

Komputerowa analiza danych doświadczalnych

Wykład 8
15.04.2016

dr inż. Łukasz Graczykowski
lgraczyk@if.pw.edu.pl

Semestr letni 2015/2016



Pobieranie próby, estymatory histogram

Próby z rozkładów cząstkowych

Próby ze skończonej populacji

Próby z rozkładu normalnego

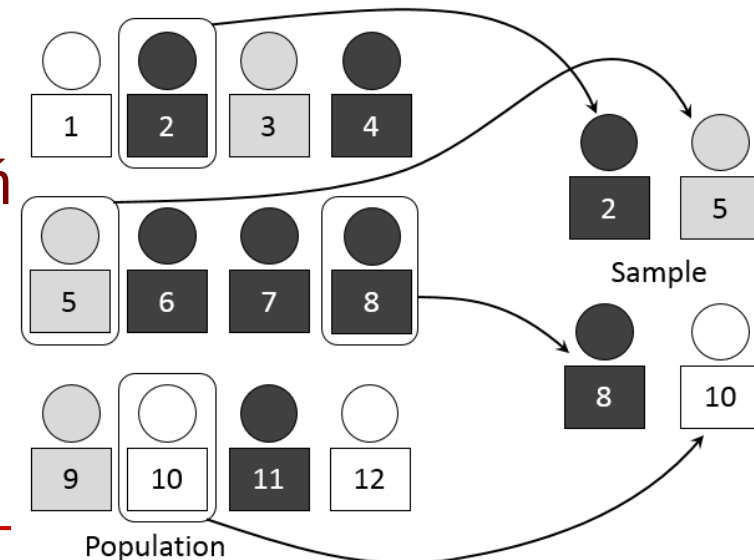
Rozkład χ^2



Pobieranie próby, estymatory

Pobieranie próby

- W przypadku pomiarów eksperymentalnych najczęściej nie znamy rozkładu prawdopodobieństwa opisującego dany pomiar (np. parametru rozkładu Poissona w rozpadach promieniotwórczych, czy parametrów rozkładu Gaussa opisującego jakąś populację)
- Te parametry chcemy wyznaczyć doświadczalnie, nie jesteśmy jednak w stanie zebrać nieskończenie wiele pomiarów
- W konsekwencji jesteśmy zmuszeni **przybliżyć rozkład gęstości za pomocą rozkładu częstości** (histogramu o skończonej liczbie wejść)
- **Próba** (*ang. sample*) nazywamy zespół doświadczeń wykonywanych w celu określenia kształtu (parametrów) poszukiwanego rozkładu:
 - próba otrzymywana jest poprzez wybór elementów z (często nieskończonego) zbioru wszystkich możliwych doświadczeń (wszystkich możliwych pomiarów), zwanego **populacją generalną**
 - próbę o n składnikach nazywamy próbą n -wymiarową



Estymatory - podsumowanie

- Podsumowując zatem **estymatory nieobciążone**:

- wartości oczekiwanej populacji → średnia z próby (**wynik doświadczenia**):

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

- wariancji populacji – wariancja z próby (aproksymowana):

$$S^2(X) = \frac{1}{n-1} ((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)$$

- wariancji wartości średniej z próby (**patrz niepewność typu A**):

$$S^2(\bar{X}) = \frac{1}{n} S^2(X) = \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$$

- wariancji (aproksymowanej) wariancji z próby

$$\sigma^2(S^2) = S^4 \left(\frac{2}{n-1} \right)$$

- odchylenia standardowego próby:

$$S = \sqrt{S^2(X)} = \frac{1}{\sqrt{n-1}} \sqrt{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}$$

- dalej możemy wyznaczać np. wariancję odchylenia std. próby...

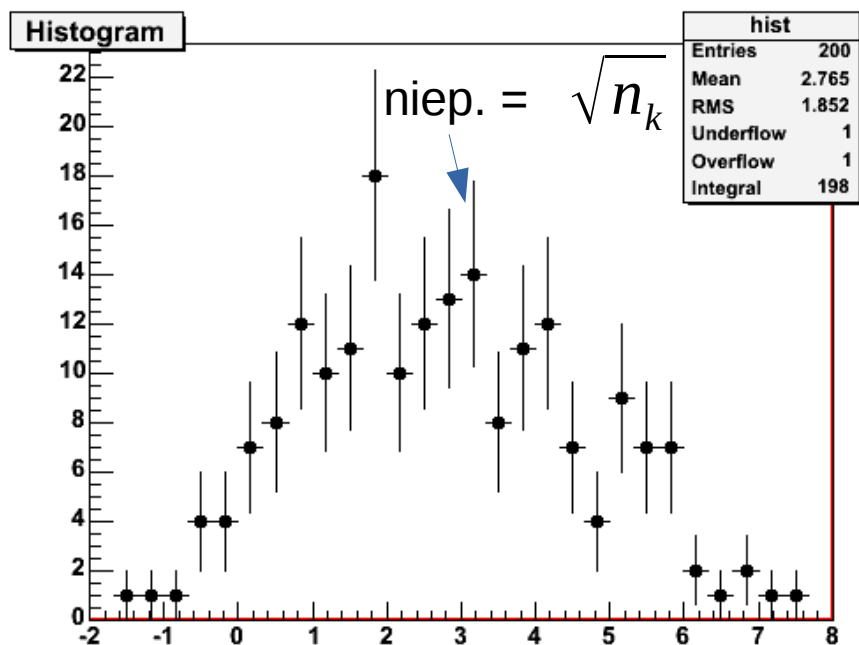
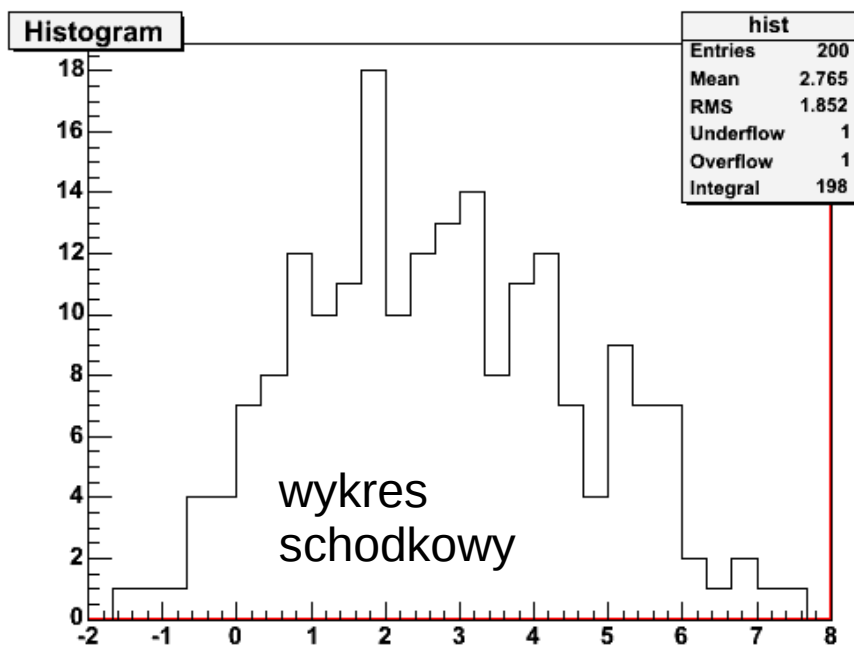
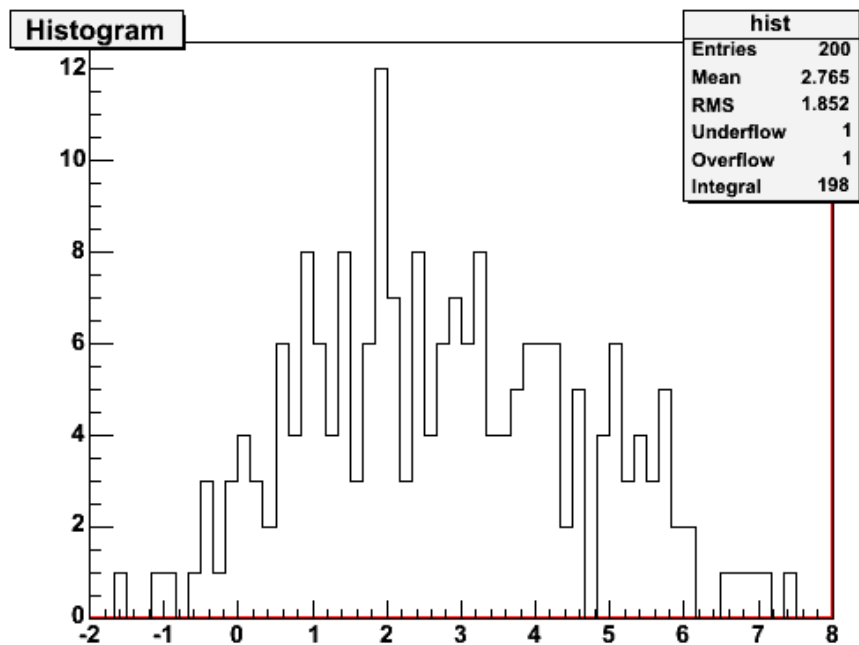
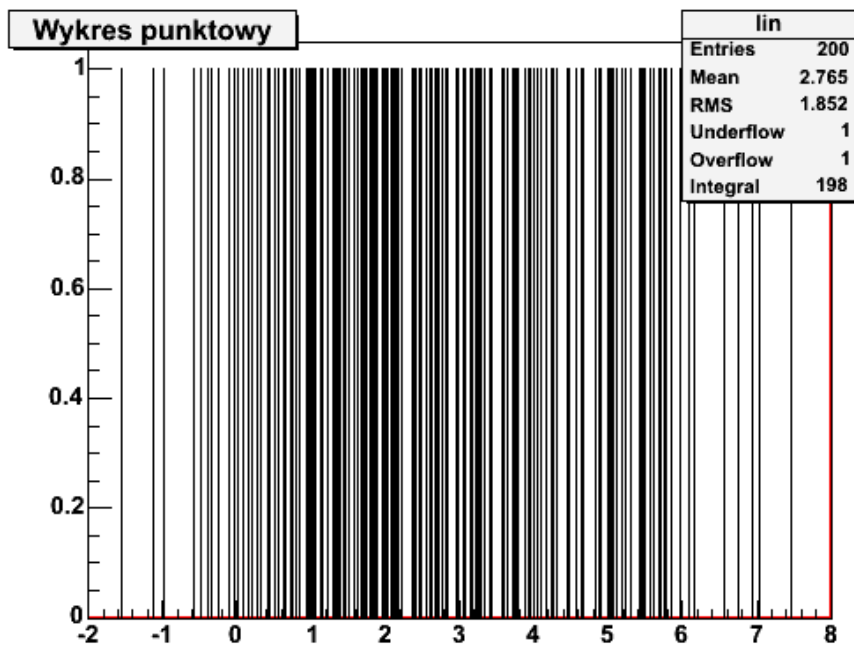


