Komputerowa analiza danych doświadczalnych

Wykład 8
12.04.2017

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

Semestr letni 2016/2017

Pobieranie próby, estymatory histogram

Próby z rozkładów cząstkowych

Próby ze skończonej populacji

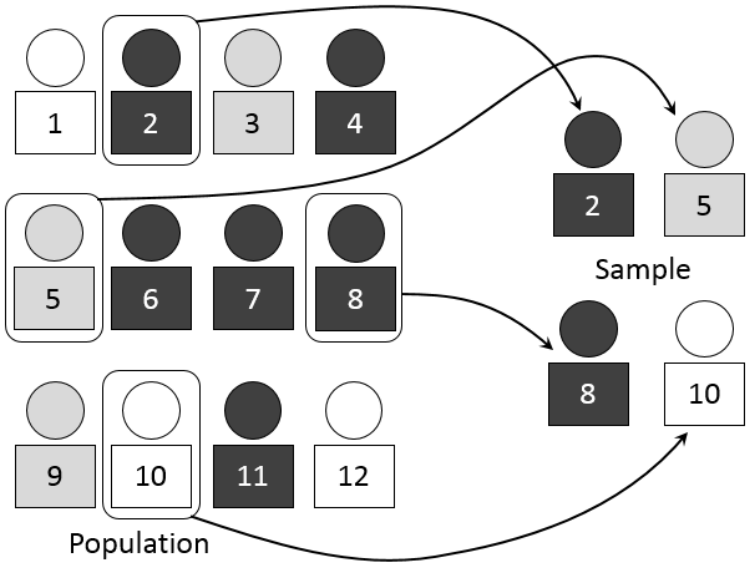
Próby z rozkładu normalnego

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 **przybliżać rozkład gęstości za pomocą rozkładu częstości** (histogramu o skończonej liczbie wejść)
- **Próbą** (*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**):

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

  - wariancji populacji – wariancja z próby (aproksymowana):

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

  - 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

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

  - odchylenia standardowego próby:

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

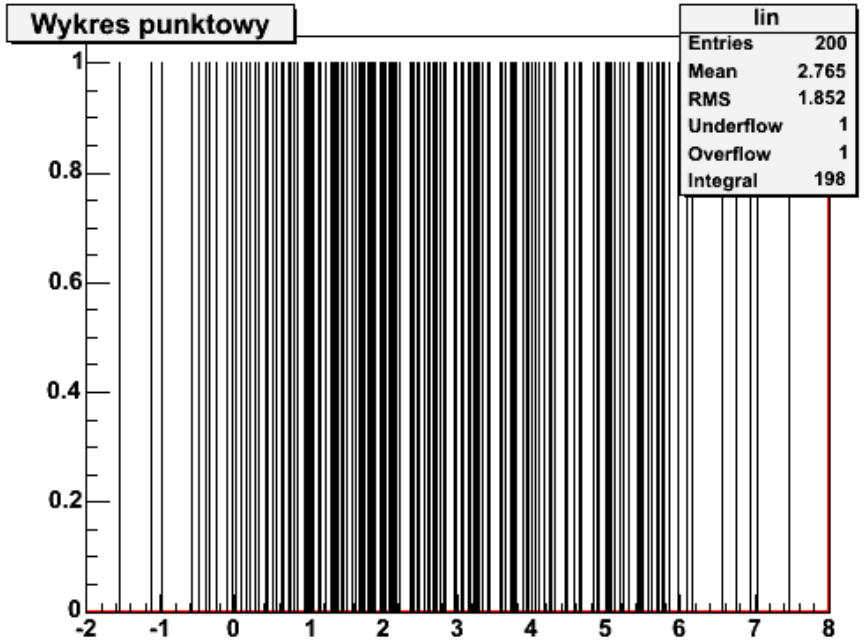
  - dalej możemy wyznaczać np. wariancję odchylenia std. próby...

