Анализ номинативных данных

Кто чаще обращается в службу знакомств: мужчины или женщины?

Зависит ли количество аварий на производстве от дня недели?

Можно ли утверждать, что водители-женщины чаще становятся участниками ДТП (дорожно-транспортных происшествий)?

Можно ли утверждать, что выигрыши в игре распределены не случайно среди проигрышей?

	Распределение:	
	эмпирическое	теоретическое
«За»	30	25
«Против»	20	25
Сумма:	50	50

Хи-квадрат

$$\chi_{э}^{2} = \sum_{i=1}^{P} \frac{(f_{э} - f_{т})^{2}}{f_{т}}$$

```scala
import org.apache.spark.ml.linalg.{Vector, Vectors}
import org.apache.spark.ml.stat.ChiSquareTest

val data = Seq(
  (0.0, Vectors.dense(0.5, 10.0)),
  (0.0, Vectors.dense(1.5, 20.0)),
  (1.0, Vectors.dense(1.5, 30.0)),
  (0.0, Vectors.dense(3.5, 30.0)),
  (0.0, Vectors.dense(3.5, 40.0)),
  (1.0, Vectors.dense(3.5, 40.0))
)

val df = data.toDF("label", "features")
val chi = ChiSquareTest.test(df, "features", "label").head
println("pValues = " + chi.getAs[Vector](0))
println("degreesOfFreedom = " + chi.getSeq[Int](1).mkString("[", ",", "]"))
println("statistics = " + chi.getAs[Vector](2))
```

Непараметрические методы

U - Манна - Уитни

Обозначим значения переменной для одной выборки X, а для другой выборки — Y и упорядочим значения обеих выборок по возрастанию.

Значения	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19
Выборка	X	X	Y	X	X	X	Y	X	X	Y	X	Y	Y	Y	Y	Y

$$U_x = mn - R_x + \frac{n(n+1)}{2},$$

$$U_y = mn - R_y + \frac{m(m+1)}{2},$$

$$U_x + U_y = mn,$$

1	Значения	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19
2	Выборка	X	X	Y	X	X	X	Y	X	X	Y	X	Y	Y	Y	Y	Y
3	Ранги	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
4	Ранги X	1	2		4	5	6		8	9		11					
5	Ранги Y			3				7			10		12	13	14	15	16

Шаг 1. Значения двух выборок объединяются в один ряд, упорядоченный в порядке возрастания или убывания. Обозначается принадлежность каждого значения к той и другой выборке (строки 1 и 2).

Шаг 2. Значения выборок ранжируются, и выписываются отдельно ранги для одной и другой выборки (строки 3–5).

Шаг 3. Вычисляются суммы рангов по $X(R_x)$ и по $Y(R_y)$: $R_x = 46$; $R_y = 90$.

Шаг 4. Вычисляются U_x и U_y по формуле 12.1:

$$U_x = 8 \cdot 8 - 46 + \frac{8(8+1)}{2} = 54, \; U_y = 8 \cdot 8 - 90 + \frac{8(8+1)}{2} = 10, \quad U_x + U_y = 64 = mn.$$

Шаг 5. Определяется p-уровень значимости: наименьшее из U сравнивается с табличным (приложение 9) для соответствующих объемов выборки $m = 8$ и $n = 8$. Значение $p \leqslant 0{,}05$ (0,01), если вычисленное $U_{\text{эмп.}} \leqslant U_{\text{табл.}}$ В нашем случае наименьшим является $U_y = 10$, которое и принимается за эмпирическое значение критерия. Оно меньше критического для $p = 0{,}05$ ($U = 13$), но больше критического для $p = 0{,}01$ ($U = 7$). Следовательно, $p < 0{,}05$.

T - Вилкоксон

Проверим гипотезу о различии значений показателя, измеренного дважды на одной и той же выборке («Условие 1» и «Условие 2»), на уровне $\alpha = 0,05$:

1	№ объекта:	1	2	3	4	5	6	7	8	9	10	11	12
2	Условие 1:	6	11	12	8	5	10	7	6	3	9	4	5
3	Условие 2:	14	5	8	10	14	7	12	13	11	10	15	16
4	Разность d_i:	–8	6	4	–2	–9	3	–5	–7	–8	–1	–11	–11
5	Ранги $\|d_i\|$:	8,5	6	4	2	10	3	5	7	8,5	1	11,5	11,5
6	Ранги d_i (+):		6	4			3						
7	Ранги d (–):	8,5			2	10		5	7	8,5	1	11,5	11,5

Шаг 1. Подсчитать разности значений для каждого объекта выборки (строка 4).

Шаг 2. Ранжировать абсолютные значения разностей (строка 5).

Шаг 3. Выписать ранги положительных и отрицательных значений разностей (строки 6 и 7).

Шаг 4. Подсчитать суммы рангов отдельно для положительных и отрицательных разностей: $T_1 = 13$; $T_2 = 65$. За эмпирическое значение критерия $T_{эмп.}$ принимается меньшая сумма: $T_{эмп} = 13$.

