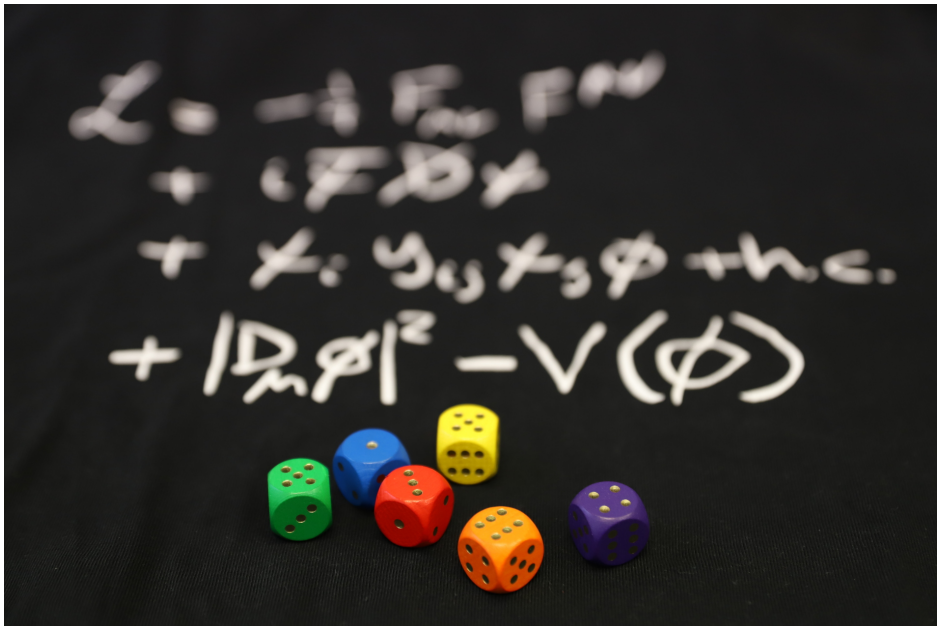
Poznámky k přednášce

Statistika ve fyzice vysokých energií (SLO/SFVE)



Jiří Kvita

Společná laboratoř optiky UP a AVČR

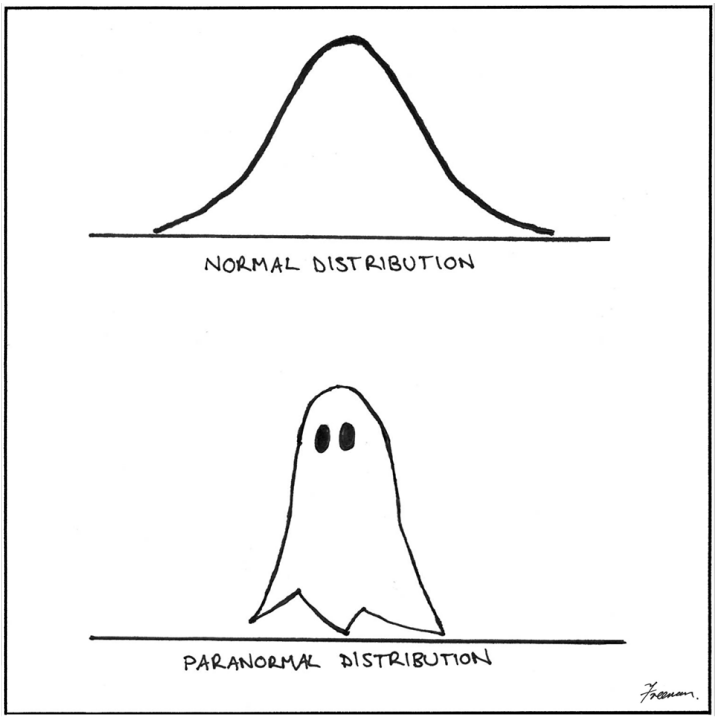
Přírodovědecká fakulta Univerzity Palackého v Olomouci

29. února 2024

Motto:

Three Statisticians Walk Into A Bar

The barman asks them, "Would you all like a drink?"
The first statistician says, "Maybe..."
The second one says, "Maybe..."
And then the third one smiles and says, "Yes!"



Obsah

1 Předmluva **5**

2 Úvod **5**
2.1 Frekventistická a Bayesovská definice pravděpodobnosti 5

3 Měření, náhodná proměnná, hustota pravděpodobnosti **6**
3.1 Pravděpodobnostní rozdělení 6
3.2 Náhodný výběr z rozdělení 6
3.3 Empirická hustota pravděpodobnosti 7
3.3.1 Spojitá rozdělení 7
3.3.2 Diskrétní rozdělení 9
3.4 Očekávaná hodnota, střední hodnota, rozptyl 11
3.4.1 Momenty rozdělení 12
3.5 Odhad měr polohy a rozptylu z nebinovaného výběru dat a z histogramu 14
3.6 Vlastnosti hustot pravděpodobnosti, odvozené veličiny 15
3.7 Více náhodných proměnných 15
3.8 Kovariance, korelace 16
3.9 Odhad měr polohy a rozptylu pro vícerozměrná data 17
3.10 Návrat k průměru 17
3.11 Data rozdělená podle dvojrozměrného Gaussova rozdělení 19
3.12 Transformace proměnných 26
3.13 Odhadování h.p. pomocí kernelu 27
3.14 Náhodná chůze 28

4 Monte Carlo metoda **30**
4.1 von Neumannova Monte Carlo metoda 30
4.2 Generování spektra pomocí transformační funkce 32

5 Funkce náhodných proměnných **34**
5.1 Funkce náhodných proměnných 34
5.2 Propagace (šíření) chyb 36
5.3 Rozptyl efektivity 40
5.4 Cross-sections ratio 41
5.5 Variance vážených dat 43

6 Odhad parametrů **44**
6.1 Zaujaté a nezaujaté odhady 44
6.2 Fitování histogramu 45

7 Věrohodnost **48**
7.1 Odhad parametrů metodou maximální věrohodnosti 48
7.2 Odhad "chyby"MLE estimátorů 50
7.2.1 Variance parametrů odhadnutých metodou maximální věrohodnosti 50
7.2.2 Grafické řešení 50
7.2.3 Cross section fit using a likelihood 51
7.3 Podíl věrohodností 52

8 Bayesův teorém **53**
8.1 Aplikace na spolehlivost testu 53
8.2 Aplikace na odhad parametrů 54

9 Testování hypotéz **56**
9.1 Testovací statistika 56
9.2 Counting Experiment 57
9.3 Neyman-Pearsonovo lemma, odhad síly signálu 59
9.4 χ^2 test pro data a fit; a mezi dvěma histogramy 60
9.5 Kompatibilita dvou měření 60

10 Klasifikace **62**
10.1 Řezy (cuty) 62
10.2 Fischerův diskriminant 62
10.3 Rozhodovací stromy 64
10.4 Umělé neurální sítě 64
10.5 Strojové učení 65

11 Aplikace **69**
11.1 Náhodné výběry z Poissonova rozdělení 69
11.2 Propagace chyb: chybový pás fitu 70
11.3 Fitování frakce 72
11.4 Výběr binování pro fit 72
11.5 Interval pokrytí Poissonova rozdělení 74
11.6 *b*-tagging 74
11.7 Kombinace měření 76
11.8 Věrohodnostní diskriminant 76
11.9 Metoda "ABCD" pro odhad pozadí 77

12 Unfolding **79**
12.1 Definice úlohy, bias a variance 79
12.2 Regularizace 79
12.3 Bayesovská metoda 79
12.4 Korekce 79

1 Předmluva

Tento text čerpá z mnoha učebnic [1] [2] [3][4] a online materiálů [5] [6] [7] [8]. Text vzniká jako poznámky k přednášce Statistika ve fyzice vysokých energií, s aktivní účastí, náměty a otázkami studentů Palackého univerzity v Olomouci, kterým patří dík za trpělivost a podněty!

2 Úvod

2.1 Frekventistická a Bayesovská definice pravděpodobnosti

Frekventistická definice pravděpodobnosti je založena na opakovatelnosti experimentu a pracuje s hustotou pravděpodobnosti, funkcí, jejíž hodnota udává očekávanou (teoretickou) franmci událostí daného typu. V jejím jazyce pracujeme s pravděpodobnostmi, že pozorujeme určitá data a jejich kompatibilitu s danou hypotézou, ale často ve smyslu toho, jaká je pravděpodobnost pozorování kompatibility, kterou pozorujeme, a ještě horší. To nutně není odpovědí na naši otázku.

Oproti tomu v realitě čelíme otázkám jiného typu, které nestojí na opakovatelnosti experimentu, a které jsou nejlépe uchopitelné pomocí Bayesova teorému a jazykem podmíněných pravděpodobností: pokud pozoruji daná data, jaká je pravděpodobnost, že pocházejí z pravděpodobnostního rozdělení o nějakých konkrétních parametrech (odpovídající např. určité hmotě nějaké částice)?

3 Měření, náhodná proměnná, hustota pravděpodobnosti

Uvažujme proces měření, kdy z experimentálního zařízení (detektoru) získáváme informaci o hodnotě nějaké pozorovatelné veličiny. Takto získaná hodnota je z principu náhodná: různé experimentální podmínky, které nutně nemáme pod kontrolou a nemůžeme je ovlivnit, vedou k různým naměřeným hodnotám. Často měříme veličiny, které jsou náhodné z principu náhodnosti měření např. v kvantové mechanice. Samotná veličina tak může nabývat různých hodnot (mít nějaké "rozdělení"), a navíc může její změřená hodnota být ovlivněna konečným rozlišením detektoru, tj. skutečností, že akt měření a použití konkrétního přístroje do procesu měření vnáší např. šum z elektroniky, ale také fluktuace v měření např. energie, které jsou dány náhodnistí fyzikálních procesů v průběhu daného měření, které opět nemůžeme ovlivnit. Měřená veličina tak nabývá hodnot z nějakého intervalu nenulové šířky, který je dán jak nahodilostí a rozdělením veličiny samotné, tak konečnou rozlišovací schopností aparatury, kterou používáme. Zkuste si výše uvedené skutečnosti promyslet pro Vaši oblíbenou experimentální techniku a veličinu.

3.1 Pravděpodobnostní rozdělení

Uvažujme náhodou veličinu X (např. energii nějaké částice), která je výsledkem nějakého měření, a která může nabývat spojitých (někdy jen diskrétních) hodnot z nějaké množiny Ω. Hodnoty veličiny si budeme značit reálnou proměnnou x, tedy $x \in \Omega$. Hovoříme o tom, že veličina X je rozdělena podle hustoty pravděpodobnosti (h.p.), také statistického rozdělení, $f(x|\boldsymbol{\theta})$, které může záviset na dalších parametrech $\boldsymbol{\theta}$ (jako je např. hmota částice, energie srážek, kterých dané částice vznikají apod.), a značíme jako

$$X \sim f(x|\boldsymbol{\theta})\,.$$

Hustota pravděpodobnosti je tedy zobrazením $f(x|\boldsymbol{\theta}) : \Omega \to \mathbb{R}_0^+$, musí být nezápornou funkcí. Mluvíme o hustotě pravděpododobnosti jako o množinové pravděpodobnostní míře.

Parametry $\boldsymbol{\theta}$ mohou mít důležitý fyzikální význam, jejich znalost může být cenná. Často jde o samotný cíl experimentu: extrahovat nejlepší možný odhad těchto parametrů z naměřených dat. Jiné parametry mohou být okrajové, nezajímavé.

3.2 Náhodný výběr z rozdělení

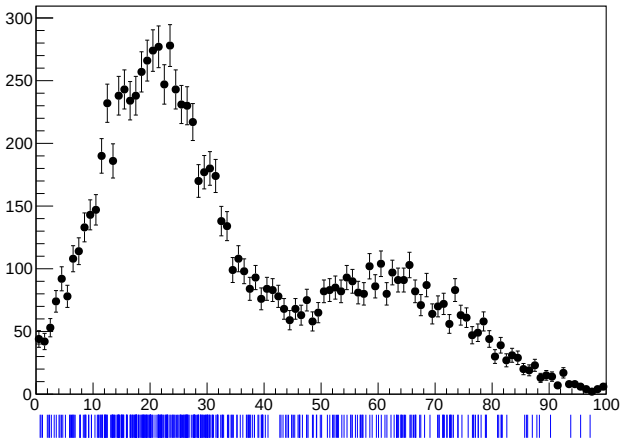
Sadu N hodnot měření náhodné veličiny X budeme nazývat náhodným výběrem z daného rozdělení, a značit jako

$$\{x_i\}_{i=1}^{N}\ .$$

Z takovéhoto výběru budeme následně často chtít extrahovat odhady veličin, jako je střední hodnota či jiné parametry rozdělení, či samotné rozdělení opravené o vliv konečné rozlišovací schopnosti měřící aparatury. Budeme chtít také třeba říci něco o tom, zda jsou dva náhodné výběry kompatibilní s hypotézou, že pocházejí ze stejné veličiny. Budeme chtít kombinovat informace získané z různých náhodných výběru, apod.

Náhodný výběr chápeme jako neredukovaná, nebinovaná data. V praxi často data redukujeme do formy histogramu, který udává počet událostí (četnost) n_i, kdy veličina X spadá do intervalu (x_i, x_{i+1}). Hodnoty $\{x_i\}_{i=1}^{n_{\text{bins}}+1}$ definují okraje intervalů, mluvíme o konkrétní formě "binningu"o n_{bins} binech, viz Obr. 1. Biny nemusí být uniformní, tj. stejné šířky. "Binováním"ztrácíme částečně informaci: místo s N hodnotami každého jednoho měření pracujeme pouze s n_{bins} souhrnnými hodnotami, informacemi o počtu

výsledků spadajících do nějakého "binu", ale současně tak redukujem množství dat, které je nutno např. dále skladovat.



Obrázek 1: Ilustrace nebinovaných dat (modré úsečky) a binovaná forma dat (histogram).

3.3 Empirická hustota pravděpodobnosti

Empirickou, tedy z dat odhadnutou, hustotu pravděpodobnosti, můžeme definovat takto: Uvažujme, že provedeme N_{exp} měření, z nichž n_{obs} je daného sledovaného typu k. Pak empirickou pravděpodobnost tohoto jevu definujeme jako

$$f_k^{\text{emp}} \equiv \frac{n_{\text{obs}}}{N_{\text{exp}}} \,.$$

Diskrétní hustotu pravděpodobnosti pak můžeme chápat jako limitní případ, kdy se počet experimentů blíží nekonečnu

$$f_k \equiv \lim_{N_{\text{exp}} \to \infty} \frac{n_{\text{obs}}}{N_{\text{exp}}} \,.$$

Obdobně lze uvažovat i pro spojitou náhodnou veličinu, kde je však potřeba uvažovat, že pozorujeme $n_{X \in (a,b)}$ případů, kdy veličina X spadá do intervalu (a, b):

$$f_{X \in (a,b)}^{\text{emp}} \equiv \frac{n_{X \in (a,b)}}{N_{\text{exp}}} \,,$$

$$P_{X \in (a,b)} \equiv \int_a^b f(x)\,\mathrm{d}x \equiv \lim_{N_{\text{exp}} \to \infty} \frac{n_{\text{X}\in(\text{a,b})}}{N_{\text{exp}}} \,.$$

3.3.1 Spojitá rozdělení

Hustota pravděpodobnosti pro spojitě rozdělenou veličinu x může záviset na dalších parametrech souhrně označených jako $\boldsymbol{\theta}$, označujeme jako $f(x|\boldsymbol{\theta})$ a musí pro ni platit normalizační podmínka

$$\int_\Omega f(x|\boldsymbol{\theta})\,\mathrm{d}x = 1 \,,$$

která má význam toho, že náhodná veličina musí nabývat (např. při měření) některé z hodnot z množiny Ω.

Hustota pravděpodobnosti $f(x|\boldsymbol{\theta})$ udává, jaká je infinitezimální pravděpodobnost toho, veličinu X naleznu v intervalu $(x, x + \mathrm{d}x)$, tj.

$$\mathrm{d}P_{X\in(x,x+\mathrm{d}x)} = f(x|\boldsymbol{\theta})\,.$$

Na intervalu konečné šířky můžeme mluvit již o pravděpodobnosti, že X nalezneme v intervalu (a, b) a psát

$$P_{X\in(a,b)} = \int_a^b f(x|\boldsymbol{\theta})\,\mathrm{d}x\,.$$

Očekávanou, střední, hodnotu dané veličiny rozdělené podle h.p. $f(x|\boldsymbol{\theta})$ definujeme jako

$$\mathrm{E}\left[x\right] \equiv \int_\Omega x \cdot f(x|\boldsymbol{\theta})\,\mathrm{d}x \equiv \mu_x\,.$$

Důležitý je také rozptyl, jehož kvadrát je definován jako

$$\sigma_X^2 \equiv \mathrm{E}\left[(x - \mathrm{E}\left[x\right])^2\right] \equiv \int_\Omega (x - \mathrm{E}\left[x\right])^2 \cdot f(x|\boldsymbol{\theta})\,\mathrm{d}x\,.$$

V praxi jsou důležitá a hojně využívána následující spojitá rozdělení

- Uniformní rozdělení
- Gaussovo rozdělení
- Chí-kvadrát rozdělení
- Cauchy/Breit-Wigner/Lorentzovo
- Relativistické Breit-Wignerovo rozdělení

$$f(m|M,\Gamma) = \frac{2\sqrt{2}}{\pi}\frac{\sqrt{1+(\Gamma/M)^2}}{\sqrt{1+\sqrt{1+(\Gamma/M)^2}}}\frac{\Gamma M}{(m^2-M^2)^2+m^2\Gamma^2}$$

- Studentovo t-rozdělení
- Gamma rozdělení
- Landauovo rozdělení
- Bifurkované Gaussovo rozdělení (viz také Obr. 2)

$$f(x|\mu,\sigma_1\sigma_2) = \sqrt{\frac{2}{\pi}}\frac{1}{(\sigma_1+\sigma_2)}\cdot\begin{cases} x<\mu: & \exp\left[-\frac{(x-\mu)^2}{2\sigma_1^2}\right] \\ x\geq\mu: & \exp\left[-\frac{(x-\mu)^2}{2\sigma_2^2}\right] \end{cases}$$

jejichž vlastnosti jsou shrnuty v Tabulce 1.
Poznámka: gama funkce je definována jako

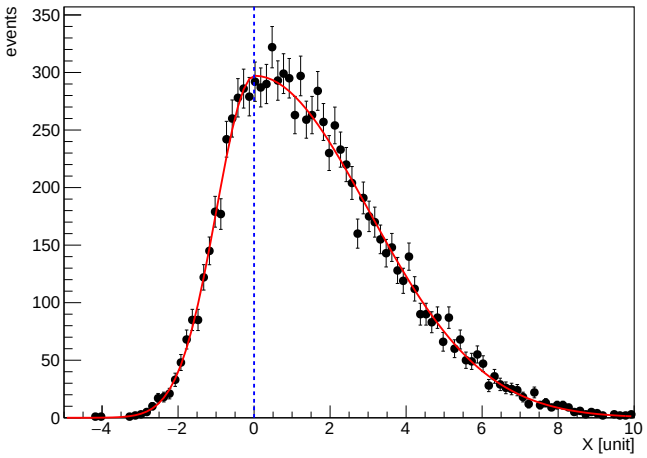
$$\Gamma(x) \equiv \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

a lze ji chápat jako rozšíření faktoriálu pro reálná čísla. Platí totiž

$$\Gamma(x+1) = x\Gamma(x)\,, \quad \Gamma(n) = (n-1)!$$

rozdělení	hustota pravděpodobnosti	$\mu \equiv \mathrm{E}\,[x]$	$\sigma^2 \equiv \mathrm{E}\left[(x - \mathrm{E}\,[x])^2\right]$
Uniformní	$f(x\|a,b) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussovo	$f(x\|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2
Exponenciální	$f(x\|\tau) = \frac{1}{\tau}\exp[-t/\tau]$	τ	τ^2
Chí-kvadrát	$f(x\|n) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{n/2-1}\exp[-x/2]$	n	$2n$
Cauchy	$f(m\|M,\Gamma) = \frac{1}{\pi}\frac{\Gamma}{(m-M)^2+\Gamma^2}$	není definováno	diverguje
Breit-Wigner	$f(m\|M,\Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(m-M)^2+\Gamma^2/4}$	není definováno	diverguje
relat. B.-W.	$f(m\|M,\Gamma) \sim \frac{2\sqrt{2}}{\pi}\frac{\Gamma M}{(m^2-M^2)^2+m^2\Gamma^2}$	není definováno	diverguje
Studentovo	$f(x\|n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} 1/\left(1+\frac{x^2}{n}\right)^{(n+1)/2}$	0	$\frac{n}{n-2}$
Gamma	$f(x\|a,b) = \frac{a}{\Gamma(b)}(ax)^{b-1}\exp[-ax]$	b/a	b/a^2
Landauovo	$f(x) = \frac{1}{\pi}\int_0^\infty \exp[-t\ln t - xt]\sin(\pi t)\mathrm{d}t$	není definováno	není definováno

Tabulka 1: Střední hodnoty a rozptyly různých diskétních hustot pravděpodobností, jako funkce jejich parametrů.



Obrázek 2: Histogram nagenerovaný podle a proložený bifurkovaným Gaussovým rozdělením o šířkách 1 a 3. Inspirováno příkladem dle M. Princové.

3.3.2 Diskrétní rozdělení

Pro náhodné veličiny, které mohou nabývat diskrétních hodnot, typicky přirozených čísel, je hustota pravděpodobnosti diskétní sada hodnot $f(k|\boldsymbol{\theta}) \equiv f_k(\boldsymbol{\theta})$ s normalizační

podmínkou

$$\sum_k f(k|\boldsymbol{\theta}) = 1\,.$$

Očekávanou, střední, hodnotu dané veličiny rozdělení podle h.p. $f(k|\boldsymbol{\theta})$ definujeme analogicky jako ve spojitém případě jako

$$\mathrm{E}\,[k] \equiv \sum_k k \cdot f(k|\boldsymbol{\theta})$$

a kvadrát rozptylu jako

$$\sigma^2 \equiv \mathrm{E}\left[(k - \mathrm{E}\,[k])^2\right] \equiv \sum_k (k - \mathrm{E}\,[k])^2 f(k|\boldsymbol{\theta})\,.$$

Můžeme mluvit o pravděpodobnosti, že velilčinu X nalezneme v rozmezí $m \ldots n$ a psát

$$P_{k=\{m,\ldots,n\}} = \sum_{k=m}^{n} f(k|\boldsymbol{\theta})\,.$$

Níže jsou shrnuty vlastnosti některých důležitých diskrétních rozdělení, viz také Tabulka 2.

- Uniformní: např. hod mincí, kde $P(k) = \frac{1}{2}$, $k = 1,\ 2$, střední hodnota je 0.5. a hod kostkou $P(k) = \frac{1}{6}$, $k = 1 \ldots 6$ dokažte si, že střední hodnota je 3.5, a spočítejte rozptyl.

- Binomické rozdělení: mějme pravděpodobnost p, že nastane daný jev. Pak pravděpodobnost, že jev nastane v k případech z n je dána binomickým rozdělením, které má hustotu pravděpodobnosti

  $$P(k|p;n) = \binom{n}{p} p^k (1-p)^{n-k}\,,$$

  kde $1 - p$ je doplňková pravděpodobnost, že daný jev nenastane. Podle očekávání se v rozdělení objevuje kombinatorický faktor. Střední hodnota je $E[k] = np$ a variance pak $\mathrm{Var}[k] = np(1-p)$.

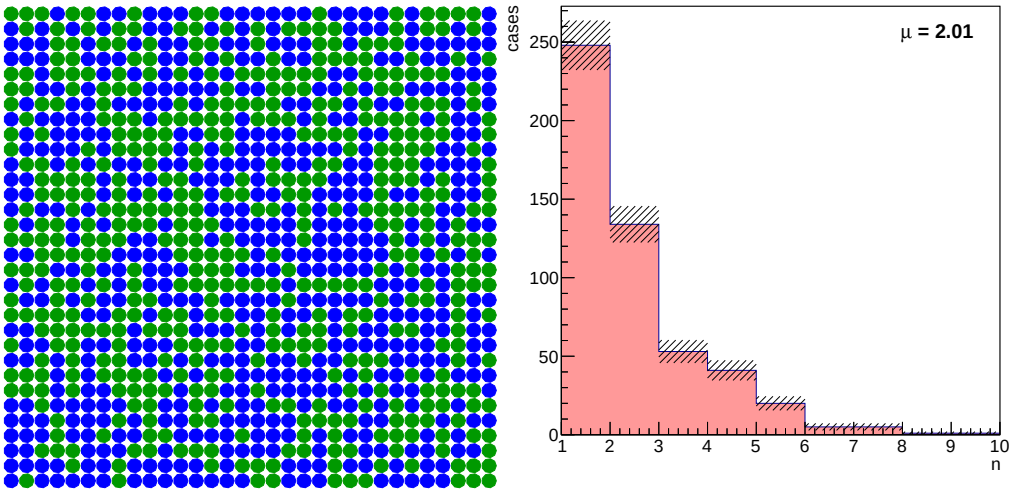
- Poissonovo rozdělení je dáno jedním parametrem, např. μ, a hustotou pravděpodobnosti

  $$P(k|\mu) = \frac{\mu^k}{k!}\mathrm{e}^{-\mu}\,,$$

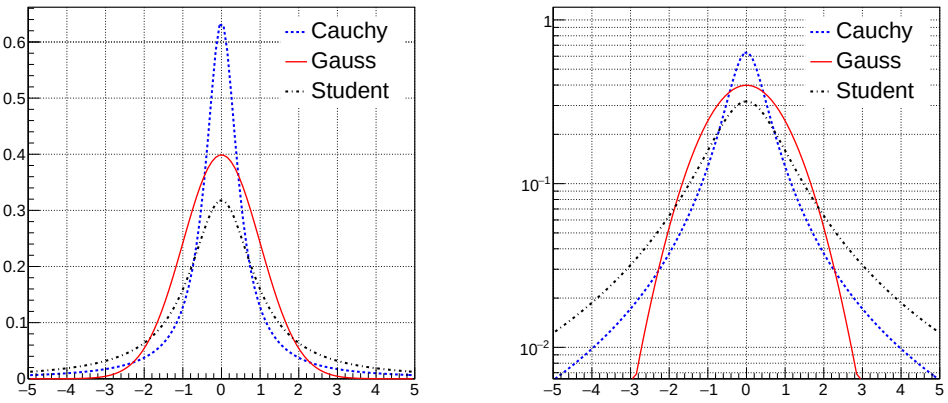
  která udává, jaká je pravděpodobnost, že pozorujeme k případů daného jevu, kde $k = 0 \ldots \infty$. Ukažte si, že jde o normovanou h.p. Střední hodnota Poissonova rozdělení je μ, variance také μ, a rozptyl je tedy $\sqrt{\mu}$.

rozdělení	hustota pravděpodobnosti	$\mu \equiv \mathrm{E}\,[k]$	$\sigma^2 \equiv \mathrm{E}\left[(k - \mathrm{E}\,[k])^2\right]$
Binomické	$P(k\|p;n) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Poissonovo	$P(k\|\mu) = \frac{\mu^k}{k!}\,\mathrm{e}^{-\mu}$	μ	μ

Tabulka 2: Střední hodnoty a rozptyly různých diskétních hustot pravděpodobností, jako funkce jejich parametrů.



Obrázek 3: Simulace náhodných sekvencí hodu mincí.Histogram četností sekvencí dané délky. Inspirováno přednáškou O. Vencálka.



Obrázek 4: Standardizované Gaussovo (plně), Cauchyho rozdělení (čárkovaně) a Studentovo t-rozdělení (čercovaně). Vůči Gaussovu rozdělení je Cauchyho rozdělení užší, Studentovo širší. Vlevo: lineární měřítko, vpravo: y-logaritmické.

3.4 Očekávaná hodnota, střední hodnota, rozptyl

Pro spojitá rozdělení je očekávaná, tj. střední, hodnota dané veličiny definována jako

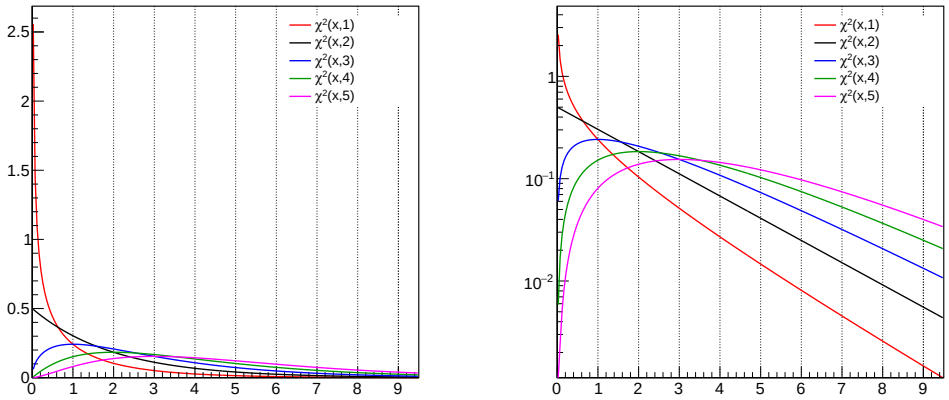
$$\mathrm{E}\left[x\right] \equiv \int_{\Omega} x \cdot f(x)\,\mathrm{d}x \equiv \mu_x\,.$$

Obecněji, střední hodnota libovolného výrazu $g(x)$ je rovna

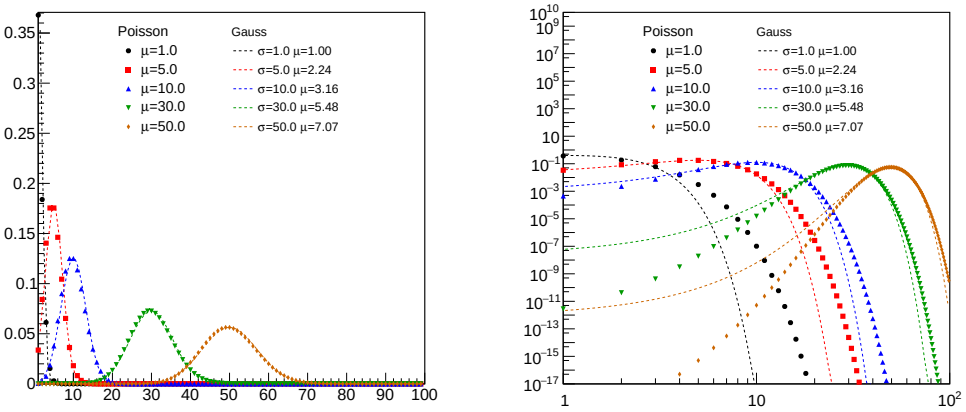
$$\mathrm{E}\left[g(x)\right] \equiv \int_{\Omega} g(x) \cdot f(x)\,\mathrm{d}x$$

Z vlastnosti integrálu je zřejmé, že $E\cdot$ je linární forma (funkci přiřazuje číslo) s vlastnostmi

$$\mathrm{E}\left[ax + b\right] = a\,\mathrm{E}\left[x\right] + b\,.$$



Obrázek 5: Chí-kvadrát rozdělení pro různý počet stupňů volnosti n. Vlevo: lineární měřítko, vpravo: y-logaritmické.