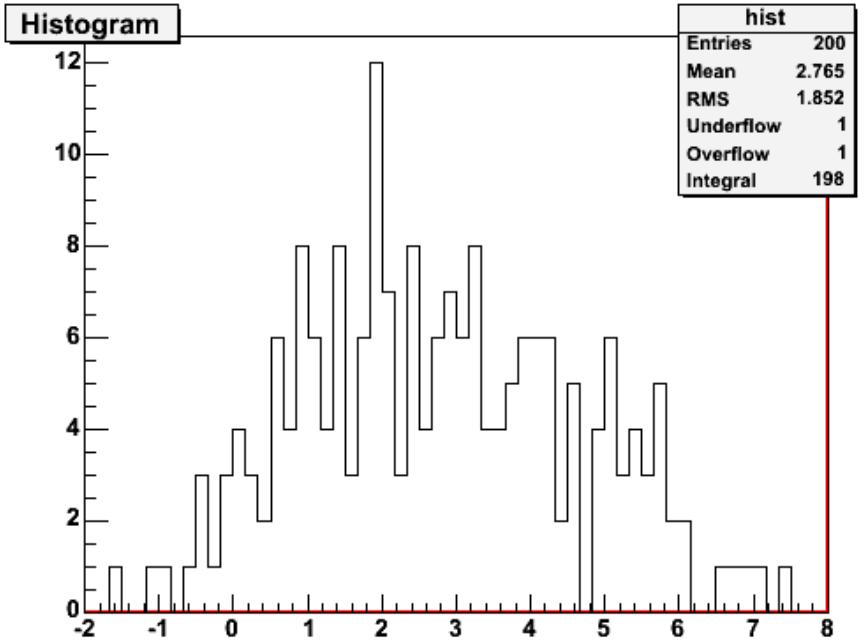
Graficzne przedstawienie próby

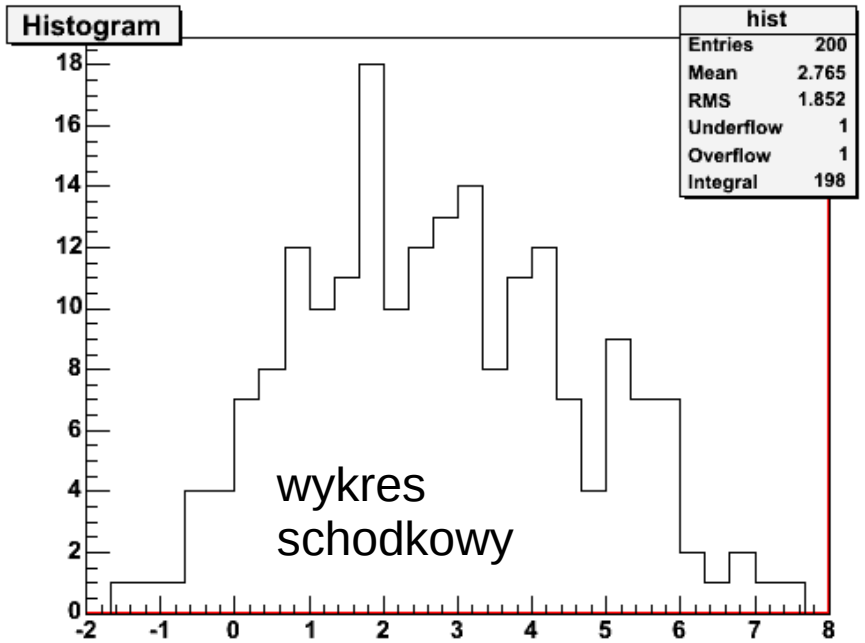
Graficzne przedstawienie próby

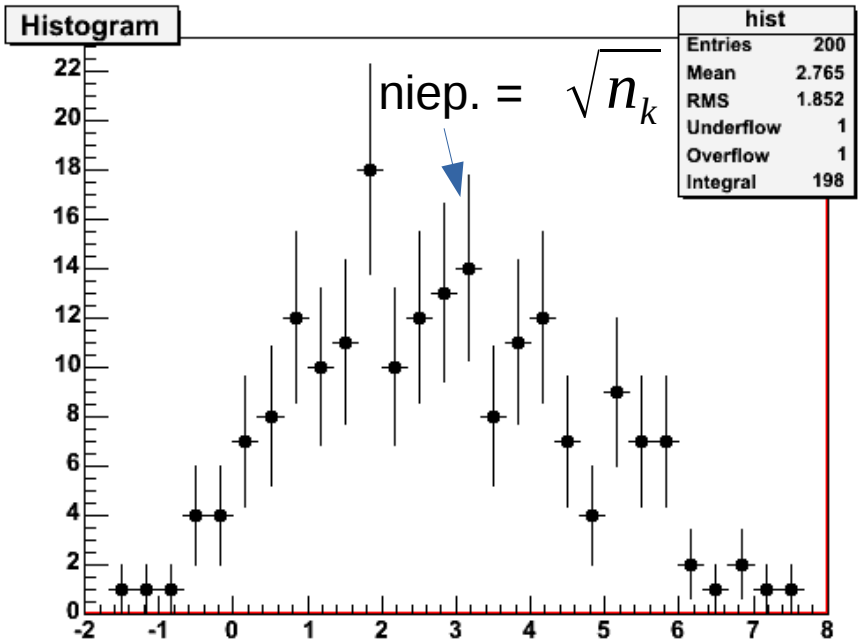
- Rozważmy próbę: $x_1, x_2, \ldots, x_n$, która zależy od jednej zmiennej losowej X
- Możemy tę próbę przedstawić jako wykres 1D – punkty na osi x – jednowymiarowy wykres punktowy
  - **wada:** co w przypadku, gdy mamy dwa takie same pomiary?
- Z reguły stosujemy zatem wykres 2D, zwany **histogramem**: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

Graficzne przedstawienie próby

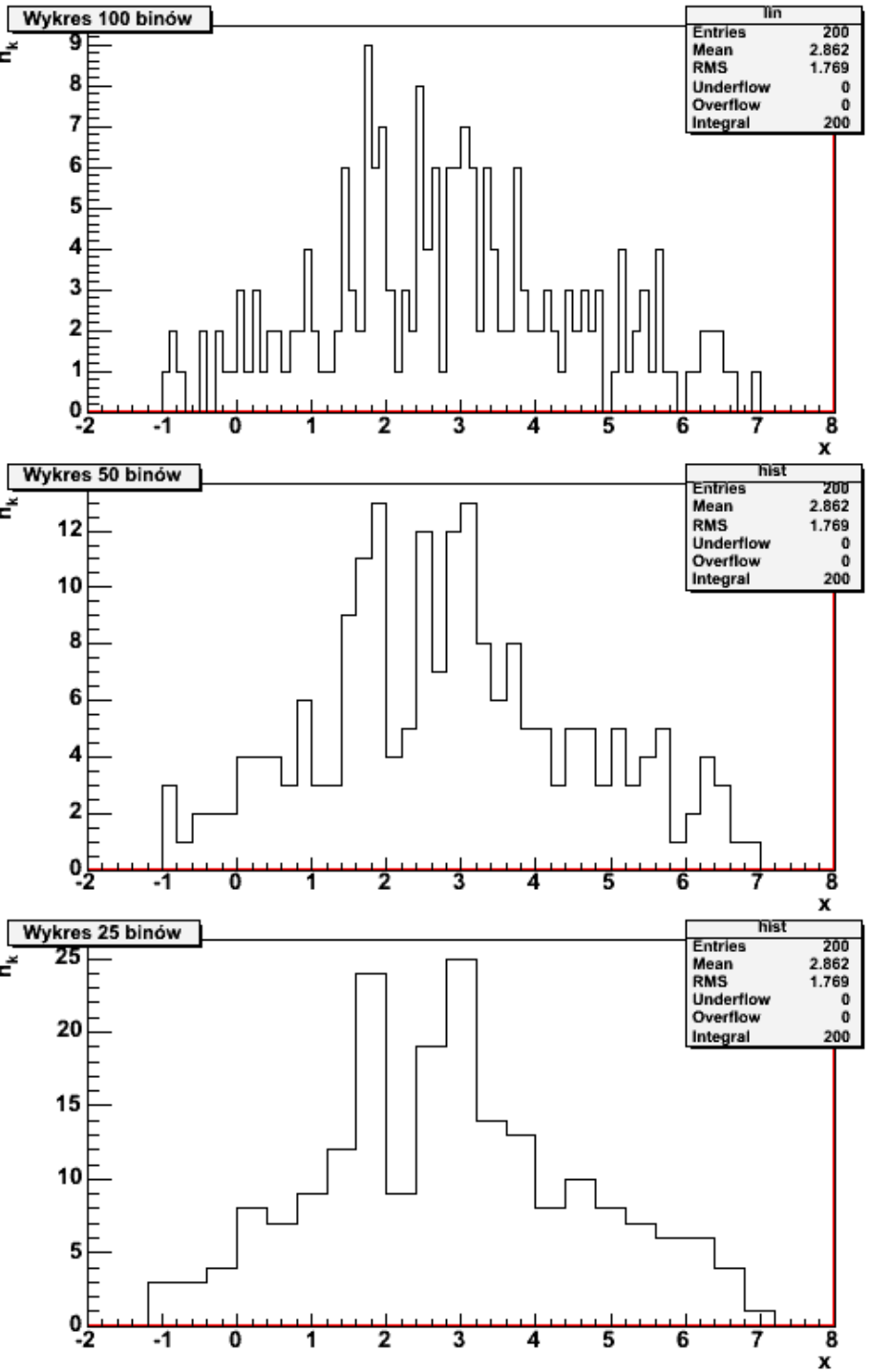








Histogram - szerokość przedziału



- Im więcej przedziałów, tym informacja o próbie jest dokładniejsza
- Większa ilość przedziałów powoduje jednak większe wahania statystyczne *od punktu do punku*
- 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 liczność (liczbę składników) r.
- Badamy zachowanie estymatorów charakterystyk rozkładu i estymatorów ich niepewności w funkcji liczności r

