Matematická statistika

Náhodná/stochastická proměnná přiřazuje pravděpodobnost/hustotu pravděpodobnosti možnému diskrétnímu/spojitému jevu z diskrétní/spojité množiny jevů.

- diskrétní příklad: hod kostkou: $p_i = 1/6$ pro $i \in \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$
- spojitý příklad: čas rozpadu jádra: $p(t) = ke^{-kt}$

Spojitou náhodnou veličinu v 1D (tj. $\mathbf{x} \in \mathbb{R}$) popisuje **distribuční funkce** (hustota pravděpodobnosti, rozdělení/rozložení pravděpodobnosti) $p(x)$:

$p(x)dx$ je pravděpodobnost, že nastane jev $\mathbf{x} \in [x, x + dx)$

Ve dvou dimenzích definujeme hustotu pravděpodobnosti $p(x, y)$ tak, že jev $x \in [x + dx)$ a zároveň $y \in [y + dy)$ nastane s pravděpodobností $p(x, y)dxdy$.

Normalizace:

$$\sum_i p_i = 1 \quad \text{nebo} \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

Kumulativní (integrální) distribuční funkce = pravděpodobnost, že padne náhodná hodnota $\mathbf{x} \leqslant x$:

$$P(x) = \int_{-\infty}^{x} p(x')dx'$$

Rozdělení pravděpodobnosti

Varování. Ve fyzice a technice nepřesně a volně zaměňujeme symbol $\mathbf{x}$ pro náhodnou veličinu a x pro její hodnotu (např. při integraci).

Střední hodnota (též *expectation value*, očekávaná hodnota; slovo průměr budeme rezervovat pro aritmetický průměr, tj. střední hodnotu výběru)

$$\mathrm{E}(\mathbf{x}) \equiv \langle \mathbf{x} \rangle \equiv \langle x \rangle_{\mathbf{x}} \overset{\text{volně}}{=} \langle x \rangle = \int x p(x) \mathrm{d}x \quad \text{nebo} \quad \sum_i x_i p_i$$

Příklad. Když hodíte na kostce ⚄, vyhrajete 5 Kč; pokud padne něco jiného, prohrajete 1 Kč. Je tato hra spravedlivá?

Ano – střední výhra je 0

Variance (též: rozptyl, fluktuace, disperze, střední kvadratická odchylka (*MSD*)

$$\mathrm{Var}(\mathbf{x}) \overset{\text{volně}}{=} \mathrm{Var}\, x = \langle (x - \langle x \rangle)^2 \rangle = \langle \Delta x^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2, \quad \text{kde } \Delta x = x - \langle x \rangle$$

Směrodatná odchylka (*standard deviation*) $= \sqrt{\mathrm{Var}(\mathbf{x})}$, ozn. $\sigma(\mathbf{x})$, δx

Příklad. Mějme rovnoměrné rozdělení $\mathbf{u}$ v intervalu $[0, 1)$; na počítači např. `rnd(0)`. Vypočtěte střední hodnotu a varianci.

$\langle \mathbf{n} \rangle = 1/2$, $\mathrm{Var}(\mathbf{n}) = 1/12$

Příklad: Giniho koeficient

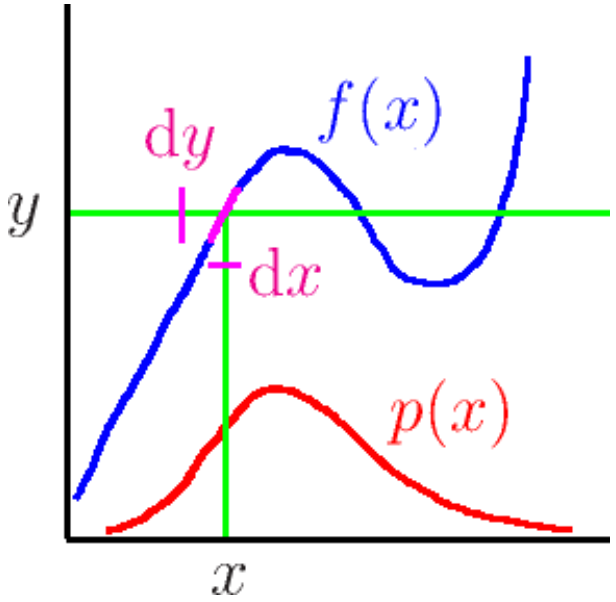
Míra nerovnosti příjmu. Příjem x s hustotou pravděpodobnosti $p(x)$, $x \geqslant 0$.

$$G = \frac{1}{2\langle x \rangle} \int_0^\infty p(x)dx \int_0^\infty p(y)dy\,|x-y|$$

