Zadania z wykładu 10

Krystian Baran 145000

18 maja 2021

Spis treści

1 Zadanie 9

Dane z próby zostały pogrupowane w tabeli:

Przedział	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 9]	(9, 10]
liczba i)	52	38	20	12	7	6	5	5	4	1
wyników ii)	33	36	19	14	9	8	4	6	5	4

Na poziomie istotności 0,02 zweryfikować hipotezę, że dane te pochodzą z rozkładu o gęstości $f(x)$ określonej wzorem (wyznaczyć a):

$$f(x) = \begin{cases} a(10-x) & \text{dla} \quad x \in [0,10] \\ 0 & \text{dla} \quad \text{w p. p.} \end{cases}$$

Wyznaczymy a z własności funkcji gęstości:

$$\begin{aligned} \int_{\mathbb{R}} f(x)dx &= 1 \\ &= \int_{\mathbb{R}} a(10-x)\mathbb{I}_{[0,10]}(x)dx \\ &= \int_0^{10} a(10-x)dx = a(10x - \frac{x^2}{2})\Big|_0^{10} \\ &= a(100-50) = 50a = 1 \\ a &= \frac{1}{50} = 0.02 \end{aligned}$$

Znając a możemy wyznaczyć dystrybuantę:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} 0.02(10-t)\mathbb{I}_{[0,10]}(t)dt \\ &= \int_0^x 0.02(10-t)dt \\ &= 0.02(10t - \frac{t^2}{2})\Big|_0^x \\ &= 0.02(10x - \frac{x^2}{2}) \end{aligned}$$

$$F(x) = \begin{cases} 0 & , \quad x < 0 \\ 0.02(10x - \frac{x^2}{2}) & , \quad x \in [0,10] \\ 1 & , \quad x > 10 \end{cases}$$

Wyznaczymy dystrybuantę z próby korzystając ze wzoru poniżej i biorąc lewe końce każdego przedziału:

$$F_n(x|X) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{(-\infty,x]}(X_k), x = 1, 2, \ldots, 10$$

Poniżej przykładowe jedno obliczenie, wystarczy wziąć liczebność danego przedziału i poprzednich i podzielić przez całkowitą liczebność.

$$F(1|X) = \frac{52}{150} \approx 0.346667$$

$$F(2|X) = \frac{52+38}{150} = \frac{90}{150} = 0.6$$

$$\dots$$

Tak wyznaczono dystrybuantę dla i i ii. Wartości zapisano poniżej wraz z wartosciami dystrybuanty wyznaczonej na początku

x	i)	ii)	F(x)
1	0.346667	0.239130	0.19
2	0.600000	0.500000	0.36
3	0.733333	0.637681	0.51
4	0.813333	0.739130	0.64
5	0.860000	0.804348	0.75
6	0.900000	0.862319	0.84
7	0.933333	0.891304	0.91
8	0.966667	0.934783	0.96
9	0.993333	0.971014	0.99
10	1.000000	1.000000	1

	SUM n
i)	150
ii)	138

Zastosujemy test Kołmogorowa, zatem musimy wyznaczyć wartość $D_n = max\{|F_n(x|X) - F(x)|\}$, zatem obliczone zostały szukane różnice i wyznaczono wartości maksymalne wynoszące:

i) 0.24

ii) 0.14

Korzystając z tablic wyznaczono wartości krytyczne:

$$\text{i)} : \frac{1.51743}{\sqrt{150}} \approx 0.123898$$
$$\text{ii)} : \frac{1.51743}{\sqrt{138}} \approx 0.129172$$

Widzimy zatem że wartości te są większe od wartości krytycznych, zatem odrzucamy hipotezę zerową mówiąca że wartości z próby mają podana dystrybuantę

2 Zadanie 13

Wygenerować próby o liczebności 100 obserwacji według rozkładów:

i) $N(900; 50)$,

ii) $TR(725; 1075)$,

Następnie

a) obliczyć podstawowe statystyki,

b) sporządzić wykresy histfit, normplot, Q-Q,

c) przeprowadzić testy losowości,

d) przeprowadzić testy normalności,

e) przeprowadzić testy zgodności z innymi rozkładami,

f) przeprowadzić test zgodności dla wygenerowanych prób.

Dane losowe wygenerowane w R za pomocą funkcji poniżej. Zaokrąglono wartości do dwóch liczb po przecinku.

- **rnorm(100, 900, 50)**
- **rtri(725, 1075, (1075-725)/2 + 725)** (pakiet "EnvStats")

Wartości prób losowych podane pod zakładką **Dane**.

2.1 a)

Korzystając z gotowych funkcji w R, obliczono średnią, wariancje i odchylenie standardowe. Wyniki zapisano poniżej.

- **mean()** - średnia
- **var()** - wariancja
- **sqrt(var())** - odchylenie standardowe

	i)	ii)
$\overline{X}$	896.18	899.5525
S^2	2485.676	4611.553

2.2 b)

Nie rysowano.

