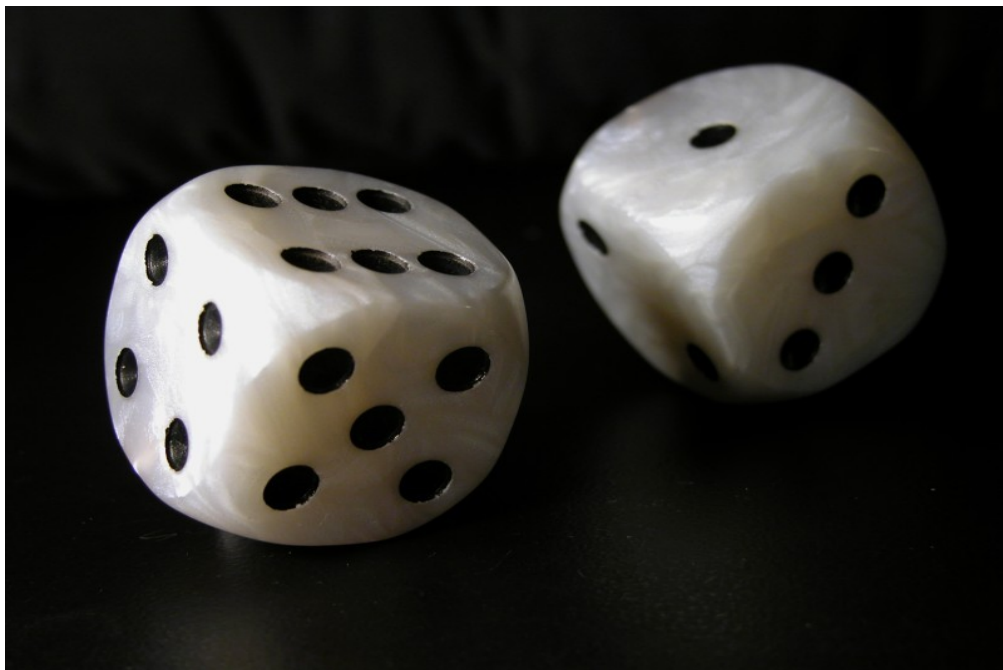


Wprowadzenie do statystyki oraz analizy danych

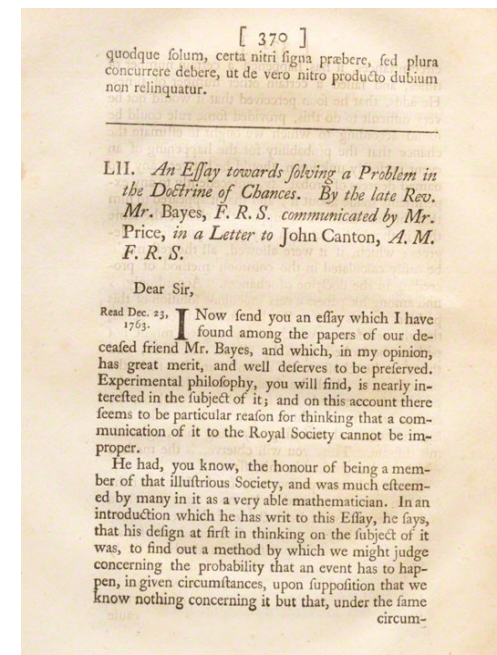
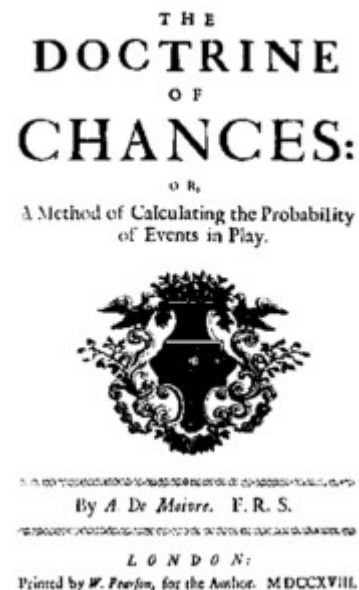
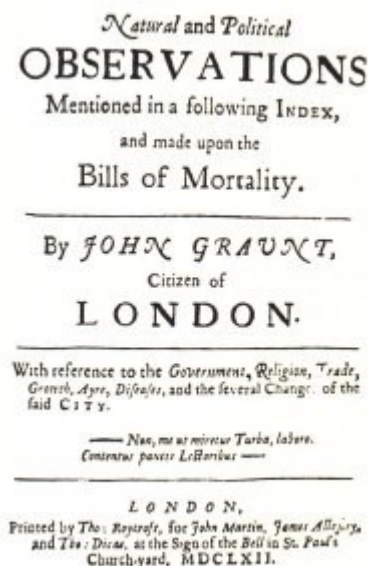


Marcin Wolter
IFJ PAN

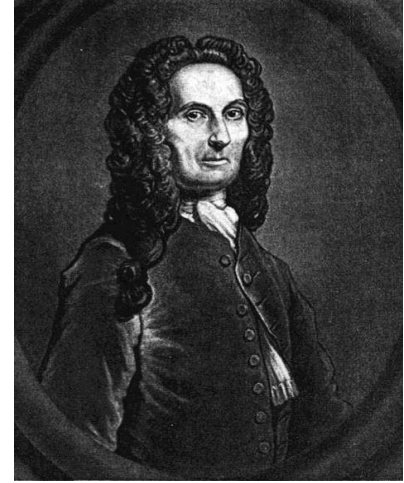
11 lipca 2014

Statystyka

- Statystyka – nauka, która bada i opisuje zjawiska losowe.
- Pierwsze prace – Al-Kindi użył statystyki do złamania szyfru Cezara (podstawieniowego), badał częstość występowania poszczególnych liter.
- Impuls do rozwoju dały badania demograficzne oraz nad grami losowymi.
- 1663 John Graunt „*Natural and Political Observations upon the Bills of Mortality*”.



Co to jest zdarzenie oraz prawdopodobieństwo?



- Zdarzeniem elementarnym będziemy nazywać każdy wynik eksperymentu (doświadczenia), którego wartość będzie zależeć od przypadku. Jest to pojęcie pierwotne teorii prawdopodobieństwa.
- **Prawdopodobieństwem (w sensie definicji częstościowej) zdarzenia A nazywamy granicę (przy N dążącym do nieskończoności) stosunku liczby prób n przy której zaszło zdarzenie A do liczby wykonanych prób N:**

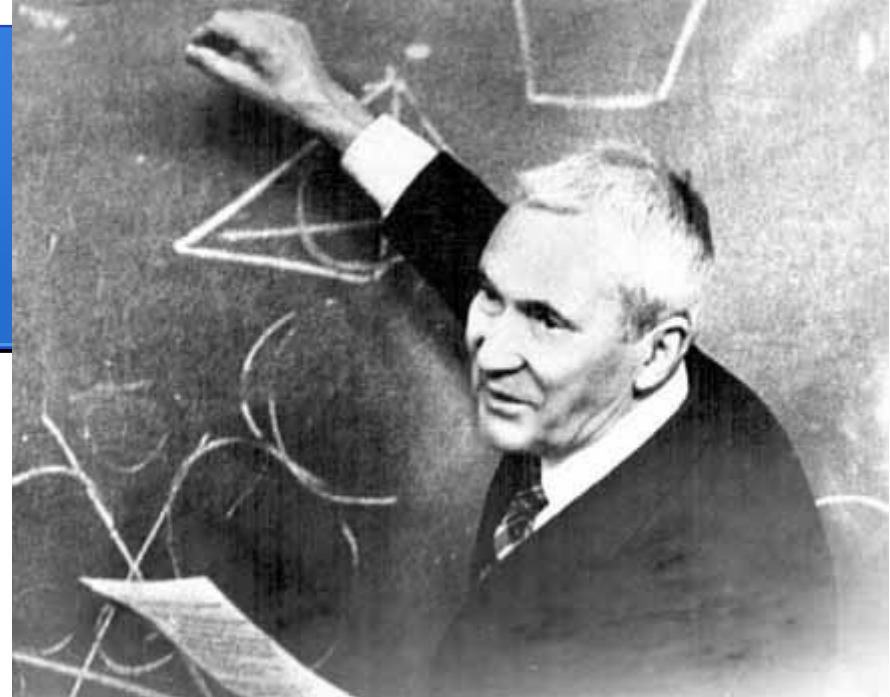
$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

Jakie jest prawdopodobieństwo wyrzucenia „6” w rzucie kostką? Takie, jaka część nieskończonej serii rzutów da nam w wyniku „6”.

- Definicja pochodzi od Abrahama de Moivre’a (1667-1754) - napisał pierwszy podręcznik statystyki „The Doctrine of Chances” (1718).

Inne definicje prawdopodobieństwa

Prawdopodobieństwo możemy definiować na wiele sposobów, niekoniecznie tak, jak de Moivre...



Андрей Никола́евич Колмогоров (1903-1987)

Definicja aksjomatyczna Kołmogorowa: prawdopodobieństwem nazywamy taką funkcję P zdefiniowaną na przestrzeni zdarzeń elementarnych, która każdemu zdarzeniu A przyporządkowuje liczbę $P(A)$, taką że:

- $P(A) \geq 0$ dla każdego zdarzenia A
- $P(A) = 1$ dla zdarzenia pewnego A
- $P(A \cup B) = P(A) + P(B)$ gdy zdarzenia A i B są rozłączne (wzajemnie wykluczają się).

Inne definicje prawdopodobieństwa



- **Definicja Bayesa:**
- Prawdopodobieństwo „a priori”, czyli bezwarunkowe, jest rozumiane jako miara przekonania, opartego na racjonalnych przesłankach, że dane wydarzenie nastąpi.
- W następnym dopiero kroku wykonujemy doświadczenia, czyli obserwacje, a ich wyniki pozwalają zmodyfikować wstępne oczekiwania. Otrzymujemy prawdopodobieństwo „a posteriori”, czyli prawdopodobieństwo wynikowe, które jest miarą oczekiwania wystąpienia danego zdarzenia po zanalizowaniu przeprowadzonych obserwacji.
- Zwolennikami podejścia Bayesa byli P. S. Laplace, H. Poincare czy też znany ekonomista John Keynes, argumentując, że w taki właśnie sposób przebiega nasze poznawanie świata.

Thomas Bayes (1702 - 1761) brytyjski matematyk i duchowny presbiteriański. Najistotniejsze dzieło: „Essay Towards Solving a Problem in the Doctrine of Chances”.

Błędy pomiarów

- Każdy pomiar wielkości fizycznej obarczony jest pewnym błędem.
- Błąd systematyczny polega na systematycznym odchyleniu wyniku pomiaru względem rzeczywistej wartości wielkości mierzonej. Czasami daje się uwzględnić istnienie tego błędu i otrzymany wynik skorygować numerycznie po pomiarze. Zawsze należy go oszacować.
- Błąd przypadkowy (statystyczny) jest miarą rozrzutu otrzymywanych wyników wokół wartości najbardziej prawdopodobnej. Błąd taki wynika albo z metody wykonywania pomiaru albo z samej natury zjawiska (np. liczba impulsów zliczanych przez detektor promieniowania).

Analiza błędów powie nam, czy i w jakim stopniu obserwowany efekt jest znaczący (np. niedawne odkrycie bozonu Higgsa), jaki jest błąd pomiaru, jakie jest prawdopodobieństwo, że dana wielkość naprawdę zawarta jest w danym przedziale.

