Измерение информации

Измерять информацию, подобно расстоянию, массе, количеству теплоты и т.д., стали относительно недавно. Поясним способ измерения информации на основе игры "Бар-Кохба". Один из участников игры должен отгадать, что задумал другой, задавая ему вопросы, допускающие только два ответа: "да"или "нет". Обычно задумывают какие-либо предметы или строки стихотворений. Количество информации, необходимое для отгадывания, можно измерить количеством вопросов, которые требуется задать при наиболее рациональной тактике дознания.

Запишем полученные ответы, заменив единицами положительные ответы и нулями отрицательные. Получим последовательность нулей и единиц

$$001100011010\ldots,$$

заменяющую последовательность ответов. Процесс замены ответов нулями и единицами называется кодированием, а сама последовательность – кодовым словом. Длина кодового слова может служить мерой количества информации.
Ответ на любой вопрос, допускающий два равно вероятных ответа "да" или "нет", содержит одну единицу информации, которую называют бит.
Иными словами, количество информации, которое можно с одинаковой вероятностью закодировать нулем или единицей, будем считать единицей информации. Заметим, что при измерении количества информации мы сознательно оставляем без внимания содержание информации.

Формула Хартли

Количество информации H, необходимое для выявления объекта, о котором только известно, что он принадлежит множеству из N объектов, вычисляется при помощи формулы Хартли:

$$H = \log_2 N. \tag{1}$$

Например, известно, что у одного из 16 студентов сегодня день рождения. Эту неопределенность можно устранить, задав несколько вопросов тому, кто знает. Сколько вопросов нужно задать, если ответы могут быть только "да" или "нет"? – Ответ: $4 = \log_2 16$.
В случае, когда количество N элементов множества не является степенью двойки, количество информации, вычисляемое по формуле Хартли, не является целым, например, $\log_2 12 = 3.58$.

Формула Шеннона

Если в множестве элементы не равновероятны, то для вычисления информации, необходимой для отгадывания элемента этого множества, используется формула Шеннона. Обозначим через

$$\xi : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}$$

случайный эксперимент, а через $x_1, \dots, x_n$ возможные исходы этого эксперимента, которые осуществляются с вероятностями $p_1, \dots, p_n$, $\sum\limits_{k=1}^{n} p_k = 1$. Количество информации $H(\xi)$, которую нужно получить, чтобы узнать, какой исход эксперимента осуществлен, вычисляется по формуле Шеннона:

$$H(\xi) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \qquad (2)$$

Существенные частные случаи 1

Пусть ξ_0 эксперимент, в результате которого вынимается красная или черная карта. Имеем два равновероятных исхода, поэтому должны получить один бит информации:

$$\xi_0 : \begin{pmatrix} \text{'красная'} & \text{'черная'} \\ 0.5 & 0.5 \end{pmatrix}$$

$$\mathrm{H}(\xi_0) = 0.5 \log_2 2 + 0.5 \log_2 2 = 0.5 + 0.5 = 1.$$

Существенные частные случаи 2

Пусть ξ_2 – некий случайный эксперимент, в результате которого из такой же колоды карт вынимается или нет туз:

$$\xi_2 : \begin{pmatrix} \text{'не туз'} & \text{'туз'} \\ \frac{8}{9} & \frac{1}{9} \end{pmatrix}$$

$$\mathrm{H}(\xi_2) = \frac{8}{9}\log_2\frac{9}{8} + \frac{1}{9}\log_2\frac{9}{1} = 0.352 + 0.151 = 0.503.$$

Необходимая для отгадывания результата случайного эксперимента ξ информация $\mathrm{H}(\xi)$ равна неопределенности этого эксперимента, поэтому часто $\mathrm{H}(\xi)$ называют энтропией распределения случайного эксперимента ξ.

