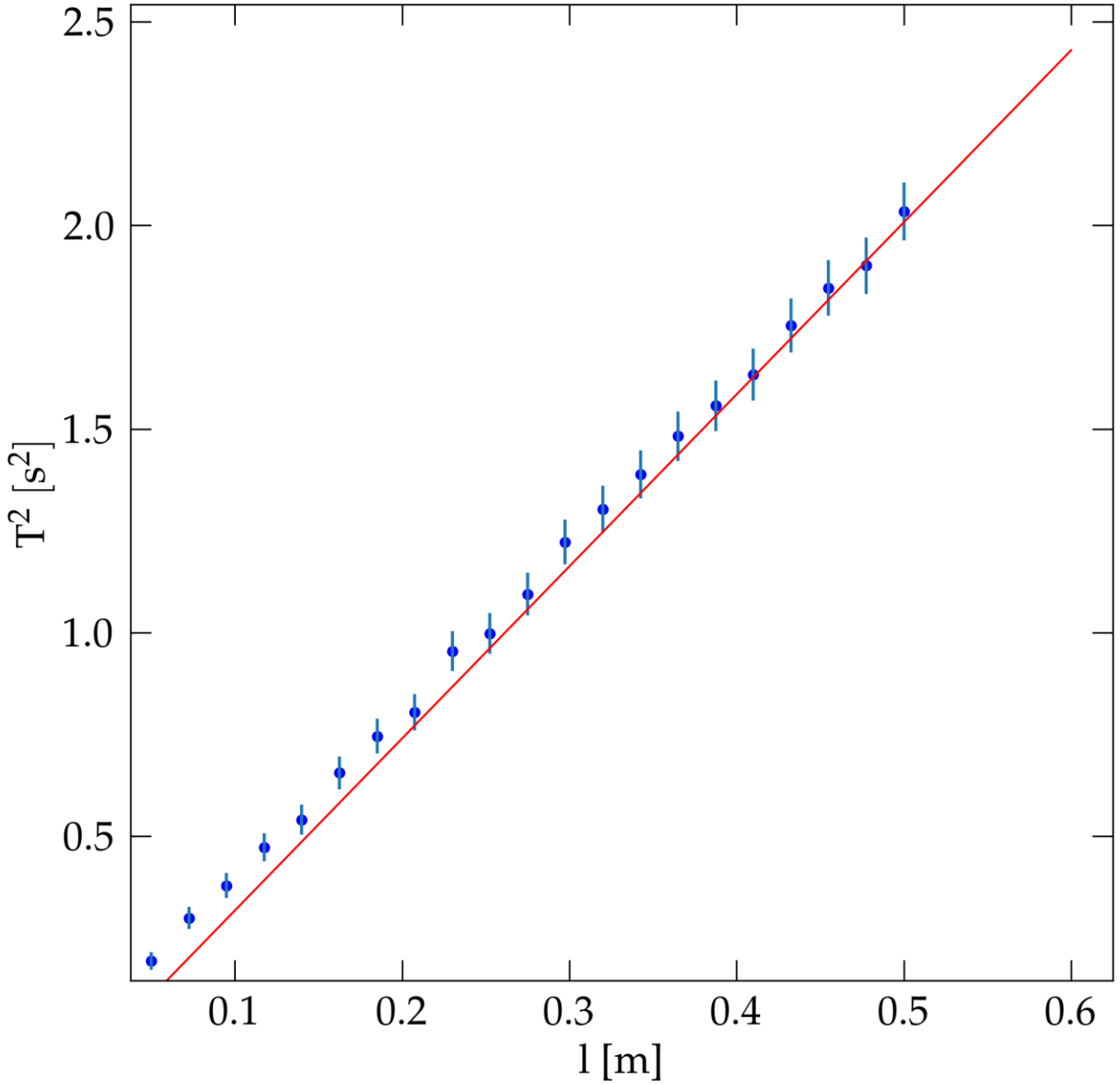
$$\epsilon = y_{\mathrm{namerane}} - y_{\mathrm{model}}$$

Odtiaľ vidíme, že ak dátové body nemajú rovnakú štandardnú odchýlku σ^2

$$\mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_i \left(y_i - (ax_i + b_i)\right)^2\right] \rightarrow \mathrm{Exp}\left[-\sum_i \left(\frac{y_i - (ax_i + b_i)}{2\sigma_i^2}\right)^2\right]$$

$$L = \sum_i \left(y_i - (ax_i + b_i)\right)^2 \rightarrow L = \sum_i \left(\frac{y_i - (ax_i + b_i)}{\sigma_i}\right)^2$$

Oponent mi ukáže takýto graf.