Czy statystyka to tylko analiza błędów?

- Opis i parametry zjawisk o charakterze losowym.
 - średnia płaca, rozkłady płacy w funkcji różnych zmiennych itp.
- Związki i korelacje pomiędzy kilkoma zjawiskami losowymi.
 - czy istnieje związek pomiędzy wzrostem pracownika a wynagrodzeniem?
- Szacowanie parametrów populacji na podstawie losowo wybranej próbki.
 - procent poparcia dla partii,
- Testowanie hipotez statystycznych.
 - czy średni wzrost Polaków i Rosjan jest taki sam?
- Prognozowanie.

zmniejszenie zużycia węgla o X powinno spowodować wzrost zużycia ropy o Y

Jednowymiarowa zmienna losowa

- Dla każdych dwóch liczb a i b takich, że $a < b$ istnieje określone prawdopodobieństwo, że zmienna X przybierze wartość z przedziału (a, b)
 - Wzrost ludzi jest zmienną losową. Dla dowolnych dwóch liczb np. 160 cm i 170 cm możemy określić prawdopodobieństwo, że przypadkowo wybrana osoba będzie mieć wzrost pomiędzy 160 cm a 170 cm.
- Funkcją rozkładu (dystribuantą) zmiennej losowej X_0 nazywamy taką funkcję $F(X_0)$, że:
$$F(X_0) = P(X < X_0) \quad -\infty < X_0 < \infty$$

Dla dowolnej wartości X_0 funkcja podaje prawdopodobieństwo, że $X < X_0$.

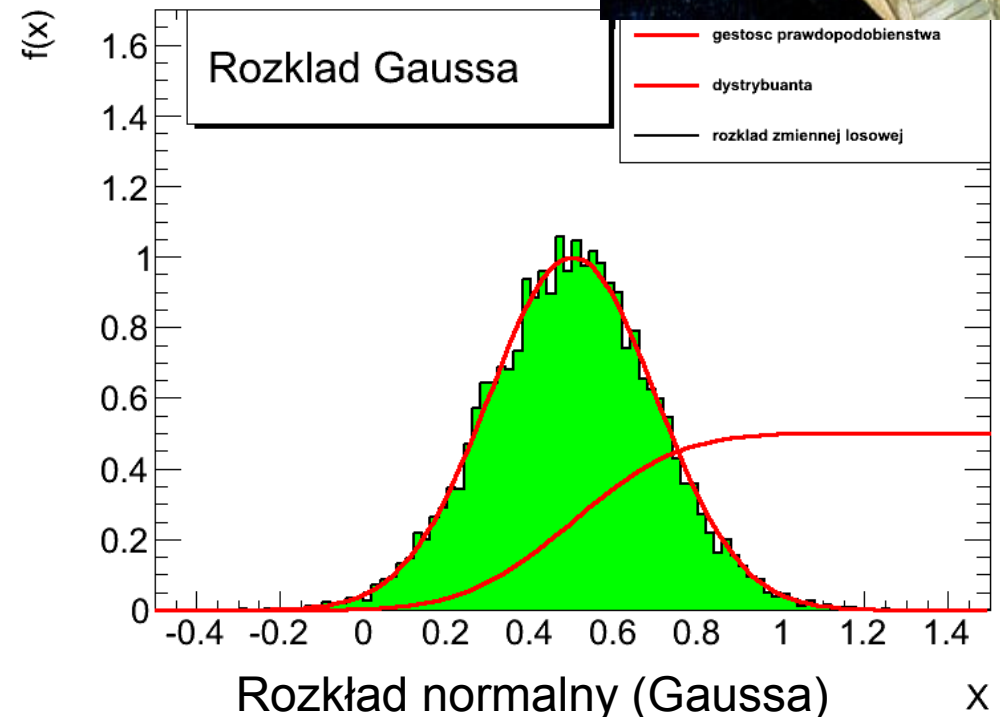
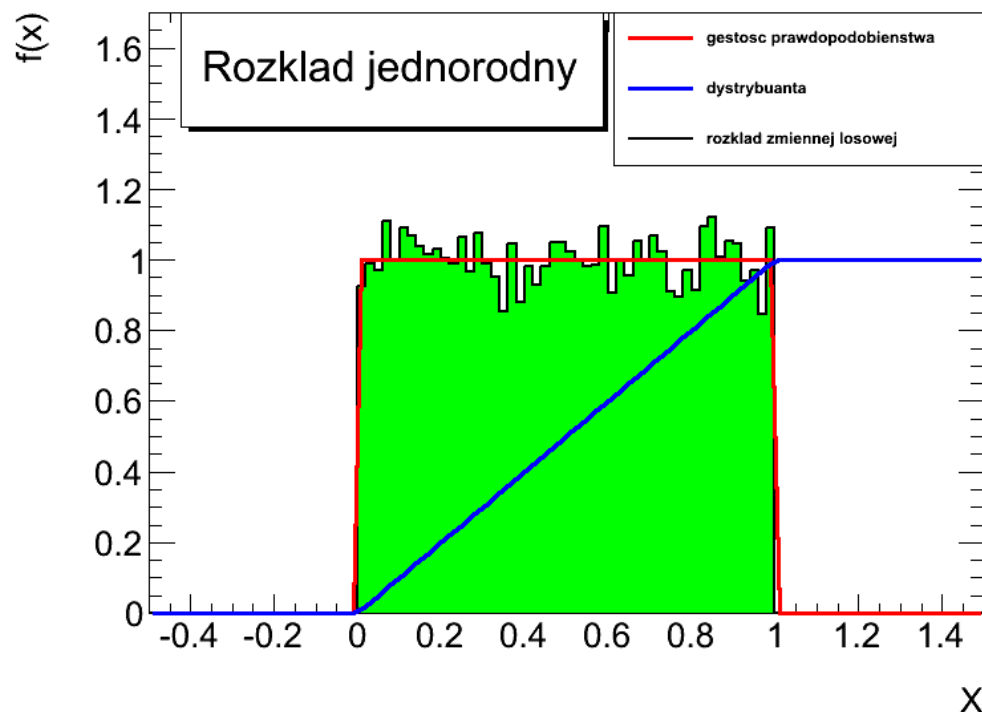
- Gęstością prawdopodobieństwa ciągłej zmiennej losowej X , nazywamy funkcję $f(X)$:

$$f(X) = \frac{dF(X)}{dX}$$

Rozkłady prawdopodobieństwa



Johann Karl Friedrich Gauss (1777 – 1855)



$$\phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Opis rozkładu

- Momentem zwykłym rzędu k zmiennej losowej X nazywamy:

- dla zmiennej losowej dyskretnej
$$\alpha_k = \sum_{i=-\infty}^{+\infty} x_i^k p_i$$

- dla zmiennej losowej ciągłej
$$\alpha_k = \int_{-\infty}^{+\infty} x^k f(x) dx$$

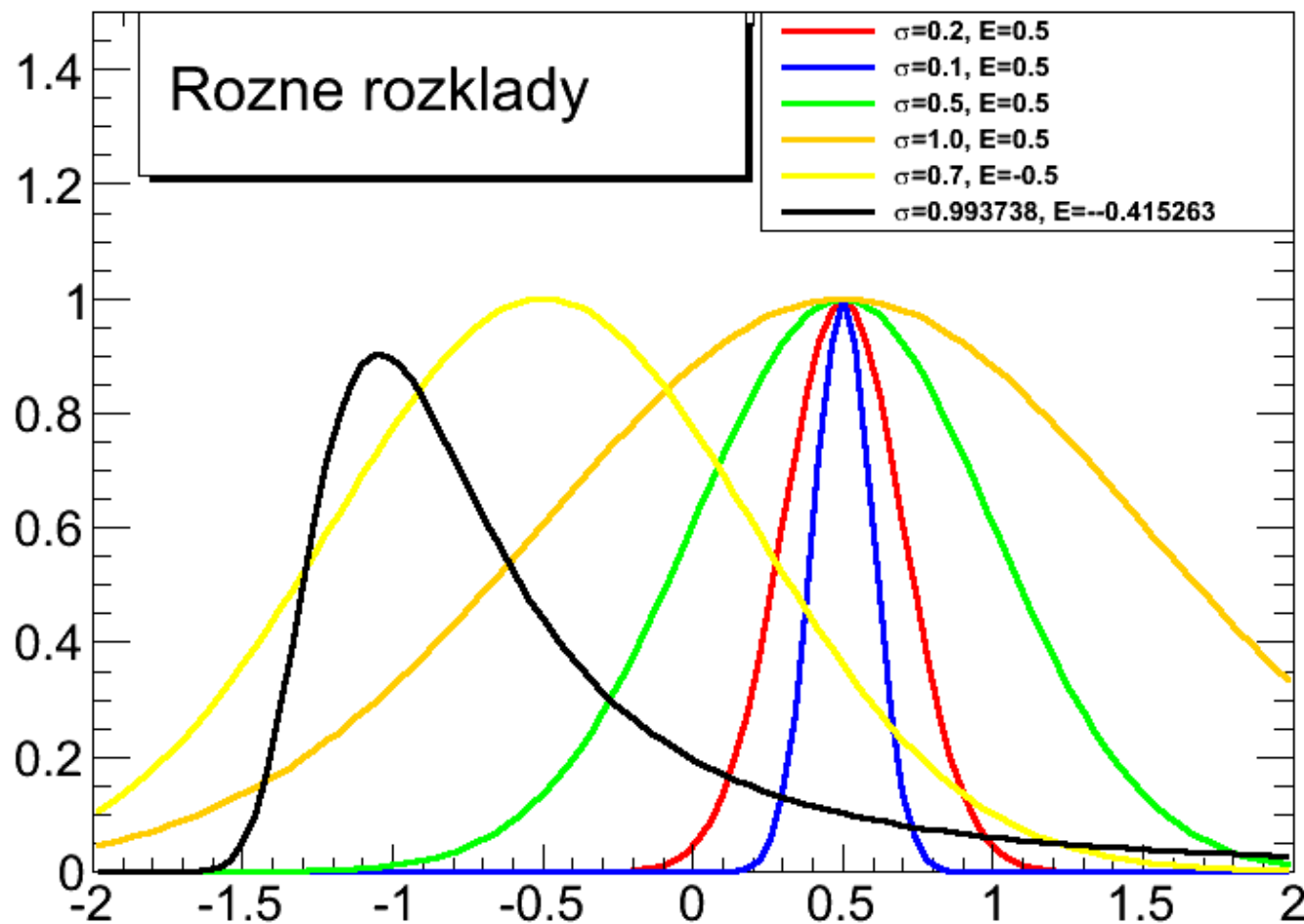
- Momentem centralnym rzędu k zmiennej losowej X nazywamy:

- dla zmiennej losowej dyskretnej
$$\mu_k = \sum_{i=-\infty}^{+\infty} (x_i - \alpha_1)^k p_i$$

- dla zmiennej losowej ciągłej
$$\mu_k = \int_{-\infty}^{+\infty} (x - \alpha_1)^k f(x) dx$$

- **Średnia: α_1 , wariancja: μ_2 . Nie są to zmienne losowe, mają konkretną, dokładnie określoną wartość liczbową!** Zmiennymi losowymi mogą być ich estymatory ustalone na podstawie losowej próby.
- Odchyleniem standardowym zmiennej losowej X nazywamy pierwiastek kwadratowy z wariancji: $\sigma = \sqrt{\mu_2}$. Jest miarą rozrzutu zmiennej losowej.