Graficzne przedstawienie próby

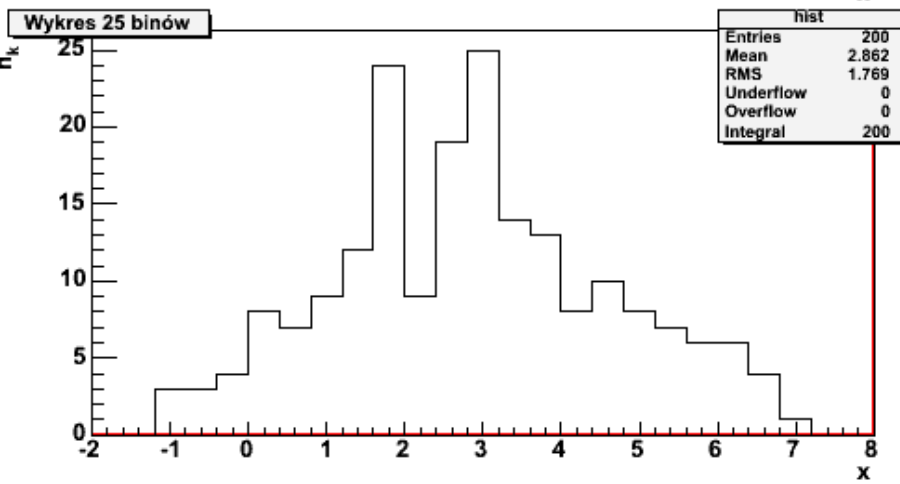
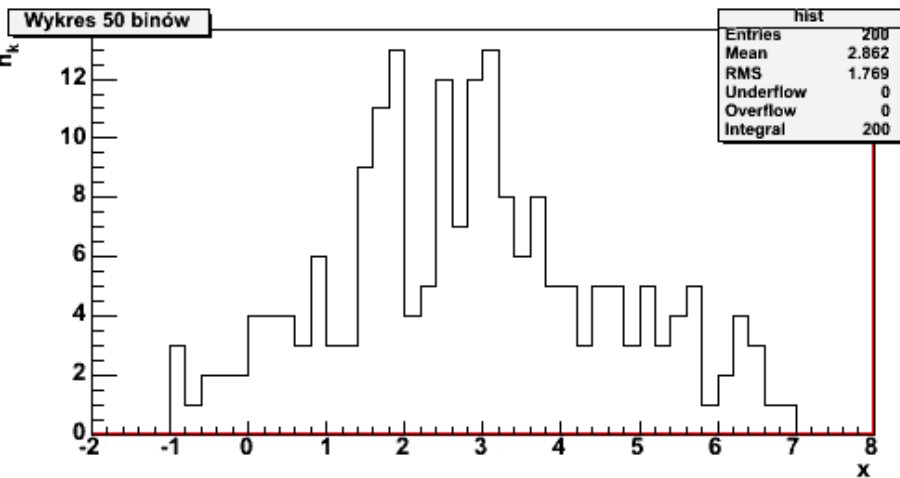
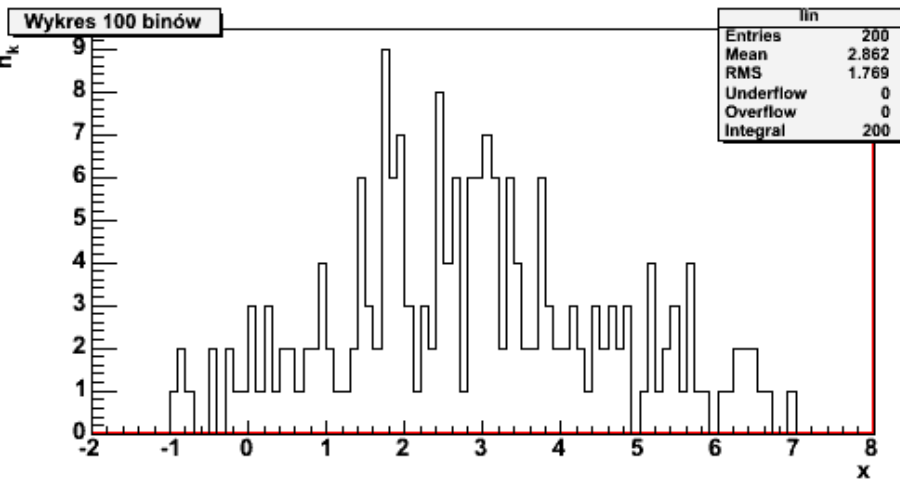
Graficzne przedstawienie próby