Существенные частные случаи 3

Пусть из перемешанной колоды 36 карт вынимается одна карта. Значениями случайного эксперимента ξ_1 являются все 36 карт, вероятность появления каждой равна 1/36:

$$\xi_1 : \begin{pmatrix} x_1 & \dots & x_{36} \\ \frac{1}{36} & \dots & \frac{1}{36} \end{pmatrix}.$$

В случае равновероятных исходов формула Шеннона совпадает с формулой Хартли

$$H(\xi_1) = \sum_{k=1}^{36} \frac{1}{36} \log_2 36 = \log_2 36 = 5.17.$$

Условная энтропия. ПРИМЕР

Известно, что за языковые способности отвечает левое полушарие. Рассмотрим таблицу сопряженности двух признаков: X — наличие у больного глиобластомы в правом или в левом полушарии головного мозга; Y — наблюдаемые нарушения речи.

	нарушения речи					
Полушарие	нет	есть	всего	(p_1, p_2)	H	q
правое	155	12	167	$(0.93, 0.07)$	0.3728	0.52
левое	94	59	153	$(0.61, 0.39)$	0.9919	0.48
всего	249	71	320	$(0.78, 0.22)$	0.7636	

Средняя условная энтропия

$$0.52 \cdot 0.3728 + 0.48 \cdot 0.9919 = 0.6556$$

Условная энтропия и количество информации

Пусть имеются случайные эксперименты

$$\xi : \begin{pmatrix} A_1 & \dots & A_N \\ p_1 & \dots & p_N \end{pmatrix} \quad \text{и} \quad \eta : \begin{pmatrix} B_1 & \dots & B_n \\ q_1 & \dots & q_n \end{pmatrix}.$$

Каждое событие B_i изменяет вероятности исходов $A_1, \dots, A_N$. Обозначим через $p_j(B_i) = P(A_j|B_i)$ соответствующие условные вероятности и определим условную энтропию

$$H_{B_i}(\xi) = -\sum_{j=1}^{N} p_j(B_i) \log_2 p_j(B_i). \tag{3}$$

Средняя условная энтропия $H_\eta(\xi)$ случайного эксперимента ξ при известном исходе эксперимента η равна

$$H_\eta(\xi) = H(\xi|\eta) = q_1 H_{B_1}(\xi) + \dots q_n H_{B_n}(\xi). \tag{4}$$

$\mathrm{H}(\eta|\xi) = \mathrm{H}(\xi,\eta) - \mathrm{H}(\xi)$

$\xi\|\eta$	y_1	$\cdots$	y_s	
x_1	p_{11}	$\cdots$	$\mathrm{p}_{1\mathrm{s}}$	p_{1*}
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_r	$\mathrm{p}_{\mathrm{r}1}$	$\cdots$	p_{rs}	$\mathrm{p}_{\mathrm{r}*}$
	p_{*1}	$\cdots$	$\mathrm{p}_{*\mathrm{s}}$	

$$\mathrm{H}(\eta|\xi) = \sum_{\mathrm{k}=1}^{\mathrm{r}} \mathrm{p}_{\mathrm{k}\cdot} \left(-\sum_{\mathrm{j}=1}^{\mathrm{s}} \frac{\mathrm{p}_{\mathrm{kj}}}{\mathrm{p}_{\mathrm{k}\cdot}} (\log_2 \mathrm{p}_{\mathrm{kj}} - \log_2 \mathrm{p}_{\mathrm{k}\cdot}) \right) =$$

$$= -\sum_{\mathrm{k}=1}^{\mathrm{r}} \sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{p}_{\mathrm{kj}} \log_2 \mathrm{p}_{\mathrm{kj}} + \sum_{\mathrm{k}=1}^{\mathrm{r}} \sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{p}_{\mathrm{kj}} \log_2 \mathrm{p}_{\mathrm{k}\cdot} =$$

$$= \mathrm{H}(\xi,\eta) + \sum_{\mathrm{k}=1}^{\mathrm{r}} \log_2 \mathrm{p}_{\mathrm{k}\cdot} \sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{p}_{\mathrm{kj}} = \mathrm{H}(\xi,\eta) - \mathrm{H}(\xi).$$

