Powtórzenie do kolokwium

Zakres materiału:

- Rozkład dwumianowy
- Rozkład normalny
- Przedziały ufności dla frakcji oraz średniej przy nieznanej wariancji
- Testy istotności: frakcji, próbkowy i dwupróbkowy test t-Studenta
- Test niezależności chi-kwadrat
- Regresja liniowa i zależność cech: współczynnik korelacji Pearsona i Spermana, prosta regresji.

Rozkład dwumianowy

- Rozkład prawdopodobieństwa
- Wykres słupkowy
- Dystrybuanta
- Parametry rozkładu
- Wartość oczekiwana (przeciętna, średnia)
- Wariancja
- Przybliżenie rozkładem normalnym

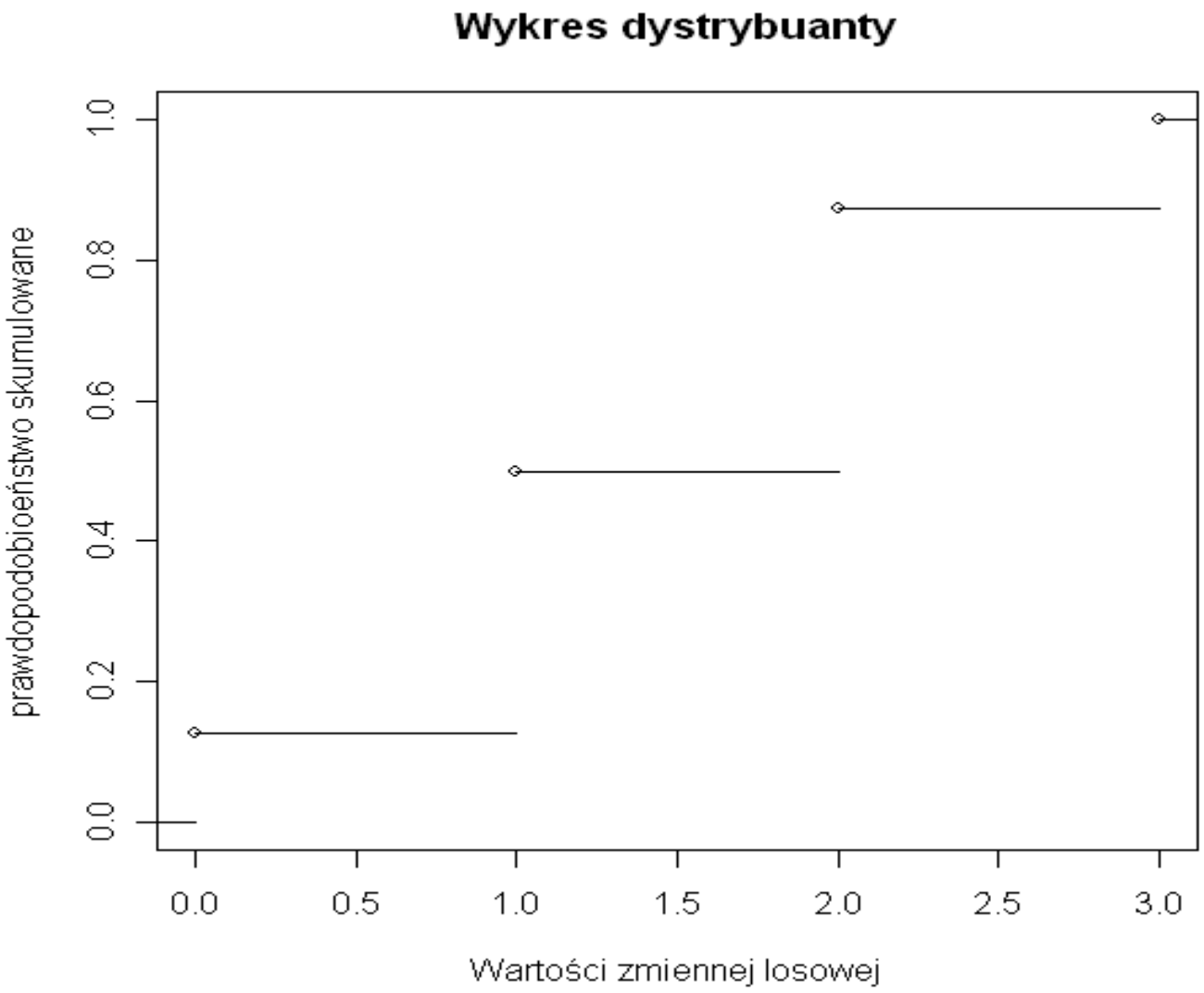
Rozkład dwumianowy-przykład

Obliczyć wartość przeciętną (oczekiwaną) zmiennej losowej dyskretnej X oznaczającej liczbę orłów w 3 niezależnych rzutach monetą.