- Rozważmy próbę: x_1, x_2, \dots, x_n , która zależy od jednej zmiennej losowej X
- Możemy tę próbę przedstawić jako wykres 1D – punkty na osi x – jednowymiarowy wykres punktowy
 - **wada:** co w przypadku, gdy mamy dwa takie same pomiary?
- Z reguły stosujemy zatem wykres 2D, zwany **histogramem**:
 - dzielimy przedział zmienności x (lub jego część) na r przedziałów o jednakowej szerokości Δx : $\xi_1, \xi_2, \dots, \xi_r$
 - środki przedziałów znajdują się w punktach: x_1, x_2, \dots, x_r
 - na osi y odkładamy liczbę elementów próby przypadającą na dany przedział: n_1, n_2, \dots, n_r
 - tak otrzymany wykres nazywamy **wykresem częstości** lub **histogramem**

Graficzne przedstawienie próby



Histogram - szerokość przedziału



- Im więcej przedziałów, tym informacja o próbie jest dokładniejsza
- Większa ilość przedziałów powoduje jednak większe wahania statystyczne *od punktu do punktu*
- Pole pod krzywą schodkową jest proporcjonalne do wielkości próby (jeśli je przeskalujemy przez $1/n$, otrzymamy częstość)

Graficzne przedstawienie próby - przykład

- Badamy “nieznany” rozkład prawdopodobieństwa przez estymatory
- Symulujemy taką sytuację poprzez generację 1000 prób z rozkładu Gaussa o wartości średniej 0 i wariancji 1. Każda próba ma licznosc (liczbę składników) r .
- Badamy zachowanie estymatorów charakterystyk rozkładu i estymatorów ich niepewności w funkcji licznosci r

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

$$\Delta \bar{X} = \sqrt{(S^2(\bar{X}))} = S(\bar{X}) = \frac{1}{\sqrt{n}} S(X)$$

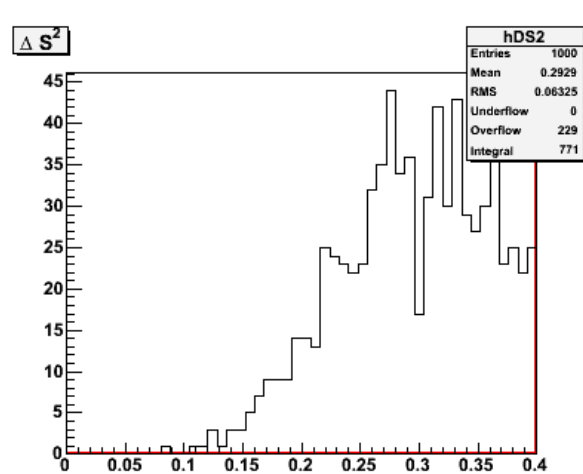
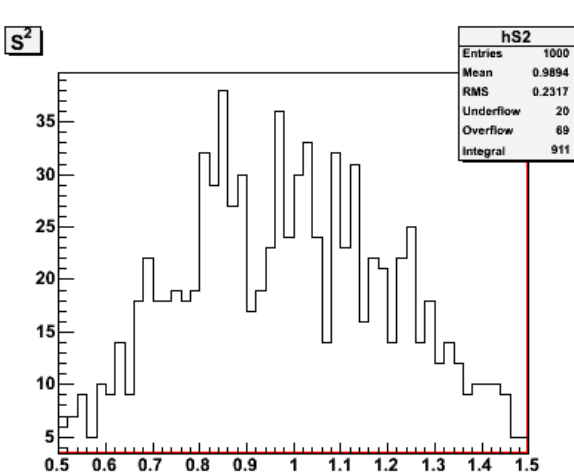
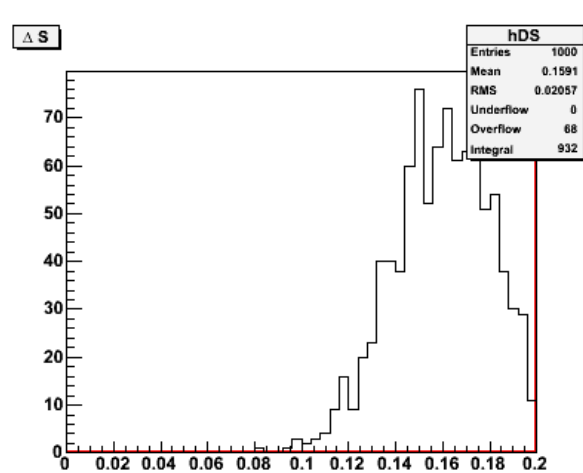
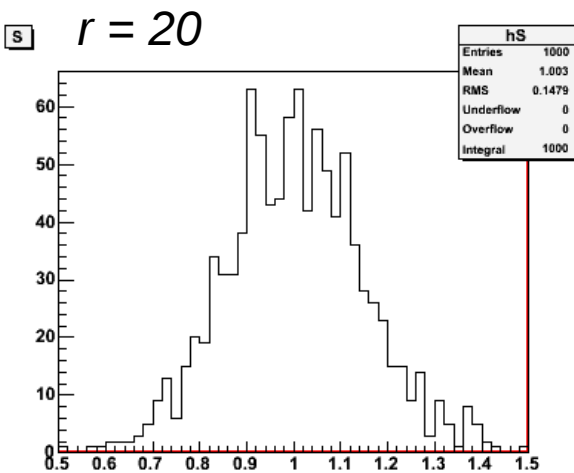
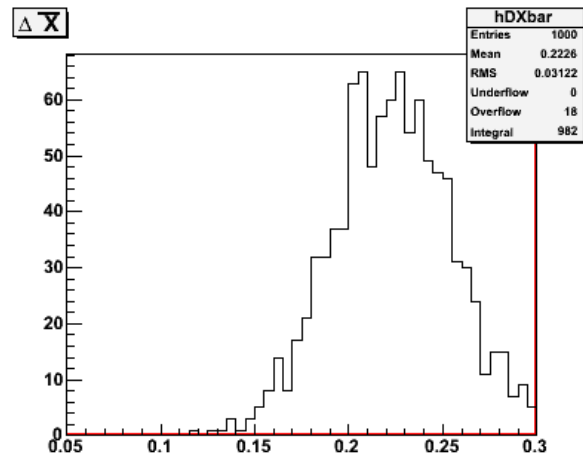
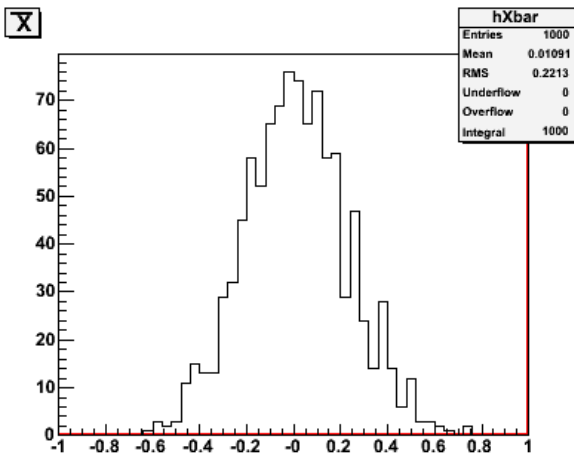
$$S = \sqrt{S^2} = \frac{1}{\sqrt{n-1}} \sqrt{\sum (X_i - \bar{X})^2}$$