Obrázek 6: Srovnání Poissonova (značky) a Gaussova (čárkovaně) rozdělení pro vybrané parametry μ, pro Gaussovo rozdělení bylo vždy voleno $\sigma = \sqrt{\mu}$. Vlevo: lineární měřítko, vpravo: dvojtě logaritmické.

Pozn.: normalizační podmínka h.p. je tedy vyjádřitelná jako $\mathrm{E}[1] = 1$.

3.4.1 Momenty rozdělení

Pro danou veličinu je n-tý moment definován jako

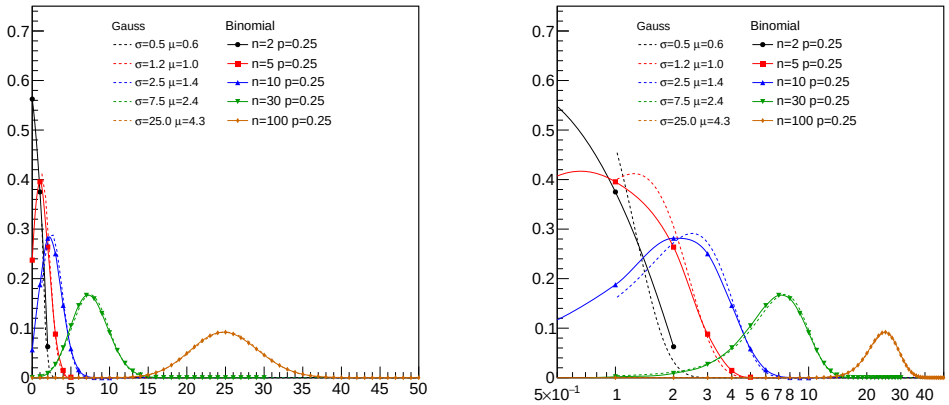
$$m_n \equiv \mathrm{E}[x^n] \ .$$

Speciálně $m_1 \equiv \mu \equiv \mathrm{E}[x]$ je střední (očekávaná, průměrná) hodnota. Centrální momenty jsou definovány okolo očekávané hodnoty

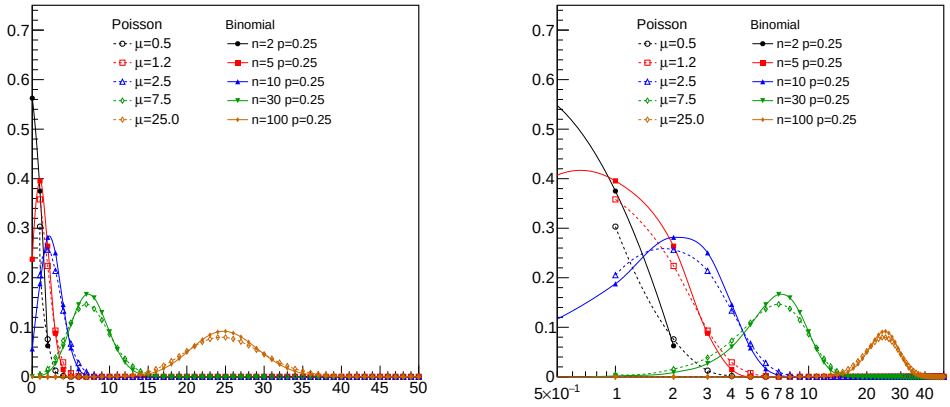
$$c_n \equiv \mathrm{E}[(x-\mu)^n]$$

a speciálně c_2 je variance, kvadrát rozptylu:

$$c_2 \equiv \mathrm{Var}[x] \equiv \sigma^2 \equiv \mathrm{E}\left[(x-\mu)^2\right] = \mathrm{E}\left[x^2\right] - (\mathrm{E}[x])^2$$



Obrázek 7: Srovnání Binomického (plné značky, plně) a Gaussova (čárkovaně) rozdělení pro vybrané hodnoty n a pro $p = 0.25$, pro Poissonovo rozdělení bylo vždy voleno $\mu = np$. Vlevo: lineární měřítko, vpravo: x-logaritmické.



Obrázek 8: Srovnání Binomického (plné značky, plně) a Poissonova (nevyplněné značky, čárkovaně) rozdělení pro vybrané hodnoty n a pro $p = 0.25$, pro Poissonovo rozdělení bylo vždy voleno $\mu = np$. Vlevo: lineární měřítko, vpravo: x-logaritmické.

(dokažte z linearity funkcionálu $\mathrm{E}[\cdot]$ poslední rovnost!) Variance je kvadratickou formou, tj. platí $\mathrm{Var}[ax] = a^2\mathrm{Var}[x]$. Rozptyl je míra šířky daného rozdělení. Jiným příkladem šířky je např. plná šířka v polovině maxima (full width at half maxomum, FWHM), pološířka apod.

Šikmost (angl. skewness) γ_1 rozdělení je bezrozměrná veličina definována jako podíl třetího centrálního momentu vůči třetí mocnině rozptylu rozdělení

$$\gamma_1 \equiv \frac{c_3}{\sigma^3} = \frac{\mathrm{E}\left[(x - \mathrm{E}[x])^3\right]}{\sqrt{\mathrm{E}\left[(x - \mathrm{E}[x])^2\right]}}\,.$$

Šikmost je nulová pro symetrické rozdělení, negativní pro rozdělení s "tailem"doleva, pozitivní pro "tail"rozdělení doprava.

Koeficient špičatosti (angl. kurtosis) je definována pomocí čtvrtého centrálního mo-

memntu jako bezrozměrná veličina

$$\gamma_4 \equiv \frac{c_4}{\sigma^4} = \frac{\mathrm{E}\left[(x - \mathrm{E}\left[x\right])^4\right]}{(E(x - \mathrm{E}\left[x\right])^2)^2}$$

a je nulový pro Gaussovo rozdělení. Rozdělení, které mají tento koeficient kladný mají výraznější "tail"než Gaussovo rozdělení, jsou tedy širší, zatímco rozdělení se záporným koeficientem jsou špičatější. Příkladem je srovnání rozdělení Gaussova Lorenntzova a studentova na Obr. 4.

3.5 Odhad měr polohy a rozptylu z nebinovaného výběru dat a z histogramu

Známý je odhad střední hodnoty veličiny z jejího výběru o N_{evts} prvcích (událostech, "events") jako aritmetický průměr

$$\hat{\mu}_X = \frac{1}{N_{\mathrm{evts}}} \sum_{i=1}^{N_{\mathrm{evts}}} x_i \,.$$

Tzv. nezaujatý odhad rozptylu, tj. standardní odchylka, střední kvadratická fluktuace je pak definován jako

$$\hat{\sigma}_X = \sqrt{\frac{1}{N_{\mathrm{evts}} - 1} \sum_{i=1}^{N_{\mathrm{evts}}} (x_i - \hat{\mu}_X)^2}$$

Pro odhad střední hodnoty veličiny z jejího histogramu o četnostech n_i analogicky definujme průměr vážený četnostmi událostí v daném binu

$$\hat{\mu}_X = \frac{1}{N_{\mathrm{evts}}} \sum_{i=1}^{N_{\mathrm{bins}}} n_i \, x_i^c \,,$$

kde $N_{\mathrm{evts}} = \sum_{i=1}^{n_{\mathrm{bins}}} n_i$ je počet událostí v histogramu a x_i^c je poloha středu i-tého binu. Pro odhad rozptylu z histogramu se používá

$$\hat{\sigma}_X = \sqrt{\frac{1}{N_{\mathrm{evts}} - 1} \sum_{i=1}^{N_{\mathrm{bins}}} n_i \, (x_i^c - \hat{\mu}_X)^2}$$

Tyto v praxi důležité odhady budou podrobněji diskutovány a ospravedlněny později.

Normalizovaný histogram, kde od četností v jednotlivých binech n_i přejdeme k frakcím událostí v každém binu $f_i \equiv n_i / N_{\mathrm{evts}}$ pak můžeme chápat jako binovanou empirickou hustotu pravděpodobnosti, a platí pro něj normalizace $\sum_{i=1}^{n_{\mathrm{bins}}} f_i = 1$. Střední hodnotu pak můžeme spočítat jednoduše jako $\hat{\mu}_X \equiv \sum_{i=1}^{n_{\mathrm{bins}}} x_i^c \, f_i$.

Pro histogram, který není uniformně binován, můžeme zadefinovat operaci vydělení šířkou binu Δ_i, čímž dojdeme k histogramu počtu událostí na šířku binu rovné $s_i \equiv n_i / \Delta_i$, pro které platí normalizační podmínka $\sum_{i=1}^{n_{\mathrm{bins}}} \Delta_i s_i = N_{\mathrm{evts}}$, a s jehož pomocí lze spočítat např. střední hodnotu jako $\hat{\mu}_X \equiv 1/N_{\mathrm{evts}} \sum_{i=1}^{n_{\mathrm{bins}}} x_i^c \, s_i \, \Delta_i$, výraz podobný integrálu, kde místo dx figuruje Δ_i

Můžeme dále uvažovat i o normalizovaném spektru s biny o obsahu $\nu_i \equiv n_i/(\Delta_i\, N_{\mathrm{evts}})$, pro které platí $\sum_{i=1}^{n_{\mathrm{bins}}} \Delta_i \nu_i = 1$ a s jehož pomocí lze spočítat střední hodnotu jako $\hat{\mu}_X \equiv \sum_{i=1}^{n_{\mathrm{bins}}} x_i^c\, \nu_i\, \Delta_i$.

3.6 Vlastnosti hustot pravděpodobnosti, odvozené veličiny

Kumulativní distribuční funkce $F(x)$ je definována jako integrál z hustoty pravděpodobnosti $f(x)$ od nejspodnější meze do určité hodnoty x, např. pro spojitou náhodnou proměnnou nabývající možných hodnot od $-\infty$ do ∞ je

$$F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'$$

Pro diskrétní případ je definice obdobná, ale v sumách, tak např. pro náhodnou diskrétní proměnnou nabývající hodnot přirozených čísel a rozdělenou podle hustoty $f(k)$

$$F(k) = \sum_{k'=0}^{k} f(k')\,.$$

Kvantil je definován následovně: pro dané $\alpha \in (0,1)$ je definován α-kvantil pomocí inverzní funkce ke kumulativní distribuční funkci

$$F^{-1}(x_\alpha) = \alpha\,.$$

Speciálním případem je medián, definován jako $\alpha_{1/2}$. Medián je takový prvek, že polovina očekávaných událostí leží symetricky vlevo a vpravo od něj. Zapište si tuto vlastnost mediánu pomocí integrálů či sum pro spojité a diskrétní rozdělení.

Mód (modus) je prvek s největši hodnotou hustoty pravděpodobnosti, tj. nejpravděpodobnější prvek.

Očekávaná hodnota, medián a modus jsou různé míry polohy, tj. nějakým způsobem významných hodnot, kterých daná veličina typicky nabývá, a jejich odhady jsou také různě citlivé na statistické fluktuace daného datového souboru.

Charakteristická funkce je definována jako Fourierova transformace hustoty pravděpodobnosti

$$\phi(u) \equiv \int_{\Omega} f(x)\exp(\mathrm{i}ux)\,\mathrm{d}x = \mathrm{E}\left[\exp(\mathrm{i}ux)\right]\,.$$

Vidíme, že momenty rozdělení lze vyjádřit jako

$$m_n = \mathrm{i}^{-n}\frac{\mathrm{d}^n\phi(u)}{\mathrm{d}u^n}\bigg|_{u=0}\,.$$

3.7 Více náhodných proměnných

Uvažujme vícerozměrnou náhodnout veličinu $\mathbf{X}$ která může reprezentovat více současně měřených veličin $\mathbf{x} = (x_1, \ldots, x_n)$, např. energie a hynosti jedné či více částic, které můžeme získat v rámci jednoho měření.

Hustotu pravděpodobnosti (tzv. sdruženou) pak píšeme $f(\mathbf{x}|\boldsymbol{\theta})$. Má vlastnost

$$\int\limits_{\Omega_1 \otimes \ldots \otimes \Omega_n} f(\mathbf{x}|\boldsymbol{\theta}) \mathrm{d}\mathbf{x} = 1\,.$$

Věta: Náhodné veličiny X a Y jsou nezávidslé právě tehdy, když je jejich sdružená hustota pravděpodobnosti je faktorizovatelná

$$f(x,y) = g(x) \cdot h(y)\,.$$

což reflektuje známé tvrzení, že náhodné veličiny jsou nezávislé, pokud pravděpodobnost, toho, že nastanou současně, je rovna součinu jejich pravděpodobností.

I pro závislé veličiny můžeme definovat tzv. marginální ("okrajové") hustoty pravděpodobnosti, kde vyintegrujeme přes veličinu, která nás např. nezajímá, a získáme h.p. pro veličinu či veličiny zbývající, pro dvě veličiny např.

$$f_X(x) = \int f(x,y|\boldsymbol{\theta}) \mathrm{d}y\,, \quad f_Y(y) = \int f(x,y|\boldsymbol{\theta}) \mathrm{d}x\,.$$

3.8 Kovariance, korelace

Kovariance mezi dvěma náhodnými veličinami x a y je definována jako

$$\mathrm{Cov}\left[x,y\right] \equiv \mathrm{E}\left[(x-\mu_x)(y-\mu_y)\right] = \mathrm{E}\left[xy\right] - \mathrm{E}\left[x\right]\mathrm{E}\left[y\right]$$

(dokažte si poslední rovnost!) a obecně pro její výpočet musíme znát sdruženou hustotu pravděpodobnosti $f(x,y)$ která obě veličiny spojuje. Lze ukázat, že pro nezávislé veličiny platí, že jejich kovariance je nula.

Pro n náhodných veličin $\{x_i\}_{i=1}^n$ můžeme definovat kovarianční matici

$$\mathrm{Cov}_{ij} \equiv \mathrm{Cov}\left[x_i, y_j\right] \equiv \mathrm{E}\left[(x_i - \mu_{x_i})(x_j - \mu_{x_j})\right]\,.$$

Všimněme si, že diagonální prvky kovarianční matice jsou variance jednotlivých náhodných veličin. Definuje se pak také korelační koeficient, např. pro dvě náhodné veličiny

$$\mathrm{Cor}\left[x,y\right] \equiv \frac{\mathrm{Cov}\left[x,y\right]}{\sqrt{\mathrm{Cov}\left[x,x\right]\mathrm{Cov}\left[y,y\right]}} = \frac{\mathrm{Cov}\left[x,y\right]}{\sqrt{\mathrm{Var}\left[x\right]\mathrm{Var}\left[y\right]}}$$

jenž nabývá hodnot v intervalu

$$\rho_{ij} \equiv \frac{\mathrm{Cov}\left[x_i,x_j\right]}{\sqrt{\mathrm{Var}\left[x_i\right]\mathrm{Var}\left[x_j\right]}} \in [-1,1]\,.$$

Příklad:
Ukažte, že inverze kovarianční matice 2×2

$$\mathrm{Cov} \equiv \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

je

$$\mathrm{Cov}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\sigma_y} \\ -\frac{\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix}$$

Všimněte si, jak je inverze nestabilní pro velkou korelaci, což mimojiné implikuje, že nemá smysl kombinovat silně korelovaná měření (v kombinaci vystupuje právě inverze kovarianční matice) resp. že kombinací silně korelovaných měření chybu nezlepšíme. Spočítejte dále zobecněný chí-kvadrát test

$$\chi^2 \equiv (\boldsymbol{x}_{\mathrm{data}} - \boldsymbol{x}_{\mathrm{theory}})^\top \cdot \mathrm{Cov}^{-1} \cdot (\boldsymbol{x}_{\mathrm{data}} - \boldsymbol{x}_{\mathrm{theory}})$$

a diskutujte jeho závislost na ρ.

3.9 Odhad měr polohy a rozptylu pro vícerozměrná data

Dvourozměrný histogram můžeme chápat jako matici četností n_{ij} pro $i = 1 \dots n^x_{\text{bins}}$, $j = 1 \dots n^y_{\text{bins}}$ kde biny jsou obdélníky definované jako $\Omega_{ij} \equiv (x_i, x_{i+1}) \otimes (y_j, y_{j+1})$.

Odhad střední hodnoty jedné z veličin, např. X, můžeme zadefinovat jako aritmetický průměr, přičemž vysčítáme přes všechny biny veličiny Y:

$$\hat{\mu}_x \equiv \frac{1}{N_{\text{evts}}} \sum_{i=1}^{n^x_{\text{bins}}} \sum_{j=1}^{n^y_{\text{bins}}} n_{ij}\, x_i \,,$$

kde $N_{\text{evts}} \equiv \sum_{i=1}^{n^x_{\text{bins}}} \sum_{j=1}^{n^y_{\text{bins}}} n_{ij}$ je počet všech událostí v histogramu.
Profil histogramu přes veličiny na ose y je graf definován jako následující "funkce"x:

$$\text{Prof}^Y_i \equiv \frac{1}{N^i_y} \sum_{j=1}^{n^y_{\text{bins}}} y^c_j\, n_{ij} \,,$$

kde $N^i_y \equiv \sum_{j=1}^{n^y_{\text{bins}}} n_{ij}$ je počet událostí v řezu histogramu v i-tém binu na ose x s y^c_j jsou středy binů osy y. Jedná se tak o vážený průměr hodnot y v každém binu osy x. Projekce na osu x podél osy y je jednorozměrný histogram s četnostmi

$$\text{Proj}^Y_i \equiv \sum_{j=1}^{n^y_{\text{bins}}} n_{ij} \,,$$

srovnejte si s definicí marginální hustoty pravděpodobnosti. Obdobně si zkuste zapsat odhady střední hodnoty veličin X a Y a odhad jejich ropztylu. Obdobně pro profil a projekci podél osy x, viz ilustrace na Obr. 9.

Pro vícerozměrný nebinovaný náhodný výběr lze definovat odhad kovariance a korelace viz Sekce 6. Pro vícerozměrná binovaná data (např. 2D histogram) můžeme obdobně definovat odhad kovariance jako

$$\widehat{\text{Cov}}[X, Y] \equiv \frac{1}{N_{\text{evts}} - 1} \sum_{i=1}^{n^x_{\text{bins}}} \sum_{j=1}^{n^y_{\text{bins}}} n_{ij}\, (x^c_i - \hat{\mu}_x)(y^c_i - \hat{\mu}_y) \,.$$

kde $N_{\text{evts}} \equiv \sum_{i=1}^{n^x_{\text{bins}}} \sum_{j=1}^{n^y_{\text{bins}}} n_{ij}$, a následně odhadnout i korelační koeficient jako

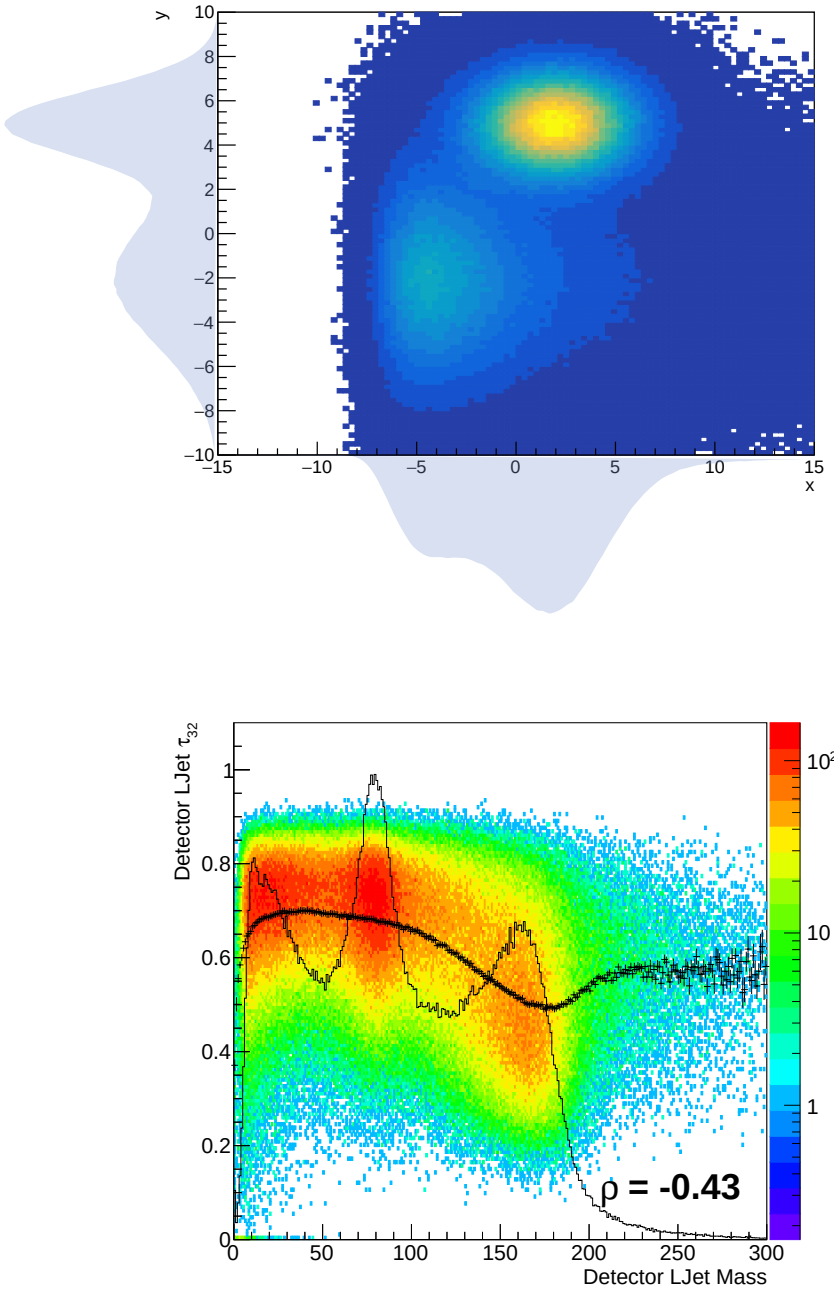
$$\hat{\rho} \equiv \frac{\widehat{\text{Cov}}[X, Y]}{\hat{\sigma}_X \hat{\sigma}_Y} \,,$$

kde $\hat{\sigma}^2_X \equiv \widehat{\text{Cov}}[X, X]$.

3.10 Návrat k průměru

Tvrzení: silně inteligentní ženy mají za partnery méně inteligentní muže? Inspirace: Ondřej Vencálek, Daniel Kahneman: Thinking Fast and Slow. Nagenerována data s korelačním koeficientem 0.65 podle 2D Gaussova rozdělení.

$$f(x|\mu_x, \sigma_x, \mu_y, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y(1-\rho^2)^{1/2}} \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma^2_x} + \frac{(y-\mu_y)^2}{\sigma^2_y} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]}$$



Obrázek 9: Příklad projekce 2D histogramu na osu x a y (nahoře, kredit: R. Přívara), a příklad projekce (černá křivka) jiného 2D histogramu (dole) na osu x a dále profilu (značky) histogramu přes hodnoty na ose y.

Pro funkční závislost lineární regrese platí

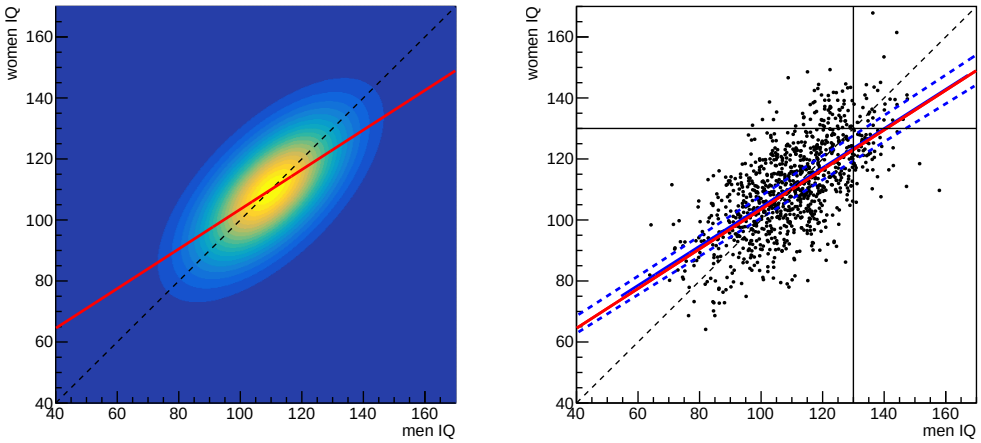
$$y(x) = \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x) + \mu_y \,.$$

Uvědomme si, že lineární regresí "měříme", jak daleko jsou body daleko ve vertikálním směru od přímky, nikoli jak daleko jsou od osy elipsy, a přímka lineární regrese se tak neshoduje s osou elipsy. Na Obr. 10 je černě čárkovaně zobrazena diagonála (*v tomto*

případě, kdy $\sigma_x = \sigma_y$ shodná s osou elipsy), červená resp. modrá přímka je teoretická resp. nafitovaná závislost lineární regrese a modře čárkovaně je zobrazen 1σ pás fitu

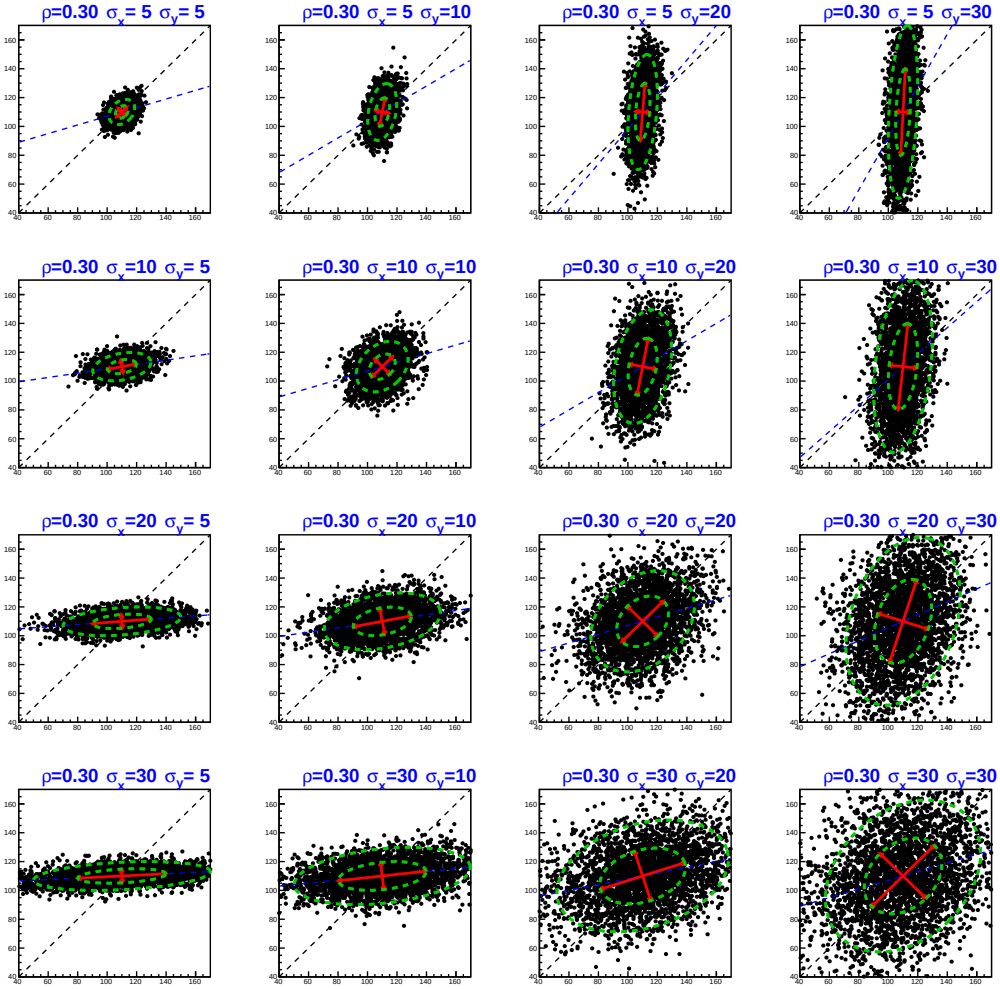
Obecně jsou hlavní osy elipsy dány vlastními vektory kovarianční matice, z které lze n-dimenzionální korelované Gaussovo rozdělení vytvořit následovně:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathrm{Cov}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\mathrm{Cov}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) .$$

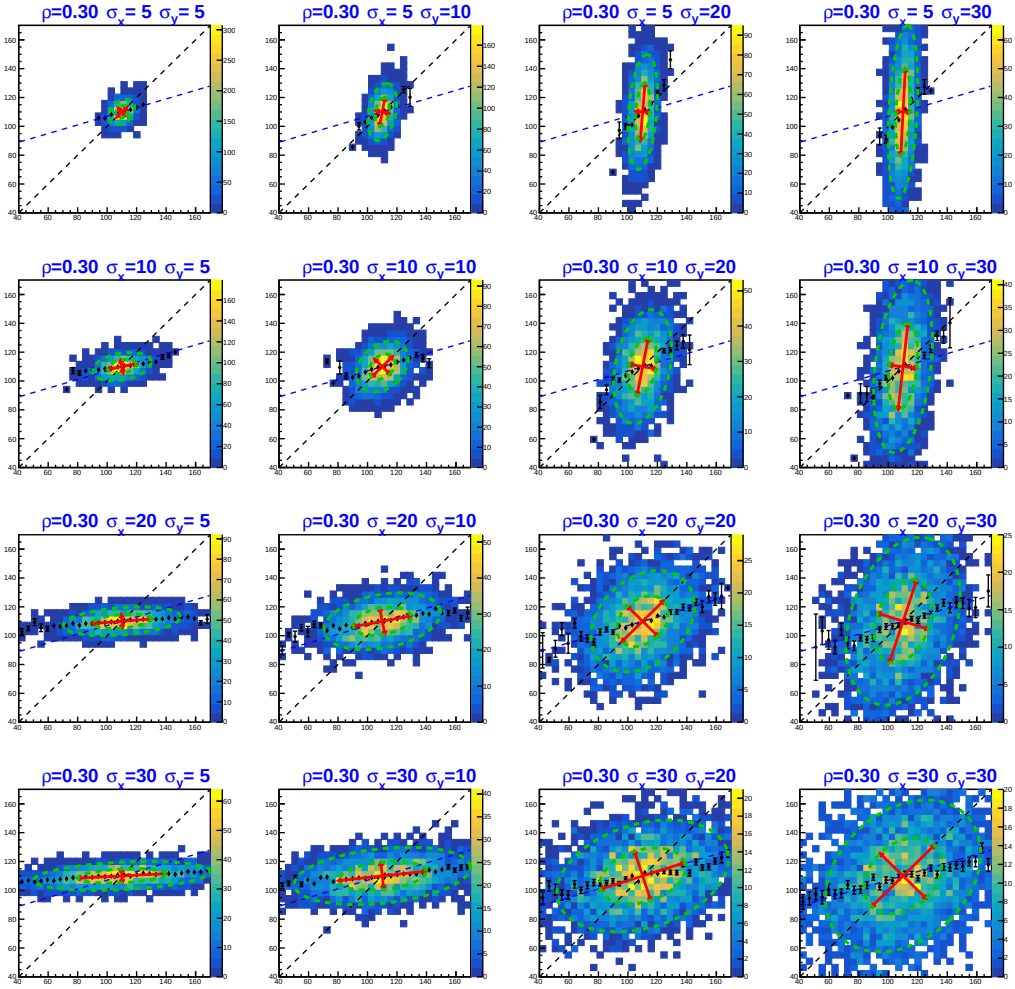


Obrázek 10: Tvrzení: silně inteligentní ženy mají za partnery méně inteligentní muže? Černě čárkovaně je zobrazena diagonála. Červená resp. modrá přímka je teoretická resp. nafitovaná závislost lineární regrese, modře čárkovaně je zobrazen 1σ pás fitu.