Количество информации

Разность между исходной энтропией $\mathrm{H}(\xi)$ и средней условной энтропией $\mathrm{H}_\eta(\xi) = \mathrm{H}(\xi|\eta)$ называется количеством информации $\mathrm{I}(\xi, \eta)$ о случайном эксперименте ξ, полученном посредством η,

$$\mathrm{I}(\xi, \eta) = \mathrm{H}(\xi) - \mathrm{H}(\xi|\eta) = \mathrm{H}(\eta) - \mathrm{H}(\eta|\xi). \tag{5}$$

Из $\mathrm{H}(\xi|\eta) = \mathrm{H}(\xi, \eta) - \mathrm{H}(\eta)$ следует

$$\mathrm{I}(\xi, \eta) = \mathrm{H}(\xi) + \mathrm{H}(\eta) - \mathrm{H}(\xi, \eta). \tag{6}$$

Совместная информация при независимых величинах

Если ξ и η независимы, то $\mathrm{I}(\xi, \eta) = 0$.

Покажем, что при $\frac{\mathrm{p_{kj}}}{\mathrm{p_{k\cdot}}} = \mathrm{p_{\cdot j}}$ для любого $\mathrm{x_k}$ имеет место $\mathrm{H}(\eta|\xi = \mathrm{x_k}) = \mathrm{H}(\eta)$. Действительно,

$$\mathrm{H}(\eta|\xi = \mathrm{x_k}) = -\sum_{\mathrm{j}=1}^{\mathrm{s}} \frac{\mathrm{p_{kj}}}{\mathrm{p_{k\cdot}}} \log_2 \frac{\mathrm{p_{kj}}}{\mathrm{p_{k\cdot}}} = -\sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{p_{\cdot j}} \log_2 \mathrm{p_{\cdot j}} = \mathrm{H}(\eta)$$

$$\mathrm{H}(\eta|\xi) = \sum_{\mathrm{k}=1}^{\mathrm{r}} \mathrm{p_{k\cdot}} \mathrm{H}(\eta|\xi = \mathrm{x_k}) = \mathrm{H}(\eta), \iff$$

$$\iff \mathrm{I}(\xi, \eta) = \mathrm{H}(\eta) - \mathrm{H}(\eta|\xi) = 0\,.$$

С помощью односторонних коэффициентов неопределенности

$$J_{X|Y} = \frac{I(X,Y)}{H(Y)} \cdot 100\% \quad \text{и} \quad J_{Y|X} = \frac{I(X,Y)}{H(X)} \cdot 100\% \tag{7}$$

измеряется доля информации одного признака, которую он разделяет с другим. Если интерес представляет не только то, как Y влияет на X, но и то, как сам X влияет на Y, то используют двусторонний или симметричный коэффициент неопределенности, равный

$$J = \frac{H(X)}{H(X)+H(Y)} J_{X|Y} + \frac{H(Y)}{H(X)+H(Y)} J_{Y|X} = \\ = \frac{2I(X,Y)}{H(X)+H(Y)} \cdot 100\% \,. \tag{8}$$

ПРИМЕР

Известно, что за языковые способности отвечает левое полушарие. Рассмотрим таблицу сопряженности двух признаков: X — наличие у больного глиобластомы в правом или в левом полушарии головного мозга; Y — наблюдаемые нарушения речи.

X \ Y	нарушения речи			
Полушарие	нет	есть	всего	P
правое	155	12	167	0.07
левое	94	59	153	0.39
всего	249	71	320	

$H(X) = 0.9986$, $H(Y) = 0.7636$, $H(X, Y) = 1.6531$, $I(X, Y) = 0.11$. Коэффициент $J_{X|Y} = 11\%$ указывает на то, какую часть информации о том, какое полушарие повреждено, мы получаем, если узнаем о том, имеются у больного нарушения речи или нет. При отсутствии речевых нарушений у 62 процентов больных поражено правое полушарие, а при нарушении речи у 83 процентов больных глиобластома в левом полушарии.

Примеры проверочных заданий

Один ученик перепрыгивает через козла с вероятностью 0.7, а второй с вероятностью 0.3. В каком случае неопределенность больше?

Ответ: одинакова.

