Математическая статистика.

Андрей Тищенко @AndrewTGk

2024/2025

Семинар 10 января

Задача 1

$x_1, \dots, x_n \sim F_\xi(x)$, найти функцию распределения для $X_{(n)},\ X_{(1)}$
$F_{X_{(n)}}(x) = P(X_{(n)} \leqslant x) = P(X_{(1)} \leqslant x, \dots, X_{(n)} \leqslant x) = P(X_1 \leqslant x, \dots, X_n \leqslant x) =$
$= P(X_1 \leqslant x) \dots P(X_n \leqslant x) = (F_\xi(x))^n$
$F_{X_{(1)}}(x) = P(X_{(1)} \leqslant x) = 1 - P(X_{(1)} > x) = 1 - P(X_{(1)} > x, \dots, X_{(n)} > x) =$
$= 1 - P(X_1 > x, \dots, X_n > x) = 1 - P(X_1 > x) \dots P(X_n > x) = 1 - \big(1 - F_\xi(x)\big)^n$

Задача 2

$x_1, \dots, x_n \sim R(0,\ 1)$. Найти $EX_{(n)},\ EX_{(1)}$.
$F_{X_{(n)}}(x) = \big(F_\xi(x)\big)^n$
$f_{X_{(n)}}(x) = \big(F_{X_{(n)}}(x)\big)' = n\big(F_\xi(x)\big)^{n-1} \cdot f_\xi(x)$
$$F_\xi(x) = \begin{cases} 0,\ x < 0 \\ x,\ x \in [0,\ 1] \\ 1,\ x > 1 \end{cases}$$
Подставим в предыдущее уравнение:

$$f_{X_{(n)}} = \begin{cases} 0,\ x < 0 \\ nx^{n-1},\ x \in [0,\ 1] \\ 0,\ x > 1 \end{cases}$$

$$EX_{(n)} = \int\limits_{-\infty}^{+\infty} x f_{X_{(n)}}(x)\,dx = \int\limits_0^1 xnx^{n-1}\,dx = n\int\limits_0^1 x^n\,dx = \frac{n}{n+1}$$

Посчитаем для $X_{(1)}$:
$F_{X_{(1)}}(x) = 1 - (1 - F_\xi(x))^n$

$$f_{X_{(1)}}(x) = \big(F_{X_{(1)}}(x)\big)' = n(1 - F_\xi(x))^{n-1}\big(F_\xi(x)\big)' = n(1 - F_\xi(x))^{n-1} f_\xi(x) = \begin{cases} 0,\ x < 0 \\ n(1-x)^{n-1},\ 0 \leqslant x \leqslant 1 \\ 0,\ x > 1 \end{cases}$$

$$EX_{(1)} = \int\limits_0^1 xn(1-x)^{n-1}\,dx = n\int\limits_0^1 x(1-x)^{n-1}\,dx = \left\langle \begin{matrix} t = 1 - x \\ x = 1 - t \end{matrix} \right\rangle = -n\int\limits_1^0 (1-t)t^{n-1}\,dt = n\int\limits_0^1 (1-t)t^{n-1}\,dt =$$

$$= n\int\limits_0^1 t^{n-1}\,dt - n\int\limits_0^1 t^n\,dt = 1 - \frac{n}{n+1}$$

Задача 3

$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$
$E\overline{x} = E\left(\frac{1}{n}\sum_{i=1}^n x_i\right) = \frac{1}{n}\sum_{i=1}^n E(x_i) = Ex_i$

$\mathcal{D}(\overline{x}) = \mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathcal{D}x_i = \frac{\mathcal{D}x_i}{n}$
Посчитаем выборочную дисперсию:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$ES^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right) = \frac{1}{n}\sum_{i=1}^{n}E(x_i - \overline{x})^2 = \mathcal{D}(x_1 - \overline{x}) = \mathcal{D}(x_1) + \mathcal{D}(\overline{x}) - 2\,\mathsf{cov}(x_1,\ \overline{x}) = \frac{(n+1)\mathcal{D}(x_1)}{n} - $
$- 2\,\mathsf{cov}(x_1,\ \overline{x})$
$\mathsf{cov}(x_1,\ \overline{x}) = \mathsf{cov}(x_1,\ \frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\,\mathsf{cov}(x_1,\ \sum_{i=1}^{n} x_i) = \frac{1}{n}\,\mathsf{cov}(x_1,\ x_1) = \frac{\mathcal{D}(x_1)}{n}$
Тогда

$$ES^2 = \frac{(n+1)\mathcal{D}(x_1)}{n} - \frac{2\mathcal{D}(x_1)}{n} = \mathcal{D}(x_1)\left(1 - \frac{1}{n}\right)$$

Несмещённая выборачная дисперсия (её математическое ожидание равняется дисперсии x_1):

$$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Посчитаем дисперсию S^2:

$$\mathcal{D}\left(x_1 - \frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mathcal{D}\left(\frac{(n-1)x_1}{n}\right) + \mathcal{D}\left(\frac{1}{n}\sum_{i=2}^{n} x_i\right) = \frac{(n-1)^2}{n^2}\mathcal{D}(x_1) + \frac{n-1}{n^2}\mathcal{D}(x_1) =$$

$$= \mathcal{D}(x_1)\left(\frac{(n-1)(n-1+1)}{n^2}\right) = \mathcal{D}(x_1)\frac{n-1}{n}$$

Семинар 17 января.

$T(x_1, x_2, \ldots, x_n) = \sqrt{\frac{\pi}{2}}\frac{1}{n}\sum\limits_{i=1}^{n}|x_i - m|$, $x_i \sim N(m, \theta^2)$

$$ET(x_1, x_2, \ldots, x_n) = \sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}E|x_i - m| = \sqrt{\frac{\pi}{2}}\,E|x_1 - m| = \sqrt{\frac{\pi}{2}}\int\limits_{-\infty}^{+\infty}|x - m|\frac{1}{\sqrt{2\pi}\theta}e^{-\frac{(x-m)^2}{2\theta^2}}dx$$

Заменим $\frac{x-m}{\theta}$ на y

$$\frac{\theta}{2}\int\limits_{-\infty}^{+\infty}|y|\cdot e^{\frac{-y^2}{2}}dy = \theta\int\limits_{0}^{+\infty} y\cdot e^{\frac{-y^2}{2}}dy = \theta(1-0) = \theta$$

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{\pi}{2}}\,|x_i - m| \xrightarrow[n\to+\infty]{\text{п. н.}} E\sqrt{\frac{\pi}{2}}\,|x_i - m|$$

Задача

$X = (X_1, \ldots,\ X_n)$, $X_i \sim R(0,\ \theta)$
$\hat{\theta} = X_{(n)}$, доказать $\lim\limits_{n\to\infty} EX_{(n)} = \theta$
$F_{X_{(n)}}(x) = (F_{X_i}(x))^n = \left(\frac{x}{\theta}\right)^n$
$f_{X_{(n)}}(x) = \frac{dF_{X_{(n)}}}{dx} = \frac{nx^{n-1}}{\theta}$

$$EX_{(n)} = \int\limits_0^\theta \frac{nx^n}{\theta^n}\,dx = \left.\frac{nx^{n+1}}{(n+1)\theta^n}\right|_0^\theta = \frac{n}{n+1}\theta \xrightarrow[n\to\infty]{} \theta$$

. То есть смещённая, но асимптотически несмещённая.
Докажем состоятельность, хотим:

$$\forall\varepsilon > 0 \quad P(|\hat{\theta} - \theta| < \varepsilon) \xrightarrow[n\to\infty]{} 1$$

$$P(-\varepsilon < X_{(n)} - \theta < \varepsilon) = F_{X_{(n)}}(\varepsilon + \theta) - F_{X_{(n)}}(\theta - \varepsilon) = 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n \xrightarrow[n\to\infty]{} 1$$

Задача

$I_n(\theta) = E\left(\frac{\delta \ln f(x,\,\theta)}{\delta\theta}\right)^2,\ I_n(\theta) = nI_1(\theta),\ x_1, \ldots,\ x_n \sim N(\theta,\ \sigma^2).$
$f(x,\ \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$
$\ln f(x,\ \theta) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}\right) = -\frac{(x-\theta)^2}{2\sigma^2} + \ln\frac{1}{\sqrt{2\pi}\sigma}$
$\frac{\delta \ln f(x,\,\theta)}{\delta\theta} = -\frac{2(x-\theta)}{2\sigma^2}\cdot(-1) = \frac{x-\theta}{\sigma^2}$
$E\left(\frac{x-\theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^4}E(x-\theta)^2 = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2} = I_1(\theta)$
$\mathcal{D}\hat{\theta} \geqslant \frac{1}{nI_1(\theta)} = \frac{\sigma^2}{n} = \mathcal{D}\overline{x}$

Семинар 24 января

Задача 4 ДЗ

$\hat{K}_{xy} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{X})(y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{X} + Ex_1 - Ex_1)(y_i - \overline{Y} + Ey_1 - Ey_1)$
$E\hat{K}_{xy} = E\frac{1}{n}\sum_{i=1}^n \left((x_i - Ex_1) - (\overline{X} - Ex_1)\right)\left((y_1 - Ey_1) - (\overline{Y} - Ey_1)\right) =$
$= E\left((x_i - Ex_1) - (\overline{X} - Ex_1)\right)\cdot\left((y_1 - Ey_1) - (\overline{Y} - Ey_1)\right) = E\left((x_1 - Ex_1)(y_1 - Ey_1) + (x_1 - Ex_1)(\overline{Y} - Ey_1) + \right.$

$= \mathrm{cov}(x,\ y) - \frac{1}{n}\,\mathrm{cov}(x,\ y) - \frac{1}{n}\,\mathrm{cov}(x,\ y) + \frac{1}{n}\,\mathrm{cov}(x,\ y)$

Задача 5 ДЗ

Решал у доски, всем gl.

Задача 1

$X_1, \ldots,\ X_n \sim \Pi(\theta)$. Проверить, что оценка $\hat{\theta} = \overline{X}$ является R-эффективной.
$E\hat{\theta} = E\frac{1}{n}\sum_{i=1}^n x_i = Ex_1 = \theta$
$\mathcal{D}\frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\theta$
$P(\xi = x_1) = \frac{e^{-\theta}\theta^{x_1}}{x_1!}$. Логарифмируем:

$$\ln\frac{e^{-\theta}\theta^{x_1}}{x_1!} = -\theta + x_1\ln\theta - \ln x_1!$$

Возьмём частную производную:

$$\frac{\delta(-\theta + x_1\ln\theta - \ln x_1!)}{\delta\theta} = -1 + \frac{x_1}{\theta}$$

Возьмём матожидание квадрата этой величины:

$$E(-1 + \frac{x_1}{\theta})^2 = \frac{1}{\theta^2}E(x_1 - \theta)^2 = \frac{\mathcal{D}x_1}{\theta^2} = \frac{1}{\theta} \Rightarrow I_n(\theta) = \frac{n}{\theta}$$

Попробуем самостоятельно подогнать оценку:
$U(x,\ \theta) = \sum_{i=1}^n -1 + \frac{x_1}{\theta} = \frac{1}{\theta}\sum_{i=1}^n (x_i - \theta) = \frac{1}{\theta}\left(-n\theta + \sum_{i=1}^n \frac{x_i}{n}\right) = \frac{n}{\theta}\left(\sum_{i=1}^n \left(\frac{x_i}{n}\right) - \theta\right)$
$\hat{\theta} - \theta = a(\theta)U(x,\ \theta) \Rightarrow a(\theta) = \frac{\theta}{n},\ \hat{\theta} = \sum_{i=1}^n \frac{x_i}{n}$

ДЗ

Задача 1

$X_1, \ldots,\ X_n \sim N(\theta, \sigma^2) \Rightarrow \forall i = \overline{1,\ n} \quad f(x_i,\ \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$

$$\ln f(x_i,\ \theta) = \ln\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\theta)^2}{2\sigma^2} = \ln\frac{1}{\sqrt{2\pi}\theta} - \frac{x^2}{2\sigma^2} + \frac{\theta x}{\sigma^2} - \frac{\theta^2}{2\sigma^2} \Rightarrow \frac{\delta}{\delta\theta}f(x_i,\ \theta) = \frac{x}{\sigma} - \frac{\theta}{\sigma^2}$$

$$U(x,\ \theta) = \sum_{i=1}^n \left(\frac{x_i}{\sigma} - \frac{\theta}{\sigma^2}\right)$$

По критерию эффективности хотим:

$$\hat{\theta} - \theta = \alpha(x)U(x,\, \theta)$$

Преобразуем: $U(x,\, \theta) = \left(\sum\limits_{i=1}^{n} \frac{x_i}{\sigma}\right) - \frac{n\theta}{\sigma^2} \Rightarrow \underbrace{\frac{\sigma^2}{n}}_{\alpha(\sigma)} U(x,\, \theta) = \underbrace{\left(\frac{1}{n}\sum\limits_{i=1}^{n} \sigma x_i\right)}_{\hat{\theta}} - \theta$

Задача 2

$X_1, \ldots,\, X_n \sim N(m,\, \theta) \Rightarrow f(x_i,\, \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-m)^2}{2\theta}}$

$$\ln f(x,\, \theta) = \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\ln\theta - \frac{(x-m)^2}{2\theta} \Rightarrow \frac{\delta}{\delta\theta} f(x,\, \theta) = -\frac{1}{2\theta} + \frac{(x-m)^2}{2\theta^2}$$

Применим критерий эффективности:

$$U(x,\, \theta) = \sum_{i=1}^{n}\left(\frac{(x-m)^2}{2\theta^2} - \frac{1}{2\theta}\right) = \sum_{i=1}^{n}\left(\frac{(x-m)^2 - \theta}{2\theta^2}\right) = \frac{1}{2\theta^2}\sum_{i=1}^{n}\left((x-m)^2 - \theta\right) =$$

$$= \frac{1}{2\theta^2}\left(\sum_{i=1}^{n}\left((x-m)^2\right) - n\theta\right) = \frac{n}{2\theta^2}\left(\frac{1}{n}\sum_{i=1}^{n}\left((x-m)^2\right) - \theta\right) \Rightarrow \underbrace{\frac{2\theta^2}{n}}_{\alpha(\theta)} U(x,\, \theta) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(x-m)^2\right)}_{\hat{\theta}} - \theta$$

Задача 3

$X_1, \ldots,\, X_n \sim G(\theta) \Rightarrow Ex = \frac{1}{\theta}$. Проверить оценку $\hat{\theta} = \frac{1}{\overline{X}}$ на несмещённость.
Хотим $E\hat{\theta} = \theta$. Попробуем по определению:

$$E\hat{\theta} = E\frac{n}{\sum\limits_{i=1}^{n} x_i} = nE\frac{1}{\sum\limits_{i=1}^{n} x_i}?$$

Для $k = 1$ Попробуем решить через функцию правдоподобия:

$$L(x_1, \ldots,\, x_n,\, \theta) = \prod_{i=1}^{n} P(\xi = x_i,\, \theta) = \prod_{i=1}^{n}(1-\theta)^{x_i - 1}\theta \approx f(x,\, \theta)$$

$$\ln f(x_i,\, \theta) = \ln\left((1-\theta)^{x_i-1}\theta\right) = (x_i - 1)\ln(1-\theta) + \ln\theta$$

$$\frac{\delta}{\delta\theta}\ln f(x,\, \theta) = \frac{1}{\theta} - \frac{x_i - 1}{1-\theta} = \frac{1 - \theta - \theta x_i + \theta}{\theta - \theta^2} = \frac{1 - \theta x_i}{\theta - \theta^2}$$

Применим критерий эффективности:

$$U(x,\, \theta) = \sum_{i=1}^{n}\frac{1-\theta x_i}{\theta - \theta^2} = \frac{1}{\theta - \theta^2}\left(n - \theta\sum_{i=1}^{n} x_i\right) = \frac{n}{\theta - \theta^2}\left(1 - \frac{\theta}{n}\sum_{i=1}^{n}\right) = \frac{n\overline{X}}{\theta - \theta^2}\left(\frac{1}{\overline{X}} - \theta\right)$$

Значит $\frac{1}{\overline{X}}$ является R-эффективной, то есть несмещённой.

Задача 4

$X_1, \ldots,\, X_n \sim Bi(k,\, \theta)$. Показать, что $\hat{\theta} = \frac{\overline{X}}{k}$ R-эффективная.
Посчитаем функцию правдоподобия:

$$L(x_1, \ldots,\, x_n,\, \theta) = \prod_{i=1}^{n} P(\xi = x_i,\, \theta) = \prod_{i=1}^{n} C_n^k \theta^{x_i} \cdot (1-\theta)^{k - x_i} \approx f(x,\, \theta)$$

$$\ln f(x_i,\, \theta) = \ln\frac{n!}{k!(n-k)!} + x_i \ln\theta + (k - x_i)\ln(1-\theta)$$

$$\frac{\delta}{\delta\theta}\ln f(x_i,\,\theta)=\frac{x_i}{\theta}+\frac{x_i-k}{1-\theta}=\frac{x_i-\theta x_i+\theta x_i-\theta k}{\theta-\theta^2}=\frac{x_i-\theta k}{\theta-\theta^2}$$

$$I_1(\theta)=E\left(\frac{x_i-\theta k}{\theta-\theta^2}\right)^2=\int\limits_{-\infty}^{+\infty}\frac{(x-\theta k)^2}{(\theta-\theta^2)^2}C_n^k\theta^x(1-\theta)^{k-x}\,dx$$

Задача 5

$$U(x,\,\theta)=\sum_{i=1}^{n}\frac{x_i-\theta k}{\theta-\theta^2}=\frac{1}{\theta-\theta^2}\left(-n\theta k+\sum_{i=1}^{n}x_i\right)=\frac{nk}{\theta-\theta^2}\left(\frac{1}{nk}\sum_{i=1}^{n}(x_i)-\theta\right)=\frac{nk}{\theta-\theta^2}\left(\frac{\overline{X}}{k}-\theta\right)$$

Получается, что $\frac{\overline{X}}{k}$ является R-эффективной

Семинар 31 января

Задача 1

$X_1,\ldots,\,X_n\sim f(x,\,\theta)$

$f(x,\,\theta)=\begin{cases}\frac{2}{\theta}xe^{-\frac{x^2}{\theta}},\ x\geqslant 0\\ 0,\ x<0\end{cases}$

Решал у доски.

Задача 2

$X_1,\ldots,\,X_n\sim R(\theta_1,\,\theta_2)$, найти оценку максимального правдоподобия.

$f(x,\,\theta_1,\,\theta_2)=\begin{cases}\frac{1}{\theta_2-\theta_1},\ x\in(\theta_1,\,\theta_2)\\ 0,\ \text{иначе}\end{cases}$

$L(x,\,\theta_1,\,\theta_2)=\prod_{i=1}^{n}f(x_i,\,\theta)=\begin{cases}\left(\frac{1}{\theta_2-\theta_1}\right)^n,\ x_i\in(\theta_1,\,\theta_2)\\ 0,\ \text{иначе}\end{cases}$

Тогда $\hat{\theta}_1=X_{(1)},\,\hat{\theta}_2=X_{(n)}$.

Попробуем по методу моментов:

$$\begin{cases}\hat{\mu}_1=\mu_1\\ \hat{\mu}_2=\mu_2=\frac{(\theta_2-\theta_1)^2}{12}+(\mu_1)^2\end{cases}$$

Распишем эту систему:

$$\begin{cases}\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{\theta_1+\theta_2}{2}\\ \frac{1}{n}\sum_{i=1}^{n}x_i^2=\frac{(\theta_2-\theta_1)^2}{12}+\left(\frac{\theta_1+\theta_2}{2}\right)^2\end{cases}\quad\Rightarrow\begin{cases}\frac{1}{n}\sum_{i=1}^{N}X_i^2=\frac{\theta_2^2+\theta_1^2-2\theta_1\theta_2}{12}+\frac{\theta_1^2+\theta_2^2+2\theta_1\theta_2}{4}=\frac{1}{3}(\theta_1^2+\theta_2^2+\theta_1\theta_2)\end{cases}$$

$$\begin{cases}2\hat{\mu}_1=\theta_1+\theta_2\\ 3\hat{\mu}_2=\theta_1^2+\theta_2^2+\theta_1\theta_2\end{cases}$$

Если решать эту систему до конца, можно получить

$$\begin{cases}\hat{\theta}_1=\overline{X}-\sqrt{3}S\\ \hat{\theta}_2=\overline{X}+\sqrt{3}S\end{cases}$$

Задача 3

$X_1,\ldots,\,X_n\sim G(\theta)$. Найдём оценку по методу моментов и по методу максимального правдоподобия:

Сначала по методу моментов:

$$\hat{\mu}_1=\mu_1=\frac{1}{\theta}\Rightarrow\hat{\theta}=\frac{1}{\overline{X}}$$

Теперь по методу максимального правдоподобия:

$$L(x,\,\theta) = \prod_{i=1}^{n} P(\xi = x_i,\,\theta) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i - 1} = \theta^n (1-\theta)^{\sum_{i=1}^{n}(x_i) - n}$$

$$\ln L(x,\,\theta) = n \ln \theta + \left(\sum_{i=1}^{n} (x_i) - n \right) \ln(1-\theta)$$

$$\frac{\delta}{\delta\theta} L(x,\,\hat{\theta}) = \frac{n}{\hat{\theta}} - \frac{\sum_i^n (x_i) - n}{1-\hat{\theta}} = 0 \Rightarrow \frac{n - n\hat{\theta} - \hat{\theta}\sum_{i=1}^{n} + n\hat{\theta}}{\hat{\theta} - \hat{\theta}^2} = 0 \Rightarrow$$

$$\Rightarrow n - \hat{\theta} \sum_{i=1}^{n} = 0 \Rightarrow \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{X}}$$

Задача 4

$X_1 \sim Bi(12,\,p)$, $X_2 \sim Bi(12,\,p)$, $X_3 \sim Bi(15,\,p)$. По методу максимального правдоподобия построим оценку p:

$$L(x_1,\,x_2,\,x_3,\,p) = \prod_{i=1}^{n} P(X_i = x_i) = P(X_1 = 5)P(X_2 = 4)P(X_3 = 4) =$$

$$= C_{12}^5 p^5 (1-p)^7 \cdot C_{12}^4 p^4 (1-p)^8 C_{15}^4 p^4 (1-p)^{11} = C_{12}^5 \cdot C_{12}^4 \cdot C_{15}^4 \cdot p^{13} \cdot (1-p)^{26}$$

$$\ln L(x_1,\,x_2,\,x_3,\,p) = \ln(C_{12}^5 \cdot C_{12}^4 \cdot C_{15}^4) + 13 \ln p + 26 \ln(1-p)$$

$$\frac{\delta}{\delta p} L(x_1,\,x_2,\,x_3,\,p) = \frac{13}{p} - \frac{26}{1-p} \Rightarrow \frac{13}{\hat{p}} - \frac{26}{1-\hat{p}} = 0 \Rightarrow \hat{p} = \frac{1}{3}$$

ДЗ к семинару 7 января

Задача из учебника №14 стр. 203

Пусть $Z_n = (X_1, \ldots,\, X_n)$ — выборка, соответствующая биномиальному распределению $Bi(10,\,\theta)$. Оценить неизвестный параметр θ методом максимального правдоподобия.

Построим функцию правдоподобия для вектора $(X_1, \ldots,\, X_n)$:

$$L(x_1, \ldots,\, x_n,\,\theta) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} C_n^{x_i} \cdot \theta^{x_i} (1-\theta)^{n-x_i}$$

Логарифмируем и дифференцируем по θ полученное произведение:

$$\frac{\delta}{\delta\theta} \ln L(x_1, \ldots,\, x_n,\,\theta) = \frac{\delta}{\delta\theta} \ln \left(\prod_{i=1}^{n} C_n^{x_i} \cdot \theta^{x_i} (1-\theta)^{n-x_i} \right) = \frac{\delta}{\delta\theta} \sum_{i=1}^{n} \left(\ln \left(C_n^{x_i} \cdot \theta^{x_i} (1-\theta)^{n-x_i} \right) \right) =$$

$$= \frac{\delta}{\delta\theta} \sum_{i=1}^{n} (\ln C_n^{x_i}) + \frac{\delta}{\delta\theta} \sum_{i=1}^{n} (x_i \ln\theta) + \frac{\delta}{\delta\theta} \sum_{i=1}^{n} \left((n - x_i) \ln(1-\theta) \right) =$$

$$= 0 + \frac{\delta}{\delta\theta} \ln\theta \sum_{i=1}^{n} x_i + \frac{\delta}{\delta\theta} \ln(1-\theta) \sum_{i=1}^{n} (n - x_i) = \frac{1}{\theta} \sum_{i=1}^{n} x_i - \frac{1}{1-\theta} \sum_{i=1}^{n} (n - x_i) =$$

$$= \frac{1}{\theta} \sum_{i=1}^{n} x_i - \frac{n^2}{1-\theta} + \frac{1}{1-\theta} \sum_{i=1}^{n} x_i = \frac{(1-\theta) n\overline{x} - \theta n^2 + \theta n\overline{x}}{\theta - \theta^2} =$$

$$= \frac{n\overline{x} - \theta n\overline{x} - \theta n^2 + \theta n\overline{x}}{\theta - \theta^2} = \frac{n\overline{x} - \theta n^2}{\theta - \theta^2} = n \frac{\overline{x} - \theta n}{\theta - \theta^2}$$

Полученную производную стоит приравнять к 0 для поиска точки экстремума. Стоит заметить, что случаи $\theta = 0$ или $\theta = 1$ интереса не представляют и количество испытаний ненулевое, иначе оценивание параметра бессмысленно, поэтому достаточно приравнять к нулю только числитель:

$$n \frac{\overline{x} - \hat{\theta} n}{\hat{\theta} - \hat{\theta}^2} = 0 \Rightarrow \overline{x} - \hat{\theta} n = 0 \Rightarrow \hat{\theta} n = \overline{x} \Rightarrow \hat{\theta} = \frac{\overline{x}}{n}$$

Ответ: ОМП для θ является $\frac{\overline{x}}{n}$.

