

Основы математической статистики

Математическая статистика позволяет обрабатывать результаты опытов, измерений и т. д. Математическая статистика использует методы теории вероятности.

Случайные события

- Событие называется детерминированным, если в результате опыта оно происходит или не происходит наверняка. В детерминированном случае мы точно знаем, что данная причина приведет к единственному, вполне определенному следствию.
- Событие называется случайным, если в результате опыта мы не можем заранее предсказать - произойдет событие или нет. При этом предполагается, что опыт можно повторять неограниченное число раз при неизменных условиях.

•События A и B называются несовместными, если появление одного исключает появление другого.

•Событие B следует из события A , если событие B происходит всегда, когда произошло событие A .

Это обозначается тем же символом, что и подмножество: $A \subset B$.

•Будем говорить о равенстве двух событий A и B , если из A следует B и из B следует A .

•Событие называется невозможным, если оно не может произойти никогда при данных условиях.

•Событие называется достоверным, если оно происходит всегда при данных условиях.

• Пусть случайный эксперимент проводится раз n , и событие A произошло m раз. Тогда говорят, что относительная частота события A есть $v(A)=m/n$.

• Частота события связана с его вероятностью.

Относительную частоту называют еще *эмпирической вероятностью* потому, что по частоте события мы оцениваем возможность его появления в будущем.

• Для любого случайного события A

$$0 \leq P_n(A) \leq 1$$

n - количество случайных экспериментов.

Две теоремы о вероятности суммы событий и произведении

1. Если события несовместны, то вероятность суммы событий равна сумме вероятностей:

$$P(A+B) = P(A) + P(B)$$

2. Если события независимы, то вероятность произведения событий равна произведению вероятностей:

$$P(A B) = P(A) P(B)$$

Классическое определение

- *Вероятностью $P(A)$ события называется отношение числа благоприятных исходов $m(A)$ к общему числу несовместных равновозможных исходов:*

$$P(A) = \frac{m(A)}{N}$$

Свойства вероятности.

- *1. Для любого случайного события A $0 \leq P(A) \leq 1$*
- *2. Пусть события A и B несовместны. Тогда $P(A+B) = P(A) + P(B)$*

Например: бросание кубика. Всего исходов 6, число исходов, благоприятных выпадению четного числа – 3. $P(A) = 1/2$

Дискретная случайная величина

$$\left(\begin{array}{cccc} \xi_1 & \xi_2 & \dots & \xi_n \\ p_1 & p_2 & \dots & p_n \end{array} \right), \quad p_1 + p_2 + \dots + p_n = 1$$

Будем предполагать, что все числа x_k различны.

Случайная величина принимает значение x_k , если произошел исход ω_k , вероятность которого равна p_k

Точнее: вероятность события $\{\xi(\omega_k)=x_k\}$ равна p_k

Дискретная случайная величина полностью

определяется своими значениями и их вероятностями.

Дисперсия

Дисперсией конечной случайной величины ξ называется число

$$D\xi = M(\xi - M\xi)^2$$

по определению математического ожидания, дисперсия вычисляется по следующей формуле

$$D\xi = \sum_i (x_i - M\xi)^2 p_i$$

Дисперсию иногда обозначают как $\sigma^2(\xi)$ или σ_ξ^2

$\sigma_\xi = \sqrt{D\xi}$ называется *среднеквадратичным отклонением* или *стандартным отклонением* случайной величины

Функция распределения

Функция действительной переменной

$$F_{\xi}(x) = P(\xi < x)$$

называется *функцией распределения* случайной величины ξ .

Свойства функции распределения

1. $P(\xi \geq x) = 1 - F_{\xi}(x)$
2. $P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a)$
3. При любом x выполняется неравенство.

$$0 \leq F_{\xi}(x) \leq 1$$

Это справедливо, поскольку функция распределения есть вероятность

4. Функция распределения есть неубывающая функция.

5. При $x \rightarrow -\infty$ событие стремится к невозможному и вероятность соответственно, стремится к нулю. При $x \rightarrow \infty$ событие становится достоверным

6. Функция распределения непрерывна слева, то есть

$$\lim_{x \rightarrow x_0 + 0} F_{\xi}(x) = F_{\xi}(x_0)$$

Случайная величина ξ называется *непрерывной случайной величиной*, если существует функция $f_{\xi}(x)$ такая, что

$$P(\xi \in [a, b]) = \int_a^b f_{\xi}(x) dx$$

Функция $f_{\xi}(x)$ называется *плотностью вероятности* или *плотностью распределения* случайной величины ξ