Шаг 5. Определяется p-уровень значимости: $T_{эмп}$ сравнивается с табличным (приложение 10) для соответствующего объема выборки. Значение $p \leqslant 0,05$ (0,01), если вычисленное $T_{эмп} \leqslant T_{табл}$ В нашем случае эмпирическое значение равно критическому значению для $p = 0,05$. Следовательно, $p = 0,05$.

Экспериментальные планы

AB

ABn

$$1-(1-\alpha)^m,$$

$m = 5, \alpha = 0,05$ она равна $\approx 22{,}6\%$

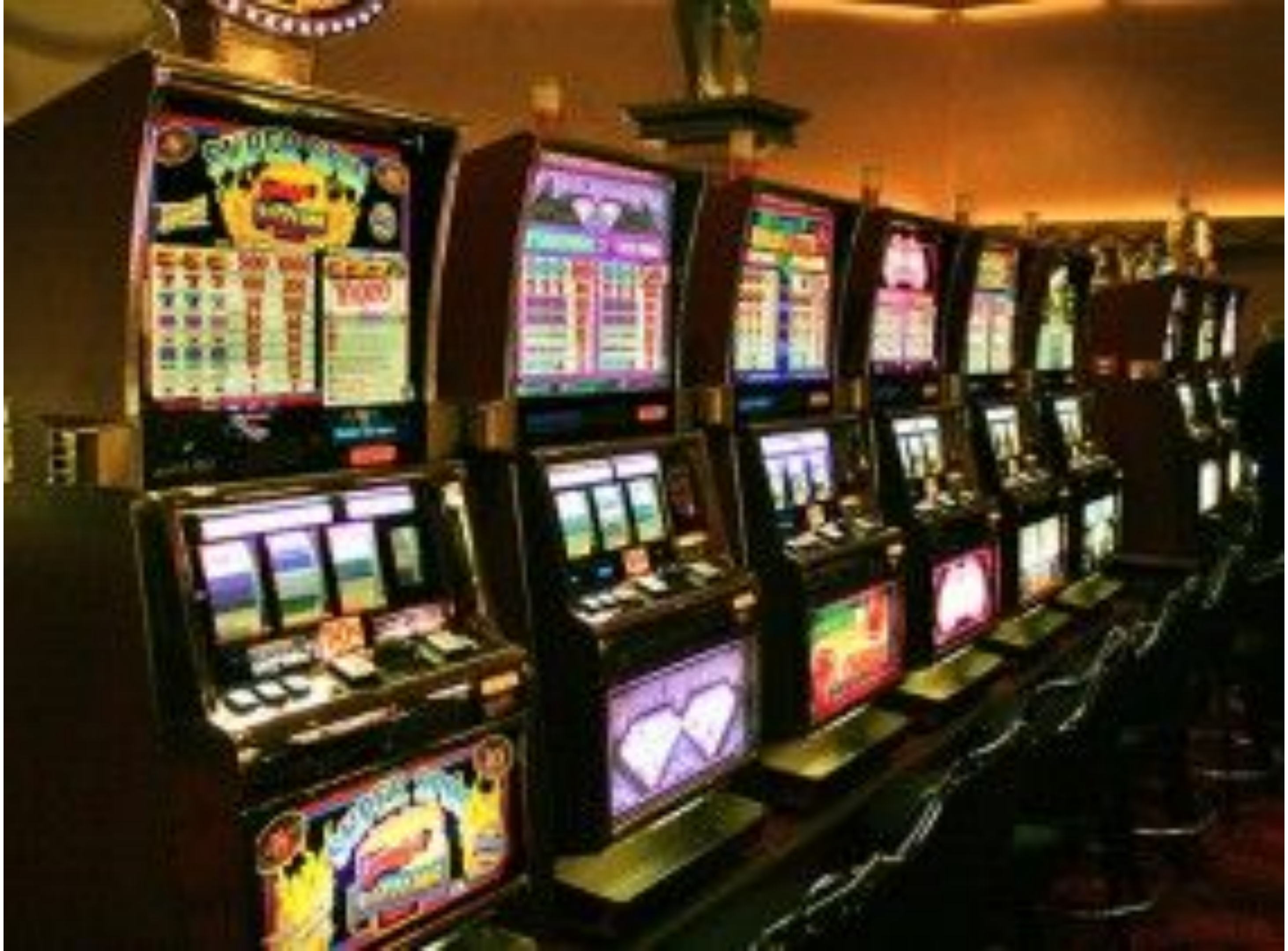
Поправка Бонферрони

$$p_i < \alpha/m$$

AABB

A-A+B-B+

Многорукие бандиты



Биномиальное распределение

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

Бернулли

$$f_a\left(y \mid \theta_a\right) = \theta_a^y (1 - \theta_a)^{1-y},$$

$$
\begin{aligned}
\theta_i &\sim \operatorname{Beta}(\alpha_i, \beta_i) \\
y_i &\sim \operatorname{Bernoulli}(\theta_i)
\end{aligned}
$$

модель легко интерпретируема, α — это количество успешных испытаний, а β — количество неуспешных испытаний; среднее значение будет $\frac{\alpha}{\alpha+\beta}$.