Примеры проверочных заданий

Проведено 30 игр: 20 на своем поле и 10 на чужом, 15 при солнце и 15 в дождь. Выиграно 20 из 30 игр. Из 20 выигранных 16 на своем поле, 4 на чужом, 10 при солнечной погоде и 10 в дождь. Из 10 проигранных 6 на чужом и 4 на своем, и 5 при солнце и 5 в дождь. Что больше влияет на результат: погода или поле?

ξ случайный результат: 0 проигыш, 1 выигрыш с p $= \frac{2}{3}$;
η_1 поле: 0 чужое, 1 свое; η_2 погода: 0 дождь, 1 солнце.

$\xi\|\eta_1$	0	1	
0	6	4	10
1	4	16	20
	10	20	30

$\xi\|\eta_2$	0	1	
0	5	5	10
1	10	10	20
	15	15	30

Свое-чужое поле

ξ случайный результат: 0 проигыш, 1 выигрыш с p $= \frac{2}{3}$;
η_1 поле: 0 чужое, 1 свое;

$\xi \mid \eta_1$	0	1	
0	6	4	10
1	4	16	20
	10	20	30

$$\mathrm{H}(\xi) = \frac{1}{3}\log_2\frac{3}{1} + \frac{2}{3}\log_2\frac{3}{2} = 0.918,$$

$$\mathrm{H}(\xi|\eta_1 = 0) = \frac{6}{10}\log_2\frac{10}{6} + \frac{4}{10}\log_2\frac{10}{4} = 0.971,$$

$$\mathrm{H}(\xi|\eta_1 = 1) = \frac{4}{20}\log_2\frac{20}{4} + \frac{16}{20}\log_2\frac{20}{16} = 0.722,$$

$$\mathrm{H}(\xi|\eta) = 0.971 \cdot \frac{10}{30} + 0.722 \cdot \frac{20}{30} = 0.805$$

Погода

ξ случайный результат: 0 проигыш, 1 выигрыш с p $= \frac{2}{3}$;
η_2 погода: 0 дождь, 1 солнце.

$\xi \mid \eta_2$	0	1	
0	5	5	10
1	10	10	20
	15	15	30

$$\mathrm{H}(\xi) = \frac{1}{3}\log_2\frac{3}{1} + \frac{2}{3}\log_2\frac{3}{2} = 0.918,$$

$$\mathrm{H}(\xi|\eta_2 = 0) = \frac{5}{15}\log_2\frac{15}{5} + \frac{10}{15}\log_2\frac{15}{10} = 0.918,$$

$$\mathrm{H}(\xi|\eta_2 = 1) = \frac{5}{15}\log_2\frac{15}{5} + \frac{10}{15}\log_2\frac{15}{10} = 0.918,$$

$$\mathrm{H}(\xi|\eta) = 0.918 \cdot \frac{10}{30} + 0.918 \cdot \frac{20}{30} = 0.918$$

Пример задачи

Обозначим через $p_{ij} = P\{\xi = x_i, \eta = y_j\}$ вероятности совместного распределения. В каком случае коэффициенты неопределенности равны нулю?

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.15 & 0.15 \\ 0.35 & 0.35 \end{bmatrix} \qquad \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.15 & 0.35 \\ 0.35 & 0.15 \end{bmatrix}$$

Распределения, связанные с нормальным

$\xi_1, \xi_2, \ldots, \xi_n$ последовательность независимых стандартно распределенных случайных величин

Распределения	хи-квадрат	Стьюдента	Фишера
Степени свободы	n	n	n и m
	$\eta = \sum\limits_{i=1}^{n} \xi_i^2$	$\zeta = \dfrac{\xi_0}{\sqrt{\frac{1}{n}\sum\limits_{k=1}^{n} \xi_k^2}}$	$\vartheta = \dfrac{\frac{1}{n}\sum\limits_{k=1}^{n} \xi_k^2}{\frac{1}{m}\sum\limits_{k=1}^{m} \eta_k^2}$

