Andrzej Sierociński

STATYSTYKA MATEMATYCZNA

Wnioskowanie statystyczne

Wykład 11

28 Model statystyczny, podstawowe problemy statystyki matematycznej

Statystyka matematyczna jest działem probabilistyki i podobnie jak w rachunku prawdopodobieństwa zajmuje się badaniem modeli matematycznych (probabilistycznych) pewnych zjawisk losowych.

Statystyka jest ściśle związana z rachunkiem prawdopodobieństwa, jednakże jej punkt widzenia jest odmienny. W rachunku prawdopodobieństwa mamy przestrzeń probabilistyczną z jednoznacznie określonym rozkładem prawdopodobieństwa, który następnie wykorzystujemy do wyznaczania prawdopodobieństw interesujących nas zdarzeń losowych. W statystyce natomiast nie zakłada się pełnej znajomości rozkładu prawdopodobieństwa, który jest cechą statystyczną elementów badanej zbiorowości (populacji generalnej).

Punktem wyjścia każdego badania statystycznego jest wylosowanie (czasem przeprowadzenie pewnych doświadczeń) z całej populacji pewnej skończonej (czasami losowej) liczby n elementów i zbadanie ich ze względu na określoną cechę (zmienną losową) X. Zawsze zakładamy, że o X posiadamy pewną wiedzę a priori, tzn. że prawdziwy rozkład prawdopodobieństwa P zmiennej losowej X należy do pewnej klasy rozkładów prawdopodobieństwa $\mathcal{P}$.

W wyniku zaobserwowania n realizacji $x_1, x_2, \ldots, x_n$ cechy X chcemy uściślić naszą wiedzę o rozkładzie $P \in \mathcal{P}$.

Przykład *Przedmiotem badania jest symetria pewnej monety. Dokonujemy n rzutów w wyniku, których otrzymujemy k, $(0 \leqslant k \leqslant n)$ orłów. Jeżeli oznaczymy przez X losową liczbę orłów uzyskanych w n niezależnych rzutach, to*

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

gdzie $p \in (0,1)$ jest (nieznanym) prawdopodobieństwem wypadnięcia orła w jednym rzucie.

Przykładowe pytania jakie możemy stawiać to:

1. *"ile wynosi p?" i*
2. *"czy moneta jest symetryczna (czy $p = 0,5$)?".*

Pierwsze pytanie jest pytaniem o ocenę wartości nieznanego parametru rozkładu prawdopodobieństwa badanej cechy. Ta część wnioskowania statystycznego, która zajmuje się odpowiedziami na tego rodzaju pytania nosi nazwę **teorii estymacji**.

Drugie pytanie jest przykładowym problemem **weryfikacji (badania prawdziwości) hipotez statystycznych**.

Dowolne dwie n-elementowe próbki z tej samej populacji są na ogół różne. Zatem wnioskowanie statystyczne, oparte na częściowej informacji, dostarcza jedynie wniosków wiarygodnych - a nie absolutnie prawdziwych.

28.1 Model statystyczny

Wygodnie jest zatem próbkę, tzn. ciąg liczbowy $x_1, x_2, \ldots, x_n$ traktować jako realizację pewnego ciągu zmiennych losowych

$$X_1, X_2, \ldots, X_n,$$

gdzie X_i, $i = 1, 2, \ldots, n$, jest zmienną losową o zbiorze wartości i-tego spośród n wylosowanych elementów.

Punktem wyjścia w naszych rozważaniach będzie zawsze pewien element losowy X (zmienna losowa, skończony lub nieskończony ciąg zmiennych losowych) odpowiadający wynikowi eksperymentu czy obserwacji, który będziemy nazywali **próbą**.

Zbiór wartości $\mathcal{X}$ elementu losowego X nazywamy **przestrzenią próby**. W dalszym ciągu będziemy zakładali, że $\mathcal{X}$ jest pewnym skończonym lub nieskończonym zbiorem przeliczalnym, albo pewnym obszarem w przestrzeni $\mathcal{R}^n$.

Niech $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ będzie rodziną rozkładów prawdopodobieństwa na przestrzeni prób $\mathcal{X}$, indeksowaną pewnym parametrem θ. Zauważmy, że dopóki nic nie zakładamy o zbiorze indeksów Θ, to parametryzacja rodziny rozkładów $\mathcal{P}$ odbywa się bez straty ogólności, ponieważ jako parametr θ rozkładu $P \in \mathcal{P}$ można przyjąć sam rozkład P.

Zawsze będziemy zakładali, że rozkłady są **identyfikowalne**, tzn. dla $\theta_1 \neq \theta_2$ mamy $P_{\theta_1} \neq P_{\theta_2}$.

Definicja Parę

$$(\mathcal{X}, \{P_\theta : \theta \in \Theta\})$$

nazywamy **przestrzenią statystyczną**, a każde odwzorowanie

$$g : \mathcal{X} \to \mathcal{R}^k$$

k-wymiarową statystyką.

Jeżeli

$$X = (X_1, X_2, \ldots, X_n),$$

gdzie $X_1, X_2, \ldots, X_n$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie prawdopodobieństwa P_θ na $\mathcal{X}$, to próbę tę nazywamy **prostą próbą losową o liczności n**, a odpowiadająca jej przestrzeń statystyczna jest przestrzenią produktową

$$(\mathcal{X}, \{P_\theta : \theta \in \Theta\})^n .$$

Przykład *Skonstruujmy przestrzeń statystyczną dla eksperymentu, w którym dokonujemy n niezależnych rzutów monetą. Wynik pojedynczego rzutu jest zmienną losową o rozkładzie dwupunktowym. Załóżmy, że prawdopodobieństwo orła w pojedynczym rzucie jest równe $\theta \in (0,1)$. Zdefiniujmy zmienną losową opisującą wynik i-tego rzutu, $1 \leqslant i \leqslant n$:*

$$X_i = \begin{cases} 0 & \text{reszka w } i-\text{tym rzucie} \\ 1 & \text{orzeł w } i-\text{tym rzucie.} \end{cases}$$

Wówczas $\mathcal{X} = \{0,1\}$, a $P_\theta(X = 1) = \theta = 1 - P_\theta(X = 0)$. Przestrzeń statystyczna jest przestrzenią produktową $(\mathcal{X}, \{P_\theta : \theta \in \Theta\})^n$.

Możliwy jest inny opis: $\mathcal{X} = \{(x_1, \ldots, x_n), x_i = 0 \vee 1\}$, a prawdopodobieństwo

$$P_\theta(X = (x_1, \ldots, x_n)) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Przykład *Dokonujemy n niezależnych pomiarów pewnej wielkości μ. Każdy pomiar jest obarczony błędem losowym ϵ, który jest zmienną losową o rozkładzie normalnym $N(0, \sigma^2)$. Skonstruować przestrzeń statystyczną.*

Jest oczywistym, że wynik i-tego pomiaru $X_i = \mu + \epsilon$, ma rozkład normalny $N(\mu, \sigma^2)$. Zatem mamy do czynienia z przestrzenią statystyczną :

$$\left(\mathcal{R}, \left\{ f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] : \mu \in \mathcal{R}, \sigma > 0 \right\}\right)^n$$

lub inaczej

$$\left(\mathcal{R}^n, \left\{ f_{\mu,\sigma}(x_1, \ldots, x_n) = \left(\sigma\sqrt{2\pi}\right)^{-n} \exp\left[-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i-\mu}{\sigma}\right)^2\right] : \mu \in \mathcal{R}, \sigma > 0 \right\}\right).$$

28.2 Dystrybuanta empiryczna i jej własności

Jeżeli $X_1, X_2, \ldots, X_n$ jest prostą próbą losową o liczności n, gdzie zmienne losowe X_i mają rozkład prawdopodobieństwa P_θ o dystrybuancie F, to wartość oczekiwaną względem tego rozkładu będziemy oznaczali przez E_θ lub E_F.

Wówczas dystrybuanta empiryczna jest statystyką, czyli zmienną losową, zdefiniowaną następująco:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}\{X_i \leqslant x\}.$$

Dla każdego $x \in \mathcal{R}$ zmienne losowe $\mathbf{I}\{X_i \leqslant x\}$, $i = 1, 2, \ldots, n$ są niezależne o jednakowym rozkładzie $b(1, F(x))$.