Задача 2

Выборка $X_1, \ldots, X_n$ порождена случайной величиной ξ с плотностью распределения

$$f_\xi(x, \theta) = \frac{1}{2} \exp(-|x - \theta|)$$

Построим оценки параметра θ по методу максимального правдоподобия и по методу моментов.

Метод максимального правдоподобия

Построим функцию правдоподобия:

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f_\xi(x_i, \theta) = \prod_{i=1}^{n} \frac{1}{2} \exp(-|x_i - \theta|) = \frac{1}{2^n} \exp\left(-\sum_{i=1}^{n} |x_i - \theta|\right)$$

Логарифмируем и продифференцируем по θ:

$$\frac{\delta}{\delta\theta} \ln L(x_1, \ldots, x_n, \theta) = \frac{\delta}{\delta\theta} \ln \frac{1}{2^n} \exp\left(-\sum_{i=1}^{n} |x_i - \theta|\right) = \frac{\delta}{\delta\theta} \ln \frac{1}{2^n} - \frac{\delta}{\delta\theta} \sum_{i=1}^{n} |x_i - \theta| =$$
$$= -\frac{\delta}{\delta\theta} \sum_{i=1}^{n} |x_i - \theta| = -\sum_{i=1}^{n} \frac{\delta}{\delta\theta} |x_i - \theta| = -\sum_{i=1}^{n} g(x_i, \theta)$$

Где $g(x, \theta) = \begin{cases} -1, & x > \theta \\ 0, & x = \theta \\ 1, & x < \theta \end{cases}$, (производная модуля).

Приравняем производную к нулю:

$$-\sum_{i=1}^{n} g(x_i, \theta) = 0 \Rightarrow \sum_{i=1}^{n} g(x_i, \theta) = 0$$

Пусть $\begin{cases} G_\theta = \{x \mid x \in (x_1, \ldots, x_n) \wedge x > \theta\} \\ E_\theta = \{x \mid x \in (x_1, \ldots, x_n) \wedge x = \theta\} \\ L_\theta = \{x \mid x \in (x_1, \ldots, x_n) \wedge x < \theta\} \end{cases}$, тогда

$$\begin{cases} \forall x \in G_\theta & g(x, \theta) = -1 \\ \forall x \in E_\theta & g(x, \theta) = 0 \\ \forall x \in L_\theta & g(x, \theta) = 1 \end{cases} \Rightarrow \sum_{i=1}^{n} g(x_i, \theta) = (-1) \cdot |G_\theta| + 0 \cdot |E_\theta| + 1 \cdot |L_\theta|$$

Преобразуем:

$$-|G_\theta| + 0|E_\theta| + |L_\theta| = 0 \Rightarrow |G_\theta| = |L_\theta|$$

То есть количество элементов больше параметра θ в выборке должно совпадать с количеством элементов меньше параметра θ.

Получается $\hat{\theta} = \begin{cases} x_{(\lfloor n/2 \rfloor)}, & n \underset{2}{\equiv} 1 \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & n \underset{2}{\equiv} 0 \end{cases}$

Метод моментов

Напишем систему уравнений для моментов (поскольку неизвестный параметр θ единственный, должно хватить одного уравнения):

$$\hat{\mu}_1 = \mu_1(\theta) \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i = E\xi$$

Посчитаем математическое ожидание случайной величины ξ:

$$E\xi = \int\limits_{-\infty}^{+\infty} x f_\xi(x,\ \theta)\, dx = \int\limits_{-\infty}^{+\infty} x \frac{1}{2} \exp\left(-|x-\theta|\right)\, dx = \left\langle \begin{matrix} a = x - \theta \\ da = dx \end{matrix} \right\rangle =$$

$$= \frac{1}{2} \int\limits_{-\infty}^{+\infty} (a+\theta) \exp(-|a|)\, da = \frac{1}{2} \underbrace{\int\limits_{-\infty}^{+\infty} a \exp(-|a|)\, da}_{=0} + \frac{\theta}{2} \int\limits_{-\infty}^{+\infty} e^{-|a|}\, da =$$

$$= \theta \int\limits_{0}^{+\infty} e^{-a}\, da = -\theta \int\limits_{0}^{+\infty} e^{-a}\, d(-a) = -\theta e^{-a} \Big|_0^{+\infty} = -\theta(0-1) = \theta$$

Итак, получаем уравнение:

$$\overline{X} = \theta$$

Его даже решать не надо, получаем $\hat{\theta} = \overline{X}$.

Ответ

По методу максимального правдоподобия: $\hat{\theta} = \begin{cases} x_{(\lfloor n/2 \rfloor)}, & n \underset{2}{\equiv} 1 \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & n \underset{2}{\equiv} 0 \end{cases}$

По методу моментов: $\hat{\theta} = \overline{X}$

Задача 3

Выборка $X_1, \ldots,\ X_n \sim \Pi(\theta) \Rightarrow \forall i \quad \begin{cases} P(X_i = k) = \frac{e^{-\theta}\theta^k}{k!} \\ EX_i = \theta \end{cases}$. Построим оценки ММ и МП для θ

Метод моментов

Снова неизвестный параметр только один, поэтому достаточно одного уравнения:

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \theta \Rightarrow \hat{\theta} = \overline{X}$$

Метод максимального правдоподобия

Функция правдоподобия:

$$L(x_1, \ldots,\ x_n,\ \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = e^{-n\theta} \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!}$$

Логарифм:

$$\ln L(x_1, \ldots,\ x_n,\ \theta) = -n\theta + \sum_{i=1}^{n} (x_i \ln \theta - \ln x_i!) = -n\theta + n\overline{X} \ln \theta - \sum_{i=1}^{n} \ln x_i!$$

Производная по θ

$$\frac{\delta}{\delta\theta} L(x_1, \ldots,\ x_n,\ \theta) = -n + \frac{n\overline{X}}{\theta}$$

Приравняем к нулю:

$$-n + \frac{n\overline{X}}{\hat{\theta}} = 0 \Rightarrow \hat{\theta} = \overline{X}$$

Ответ: оценки МП и ММ равны $\overline{X}$

Задача 4

Ученик и тренер стреляют в цель до первого попадания (геометрическое распределение). Известно, что тренер попадает в цель с вероятностью в два раза большей, чем ученик. В ходе соревнования тренер попал в цель при втором выстреле, а ученик — при пятом. Построить ОМП для вероятности попадания учеником в цель при единичном выстреле.

Пусть ξ — количество выстрелов, необходимых тренеру для попадания. Знаем $\xi \sim G(\theta_1)$.
Пусть η — количество выстрелов, необходимых ученику для попадания. Знаем $\eta \sim G(\theta_2)$.
Также знаем, что $\theta_1 = 2\theta_2$.
(Неоднородная выборка???) $(\xi,\ \eta)$ получила реализацию $(x_1,\ x_2) = (2,\ 5)$. Нужно построить оценку максимального правдоподобия для параметра θ_2.

Функция правдоподобия:

$$L(x_1,\ x_2,\ \theta_1,\ \theta_2) = P(\xi = 2) \cdot P(\eta = 5) = (1-\theta_1) \cdot \theta_1 \cdot (1-\theta_2)^4 \cdot \theta_2 = 2(1-2\theta_2) \cdot (1-\theta_2)^4 \cdot \theta_2^2$$

Логарифмируем:

$$\ln L(x_1,\ x_2,\ \theta_1,\ \theta_2) = \ln 2 + \ln(1-2\theta_2) + 4\ln(1-\theta_2) + 2\ln\theta_2$$

Продифференцируем:

$$\begin{aligned}\frac{\delta}{\delta\theta_2}\ln L(x_1,\ x_2,\ \theta_1,\ \theta_2) &= -\frac{2}{1-2\theta_2} - \frac{4}{1-\theta_2} + \frac{2}{\theta_2} = \frac{2(1-2\theta_2)(1-\theta_2) - 4\theta_2\cdot(1-2\theta_2) - 2\theta_2\cdot(1-\theta_2)}{(\theta_2 - 2\theta_2^2)(1-\theta_2)} = \\ &= \frac{2(1-3\theta_2+2\theta_2^2) - 4(\theta_2-2\theta_2^2) - 2(\theta_2-\theta_2^2)}{\theta_2 - 3\theta_2^2 + 2\theta_2^3} = \frac{2-6\theta_2+4\theta_2^2-4\theta_2+8\theta_2^2-2\theta_2+2\theta_2^2}{\theta_2-3\theta_2^2+2\theta_2^3} = \\ &= \frac{14\theta_2^2 - 12\theta_2 + 2}{2\theta_2^3 - 3\theta_2^2 + \theta_2}\end{aligned}$$

Приравняем к нулю:

$$\frac{14\hat\theta_2^2 - 12\hat\theta_2 + 2}{2\hat\theta_2^3 - 3\hat\theta_2^2 + \hat\theta_2} = 0 \Rightarrow 14\hat\theta_2^2 - 12\hat\theta_2 + 2 = 0 \Rightarrow 7\hat\theta_2^2 - 6\hat\theta_2 + 1 = 0 \Rightarrow \mathcal{D}' = 9 - 7 = 2 \Rightarrow \left[\begin{array}{l}\hat\theta_2 = \dfrac{3+\sqrt{2}}{7} \Rightarrow \theta_1 > 1 \\ \hat\theta_2 = \dfrac{3-\sqrt{2}}{7}\end{array}\right.$$

Ответ: $\hat\theta_2 = \frac{3-\sqrt{2}}{7} \approx 0.22654$

Задача 5

Выборка $X_1, \ldots,\ X_n$ порождена случаной величиной X с плотностью распределения:

$$f(x,\ \theta) = \begin{cases}\frac{1}{\theta}x^{\frac{1-\theta}{\theta}}, & x \in (0,\ 1) \\ 0, & x \notin (0,\ 1)\end{cases}$$

Построим оценку максимального правдоподобия для параметра θ и исследуем его на несмещённость.
Построим функцию правдоподобия:

$$L(x_1, \ldots,\ x_n,\ \theta) = \prod_{i=1}^{n} f(x_i,\ \theta) = \theta^{-n}\prod_{i=1}^{n} x_i^{\frac{1-\theta}{\theta}}$$

Логарифмируем функцию правдоподобия:

$$\ln L(x_1, \ldots,\ x_n,\ \theta) = \sum_{i=1}^{n}\left(\frac{1-\theta}{\theta}\ln x_i\right) - n\ln\theta = \frac{1}{\theta}\sum_{i=1}^{n}(\ln x_i) - \sum_{i=1}^{n}(\ln x_i) - n\ln\theta$$

Продифференцируем логарифм по θ:

$$\frac{\delta}{\delta\theta}L(x_1, \ldots,\ x_n,\ \theta) = -\frac{n}{\theta} - \frac{1}{\theta_2}\sum_{i=1}^{n}\ln x_i = \frac{-n\theta - \sum\limits_{i=1}^{n}\ln x_i}{\theta^2}$$

Приравняем к нулю:

$$-n\hat{\theta} - \sum_{i=1}^{n} \ln x_i = 0 \Rightarrow \hat{\theta} = -\frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

Проверим на несмещённость:

$$E\hat{\theta} = -E\ln x_1 = -\int\limits_0^1 \ln(x)\cdot\frac{1}{\theta}x^{\frac{1-\theta}{\theta}}\,dx = \left\langle \begin{matrix} a = x^{\frac{1}{\theta}},\ \dfrac{d}{dx}x^{\frac{1}{\theta}} = \dfrac{1}{\theta}x^{\frac{1}{\theta}-1} \\ da = \dfrac{1}{\theta}x^{\frac{1-\theta}{\theta}}dx,\ x = a^\theta \end{matrix} \right\rangle = -\int\limits_{0^{\frac{1}{\theta}}}^{1^{\frac{1}{\theta}}} \ln(a^\theta)\,da =$$

$$= -\theta\int\limits_0^1 \ln(a)\,da = -\theta\,(a\ln a - a)\Big|_0^1 = \theta$$

Несмещённая.

Семинар 7 февраля

$X_1, \ldots, X_n \sim F(x,\,\theta)$. Считается, что $\big(T_1(x_1, \ldots, x_n),\, T_2(x_1, \ldots, x_n)\big)$ является доверительным интервалом уровня $1-\alpha$, если:

$$P\big(T_1(x_1, \ldots, x_n) < \theta < T_2(x_1, \ldots, x_n)\big) = 1-\alpha$$

Например, для $X_1, \ldots, X_n \sim N(m,\,\sigma^2)$, σ известна.
$\hat{m} = \overline{X},\ \mathcal{D}\overline{X} = \frac{\sigma^2}{n} \Rightarrow \frac{\sqrt{n}(\overline{X}-m)}{\sigma} \sim N(0,\,1)$. Для построения доверительного интервала нужно оценить вероятность попадания опорной статистики на интервал:

$$P\left(Z_{\alpha/2} < \frac{\sqrt{n}(\overline{X}-m)}{\sigma} < Z_{1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\overline{X} - \frac{\sigma Z_{1-\alpha/2}}{\sqrt{n}} < m < \overline{X} + \frac{\sigma Z_{1-\alpha/2}}{\sqrt{n}}\right) = 1-\alpha$$

Если σ тоже неизвестна, то подставляем её оценку $\tilde{S} = \frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\overline{X})^2$ и получаем распределение Стьюдента, значит стоит брать его квантили.

$$\frac{\sqrt{n}(\overline{X}-m)}{\tilde{S}} = \frac{\sqrt{n}(\frac{\overline{X}-m}{\sigma})}{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{X}}{\sigma}\right)^2}$$

То есть стандартное гауссовское делим на корень из χ^2.
Итого:

$$P\left(\overline{X} - \frac{\tilde{S}t_{1-\alpha/2,\,n-1}}{\sqrt{n}} < m < \overline{X} + \frac{\tilde{S}t_{1-\alpha/2,\,n-1}}{\sqrt{n}}\right) = 1-\alpha$$

Если математическое ожидание известно, но мы хотим интервал для дисперсии:

$$\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^2} \sim \chi^2(n)$$

$$P\left(\chi^2_{n,\,1-\alpha/2} < \frac{\sum(x_i-m)^2}{\sigma^2} < \chi^2_{n,\,1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\frac{\sum_{i=1}^{n}(x_i-m)^2}{\chi^2_{n,\,1-\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i-m)^2}{\chi^2_{n,\,1-\alpha/2}}\right) = 1-\alpha$$

Если неизвестны оба:

$$\sum_{i=1}^{n}\frac{(x_i-\overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

$$P\left(\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{\chi^2_{n-1,\,1-\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{\chi^2_{n-1,\,1-\alpha/2}}\right) = 1-\alpha$$

Задача

Импортёр упаковывает чай в пакеты с номинальным весом 125 грамм. Известно, что упаковочная машина работает с известным среднеквадратическим отклонением 10 грамм. Выбрали 50 пакетов чая, выборочное среднее их веса оказалось равно $125,8$.
То есть $n = 50,\ \overline{X} = 125,8,\ X_1, \ldots,\ X_n \sim N(m,\ 100)$.
$\overline{X} \sim N\left(m,\ \frac{\sigma^2}{n}\right) \Rightarrow \frac{\sqrt{n}(\overline{X}-m)}{\sigma} \sim N(0,\ 1) \Rightarrow$

$$P\left(Z_{0,025} < \frac{\sqrt{n}(\overline{X}-m)}{\sigma} < Z_{0,95}\right) = 0,95$$

$$P\left(\overline{X} - \frac{\sigma Z_{0,95}}{\sqrt{n}} < m < \overline{X} + \frac{\sigma Z_{0,95}}{\sqrt{n}}\right) = 0,95$$

$$P(123,028 < m < 128,571) = 0,95$$

125 лежит в этом интервале, поэтому всё хорошо.
Длина интервала получается $\frac{2\sigma Z_{0,95}}{\sqrt{n}}$, хотим, чтбы это равнялось 2

$$\sqrt{n} = \sigma Z_{0,95} \Rightarrow n \approx 384$$

ДЗ 14 февраля

Задача 1

10 изделий сделано за 79, 74, 112, 95, 83, 96, 77, 84, 70, 90 минут. Построить ДИ уровня 0.95 для среднего времени сборки.
Получаем $X_i \sim N(m,\ \sigma)$, просят доверительный интервал для m. С прошлого семинара:

$$P\left(\overline{X} - \frac{\tilde{S}t_{1-\alpha/2,\ n-1}}{\sqrt{n}} < m < \overline{X} + \frac{\tilde{S}t_{1-\alpha/2,\ n-1}}{\sqrt{n}}\right) = 1-\alpha$$

Здесь:

$$n = 10$$
$$\alpha = 0.05$$
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$
$$t_{1-\alpha/2,\ n-1} = \text{так и не понял где посмотреть (квантиль распределения Стьюдента)}$$

Задача 2

Теперь ДИ для дисперсии уровня 0.9, опять воспользуемся записями семинара:

$$P\left(\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{\chi^2_{n-1,\ 1-\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{\chi^2_{n-1,\ 1-\alpha/2}}\right) = 1-\alpha$$

Задача 3

Тоже построить ДИ для матожидания и дисперсии гауссовской величины, только с другими значениями. Из сложностей только $\tilde{S}^2 = \frac{n}{n-1}S^2$

Задача 4

Показать, что $S^2 = \hat{\mu}_2 - (\hat{\mu}_1)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\sum_{i=1}^{n} x_i^2 - \frac{2}{n^2}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2\overline{X}}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2$$

Семинар 14 февраля

Даны две выборки:

$$\begin{cases} X_1, \ldots, X_n \sim N(m_1,\, \sigma_1^2) \\ Y_1, \ldots, Y_n \sim N(m_2,\, \sigma_2^2) \end{cases}$$

$\sigma_1,\ \sigma_2$ известны, тогда для построения ДИ $\theta = m_1 - m_2$:

$$\frac{\overline{X} - \overline{Y} - \theta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,\, 1)$$

Если дисперсии неизвестны, но одинаковы:

$$\hat{\mathcal{D}}(\overline{X} - \overline{Y}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Дисперсию не знаем, поэтому подставим оценку:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2 + \sum_{i=1}^{n}(y_i - \overline{Y})^2}{n_1 + n_2 - 2}$$

Тогда можно сказать

$$\frac{\overline{X} - \overline{Y} - \theta}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

Задача

$\overline{X} = -11.87,\ \overline{Y} = -13.75,\ \sigma_1^2 = 20,\ \sigma_2^2 = 22,\ n_1 = n_2 = 13,\ \alpha = 0.05$
$X \sim N(m_1,\, 20),\ Y \sim N(m_2,\, 22)$ Знаем матожидания и дисперсию, тогда ДИ для $\theta = m_2 - m_1$

$$P\left((\overline{X} - \overline{Y}) - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leqslant \theta \leqslant (\overline{X} - \overline{Y}) + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 0.95$$

$$P\left(-1.64 \leqslant \theta \leqslant 5.4\right) = 0.95$$

Модифицирем задачу. σ теперь неизвестны, но мы считаем их одинаковыми, тогда

$$P\left((\overline{X} - \overline{Y}) - 2.06S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leqslant \theta \leqslant (\overline{X} - \overline{Y}) + 2.06S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 0.95$$

Если у нас посчитано ${S_X}^2$ и S_Y^2, то можем посчитать S:

$$S^2 = \frac{n_1 S_X^2 + n_2 S_Y^2}{n_1 + n_2 - 2}$$

Если посчитать, то получаем

$$P\left(-1.98 \leqslant \theta \leqslant 5.74\right) = 0.95$$

Задача

$X_1, \ldots, X_n \sim \Pi(\theta)$. Построим асимптотический доверительный интервал.
Для распределения Пуассона верно: $\hat{\theta} = \overline{X},\ \mathcal{D}\overline{X} = \frac{\sigma^2}{n} = \frac{\theta}{n}$
Тогда при больших n:

$$\frac{(\hat{\theta} - \theta)}{\sqrt{\frac{\hat{\theta}}{n}}} \sim N(0,\ 1)$$

$$P\left(Z_{1-\alpha/2} \leqslant \frac{(\overline{X} - \theta)}{\sqrt{\frac{\overline{X}}{n}}} \leqslant Z_{1-\alpha/2}\right) = 1 - \alpha$$

$$P\left(\sqrt{\frac{n}{\overline{X}}} Z_{1-\alpha/2} \leqslant (\overline{X} - \theta) \leqslant \sqrt{\frac{n}{\overline{X}}} Z_{1-\alpha/2}\right) = 1 - \alpha$$

$$P\left(\overline{X} - \sqrt{\frac{n}{\overline{X}}} Z_{1-\alpha/2} \leqslant \theta \leqslant \overline{X} + \sqrt{\frac{n}{\overline{X}}} Z_{1-\alpha/2}\right) = 1 - \alpha$$

ДЗ на 21 февраля

Задача 1

Имеются данные о доходах Центрального федерального округа:

$$10043;\ 9596;\ 10305;\ 8354;\ 9413;\ 19776;\ 9815;\ 11311;\ 11253;\ 10856;\ 11389 \Rightarrow n_x = 11,\ \overline{X} = 11\,101$$

И Приволжского федерального округа:

$$14253;\ 7843;\ 9581;\ 8594;\ 16119;\ 10112;\ 10173;\ 9756 \Rightarrow n_y = 8,\ \overline{Y} = 10\,803.875$$

Построить ДИ уровня 0.95 для разности значений среднедушевных доходов населения Центрального и Приволжского федеральных округов. Предполагается, что все наблюдения имеют гауссовское распределение и одинаковые дисперсии.
$X \sim N(m_1,\ \sigma^2)$ (доход в ЦФО), $Y \sim N(m_2,\ \sigma^2)$ (доход в ПФО). Оценим величину $m = m_1 - m_2$
Для гауссовских величин хорошей оценкой m будет величина $\overline{X} - \overline{Y}$.

$$E\left(\overline{X} - \overline{Y}\right) = m$$

$$\hat{\mathcal{D}}\left(\overline{X} - \overline{Y}\right) = \sigma^2 \left(\frac{1}{n_x} + \frac{1}{n_y}\right)$$

σ не знаем, подставим оценку $S^2 = \dfrac{\sum\limits_{i=1}^{n_x}(x_i - \overline{X})^2 + \sum\limits_{i=1}^{n_y}(y_i - \overline{Y})^2}{n_x + n_y - 2}$

Теперь мы можем составить хорошую случайную величину:

$$\frac{\left(\overline{X} - \overline{Y}\right) - m}{\sqrt{\hat{\mathcal{D}}\left(\overline{X} - \overline{Y}\right)}} = \frac{\overline{X} - \overline{Y} - m}{S\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t(n_x + n_y - 2) = t(17)$$

Сейчас сделаю фокус, чтобы было понятнее, почему это Стьюдент:

$$\frac{\overline{X} - \overline{Y} - m}{S\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \frac{\frac{\overline{X} - \overline{Y} - m}{\sigma}}{\frac{S}{\sigma}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

Далее во избежание страшных дробей я распишу числитель и знаменатель отдельно. Начнём с числителя:

$$\frac{\overline{X} - \overline{Y} - m}{\sigma} : \begin{cases} E\left(\frac{\overline{X} - \overline{Y} - m}{\sigma}\right) = 0 \\ \mathcal{D}\left(\frac{\overline{X} - \overline{Y} - m}{\sigma}\right) = 1 \end{cases} \Rightarrow \frac{\overline{X} - \overline{Y} - m}{\sigma} \sim N(0,\ 1)$$