Średnia i odchylenie standardowe



Estymatory wielkości

- **Pytanie: jaki jest średni wzrost dorosłych Polaków?**
- Bierzemy wszystkich dorosłych Polaków, mierzymy wzrost każdego z nich, liczymy średnią. Postępowanie tyleż proste co bezużyteczne.
- Ale: **TYLKO TA METODA POZWOLI ZNALEŹĆ RZECZYWISTY ŚREDNI WZROST POLAKÓW**, każda inna jest oszacowaniem na podstawie wylosowanej próby, a zatem obarczonym błędem. **Nie znamy przecież gęstości prawdopodobieństwa w funkcji wzrostu: $f(x)$.**
- Losujemy przypadkową próbę 100 Polaków, używamy *estymatora* średniej:

$$\bar{x} = 1/n \sum_i x_i$$

UWAGA: próba powinna być nieobciążona, nie wybierać do pomiaru drużyny koszykówki ani zespołu dżokejów!

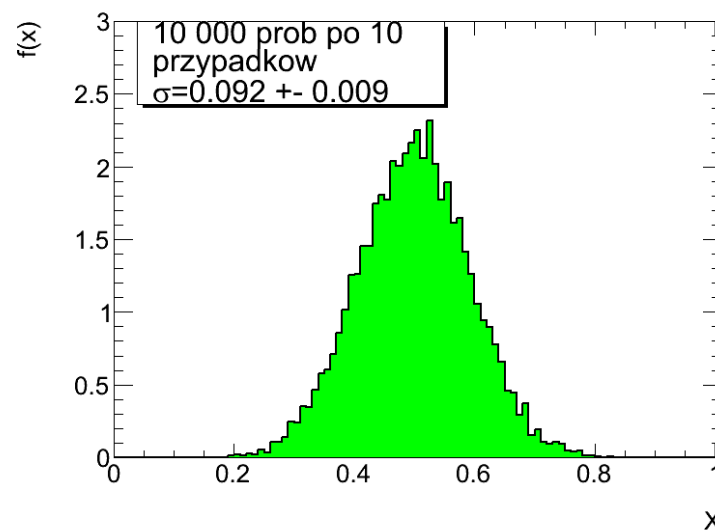
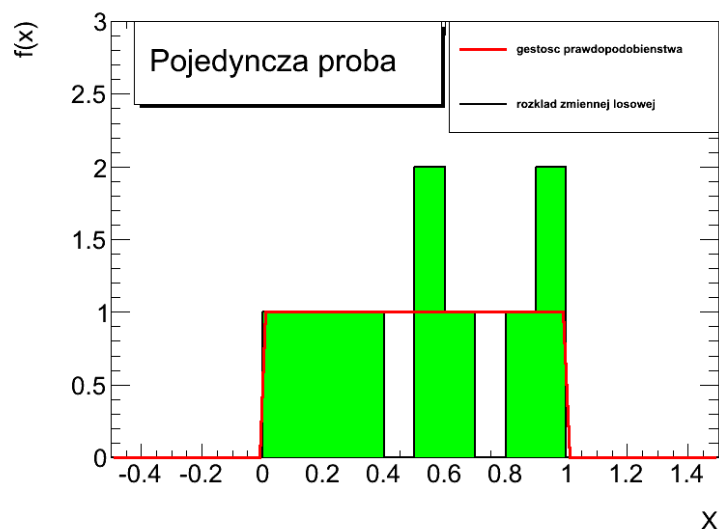
- **Staramy się uzyskać informacje o nieznanym rozkładzie prawdopodobieństwa na podstawie skończonej próby pomiarów.**

Estymatory – parę definicji

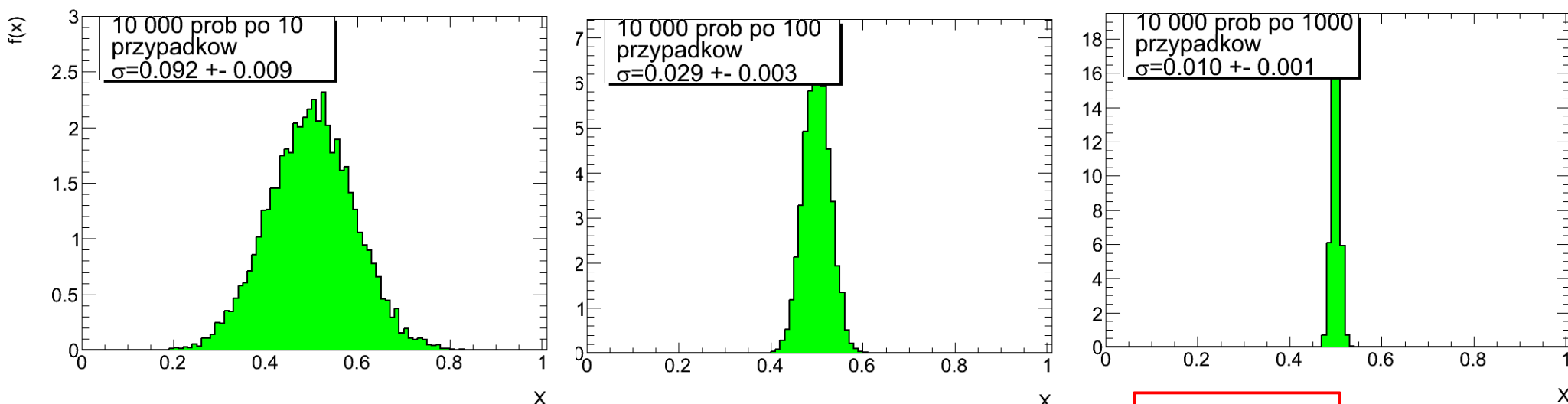
- **Estymator zgodny** - przy dużych próbach losowych prawdopodobieństwo, że wartość estymatora będzie się znacząco różnić od estymowanego parametru jest zerowe.
- **Estymator konsystentny** - przy dużych próbach rozkład wartości estymatora staje się bardzo wąski, a przy n dążącym do nieskończoności rozkład staje się nieskończenie wąską "szpilką".
- **Estymator nieobciążony** - wartość oczekiwana uzyskana z rozkładu wartości estymatora zawsze (dla dowolnego n) pokrywa się z parametrem estymowanym.

Estymator średniej

- Estymator średniej $\bar{x} = 1/n \sum x_i$ jest nieobciążony, czyli jego wartość oczekiwana jest zgodna z prawdziwą wartością średniej.
- Dla dużej próby otrzymamy estymowaną wartość średniej równą prawdziwej średniej. Jaki jest rozkład prawdopodobieństwa estymowanej średniej dla niewielkiej próby?
- Zróbmy „eksperyment” losując wielokrotnie 10 przypadków z rozkładu jednorodnego i licząc dla średnią dla każdej próby.



Estymator średniej



- Błąd estymacji średniej maleje ze wzrostem statystyki
- Zauważmy, że rozkład średniej wygląda jak rozkład normalny (Gausa), pomimo że wyjściowy rozkład był rozkładem jednorodnym.
- Okazuje się, że tak będzie dla każdego rozkładu wyjściowego – *centralne twierdzenie graniczne*.

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

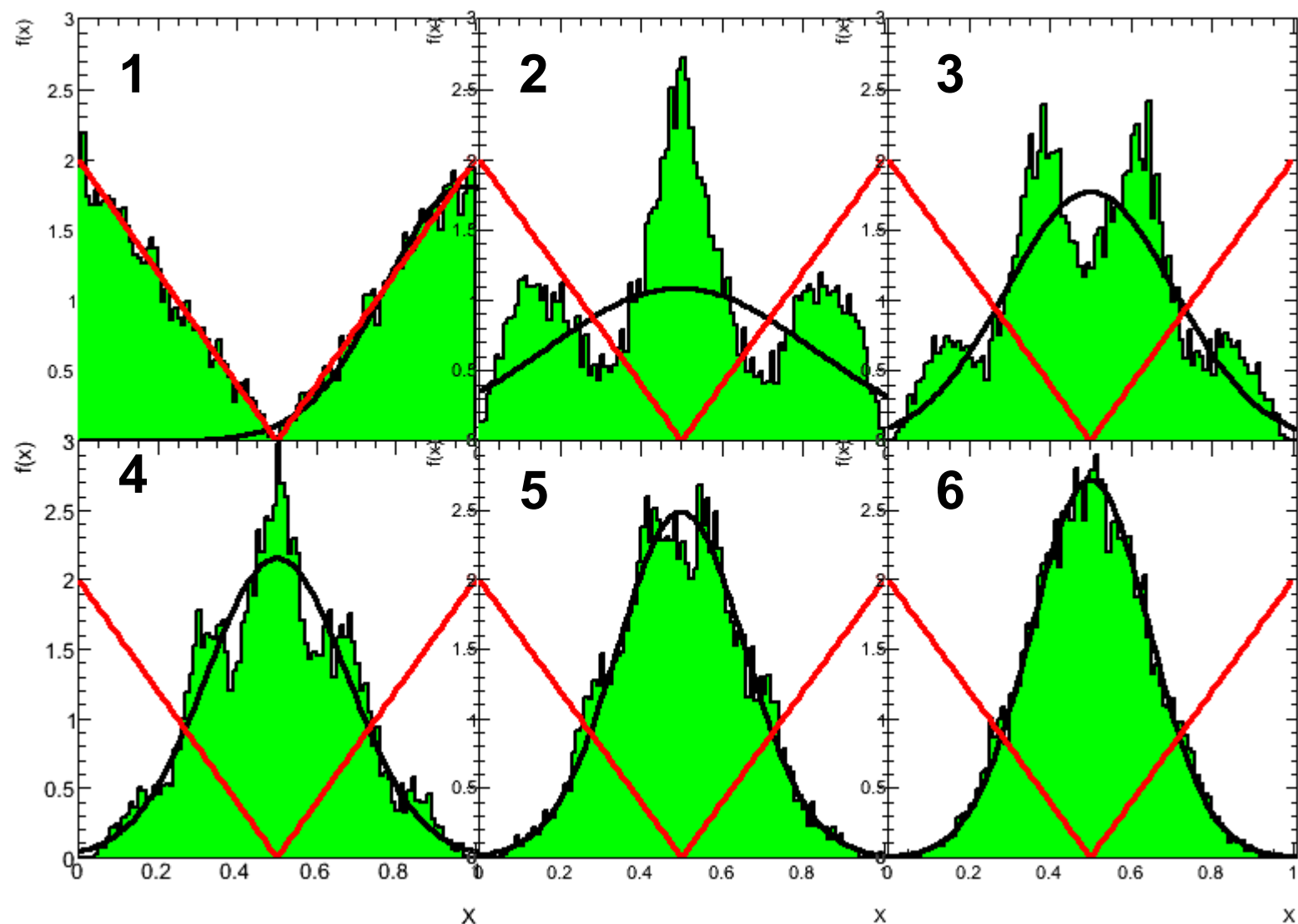
Centralne twierdzenie graniczne

- Jedno z najważniejszych twierdzeń rachunku prawdopodobieństwa, uzasadniające powszechne występowanie w przyrodzie rozkładów zbliżonych do rozkładu normalnego.
- Jeśli X_i są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, takiej samej wartości oczekiwanej (średniej) \bar{x} i skończonej wariancji σ^2 większej od zera to zmienna losowa o postaci:

$$Y = \frac{\sum_i (X_i - \bar{x})}{\sigma \sqrt{n}}$$

zbiega się do rozkładu normalnego, gdy n rośnie do nieskończoności.