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

estymator wartości oczekiwanej populacji
średnia z próby

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

niepewność wart. średniej - estymator odch. st.wartości średniej z próby (estymatora wart. oczekiwanej)

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

estymator **odch. std.** populacji

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

niepewność estymatora odch. std. populacji – estymator odch. std. estymatora **odch. std.** populacji

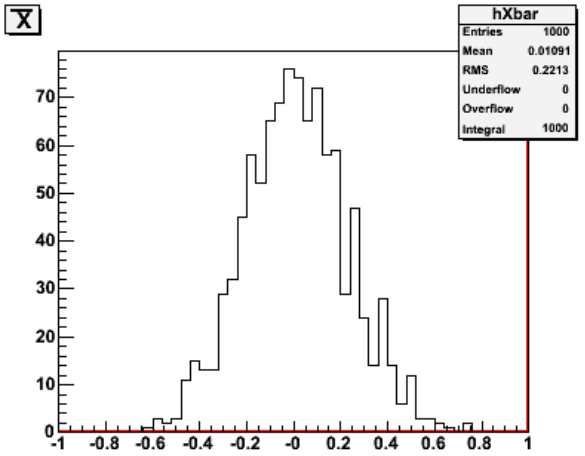
$$S^2(X) = \frac{1}{n-1}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2\right\}$$

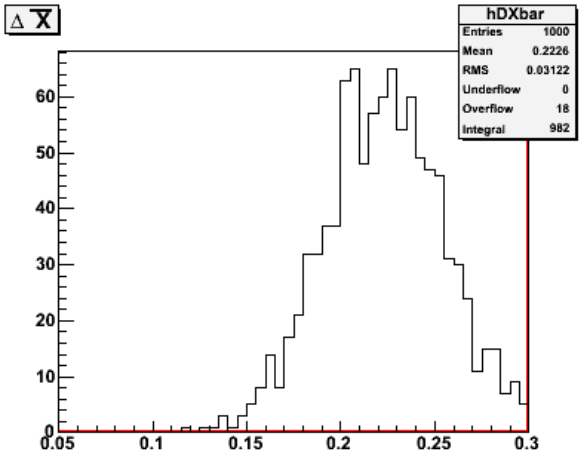
estymator **wariancji** populacji

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

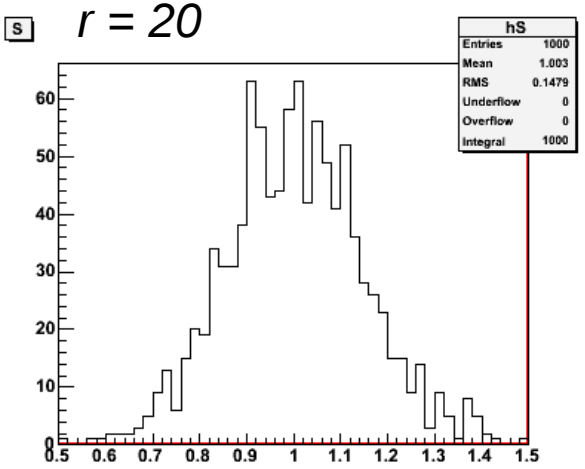
niepewność estymatora wariancji populacji – estymator odch. std. estymatora **wariancji** populacji