Korzystając z własności rozkładu Bernoulliego oraz stosując do ciągu $\mathbf{I}\{X_i \leqslant x\}$, $i = 1, 2, \ldots$ mocne prawo wielkich liczb oraz centralne twierdzenie graniczne otrzymujemy następujące własności:

Dla dowolnego $x \in \mathcal{R}$

1. $E_F\left(F_n(x)\right) = F(x)$,
2. $P_F\left\{\lim\limits_{n\to\infty} F_n(x) = F(x)\right\} = 1$,
3. $\lim\limits_{n\to\infty} P_F\left\{\sqrt{n}\frac{F_n(x)-F(x)}{\sqrt{F(x)[1-F(x)]}} \leqslant t\right\} = \Phi(t)$, dla każdego $t \in \mathcal{R}$, gdzie Φ oznacza dystrybuantę standardowego rozkładu normalnego.

Można powiedzieć, że własności te wyjaśniają sens w jakim próba losowa $X_1, X_2, \ldots, X_n$ odtwarza rozkład, z którego pochodzi.

29 Estymacja

29.1 Estymacja punktowa - sformułowanie problemu

Niech cecha X ma rozkład prawdopodobieństwa P_θ z pewnej rodziny rozkładów $\{P_\theta : \theta \in \Theta\}$, gdzie θ jest nieznanym parametrem. Naszym zadaniem jest wskazanie tego rozkładu, tzn. oszacowanie nieznanej wartości parametru θ.

Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu P_θ. Jak wiadomo z własności dystrybuanty empirycznej próba losowa wraz ze wzrostem liczby obserwacji coraz lepiej przybliża nieznany rozkład. Zatem jedyne, co możemy zrobić, to znaleźć oszacowanie parametru θ na podstawie zaobserwowanych wartości próby losowej. Zadanie to można sformułować nieco ogólniej jako zadanie szacowania wartości pewnej funkcji g od parametru θ.

W dalszym ciągu będziemy rozważali jedynie przypadek, gdy funkcja g jest funkcją rzeczywistą o wartościach w $\mathcal{R}$, $g : \Theta \to \mathcal{R}$.

Definicja
Każdą statystykę

$$\hat{g}_\theta = \hat{g}_\theta(X_1, X_2, \ldots, X_n)$$

służącą do oceny wartości funkcji $g_\theta = g(\theta)$, nazywamy **estymatorem parametru** g_θ.

Oczywiście nie wszystkie statystyki, które mogą być używane do estymacji g_θ są jednakowo dobre. Podstawowym czynnikiem, który będzie decydował o tym czy dany estymator jest lepszy od drugiego estymatora będzie odpowiednio zdefiniowany **błąd estymacji**, czyli odległość estymatora od wartości estymowanej. W dalszym ciągu ograniczymy się do przypadku tzw. **błędu średniokwadratowego**, najczęściej używanego w teorii estymacji.

Definicja
Błędem średniokwadratowym estymatora $\hat{g}_\theta$ parametru g_θ, nazywamy wyrażenie

$$MSE(g_\theta) = E_\theta \left\{ |\hat{g}_\theta - g_\theta|^2 \right\}.$$

W teorii estymacji błąd średniokwadratowy nosi nazwę **ryzyka** estymatora przy kwadratowej funkcji straty $L(\hat{g}_\theta, g_\theta) = |\hat{g}_\theta - g_\theta|^2$. Ideałem byłoby wyznaczenie takiego estymatora, który minimalizowałby błąd średniokwadratowy jednostajnie dla wszystkich rozkładów prawdopodobieństwa z rodziny $\{P_\theta : \theta \in \Theta\}$. Niestety, przy tak ogólnym sformułowaniu problemu jest to niemożliwe.

Istotnie, wystarczy zauważyć, że estymatory stałe, postaci $\hat{g}_\theta = \theta_0$ dają dla $\theta = \theta_0$ ryzyko równe 0, także przy innej (niekoniecznie kwadratowej) funkcji straty. Problem ten można rozwiązać przez odpowiednie ograniczenie klasy rozważanych estymatorów tak, aby w nowej klasie minimum funkcji ryzyka istniało. Jest to znany zabieg jaki stosuje się w wielu problemach optymalizacyjnych. W statystyce zwykle nakłada się na estymatory wymaganie tzw. **nieobciążoności**.

Definicja
Estymator $\hat{g}_\theta$ parametru g_θ, nazywamy **estymatorem nieobciążonym (EN)**, jeżeli dla każdego $\theta \in \Theta$ mamy

$$E_\theta \left[\hat{g}_\theta(X_1, X_2, \ldots, X_n)\right] = g_\theta.$$

Warunek ten mówi, że średnio estymator daje wartość estymowanego parametru. Oczywiście klasa estymatorów nieobciążonych nie zawiera estymatorów stałych, które z praktycznego punktu widzenia są niepotrzebne. Niestety, w pewnych przypadkach, założenie nieobciążoności eliminuje także estymatory, które moglibyśmy uznać za dobre. Zwróćmy uwagę na fakt, że dla estymatora nieobciążonego jego błąd średniokwadratowy jest po prostu jego wariancją. Tym samym w klasie estymatorów nieobciążonych problem wyznaczenia estymatora, dla którego błąd średniokwadratowy jest najmniejszy jest problemem wyznaczenia **estymatora o minimalnej wariancji (ENMW).**

Uwaga. Średnia z próby

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

jest estymatorem nieobciążonym wartości oczekiwanej populacji (o ile istnieje). Istotnie

$$E_\theta\left(\bar{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n} E_\theta\left(X_i\right) = E_\theta\left(X_1\right) \quad \forall\theta \in \Theta.$$

Uwaga. Wariancja empiryczna z próby

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

jest estymatorem nieobciążonym wariancji populacji (o ile istnieje).

Istotnie

$$\begin{aligned} E_\theta\left(s^2\right) &= E_\theta\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)^2\right) = \\ &= E_\theta\left(\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2-n\bar{X}_n^2\right]\right) = \\ &= \frac{1}{n-1}\left[n(\sigma^2+\mu^2)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right]=\sigma^2. \end{aligned}$$

Uwaga. Dla próby losowej z rozkładu Bernoulliego z rodziny $\{b(1,p), p\in(0,1)\}$ średnia z próby

$$\hat{p}_n = \bar{X}_n$$

jest estymatorem nieobciążonym o minimalnej wariancji parametru p. ($\hat{p}_n$ jest $ENMW(p)$).

Uwaga. Dla próby losowej z rozkładu normalnego z rodziny $\{N(\mu,\sigma^2), \mu\in\mathcal{R}, \sigma\in\mathcal{R}_+\}$ (μ,σ nieznane) średnia z próby oraz wariancja z próby są estymatorami nieobciążonymi o minimalnej wariancji, tzn.

$$\begin{array}{rll} \bar{X}_n & \text{jest} & ENMW(\mu) \quad \text{oraz} \\ s^2 & \text{jest} & ENMW(\sigma^2). \end{array}$$

W literaturze bardzo często dla estymatora ENMW używa się określenia **estymator najefektywniejszy**. Wiąże się to z tzw. pojęciem **efektywności estymatorów nieobciążonych**. Przy pewnych dość ogólnych założeniach o rodzinie rozkładów, można wyznaczyć ograniczenie dolne na wariancję estymatorów nieobciążonych. Możliwe jest także porównanie wariancji każdego badanego estymatora z kresem dolnym wariancji estymatorów nieobciążonych.

Z praktycznego punktu widzenia dwa modele:

1. dwumianowy oraz
2. normalny

odgrywają najistotniejszą rolę we wnioskowaniu statystycznym.

Dla modelu dwumianowego mamy $n\cdot\hat{p}_n \sim b(n,p)$, co pozwala wyznaczać wszystkie prawdopodobieństwa dla statystyki $\hat{p}_n$ dla małych $n\leqslant 25$ lub w oparciu o przybliżenie normalne bądź Poissona dla $n>25$.