Odhad chyby parametrov fitu

Využitím pravidiel pre šírenie chýb a dosadením do rovníc pre koeficienty **a** a **b** (viď literatúra)

$$\sigma_a^2 = \frac{\sigma_y^2}{\sum_i (x_i - \overline{x})^2}$$

$$\sigma_b^2 = \sigma_y^2 \left[\frac{1}{N} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2} \right]$$

Je to užitočné?

$$\sigma_y^2 \approx u_{\text{res}}^2 = \frac{\sum_i (y_i - (ax_i + b))^2}{n - 2}$$

Hodnotenie kvality fitu

Využitím pravidiel pre šírenie chýb a dosadením do rovníc pre koeficienty **a** a **b** (viď literatúra)

$$R^2 = 1 - \frac{S_r}{S_b}$$

$$S_t = \sum_i (y_i - \overline{y})^2$$

$$S_r = \sum_i (y_i - (ax_i + b))^2$$

Príklad – meranie tiažového zrýchlenia analýzou voľného pádu

Tab. 12 Výpočet tíhového zrychlení regresní analýzou

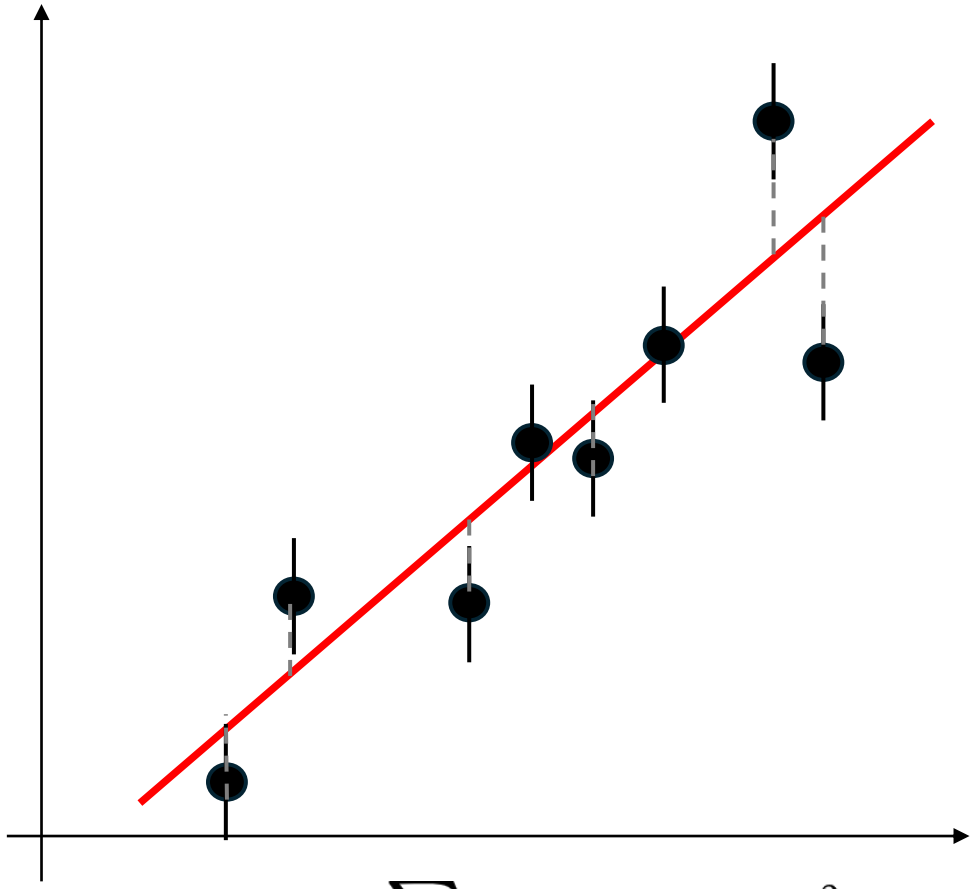
$\frac{t_i}{\mathrm{s}}$	$\frac{x_i}{\mathrm{s}^2}$	$\frac{y_i}{\mathrm{m}}$
0	0	0
0,90	0,405	4,0
1,03	0,53045	5,2
1,10	0,605	6,0
1,20	0,720	7,0
1,28	0,8192	8,0
1,35	0,91125	9,0
1,43	1,02245	10,0
1,50	1,125	11,0

$$\sum x_i^2 = 5{,}14232\ \mathrm{s}^4$$
$$\sum x_i y_i = 50{,}40269\ \mathrm{m\cdot s^{-2}}$$
$$\sum y_i^2 = 494{,}04\ \mathrm{m}^2$$
$$b^* = g = \frac{50{,}40269}{5{,}14232}\ \mathrm{m\cdot s^{-2}} = 9{,}8015\ \mathrm{m\cdot s^{-2}}$$
$$S_\mathrm{e} = \left(494{,}04 - \frac{50{,}40269^2}{5{,}14232}\right)\ \mathrm{m}^2 = 0{,}01568\ \mathrm{m}^2$$
$$s = \sqrt{\frac{0{,}01568}{8}}\ \mathrm{m} = 0{,}0443\ \mathrm{m} \doteq 0{,}05\ \mathrm{m}$$
$$s_{b^*} = \frac{0{,}0443}{\sqrt{5{,}142}}\ \mathrm{m\cdot s^{-2}} = 0{,}0195\ \mathrm{m\cdot s^{-2}} \doteq 0{,}02\ \mathrm{m\cdot s^{-2}}$$
$$\boxed{g = (9{,}80 \pm 0{,}02)\ \mathrm{m\cdot s^{-2}}}$$