Квантили распределения хи-квадрат

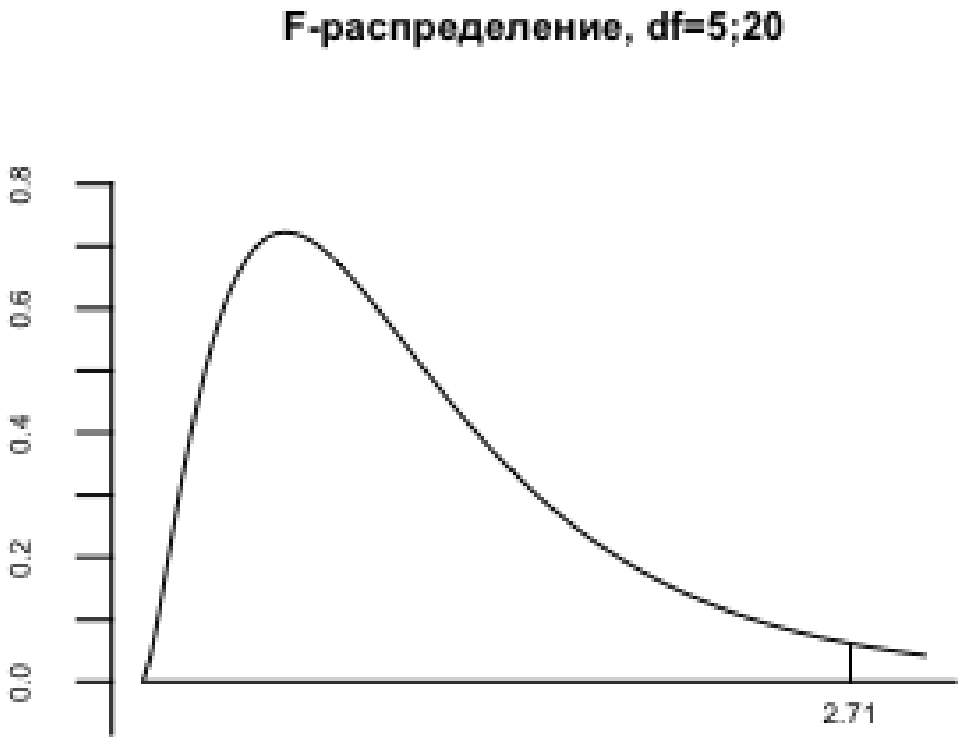
P-квантиль распределения сл.вел. ξ – это число x_p такое, что $P\{\xi \leq x_p\} = P$. Например, 0.95-квантиль распределения $\chi^2(3)$ со степенью свободы 3 равна 7.81, то есть $P\{\chi^2(3) < 7.81\} = 0.95$, $P\{\chi^2(3) > 7.81\} = 0.05$. В R можно использовать функцию qchisq(0.95, 3). Наоборот pchisq(7.81, 3) приведет к значению 0.95.

Квантиль распределения Стьюдента

Например, 0.95-квантиль распределения Стьюдента со степенью свободы 20 равна 1.72, то есть $\mathrm{P}\{\mathrm{t} < 1.72\} = 0.95$. В R можно использовать функцию $\mathrm{qt}(0.95, 20)$. Наоборот $\mathrm{pt}(1.72, 20)$ приведет к значению 0.95.

P-квантили распределения Фишера

P-квантили распределения Фишера с n_1, n_2 степенями свободы ищутся при помощи функции $\mathrm{qf}(P, n_1, n_2)$. Если по некоторой x квантили нужно найти соответствующую вероятность, то используется функция $\mathrm{pf}(x, n_1, n_2)$, например, функция $\mathrm{pf}(2.71, 5, 20)$ приводит к 0.95.



Проверка статистичесикх гипотез

- Для проверки статистической гипотезы H_0 относительно параметров θ или других свойств генеральной совокупности с функцией распределения $F(x|\theta)$ случайной величины ξ используются выборочные наблюдения $x_1, \ldots, x_n$. Альтернативную гипотезу будем обозначать через H_1.
- Гипотеза называется простой, если она однозначно определяет функцию распределения, например, $H_0 : \theta = \theta_0$, иначе сложной, например, $H_0 : \theta > \theta_0$.
- Правило, согласно которому отвергается гипотеза, называется статистическим критерием, а используемая для проверки гипотезы функция от выборочных наблюдений $x_1, \ldots, x_n$ называется статистикой критерия.

