Zadanie1: Objaśnij co robią poniższe polecenia w R

- rnorm(4, 2, 3)
- dnorm(4, 2, 3)
- pnorm(4, 2, 3)
- qnorm(0.4, 2, 3)

Wstaw ?? aby zachodziła równość że 0.2 = pnorm(qnorm(??, ??, ??), 2, 3)

Zadanie2: Rozważamy rozkłady $Unif(\{1,3\})$ oraz $N(2,1)$. Losujemy jedną wartość z jednego z tych rozkładów. Czy można stwierdzić z którego rozkładu było losowanie? Odpowiedź uzasadnij.

- 2.871
- 3.014

Zadanie3: Wyznacz wartość oczekiwaną, wariancję i odchylenie dla zmiennej losowej, której rozkład opisuje następująca tabela:

X	7	6	0
p	0.2	0.1	0.7

Zadanie4: Kierując się kryterium wartości oczekiwanej, ile złotych byłbyś w stanie położyć aby zagrać w następującą grę:

Rzucamy jednocześnie dwoma kostkami. Jako wynik przyjmujemy $a * b + b$, gdzie a i b to wyniki z pierwszej i drugiej kostki odpowiednio. Ten wynik jest następnie nam wypłacany. Przykładowo, jeżeli wypadnie 4 i 5 to otrzymujemy 4*5+5 = 25 zł.

Zadanie5: W drugiej turze wyborów prezydenckich, w dniu wyborów, po wyjściu z lokali wyborczych zapytano głosujących o poparcie w stosunku do kandydatów A i B. Otrzymano 600 głosów za A oraz 580 głosów za B; część osób nie udzieliło odowiedzi. Z innych, wcześniejszych badań wiadomo, że 10 % osób popierających A niechętnie odpowiada na ankiety; w przypadku osób popierających B jest to 15 %. Na podstawie przeprowadzonej sondy odpowiedz na pytania:

- Ile osób nie odpowiedziało na ankietę? W dalszej części używaj liczebności skorygowanych.
- Kto najprawdopodobniej wygra wybory? O ile punktów procentowych? O ile procent?
- Jakie jest prawdopodobieństwo, że kandydat A uzyska ponad 51 % wszystkich głosów?
- Jaki jest przedział ufności dla A na poziomie 95 %? Co można zrobić żeby niepewność była mniejsza?

Napisz na jakie aspekty muszą zwrócić szczególną uwagę osoby przeprowadzające tego typu ankiety.

Zadanie6: Zbudowano model $Y = 6X + 3 + \epsilon$, gdzie $\epsilon \sim N(\mu = 0, \sigma = 1.2)$. Zinterpretuj wyestymowane współczynniki, a następnie wyznacz:

- $E(Y|X = 2) =$
- $P(Y > 12|X = 2) =$
- takie a, że $P(13 - a \leq Y|X = 2 \leq 13 + a) \approx 0.95$ (Wskazówka: można skorzystać z reguły dwóch sigm).

Jak zmieni się wartość Y jeżeli wartość cechy X wzrośnie o 4?

Zadnie7. Operon laktozowy u bakterii E. coli odpowiada za produkcję enzymów umożliwiających rozkład laktozy. Ekspresja tego operonu jest regulowana przez obecność dwóch cukrów:

- Laktoza – induktor: wyłącza represor i umożliwia transkrypcję.
- Glukoza – preferowane źródło energii: jej obecność obniża poziom cAMP, co uniemożliwia aktywację operonu przez kompleks cAMP-CRP.

Zbudowano model liniowy określający poziom ekspresji operonu laktozowego (Y) w zależności od obecności laktozy (L) i glukozy (G).

$$Y = 0.1 + 1.5L + 2C + 4(L * C)$$

gdzie $C = 1 - G$. Zmienne L oraz G przyjmują poziomy 0 (brak) lub 1 (obecność).

- Zinterpretuj zbudowany model (interpretacja statystyczna i przyczynowa)
- Wyznacz wartość Y przy każdej kombinacji zmiennych objaśniających. Kiedy operon laktozowy wykazuje najwyższą ekspresję? Ile ona wtedy wynosi?