2.3 c)

Aby zbadać losowość próby zastosujemy test Walda Wolfowitza. Wyznaczymy statystykę następująco:

$$Z = \frac{R - \mu_R}{\sigma_R}$$

Statystyka ta ma rozkład statystyki $\sim N(0,1)$. R jest liczbą serii która wyznaczamy jako ilość liczb mniejszych od mediany.

$$\begin{aligned}\mu_R &= \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1 \\ &= \frac{2 \cdot 50 \cdot 50}{50 + 50} + 1 \approx 51\end{aligned}$$

$$\begin{aligned}\sigma_R^2 &= \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \\ &= \frac{2 \cdot 50 \cdot 50 (2 \cdot 50 \cdot 50 - 50 = 50)}{(50 + 50)^2 (50 + 50 - 1)} \approx 24.747475\end{aligned}$$

$$\sigma_R = \sqrt{\sigma_R^2} \approx 4.974683$$

Wtedy, ponieważ R jest równe dla oby danych i wynosi 50:

$$Z = \frac{50 - 51}{4.974683} \approx 0.201018$$

Obliczymy *p-value* dla prawostronnej i lewostronnej hipotezy o losowości:

$$\text{p-value}_1 \overset{R}{=} pnorm(0.201018, 0, 1) \approx 0.5796578$$

$$\text{p-value}_2 \overset{R}{=} 1 - pnorm(0.201018, 0, 1) \approx 0.4203422$$

W oby przypadkach nie możemy odrzucić hipotezę że wartości pochodzą z próby losowej.

2.4 d)

Test normalności został przeprowadzony za pomocą funkcji w R **ks.test(x,test)**, gdzie **test** jest dystrybuantą:

- $F_{N(900,50)}$ dla i)
- $F_{N(\overline{X},s)}$ dla ii)

Test jest dwustronny i oddaje wartości *p-value*, odpowiednio, 0.7774 dla **i)**, 0.8638 dla **ii)**. Zatem, przyjmując $\alpha = 0.05$ nie możemy odrzucić hipotezę o normalności. Wnioskujemy że oba rozkłady są normalne.
Ponieważ drugi rozkład pochodzi od rozkładu trójkątnego, możemy powiedzieć że tego typu rozkład jest zbliżony do normalnego.

2.5 e)

Nie przeprowadzono testu.

2.6 f)

Jak w podpunkcie **d**, zastosujemy test Kołmogorowa w R następująco: **ks.testx,y**. Gdzie x są dane pierwszej próby a y są dane drugiej próby. Orzymano następujący wynik:
Two-sample Kolmogorov-Smirnov test
data: x and y
D = 0.13, p-value = 0.3667
alternative hypothesis: two-sided

p-value jest większe od α, zatem wnioskujemy że rozkłady są do siebie podobne. Natomiast, z tabeli w **Tabele** weźmiemy wartość D_n dla n = 100, uzyskamy:

$$D_n = \frac{1.35810}{\sqrt{100}} = 0.13581$$

Zatem widzimy że rozkłady są bardzo blisko bycia rożnych, ponieważ gdy $D_n < D$ to odrzucamy hipotezę równania się rozkładów.

3 Zadanie 18

Wygenerować dużą próbę według jednego z rozkładów: beta, gamma, Weibulla lub logarytmiczno-normalnego i przekazać uzyskane dane drugiej osobie do identyfikacji rozkładu – nie informując o mechanizmie generowania. Dokonać oceny jakości dokonanej identyfikacji.

Uzyskane dane załadowano w R i, za pomocą funkcji w pakiecie "moments" obliczono współczynnik asymetrii.

$$\tilde{\mu}_3 = \frac{\mu_3}{\sigma^3} \stackrel{R}{=} skewness(data) \approx 3.454167$$

Jest to wartość dodatnia, zatem rozkład danych może być logarytmiczno-normalny, Weibulla lub gamma.

Aby przeprowadzić testy skorzystano z następującego skryptu w R, który oblicza dystrybuantę dla podanych danych, dokonuje "fitdistr" dla danego rozkładu testowego, oblicza wartość $D_{n,\alpha}$ na poziomie $\alpha = 0.05$ według tablicy w **Tablice** i oblicza $D_n = max\{|F_n(x) - F(x)|\}$.

```
library("moments")
library("MASS")
data = read.csv("data.csv")
x = sort(data[[1]])
D = 1.3581/sqrt(length(x))
cat("Skewness = ", skewness(x), "\ n")
# Dystrybuanta empiryczna
X = c()
s = 0
for(i in x) { s = length(which(x <= i))/length(x)
   X = c(X, s)
}

# Weibull
estimate = fitdistr(x,"weibull")
k = estimate[[1]][1]
lambda = estimate[[1]][2]
Y = pweibull(x, k, lambda)
d0 = max(abs(X-Y))
cat("D0 = ", d0, " D = ", D, "\n")

# gamma
estimate = fitdistr(x,"gamma")
alpha = estimate[[1]][1]
sigma = estimate[[1]][2]
Y = pgamma(x, alpha, rate = sigma)
d0 = max(abs(X-Y))
```

```
cat("D0 = ", d0, " D = ", D, "\ n")

#lognormal
estimate = fitdistr(x,"lognormal")
meanlog = estimate[[1]][1]
sdlog = estimate[[1]][2]
Y = plnorm(x, meanlog, sdlog)
d0 = max(abs(X-Y))
cat("D0 = ", d0, " D = ", D, "\n")
```

