Основные термины и формулы теории вероятностей

СЛУЧАЙНОЕ СОБЫТИЕ -- результат опыта или наблюдения. Случайным мы называем такое событие, которое не можем достоверно предсказать.

ПОЛНОЕ ПОЛЕ СОБЫТИЙ -- перечень всех событий, которые могут произойти в данном опыте. Других событий не может быть.

ВЕРОЯТНОСТЬ ОДНОВРЕМЕННОГО ПОЯВЛЕНИЯ ДВУХ СОБЫТИЙ A и B равна

$$Pr(A \cap B) = Pr(A)Pr(B|A) = Pr(B)Pr(A|B).$$

Здесь $Pr(B|A)$ -- вероятность наступления события B при условии, что событие A уже произошло (условная вероятность).

УСЛОВНАЯ ВЕРОЯТНОСТЬ $Pr(B|A) = Pr(A \cap B)/Pr(A)$.

ФОРМУЛА ПОЛНОЙ ВЕРОЯТНОСТИ

Пусть событие A может произойти одновременно с каким-нибудь из попарно независимых событий $B_1, B_2,...,B_k$, тогда

$$Pr(A)=\Sigma_i\, Pr(B_i)Pr(A|B_i),$$

где $Pr(B_i)$ -- вероятность события B_i, а $Pr(A|B_i)$ -- условная вероятность.

Формулы теории вероятностей

$$Pr(B|A) = Pr(A \cap B)/Pr(A) \qquad Pr(A)=\Sigma_i\, Pr(B_i)Pr(A|B_i)$$

Частота мутаций BRCA1 и BRCA2, вызывающих рак молочной железы, составляет 0.006. Риск развития болезни у носителей любой из этих мутаций 82%. В отсутствии мутации риск болезни составляет 8%.

Задача 1: Найти частоту болезни в популяции.

Решение: Пусть событие A – наличие болезни, а события B_1 и B_2 – наличие или отсутствие мутаций BRCA1 и BRCA2, соответственно.

Генотип	Мут.	Норм.
$Pr(B_i)$	0.006	0.994
$Pr(A\|B_i)$	0.80	0.08
$Pr(A\cap B_i) = Pr(B_i)\, Pr(A\|B_i)$	0.0048	0.0795

Частота болезни в популяции, $Pr(A)$, определяется по формуле полной вероятности: $Pr(A)=\Sigma Pr(B_i)Pr(A|B_i)$ = 0.0048 + 0.0795 = 0.0843

Задача 2: Найти вероятность того, что больной несет мутацию BRCA1 или BRCA2.

Решение: Надо найти $Pr(B_1|A)$

$$\mathbf{Pr(B_1 \mid A)} = \frac{\mathbf{Pr(B_1 \cap A)}}{\mathbf{Pr(A)}} = \frac{\mathbf{Pr(B_1)\,Pr(A \mid B_1)}}{\mathbf{Pr(A)}} = \frac{0.006 \times 0.80}{0.0843} = 0.057$$

Основные термины матстатистики

СЛУЧАЙНАЯ ВЕЛИЧИНА -- такая величина, которая в каждом конкретном случае принимает случайное значение, но во всех испытаниях есть закономерность проявления этой случайной величины.

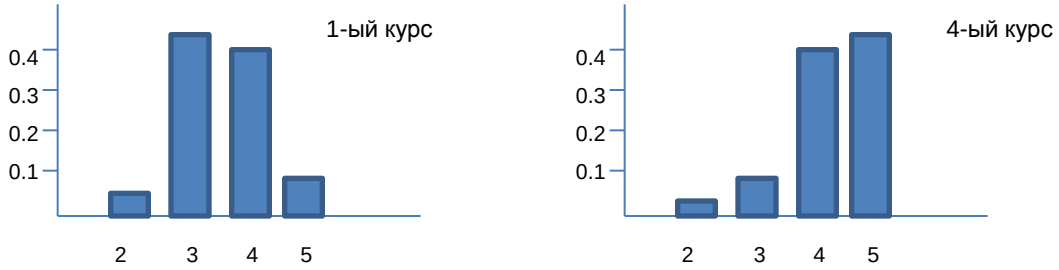
ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ -- все значения случайной величины, существующие в природе.

ВЫБОРКА -- это часть генеральной совокупности, доступная для анализа. Выборка репрезентативна, если закономерности, установленные для нее, справедливы для генеральной совокупности.

Случайные величины бывают дискретные и континуальные.

Основная характеристика случайной величины -- это ее распределение.

Распределение дискретной случайной величины задается набором ее возможных значений и вероятностей, с которыми они могут быть реализованы.