Теперь знаменатель:

$$\frac{S}{\sigma}\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}=\frac{\sqrt{\sum\limits_{i=1}^{n_x}(x_i-\overline{X})^2+\sum\limits_{i=1}^{n_y}(y_i-\overline{Y})^2}}{\sigma\sqrt{n_x+n_y-2}}\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}=\frac{\sqrt{\sum\limits_{i=1}^{n_x}\frac{(x_i-\overline{X})^2}{\sigma^2}+\sum\limits_{i=1}^{n_y}\frac{(y_i-\overline{Y})^2}{\sigma^2}}}{\sqrt{n_x+n_y-2}}\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}=$$

$$=\frac{\sqrt{\sum\limits_{i=1}^{n_x}\left(\frac{x_i-\overline{X}}{\sigma}\right)^2+\sum\limits_{i=1}^{n_y}\left(\frac{y_i-\overline{Y}}{\sigma}\right)^2}}{\sqrt{n_x+n_y-2}}\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$$

Это сумма квадратов центрированных и нормированных гауссовских величин, то есть знаменатель распределён по χ^2.
Получается, что наша случайная величина получается в результате деления $N(0,\ 1)$ на χ^2, то есть это по определению распределение Стьюдента.
Перед построением доверительного интервала введём обозначение $\tau = t_{17,\ 0.975} = -t_{17,\ 0.025} \approx 2.11$.

$$P\left(-\tau<\frac{\overline{X}-\overline{Y}-m}{S\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}<\tau\right)=0.95$$

$$P\left(\left(\overline{X}-\overline{Y}\right)-\tau S\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}<m<\left(\overline{X}-\overline{Y}\right)+\tau S\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}\right)=0.95$$

$$P\left(-2\,446.617<m<3\,040.867\right)=0.95$$

Задача 2

Для проверки качества деталей из большой партии выбрали 200 деталей. Среди них оказалось 12 бракованных. Построить асимптотический доверительный интервал уровня надёжности 0.95 для доли бракованных деталей.
Полагаем, что количество бракованных деталей имеет распределение $Bi(200,\ p)$, где p и будет искомой долей бракованных деталей. Оценкой максимального правдободоия для p является $\hat{p} = \frac{\overline{X}}{n}$ (было в домашке за 7 января). 200 тяжело сосчитать на пальцах, поэтому считаем его достаточно большим, чтобы применить теорему Муавра-Лапласа:

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\sim N(0,\ 1)$$

Теперь можно очень просто построить доверительный интервал ($z = Z_{0.975} = -Z_{0.025} = 1.96$):

$$P\left(-z<\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}<z\right)=0.95$$

$$P\left(\hat{p}-z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}<p<\hat{p}+z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)=0.95$$

$$P\left(0.027<p<0.093\right)=0.95$$

Задача 3

$X\sim Bi(n_1,\ p_1),\ Y\sim Bi(n_2,\ p_2)$. По условию $n_1,\ n_2$ большие. Построить асимптотический доверительный интервал для $p = p_1 - p_2$. (Показать, что статистика $\frac{\hat{p}_1-\hat{p}_2-(p_1-p_2)}{\sqrt{\hat{\mathcal{D}}(\hat{p}_1-\hat{p}_2)}}$, где $\hat{\mathcal{D}}\left(\hat{p}_1-\hat{p}_2\right)=\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$, имеет

асимптотически стандартное нормальное распределение$\Big)$.
Найдём оценку максимального правдоподобия для p:

$$\xi = X - Y \Rightarrow \begin{cases} P(\xi = 1) = P(X = 1) \cdot P(Y = 0) = p_1 q_2 \\ P(\xi = 0) = P(X = 1) \cdot P(Y = 1) + P(X = 0) \cdot P(Y = 0) = p_1 p_2 + q_1 q_2 \\ P(\xi = -1) = P(X = 0) \cdot P(Y = 1) = q_1 p_2 \end{cases}$$

Построим функцию правдоподобия для реализации вектора $(Z_1, \ldots, Z_n)$, порождённого случайной величиной ξ:

$$L(z_1, \ldots, z_n, p) = \prod_{i=1}^{n} P(\xi = z_i)$$

Дальше непонятно, значит всё-таки надо воспользоваться подсказкой. Обозначим $T(\hat{p}_1 - \hat{p}_2) = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)}}$
Из предыдущей задачи:

$$\hat{p}_1 = \frac{\overline{X}}{n_1},\ \hat{p}_2 = \frac{\overline{Y}}{n_2}$$

$E(\hat{p}_1 - \hat{p}_2) = E\hat{p}_1 - E\hat{p}_2 = p_1 - p_2$
$\mathcal{D}(p_1 - p_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
$\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2) = \hat{\mathcal{D}}(\hat{p}_1) + \hat{\mathcal{D}}(\hat{p}_2)$
Тогда при больших $n_1,\ n_2$ (в нашем случае это так) должно выполняться:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)}} \sim N(0,\ 1)$$

Что-то очень странное, надо будет уточнить на семинаре.
Если это верно, тогда доверительный интервал уровня $1 - \alpha$ выглядит так:

$$P\left(\hat{p}_1 - \hat{p}_2 - Z_{1-\alpha/2}\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + Z_{1-\alpha/2}\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)}\right) = 1 - \alpha$$

Задача 4

Два года назад у 252 студентов было 29 неудов. В прошлом году у 286 оказалось 42 неуда. Построить доверительный интервал уровня надёжности 0.95 для разности вероятностей неудов в этих двух выборках.
Если пользоваться результатом задачи 3:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)}} \sim N(0,\ 1)$$

Где $\hat{p}_1 = \frac{29}{252},\ \hat{p}_2 = \frac{42}{286},\ n_1 = 252,\ n_2 = 286$, тогда:

$$P\left(-z < \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)}} < z\right) = 0.95$$

$$P\left(\hat{p}_1 - \hat{p}_2 - z\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z\sqrt{\hat{\mathcal{D}}(\hat{p}_1 - \hat{p}_2)}\right) = 0.95$$

$$P(-0.087 < p_1 - p_2 < 0.025) = 0.95$$

Задача 5

Из 500 опрошенных клиентов магазина 100 человек довольны обслуживанием. Построить асимптотический доверительный интервал уровня надёжности 0.95 для доли покупателей, довольных обслуживанием.

Полагаем, что количество довольных клиентов распределено как $Bi(500,\ p)$.
Считаем 500 ОГРОМНЫМ числом, поэтому:

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,\ 1)$$

Здесь $\hat{p} = \frac{100}{500}$, получаем:

$$P\left(-z < \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z\right) = 0.95$$

$$P\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

$$P\left(0.165 < p < 0.235\right) = 0.95$$

Задача 6

Из 400 опрошенных клиентов другого магазина 70 человек довольны обслуживанием. Построить асимптотический доверительный интервал уровня надёжности 0.98 для разности долей довольных клиентов (в этой задаче и предыдущей).
Пользуясь результатом задачи 3 получаем:

$$P\left(\hat{p}_1 - \hat{p}_2 - Z_{0.99}\sqrt{\hat{\mathcal{D}}\left(\hat{p}_1 - \hat{p}_2\right)} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + Z_{0.99}\sqrt{\hat{\mathcal{D}}\left(\hat{p}_1 - \hat{p}_2\right)}\right) = 0.98$$

Здесь $\hat{p}_1 = \frac{100}{500},\ \hat{p}_2 = \frac{70}{400},\ Z_{0.99} \approx 2.326,\ \sqrt{\hat{\mathcal{D}}\left(\hat{p}_1 - \hat{p}_2\right)} \approx 0.026$, подставим и получим:

$$P\left(-0.036 < p_1 - p_2 < 0.086\right) = 0.98$$

Семинар 21 февраля

Задача

Из 200 деталей 12 бракованных. Проверить гипотезу о том, что 5% деталей бракованные.
$X_1, \ldots,\ X_{200} \sim Bi(1,\ p)$
$H_0 : p = 0.05 = p_0$ против $H_1 : p > 0.05$ при $\alpha = 0.05$.
$X = \sum_{i=1}^{200} x_i$ — количество успехов
$T(x) = \frac{x - 200 \cdot p_0}{\sqrt{200 p_0 (1-p_0)}}$
$T(x)\big|_{H_0} \sim N(0,\ 1)$
Тогда доверительная область $(-\infty,\ Z_{0.95}) = (-\infty,\ 1.64)$
$T(x) = \frac{12-10}{\sqrt{200 \cdot 0.05 \cdot 0.95}} = 0.649 \Rightarrow$ попали в доверительную область, значит верим H_0.

Задача

Проводится тестирование по английскому языку. Предлагается 100 вопросов, на каждый из которых 4 ответа, 1 из них правильный. Один студент ответил правильно на 30 вопросов. Можно ли считать при $\alpha = 0.05$, что этот студент не знает английский язык?
$X_1, \ldots,\ X_{100} \sim Bi(1,\ p)$
$H_0 : p = 0.25 = p_0$ (то есть студент угадывает ответы $\Rightarrow$ не знает).
$H_1 : p > 0.25$
Статистику возьмём такую же, как в прошлой задаче $T(x) = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$. Тогда:

$$T(x)\big|_{H_0} \sim N(0,\ 1)$$

Критической областью является $(Z_{0.95},\ +\infty)$. Теперь посчитаем статистику:

$$T(30) = \frac{30 - 25}{\sqrt{100 \cdot 0.25 \cdot 0.75}} = 1.15 < Z_{0.95}$$

Статистика попала в доверительную область, значит студент не знает английский язык.

Задача

Проведено исследование по выведению факторов риска заболеваемости туберкулёзом. Одним из факторов считается низкий доход в семье. Среди 300 семей с низким доходом 12 больных, среди 100 семей с высоким доходом 2 больных. Можно ли сказать, что низкий доход влияет на заболеваемость.

ДЗ к 28 февраля

Задача 1

Для прохода в парламент необходимо 7% голосов избирателей. Опросили 1 000 человек, 68 из которых собираются голосовать за партию A. Можно ли на уровне значимости 0.05 считать, что партия A пройдёт в парламент?

Решение

Считаем, что для случайной величины $\xi = \{$"Количество проголосовавших за A"$\}$ справедливо

$$\xi \sim Bi(1\,000,\ p)$$

Составим две гипотезы про полученное распределение:

$$H_0 : p = p_0 = 0.07, \text{ против } H_1 : p < 0.07$$

Если верна H_0, то мы можем считать, что партия прошла в парламент.
Уровень значимости указан в задаче и равен 0.05.
Выбираем следующую статистику:

$$T(x) = \frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{1\,000}}}$$

В общем случае про эту статистику мы ничего сказать не можем, но при условии верности H_0 она становится центрированной и нормированной, потому что:

$$\overline{X} = \hat{p}$$
$$E\left(\hat{p}\right)\big|_{H_0} = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = Ex_i = p_0$$
$$\mathcal{D}\left(\hat{p}\right)\big|_{H_0} = \mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathcal{D}\left(x_i\right) = \frac{\mathcal{D}(x_i)}{n} = \frac{p_0(1-p_0)}{n} = \frac{p_0(1-p_0)}{1\,000}$$

В этих уравнениях я преобразовал ξ в случайный вектор X состоящий из $1\,000$ случайных величин $Be(p_0)$.
Итак, $T(x)$ действительно центрированная и нормированная случайная величина, значит

$$T(x) \xrightarrow[n\to\infty]{} U \sim N(0,\ 1)$$

В нашем случае $\overline{X} = 68$, посчитаем эту статистику:

$$T(x) = \frac{\frac{68}{1\,000} - 0.07}{\sqrt{\frac{p_0(1-p_0)}{1\,000}}} = -0.248$$

Доверительным интервалом в нашем случае является $(Z_{0.05},\ +\infty)$, где $Z_{0.05} = -1.64$.
Получаем $T(68) \in (-1.64,\ +\infty) \Rightarrow$ принимаем H_0

Ответ

Да, можно так считать.

Задача 2

Известно, что женщины-водители составляют 30% от общего числа водителей. Зафиксировали $n = 635$ ДТП, 132 из которых произошли по вине женщин-водителей. Можно ли на уровне значимости 0.01 считать, что женщины водят машину аккуратнее (реже попадают в ДТП)?

Решение

Пусть случаный вектор X, каждый элемент которого $x \sim Be(p)$ равен 1, если причиной ДТП была женщина и 0 иначе. Составим две гипотезы

$$H_0 : p = p_0 = 0.3, \text{ против } H_1 : p < 0.3$$

Принятие H_0 означает, что женщины водят не аккуратнее мужчин (отрицательный ответ на вопрос задачи).
Требуемый уровень значимости: 0.01
Составим статистику:

$$T(x) = \frac{\overline{X} - np_0}{\sqrt{np_0q_0}}$$

Если считать H_0 верной, то $np_0 = EX$, а $\sqrt{np_0q_0} = \sqrt{\mathcal{D}X}$, значит:

$$T(x)\big|_{H_0} \xrightarrow[n\to\infty]{} U \sim N(0,\ 1)$$

В нашем случае $\overline{X} = \{\text{“Количество ДТП из-за женщин”}\} = 132$.
Доверительный интервал $(Z_{0.01},\ +\infty) \approx (-2.326,\ +\infty)$.

$$T(x) = \frac{132 - 635 \cdot 0.3}{\sqrt{635 \cdot 0.3 \cdot 0.7}} \approx -5.066$$

Попали в критическую область, значит принимает альтернативную гипотезу.

Ответ

Да, можно так считать.

Задача 3

Есть два пресса, штампующих одинаковые детали. Из $n_1 = 1\,000$ деталей первого пресса оказалось 25 бракованных. Из $n_2 = 800$ деталей второго пресса оказалось 18 бракованных. Можно ли на уровне значимости 0.01 считать, что доля брака у этих прессов одинакова?

Решение

Объявим две случайные величины $\xi \sim Be(p_1),\ \eta \sim Be(p_2)$.
Случаный вектор X, порождённый 1 000 случаных величин ξ, и случаный вектор Y, порождённый 800 случайными величинами η.
Гипотезы:

$$H_0 : p_1 = p_2, \text{ против } H_1 : p_1 \neq p_2$$

Принятие H_0 означает положительный ответ на задачу.
Требуемый уровень значимости 0.01
Перефразируем гипотезы:

$$H_0 : p_1 - p_2 = 0, \text{ против } H_1 : p_1 - p_2 \neq 0$$

Теперь можно составить статистику:

$$T(x,\ y) = \frac{\overline{X} - \overline{Y} - (p_1 - p_2)}{\sqrt{\mathcal{D}(\overline{X} - \overline{Y})}}$$

Если считать H_0 верной, можно сделать следующие преобразования:

$$p_1 - p_2 = 0;$$

$$\mathcal{D}(\overline{X} - \overline{Y}) = \mathcal{D}(\overline{X}) + \mathcal{D}(\overline{Y}) = \frac{\left(\mathcal{D}(\xi)\right)^2}{n_1} + \frac{\left(\mathcal{D}(\eta)\right)^2}{n_2} = \left\langle \begin{matrix} H_0 : p_1 = p_2 \Rightarrow \xi \sim \eta \\ \mathcal{D}(\xi) = \mathcal{D}(\eta) = \sigma^2 \end{matrix} \right\rangle =$$

$$= \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right);$$

$$\sigma^2 = (n_1 + n_2)\hat{p}\,(1 - \hat{p}) = (n_1 + n_2)\frac{\overline{X} + \overline{Y}}{n_1 + n_2}\left(1 - \frac{\overline{X} + \overline{Y}}{n_1 + n_2}\right) = \left(\overline{X} + \overline{Y}\right)\left(1 - \frac{\overline{X} + \overline{Y}}{n_1 + n_2}\right);$$

$$\mathcal{D}\left(\overline{X} - \overline{Y}\right) = \left(\overline{X} + \overline{Y}\right)\left(1 - \frac{\overline{X} + \overline{Y}}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = (25 + 18)\left(1 - \frac{25 + 18}{1\,000 + 800}\right)\left(\frac{1}{1\,000} + \frac{1}{800}\right) \approx$$

$$\approx 0.0944 \Rightarrow \sqrt{\mathcal{D}\left(\overline{X} - \overline{Y}\right)} \approx 0.307$$

Снова получаем что-то сходящееся к центрированной и нормированной гауссовской величине:

$$T(x,\, y) = \frac{\overline{X} - \overline{Y} - 0}{\sqrt{\mathcal{D}\left(\overline{X} - \overline{Y}\right)}} \xrightarrow[n \to \infty]{} U \sim N(0,\, 1)$$

При подстановке наших чисел получаем:

$$T(x,\, y) = \frac{\frac{25}{1000} - \frac{18}{800}}{0.307} \approx 0.008$$

Доверительный интервал в данной задаче $(Z_{0.005},\, Z_{0.995}) = (-2.576,\, 2.576)$. Статистика попала в доверительный интервал, значит мы верим H_0

Ответ

Да, можно

Задача 4

В Москве 66 человек из $n_1 = 600$ недовольны своей работой. В Московской области 60 человек из $n_2 = 500$ недовольны своей работой. Можно ли на уровне значимости 0.05 считать, что в области доля недовольных выше?

Решение

Вспомним товарища Дашкова:
$X = (x_1, \ldots,\, x_{600}),\ \forall x \quad x \in X \Rightarrow x \sim Be(p_1)$
$Y = (y_1, \ldots,\, y_{500}),\ \forall y \quad y \in Y \Rightarrow y \sim Be(p_2)$
Гипотезы:

$$H_0 : p_1 = p_2,\ \text{против}\ H_1 : p_1 < p_2$$

Принятие H_0 влечёт отрицательный ответ на задачу.
Уровень значимости 0.05
Снова пошаманим с гипотезами:

$$H_0 : p_1 - p_2 = 0,\ \text{против}\ H_1 : p_1 - p_2 < 0$$

Составляем статистику:

$$T(x,\, y) = \frac{\overline{X} - \overline{Y} - (p_1 - p_2)}{\sqrt{\mathcal{D}\left(\overline{X} - \overline{Y}\right)}}$$

Это идентично предыдущей задаче, поэтому сразу запишу всё получаемое при условии верности H_0:

$$p_1 - p_2 = 0;$$
$$\mathcal{D}\left(\overline{X} - \overline{Y}\right) = \left(\overline{X} + \overline{Y}\right)\left(1 - \frac{\overline{X} + \overline{Y}}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) =$$
$$= (66 + 60)\left(1 - \frac{66 + 60}{600 + 500}\right)\left(\frac{1}{600} + \frac{1}{500}\right) \approx 0.409;$$
$$\sqrt{\mathcal{D}\left(\overline{X} - \overline{Y}\right)} \approx 0.634$$

Думаю уже понятно, что полученная нами центрированная и нормированная статистика сходится к $N(0,\ 1)$. Подставим числа в статистику:

$$T(x,\ y) = \frac{\frac{66}{600} - \frac{60}{500}}{0.634} \approx -0.016$$

Доверительный интервал в нашем случае $(Z_{0.05},\ +\infty) \approx (-1.645,\ +\infty)$. Статистика лежит в доверительном интервале, значит верна $H_0 \Rightarrow$ ответ на задачу отрицательный.

Ответ

Нет, нельзя

Задача 5

Вероятность рождения мальчика $p_0 = 0.52$, в случайной выборке из $n = 5\,000$ людей от 30 до 60 лет оказалось 2 500 мужчин и 2 500 женщин. Можно ли на уровне значимости 0.05 считать, что смертность мужчин и женщин одинакова.

Решение

$X = (x_1, \ldots,\ x_{5\,000}),\ \forall x \in X \quad x \sim Be(p)$, в таком случае $\overline{X}$ — частота события {“Встретить мужчину”}
Гипотезы:

$$H_0 : p = p_0 = 0.52, \text{ против } H_1 : p < 0.52$$

Принятие H_0 означает, что смертность одинакова, то есть положительный ответ на задачу. Уровень значимости 0.05
Составим статистику:

$$T(x) = \frac{\overline{X} - p_0}{\sqrt{\mathcal{D}\left(\overline{X}\right)}}$$

При верности H_0 можем сказать, что $T(x)$ сходится к центрированной и нормированной гауссовской величине, помимо этого:

$$\sqrt{\mathcal{D}(\overline{X})} = \sqrt{\mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)} = \frac{1}{n}\sqrt{n\mathcal{D}(x_1)} = \sqrt{\frac{\mathcal{D}(x_1)}{n}}\Bigg|_{H_0} = \sqrt{\frac{p_0(1-p_0)}{n}} \approx 0.007$$

При подстановке в статистику получаем

$$T(x) = \frac{\frac{2\,500}{5\,000} - 0.52}{0.007} \approx -2.83$$

Доверительный интервал в нашем случае $(Z_{0.05},\ +\infty) \approx (-1.645,\ +\infty)$. Статистика попадает в критическую область, значит мы принимаем H_1.

Ответ

Нет, нельзя.

Семинар 28 февраля.

Полезная информация

$x_1, \ldots, x_n \sim N(m, \sigma^2)$, дисперсия известна.
Проверяем гипотезу $H_0 : m = m_0$. Для этого подходит статистика:

$$T(x) = \frac{(\overline{X} - m_0)\sqrt{n}}{\sigma} \sim N(0, 1)$$

Если дисперсия неизвестна, то подойдёт:

$$T(x) = \frac{(\overline{X} - m_0)\sqrt{n-1}}{S} = \frac{(\overline{X} - m_0)}{\tilde{S}} \sim t(n-1)$$

Задача

$n = 50$, $X = (x_1, \ldots, x_n) \sim N(m, 10^2)$, $\overline{X} = 125.8$. Хотим матожидание 125.
Проверяем гипотезу $H_0 : m = m_0 = 125$ против $H_1 : m \neq 125$. Составим статистику:

$$T(x) = \frac{(\overline{X} - m_0)\sqrt{n}}{\sigma} = \frac{(125.8 - 125)\sqrt{50}}{10} \approx \frac{1}{2}$$

Доверительный интервал $(Z_{0.025}, Z_{0.975}) \approx (-1.96, 1.96)$. Статистика попала в доверительный интервал.

Задача

Есть выборка $X = (x_1, \ldots, x_n)$, $n = 10$, $\forall x \in X \quad x \sim N(m, \sigma^2)$. Никакие параметры неизвестны. Хотим проверить гипотезу $H_0 : m = m_0 = 90$ против $H_1 : m < 90$
$\overline{X} = 86$, $n = 10$, $\tilde{S} = 12.55$, $\alpha = 0.05$. Рассмотрим статистику:

$$T(x) = \frac{(86 - 90)\sqrt{10}}{12.55} = \frac{-4\sqrt{10}}{12.55} \approx -1.008$$

Доверительный интервал $(t_{9,\, 0.05}, +\infty) = (-1.833, +\infty)$, то есть попали в доверительный, значит принимаем H_0.

Ещё полезной информации

Если проверяем гипотезу $H_0 : \sigma^2 = \sigma_0^2$:
С известным математическим ожиданием:

$$\frac{\sum_{i=1}^{n} (x_i - m)^2}{\sigma_0^2}\Big|_{H_0} \sim \chi^2(n)$$

С неизвестным математическим ожиданием:

$$\frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}{\sigma_0^2}\Big|_{H_0} \sim \chi^2(n-1)$$

Задача

$n = 25$, $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.02$, $\alpha = 0.05$. Проверяем гипотезу $H_0 : \sigma^2 = \sigma_0^2 = 0.01$, альтернатива $H_1 : \sigma^2 > 0.01$.
Люди, которые для нас посчитали сумму знали математическое ожидание случайной величины, поэтому используем формулу, в которой оно известно:

$$T(x) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma_0^2} = \frac{0.5}{0.01} = 50$$

Доверительная область $(-\infty, \chi^2_{25,\, 0.95}) \approx (-\infty, 37.65)$. В доверительную область статистика не попала, значит принимаем H_1.

Задача

Станок штампует валики, в выборке объёма $n = 17$, выборочное среднее получилось 20.5, выборочная дисперсия $S^2 = 16$. Проверить на уровне значимости 0.05 гипотезу $H_0 : \sigma^2 = 18$ (альтернативой будет $H_1 : \sigma^2 \neq 18$).
Математическое ожидание и дисперсию мы не знаем, поэтому берём статистику

$$T(x) = \frac{\overbrace{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}^{=nS^2}}{18} = \frac{17 \cdot 16}{18} \approx 15.11 \sim \chi^2(16)$$

Доверительный интервал $(\chi^2_{0.25,\ 16},\ \chi^2_{0.975,\ 16}) \approx (6.9,\ 28.84) \Rightarrow$ статистика не попала, принимаем H_1.