Rozkład prawdopodobieństwa

Mamy 3 próby Bernoulliego z p=0.5
P(X=0)=choose(3,0)*0.5^0*0.5^3
#[1] 0.125
P(X=1)=choose(3,1)*0.5^1*0.5^2
#[1] 0.375
P(X=2)=choose(3,2)*0.5^2*0.5^1
#[1] 0.375
P(X=3)=choose(3,3)*0.5^3*0.5^0
#[1] 0.125

Dystrybuanta



Dystrybuanta

Narysuj dystrybuantę zmiennej X (liczby orłów w 3 rzutach)

Aby narysować dystrybuantę użyj następujących komend:

```
x=c(0,1,2,3)
y=c(0, 0.125, 0.5, 0.875, 1)
a=stepfun(x,y)
plot.stepfun(a,verticals=FALSE,main="Wykres dystrybuanty")
```

Wartość oczekiwana i wariancja

$EX=0*0.125+1*0.375+2*0.375+3*0.125$

EX

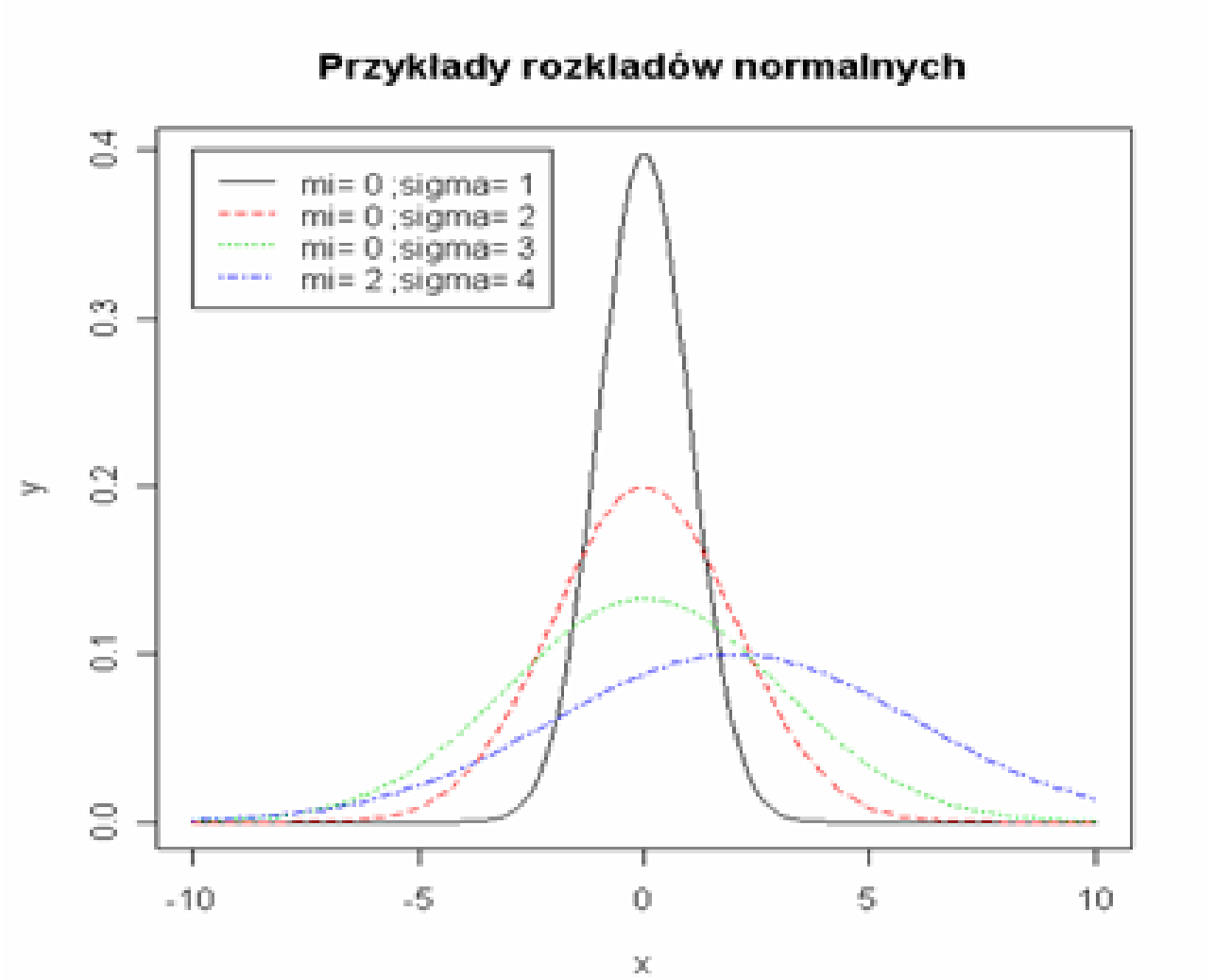
1.5