Centralne twierdzenie graniczne



Estymator wariancji

- Podobnie na podstawie skończonej próby możemy estymować wariancję rozkładu:

$$\bar{\mu}_2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

Estymator ten jest nieobciążony i zgodny.

- Estymator odchylenia standardowego:

$$\bar{\sigma} = \sqrt{\bar{\mu}_2} = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$$

jest zgodny, ale obciążony.

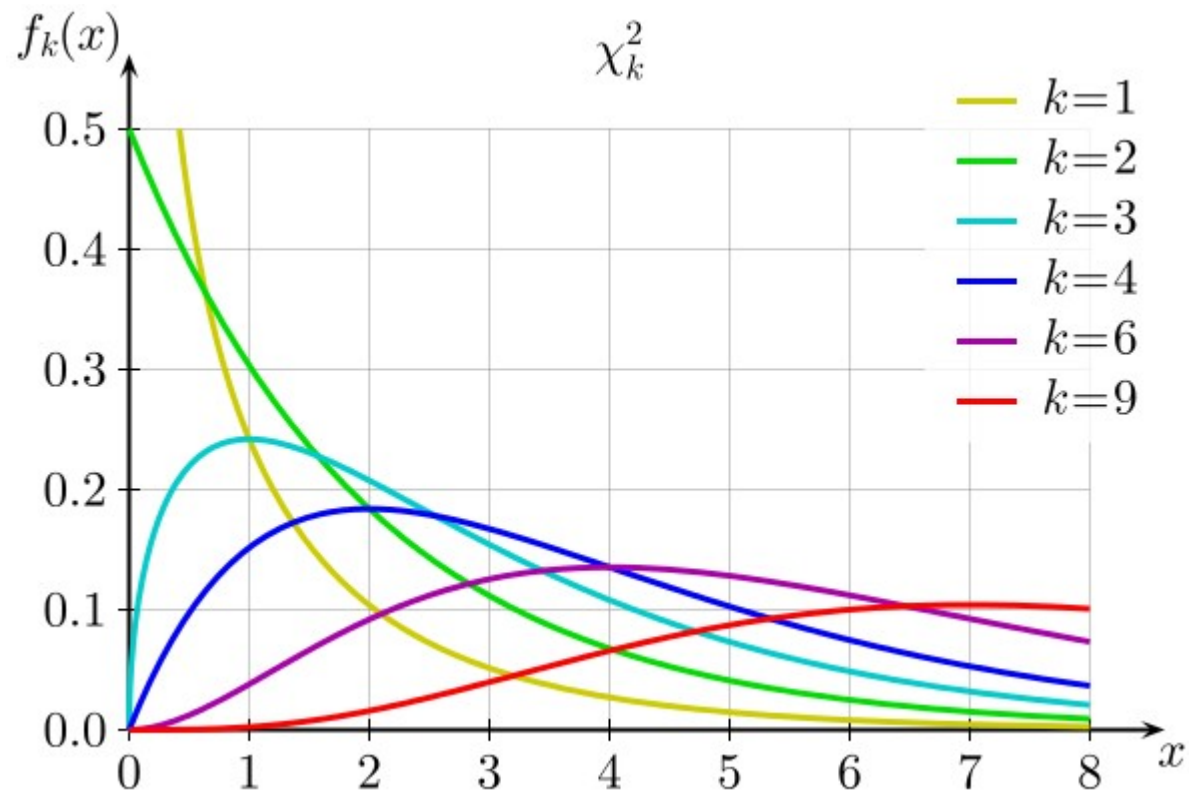
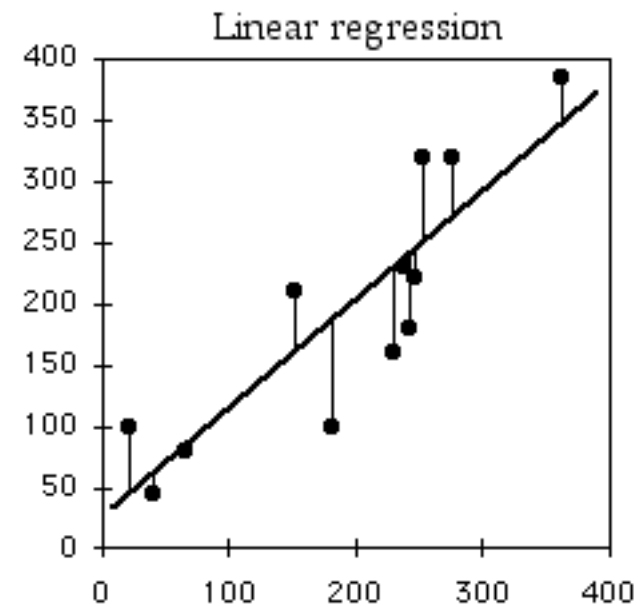
Test χ^2

- Test χ^2 mówi nam, jak dobrze wielkości mierzone x_i pasują do naszej hipotezy, czyli wielkości oczekiwanych

$$\chi^2 = \sum_i \left(\frac{x_i - E_i}{\sigma_i} \right)^2$$

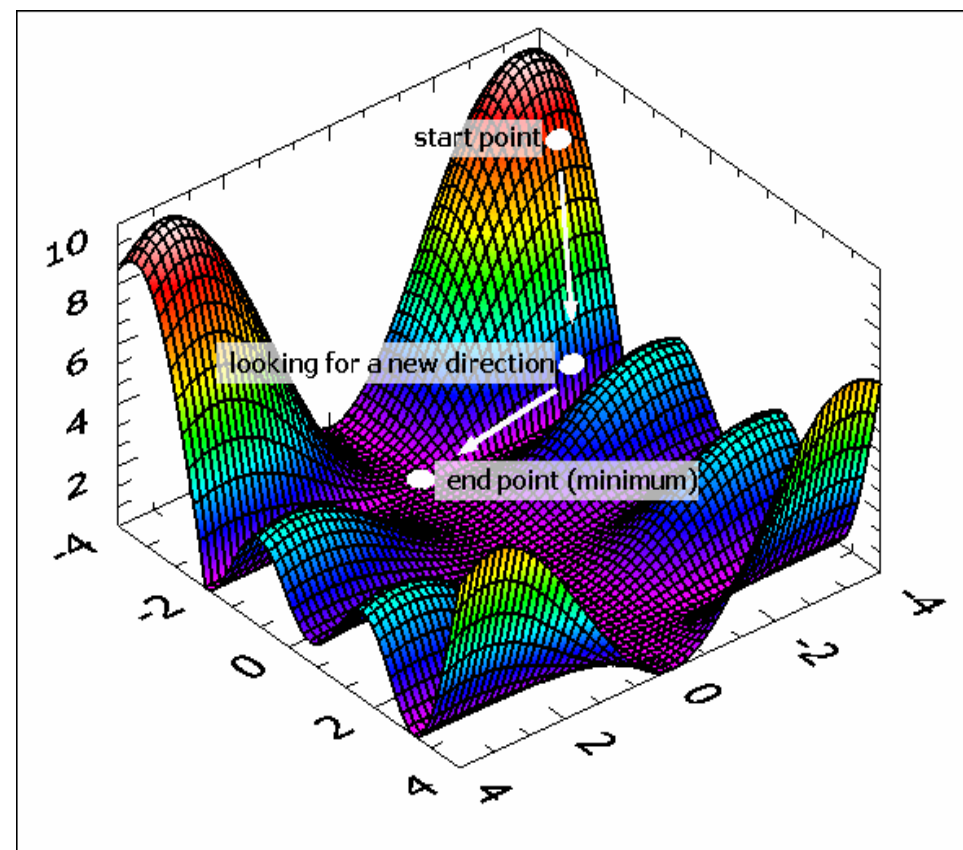
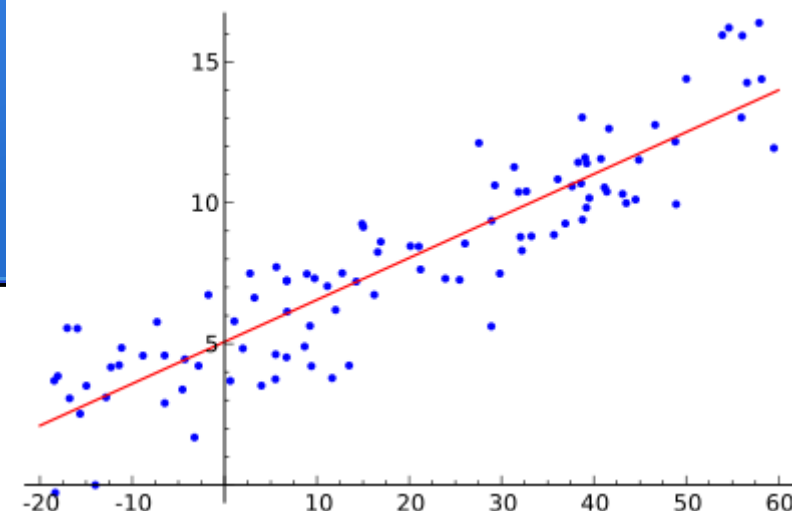
Duże wartości χ^2 świadczą, że hipoteza nie pasuje do danych, zbyt małe, że pasuje zbyt dobrze.

- Średnia wartość rozkładu χ^2 równa się liczbie stopni swobody k , a więc spodziewamy się $\chi^2/k=1$.