Podstawą wszystkich procedur wnioskowania statystycznego dla modelu normalnego jest twierdzenie, którego tezy są częściowo znane, ale podamy je bez dowodu.

Twierdzenie.
Jeżeli $X_1, X_2, \ldots, X_n$ jest prostą próbą losową z populacji o rozkładzie normalnym $N(\mu, \sigma^2)$, to

(a) $\bar{X}_n$ ma rozkład normalny $N(\mu, \frac{\sigma^2}{n})$;

(b) $\frac{(n-1)s^2}{\sigma^2}$ ma rozkład chi-kwadrat $\chi^2[n-1]$ z $n-1$ stopniami swobody;

(c) $\sqrt{n}\frac{\bar{X}_n - \mu}{s}$ ma rozkład $t[n-1]$ t-Studenta z $(n-1)$ stopniami swobody;

(d) statystyki $\bar{X}_n$ i s^2 są niezależne.

W przypadku próby losowej z populacji o innym rozkładzie prawdopodobieństwa do oceny nieznanej wartości oczekiwanej i wariancji również używa się średniej i wariancji z próby. Z MPWL Kołmogorowa wynika, że oba te estymatory, wraz ze wzrostem liczności, są zbieżne z prawdopodobieństwem 1 do parametrów teoretycznych. O estymatorach posiadających tę własność mówimy, że są **(mocno) zgodne**.

29.2 Estymacja przedziałowa - Przedziały ufności

Szansa na to, że ocena punktowa "trafi" we właściwą, prawdziwą wartość nieznanego parametru jest na ogół bardzo mała. Dlatego też zwykle szacujemy nieznaną wartość parametru za pomocą pewnego losowego przedziału, który pokrywa nieznaną wartość parametru z założonym z góry dużym prawdopodobieństwem.

Takie przedziały nazywamy **przedziałami ufności**, a prawdopodobieństwo zdarzenia, że nieznana wartość parametru trafi do tego przedziału nazywamy **poziomem ufności**.

Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z populacji o rozkładzie z rodziny $\{P_\theta : \theta \in \Theta\}$, gdzie θ jest nieznanym parametrem. Załóżmy, że istnieją takie dwie statystyki służące do oceny nieznanego parametru g_θ

$$U_1 = U_1(X_1, X_2, \ldots, X_n) \quad \text{oraz} \quad U_2 = U_2(X_1, X_2, \ldots, X_n),$$

że $U_1 \leqslant U_2$.

Przedział ufności dla parametru g_θ
Przedział $[U_1, U_2]$ nazywamy **przedziałem ufności dla g_θ na poziomie ufności** $1-\alpha$, jeżeli

$$P\left(U_1 \leqslant g_\theta \leqslant U_2\right) \geqslant 1-\alpha,$$

gdzie $\alpha \in (0,1)$.

Problem sprowadza się do wyznaczenia dwóch statystyk U_1 i U_2. Najprostszym rozwiązaniem jest wyznaczenie dla danego parametru g_θ tzw. **funkcji centralnej**

$$t(X_1, X_2, \ldots, X_n; g_\theta),$$

tzn. funkcji spełniającej następujące dwa warunki

- Funkcja centralna jest funkcją próby oraz estymowanego parametru i jest monotoniczna ze względu na ten parametr.
- Funkcja centralna jest zmienną losową o znanym rozkładzie prawdopodobieństwa.

Sposób wyboru takiej funkcji prześledzimy na kilku podstawowych przykładach.

29.2.1 Przedziały ufności dla wartości oczekiwanej

Załóżmy, że cecha $X \sim N(\mu, \sigma^2)$, gdzie σ jest znane. Wówczas $\theta = \mu = g_\theta$ oraz $\Theta = \mathcal{R}$.

Jako funkcję centralną przyjmujemy wówczas

$$t(X_1, X_2, \ldots, X_n; \mu) = \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n}.$$

Przy przyjętych założeniach t ma rozkład $N(0,1)$. Zatem, dla dowolnego $x \in \mathcal{R}$ mamy

$$P\left(\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \leqslant x\right) = \Phi(x),$$

gdzie Φ jest dystrybuantą standardowego rozkładu normalnego.

MODEL I - rozkład normalny σ znane
Przedziałem ufności na poziomie ufności $1-\alpha$ dla wartości oczekiwanej μ dla populacji o rozkładzie normalnym ze znaną wariancją σ^2 jest przedział

$$\left[\bar{X}_n - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right],$$

gdzie $\alpha \in (0,1)$ oraz $u_{1-\frac{\alpha}{2}}$ jest $\left(1-\frac{\alpha}{2}\right)$-kwantylem standardowego rozkładu normalnego.

Dowód wynika z faktu, iż

$$P\left(\left|\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n}\right| \leqslant u_{1-\frac{\alpha}{2}}\right) = 2\Phi(u_{1-\frac{\alpha}{2}}) - 1 = 1-\alpha.$$

MODEL II - rozkład normalny σ nieznane Przedziałem ufności na poziomie ufności $1-\alpha$ dla wartości oczekiwanej μ dla populacji o rozkładzie normalnym z nieznaną wariancją σ^2 jest przedział

$$\left[\bar{X}_n - t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{X}_n + t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right],$$

gdzie $\alpha \in (0,1)$, $t_{1-\frac{\alpha}{2},n-1}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ rozkładu $t[n-1]$ oraz $s = \sqrt{s^2}$.

Dowód wynika z faktu, iż

$$P\left(\left|\frac{\bar{X}_n - \mu}{s}\sqrt{n}\right| \leqslant t_{1-\frac{\alpha}{2},n}\right) = 1-\alpha.$$

Dla wartości liczby stopni swobody $n > 40$ możemy skorzystać z przybliżenia kwantylami standardowego rozkładu normalnego:

$$t_{1-\frac{\alpha}{2},n} \approx u_{1-\frac{\alpha}{2}}.$$

MODEL III - rozkład dowolny σ nieznane, $n \geqslant 50$
Przedziałem ufności na poziomie ufności $1-\alpha$ dla wartości oczekiwanej μ dla populacji o rozkładzie z nieznaną wariancją σ^2 jest przedział

$$\left[\bar{X}_n - u_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X}_n + u_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right],$$

gdzie $\alpha \in (0,1)$, $u_{1-\frac{\alpha}{2}}$ jest $\left(1-\frac{\alpha}{2}\right)$-kwantylem rozkładu normalnego oraz $s = \sqrt{s^2}$.

Dowód wynika z Centralnego Twierdzenia Granicznego, ponieważ

$$\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \approx N(0,1).$$

Ponadto, dla dużych n mamy $\sigma \approx s$.

MODEL IV - Przedział ufności dla wskaźnika struktury p rozkład $B(1,p)$, $p \in (0,1)$, $n \geqslant 100$
Przedziałem ufności na poziomie ufności $1-\alpha$ dla prawdopodobieństwa sukcesu p w schemacie Bernoulli'ego jest przedział

$$\left[\hat{p}_n - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right],$$

gdzie $\alpha \in (0,1)$, $u_{1-\frac{\alpha}{2}}$ jest $\left(1-\frac{\alpha}{2}\right)$-kwantylem rozkładu normalnego oraz $\hat{p}_n = \bar{X}_n$.

Dowód wynika z Centralnego Twierdzenia Granicznego oraz z faktu, iż dla dużych n mamy $\sigma \approx \sqrt{\hat{p}_n(1-\hat{p}_n)}$.

29.2.2 Przedziały ufności dla wariancji

Korzystając z tego, iż w modelu normalnym ze znaną wartością oczekiwaną μ estymatorem nieobciążonym o minimalnej wariancji dla σ^2 jest

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

oraz

$$\frac{n\hat{\sigma}_0^2}{\sigma^2} \sim \chi^2[n],$$

można wyznaczyć przedział ufności dla wariancji oraz odchylenia standardowego o założonym poziomie ufności $(1-\alpha)$.