Zdroj: Bohumil Vybíral **Zpracovaní dat fyzikálnych merení**

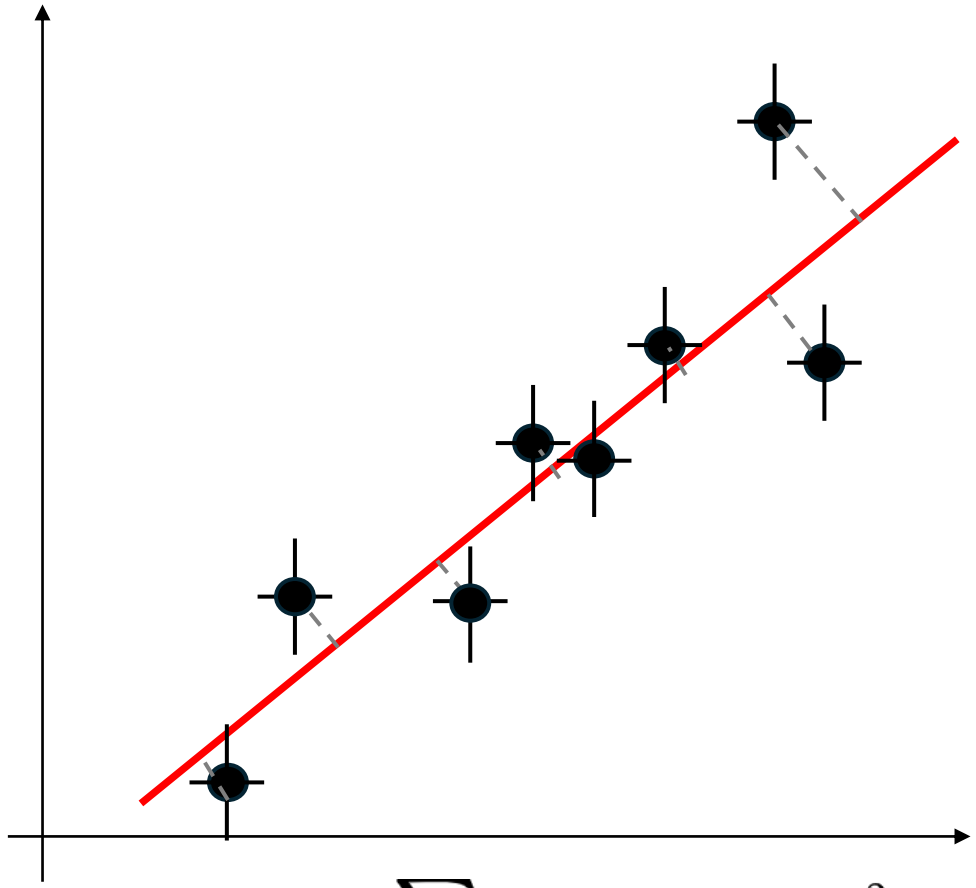
Chyba merania v meranej veličine vs chyba merania (aj) v vstupnej veličine

"Ordinary least squares"



$$\mathcal{L} = \sum_i (y_i - (ax_i + b))^2$$

"Total least squares"



$$\mathcal{L} = \sum_i (ax_i + by_i + c)^2$$

Tiež analyticky riešiteľný problém

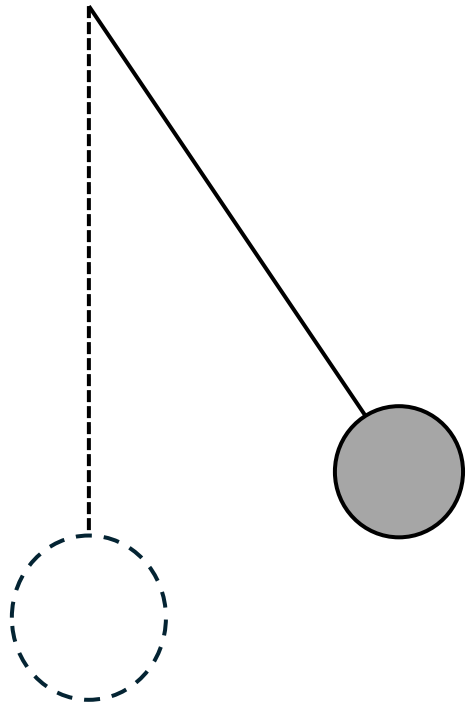
Fitovanie polynomiálnej závislosti a iné

