

Komputerowa analiza danych doświadczalnych

Wykład 8
12.04.2019

dr inż. Łukasz Graczykowski
lukasz.graczykowski@pw.edu.pl

Semestr letni 2018/2019



Estymatory

Próby z rozkładów cząstkowych

Próby z rozkładu normalnego

Rozkład χ^2



Pobieranie próby, estymatory

Pobieranie próby

- **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

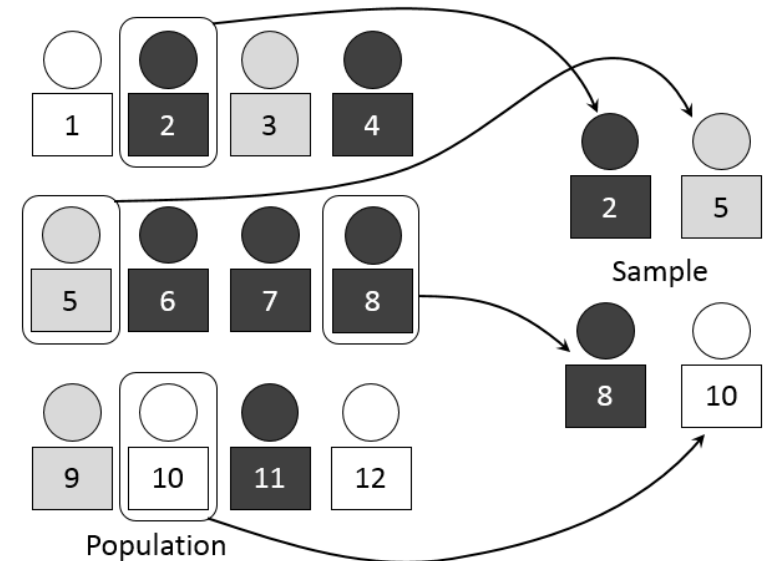
- Właściwości próby losowej:
 - wybór każdego elementu próby jest niezależny
 - każdy element próby jest zmienną losową o takim samym rozkładzie prawdopodobieństwa jak populacja

$$g_1(X_1) = g_2(X_2) = \dots = g_n(X_n) = f(X)$$

czyli

$$E(X_1) = E(X_2) = \dots = E(X_n) = E(X)$$

$$\sigma^2(X_1) = \sigma^2(X_2) = \dots = \sigma^2(X_n) = \sigma^2(X)$$



Estymatory

- Każda funkcja elementów próby nazywana jest **statystyką**
 $S = S(X_1, X_2, \dots, X_n)$
- Poszukujemy takich statystyk (**estymacja**), które pozwolą nam na podstawie próby wnioskować o własnościach (parametrach) całej populacji
- **Estymator** to statystyka, która pozwala nam na wnioskowanie o zadanym parametrze rozkładu populacji
- Oczekujemy, aby dany estymator był:

- **nieobciążony**, jeżeli niezależnie od liczebności próby jego wartość oczekiwana jest równa wartości estymowanego parametru:

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

- **zgodny**, jeżeli jego wariancja znika:

$$\lim_{n \rightarrow \infty} \sigma(S(X_1, X_2, \dots, X_n)) = 0$$

Estymatory

- Zadaniem badań eksperymentalnych jest wyznaczanie nieznanymi parametrów rozkładów, np. opisujących różne prawa fizyczne
- Możemy jedynie otrzymywać przybliżone oszacowania (*estymacje*) tych parametrów poprzez przeprowadzanie eksperymentów (*wybór prób losowych*)
- Parametry badanych rozkładów (praw fizycznych) wyznaczamy przez estymatory
- Jeżeli uda nam się znaleźć dla danego parametru estymator nieobciążony i zgodny, to jesteśmy w stanie określić ten parametr z dowolną dokładnością (tym większą, im większa jest liczebność próby losowej – liczba pomiarów)
- **Nigdy nie jesteśmy w stanie podać w 100% dokładnych wartości parametrów badanego rozkładu (prawa fizycznego)**