Funkce náhodné veličiny

Mějme reálnou náhodnou veličinu $\mathbf{x}$ s rozdělením $p(x)$ a reálnou funkci $f(x)$. Veličina (pozorovatelná) $f(\mathbf{x})$ má rozdělení (sčítá se přes všechny kořeny):

$$p_f(y) = \sum_{x:f(x)=y} \frac{p(x)}{|f'(x)|}$$



Příklad. Mějme rovnoměrné rozdělení u v intervalu $[0,1)$. Jaké rozdělení má $t = -\ln u$?

exp(−t): např. čas rozpadu 1 atomu s k = 1

Funkce náhodné veličiny: střední hodnota

Střední hodnota veličiny $\mathbf{f}$:

$$\langle f \rangle = \int f(x)p(x)dx$$

nebo z nové náhodné proměnné $\mathbf{f} = f(\boldsymbol{x})$:

$$\langle f \rangle = \int y p_f(y) dy$$

Obě střední hodnoty jsou stejné:

$$\langle f \rangle = \int f(x)p(x)dx \overset{\text{subst. } y=f(x)}{=} \int \frac{yp(x)}{f'(x)}dy = \int y p_f(y) dy$$

kde v 2. integrálu $x =$ řešení rovnice $f(x) = y$, které zde pro jednoduchost uvažujeme jen jedno a také předpokládáme, že funkce f je rostoucí.

Kovariance

- Kovariance x a y dvojrozměrného rozdělení $p(x, y)$

$$\mathrm{Cov}\,(\mathbf{x}, \mathbf{y}) = \langle \Delta x \Delta y \rangle = \int \Delta x \Delta y p(x, y) dx dy$$

- Kovariance dvou veličin $f(x)$ a $g(x)$ (obdobně u diskrétného či víceroz-měrného rozdělení):

$$\mathrm{Cov}\,(f, g) = \langle \Delta f \Delta g \rangle = \int \Delta f \Delta g\, p(x) dx$$

Nezávislé náhodné veličiny

Náhodné veličiny $\mathbf{x}$ (s rozdělením $p_1(x)$) a $\mathbf{y}$ (s rozdělením $p_2(y)$):

$$p(x, y) = p_1(x) p_2(y)$$

V diskrétním případě (např. dva hody kostkou, $p_{ij} = 1/36$):

$$p_{ij} = p_{1,i} p_{2,j}$$

Kovariance nezávislých náhodných veličin je nula:

$$\mathrm{Cov}\,(\mathbf{x}, \mathbf{y}) = \langle \Delta x \Delta y \rangle_{\mathbf{x}+\mathbf{y}} = \int dx \int dy\, \Delta x p_1(x) \Delta y p_2(y) = \langle \Delta x \rangle_{\mathbf{x}} \langle \Delta y \rangle_{\mathbf{y}} = 0$$

Korelační koeficient

$$r(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}\,(x)\mathrm{Var}\,(y)}}$$

Příklad. Nechť $\mathbf{u}_1$ a $\mathbf{u}_2$ jsou dvě nezávislá rovnoměrný rozdělení v [0,1].
Calculate:

Vypočtěte: a) $r(u_1, -u_1)$

b) $r(u_1^2, u_1^2)$

c) $r(u_1, u_2 + u_1)$ (viz Maple)

a) -1, b) 1, c) $1/\sqrt{2}$

```
tab 1 100000 | tabproc "rnd(0)" "rnd(0)" | tabproc A A+B | lr
```

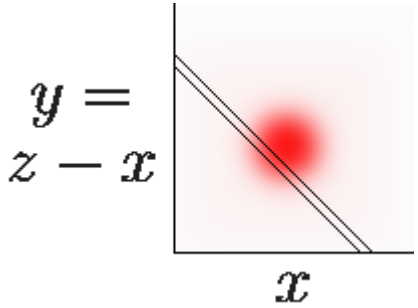
Součet náhodných proměnných

Nechť $\mathbf{x}$ a $\mathbf{y}$ jsou dvě spojité náhodné proměnné s rozdělením $p(x,y)$. Rozdělení součtu $\mathbf{x}+\mathbf{y}$ je

$$p_{\mathbf{x}+\mathbf{y}}(z)dz = \iint_{x+y\in(z,z+dz)} p(x,y)dxdy \overset{y:=z-x}{=} \int p(x,z-x)dxdz$$

$\Rightarrow$

$$p_{\mathbf{x}+\mathbf{y}}(z) = \int p(x,z-x)dx$$



Nechť nyní $p(x,y) = p_1(x)p_2(y)$. Pak

$$p_{\mathbf{x}+\mathbf{y}}(z) = \int p_1(x)p_2(z-x)dx \equiv (p_1 * p_2)(z)$$

$p_1 * p_2$ se nazývá **konvoluce**

Diskrétní příklad: Hodíme dvojicí kostek. Jaké rozdělení má součet ok?