MODEL I - rozkład normalny, μ znane
Przedziałem ufności na poziomie ufności $1-\alpha$ dla wariancji σ^2 dla populacji o rozkładzie normalnym ze znaną wartością oczekiwaną μ jest przedział

$$\left[\frac{n\hat{\sigma}_0^2}{\chi^2_{1-\frac{\alpha}{2},n}}, \frac{n\hat{\sigma}_0^2}{\chi^2_{\frac{\alpha}{2},n}}\right],$$

gdzie $\alpha \in (0,1)$, $\chi^2_{1-\frac{\alpha}{2},n}$ oraz $\chi^2_{\frac{\alpha}{2},n}$ są $\left(1-\frac{\alpha}{2}\right)$ oraz $\left(\frac{\alpha}{2}\right)$ kwantylami rozkładu $\chi^2[n]$.

Dla $n > 40$ dla kwantyli rozkładu χ^2 stosujemy przybliżenie Fishera kwantylami rozkładu normalnego, mianowicie

$$\chi^2_{\alpha,n} \approx \frac{1}{2}\left(\sqrt{2n-1}+u_\alpha\right)^2.$$

Podobnie dla modelu normalnego z nieznaną wartością oczekiwaną μ wykorzystujemy fakt, iż

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2[n-1].$$

MODEL II - rozkład normalny, μ nieznane
Przedziałem ufności na poziomie ufności $1-\alpha$ dla wariancji σ^2 dla populacji o rozkładzie normalnym z nieznaną wartością oczekiwaną μ jest przedział

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right],$$

gdzie $\alpha \in (0,1)$, $\chi^2_{1-\frac{\alpha}{2},n-1}$ oraz $\chi^2_{\frac{\alpha}{2},n-1}$ są $\left(1-\frac{\alpha}{2}\right)$ oraz $\left(\frac{\alpha}{2}\right)$ kwantylami rozkładu $\chi^2[n-1]$.

Korzystając z twierdzenia granicznego, dla dużych n, można skonstruować przybliżony przedział ufności dla wariancji, dla rozkładu normalnego dla dużych n ($n > 40$).

MODEL II (cd.) - rozkład normalny, μ nieznane, $n > 40$
Przedziałem ufności na poziomie ufności $1-\alpha$ dla wariancji σ^2 dla populacji o rozkładzie normalnym z nieznaną wartością oczekiwaną μ i dużą licznością próby jest przedział

$$\left[\frac{2ns^2}{\left(\sqrt{2n-3}+u_{1-\frac{\alpha}{2}}\right)^2}, \frac{2ns^2}{\left(\sqrt{2n-3}-u_{1-\frac{\alpha}{2}}\right)^2}\right],$$

gdzie $\alpha \in (0,1)$ oraz $u_{1-\frac{\alpha}{2}}$ jest $\left(1-\frac{\alpha}{2}\right)$-kwantylem standardowego rozkładu normalnego.

Korzystając z twierdzenia granicznego, dla dużych n, można skonstruować przedział ufności dla wariancji, dla dowolnego rozkładu.

MODEL III - rozkład dowolny, μ nieznane, $n \geqslant 100$
Przedziałem ufności na poziomie ufności $1-\alpha$ dla wariancji σ^2 dla populacji

o rozkładzie z nieznaną wartością oczekiwaną μ jest przedział

$$\left[\frac{s^2}{1+\sqrt{\frac{2}{n}}\cdot u_{1-\frac{\alpha}{2}}}, \frac{s^2}{1-\sqrt{\frac{2}{n}}\cdot u_{1-\frac{\alpha}{2}}}\right],$$

gdzie $\alpha \in (0,1)$ oraz $u_{1-\frac{\alpha}{2}}$ jest $\left(1-\frac{\alpha}{2}\right)$-kwantylem standardowego rozkładu normalnego.

Dowód wynika stąd, że dla dostatecznie dużych n na podstawie CTG mamy

$$\frac{s^2-\sigma^2}{\sigma^2}\sqrt{\frac{n}{2}} \approx N(0,1).$$

30 Weryfikacja hipotez statystycznych

Hipotezą statystyczną nazywamy dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa badanej cechy.

H_0 - hipoteza zerowa,

H_1 - hipoteza alternatywna.

Testem hipotezy statystycznej nazywamy postępowanie, które każdej możliwej realizacji próby losowej $X_1, X_2, \ldots, X_n$ przyporządkowuje - z ustalonym prawdopodobieństwem - decyzję przyjęcia albo odrzucenia sprawdzanej hipotezy (H_0).

Statystyką testową nazywamy funkcję próby $\delta(X_1, X_2, \ldots, X_n)$, na podstawie której wnioskuje się o odrzuceniu lub nie hipotezy statystycznej H_0.

Zbiorem krytycznym nazywamy podzbiór W przestrzeni prób $\mathcal{X}$, do którego należą wszystkie realizacje próby, dla których podejmujemy decyzję o odrzuceniu weryfikowanej hipotezy.

Błąd I rodzaju
Błędem I rodzaju nazywamy prawdopodobieństwo odrzucenia hipotezy H_0 kiedy jest ona prawdziwa, tzn.

$$\alpha = P(W|H_0 \text{ prawdziwa}).$$

Dopełnienie $\overline{W}$ zbioru krytycznego W nazywamy zbiorem przyjęć hipotezy zerowej H_0.

Błąd II rodzaju
Błędem II rodzaju nazywamy prawdopodobieństwo przyjęcia hipotezy H_0 kiedy jest ona fałszywa, tzn.

$$\beta = P(\overline{W}|H_1 \text{ prawdziwa}).$$

	Decyzja	
	odrzucamy H_0	**nie odrzucamy** H_0
H_0 **prawdziwa**	Błąd I rodzaju α	Decyzja poprawna $1-\alpha$
H_0 **nieprawdziwa**	Decyzja poprawna $1-\beta$	Błąd II rodzaju β

Nie jest możliwe jednoczesne minimalizowanie błędów I i II rodzaju. Jeżeli liczność próby jest ustalona, to zmniejszając błąd I rodzaju powodujemy zwiększenie błędu II rodzaju i odwrotnie.

Znacznie trudniejsze jest kontrolowanie błędu II rodzaju, gdyż na ogół rozkład statystyki testowej jest dokładnie znany jedynie przy założeniu prawdziwości hipotezy zerowej. Natomiast, jeżeli prawdziwa jest hipoteza alternatywna, rozkład ten jest bardziej skomplikowany i wyznaczenie błędu II rodzaju jest znacznie trudniejsze. Dlatego, zwykle stosuje się procedurę uproszczoną, polegającą na tym, że zakłada się wielkość błędu I rodzaju α i spośród wszystkich testów wybiera się taki, że błąd II rodzaju β jest najmniejszy. O takich testach mówimy, że są **najmocniejsze**. Istnienie takich testów, przy pewnych założeniach, wynika z lematu Neymana - Pearsona. Postępowanie takie, niesymetrycznie traktuje hipotezę zerową i hipotezę alternatywną.

30.1 Parametryczne testy istotności

Hipotezą parametryczną nazywamy dowolne przypuszczenie dotyczące wartości nieznanych parametrów rozkładu badanej cechy.

Każdą hipotezę statystyczną, która nie jest hipotezą parametryczną nazywamy **hipotezą nieparametryczną**, np. hipotezę dotyczącą postaci rozkładu prawdopodobieństwa badanej cechy lub niezależności dwóch cech, itp..

W dalszym ciągu skoncentrujemy się na skonstruowaniu testów parametrycznych dotyczących wartości oczekiwanej i wariancji badanej cechy.

Z uwagi na trudności z wyznaczaniem błędu II rodzaju (prawdopodobieństwo przyjęcia H_0, kiedy jest fałszywa), jako hipotezę zerową wybiera się zwykle przypuszczenie, co do którego prawdziwości mamy poważne zastrzeżenia. Odrzucając hipotezę zerową możemy popełnić co najwyżej błąd I rodzaju, który jesteśmy w stanie kontrolować. Oczywiście, w przypadku, gdy nie możemy odrzucić hipotezy zerowej i jednocześnie nie znamy wielkości błędu II rodzaju, decyzja o przyjęciu hipotezy zerowej jest co najmniej wątpliwa. W takiej sytuacji, mówimy jedynie, że **nie ma podstaw do odrzucenia hipotezy zerowej**. Tego rodzaju testy nazywamy **parametrycznymi testami istotności** a błąd I rodzaju α nazywamy **poziomem istotności testu**.