Estymator wartości oczekiwanej

Populacja

- opisana funkcją gęstości:

$$f(x) = P(X = x)$$

- posiada **wartość oczekiwaną**:

$$E(X) = \hat{x} = \int_{-\infty}^{\infty} x f(x) dx$$

- wartość oczekiwana rozkładu to **jedna liczba**
 - nie jest zmienną losową
 - chcemy ją zmierzyć doświadczalnie
- np. dla rozkł. Gaussa:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \mu$$

Próba losowa

- zakładamy, że **średnia arytmetyczna** z elementów próby jest estymatorem wartości oczekiwanej

- **średnia arytmetyczna** jest statystyką:

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

- **jest zmienną losową** (zależy od elementów próby)
- posiada swoją **wartość oczekiwaną** oraz **wariancję**
- oczekujemy, że będzie ona *estymatorem nieobciążonym* i *zgodnym* wartości oczekiwanej populacji:

$$E(\bar{X}) = E(X) = \hat{x}, \text{ dla każdego } n$$

$$\lim_{n \rightarrow \infty} \sigma(\bar{X}) = 0$$

- **jak to sprawdzić?**

Estymator wartości oczekiwanej

Próba losowa

- **średnia arytmetyczna** z elem. próby jest statystyką:

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

- **wartość oczekiwana** średniej arytmetycznej:

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n} E(X_1 + X_2 + \dots + X_n) = \frac{1}{n} (E(X_1) + E(X_2) + \dots + E(X_n)) \\ &= \frac{1}{n} n E(X) = E(X) = \hat{x}, \text{ dla każdego } n \end{aligned}$$

Czyli wykazaliśmy, że estymator jest **nieobciążony**

- **wariancja** średniej arytmetycznej:

$$\begin{aligned} \sigma^2(\bar{X}) &= E\{\bar{X} - E(\bar{X})\}^2 = E\left\{\left(\frac{X_1 + X_2 + \dots + X_n}{n} - \hat{x}\right)^2\right\} \\ &= \frac{1}{n^2} E\{[(X_1 - \hat{x}) + (X_2 - \hat{x}) + \dots + (X_n - \hat{x})]^2\} \end{aligned}$$

Musimy pokazać, że to wyrażenie zbiega do 0 dla dużych n

Estymator wartości oczekiwanej

Próba losowa

- **wariancja** średniej arytmetycznej:

$$\begin{aligned}\sigma^2(\bar{X}) &= E\left\{\bar{X} - E(\bar{X})\right\}^2 = E\left\{\left(\frac{X_1 + X_2 + \dots + X_n}{n} - \hat{x}\right)^2\right\} \\ &= \frac{1}{n^2} E\left\{[(X_1 - \hat{x}) + (X_2 - \hat{x}) + \dots + (X_n - \hat{x})]^2\right\}\end{aligned}$$

- ponieważ elementy próby losowej są niezależne:

$$\text{Cov}(X_i, X_j) = E\left\{(X_i - \hat{x})(X_j - \hat{x})\right\} = 0$$

- to upraszczamy nawias:

$$\begin{aligned}E\left\{[(X_1 - \hat{x}) + (X_2 - \hat{x}) + \dots + (X_n - \hat{x})]^2\right\} &= E\left\{(X_1 - \hat{x})^2 + (X_2 - \hat{x})^2 + \dots + (X_n - \hat{x})^2\right\} \\ &= E(X_1 - \hat{x})^2 + E(X_2 - \hat{x})^2 + \dots + E(X_n - \hat{x})^2 = \sigma^2(X_1) + \sigma^2(X_2) + \dots + \sigma^2(X_n) = n\sigma^2(X)\end{aligned}$$

- zatem ostatecznie wariancja średniej arytmetycznej:

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

- ponieważ wariancja rozkładu populacji jest jedną liczbą, to:

$$\lim_{n \rightarrow \infty} \sigma(\bar{X}) = 0$$

Wykazaliśmy, że estymator jest **zgodny**

Estymator wartości oczekiwanej

Populacja

- opisana funkcją gęstości:

$$f(x) = P(X = x)$$

- posiada **wartość oczekiwaną**:

$$E(X) = \hat{x} = \int_{-\infty}^{\infty} x f(x) dx$$

- wartość oczekiwana rozkładu to **jedna liczba**
 - nie jest zmienną losową

- np. dla rozkł. Gaussa:

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \mu$$

Próba losowa

- **średnia arytmetyczna** z elem. próby jest statystyką:

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

- jest **estymatorem nieobc. i zgodnym** wartości oczekiwanej populacji

- **wartość oczekiwana** średniej arytmetycznej:

$$E(\bar{X}) = E(X) = \hat{x}, \text{ dla każdego } n$$

- **wariancja** (niepewność wyznaczenia) średniej arytmetycznej:

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

Ale... żeby wyznaczyć jej wartość musimy znaleźć estymator wariancji populacji $\sigma^2(X)$

Estymatory - podstawowe

- Przykładowe **estymatory nieobciążone**:

- **wartości oczekiwanej populacji** → *średnia arytmetyczna z próby*:

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

- **wariancji populacji** → *średnia odchyleń kwadratowych*:

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

- Wariancje (niepewności) estymatorów:

- **wariancja średniej arytmetycznej**:

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

- **wariancja średniej odchyleń kwadratowych**:

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

- **Uwaga!** Wariancje estymatorów są również estymatorami – możemy więc liczyć np. wariancję wariancji średniej arytmetycznej, itd.

- **Czy te wzory coś przypominają?**

Przykład 1 – pomiar bezpośredni

- Mierzymy przy pomocy suwmiarki bok pręta d o przekroju kwadratowym
- Dokładność suwmiarki (niepewność wzorcowania):

<http://www.kreocen.pl/img/p/1057776/1/TESA-Suwmiarka-STANDARD-005-mm.jpg>
http://www.drut.com.pl/images/com_sobi2/gallery/69/69_image_1_bml.jpg



$$\Delta d = 0,1 \text{ mm}$$

- Seria $n=11$ pomiarów (w mm): 12,5; 12,3; 12,6; 12,5; 12,6; 12,5; 12,4; 12,3; 12,5; 12,4; 12,6

- Wynik (średnia arytmetyczna):

$$\bar{d} = \sum_{i=1}^n d_i = 12,4727 \text{ mm}$$

- Niepewność typu A:

$$u_A(d) = \sqrt{\sum_{i=1}^n \frac{1}{n(n-1)} (d_i - \bar{d})^2} = 0,033278 \text{ mm}$$

- Niepewność typu B:

$$u_B(d) = \frac{\Delta d}{\sqrt{3}} = 0,057735 \text{ mm}$$

- Niepewność całkowita:

$$u(d) = \sqrt{u_A^2(d) + u_B^2(d)} = 0,06639 \text{ mm} \approx 0,066 \text{ mm}$$

- Wynik: $d = 12,473(66) \text{ mm}$

WYKŁAD 1



Stopnie swobody

Stopnie swobody

- Wariancję populacji określamy przez sumę kwadratów różnic:

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

 Dlaczego $1/(n-1)$? To jest średnia różnicy kwadratów, powinno być $1/n$

- Zajmijmy się kwadratami różnic: $\sum_{i=1}^n (X_i - \bar{X})^2$

- Wartości X_i mogą przybierać dowolne wartości
- Dowolność ta jest jednak ograniczona ograniczony warunkiem (**wiązanie**) istnienia wartości średniej:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \Rightarrow \sum_{i=1}^n (X_i - \bar{X}) = 0$$

- Mówimy, że **liczba stopni swobody** dla sumy kwadratów wynosi $n-1$
- Suma kwadratów dzielona przez liczbę stopni swobody to **odchylenie średnie kwadratowe (root-mean square - RMS)**