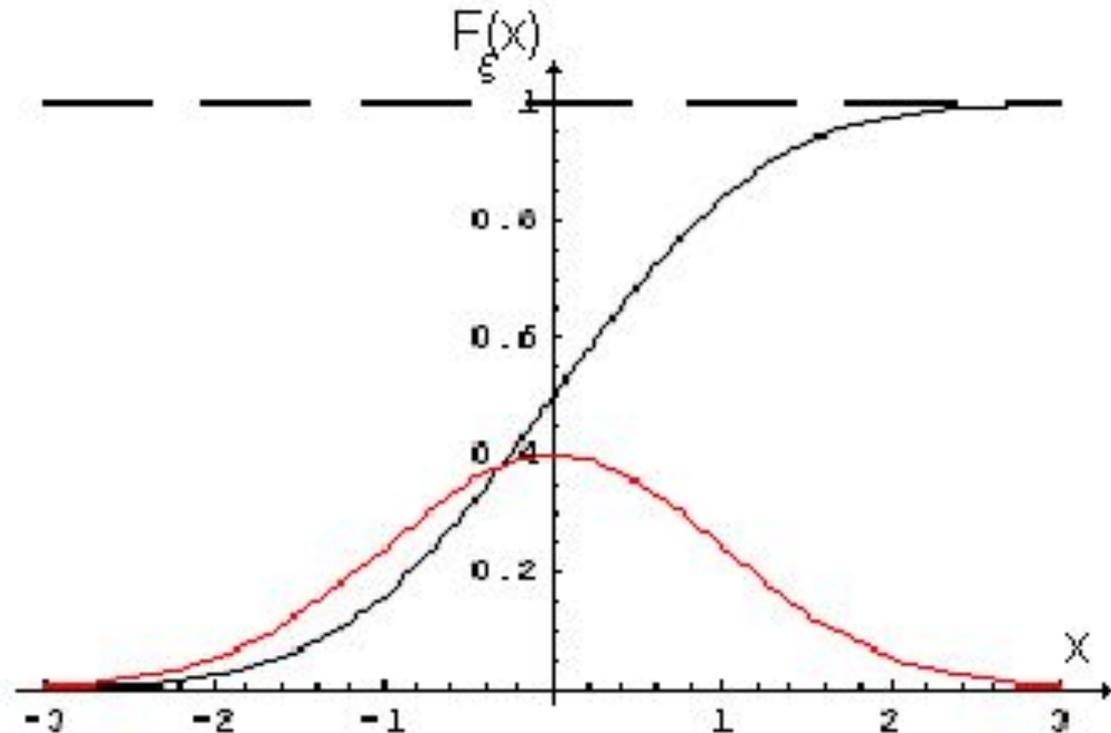
Распределение Гаусса

Говорят, что случайная величина ξ , распределена по *нормальному закону* (имеет *нормальное распределение*) с параметрами m и σ , ($\sigma > 0$) если она имеет плотность распределения

$$f_{\xi}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

$$F_{\xi}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

На рисунке представлены графики стандартного (при $m=0$ и $\sigma=1$) нормального распределения Гаусса (черный) и его плотности (красный)



Статистика

- Генеральной совокупностью называется вся совокупность исследуемых объектов
- Выборочной совокупностью или просто выборкой называют совокупность случайно отобранных из генеральной совокупности объектов
- Объемом совокупности называют число объектов этой совокупности

Способы формирования выборочной совокупности

- Повторный – после измерений объект возвращают в генеральную совокупность
- Бесповторный – после измерений объект в генеральную совокупность не возвращается

Выборка должна быть репрезентативной - представительной. Для этого объекты из генеральной совокупности должны отбираться случайно.

Выборка и ее обработка

- Упорядочивание. Элементы выборки x_1, x_2, \dots, x_n располагаются в порядке возрастания.
- Частотный анализ. Пусть выборка содержит k различных значений z_1, z_2, \dots, z_k , причем z_i встречается n_i ($i=1, 2, \dots, k$) Число n_i называют частотой элемента z_i ,

$$\sum_{i=1}^k n_i = n$$

- Совокупность пар (z_i, n_i) называют статистическим рядом выборки. Часто его представляют в виде таблицы – в первой строке z_i , во второй n_i .
- Величина $v_i = n_i / n$ называется относительной частотой
- Накопленная частота значения z_i равна $n_1 + n_2 + \dots + n_i$.
- Относительная накопленная частота $v_1 + v_2 + \dots + v_i$

Эмпирическая функция распределения

Каждой выборке $\{x_1, x_2, \dots, x_n\}$ можно поставить в соответствие конечную случайную величину, принимающую эти значения с равными вероятностями $1/n$

$$\xi_n = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1/n & 1/n & \dots & 1/n \end{pmatrix}$$

Это распределение называется *выборочным, или эмпирическим, распределением*. Как и для любой конечной случайной величины, для эмпирической случайной величины можно построить ступенчатую функцию распределения; она называется *выборочной функцией распределения*. Кроме того, можно вычислить числовые характеристики выборочной случайной величины ξ_n - математическое ожидание, дисперсию.