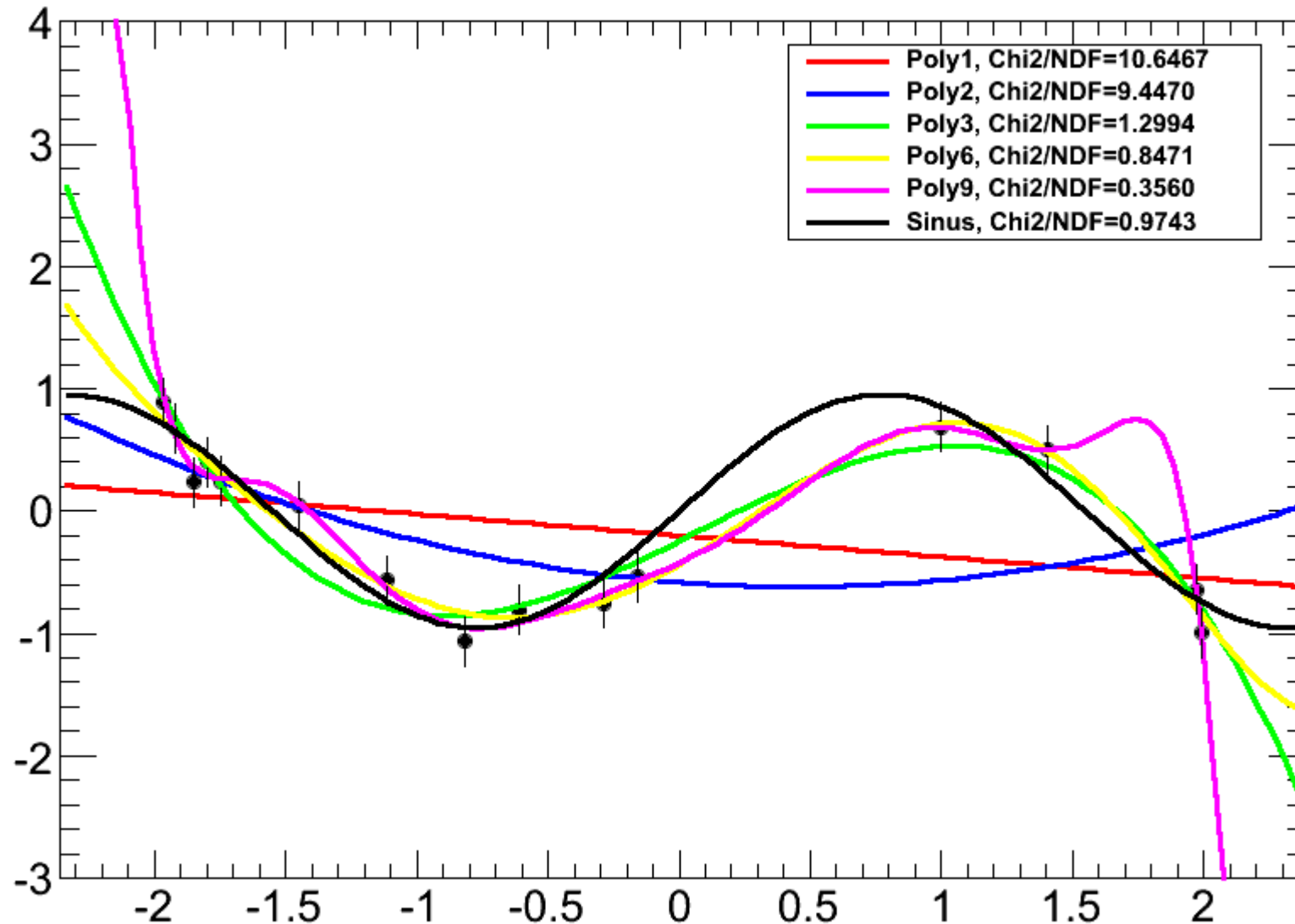


Dopasowanie funkcji

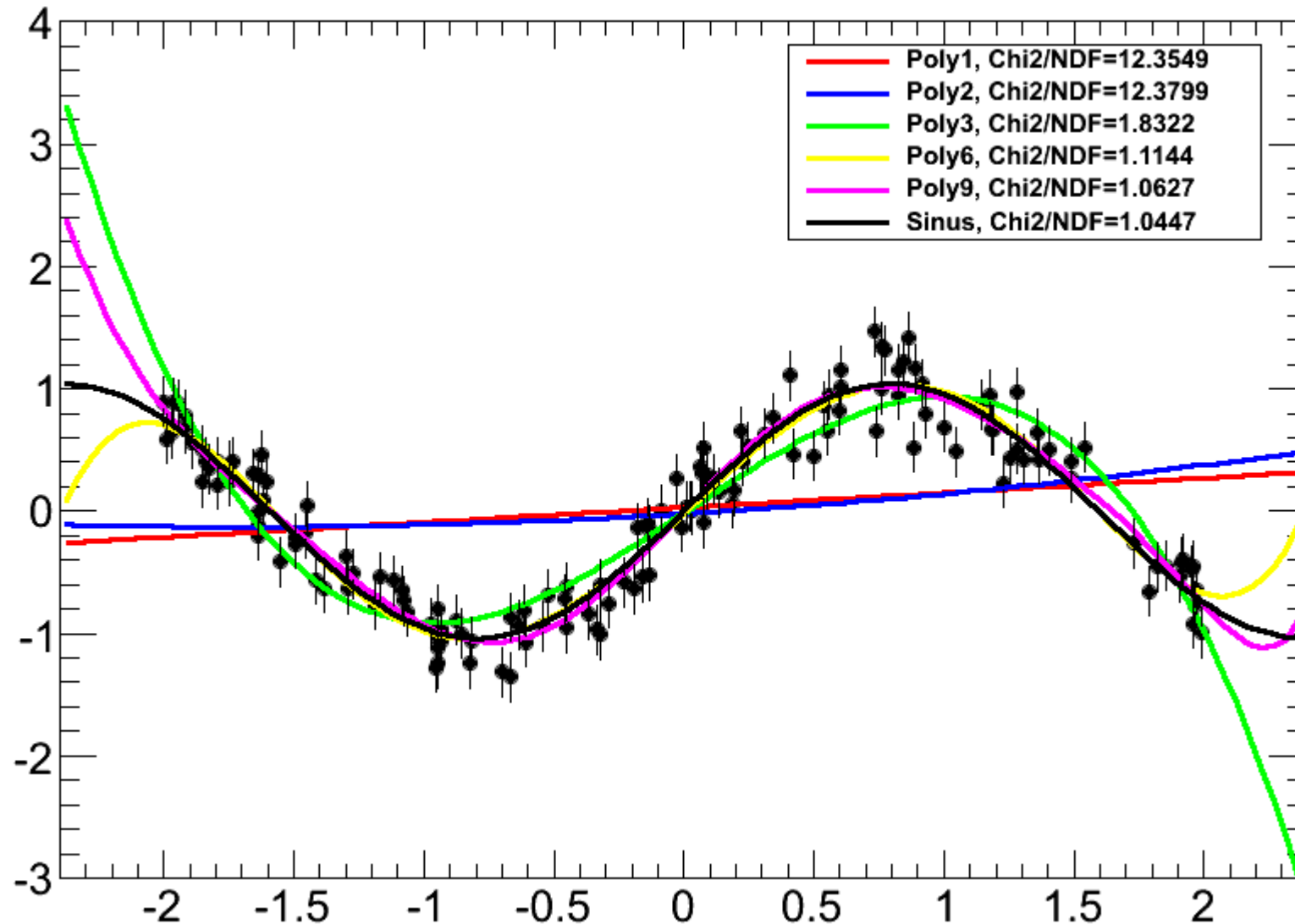
- Zadanie: do punktów doświadczalnych jak najlepiej dopasować funkcję zadanej klasy. Sprawdzić, czy funkcja opisuje dane doświadczalne.
- Najczęściej używana metoda: znaleźć takie parametry funkcji, aby zminimalizować wartość χ^2 – metoda najmniejszych kwadratów.
- W przypadku funkcji liniowej można to zrobić algebraicznie (przepis podał Gauss), przy bardziej skomplikowanych funkcjach numerycznie (np. minimalizacja metodą największego gradientu).



Dopasowanie funkcji



Dopasowanie funkcji



Przedział ufności

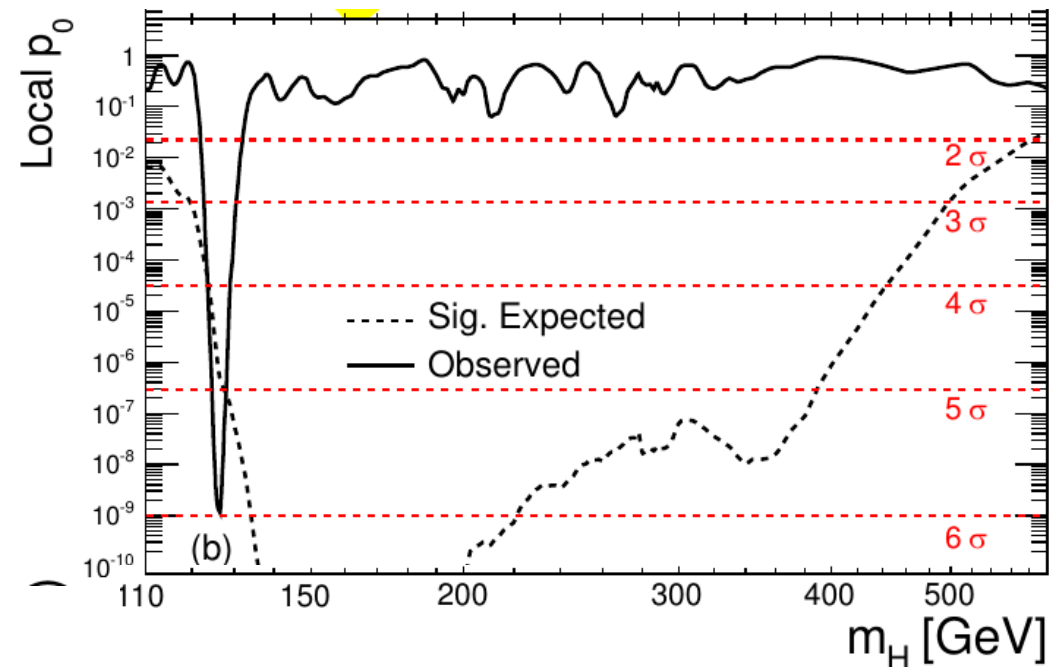
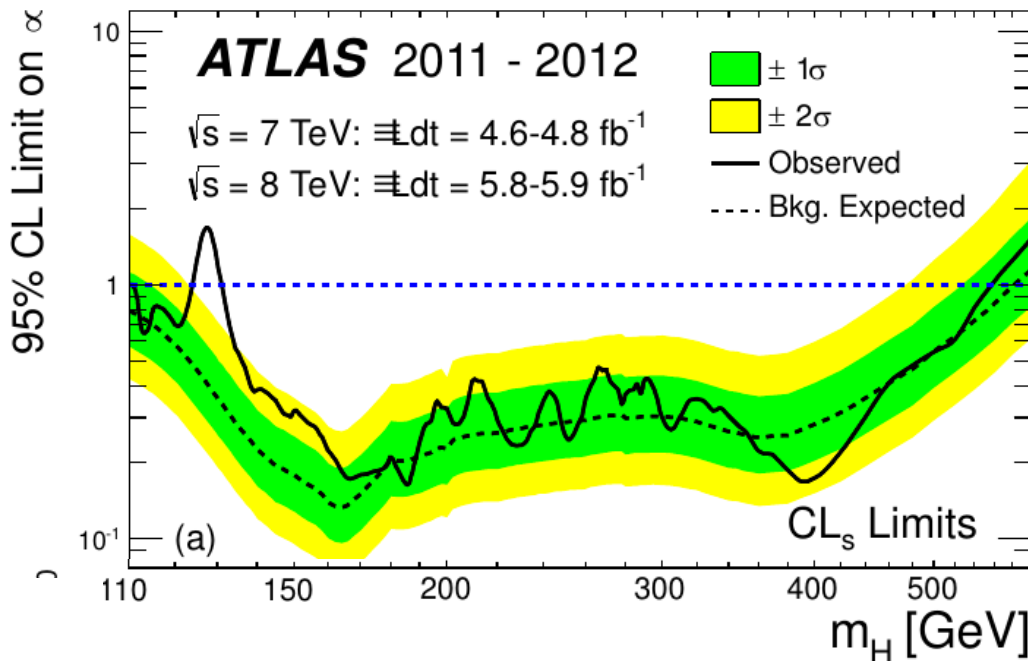
- Przedziałem ufności na poziomie ufności $1-\alpha$ dla nieznanego parametru Q nazywamy taki przedział (Q_1, Q_2) , dla którego prawdopodobieństwo, że przedział ten zawiera w sobie rzeczywistą wartość estymowanego parametru Q wynosi $1-\alpha$: $P(Q_1 < Q < Q_2) = 1 - \alpha$
- Poszukujemy takich dwóch wartości Q_1 i Q_2 , aby przedział przez nie wyznaczony z zadanyam prawdopodobieństwem zawierał w sobie rzeczywistą wartość parametru Q :

Q ma konkretną wartość, choć nieznaną. Wyznaczamy Q_1 i Q_2 dla każdej próby statystycznej. Chcemy, aby z prawdopodobieństwem $1-\alpha$ w tak wyznaczonym przedziale znajdował się nieznanany parametr Q .