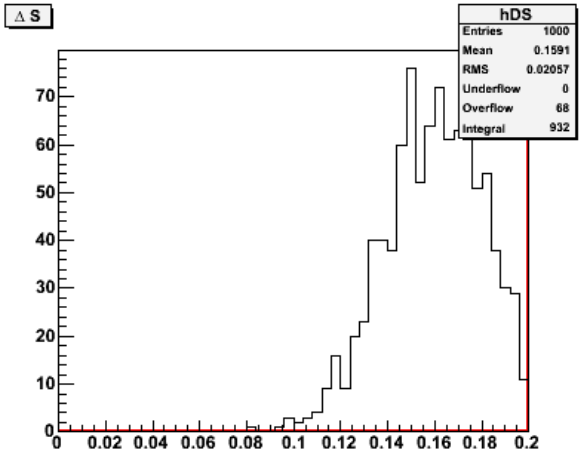
Estymatory - histogramy

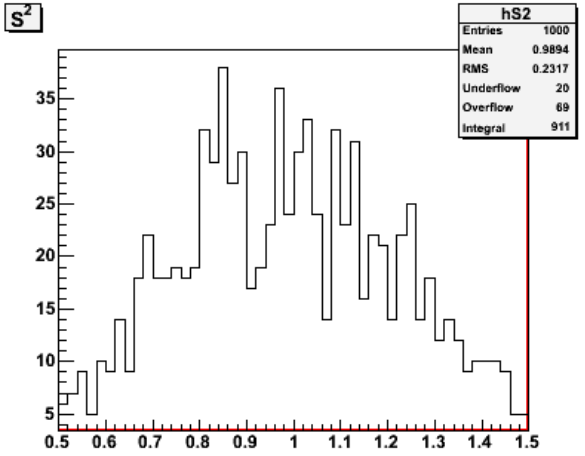


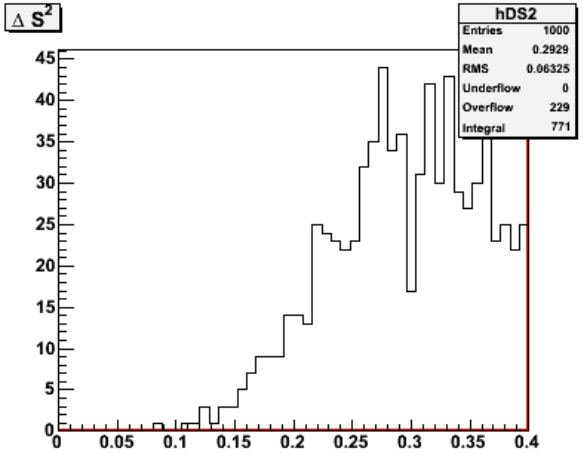


r = 20

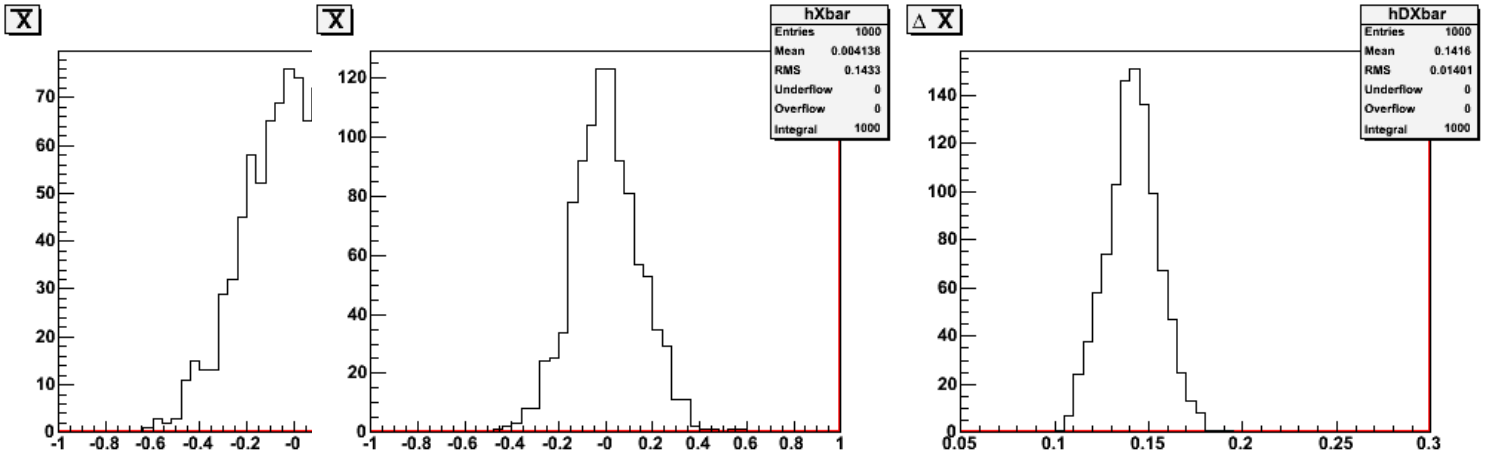




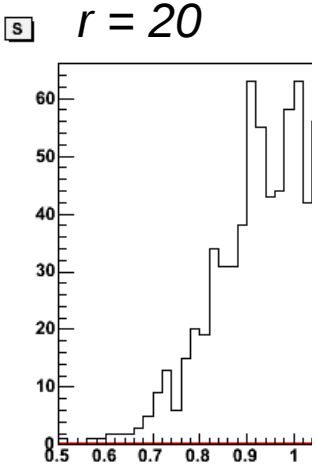




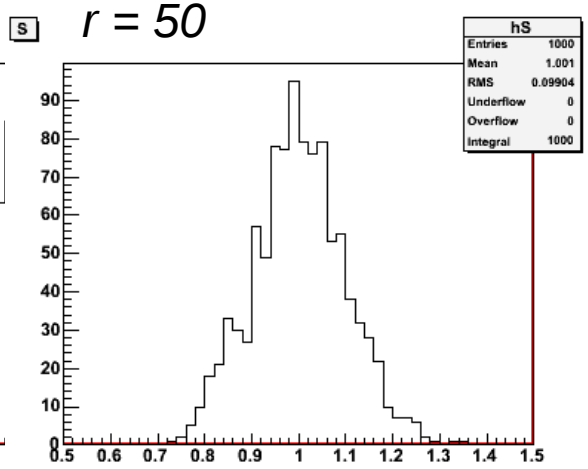
