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\operatorname{cov}(x_1,\ \overline{x}) = \frac{(n+1)\mathcal{D}(x_1)}{n} - $
$- 2\operatorname{cov}(x_1,\ \overline{x})$
$\operatorname{cov}(x_1,\ \overline{x}) = \operatorname{cov}(x_1,\ \frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\operatorname{cov}(x_1,\ \sum_{i=1}^{n} x_i) = \frac{1}{n}\operatorname{cov}(x_1,\ x_1) = \frac{\mathcal{D}(x_1)}{n}$
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

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

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

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

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

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

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

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

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

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

Логарифмируем и дифференцируем по θ полученное произведение:

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

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

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

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

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

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:

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

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:

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

$\sigma_1,\ \sigma_2$ 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:

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

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

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

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

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

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:

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

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

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

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

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

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:

$$p_1 - p_2 = 0;$$

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

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

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

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

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

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:

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

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.

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.