Семинар 7 марта

Задача

В первых 10 000 знаках после запятой числа π справедлива такая таблица

Цифра	Количество вхождений ν_i
0	968
1	1026
2	1021
3	974
4	1014
5	1046
6	1021
7	970
8	948
9	1012

Проверяем $H_0 : p_1 = \cdots = p_{10} = \frac{1}{10}$
$\hat{\chi}^2 = \sum\limits_{k=1}^{10} \frac{n}{p_k^{(0)}} \left(\hat{p}_k - p_k^{(0)}\right)\Big|_{H_0} \sim \chi^2(9)$. $\hat{p}_k = \frac{\nu_k}{n},\ p_k\big|_{H_0} = \frac{1}{10}$.
В нашем случае $\hat{\chi}^2 = \sum_{i=1}^{10} \frac{\left(\frac{\nu_i}{n} - p_k\right)^2}{p_k} = 9.318$, квантиль $\chi^2_{9,\ 0.95} = 16.92$. Попали в доверительную облатсь, значит H_0 верна.

Задачи на критерий Стьюдента

$X_1, \ldots,\ X_m \sim N(m_1,\ \sigma^2)$
$Y_1, \ldots,\ Y_n \sim N(m_2,\ \sigma^2)$
$X,\ Y$ независимы, σ неизвестны, но одинаковы (если сомневаемся, то проверяем критерием Фишера). Тогда для проверки гипотезы $H_0 : \theta = m_1 - m_2 = 0$ можно использовать статистику

$$T(x,\ y) = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}\Big|_{H_0} \sim t(m + n - 2)$$

$S^2 = \frac{\sum_{i=1}^{m}(x_i - \overline{X})^2 = \sum_{i=1}^{n}(y_i - \overline{Y})^2}{m+n-2}$

Задача

$X_1, \ldots,\ X_{29} \sim N(m_1,\ \sigma^2)$
$Y_1, \ldots,\ Y_{16} \sim N(m_2,\ \sigma^2)$
$H_0 : m_1 = m_2$ против $H_1 : m_1 > m_2$.

Дано $\overline{X} = 2.5,\ \overline{Y} = 2.06,\ S_x^2 = 0.67,\ S_y^2 = 0.42$
$T(x,\ y) = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$
$S^2 = \frac{\sum_{i=1}^{29}(x_i - \overline{X})^2 + \sum_{i=1}^{16}(y_i - \overline{Y})}{m+n-2} = \frac{nS_x^2 + mS_y^2}{m+n-2} = 0.608 \Rightarrow S = 0.78$
Подставим числа в статистику, получаем $T(x,\ y) = 1.81$, а квантиль $t_{43,\ 0.95} \approx 1.68$. Критическая область находится справа, значит статистика попала в критическую область и мы принимаем H_1.

Задача про кобальт и кроликов

$X_1, \ldots,\ X_m \sim F(t)$, контрольная выборка
$Y_1, \ldots,\ Y_n \sim F(t - \theta)$, опытная группа.
Проверяем гипотезу $H_0 : \theta = 0$ против $H_1 : \theta > 0$. На лекции доказывалось, что $\theta = EY - EX$.
Распределение мы не знаем, значит не знаем ничего про равенство дисперсий, значит критерий Стьюдента применять мы здесь не можем, воспользуемся критерием Вилкоксона.

$$W = \sum_{i=1}^{n} R_i,\ \text{сумма рангов } y \text{ в вариационном ряду}$$

Контрольная группа: 560, 580, 600, 420, 530, 490, 580, 740
Опытная группа: 692, 700, 621, 640, 561, 680, 630
Упорядочим: $420_x,\ 490_x,\ 530_x,\ 560_x,\ 561_y,\ 580_x,\ 580_x,\ 600_x,\ 621_y,\ 630_y,\ 640_y,\ 680_y,\ 692_y,\ 700_y,\ 740_x$
Итого ранги y : 5, 9, 10, 11, 12, 13, 14. Сумма рангов 74. Если хотим проверить на уровне значимости 0.05 столкнёмся с проблемой дискретности распределения, поэтому придётся брать квантиль уровня 0.047, который равен 71. Попали в критическую область, значит кобальтовые добавки влияют на вес кроликов.
$EW_{8,\ 7} = (8 + 7 + 1)\frac{7}{2} = 56,\ \mathcal{D}W_{8,\ 7} = \frac{7 \cdot 8}{12} \cdot (8 + 7 + 1) \approx 74.667$
Если брать аппроксимацию гауссовской величиной:

$$W^* = \frac{W - 56}{\sqrt{72}} \sim N(0,\ 1)$$

В наших числах $W^* \approx 2,12$. На уровне значимости 0.95, а критическая область начинается с 1.64, то есть тут тоже приняли бы альтернативу.

Домашнее задание к 14 марта.

Задача 1

Наблюдались показания $n = 500$ наугад выбранных часов, выставленных в витринах часовщиков. Пусть i - номер промежутка от i – го часа до $(i+1)$-го часа, $i = 0,\ 1, \ldots,\ 11$.

i	0	1	2	3	4	5	6	7	8	9	10	11
n_i	41	34	54	39	49	45	41	33	37	41	47	39

Согласуются ли эти данные с гипотезой о том, что показания часов распределены равномерно в интервале (0;12)? Уровень значимости принять равным 0.05.

Решение

Нам дана гипотеза $H_0 : X_1, \ldots,\ X_{500} \sim R(0,\ 12) \Rightarrow F(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{12}, & x \in [0,\ 12] \\ 1, & x > 12 \end{cases}$
На интервалы уже всё разбили, неизвестных параметров в распределении нет, значит оценивать их не надо. Посчитаем частоты попадания на интервалы:

i	0	1	2	3	4	5	6	7	8	9	10	11
$\hat{p}_i$	$\frac{41}{500}$	$\frac{34}{500}$	$\frac{54}{500}$	$\frac{39}{500}$	$\frac{49}{500}$	$\frac{45}{500}$	$\frac{41}{500}$	$\frac{33}{500}$	$\frac{37}{500}$	$\frac{41}{500}$	$\frac{47}{500}$	$\frac{39}{500}$

При справедливости H_0 вероятность попадания на все интервалы должна быть одинаковой (так как интервалы поровну делят область некоторого равномерного распределения), то есть $\forall i = \overline{0,\ 11} \quad p_i^{(0)} = \frac{1}{12}$
По критерию хи-квадрат:

$$\hat{\chi}^2 = \sum_{i=0}^{11} \frac{500}{p_i^{(0)}} \left(\hat{p}_i - p_i^{(0)}\right)^2 \Big|_{H_0} \sim \chi^2\,(11)$$

Доверительным интервалом будет $(0,\ \chi^2_{0.95,\ 11}) = (0,\ 19.675)$. Наша оценка примерно равна 10, значит попала в доверительную область.

Ответ

Согласуются.

Задача 2

В некоторой компании работает $n = 500$ продавцов, на каждого из которых может поступить жалоба. За последний месяц на 275 продавцов жалоб не поступало, на 150 поступило по одной жалобе, на 50 – по две жалобы, на остальных – три или более жалоб. С помощью критерия хи-квадрат проверьте гипотезу о том, что количество жалоб на продавца есть случайная величина подчиняющаяся распределению Пуассона со средним значением одна жалоба в месяц. Уровень значимости считать равным 0.05.

i	0	1	2	3+
n_i	275	150	50	25

Решение

Нужно проверить гипотезу $H_0 : X_1, \ldots,\ X_{500} \sim \Pi(1) \Rightarrow P(X_i = k) = \frac{e^{-1}1^k}{k!} = \frac{1}{k! \cdot e}$
Разбили на 4 интервала (для продавцов с 0, 1, 2, 3+ жалобами), неизвестных параметров в распределении нет, значит ничего не надо оценивать. Посчитаем частоты:

i	0	1	2	3+
$\hat{p}_i$	$\frac{275}{500}$	$\frac{150}{500}$	$\frac{50}{500}$	$\frac{25}{500}$

При справедливости H_0 они должны выглядеть так:

i	0	1	2	3+
$p_i^{(0)}$	$\frac{1}{e}$	$\frac{1}{e}$	$\frac{1}{2e}$	$1 - \frac{5}{2e}$

По критерию хи-квадрат должно выполняться:

$$\hat{\chi}^2 = \sum_{i=0}^{3} \frac{500}{p_i^{(0)}} \left(\hat{p}_i - p_i^{(0)}\right)^2 \Big|_{H_0} \sim \chi^2\,(3)$$

Доверительная область будет $(0,\ \chi^2_{0.95,\ 3}) = (0,\ 7.814)$. Оценка получилась 76.21, походу мимо.

Ответ

Гипотеза отвергается.

Задача 3

Имеется набор данных «ирисы Фишера». Эти данные собраны ботаником Эдгаром Андерсоном. Они включают длину и ширину чашелистиков, длину и ширину лепестков трёх видов ирисов (setosa, versocolor и виргинский). Выберем из этих данных случайным образом по 10 измерений длин чашелистиков цветов вида setosa и цветов вида versocolor.
Длина (в мм) чашелистиков у выбранных цветов вида setosa:

$$X = (5.1,\ 4.9,\ 4.7,\ 4.6,\ 5.0,\ 5.4,\ 4.6,\ 4.4,\ 4.8,\ 4.8) \Rightarrow \overline{X} = 4.83$$

Длина (в мм) чашелистиков у цветов вида versocolor:

$$Y = (5.7,\ 6.3,\ 4.9,\ 6.6,\ 5.2,\ 5.0,\ 5.9,\ 6.0,\ 5.6,\ 5.8) \Rightarrow \overline{Y} = 5.7$$

Можно ли считать, опираясь на эти данные, что длина чашелистиков у цветов вида versocolor в среднем больше, чем у цветов вида setosa? Решите данную задачу в ситуации, когда:

а) предполагается, что наблюдаемый признак имеет гауссовское распределение;

б) нет предположения о виде распределения наблюдаемого признака.

Решение а)

$X \sim N(m_1,\ \sigma_1^2),\ Y \sim N(m_2,\ \sigma_2^2)$. Дисперсии неизвестны, могут быть различны. Хотим проверить гипотезу $H_0 : m_1 = m_2$ против $H_1 : m_2 > m_1$

$$T(x,\ y) = \frac{\overline{Y} - \overline{X} - (m_2 - m_1)}{\sqrt{\mathcal{D}\left(\overline{Y} - \overline{X}\right)}} = \left.\frac{\overline{Y} - \overline{X} - (m_2 - m_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right|_{H_0} = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Дисперсии не знаем, оценим их:

$$\sigma_1^2 \to \tilde{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \overline{X}\right)^2 \approx 0.082$$

$$\sigma_2^2 \to \tilde{S}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(y_i - \overline{Y}\right)^2 \approx 0.3$$

При справедливости H_0 должны получить распределение Стьюдента $t(n_1 + n_2 - 2)$. Посчитаем значение статистики:

$$T(x,\ y) \approx \frac{5.7 - 4.83}{\sqrt{\frac{0.082}{9} + \frac{0.3}{9}}} \approx 4.223$$

Доверительный интервал $(-\infty,\ t_{18,\ 0.95}) = (-\infty,\ 1.734)$. Попали в критическую область, значит принимаем H_1, то есть длина чашелистиков у versocolor действительно в среднем больше.

Решение б)

$X \sim F(t),\ Y \sim F(t - \theta)$, проверяем гипотезу $H_0 : \theta = 0$ против $H_1 : \theta > 0$
Про дисперсии ничего не знаем, поэтому надо пользоваться ранговым критерием Вилкоксона.
Ранжируем данные:

$$\underbrace{4.4_X}_{1},\ \underbrace{4.6_Y,\ 4.6_X}_{2.5},\ \underbrace{4.7_X}_{4},\ \underbrace{4.8_X,\ 4.8_X}_{5.5},\ \underbrace{4.9_X,\ 4.9_Y}_{7.5},\ \underbrace{5.0_X,\ 5.0_Y}_{9.5},$$

$$\underbrace{5.1_X}_{11},\ \underbrace{5.2_Y}_{12},\ \underbrace{5.4_X}_{13},\ \underbrace{5.6_Y}_{14},\ \underbrace{5.7_Y}_{15},\ \underbrace{5.8_Y}_{16},\ \underbrace{5.9_Y}_{17},\ \underbrace{6.0_Y}_{18},\ \underbrace{6.3_Y}_{19},\ \underbrace{6.6_Y}_{20}$$

Ранги y: 2.5, 7.5, 9.5, 12, 14, 15, 16, 17, 18, 19, 20. Сумма получается 150.5. Если размеры равны, то ожидается

$$EW_{10,\ 10}\Big|_{H_0} = (10 + 10 + 1)\cdot 5 = 105, \quad \mathcal{D}W_{10,\ 10}\Big|_{H_0} = \frac{100}{12}(10 + 10 + 1) = \frac{2100}{12} = 175$$

Следующая штука должна иметь стадартное гауссовское распределение:

$$W^* = \frac{W_{10,\ 10} - EW_{10,\ 10}}{\sqrt{\mathcal{D}W_{10,\ 10}}} = \frac{150.5 - 105}{\sqrt{175}} \approx 3.439$$

$Z_{0.95} = 1.64$, доверительный интервал $(-\infty,\ 1.64)$, значит попали в критическую, то есть Y в среднем больше, результаты совпали.

Ответ

По обоим методам можно так считать.

Задача 4

Деятельность отделения банка характеризуется некоторым показателем X. Для проверки была случайным образом сделана выборка 10 однотипных отделений банка. Показатель X у этих отделений составил:

$$Y = (258,\ 588,\ 477,\ 577,\ 619,\ 614,\ 641,\ 543,\ 517,\ 593)$$

После экономического кризиса показатель X у 9 случайным образом выбранных отделений банка составил:

$$Z = (537,\ 398,\ 256,\ 440,\ 376,\ 524,\ 527,\ 589,\ 479)$$

Можно ли считать, опираясь на эти данные, что экономический кризис привёл к снижению показателя X. Уровень значимости принять равным 0.05.

Решение

$Y \sim F(t),\ Z \sim F(t-\theta)$, гипотеза $H_0 : \theta = 0$, против $H_1 : \theta < 0$
Про распределение и его дисперсию мы ничего не знаем, поэтому воспользуемся критерием Вилкоксона.
Сумма рангов Z получилась 65.

$$EW_{10,\,9}\Big|_{H_0} = (10+9+1)\cdot\frac{9}{2} = 45, \quad \mathcal{D}W_{10,\,9}\Big|_{H_0} = \frac{90}{12}(20) = 150$$

Построим статистику (она, кстати, номральной гауссовской должна быть):

$$W^* = \frac{65-45}{\sqrt{150}} \approx 1.633$$

На удивление, реально попали в $(-\infty,\ 1.64)$, то есть принимаем гипотезу о равенстве.

Задача 5

Показать, что если $X \sim F(t)$, а $Y \sim F(t-\theta)$, то $H_0 : \theta < 0 \Rightarrow EY < EX$
Распишем по определению:

$$EX = \int\limits_{-\infty}^{+\infty} t f(t)\,dt$$

$$EY = \int\limits_{-\infty}^{+\infty} t f(t-\theta)\,dt = \left\langle \begin{matrix} z = t-\theta \\ dt = dz \end{matrix} \right\rangle = \int\limits_{-\infty}^{+\infty} (z+\theta) f(z)\,dz = \int\limits_{-\infty}^{+\infty} z f(z)\,dz + \theta \underbrace{\int\limits_{-\infty}^{+\infty} f(z)\,dz}_{=1} = EX + \theta$$

$$EY = EX + \theta\Big|_{H_0} < EX + 0 = EX$$

Доказали.

Семинар 14 марта

Справка

$X_1, \ldots,\ X_m \sim F(t-\mu)$ и $Y_1, \ldots,\ Y_n \sim F\left(\frac{t-\mu}{\Delta}\right),\ \Delta > 0$
Проверяем $H_0 : \Delta = 1$
Если нам известно, что
$X_1, \ldots,\ X_m \sim N(m_1,\ \sigma_1^2)$

$Y_1, \ldots, Y_n \sim N(m_2, \sigma_2^2)$
Проверяем гипотезу $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$

$$T(x, y) = \frac{\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{X})^2}{\frac{1}{n-1}\sum_{i=0}^{n}(y_i - \overline{Y})^2} \sim F(m-1, n-1)$$

Против гипотезы $H_1 : \sigma_1 < \sigma_2$:
Смотрим на отношение выборочных дисперсий $\frac{S_y^2}{S_x^2}\Big|_{H_0} \sim F(n-1, m-1)$. Критическая область справа.
Против гипотезы $H_2 : \sigma_1 > \sigma_2$:
Смотрим на отношение выборочных дисперсий $\frac{S_x^2}{S_y^2}\Big|_{H_0} \sim F(m-1, n-1)$. Критическая область также справа (однако мы поменяли числитель со знаменателем).
Против гипотезы $H_3 : \sigma_1 \neq \sigma_2$:
Также ставим большее в знаменатель, получаем соответствующее распределение Фишера. Критическая область также справа, границей будет квантиль фишера уровня $1 - \alpha/2$.

Задача

$X_1, \ldots, X_{29} \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_{16} \sim N(m_2, \sigma_2^2)$
$H_0 : m_1 = m_2$ против $H_1 : m_1 > m_2$.
Дано $\overline{X} = 2.5,\ \overline{Y} = 2.06,\ S_x^2 = 0.67,\ S_y^2 = 0.42 \Rightarrow \tilde{S}_x^2 = \frac{29}{28}S_x^2 = 0.694,\ \tilde{S}_y^2 = \frac{16}{15}S_y^2 = 0.448$
Нужно проверить гипотезу о том, что дисперсии одинаковы $H_0 : \sigma_1 = \sigma_2 = \sigma$ против $H_1 : \sigma_1 \neq \sigma_2$
Возьмём статистику $T(x, y) = \frac{\tilde{S}_x^2}{\tilde{S}_y^2} = \frac{0.694}{0.448} < t_{28,\ 15,\ 0.95} \Rightarrow$ принимаем H_0

Задача

Есть станок, который производит детали с параметрами, распределёнными по неизвестному распределению с неизвестным (но одинаковым) матожиданием и дисперсией. Его наладили, нужно проверить гипотезу о том, что дисперсия уменьшилась.
До наладки было:

$$X = (52.4,\ 56.1,\ 48.6,\ 46.5,\ 46,\ 42.2,\ 48.8,\ 56.6,\ 59.8,\ 49.7,\ 51.6)$$

После наладки стало:

$$Y = (49.3,\ 47.7,\ 52.8,\ 48.3,\ 49.1,\ 46.4,\ 47,\ 52,\ 51.5,\ 51.2,\ 49.8)$$

Предполагаем, что $X \sim F(t - \mu),\ Y \sim F\left(\frac{y-\mu}{\Delta}\right)$
Проверяем гипотезу $H_0 : \Delta = 1$ против $H_1 : \Delta < 1$.
Применяем Ансари-Брейли. $A_{11,\ 11} = 53,\ EA_{11,\ 11} = 66,\ \mathcal{D}A_{11,\ 11} = 29.49$
$A^* = \frac{53-66}{\sqrt{29.49}} \approx -2.5$
Граница критической области находится в точке $Z_{0.05} = -1.64 \Rightarrow$ попали в критическую область, то есть дисперсия действительно уменьшилась.

ДЗ к семинару 21 марта

Задача 1

Имеется таблица температур в двух городах за $n = m = 13$ лет.

Год	Саратов	Алатырь
1891	-19.2	-21.8
1892	-14.8	-15.4
1893	-19.6	-20.8
1894	-11.1	-11.3
1895	-9.4	-11.6
1896	-16.9	-19.2
1897	-13.7	-13.0
1899	-4.9	-7.4
1911	-13.9	-15.1
1912	-9.4	-14.4
1913	-8.3	-11.1
1914	-7.9	-10.5
1915	-5.3	-7.2

Необходимо проверить равенство дисперсий температур в этих городах. Вообще про распределение мы ничего не знаем, поэтому стоило бы сразу использовать критерий Ансари-Бредли, но это требуется во второй задаче, то есть здесь, скорее всего, подразумевается критерий Фишера. Для его применения мы должны предположить, что температуры распределены нормально, например $N(m_1,\ \sigma_1^2),\ N(m_2,\ \sigma_2^2)$ для температуры Саратова и Алатыря соответственно.

Решение

Нужно проверить гипотезу:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2,\ \text{здесь } \sigma^2 \text{ — некое число, введённое для удобства обозначения}$$

Альтернативой будет $H_1 : \sigma_1^2 \neq \sigma_2^2$ Запишем статистику T:

$$T(x,\ y) = \left.\frac{\frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - \overline{X}\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}\right|_{H_0} \sim F(m-1,\ n-1)$$

Числителем здесь является $\tilde{S}_x^2$, знаменателем — $\tilde{S}_y^2$. Посчитаем их:

$$\overline{X} \approx -11.88$$
$$\tilde{S}_x^2 = \frac{1}{12}\sum_{i}^{13}\left(x_i - \overline{X}\right)^2 \approx 23.99$$
$$\overline{Y} \approx -13.75$$
$$\tilde{S}_y^2 = \frac{1}{12}\sum_{i}^{13}\left(y_i - \overline{Y}\right)^2 \approx 21.76$$

$\tilde{S}_x^2$ получилась больше, значит статистику мы не меняем (если бы $\tilde{S}_y^2$ оказался больше, то пришлось бы рассматривать $\frac{1}{T(x,\ y)}$).
Статистика должна иметь распределение $F(12,\ 12)$, квантиль уровня значимости 0.05 для такого распределения равен 2.69, а критическая область находится справа, то есть доверительный интервал: $(-\infty,\ 2.69)$.
Посчитаем значение нашей статистики:

$$T(x,\ y) = \frac{23.99}{21.76} \approx 1.1$$

Попали в доверительную область, значит принимаем H_0.

Ответ

Дисперсии действительно можно считать одинаковыми

Задача 2

Решить предыдущую задачу, пользуясь критерием Ансари-Бредли.

Решение

Полагаем, что выборка $X_1, \ldots, X_{13} \sim F(t-\mu)$ и выборка $Y_1, \ldots, Y_{13} \sim F\left(\frac{t-\mu}{\Delta}\right)$, $\Delta > 0$ и уровень доверия $\alpha = 0.05$
Критерий Ансари-Бредли требует $F(\mu) = 0.5$ (в условии задачи есть подсказка на этот счёт, нужно центрировать данные выборочной медианой).
Теперь нужно проверить гипотезу $H_0 : \Delta = 1$ против $H_1 : \Delta \neq 1$
Согласно критерию следует рассмотреть статистику:

$$A_{m,\,n} = \sum_{i=1}^{m+n} \left(\frac{m+n+1}{2} - \left| R_i - \frac{m+n+1}{2} \right| \right)$$

С нашими числами будет выглядеть так:

$$A_{13,\,13} = \sum_{i=1}^{26} \left(\frac{27}{2} - \left| R_i - \frac{27}{2} \right| \right) = 83$$

Здесь R_i — ранг X_i в объединённой выборке $\big($в общем случае берём ту, распределение которой мы считаем $F(t-\mu)\big)$.
На лекции говорилось, что есть таблица точных значений квантилей этой статистики при $n + m \leqslant 20$ (не наш случай), поэтому придётся апроксимировать гауссовским распределением:

$$A^* = \frac{A_{m,\,n} - EA_{m,\,n}}{\sqrt{\mathcal{D}A_{m,\,n}}}\bigg|_{H_0} \sim N(0,\,1)$$

Формулы для математического ожидания и дисперсии статистики известны:

$$EA_{m,\,n} = \begin{cases} \frac{m(m+n+2)}{4}, & m+n \underset{2}{\equiv} 0 \\ \frac{m(m+n+1)^2}{4(m+n)}, & m+n \underset{2}{\equiv} 1 \end{cases}$$

$$\mathcal{D}A_{m,\,n} = \begin{cases} \frac{mn(m+n+2)(m+n-2)}{48(m+n-1)}, & m+n \underset{2}{\equiv} 0 \\ \frac{mn\Big((m+n)^2+3\Big)(m+n+1)}{48(m+n)^2}, & m+n \underset{2}{\equiv} 1 \end{cases}$$

В нашем случае $m + n = 13 + 13 = 26 \underset{2}{\equiv} 0 \Rightarrow$ используем соответствующие формулы:

$$EA_{13,\,13} = \frac{13(13+13+2)}{4} = \frac{13 \cdot 28}{4} = 13 \cdot 7 = 91$$

$$\mathcal{D}A_{13,\,13} = \frac{(13 \cdot 13)(13+13+2)(13+13-2)}{48(13+13-1)} = \frac{169 \cdot 28 \cdot 24}{48 \cdot 25} \approx 94.64$$

Теперь можно посчитать A^*:

$$A^* = \frac{83 - 91}{\sqrt{94.64}} \approx -0.822$$

Доверительный интервал: $(Z_{0.025},\ Z_{0.975}) = (-1.96,\ 1.96)$. Статистика попала в доверительный интервал, значит принимается H_0

Ответ

По этому критерию принимаем $\Delta = 1$, а $\Delta^2 = \frac{\mathcal{D}Y}{\mathcal{D}X} \Rightarrow \mathcal{D}Y = \mathcal{D}X$

Задача 3

За последние 5 лет выборочная дисперсия доходности актива А составила 0.04, а выборочная дисперсия доходности актива Б составила 0.05. Есть ли основание утверждать (на уровне значимости 0.05), что вложения в актив А менее рискованны, чем вложения в актив Б? Предполагается, что доходности активов являются гауссовскими СВ.