Dla rodziny parametrycznej rozkładów $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ zbiór możliwych parametrów Θ rozbijamy na dwa rozłączne podzbiory: ω oraz $\Theta - \omega$.

Wówczas hipotezy przyjmują postać

$$H_0 : \theta \in \omega \quad \text{oraz} \quad H_1 : \theta \in \Theta - \omega.$$

Jeżeli zbiór ω składa się z jednego elementu, to taką hipotezę nazywamy **hipotezą prostą**. W przeciwnym przypadku jest to **hipoteza złożona**.

Kluczową sprawą przy konstrukcji parametrycznego testu istotności jest znalezienie statystki testowej, której rozkład jest znany przy hipotezie zerowej. Wówczas dla danego poziomu istotności α można znaleźć taki zbiór krytyczny W, że

$$\forall \theta \in \omega \quad P_\theta(W) \leqslant \alpha,$$

tzn. prawdopodobieństwo podjęcia błędnej decyzji o odrzuceniu prawdziwej hipotezy H_0 jest mniejsze od α.

30.1.1 Testy dla wartości oczekiwanej

Test hipotezy $H_0 : \mu = \mu_0$ na poziomie istotności α

MODEL I Rozkład normalny $N(\mu, \sigma^2)$, σ znane
Statystyka testowa:

$$U = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim_{|H_0} N(0,1).$$

Hipoteza alternatywna:

$$\text{(a) } H_1 : \mu \neq \mu_0, \quad \text{(b) } H_1 : \mu > \mu_0, \quad \text{(c) } H_1 : \mu < \mu_0.$$

Decyzja - odrzucamy H_0 jeżeli:

$$\text{(a) } |U| \geqslant u_{1-\frac{\alpha}{2}}, \quad \text{(b) } U \geqslant u_{1-\alpha}, \quad \text{(c) } U \leqslant u_{\alpha}.$$

$\alpha \in (0,1)$, u_α: α-kwantyl standardowego rozkładu normalnego.

MODEL II Rozkład normalny $N(\mu, \sigma^2)$, σ nieznane

Statystyka testowa:

$$t = \sqrt{n}\frac{\bar{X}_n - \mu_0}{s} \sim_{|H_0} t[n-1].$$

Hipoteza alternatywna:

$$\text{(a) } H_1 : \mu \neq \mu_0, \quad \text{(b) } H_1 : \mu > \mu_0, \quad \text{(c) } H_1 : \mu < \mu_0.$$

Decyzja - odrzucamy H_0 jeżeli:

$$\text{(a) } |t| \geqslant t_{1-\frac{\alpha}{2},n-1}, \quad \text{(b) } t \geqslant t_{1-\alpha,n-1}, \quad \text{(c) } t \leqslant t_{\alpha,n-1} = -t_{1-\alpha,n-1}.$$

$\alpha \in (0,1)$, $t_{\alpha,n-1}$ α-kwantyl rozkładu $t[n-1]$.

MODEL III Rozkład dowolny $n \geqslant 100$, σ nieznane

Statystyka testowa:

$$Z = \sqrt{n}\frac{\bar{X}_n - \mu_0}{s} \approx_{|H_0} N(0,1).$$

Hipoteza alternatywna:

$$\text{(a) } H_1 : \mu \neq \mu_0, \quad \text{(b) } H_1 : \mu > \mu_0, \quad \text{(c) } H_1 : \mu < \mu_0.$$

Decyzja - odrzucamy H_0 jeżeli:

$$\text{(a) } |Z| \geqslant u_{1-\frac{\alpha}{2}}, \quad \text{(b) } Z \geqslant u_{1-\alpha}, \quad \text{(c) } Z \leqslant u_{\alpha}.$$

$\alpha \in (0,1)$, u_α: α-kwantyl standardowego rozkładu normalnego.

Test hipotezy $H_0 : p = p_0$ na poziomie istotności α

MODEL IV Rozkład Bernoulliego $b(1,p)$ $n \geqslant 100$

Statystyka testowa:

$$Z = \sqrt{n}\frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)}} \approx_{|H_0} N(0,1).$$

Hipoteza alternatywna:

$$\text{(a) } H_1 : p \neq p_0, \quad \text{(b) } H_1 : p > p_0, \quad \text{(c) } H_1 : p < p_0.$$

Decyzja - odrzucamy H_0 jeżeli:

$$\text{(a) } |Z| \geqslant u_{1-\frac{\alpha}{2}}, \quad \text{(b) } Z \geqslant u_{1-\alpha}, \quad \text{(c) } Z \leqslant u_{\alpha}.$$

$\alpha \in (0,1)$, u_α: α-kwantyl standardowego rozkładu normalnego, $\hat{p}_n = \bar{X}_n$.

30.1.2 Testy dla wariancji

Test hipotezy $H_0 : \sigma^2 = \sigma_0^2$ na poziomie istotności α

MODEL I Rozkład normalny $N(\mu, \sigma^2)$, μ znane

Statystyka testowa:

$$\chi_0^2 = \frac{n\hat{\sigma}_0^2}{\sigma_0^2} \sim_{|H_0} \chi^2[n]$$

Hipoteza alternatywna:

$$\text{(a) } H_1 : \sigma^2 \neq \sigma_0^2, \quad \text{(b) } H_1 : \sigma^2 > \sigma_0^2, \quad \text{(c) } H_1 : \sigma^2 < \sigma_0^2.$$

Decyzja - odrzucamy H_0 jeżeli:

$$\text{(a) } \chi_0^2 \geqslant \chi^2_{1-\frac{\alpha}{2},n} \ \vee \ \chi_0^2 \leqslant \chi^2_{\frac{\alpha}{2},n} \text{ (b) } \ \chi_0^2 \geqslant \chi^2_{1-\alpha,n}, \text{ (c) } \ \chi_0^2 \leqslant \chi^2_{\alpha,n}.$$

$\alpha \in (0,1)$, $\chi^2_{\alpha,n}$: α-kwantyl rozkładu $\chi^2[n]$, $\hat{\sigma}_0^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \mu)^2$.

MODEL II Rozkład normalny $N(\mu, \sigma^2)$, μ nieznane

Statystyka testowa:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim_{|H_0} \chi^2[n-1]$$

Hipoteza alternatywna:

$$\text{(a) } H_1 : \sigma^2 \neq \sigma_0^2, \quad \text{(b) } H_1 : \sigma^2 > \sigma_0^2, \quad \text{(c) } H_1 : \sigma^2 < \sigma_0^2.$$

Decyzja - odrzucamy H_0 jeżeli:

$$\text{(a) } \chi^2 \geqslant \chi^2_{1-\frac{\alpha}{2},n-1} \ \vee \ \chi^2 \leqslant \chi^2_{\frac{\alpha}{2},n-1}$$

$$\text{(b) } \chi^2 \geqslant \chi^2_{1-\alpha,n-1}, \quad \text{(c) } \chi^2 \leqslant \chi^2_{\alpha,n}.$$

$\chi^2_{\alpha,n-1}$: α-kwantyl rozkładu $\chi^2[n-1]$, $s^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$.

MODEL III Rozkład dowolny $n \geqslant 100$, μ nieznane

Statystyka testowa:

$$Z = \frac{s^2 - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{n}{2}} \approx_{|H_0} N(0,1).$$

Hipoteza alternatywna:

$$\text{(a) } H_1 : \sigma^2 \neq \sigma_0^2, \quad \text{(b) } H_1 : \sigma^2 > \sigma_0^2, \quad \text{(c) } H_1 : \sigma^2 < \sigma_0^2.$$

Decyzja - odrzucamy H_0 jeżeli:

$$\text{(a) } |Z| \geqslant u_{1-\frac{\alpha}{2}}, \quad \text{(b) } Z \geqslant u_{1-\alpha}, \quad \text{(c) } Z \leqslant u_\alpha.$$

$\alpha \in (0,1)$, u_α: α-kwantyl standardowego rozkładu normalnego.