Принцип маловероятных событий

- Проверка гипотез основана на принципе маловероятных событий, согласно которому события, вероятность которого мала, считаются невозможными.
- В качестве „малой“ принимают $\alpha = 0.05$ — вероятность выигрыша в игре в кости.
- Гипотезу отвергают тогда, когда наблюдаемое значение статистики попадает в критическую область V, которую выбирает так, чтобы вероятность $\mathrm{P}(\mathrm{V}|\mathrm{H}_0)$ была мала,

$$\mathrm{P}(\mathrm{V}|\mathrm{H}_0) \leq \alpha\,.$$

Построение критерия

- Пусть имеется выборка наблюдений $x_1, \ldots, x_n$ случайной величины, имеющей нормальное распределение со средним μ и дисперсией σ^2.
- Статистика – выборочное среднее $\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$.
- Можно показать, $\bar{x}$ при увеличении n имеет нормальное распределение с тем же средним μ и дисперсией σ^2/n.
- Статистика критерия $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$

ПРИМЕР.

На $n = 25$ автомобилях с усовершенствованным двигателем средний расход бензина составил $\bar{x} = 9.3$ л на 100 км. Считая выборку нормальной $\mathcal{N}(\mu, \sigma)$ с дисперсией $\sigma^2 = 4$, выясним, нельзя ли наблюдаемое улучшение технической характеристики считать случайным.

- Проверим гипотезу $H_0 : \mu = 10$ о том, что расход топлива не изменился, В качестве альтернативной рассмотрим гипотезу $H_1 : \mu < 10$.
- Статистика критерия $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$
- Критическая область $V = (-\infty; -1.645]$, $P(V) = 0.05$.
- Значение статистики $Z = -1.75 \in V$, следовательно, гипотеза H_0 отвергается в пользу альтернативной H_1, уменьшение топлива значимо и не может быть объяснено случайностью.

Доверительный уровень вероятности или значимость, p-значение

По заданному значению x_* статистики X считается доверительный уровень вероятности или p-значение.

p-значение в случае односторонней альтернативы $H_1 : \Theta > \Theta_0$

равно вероятности получить случайное значение больше наблюдаемого, $p = P\{X > x_*\}$.

p-значение в случае односторонней альтернативы $H_1 : \Theta < \Theta_0$

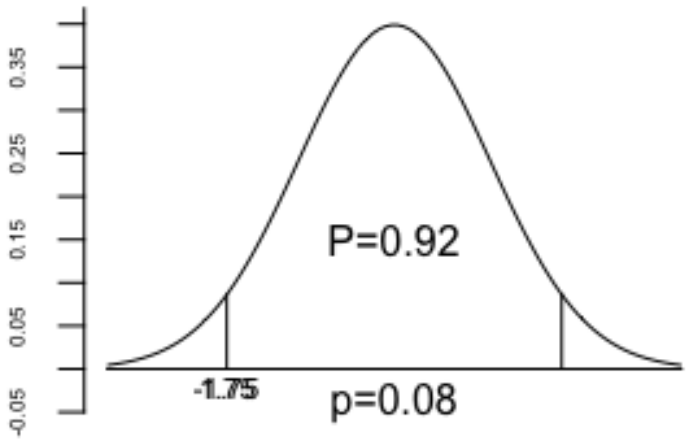
равно вероятности получить случайное значение меньше наблюдаемого, $p = P\{X < x_*\}$.

Нулевая гипотеза отвергается в пользу альтернативной в случае, когда p меньше заданного уровня значимости α, то есть при $p < \alpha$. В качестве уровня значимости обычно берется значение 0.05.

p-значение в случае односторонней альтернативы $H_1 : \Theta \neq \Theta_0$

равно вероятности получить случайное значение больше наблюдаемого по абсолютной величине, $p = P\{|X| > |x_*|\}$.

В примере при альтернативе $H_1 : \mu \neq 10$ значимость равна $p = 0.08$, следовательно, при $\alpha = 0.1$ нулевая гипотеза $H_0 : \mu = 10$ отвергается, а при $\alpha = 0.05$ нет оснований отвергнуть $H_0 : \mu = 10$.