выборочное математическое ожидание (его обычно называют *выборочным средним*), *выборочная дисперсия*, *выборочная медиана* и т.д. Например, выборочное среднее (его обозначают через \bar{x}) есть не что иное как среднее арифметическое значений выборки \bar{x}

$$M\xi_n = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Соответственно выборочная дисперсия s^2 равна

$$D\xi_n = s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

ОЦЕНКА ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Пусть в нашем распоряжении имеется выборка $\{x_1, x_2, \dots, x_n\}$ из генеральной совокупности с функцией распределения $F(x)$. Функция распределения $F_n^*(x)$ эмпирической случайной величины

$$\xi_n = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1/n & 1/n & & 1/n \end{pmatrix}$$

Поскольку каждое значение из выборки есть случайная величина с функцией распределения, то вероятность успеха равна $p=F(x)$. Число успехов равно $\mu_n(x)$, а относительная частота успеха равна $\mu_n(x)/n$ и совпадает с выборочной функцией распределения.

Следовательно, выборочная функция распределения представляет собой относительную частоту успеха, а функция распределения генеральной совокупности - вероятность успеха. Из предыдущего нам известно, что относительная частота есть несмещенная состоятельная оценка вероятности. Значит, выборочная функция распределения действительно является несмещенной, состоятельной и эффективной оценкой функции распределения:

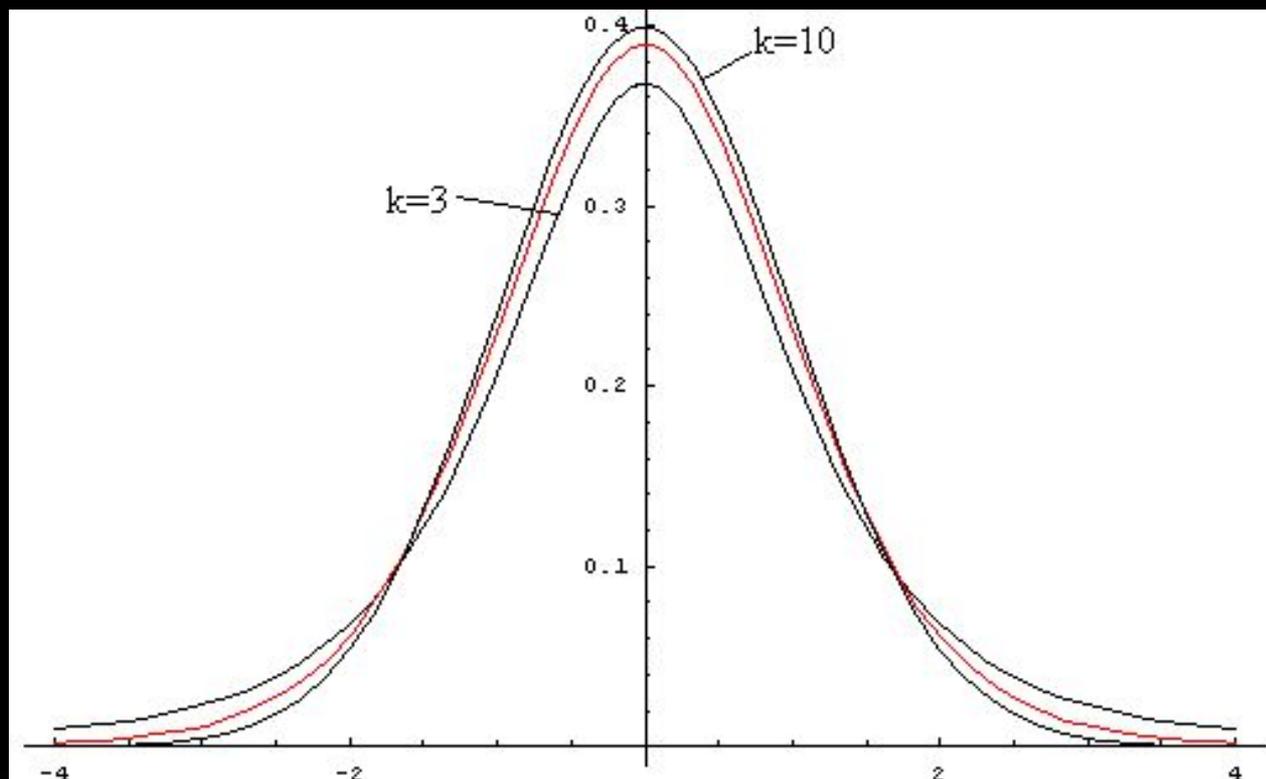
$$MF_n^*(x) = F(x)$$
$$\lim_{n \rightarrow \infty} P\left(\left|F_n^*(x) - F(x)\right| < \varepsilon\right) = 1$$

Выборочные квантили

Выборочный квантиль определяется по выборке.