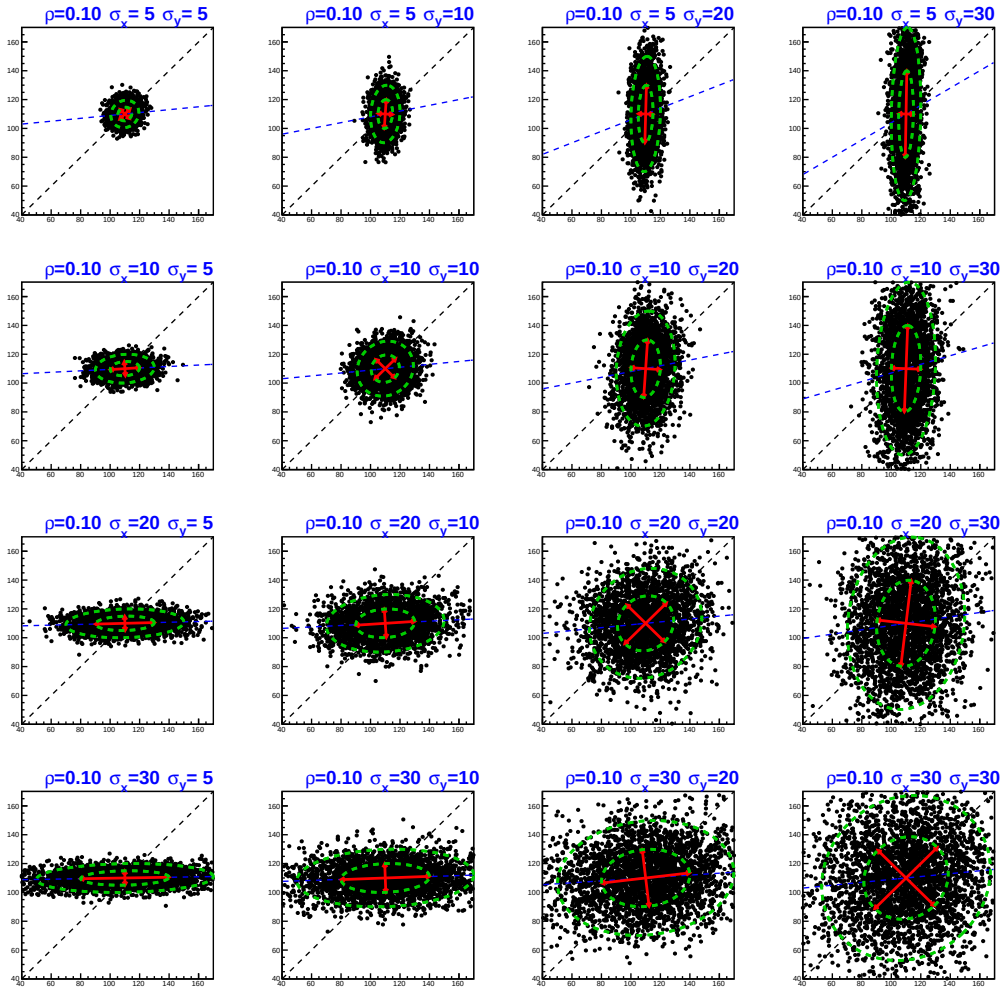
3.11 Data rozdělená podle dvojrozměrného Gaussova rozdělení



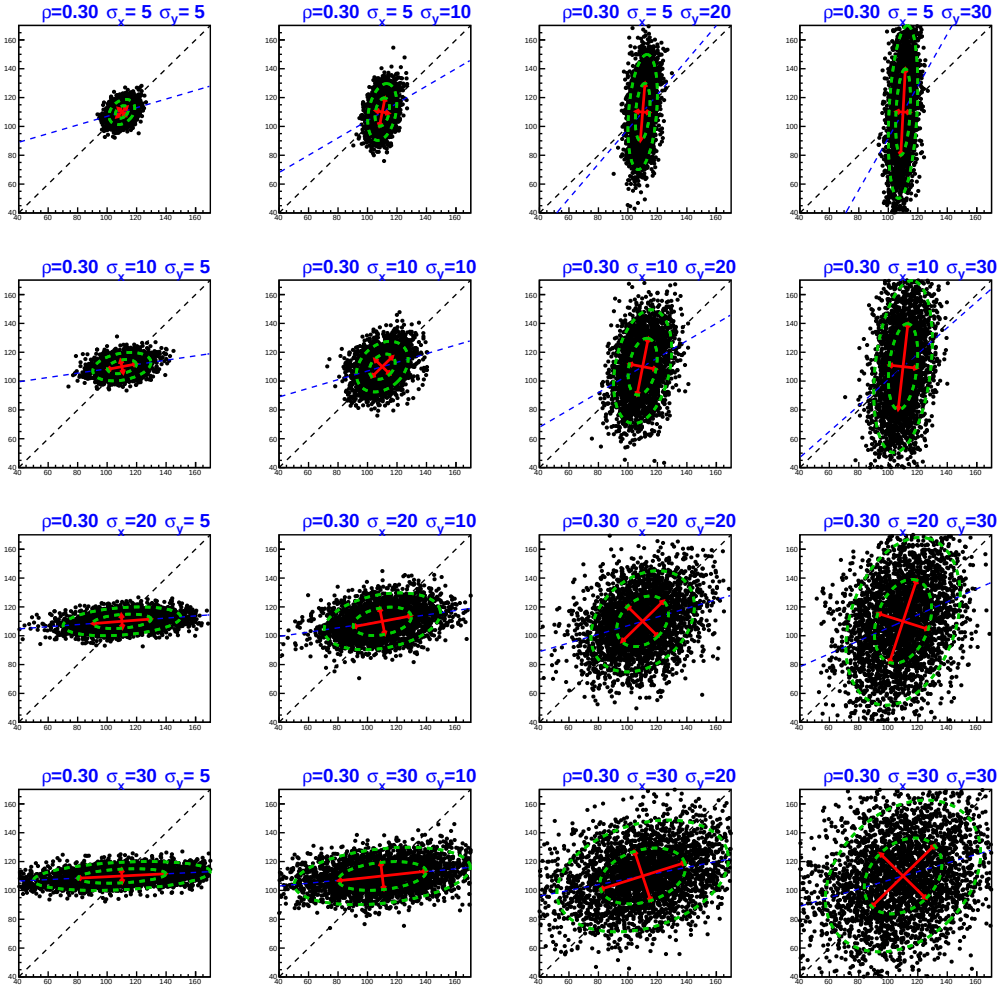
Obrázek 11: Data nagenerována podle dvojrozměrného Gaussova rozdělení pro korelaci $\rho = 0.30$ a $\sigma_x = 10$ a $\sigma_y = 20$. Černě čárkovaně je zobrazena diagonála, modrá čárkovaná přímka je teoretická závislost lineární regrese, červeně jsou zobrazeny osy elipsy délky σ_x a σ_y. Směry os jsou vlastními vektory kovarianční matice. Zeleně jsou zobrazeny 1- a 2-σ elipsy.



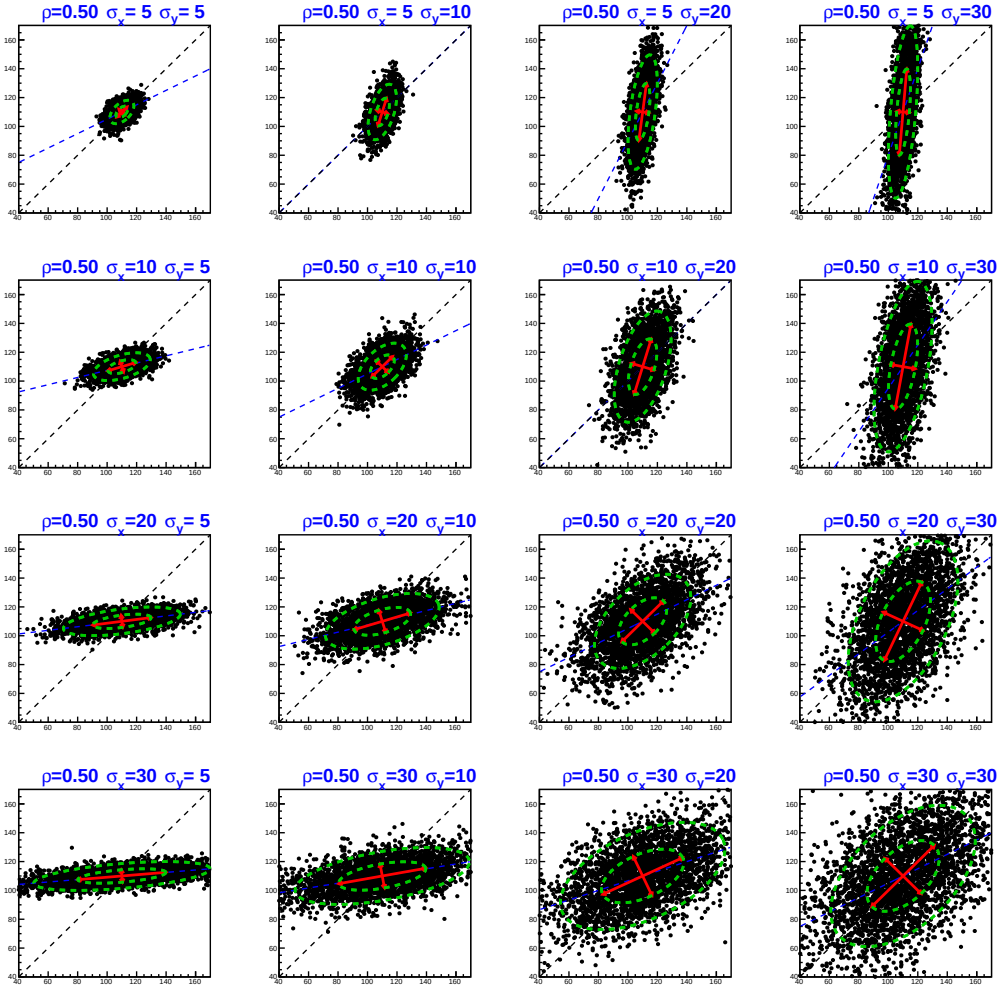
Obrázek 12: Binovaná verze dat nagenerovaných podle dvojrozměrného Gaussova rozdělení pro korelaci $\rho = 0.10$ a parametrů $\sigma_x = 10$ a $\sigma_y = 30$. Černě čárkovaně je zobrazena diagonála, modrá čárkovaná přímka je teoretická závislost lineární regrese, červeně jsou zobrazeny osy elipsy délky σ_x a σ_y. Směry os jsou vlastními vektory kovarianční matice. Zeleně jsou zobrazeny 1- a 2-σ elipsy. Černé body jsou profilem 2D histogramu vykresleného v barevné škále.



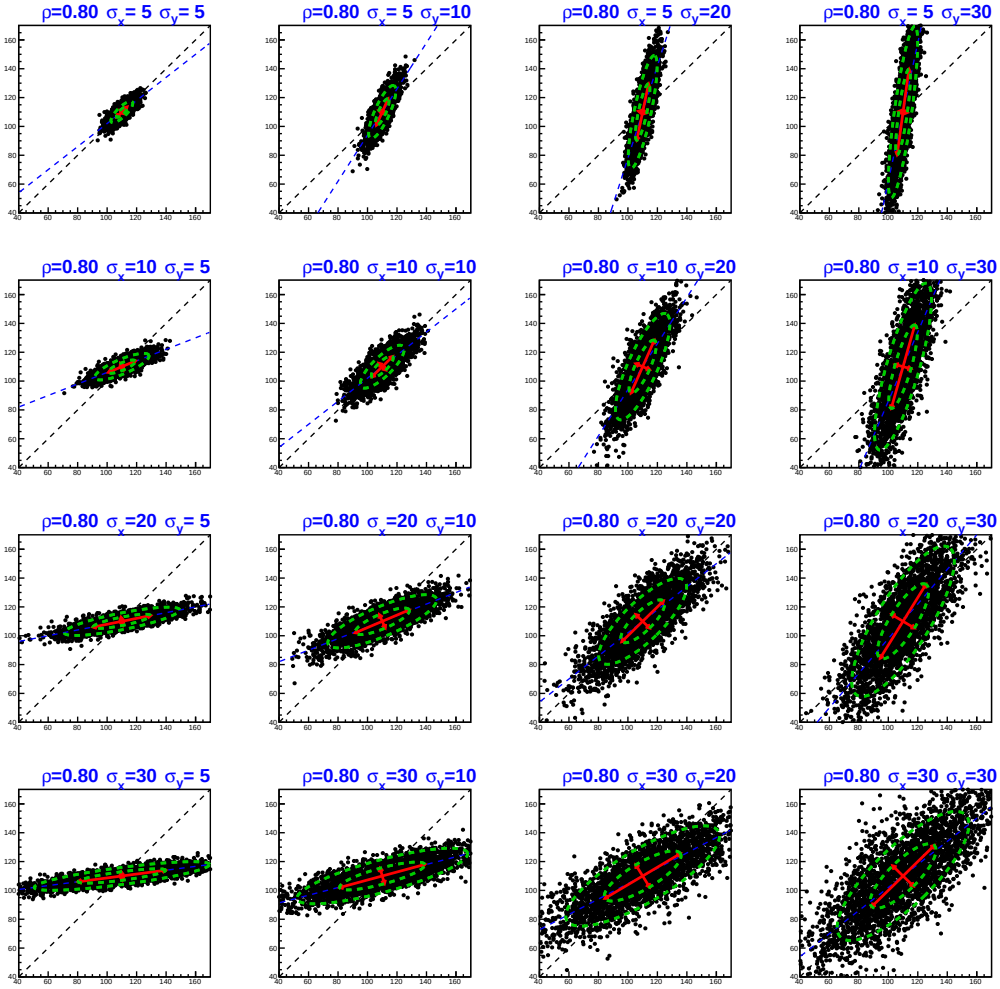
Obrázek 13: Data nagenerována podle dvojrozměrného Gaussova rozdělení pro korelaci $\rho = 0.10$ pro různé hodnoty parametrů σ_x a σ_y. Černě čárkovaně je zobrazena diagonála, modrá čárkovaná přímka je teoretická závislost lineární regrese, červeně jsou zobrazeny osy elipsy délky σ_x a σ_y. Směry os jsou vlastními vektory kovarianční matice. Zeleně jsou zobrazeny 1- a 2-σ elipsy.



Obrázek 14: Data nagenerována podle dvojrozměrného Gaussova rozdělení pro korelaci $\rho = 0.30$ pro různé hodnoty parametrů σ_x a σ_y. Černě čárkovaně je zobrazena diagonála, modrá čárkovaná přímka je teoretická závislost lineární regrese, červeně jsou zobrazeny osy elipsy délky σ_x a σ_y. Směry os jsou vlastními vektory kovarianční matice. Zeleně jsou zobrazeny 1- a 2-σ elipsy.



Obrázek 15: Data nagenerována podle dvojrozměrného Gaussova rozdělení pro korelaci $\rho = 0.50$ pro různé hodnoty parametrů σ_x a σ_y. Černě čárkovaně je zobrazena diagonála, modrá čárkovaná přímka je teoretická závislost lineární regrese, červeně jsou zobrazeny osy elipsy délky σ_x a σ_y. Směry os jsou vlastními vektory kovarianční matice. Zeleně jsou zobrazeny 1- a 2-σ elipsy.



Obrázek 16: Data nagenerována podle dvojrozměrného Gaussova rozdělení pro korelaci $\rho = 0.80$ pro různé hodnoty parametrů σ_x a σ_y. Černě čárkovaně je zobrazena diagonála, modrá čárkovaná přímka je teoretická závislost lineární regrese, červeně jsou zobrazeny osy elipsy délky σ_x a σ_y. Směry os jsou vlastními vektory kovarianční matice. Zeleně jsou zobrazeny 1- a 2-σ elipsy.

3.12 Transformace proměnných

Je-li X náhodná proměnná s hustotou pravděpodobnosti $f(x)$, jaká je hustota pravděpodobnosti $g(y)$ pro náhodnou proměnnou $y = h(x)$? Omezíme se zatím na případ, že jde vzájemně jednoznačné zobrazení, tj. že existuje $x = h^{-1}(y)$. Porovnáním infinitezimálních pravděpodobností, které si musejí být rovny z monotónnosti zobrazení $y = h(x)$ obdržíme

$$f(x)\mathrm{d}x = g(y)\mathrm{d}y$$

$$g(y) = f(x)\frac{\mathrm{d}x}{\mathrm{d}y}\,.$$

Fyzik má tendenci zapsat

$$g(y) = f(x(y))\frac{\mathrm{d}x(y)}{\mathrm{d}y} = f(x(y))\frac{1}{\frac{\mathrm{d}y}{\mathrm{d}x}} = f(x(y))\frac{1}{\frac{\mathrm{d}h}{\mathrm{d}x}}$$

či snad o něco lépe

$$g(y) = f(h^{-1}(y))\frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y}$$

správněji však

$$g(y) = \frac{f(h^{-1}(y))}{\left|\frac{\mathrm{d}h(h^{-1}(y))}{\mathrm{d}y})\right|} = f(h^{-1}(y))\left|\frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y}\right|$$

Jde vlastně o větu o substituci, s funkcí nakládáme jako s výrazem pod integrálem. Výraz $\frac{\mathrm{d}x}{\mathrm{d}y}$ je Jakobiánem $\mathcal{J}$ dané transformace proměnných.

Příklad:
Má-li náhodná veličina $x \in (-1, 1)$ hustotu pravděpodobnosti $f(x)$, jakou hustotu pravděpodobnosti má náhodná veličina $y = h(x) = \cos x$?

Ve více rozměrech pak pro funkci $\mathbb{R}^n \to \mathbb{R}^n$: $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$, čímž rozumíme n funkcí $y_i = h_i(\boldsymbol{x})$ a jejich inverze $x_i = h_i^{-1}(\boldsymbol{y})$ příp. ve zkratce jako $x_i = x_i(\boldsymbol{y})$

$$\mathcal{J}[\boldsymbol{h}^{-1}] = \det\begin{pmatrix} \partial x_1/\partial y_1 & \cdots & \partial x_1/\partial y_n \\ \vdots & \ddots & \vdots \\ \partial x_n/\partial y_1 & \cdots & \partial x_n/\partial y_n \end{pmatrix}$$

Příklad:
spočítejte si Jakobián pro přechod od kartézských k polárním a sférickým souřadnicím. Všimněte si, že máme výhodu, že obvyklá definice těchto nových souřadnic je přímo sadou inverzních funkcí, které potřebujeme na výpočet Jakobiánu, zatímco obecně (a v definici $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$) jsou "nové"sořadnice funkcí "starých"a je potřeba najít inverzní transformaci, anebo použít větu o derivaci inverzní funkce.

Příklad:
Z Planckova zákona pro spektrální hustotu černého tělesa jako funce kruhové frekvence

$$\frac{\mathrm{d}N}{\mathrm{d}\omega} = \frac{1}{\pi^2 c^3}\frac{\hbar\omega^3}{\exp\left(\frac{\hbar\omega}{kT}\right) - 1}$$

odvoďte obdobný výraz jako funkci vlnové délky.

3.13 Odhadování h.p. pomocí kernelu

Odhad hustoty pravděpodobnosti na základě binovaných dat pomocí kernelu [9].

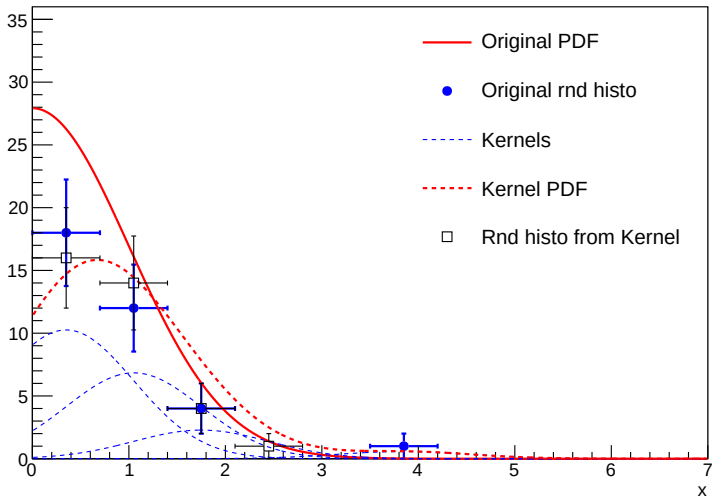
Nebinovaná verze pro N událostí s volitelným zjemňovacím parametrem h

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}h} \exp\left[-\frac{(x - x_i)^2}{2h^2}\right] .$$

Binovaná verze pro n_b binů s volitelným zjemňovacím parametrem h

$$f(x) = \frac{1}{N_y} \sum_{i=1}^{n_b} \frac{y_i}{\sqrt{2\pi}h} \exp\left[-\frac{(x - x_i)^2}{2h^2}\right] ,$$

kde y_i je zaznamenaná četnost událostí v i-tém binu o středu x_i a $N_y = \sum_{i=1}^{n_b} y_i$. Příklad viz Obr. 17 pro histogram nagenerovaný z původní h.p. a z kernel PDF pro h rovno šířce binu. Diskutujte vznik a existenci píku blízko počátku takto kernelovsky odhadnuté h.p. na tomto příkladě.

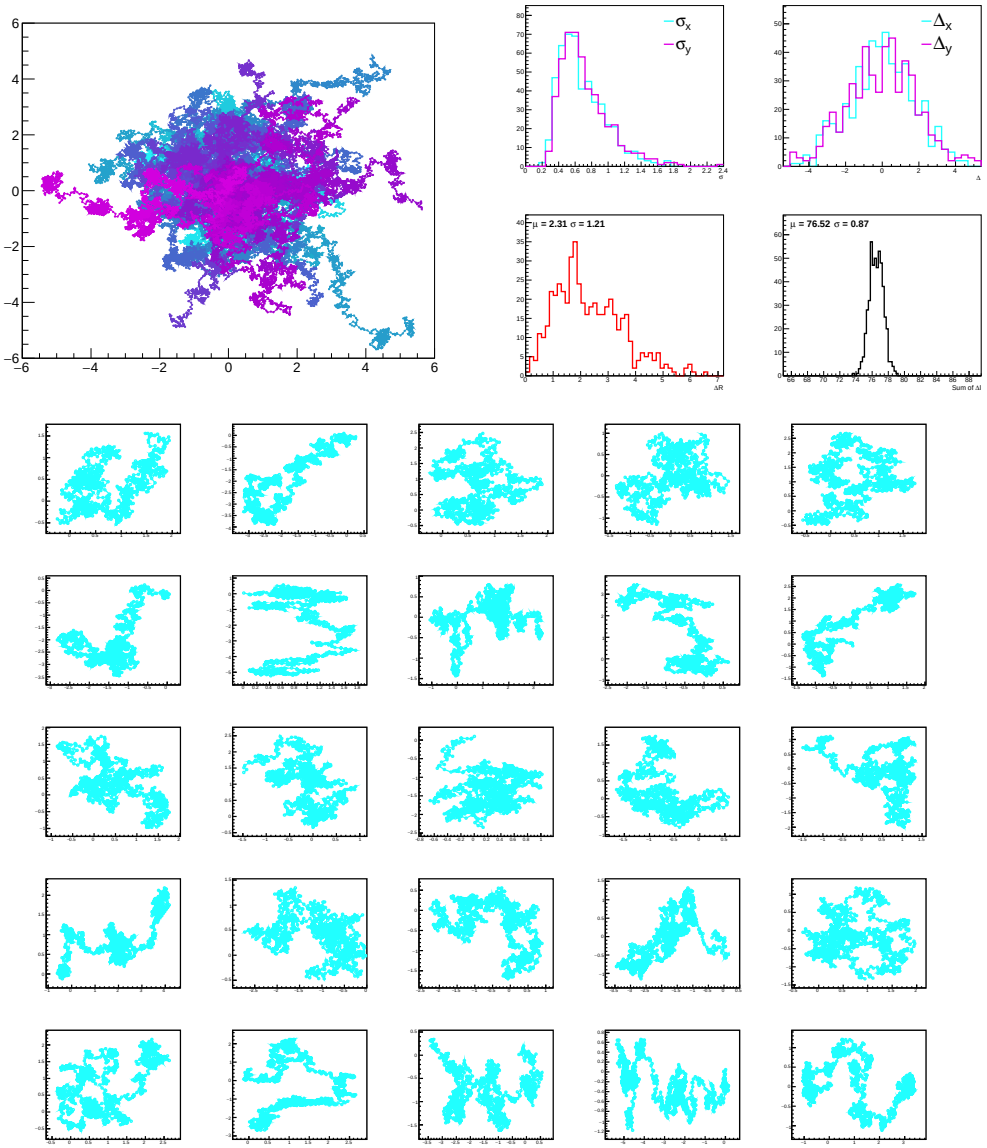


Obrázek 17: Ilustrace odhadu hustoty pravděpodobnosti z histogramu pomocí gaussovských kernelů.

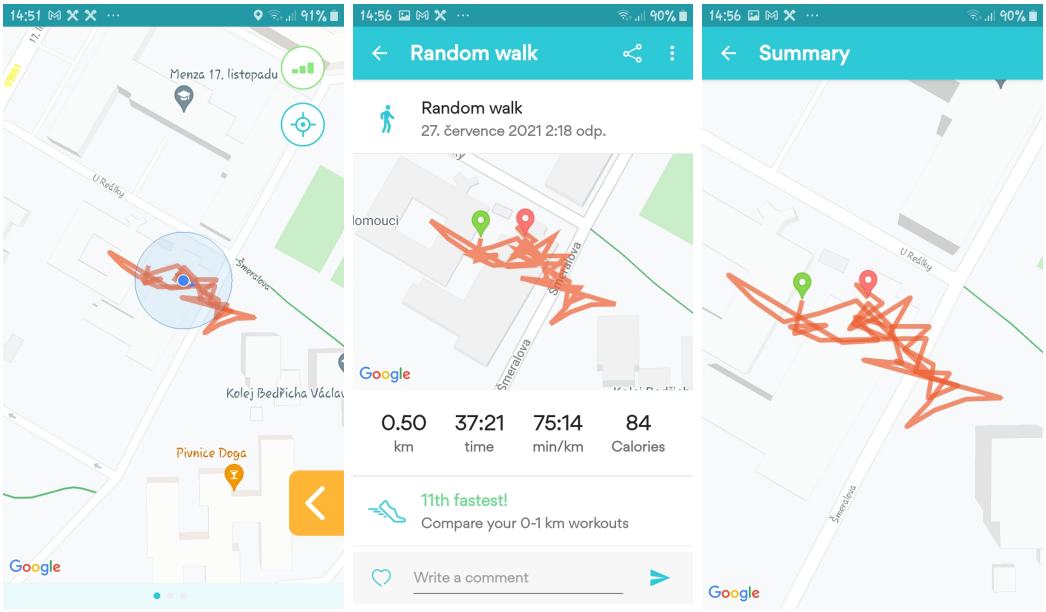
3.14 Náhodná chůze

Náhodná chůze alias Random walk. Ukázka různých náchodných 2D procházek s 1000 kroky, s délkou kroku dle náhodného výběru z uniformního rozdělení na intervalu $[-0.1, 0.1]$, je na Obr. 18.

Náhodná procházka, jak ji zaznamenala aplikace RunKeeper z mobilního telefonu dle jeho GPS je na Obr. 19, s vyznačenou chybovou elipsou. Za 37 min takto telefon na parapetu "ušel"půl kilometru. Všimněte si protáhlosti procházky v jednom směru.



Obrázek 18: Ukázka 50 různých náchodných 2D procházek s 1000 kroky, s délkou kroku dle náhodného výběru z uniformního rozdělení na intervalu $[-0.1, 0.1]$ (nahoře vlevo). Individuální procházky jsou zobrazeny dole, vpravo nahoře pak rozptyly rozdělení souřadnic x a y mezikroků přes 500 procházek, střední hodnota výsledné pozice podél obou os, výsledné posunutí na konci procházky ΔR, a součet délky segmentů drah, opět jako rozdělení přes 500 různých procházek.



Obrázek 19: Náhodná procházka, jak ji zaznamenala aplikace RunKeeper z nehybného mobilního telefonu dle jeho GPS, s výhledem z okna budovy směrem na severozápadní stranu, s vyznačenou chybovou elipsou.

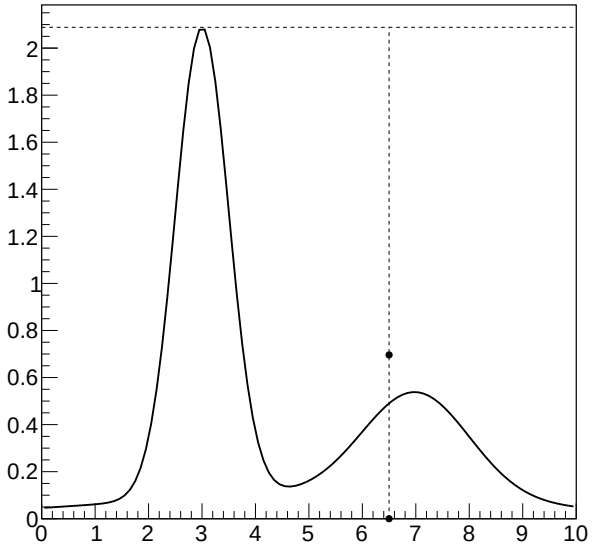
4 Monte Carlo metoda

4.1 von Neumannova Monte Carlo metoda

Cílem je nagenerovat sadu (pseudo)náhdných čísel $\{x_i\}$ podle dané h.p. $f(x)$. Metoda je založena na znalosti této funkce a jejího maxima, nebo alespoň horní meze, $f_{\max}$, a možnosti generovat rovnoměrně rozdělenou náhodnou veličinu. Algoritmus je následující:

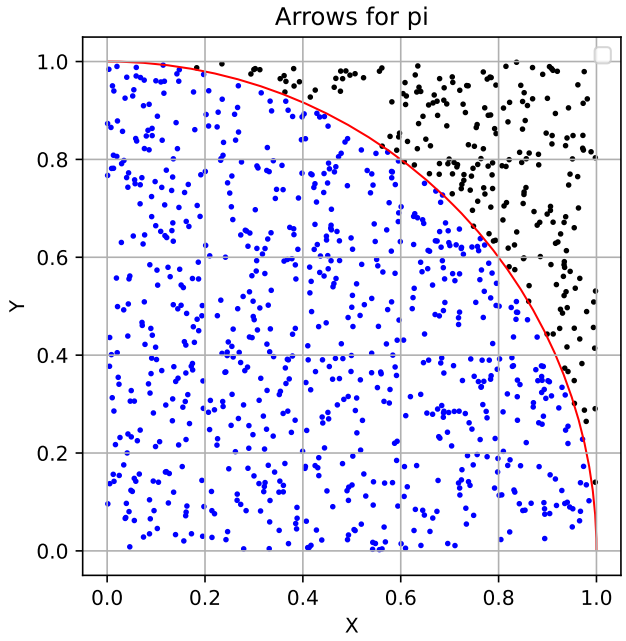
1. Vyber x z rovnoměrného rozdělení $x \sim U[x_{\min}, x_{\max}]$
2. Vyber y z rovnoměrného rozdělení $y \sim U[0, f_{\max}]$
3. Je-li $f(x) > y$, akceptuj x, jinak zahoď.