$p(2) = 1/36, p(3) = 2/36, \ldots p(7) = 6/36, \ldots p(12) = 1/36$

Příklad. Vypočtěte rozdělení $\mathbf{u}_1 - \mathbf{u}_2$

0 for $|x| > 1$, $1 - |x|$ otherwise

```
tab 1 100000 | tabproc "rnd(0)-rnd(0)" | histogr -1.5 1.5 .1 | plot -
```

Součet nezávislých náhodných proměnných

Střední hodnota i variance součtu nezávislých náhodných veličin jsou aditivní. S epickou šíří:

$$\mathsf{E}\left(\mathbf{x}+\mathbf{y}\right) = \int z p_{\mathbf{x}+\mathbf{y}}(z)\mathrm{d}z = \int z p_1(x) p_2(z-x)\mathrm{d}x\mathrm{d}z$$

$$\overset{y:=z-z}{=} \int (x+y) p_1(x) p_2(y)\mathrm{d}x\mathrm{d}y = \langle x\rangle_1 + \langle y\rangle_2 = \mathsf{E}\left(\mathbf{x}\right) + \mathsf{E}\left(\mathbf{y}\right)$$

Přímo:

$$\mathsf{E}\left(\mathbf{x}+\mathbf{y}\right) = \int p_1(x) p_2(y)(x+y)\mathrm{d}x\mathrm{d}y$$

$$= \int p_1(x) p_2(y) x\mathrm{d}x\mathrm{d}y + \int p_1(x) p_2(y) y\mathrm{d}x\mathrm{d}y = \int p_1(x) x\mathrm{d}x + \int p_2(y) y\mathrm{d}y = \mathsf{E}\left(\mathbf{x}\right) + \mathsf{E}\left(\mathbf{y}\right)$$

$$\mathsf{Var}\left(\mathbf{x}+\mathbf{y}\right) = \langle(\Delta x+\Delta y)^2\rangle_{\mathbf{x}+\mathbf{y}}$$

$$= \langle(\Delta x)^2\rangle_{\mathbf{x}+\mathbf{y}} + 2\langle\Delta x\Delta y\rangle_{\mathbf{x}+\mathbf{y}} + \langle(\Delta y)^2\rangle_{\mathbf{x}+\mathbf{y}} = \mathsf{Var}\left(\mathbf{x}\right) + \mathsf{Var}\left(\mathbf{y}\right)$$

Centrální limitní věta

Součet n stejných nezávislých rozdělení s konečnou střední hodnotou a konečnou variancí je pro velké n rovno Gaussově rozdělení se střední hodnotou $n\langle x\rangle$ a variancí $n\mathrm{Var}\, x$.

Příklad. Uvažujme diskrétní rozdělení $\mathbf{b}$: $p(-1/2) = p(1/2) = 1/2$. Aproximujte součet n takových rozdělení.

$$\begin{aligned}
n = 1 \quad & p(-1/2) = 1/2,\ p(1/2) = 1/2,\ \mathrm{Var}\,\mathbf{b} = 1/4 \\
n = 2 \quad & p(-1) = 1/4,\ p(0) = 1/2,\ p(1) = 1/4,\ \mathrm{Var}\,\mathbf{b}^2 = 2/4 \\
n = 3 \quad & p(\pm 3/2) = 1/8,\ p(\pm 1/2) = 3/8,\ \mathrm{Var}\,\mathbf{b}^3 = 3/4
\end{aligned}$$

Pro jednoduchost uvažujme jen sudé n. Pak pro $k = -n/2 \,..\, n/2$:

$$p(k) = \binom{n}{n/2+k} 2^{-n} \approx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{k^2}{2\sigma^2}\right), \quad \sigma^2 = \mathrm{Var}\,(\mathbf{b}^n) = \frac{n}{4}$$

Důkaz: potřebujeme Stirlingův vzorec ve tvaru $n! \approx n^n e^{-n}/\sqrt{2\pi n}$

Ověření centrální limitní věty

$$\binom{n}{\frac{n}{2}+1} = \frac{n!}{(\frac{n}{2}-1)!(\frac{n}{2}+1)!} = \frac{n!}{(\frac{n}{2})!/(\frac{n}{2})\cdot(\frac{n}{2})!(\frac{n}{2}+1)} = \binom{n}{\frac{n}{2}} \times \frac{\frac{n}{2}}{\frac{n}{2}+1}$$

$$\ln p(\frac{n}{2},1) = \ln p(\frac{n}{2},0) + \ln \frac{\frac{n}{2}}{\frac{n}{2}+1} \approx \ln p(\frac{n}{2},0) - \frac{2}{n}$$

Další člen

$$\ln p(\frac{n}{2},2) = \ln p(\frac{n}{2},1) + \ln \frac{\frac{n}{2}-1}{\frac{n}{2}+2} \approx \ln p(\frac{n}{2},1) - \frac{6}{n}$$

a obecně

$$\ln p(n,k) \approx \ln p(n,0) - 2\sum_{j=1}^{k} \frac{2k-1}{n}, \quad \sum_{j=1}^{k}(2k-1) \approx \int_0^k (2k-1)dk = k(k-1) \approx k^2$$

Obdobně pro záporná k. V limitě velkých k a n tedy

$$p(n,k) \approx p(n,0)\exp\left(-\frac{k^2}{n/2}\right)$$

Po normalizaci dostaneme kýžené

Matematická statistika a metrologie

Názvosloví kolísá podle oboru...