Estymatory - histogramy

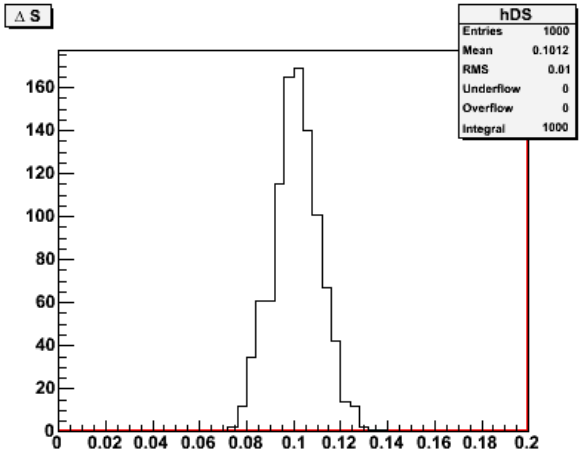


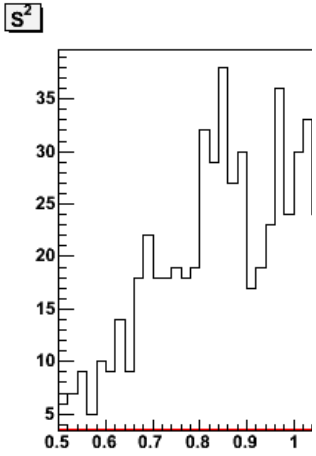
r = 20

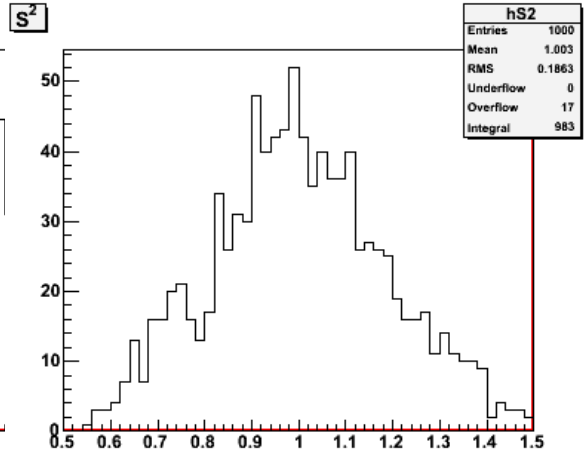


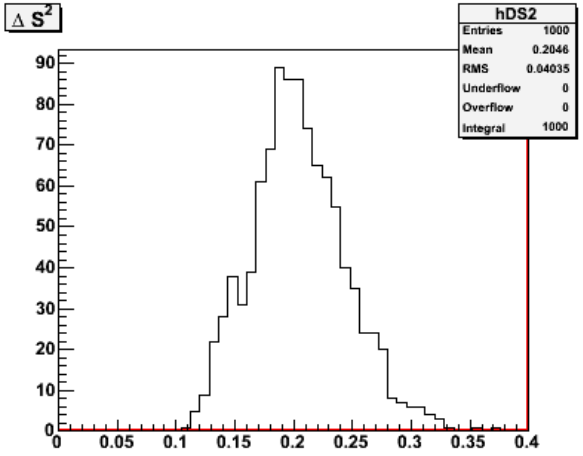
r = 50



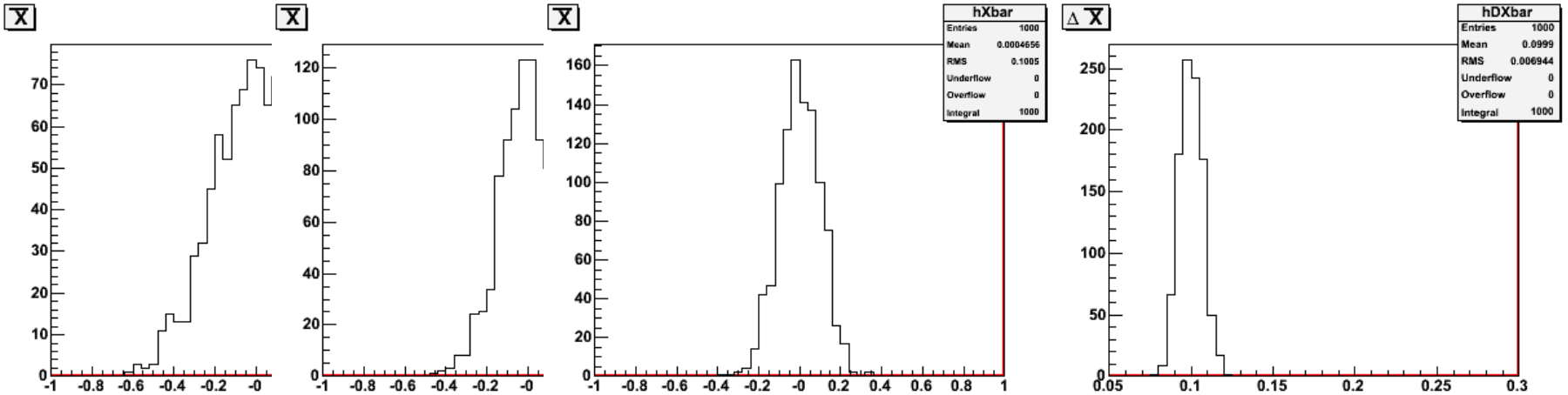


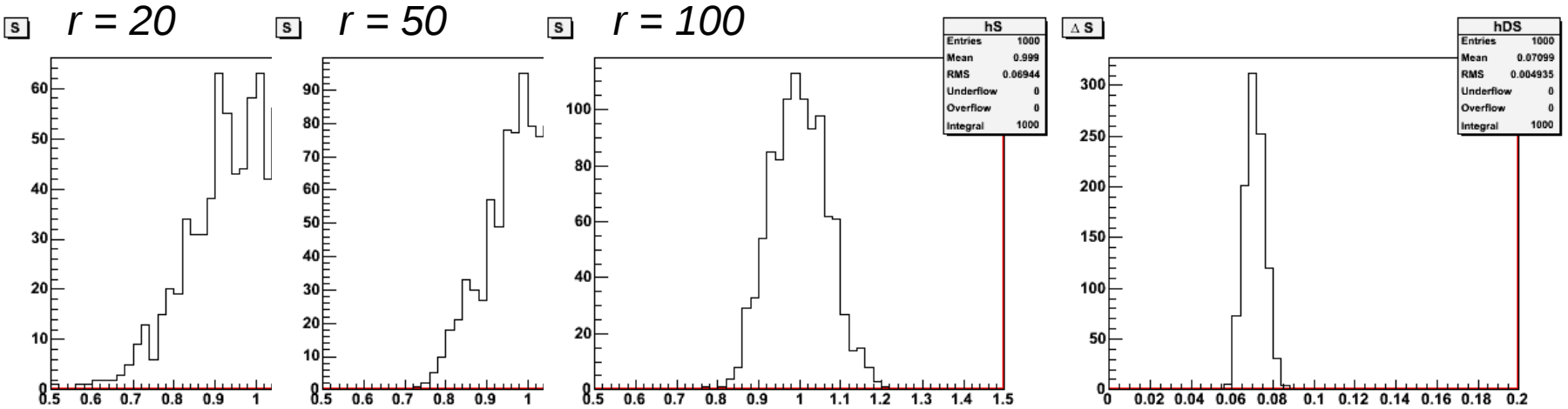


