Задача 2.3

Известно, что электричка "Вашингтон-Петушки" аварийно останавливается раз в несколько дней. Аналитики РЖД проанализировали, сколько дней электричка едет без поломок, и составили выборку: x = (3, 22, 13, 6, 18, 5, 6, 10, 7, 15).

РЖД хочет проверить гипотезу, что дисперсия распределения равна 9 против правосторонней альтернативы.

Требуется:

- Ввести предположение, каким распределением описывается данная выборка.
- Записать задачу формально.
- Предложить критерий для оценки дисперсии распределения.
- Проверить гипотезу о значении дисперсии распределения для уровня значимости $\alpha = 0.05$ аналитически.
- Вывести и получить доверительный интервал для значения дисперсии при $\alpha = 0.05$.

Распределение

Если

- обозначить день без поломки успехом для РЖД (или неудачей для ленивого машиниста),
- а день с поломкой - неудачей для РЖД (успехом для лентяя-машиниста),
- поломка в отдельно взятый лень - бернуллиевская случайная величина,

то дни без поломки - число идущих подряд успешных дней до первого неуспешного дня, или привычнее число неуспешных испытаний подряд в серии Бернулли до первого успеха.

То есть выборка - из геометрического распределения, $P(X = n) = (1-p)^n p$

$\mathbb{E}(X) = \frac{1-p}{p} \quad \mathbb{V}(X) = \frac{1-p}{p^2}$

Формальная задача

$X \sim Geom(p)$

$H_0 : \mathbb{V}(X) = 9$

$H_1 : \mathbb{V}(X) \geq 9$

Поскольку дисперсия дает квадратное уравнение, имеющие корни разных знаков (отрицательный свободный член), является убывающей функцией от вероятности, то задача эквивалентна следующей:

$X \sim Geom(p)$

$H_0: p = p_0$

$H_1: p < p_0$

Критерий

Поскольку не выполнена нормальность данных, то нужен соответствующий критерий, например, критерий меток.

Для него статистика имеет вид

$Z\left(X^n\right) = \frac{S(\theta_0)}{\sqrt{I(\theta_0)}} \sim N(0,1)$

$$S(p) = \frac{\partial}{\partial p} \log L(X^n, p) = \frac{\partial}{\partial p} \sum_1^n \log P(X_i|p)$$

$$= \frac{\partial}{\partial p} \sum_1^n (x_i \log(1-p) + \log p) = -\frac{n\overline{X}}{1-p} + \frac{n}{p}$$

$I(p) = -\mathbb{E}\frac{\partial^2}{\partial p^2} \log L(X^n, p) = \mathbb{E}\left[\frac{n\overline{X}}{(1-p)^2} + \frac{n}{p^2}\right] = n\left(\frac{1}{p^2} + \frac{1}{p(1-p)}\right) = \frac{n}{p^2(1-p)}$

$Z(X^n) = \sqrt{n}\left(\sqrt{1-p} - \frac{p\overline{X}}{\sqrt{1-p}}\right) = \frac{\mathbb{E}(X) - \overline{X}}{\sqrt{\frac{\mathbb{V}X}{n}}}$

Уравнение на p_0:

$9p^2 + p - 1 = 0$

$p_0 = \frac{1+\sqrt{37}}{18}$

```
def z(p, x):
  expectation = (1 - p)/p
  variance = (1-p)/(p**2)
  mean = np.mean(x)

  return np.sqrt(len(x))*(expectation - mean)/np.sqrt(variance)

def p0():
  return (1+np.sqrt(37))/18

x = (3, 22, 13, 6, 18, 5, 6, 10, 7, 15)

z_value = z(p0(), x)
print('z: {}'.format(z_value))
print('z_alpha: {}'.format(st.norm.ppf(0.05)))
# Достигаемый уровень значимости
```

```
st.norm.cdf(z_value)
>>> z: -14.3136734169704
>>> z_alpha: -1.6448536269514729
>>> 8.98808242910111e-47
```

Видим, что гипотезу можно смело отвергать.

Доверительный интервал

Для построения доверительного интервала удобно использовать критерий Вальда.

$$\frac{p_{MLE} - p_0}{\sqrt{\mathbb{V} p_{MLE}}} \sim N(0,1)$$

$$p_{MLE} = \frac{n}{\sum_1^n x_i} \approx 0.095$$

$$p_0 \in \left[p_{MLE} - z *_{1-\alpha/2} \sqrt{I^{-1}\left(p_{MLE}\right)}, p_{MLE} + z_{1-\alpha/2} \sqrt{I^{-1}\left(p_{MLE}\right)}\right]$$

```
def I(p, n):
  return n/(p**2 * (1-p))

def interval(x, alpha=0.05):
  p_mle = 1 / np.mean(x)
  z = st.norm.ppf(1-alpha/2)
  I_value = I(p_mle, len(x))
  return p_mle - z*np.sqrt(1/I_value), p_mle + z*np.sqrt(1/I_value)

inter = interval(x)
inter
>>> (0.03909117426819991, 0.15138501620799055)
```

Для дисперсии (в силу ее монотонности от p) интервал есть:

```
(1-inter[1])/inter[1]**2, (1-inter[0])/inter[0]**2
>>> (37.029249706360005, 628.8176877798868)
```