Решение

Полагаем, что $X \sim N(m_1,\, \sigma_x^2)$ — выборка доходностей предприятия А, Y — выборка доходностей предприятия Б. Нам дано $S_x^2 = 0.04$ и $S_y^2 = 0.05$ (дана обычная выборочная дисперсия, про несмещённость не сказано). Сформулируем гипотезы для проверки:

$$H_0 : \sigma_x = \sigma_y \text{ против } H_1 : \sigma_x < \sigma_y$$

Если верна H_0, то вложения в любой из активов влекут за собой одинаковые риски. Принятие H_1 будет означать, что вкладываться в актив A безопаснее.

Для всех известных критериев оценки дисперсии необходимо знать размер выборок. В нашем случае предположу $n = m = 5$ (доходность замерялась ежегодно).

В нашем случае мы сравниваем дисперсии гауссовских величин (в условии написано считать их таковыми), поэтому пользоваться будем критерием Фишера. В этом критерии необходима несмещённая выборочная дисперсия, поэтому стоит получить её из данной нам смещённой:

$$\tilde{S}_x^2 = \frac{n}{n-1} S_x^2 = \frac{5}{4} \cdot 0.04 = 0.05$$
$$\tilde{S}_y^2 = \frac{m}{m-1} S_y^2 = \frac{5}{4} \cdot 0.05 = 0.0625$$

$\tilde{S}_y^2$ больше, значит нужно рассматривать статистику:

$$T(x,\, y) = \left.\frac{\tilde{S}_y^2}{\tilde{S}_x^2}\right|_{H_0} \sim F(m-1,\, n-1) \sim F(4,\, 4)$$

Подставив числа в статистику, получаем:

$$T(x,\, y) = \frac{5}{4}$$

Квантилем уровня 0.95 для $F(4,\, 4)$ является 6.4, значит доверительным интервалом будет $(0,\, 6.4)$. Статистика попала в доверительный интервал, значит принимаем H_0.

Ответ

Оснований утверждать о меньшем количестве риска нет.

Задача 4

Два завода изготавливают электролампы одинакового типа. Из продукции завода №1 случайным образом выбрано 10 ламп, из продукции завода №2 — 12 ламп. Испытания по длительности горения ламп (в часах) следующие:

Для завода №1: 1243, 1238, 1253, 1243, 1254, 1260, 1251, 1246, 1255, 1237.
Для завода №2: 1244, 1255, 1258, 1266, 1249, 1257, 1260, 1247, 1256, 1271, 1252, 1259.

Проверьте гипотезу об однородности двух выборок, применяя критерий Колмогорова-Смирнова. (Указание. Квантиль уровня $1 - 0.049$ распределения статистики Колмогорова-Смирнова для выборок объёма 10 и 12 равна $\frac{33}{60}$).

Решение

Будем работать с гипотезами:

$$H_0 : \forall t \quad F(t) = G(t), \text{ против } H_1 : \exists t \quad F(t) \neq G(t)$$

Статистикой будет:

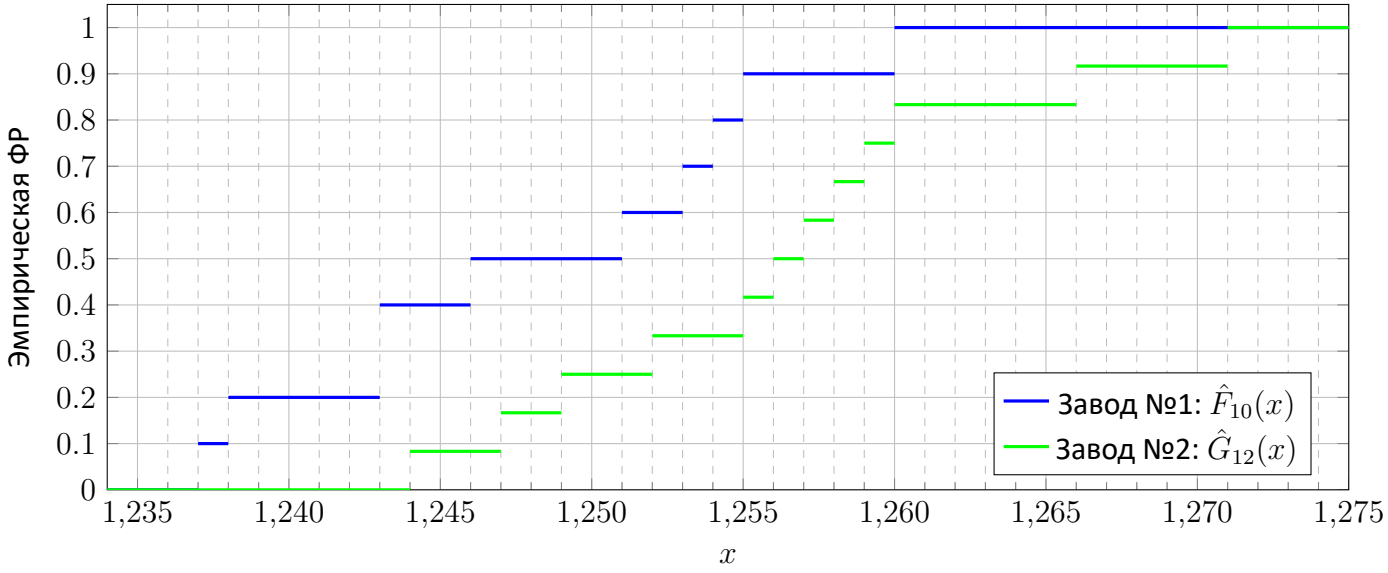
$$D_{m,\,n} = \max_{1 \leqslant i \leqslant m+n} \left| \hat{F}_m(z_i) - \hat{G}_n(z_i) \right|$$

В этой статистике z_i — i-ый элемент объединённой выборки. Точный квантиль этой статистики дан в указании к заданию и равен $\frac{33}{60}$. То есть доверительным интервалом будет $\left(0,\ \frac{33}{60}\right) = (0,\ 0.55)$.

Для применения критерия Колмогорова-Смирнова необходимо построить эмпирические функции распределения для обоих заводов. Для этого стоит упорядочить обе выборки:

Для завода №1: 1237, 1238, 1243, 1243, 1246, 1251, 1253, 1254, 1255, 1260.
Для завода №2: 1244, 1247, 1249, 1252, 1255, 1256, 1257, 1258, 1259, 1260, 1266, 1271.

А теперь рисуем:



По графику видно, что наибольшее различие между функциями находится в точке $x = 1255$:

$$D_{10,\,12} = \left| \frac{9}{10} - \frac{5}{12} \right| = 0.483$$

Попали в доверительный интервал, значит принимаем H_0 и говорим, что распределения одинаковы.

Ответ

Гипотеза однородности принимается по критерию Колмогорова-Смирнова.

Задача 5

Пусть выборка $X_1, \ldots,\ X_n$ порождена случайной величиной X с непрерывным распределением $F(t - \mu)$, а выборка $Y_1, \ldots,\ Y_n$ — случаной величиной Y с распределением $F\left(\frac{t-\mu}{\Delta}\right),\ \Delta > 0$. Предполагается, что $\mathcal{D}X < \infty$ и выполняется равенство $\int\limits_{-\infty}^{+\infty} t f(t)\, dt = 0$. Показать, что из справедливости $H_0 : \Delta < 1$ следует неравенство $\mathcal{D}X > \mathcal{D}Y$.

Решение

Чтобы посчитать дисперсии, нужно сначала посчитать математические ожидания, а для подсчёта матожиданий нужны функции плотностей вероятности:

$$f_x(t) = (F(t-\mu))' = f(t-\mu)\cdot(t-\mu)' = f(t-\mu)$$
$$f_y(t) = \left(F\left(\frac{t-\mu}{\Delta}\right)\right)' = f\left(\frac{t-\mu}{\Delta}\right)\cdot\left(\frac{t-\mu}{\Delta}\right)' = \frac{1}{\Delta}\cdot f\left(\frac{t-\mu}{\Delta}\right)$$

Теперь можно считать математические ожидания:

$$EX = \int\limits_{-\infty}^{+\infty} t f_x(t)\,dt = \int\limits_{-\infty}^{+\infty} t f(t-\mu)\,dt = \left\langle \begin{matrix} a = t-\mu \\ da = dt \end{matrix} \right\rangle = \int\limits_{-\infty}^{+\infty} (a+\mu) f(a)\,da =$$
$$= \underbrace{\int\limits_{-\infty}^{+\infty} a f(a)\,da}_{=0,\ \text{по усл.}} + \mu \underbrace{\int\limits_{-\infty}^{+\infty} f(a)\,da}_{=1} = \mu$$

$$EY = \int\limits_{-\infty}^{+\infty} t f_y(t)\,dt \int\limits_{-\infty}^{+\infty} \frac{t}{\Delta} f\left(\frac{t-\mu}{\Delta}\right)\,dt = \left\langle \begin{matrix} z = \dfrac{t-\mu}{\Delta} \\ dz = \dfrac{dt}{\Delta} \end{matrix} \right\rangle = \int\limits_{-\infty}^{+\infty} \frac{\Delta\cdot z+\mu}{\Delta} f(z)\Delta\,dz =$$
$$= \Delta \int\limits_{-\infty}^{+\infty} z f(z)\,dz + \mu \int\limits_{-\infty}^{+\infty} f(z)\,dz = \mu$$

И, наконец, дисперсии:

$$\mathcal{D}X = \int\limits_{-\infty}^{+\infty} (t-EX)^2 f_x(t)\,dt = \int\limits_{-\infty}^{+\infty} (t-\mu)^2 f(t-\mu)\,dt = \left\langle \begin{matrix} a = t-\mu \\ da = dt \end{matrix} \right\rangle = \int\limits_{-\infty}^{+\infty} a^2 f(a)\,da$$
$$\mathcal{D}Y = \int\limits_{-\infty}^{+\infty} (t-EY)^2 f_y(t)\,dt = \int\limits_{-\infty}^{+\infty} (t-\mu)^2 \frac{1}{\Delta} f\left(\frac{t-\mu}{\Delta}\right)\,dt = \int\limits_{-\infty}^{+\infty} \Delta^2 \left(\frac{t-\mu}{\Delta}\right)^2 f\left(\frac{t-\mu}{\Delta}\right)\frac{dt}{\Delta} =$$
$$= \left\langle \begin{matrix} z = \dfrac{t-\mu}{\Delta} \\ dz = \dfrac{dt}{\Delta} \end{matrix} \right\rangle = \int\limits_{-\infty}^{+\infty} \Delta^2 z^2 f(z)\,dz = \Delta^2 \int\limits_{-\infty}^{+\infty} z^2 f(z)\,dz = \Delta^2 \mathcal{D}X$$

Получаем соотношение: $\mathcal{D}Y = \Delta^2 \mathcal{D}X \Rightarrow \dfrac{\mathcal{D}Y}{\mathcal{D}X} = \Delta^2\Big|_{H_0} < 1 \Rightarrow \dfrac{\mathcal{D}Y}{\mathcal{D}X} < 1 \Rightarrow \mathcal{D}X > \mathcal{D}Y$, ч. т. д.

Семинар 21 марта

Задача 1

Случанйым образом выбраны 14 человек одного возраста, проживающих в одном городе, но имеющих различный уровень образования. Их опрашивли про доход:
Неполное среднее: 37, 19, 26, 42 $\Rightarrow \overline{X}_{\bullet 1} = 31$
Среднее специальное: 47, 39, 52, 41, 51 $\Rightarrow \overline{X}_{\bullet 2} = 46$
Высшее: 64, 78, 59, 71, 63 $\Rightarrow \overline{X}_{\bullet 3} = 67$
$\overline{X}_N = 49.2$

Проверить гипотезу о том, что средние доходы в группах одинаковые, против альтернативы о том, что доходы неодинаковые.
В этой задаче фактор: образование. Уровни фактора: неполное среднее, среднее специальное, высшее.

Первый случай

$$x_{i\,j} = \theta + \tau_j + \varepsilon_{i\,j},\ j = 1,\ 2,\ 3$$

Предполагаем, что $\varepsilon_{i\,j} \sim N(0,\ \sigma^2)$
Проверяем $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$ против $H_1 : \exists j : \tau_j \neq 0$

Решение

$SS_{\text{ур. ф.}} = \sum_{j=1}^{n} n_j \left(\overline{X}_j - \overline{X}_N\right)^2 = 4 \cdot 18.2^2 + 5 \cdot 3.2^2 + 5 \cdot 18.2^2 \approx 2\,960$
$SS_{\text{случ.}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{i\,j} - \overline{X}_{\bullet\,j}\right)^2 = 12^2 + 6^2 + 5^2 + 11^2 + 5^2 + 6^2 + 1^2 + 7^2 + 5^2 + 3^2 + 11^2 + 8^2 + 4^2 + 4^2 = 688$
Берём статистику:

$$\hat{f} = \frac{\frac{1}{k-1} SS_{\text{ур. ф.}}}{\frac{1}{N-k} SS_{\text{случ.}}} = \frac{\frac{3032}{2}}{\frac{688}{11}} = \frac{1480}{62.56} \approx 23$$

При справедливости H_0:

$$\hat{f}\Big|_{H_0} \sim F(2,\ 11)$$

Квантиль $F_{0.05,\ 2,\ 11} = 3.98$, критическая область справа. Статистика попала в критическую область, значит принимается альтернатива.

Второй случай

Не пользуемся предположение о гауссовости распределения. Будем применять критерий Краскела-Уоллиса.

$$x_{i\,j} = \theta_j + \varepsilon_{i\,j}$$

Необходимо посчитать средний ранг для каждого уровня фактора:
Неполное среднее: $r_{\bullet\,1} = 3$
Среднее специальное: $r_{\bullet\,2} = 6.6$
Высшее: $r_{\bullet\,3} = 12$
Статистика будет выглядеть вот так:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{3} n_j \left(\overline{r}_{\bullet\,j} - \frac{N+1}{2}\right)^2 = \frac{12}{14 \cdot 15} \left(4 \cdot 4.5^2 + 5 \cdot 0.9^2 + 5 \cdot 4.5^2\right) \approx 10.65$$

При справедливости H_0 для этой статистики справедливо:

$$H\Big|_{H_0} \sim \chi^2(2)$$

Квантиль $\chi^2_{2\,0.95} = 5.99$, значит попали в критическую область, значит здесь тоже принимаем H_2.

Третий случай

Проверяем гипотезу $H_0 : \theta_1 = \theta_2 = \theta_3$ против $H_1 : \theta_1 \leqslant \theta_2 \leqslant \theta_3$, где хотя бы одно неравенство строгое.
Введём функции:

$$\varphi(y,\ z) = \begin{cases} 1, & y < z \\ 0.5, & y = z \\ 0, & y > z \end{cases}, \quad U_{l,\ m} = \sum_{i=1}^{n_l} \sum_{j=1}^{n_m} \varphi(x_{i\,l},\ x_{i\,m})$$

Возьмём статистику:

$$J = \sum_{1 \leqslant L < m \leqslant k} U_{l,\ m}$$

$U_{1,\ 2} = 5 + 5 + 5 + 3 = 18$
$U_{1,\ 3} = 5 \cdot 4 = 20$
$U_{2,\ 3} = 5 \cdot 5 = 25$, теперь можем посчитать статистику $J = 18 + 20 + 25 = 63$
Тогда справедиво:

$$J^* = \frac{J - EJ}{\sqrt{\mathcal{D}J}}\Bigg|_{H_0} \sim N(0,\ 1)$$

Из лекции берём формулы для математического ожидания и дисперсии:

$$EJ = \frac{1}{4}\left(N^2 - \sum_{i=1}^{k} n_i^2\right) = 32.5$$

$$\mathcal{D}J = \frac{1}{72}\left(N^2(2N+3) - \sum_{i=1}^{k} n_j^2(2n_j+3)\right) \approx 73 \Rightarrow \sqrt{\mathcal{D}J} \approx 8.5$$

Тогда значение центрированной и нормированной статистики:

$$J^* = \frac{63 - 32.5}{8.5} \approx 3.6$$

Квантиль $Z_{0.95} = 1.64$, то есть статистика попала в критическую область.

ДЗ к 4 апреля

Задача 1

Изучается влияние денежного стимулирования на производительность труда. Шести однородным группам, по 5 человек, раздали задачи одинаковой сложности. Задачи были выданы каждому члену группы независимо от остальных. Группы различаются только по денежному вознаграждению за каждую решенную задачу. Величина вознаграждения зависит от номера группы: чем больше номер группы, тем больше вознаграждение. Каждой группе известна цена вознаграждения за решенную задачу. В таблице представлено количество решенных задач каждым членом группы.

Группа 1	Группа 2	Группа 3	Группа 4	Группа 5	Группа 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Влияет ли вознаграждение на количество решенных задач?
Решите данную задачу, предполагая, что:

а) все наблюдения имеют нормальное распределение с одинаковыми дисперсиями;

б) наблюдения имеют некоторые неизвестные непрерывные распределения, которые могут различаться только математическим ожиданием;

в) наблюдения имеют некоторые неизвестные непрерывные распределения, которые могут различаться только математическим ожиданием, и имеется априорное предположение о том, что с ростом вознаграждения растёт количество решённых задач.

Решение а)

В этом случае можно воспользоваться критерием Фишера.
Полагаем, что каждое наблюдение x_{ij} (числа в таблице выше) может быть представлено в виде:

$$x_{ij} = \theta + \tau_j + \varepsilon_{ij}$$

Где:

θ – некоторое неизвестное общее среднее

τ_j – отклонение, зависящее от фактора

ε_{ij} – случайное отклонение

Полагаем, что $\varepsilon_{ij} \sim N(0,\, \sigma^2)$ (дисперсия везде одинаковая).
Критерий Фишера проверяет гипотезу $H_0 : \tau_1 = \cdots = \tau_k = 0$ против $H_1 : \exists i : \tau_i \neq 0$. В нашем случае $k = 6$,

$N = 30$ (общее количество испытаний со всеми уровнями фактора).
Теперь для подсчёта статистики необходимо подсчитать $SS_{\text{случ.}}$ и $SS_{\text{ур. ф.}}$:

$$\frac{SS_{\text{случ.}}}{\sigma^2} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\frac{x_{ij} - \overline{X}_{\bullet j}}{\sigma}\right)^2$$
$$\frac{SS_{\text{ур. ф.}}}{\sigma^2} = \sum_{j=1}^{k} n_j\left(\frac{\overline{X}_{\bullet j} - \overline{X}_N}{\sigma}\right)^2$$

Здесь:

$$\overline{X}_N = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij} = 15.6$$
$$\overline{X}_{\bullet 1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1} = 10$$
$$\overline{X}_{\bullet 2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2} = 11.8$$
$$\overline{X}_{\bullet 3} = \frac{1}{n_3}\sum_{i=1}^{n_3} x_{i3} = 13.6$$
$$\overline{X}_{\bullet 4} = \frac{1}{n_1}\sum_{i=1}^{n_4} x_{i4} = 15.6$$
$$\overline{X}_{\bullet 5} = \frac{1}{n_5}\sum_{i=1}^{n_5} x_{i5} = 20$$
$$\overline{X}_{\bullet 6} = \frac{1}{n_6}\sum_{i=1}^{n_6} x_{i6} = 22.6$$

Теперь можно подставить (σ здесь не нужна):

$$SS_{\text{случ.}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{X}_{\bullet j}\right)^2 = 224.4$$
$$SS_{\text{ур. ф.}} = \sum_{j=1}^{k} n_j\left(\overline{X}_{\bullet j} - \overline{X}_N\right)^2 = 590.8$$

$SS_{\text{ур. ф.}}$ оказалось больше, тогда должно быть справедливо:

$$\frac{N-k}{k-1}\cdot\frac{SS_{\text{ур. ф.}}}{SS_{\text{случ.}}} \sim F(k-1,\ N-k) \sim F(5,\ 24)$$

Нужный нам квантиль уровня 0.95 этого распределения равен примерно 2.621, наша статистика же имеет значение:

$$\frac{24}{5}\cdot\frac{590.8}{224.4} \approx 12.6$$

Критической областью здесь будет $(2.621,\ +\infty)$, то есть уверенно находимся в критической областью, значит H_0 опровергается.

Решение задачи 2

Сразу же разберёмся с контрастами в этой задаче. Нам нужно построить три оценки для контрастов:

$$\gamma_1 = \theta_2 - \theta_1 \Rightarrow c_2 = 1,\ c_1 = -1$$
$$\gamma_2 = \theta_4 - \theta_1 \Rightarrow c_4 = 1,\ c_1 = -1$$
$$\gamma_3 = \theta_6 - \theta_1 \Rightarrow c_6 = 1,\ c_1 = -1$$

На лекции давалась следующая оценка контраста:

$$\hat{\gamma} = \sum_{j=1}^{k} c_j \cdot \overline{X}_{\bullet j}$$

То есть в нашем случае:

$$\hat{\gamma}_1 = \overline{X}_{\bullet 2} - \overline{X}_{\bullet 1} = 1.8$$
$$\hat{\gamma}_2 = \overline{X}_{\bullet 4} - \overline{X}_{\bullet 1} = 5.6$$
$$\hat{\gamma}_3 = \overline{X}_{\bullet 6} - \overline{X}_{\bullet 1} = 12.6$$

Также из лекции возьмём:

$$\hat{\gamma} \sim N\left(\gamma,\, \sigma^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}\right)$$

Замечание: индекс здесь опущен, так как это верно для всех подобных оценок.

В нашем случае $\sum\limits_{j=1}^{k} \frac{c_j^2}{n_j} = \frac{2}{6}$. Строить доверительные интервалы гауссовских величин мы умеем:

$$\frac{\hat{\gamma} - \gamma}{\hat{\sigma}\sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}}} = \frac{\hat{\gamma} - \gamma}{\hat{\sigma}\sqrt{\frac{2}{6}}} \sim t(N-k) \sim t(24)$$

Здесь $\hat{\sigma} = \frac{1}{N-k} SS_{\text{случ.}} = \frac{224.4}{24} \approx 9.35$.

Теперь можно построить интервал в общем виде:

$$P\left(\hat{\gamma} - t_{0.975,\,24}\hat{\sigma}\sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}} < \gamma < \hat{\gamma} + t_{0.975,\,24}\hat{\sigma}\sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}}\right) = 0.95$$

Подсмотрели $t_{0.975,\,24} \approx 2.064$.

Посчитаем интервалы для каждого контраста:

1. Для γ_1:
$$P\left(-9.342 < \gamma < 12.942\right) = 0.95$$
2. Для γ_2:
$$P\left(-5.542 < \gamma < 16.742\right) = 0.95$$
3. Для γ_3:
$$P\left(1.458 < \gamma < 23.742\right) = 0.95$$

Решение б)

Знаем только, что функции распределения величин непрерывны и могут различаться только матожиданием.

Значит будем применять критерий Краскела-Уоллиса.

Проверяем гипотезу $H_0 : \theta_1 = \cdots = \theta_k = \theta$ против $H_1 : \exists i : \theta_i \neq \theta$

Нам понадобится таблица рангов:

Группа 1	Группа 2	Группа 3	Группа 4	Группа 5	Группа 6
5.5	2	9	9	27.5	23.5
7	5.5	20	14	17	21.5
3.5	17	13	17	26	30
11.5	11.5	3.5	17	21.5	29
1	9	17	23.5	25	27.5

Пользуясь этими рангами, составляем статистику:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{r}_{\bullet j} - \frac{N+1}{2}\right)^2$$

Посчитаем средние ранги по столбцам:

$$\overline{r}_{\bullet\,1} = \frac{1}{n_1}\sum_{i=1}^{n_1} r_{i\,1} = 5.7$$

$$\overline{r}_{\bullet\,2} = \frac{1}{n_2}\sum_{i=1}^{n_2} r_{i\,2} = 9$$

$$\overline{r}_{\bullet\,3} = \frac{1}{n_3}\sum_{i=1}^{n_3} r_{i\,3} = 12.5$$

$$\overline{r}_{\bullet\,4} = \frac{1}{n_4}\sum_{i=1}^{n_4} r_{i\,4} = 16.1$$

$$\overline{r}_{\bullet\,5} = \frac{1}{n_5}\sum_{i=1}^{n_5} r_{i\,5} = 23.4$$

$$\overline{r}_{\bullet\,6} = \frac{1}{n_6}\sum_{i=1}^{n_6} r_{i\,6} = 26.3$$

Подставим подсчитанные значения в статистику и получим:

$$H \approx 21.077$$

При справедливости H_0 должно выполняться:

$$H \sim \chi^2(k-1) \sim \chi^2(5)$$

Нужный нам квантиль $\chi^2_{0.95,\,5} \approx 11.07$ Критическая область справа, статистика туда попала, значит в H_0 мы не верим.