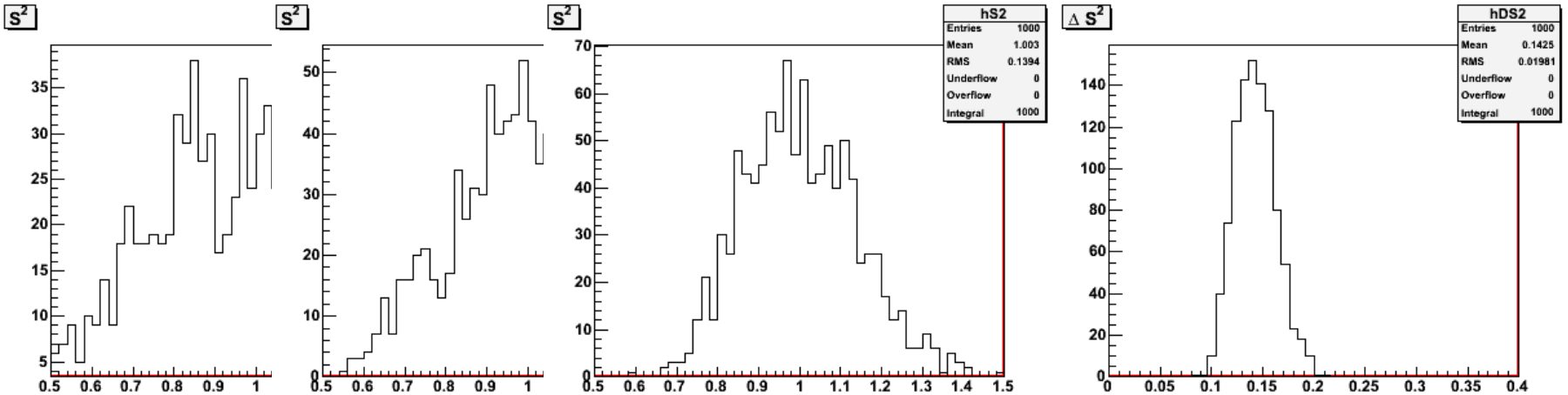




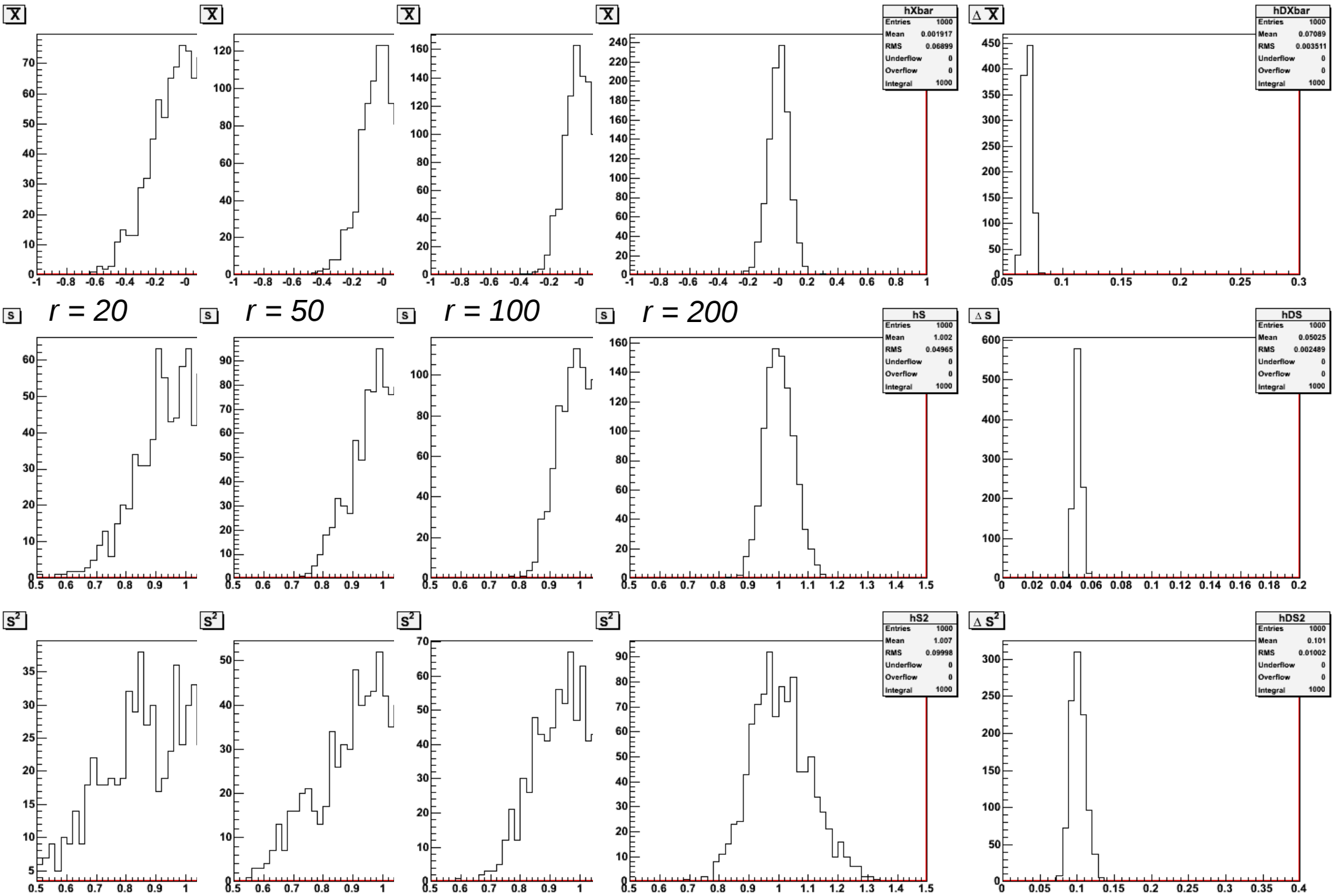
Estymatory - histogramy



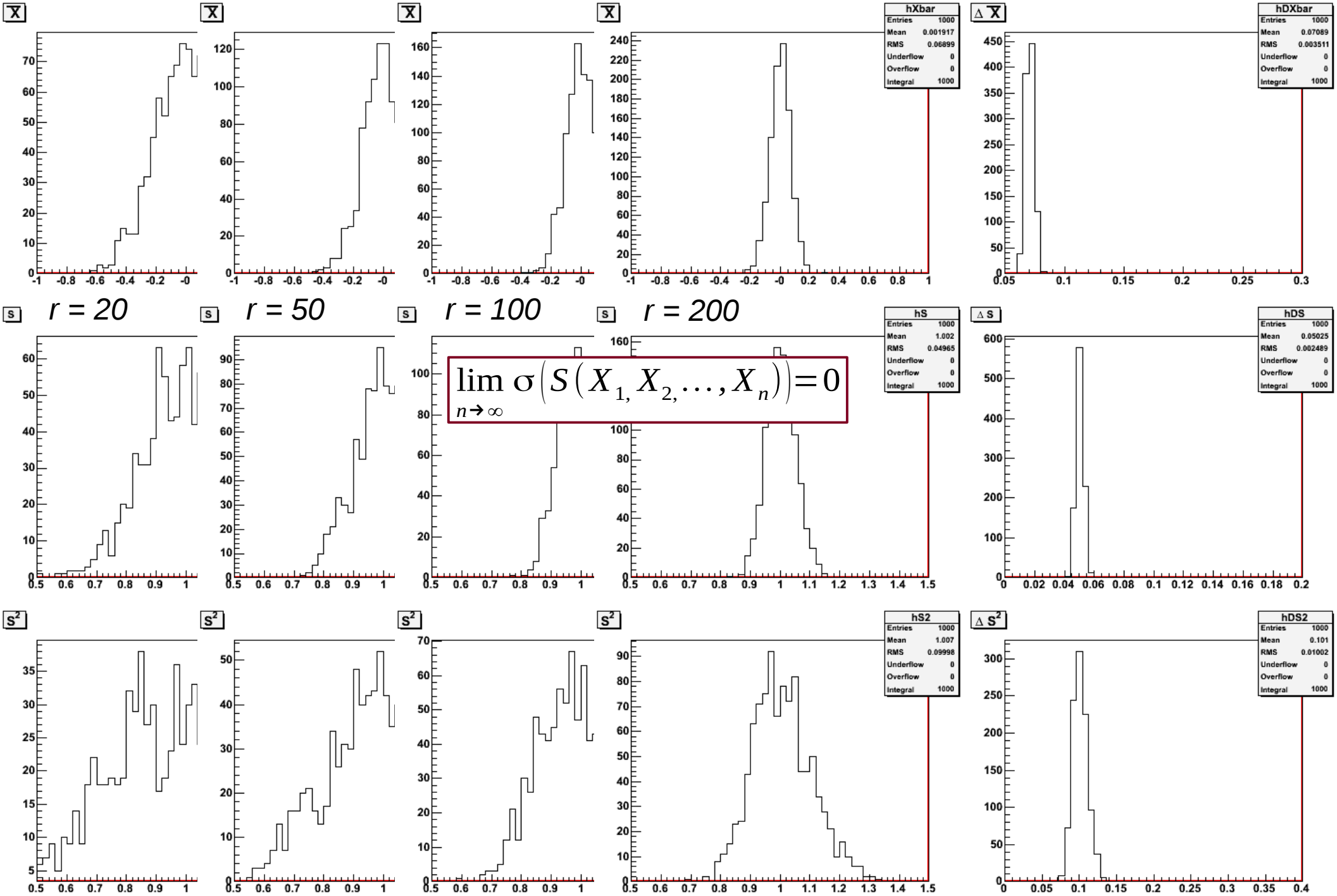




Estymatory - histogramy



Estymatory - histogramy



$$\lim_{n \to \infty} \sigma\left(S\left(X_1, X_2, \ldots, X_n\right)\right) = 0$$