$$\Delta S = \frac{S}{\sqrt{2(n-1)}}$$

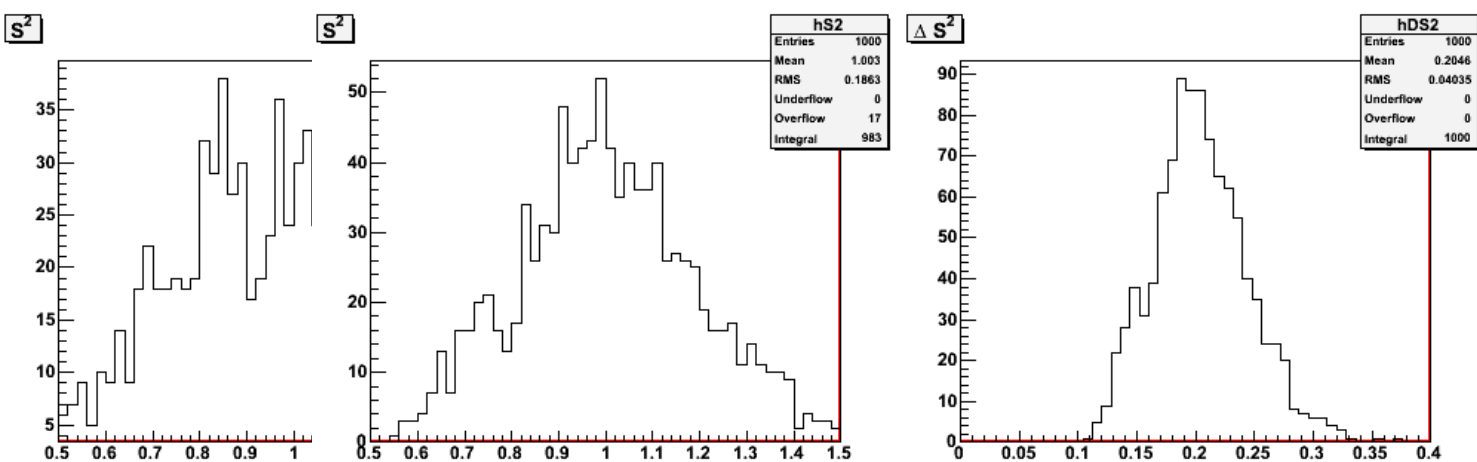
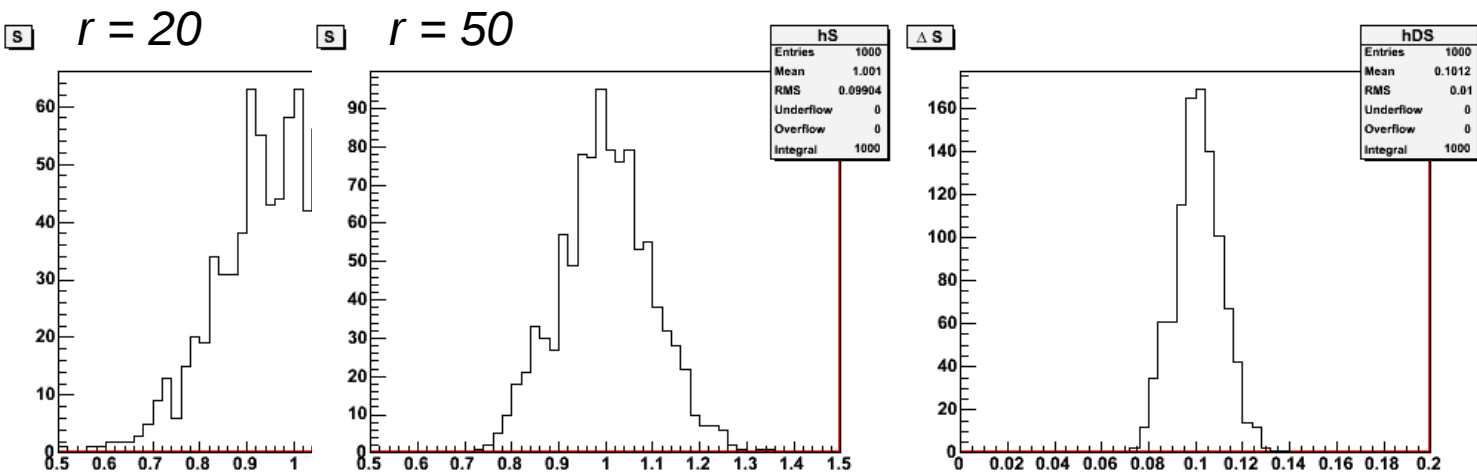
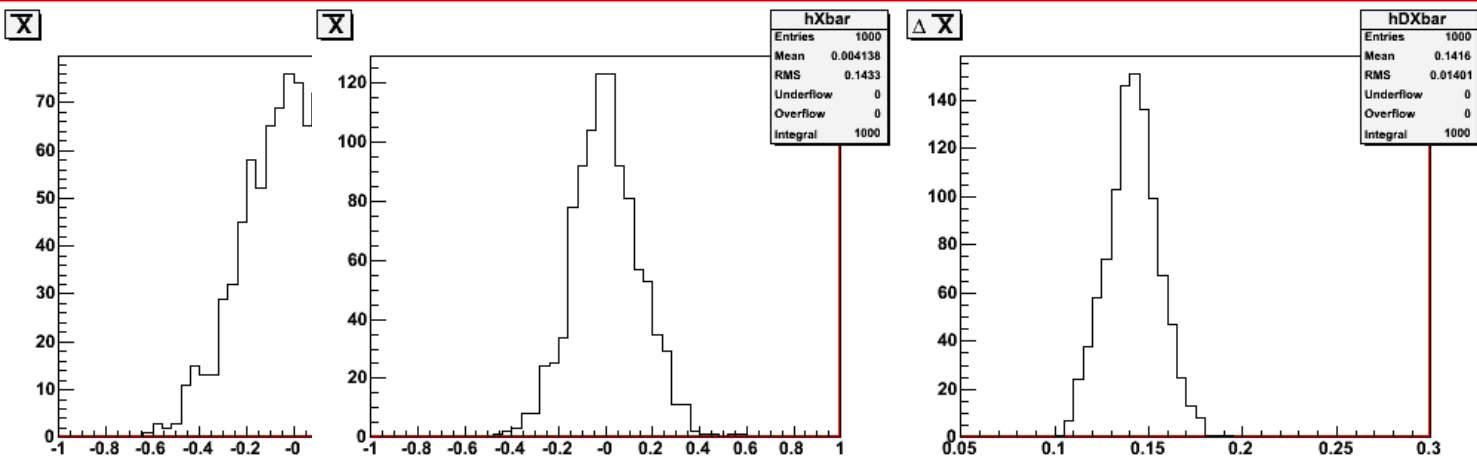
$$S^2 = \frac{1}{n-1} \{ (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \}$$