Nakreslete si obrázek a nazřete, viz Obr. 20. Generování touto metodou je jednoduché, ale má malou efektivitu např. pro rychle klesající $f(x)$, kdy generujeme x a y, abychom je ve většině případů zase zahodili. Proto se hodí rozdělení anebo x vhodně transformovat.



Obrázek 20: Ilustrace von Neumannovy Monte Carlo metody.

Pomocí této metody lze získat odhad čísla π výpočtem frakce událostí uvnitř kruhu za použití bodů z výběru z uniformního rozděleni, viz Obr. 21 V tabulce 3 jsou srhnuty výsledky pro různé počty nagenerovaných událostí. Zkuste si sami ve svém oblíbeném programovacím jazyce!



Obrázek 21: Ilustrace náhodných událostí vybraných z uniformního rozdělení pro výpočet odhadu čísla π.

N	odhad π
10	3.200000
100	2.840000
1,000	3.148000
10,000	3.147600
100,000	3.146680
1,000,000	3.144896
10,000,000	3.141090
100,000,000	3.141648
1,000,000,000	3.141605

Tabulka 3: Výsledky odhadu čísla π pro různé počty nagenerovaných událostí.

4.2 Generování spektra pomocí transformační funkce

Kredit: Oldřich Kepka.
Cílem je najít transformační funkci

$$u = u(x)$$

která transformuje např. uniformně rozdělenou náhodnou veličinu x na námi požadované rozdělení pro náhodnou veličinu u.Veličina x je rozdělena obecně podle hustoty pravděpodobnosti $f(x)$, zatímco veličina u má požadované rozdělení $g(u)$. Ekvivalentně hledáme inverzi vztahu $u = u(x)$, neb pro dané uniformně rozdělené x chceme najít odpovídající u.

Jde o transformaci proměnných, na kterou se nyní podíváme jazykem kumulativních distribučních funkcí: je-li u monotónní funkcí x, pak musí platit, že pravděpodobnost toho, že $x' < x$ a $u' < u$ kde $x' = x(u')$ a $x = x(u)$ (trochu neobratně zde symbol x používáme i pro funkci), tj. musí být rovna

$$F(x) = G(u)\,,$$

Z rovnosti distribučních funkcí dostáváme

$$u(x) = G^{-1}\left(F(x)\right)\,.$$

Ve speciálním případě uniformně rozdělené veličiny x je tedy $f(x) = U(x; x_1, x_2)$, je primitivní funkce $F(x) = \frac{x-x_1}{x_2-x_1}$.

Intuitivní "fyzikální postup"spočívá ve zderivování podle u (na levé straně jako složenou funkci), kdy dostaneme

$$g(u) = f(x(u))\frac{\mathrm{d}x}{\mathrm{d}u} = f(x(u))\,x'(u)\,.$$

Ekvivalentně dojdeme ke stejnému výsledku o něco méně exaktněji porovnáním infinitezimálních pravěpodobností $f(x)\mathrm{d}x = g(u)\mathrm{d}u$.

$$\int_{x_1}^{x(u)} f(x')\mathrm{d}x' = \int_{u_1}^{u} g(u')\,\mathrm{d}u'\,,$$

odkud je potřeba si vyjádřit $u(x)$.

Následně hledejme předpisu pro transformaci od x k u, známe-li rozdělení $g(u)$, jakému má u "podléhat".

Příklad:
Nalezněte $u = u(x)$ pro transformaci $g(u) := \frac{1}{\tau}\mathrm{e}^{-u/\tau}$ pro uniformně rozdělené $x \in (x_1, x_2)$ a pro $u \in (0, \infty)$. Výsledky:

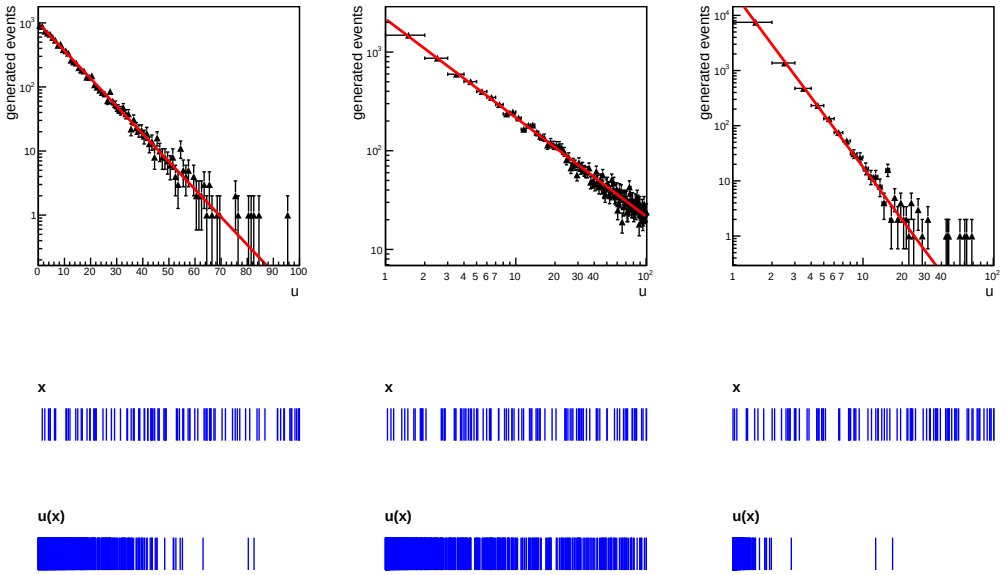
$$G(u) = 1 - \mathrm{e}^{-u/\tau}\,, \quad u(x) = \tau\,\ln\frac{x_2 - x_1}{x_2 - x}$$

Příklad:
Nalezněte $u = u(x)$ pro transformaci $g(u) := C/u$ pro uniformně rozdělené $x \in (x_1, x_2)$ a pro $u \in (u_1, u_2)$ (musíme se omezit na konečný interval z důvodu konečnosti normalizace). Nejprve nalezněte normalizační konstantu C. Výsledky:

$$G(u) = \frac{\ln\frac{u}{u_1}}{\ln\frac{u_2}{u_1}}\,, \quad u(x) = u_1\left(\frac{u_2}{u_1}\right)^{\frac{x-x_1}{x_2-x_1}}$$

Příklad:
Nalezněte $u = u(x)$ pro transformaci $g(u) := C/u^\alpha$ $(\alpha > 1)$ pro uniformně rozdělené $x \in (x_1, x_2)$ a pro $u \in (u_1, \infty)$. Nejprve opět nalezněte normalizační konstantu C. Výsledky:

$$G(u) = 1 - \left(\frac{u_1}{u}\right)^{\alpha-1}, \quad u(x) = u_1 \left(\frac{x_2 - x}{x_2 - x_1}\right)^{-\frac{1}{\alpha-1}}$$



Obrázek 22: Rozdělení (zleva doprava) $\sim \exp[-u/\tau]$, $\sim 1/u$ a $\sim 1/u^\alpha$, $\alpha > 1$ nagenerovaná pomocí transformační funkce z uniformního rozdělení. Rozmyslete si linearitu mocninných rozdělení v dvojtě logaritmickém měřítku.

5 Funkce náhodných proměnných

5.1 Funkce náhodných proměnných

Mějme dvě nezávislé náhodné veličiny X a Y rozdělené podle nějakých hustot pravděpodobnosti, a uvažujme také jejich kumulativní distribuční funkce

$$X \sim f(x)\,, \quad F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'\,,$$

$$Y \sim f(y)\,, \quad F(y) = \int_{-\infty}^{y} f(y')\,\mathrm{d}y'\,.$$

Zajímá nás hustota pravděpodobnosti funkce náhodných veličin $z = z(x, y)$, tj. hledáme $h(z)$ tak, že

$$Z \sim h(z)\,, \quad H(z) = \int_{-\infty}^{z} h(z')\,\mathrm{d}z'$$

Příklad:
Hustota pravděpodobnosti součtu dvou náhodných proměnných je dána konvolucí jejich hustot pravděpodobnosti, tj, pro $z = x + y$ platí, že její h.p. $h(z)$ je

$$h(z) = (f * g)(z) = \int_{\mathbb{R}} f(x)g(z - x)\,\mathrm{d}x = \int_{\mathbb{R}} f(z - y)g(y)\,\mathrm{d}y$$

Tj. konvoluce je symetrická, a má další zajímavé vlastnosti, např. její Fourierův obraz je roven součinu jednotlivých Fourierových obrazů $\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g)$.

Odvození konvoluce

1. pomocí kumulativní distribuční funkce:

   $$h(z) = H'(z)$$

   $$H(z) = \int_{\mathbb{R}} \int_{y'<y(x)} f(x)g(y')\,\mathrm{d}y'\mathrm{d}x = \int_{\mathbb{R}} f(x) \int_{-\infty}^{z-x} g(y')\,\mathrm{d}y'\mathrm{d}x = \int_{\mathbb{R}} f(x)G(z-x)\mathrm{d}x$$

   a tedy

   $$h(z) = \frac{\mathrm{d}}{\mathrm{d}z} \int_{\mathbb{R}} f(x)G(z - x)\,\mathrm{d}x = \int_{\mathbb{R}} f(x)\frac{\partial}{\partial z}G(z - x)\,\mathrm{d}x = \int_{\mathbb{R}} f(x)g(z - x)\,\mathrm{d}x$$

   kde podmínka na y je dána dle $z \geq x + y$ vztahem $y < z - x$

2. pomocí delta funkce, která zaručí vztah $z - x - y = 0$

   $$h(z) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x)g(y)\delta(z - x - y)\mathrm{d}x\mathrm{d}y = \int_{\mathbb{R}} f(x)g(x - z)\,\mathrm{d}x$$

3. pomocí infinitezimálních pravděpodobností: integrujeme plochu mezi přímkami $y = z - x$ a $y = z + \mathrm{d}z - x$

   $$\mathrm{d}P_{z\in(z',z'+\mathrm{d}z')} \equiv h(z')\,\mathrm{d}z' = \int_{\mathbb{R}} \int_{z'-x}^{z'+\mathrm{d}z'-x} f(x)g(y)\,\mathrm{d}y\mathrm{d}x = \left(\int_{\mathbb{R}} f(x)g(z' - x)\,\mathrm{d}x \right) \mathrm{d}z'$$

   a tedy opět

   $$h(z) = \int_{\mathbb{R}} f(x)g(z - x)\,\mathrm{d}z$$

f	g	$f * g$	aplikace
Gauss	Gauss	Gauss	Gaussovské signály
Gauss	Heaviside	error function	trigger turn-on
Gauss	Lorentz (Breit-Wigner)	Voigt	spektroskopie
Gauss	Landau	nelze analyt. funkcemi	silně fluktuující signál
Poisson$(\cdot\|\mu)$	Binomial$(\cdot\|p;N)$	Poisson$(\cdot\|p\mu)$	

Tabulka 4: Příklady konvolucí a jejich praktické aplikace.

V praxi se často setkáváme s konvolucemi několika významných hustot pravděpodobnosti, jejichž výsledky a aplikace jsou uvedeny v Tabulce 4.

Příklad:
Ukažte, že konvoluce dvou Gaussových rozdělení je opět Gaussovo rozdělení s parametry $\mu = \mu_1 + \mu_2$ a $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$, tj. že hustota pravděpodobnosti pro náhodou veličinu, jež je součtem dvou gaussovsky rozdělených náhodných veličin, je

$$(g(\cdot|\mu_1,\sigma_1) * g(\cdot|\mu_2,\sigma_2))(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left[-\frac{(x-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]$$

Srovnejte výsledný rozptyl s výsledkem pro propagaci chyb součtu dvou náhodných veličin.

Příklad:
Ukažte, že konvoluce dvou normalizovaných exponenciálních funkcí $f(x) = a\,\exp(-ax)$ a $g(x) = b\,\exp(-bx)$, kde $x \in (0,\infty)$ je hustota pravděpodobnosti

$$h(x) = (f * g)(x) = \frac{ab}{b-a}\left[\exp(-ax) - \exp(-bx)\right]$$

která může např. popisovat rozpad částice spolu s konečnou dobou života excitovaného stavu ve scintilátoru, kterým detekujeme rozpad. Nápověda: uvědomte si, že h.p. můžeme zapsat také jako $f(x) = \Theta(x)\,a\,\exp(-ax)$ a $g(x) = \Theta(x)\,b\,\exp(-bx)$, kde $x \in (\infty,\infty)$ a skoková funkce $\Theta(x)$ je rovna jedné pro $x > 0$ a jinak je nulová.

Příklad:
Motivace: velikost chybějící příčné energie v kalorimetru je dána velikostí 2D vektoru chybějící energie v příčné rovině experimentů na LHC:

$$\not{E}_{\mathrm{T}} \equiv \sqrt{\not{E}_{\mathrm{x}}^2 + \not{E}_{\mathrm{y}}^2}\,.$$

Jaké je rozdělení $\not{E}_{\mathrm{T}}$, jsou-li její složky rozděleny Gaussovsky?
Ukažte tedy, že pro náhodnout veličinu $z = \sqrt{x^2+y^2}$ sestavenou z náhodných veličin x a y z nichž každá je rozdělena podle Gaussova rozdělení se středem v nule a o šířce sigma, tj. $f(x) = (2\pi\sigma^2)^{-1/2}\,\exp[-x^2/(2\sigma^2)]$, lze získat analytickou hustotu pravděpodobnosti $h(z)$. Přesvědčte se, že jde o normalizovanou h.p. Postupujte následovně:

$$h(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(z - \sqrt{x^2+y^2}) f(x)\,f(y))\,\mathrm{d}x\mathrm{d}y\,.$$

Výsledek:

$$E_{\mathrm{T,y}}^{\mathrm{miss}} \sim h(z) \propto z\,\exp\left(-\frac{z^2}{2\sigma^2}\right)\,.$$

Příklad:
Ukažte, že pro náhodnou veličinu, která je součinem náhodných veličin, tj. $z = xy$, je hustota pravděpodobnosti dána tzv. Mellinovou konvolucí

$$h(z) = \int_{\mathbb{R}} f(x)g(z/x)\frac{\mathrm{d}x}{|x|}$$

Pokud budete počítat přes delta funkci, využijte toho, že $\delta(ax) = \frac{1}{|a|}\delta(x)$.
Příklad:
Odvoďte hustotu pravděpodobnosti pro podíl dvou náhodných veličin $Z = X/Y$ a ukažte, že pro normálně rozdělené veličiny $X, Y \sim \mathrm{Gauss}(0, \sigma)$ je hustota pravděpodobnosti veličiny Z dána Cauchyho rozdělením.
Příklad:
Spočítejte si hustotu pravděpodobnosti pro náhodnou veličinu, jež je dána součtem příspěvků od signálu a šumu, tj. konvoluci ideálního signálu rozděleného jako $X \sim \delta(x - x_0)$ se šumem rozděleného podle Gaussova rozdělení $Y \sim \mathrm{Gauss}(0, \sigma)$. Zauvažujte, jak by vypadala konvoluce reálného signálu (Lorentzovská čára) a šumu.

5.2 Propagace (šíření) chyb

Tímto zavádějícím titulem myslíme řešení následujícího problému:
Nechť z je funkcí náhodných proměnných x_i. Známe-li variance ("chyby") σ_{x_i}, jaká je "chyba"(variance) veličiny z? Jde tedy o výpočet variance či rozptylu funkce náhodných proměnných. Motivační příklad: měřením hybnosti v dráhovém detektoru a energie v kalorimetru můžeme usoudit na hmotu pozorované nabité částice. Zajímá nás, jaký je rozpyl hmoty, je-li

$$m \equiv m(p, E) = \sqrt{E^2/c^4 - p^2/c^2}\,.$$

Obecné řešení: Funkci $z \equiv g(\boldsymbol{x})$ si rozvedeme do Taylorovy řady kolem střední hodnoty. Taylorův rozvoj funkce z kolem bodu $\boldsymbol{\mu}$ pak zapíšeme jako

$$z \equiv g(\boldsymbol{x}) = g(\boldsymbol{\mu}) + \sum_i \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)\cdot(x_i - \mu_i) + \mathcal{O}\left((x_i - \mu_i)^2\right) = g(\mu) + \nabla g(\mu)\cdot(\boldsymbol{x} - \boldsymbol{\mu}) + \mathcal{O}\left((x_i - \mu_i)^2\right)$$

(poslední výraz je zapsán za použití skalárního součinu gradientu funkce g a vektoru posunutí $\boldsymbol{x} - \boldsymbol{\mu}$).

Nejprve budeme potřebovat přibližný výraz pro $\mathrm{E}[z]$, kde, jak se ukáže, bude potřeba jít do druhého řádu Taylorova rozvoje (a pro výpočet variance jej nakonec stejně zanedbáme). Pro funkci jedné proměnné máme

$$\mathrm{E}[z] = \mathrm{E}\left[g(\mu) + g'(\mu)(x - \mu) + \frac{1}{2!}g''(\mu)(x - \mu)^2 + \mathcal{O}\left((x - \mu)^3\right)\right].$$

Druhý člen je však nulový (derivace je vyhodnocena v bodě $x = \mu$ a je tedy konstanta, a $\mathrm{E}[x] = \mu$. Střední hodnota z je tedy $g(\mu)$. Se zahrnutím korekce druhého řádu je přesněji

$$\mathrm{E}[z] \approx g(\mu) + \frac{1}{2}g''(\mu)\mathrm{Var}[x]\,.$$

Tato skutečnost platí i ve více dimenzích, čehož využijeme.
Variance z je dle definice

$$\mathrm{Var}[z] = \mathrm{E}\left[(z - \mathrm{E}[z])^2\right]$$

$$\mathrm{Var}\left[z\right] = \mathrm{E}\left[\left(g(\boldsymbol{\mu}) + \sum_i \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)(x_i - \mu_i) - \mathrm{E}\left[z\right]\right)^2\right] .$$

Prostým roznásobením pak

$$\mathrm{Var}\left[z\right] = \mathrm{E}\left[\left(g(\boldsymbol{\mu}) + \sum_i \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)(x_i - \mu_i) - \mathrm{E}\left[z\right]\right)\left(g(\mu) + \sum_j \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_j}\right)(x_j - \mu_j) - \mathrm{E}\left[z\right]\right)\right]$$

a rozepsáním $\mathrm{E}\left[z\right]$ do prvního řádu v $x_i - \mu_i$, tj. $\mathrm{E}\left[z\right] \approx g(\boldsymbol{\mu})$

$$\mathrm{Var}\left[z\right] \approx \mathrm{E}\left[\left(\sum_i \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)(x_i - \mu_i)\right) \cdot \left(\sum_j \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_j}\right)(x_j - \mu_j)\right)\right]$$

a tedy

$$\mathrm{Var}\left[z\right] \approx \sum_{i,j} \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)\left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_j}\right) \mathrm{E}\left[(x_i - \mu_i)(x_j - \mu_j)\right]$$

$$\mathrm{Var}\left[z\right] \approx \sum_{i,j} \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)\left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_j}\right) \mathrm{Cov}\left[x_i, x_j\right] .$$

Tento známý výsledek můžeme zapsat do přehledné formy

$$\sigma_z \approx \sqrt{\sum_{i,j} \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)\left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_j}\right) \mathrm{Cov}\left[x_i, x_j\right]} .$$

Pro nekorelované veličiny, tj. takové, pro které $\mathrm{Cov}\left[x_i, x_j\right] = 0$ pro $i \neq j$ a $\mathrm{Cov}\left[x_i, x_i\right] = \sigma_{x_i}^2$, nebo také souhrnně $\mathrm{Cov}\left[x_i, x_j\right] = \delta_{ij}\sigma_{x_i}^2$, kde Kroneckerův symbol

$$\delta_{ij} \equiv \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} ,$$

pak platí známý "zákon šíření chyb"

$$\sigma_z \approx \sqrt{\sum_i \left(\frac{\partial g(\boldsymbol{\mu})}{\partial x_i}\right)^2 \sigma_{x_i}^2} ,$$

kde derivace jsou vyhodnoceny v bodě očekávané hodnoty $\boldsymbol{x} = \boldsymbol{\mu}$. Všimněte si, že jde o přiblížení prvního řádu, kde funkci náhodných proměnných $g(\boldsymbol{x})$ aproximujeme tečnou nadrovinou. Obecně však $\mathrm{Cov}\left[x_i, x_j\right] = \rho_{ij}\sigma_{x_i}\sigma_{x_j}$.

Příklad:

Dle "zákona" o šíření chyb si spočítejte rozptyl součtu, součinu a podílu náhodných veličin X a Y, spojitých proměnných jako funkce (jednotlivě) x a y, tj. "proderivujte" si a získejte "chybu" veličin

$$z = x + y\,, \quad z = xy\,, \quad z = x/y\,, \quad (y \neq 0)$$

ale pak také $z = x^n$ a $z = 1/x^n$ (n přirozené číslo), známe-li rozptyly σ_x a σ_y. Upravte na co nejkompaktnější tvar, "z pod" odmocniny vytkněte, co jde, a vyjádřete si vždy i relativní "chybu" σ_z/z. Diskutujte nejprve případ nezávislých veličin, ale pak si propčtěte i případy s obecnou korelací ρ, a nakonec zjednodušte výsledky pro extrémní případy $\rho = \pm 1$ (často lze odmocnit).

Příklad:
Ukažte, že variance následující sumy náhodných veličin X_i

$$\bar{X} \equiv \frac{1}{N}\sum_{i=1}^{N} X_i$$

"podléhající"stejné hustotě pravděpodobnosti, a tedy i střední hodnotě a rozptylu μ, σ, je známý výsledek $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$, tj. chyba průměru N měření se zlepšuje faktorem $1/\sqrt{N}$.
Příklad:
Dokončete propagaci chyb pro příklad $m \equiv m(p, E)$, tj. spočtěte σ_m znáte-li σ_E a σ_p.
Příklad:
BMI, tzv. body mass index, je funkce váhy w v kg a výšky člověka h v m, definována jako BMI $\equiv w/h^2$. Na Obrázku 23 jsou projekce z dvojdimenzionálního výběru jako funkce váhy či výšky, kdy v populaci byla získána současně data o výšce i váze člověka. 2D data jsou vizualizována 2D histogramem (scatter-plot), a dále je spočítán a zobrazen BMI. Data kredit: O. Kepka.
Pozorovaná korelace mezi výškou a váhou v těchto datech je $0,7098$. Podle propagace chyb bez započtení korelace by očekávaný rozptyl veličiny $b \equiv$ BMI byl (ověřte si!)

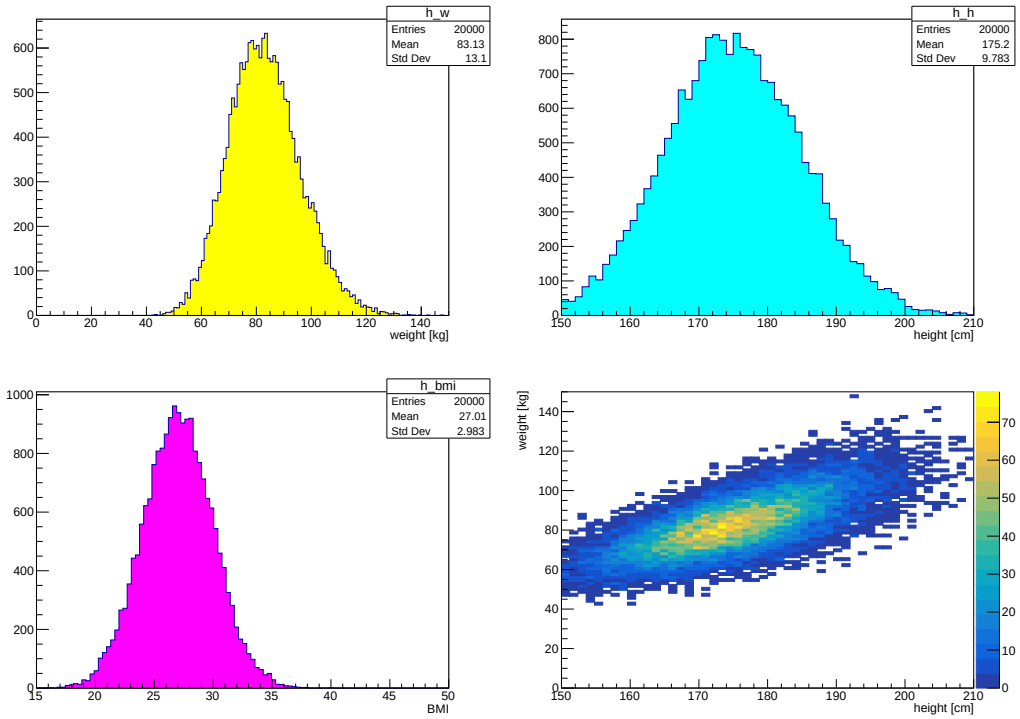
$$\hat{\sigma}_b = b\sqrt{\left(\frac{\sigma_w}{w}\right)^2 + 4\left(\frac{\sigma_h}{h}\right)^2}$$

což pro pozorované hodnoty (např. "odečteno"z odhadů střední hodnoty a rozptylu histogramu, viz. Obr. 23) $\sigma_w = 13,10$ a $\sigma_h = 0,09783$ dává $\hat{\sigma}_B = 5,22$.
Se započtením korelace je správně výraz

$$\hat{\sigma}_{\mathrm{b}} = \sqrt{\left(\frac{\partial b}{\partial w}\right)^2 \sigma_w^2 + 2\rho\left(\frac{\partial b}{\partial w}\right)\left(\frac{\partial b}{\partial h}\right)\sigma_w\sigma_h + \left(\frac{\partial b}{\partial h}\right)^2 \sigma_h^2}$$

$$\hat{\sigma}_{\mathrm{b}} = b\sqrt{\left(\frac{\sigma_w}{w}\right)^2 - 4\rho\frac{\sigma_w}{w}\frac{\sigma_h}{h} + 4\left(\frac{\sigma_h}{h}\right)^2},$$

kde se za w a h dosazují pozorované střední hodnoty.
Ačkoli je korelace mezi veličinmi pozitivní, díky podílové závislosti je jedna z derivací negativní, což vede k menší správné hodnotě odhadu rozptylu $2,979$, která je navíc konzistetní se rozptylem odečteným z histogramu $2,983$. Interpretace je taková, že pokud jedna veličina statisticky fluktuuje k vyšším hodnotám, druhá, korelovaná, veličina také, ale v podílu se tento efekt z části vyruší.



Obrázek 23: Rozdělení váhy w, výšky h, BMI $\equiv w/h^2$ a také 2D rozdělení váhy a výšky člověka v populaci. Data kredit: O. Kepka.

5.3 Rozptyl efektivity

Uvažujme, že měříme efektivitu toho, že událostí procházejí nějakým výběrovým Kritériem. Pozitivně zaznamenáme k_{obs} případů z n a spočítáme si odhad efektivity ε jako $\hat{\varepsilon} = k_{\text{obs}}/n$. Jaký je rozptyl veličiny $\varepsilon = k/n$ popř. odhad jejího rozptylu?

Pro počet pozorovaných případů k můžeme uvažovat binomickou hustotu pravděpodobnosti

$$P(k|p;n) = \binom{n}{k} p^k (1-p)^{n-k} \,,$$

která má známou střední hodnotu $\mathrm{E}[k] = np$ a varianci $\mathrm{Var}[k] = \mathrm{E}\left[(k - \mathrm{E}[k])^2\right] = np(1-p)$.

Pro náhodnou veličinu $\varepsilon = k/n$ nemůžeme použít přímočarý vzorec pro "propagaci chyb", neb počet událostí k je podmnožinou všech událostí n, a nejsou tak nezávislými veličinami. Nicméně střední hodnota je rovna

$$\mathrm{E}[\varepsilon] = \frac{\mathrm{E}[k]}{n} = \frac{np}{n} = p$$

a pro odhad p můžeme použít $\hat{p} = \hat{\varepsilon}$. Dále pak variance je rovna

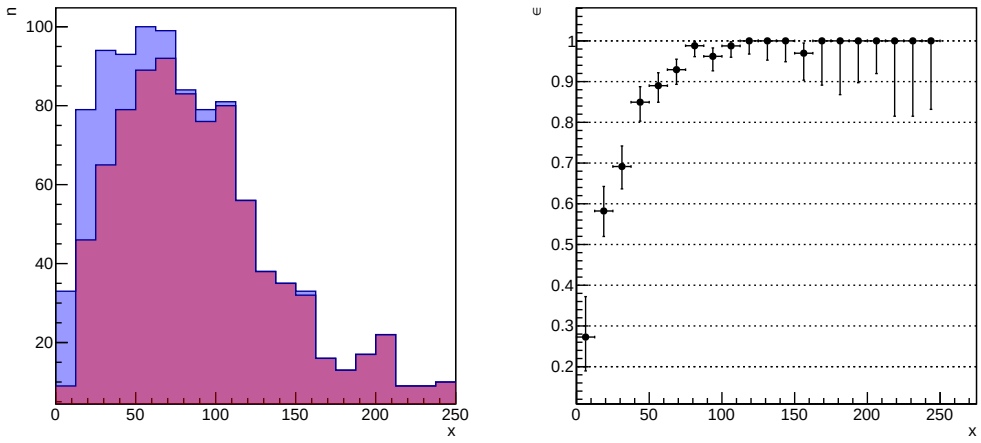
$$\mathrm{Var}[\varepsilon] = \frac{\mathrm{Var}[k]}{n} = \frac{np(1-p)}{n} = p(1-p) \equiv \sigma_\varepsilon^2$$

a pro odhad variance použijeme

$$\hat{\sigma}_\varepsilon^2 = \hat{p}(1-\hat{p}) = \hat{\varepsilon}(1-\hat{\varepsilon}) \,,$$

$$\hat{\sigma}_\varepsilon = \sqrt{\hat{\varepsilon}(1-\hat{\varepsilon})} \,,$$

což je známý vzorec pro binomický odhad neurčitosti v měření efektivity. Všimněme si, že rozptyl efektivity se v tomto binomickém modelu blíží nule pro efektivitu blízkou nule či jedné. Tyto "chyby" jsou jednou z možností, jak prezentovat neurčitosti v rozděleních, které jsou podílem dvou histogramů po a před nějakými výběrovými pravidly, např. v analýze efektovity triggeru či "cutů" v analýze, viz ilustrace na Obr. 24.



Obrázek 24: Vpravo: model celkového histogramu (modře), histogramu po modelovém výběrovém kritériu (červeně), a výsledná efektivita s binomickými chybami (vpravo).