Математическое ожидание дискретной случайной величины X, принимающей значения x_1, x_2,...,x_n с вероятностями $Pr(x_1)$, $Pr(x_2)$,...,$Pr(x_n)$, равно

$$\mu = E(X) = \Sigma_i \, x_i Pr(x_i),$$

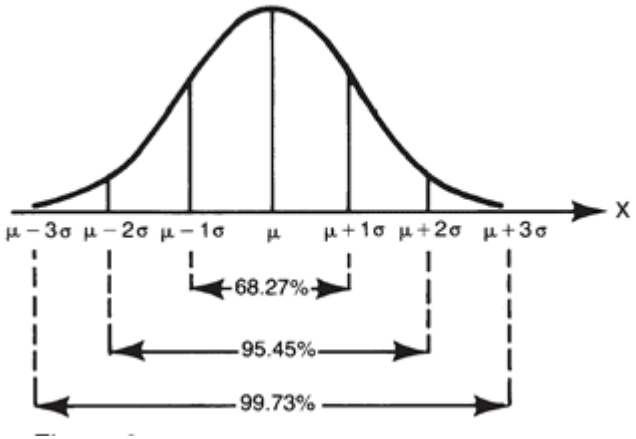
А дисперсия этой случайной величины X определяется выражением

$$\sigma^2 = E((X - \mu)^2) = E(X^2) - \mu^2.$$

Стандартные распределения

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ описывает распределение континуальных случайных величин. Универсальность этого распределения следует из центральной предельной теоремы. Пусть $x_1,x_2,...,x_n$ -- случайные величины с произвольными распределениями. Если эти случайные величины независимы и их число велико, то величина $X = \Sigma x_i$ имеет нормальное распределение, плотность которого описывается формулой

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Свойства:

- распределение симметрично;
- максимальная плотность распределения находится в точке μ;
- ширина в точке перегиба составляет 2σ, и в этот интервал попадает примерно 68 % наблюдений;
- в интервал шириной 4σ попадает около 95 % наблюдений;
- в интервал шириной 6σ попадает приблизительно 99 % наблюдений.

Стандартные распределения

Распределение хи-квадрат (χ^2) обычно используется для тестирования гипотез. Пусть величина y имеет нормальное распределение с параметрами $\mu = 0$, $\sigma = 1$ и есть выборка независимых значений размером n: $y_1, y_2, \ldots, y_n$. Введем новую случайную величину X, равную $X = y_1^2 + y_2^2 + \ldots + y_n^2$.

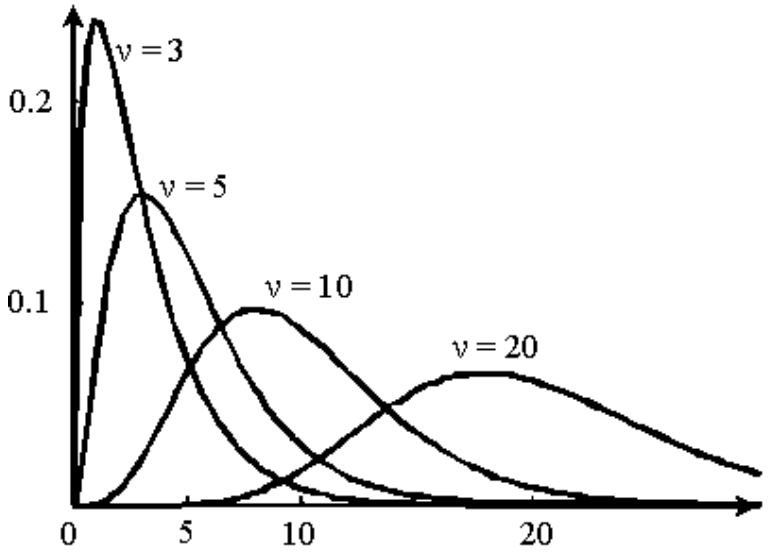
Эта величина имеет распределение χ^2 с числом степеней свободы n.

Плотность этого распределения описывается формулой

$$f(x) = \frac{(1/2)^{\nu/2} x^{\nu/2-1} e^{-x/2}}{\Gamma(\nu/2)},$$

где ν – число степеней свободы (df), $\Gamma(u)$ – гамма функция.

$\mu = \nu$,
$\sigma^2 = 2\nu$



Статистические проблемы конкретного эксперимента

1. Оценка случайной величины
2. Сравнение случайных величин
3. Взаимодействие случайных величин

Статистические проблемы конкретного эксперимента

1. Оценка случайной величины
2. Сравнение случайных величин
3. Взаимодействие случайных величин