Теорема о согласии распределений

H_0 — выборка $x_1, \ldots, x_n$ сл.вел. ξ с функцией распределения $F(x)$. S — множество значений случайной величины.

$$\cup_{i=1}^{r} S_i = S, \quad S_i \cap S_j = \emptyset, \ i \neq j,$$

$$p_i = P\{\xi \in S_i\} > 0, \quad \sum_{i=1}^{r} p_i = 1.$$

Соответствующие эмпирические частоты равны ν_i, $\sum\limits_{i=1}^{r} \nu_i = n$.

Theorem

При справедливости H_0 статистика

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \tag{9}$$

имеет распределение $\chi^2(r-1)$ хи-квадрат с числом степеней свободы, равным $df = r - 1$. Если s параметров оцениваются по выборке, то $df = r - s - 1$.

Согласие выборочного распределения с нормальным

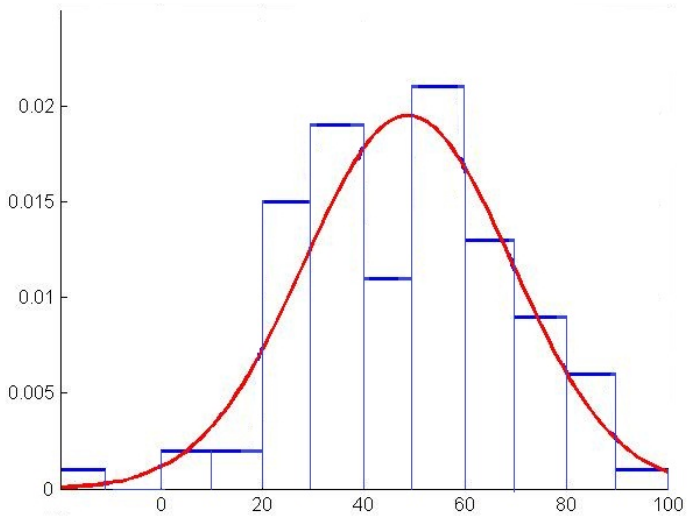


Рис.: Выборочная гистограмма и плотность $\mathcal{N}(\mu, \sigma)$.

Оценки $\hat{\mu} = \bar{\mathrm{x}} = 48.72$, $\hat{\sigma} = \mathrm{S} = 20.47$. Наблюдаемое значение статистики равно $\chi^2_* = 8.356$, p-значение имеет вид $\mathrm{p} = \mathrm{P}\{\chi^2(5) > \chi^2_*\}$, гипотеза о согласии эмпирического распределения с нормальным не отвергается с уровнем значимости $\alpha = 0.05$.

i	$(z_{i-1}; z_i]$	ν_i	$\Phi(\frac{z_i-\hat{\mu}}{\hat{\sigma}})$	p_i	$\frac{(\nu_i-np_i)^2}{np_i}$
1	$(-\infty; 20]$	5	0.080	0.080	1.144
2	$(20; 30]$	15	0.180	0.100	2.510
3	$(30; 40]$	19	0.335	0.155	0.799
4	$(40; 50]$	11	0.525	0.190	3.360
5	$(50; 60]$	21	0.709	0.184	0.359
6	$(60; 70]$	13	0.851	0.142	0.094
7	$(70; 80]$	9	0.937	0.086	0.018
8	$(80; +\infty)$	7	1	0.063	0.072

Критерий хи-квадрат независимости

X и Y – качественные признаки с градациями $x_1, \ldots, x_r$ и $y_1, \ldots, y_s$.

ПРИМЕР 1.1. X – это случайный ответ на вопрос „У меня бывают периоды плача и близости к слезам“, допускающий четыре градации x_1 – „совершенно не согласен“, x_2 – „скорее не согласен“, x_3 – „скорее согласен“, x_4 – „совершенно согласен“.

Признак Y указывает на пол: y_1 – женский, y_2 – мужской.