$EX=n*p=3*0.5=1.5$

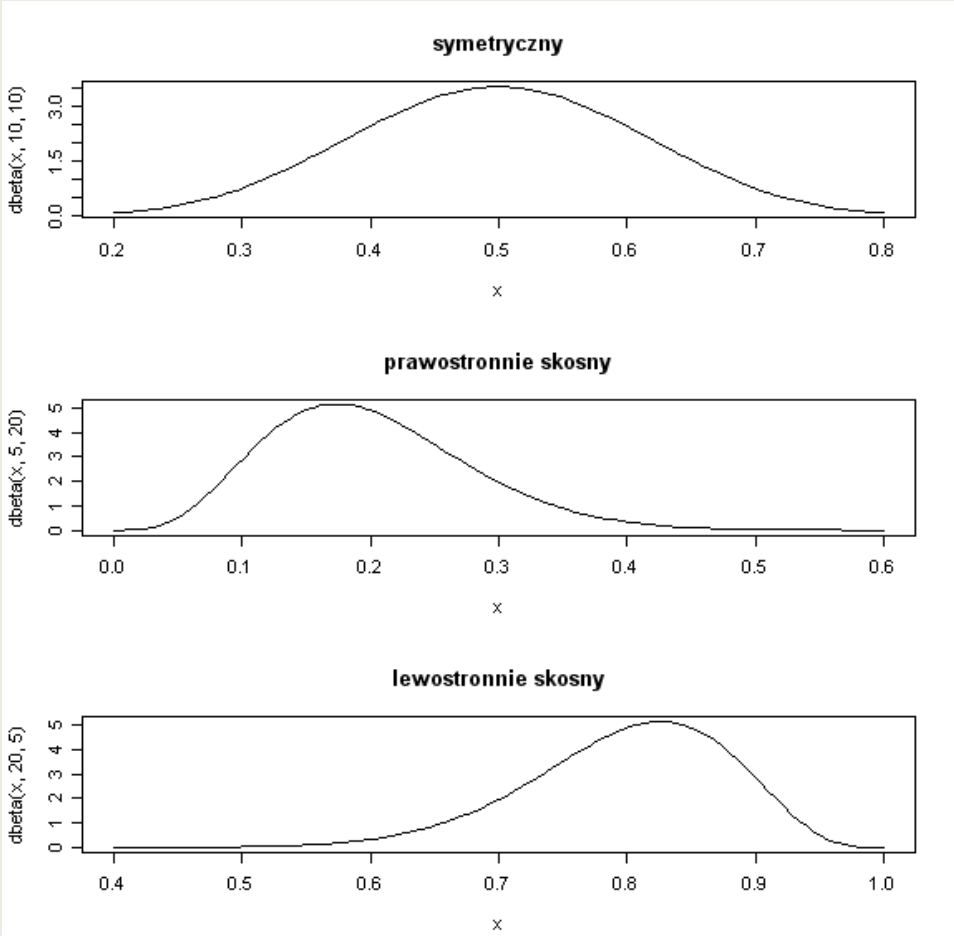
$Var(X)= n*p*(1-p)=3*0.5*0.5=0.75$

Rozkład normalny

- Typ zmiennej losowej
- Rozkład prawdopodobieństwa
- Parametry rozkładu
- Przykłady rozkładów normalnych
- Standaryzacja
- Odczyt z tablic


Przyklady rozkladów normalnych
mi= 0 ;sigma= 1
mi= 0 ;sigma= 2
mi= 0 ;sigma= 3
mi= 2 ;sigma= 4
y
0.0
0.1
0.2
0.3
0.4
-10
-5
0
5
10
x

Rozkłady prawdopodobieństwa



Rozkład normalny-przykład

Podaj w przybliżeniu ile osób z grupy 200 osób waży od 65 do 75 kilogramów, jeśli wiadomo, że rozkład wagi tej grupy osób ma rozkład normalny o średniej 70 kg i odchyleniu 5 kg.