1.1. Точечная оценка
1.2. Интервальная оценка
1.3. Планирование объема выборки

Оценка случайной величины

Анализ	Случайная величина	
	Дискретная	Континуальная
Точечная оценка	$\hat{p} = \frac{k}{n}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Интервальная оценка	$\hat{p} - 1.96\sigma_{\hat{p}} \le \hat{p} \le \hat{p} + 1.96\sigma_{\hat{p}}$ $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\bar{x} - 1.96\sigma_{\bar{x}} \le \bar{x} \le \bar{x} + 1.96\sigma_{\bar{x}}$ $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n}}$
Планиро-вание	$\hat{p}(1-\Delta) \le \hat{p} \le \hat{p}(1+\Delta)$ $n = \frac{1.96^2(1-p)}{p\Delta^2}$	$\bar{x}(1-\Delta) \le \bar{x} \le \bar{x}(1+\Delta)$ $n = \frac{1.96^2\sigma^2}{\bar{x}^2\Delta^2}$

Оценка случайной величины

	Случайная величина	
	Дискретная	Континуальная
Исключение	$\hat{p} \le 0.05, \hat{p} \ge 0.95$	$x \ne N(\mu, \sigma)$
Выход	$\hat{q} = 2 \arcsin \sqrt{\hat{p}}$ $q \approx N(\hat{q}, \frac{1}{n})$	Трансформация Ресамплинг

Пример:
k = 7
n = 500

$$\hat{\boldsymbol{p}} = \frac{7}{500} = 0.014$$

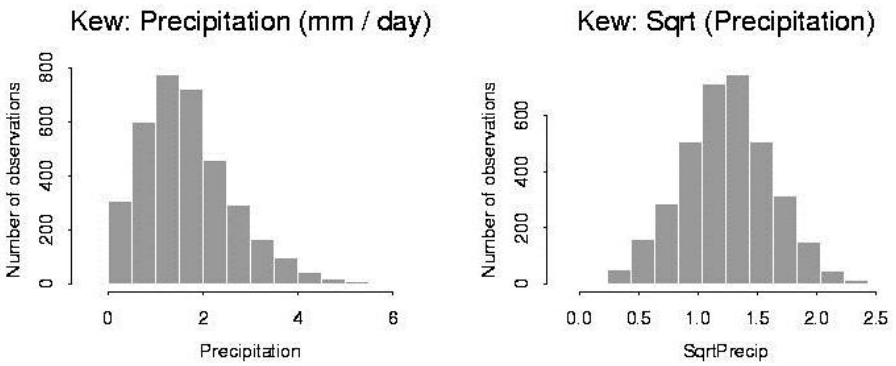
$$\hat{\boldsymbol{q}} = 2\,\boldsymbol{arcsin}\,\sqrt{0.014} = 0.2372$$

$$\boldsymbol{q}_{min} = \hat{\boldsymbol{q}} - 1.96 \times \frac{1}{\sqrt{500}} = 0.1495$$

$$\boldsymbol{q}_{min} = \hat{\boldsymbol{q}} + 1.96 \times \frac{1}{\sqrt{500}} = 0.3248$$