$$\Delta S^2 = S^2 \sqrt{\frac{2}{n-1}}$$

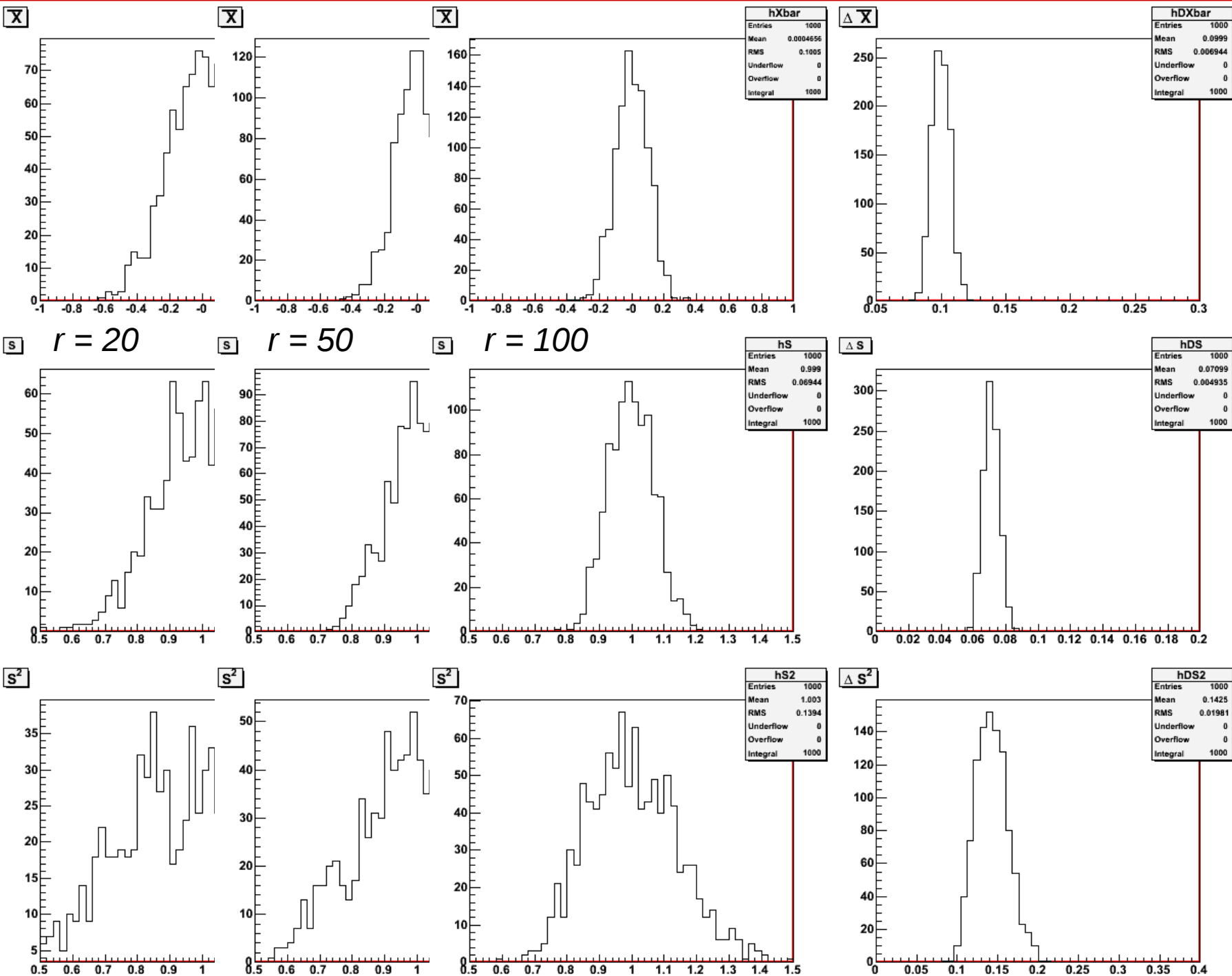
Estymatory - histogramy



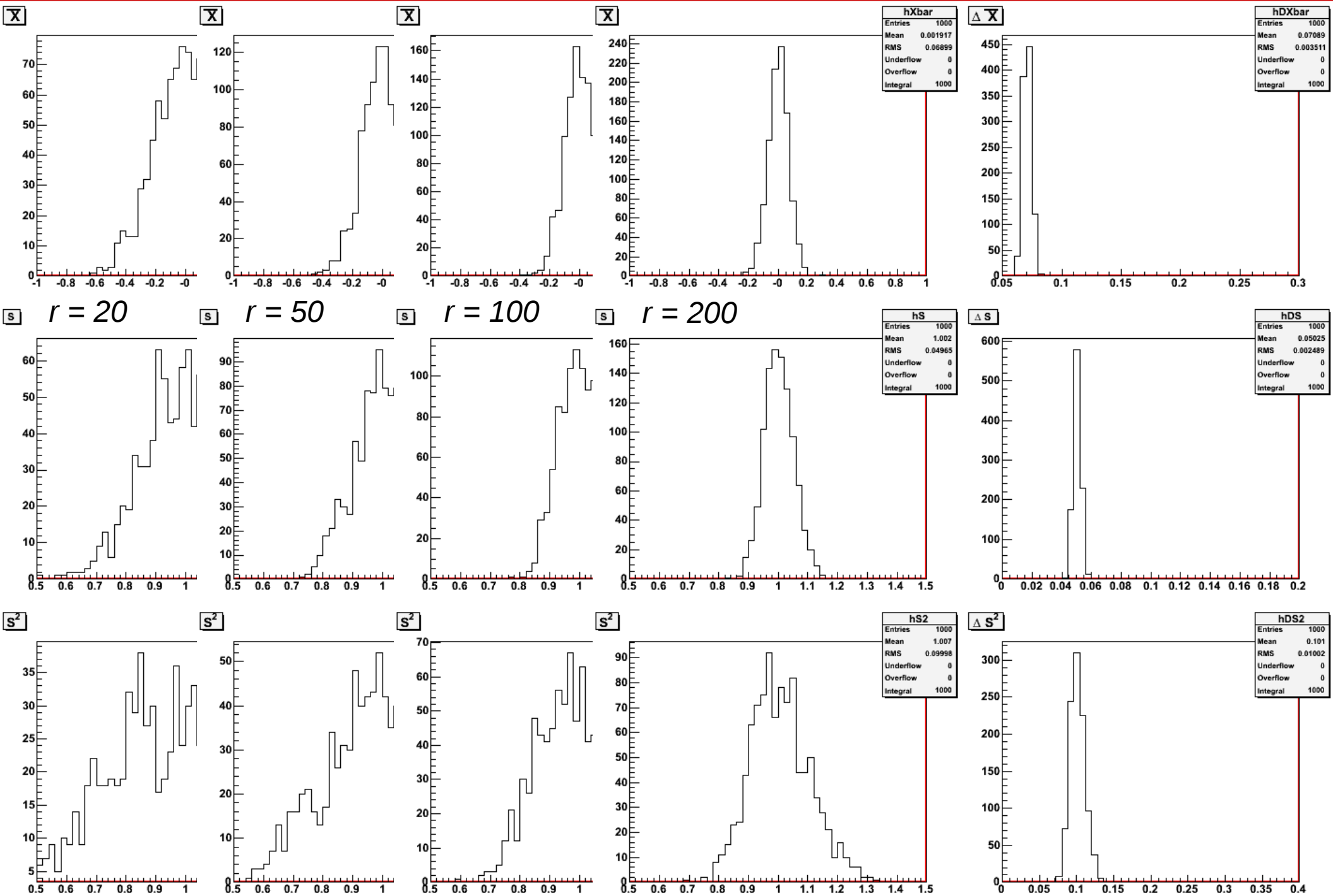
Estymatory - histogramy



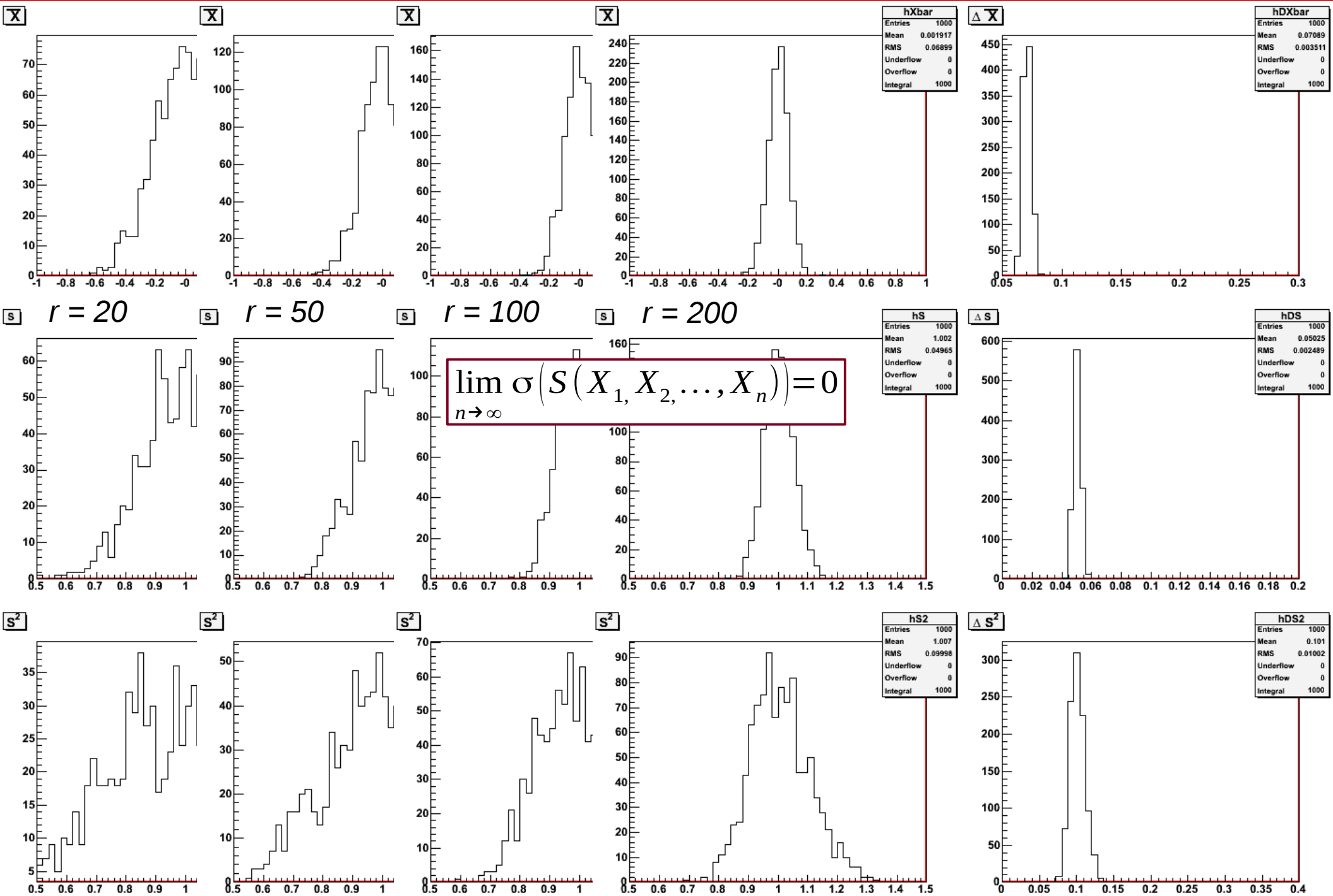
Estymatory - histogramy



Estymatory - histogramy



Estymatory - histogramy





Pobieranie próby z rozkładów cząstkowych

Próby z rozkładów cząstkowych

- Często wygodnie jest podzielić populację G (np. wszystkich studentów w Europie) na t podpopulacji G_i (np. studentów na poszczególnych uniwersytetach)

- Podpopulacje G_i , które są opisane gęstościami prawdopodobieństwa $f_i(x)$ i mają odpowiednie dystrybuanty:

$$F_i(x) = \int_{-\infty}^x f_i(x) dx = P(X \leq x | x \in G_i)$$

- Dla całej populacji mamy zatem:

$$F(x) = P(X \leq x | x \in G) = \sum_{i=1}^t P(X < x | X \in G_i) P(X \in G_i) = \sum_{i=1}^t P(X \in G_i) F_i(x)$$

- Możemy wprowadzić oznaczenie: $P(X \in G_i) = p_i$

- Obliczamy wartość średnią populacji G :

$$\hat{x} = E(X) = \int_{-\infty}^{\infty} xf(x) dx = \sum_{i=1}^t p_i \int_{-\infty}^{\infty} xf_i(x) dx = \sum_{i=1}^t p_i \hat{x}_i$$

- **Wniosek:** wartość średnia z populacji to średnia ważona wartości średnich podpopulacji pomnożonych przez ich prawdopodobieństwa

- **Uwaga! Oznaczenia:** średnia z populacji: \hat{x} średnia z próby: \bar{X}

Próby z rozkładów cząstkowych

- Wariancja z populacji G (uwzględniając niezależność elementów X_i):

$$\sigma^2(X) = E((x - \hat{x})^2) = \sum_{i=1}^t p_i E\left(\left[(x - \hat{x}_i) + (\hat{x}_i - \hat{x})\right]^2\right) = \sum_{i=1}^t p_i \left(\sigma_i^2 + (\hat{x}_i - \hat{x})^2\right)$$

- Wniosek:** wariancja jest średnią ważoną wariancji z podpopulacji i wariancji wartości średniej podpopulacji względem wartości średniej z całej populacji