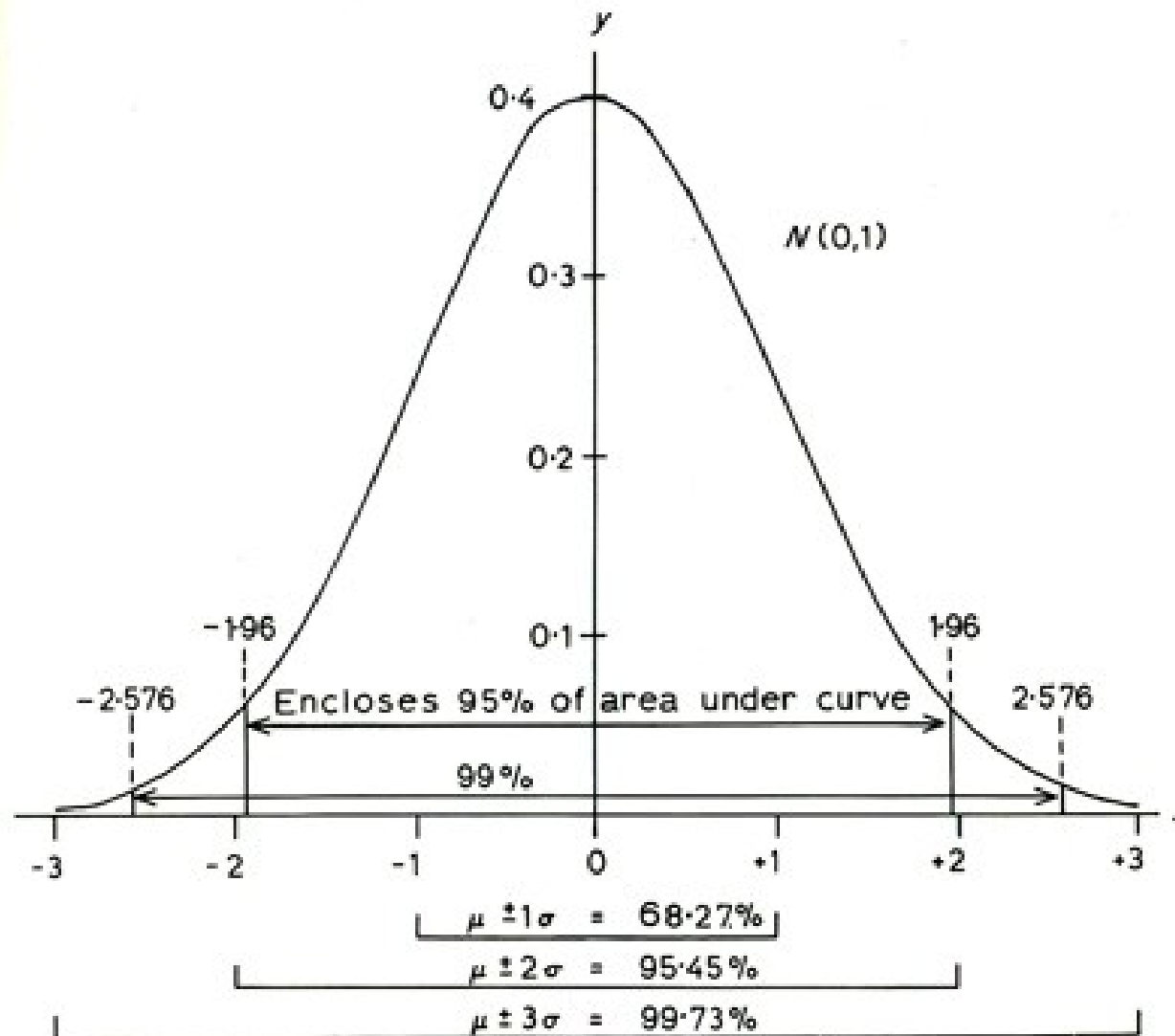
Odkrycie bozonu Higgsa

Clear evidence for the production of a neutral boson with a measured mass of **126.0 ± 0.4 (stat) ± 0.4 (sys) GeV** is presented. This observation, which has a significance of **5.9 standard deviations**, corresponding to a background fluctuation probability of **1.7×10^{-9}** , is compatible with the production and decay of the Standard Model Higgs boson.

„Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC”, ATLAS Collaboration



Rozkład normalny a przedziały ufności



Dwie zmienne - korelacje

- Zmienne x i y są niezależne, jeśli zachodzi związek:

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

czyli wartość jednej zmiennej nie zmienia prawdopodobieństwa wystąpienia danej wartości drugiej zmiennej.

- Kowariancją nazywamy:

$$\text{cov}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})(y - \bar{y}) f(x, y) dx dy$$

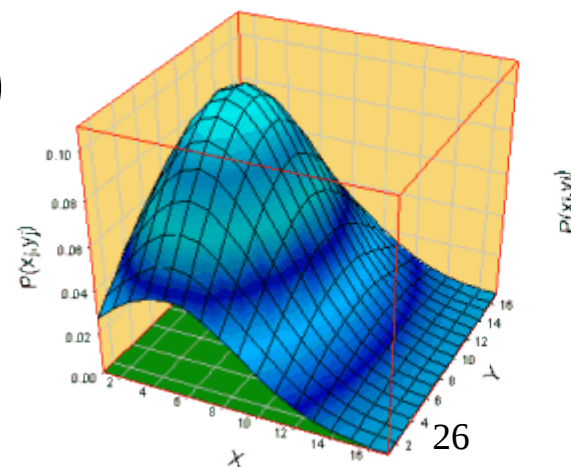
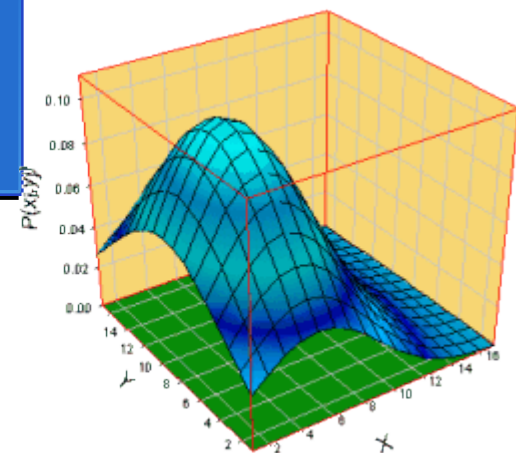
a współczynnikiem korelacji (liniowej):

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma(x) \sigma(y)}$$

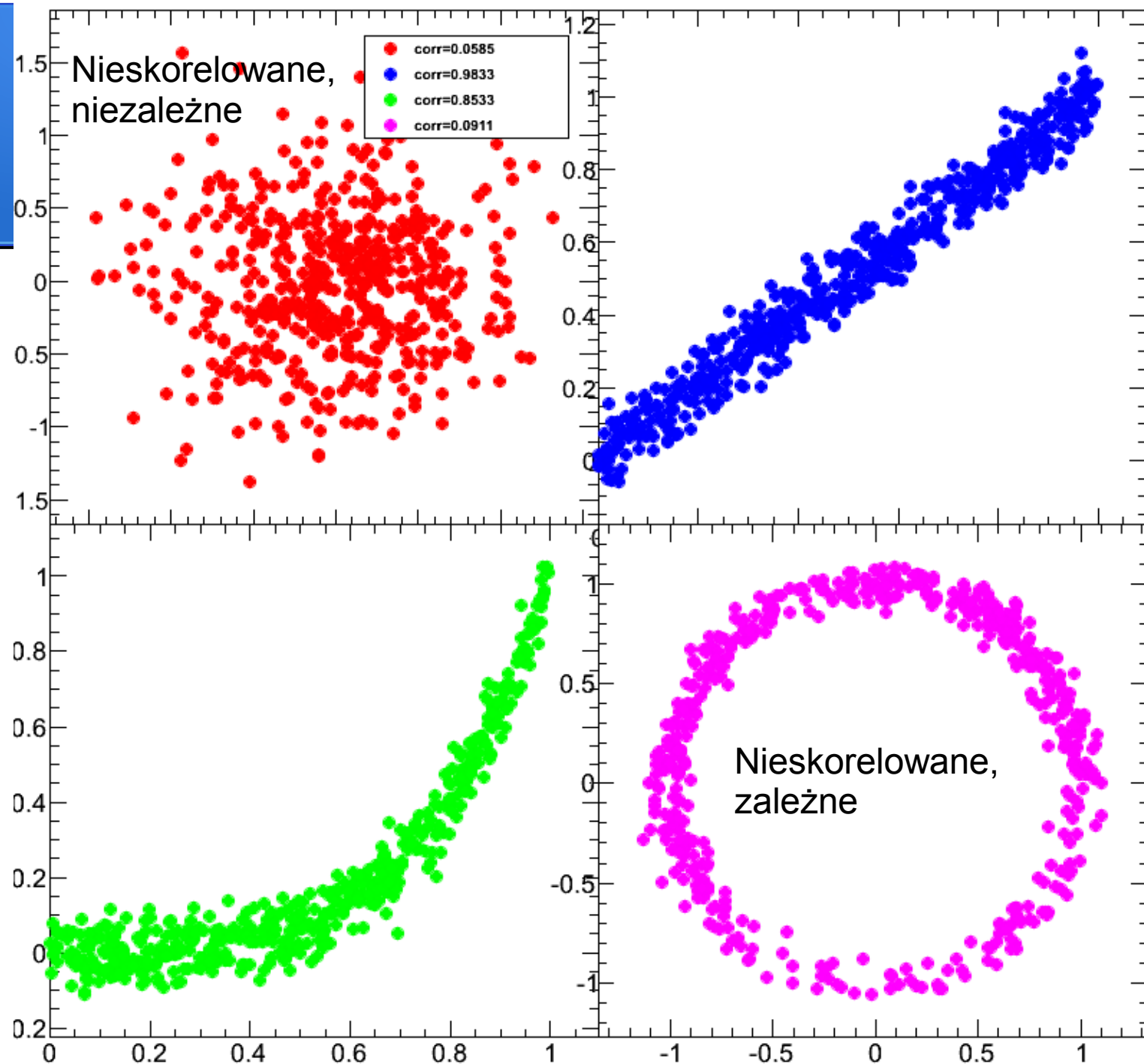
- Estymator kowariancji:

$$\text{cov}(x, y) = \frac{1}{N-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- UWAGA:** zmienne niezależne mają zerowy współczynnik korelacji, ale zmienne nieskorelowane mogą być zależne!