Из $n = 36$ студентов, участвующих в анкетировании, 14 студентов „совершенно не согласны“, из них 6 девушек и 8 юношей, „скорее не согласны“ пятеро (4 девушки и 1 юноша), 7 человек „скорее согласны“ (1 юноша и 6 девушек) и 10 „совершенно согласны“ (один юноша и 9 девушек).

Таблицы сопряженности

$X \setminus Y$	y_1	y_2	$\dots$	y_s	сумма
x_1	n_{11}	n_{12}	$\dots$	n_{1s}	n_{1*}
x_2	n_{21}	n_{22}	$\dots$	n_{2s}	n_{2*}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	n_{r*}
сумма	n_{*1}	n_{*2}	$\dots$	n_{*s}	n

$X \setminus Y$	y_1	y_2	сумма
x_1	6	8	14
x_2	4	1	5
x_3	6	1	7
x_4	9	1	10
сумма	25	11	36

Представим эти данные в виде таблицы сопряженности, где строки определяются градациями признака X, столбцы – градациями признака Y, а элементы таблицы равны количеству n_{ij} наблюдений с соответствующими градациями: $X = x_i$, $Y = y_j$, $i = 1, 2, \dots, r$, $j = 1, 2, \dots, s$. Через n_{i*} и n_{*j} обозначены соответствующие суммы:

$$n_{i*} = \sum_{j=1}^{s} n_{ij}, \quad n_{*j} = \sum_{i=1}^{r} n_{ij}.$$

Обозначим через

$$p_{ij} = P(X = x_i, Y = y_j),\ p_{i*} = P(X = x_i),\ p_{*j} = P(Y = y_j).$$

Если X и Y независимы, тогда для любых $i = 1, 2, \ldots, r$ и $j = 1, 2, \ldots, s$ справедливо

$$p_{ij} = p_{i*}p_{*j}. \tag{10}$$

При конечном количестве наблюдений оценками вероятностей являются отношения

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}, \quad \widehat{p}_{i*} = \frac{n_{i*}}{n}, \quad \widehat{p}_{*j} = \frac{n_{*j}}{n}, \tag{11}$$

и крайне редко удается получить равенства вида $\frac{n_{ij}}{n} = \frac{n_{i*}}{n} \cdot \frac{n_{*j}}{n}$, например $\widehat{p}_{11} = \frac{6}{36} \neq \frac{14}{36} \cdot \frac{25}{36} = \widehat{p}_{1*}\widehat{p}_{*1}$.
Важно определить, насколько случайным является это нарушение.

Статистика Пирсона χ^2 при независимости X и Y

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\frac{n_{i*}}{n}\frac{n_{*j}}{n})^2}{n\frac{n_{i*}}{n}\frac{n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2 - 2n_{ij}\frac{n_{i*}n_{*j}}{n} + \left(\frac{n_{i*}n_{*j}}{n}\right)^2}{\frac{n_{i*}n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s} \left(\frac{n \cdot n_{ij}^2}{n_{i*}n_{*j}} - 2n_{ij} + \frac{n_{i*}n_{*j}}{n} \right) = n \left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i*}n_{*j}} - 1 \right)$$

и при $n \to \infty$ имеет распределение хи-квадрат с числом степеней свободы, равным

$$rs - 1 - (r-1) - (s-1) = (r-1)(s-1),$$

так как $\sum\limits_{i=1}^{r} p_{i*} = \sum\limits_{j=1}^{s} p_{*j} = 1$, и количество оцениваемых параметров по выборке равно $r + s - 2$.

Пример

X \ Y	y_1	y_2	сумма	P
соверш. не согл.	6	8	14	0.57
скорее не согл.	4	1	5	0.20
скорее согл.	6	1	7	0.14
соверш. согл.	9	1	10	0.10
сумма	25	11	36	0.31

$\chi^2_* = 7.79$ с числом степеней свободы, равным $(4-1)(2-1) = 3$.

В электронных таблицах Excel для получения P-значения можно воспользоваться встроенной функцией $\text{P} = \text{ХИ2РАСП}(7.79; 3) = 0.05$. В R функция pchisq.

Следовательно, независимость признаков отвергается, и то, что есть юноши и девушки по-разному отвечают на поставленный вопрос.