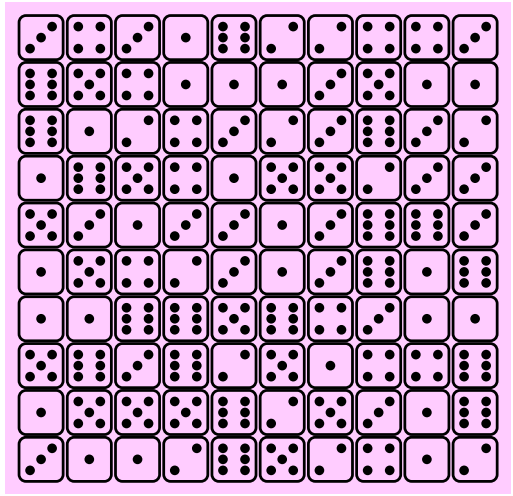
Statistika, odhad, „statistický algoritmus“, (úžeji) „statistický funkcionál, (v metrologii „měřicí funkce“, *measurement function*) je vzorec/algoritmus, podle kterého počítáme výsledek z (vzorku) náhodných veličin (v metrologii z dat). Statistika je také náhodnou veličinou.

Příklady: aritmetický průměr, parametry modelu při fitování metodou nejmenších čtverců.

Standardní chyba statistiky = směrodatná (standardní) odchylka (odmocnina variance) rozdělení (rozdělovací funkce) této statistiky.

Nejistota (*uncertainty*) v metrologii zahrnuje kritické posouzení systematických, náhodných, diskretizačních aj. chyb. Obdobně „standardní nejistota“.

Aritmetický průměr jako příklad statistiky



Mějme **vzorek** (*sample*) náhodné veličiny (výběr, trajektorii v simulacích), např. 100× hodíme kostkou.

Aritmetický průměr jako odhad střední hodnoty:

$$\langle x\rangle \approx \bar{x}_n \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \frac{1}{n}\sum_i x_i$$

Spočítejme varianci náhodné veličiny $\bar{x}_n$:

$$\mathrm{Var}\,(\bar{x}_n) = \langle(\bar{x}_n - \langle x\rangle)^2\rangle = \left\langle\left(\frac{\sum_i \Delta x_i}{n}\right)^2\right\rangle = \frac{\mathrm{Var}\,x}{n}$$

Použili jsme nezávislost, tj. $\mathrm{Cov}\,(x_i, x_j) = 0$ pro $i \neq j$. Spočítejme:

$$\left\langle \sum_i \left(x_i - \frac{1}{n}\sum_j x_j\right)^2 \right\rangle = n\left[\left(\left(1-\frac{1}{n}\right)x_1 - \frac{1}{n}x_2 + \cdots\right)^2\right] = (n-1)\mathrm{Var}\,x \quad (1)$$

A proto **odhady** jsou (1 = počet stupňů volnosti):

$$\mathrm{Var}\,x \approx \sigma_n^2(x) \equiv \frac{\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2}{n-1} = \frac{\sum_i \Delta x_i^2}{n-1}, \quad \mathrm{Var}\,(\bar{x}_n) = \frac{\mathrm{Var}\,x}{n}$$

Statistika $\sigma_n^2(x)$ = (korigovaný/nestranný/nevychýlený) výběrový rozptyl

Rozlišujte

Pro zpracování nekorelovaných dat metodou aritmetického průměru, se stejnými vahami dat:

- směrodatná odchylka veličiny (standardní nejistota jednoho měření)
  jejím (nestranným) odhadem z dat je σ_n
- standardní chyba (nejistota) aritmetického průměru, tj. nepřesnost, se kterou známe $\langle x \rangle$
  jejím (nestranným) odhadem z dat je $\sigma_n/\sqrt{n}$

Odhad je nestranný (nevychýlený, *unbiased*), jestliže se jeho střední hodnota rovná střední hodnotě hledané veličiny (parametru) (viz (1) pro σ_n^2)

Zvyky

Výsledky statistického zpracování píšeme takto:

$$\underline{\text{veličina}} = \underline{\text{odhad veličiny}} \pm \underline{\text{odhad chyby}^\dagger}$$

Fyzika: $\underline{\text{odhad chyby}^\dagger} = \sigma =$ odhadnutá směrodatná (standardní) chyba†; nepřesně (odhadnutá) chyba†, směrodatná (standardní) odchylka (rozumí se aritmetického průměru či jiné statistiky)

Obvyklá notace: $123.4 \pm 0.5 \equiv 123.4(5)$

Biologie, ekonomie, inženýrství: Zpravidla se používá hladina významnosti (confidence level) 95 % (data jsou s pravděpodobností 95 % uvnitř mezí). V případě Gaussova rozdělení:

$$\underline{\text{odhad chyby}^\dagger} = 2 \times (\text{standardní chyba})$$

Chemie: často ignorováno; pokud udáno, tak nikdo neví, zda σ či 2σ...