Sústavu dvoch lineárnych rovníc vieme zovšeobecniť aj na sústavu lineárnych rovníc (viď vš. skriptá)

$$y_{\text{model}} = a + b_1 x + b_2 x^2 + \ldots + b_N x^N$$

Viac parametrov pri štúdiu na Matfyze

Linearizácia – pr. matematické kyvadlo

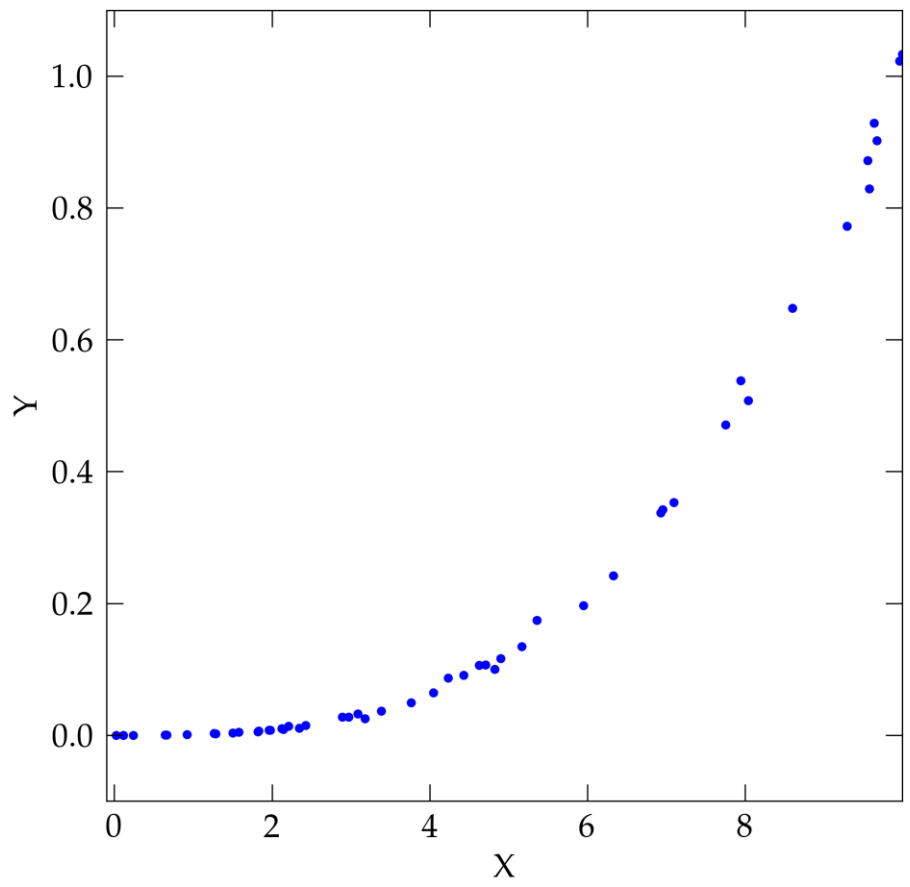


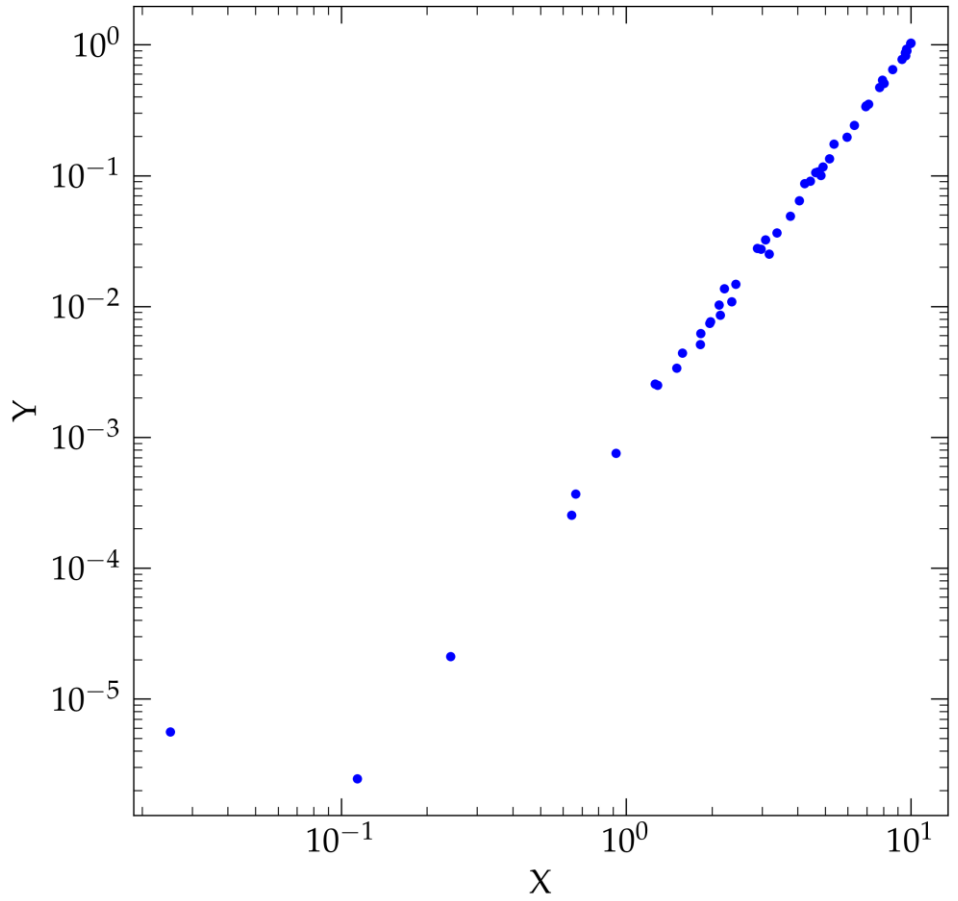
$$T = 2\pi\sqrt{\frac{l}{g}} \rightarrow T^2 = 4\pi^2\frac{l}{g}$$