31 Testy nieparametryczne

31.1 Test zgodności chi-kwadrat Pearsona

Problem Coca-Coli
25 osób poddało się testowi, w którym miało ocenić jakość 3 napojów:

1. *Coca Coli wg starej receptury,*
2. *Coca Coli wg nowej receptury oraz*
3. *Pepsi Coli.*

W wyniku testu 12 osób opowiedziało się za starą recepturą, 7 za nową i tylko 6 za Pepsi Colą. Jedynie 3 osoby prawidłowo rozpoznały wszystkie 3 napoje.

Czy wyniki te mogą być zinterpretowane na korzyść starej formuły? Czy jest to wynik czystego przypadku?

W przykładzie mamy do czynienia z ocenami. Otrzymane oceny można traktować jako obserwacje pewnej zmiennej losowej, która przyjmuje 3 wartości: najlepszy, średni, najgorszy. Nie jest to zmienna losowa z jaką mieliśmy do czynienia do tej pory, generująca dane ilościowe. Takie dane będziemy nazywali **danymi jakościowymi**.

Dane jakościowe mogą być dwojakiego typu:

- o wartościach uporządkowanych lub
- nominalne.

Wszelkiego typu oceny, np. oceny szkolne, oceny dotyczące stopnia sympatii, dla których można zastosować jakąś relację porządkującą są **danymi o wartościach uporządkowanych**. Są jednak dane takie jak kolor oczu, grupa krwi, wyznanie religijne, których w żaden sposób nie da się uporządkować. Dane takie nazywamy **danymi nominalnymi**.

Należy zwrócić uwagę, że często cechy jakościowe powstają poprzez **dyskretyzację jakiejś cechy ilościowej**, np.

- cecha opisująca liczbę mieszkańców miast:
  małe miasta, średnie miasta, duże miasta;
- liczba dziennie wypalanych papierosów:
  do 5, 5-10, 10-20 i pow. 20.

Poszczególne wartości (poziomy) zmiennej jakościowej nazywamy **kategoriami**. Test statystyczny służący do weryfikacji hipotezy o postaci rozkładu nazywamy **testem zgodności**.

Przedstawimy test zgodności χ^2-Pearsona dla rozkładów jakościowych, który przy odpowiednim zastosowaniu dyskretyzacji (utworzeniu kategorii), może być z powodzeniem wykorzystany również w przypadku zmiennych ilościowych.

31.1.1 Weryfikacja prostej hipotezy o zgodności

Niech zmienna jakościowa X przyjmuje k wartości (kategorii) $x_1, x_2, \ldots, x_k$ i niech $P(X = x) = p_i$, $i = 1, 2, \ldots, k$ gdzie $\sum_{i=1}^{k} p_i = 1$.

Jednym z najważniejszych problemów dla tego typu zmiennych jest badanie zgodności ich rozkładu z zadanym rozkładem prawdopodobieństwa $p_1^0, p_2^0, \ldots, p_k^0$ gdzie $p_i^0 \geqslant 0$.

Do tego celu wykorzystywany jest test χ^2-Pearsona oparty na prostym twierdzeniu granicznym dla schematu wielomianowego.

Schemat wielomianowy

Schematem wielomianowym nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, ..., X_n$ o jednakowych dyskretnych rozkładach prawdopodobieństwa $P(X_j = x_i) = p_i$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$, gdzie k jest dowolną liczbą naturalną, a $\sum_{i=1}^{k} p_i = 1$ (dla $k = 2$ mamy schemat dwumianowy).

Prawdopodobieństwo tego, że w schemacie wielomianowym zaobserwujemy n_i razy wynik x_i, $i = 1, ..., k$ oraz $\sum_{i=1}^{k} n_i = n$ jest równe

$$P_n(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!} p_1^{n_1} \cdot p_2^{n_2} \cdot \ldots \cdot p_k^{n_k}.$$

Prawdopodobieństwa te uzyskujemy jako kolejne człony rozwinięcia wielomianu $(p_1 + p_2 + \ldots + p_k)^n$.

Oznaczmy przez $n_i^0 = n \cdot p_i^0$ oczekiwaną liczbę wypadnięć kategorii x_i, wówczas prawdziwe jest następujące twierdzenie graniczne.

Twierdzenie. *Dla schematu wielomianowego n niezależnych doświadczeń, w których dla wszystkich $i = 1, 2, \ldots, k$, $p_i > 0$ mamy przy $n \longrightarrow \infty$*

$$\chi_*^2 = \sum_{i=1}^{k} \frac{(n_i - n_i^0)^2}{n_i^0} \Longrightarrow \chi^2[k-1],$$

gdzie n_i oznacza obserwowaną a n_i^0 oczekiwaną liczbę wypadnięć wartości x_i (i-tej kategorii).

Jeśli pobrana próba losowa o liczności n została sklasyfikowana w k kategorii $x_1, x_2, \ldots, x_k$ oraz n_i razy zaobserwowano kategorię x_i, $i = 1, ..., k$, to na poziomie istotności α odrzucamy hipotezę

$$\begin{aligned} H_0: \quad & \text{prawdopodobieństwo } p_i \text{ zakwalifikowania obserwacji} \\ & \text{do kategorii } x_i \text{ wynosi } p_i^0,\ i = 1, 2, \ldots, k,\ \textstyle\sum_{i=1}^{k} p_i^0 = 1 \end{aligned}$$

jeżeli statystyka

$$\chi_*^2 = \sum_{i=1}^{k} \frac{(n_i - n_i^0)^2}{n_i^0} \geqslant \chi_{1-\alpha;k-1}^2,$$

gdzie $n_i^0 = n \cdot p_i^0$, a $\chi_{1-\alpha;k-1}^2$ jest $(1-\alpha)$-kwantylem rozkładu chi-kwadrat z $(k-1)$ stopniami swobody.

Przykład
W przychodni psychologicznej w pewnym mieście, zajmującej się leczeniem lęku przed ciemnością zarejestrowanych było 200 pacjentów w wieku od 5 do 33 lat. W tabeli poniżej podano liczby pacjentów dla 6 kategorii wiekowych.

Grupa wiekowa	*5-9*	*10-14*	*15-19*	*20-24*	*25-29*	*30-34*
Liczność	*31*	*53*	*41*	*29*	*24*	*22*

Jednocześnie znana jest struktura wiekowa mieszkańców tej miejscowości. W tabeli podano procent mieszkańców dla tych samych 6 kategorii wiekowych.

Grupa wiekowa	*5-9*	*10-14*	*15-19*	*20-24*	*25-29*	*30-34*
Procent	*6,1*	*7,3*	*9,3*	*8,5*	*9,4*	*7,9*

Czy rozkład wieku w próbce w sposób istotny różni się od rozkładu wieku w populacji? Przyjąć poziom istotności $\alpha = 0,01$.

Rozwiązanie.
Nietrudno zauważyć, że grupa wiekowa od 5 do 33 lat stanowi $48,5\%$ całości populacji tej miejscowości. Zatem przyjmując oznaczenie d_i na procentowość i-tej kategorii wiekowej otrzymujemy, że oczekiwana liczność i-tej kategorii wiekowej w 200-elementowej próbie osób w wieku od 5 do 33 lat wynosi

$$n_i^0 = \frac{d_i}{48,5} \cdot 200$$

Wszystkie obliczenia zestawione są w poniższej tabelce.

nr	Wiek	n_i	n_i^0	$(n_i - n_i^0)^2$	$(n_i - n_i^0)^2/n_i^0$
1	5-9	31	25,15	34,2225	1,361
2	10-14	53	30,10	524,4100	17,422
3	15-19	41	38,35	7,0225	0,183
4	20-24	29	35,05	36,6025	1,044
5	25-29	24	38,76	217,8576	5,621
6	30-33	22	32,58	111,9364	3,436
				χ_*^2	29,077

Ponieważ $\chi^2_{0,99;5} = 15,9$ to hipotezę o zgodności rozkładu wieku wśród pacjentów z rozkładem wieku w całej miejscowości należy odrzucić.

31.1.2 Weryfikacja złożonej hipotezy o zgodności

Przykład *Powiedzmy, że badamy liczbę dziennych reklamacji i chcemy wykazać, że są w pełni losowe. Model, który to opisuje jest modelem poissonowskim i prawdopodobieństwa zaliczenia do poszczególnych kategorii (określona liczba reklamacji) zależą od średniej liczby reklamacji λ, która nie jest znana i musi zostać wyestymowana na podstawie uzyskanej próby.*

W opisanym powyżej przykładzie mówimy o złożonej hipotezie o zgodności.