$$0.0056 \le \hat{\boldsymbol{p}} \le 0.025$$

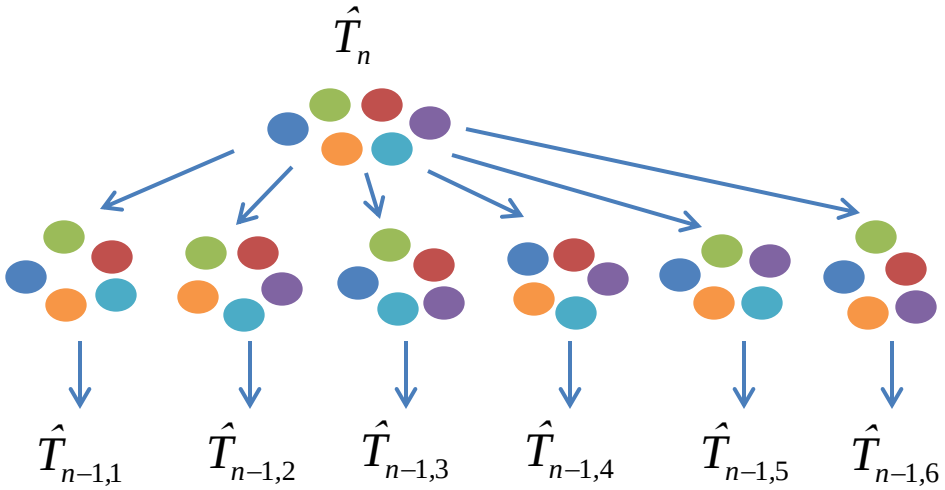
Пример трансформации

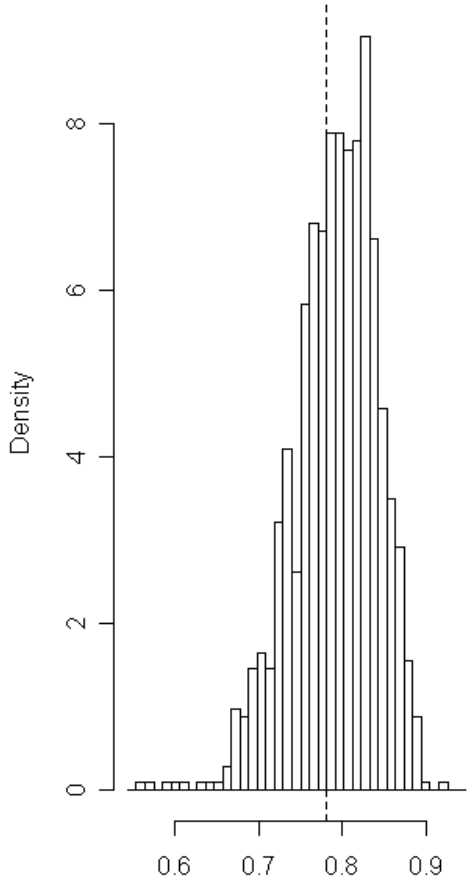


ТЕХНИКА РАЗМНОЖЕНИЯ ВЫБОРОК (RESAMPLING)

Уточнение распределений и оценок параметров

1. Техника складного ножа (Jacknife)

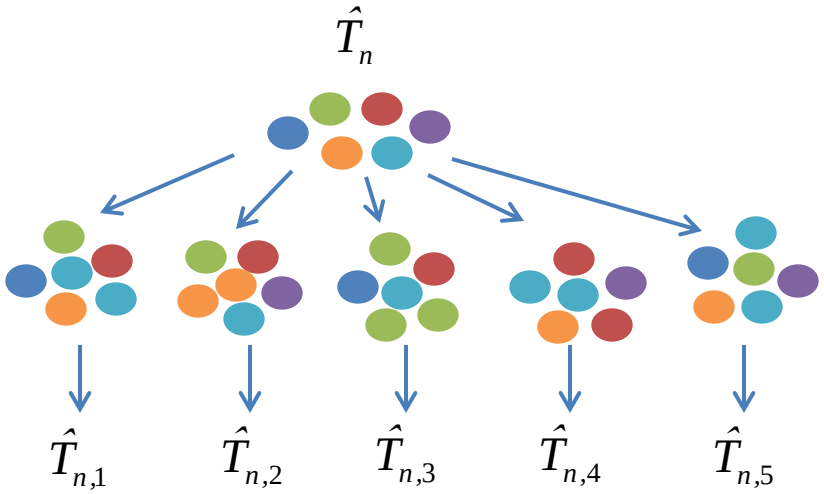


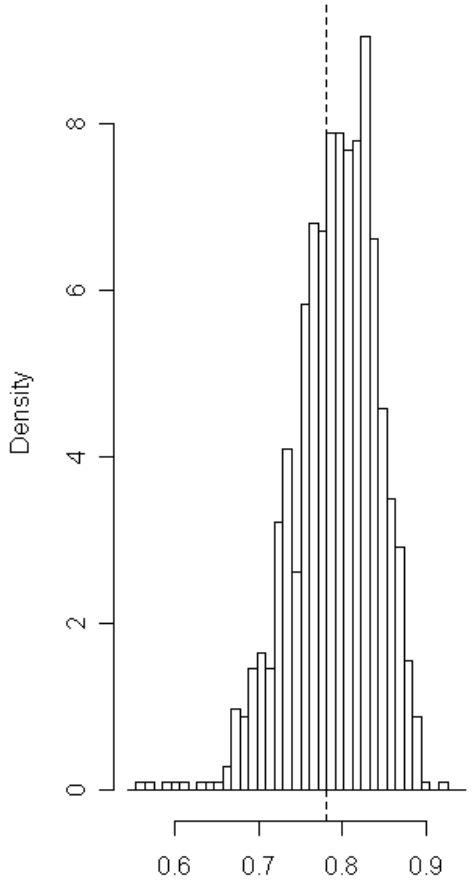


ТЕХНИКА РАЗМНОЖЕНИЯ ВЫБОРОК (RESAMPLING)

Уточнение распределений и оценок параметров

2. Техника бутстреп (Bootstrap)





Статистические проблемы конкретного эксперимента

1. Оценка случайной величины
2. Сравнение случайных величин
3. Взаимодействие случайных величин

} Тестирование гипотез

2-3.1. Формулировка гипотез
2-3.2. Выбор и подсчет статистики
2-3.3. Тестирование гипотезы

Формулировка нулевой и альтернативной гипотез

1. Формулировка биологических гипотез
2. Перевод биологических гипотез в статистические
3. Выбор нулевой гипотезы

Нулевая гипотеза:

- априори считается верной;
- должна быть максимально полно определена, т.е. известны значения параметров модели;
- предпочтительно должна быть выбрана так, чтобы исследователь был заинтересован в ее отвержении.

Альтернативная гипотеза:

- отличается от нулевой значениями параметров
  - ✓ простая
  - ✓ сложная

Выбор и подсчет статистики

Статистика – это величина, определенная как функция от всех нализируемых эмпирических данных. Значение этой функции может быть предсказано со значительно большей точностью, чем результат отдельного измерения.