Rozwiązanie:

X-waga losowo wybranej osoby

$P(65<X<75)=P(X<75)-P(X<65)$

Dokonujemy standaryzacji zmiennej

$Z=(X-70)/5$

czyli

$P(65<X<75)=P(Z<(75-70)/5)-P(Z<(65-70)/5)$

$=P(Z<1)-P(Z<-1)=2P(Z<1)-1=0.68$

Stąd w przybliżeniu około $200*0.68=136$ osób z tej grupy waży między 65 a 75 kg.

Przybliżenie rozkładu dwumianowego rozkładem normalnym

- Przybliżony przedział ufności dla frakcji elementów wyróżnionych w populacji
- Test dla frakcji elementów wyróżnionych w populacji

Przedział ufności dla frakcji wyróżnionej w populacji- przykład

Podaj 95% przedział ufności dla frakcji osób chorych na ropniaka płuc wśród wszystkich chorych na choroby płuc, jeśli na podstawie próby reprezentacyjnej 1000 osób ustalono, że na ropniaka płuc chorowało 10% pacjentów poradni chorób płuc.

Rozwiązanie:

Korzystamy ze wzoru:

$$\left[\hat{p} - \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} * z, \hat{p} + \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} * z \right]$$

k=100 # ilość osób z ropniakiem wśród 1000 chorych poradni chorób płuc.

n=1000 # rozmiar próby

$$\hat{p} = k / n$$

= 100/1000=0.1 # oszacowana częstość występowania ropniaka wśród chorób płuc

z=z(1-alfa/2) # kwantyl odpowiedniego rzędu z rozkładu normalnego N(0,1), alfa=0.05,
z=z(0.975)=1.96
n=1000, phat=0.1, z=1.96

Podstawiając do wzoru otrzymujemy:

L=phat-sqrt(phat*(1-phat)/n)*z= 0.0814058
P=phat+sqrt(phat*(1-phat)/n)*z= 0.1185942

Stąd z 95 % ufnością możemy ocenić, że chorzy na ropniaka płuc stanowią od 8.14% do 11.86% wszystkich chorób płuc.

Przedział ufności dla średniej w populacji normalnej o nieznanej wariancji

Czas wykonania pewnej analizy możemy traktować jako zmienną losową o rozkładzie normalnym. Podać 90% przedział ufności dla średniego czasu pewnej analizy na podstawie poniższej próby (w sek.)

x=c(10.3, 15.1, 13.8, 16.4, 13, 15.2, 14.8, 16.4, 16.1, 15.1)

Korzystamy ze wzoru:

$$\left[\overline{X} - t_{tab}(\alpha; n-1)\frac{S}{\sqrt{n}}, \overline{X} + t_{tab}(\alpha; n-1)\frac{S}{\sqrt{n}} \right]$$

```
n=10
x=c(10.3, 15.1, 13.8, 16.4, 13, 15.2, 14.8, 16.4,
   16.1, 15.1)
 srednia=mean(x)=14.62
odchylenie=sd(x)=1.86
alfa=0.1
t=qt(1-alfa/2,n-1)=1.83 # odpowiedni kwantyl
          z rozkładu t-Studenta o n-1 stopniach
          swobody
```

Obliczamy końce przedziału:

- L=mean(x)-t*odchylenie/sqrt(n)=13.54
- P=mean(x)+t*odchylenie/sqrt(n)=15.70

Z 90% ufnością możemy twierdzić, że średni czas analizy wynosi od 13.54 do 15.70 sekund.

Testowanie hipotez statystycznych

- Test dla frakcji (proporcji, odsetka) elementów wyróżnionych w populacji
- Jednopróbkowy test t-Studenta
- Test t-Studenta dla par powiązanych
- Test t-Studenta dla par niepowiązanych (jednakowe wariancje)
- Test niezależności chi-kwadrat

Test dla frakcji - przykład

Czy można twierdzić (na poziomie istotności 0.05) na podstawie danych o ropniaku płuc, że występuje u 10% pacjentów chorych na płuca?