Решение в)

В этом случае будем пользоваться критерием Джонкхиера.
Проверяем гипотезу $H_0 : \theta_1 = \cdots = \theta_k = \theta$ против альтернативы $H_1 : \theta_1 \leqslant \ldots \theta_k$, где хотя бы одно неравенство строгое.
Для подсчёта статистики понадобится ввести функцию:

$$\varphi(y,\,z) = \begin{cases} 1, & y < z \\ 0.5, & y = z \\ 0, & y > z \end{cases}$$

С её помощью зададим ещё одну функцию:

$$U_{l,\,m} = \sum_{i=1}^{n_l}\sum_{j=1}^{n_m} \varphi\left(x_{i\,j},\,x_{j\,m}\right)$$

Статистика в данном критерии выглядит следующим образом:

$$J = \sum_{1 \leqslant l < m \leqslant k} U_{l,\,m}$$

Посчитаем функции U:

$$U_{0\,1} = 17.0$$
$$U_{0\,2} = 20.5$$
$$U_{1\,2} = 17.0$$
$$U_{0\,3} = 24$$
$$U_{1\,3} = 20.5$$
$$U_{2\,3} = 16.5$$
$$U_{0\,4} = 25$$
$$U_{1\,4} = 24.5$$
$$U_{2\,4} = 23.5$$
$$U_{3\,4} = 22.0$$
$$U_{0\,5} = 25$$
$$U_{1\,5} = 25$$
$$U_{2\,5} = 25$$
$$U_{3\,5} = 23.5$$
$$U_{4\,5} = 18.0$$

Сумма этого добра будет статистикой J:

$$J = 327$$

Из лекции известно, что при $\min(n_1, \ldots, n_k) \to \infty$ справедливо:

$$J^* = \frac{J - EJ}{\sqrt{\mathcal{D}J}} \sim N(0,\ 1)$$

Поскольку не все люди могут пальцами одной руки посчитать 5, мы будем пользоваться этой апроксимацией. Из лекции знаем следующие прикольные факты:

$$\begin{cases} EJ = \frac{1}{4}\left(N^2 - \sum\limits_{i=1}^{k} n_i^2\right) = 187.5 \\ \mathcal{D}J = \frac{1}{72}\left(N^2(2N+3) - \sum\limits_{i=1}^{k} n_i^2(2n_i+3)\right) \approx 760.417 \Rightarrow \sqrt{\mathcal{D}J} \approx 27.576 \end{cases}$$

Подставляем:

$$J^* = \frac{327 - 187.5}{27.576} \approx 5.059$$

Критическая область здесь $(Z_{0.95},\ +\infty)$, знаем $Z_{0.95} \approx 1.64$, то есть находимся глубоко в критической области и снова не верим в H_0.

Семинар 4 апреля

Работаем с величинами, измерямые в номинальных шкалах.

$A \backslash B$	B_1	$\ldots$	B_k	Σ
A_1	$n_{1\,1}$	$\ldots$	$n_{1\,k}$	$n_{1\,\bullet}$
$\vdots$	$\vdots$	$n_{i\,j}$	$\vdots$	$n_{i\,\bullet}$
A_m	$n_{m\,1}$	$\ldots$	$n_{m\,k}$	$n_{m\,\bullet}$
Σ	$n_{\bullet\,1}$	$\ldots$	$n_{\bullet\,k}$	n

Проверяем $H_0 : P(A = A_i,\ B = B_j) = p_{i\,\bullet} \cdot p_{j\,\bullet}$ против $H_1 : P(A = A_i,\ B = B_j) \neq p_{i\,\bullet} \cdot p_{\bullet\,j}$.
Используется статистика

$$\hat{\chi}^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{n\left(n_{i\,j} - \frac{n_{i\,\bullet} n_{\bullet\,j}}{n}\right)^2}{n_{i\,\bullet}\, n_{\bullet\,j}}$$

Для этой статистики справедливо:

$$\chi^2\Big|_{H_0} \sim \chi^2\big((k-1)(m-1)\big)$$

Задача

Исследуем влияние вакцины от холеры. Было 1630 привитых человек, из них 5 человек заболело. 1033 непривитых человека, из них заболело 11. Получаем два признака A — привит (П) / не привит (НП) и B — заболел (Б) / не заболел (НБ). Составим табилцу:

П\Б	Б	НБ	Σ
П	5	1625	1630
НП	11	1022	1033
Σ	16	2647	2663

Гипотезы $H_0,\ H_1$ написаны выше. Для таблиц 2×2 есть простая формула:

$$\hat{\chi}^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}\, n_{2\bullet}\, n_{\bullet 1}\, n_{\bullet 2}} \approx 6.02$$

Знаем, что должно выполняться:

$$\hat{\chi}^2\Big|_{H_0} \sim \chi^2(1)$$

Разделять доверительную и критическую область здесь будет квантиль $\chi^2_{0.95,\,1} \approx 3.84$. То есть статистика попала в критическую область.
Посчитаем коэффициенты контингенции и Юла соответственно:

$$\begin{cases} \Phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}\, n_{2\bullet}\, n_{\bullet 1}\, n_{\bullet 2}}} \approx -0.048 \\ Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \approx -0.56 \end{cases}$$

Задача

Собраны данные по ангийским ВУЗам. Рассматриваются несколько факультетов и пол студента. Имеется ли зависимость между полом студента и выбранной специализацией?

спец.\пол	М	Ж	Σ
Искусствоведение	197	223	420
Естественные науки	168	92	260
Соц.-эк. науки	115	105	220
Σ	480	420	900

Гипотезы не меняются. Строим статистику:

$$\hat{\chi}^2 = \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{n\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}} \approx 20.36$$

Разделять области будет $\chi^2_{0.95,\,2} \approx 5.99$, то есть попали в критическую область.
Посчитаем коэффициенты Пирсона и Краммера соответственно:

$$\begin{cases} P = \sqrt{\frac{\hat{\chi}^2}{\hat{\chi}^2 + n}} \approx 0.148 \\ C = \sqrt{\frac{\hat{\chi}^2}{n \cdot \min\{(k-1),\,(m-1)\}}} \approx 0.15 \end{cases}$$

То есть зависимость между полом и направлением имеется, но она не очень большая.

ДЗ к 11 апреля

Задача 1

Исследовалась зависимость между удовлетворенностью образом жизни и материальным положением семьи. На вопрос об удовлетворённости образом жизни каждый респондент должен дать один из ответов — не удовлетворён, удовлетворён, на вопрос о материальном положении — имеет низкий уровень или имеет высокий уровень. Было опрошено 665 респондентов. 68 из 225 респондентов, имевших низкий уровень материального положения, были удовлетворены своим образом жизни; среди 430 респондентов с высоким материальным положением своим образом жизни были удовлетворены 357. Являются ли показатели «материальное положение семьи» и «удовлетворенность образом жизни» зависимыми? Вычислите коэффициенты контингенции и ассоциации Юла этих показателей.

Решение

Исследуется связь между материальным положением семьи (A) и удовлетворённостью образом жизни (B). Данные:

Таблица 1: Таблица сопряжённости

$A\backslash B$	Не удовлетворён	Удовлетворён	Всего
Низкий уровень	157	68	225
Высокий уровень	73	357	430
Всего	230	425	665

H_0 : переменные независимы против H_1 : переменные зависимы.
Посчитаем статистику:

$$\hat{\chi}^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} = \frac{665 \cdot (157 \cdot 357 - 68 \cdot 73)^2}{230 \cdot 425 \cdot 225 \cdot 430} \approx 180.74$$

Критическая область в нашем случае начинается с точки $\chi^2_{0.95\,1} = 3.84$, то есть мы глубоко в критической области.
Посчитаем коэффициент контингенции и ассоциации соответственно:

$$\Phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}} \approx 0.525, \quad Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \approx 0.837.$$

Интерпретация: Сильная положительная связь ($Q \approx 0.84$), материальное положение влияет на удовлетворённость.

Задача 2

В результате проведенного исследования было установлено, что у 782 светлоглазых отцов сыновья тоже имеют светлые глаза, а у 89 светлоглазых отцов сыновья — темноглазые. У 50 темноглазых отцов сыновья также темноглазые, а у 79 темноглазых отцов сыновья — светлоглазые. Проверить гипотезу о независимости между цветом глаз отцов (СВ X) и цветом глаз их сыновей (СВ Y). Уровень доверия принять равным 0.99

Решение

$Y\backslash X$	Тёмный	Светлый	Всего
Тёмный	50	89	139
Светлый	79	782	861
Всего	129	871	1000

Проверяем H_0: переменные независимы против H_1: переменные зависимы.
Считаем статистику:

$$\hat{\chi}^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} = \frac{1000 \cdot (50 \cdot 782 - 89 \cdot 79)^2}{129 \cdot 871 \cdot 861 \cdot 139} \approx 76.479$$

Критическая область в нашем случае начинается с числа $\chi^2_{0.99,\,1} \approx 6.635$, то есть мы попали глубоко в критическую область, значит гипотеза о независимости отвергается.

Задача 3

Утверждается, что результат действия лекарства зависит от способа его применения. Проверьте эту гипотезу по данным, представленным в таблице. Уровень доверия принять равным 0.95.

$X\backslash Y$	A	B	C	Всего
неблагоприятный	11	17	16	44
благоприятный	20	23	19	62
Всего	31	40	35	106

Проверяем H_0 : переменные независимы против H_1 : переменные зависимы.
Посчитаем статистику для табилцы 2×3:

$$\hat{\chi}^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{n\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}} = 106 \cdot \left(\frac{\left(11 - \frac{31\cdot 44}{106}\right)^2}{31 \cdot 44} + \frac{\left(17 - \frac{40\cdot 44}{106}\right)^2}{40 \cdot 44} + \frac{\left(16 - \frac{35\cdot 44}{106}\right)^2}{35 \cdot 44} \right) \approx 0.735$$
$$+ 106 \cdot \left(\frac{\left(20 - \frac{31\cdot 62}{106}\right)^2}{31 \cdot 62} + \frac{\left(23 - \frac{40\cdot 62}{106}\right)^2}{40 \cdot 62} + \frac{\left(19 - \frac{35\cdot 62}{106}\right)^2}{35 \cdot 62} \right)$$

Критическая область в нашем случае начинается с точки $\chi^2_{2\,0.95} \approx 5.992$, то есть попали в доверительную область, можно принять гипотезу о независимости переменных.

Задача 4

Изучается влияние денежного стимулирования на производительность труда. Шести однородным группам, по 5 человек, раздали задачи одинаковой сложности. Задачи были выданы каждому члену группы независимо от остальных. Группы различаются только по денежному вознаграждению за каждую решенную задачу. Величина вознаграждения зависит от номера группы: чем больше номер группы, тем больше вознаграждение. Каждой группе известна цена вознаграждения за решенную задачу. В таблице представлено количество решенных задач каждым членом группы.

Группа 1	Группа 2	Группа 3	Группа 4	Группа 5	Группа 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Нам нужно построить три оценки для контрастов:

$$\gamma_1 = \theta_2 - \theta_1 \Rightarrow c_2 = 1,\ c_1 = -1$$
$$\gamma_2 = \theta_4 - \theta_1 \Rightarrow c_4 = 1,\ c_1 = -1$$
$$\gamma_3 = \theta_6 - \theta_1 \Rightarrow c_6 = 1,\ c_1 = -1$$

На лекции давалась следующая оценка контраста:

$$\hat{\gamma} = \sum_{j=1}^{k} c_j \cdot \overline{X}_{\bullet j}$$

То есть в нашем случае:

$$\hat{\gamma}_1 = \overline{X}_{\bullet 2} - \overline{X}_{\bullet 1} = 1.8$$
$$\hat{\gamma}_2 = \overline{X}_{\bullet 4} - \overline{X}_{\bullet 1} = 5.6$$
$$\hat{\gamma}_3 = \overline{X}_{\bullet 6} - \overline{X}_{\bullet 1} = 12.6$$

Также из лекции возьмём:

$$\hat{\gamma} \sim N\left(\gamma,\ \sigma^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}\right)$$

Замечание: индекс здесь опущен, так как это верно для всех подобных оценок.

В нашем случае $\sum_{j=1}^{k} \frac{c_j^2}{n_j} = \frac{2}{6}$. Строить доверительные интервалы гауссовских величин мы умеем:

$$\frac{\hat{\gamma} - \gamma}{\hat{\sigma}\sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}}} = \frac{\hat{\gamma} - \gamma}{\hat{\sigma}\sqrt{\frac{2}{6}}} \sim t(N-k) \sim t(24)$$

Здесь $\hat{\sigma} = \frac{1}{N-k} SS_{\text{случ.}} = \frac{224.4}{24} \approx 9.35$.

Теперь можно построить интервал в общем виде:

$$P\left(\hat{\gamma} - t_{0.975,\,24}\hat{\sigma}\sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}} < \gamma < \hat{\gamma} + t_{0.975,\,24}\hat{\sigma}\sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}}\right) = 0.95$$

Подсмотрели $t_{0.975,\,24} \approx 2.064$.

Посчитаем интервалы для каждого контраста:

1. Для γ_1:
$$P\left(-9.342 < \gamma < 12.942\right) = 0.95$$
2. Для γ_2:
$$P\left(-5.542 < \gamma < 16.742\right) = 0.95$$
3. Для γ_3:
$$P\left(1.458 < \gamma < 23.742\right) = 0.95$$

Семинар 11 апреля

Задача 1

В 2009г. центром исследования гражданского общества и некоммерческого сектора НИУ ВШЭ была сформирована репрезентативная выборка из 2000 респондентов. Среди ста вопросов анкеты были, в частности, такие: 1) какое из шести перечисленных описаний точнее всего соответствует материальному положению вашей семьи; 2)удовлетворены ли вы своим здоровьем. На первый вопрос предлагались ответы:
– денег не хватает даже на питание (категория A1);
– на питание денег хватает, но не хватает на покупку одежды и обуви (категория A2);
– на покупку одежды и обуви денег хватает, но не хватает на покупку бытовой техники (категория A3);
– денег вполне хватает на покупку крупной бытовой техники, но не можем купить новый автомобиль (категория A4);
– денег хватает на все, кроме таких дорогих приобретений, как квартира, дом (категория A5);
– материальных затруднений не испытываем, при необходимости могли бы приобрести квартиру, дом (категория A6).
Ответы на второй вопрос: удовлетворен (категория B1) и не удовлетворен (категория B2). Результаты опроса представлены в таблице сопряженности признаков A (материальное положение семьи) и B (удовлетворенность состоянием своего здоровья).

	b_1	b_2	Всего
a_1	83	154	237
a_2	278	354	632
a_3	478	299	769
a_4	204	76	280
a_5	46	20	66
a_6	13	3	16
Всего	1094	906	2000

Посчитаем λ меры прогнозов:

$$\hat{p}_1^{(B)} = 1 - \frac{b_{1\bullet}}{n} = \frac{906}{2\,000} \approx 0.453$$

$$\hat{p}_2^{(B)} = 1 - \frac{154 + 354 + 478 + 204 + 46 + 13}{n} = \frac{759}{2\,000}$$

$$\lambda_B = \frac{\hat{p}_1^{(B)} - \hat{p}_2^{(B)}}{\hat{p}_1^{(B)}} = \frac{147}{906} \approx 0.16$$

$$\hat{p}_1^{(A)} = 1 - \frac{a_{\bullet 3}}{n} = \frac{1\,231}{2\,000}$$

$$\hat{p}_2^{(A)} = 1 - \frac{478 + 354}{n} = \frac{1\,176}{2\,000}$$

$$\lambda_A = \frac{\hat{p}_1^{(A)} - \hat{p}_2^{(A)}}{\hat{p}_1^{(A)}} = \frac{55}{1231} \approx 0.045$$

Задача 2

У 12 школьников изучались две характеристики: оценки IQ, определённые с помощью шкалы интеллекта Стенфорда-Бине (показатель X) и успеваемость по химии, оцененная на основе теста из 35 вопросов (показатель Y). Данные внесены в следующую таблицу.
Проверяем гипотезу $H_0 : \tau_{X\,Y} = 0,\ H_1 : \tau_{X\,Y} > 0$.

X	Y
122	31
112	25
110	19
120	24
103	17
126	28
113	18
114	20
106	16
108	15
128	27
109	21

Решение через критерий Спирмена

Если посчитать головой ранги, то можно получить:

$$\sum_{i=1}^{12} (R_i - S_i)^2 = 4 + 9 + 0 + 1 + 4 + 0 + 9 + 4 + 0 + 4 + 4 + 9 = 48$$

Тогда статистика:

$$\rho_S = 1 - \frac{6 \cdot 48}{11^3 - 11} \approx 0.84$$

Альтернатива односторонняя, поэтому критичесая область будет $(S_{0.95,\,12},\ +\infty) = (0.503,\ +\infty)$, то есть принимаем альтернативную гипотезу.

Решение через критерий Кендала

Перерисуем таблицу через ранги:

X	Y
1	3
2	2
3	1
4	7
5	5
6	9
7	4
8	6
9	8
10	12
11	11
12	10

Произведём подсчёты:

$$K = 2 + 1 + 0 + 3 + 1 + 3 + 0 + 0 + 0 + 2 + 1 + 0 = 13$$

Теперь можно посчитать статистику:

$$\hat{\tau} = 1 - \frac{4 \cdot 13}{12 \cdot 11} \approx 0.61$$

Критическая область здесь $(\tau_{12,\,0.95},\ +\infty) = (0.394,\ +\infty)$, то есть попали в критическую область и снова опровергаем H_0.

ДЗ к 18 апреля

Задача 1

Имеются данные об индексе счастья за 2023 год (показатель X) и ВВП на душу населения по паритету покупательской способности (в долларах США) за 2022 год согласно оценке Всемирного Банка (показатель Y) по 18 странам

Страна	X	Y
Финляндия	7.84	59.03
Дания	7.62	74.0
Швейцария	7.57	83.6
Исландия	7.55	69.08
Германия	7.16	63.15
Канада	7.1	58.4
Великобритания	7.06	54.6
США	6.95	76.33
Франция	6.69	55.49
Италия	6.48	51.87
Узбекистан	6.18	9.53
Латвия	6.03	39.96
Аргентина	5.93	26.51
Греция	5.72	36.84
Россия	5.48	34.64
Китай	5.34	21.48
Индия	4.06	8.38
Афганистан	2.4	1.67

Можно ли считать (на уровне значимости 0.01), что показатели индекса счастья и ВВП на душу населения зависимы? Если да, то охарактеризуйте эту зависимость.

Решение

Для решения этой задачи воспользуемся критерием Спирмена. Чтобы применять критерий, нужно ранжировать данные: Проверяем гипотезу $H_0 : \rho_S = 0$ против $H_1 : \rho_S \neq 0$

Страна	R	S	$(R-S)^2$
Финляндия	1	4	9
Дания	2	2	0
Швейцария	3	1	4
Исландия	4	3	1
Германия	5	5	0
Канада	6	6	0
Великобритания	7	7	0
США	8	2	36
Франция	9	8	1
Италия	10	9	1
Узбекистан	11	15	16
Латвия	12	10	4
Аргентина	13	12	1
Греция	14	11	9
Россия	15	13	4
Китай	16	14	4
Индия	17	16	1
Афганистан	18	18	0

С помощью этих рангов можем посчитать коэффициент ранговой корреляции Спирмена:

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n(n^2-1)} = 1 - \frac{6 \cdot 91}{18(324-1)} = 1 - \frac{546}{5814} \approx 0.906$$

Ищем квантиль $\rho_{18,\,0.995}$ (потому что интервал двусторонний, но я смог найти только $\rho_{18,\,0.99} \approx 0.6$), то есть считаем доверительным интервал $(-0.6,\ 0.6)$, куда мы не попали, то есть гипотеза о независимости опровергается.

Ответ

Присутствует сильная положительная зависимость

Задача 2

Задачу №1 дополните (используя ссылки ниже) данными по 20 странам (на Ваш выбор) об индексе счастья за 2023 год (показатель X) и ВВП на душу населения по паритету покупательской способности (в долларах США) за 2022 год согласно оценке Всемирного Банка (показатель Y). Вычислите выборочный коэффициент корреляции и ранговый коэффициент корреляции Спирмена для новой выборки объема 38.

Решение

Для начала добавим двадцать новых стран (см. следующую страницу). На самом деле я не имею понятия, откуда были взяты числа из предыдущего задания, поэтому попросил ИИшку составить всю таблицу заново.

Таблица 2: Точные данные индекса счастья (X) и ВВП на душу населения (Y) для 38 стран

Страна	X (Happiness Index 2023)	Y (GDP per capita 2023, $)
Исходные данные (18 стран)		
Финляндия	7.804	59.030
Дания	7.586	74.000
Швейцария	7.571	83.600
Исландия	7.553	69.080
Германия	7.312	63.150
Канада	7.032	58.400
Великобритания	6.944	54.600
США	6.894	76.330
Франция	6.661	55.490
Италия	6.467	51.870
Узбекистан	6.174	9.530
Латвия	6.032	39.960
Аргентина	5.932	26.510
Греция	5.723	36.840
Россия	5.477	34.640
Китай	5.339	21.480
Индия	4.036	8.380
Афганистан	2.404	1.670
Дополнительные данные (20 стран)		
Нидерланды	7.464	64.572
Швеция	7.395	66.209
Норвегия	7.315	106.594
Новая Зеландия	7.123	48.424
Австрия	7.097	56.788
Австралия	7.057	65.366
Бельгия	6.927	53.656
Ирландия	6.911	102.217
Япония	6.392	33.950
Южная Корея	6.301	36.193
Польша	6.295	18.741
Чехия	6.290	27.638
Бразилия	6.202	10.412
Мексика	6.168	11.091
Турция	5.850	10.590
ЮАР	5.536	6.531
Индонезия	5.466	4.871
Филиппины	5.200	3.976
Вьетнам	5.033	4.284
Нигерия	4.552	2.184

Данные здесь неточные (почему-то deepseek не хочет копировать реальные значение и ставит какие-то похожие на них), но вручную 38 стран переписывать с двух сайтов я точно не буду.
Выборочный коэффициент корреляции считается по формуле:

$$\hat{\rho}_{X\,Y} = \frac{\sum\limits_{i=1}^{n} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(x_i - \overline{X}\right)^2 \sum\limits_{i=1}^{n} \left(y_i - \overline{Y}\right)^2}} \approx 0.576$$

Посчитал в уме.

Ранговый коэффициент корреляции Спирмена (ранги и сам коэффициент я тоже посчитал в уме):

$$\rho_S = 1 - \frac{6\sum\limits_{i=1}^{n}(R_i - S_i)^2}{n(n^2-1)} \approx 0.685$$

Задача 3

Имеются сведения о возрасте X (в годах) и среднемесячной заработной плате Y (в тыс. рублей) 30 сотрудников некоторой организации.

X	Y
19	8,5
20	9,2
22	11,2
24	10,4
28	16,2
30	17,4
31	14,3
32	24,9
34	22,8
37	20,8
39	34,1
40	30,4
41	28,8
43	33,8
45	35,3
46	34,4
47	32,3
48	29,4
50	31,8
51	30,3
53	28,7
54	31,9
55	25,5
58	19,9
60	22,3
62	20,6
65	18,3
68	14,7
70	14,1
72	15,0

Применяя критерий Спирмена, проверьте гипотезу о том, что возраст сотрудника и величина его заработной платы независимы. Уровень значимости считать равным 0,05.

Решение

Просят применить критерий Спирмена, значит надо ранжировать $\left(\top \frown \top\right)$

№	X	Y	Ранг X (R_X)	Ранг Y (R_Y)	$(R_X - R_Y)^2$
1	19	8,5	1	1	0
2	20	9,2	2	2	0
3	22	11,2	3	4	1
4	24	10,4	4	3	1
5	28	16,2	5	9	16
6	30	17,4	6	10	16
7	31	14,3	7	6	1
8	32	24,9	8	17	81
9	34	22,8	9	16	49
10	37	20,8	10	14	16
11	39	34,1	11	28	289
12	40	30,4	12	23	121
13	41	28,8	13	20	49
14	43	33,8	14	27	169
15	45	35,3	15	30	225
16	46	34,4	16	29	169
17	47	32,3	17	26	81
18	48	29,4	18	21	9
19	50	31,8	19	24	25
20	51	30,3	20	22	4
21	53	28,7	21	19	16
22	54	31,9	22	25	9
23	55	25,5	23	18	25
24	58	19,9	24	12	144
25	60	22,3	25	15	100
26	62	20,6	26	13	169
27	65	18,3	27	11	256
28	68	14,7	28	7	441
29	70	14,1	29	5	576
30	72	15,0	30	8	484

Сумма квадратов разностей рангов:

$$\sum_{i=1}^{n} (R_X - R_Y)^2 = 3\,542$$

Посчитаем ранговый коэффициент корреляции Спирмена:

$$\rho_S = 1 - \frac{6 \cdot 3\,542}{30 \cdot 899} \approx 0.212$$

Опять-таки критическая область двусторонняя, значит в идеале надо искать $\rho_{30,\,0.975}$, но ничего лучше $\rho_{30,\,0.95} \approx 0.36$ я не нашёл, то есть доверительным считаем интервал $(-0.36,\ 0.36)$, сюда попали, значит принимаем гипотезу об отсутствии корреляции.