Примеры:

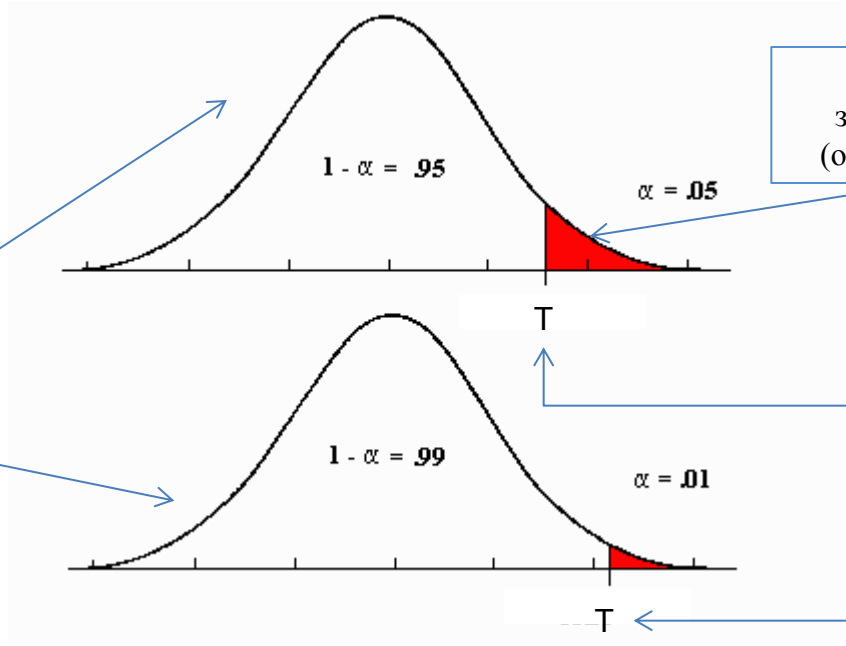
$$\chi_1^2 = \frac{(k - np)^2}{np(1-p)}$$

$$\chi_{k-1}^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$t_{n_1+n_2-2} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{(\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_1}^2)}}$$

$$F_{n_1-1,n_2-1} = \frac{\sigma_{x_1}^2}{\sigma_{x_2}^2}$$

Нулевая гипотеза



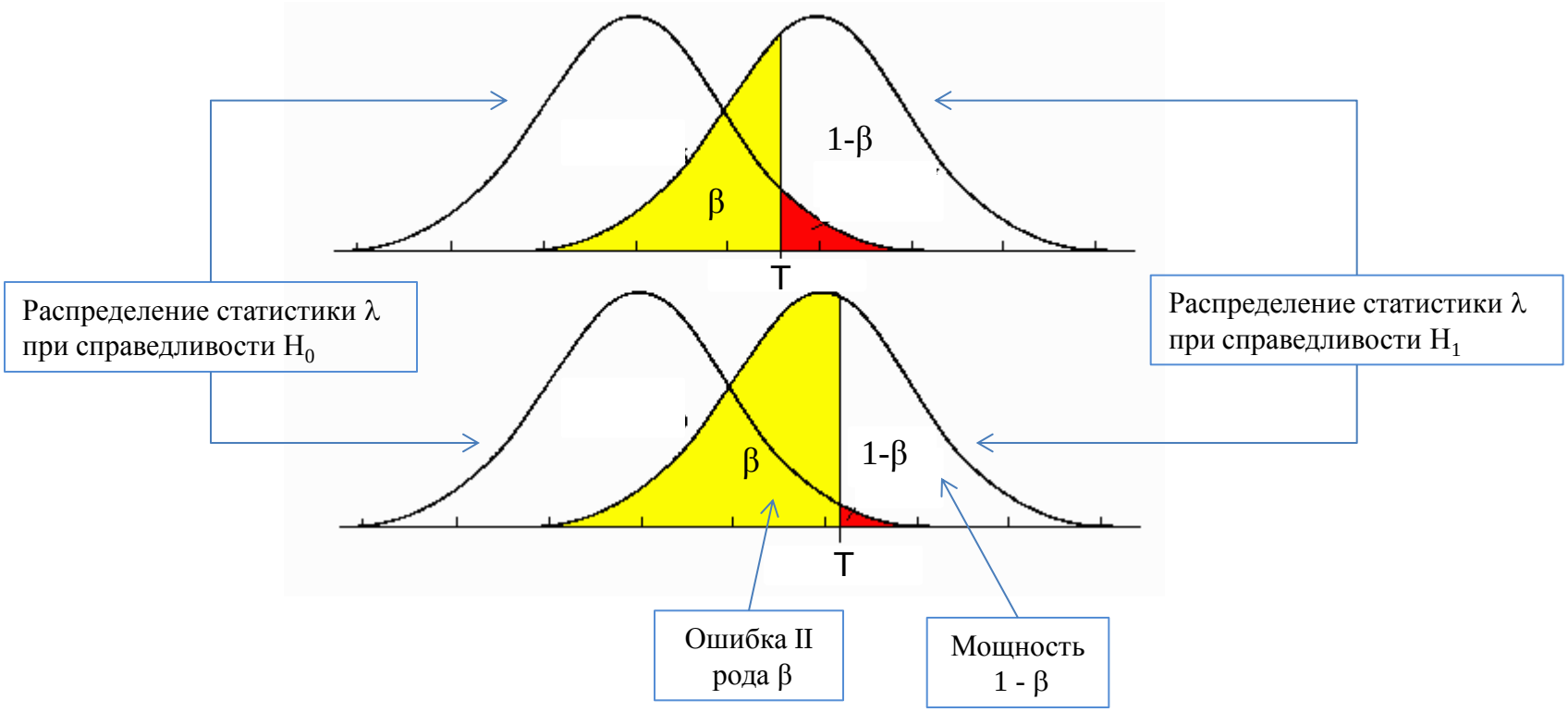
Распределение статистики λ при справедливости H_0