Wynik tego skryptu jest następujący:

- Skewness = 3.454167
- (Weibull) D0 = 0.06315938 D = 0.03036804
- (gamma) D0 = 0.07382558 D = 0.03036804
- (lognormal) D0 = 0.008872974 D = 0.03036804

Aby wiedzieć czy dane mają podaną rozkład porównujemy D z $D0$, jeżeli D jest większe od $D0$ to przyjmujemy że dane mają podany rozkład, w przeciwnym wypadku nie mają danego rozkładu. Widzimy zatem że dla Rozkładu Weibulla i gamma $D < D0$ zatem dane nie mają żadne z tych danych; natomiast widzimy że dla rozkładu logarytmiczno normalnego $D > D0$, zatem wnioskujemy że dane mają rozkład logarytmiczno normalny.

4 Tablice

Tablica wartości $D_{n,\alpha}$ testu Kołmogorowa.

n\α	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

5 Dane

Lp	x
1	788.93
2	935.01
3	835.42
4	860.39
5	878.83
6	849.94
7	908.35
8	968.99
9	870.11
10	892.63
11	897.37
12	920.72
13	966.1
14	866.96
15	832.1
16	855.79
17	804.91
18	853.83
19	947.65
20	955.98
21	800.77
22	934.35
23	931.69
24	870.11
25	888.78
26	871.46
27	935.5
28	904.21
29	842.15
30	866.94
31	885.97
32	897.12
33	882.85
34	979.81
35	941.73
36	828.26
37	853.63
38	865.62
39	937.56
40	890.47
41	794.4
42	810.03
43	881.28
44	814.12
45	892.84
46	938.28
47	1025.62
48	822.67
49	881.54
50	889.49
51	917.36
52	951.67
53	931.21
54	865.7
55	951.29
56	922.04
57	930.19
58	904.3
59	961.71
60	791.94
61	938.9
62	867.11
63	925.02
64	948.97
65	889.95
66	904.09
67	970.95
68	828.7
69	908.15
70	881.51
71	888.93
72	920.37
73	933.91
74	863.51
75	888.06
76	967.33
77	957
78	955.78
79	989.64
80	905.95
81	940.8
82	924.3
83	854.09
84	863.63
85	869.84
86	939.85
87	862.09
88	836.87
89	903.25
90	928.97
91	928.12
92	969.11
93	893.51
94	944.17
95	913.88
96	906.93
97	857.38
98	933.96
99	819.42
100	817.33

Lp	x
1	842.23
2	859.28
3	951.14
4	839.27
5	964.35
6	901.59
7	840.32
8	901.09
9	878.21
10	899.32
11	1069.14
12	809.81
13	882.93
14	892.01
15	919.61
16	939.53
17	804.38
18	731.17
19	998.84
20	979.77
21	831.48
22	901.58
23	954.37
24	879.23
25	856.04
26	942.74
27	946.08
28	898.02
29	995.29
30	819.69
31	868.23
32	867.35
33	766.67
34	787.84
35	945.2
36	1036.35
37	825.6
38	760.42
39	963.32
40	896.01
41	809.51
42	938.51
43	829.88
44	955.69
45	887.25
46	928.08
47	808.55
48	898.84
49	879.96
50	798.1
51	936.57
52	928.21
53	904.59
54	870.65
55	945.23
56	911.54
57	952.39
58	786.45
59	980.29
60	859.18
61	853.97
62	906.85
63	905.28
64	911.85
65	1000.95
66	864.62
67	823.08
68	1026.75
69	872.34
70	869.66
71	930.93
72	1028.37
73	1001.82
74	875.69
75	884.66
76	776.26
77	952.72
78	949.97
79	848.93
80	861.74
81	886.01
82	986.02
83	869.18
84	887.4
85	954.19
86	847.53
87	846.72
88	902.28
89	903.08
90	961.64
91	977.29
92	995.46
93	880.4
94	844.93
95	891.68
96	870.76
97	829.35
98	966.82
99	908.08
100	1045.02

6 Bibliografia

- https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test
- https://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/
- https://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/kolmogorov-smirnov-test/
- https://kindsonthegenius.com/blog/how-to-perform-wald-wolfowitz-test-testing-for-homogeneity-with-run-test/