- Teraz z każdej podpopulacji wybierzmy próbkę o liczności n_i , w sumie n elementów: $n = \sum_{i=1}^t n_i$. Średnia arytmetyczna z całej próby wynosi wtedy:

$$\bar{X}_p = \frac{1}{n} \sum_{i=1}^t \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^t n_i \bar{X}_i$$

- Wartość oczekiwana i wariancja średniej z całej próby:

$$E(\bar{X}_p) = \frac{1}{n} \sum_{i=1}^t n_i \hat{x}_i$$

$$\sigma^2(\bar{X}_p) = \frac{1}{n^2} \sum_{i=1}^t n_i^2 E\left(\left(\bar{X}_i - \hat{x}_i\right)^2\right) = \frac{1}{n^2} \sum_{i=1}^t n_i^2 \sigma^2(\bar{X}_i) = \frac{1}{n} \sum_{i=1}^t \frac{n_i}{n} \sigma_i^2$$

Estymatory dla rozkładów cząstkowych

- Zauważmy jednak, że wartość średnia \bar{X}_p nie może być estymatorem wartości średniej z całej populacji \hat{x} , gdyż zależy ona od dowolnego wyboru wielkości n_i próbek cząstkowych (mamy n_i)

$$E(S(X_1, X_2, \dots, X_n)) = \lambda, \text{ dla każdego } n$$

- Jeśli jednak porównamy wzory na średnią z populacji i całej próby:

$$\hat{x} = \sum_{i=1}^t p_i \hat{x}_i \qquad \bar{X}_p = \frac{1}{n} \sum_{i=1}^t n_i \hat{x}_i$$

- To widać, że jeżeli warunek $p_i = n_i/n$ jest spełniony, to wartość średnia z populacji \hat{x} może być estymowana przez wyznaczenie najpierw wartości średnich poszczególnych prób \bar{X}_i , wewnątrz poszczególnych podpopulacji, a potem przez wyrażenie:

$$\tilde{X} = \sum_{i=1}^t p_i \bar{X}_i \quad \text{- estymator wartości średniej z populacji}$$

zależny od wartości średnich z prób

- Wariancja powyższego estymatora:

$$\sigma^2(\tilde{X}) = \sum_{i=1}^t p_i^2 \sigma^2(\bar{X}_i) = \sum_{i=1}^t \frac{p_i^2}{n_i} \sigma_i^2$$