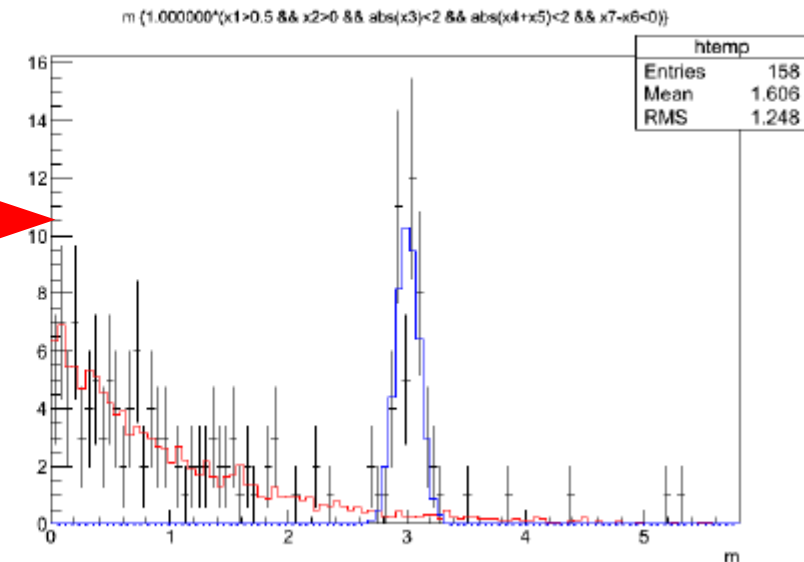
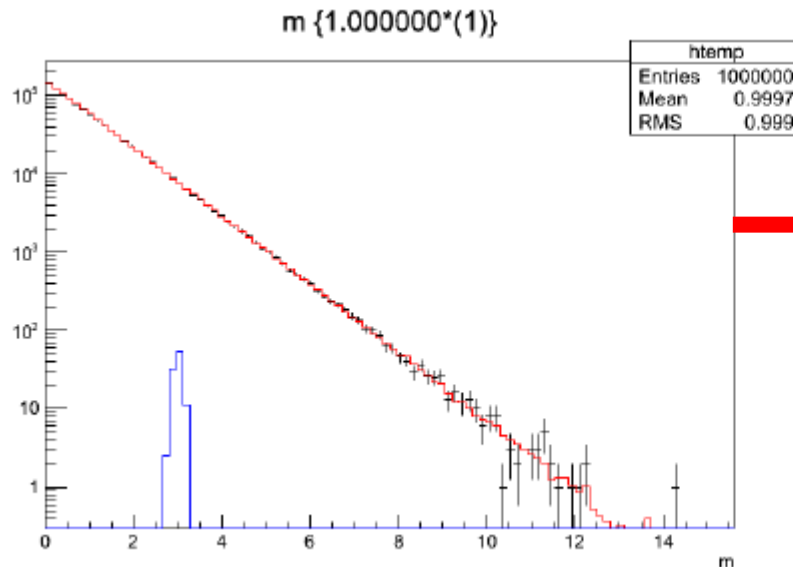
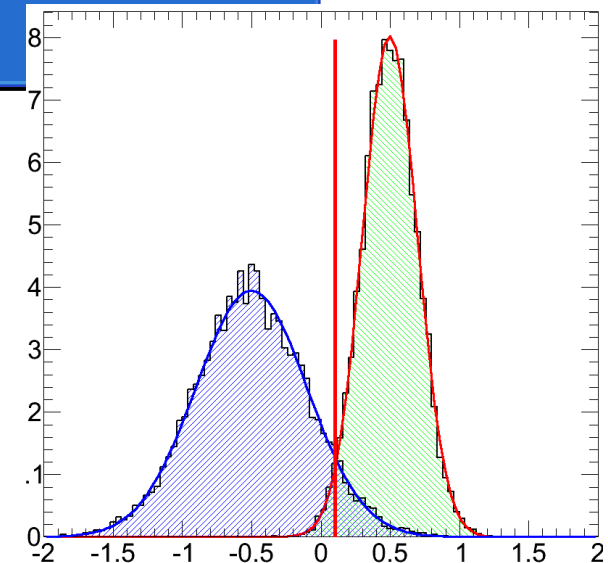


Korelacje



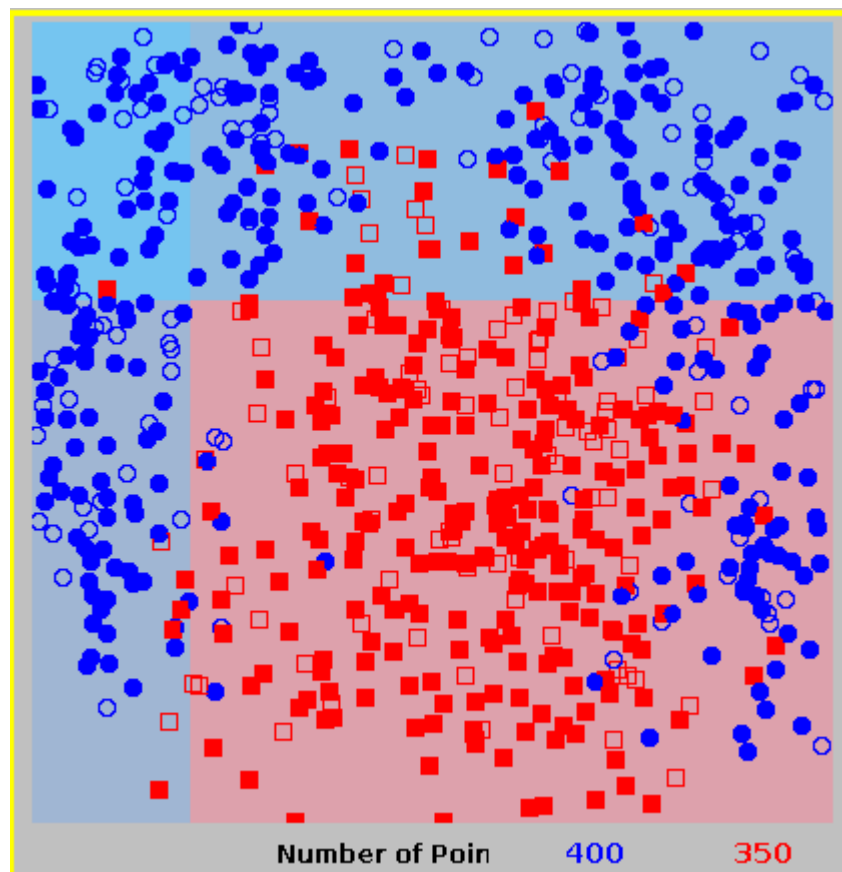
Analiza wielowymiarowa separacja sygnału i tła

- Wczorajsze ćwiczenie – za pomocą cięcia na zmiennych $x_1 \dots x_7$ można bardzo silnie zredukować tło i wydobyć z danych sygnał.
- Czy cięcia to optymalna metoda selekcji?

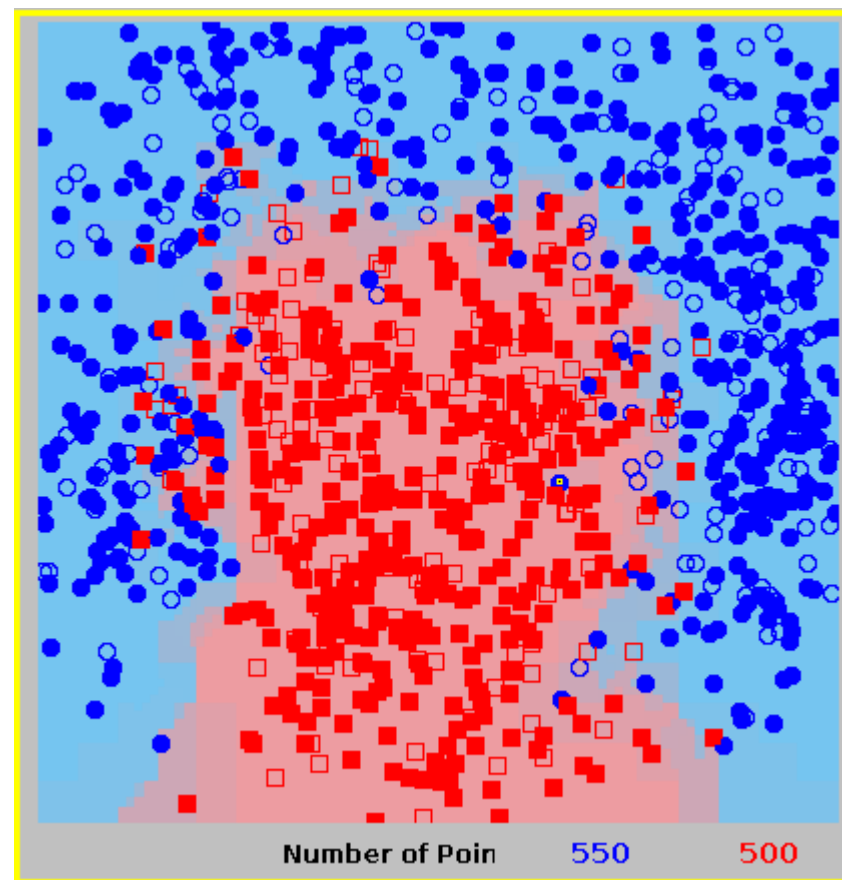


Cięcie vs separacja nieliniowa

Cięcia



Separacja nieliniowa

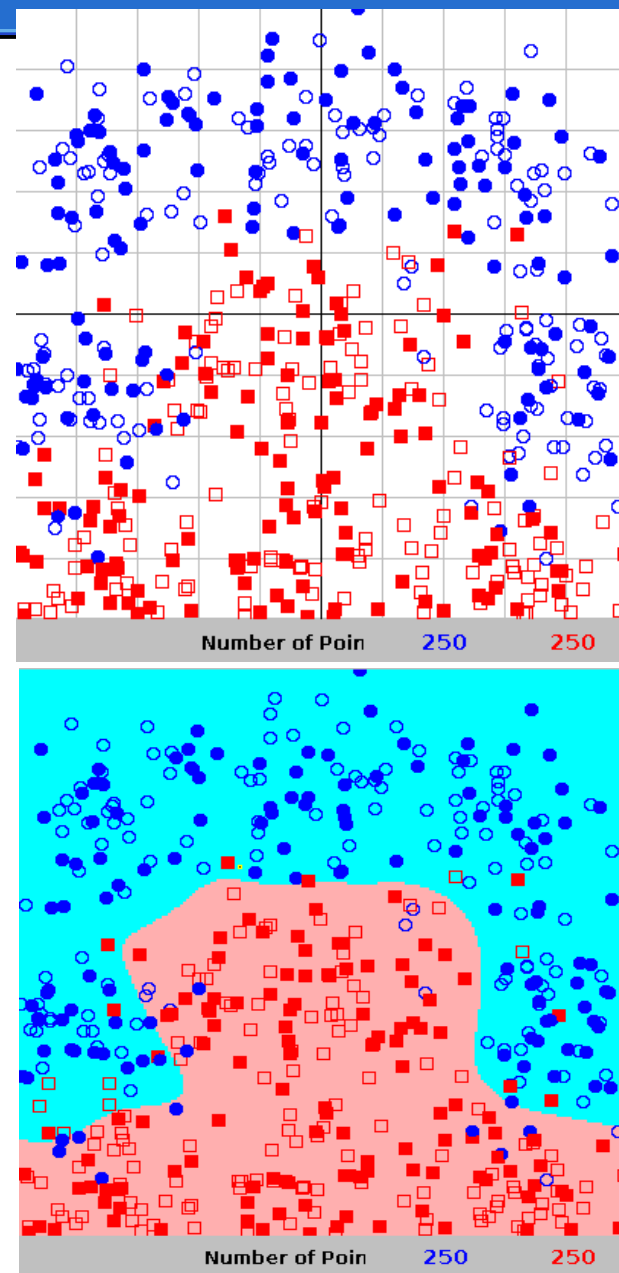


?

Np. sieci neuronowe, boosted
decision trees, itd

Jak działa algorytm uczący się?

Np. Sieć neuronowa, czy BDT.