Zadanie8: Plik tekstowy zawiera cztery kolumny x_1-zmienna numeryczna ciągła, x_2-zmienna numeryczna ciągła, x_3-zmienna numeryczna przyjmująca wartości ze zbioru {3,4,5} oraz y.

Objaśnij następujące modele:

- lm(y~x1)
- lm(y~0+x1)
- lm(y~x1+x2)
- lm(y~x1*x2)
- lm(y~I(x1*x2))
- lm(y~x1+factor(x3))
- lm(y~I(x1^2)+x2)
- lm(y~.)
- lm(y~1)

Zadanie9: Oceń prawdziwość bądź fałszywość poniższych sformułowań. Uzasadnij odpowiedź.

- Jeżeli buduje 2 modele liniowe na tym samym zbiorze danych, to kierując się R^2 ten model będzie lepszy, który będzie posiadał większą wartość R^2.
- Jeżeli buduje 2 modele liniowe na tym samym zbiorze danych, to kierując się AIC ten model będzie lepszy, który będzie posiadał większą wartość AIC.
- Jeżeli współczynnik korelacji między zmienną x_2 a y ma wartość bliska zero, to zmienna x_2 nie powinna być uwzględniana przy budowie modelu liniowego, modelując zmiennną y w oparciu o pewne zmienne (i tak nie będzie istotna statystycznie).

Zadanie10:

Dane składają się z 2 kolumn oraz 10 rekordów: pierwsza (x) zawiera informacje odnośnie kategorii (A, B lub C), druga (y) zawiera pewne wartości numeryczne. Wyznaczono średnią dla drugiej kolumny względem kategorii, otrzymano $x_A = 4$ $x_B = -2$, $x_C = 1$. Zbudowano model $y \sim x$. Uzupełnij pola z ????. Czy Adjusted R-squared jest większe czy mniejsze niż 0.92?

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
 -1.00  -0.75   0.00   0.75   1.00

Coefficients:
```

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-1.00  -0.75   0.00   0.75   1.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  ????     0.4629     ????  5.55e-05 ***
x8           ????     0.7071     ????  6.25e-05 ***
xC           ????     0.7071     ????  0.00383 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9258 on 7 degrees of freedom
Multiple R-squared:  0.9119,    Adjusted R-squared:
F-statistic: 36.23 on 2 and 7 DF,  p-value: 0.000203
```

Zadanie11:

Zmienną $Y|x_1, x_2$ można opisać rozkładem Poissona z parametrem $\lambda = e^{2x_1 - 3x_2 + 1}$. Zaobserwowano, że $x_1 = 3$ a $x_2 = 1$.

- Ile wynosi wartość oczekiwana Y?
- Jakie jest prawdopobieństwo, że zmienna Y przyjmie wartość większą niż 55?

Jak zmieni się wartość średnia Y jeżeli wartość cechy x_2 wzrośnie o 1?

Jak zmieni się wartość średnia Y jeżeli wartość cechy x_2 wzrośnie o 1.5?

Zadanie12: Rozważ dane $AD.csv$. Zbuduj model przewidujący czy dany pacjent choruje na Alzheimera (zmienna DX_bl) w oparciu o zmienne AGE, HippoNV, MMSCORE, TOTAL13, FDG, AV45.