Podobne vieme linearizovať aj iné vzťahy a zistiť meranú závislosť len analýzou dát

Linearizácia – log log vs log lin grafy, a pod.

$$y \sim Ax^n \rightarrow \log(y) \sim \log(A) + n\log(x)$$



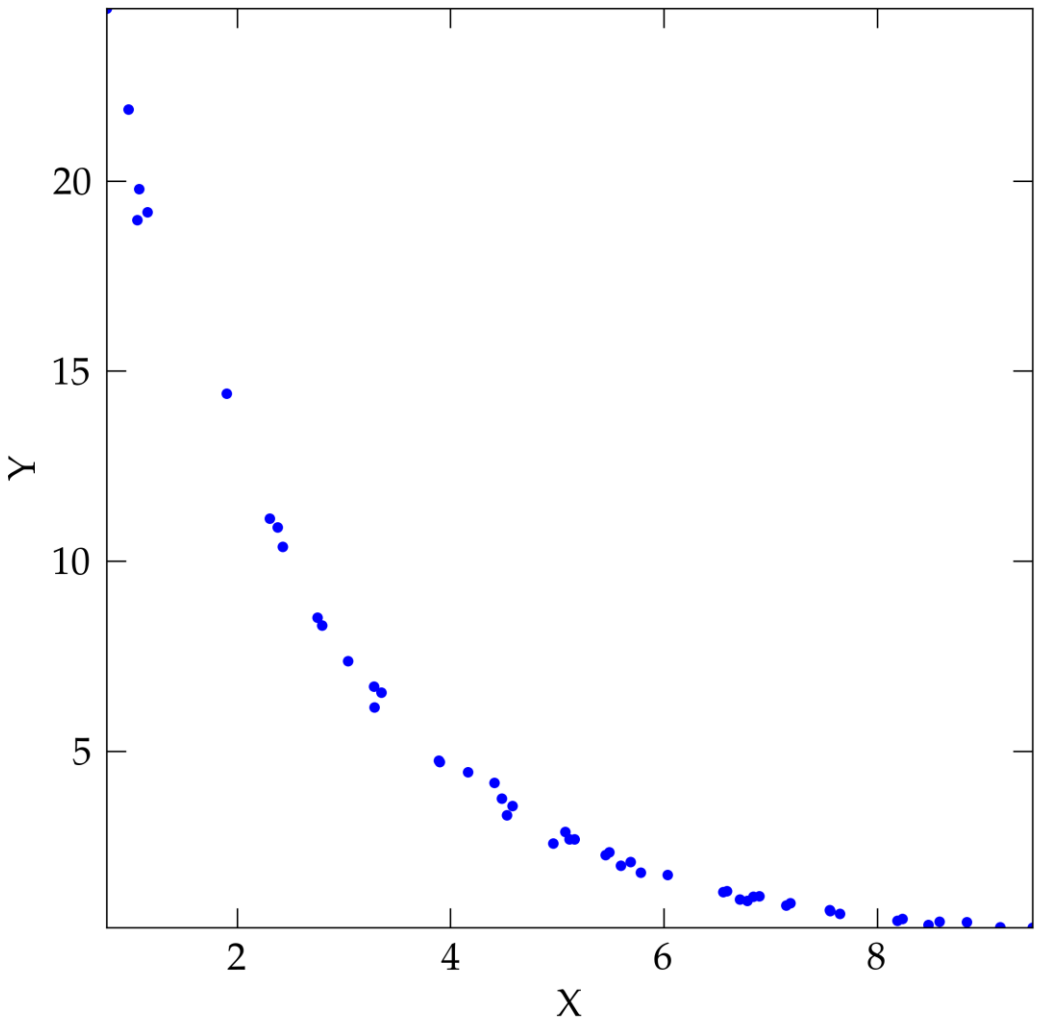


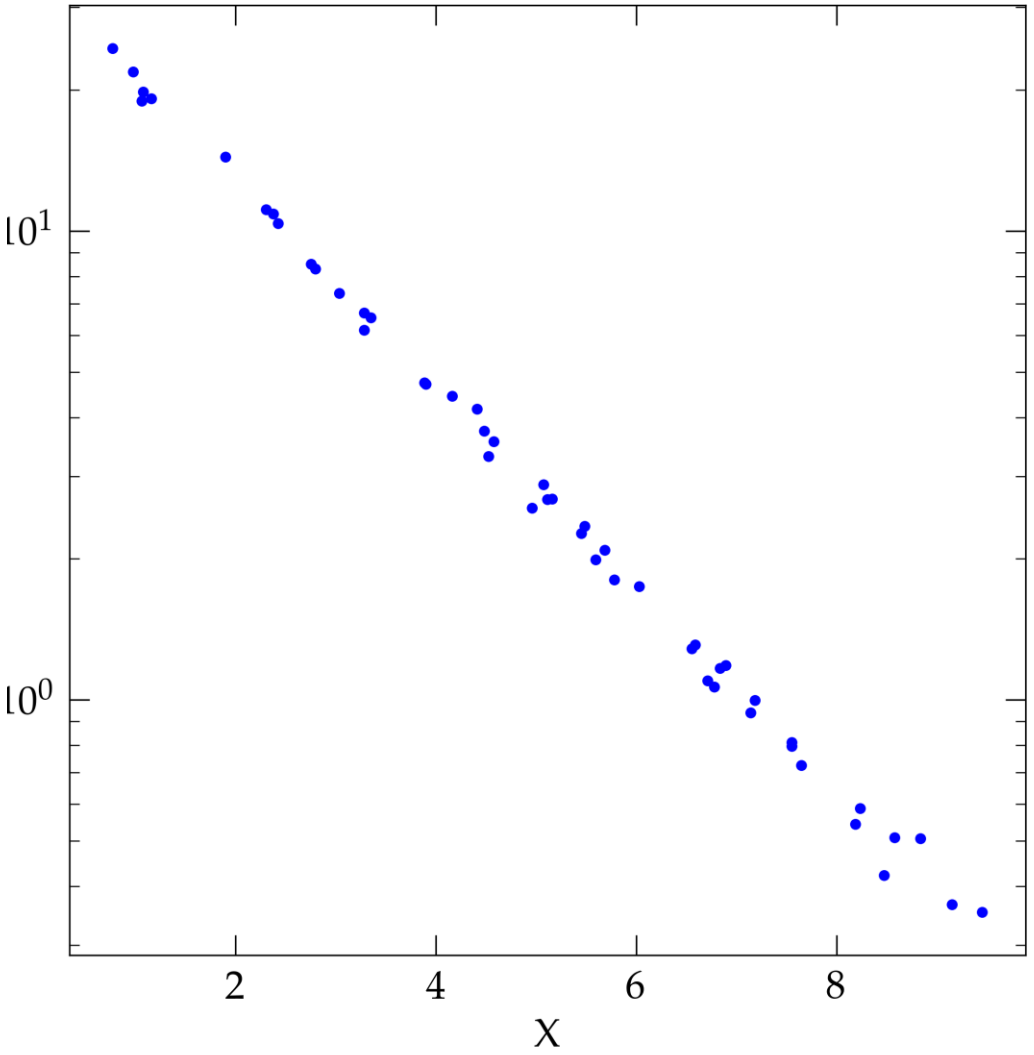
Pozor!

$$y \sim Ax^n + B \text{ nevedie na } \log(y) \sim \log(A) + n\log(x) + \log(B)$$

Linearizácia – log log vs log lin grafy, a pod.

$$y \sim A\mathrm{Exp}[-\alpha x] \Rightarrow \ln(y) = \ln(A) - \alpha x$$





Tichý nedostatok linearizácie

Problém

$$y_i = A\mathrm{Exp}[-\alpha x_i] + \epsilon_i \nRightarrow \ln(y_i) = \ln(A) - \alpha x_i + \ln(\epsilon_i)$$

Výsledná distribúcia chýb nemusí byť opísateľná normálovým rozdelením

Optimum, ktoré nájdeme regresiou na linearizovaných dátach nemusí byť skutočným optimom pre pôvodný model!

Nelineárna regresia

V praxi sa riešia nelineárne optimalizačné úlohy

$$L = \sum_i (y_i - f(x_i; \text{parametre modelu}))^2$$

Rôzne algoritmy na optimalizáciu týchto úloh, ktoré rôznym spôsobom „kvadraticky“ aproximujú (viď napr. **Algorithms for Optimization**)

- Konjugované gradienty
- Broyden–Fletcher–Goldfarb–Shanno algorithm
- …

Implementované v knižniciach
Vstupné dáta môžu byť takisto usporiadaná n-tica t.j. vektor!

Ako si poradiť so šumom?

Čo ak je v dátach veľa šumu? Robustifikácia

Niekedy pravdepodobnostné rozdelenie obsahuje tzv. „ťažké chvosty“, t.j. pravdepodobnostné rozdelenie klesá pomalšie ako $p(x) \propto \exp(-x^2)$, pr. $p(x) \propto \exp(-|x|)$. Vtedy prestáva byť splnený predpoklad o normálnom rozdelení, ktorý sme si ukázali pri *maximum likelihood estimation,* dá sa ale stále použiť táto metóda.