Уровень значимости α (ошибка I рода)

$\alpha = Pr(\lambda > T|H_0)$

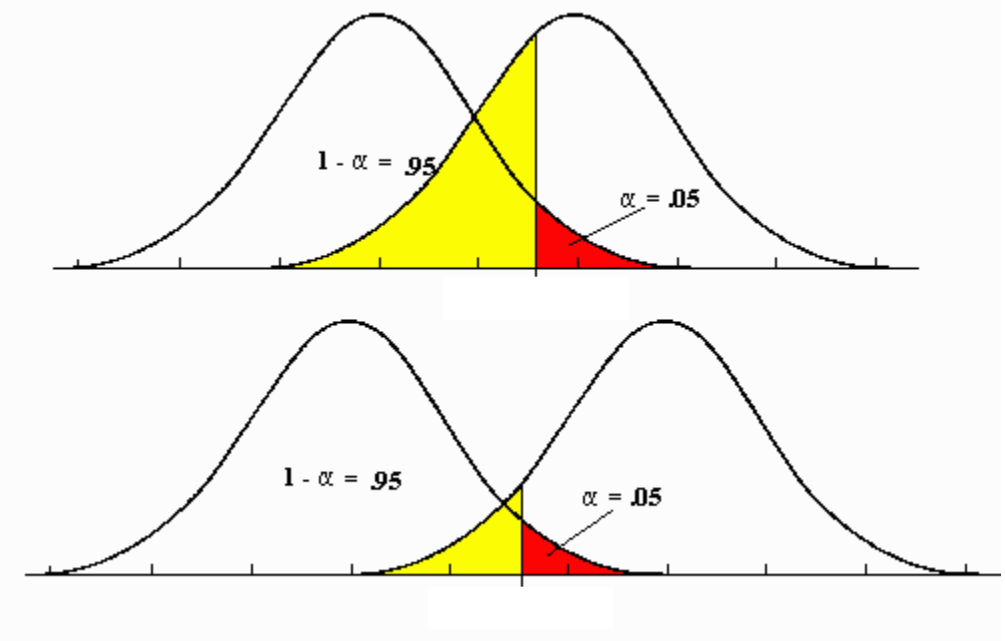
Граничное значение T

Нулевая и альтернативная гипотезы

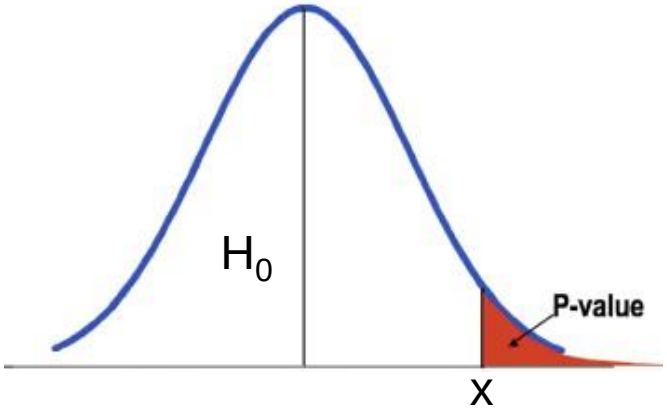


$$\beta = Pr(\lambda<T|H_1)$$
$$1-\beta = Pr(\lambda>T|H_1)$$

Нулевая и альтернативная гипотезы



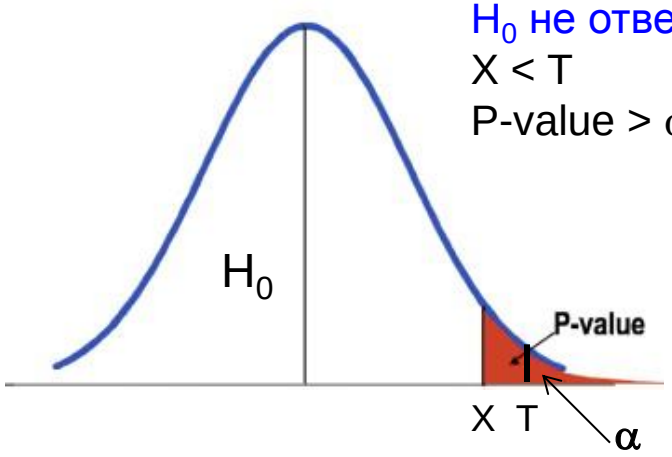
Тестирование гипотез



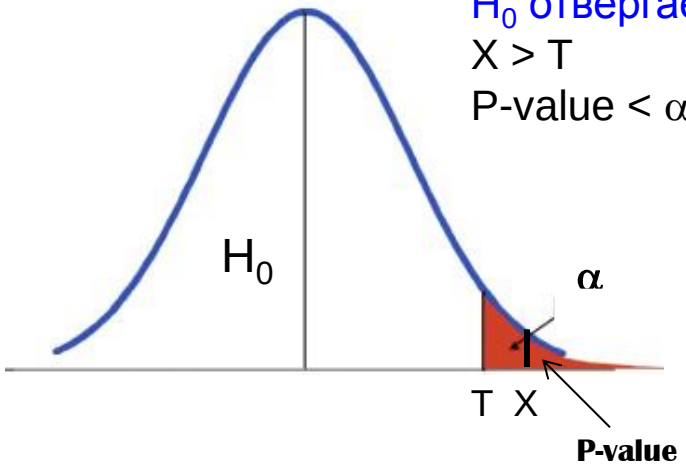
x – значение статистики в конкретном эксперименте