- Jaka jest optymalna wielkość próbek n_i , która pozwala na minimalizację powyższej wariancji (niepewności estymatora)? $n_i = n p_i \sigma_i / \sum_{i=1}^t p_i \sigma_i$

- optymalna wielkość n_i próby wewnątrz podpopulacji i , jest prop. do prawdopod. p_i ważonej z odpowiednim odchyleniem standardowym



Pobieranie próby ze skończonej populacji bez uzupełnień Stopnie swobody

Rozkład hipergeometryczny – przypomnienie

- W urnie jest N kul – k białych i $N-K$ czarnych
- W n próbach wyciągamy (bez zwracania) k kul białych i $n-k=l$ kul czarnych. Jakie jest prawdopodobieństwo wyciągnięcia k kul białych?
- Wylosowanie kolejnej kulki zmienia proporcje kul białych do czarnych i wpływa na wynik kolejnego losowania – rozkład dwumianowy nie ma tu zastosowania. Mamy jednak:

- liczbę możliwości wylosowania n z N kulek: $\binom{N}{n}$
- prawdopodobieństwo takiego zdarzenia: $1/\binom{N}{n}$
- możliwość wylosowania k spośród K białych i l spośród L czarnych kulek wynoszą: $\binom{K}{k} \binom{L}{l}$
- prawdopodobieństwo szukane wynosi zatem: $W_k = \frac{\binom{K}{k} \binom{L}{l}}{\binom{N}{n}}$

Próba ze skończonej populacji

- Poszczególne pobrania w rozkładzie hipergeometrycznym nie są niezależne → jest to **przykład pobierania elementów ze skończonej (o wartościach dyskretnych) populacji bez jej uzupełniania**, nie jest to więc prawdziwie losowe pobieranie próby
- Rozpatrzmy populację składającą się z N elementów y_1, y_2, \dots, y_N
- Pobieramy z niej próbę o liczebności n i elementach X_1, X_2, \dots, X_N
 - w przypadku rozkładu hipergeometrycznego zarówno y_i i X_i mogą przyjmować tylko wartości 0 i 1
- Ponieważ prawdopodobieństwo pobrania jakiegokolwiek elementu populacji jest jednakowe:
$$E(y) = \hat{y} = \bar{y} = \frac{1}{N} \sum_{j=1}^n y_j$$
 - średnia arytmetyczna z populacji składającej się ze skończonej liczby elementów nie jest jednak zmienną losową
- Ponieważ mamy skończoną populację, definicja wariancji będzie miała ten sam “problem” co estymator wariancji z ostatniego wykładu

Próba ze skończonej populacji

- Analogicznie jak w przypadku estymatora wariancji, określamy zatem: wariancję populacji przez sumę kwadratów różnic:

$$\sigma^2(y) = \frac{1}{N-1} \sum_{j=1}^N (y_j - \bar{y})^2 = \frac{1}{N-1} \left(\sum_{j=1}^N y_j^2 - \frac{1}{N} \left(\sum_{j=1}^N y_j \right)^2 \right) \quad \bar{y} = \frac{1}{N} \sum_{j=1}^N y_j$$

- Zajmijmy się kwadratami różnic:

$$\sum_{j=1}^N (y_j - \bar{y})^2$$

- Ponieważ nie ograniczyliśmy populacji w żaden sposób, wartości y_i mogą przybierać dowolne wartości
- Czyli pierwszy składnik sumy, może przybierać dowolną wartość, analogicznie 2, 3, ..., (N-1)-szy składnik
- Natomiast N-ty (któryś z nich) składnik sumy będzie ograniczony warunkiem:

$$\sum_{j=1}^N (y_j - \bar{y}) = N \cdot \bar{y} - N \cdot \bar{y} = 0$$

- Warunek jest konieczny, aby zgodziła się średnia
- Mówimy, że **liczba stopni swobody dla sumy kwadratów wynosi N-1**

Próba ze skończonej populacji

- Mówimy, że liczba stopni swobody dla sumy kwadratów wynosi $N-1$
- Suma kwadratów dzielona przez liczbę stopni swobody to **odchylenie średnie kwadratowe**:
$$\sigma^2(y) = \frac{1}{N-1} \sum_{j=1}^N (y_j - \bar{y})^2$$
- Pierwiastek z tej wielkości (mający wymiar wielkości mierzonej) to pierwiastek ze średniego odchylenia kwadratowego (*ang. RMS – root-mean square deviation*)



Pobieranie próby z rozkładu normalnego

Pobieranie próby z rozkładu normalnego

- Badamy populację opisaną rozkładem Gaussa o wartości średniej a i wariancji σ^2 . Z tej populacji wybieramy próbę o liczności n

- Napiszmy funkcję charakterystyczną tego rozkładu:

$$\varphi_X(t) = \exp(it a) \exp(-\sigma^2 t^2 / 2) \Rightarrow \varphi_{\bar{X}}(t) = \left\{ \exp\left(i \frac{t}{n} a\right) \exp\left(-\frac{\sigma^2}{2} \left(\frac{t}{n}\right)^2\right) \right\}^n$$

- Rozpatrując zmienną $\bar{X} - a = \bar{X} - \hat{x}$ zamiast X dostaniemy:

$$\varphi_{\bar{X}-a}(t) = \exp\left(-\frac{\sigma^2 t^2}{2n}\right)$$

- Czyli funkcję charakterystyczną rozkładu normalnego o tej samej średniej ale zmienionej wariancji: $\sigma^2(\bar{X}) = \sigma^2(X)/n$

- Rozpatrzmy najprostszyp przypadek – rozkład normalny (średnia 0 i odchylenie standardowe 1): $\varphi_{\bar{X}}(t) = \exp(-t^2 / 2n)$

- Pobieramy z niego próbę n -elementową i tworzymy sumę kwadratów:

$$\chi^2 \equiv X^2 = X_1^2 + X_2^2 + \dots + X_n^2$$

- Można udowodnić, że dystrybuanta ma postać: $F(\chi^2) = \frac{1}{\Gamma(\lambda) 2^\lambda} \int_0^{\chi^2} u^{\lambda-1} e^{-1/2 u} du$

Pobieranie próby z rozkładu normalnego

- Można udowodnić, że dystrybuanta ma postać:

$$F(\chi^2) = \frac{1}{\Gamma(\lambda)2^\lambda} \int_0^{\chi^2} u^{\lambda-1} e^{-1/2u} du$$

– gdzie $\lambda = 1/2 n$ a n to liczba stopni swobody

- Wprowadzając oznaczenie: $k = \frac{1}{\Gamma(\lambda)2^\lambda}$
- Otrzymujemy rozkład gęstości prawdopodobieństwa:

$$f(\chi^2) = k \cdot (\chi^2)^{\lambda-1} e^{-1/2\chi^2}$$

- Funkcja charakterystyczna rozkładu χ^2 ma postać:

$$\varphi_{\chi^2}(t) = (1 - 2it)^{-\lambda}$$

- Korzystając z jej własności funkcji charakterystycznej otrzymujemy natychmiast, że suma dwóch różnych χ^2 o n_1 i n_2 stopniach swobody daje również rozkład χ^2 o $n = n_1 + n_2$ stopniach swobody