Pobieranie próby z rozkładów cząstkowych

Próby z rozkładów cząstkowych

- Często wygodnie jest podzielić populację G (np. wszystkich studentów w Europie) na t podpolujacji G_i (np. studentów na poszczególnych uniwersytetach)
- Podpopulacje G_i, które są opisane gęstościami prawdopodobieństwa $f_i(x)$ i mają odpowiednie dystrybuanty:

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

- Dla całej populacji mamy zatem:

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

- Możemy wprowadzić oznaczenie: $P(X \in G_i) = p_i$
- Obliczamy wartość średnią populacji G:

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

- **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):

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

- **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: $n=\sum_{n=1}^{t} n_i$. 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:

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

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

Estymatory dla rozkładów cząstkowych

- Zauważmy jednak, że wartość średnia $\overline{X}_p$ 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)

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

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

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

- 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 $\overline{X}_i$, wewnątrz poszczególnych podpopulacji, a potem przez wyrażenie:

$$\widetilde{X} = \sum_{i=1}^{t} p_i \overline{X}_i$$

- estymator wartości średniej z populacji zależeny od wartości średnich z prób

- Wariancja powyższego estymatora:

$$\sigma^2\left(\widetilde{X}\right) = \sum_{i=1}^{t} p_i^2 \sigma^2\left(\overline{X}_i\right) = \sum_{i=1}^{t} \frac{p_i^2}{n_i} \sigma_i^2$$

- 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

Estymatory dla rozkładów cząstkowych

- Gdzie my to wszystko wykorzystujemy i po co?
  - Możemy sobie wyobrazić badania społeczne, gdzie próbujemy wnioskować na temat całej populacji poprzez analizy poszczególnych podgrup (podpopulacji) – np. analizujemy dane na temat wszystkich studentów w Europie poprzez poszczególne analizy studentów poszczególnych typów uczelni (np. osobno techniczne, medyczne, ogólne)
  - Albo rozkład powierzchni zajmowanej przez pewien gatunek trawy

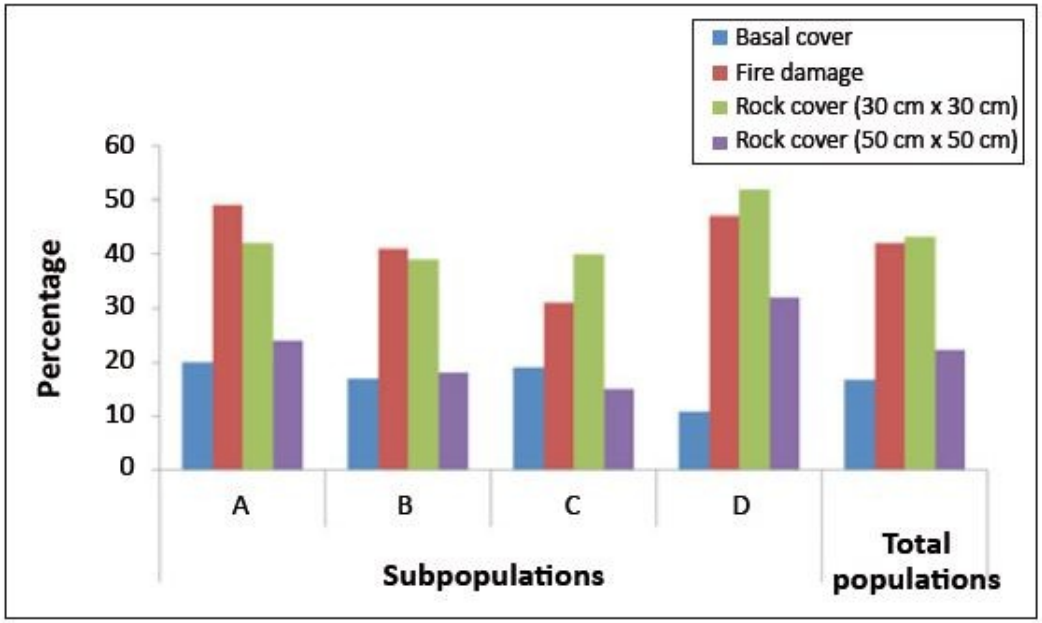


FIGURE 6: Percentage basal grass cover, fire damage and rock cover per subpopulation and total population of *Haworthia koelmaniorum* var. *Mcmurtryi* within the study area.

Stopnie swobody

Stopnie swobody

- Jak pamiętamy, wariancję populacji określamy przez sumę kwadratów różnic:

$$\sigma^2(X)=\frac{1}{N-1}\sum_{j=1}^{N}\left(X_i-\bar{X}\right)^2=\frac{1}{N-1}\left(\sum_{j=1}^{N}X_j^2-\frac{1}{N}\left(\sum_{j=1}^{N}X_j\right)^2\right) \qquad \bar{X}=\frac{1}{N}\sum_{j=1}^{n}X_j$$

- Zajmijmy się kwadratami różnic:

$$\sum_{j=1}^{N}\left(X_i-\bar{X}\right)^2$$

- Ponieważ nie ograniczyliśmy populacji w żaden sposób, wartości X_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 (**więzem**):

$$\sum_{j=1}^{N}\left(X_i-\bar{X}\right)=N\cdot\bar{X}-N\cdot\bar{X}=0$$

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

Stopnie swobody

- Mówimy, że **liczba stopni swobody dla sumy kwadratów wynosi** *N-1*
- Suma kwadratów dzielona przez liczbę stopni swobody to **odchylenie średnie kwadratowe**: $\sigma^2(X) = \frac{1}{N-1} \sum_{j=1}^{N} \left(X_i - \bar{X} \right)^2$
- 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:

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

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

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

wartość średnia pobranej próby (czyli estymator wielkości a :))

- Czyli funkcję charakterystyczną rozkładu normalnego o tej samej średniej ale zmienionej wariancji: $\sigma^2(\bar{X})=\sigma^2(X)/n$
- Rozpatrzmy najprostszy przypadek – rozkład normalny (średnia 0 i odchylenie standardowe 1): $\varphi_{\bar{X}}(t)=\exp\left(-t^2/2n\right)$
- Pobieramy z niego próbę n-elementową i tworzymy sumę kwadratów:

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

- Można udowodnić, że dystrybuanta ma postać:

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

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:

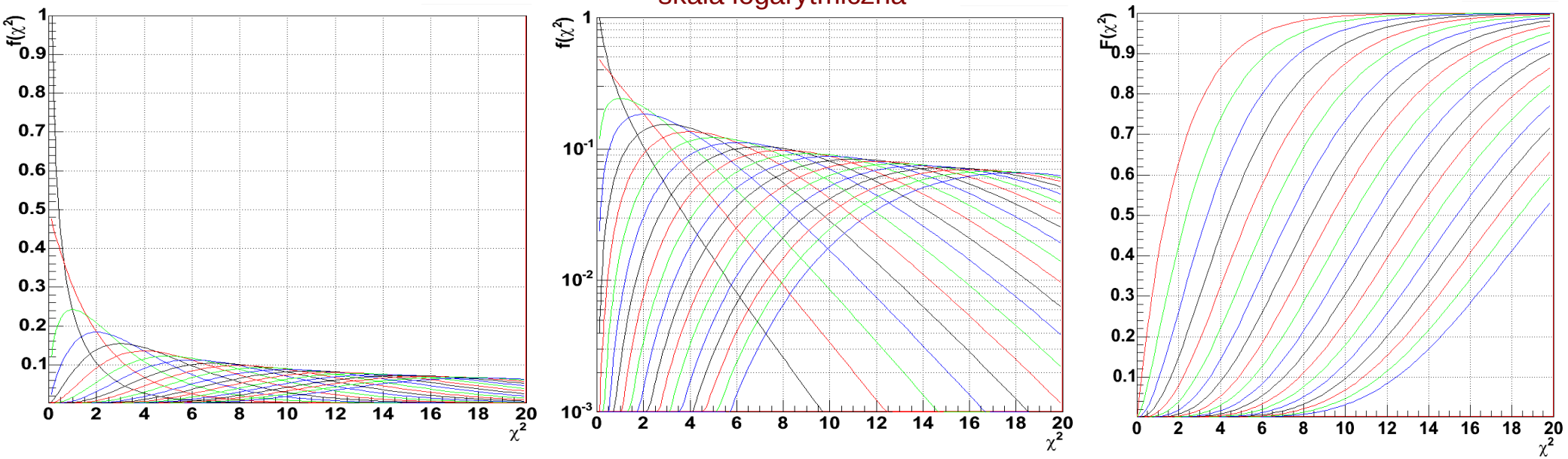
$$E\left\{X^2\right\} = -i\,\varphi_{\chi^2}'(0) = 2\lambda \equiv n$$

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

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

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

skala logarytmiczna



- 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

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

- a w ogólnym przypadku wielowymiarowym, gdy zmienne są zależne:

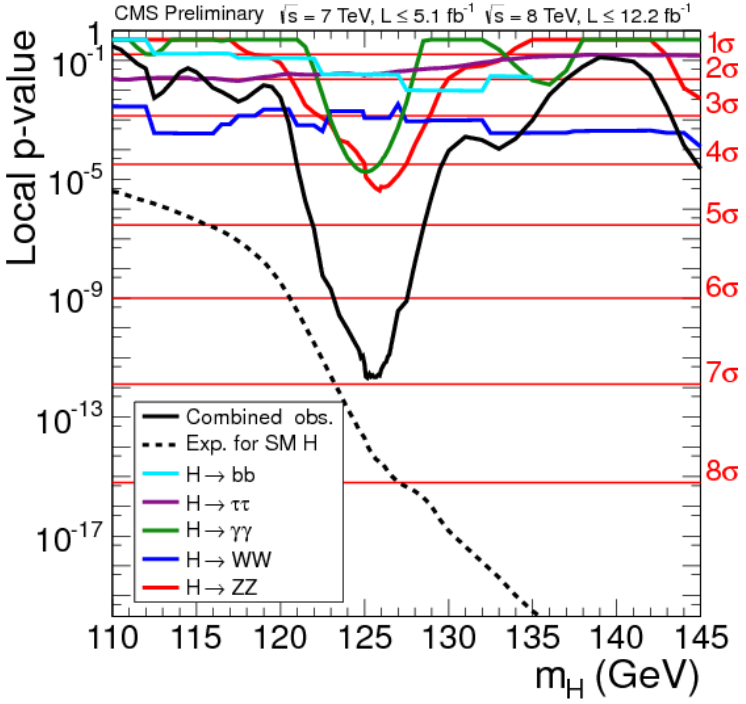
$$X^2 = (\boldsymbol{X} - \boldsymbol{a})^T B (\boldsymbol{X} - \boldsymbol{a})$$

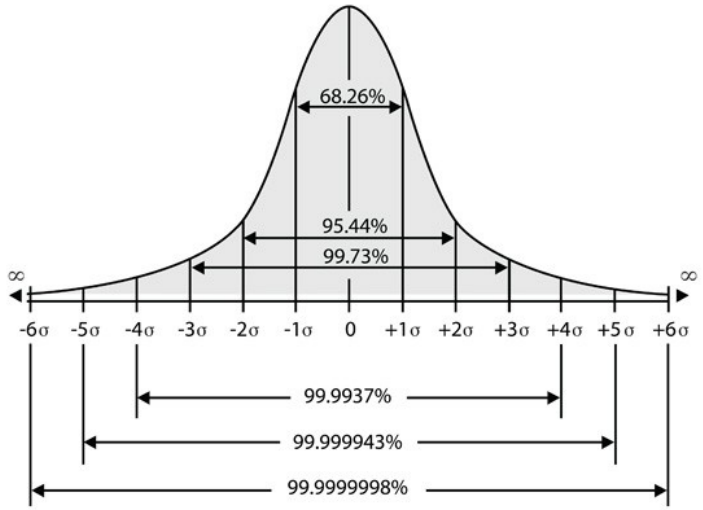
- Nieobciążony i zgodny estymator wariancji z populacji to:

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

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 (więzy) pomiędzy wyrażeniami $(X_i-\bar{X})^2$ redukuje liczbę stopni swobody o 1
- Poziomy ufności – przykład:





$$P(|Y-a|\leqslant\sigma)=68{,}3\% \quad P(|Y-a|>\sigma)=31{,}7\%$$
$$P(|Y-a|\leqslant 2\sigma)=95{,}4\% \quad P(|Y-a|>2\sigma)=4{,}6\%$$
$$P(|Y-a|\leqslant 3\sigma)=99{,}8\% \quad P(|Y-a|>3\sigma)=0{,}2\%$$


WYDZIAŁ FIZYKI
POLITECHNIKA WARSZAWSKA
POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY

KONIEC