W przypadku, gdy prawdopodobieństwa p_i^0 są znane z dokładnością do m nieznanych parametrów, to parametry te musimy oszacować na podstawie posiadanej próby.

Załóżmy, że dla każdego i, $i = 1, 2, \ldots, k$ prawdopodobieństwa zakwalifikowania do kategorii x_i zależą od m nieznanych parametrów $p_i^0(\theta_1, \ldots, \theta_m)$. Wówczas należy wyznaczyć ich estymatory uzyskane poprzez maksymalizację funkcji wiarogodności

Funkcja wiarogodności

$$L = \prod_{i=1}^{k} [p_i^0(\theta_1, \ldots, \theta_m)]^{n_i}.$$

Twierdzenie. *Dla schematu wielomianowego n niezależnych doświadczeń, w których dla wszystkich $i = 1, 2, \ldots, k$, $p_i(\theta_1, \ldots, \theta_m) > 0$ zależne od m nieznanych parametrów są wyznaczone przez maksymalizację funkcji wiarogodności mamy przy $n \longrightarrow \infty$*

$$\chi_*^2 = \sum_{i=1}^{k} \frac{(n_i - n_i^0)^2}{n_i^0} \Longrightarrow \chi^2[k - 1 - m],$$

gdzie n_i oznacza obserwowaną a n_i^0 oczekiwaną liczbę wystąpień wartości x_i (i-tej kategorii).

Jeżeli hipoteza zerowa nie specyfikuje pewnych parametrów $\theta_1, \ldots, \theta_m$ rozkładu kategorii x_i, to statystykę testową χ^2 wyznacza się w oparciu o liczności oczekiwane obliczone na podstawie prawdopodobieństw $\hat{p}_i^0$ otrzymanych przez maksymalizację funkcji wiarogodności.

Hipotezę zerowa odrzucamy jeżeli

$$\chi^2_* = \sum_{i=1}^{k} \frac{(n_i - n_i^0)^2}{n_i^0} \geqslant \chi^2_{1-\alpha;k-1-m},$$

gdzie $n_i \geqslant 5$.

Jeżeli liczność klasy jest mniejsza od 5 to łączymy tę klasę z klasą sąsiednią.

Przykład

W pewnej firmie używane są 3 *kserografy tego samego typu. W ciągu* 300 *roboczych dni odnotowano liczbę uszkodzonych urządzeń.*

Liczba uszkodzeń	*0*	*1*	*2*	*3*
Liczba dni	*147*	*113*	*35*	*5*

Zweryfikować hipotezę o losowym charakterze uszkodzeń. Przyjąć poziom istotności $\alpha = 0,1$.

Rozwiązanie. Przyjmując założenie o losowym charakterze uszkodzeń otrzymujemy, że liczba uszkodzeń ma rozkład dwumianowy $b(3;\theta)$, gdzie θ jest nieznanym prawdopodobieństwem uszkodzenia w jednym dniu pojedynczego urządzenia. Wówczas

Liczba uszkodzeń	0	1	2	3
Prawdopodobieństwo $p_i^0(\theta)$	$(1-\theta)^3$	$3\theta(1-\theta)^2$	$3\theta^2(1-\theta)$	θ^3

Jeżeli przez n_i oznaczymy liczność i-tej kategorii, to funkcja wiarogodności przyjmuje postać

$$\begin{aligned} L &= \prod_{i=1}^{k} [p_i^0(\theta)]^{n_i} = [(1-\theta)^3]^{n_1} \cdot [3\theta(1-\theta)^2]^{n_2} \cdot [3\theta^2(1-\theta)]^{n_3} \cdot [\theta^3]^{n_4} \\ &= 3^{n_2+n_3} \cdot \theta^{n_2+2n_3+3n_4} \cdot (1-\theta)^{3n_1+2n_2+n_3}. \end{aligned}$$

Wprowadźmy następujące oznaczenia

$$N = 3(n_1 + n_2 + n_3 + n_4)$$

- całkowita liczba doświadczeń Bernoulliego,

$$N_1 = n_2 + 2n_3 + 3n_4$$

- całkowita liczba sukcesów.

Wówczas

$$\frac{\partial \ln L}{\partial \theta} = \frac{N_1}{\theta} - \frac{N - N_1}{1-\theta}.$$

Stąd otrzymujemy, że

$$\hat{\theta} = \frac{N_1}{N} = \frac{198}{900} = 0,22.$$

Wszystkie obliczenia zestawione są w poniższej tabelce.

nr	Liczba awarii	n_i	p_i^0	n_i^0	$(n_i - n_i^0)^2/n_i^0$
1	0	147	0,4746	142,37	0,15
2	1	113	0,4015	120,46	0,46
3	2	35	0,1133	33,98	0,03
4	3	5	0,0106	3,19	1,02
				χ^2_*	1,6649

Ponieważ kwantyl wynosi $\chi^2_{0,9;2} = 4,6052$, to nie ma podstaw do odrzucenia hipotezy o tym, że dzienna liczba awarii ma rozkład dwumianowy. Warto dodatkowo zauważyć, że p-wartość jest równa 0,434982.

Test do weryfikacji hipotezy o zgodności rozkładu prawdopodobieństwa dowolnej cechy X o dystrybuancie F.

Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu o dystrybuancie F. Dla zweryfikowania hipotezy

$$H_0 : F(x) = F_0(x) \quad \forall x \in \mathcal{R},$$

gdzie F_0 jest daną dystrybuantą, dokonujemy rozbicia zbioru wartości zmiennej losowej X na k rozłącznych podzbiorów (klas).

Niech n_i oznacza liczbę obserwacji należących do i-tej klasy, p_i - prawdopodobieństwo wpadnięcia obserwacji zmiennej losowej X o dystrybuancie $F_0(x)$ do i-tej klasy. Zakładamy $n_i \geqslant 5$.

Statystyka testowa testu zgodności χ^2-Pearsona

$$\chi^2_* = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}.$$

Jeżeli hipoteza H_0 nie specyfikuje pewnych parametrów $\theta_1, \ldots, \theta_m$ rozkładu zmiennej losowej X, to prawdopodobieństwa $p_i = p_i(\theta_1, \ldots, \theta_m),\ i = 1, ..., k$ zastępuje się ich estymatorami uzyskanymi przez maksymalizację funkcji wiarogodności

$$L = \prod_{i=1}^{k} [p_i(\theta_1, \ldots, \theta_m)]^{n_i}.$$

Hipotezę H_0 odrzucamy, jeżeli obliczona wartość statystyki χ^2_* przekracza wartość krytyczną - kwantyl $\chi^2_{1-\alpha,k-m-1}$, gdzie m jest liczbą oszacowanych parametrów.

Uwaga 1. Jeżeli zmienna losowa X ma rozkład Poissona z nieznanym parametrem λ i pogrupujemy obserwacje w k klas

$$(-\infty, a],\ (a+i-1, a+i],\ i=1,...,k-2,\ (a+k-2, \infty),$$

gdzie a jest pewną liczbą naturalną, to estymatory prawdopodobieństw p_i wyrażają się wzorami:

$$\hat{p}_1 = \sum_{i=0}^{a} P(i;\overline{\lambda}),\quad \hat{p}_k = \sum_{i=a+k-1}^{\infty} P(i;\overline{\lambda}),$$

$$\hat{p}_j = P(j;\overline{\lambda}),\quad j=2,\ldots,k-1,$$

gdzie

$$\overline{\lambda} = \sum_{i=0}^{n} X_i,\quad a\quad P(j;\overline{\lambda}) = e^{-\overline{\lambda}}\frac{\overline{\lambda}^j}{j!}.$$

Uwaga 2. Jeżeli zmienna losowa X ma rozkład normalny z nieznanymi parametrami μ i σ i pogrupujemy obserwacje w k klas

$$\left(-\infty, a_1+\frac{h}{2}\right],\ \left(a_i-\frac{h}{2}, a_i+\frac{h}{2}\right],\ i=2,...,k-1,\ \left(a_k-\frac{h}{2}, \infty\right),$$

gdzie $a_{i+1} = a_i + h$, $a_1 \in \mathcal{R}, h > 0$, to estymatory prawdopodobieństw p_i wyrażają się wzorami:

$$\hat{p}_1 = \Phi\left(\frac{a_1+\frac{h}{2}-\hat{\mu}}{\hat{\sigma}}\right),\quad \hat{p}_k = 1-\Phi\left(\frac{a_k-\frac{h}{2}-\hat{\mu}}{\hat{\sigma}}\right),$$

$$\hat{p}_i = \Phi\left(\frac{a_i+\frac{h}{2}-\hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{a_i-\frac{h}{2}-\hat{\mu}}{\hat{\sigma}}\right),\quad i=2,\ldots,k-1,$$

gdzie

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k} n_i a_i,\quad a\quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (a_i-\hat{\mu})^2}.$$

Uwaga 3. Jeżeli zaobserwujemy $n_i < 5$ dla pewnego i, to należy połączyć sąsiednie klasy.