Vždy nutno udat typ chyby

†nebo nejistoty

Studentovo t-rozdělení

Odvodili jsme, že náhodná veličina $\overline{x}_n$ má Gaussovo rozdělení s parametry:

$$\langle \overline{x}_n \rangle = \langle x \rangle, \quad \mathrm{Var}\,(\overline{x}_n) = \frac{\mathrm{Var}\, x}{n} \approx \sigma^2(\overline{x}_n) \equiv \frac{\sum_i \Delta x_i^2}{n(n-1)}$$

Ale odhad standardní chyby průměru, $\sigma(\overline{x}_n)$, je počítán z málo dat!

Definujme Studentovo t-rozdělení s parametrem ν (počet stupňů volnosti) jako rozdělení náhodné veličiny

$$\frac{\overline{x}_{\nu+1} - \langle x \rangle}{\sigma(\overline{x}_{\nu+1})}$$

Tvrzení: distribuční funkce je

$$t_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Platí, že limita je normalizované Gaussovo rozdělení:

$$\lim_{\nu\to\infty} t_\nu(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Příklad – Gauss a Student

a) Spočítali jsme ze 1 000 000 měření, že $\overline{X} = 1.50 \pm 0.10$ (za $\pm$ je standardní chyba). S jakou pravděpodobností neplatí $\langle X \rangle \in (1.20, 1.80)$?

0.27 %

b) Spočítali jsme z 10 měření, že $\overline{X} = 1.50 \pm 0.10$. S jakou pravděpodobností neplatí $\langle X \rangle \in (1.20, 1.80)$?

1.5 %

Porovnání dvou výběrů

Porovnáváme 2 výběry (n a m dat) ze stejného souboru.

Tvrzení. Náhodná veličina

$$\frac{\bar{x}_n - \bar{x}_m}{s\sqrt{1/n + 1/m}}, \quad \text{kde } s^2 = \frac{(n-1)\sigma_n^2(x) + (m-1)\sigma_m^2(x)}{n+m-2}$$

má Studentovo t-rozdělení.

NB: σ_n je výběrový rozptyl dat (ne chyba průměru)

Váhy

Vážený průměr (váhy w_i nemusí být normalizované)

$$\overline{x} = \frac{\sum_i x_i w_i}{\sum_i w_i}$$

Známe x_i (nezávislé) s chybami σ_i. Jaké máme volit váhy?

Odvodíme pro 2 veličiny:

$$\overline{x} = w x_1 + (1-w) x_2$$

$$\sigma^2(\overline{x}) = \langle (\overline{x} - \langle x \rangle)^2 \rangle = \langle (w\Delta x_1 + (1-w)\Delta x_2)^2 \rangle = w^2\sigma_1^2 + (1-w)^2\sigma_2^2$$

Minimum nastane pro

$$w = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}, \quad 1 - w = w_2 = \frac{1/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}$$

Tedy (a platí obecně):

$$w_i = \frac{1}{\sigma_i^2}$$

Ale problém může být, pokud neznáme σ_i přesně.

Průměrování nezávislých měření

Máme m nezávislých měření x_i, $i = 1..m$, vč. odhadů standardní chyby σ_i. Chceme odhad střední hodnoty $\overline{x}$ a standardní chyby σ.

Malá množství dat. Počet dat při výpočtu x_i je $= n_i$

$$\overline{x} = \frac{\sum_i n_i x_i}{\sum n_i}, \quad \sigma = \sqrt{\frac{\sum_i n_i(n_i-1)\sigma_i^2}{\sum_i (n_i-1)\sum_i n_i}}$$

Známé váhy. Známe (přesně) váhy w_i, dat je „mnoho“:

$$\overline{x} = \frac{\sum_i w_i x_i}{\sum w_i}, \quad \sigma = \frac{\sqrt{\sum_i w_i^2 \sigma_i^2}}{\sum w_i}$$

Neznámé váhy. Pak $w_i = 1/\sigma_i^2$ (za předpokladu, že σ_i jsou dostatečně přesné) a z výše uvedeného vzorce dostaneme