Príklady zdroja outlierov: *počítačové spracovanie obrazu – hot pixels, EM šum, ...*

„Kuchynský recept“:

Štatistický súbor

Priemer → Medián

Smerodajná odchýlka → „Median absolute deviation“

$$\hat{\sigma} = \frac{1}{0.674}\mathrm{MED}\left[|x - \mathrm{MED}[\mathrm{x}]|\right]$$

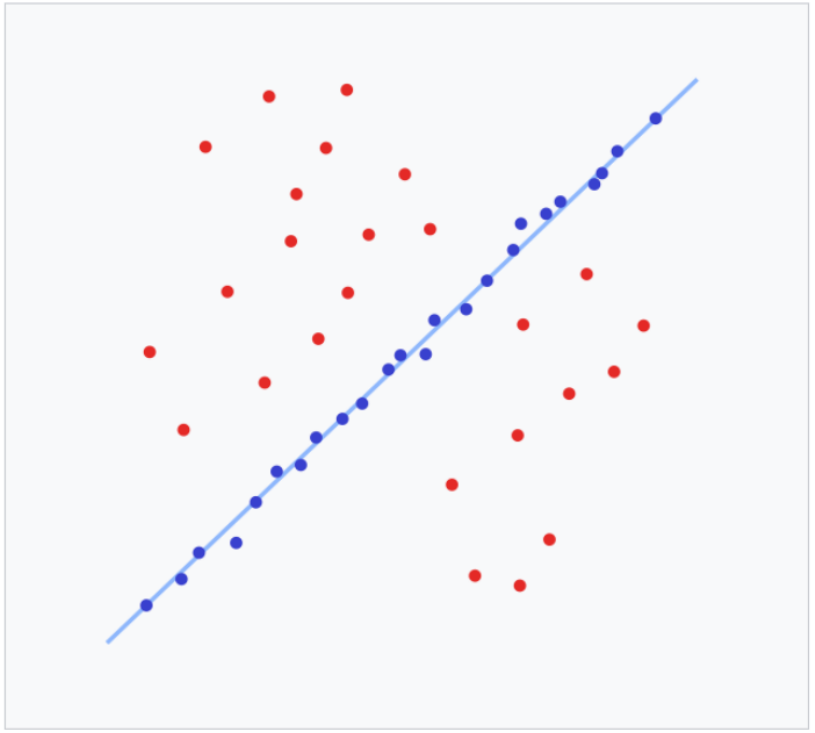
Regresia

Minimalizácia súčtu najmenších štvorcov odchýlok → Minimalizácia súčtu absolútnych rozdielov odchýlok (tzv. L1 norma)

Pozor už nie je analyticky riešiteľný problém, ale existujú rýchle algoritmy! [DOPLNIT]

RANSAC – Ako odfiltrovať šum

- Vyberieme **náhodnú** podčasť originálnych dát
- Získame model na tejto podčasti dát
- Otestujeme model voči všetkým dátam a vyberieme tie,
  ktoré sú dostatočne blízko dátam tieto nazväme tzv. *inlieri*
- Získame model voči všetkým dátam, ktoré sú dostatočne blízko