Rozwiązanie problem Coca-Coli
Jeżeli postawimy hipotezę, że wybór napoju jest przypadkowy, to hipoteza zerowa ma postać

$$H_0 : p_1 = p_2 = p_3.$$

Zatem

$$n_1^0 = n_2^0 = n_3^0 = \frac{25}{3}.$$

Proste obliczenia dają

$$\chi_*^2 = \sum_{i=1}^{3} \frac{(n_i - n_i^0)^2}{n_i^0} = \frac{3}{25}\left[(12 - \frac{25}{3})^2 + (7 - \frac{25}{3})^2 + (6 - \frac{25}{3})^2\right] = 2,48.$$

Wartość krytyczna $\chi^2_{0,9;2} = 4,605$*, zatem nie ma podstaw do odrzucenia hipotezy zerowej. Ponadto, faktyczny poziom istotności, tzw. p-wartość, wynosi*

$$P(\chi^2[2] \geqslant 2,48) = 0,2894.$$

Przy założeniu, że mamy do czynienia z czystym przypadkiem, liczba osób, które prawidłowo rozpoznają wszystkie 3 napoje jest zmienną losową o rozkładzie dwumianowym

$$S_{25} \sim B\left(25, \frac{1}{6}\right).$$

Wykorzystując informację, że jedynie 3 osoby poprawnie rozpoznały wszystkie 3 napoje, możemy zauważyć, że prawdopodobieństwo

$$P(S_{25} \leqslant 3) = 0,3816.$$

Jest to dodatkowa przesłanka za tym, żeby traktować wynik eksperymentu jako czysto przypadkowy.

31.2 Weryfikacja hipotezy o niezależności

Załóżmy, że mamy n-elementową próbę dwuwymiarowej zmiennej losowej (X, Y), gdzie zmienna X może przyjąć jedną z k kategorii x_i, $i = 1, \ldots, k$ oraz zmienna Y jedną z l kategorii y_j, $j = 1, \ldots, l$.

Łącznie mamy $k \cdot l$ par kategorii (x_i, y_j).

Niech n_{ij} będzie zaobserwowaną z próby licznością kategorii (x_i, y_j). Macierz $(n_{ij})_{k \times l}$ nazywamy **tablicą kontyngencji**.

Chcemy skonstruować test pozwalający zweryfikować hipotezę o niezależności zmiennych X i Y.

Hipoteza o niezależności X i Y

$$H_0: \quad \forall (i,j)\ p_{ij} = p_{i\bullet} \cdot p_{\bullet j}, \quad i = 1, \ldots, k,\ j = 1, \ldots, l,$$

gdzie p_{ij} oznacza prawdopodobieństwo zakwalifikowania obserwacji do kategorii (x_i, y_j), a $p_{i\bullet}$ oraz $p_{\bullet j}$ oznaczają prawdopodobieństwa brzegowe.

Niezbędne jest wyestymowanie nieznanych prawdopodobieństw brzegowych. Wprowadźmy następujące oznaczenia

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij}, \quad i = 1, \ldots, k \quad \text{oraz} \quad n_{\bullet j} = \sum_{i=1}^{k} n_{ij}, \quad j = 1, \ldots, l$$

na liczność kategorii x_i zmiennej X oraz na liczność kategorii y_j zmiennej Y, odpowiednio.

Otrzymujemy zatem następującą tablicę kontyngencji

Y / X	y_1	y_2	$\cdots$	y_l	$n_{i\bullet}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2l}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	$n_{k\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet l}$	n

gdzie

$$\sum_{j=1}^{l} n_{\bullet j} = \sum_{i=1}^{k} n_{i\bullet} = n.$$

Nietrudno pokazać, że estymatorami największej wiarogodności prawdopodobieństw brzegowych są

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}$$

oraz

$$\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}.$$

Oczekiwana wartość liczby obserwacji wpadających do kategorii (x_i, y_j), przy założeniu niezależności zmiennych X i Y wynosi

Oczekiwana liczba obserwacji dla kategorii (x_i, y_j)

$$n_{ij}^0 = n \cdot \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}.$$

Możemy zatem utworzyć statystykę χ^2

Statystyka χ^2 dla testu niezależności

$$\chi^2_* = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - n^0_{ij}\right)^2}{n^0_{ij}}$$

Statystyka testowa, przy prawdziwości hipotezy zerowej ma rozkład χ^2 z $(k-1)(l-1)$ stopni swobody, ponieważ

$$k \cdot l - 1 - (l-1) - (k-1) = (k-1)(l-1).$$

Tak uzyskany test jest testem jednostronnym i odrzucamy hipotezę zerową jeśli wartość statystyki zaobserwowana z próby

$$\chi^2_* \geqslant \chi^2_{1-\alpha;(k-1)(l-1)}$$

przekracza dla zadanego poziomu istotności α wartość krytyczną ($(1-\alpha)$-kwantyl rozkładu $\chi^2[(k-1)(l-1)]$).

Przykład

W ciągu 3 miesięcy zaobserwowano 145 awarii maszyn. W tabelce podano liczby awarii poszczególnych maszyn w czasie każdej zmiany.

	Maszyna			
Zmiana	*A*	*B*	*C*	*D*
1	*9*	*5*	*11*	*12*
2	*9*	*11*	*18*	*20*
3	*12*	*9*	*12*	*17*

Czy liczba awarii maszyn jest niezależna od zmiany?
Przyjąć poziom istotności $\alpha = 0,05$.

Rozwiązanie.

W tabeli poniżej podano liczności brzegowe oraz liczności oczekiwane dla wszystkich par kategorii.

	n^0_{i1}	n^0_{i2}	n^0_{i3}	n^0_{i4}	$n_{i\bullet}$
n^0_{1j}	7,66	6,38	10,46	12,50	37
n^0_{2j}	12,00	10,00	16,40	19,60	58
n^0_{3j}	10,34	8,62	14,14	16,90	50
$n_{\bullet j}$	30,00	25,00	41,00	49,00	145

Proste obliczenia dają nam wartość statystyki testowej

$$\begin{aligned}\chi^2_* = & \frac{(9-7,66)^2}{7,66} + \frac{(5-6,38)^2}{6,38} + \frac{(11-10,46)^2}{10,46} + \frac{(12-12,5)^2}{12,5} + \\ + & \frac{(9-12)^2}{12} + \frac{(11-10)^2}{10} + \frac{(18-16,4)^2}{16,4} + \frac{(20-19,6)^2}{19,6} + \\ + & \frac{(12-10,34)^2}{10,34} + \frac{(9-8,62)^2}{8,62} + \frac{(12-14,14)^2}{14,14} + \frac{(17-16,9)^2}{16,9} = 2,20.\end{aligned}$$

Statystyka testowa ma $(4-1)(3-1) = 6$ stopni swobody.

Ponieważ wartość krytyczna $\chi^2_{0,95;6} = 12,5916$, to stwierdzamy, że nie ma podstaw do odrzucenia hipotezy o tym, że liczba awarii jest niezależna od zmiany. Dodatkowo możemy wyznaczyć p-wartość, która jest równa $0,9002$.