P-value = $Pr(\lambda > x | H_0)$

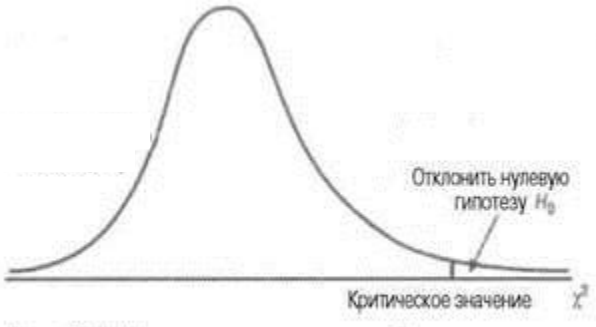
H_0 не отвергается, если:
X < T
P-value > α



H_0 отвергается, если:
X > T
P-value < α



Тестирование гипотез



Критическое значение T = 3.84
Уровень значимости α = 0.05

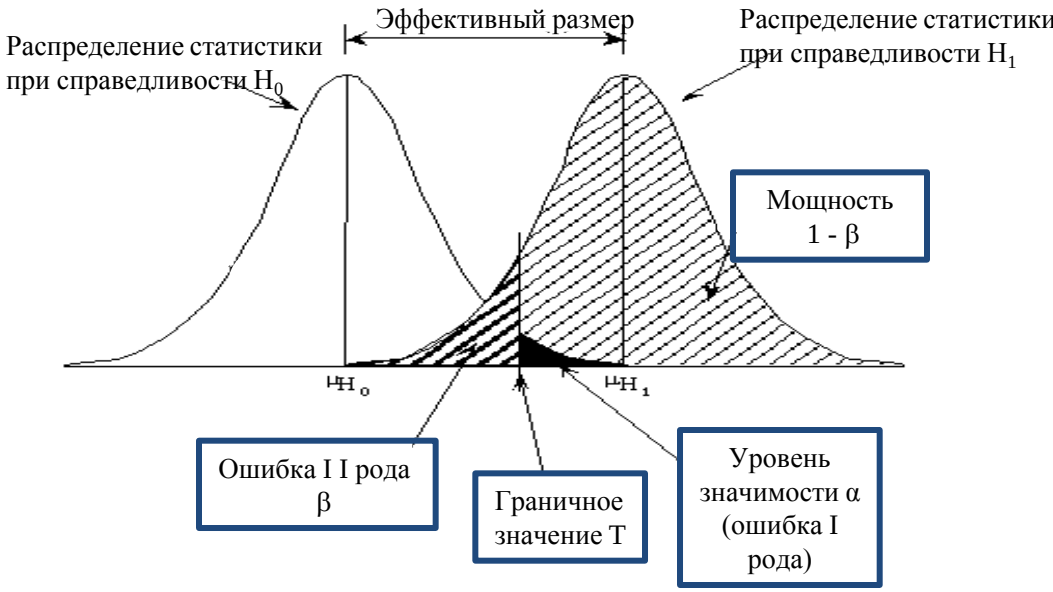
Результат эксперимента	Интерпретация
1 $\chi^2 = 3.55$	
2 p-value = 0.04	
3 $\chi^2 = 6.21$	
4 p-value = 0.11	
5 $\chi^2 = 1.78$	