Rozwiązanie:

Dane: n=1000, k=100, phat=k/n=0.1

Weryfikujemy H_0: p=0.1 przeciw H_1: p jest różne od 0.1

Statystyka testowa jest postaci

$$Z = \sqrt{n}\,\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}$$

gdzie p_0 jest hipotetyczną wielkością frakcji, u nas wynosi 0.1

U nas Z=0, alfa=0.05, więc odpowiedni kwantyl z rozkładu normalnego wynosi z(1-alfa/2)=1.96, zbiorem krytycznym dla H_0 jest suma przedziałów:

$$(-\infty,-1.96)\cup(1.96,+\infty)$$

Nie ma podstaw do odrzucenia H_0. Można twierdzić, że chorzy na ropniaka stanowią 10% populacji chorych na płuca.

Jednopróbkowy test t-Studenta

Zważono 81 chomików uzyskując następujące wyniki (w gramach): średnia 54 oraz odchylenie standardowe 15.4. Czy na poziomie istotności 0.05 można twierdzić, że średnia waga chomika wynosi więcej od 50 gram?

Rozwiązanie:

Testujemy hipotezę H_0: $\mu = 50$

przeciw hipotezie H_1: $\mu > 50$

Statystyka testowa jest postaci

$$T = \sqrt{n}\,\frac{\overline{X} - 50}{S}$$

gdzie S=15, $\overline{X} = 54$

Stąd T=2.4

Odpowiedni kwantyl z rozkładu t-Studenta o n-1 stopniach swobody wynosi t=1.66.

Zbiór krytyczny dla H_0 jest postaci:

$$[t(2\alpha, n-1), +\infty) = [1.66, +\infty)$$

odrzucamy hipotezę H_0 (na poziomie istotności 0.05)

waga chomików wynosi więcej niż 50 gram.

Test t-Studenta dla par powiązanych

Badano skuteczność diety odchudzającej na 7 pacjentkach. Wyniki wagi ciała (w kg) były następujące:

Przed dietą: 78, 84, 68, 74, 94, 78, 79

Po diecie: 73, 75, 68, 70, 92, 80, 68

Czy na poziomie istotności 0.05 można sądzić, że dieta była skuteczna? Przyjąć, że waga ciała ma rozkład normalny.

Rozwiązanie:

Weryfikujemy hipotezę H_0: $\mu_1 = \mu_2$

wobec hipotezy H_1: $\mu_1 > \mu_2$

x=c(78, 84, 68, 74, 94, 78, 79)

y=c(73, 75, 68, 70, 92, 80, 68)

d=x-y

5 9 0 4 2 -2 11

mean(d)=4.14 , sd(d)=4.67

n=7

Statystyka testowa ma postać:

T=sqrt(n)*mean(d)/sd(d)=2.35

Odrzucamy H_0 gdy *T*>*t*, gdzie *t*=*t*(2alfa,n-1) # odpowiedni kwantyl. Z tablic *t*=qt(0.95,6)=1.94, więc *T*>*t* czyli należy odrzucić H0. Można uznać, że dieta jest skuteczna.

Test t-Studenta dla par niepowiązanych (jednakowe wariancje)

Na terenie Puszczy Niepołomickiej odłowiono po 9 samców i 9 samic nornicy rudej. Po przeniesieniu do laboratorium u każdego osobnika zmierzono masę ciała:

samce: 35 30 26 29 22 31 25 19 31

samice: 21 27 18 24 21 23 34 16 28

Czy samce różniły się od samic średnią masą ciała (przy poziomie istotności 0,05)?

Rozwiązanie:

Weryfikujemy hipotezy: H_0: $\mu_1 = \mu_2$

wobec hipotezy H_1: $\mu_1 \neq \mu_2$

Statystyka testowa jest postaci:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

x=c(35, 30, 26, 29, 22, 31, 25, 19, 31)

y=c(21, 27, 18, 24, 21, 23, 34, 16, 28)

Odrzucamy H_0 gdy $|T| > t$

gdzie *t* jest kwantylem rzędu 1-alfa/2 z rozkładu t-Studenta o n_1+n_2-2 stopniach swobody.