Pobieranie próby z rozkładów cząstkowych

Próby z rozkładów cząstkowych

- Czasem mamy do czynienia z sytuacją, gdy nie da się wybrać losowo próby z populacji generalnej
- Możemy jednak podzielić populację generalną na **podpopulacje**:
- Przykład:
 - badamy jakąś cechę wszystkich studentów w Europie – populacja generalna
 - najłatwiej to zrobić poprzez badanie tej cechy na poszczególnych uniwersytetach – podpopulacje
 - rozkład podpopulacji **nie jest** taki sam jak rozkład populacji generalnej
 - podpopulacje są jednak związane z całą populacją
- W jaki sposób wnioskować o całej populacji na podstawie prób losowych wybranych z podpopulacji?

Próby z rozkładów cząstkowych

- Dzielimy populację generalną G na t podpopulacji G_i
- Prawdopodobieństwo, że zmienna losowa należy do G_i : $P(X \in G_i) = p_i$
- Podpopulacje G_i , które są opisane gęstościami prawdopodobieństwa $f_i(x)$ i mają odpowiednie dystrybuanty $F_i(x)$:

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

- Dla całej populacji mamy zatem:

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

- Dla gęstości prawdopodobieństwa populacji: $f(x) = \sum_{i=1}^t p_i f_i(x)$

- Obliczamy wartość oczekiwaną populacji G :

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

Próby z rozkładów cząstkowych

- Wartość oczekiwana 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 oczekiwana z populacji to średnia ważona wartości oczekiwanych podpopulacji pomnożonych przez ich prawdopodobieństwa
- Jak zatem wyznaczyć **wariancję populacji** na podstawie podpopulacji?

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

- Wniosek:** wariancja populacji jest średnią ważoną wariancji z podpopulacji σ_i i wariancji wartości średniej podpopulacji $\hat{x}_i$ względem wartości średniej z całej populacji $\hat{x}$
- To nie koniec... Teraz musimy wybrać **próby losowe z podpopulacji i policzyć estymatory**

Próby z rozkładów cząstkowych

- 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 (niepewność) średniej z całej próby:

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

próbki z poszczególnych podpopulacji

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

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

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

- 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

$$p_i = \frac{n_i}{n}$$

- Wariancja powyższego estymatora: $\sigma^2(\tilde{X}) = \sum_{i=1}^t p_i^2 \sigma^2(\bar{X}_i) = \sum_{i=1}^t \frac{p_i^2}{n_i} \sigma_i^2$
- Oczywiście, chcielibyśmy tak dobierać próbki, by była ona jak najmniejsza: $n_i = n p_i \sigma_i / \sum_{i=1}^t p_i \sigma_i$

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
 - Czy przeprowadzamy badania kliniczne leku w wielu ośrodkach w różnych krajach