Ответ

Корреляция весьма мала, то есть монотонная связь этих величин слаба. Нельзя утверждать об отсутсвии зависимостей другого характера.

Задача 4

Изучается взаимосвязь между зоркостью правого и левого глаза. Зоркость каждого глаза соответствует одной из четырёх категорий – высшая, вторая, третья и низшая. По результатам обследования 3242 мужчин в воз-

расте 30-39 лет получены следующие данные о зоркости.

Правый/Левый	высшая	вторая	третья	низшая	
высшая	821	112	85	35	1053
вторая	116	494	145	27	782
третья	72	151	583	87	893
низшая	43	34	106	333	516
	1052	791	919	482	3244

Оцените меры прогноза Гутмана. Прокомментируйте полученный результат (суммы я добавил сам, получилось, что обследовали на два человека больше, чем обещали).

Решение

Если будем угадывать зоркость левого (L) глаза без знания зоркости правого (R), вероятность ошибки будет:

$$\hat{p}_1^{(L)} = 1 - \frac{1052}{3244} \approx 0.676$$

Зная зоркость правого глаза:

$$\hat{p}_2^{(L)} = 1 - \frac{821 + 494 + 583 + 333}{3244} \approx 0.312$$

Теперь можно посчитать коэффициент Гутмана:

$$\lambda_L = \frac{\hat{p}_1^{(L)} - \hat{p}_2^{(L)}}{\hat{p}_1^{(L)}} \approx 0.5385$$

Теперь для другого глаза:

$$\hat{p}_1^{(R)} = 1 - \frac{1053}{3244} \approx 0.675$$

Со знанием зоркости другого глаза:

$$\hat{p}_2^{(R)} = 1 - \frac{821 + 494 + 583 + 333}{3244} \approx 0.538$$

Посчитаем коэффициент Гутмана:

$$\lambda_R = \frac{\hat{p}_1^{(R)} - \hat{p}_2^{(R)}}{\hat{p}_1^{(R)}} \approx 0.5378$$

Ответ

Зная зоркость второго глаза, можно вдвое чаще угадывать зоркость первого, может быть связано с тем, что зоркость глаз теряется одновременно.

Семинар 18 апреля

Задача

В задаче про заработную плату попробуем воспользоваться критерием хи-квадрат. Для этого нужно разбить значения на интервалы:

$X \backslash Y$	$0-21$	$21+$	
< 35	7	2	9
$35 \dots 55$	1	13	14
> 55	6	1	7
	14	16	30

Проверяем гипотезу $H_0 : \forall x,\ y \quad F_{x\,y}(x,\ y) = F_x(x) \cdot F_y(y)$ против $H_1 : \exists x,\ y \quad F_{x\,y}(x,\ y) \neq F_x(x) \cdot F_y(y)$
Применяем критерий хи-квадрат:

$$\hat{\chi}^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{n\left(n_{i\,j} - \frac{n_{i\,\bullet} n_{\bullet\,j}}{n}\right)^2}{n_{i\,\bullet} n_{\bullet\,j}} \approx 16.47$$

Критическая область начинается с числа $\chi^2_{0.95,\,2} \approx 5.99$, то есть статистика попала в критическую область, значит принимается альтернативная гипотеза.

Задача

Рассматриваем задачу про связь индекса счастья с ВВП, проверяем независимость этих показателей.
$H_0 : \rho_{XY} = 0$ против $H_1 : \rho_{XY} > 0$.
По данным можно было посчитать $\hat{\rho}_{XY} = 0.86$.
С помощью этого коэффициента можно составить статистику:

$$T(x,\, y) = \frac{\sqrt{n-2}\hat{\rho}_{XY}}{\sqrt{1-\hat{\rho}^2_{XY}}}\Big|_{H_0} \sim t(n-2)$$

Если подставить наши числа, можно получить:

$$T(x,\, y) \approx 6.2$$

Квантиль $t_{0.95,\,18} \approx 1.73$, то есть мы уверенно попали в критическую область.
Попробуем построить доверительный интервал:

$$P\left(Z_{0.025} < \frac{\hat{\rho}_{xy} - E\hat{\rho}_{xy}}{\sqrt{\mathcal{D}\hat{\rho}_{xy}}} < Z_{0.975}\right) = 0.95$$

$$P\left(\hat{\rho}_{xy}\frac{1-\hat{\rho}^2_{xy}}{2n} + \hat{\rho}_{xy} - Z_{0.975}\frac{1-\hat{\rho}^2_{xy}}{\sqrt{n}} < \rho_{xy} < \hat{\rho}_{xy}\frac{1-\hat{\rho}^2_{xy}}{2n} + \hat{\rho}_{xy} + Z_{0.975}\frac{1-\hat{\rho}^2_{xy}}{\sqrt{n}}\right) = 0.95$$

$Z_{0.975} \approx 1.96$ (гауссовский квантиль). Подставим числа:

$$P\left(0.746 < \rho_{xy} < 0.987\right) = 0.95$$

Попробуем побаловаться с z-преобразованиями Фишера.

$$\hat{z} = \frac{1}{2}\ln\frac{1+\hat{\rho}_{xy}}{1-\hat{\rho}_{xy}}$$

Построим доверительный интервал:

$$P\left(Z_{0.025} < \frac{\hat{z} - E\hat{z}}{\sqrt{\mathcal{D}\hat{z}}} < Z_{0.975}\right)$$

$$P\left(-\frac{\hat{\rho}_{xy}}{2(n-1)} + \hat{z} - Z_{0.975}\sqrt{\frac{1}{n-3}} < \operatorname{arcth}\rho_{xy} < -\frac{\hat{\rho}_{xy}}{2(n-1)} + \hat{z} - Z_{0.975}\sqrt{\frac{1}{n-3}}\right)$$

Подставим числа:

$$P\left(0.762 < \operatorname{arcth}\rho_{xy} < 1.774\right) = 0.95$$

Применим к этому неравенству гиперболический тангенс:

$$P\left(0.642 < \rho_{xy} < 0.944\right) = 0.95$$

ДЗ к 25 апреля

Задача 1

Показать, что выборочный коэффициент корреляции можно преобразовать к виду (опечатка в оригинальном условии исправлена)

$$\hat{\rho}_{xy} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{X}\cdot\overline{Y}}{\sqrt{\sum\limits_{i=1}^{n} x_i^2 - n\overline{X}^2}\sqrt{\sum\limits_{i=1}^{n} y_i^2 - n\overline{Y}^2}}$$

Задача 1

Оригинальная формула выглядит вот так:

$$\hat{\rho}_{x\,y} = \frac{\sum\limits_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(x_i - \overline{X}\right)^2 \cdot \sum\limits_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}}$$

Преобразуем числитель дроби:

$$\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = \sum_{i=1}^{n}\left(x_i y_i - x_i\overline{Y} - y_i\overline{X} + \overline{X}\cdot\overline{Y}\right) = \sum_{i=1}^{n}\left(x_i y_i + \overline{X}\cdot\overline{Y}\right) - \overline{Y}\sum_{i=1}^{n}x_i - \overline{X}\sum_{i=1}^{n}y_i =$$
$$= \sum_{i=1}^{n}\left(x_i y_i\right) - n\overline{X}\cdot\overline{Y} - n\overline{X}\cdot\overline{Y} + n\overline{X}\cdot\overline{Y} = \sum_{i=1}^{n}\left(x_i y_i\right) - n\overline{X}\cdot\overline{Y}$$

Преобразуем знаменатель:

$$\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2} = \sqrt{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{X} + \overline{X}^2\right)\cdot\sum_{i=1}^{n}\left(y_i^2 - 2y_i\overline{Y} + \overline{Y}^2\right)} =$$
$$= \sqrt{\left(\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}2x_i\overline{X} + n\overline{X}^2\right)\left(\sum_{i=1}^{n}y_i^2 - \sum_{i=1}^{n}2y_i\overline{Y} + n\overline{Y}^2\right)}$$
$$= \sqrt{\left(\sum_{i=1}^{n}x_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right)\left(\sum_{i=1}^{n}y_i^2 - 2n\overline{Y}^2 + n\overline{Y}^2\right)}$$
$$= \sqrt{\sum_{i=1}^{n}x_i^2 - n\overline{X}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\overline{Y}^2}$$

Тут можно понять, что в оригинальном условии должна быть опечатка (символы такие же, но от корня никак не избавиться, да и в условии получается несимметричная формула, хотя должно быть $\hat{\rho}_{x\,y} = \hat{\rho}_{y\,x}$)

Задача 2

В ДЗ семинара №2 требовалось вычислить выборочный коэффициент корреляции между экзаменационными оценками студентов вашей группы по двум (на ваш выбор) предметам. Используя эти данные, проверьте гипотезу: а) о некоррелированности показателей оценок по двум дисциплинам; б) о независимости оценок по двум дисциплинам.

Решение

(Я фанат сводных таблиц, кстати). Вот уже разделённые на отрезки оценки:

$C++\backslash AaDS$	Уд	Хор	Отл	
Уд	5	2	0	7
Хор	3	8	8	19
Отл	1	1	1	3
	9	11	9	29

А вот просто табличка с оценками:

Mark \ Subj	$C++$	$AaDS$
4	2	3
5	5	6
6	11	8
7	8	3
8	3	5
9	0	2
10	0	2

а)

Проверяем гипотезу $H_0 : \rho_{xy} = 0$ против $H_1 : \rho_{xy} \neq 0$ Для определения коррелированности посчитаем выборочную корреляцию по формуле, доказанной в задаче 1:

$$\hat{\rho}_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{X} \cdot \overline{Y}}{\sqrt{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(y_i - \overline{Y}\right)^2}}$$

Посчитаем всё необходимое (считал по таблице индивидуальных оценок, то есть $x_i y_i$ это произведение оценок одного и того же человека):

$$\overline{X} = \frac{1}{29}\left(2 \cdot 4 + 5 \cdot 5 + 11 \cdot 6 + 8 \cdot 7 + 3 \cdot 8 + 0 \cdot 9 + 0 \cdot 10\right) \approx 6.172$$

$$\overline{Y} = \frac{1}{29}\left(3 \cdot 4 + 6 \cdot 5 + 8 \cdot 6 + 3 \cdot 7 + 5 \cdot 8 + 2 \cdot 9 + 2 \cdot 10\right) \approx 6.517$$

$$\sum_{i=1}^{n} x_i y_i = 1193$$

$$\sqrt{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2} \approx 5.669$$

$$\sqrt{\sum_{i=1}^{n} \left(y_i - \overline{Y}\right)^2} \approx 9.124$$

Подставив посчитанные числа, получаем:

$$\hat{\rho} \approx 0.511$$

По материалам лекции знаем, что выполняется следующее

$$T(x,\ y) = \left.\frac{\hat{\rho}_{xy}\sqrt{n-2}}{\sqrt{1-\hat{\rho}_{xy}^2}}\right|_{H_0} \sim t(n-2)$$

Посчитаем статистику:

$$T(x,\ y) \approx 3.089$$

Доверительной областью в нашем случае является $(t_{0.025,\ 27},\ t_{0.975,\ 27}) \approx (-2.052,\ 2.052)$, сюда мы не попали, значит оценки коррелированы.

б)

Теперь проверяем гипотезу $H_0 : \forall i,\ j \quad P(x = x_i,\ y = y_i) = P(x = x_i) \cdot P(y = y_i)$ против $H_1 : \neg H_0$ (отрицание)
Можем воспользоваться критерием хи-квадрат. Посчитаем статистику:

$$\hat{\chi}^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{n\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{n_{i\bullet} n_{\bullet j}} \approx 8.302$$

Критическая область начинается с точки $\chi^2_{0.95,\ 4} \approx 9.488$, то есть мы находимся в доверительной области, значит принимается гипотеза о независимости.

Ответ

а) коррелированы, б) независимы

Задача 3

Постройте асимптотический ДИ уровня надежности 0.95 для коэффициента корреляции показателей оценок по двум дисциплинам. Постройте ДИ уровня надежности 0.95 для коэффициента корреляции этих показателей, используя преобразование Фишера.

Решение

В прошлом задании посчитали $\hat{\rho}_{x\,y} \approx 0.511$, из лекций знаем, что выполняется

$$\frac{\hat{\rho}_{x\,y} - E\hat{\rho}_{x\,y}}{\sqrt{\mathcal{D}\hat{\rho}_{x\,y}}} \sim N(0,\ 1)$$

Математическое ожидание и дисперсия $\hat{\rho}_{x\,y}$ записаны в лекции, подставив их и сделав преобразования, получаем

$$P\left(\hat{\rho}_{x\,y}\frac{1-\hat{\rho}_{x\,y}^2}{2n} + \hat{\rho}_{x\,y} - Z_{0.975}\frac{1-\hat{\rho}_{x\,y}^2}{\sqrt{n}} < \rho_{x\,y} < \hat{\rho}_{x\,y}\frac{1-\hat{\rho}_{x\,y}^2}{2n} + \hat{\rho}_{x\,y} + Z_{0.975}\frac{1-\hat{\rho}_{x\,y}^2}{\sqrt{n}}\right) = 0.95$$

$$P\left(0.249 < \rho_{x\,y} < 0.786\right) = 0.95$$

Можно побаловаться z-преобразованиями:

$$\hat{z} = \frac{1}{2}\ln\frac{1+\hat{\rho}_{x\,y}}{1-\hat{\rho}_{x\,y}}$$

На семинаре считали ДИ:

$$P\left(-\frac{\hat{\rho}_{x\,y}}{2(n-1)} + \hat{z} - Z_{0.975}\sqrt{\frac{1}{n-3}} < \operatorname{arcth}\rho_{x\,y} < -\frac{\hat{\rho}_{x\,y}}{2(n-1)} + \hat{z} + Z_{0.975}\sqrt{\frac{1}{n-3}}\right) = 0.95$$

Посчитаем

$$P\left(0.171 < \operatorname{arcth}\rho_{x\,y} < 0.939\right) = 0.95$$

Применим гиперболический тангенс

$$P\left(0.169 < \rho_{x\,y} < 0.735\right) = 0.95$$

Задача 4

Согласно переписи населения Швеции 1936 года из совокупности супружеских пар была получена выборка 25263 пары, вступивших в брак в 1931-1936 годах. В таблице приведено распределение годовых доходов (в тыс. крон) и количество детей (показатель X) у супружеских пар в этой выборке.

$X\backslash Y$	$0-1$	$1-2$	$2-3$	>3	
0	2 161	3 577	2 184	1 636	9 558
$1-2$	3 691	6 834	2 862	1 358	14 745
$\geqslant 3$	264	517	127	52	960
	6 116	10 928	5 173	3046	25263

Являются ли зависимыми уровень дохода семьи и количество детей в этой семье? Уровень значимости считать равным 0.05

Решение

Будем использовать критерий хи-квадрат, так как данные уже за нас разбили на интервалы.
Проверяем гипотезу $H_0 : \forall x,\ y \quad F_{x\,y} = F_x(x) \cdot F_y(y)$ против $H_1 : \exists x,\ y \quad F_{x\,y} \neq F_x(x) \cdot F_y(y)$
Посчитаем статистику:

$$\hat{\chi}^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{n\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}} \approx 547.467$$

Критическая область начинается с точки $\chi^2_{0.95,\,6} \approx 12.59$, поэтому принимаем гипотезу о зависимости.

Семинар 25 апреля

Задача

$\hat{\rho}_{1\,2} = 0.2, \hat{\rho}_{1\,3} = 0.41,\ \hat{\rho}_{2\,3} = 0.82$. Хотим узнать, коррелированы ли показатели 1 и 2.

Решение

Наблюдали за 100 рабочих, исследовали три показателя возраст, производительность, стаж. Проверяем H_0 : $\rho_{1\,2} = 0$ против $H_1 : \rho_{1\,2} > 0$
Для проверки таких гипотез используется статистика:

$$\left.\frac{\hat{\rho}_{1\,2}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2_{1\,2}}}\right|_{H_0} \sim t(n-2)$$

С нашими числами:

$$\frac{0.2 \cdot \sqrt{98}}{\sqrt{1-(0.2)^2}} = 2.02 > t_{0.95,\,98} \approx 1.65$$

Принимаем альтернативную гипотезу.
Теперь можем посчитать корреляцию при фиксированном третьем признаке:

$$\hat{\rho}_{1\,2;\,3} = \frac{\hat{\rho}_{1\,2} - \hat{\rho}_{2\,3}\hat{\rho}_{1\,3}}{\sqrt{(1-\hat{\rho}^2_{2\,3})(1-\hat{\rho}^2_{1\,3})}} \approx -0.26$$

Теперь проверим гипотезу $H_0 : \rho_{1\,2;\,3} = 0$ против $H_1 : \rho_{1\,2;\,3} \neq 0$:

$$\left.\frac{\hat{\rho}_{1\,2;\,3}\sqrt{n-3}}{\sqrt{1-\hat{\rho}_{1\,2;\,3}}} \approx -2.65\right|_{H_0} \sim t(n-3)$$

Критическая область $(-\infty,\ -1.65)$, мы в неё попали.

Задача

Автосалон предоставил сведенья о цене, ширине, длине, массе автомобилей. По ним посчитаны парные коэффициенты корреляции:

$$\hat{\rho}_{1\,2} = 0.33,\ \hat{\rho}_{1\,3} = 0.16,\ \hat{\rho}_{1\,4} = 0.53,\ \hat{\rho}_{2\,3} = 0.71,\ \hat{\rho}_{2\,4} = 0.72,\ \hat{\rho}_{3\,4} = 0.63$$

Нужно посчитать множественный коэффициент корреляции. Для этого сначала необходимо составить матрицу выборочных коэффициентов корреляции:

$$\hat{R} = \begin{pmatrix} 1 & 0.33 & 0.16 & 0.53 \\ 0.33 & 1 & 0.71 & 0.72 \\ 0.16 & 0.71 & 1 & 0.63 \\ 0.53 & 0.72 & 0.63 & 1 \end{pmatrix},\ \det \hat{R} \approx 0.15,\ \det \mathbb{R}_{1\,1} = 0.22$$

Можем оценить множественный коэффициент корреляции:

$$\hat{R}_{1\,(2,\,3,\,4)} = \sqrt{1 - \frac{0.15}{0.22}} = 0.564$$

Проверяем гипотезу $H_0 : R_{1\,(2,\,3,\,4)} = 0$ против $H_1 : R_{1\,(2,\,3,\,4)} > 0$.
Пользуемся статистикой с лекции:

$$\left.\frac{\frac{1}{l-1}\hat{R}^2_{1\,(2,\ldots l)}}{\frac{1}{n-l}\left(1-R^2_{1\,(2,\ldots l)}\right)}\right|_{H_0} \sim F(l-1,\ n-l)$$

В нашем случае $l = 4,$ (количество признаков) $n = 34$ (количетсво испытаний). А статистика чуть больше 4, критическая область начинается с точки 2.92, поэтому принимается альтернативная гипотеза.

Задача

Дана таблица с ранжированием бизнес проектов тремя различными экспертами:

A	1	4	2	5	3	7	6
B	2	1	3	4	5	6	7
C	2	1	4	5	3	7	6

Нужно посчитать коэффициент конкардации данных экспертов (обладают ли они одинаковой системой предпочтений). Формально проверяется гипотеза $H_0 : \forall x_1,\ x_2,\ x_3 \quad F_\xi(x_1,\ x_2,\ x_3) = \prod\limits_{i=1}^{3} F_{\xi_i}(x_i)$ против гипотезы $H_1 : \exists x_1,\ x_2,\ x_3 \quad F_\xi(x_1,\ x_2,\ x_3) \neq \prod\limits_{i=1}^{3} F_{\xi_i}(x_i)$

$$\hat{W}_n(m) = \frac{12}{m^2(n^3-n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} R_{i\,j} - m\frac{n+1}{2} \right) \approx 0.84$$

Далее применяем данные с лекции:

$$m(n-1)\hat{W}_n(m)\Big|_{H_0} \sim \chi^2(n-1)$$

Статистика принимает значение больше $\chi^2_{0.95,\,6} \approx 12.59$, значит мы принимаем H_1, то есть они обладают одинаковой системой предпочтений.

Домашнее задание к семинару 16 мая

Задача 1

Имеются данные по $n = 88$ странам за 2022 год о выборочных коэффициентах корреляции показателей ВВП на душу населения (X_1), проценте взрослого населения, страдающего ожирением (X_2) и уровнем миграции (X_3):

$$\hat{\rho}_{1\,2} = 0.463,\ \hat{\rho}_{1\,3} = 0.639,\ \hat{\rho}_{2\,3} = 0.148$$

Вычислите частный коэффициент корреляции показателей X_2 и X_3 при условии что, показатель X_1 фиксирован. Проверьте гипотезу о том, что $\rho_{2\,3} = 0$ и гипотезу $\rho_{2\,3;1} = 0$.

Решение

Посчитаем частный коэффициент корреляции при условии фиксированного первого признака (поскольку настоящие коэффициенты корреляции нам неизвестны, будем считать по выборке):

$$\hat{\rho}_{2\,3;1} = \frac{\hat{\rho}_{2\,3} - \hat{\rho}_{1\,2}\hat{\rho}_{1\,3}}{\sqrt{(1-\hat{\rho}^2_{1\,2})\,(1-\hat{\rho}^2_{1\,3})}} \approx -0.217$$

Теперь проверяем требуемые гипотезы (альтернатива не указана, поэтому возьмём общего вида):

$$H_0 : \rho_{2\,3} = 0,\ \text{против}\ H_1 : \rho_{2\,3} \neq 0$$

Для проверки такой гипотезы подходит следующая статистика:

$$\frac{\hat{\rho}_{2\,3}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2_{2\,3}}} \approx 1.388\Bigg|_{H_0} \sim t(n-2)$$

В нашем случае степеней свободы достаточно, чтобы принять распределение Стьюдента за гауссовское, поэтому квантиль $t_{0.975,\,86} \approx Z_{0.975} \approx 1.96$, то есть произошло попадание в доверительную область, значит принимаем H_0.
Осталось проверить:

$$H_0 : \rho_{2\,3;1} = 0,\ \text{против}\ H_1 : \rho_{2\,3;1} \neq 0$$

Для этих гипотез подойдёт статистика:

$$T = \frac{\hat{\rho}_{2\,3;1}\sqrt{n-2-a}}{\sqrt{1-\hat{\rho}^2_{2\,3;1}}}\Bigg|_{H_0} \sim t(n-2-a)$$

Здесь a — количество фиксированных элементов (в нашем случае такой 1)

$$T \approx -2.049$$

Попали в критическую область, значит верим в H_1

Ответ

$\hat{\rho}_{2\,3;1} \approx -0.217$, гипотеза $\rho_{2\,3} = 0$ принята, гипотеза $\rho_{2\,3;1} = 0$ опровергнута.

Задача 2

Имеются данные по 88 странам за 2022 год о выборочных коэффициентах корреляции показателей ВВП на душу населения (X_1), проценте взрослого населения, страдающего ожирением (X_2), коэффициенте чистой миграции (X_3) и проценте населения, находящемся за чертой бедности (X_4):

$$\hat{\rho}_{1\,2} = 0.463,\ \hat{\rho}_{1\,3} = 0.639,\ \hat{\rho}_{1\,4} = -0.4,\ \hat{\rho}_{2\,3} = 0.148,\ \hat{\rho}_{2\,4} = -0.34,\ \hat{\rho}_{3\,4} = -0.183$$

Оценить множественный коэффициент корреляции $R_{1\,(2,3,4)}$. Проверить гипотезу о том, что $R_{1\,(2,3,4)} = 0$, предполагая, что *данные имеют гауссовское распределение* (зачем?). Прокомментировать полученный результат.