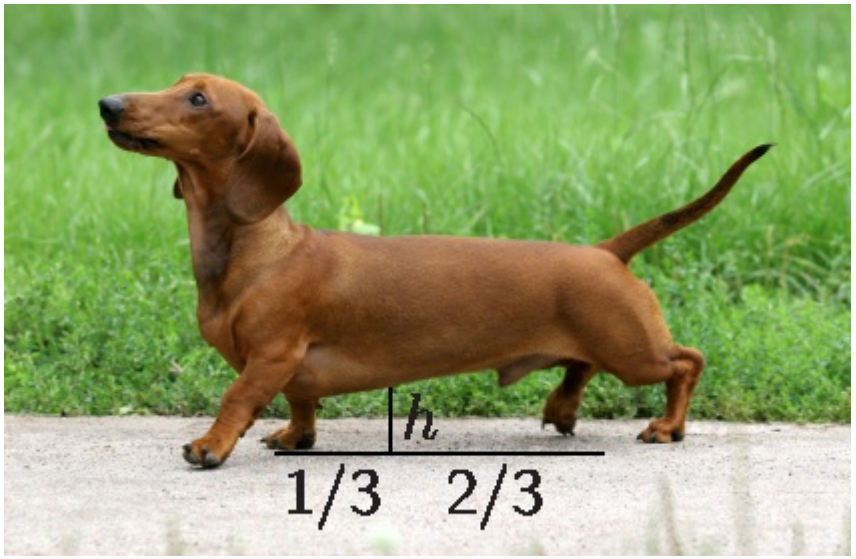
$$\overline{x} = \frac{\sum_i x_i/\sigma_i^2}{\sum 1/\sigma_i^2}, \quad \sigma = \frac{1}{\sqrt{\sum_i 1/\sigma_i^2}}$$

Příklad (viz mat-num2.mw)

Firma vyrábí podpěry pro příliš dlouhé jezevčíky. Zadala dvěma agenturám měření spodní výšky jezevčíka. Agentura A zjistila výšku podpěry $h = 12.6 \pm 1.1$ cm (standardní chyba), agentura B naměřila $h = 9.2 \pm 1.3$ cm.

a) Stanovte co nejpřesněji výšku podpěry vč. odhadu nejistoty. Předpokládejte, že obě agentury použily dostatečné množství psů.
b) Jsou oba výsledky v souladu (na 95% hladině významnosti)?
c) Opakujte oba výpočty, jestliže víte, že Agentura A měřila 12 jezevčíků, zatímco Agentura B jen 6 jezevčíků.

a) 11.2(8); b) ne; c+a) 11.5(9); c+b) ano



Metoda nejmenších čtverců

$\vec{x}_i$ = nezávisle proměnné (n vektor; libovoln0 dimenze, $i = 1..n$)
y_i = závisle proměnné (reálná čísla)
$1/\sigma_i^2$ = váhy
$\vec{a}$ = parametry (p hodnot zapíšeme jako vektor), $p \leqslant n$, nejlépe $p \ll n$)

Hledáme funkci $f_{\vec{a}}(\vec{x})$ závislou na p parametrech vystihující data $(\vec{x}_i, y_i)$. Parametry $\vec{a}$ budeme hledat z podmínky minima součtu kvadrátů odchylek:

$$\min_{\vec{a}} S^2, \quad S^2 = \sum_i \left[\frac{f_{\vec{a}}(\vec{x}_i) - y_i}{\sigma_i}\right]^2$$

Věta (Gauss–Markov): pro funkci $f_{\vec{a}}$ lineárně závisející na $\vec{a}$ je metoda nejmenších čtverců:

Best (dává nejmenší varianci odhadnutých$\vec{a}$)
Linear (= předpoklad)
Unbiased ($\langle\vec{a}\rangle$ je správné)
Estimate (BLUE).

$\langle S^2\rangle = n - p$

- V limitě $n \to \infty$ platí $s = \sqrt{S^2/(n-p)} \to 1$ (posouzení fitu)

Metoda nejmenších čtverců

Příklad. Pro $f_a(x) = a$ (konstanta) a $\sigma_i = 1$ najděte odhad a

$\underline{\hbar} = \mathfrak{v}$

Výsledkem **fitování** (korelace, regrese, prokládání) je:

- odhad $\vec{a}$
- odhad standardních chyb
- odhad korelací mezi parametry
- odhad hodnoty nějaké funkce $g(\vec{a})$ (vč. odhadu standardní chyby)

V limitě $n \to \infty$ platí $s = \sqrt{S^2/(n-p)} \to 1$ (posouzení fitu)

Metoda nejmenších čtverců – linearizace

Nechť $\vec{a}_0$ je přesná (hledaná) hodnota parametrů. Pro každé $\vec{x}$ rozvineme:

$$f_{\vec{a}}(\vec{x}) \approx f_{\vec{a}_0}(\vec{x}) + \sum_{j=1}^{p} \Delta a_j f_j(\vec{x}), \quad f_j(\vec{x}) = \frac{\partial f_{\vec{a}_0}(\vec{x})}{\partial a_j}$$

kde $\vec{a} = \vec{a}_0 + \Delta\vec{a}$.