5.4 Cross-sections ratio

An interesting question arises in measurements of a ratio of two cross sections. The application can be a cross-section ratio for a given process at different collider energies, a ratio between cross sections of two different processes at the same collider energy and data set, or a ratio of a cross section of a given process in nuclear (AA or pA) collisions to the same measured quantity but in a reference pp environment. In all these cases, the benefit is a partial cancellation of systematic uncertainties which can be, depending on the case, the luminosity uncertainties, theory or modelling uncertainties or systematic uncertainties if evaluated using the same or similar methods. For a ratio of two quantities $z = \frac{x}{y}$ the uncertainties propagation can be expressed as split in those uncorrelated or partially correlated, ρ being an additional or effective correlation coefficient between the systematics in x and y, as

$$\frac{\sigma_z^2}{z^2} = - 2\rho \left(\frac{\sigma_x^{\text{syst,corr}}}{x}\right)\left(\frac{\sigma_y^{\text{syst,corr}}}{y}\right) + \left(\frac{\sigma_x^{\text{syst,uncorr}}}{x}\right)^2 + \left(\frac{\sigma_y^{\text{syst,uncorr}}}{y}\right)^2 \quad (1)$$

$$+ \left(\frac{\sigma_x^{\text{stat}}}{x}\right)^2 + \left(\frac{\sigma_y^{\text{stat}}}{y}\right)^2 + \left(\frac{\sigma_x^{\text{syst,corr}}}{x}\right)^2 + \left(\frac{\sigma_y^{\text{syst,corr}}}{y}\right)^2 . \quad (2)$$

This formula can be verified using "toy" experiments, in which x_i and y_i are computed in each i-th experiment by smearing the central values x_0 and y_0 by random numbers drawn from Gaussian distributions as follows:

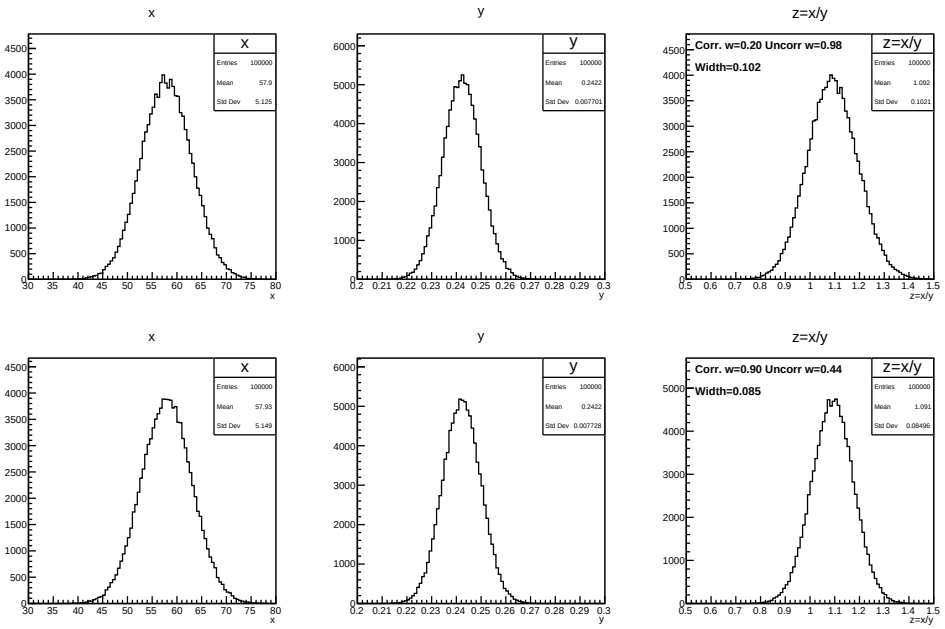
- for uncorrelated uncertainty sources like the statistical or uncorrelated systematics, different random numbers are drawn for x and y;
- for correlated uncertainties, a single random number is drawn to smear x and y coherently.

This can be expressed as follows, with λ_i being random numbers drawn from standard normal distribution, *i.e.* $\lambda \sim \mathcal{G}(\lambda|0,1)$

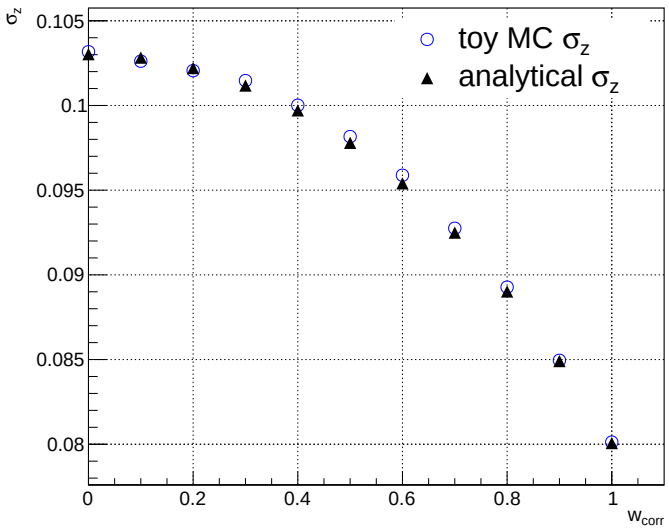
$$x_i = x_0 + \sum_j \lambda_j^{(x)} \sigma_j^{\text{x,uncorr}} + \sum_k \lambda_k \sigma_k^{\text{x,corr}} \quad (3)$$

$$y_i = y_0 + \sum_j \lambda_j^{(y)} \sigma_j^{\text{y,uncorr}} + \sum_k \lambda_k \sigma_k^{\text{y,corr}} . \quad (4)$$

The event-by-event ratio $z_i = x_i/y_i$ is then computed in each pseudo-experiment, an histogram of z is filled and its width extracted and compared to the standard deviation computed using the analytical formula. This test is illustrated in Figure 26 where the "toy" and analytical-based uncertainties in z are plotted in the same picture, as a function of the weight (a fraction) given to the correlated systematic uncertainty. As expected, larger fraction of uncertainties treated as correlated leads to their partial cancellation in the ratio and a small total uncertainty in the ratio.



Obrázek 25: Evolution of the total uncertainty in a model ratio variable depending on the weight (fraction) of correlated uncertainties in the numerator and denominator, evaluated using 100k statistical toys (open circles) compared to the result of an analytical formula (filled triangles).



Obrázek 26: Evolution of the total uncertainty in a model ratio variable depending on the weight (fraction) of correlated uncertainties in the numerator and denominator, evaluated using 100k statistical toys (open circles) compared to the result of an analytical formula (filled triangles).

5.5 Variance vážených dat

Ukládáme-li si např. do histogramu četnosti váhované nějakým faktorem w_i, je potřeba definovat "chybu"celkové sumy událostí v takovémto binu. Pro Poissonovsky rozdělená data (tj. neváhovaná) platí

$$\hat{\nu} = n\,, \quad \hat{\sigma}_{\hat{\nu}} = \sqrt{n}\,, \quad \text{a tedy} \quad n = \left(\frac{\hat{\nu}}{\hat{\sigma}_{\hat{\nu}}}\right)^2 .$$

Pro váhovaná data použijeme zřejmě

$$\hat{\nu} = \sum w_i$$

a pro hledanou varianci $\hat{\nu}$ použijeme podle "propagace chyb"

$$\hat{\sigma}^2_{\hat{\nu}} = \sum \mathrm{Var}\left[w_i\right] .$$

Protože ale uvažujeme jednotlivé události vážené faktorem w_i, bude z vlastností variance jako kvadratické formy $\mathrm{Var}\left[w_i\right] = \mathrm{Var}\left[w_i \cdot 1\,\mathrm{evt}\right] = w_i^2\,\mathrm{Var}\left[1\,\mathrm{evt}\right] = w_i^2\sqrt{1}$ a tedy

$$\hat{\sigma}_{\hat{\nu}} = \sqrt{\sum w_i^2}\,.$$

Nakonec lze v této v analogii s Poissonovým rozdělením nevážených dat definovat tzv. efektivní počet událostí dat váhovaných

$$n_{\mathrm{eff}} = \frac{\left(\sum w_i\right)^2}{\sum w_i^2}\,,$$

který se občas používá k definici a porovnání statistické významnosti váhovaného počtu událostí.

6 Odhad parametrů

6.1 Zaujaté a nezaujaté odhady

Nechť X je náhodná veličina s (často neznámou) hustotou pravděpodobnosti $f(x;\mu,\sigma,\ldots)$, která závisí na nějakých parametrech μ,σ aj. kde např. může být μ rovno (opět en nutně známé) střední hodnotě rozdělení a σ jeho rozptylu, popř. rozdělení může mít další parametry.

Tuto veličinu N-krát "pozoruji"("měřím"), a získám tak tzv. náhodný výběr z rozdělení $f(x;\mu,\sigma,\ldots)$ popř. náhodný výběr veličiny X, značíme jako množinu pozorovaných čísel $\{x_i\}_{i=1}^N$.

Jako odhad střední hodnoty se často používá aritmetický průměr

$$\hat{\mu} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i \,.$$

Jak ale víme, že jde o nejlepší odhad skrytého parametru přírody μ? Pokud bychom opakovali výběr, a získali mnoho takových N-tic (výběrů), spočítali bychom si mnoho odhadů $\hat{\mu}$, a "rádi bychom", aby tato střední hodnota $\hat{\mu}$ přes mnoho takovýchto ensemblů byla blízká skrytému parametru μ. Pohlížejme tedy nyní na každou hodnotu x_i jako na náhodnou veličinu, samozřejmě také rozdělenu podle $f(x;\mu,\sigma,\ldots)$, a považujme nyní i $\hat{\mu} = \hat{\mu}(\boldsymbol{x})$ za náhodnou veličinu (lineární funkci náhodných veličin x_i), a podívejme se na tzv. zaujatost (bias) $b_{\hat{\mu}}$ odhadu parametru $\hat{\mu}$ definované jako rozdíl střední hodnoty odhadu parametru a jeho skutečné (skryté) hodnoty

$$b_{\hat{\mu}} \equiv \mathrm{E}\left[\hat{\mu}\right] - \mu \,.$$

V našem konkrétním případě získáme

$$b_{\hat{\mu}} = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] - \mu = \frac{1}{N}N\mu - \mu = 0 \,,$$

kde jsme využili toho, že každá veličina x_i má stejnou střední hodnotu μ: $\mathrm{E}\left[x_i\right] = \mu$. Dostáváme tedy, že aritmetický průměr je skutečně nezaujatým odhadem parametru μ. Zkusme nyní, jak odhadnout rozptyl, resp. jeho kvadrát. Inspirováni analytickou definicí $\sigma^2 \equiv \mathrm{E}\left[(x_i-\mu)^2\right]$ zkusme definovat

$$\hat{\sigma}^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N} x_j\right)^2 \,,$$

tj. průměr kvadratických odchylek měřených hodnot x_i od našeho nejlepšího odhadu jejich střední hodnoty $\hat{\mu}$ (nezapomínejme, že $\hat{\mu}$ je funkcí konkrétních dat!). Podrobnou analýzou zaujatosti tohoto odhadu však získáme

$$b_{\hat{\sigma}^2} = \mathrm{E}\left[\hat{\sigma}^2\right] - \sigma^2 = \sigma^2 - \frac{2}{N}N\sigma^2 + \frac{1}{N^3}N^2\sigma^2 = -\frac{\sigma^2}{N} \,.$$

Tj. nejde o nezaujatý odhad, ačkoli se jeho zaujatost (rozdíl od skutečného kvadrátu rozptylu) v průměru snižuje s velikostí vzorku dat N jako $1/N$, zaujatost rozptylu však pouze jako $\sqrt{1/N}$.

Správný, nezaujatý, odhad čtverce rozptylu je známá definice

$$\hat{\sigma}^2_{\text{unb.}} \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j\right)^2 .$$

Ukažte si, že jde opravdu o nezaujatý odhad.
Nakonec, je-li hodnota μ známa, je nezaujatým odhadem rozptylu přímo

$$\hat{\sigma}^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 .$$

Obdobně, pro měření dvojic náhodných veličin X a Y (např. hybnost částice a nějaký úhel v detektoru) a pro konkrétní výběr (měření) hodnot $\{[x_i, y_i]\}_{i=1}^{N}$ lze definovat odhad jejich kovariance

$$\widehat{\text{Cov}}[X,Y] \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{\mu}_x)(y_i-\hat{\mu}_y) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j\right)\left(y_i - \frac{1}{N}\sum_{k=1}^{N}y_k\right) .$$

a následně odhadnout i korelační koeficient jako

$$\hat{\rho} \equiv \frac{\widehat{\text{Cov}}[X,Y]}{\hat{\sigma}_X\hat{\sigma}_Y} ,$$

kde $\hat{\sigma}^2_X \equiv \widehat{\text{Cov}}[X,X]$.

6.2 Fitování histogramu

Pearsonův χ^2 chí-kvadrát je statistika (funkce dat) závisející naparametrech fitu

$$\chi^2_{\text{fit;data}} \equiv \chi^2_{\text{fit}}(\boldsymbol{\theta}|\mathbf{d}) = \sum_{i=1}^{n_{\text{bins}}}\frac{[d_i - f(x_i^c|\boldsymbol{\theta})]^2}{d_i} ,$$

kde d_i je počet pozorovaných událostí v i-tém binu veličiny X, $\boldsymbol{\theta}$ jsou parametry fitu a x_i^c je souřadnice středu binu i. Odhad rozptylu pozorovaného počtu událostí d_i je podle Poissonovské statistiky $\sqrt{d_i}$ a variance tohoto počtu je tedy opět d_i.

V případě nepoissonovských neurčitostí se používá výraz

$$\chi^2_{\text{fit;data}} \equiv \chi^2_{\text{fit}}(\boldsymbol{\theta}|\mathbf{d}) = \sum_{i=1}^{n_{\text{bins}}}\frac{[d_i - f(x_i^c|\boldsymbol{\theta})]^2}{\sigma_i^2} ,$$

kde σ_i jsou neurčitosti datových bodů.

Jedná se tedy o sumu kvadratických výrazů, čtverců, a výraz $\chi^2_{\text{fit;data}}$ je v průběhu fitování minimalizován za účelem nalezení nejlepších parametrů fitu $\hat{\boldsymbol{\theta}}$. V principu jde o nalezení stacionírních bodů, tj. o hledání

$$\nabla_{\boldsymbol{\theta}}\ \chi^2_{\text{fit;data}}(\boldsymbol{\theta}|\hat{\mathbf{d}}) = 0$$

kde gradient počítáme jako derivace podle jednotlivých parametrů $\boldsymbol{\theta}$, tj.

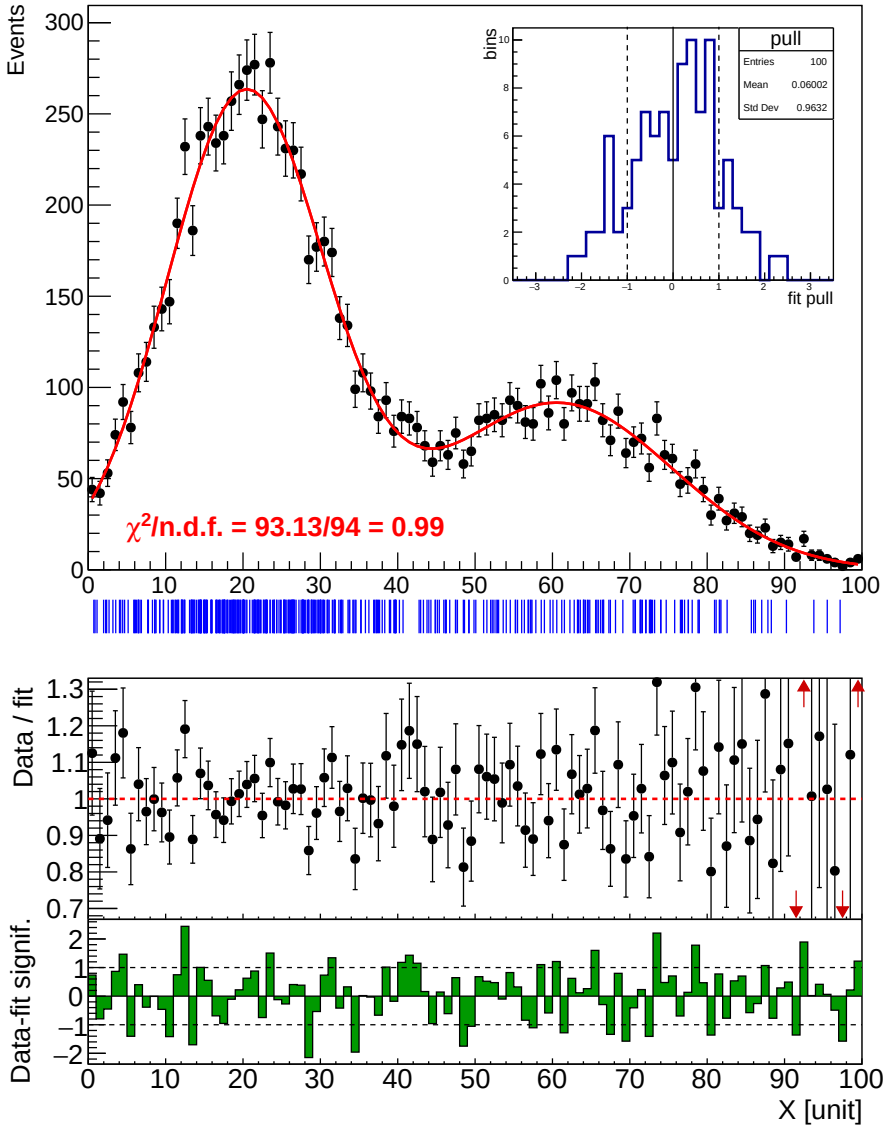
$$\nabla_{\boldsymbol{\theta}} \equiv \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_{n_p}}\right)$$

kde n_p je počet parametrů fitovací funkce.

Můžeme definovat signifikanci rozdílu mezi hodnotou naměřené veličiny a nejlepší předpovědí z fitu jako rozdíl těchto hodnot v jednotkách neurčitosti dat právě jako

$$S_i \equiv \frac{\left[d_i - f(x_i^c|\hat{\boldsymbol{\theta}})\right]^2}{\sigma_i^2} \,.$$

V případě fitu se této veličině také anglicky říká pull, a histogram těchto hodnot přes všechny biny, kde proběhl fit, má očekávanou hodnotu nula a standardní odchylku rovnu jedné. Viz ilustrace na Obr. 27, kde n.d.f. (ang. number of degrees of freedom) je očet stupňů volnosti fitu a je počtu binů fitovaného histogramu mínus počet volných parametrů fitu, n.d.f.$= n_{\text{bins}} - n_p$.



Obrázek 27: Histogram z nebinovaných hodnot (modře, jen vybrané) je proložen fitem (červená křivka). Níže je spočítán rozdíl četností v histogramu oproti fitu, vztaženo na předpověď z fitu. Červené šipky poukazují na body ležící mimo vymezenou oblast podílu. Nejníže je pak spočítaná signifikance rozdílu (zeleně), tj. rozdíl mezi daty a fitem v jednotkách neurčitosti datových bodů, s vodícími linkami odpovídajícím $\pm 1\sigma$ intervalům. Nahoře vložen histogram této signifikance, zde také interpretována jako tzv. "pull", který má podle očekávání střední hodnotu nula a jednotkovou šířku.

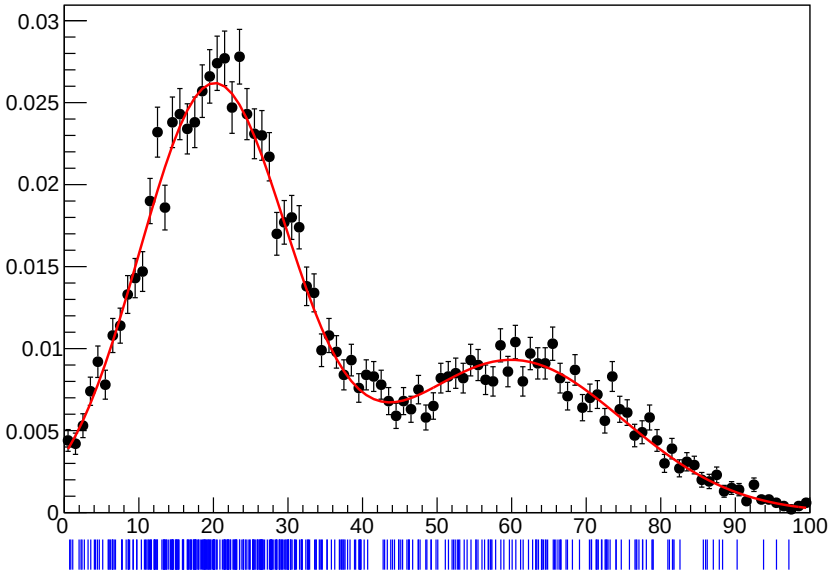
7 Věrohodnost

7.1 Odhad parametrů metodou maximální věrohodnosti

Pro n měření náhodné veličiny X, tj. sadu hodnot $\{x_i\}_{i=1}^N \equiv \boldsymbol{x}$ budeme uvažovat, že odpovídají výběrům z nějaké hustoty pravděpodobnosti, tj. uvažjeme nějaký *model* pro naše data. Soulad jednoho měření s danou h.p. můžeme vyjádřit vyhodnocením dané h.p. v bodě x_i, tj. $p_i \equiv f(x_i|\boldsymbol{\theta})$. Pozor, p_i není pravděpodobnost, to bychom museli přeintegrovat přes nějaký konečný interval x. Hustota pravděpodobnosti závisí na nějakých parametrech $\boldsymbol{\theta}$, a naším cílem je extrahovat (odhadnout) z dat takovou sadu parametrů $\hat{\boldsymbol{\theta}}$, kterů jsou v nejlepší shodě s daty, tj. takové, které dají nejlepší h.p., která by mohla data popisovat. Některé z parametrů mohou být důležité fyzikální veličiny jako hmota částice, frakce signálu, jiné mohou být rozlišení či jiné pomocné parametry. Je-li každé měření nezávislé, můžeme sestrojit veličinu, která má význam jakési hustoty pravděpodobnsoti naměření daného vektoru $\boldsymbol{x}$ v datech, na který však pro daná fixní data pohlížíme jako na funkci parametrů $\boldsymbol{\theta}$, a hovoříme o věrohodnostní funkci (angl. likelihood)

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) \equiv \prod_{i=1}^{N} f(x_i|\boldsymbol{\theta})$$

Evidentně, pokud se s parametry $\boldsymbol{\theta}$ "trefíme"blízko ideálním parametrům $\boldsymbol{\theta}_{\text{true}}$, podle kterých jsou data rozdělena, bude hodnota věrohodnosti vyšší než pro parametry, které realitě neodpovídají. Podoba takovýchto dat je ilustrována na Obr. 28.



Obrázek 28: Ilustrace části nebinovaných dat (modré úsečky) rozdělených podle funkce (červeně), a dále normalizovaná binovaná forma dat (histogram).

V praxi se častěji pracuje s logaritmem, který převádí součin na sumy a vede k numericky stabilnějším proměnným v kódu i jednodušším symbolickým úpravám, jak brzy uvidíme. Pak se maximalizuje $\ln \mathcal{L}$ nebo minimalizuje $-\ln \mathcal{L}$ vzhledem k parametrům rozdělení $\boldsymbol{\theta}$, které chceme z dat odhadnout. Funkce $\mathcal{L}$ i $\ln \mathcal{L}$ mají stejné stacionární body a obecně hledáme takovou sadu parametrů $\hat{\boldsymbol{\theta}}$, pro které je derivace (obecně gradient)

podle $\boldsymbol{\theta}$ stacionární, tj.

$$\nabla_{\boldsymbol{\theta}}\,\mathcal{L}(\hat{\boldsymbol{\theta}}) = 0\,.$$

Výběr modelu se často provádí "pohledem"na binovaná data, tj. na histogram, ale fit parametrů je nejlepší provést na datech nebinovaných, tj. maximalizováním věrohodnostní funkce.

V praxi je často věrohodnost složitou mnohoparametrickou funkcí a je obtížné ji opakovaně vyhodnocovat jako funkci spojých parametrů $\boldsymbol{\theta}$ (jeden z parametrů může navíc být např. hmotou částice vystupující v simulaci, a je tak možné nageneroavt jen konečný počed sad vzorků s různou hmotou, přičemž pro kažkou hmotu je vyhodnocen souhlas, např právě věrohodnostu, s daty pomocí nějaké kinematické veličiny). V tom případě je věrohodnost napočítána jako funkce několika diskrétních sad $\boldsymbol{\theta}_j$ a minimum je nalezeno fitováním diskétního grafu (angl. binned likelihood fit). V následujícím si však ukážeme některé analytické výrazy dávající odhad nejlepších parametrů metodou maximální věrohodnosti pro některé základní modely hustot pravděpodobnosti.

Příklad:

Gausovská hustota pravděpodobnosti. Pro uvažovanou h.p.

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma_i^2}\right]\,,$$

tj. dovolíme-li dokonce, aby každé měření mělo obecně jiný rozptyl σ_i, bude záporný logaritmus věrohodnosti roven $-\ln\mathcal{L} = \sum\frac{(x-\mu)^2}{\sigma_i^2}$ a pro odhad parametru μ položením

$$\frac{\partial}{\partial\mu}(-\ln\mathcal{L})\Big|_{\mu=\hat{\mu}} = 0 \quad\Rightarrow\quad \sum_i \frac{x_i-\hat{\mu}}{\sigma_i} = 0 \quad\Rightarrow\quad \hat{\mu} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}} \equiv \frac{\sum_i w_i\,x_i}{\sum_i w_i}$$

tj. do odhadu střední hodnoty váhujeme každou změřenou hodnotu faktorem $w_i \equiv 1/\sigma^2$, tj. převrácenou hodnotou její variance. Pro shodné variance všech naměřených hodnot pak dostaneme

$$-\ln\mathcal{L} = \sum_i \frac{(x-\mu)^2}{\sigma^2} \quad\Rightarrow\quad \hat{\mu} = \frac{1}{N}\sum_i x_i\,,$$

o kterém již výme, že jde o nezaujatý odhad střední hodnoty. Obdobně, známe-li μ, můžeme odhadnout rozptyl podle

$$-\ln\mathcal{L} = \sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - N\ln(\sqrt{2\pi}\sigma)$$

$$\frac{\partial}{\partial\sigma}(-\ln\mathcal{L}) = \sum_i \frac{(x_i-\mu)^2}{\sigma^3} - \frac{N}{\sigma}$$

a položením

$$\frac{\partial}{\partial\sigma}(-\ln\mathcal{L})\Big|_{\sigma=\hat{\sigma}} = 0 \quad\Rightarrow\quad \hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i-\mu)^2\,,$$

což je však zaujatý odhad rozptylu!

Příklad:

Poissonovská hustota pravděpodobnosti. Odvoďte dále odhad střední hodnoty parametrů, že data uvažujeme rozdělena podle Poissonova rozdělení.

Příklad:

Zopakujte si totéž pro případ exponenciální hustoty pravděpodobnosti.

7.2 Odhad "chyby"MLE estimátorů

7.2.1 Variance parametrů odhadnutých metodou maximální věrohodnosti

Podle tzv. Rao-Cramér-Frechet (RCF) podmínky platí následující spodní odhad variance odhadu parametru metodou maximální věrohodnosti

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \frac{1}{\mathrm{E}\left[-\frac{\partial^2 L}{\partial \theta^2}\right]}$$

kde $b(\hat{\theta})$ je bias parametru $b(\hat{\theta}) \equiv \mathrm{E}\left[\hat{\theta}\right] - \theta$. Často je parametr nezaujatý, a za obvyklých podmínek také dokonce platí přímo

$$\mathrm{Var}\left[\hat{\theta}\right] = \frac{1}{\mathrm{E}\left[-\frac{\partial^2 L}{\partial \theta^2}\right]}$$

7.2.2 Grafické řešení

Ilustrace průběhu věrohodnost jako kvadratická funkce v okolí minima je na Obr. 29. Podrobněji, z Taylorova rozvoje zíkáváme pro rozvoj $-\ln \mathcal{L}$ okolo minima $\hat{\theta}$

$$-\ln \mathcal{L}(\theta) \approx -\ln \log \mathcal{L}(\hat{\theta}) + \frac{1}{1!}\frac{\partial[-\ln \mathcal{L}]}{\partial \theta}(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2!}\frac{\partial^2[-\ln \mathcal{L}]}{\partial \theta^2}(\hat{\theta})(\theta - \hat{\theta})^2$$

Druhý člen na pravé straně je však nula (první derivace v minimu je nula) a tedy

$$-\ln \mathcal{L}(\theta) \approx -\ln \mathcal{L}(\hat{\theta}) + \frac{1}{2}\frac{\partial^2[-\ln \mathcal{L}]}{\partial \theta^2}(\hat{\theta})(\theta - \hat{\theta})^2 .$$

Současně, dle RCF teorému můžeme odhadnout

$$\hat{\sigma}_\theta^2 = -1/\frac{\partial^2[\ln \mathcal{L}]}{\partial \theta^2}(\hat{\theta})$$

a tedy

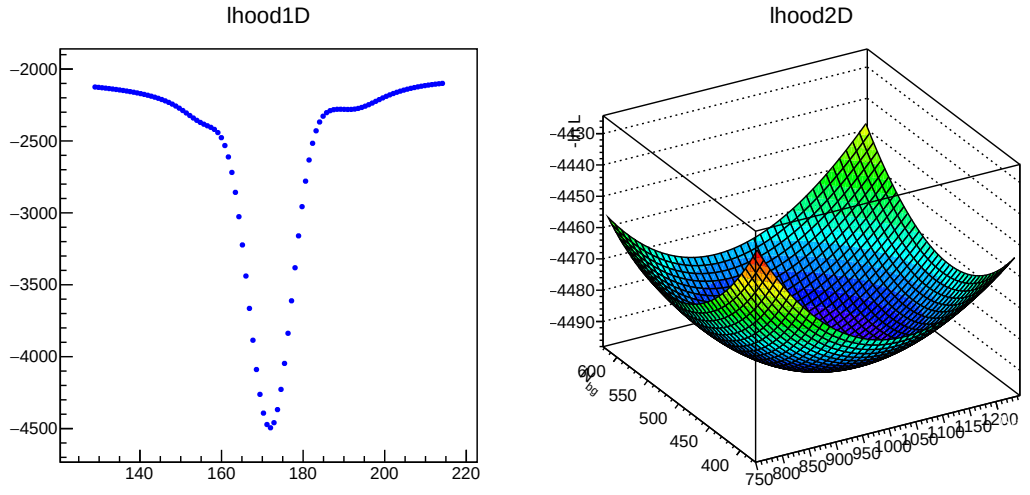
$$-\ln \mathcal{L}(\theta) \approx -\ln \mathcal{L}(\hat{\theta}) + \frac{1}{2\hat{\sigma}_\theta^2}(\theta - \hat{\theta})^2 .$$

Vyhodnocením v posunutích σ_θ okolo minimální hodnoty $\hat{\theta}$ dostáváme

$$-\ln \mathcal{L}(\hat{\theta} \pm \sigma_\theta) \approx -\ln \mathcal{L}(\hat{\theta}) + \frac{1}{2\hat{\sigma}_\theta^2}(\hat{\theta} \pm \hat{\sigma}_\theta - \hat{\theta})^2 = -\ln \mathcal{L}(\hat{\theta}) + \frac{1}{2}$$

a tedy negativní logarotmus věrohodnosti se na 1σ intervalu okolo $\hat{\theta}$ mění o 1/2. Toho lze opačně využít ke grafickému odečtu $\hat{\sigma}_\theta$.