$n_1=n_2=9$

mean(x)=27.5

mean(y)=23.5

sd(x)=5

sd(y)=5.05

T =(27.5-23.5)/sqrt((8*5^2+8*5.5^2)*(1/9+1/9)/(9+9-2))= 1.614

t=qt(0.975,16)=2.12, więc nie ma podstaw do odrzucenia H_0. Samice i samce średnio tyle samo ważą.

Przykład: chi-kwadrat niezależności

H_0: brak zależności między cechami *X* i *Y*

przeciw hipotezie alternatywnej

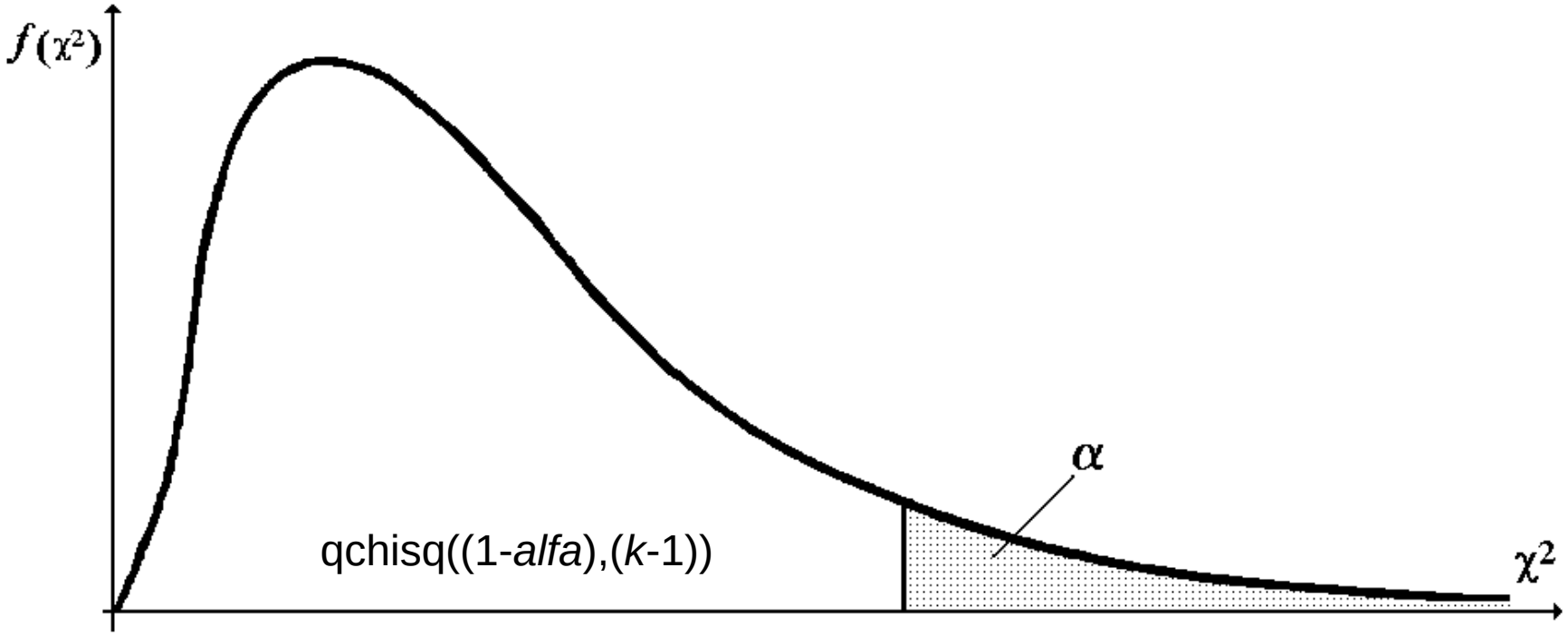
H_1: cechy są zależne

Statystyka testowa jest postaci:

Przykład: chi-kwadrat niezależności

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - n_{i\bullet} n_{\bullet j} / n\right)^2}{n_{i\bullet} n_{\bullet j} / n}$$

Rozkład chi-kwadrat



Wartości statystyki > qchisq(0.95,(*k*-1)*(*l*-1)) świadczą przeciw Ho

Przykład

W trzech szpitalach zastosowano nowa metodę leczenia pewnej choroby.

W szpitalu A na n_1=100 leczonych zaobserwowano 80 przypadków poprawy,

w szpitalu B na n_2=50 leczonych - 30 przypadków poprawy, a

w szpitalu C na n_3=80 leczonych - 60.