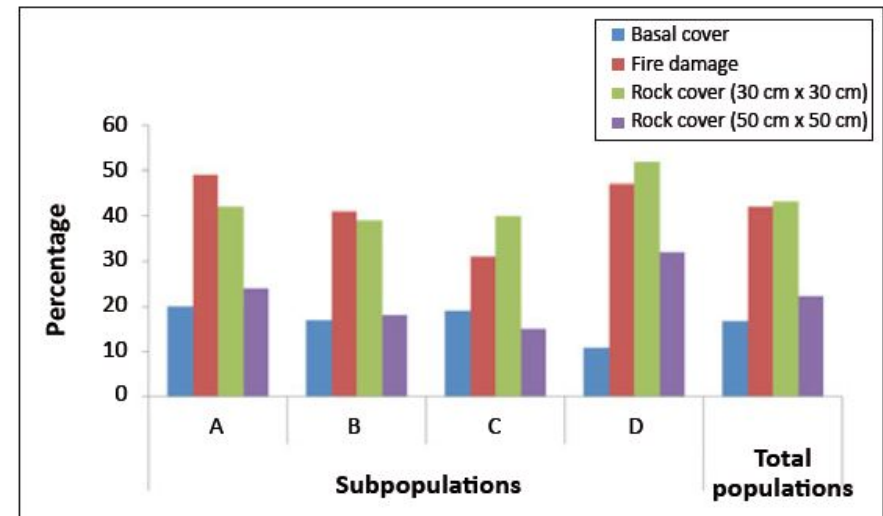


FIGURE 6: Percentage basal grass cover, fire damage and rock cover per subpopulation and total population of *Haworthia koelmaniorum* var. *Mcmurtryi* within the study area.



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:

$$\phi_X(t) = \exp(it a) \exp(-\sigma^2 t^2 / 2)$$

- Oraz funkcję char. dla wartości średniej (patrz Brandt):

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

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

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

wartość średnia pobranej próby (czyli estymator wielkości a :)

- Czyli funkcję charakterystyczną rozkładu normalnego **zmienionej wariancji**: $\sigma^2(\bar{X}) = \sigma^2(X)/n$

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

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

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

Pobieranie próby z rozkładu normalnego

- Można udowodnić, że dystrybuanta ma postać:

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

– **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:

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

- 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

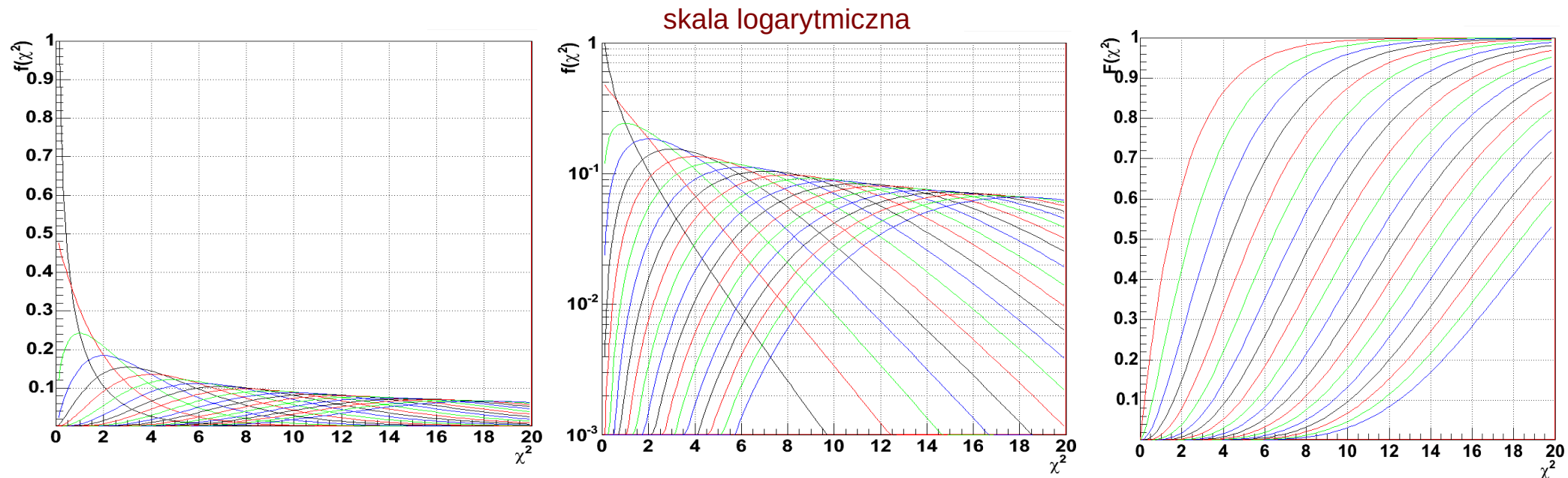
- Różniczkując funkcję charakterystyczną możemy wyznaczyć wartość oczekiwaną i wariancję rozkładu χ^2 :

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

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

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

- 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 % ilości odchyłeń standardowych rozkładu normalnego σ)