Druhý pohled je ze vztahu věrohodnosti a chi-kvadrát výrazu pro gaussovskou hustotu pravděpodobnosti, kdy platí $\chi^2 = -\frac{1}{2}\ln \mathcal{L}$ a změna dat o jednu standardní odchylku je ekvivalentní změně $\Delta\chi^2 = 1$ a tedy $\Delta\left[-\ln \mathcal{L}\right] = \frac{1}{2}$ a tedy opět jde o posun o 1/2.
Příklad:



Obrázek 29: Ilustrace kvadratického profilu $-\ln \mathcal{L}$ v okolí minima. Jako funkce jednoho parametru (vlevo) a dvou parametrů (vpravo).

7.2.3 Cross section fit using a likelihood

Advantage: it is easy to combine channels by multiplying the likelihoods.

Example on extracting a cross-section σ_X of a proces X by fitting additional (nuisance) parameters $\vec{\alpha}$ (allowing for a freedom in measured efficiencies and calibrations).

The normalisation part of likelihood depends on the cross-section to be extracted, the expected total expected number of measure entries $\nu \equiv \nu_S(m_X, \vec{\alpha}) + \nu_B(\vec{\alpha})$ depends on fyzikálním parametru m_X (hmota hledané rezonance) and nuisance parameters and on the background prediction.

$$\mathcal{L}(m_X, \vec{\alpha}) \equiv \mathcal{L}_{\mathrm{S,Bg}}(m_X, \vec{\alpha}) \cdot \mathcal{L}_{\mathrm{nuisance}}(\vec{\alpha})$$

zatímco v případě binned likelihood může jít o odnoty signálu a pozadí v binech i (jejich tvar)

$$\mathcal{L}_{\mathrm{S,Bg}}(m_X, \vec{\alpha}) \equiv \prod_i^n f_S \, g_S(x_i|, m_X, \vec{\alpha}) + (1 - f_S) \, g_B(x_i|\vec{\alpha})$$

kde f_S je frakce signálu (další z parametrů)

Usually one fixes the total number of observed events by a Poisson term (extended likelihood)

$$\mathcal{P}(N_{\mathrm{obs}}|\nu(m_X, \vec{\alpha}))$$

a tedy $f_S = f_S(m_X)$, což může pomoci omezit některý z parametrů. Nehledáme např. jen pozici peaku odpovídající nějaké rezonanci X, ale příslušný účinný průřez σ_X, a tedy i očekávaný počet událostí (tj. plocha pod peakem), může záviset na hmotě m_X a celkový počet událostí je další svazující podmínka závislá na m_X.

Nuisance parameters are often assumed to be Gaussian, Gamma-function or log-normal distributed.

$$\mathcal{L}_{\mathrm{nuisance}}(\vec{\alpha}) \equiv \prod_i^n \mathcal{G}(\vec{\beta}_i) \,, \qquad \{\vec{\alpha}\} \equiv \bigcup_i \{\vec{\beta}_i\} \,.$$

The likelihood $\mathcal{L}$ is maximised w.r.t. the free parameters σ_X and $\vec{\alpha}$. In practice, due to limited numerical precision, it is always the logarithm of the likelihood one works with, making products sums and usually $[-\ln \mathcal{L}]$ is minimised.

Excercise I: express (expand) the logarithm of the likelihood assuming Poisson and Gauss p.d.f.'s for event counts and for nuisance parameters.
Hint: Poisson distribution probability density for observing n events with the mean of μ is given by

$$\text{Poisson}(n, \mu) = \frac{\mu^n}{n!} e^{-\mu} \,.$$

As one usually works with logarithms of likelihoods, the following simplification is useful:

$$\ln \text{Poisson}(n, \mu) = n \ln \mu - \mu + \text{const} \,,$$

as n is from measurement and is thus constant in minimization, which is usually w.r.t. to μ. See how this maximizes as funstion of μ, and also examine and compare to the similar expression for the gaussian p.d.f.

Excercise II: Write a simple toy Monte Carlo tool (e.g. in ROOT) and try to fit the input "true" fraction. Note the numerical danger of working with the likelihood instead of the logarithms.

7.3 Podíl věrohodností

odhad parametrů pomocí profile likelihood funkce se zahrnutím systematických parametrů (nuisance parameters) [10, 11]

$$q(\mu) \equiv -2 \ln \frac{\mathcal{L}(\mu, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})}$$

konvergence testovací statistiky $q(\mu)$ ke Gaussovu rozdělení

8 Bayesův teorém

Podmíněnou pravděpodobnost $P(A|B)$ jevu A *za předpokladu*, že nastal jev B můžeme použít k výpočtu pravděpodobnosti toho, že nastane jev A i B současně:

$$P(A \cap B) \equiv P(A \text{ a } B) \equiv P(A|B)P(B)\,.$$

Ze symetrie průniku můžeme psát také

$$P(B \cap A) \equiv P(B|A)P(A)\,.$$

Události A a B jsou podle definice nezávislé pokud $P(A|B) = P(A)$. V takovém to případě pak

$$\Rightarrow P(A \cap B) = P(A)P(B) \quad \text{pro nezávislé } A, B\,.$$

Srovnáním prvních dvou rovnic pak dostáváme Bayesův teorém (Thomas Bayes, 1701–1761) pro podmíněné pravděpodobnosti

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}\,.$$

Poznámka: čteme tedy $P(A|B)$ také jako pravděpodobnost toho, že když nastal jev B, nastane jev A.

8.1 Aplikace na spolehlivost testu

Zajímá nás odpověď na následující otázku: jaká je pravděpodobnost, že pozitovní výsledek laboratorního testu na přítomnost nějakého patogenu např. v krvi opravdu znamená, že dotyčný pacient je nemocný? Pro vyjádření spolehlivosti testu potřebujeme znát několik nezávislých informací.

- Nechť je v populaci n lidí (množina Ω) n_N osob nemocných (množina Ω_N) a $n_Z = n - n_N$ zdravých (množina Ω_Z). Frakce nemocných osob je tedy $\xi \equiv n_N/(n_Z + n_N)$.
- Dále potřebujeme znát (pravou) efektivitu testu v případě, že je osoba skutečně nemocná, označme si jako podmíněnou pravděpodobnost pozitivního výsledku testu za předpokladu, že šlo o člověka z nemocné části populace: $\epsilon_N \equiv P_+(\Omega_N) = P(+|N)$.
- Test však bohužel může dát pozitivní výsledek i v případě zdravé osoby. Tuto falešnou efektivitu si můžeme označit jako $\epsilon_Z \equiv P_+(\Omega_Z) = P(+|Z)$
- Pravděpodobnost toho, že z populace náhodným výběrem získáme k laboratornímu testu nemocnou osobu je dána $P(N) = n_N/(n_Z + n_N) = \xi$.
- Pavděpodobnost, že vybereme zdravou osobu je dána $P(Z) = n_Z/(n_Z + n_N) = 1 - \xi$.
- Nyní se podívejme na pravděpodobnost toho, že dostaneme pozitivní výsledek testu pro náhodně vybranou osobu. Půjde o součet dvou příspěvků daných nenulovou efektivitou testu pro nemocné i zdravé osoby:

$$P(+) = P(N)P(+|N) + P(Z)P(+|Z) = \frac{n_N}{n_Z + n_N}\epsilon_N + \frac{n_Z}{n_Z + n_N}\epsilon_Z\,,$$

$$P(+) = \xi\epsilon_N + (1 - \xi)\epsilon_Z\,.$$

Nyní konečně definujme spolehlivst testu, tj. pravděpodobnost toho, že pokud dal test pozitivní výsledek, jde skutečně o nemocného člověka (a má smysl začít s nákladnou či pro pacienta náročnou léčbou). Jde o podmíněnou pravděpodobnost $P(N|+)$, kterou si vyjádříme pomocí Bayesova teorému:

$$P(N|+) = P(+|N)\frac{P(N)}{P(+)}$$

s využitím výše zadefinovaných veličin a po drobných úpravách získáme

$$P(N|+) = P(+|N)\frac{P(N)}{P(+)} = \frac{1}{1+\frac{\epsilon_Z}{\epsilon_N}\frac{n_Z}{n_N}}\,,$$

$$P(N|+) = P(+|N)\frac{P(N)}{P(+)} = \frac{1}{1+\frac{\epsilon_Z}{\epsilon_N}\frac{1-\xi}{\xi}}\,.$$

Všimněte si, že

- Výsledek předpokládá, že známe efektivitu (pravou i falešnou) testu $\epsilon_{N,Z}$, např. z jiného nezávislého (ale třeba nákladného) testu nebo z nějakého jiného jasného klinického projevu (ale např. náročně proveditelného).
- Spolehlivost testu je funkcí složení populace zdravých a nemocných lidí: n_Z/n_N.
- Falešnou efektivita testu ϵ_Z je potřeba zajistit na co nejnižší hodnotě, aby byla zajištěna dostatečná spolehlivost testu.

Spočtěte si spolehlivost testu pro sady parametrů

- $\epsilon_N = 0.99\,, \quad \epsilon_Z = 0.01\,, \quad \xi = 0.20$
- $\epsilon_N = 0.99\,, \quad \epsilon_Z = 0.01\,, \quad \xi = 0.05$
- $\epsilon_N = 0.99\,, \quad \epsilon_Z = 0.001\,, \quad \xi = 0.05$
- $\epsilon_N = 0.999\,, \quad \epsilon_Z = 0.01\,, \quad \xi = 0.05$

Nakreslete $P(1|+)$ jako funkci proměnných ϵ_0, ϵ_1, a $p \equiv n_N/n_Z$ či ξ.

8.2 Aplikace na odhad parametrů

Tzv. Bayesovská inference, odhad parametrů $\boldsymbol{\theta}$ z dat $\boldsymbol{x}$ na základě Bayesova teorému, je založena na přepisu podmíněných pravděpodobností

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{\theta})\,\frac{P(\boldsymbol{\theta})}{P(\boldsymbol{x})}\,,$$

kde $P(\boldsymbol{x}|\boldsymbol{\theta})$ je pravděpodobnost pozorování daných dat $\boldsymbol{x}$ podmíněná platností parametrů $\boldsymbol{\theta}$. Nejde tedy o nic jiného než o věrohodnostní funkci (likelihood)! V ní je již učiněn výběr konkrétní hustoty pravděpodobnosti (modelu) s konktrétní parametrizací pomocí $\boldsymbol{\theta}$. $P(\boldsymbol{x})$ je pravděpodobnost toho, že pozorujeme daná data nezávisle na hodnotě parametrů $\boldsymbol{\theta}$, což znamená, že přes ně musíme vystředovat, tj. vyintegrovat přes nějakou omezenou množinu $\Omega_{\boldsymbol{\theta}}$, kde prior definujeme nenulový

$$P(\boldsymbol{x}) \equiv \int\limits_{\Omega_{\boldsymbol{\theta}}} P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

a tedy

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{\theta})\,\frac{P(\boldsymbol{\theta})}{\int\limits_{\Omega_{\boldsymbol{\theta}}} P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}\,.$$

Výsledek, $P(\boldsymbol{\theta}|\boldsymbol{x})$, bývá nazýván jako *posterior* a vyjadřuje pravděpodobnost hodnot parametrů $\boldsymbol{\theta}$ pro daná data, a lze ji chápat jako (konkrétními daty podmíněnou) hustotu pravděpodobnosti pro parametry $\boldsymbol{\theta}$. Ty nejlepší parametry pak můžeme určit např. jako takové, pro které je posterior maximální, popř. vzít průměr či jinou míru polohy.

Nakonec, $P(\boldsymbol{\theta})$ je tzv. prior, určitá *a priori* informace o hodnotách parametrů, které hledáme. Tato informace, kterou musíme do teorému dodat, je občas kritizována za svou subjektivitu, současně ale může jít např. o výsledek předchozího experimentu s větší variancí, který naopak dobře a správně může posloužit jako vodítko pro odhad parametrů z nových dat. Volba prioru je často obtížná, heuristická, subjektivní. Nabízí se zvolit konstantu, tj. uniformí rozdělení na nějakém dostatečně širokém intervalu. Jakkoli správnou a nezaujatou volbou se může zdát, není bohužel invariatní vůči transformaci proměnných, tj. tzv. původní *flat* prior nutně není flat po transformaci. Vidíme však důležitou souvislost, že pro konstantní prior je Bayesovský odhad parametrů shodný s metodou maximální věrohodnosti, tj. že nabývá maxima.

9 Testování hypotéz

9.1 Testovací statistika

Definice: statistika $t(\boldsymbol{x})$ je libovolná funkce dat $\boldsymbol{x}$. Cílem je, aby tato funkce separovala základní (H_0) a alternativní (H_1) hypotézu. Testovací statistikou může být jednoduše samotný počet událostí, ale i složitější funkce jako např. χ^2 výraz, věrohodnost či poměr věrohodnostních funkcí.

Základní myšlenkou testování hypotéz je definování kritické hodnoty testovací statistiky t_c tak, aby testovací statistika separovala mezi hypotézami H_0 (např aby se "kupila"vlevo) a H_1 (aby se "kupila"vpravo). Známe-li hustoty pravděpodobnosti $g(t|H_k)$ $(k = 0, 1)$ pro obě hypotézy (ať už analyticky či pomocí nějaké Monte Carlo metody), můžeme spočítat veličinu

$$\alpha \equiv \int_{t_c}^{\infty} g(t|H_0)\,\mathrm{d}t$$

nazývanou *velikost testu*. V jednom konkrétním experimentu (např. během detektoru ATLAS v průběhu několika let a analýzou všech zaznamenaných dat za účelem hledání Higgsova bosonu) následně získáme jednu konkrétní hodnotu testovací statistiky t_{obs}. Bude-li $t_{\mathrm{obs}} > t_c$, zamítneme hypotézu H_0 (statistika t by v případě její platnosti měla nabývat spíše nižších hodnot). V případě, že $t_{\mathrm{obs}} < t_c$, hypotézu H_0 akceptujeme (nezamítneme). Obdobně se definuje

$$\beta \equiv \int_{-\infty}^{t_c} g(t|H_I)\,\mathrm{d}t$$

a z ní odvozená veličina $1 - \beta$ nazývaná *síla testu*.

Cílem je mít testovací statistiku takovou, že můžeme vybrat t_c tak, že α i β jsou malé.

V praxi se často vybírá velikost testu α, např. 5%, podle které se určí t_c. Parametr α definuje pravděpodobnost, s jakou při platnosti hypotézy H_0 zpozorujeme t v kritick0 oblasti $(t > t_c)$.

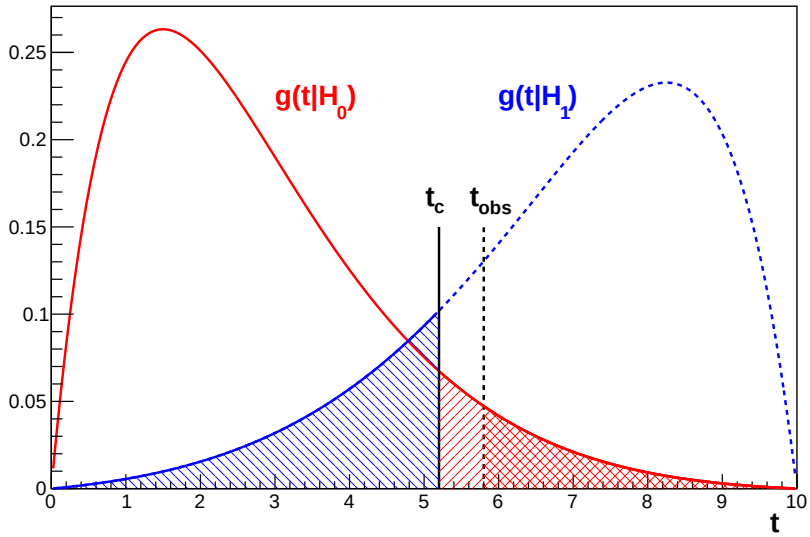
Na základě t_{obs} se pak počítá tzv. p-value, tj pravděpodobnsot toho, že při platnosti H_0 dostaneme výsledek pro t v takovémto či ještě horším nesouladu s očekáváním

$$p{-}\mathrm{val} \equiv \int_{t_{\mathrm{obs}}}^{\infty} g(t|H_0)\,\mathrm{d}t\,.$$

Na Obr. 30 jsou ilustrovány výše uvedené pojmy, Tabulka 5 shrnuje případy dobrých a správných rozhodnutí a jejich pravděpodobnosti.

platí hypotéza	zamítli jsme H_0	akceptovali jsme H_0
H_0	$P = \alpha$, chyba I. druhu	$P = 1 - \alpha$, správné rozhodnutí
H_1	$P = 1 - \beta$, správné rozhodnutí	$P = \beta$, chyba II. druhu

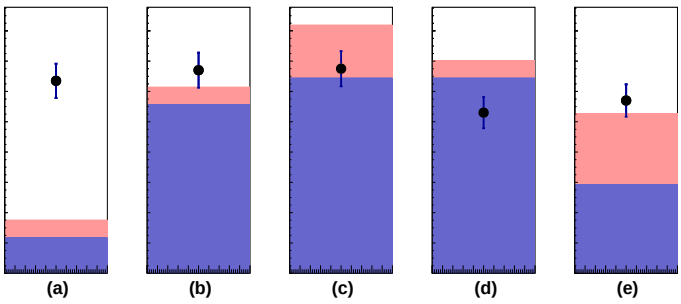
Tabulka 5: Tabulka možných případů testování hypotéz a jejich pravděpodobností P.



Obrázek 30: Ilustrace pojmů kritická oblast $t = t_c$, velikost testu α (všechna červeně šrafovaná oblast) a parametr β (modře šrafovaná oblast) a pozorované hodnoty testovací statistiky ($t = t_{\text{obs}}$) a p-value (červená hustě šrafovaná oblast) pro testovanou hypotézu H_0 (červeně) a alternativní hypotézu H_1 (modře).

9.2 Counting Experiment

V případě prostého počítání událostí daného typu hovoříme o counting experimentu. Data srovnáváme s předpovědí, která je dána součtem předpovědí počtu událostí z procesů z pozadí a ze signálního procesu, viz. Obr. 31 pro ilustraci různých možných případů.

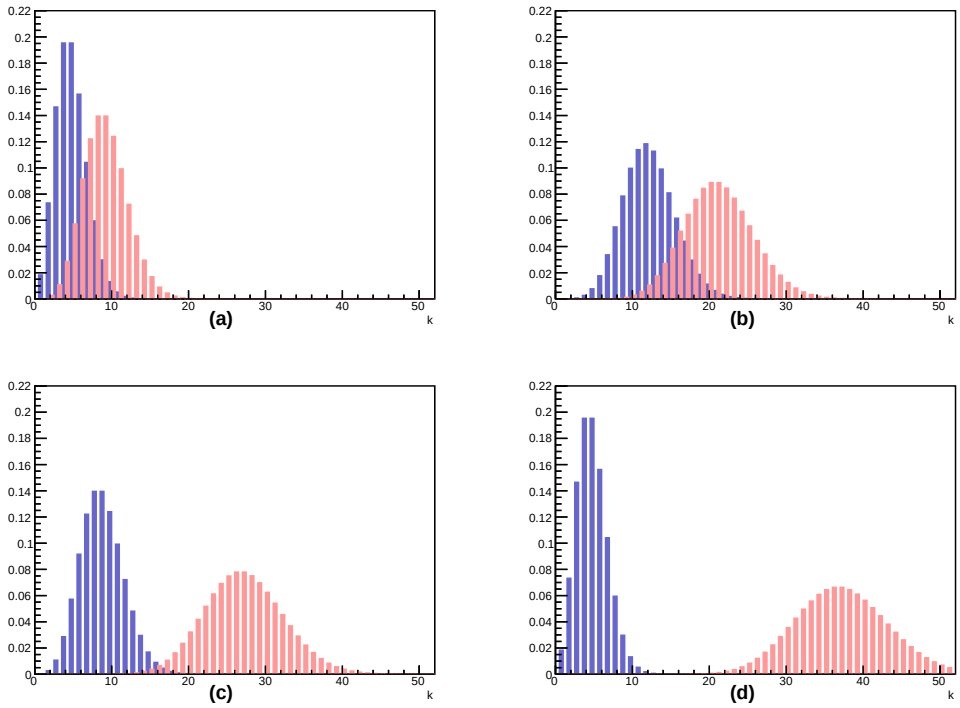


Obrázek 31: Ilustrace různých stupňů souhlasu dat a předpovědí při různých očekávaných poměrech počtu událostí z pozadí (modře) a signálního procesu (červeně), jejichž součet (stack) srovnáváme s daty.

Při rozhodování mezi hypotézami může jako testovací statistika posloužit i prostý počet pozorovaných případů. Pozorujeme-li d událostí, pak s uvážením Poissonovy statistiky pro předpovězený počet případů podle každé z hypotéz je možné spočítat p_0 value následovně

$$p_0 = \sum_{k=d}^{\infty} \frac{\nu_b^k}{k!} \mathrm{e}^{-\nu_b} = 1 - \sum_{k=0}^{d-1} \frac{\nu_b^k}{k!} \mathrm{e}^{-\nu_b} \,.$$

Viz také Obr. 32 pro ilustraci různých možných případů, jak mohou být obě hypotézy H_0 a H_1 odděleny. Pozor, zde H1 značí model, že do daného pozorování přispívá jak pozadí, tak signál.



Obrázek 32: Ilustrace různých překryvů hustot pravděpodobnsoti pro různá Poissonova rozdělení dvou různých hypotéz.

Užitečné jsou také relace [12] definující normalizovanou nekompletní "dolní" gama funkci, gama funkci, a jejich využití pro vyhodnocení sumy Poissonovských členů pro p_0 hodnotu výše, a to

$$\Gamma(x,y) \equiv \frac{1}{\Gamma(x)} \int_0^y \xi^{x-1} \mathrm{e}^{-\xi}\, \mathrm{d}\xi \tag{5}$$

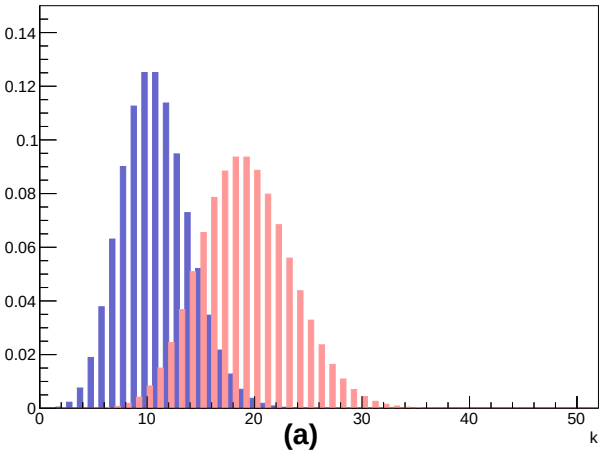
$$\Gamma(x) \equiv \int_0^\infty \xi^{x-1} \mathrm{e}^{-\xi}\, \mathrm{d}\xi \quad \Rightarrow \quad \lim_{y\to\infty} \Gamma(x,y) = 1 \tag{6}$$

$$\sum_{k=d}^{\infty} \frac{\nu_b^k}{k!} \mathrm{e}^{-\nu_b} = \Gamma(d,\nu_b)\,. \tag{7}$$

Příklad:

Dle agenturní zprávy ČTK ze dne 2.2.2024 "Počet nehod na železničních přejezdech byl letos v lednu oproti loňsku téměř dvojnásobný. V tomto roce se v lednu stalo 18 střetů na přejezdech, zatímco v lednu 2023 deset. Vyplývá to ze statistiky Drážní inspekce (DI) zveřejněné na jejím webu."stanovte signifikanci rozdílu v počtu nehod na přejezdech v letech 2024 a 2023. Jak byste navrhli spočítat nějakou pravděpodobnost (ne)kompatibility obou výsledků? Viz také Obr. 33.

Úvahy



Obrázek 33: Srovnání Poissonových rozdělení se střední hodnotou 10 (modře) a 18 (červeně).

- Mlčky předpokládáme, že jde o dva náhodné výběry pro stejnou náhodnou veličinu, a ze stejného Poissonova rozdělení. Je to ale oprávněné, příp. jaké jsou limitace?
- Můžeme nadále testovat kompatibilitu, že jde o jedno či různá Poissonova rozdělení.
- I tak, jedná se skutečně o náhodné veličiny? Je možné, že někdo se může pokoušet statistiky cíleně ovlivňovat nějakou cílenou akcí např. v zabezpečovacím systému. Tj. jednotlivé případy nemusejí být náhodné.
- Mohlo dojít k opravě nezanedbatelné frakce přejezdů na zabezpečenější, mohl být schválen zákon o úpravě rychlosti na přejezdech, auta mohla získat do povinné výbavy automatické varovné systémy apod.
- Je v ČR v obou letech stejný počet lidí, aut, vlaků a přejezdů? Jezdí auta a vlaky stejně často?
- Je stejné počasí? Měnilo by pouze očekávanou hodnotu Poissonova rozdělení nebo i tvar rozdělení?
- Nemáme například příliš posunutý kalendář oproti oběhu Země okolo Slunce, s nezanedbatelným vlivem na počasí, délku dne apod?
- Experiment nelze se stejnými podmínkami opakovat, tj. tvrzení nelze jednoduše přímočaře podložit dalšími daty.

9.3 Neyman-Pearsonovo lemma, odhad síly signálu

9.4 χ^2 test pro data a fit; a mezi dvěma histogramy

Pearsonův χ^2 test je založen na vyhodnocení výrazu, kterému se říká chí-kvadrát

$$\chi^2_{\text{fit;data}} \equiv \chi^2_{\text{fit}}(\boldsymbol{\theta}|\mathbf{d}) = \sum_{i=1}^{n_{\text{bins}}} \frac{[d_i - f(x_i|\boldsymbol{\theta})]^2}{d_i}\,,$$

kde d_i je počet pozorovaných událostí v i-tém binu, $\boldsymbol{\theta}$ jsou parametry fitu a x_i je souřadnice středu binu i. Odhad rozptylu pozorovaného počtu událostí d_i je podle Poissonovské statistiky $\sqrt{d_i}$ a variance tohoto počtu je tedy opět d_i.

Výraz $\chi^2_{\text{fit}}(\boldsymbol{\theta}|\mathbf{d})$ by šlo minimalizovat, a fit provést hledáním nejlepších parametrů fitu v tomto smyslu. Testovací statistika $\chi^2_{\text{fit}}(\boldsymbol{\theta}|\mathbf{d})$ je rozdělena podle chí-kvadrát rozdělení a pro veličinu označenou např. u má tvar

$$\chi^2(u|N) = \frac{u^{N/2-1}}{\Gamma\left(\frac{N}{2}\right) 2^{N/2}} e^{-u/2}$$

a střední hodnotu N. Tedy

$$E[\chi^2(\cdot|N)]/N = 1\,,$$

což je známý základní test toho, zda je např. χ^2_{obs} nějakého fitu kompatibilní s očekávanou hodnotou. Tj. není pravda, že dobrý fit se vyznačuje co nejmenší hodnotou výrazu χ^2_{obs}, každý bod může v principu fluktuovat okolo jedné standardní odchylky.

Hypotézu, zda jsou data kompatibilní s tvarem daného fitu, můžeme dále testovat tak, že spočítáme pravděpodobnost, že obdržíme χ^2 rovno χ^2_{obs} **anebo horší**, a to spočítáním integrálu

$$\int_{\chi^2_{\text{obs}}}^{\infty} \chi^2(u|N)\,\mathrm{d}u \equiv p_0$$

Tyto integrály jsou tabelovány (závisí nejen na χ^2_{obs} ale i na N!) nebo se dají spočítat pomocí funkcí v balíčku ROOT jako tzv. χ^2 probability. Počet stupňů volnosti fitu je počet binů fitovaného histogramu mínus počet volných parametrů fitu. Na místě je však obezřetnost: primárně náz zajímá otázka, zda na základě dat můžeme říci, že tato pocházejí z daného pravděpodobnostního rozdělení, popř. jaká je pravděpodobnost toho souladu. Výše uvedený postup však odpovídá na otázku, jaká jepravděpodobnost toho, že data jsou nekompatibilní takto anebo ještě více. Nicméně jedná se o v praxi nadále hojně používaný přístup, který má svá úskalí a je citlivý na statistické fluktuace.

Pro posouzení souladu dvou histogramů, např. z data a nějaké teorie z Monte Carlo simulace (MC), se používá testovací statistika

$$\chi^2_{\text{MC;data}} \equiv= \sum_{i=1}^{n_{\text{bins}}} \frac{\left[d_i - y_i^{\text{MC}}\right]^2}{\left(\sigma_i^{\text{data}}\right)^2 + \left(\sigma_i^{\text{MC}}\right)^2}\,,$$

která reflektuje obecně nepoissonovské neurčitosti dat, ale také druhého histogramu (může jít například o statistické či modelovací neurčitosti z Monte Carlo metody, pomocí které byl historam nagenerován).

9.5 Kompatibilita dvou měření

Pro Gaussovsky rozdělenou veličinu je pravděpodobnost, že nalezneme hodnotu o x_0 nebo ještě vzdálenější od středu rozdělení μ dána funkcí

$$p(x_0) = 2\frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}-(\mu-x_0,\mu+x_0)} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x = \operatorname{erfc}\left(\frac{x_0-\mu}{\sqrt{2}\sigma}\right)\,,$$

kde σ je šířka rozdělení a kde

$$\mathrm{erfc}(y) \equiv \frac{2}{\sqrt{\pi}} \int_{y}^{+\infty} \mathrm{e}^{-x^2} \mathrm{d}x \,.$$

Ilustrace několika důležitých hodnot Gaussovských $p-$value hodnot viz Tabulka 6.

n	pozn.	p-val	$1-p$
1		0.31731	0.68269
2		0.04550	0.95450
3	evidence	0.00270	0.99730
5	objev	0.000000573	0.999999427

Tabulka 6: Tabulka hodnot p a $1-p$ pro p-values pro Gaussovo rozdělení pro různé hodnoty n.

Příklad:
Mějme následující dvě měření téže veličiny [13, 14]

$$\sigma(pp \to X) = 95.35 \pm 0.38\,(\mathrm{stat.}) \pm 1.25\,(\mathrm{exp.}) \pm 0.37\,(\mathrm{extr.})\,\mathrm{mb}$$

$$\sigma(pp \to X) = 96.07 \pm 0.18\,(\mathrm{stat.}) \pm 0.85\,(\mathrm{exp.}) \pm 0.31\,(\mathrm{extr.})\,\mathrm{mb}$$

Jaká je kompatibilita těchto dvou měření v rámci statistické a statistické+systematické chyby? (jakkoli jde o měření při různých težišťových energiích srážek). V tomto případě jsou výsledky dle výše uvedeného postupu uvedeny v Tabulce 7.