- Potrzebne są **dane treningowe**, dla których znamy poprawną odpowiedź, czy jest to sygnał czy tło, dzielimy je na zbiór treningowy i testowy.
- Znajdujemy funkcję $f(\mathbf{x})$ najlepiej opisującą prawdopodobieństwo, że dany przypadek jest klasy „sygnał” - minimalizacja tzw. funkcji straty (np. χ^2).
- Poszczególne algorytmy różnią się: klasą funkcji $f(\mathbf{x})$ (np. liniowe, nieliniowe), funkcją straty, sposobem jej minimalizacji.
- Wszystkie aproksymują nieznaną bayesowską funkcję decyzyjną BDF na podstawie skończonego zbioru danych treningowych.

BDF -idealna funkcja klasyfikująca, określona przez nieznaną gęstość prawdopodobieństwa sygnału i tła.

Popularne metody

- **Metody liniowe**

- Cięcia
- Dyskryminanty Fishera

- **Metody nieliniowe**

- Naiwny klasyfikator bayesowski,
- Estymatory gęstości prawdopodobieństwa, (Probability Density Estimator),
- Metoda najbliższych sąsiadów,
- PDE_RS,
- Sieci neuronowe

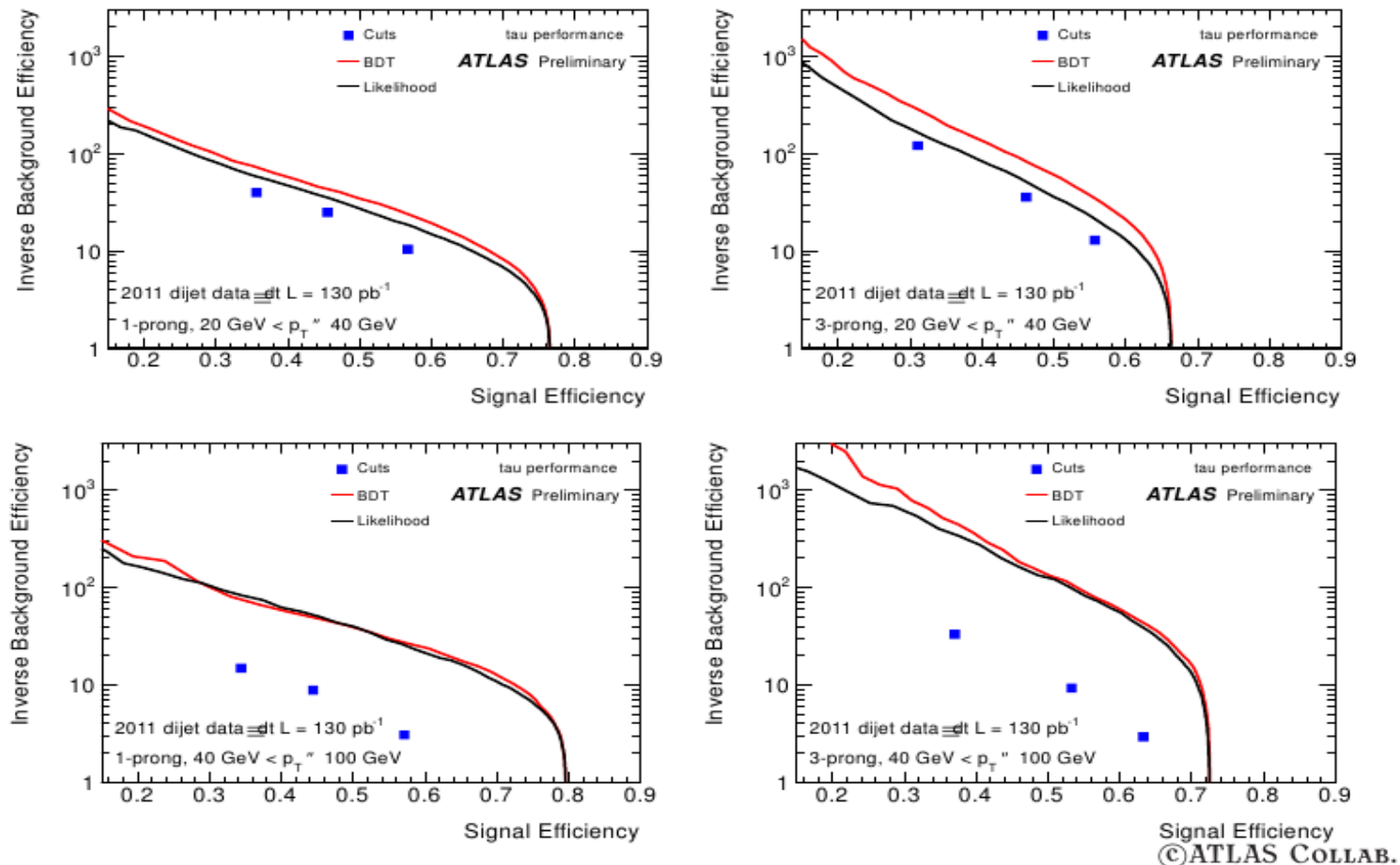
- Wzmocnione drzewa decyzyjne (Boosted Decision Trees),
- Maszyna wektorów wspierających (Support Vector Machine).

- **I wiele innych...**

- analiza składowych, niezależnych (Independent Component Analysis)
- analiza składowych głównych.

Wszystkie opierają się na tej samej zasadzie: poszukują najlepszej funkcji dyskryminującej w oparciu o próbkę treningową.

Identyfikacja hadronowych rozpadów leptonów tau w eksperymencie ATLAS



- Szereg zmiennych identyfikujących, żadna z nich pojedynczo nie daje dobrej identyfikacji.
- **Użycie metod wielu zmiennych zwiększa skuteczność identyfikacji.**

Applet pokazujący działanie różnych metod klasyfikujących

- <http://www.cs.technion.ac.il/~rani/LocBoost/>

Classification Applet

Generate 2-dimentional, 2-class classification problems and solve them using a number of classification schemes, as well as **LocBoost** - a new boosting algorithm

Point Editing

☐ Add single points

☒ Add: points, using Distribution

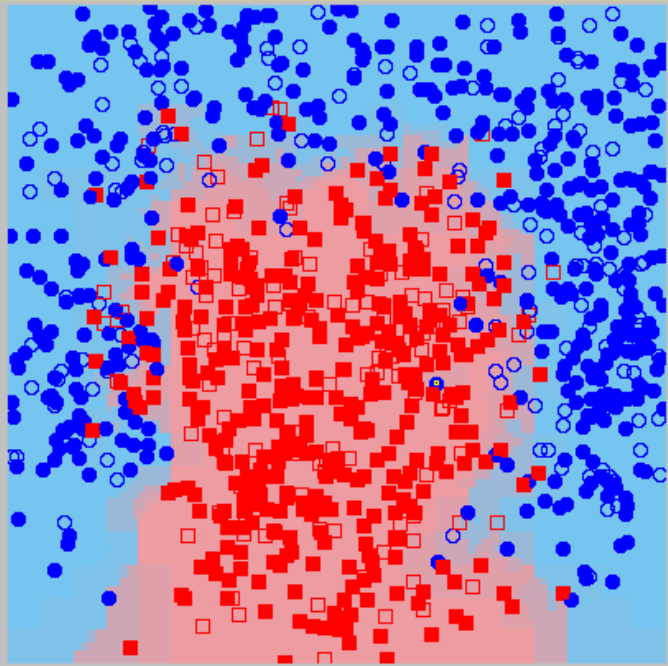
Sigma:

Use % of the data for training

Classification

Classifier:

Neighbors



Number of Poin 550 500

Test Err 0.099

Train Err 0.053

Decision Boundary smoothness (1-100)

☒ Use gradient colors

Podsumowanie (jeśli jeszcze wszyscy nie śpią)

**Trochę statystyki jest niezbędne, aby
prawidłowo zinterpretować dane pomiarowe**

Dodatkowe transparencje

Estymator wariancji

Sprawdźmy, dlaczego estymator wariancji, ma taką nie intuicyjną postać ($n-1$ w mianowniku).

Wartość średnią (oczekianą) zmiennej losowej liczymy jako:

$$\bar{x} = \frac{1}{n} \cdot (x_1 + \dots + x_n)$$

Estymatorem wartości średniej jest:

$$\bar{x} = E(\bar{x}) = \frac{1}{n} \cdot (E(x_1) + \dots + E(x_n)) = \frac{1}{n} \cdot (\bar{x}_1 + \dots + \bar{x}_n)$$

Załóżmy, dla potrzeb przykładu, że estymator wariancji - S_x^2 - zdefiniujemy poprzez analogię z estymatorem wartości średniej:

$$S_x^2 = \frac{1}{n} \left((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right) = \frac{1}{n} E \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)$$

Przypomnijmy. Estymator jest zmienną losową. Jeżeli policzymy wartość S_x^2 dla wielu różnych prób, to otrzymamy różne wartości. Zgodnie z definicją estymator jest nieobciążony, gdy wartość oczekiwana z estymatora jest równa wartości estymowanego parametru Θ . Sprawdźmy więc, czy tak jest w tym przypadku:

$$\begin{aligned} E(S_x^2) &= \frac{1}{n} E \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) = \frac{1}{n} E \left(\sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \bar{x})^2 \right) = \\ &= \frac{1}{n} \sum_{i=1}^n \left(E \left((x_i - \bar{x})^2 \right) - E \left((\bar{x} - \bar{x})^2 \right) \right) \end{aligned}$$

Zauważmy, że:

$$E \left((x_i - \bar{x})^2 \right) = \sigma_x^2$$

$$\begin{aligned} E \left((\bar{x} - \bar{x})^2 \right) &= \text{var}(\bar{x}) = \text{var} \left(\frac{1}{n} \sum_{i=1}^n x_i \right) = \\ &= \frac{1}{n^2} \cdot \text{var} \left(\sum_{i=1}^n x_i \right) \star \frac{1}{n^2} \cdot \sum_{i=1}^n \text{var}(X) = \frac{1}{n} \cdot \sigma^2 \end{aligned}$$

przejście oznaczone czerwoną gwiazdką, nie jest oczywiste i zostało wyjaśnione w przypisie

Tak więc ostatecznie możemy zapisać, że:

$$\begin{aligned} E(S_x^2) &= \frac{1}{n} \cdot \left[n \cdot \sigma^2(x) - n \cdot \left(\frac{1}{n} \cdot \sigma^2(x) \right) \right] \\ E(S_x^2) &= \frac{n-1}{n} \cdot \sigma^2(x) \end{aligned}$$

Jak widać, wartość oczekiwana naszego estymatora zależy od n . Nie może więc być równa wartości estymowanej, bo ta **na pewno nie zależy** od n . **Tak więc, obrany przez nas estymator, jest obciążony.** Z postaci wyrażenia na $E(S_x^2)$ możemy wywnioskować, jak powinien wyglądać nieobciążony estymator wariancji. Po prostu powinien on być dzielony przez $(n-1)$ a nie n !

$$S^2 = \tilde{\sigma}^2 = \frac{1}{n-1} \cdot \left((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right)$$

Ot i cała tajemnica dziwnej postaci estymatora wariancji.

* Przypis

$$\text{var} \left(\sum x_i \right) = \text{var}(x_1 + \dots + x_n) = \text{var}(x_1) + \dots + \text{var}(x_n) =$$

Każda zmienna x_i podlega temu samemu rozkładowi. Jest to rozkład zmiennej losowej X . A więc wariancja każdej ze zmiennych x_i jest równa wariancji zmiennej losowej X :

$$= \underbrace{\text{var}(X) + \dots + \text{var}(X)}_n = \sum \text{var}(X)$$