Czy szansa wyleczenia zależy od szpitala?

Przyjąć poziom istotności równy 0.05.

	A	B	C	suma
Brak_p	20	20	20	60
popr	80	30	60	170
suma	100	50	80	230

Wartości obserwowane (empiryczne)

obliczamy wartości brzegowe

obliczamy wartości oczekiwane

Brak_p	26,08696	13,04348	20,86957
popr	73,91304	36,95652	59,13043

60*100/230
26.08696

kwadraty reszt jako składniki Chi kwadrat

((20-26.08696)^2)/26.08696
1.420291

Brak_p	1,42029	3,710145	0,036232
popr	0,501279	1,309463	0,012788

chi-kwadrat obliczona statystyka	6,990196

Suma kwadratów reszt

chi-kwadrat z tablic 5.99

zbiór krytyczny [5.99,+niesk)

df= (w-1)*(k-1)=(2-1)*(3-1)=2

alfa=0.05

Ho: wiersze i kolumny niezależne (nie ma zalenożności miedzy stanem pacjenta a szpitalami)

Decyzja: statystyka obliczona 6.99 wpada do zbioru krytycznego [5.99,+niesk) odrzucamy Ho na korzyść H1: wykryto zależność (miedzy szpitalami a stanem zdrowia pacjenta) na poziomie istotności 0.05

Pearson's Chi-squared test

data: rbind(niepopr, popr)
X-squared = 6.9902, df = 2, p-value = 0.03035 p-value <0.05

Decyzja: p-value <0.05 odrzucamy H_0

W przypadku, gdy test niezależności chi kwadrat odrzuci hipotezę o niezależności cech o sile i kierunku zależności między cechami możemy dowiedzieć z współczynników korelacji.

Własności i interpretacja współczynników korelacji Spearmana i Pearsona.

Kiedy je stosujemy.

Współczynnik korelacji Spermana

Ustalić natężenie współzależności między opiniami o nauczycielach dyrektora szkoły i wizytatora. Opinie te zostały wydane na podstawie kontroli całokształtu pracy zawodowej i kwalifikacji nauczycieli. Wyniki kontroli ujęto w punktach:

Dyrektora: 81, 65, 75, 73, 65, 87, 78, 93, 83, 75

Wizytatora: 78, 64, 74, 69, 67, 87, 83, 92, 79, 71

Rozwiązanie:

```
x=c(81, 65, 75, 73, 65, 87, 78, 93, 83, 75)
y=c(78, 64, 74, 69, 67, 87, 83, 92, 79, 71)
rank(x)
7.0  1.5  4.5  3.0  1.5  9.0  6.0 10.0  8.0  4.5
rank(y)
 6  1  5  3  2  9  8 10  7  4
d=rank(x)-rank(y)
 1.0  0.5 -0.5  0.0 -0.5  0.0 -2.0  0.0  1.0  0.5
```

d^2

1.00 0.25 0.25 0.00 0.25 0.00 4.00 0.00 1.00 0.25

n=10

rsperm=1-6*sum(d^2)/(n*(n^2-1))=0.96

Otrzymany wynik wskazuje, że współzależność opinii dyrektora i wizytatora jest bardzo silna. Oceniając nauczycieli, zarówno dyrektor, jak i wizytator kierowali się podobnymi kryteriami.

Współczynnik korelacji Pearsona

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2\right)}}$$

Współczynnik korelacji Pearsona

Wartość ***r*** zawiera się zawsze w przedziale <-1,+1>,
Pozwala ocenić kierunek i siłę współzależności liniowej między dwiema cechami,
Kierunek współzależności:
***r* >0** wtedy większej wartości jednej cechy odpowiada większa wartość drugiej. Mówi się, że cechy korelują dodatnio.
***r*<0** wtedy większej wartości jednej cechy odpowiada mniejsza wartość drugiej. Mówi się, że cechy korelują ujemnie.

Siła współzależności:
Im bardziej **| *r* |** różni się od zera tym większa jest współzależność liniowa badanych cech.
Wartość ***r* = 0** oznacza brak zależności liniowej między cechami.
Wartość ***r* = 1** lub ***r* = -1** oznacza, że między cechami zachodzi zależność liniowa.
Każdą wartość jednej cechy można obliczyć mając wartość drugiej cechy – wg równania ***y=a+bx*.**
r jest miarą przybliżenia wykresu punktów indywidualnych linią prostą (im bardziej **| *r* |** jest bliskie 1 tym bardziej wykres punktów indywidualnych jest bliski linii prostej.