V rámci chyby	signifikance	p-val
Statistické	1.83	0.06706
Statistické $\oplus$ systematické	0.49	0.62360

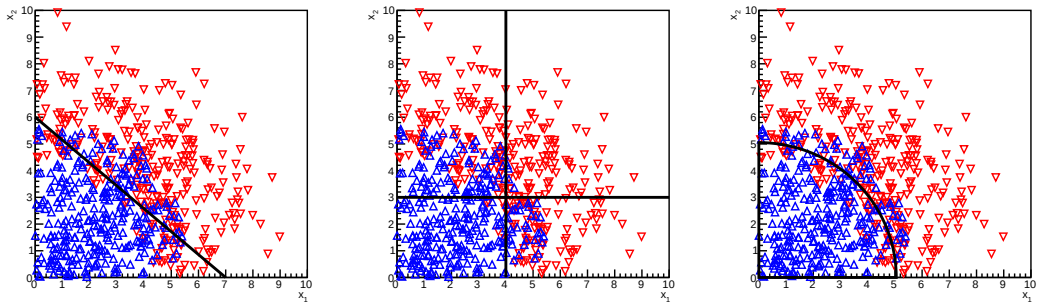
Tabulka 7: Tabulka hodnot signifikace a p-value pro srovnání dvou výsledků experimentu v rámci pouze statistické nebo statistické a systematické neurčitosti.

10 Klasifikace

Častým praktickým úkolem je rozhodnout se, zda daná událost je klasifikovatelná jako nějaký signál či pozadí. Vpodstatě se rozhodujeme, zda daná událost je více konzistentní či odpovídá jedné hyptéze či nějaké jiné. V praxi se často používají řezy v prostoru změřených hodnot nějakých pozorovatelných veličin x_i $(i = 1 \dots n)$, které o každé události $j = 1 \dots N_{\text{exp}}$ máme.

10.1 Řezy (cuty)

Na základě nějakého teoretického či empirického očekávání se provede volba na základě hodnoty jedné vybrané veličiny x_i, a to např. jako $x_i < t_i \Rightarrow$ akceptuj jako signál, jinak zavrhni (klasifikuj jako pozadí). Cílem je separovat pozadí od signálu. Vpodstatě jde o využití marginalizovaní hustoty pravděpodobnosti (vyintegrovanou přes všechny proměnné kromě jedné, podle které se rozhodneme) ať už známé teoreticky, ze simulace, či empiricky z dat. Opakováním takovýchto výběrů se provádějí další "pravoúhlé"řezy do prostoru možných hodnot pozorovatelných veličin, reprezentovaných náhodnými veličinami x_i Jde tedy o soubor (často lineárních) kritérií, které nutně nemusí být optimální. Pro ilustraci viz. Obr. 34. V dalších sekcích si ukážeme, jak lze klasifikaci dále rozvinout.



Obrázek 34: Ilustrace různých způsobů výběru hraniční oblasti pro klasifikaci (separaci) datových bodů příslušejících k různým hypotézám: vlevo řez na základě lineárního kritéria, uprostřed dva pravoúlé řezy, vpravo nelineární kritérium (kruh).

10.2 Fischerův diskriminant

Myšlenkou je najít takovou statistiku t lineární v proměnných x_i (připomeňme, že sadu naměřených hodnot $\{x_i\}_{i=1}^n$ máme k dispozici pro každé měření $j = 1 \dots N_{\text{exp}}$), tak, aby byla maximalizována separaci mezi hypotézami H_0 a H_1. Tuto statistiku si zapíšeme jako

$$t(\boldsymbol{x}) \equiv \sum_{i=1}^{n} a_i x_i = \boldsymbol{a} \cdot \boldsymbol{x} = \boldsymbol{a}^\top \boldsymbol{x}$$

Známe-li pro obě hypotézy střední hodnoty

$$(\mu_k)_i = \mathrm{E}\left[x_i | H_k\right] \equiv \int x_i \, f(\boldsymbol{x}|H_k) \mathrm{d}^n \boldsymbol{x} \,, \quad k = 0, 1$$

a kovariance

$$(V_k)_{ij} = \mathrm{Var}\left[x_i | H_k\right] = \mathrm{E}\left[(x - \mu_k)_i (x - \mu_k)_j | H_k\right] \equiv \int (x - \mu_k)_i (x - \mu_k)_j \, f(\boldsymbol{x}|H_k) \mathrm{d}^n \boldsymbol{x} \,, \quad k = 0, 1$$

pro všechny proměnné x_i (kde $f(\boldsymbol{x}|H_k)$ je hustota pravděpodobnosti náhodného vektoru $\boldsymbol{x}$ pokud platí hypotéza H_k), můžeme obdobně spočítat očekávané hodnoty statistiky t

$$\tau_k = \mathrm{E}\left[t|H_k\right] \equiv \int t\, g(t|H_k)\,\mathrm{d}t\,, \quad k = 0, 1$$

a její variance v případě, že platí H_0 či H_1

$$\Sigma_k^2 = \int (t-\tau_k)^2\, g(t|H_k)\,\mathrm{d}t\,, \quad k = 0, 1$$

Separace bude maximální, bude-li maximální rozdíl $\tau_0 - \tau_1$, a to "v jednotkách chyb"(přesněji variancí Σ_k^2) statistik t_k pro jednotlivé hypotézy, tj. zkusme minimalizovat výraz

$$J(\boldsymbol{a}) \equiv \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}\,.$$

Čitatel lze dále rozepsat jako

$$(\tau_0 - \tau_1)^2 = \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j = \sum_{i,j=1}^{n} a_i B_{ij} a_j = \boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a}$$

Podobně jmenovatel lze přepsat jako

$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1) = \boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}$$

kde matice $\boldsymbol{W} \equiv \boldsymbol{V}_0 + \boldsymbol{V}_1$ je sumou kovariančních matic náhodného vektoru $\boldsymbol{x}$ pro dvě uvažované hypotézy. Jest tedy

$$J(\boldsymbol{a}) = \frac{\boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}}$$

a minimalizací $J(\boldsymbol{a})$ vůči hledaným koeficientům $\boldsymbol{a}$ lze získat

$$\boldsymbol{a} \sim \boldsymbol{W}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\,,$$

tj. koeficienty jsou určeny až na libovolnou normalizační konstantu, a takto definované statistice t se říká Fischerův lineární diskriminant.

Pro specální případ, kdy jsou data rozdělena podle 2D Gaussova rozdělení

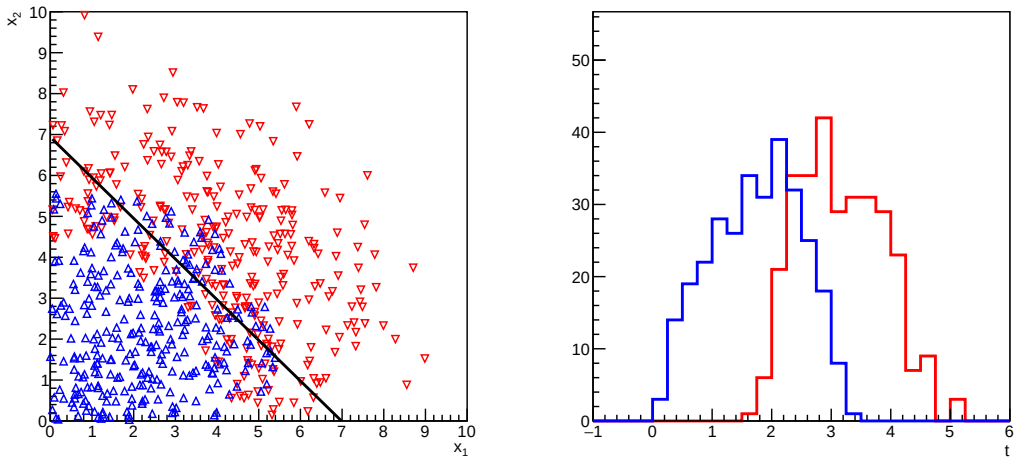
$$f(\boldsymbol{x}|H_k) = \frac{1}{(2\pi)^{n/2}\sqrt{|V|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{V}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right]\,, \quad k = 0, 1$$

přičemž současně jsou shodné i kovarianční matice $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{V}$, lze ukázat, že testovací statistika má tvar

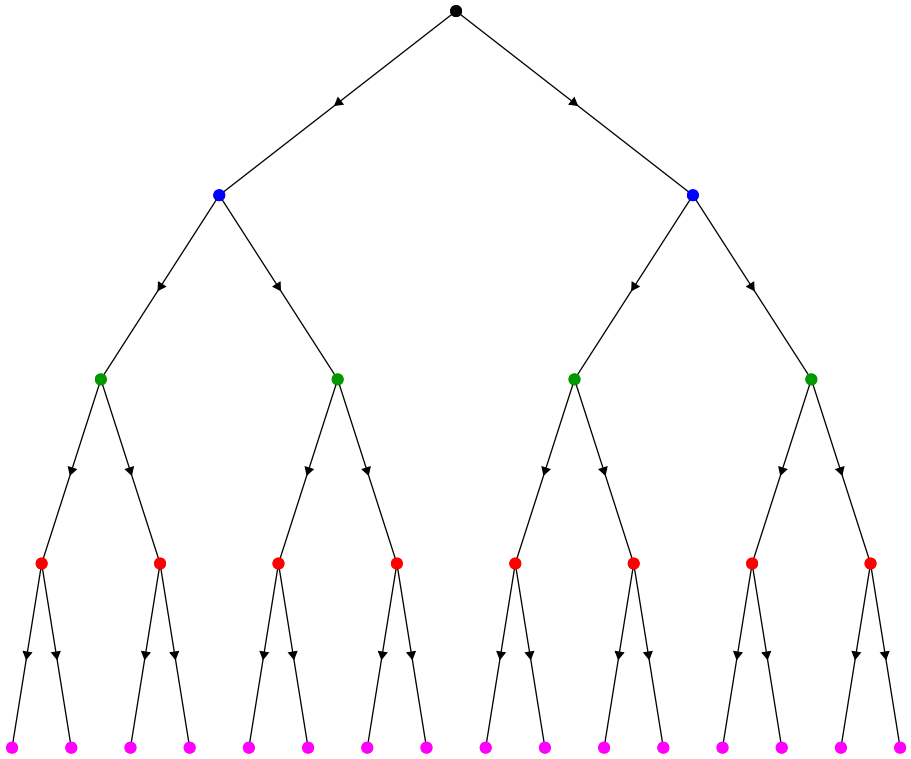
$$t(\boldsymbol{x}) = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{V}^{-1}\boldsymbol{x}$$

Následně lze definovat kritickou hodnotu t_c, kterou použijeme k výběru událostí konzistentích s hypotézou 0 či 1, a např. klasifikovat události jako spíše signální, pokud třeba $t > t_c$. Hranice oblasti $t = t_c$ je lineární funkce dat, a speciálně ve dvou rozměrech jde o funkci $a_1 x_1 + a_2 x_2 = t_c$ a tedy rovnici přímky v rovině x_1–x_2.

Pro ilustraci takovéto linární separace viz. Obr. 34, kde pseudodata byla nagenerována z 2D Gaussových rozdělení modifikovaných pro případ signálu (červeně) kvadratickou funkcí s korelací mezi x_1 a x_2 rovnou 0.45.



Obrázek 35: Ilustrace schopnosti separace a rozdělení ad-hoc lineárního diskriminantu (statistiky) t pro případ dat z Obr. 34.



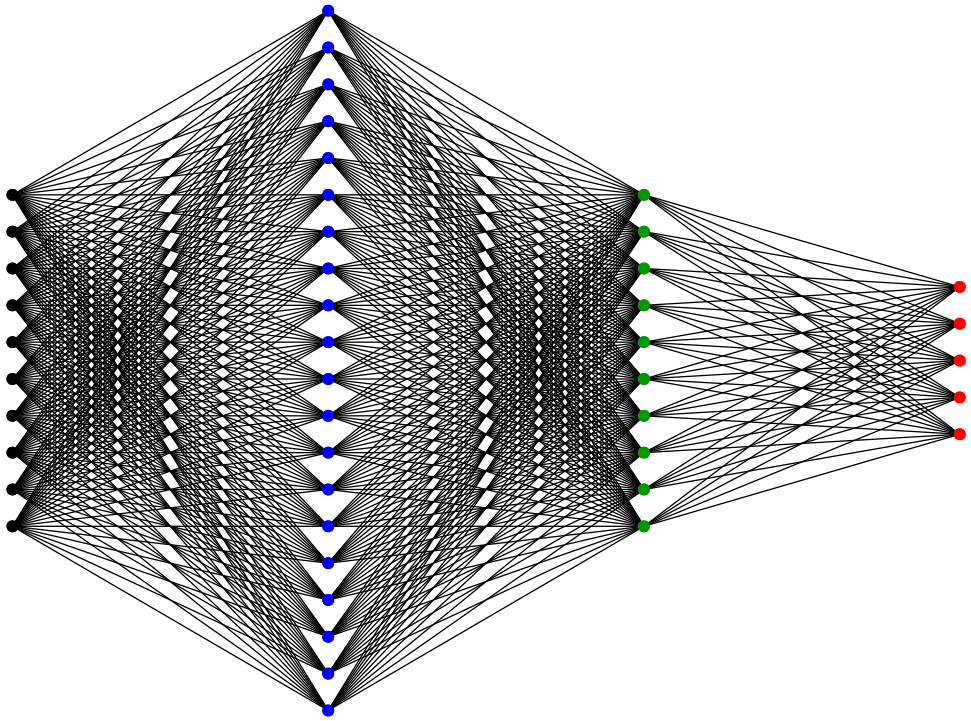
Obrázek 36: Ukázka architektury binárního stromu.

10.3 Rozhodovací stromy

Architektura

10.4 Umělé neurální sítě

Architektura

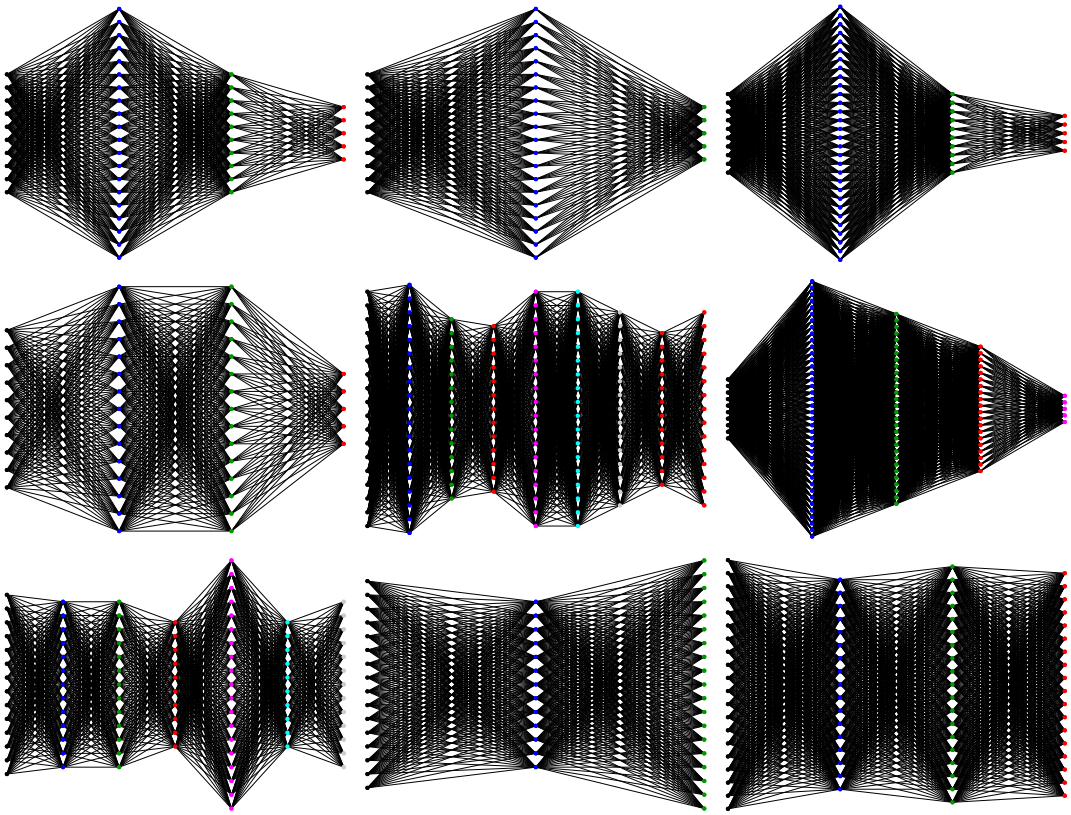


Obrázek 37: Ukázka možné architektury umělé neurální sítě.

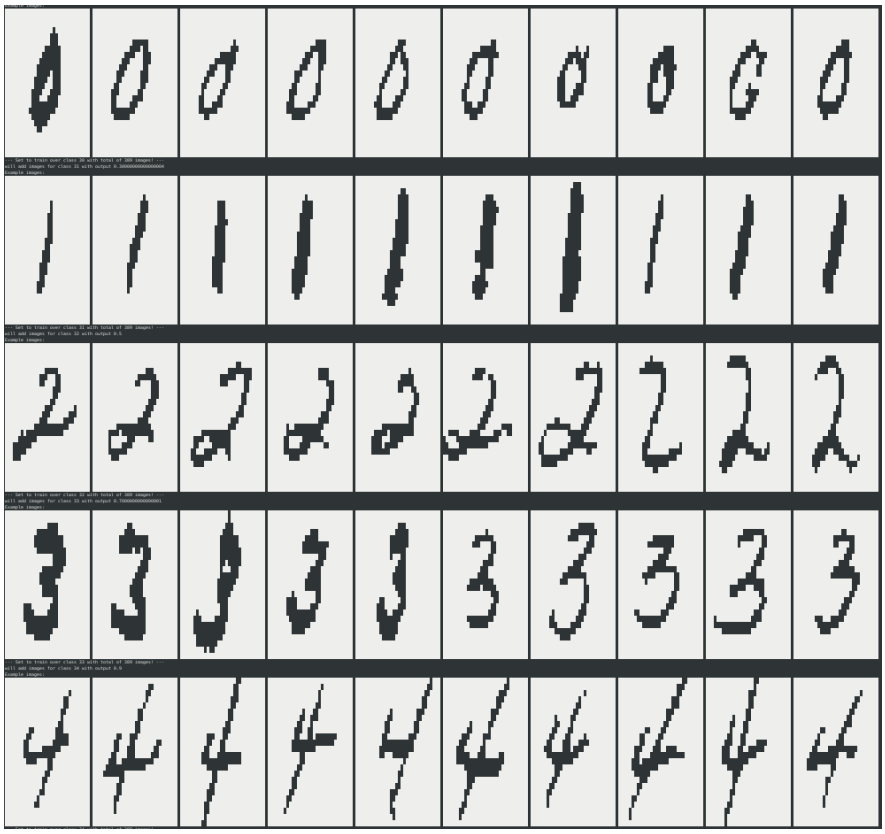
10.5 Strojové učení

učení s učitelem a bez učitele

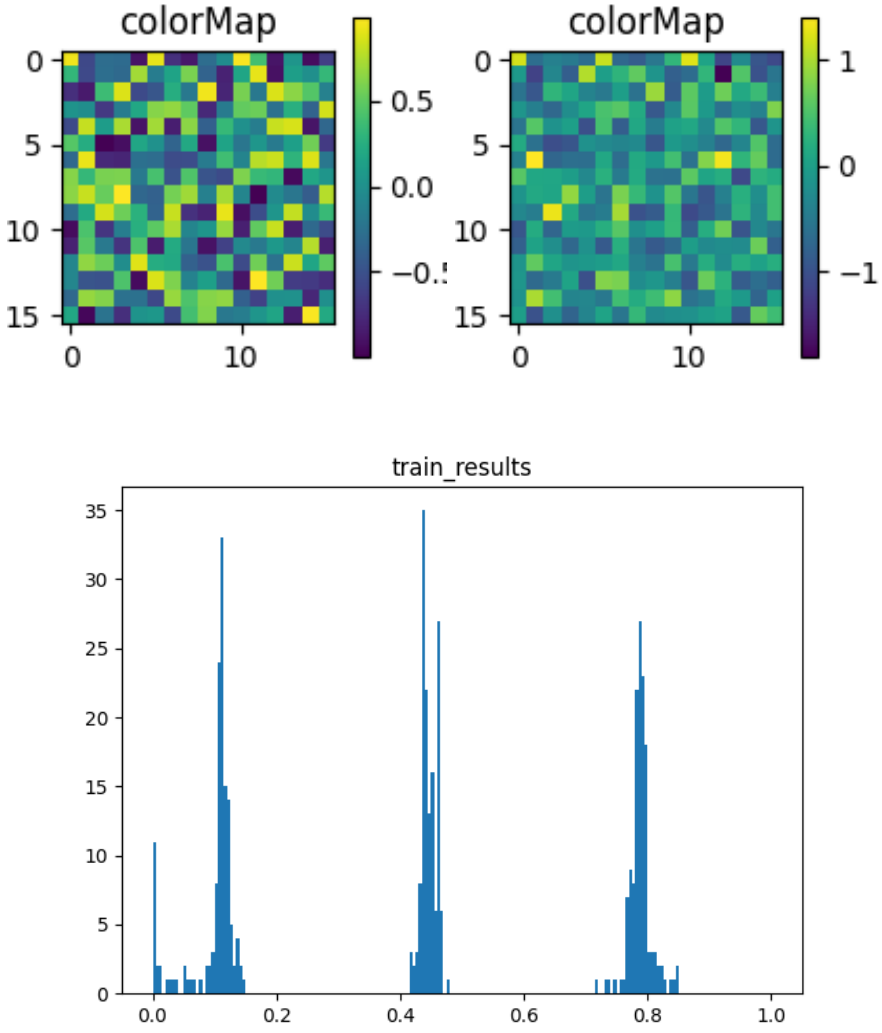
ROC křivka, AUC



Obrázek 38: Ukázky možných architektur umělé neurální sítě.



Obrázek 39: Ukázka z NIST databáze rukou saných číslic.



Obrázek 40: Nahoře: porování počátečních náhodných (clevo) a optimalizovaných vah (vpravo) mezi dvěma vrstvami NN. Dole: výstup NN diskriminantu natrénovaného na 3 různé číslice.

11 Aplikace

11.1 Náhodné výběry z Poissonova rozdělení

Nechť náhodná veličina $\mathcal{N}$ je rozdělena dle Poissonova rozdělení o parametru μ. Nechť N_i jsou náhodné výběry (měření) veličiny $\mathcal{N}$ pro $i = 1..n$. Již víme, že

$$\bar{N} \equiv \hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^{n} N_i$$

je nezaujatý odhad střední hodnoty $\mathrm{E}[\mathcal{N}] = \mu$ a nezaujatý odhad rozptylu je

$$\hat{\sigma}^2 \equiv \frac{1}{n-1}\sum_{i=0}^{n}(N_i - \bar{N})^2$$

kde rozptyl samotný je pro Poissonovo rozdělení opět

$$\sigma^2 \equiv \mathrm{Var}[\mathcal{N}] \equiv \mu$$

Zajímá nás, jaká je chyba průměru z měření $\bar{N}$, tj. jaký je odhad $\hat{\sigma}_{\bar{N}}$. Z věty XYZ již víme, že pro n opakovaných měření se rozptyl průměru zlepšuje s n jako

$$\hat{\sigma}_{\bar{N}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=0}^{n}(N_i - \bar{N})^2}\,.$$

Na druhou stranu bychom na sumu

$$S \equiv \sum_{i=0}^{n} N_i = n\bar{N}$$

mohli také pohlížet jako na (jeden) výběr z náhodné proměnné podléhající Poissonově statistice, a tedy pro odhad jejího rozptylu můžeme nezaujatě vzít

$$\hat{\sigma}_S = \sqrt{S}$$

a pro odhad rozptylu veličiny $\bar{N}$ tedy odtud můžeme také napsat

$$\hat{\sigma}'_{\bar{N}} = \frac{\sqrt{S}}{n} = \sqrt{\frac{\bar{N}}{n}}\,.$$

Podívejme se na poměr kvadrátů těchto zdánlivě, alespoň algebraicky, různých odhadů chyb $\hat{\sigma}_{\bar{N}}$ a $\hat{\sigma}'_{\bar{N}}$:

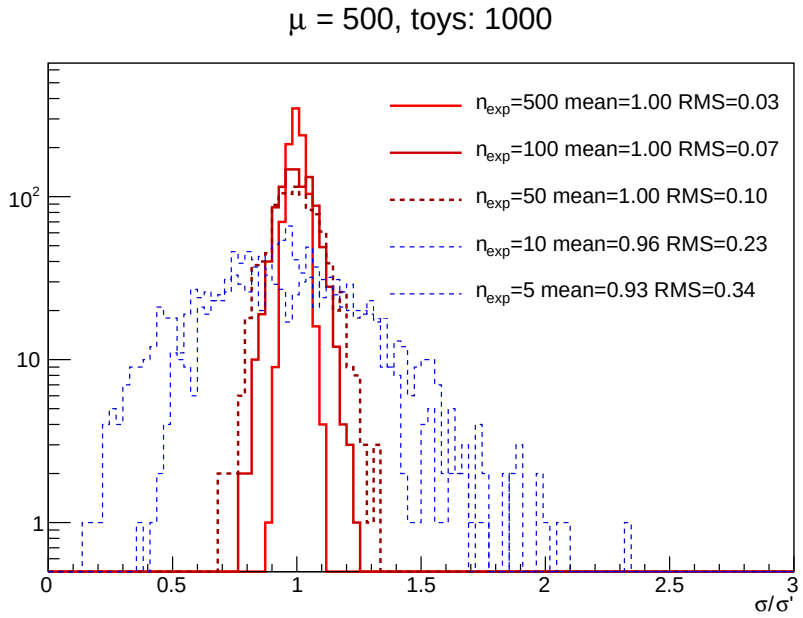
$$\frac{\sigma\mathcal{N}^2}{(\hat{\sigma}'_{\bar{N}})^2} = \frac{\frac{1}{n(n-1)}\sum_{i=0}^{n}(N_i - \bar{N})^2}{\frac{\bar{N}}{n}}\,.$$

Jakkoli čitatel a jmenovatel vypadají rozdílně, v obou případech výraz nabývá hodnoty odhadu střední hodnoty a tedy zároveň i rozptylu (odmocnině z variance) Poissonova rozdělení, a je tedy roven a konverguje

$$\frac{\sigma\mathcal{N}^2}{(\hat{\sigma}'_{\bar{N}})^2} = \frac{\frac{1}{n(n-1)}\sum_{i=0}^{n}(N_i - \bar{N})^2}{\frac{1}{n^2}\sum_{i=1}^{n} N_i} = \frac{\frac{1}{n}\hat{\sigma}^2}{\frac{1}{n}\hat{\mu}} \to \frac{\mu}{\mu} = 1$$

a oba rozptyly jsou si tedy v limitě velké statistiky (velkého n) ekvivalentní, viz též Obr. 41, přičemž odhad rozptylu $\hat{\sigma}'_{\bar{N}}$ je zjevně výpočetně jednodušší.

Hlouběji jde o to, že Poissonovo rozdělení má jen jeden parametr $\mu = \sigma^2$. Současně Poissonovo rozdělení pro velké μ přechází v Gaussovo rozdělení, ale takové, pro které je právě $\sigma = \sqrt{\mu}$. Limitně jde tedy o Gaussovu křivku, která ale tedy "ví, kde je"ve smyslu, že její šířka je odmocninou její vzdálenosti od počátku.



Obrázek 41: Ilustrace konzistence odhadů rozptylu průměru z Poissonova rozdělení o parametru $\mu = 500$ pro různé počty experimentů, poměr rozptylů studován a naplněn jako histogram přes 1000 pseudoexperimentů.

11.2 Propagace chyb: chybový pás fitu

Assume we have 1D-fit function

$$f \equiv f(x; \boldsymbol{a})$$

which can e.g. describe some data, and that we have a full covariance matrix $\mathrm{Cov}(a_i, a_j)$ of the best-fit parameters $\boldsymbol{a}$.

In every point x the fit uncertainty due to error and correlation of the fit parameters $\boldsymbol{a}$ is in the first order

$$\sigma_f^2 \equiv \sum_{i,j} \frac{\partial f}{\partial a_i} \frac{\partial f}{\partial a_j} \mathrm{Cov}(a_i.a_j) = \sum_i \left(\frac{\partial f}{\partial a_i} \right)^2 \sigma_{a_i}^2 + \sum_{i<j} 2 \frac{\partial f}{\partial a_i} \frac{\partial f}{\partial a_j} \mathrm{Cov}(a_i.a_j)$$

$$\sigma_f^2 = \sum_i \left(\frac{\partial f}{\partial a_i} \right)^2 \sigma_{a_i}^2 + \sum_{i<j} 2 \frac{\partial f}{\partial a_i} \frac{\partial f}{\partial a_j} \sigma_{a_i} \sigma_{a_j} \mathrm{corr}(a_i.a_j)$$

Alternativelly, having a fit function describing the fitted data

$$y(x) = f(x; \{a_i\})$$

together with the fit error matrix cov(i, j) one can derive the formula for the n-σ fit error boundary

$$y_\pm(x) = f(x; \{a_i\}) \pm n \Delta f(x; \{a_i\}) \,,$$

where

$$\Delta f(x;\{a_i\}) = \left[\sum_{i,j=0}^{n_{\rm par}} \left(\frac{\partial f}{\partial a_i}\right)\left(\frac{\partial f}{\partial a_j}\right) {\rm cov}^2(i,j)\right]^{\frac{1}{2}} .$$

Only in the case of independent (orthogonal) fit parameters and diagonal error matrix, this simplifies to

$$\Delta f(x;\{a_i\}) = \left[\sum_{i=0}^{n_{\rm par}} \left(\frac{\partial f}{\partial a_i}\right)^2 \sigma_i^2\right]^{\frac{1}{2}} .$$

Příklad:
Pro jednoduchý případ lineárního fitu funkcí $f(x;a,b) = ax+b$ s parametry a a b máme

$$\Delta f(x;a,b) = \sqrt{\sigma_a^2 x^2 + \sigma_b^2}\,.$$

a tedy horní hranice 1-sigma pásu lineárnho fitu je dána funkcí

$$f_+(x;a,b,\sigma_a,\sigma_b) = ax + b + \sqrt{\sigma_a^2 x^2 + \sigma_b^2}$$

a dolní hranice 1-sigma pásu lineárnho fitu je dána funkcí

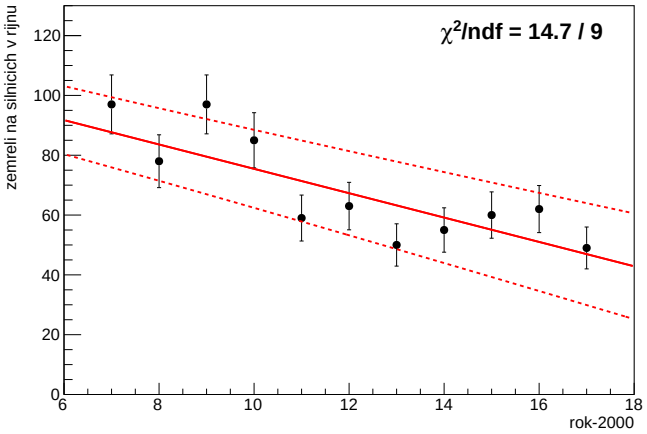
$$f_-(x;a,b,\sigma_a,\sigma_b) = ax + b - \sqrt{\sigma_a^2 x^2 + \sigma_b^2}\,.$$

Vidíme, že funkce definující horní a dolní hranici chybového pásu je oproti linární funkci modifikována odmocninou z kvadratické funkce. Ilustrace příslušného chybového pásu je na Obrázku 42.
Příklad:
Pro případ Gaussova rozdělení máme užitečné výrazy

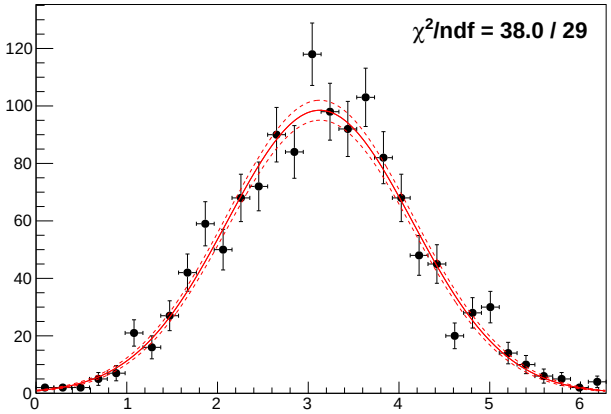
$$f(x;a,\mu,\sigma) = a\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$\frac{\partial f}{\partial a}(x) = \frac{f(x)}{a}\,, \quad \frac{\partial f}{\partial \mu}(x) = \frac{x-\mu}{\sigma^2} f(x)\,, \quad \frac{\partial f}{\partial \sigma}(x) = \frac{(x-\mu)^2}{\sigma^3} f(x)$$

Ilustrace příslušného chybového pásu je na Obrázku 43.