Решение

Оценим множественный коэффициент корреляции. Для этого нужно составить матрицу выборочных корреляций (здесь $\xi = (\xi_1, \ldots, \xi_4)$ — случайная величина, порождающая данные по одному городу):

$$\hat{R}_\xi = \begin{pmatrix} 1 & 0.463 & 0.639 & -0.4 \\ 0.463 & 1 & 0.148 & -0.34 \\ 0.639 & 0.148 & 1 & -0.183 \\ -0.4 & -0.34 & -0.183 & 1 \end{pmatrix},\ \det \hat{R}_\xi \approx 0.357,\ \det \hat{\mathbb{R}}_{1\,1} \approx 0.847$$

Для подсчёта оценки множественной корреляции существует формула:

$$\hat{R}_{1\,(2,3,4)} = \sqrt{1 - \frac{\det \hat{R}_\xi}{\hat{\mathbb{R}}_{1\,1}}} \approx 0.421$$

Проверим гипотезу (выбрал такую альтернативу, потому что оценка сильно больше 0):

$$H_0 : R_{1\,(2,3,4)} = 0,\ \text{против}\ H_1 : R_{1\,(2,3,4)} > 0$$

Для этого подойдёт статистика:

$$T = \frac{n-l}{l-1} \cdot \frac{\hat{R}^2_{1\,(2,3,4)}}{1-\hat{R}^2_{1\,(2,3,4)}}\Bigg|_{H_0} \sim F(l-1,\ n-l)$$

Здесь l — количество элементов, от которых рассматривается зависимость (в нашем случае 3). Подставим числа:

$$T \approx 9.181 \sim F(2,\ 85)$$

Квантиль распределения фишера $f_{0.95,\ 2,\ 85} \approx 3.104$, то есть попали в критическую область. Принимается гипотеза H_1

Ответ

Оценка множественной корреляции $\hat{R}_{1\,(2,3,4)} \approx 0.421$. Гипотеза $R_{1\,(2,3,4)} = 0$ опровергается, то есть существует значимая зависимость между ВВП страны и комбинацией остальных признаков.

Задача 3

Четверо судей оценивают выступление фигуристов, вышедших в финал соревнований. Результаты распределения мест указаны в таблице

Судья 1	1	3	4	6	2	5
Судья 2	2	4	3	5	1	6
Судья 3	2	3	1	5	4	6
Судья 4	3	4	2	6	1	5

Можно ли считать, что данная судейская коллегия обладает общей системой предпочтений?

Решение

Формально на такой вопрос можно ответить, проверив гипотезы:

$$H_0 : \forall x_1, \ldots, x_m \in \mathbb{R}^1 \quad F_\xi(x_1, \ldots, x_m) = \prod_{i=1}^{m} F_{\xi_i}(x_i), \text{ против}$$

$$H_1 : \exists x_1, \ldots, x_m \in \mathbb{R}^1 \quad F_\xi(x_1, \ldots, x_m) \neq \prod_{i=1}^{m} F_{\xi_i}(x_i)$$

Для ответа на данный вопрос необходимо вспомнить товарища Кендалла:

$$\hat{W}(m) = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} R_{ij} - m\frac{n+1}{2} \right)^2$$

Здесь m — количество судей, n — количество оцениваемых объектов. С помощью языка программирования python данная статистика была посчитана:

$$\hat{W}(m) \approx 0.7$$

Таблицу с точными квантилями я не нашёл, поэтому буду аппроксимировать распределением χ^2:

$$m(n-1)\hat{W}(m) = 14\Big|_{H_0} \sim \chi^2(n-1)$$

Квантиль $\chi^2_{0.95,\,5} \approx 11.07$, то есть статистика попала в критическую область, значит принмается H_1

Ответ

Так считать нельзя.

Семинар 16 мая

Работаем с моделями вида

$$y_i = \theta x_i + \varepsilon_i,\ i = \overline{1,\ n},\ E\varepsilon_i = 0,\ K_\varepsilon = \sigma^2$$

Метод наименьших квадратов

Метод заключается в поиске такой θ, чтобы:

$$S(\theta) = \sum_{i=1}^{n} (y_i - \theta x_i)^2 \to \min_\theta$$

Для поиска минимума нужно взять частную производную:

$$\frac{\partial}{\partial\theta} S(\theta) = \sum_{i=1}^{n} -2(y_i - \hat{\theta}x_i)x_i = 0 \Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Задача 1

y_i — себестоимость, x_i — тираж в тысячах экземляров:

тираж	1	2	3	4	5
себестоимость	6	5	4	4	3

Известно:

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$$
$$E\varepsilon_i = 0$$
$$K_\varepsilon = \sigma^2 I$$
$$\varepsilon \sim N(0,\ \sigma^2 I)$$

Необходимо оценить $\theta_0,\ \theta_1$, проверить гипотезу $H_0 : \theta_1 = 0$, оценить σ_ε^2
Составим матрицы по данным:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix},\ Y = \begin{pmatrix} 6 \\ 5 \\ 4 \\ 4 \\ 3 \end{pmatrix},\ \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$$

На лекции вывели крутую формулу дял оценки θ:

$$\hat{\theta} = \left(X^T X\right)^{-1} X^T Y = \begin{pmatrix} 6.5 \\ -0.7 \end{pmatrix}$$

Пользуясь оценкой θ можем построить оценку y:

$$\hat{y}_i = 6.5 - 0.7 x_i \Rightarrow \hat{\varepsilon}_i = \begin{pmatrix} 0.2 \\ -0.1 \\ -0.4 \\ 0.3 \\ 0 \end{pmatrix}$$

Оценим σ_ε:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{5-(1+1)} \sum_{i=1}^{5} \hat{\varepsilon}_i^2 \approx \frac{1}{10}$$

Теперь проверяем гипотезу $H_0 : \theta = 0$

$$T = \frac{\hat{\theta}}{\sqrt{\hat{\sigma}_{\hat{\theta}}^2}}$$

$$K_{\hat{\theta}} = \sigma_\varepsilon^2 \left(X^T X\right)^{-1},\ \hat{\sigma}_{\hat{\theta}_1} = K_{\hat{\theta}}[2,\ 2] = \frac{1}{50} \cdot 5\hat{\sigma}_\varepsilon^2 = \frac{1}{100} \Rightarrow T = -7\Big|_{H_0} \sim t(n-(p+1))$$

Здесь $K_{\hat{\theta}}[2,\ 2]$ — элемент с индексом $(2,\ 2)$ из матрицы. Доверительной областью здесь будет $(t_{3,\ 0.025},\ t_{3,\ 0.975}) = (-3.182,\ 3.182)$. Статистика в доверительную область не попадает, значит H_0 опровергается.
Можно построить доверительный интервал:

$$f(x) = \theta_0 + \theta_1 x,\ \hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 x = 6.5 - 0.7x$$

$$E\hat{f}(x) = \theta_0 + \theta_1 x$$

$$\mathcal{D}\hat{f}(x) = \mathcal{D}\left(\hat{\theta}_0 + \hat{\theta}_1 x\right) = \mathcal{D}\left(\overline{Y} - \hat{\theta}_1 \overline{X} + \hat{\theta}_1 x\right) = \mathcal{D}\overline{Y} + (x - \overline{X})^2 \mathcal{D}\hat{\theta}_1 = \frac{1}{n^2}\mathcal{D}\varepsilon + \left(x - \overline{X}\right)^2 \sigma_\varepsilon^2 K_\theta[2,\ 2]$$

Все наблюдения гауссовские, поэтому

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\mathcal{D}\hat{f}(x)}} \sim N(0,\ 1)$$

Однако при подсчёте $\mathcal{D}\hat{f}(x)$ необходимо подставлять оценку $\hat{\sigma}_\varepsilon$, поэтому получается Стьюдент:

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\hat{\mathcal{D}}\hat{f}(x)}} = \frac{\hat{f}(x) - f(x)}{\hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \left(x - \overline{X}\right)^2 K_{\hat{\theta}}[2,\ 2]}} \sim t(n - (p+1))$$

С помощью этой статистики мы можем получить оценку нашей функции во всех точках. Например:

$$\hat{f}(6) = 2.3,\ G = t_{0.025}\sqrt{\frac{1}{50} + (6-3)^2 \frac{1}{10}}$$

ДЗ к семинару 23 мая

Задача 1

В таблице указана динамика веса поросят:

Возраст X	0	1	2	3	4	5	6
Вес Y	1.2	2.5	3.9	5.2	6.4	7.7	9.2

Предполагается, что справедлива модель вида:

$$Y_i = a + bX_i + \varepsilon_i,\ \varepsilon \sim N(0,\ \sigma^2 I)$$

МНК оценки a, b

МНК оценка для вектора $\theta = \begin{pmatrix} a \\ b \end{pmatrix}$:

$$\hat{\theta} = \left(X^T X\right)^{-1} X^T \cdot Y \approx \begin{pmatrix} 1.204 \\ 1.318 \end{pmatrix}$$

Оценка дисперсии

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Проверка гипотезы

$H_0 : b = 0$ против $H_1 : b \neq 0$

$$\frac{\hat{b}}{\hat{\sigma}_\varepsilon \sqrt{C_{2,\,2}}}\Big|_{H_0} \sim t(n-p-1)$$

Построить точечную и интервальную оценку

$$\hat{f}(x) = \hat{a} + \hat{b}x$$

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\hat{\mathcal{D}}\hat{f}(x)}} \sim t(n-p-1)$$

Задача 2

Имеются данные о количестве $NaNO_3$, которое можно растворить в 100 граммах воды в зависимости от температуры t:

t	Y
0	66.7
4	71
10	76.3
15	80.6
21	85.7
29	92.9
36	99.4
51	113.6
68	125.1

Модель $y_i = \theta_0 + \theta_1 t_i + \theta_2 t_i^2 + \varepsilon_i,\ \varepsilon \sim N(0,\ \sigma^2 I)$

Оценка параметров

$$\hat{\theta} = \begin{pmatrix} 66.706 \\ 0.96 \\ -0.001 \end{pmatrix}$$

Проверка гипотезы

Проверяем $H_0 : \theta_2 = 0$ против $H_1 : \theta_2 \neq 0$.
Посчитаем матрицу C:

$$C = \sigma_\varepsilon^2 \left(X^T X\right)^{-1} = \begin{pmatrix} 0.586 & -0.036 & 0.0004 \\ -0.036 & 0.0035 & -0.00005 \\ 0.0004 & -0.00005 & 0.0000007 \end{pmatrix}$$

Для этого необходимо оценить σ_ε:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y})^2 \approx 1.202$$

$$\frac{\hat{\theta}_2}{\hat{\sigma}_\varepsilon \sqrt{C_{3,\,3}}} \approx -1.07$$

То есть попадаем в доверительную область, значит можно перейти к более короткой

3 часть задания

$$SS_{\text{общ}} = \sum_{i=1}^{n} \left(y_i - \overline{Y}\right)^2 \approx 3083.98$$

Проверим значимость самой модели (в задании такого не было):

$$T = \frac{SS_{\text{регр.}}}{p} \cdot \frac{n-p-1}{SS_{\text{случ.}}} = \approx 12000 > f_{0.95,\,1,\,7}$$

Попали в критическую область, значит модель значимая.

$$R^2_{y\,x} = 1 - \frac{SS_{\text{случ.}}}{SS_{\text{общ.}}} = 0.998$$

$$R^2_{\text{adj}} \approx 0.9976$$

Задача 3

Бюджетное обследование пяти случайным образом выбранных семей дало следующие результаты (в тыс. у. е.)

Семья	Накопления Y	Доход X	Имущество Z
1	3.0	40	60
2	6.0	55	36
3	5.0	45	36
4	3.5	30	15
5	1.5	30	90

Предполагается, что справедлива модель вида:

$$Y_i = a + b_1 X_i + b_2 Z_i + \varepsilon_i$$

а) Спрогнозировать накопления семьи, имеющей доход $X = 40$ тыс. у. е. и имущество $Z = 25$ тыс. у. е.

б) Предположим, что доход семьи вырос на 10 тыс. у. е. в то время как стоимость имущества не изменилась. Оцените, как вырастут накопления семьи.

в) Оцените, как вырастут накопления семьи, если её доход увеличился на 5 тысяч, а имущество на 15 тысяч.

Решение

В этом случае матирица регрессоров X будет состоять из столбцов 1, X (дохода) и Z (имущества):

$$X = \begin{pmatrix} 1 & 40 & 60 \\ 1 & 55 & 36 \\ 1 & 45 & 36 \\ 1 & 30 & 15 \\ 1 & 30 & 90 \end{pmatrix}$$

Теперь можно оценить вектор $\theta = \begin{pmatrix} a \\ b_1 \\ b_2 \end{pmatrix}$:

$$\hat{\theta} = \left(X^T X\right)^{-1} X^T Y = \begin{pmatrix} \hat{a} \\ \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} \approx \begin{pmatrix} 0.2787 \\ 0.1229 \\ -0.0294 \end{pmatrix}$$

С этими знаниями можно оценивать накопления.

Пункт а

В этом случае $X = 40$, $Z = 25$.

$$\hat{y} = \hat{a} + \hat{b}_1 X + \hat{b}_2 Z = 0.2787 + 0.1229 \cdot 40 - 0.0294 \cdot 25 = 4.87$$

Пункт б

У какой-то семьи доход был X_1, а стал $X_1 + \Delta X$, где $\Delta X = 10$. Стоимость имущества была Z_1, стала $Z_1 + \Delta Z$, где $\Delta Z = 0$. Посчитаем разницу в накоплениях:

$$\text{dif} = \hat{a} + \hat{b}_1 (X_1 + \Delta X) + \hat{b}_2 (Z_1 + \Delta Z) - \left(\hat{a} + \hat{b}_1 X_1 + \hat{b}_2 Z_1\right) = \hat{b}_1 \cdot \Delta X = 1.229$$

Пункт в

Аналогично предыдущему пункту, но $\Delta X = 5$, $\Delta Z = 15$.

$$\text{dif} = \hat{a} + \hat{b}_1(X_1 + \Delta X) + \hat{b}_2(Z_1 + \Delta Z) - \left(\hat{a} + \hat{b}_1 X_1 + \hat{b}_2 Z_1\right) = \hat{b}_1 \cdot \Delta X + \hat{b}_2 \cdot \Delta Z = 0.173$$

Семинар 23 мая

На семинаре дополняли задачи из дз выше.

ДЗ к семинару 30 мая

Задача 1

В задаче 3 из ДЗ к семинару 23 мая необходимо проверить:

1. $H_0 : b_1 = b_2 = 0$ против $H_1 : b_1,\ b_2 \neq 0$
2. $H_0 : b_1 = 0$ против $H_1 : b_1 \neq 0$
3. $H_0 : b_2 = 0$ против $H_1 : b_2 \neq 0$
4. $H_0 : b_1 = 0.8$ против $H_1 : b_1 \neq 0.8$ (такое значение было вычислено согласно данным по другой стране).

Решение

Восстановим необходимые данные из задачи 3.
Составим матрицу X:

$$X = \begin{pmatrix} 1 & 40 & 60 \\ 1 & 55 & 36 \\ 1 & 45 & 36 \\ 1 & 30 & 15 \\ 1 & 30 & 90 \end{pmatrix}$$

И матрицу Y:

$$Y = \begin{pmatrix} 3 \\ 6 \\ 5 \\ 3.5 \\ 1.5 \end{pmatrix}$$

Найдём оценку вектора θ:

$$\hat{\theta} = \left(X^T X\right)^{-1} X^T Y = \begin{pmatrix} \hat{a} \\ \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} \approx \begin{pmatrix} 0.2787 \\ 0.1229 \\ -0.0294 \end{pmatrix}$$

Первый пункт

Для разбора гипотезы о тривиальности модели необходимо посчитать $SS_{\text{случ.}}$ и $SS_{\text{регр.}}$:

$$SS_{\text{регр.}} = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{Y}\right)^2 \approx 12.019$$

$$SS_{\text{случ.}} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \approx 0.281$$

Теперь необходимо посчитать дисперсию σ_ε:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - p - 1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \approx 0.141$$

Теперь можно проверять гипотезы, используя следующую статистику:

$$\frac{n-p-1}{p} \cdot \frac{SS_{\text{регр.}}}{SS_{\text{случ.}}} \approx 42.753 \sim F(p,\ n-p-1)$$

Доверительной областью для этой статистики будет $(0,\ F_{0.95,\,2,\,2})$, что приблизительно равно $(0,\ 19)$. Статистика попадает в критическую область, значит гипотеза о тривиальности модели отвергается.

Второй пункт

Для этого необходимо посчитать ковариационную матрицу:

$$C = \left(X^T X\right)^{-1} = \begin{pmatrix} 5.6916 & -0.1074 & -0.0252 \\ -0.1074 & 0.0024 & 0.0002 \\ -0.0252 & 0.0002 & 0.0003 \end{pmatrix}$$

Проверяем гипотезу $H_0 : b_1 = 0$ против $H_1 : b_1 \neq 0$.

$$\frac{\hat{b}_1}{\hat{\sigma}_\varepsilon \sqrt{C_{2,\,2}}} \approx 17.848\Big|_{H_0} \sim t(n-p-1)$$

Доверительной областью для этой статистики будет $(t_{0.025,\,2},\ t_{0.975,\,2})$, что приблизительно равно $(-4.303,\ 4.303)$. Статистика попадает в критическую область, значит гипотеза опровергается.

Третий пункт

Проверяем гипотезу $H_0 : b_2 = 0$ против $H_1 : b_2 \neq 0$.

$$\frac{\hat{b}_2}{\hat{\sigma}_\varepsilon \sqrt{C_{3,\,3}}} \approx -11.537\Big|_{H_0} \sim t(n-p-1)$$

Доверительной областью для этой статистики будет $(t_{0.025,\,2},\ t_{0.975,\,2})$, что приблизительно равно $(-4.303,\ 4.303)$. Статистика попадает в критическую область, значит гипотеза опровергается.

Четвёртый пункт

Проверяем гипотезу $H_0 : b_1 = 0.8$ против $H_1 : b_1 \neq 0.8$.
Для проверки этой гипотезы необходимо вспомнить способ получения статистики, используемой в предыдущих пунктах. Изначально для оценки θ_k исползуется следующая формула:

$$\frac{\hat{\theta}_k - \theta_k}{\sigma_\varepsilon \sqrt{C_{k+1,\,k+1}}} \sim N(0,\ 1)$$

В нашем случае имеются подозрения $\theta_k = 0.8$. Поэтому в формулу подставляем $\theta_k = 0.8$. Дисперсию σ_ε мы не знаем, поэтому подставляем $\hat{\sigma}_\varepsilon$, что меняет тип распределения на Стьюдента:

$$\frac{\hat{b}_1 - 0.8}{\hat{\sigma}_\varepsilon \sqrt{C_{2,\,2}}} \approx -98.351\Big|_{H_0} \sim t(n-p-1)$$

Доверительной областью для этой статистики будет $(t_{0.025,\,2},\ t_{0.975,\,2})$, что приблизительно равно $(-4.303,\ 4.303)$. Статистика попадает в критическую область, значит гипотеза опровергается. Потенциально это может означать, что в разных странах значения коэффициентов могут сильно различаться.

Семинар 30 мая

Задача 1

Имеются ежеквартальные данные (Доугерти с. 274) Y_t о расходах потребителей на газ и электричество в США (в млрд долларов сопоставимых ценах 1972 года) за 1977-1982 годы:

Квартал/Год	1977	1978	1979	1980	1981	1982
1	7.33	7.65	7.96	7.74	8.04	8.26
2	4.7	4.92	5.01	5.1	5.27	5.51
3	5.1	5.15	5.05	5.67	5.51	5.41
4	5.46	5.55	5.59	5.92	6.04	5.83

Предполагается, что функция расходов на газ и электричество зависит от номера квартала. После введения фиктивных бинарных переменных $D_2,\ D_3,\ D_4$, соответствующих второму, третьему и четвёртому кварталам, были построены МНК-оценки параметров следующей регрессии:

$$Y_t = a + bt + \delta_2 D_{t_2} + \delta_3 D_{t_3} + \delta_4 D_{t_4} + \varepsilon_t$$

В таблице приведены оценки параметров и среднее квадратичное отклонение оценок (СКО):

параметр	a	b	δ_2	δ_3	δ_4
оценка	7.5	0.03	-2.78	-2.58	-2.19
СКО	0.09	0.005	0.09	0.1	0.1

1. Определите фиктивные переменные $D_2,\ D_3,\ D_4$
2. Верно ли, что регрессионные зависимости для разных кварталов различны?
3. Запишите регрессионные уравнения для Y_t, соответствующие первому, второму, третьему и четвёртому кварталам.

Решение

Составим оценку модели:

$$\hat{Y_t} = 7.5 + 0.03t + 2.78D_{t_2} - 2.58D_{t_3} - 2.19D_{t_4}$$

Первый пункт

Фиктивные переменные $D_2,\ D_3,\ D_4$ определяются следующим образом:

$$D_{t_i} = \begin{cases} 1, & \text{если месяц из } i\text{-го квартала } i \in [2,\ 4] \\ 0, & \text{иначе} \end{cases}$$

Второй пункт

Проверяем различность регрессионных зависимостей:

1. Первый со вторым: $H_0 : \theta_2 = 0,\ H_1 : \theta_2 \neq 0$

$$T = \frac{\hat{\theta}_2}{\hat{\sigma}_{\hat{\theta}_2}} = \frac{2.78}{0.09} \sim t_{n-p-1}$$

   Попали в критическую область, значит регрессиии для первого и второго кваратала различны

2. Все остальные проверяются аналогично, все попадают в критическую область (проверяли только 1 со 2, 1 с 3 и 1 с 4)

Задача 2

Y — вес новорождённого, X — количество сигарет, выкуренных будущей матерью. Провели $n = 964$ наблюдения. В скобках под числами написаны средние квадратические отклонения.

$$\hat{Y} = \underset{(14)}{3418} - \underset{(2.1)}{72}\, x, \quad \hat{R}^2 = 0.012,\ SS_{\text{случ.}} = 158.6 \cdot 10^6$$

Далее наблюдения разделили на две подгруппы:

1. У матери первый ребёнок.

$$\hat{Y} = \underset{(18)}{3363} - \underset{(2.8)}{4}\, x, \quad SS^1_{\text{случ.}} = 91.2 \cdot 10^6,\ n = 584$$

2. У матери не первый ребёнок

$$\hat{Y} = \underset{(23)}{3506} - \underset{(3.1)}{12.1} x, \quad SS^2_{\text{случ.}} = 63.5 \cdot 10^6,\ n = 380$$

Необходимо проверить значимость регрессии (в данном случае достаточно проверить, что параметр при x ненулевой, есть ещё общее решение для проверки всех коэффициентов сразу).
$H_0 : \theta_1 = 0,\ H_1 : \theta_1 \neq 0.$
Составим статистику как в прошлой задаче:

$$T = \frac{\hat{\theta}_1}{\hat{\sigma}_{\hat{\theta}_1}} = \frac{-7.2}{2.1} \approx -3.429 \sim t(964 - 2) \sim N(0,\ 1)$$

Доверительной областью является $(-1.96,\ 1.96)$, статистика попала в критическую область, значит регрессия значимая. Теперь используем второй способ проверки значимости:

$$R^2 = \frac{SS_{\text{регр.}}}{SS_{\text{общ.}}}$$
$$SS_{\text{общ.}} = SS_{\text{регр.}} + SS_{\text{случ.}}$$
$$SS_{\text{случ.}} = 158.6 \cdot 10^6, \text{ было в условии}$$

Теперь можно составить статистику:

$$T = \frac{n - p - 1}{p} \frac{SS_{\text{регр.}}}{SS_{\text{случ.}}} = \frac{n - p - 1}{p} \frac{R^2}{(1 - R)^2} \approx 3.887 \sim F(0.95,\ 982,\ 3)$$

Необходимый квантиль равен 3.85, статистика попала в критическую область.
Теперь хотим проверить, стоит ли разбивать регрессию на две по признаку первенства ребёнка.
Для этого составим статистику (пользуемся критерием Чоу):

$$T = \frac{SS_{\text{случ.}} - \left(SS^1_{\text{случ.}} + SS^2_{\text{случ.}}\right)}{\left(SS^1_{\text{случ.}} + SS^2_{\text{случ.}}\right)} \cdot \frac{n + m - 2p - 2}{p + 1} \approx 12.101 \sim F(0.95,\ 2,\ 960)$$

Доверительный интервал выглядит так: $(0,\ 3.065)$. Статистика попала в критическую область, значит выборку действительно стоит разбить на эти две подвыборки, так как регрессоры у них различаются.
Теперь проверяем на влияние курение для разных подгрупп:
$H_0 : \theta^1_1 = 0$

$$T = \frac{\hat{\theta}^1_1}{\hat{\sigma}\hat{\theta}^1_1} = \frac{-4}{2.8} \approx 1.429 \sim t(582) \sim N(0,\ 1)$$

Попали в доверительную область, значит на первого ребёнка сигареты не сильно влияют.
$H_0 : \theta^2_1 = 0$

$$T = \frac{\hat{\theta}^2_1}{\hat{\sigma}\hat{\theta}^2_1} = \frac{-12.1}{3.1} \approx 3.903 \sim t(582) \sim N(0,\ 1)$$

Попали в критическую область, значит для этой группы сигареты влияют.