Regresja liniowa

Badano zależność między dawką preparatu X (w mg) a poziomem stężenia we krwi pewnego hormonu Y (w mg). Otrzymano następujące wyniki:

x=c(10, 15, 20, 25, 30, 35, 40)

y=c(5, 7, 6, 11, 15, 14, 20)

Oblicz współczynnik korelacji Pearsona i podaj prostą regresji y=a+bx

n=7

w1=sum(x*y)-n*mean(x)*mean(y)

w2=sqrt(sum(x^2)-n*mean(x)^2)*

sqrt*(sum(y^2)-n*mean(y)^2)

r=w1/w2=0.95 # bardzo silna korelacja dodatnia

y=a+bx

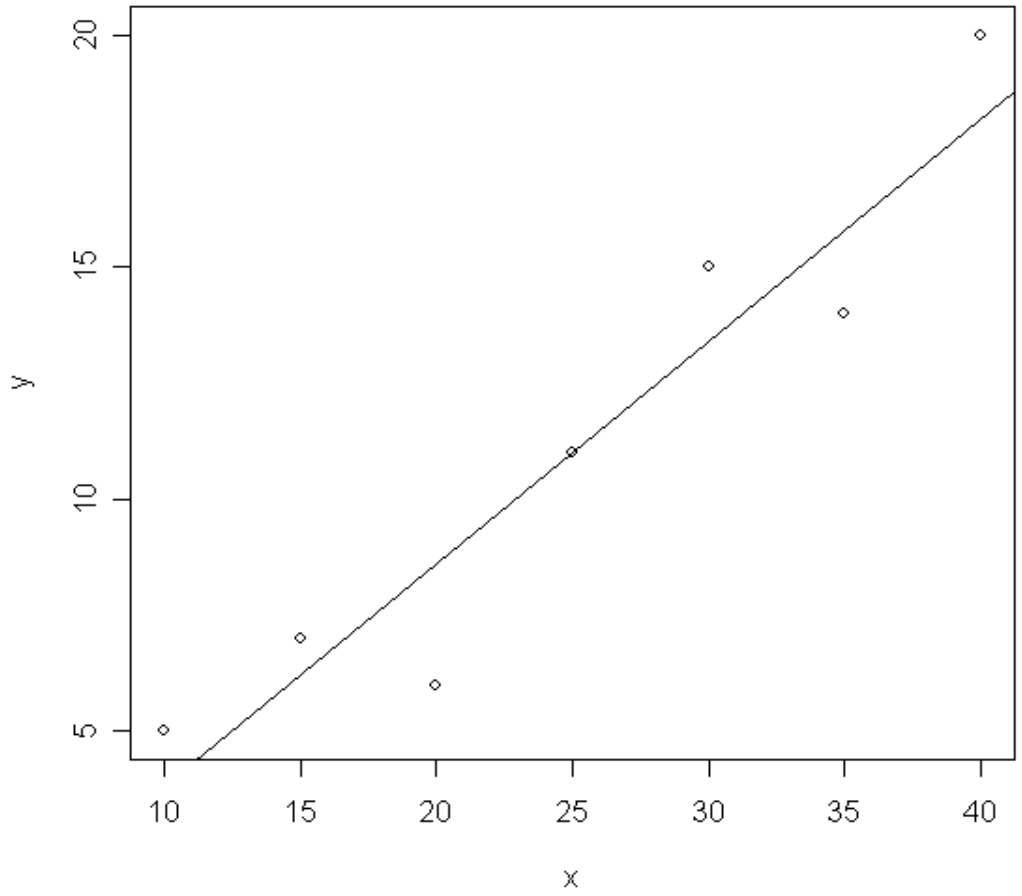
r=0.95

b=r*sd(y)/sd(x)=0.48

a=mean(y)-b*mean(x)=-1

Prosta regresji: y=-1+0.48*x

Prosta regresji i wykres rozproszenia danych



Oszacuj jakiego poziomu hormonu Y możemy się spodziewać stosując 22 mg substancji X.

Rozwiązanie:

$y^* = -1 + 0.48 * 22 = 9.56$

Odp. 9.56 mg.