- Czy wszystkie zmienne są istotne statystycznie?
- Z użyciem $step$ zredukuj zbudowany wcześniej model - otrzymując model $m2$.
- Wyznacz dokładność, czułość i specyficzność dla $m2$.
- Czy można wskazać zmienną najbardziej istotną? Dlaczego?
- Dokonaj standaryzacji zmiennych objaśniających (bezpośrednio lub możesz zastosować polecenie scale())
- Czy można wskazać zmienną najbardziej istotną? Jeżeli tak, to która to zmienna. Jaki jest jej efekt?

```
df = read.csv("AD.csv")
df$DX_bl = as.factor(df$DX_bl)

model = glm(DX_bl~AGE+HippoNV+MMSCORE+TOTAL13+FDG+AV45,
            df, family = "binomial")
summary(model)

model = glm(DX_bl~scale(AGE)+scale(HippoNV)+scale(MMSCORE)
            +scale(TOTAL13)+scale(FDG)+scale(AV45),
            df, family = "binomial")
summary(model)

model2 = step(model, trace = 0)
summary(model2)
knitr::kable(table(model2$y, model$fitted.values>0.5))
```

```
Call:
glm(formula = DX_bl ~ AGE + HippoNV + MMSCORE + TOTAL13 + FDG +
    AV45, family = "binomial", data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 61.49198    7.82389   7.860 3.86e-15 ***
AGE         -0.03229    0.02936  -1.100    0.271
HippoNV     -27.30892    4.12601  -6.619 3.62e-11 ***
MMSCORE     -0.96875    0.15562  -6.225 4.81e-10 ***
TOTAL13      0.31964    0.04841   6.603 4.04e-11 ***
FDG         -3.66130    0.52029  -7.037 1.96e-12 ***
AV45         0.35159    1.01241   0.347    0.728
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 711.27  on 516  degrees of freedom
Residual deviance: 197.53  on 510  degrees of freedom
AIC: 211.53

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = DX_bl ~ scale(AGE) + scale(HippoNV) + scale(MMSCORE) +
    scale(TOTAL13) + scale(FDG) + scale(AV45), family = "binomial",
    data = df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.80720    0.23356   3.456 0.000548 ***
scale(AGE)    -0.23144    0.21047  -1.100 0.271499
scale(HippoNV) -2.08995    0.31576  -6.619 3.62e-11 ***
scale(MMSCORE) -2.05537    0.33017  -6.225 4.81e-10 ***
scale(TOTAL13)  2.25161    0.38029   6.603 4.04e-11 ***
scale(FDG)     -2.43302    0.34546  -7.037 1.96e-12 ***
scale(AV45)     0.07379    0.21249   0.347 0.728386
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 711.27  on 516  degrees of freedom
Residual deviance: 197.53  on 510  degrees of freedom
AIC: 211.53

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = DX_bl ~ scale(HippoNV) + scale(MMSCORE) + scale(TOTAL13) +
    scale(FDG), family = "binomial", data = df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.7584     0.2261   3.354 0.000796 ***
scale(HippoNV) -2.0209     0.3043  -6.642 3.10e-11 ***
scale(MMSCORE) -2.0701     0.3302  -6.268 3.65e-10 ***
scale(TOTAL13)  2.5479     0.3858   6.604 4.01e-11 ***
scale(FDG)     -2.3847     0.3394  -7.025 2.13e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 711.27  on 516  degrees of freedom
Residual deviance: 198.62  on 512  degrees of freedom
AIC: 208.62

Number of Fisher Scoring iterations: 8
```

	FALSE	TRUE
0	270	15
1	23	209

Zadanie13:

- Funkcja przeżycia dana jest przez $S(t) = e^{-0.05t}$. Oblicz i zinterpretuj wartość $S(2)$.
- Funkcja hazardu dana jest przez $h(t) = 0.08t^3$. Wyznacz i zinterpretuj wartość $\frac{h(2)}{h(1)}$
- Zbudowano model Coxa, modelując czas potrzebny do pojawienia się pierwszych owoców od momentu zasadzenia drzewa w oparciu o pewne czynniki: X_1 ilość światła (%), X_2 zawartość azotu w glebie (mg/kg), X_3 opady (mm/miesiąc), X_4 nawóz (1-tak, 0-nie)

$$h(t|X_1, X_2, X_3, X_4) = h_0(t) \cdot e^{\beta_1 X_1 + 0.2X_2 + 0.3X_3 + 0.4X_4}$$

Wiadomo, że każdy dodatkowy 1% światła zwiększa prawdopodobieństwo owocowania o 5%. Ile wynosi β_1? Ile razy zwiększy się (lub zmniejszy) wartość prawdopodobieństwa owocowania jeżeli X_2 wzrośnie o 1, a równocześnie X_3 zmniejszy się o 2?