Определение уровня значимости

λ- случайное значение статистики

$Pr(\lambda>T|H_0) = \alpha$

$Pr(\lambda>T|H_1) = 1-\beta$



Правило выбора α и β

$$Pr(H_0|\lambda > T) << Pr(H_1|\lambda > T)$$

$$Pr(H_0 \mid \lambda > T) = \frac{Pr(H_0)Pr(\lambda > T \mid H_0)}{Pr(H_0)Pr(\lambda > T \mid H_0) + Pr(H_1)Pr(\lambda > T \mid H_1)} = \frac{Pr(H_0)\alpha}{Pr(H_0)\alpha + Pr(H_1)(1-\beta)}$$

$$Pr(H_1 \mid \lambda > T) = \frac{Pr(H_1)(1-\beta)}{Pr(H_0)\alpha + Pr(H_1)(1-\beta)}$$

$$Pr(H_0)\alpha << Pr(H_1)(1-\beta)$$

Определение уровня значимости

$$Pr(H_0)\alpha << Pr(H_1)(1-\beta)$$

$Pr(H_0) = Pr(H_1)$

$\alpha = 0.05$

$1-\beta = 0.8$

$Pr(H_0) \neq Pr(H_1)$

$$Pr(H_1) = \sum_{i=1}^{22}\left(\frac{l_i}{L}\right)^2 \approx 0.05$$

$$Pr(H_0) \approx 0.95$$

Анализ сцепления

Если $\alpha = 0.05$ и $1-\beta = 0.8$, то

$$Pr(H_0)\alpha \approx Pr(H_1)(1-\beta)$$

Чтобы $Pr(H_0)\alpha << Pr(H_1)(1-\beta)$,

при анализе сцепления принимают

$\alpha = 0.001$ и $1-\beta = 0.99$

УРОВЕНЬ ЗНАЧИМОСТИ ПРИ МНОЖЕСТВЕННОМ ТЕСТИРОВАНИИ

Если тесты независимы и в одном испытании уровень значимости равен

$$\alpha = Pr(\lambda > T \mid H_0),$$

то вероятность того, что статистика λ превысит граничное значение критерия хотя бы в одном из n испытаний, равна

$$\alpha^* = Pr(\lambda > T \mid H_0, n) = 1 - (1 - \alpha)^n.$$

Если α мало, вводится поправка Бонферрони:

$$\alpha^* \approx n\alpha.$$

Для зависимых тестов либо специальная аналитическая оценка (при полногеномном картировании признаков человека $\alpha = 5 \times 10^{-8}$), либо эмпирическая оценка с использованием различных техник размножения выборки (ресамплинга).

Планирование объема выборки

$$\chi_1^2 = \frac{(k - np)^2}{np(1-p)}$$

$$t_{n_1+n_2-2} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{(\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_1})}}$$

Задание альтернативной гипотезы:

$$\hat{p} = \frac{k}{n}$$

$$\left|\bar{x}_1 - \bar{x}_2\right| = \Delta_x$$

$$\sigma^2_{x_1} = \sigma^2_{x_1} = \sigma^2$$

Задание граничного значения критерия:

$$\chi_1^2 = 3.84$$

$$t_{n_1+n_2-2} = 1.96$$

Оценка объема выборки:

$$3.84 = \frac{(\hat{n}\hat{p} - \hat{n}p)^2}{\hat{n}p(1-p)}$$

$$\hat{n} = \frac{3.84\,p(1-p)}{(\hat{p} - p)^2}$$

$$1.96 = \frac{\Delta_x}{\sqrt{\left(\frac{\sigma^2}{\hat{n}_1} + \frac{\sigma^2}{\hat{n}_2}\right)}}$$

Оценка значения параметра при фиксированном объеме выборки:

$$\Delta = \left|\hat{p} - p\right|$$

$$\hat{\Delta} = \sqrt{\frac{3.84\,p(1-p)}{n}}$$

$$\hat{\Delta}_x = \frac{1.96}{\sqrt{\left(\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)}}$$

На дом:

1. Сформулировать статистическую проблему каждого конкретного эксперимента.
2. Для экспериментов, направленных на получение оценки случайной величины:
    - а) предсказать распределение этой величины
    - б) задать необходимую точность оценки
    - в) определить объем выборки, необходимый для получения такой оценки
3. Для экспериментов, предполагающих тестирование гипотез:
    - а) сформулировать биологические и статистические гипотезы
    - б) обосновать выбор нулевой гипотезы
    - в) обосновать выбор статистики
    - г) определить объем выборки, необходимый для различения статистических гипотез