Pokud jsou změny parametrů $\vec{a}$ malé, bez újmy na obecnosti stačí studovat lineární model

$$f_{\vec{a}}(\vec{x}) = \sum_{j=1}^{p} a_j f_j(\vec{x}),$$

kde $\{f_j(\vec{x})\}_{j=1}^{p}$ je (obecně neortogonální) báze.

Metoda nejmenších čtverců – lineární model

$$f_{\vec{a}}(\vec{x}) = \sum_{j=1}^{p} a_j f_j(\vec{x})$$

Dále předpokládejme stejné váhy. Data s chybami můžeme zapsat jako:

$$y_i = \sum_{j=1}^{p} a_j f_j(\vec{x}) + \delta y_i, \qquad \langle \delta y_i \rangle = 0, \;\; \langle \delta y_i \delta y_j \rangle = \sigma^2 \delta_{ij}$$

kde $\delta_{ij} = 1$ pro $i = j$ a $\delta_{ij} = 0$ pro $i \neq j$ (Kroneckerovo delta).

$$S^2 = \sum_{i=1}^{n} \left[\sum_{j=1}^{p} a_j f_j(\vec{x}_i) - y_i \right]^2$$

Hledáme minimum, tedy spočítáme derivaci a položíme $= 0$:

$$\frac{1}{2} \frac{\partial S^2}{\partial a_k} = \sum_i f_k(\vec{x}_i) \left[\sum_{j=1}^{p} a_j f_j(\vec{x}_i) - y_i \right] = (A \cdot \vec{a} - \vec{b})_k \stackrel{!}{=} 0$$

kde $A = F \cdot F^T$, $\vec{b} = F \cdot \vec{y}$, $F_{ki} = f_k(x_i)$ (matice $p \times n$) a $\cdot =$ maticové násobení.

Metoda nejmenších čtverců – kovarianční matice

$$A \cdot \vec{a} = \vec{b}, \quad \vec{a} = A^{-1} \cdot \vec{b} = A^{-1} \cdot F \cdot \vec{y}$$

Zbývá spočítat chyby odhadů a korelace (kovariance) mezi parametry; pravidlo: sčítáme vždy přes dvojice stejných indexů:

$$\begin{aligned}
\mathrm{Cov}\,(a_i, a_j) = \langle \Delta a_i \Delta a_j \rangle &= \sum A^{-1}_{i\alpha} F_{\alpha k} \delta y_k \, A^{-1}_{j\beta} F_{\beta l} \delta y_l \\
&= \sum A^{-1}_{i\alpha} F_{\alpha k} \, A^{-1}_{j\beta} F_{\beta l} \sigma^2 \delta_{kl} \\
&= \sum A^{-1}_{i\alpha} F_{\alpha k} \, A^{-1}_{j\beta} F_{\beta k} \sigma^2 \\
&= \sum A^{-1}_{i\alpha} A_{\alpha\beta} A^{-1}_{j\beta} \sigma^2 \\
&= \sum A^{-1}_{i\alpha} A_{\alpha\beta} A^{-1}_{\beta j} \sigma^2 \\
&= A^{-1}_{ij} \sigma^2
\end{aligned}$$

Pokud neznáme σ, odhadneme ho takto (analogie s průměrem, kdy $p = 1$):

$$\sigma^2 = \frac{S^2}{n - p}$$

Výsledkem fitování nejsou jen odhady chyb parametrů (na diagonále), ale i korelace (kovariance)!

Metoda nejmenších čtverců – funkce parametrů

Potřebujeme spočítat $g(\vec{a})$ s chybou

$$g_{\vec{a}} \approx g_{\vec{a}_0} + \sum_{j=1}^{p} \Delta a_j g_j(\vec{x}), \quad g_j = \frac{\partial g_{\vec{a}_0}}{\partial a_j}$$

$$\langle (g_{\vec{a}} - g_{\vec{a}_0})^2 \rangle = \langle \sum_{ij} \Delta a_i g_i \Delta a_j g_j \rangle = \sum_{ij} g_i A^{-1}_{ij} \sigma^2 g_j$$

Příklady $g(\vec{a})$: a_i (jeden z parametrů), $\int_{x_0}^{x_1} f(x)dx$

Metoda nejmenších čtverců – chyby MC metodou

- Minimalizujeme $S^2 \Rightarrow$ získáme $\vec{a}_0$ a $g(\vec{a}_0)$
- Pro $k = 1..m$ provedeme:
  - Vyrobíme falešná data

    $$y_i^{(k)} = f_{\vec{a}_0}(\vec{x}_i) + \sigma_i u$$

    kde u je náhodné číslo s normalizovaným Gaussovým rozdělením (chyby σ_i dat y_i známe; pokud ne, použijeme $\sigma_i = \sqrt{S^2/(n-p)}$)
  - Vypočteme metodou nejmenších čtverců parametry $\vec{a}^{(k)}$
  - Vypočteme $g(a^{(k)})$
- Výsledky $g(a^{(k)})$ pro $k = 1..m$ zpracujeme jako nezávislá data a dostaneme odhad standardní chyby $\sigma(g)$