– wielkość α jest nazywana **poziomem istotności**

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

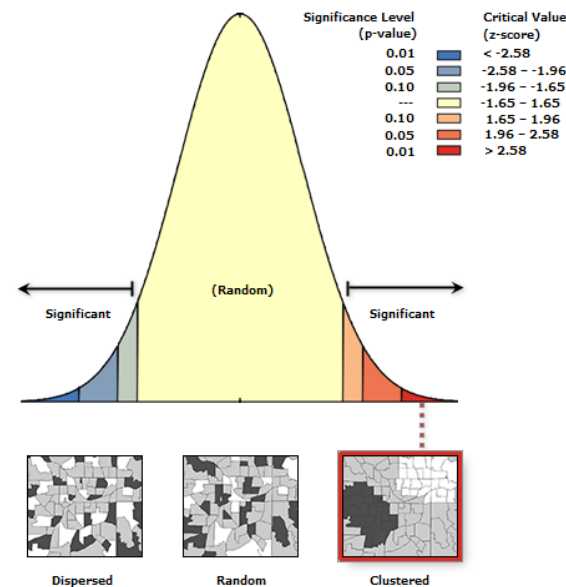
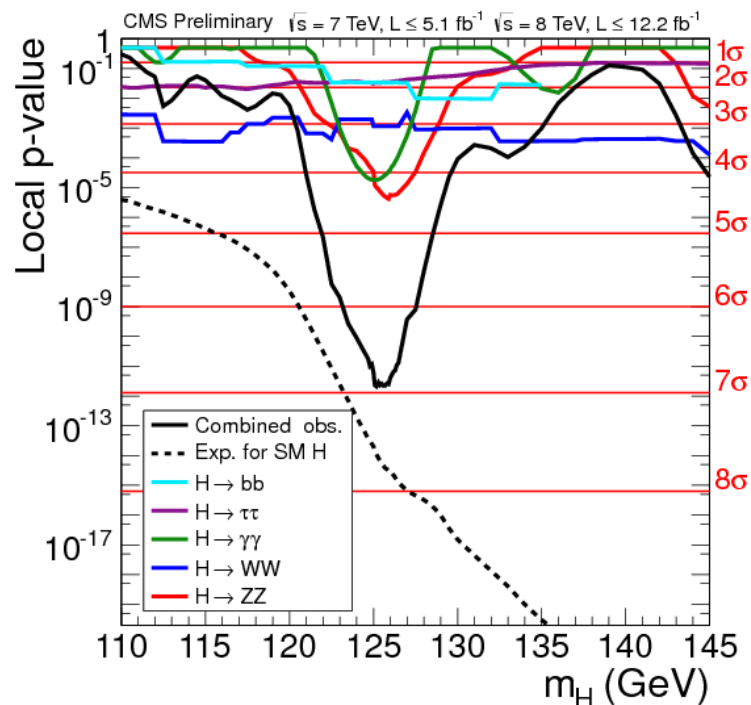
$$\chi^2 = X^2 = \frac{(X_1 - a)^2 + (X_2 - a)^2 + \dots + (X_n - a)^2}{\sigma^2}$$

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

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

Rozkład χ^2 a estymator wariancji

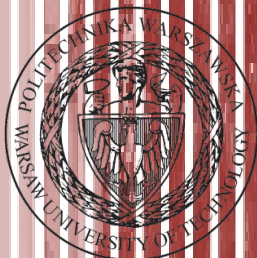
- Można udowodnić, że zmienna losowa: $\frac{n-1}{\sigma^2} S^2$ ← estymator wariancji
- ma rozkład χ^2 z $f=n-1$ stopniami swobody. Wynika to stąd, że wyrażenia $(X_i - \bar{X})^2$ nie są liniowo niezależne (co już wiemy). Każde dodatkowe równanie (więzy) pomiędzy wyrażeniami $(X_i - \bar{X})^2$ redukuje 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