Obrázek 42: Příklad chyby fitu pro případ předpokládané lineární závislosti. Čárkovaně jsou zobrazeny 1-σ chybové pásy fitu. Většina fitovaných bodů leží uvnitř pásu.

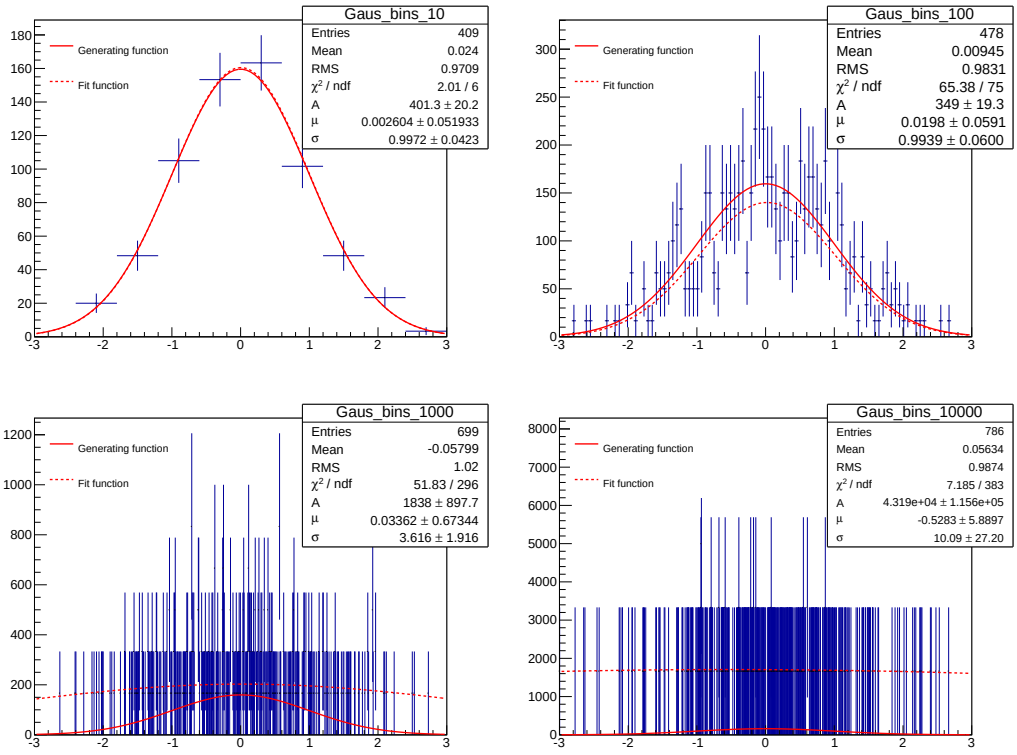


Obrázek 43: Příklad chyby fitu pro případ Gaussova rozdělení. Čárkovaně jsou zobrazeny 1-σ chybové pásy fitu. Většina fitovaných bodů leží v rámci očekávané statistické chyby uvnitř pásu.

11.3 Fitování frakce

11.4 Výběr binování pro fit

Rozdělení 400 bodů v historamech různě hustého binování jsou na Obrázku 44.



Obrázek 44: Příklady fitů pseudo dat vygenerovaných ze standardního Gaussova rozdělení pro počet nagenerovaných případů roven 400 pro různý počet dělení: 10, 10, 100, 10000, a výsledné hodnoty fitovaných parametrů. Počet událostí v každém binu byl vydělen šířkou binu.

11.5 Interval pokrytí Poissonova rozdělení

Pokrytí intervalu Poissonovsky rozdělených dat, aneb asymetrické chyby v případě malého počtu událostí. Pearsonova χ^2 funkce je v případě jednoho výběru z Poissonova rozdělení dána výrazem

$$\chi^2(\mu, k) = \frac{(k-\mu)^2}{\mu}\,,$$

kde data, k, tj. počet pozorovanýžch událostí, je dáno. Minimum této funkce nastává pro $\hat{\mu} = k$ a interval pokrytí a jeho okraje μ_1 a μ_2 můžeme definovat tak, že χ^2 se změní maximálně o jedničku, popř. obecně o Δ

$$\mu \in (\mu_1,\, \mu_2): \quad \chi^2(\mu, k) \leq \Delta\,,$$

odkud lze snadno odvodit dvě řešení

$$\mu_{1,2} = k + \frac{\Delta}{2} \pm \sqrt{k\Delta + \frac{\Delta^2}{4}}$$

a pro asymetrické šířky intervalu pokrytí okolo $\hat{\mu}$

$$\sigma_2 \equiv \sigma_{\text{up}} = \mu_2 - \hat{\mu}$$

$$\sigma_1 \equiv \sigma_{\text{down}} = \hat{\mu} - \mu_1$$

dostáváme (volbou znamének tak, aby $\sigma_i > 0$)

$$\sigma^{\text{up}}_{\text{down}} = \pm\frac{\Delta}{2} + \sqrt{k\Delta + \frac{\Delta^2}{4}}$$

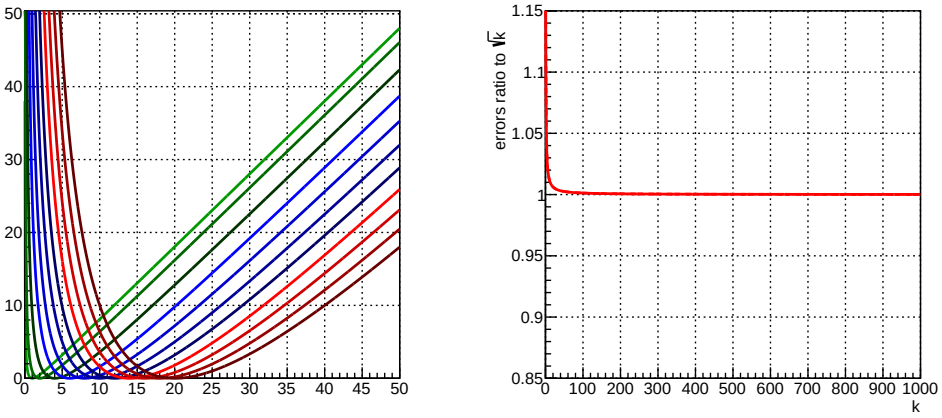
a pro obvyklou volbu "1-σ"pokrytí $\Delta = 1$ konečně

$$\sigma^{\text{up}}_{\text{down}} = \pm\frac{1}{2} + \sqrt{k + \frac{1}{4}}\,.$$

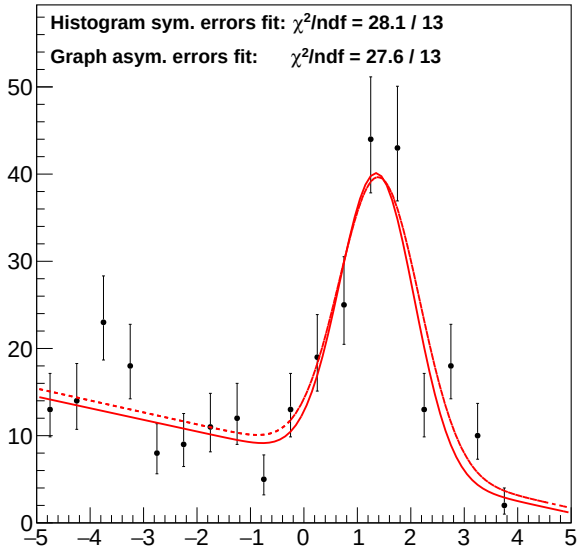
Obrázek 45 zobrazuje rozdělení $\chi^2(\mu, k)$ pro různé hodnoty k a dále poměr asymetrických chyb vůči standardní chybě Poissonova rozdělení $\sqrt{k}$.

11.6 b-tagging

ROC curve, operation point, efektivita pozadí a signálu, faktor potlačení pozadí



Obrázek 45: Rozdělení $\chi^2(\mu, k)$ pro různé hodnoty k (vlevo) a poměr chyb $\sigma^{\text{up}}_{\text{down}}/\sqrt{k}$ (vpravo).



Obrázek 46: Ilustrace velikosti asymetrických chyb v případě nízké statistiky na základě pokrytí intervalu $\chi^2(\mu, k) \leq 1$. Fit grafu s asymetrickými chybami resp. histogramu se symetrickými chybami je zobrazen plnou resp. čárkovanou čárou.

11.7 Kombinace měření

Příklad: nekorelované veličiny Uvažujme dva způsoby určení hmotnosti jetů v částicových experimentech, a to pomocí informace z kalorimetru anebo z dráhového detektoru. Takzvanou kalorimetrickou (Calo) a track-assisted (TA) hmotu můžeme v prvním přiblížení považovat za nezávislé veličiny, a lze provést jejich kombinaci, která bude preferoavt způsob, který bude pro danou událost přesnější. Intuitivně, měření s větši chybou (kterou lze parametriovat jako funkci rapidity či p_{T} jetu a odhanout např. v simulaci či v jiné nezávislé studii) by měl přispívat méně. Odtud lze odhadnout, že kombinaci můžeme provést např následovně:

$$m_{\mathrm{comb}} = \frac{1/\sigma^2_{\mathrm{Calo}}}{1/\sigma^2_{\mathrm{Calo}} + 1/\sigma^2_{\mathrm{TA}}}\, m_{\mathrm{Calo}} + \frac{1/\sigma^2_{\mathrm{TA}}}{1/\sigma^2_{\mathrm{Calo}} + 1/\sigma^2_{\mathrm{TA}}}\, m_{\mathrm{TA}}\,,$$

tj. vážený průměr obou odhadů, kde váhy jsou rovny kvardrátu převrácené hodnotě očekávaného rozptylu pro danou veličinu.

Zobecnění a odvození: technika BLUE Inverze kovarianční matice.

11.8 Věrohodnostní diskriminant

This is one of so-called multivariate techniques, trying to separate the signal and background based on differences in shapes of selected kinematical variables, ranging from angular correlations, multiplicity spectra of jets, tracks, to energy, momentum, mass spectra etc., and any combination of the aforementioned.

Assuming uncorrelated topological variables, one can express the probability of the event being signal-like as

$$\prod_{\mathrm{i}\in\mathrm{var}} \mathcal{S}_i^{\mathrm{fit}}(x_i)$$

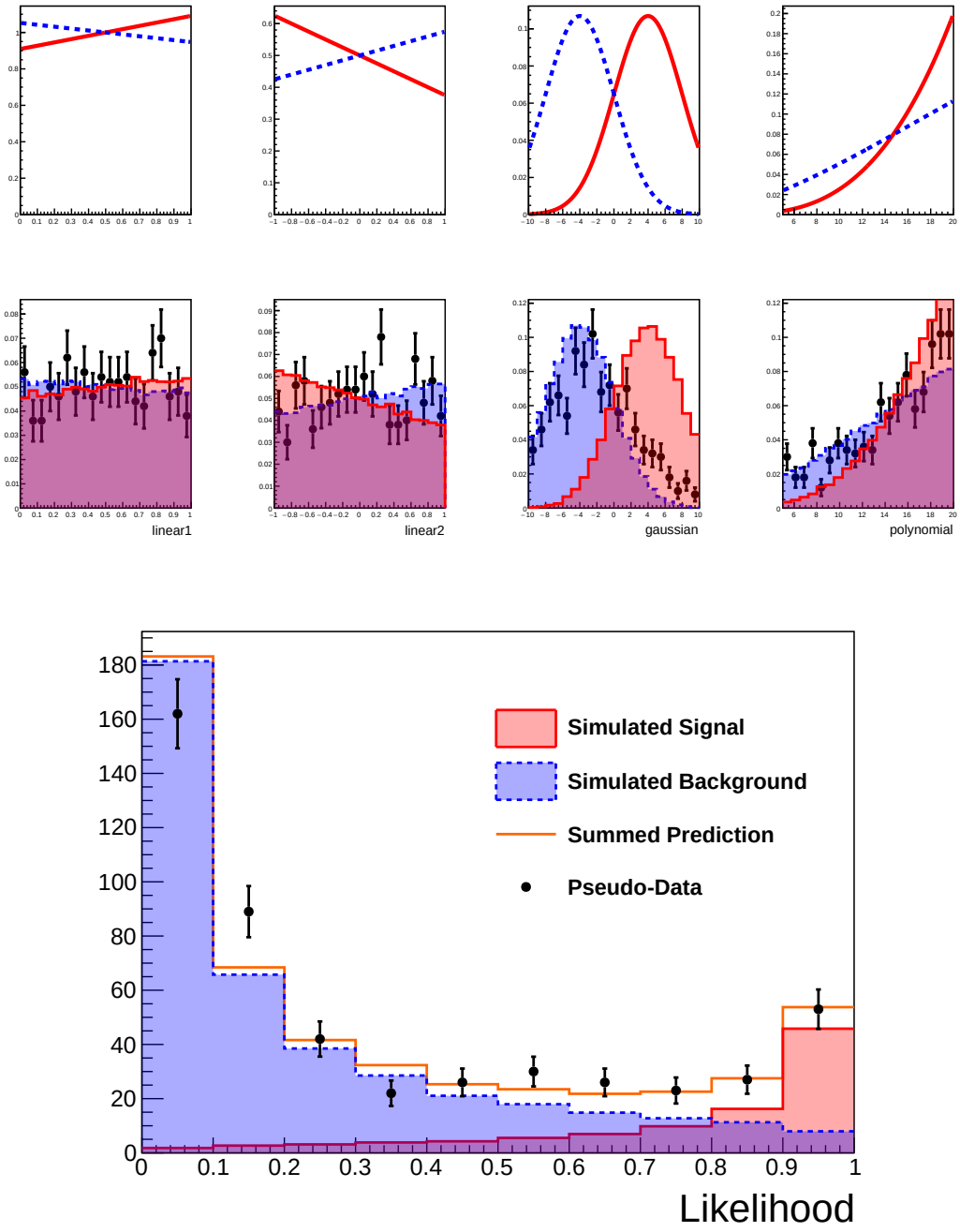
where the product runs through the variables and $\mathcal{S}_i$ is the fitted shape of the i-th variable (normalised to unit area) evaluated at the event's variable value x_i. The same fitting procedure can be done for background samples, i.e. one relies on signal and background shapes from simulation. For each event, one can compute the event topological likelihood as

$$\mathcal{L}_{\mathrm{event\,j}} \equiv \frac{\prod\limits_{\mathrm{i}\in\mathrm{var}} \mathcal{S}_i^{\mathrm{fit}}(x_i)}{\prod\limits_{\mathrm{i}\in\mathrm{var}} \mathcal{S}_i^{\mathrm{fit}}(x_i) + \prod\limits_{\mathrm{i}\in\mathrm{var}} \mathcal{B}_i^{\mathrm{fit}}(x_i)}$$

$$\mathcal{L}_{\mathrm{event\,j}} \equiv \frac{\xi}{1+\xi}\,, \qquad \xi \equiv \prod_{\mathrm{i}\in\mathrm{var}} \frac{\mathcal{S}_i^{\mathrm{fit}}(x_i)}{\mathcal{B}_i^{\mathrm{fit}}(x_i)}$$

This event-by-event likelihood can be filled into an histogram and fitted (binned likelihood fit) using likelihood shapes for signal and background from simulation to measure the signal fraction in data. The idea is that signal-like events accumulate closer to 1., while background closer to 0., aiming to separate the two. A cut on the likelihood value on an event-by-event basis is posible to enhance the signal purity, or just the signal fraction can be extracted e.g. to determine a cross-section.

Note that for this technique, the important thing is to choose variables with as little correlation as possible so that the multiplication of probabilities makes sense (holds only for independent observables). This can also be taken care by diagonalizing the correlation matrix in the variables space. When preparing the templates for each varible, one can either fit separately $\mathcal{S}_i$ and $\mathcal{S}_i$, or directly the ration $\mathcal{S}_i/\mathcal{B}_i$.



Obrázek 47: Vstupní proměnné pro signál a pozadí, a výsledná věrohodnost založená na odvozených hustotách pravděpodobnosti. Je patrná dobrá separace signálu a pozadí.

11.9 Metoda "ABCD" pro odhad pozadí

Using a 2D distribution of data events and MC signal events, it is possible to estimate the signal fraction on one region C of a 2D plot, where one expectes enhanced signal (with background still able to contaminate it). Control regions A, B, D are expected to be populated mostly by backroud, altough the correction for signal leakage to these regions is often needed.

The main assumption and the choice of the discriminating variables for the x and

y axes is that these variables should be as independent as possible, i.e. with a small correlation, at least for the background sample (which can be again verified using a background MC sample).

If independent, than the efficiency of the background to pass the isolation should be independent on the other variable, therefore for number of background events in each region A, B, C, B one can write

$$\frac{C_{\rm bg}}{A_{\rm bg}} = \frac{D_{\rm bg}}{B_{\rm bg}} \,.$$

Therefore the estimated bg. fraction in the signal region C is

$$\hat{C}_{\rm bg} = \frac{A_{\rm bg}\, D_{\rm bg}}{B_{\rm bg}}$$

and the desired estimated signal contribution to the C region is

$$\hat{C}_{\rm sig} \equiv C - \hat{C}_{\rm bg} \,.$$

Now to acount for the signal leakage from C to the bg. regions, one can evaluate on the signal MC sample the signal leakage fractions w.r.t. to the signal regions as

$$c_X \equiv \frac{X^{\rm MC,sig}}{C^{\rm MC,sig}} \,, \qquad X \in \{A, B, D\}$$

and apply these corrections coefficients on the measured data sample

$$C_{\rm sig} \equiv C - \frac{(A_{\rm bg} - c_A C_{\rm sig})\,(D_{\rm bg} - c_D C_{\rm sig})}{(B_{\rm bg} - c_B C_{\rm sig})}$$

which can be solved in terms of the desired $C_{\rm sig}$.

Applicattions: photon purity estimation using two separate photon isolation criteria; QCD bg. estimation using two lepton isolations nad missing transverse energy or transverse W mass variable; new physics signal searches.

12 Unfolding

12.1 Definice úlohy, bias a variance

12.2 Regularizace

12.3 Bayesovská metoda

Fully Bayesian Unfolding [15]

12.4 Korekce

Unfolding spektra získaného z dat na detektorové úrovni lze shrnout jako

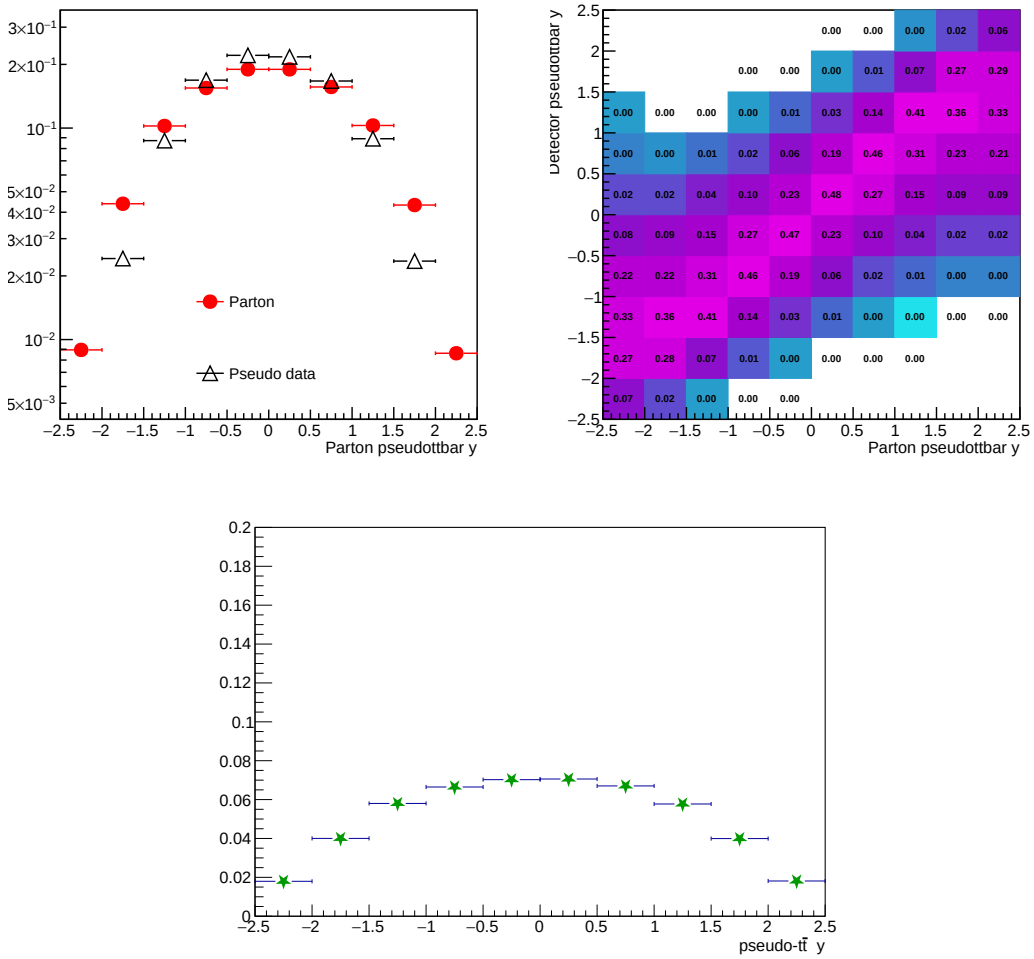
$$\mathrm{Unf}_i^{\mathrm{ptcl}} \equiv f_i^{\mathrm{eff}} \mathcal{M}_{ij}^{-1} \left[f_j^{\mathrm{acc}} \left(\mathrm{D} - \mathrm{Bg} \right)_j \right]$$

popř. detailněji jako

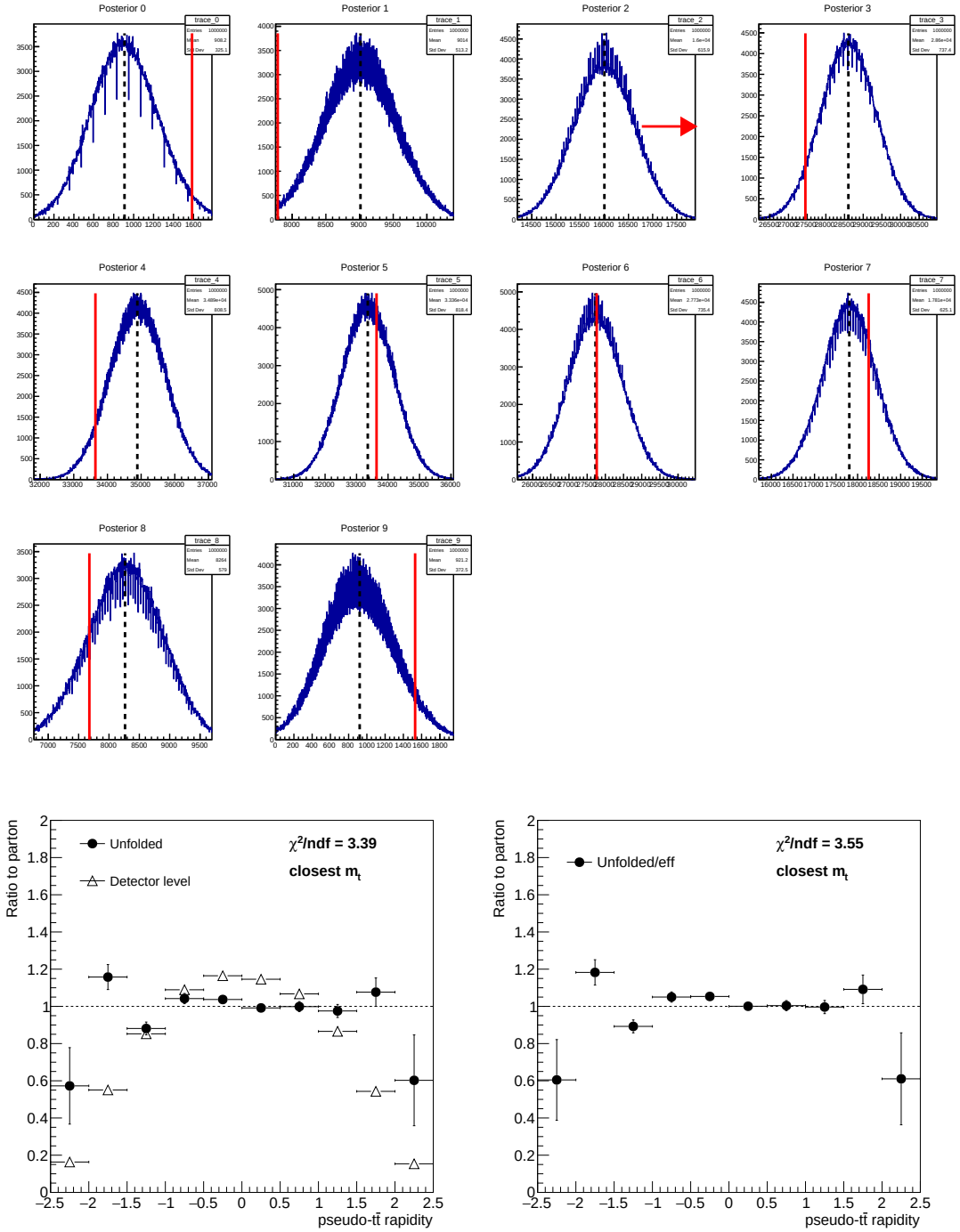
$$\mathrm{Unf}_i^{\mathrm{ptcl}} \equiv \underbrace{\left(\frac{\mathrm{Ptcl}^{\text{passed part}}_{\text{incl. det over\&under flows}}}{\mathrm{Ptcl}^{\text{passed det\&ptcl}}_{\text{w/o det over\&under flows}} \right)_i}_{\text{efficiency correction}} \underbrace{\left(\mathcal{M}_{ij}^{\text{passed det\&ptcl}} \right)^{-1}}_{\text{unfolding}} \left[\underbrace{\left(\frac{\mathrm{Det}^{\text{passed det\&part}}_{\text{w/o ptcl over\&under flows}}}{\mathrm{Det}^{\text{passed det}}_{\text{incl. ptcl over\&under flows}} \right)_j}_{\text{acceptance correction}} \left(\mathrm{D} - \mathrm{Bg} \right)_j \right],$$

kde

- $\mathrm{Ptcl}^{\text{passed part}}_{\text{incl. det over\&under flows}}$ je spektrum na *částicové úrovni*, tj. v tzv. částicovém fázovém prostoru, popř. po řezech, které tento prostor definují (passed part.), a zahrnuje tedy i události pod či nad rozsah osy na detektorové úrovni;
- $\mathrm{Ptcl}^{\text{passed det\&ptcl}}_{\text{w/o det over\&under flows}}$ je projekcí response matice na osu příslušející *částicové* úrovni, a to bez zahrnutí událostí, které na *detektorové* úrovni spadají pod či nad rozsah rozsah osy;
- $\mathrm{Det}^{\text{passed det\&part}}_{\text{w/o ptcl over\&under flows}}$ je projekcí response matice na osu příslušející *detektorové* úrovni, a to bez zahrnutí událostí, které na *částicové* úrovni spadají pod či nad rozsah rozsah osy;
- $\mathrm{Det}^{\text{passed det}}_{\text{incl. ptcl over\&under flows}}$ je spektrum na *detektorové úrovni*, tj. v tzv. detektorovém fázovém prostoru, tj. po řezech, které tento prostor definuji, a zahrnuje tedy i události pod či nad rozsah osy na částicové úrovni.



Obrázek 48: Ingredience FBU unfoldingu.



Obrázek 49: Příklad posteriorů FBU unfoldingu a closure test ratios.

Reference

[1] Glen Cowan. Statistical Data Analysis. 1998, 2002. Oxford Science Publications, 1998, reprinted 2002.

[2] Olaf Behnke, Kevin Kroninger, Gregory Schott, Thomas Schorner-Sadenius. Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods. 2013.

[3] Gerhard Bohm, Günter Zech. Introduction to Statistics and Data Analysis for Physicists. 2017. http://inspirehep.net/record/704473/files/vstatmp_e17.pdf.

[4] Frederick James. Statistical Methods in Experimental Physics. 2006.

[5] Glen Cowan. Statistical Data Analysis Lectures. 2017. http://www.pp.rhul.ac.uk/~cowan/stat_course.html.

[6] Wikimedia. Poisson limit theorem. *Wikipedia*, 2017. https://en.wikipedia.org/wiki/Poisson_limit_theorem.

[7] Particle Data Group. Particle Data Group, Probability. *PDG*, 2016. http://pdg.lbl.gov/2017/reviews/rpp2016-rev-probability.pdf.

[8] Particle Data Group. Particle Data Group, Statistics. *PDG*, 2016. http://pdg.lbl.gov/2017/reviews/rpp2016-rev-statistics.pdf.

[9] Kyle S. Cranmer. Kernel estimation in high-energy physics. *Comput. Phys. Commun.*, 136:198–207, 2001.

[10] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2), Feb 2011.

[11] Procedure for the LHC Higgs boson search combination in Summer 2011. Technical report, CERN, Geneva, Aug 2011.

[12] Georgios Choudalakis. On hypothesis testing, trials factor, hypertests and the BumpHunter. In *PHYSTAT 2011*, 1 2011. https://arxiv.org/abs/1101.0390.

[13] Morad Aaboud et al. Measurement of the total cross section from elastic scattering in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Phys. Lett.*, B761:158–178, 2016.

[14] Georges Aad et al. Measurement of the total cross section from elastic scattering in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Nucl. Phys.*, B889:486–548, 2014.

[15] Georgios Choudalakis. Fully Bayesian Unfolding, 2012. https://arxiv.org/abs/1201.4612.