Pobieranie próby z rozkładu normalnego

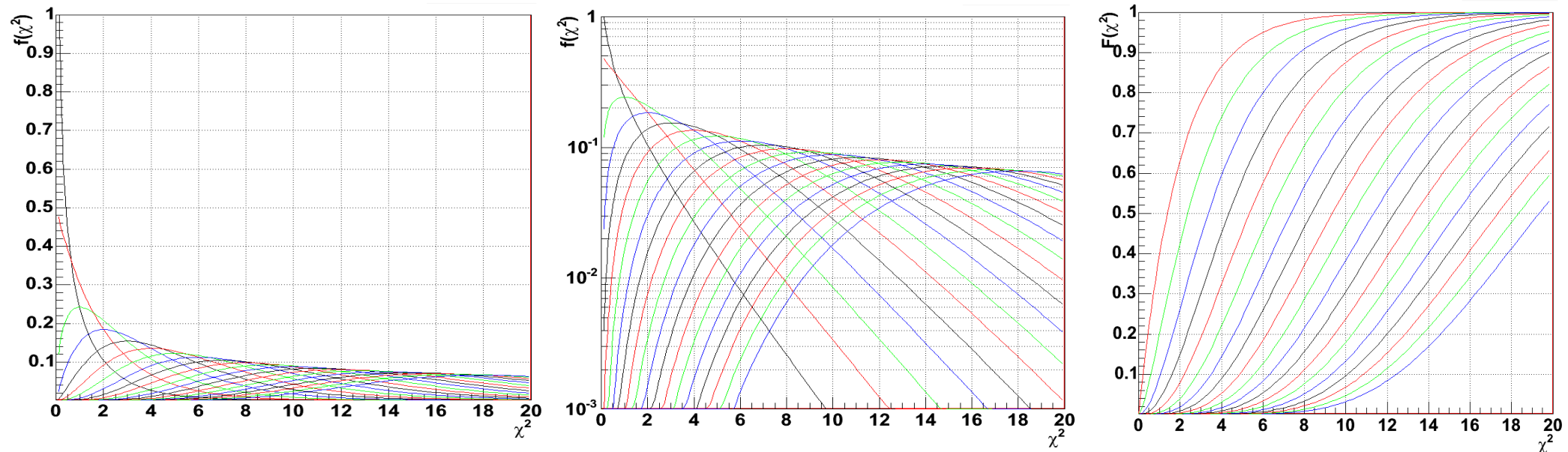
- Różniczkując funkcję charakterystyczną możemy wyznaczyć wartość oczekiwaną i wariancję rozkładu χ^2 :

$$E\{X^2\} = -i\varphi_{\chi^2}'(0) = 2\lambda \equiv n$$

$$E\{(X^2)^2\} = -i\varphi_{\chi^2}''(0) = 4\lambda^2 + 4\lambda$$

$$\sigma^2(X^2) = E\{(X^2)^2\} - (E\{X^2\})^2 = 4\lambda \equiv 2n$$

- Czyli wartość średnia z rozkładu χ^2 wynosi n a wariancja $2n$



- Wykresy rozkładu i dystrybuanty rozkładu χ^2 dla n od 1 do 20

Rozkład χ^2 – zastosowanie

- Rozkład χ^2 stosuje się jako miarę ufności uzyskanego wyniku (odchylenia elementów próby od wartości średniej populacji). Im mniejsza wartość χ^2 tym pozornie słuszniejszy wynik. Jako miary zaufania do wyniku używa się wielkości: $W(\chi^2) = 1 - F(\chi^2) \equiv p = 1 - \alpha$

nazywanej **poziomem ufności** (zwykle podawanym w %)

- W rzeczywistych przypadkach mamy do czynienia z pełnym rozkładem Gaussa o dowolnym a i σ . Wprowadzamy wtedy odpowiednie przeskalowanie

$$X^2 = \frac{(X_1 - a)^2 + (X_2 - a)^2 + \dots + (X_n - a)^2}{\sigma^2}$$

- a w ogólnym przypadku gdy zmienne są zależne:

$$X^2 = (\mathbf{X} - \mathbf{a})^T \mathbf{B} (\mathbf{X} - \mathbf{a})$$

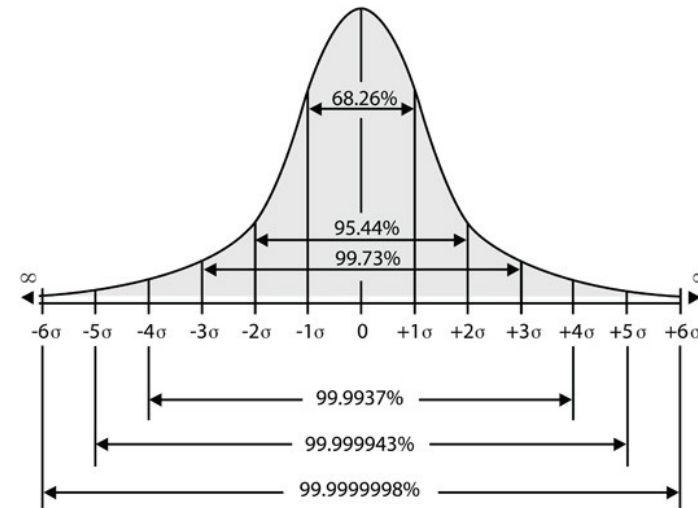
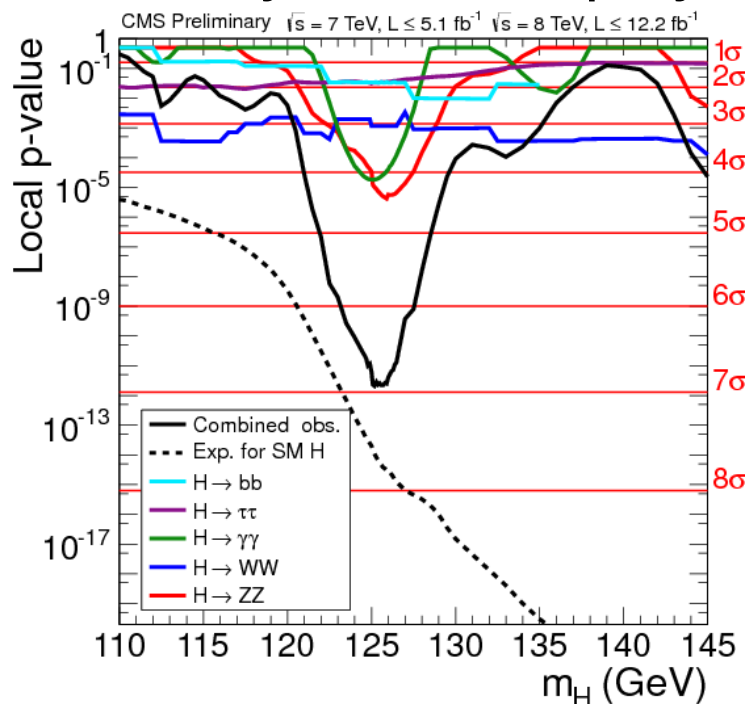
- Nieobciążony i zgodny estymator wariancji z populacji to:

$$S^2 = \frac{1}{n-1} \left\{ (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right\}$$

Rozkład χ^2 a estymator wariancji

- Można udowodnić, że zmienna losowa: $\frac{n-1}{\sigma^2} S^2$
- ma rozkład χ^2 z $f=n-1$ stopniami swobody. Wynika to stąd, że wyrażenia $(X_i - \bar{X})^2$ nie są liniowo niezależne, gdyż zawierają czynnik \bar{X} , który zależy od wszystkich wartości X_i . Każde dodatkowe równanie pomiędzy wyrażeniami $(X_i - \bar{X})^2$ redukuje liczbę stopni swobody o 1

- Poziomy ufności – przykład:



$$P(|Y - a| \leq \sigma) = 68,3\% \quad P(|Y - a| > \sigma) = 31,7\%$$

$$P(|Y - a| \leq 2\sigma) = 95,4\% \quad P(|Y - a| > 2\sigma) = 4,6\%$$

$$P(|Y - a| \leq 3\sigma) = 99,8\% \quad P(|Y - a| > 3\sigma) = 0,2\%$$



KONIEC