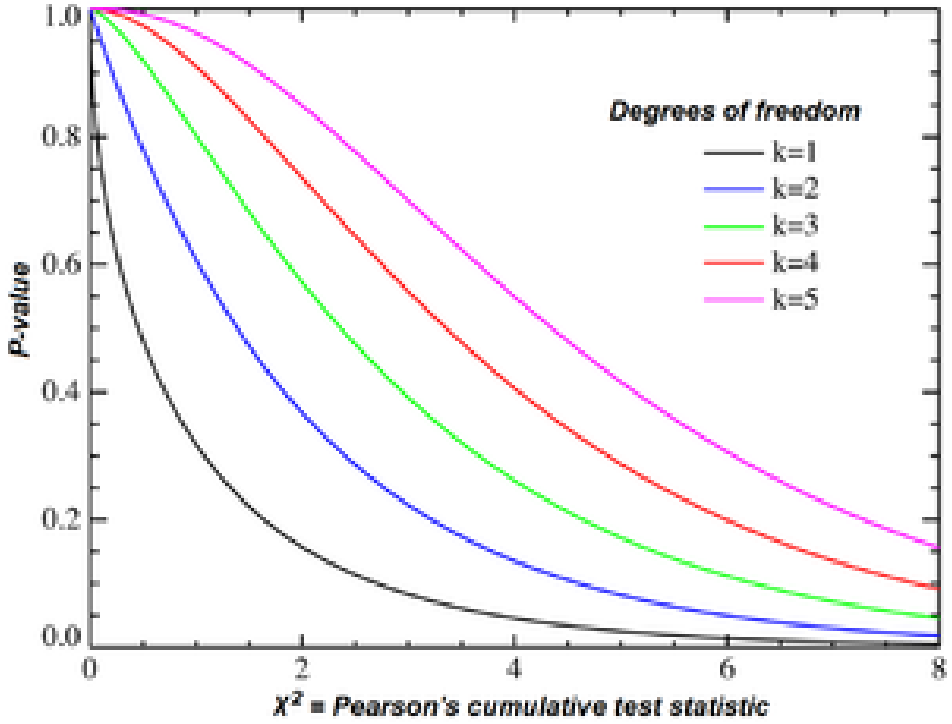
Testovanie štatistických hypotéz

Testovanie štatistických hypotéz

- Chi kvadrát – používaný na porovnanie či pozorovaný štatistický súbor prislúcha teoretickému modelu

$$\chi^2 = \sum_{\text{pocet kategorii}} \frac{(O_i - E_i)^2}{E_i}$$

Porovnávame s tabulovanými dátami, viď numerické knižnice v Pythone a porovnáme hodnotu koeficientu k prislúchajúcej p-hodnote



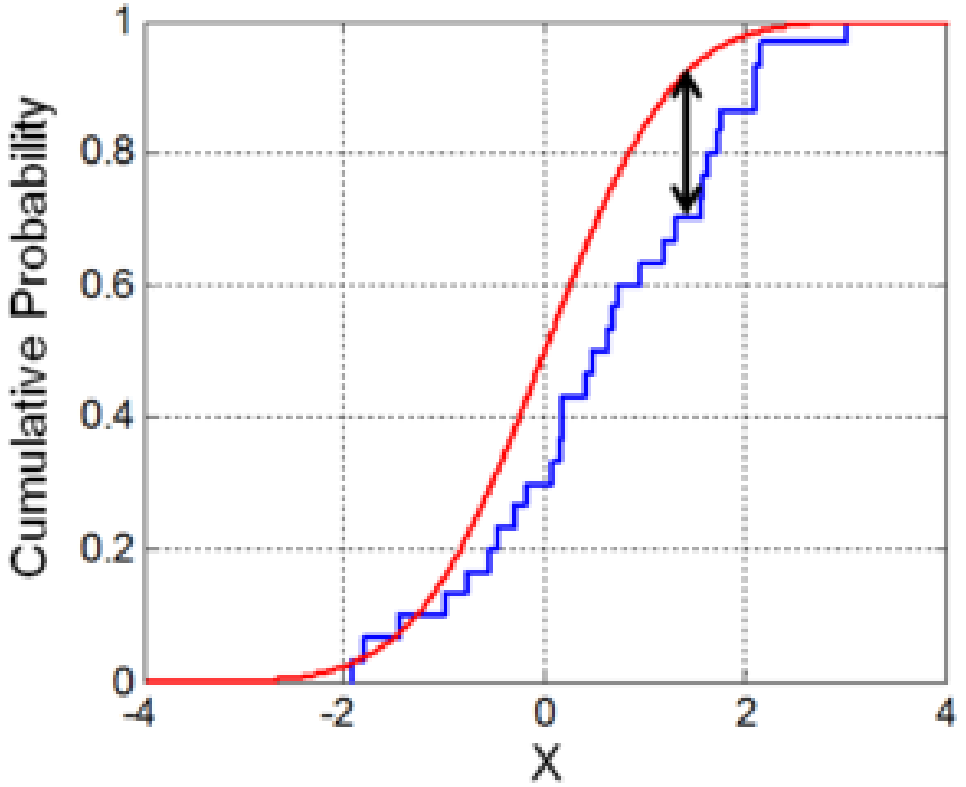
Zdroj: https://en.wikipedia.org/wiki/Pearsons_chi-squared_test

Testovanie štatistických hypotéz

- Kolmogorov Smirnoff test – porovnanie či distribúcia testovanej vzorky pochádza z danej distribúcie alebo nie

$$CDF[x] = \frac{\text{pocet merani s hodnotou} \leq \text{x}}{N}$$

$$D_n = \max_x |CDF_n(x) - CDF(x))|$$



Zdroj: https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test

Ako na to prakticky … napr. v Pythone

Užitočné knižnice

Kde sa dozvedieť viac?

- Doc. RNDr. František Kundracik, CSc.
  Spracovanie experimentálnych dát – skriptá FMFI UK

- Studijní texty (českej) FO
  Bohumil Vybíral – **Zpracovaní dat fyzikálnych merení**

Pre pokročilých:

- W.H. Press, S.A. Teukolsky, W. T. Vetterling, B.P. Flannery – **Numerical Recipes: The Art of Scientific Computing**, Cambridge University Press

- M. J. Kochenderfer, T. A. Wheeler, **Algorithms for Optimization** - MIT Press Cambridge, Massachusetts; London, England. 2019

- R.C. Maronna, R.D. Martin, V.J.Yohai, **Robust statistics**. John Wiley & Sons, 2006