Metoda nejmenších čtverců – jak na to

Lineární model: řešitelný metodou lineární algebry, nebývají problémy (pokud jsou, stačí ortonormalizovat bázi)

Nelineární model:

Problém 1: můžeme mít více lokálních minim, některá pro nevlastní hodnoty parametrů (= divergence)

Problém 2: dlouhá zakřivená údolí při minimalizaci (pomalé)

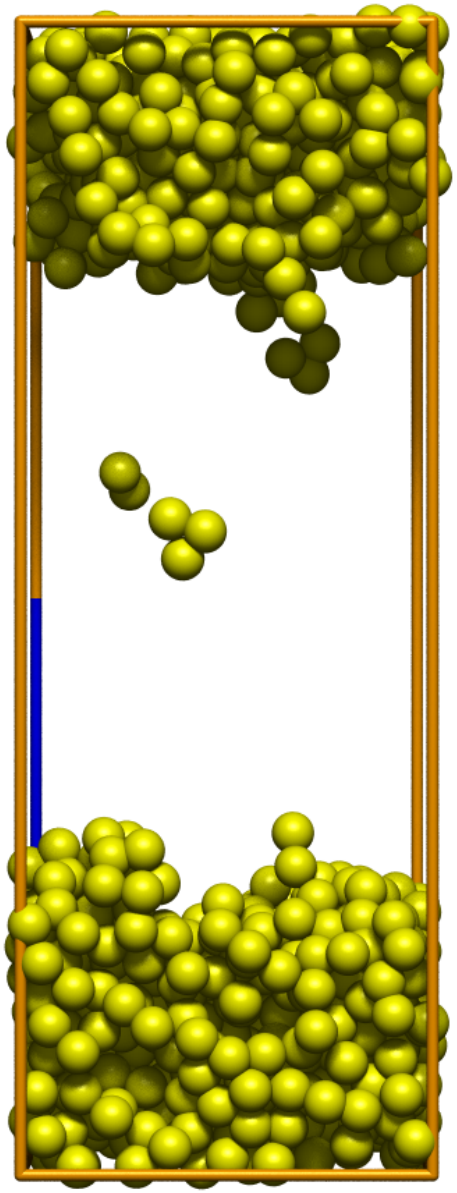
Minimalizace nelineární funkce mnoha proměnných:

- metoda největšího spádu (steepest descent, greedy)
- konjugované gradienty
- amoeba (Nelder–Mead)
- Monte Carlo search (na začátku)
- (Gaussova–)Newtonova metoda (jsme-li blízko řešení)
- (Levenbergova–)Marquardtova metoda (kombinace Newton, gradient, tlumení)

Příklad z praxe

Při simulaci modelu platiny ve slab-geometrii byla získána data pro tlak ve směru kolmém na vrstvu:

T/K	p/bar	stderr/bar
3700	14.7	2.2
3750	11.9	1.4
3800	14.9	2.6
3850	18.9	2.8
3900	16.3	1.8
3950	16.5	3.2
4000	26.5	3.3
4050	24.3	2.6
4100	30.6	2.6
4150	28.5	3.5
4200	34.5	3.5
4250	43.4	2.6
4300	48.0	3.1



Vypočtěte bod varu platiny za tlaku 1 bar a odhadněte chybu.

Řešení

Budeme předpokládat platnost Clausiovy–Clapeyronovy rovnice a konstantní výparnou enthalpii.

$$\ln p = a + b/T$$

kde a a b jsou konstanty, které budeme fitovat. Pak stanovíme hodnotu funkce g, která je řešením rovnice $\ln p = a + b/T$ pro $p = 1$ bar.

- Přímé fitování na $p = \exp(a + b/T)$:
  $s = 1.067$, $T_{vap} = 3021(55)$ K, přeškálováno s (59)
- Fitujeme $\ln(p)$ vs. $1/T$ (lineární regrese):
  $s = 1.081$, $T_{vap} = 2992(53)$ K, přeškálováno s (57)
- Nepředpokládáme-li znalost standardních chyb měření, vyjde:
  $T_{vap} = 3015(74)$ K
- Protože data jsou založena na stejně dlouhých trajektorích, můžeme chyby jednotlivých měření vyrovnat. Vyjde:
  $s = 1.138$, $T_{vap} = 2965(63)$ K, přeškálováno s (72)

Závěr: $T_{vap} = 2965(72)$