- Квантиль – левее должно располагаться кол-во значений, соответствующее индексу квантили. Например, для квантили $x_{0.8}$ левее должно располагаться 80% значений выборки.

Распределение Стьюдента



На рисунке красным выделено нормальное распределение, черным – распределение Стьюдента.

Свойства распределения Стьюдента

Распределение Стьюдента симметрично, причем $Mt(k) = 0$.

При больших k распределение Стьюдента близко к стандартному нормальному распределению $N(0, 1)$.

Доверительный интервал математического ожидания.

Случайная величина U распределена по нормальному закону

$$\frac{\bar{x} - m}{\sigma/\sqrt{n}} \sim N(0,1)$$

Случайная величина

$$\frac{\bar{x} - m}{S/\sqrt{n}} \sim t(n-1)$$

распределена по закону Стьюдента, а доверительный интервал математического ожидания примет вид (τ_α - квантиль распределения Стьюдента, $\alpha = (1 + \gamma)/2$)

$$\left(\bar{x} - \tau_\alpha \frac{S}{\sqrt{n}}, \bar{x} + \tau_\alpha \frac{S}{\sqrt{n}} \right)$$

Пример

Вычислим доверительные интервалы для нашей выборки.

Интервал для математического ожидания. Случай 1. Будем считать, что несмещенная оценка дисперсии – точное значение.

Выберем уровень значимости $\gamma = 0.95$. По таблице найдем квантиль стандартного распределения $u_{0.975} = 1.96$. Подставим в формулу

$$u_{(1+\gamma)/2} \frac{\sigma}{\sqrt{n}}$$

$m=0.51735$, $\sigma=0,288955$, $n=49$. После вычислений получим $0,0809074$.

Интервал будет $0.51735 - 0,0809074 < m < 0.51735 + 0,0809074$
 $0,4364426 < m < 0,5982574$.

Пример. Интервал для дисперсии

$$S^2=0,083495$$

$$\frac{S^2(n-1)}{\chi_{(1+\gamma)/2}^2(n-1)} < \sigma^2 < \frac{S^2(n-1)}{\chi_{(1-\gamma)/2}^2(n-1)}$$

Находим квантили распределения $\chi_{(1+\gamma)/2}^2$ и $\chi_{(1-\gamma)/2}^2$.

$$\chi_{0.975}^2 = 71.4 \quad \chi_{0.025}^2 = 42.85$$

Находим интервал $0,056131 < \sigma^2 < 0,09353$

Интервал для математического ожидания. Случай 2.

Используем распределение Стьюдента. Формула та же, что и раньше, но вместо квантиля нормального распределения используется квантиль распределения Стьюдента. $t(48)_{0.975} = 2.0105$

После вычислений получим

$$0.51735 - 0,082992 < m < 0.51735 + 0,082992$$

$$0,434358 < m < 0,600342$$

Статистическая гипотеза

- Любое утверждение о виде или свойствах закона распределения наблюдаемых случайных величин
- Всякий раз предполагаем, что у нас имеются две взаимоисключающие гипотезы:

основная и альтернативная

Нулевой (основной) гипотезой - H_0
называют какое-либо конкретное
предположение о теоретической функции
распределения или предположение,
влекущее за собой важные практические
последствия

Альтернативная гипотеза H_1 - любая
гипотеза, исключая нулевую

Задача проверки статистической гипотезы состоит в том, чтобы, используя статистические данные (выборку)

$$X_1, X_2, \dots, X_n,$$

принять или отклонить нулевую гипотезу

Нулевые и альтернативные гипотезы формулируются как утверждение о принадлежности функций распределения некоторой случайной величины определенному классу распределений

$$\Phi_0, \Phi_1 \in \Phi, \quad \Phi_0 \cup \Phi_1 = \Phi$$

$$\Phi_0 \cap \Phi_1 = \emptyset$$

$$H_0 : F_x \in \Phi_0; \quad H_1 : F_x \in \Phi_1$$

Гипотеза называется **простой**, если соответствующий класс распределений содержит лишь **одно** распределение, в противном случае гипотеза будет **сложной**.

Гипотезы о параметрах распределений называются **параметрическими**

Статистикой критерия

называется функция от выборки

$$T(X) \in \tau$$

значение которой для заданной
выборки служит основанием принятия
или отклонения основной гипотезы

- **Статистический критерий** –
правило, позволяющее только по
результатам наблюдений

- X_1, X_2, \dots, X_n

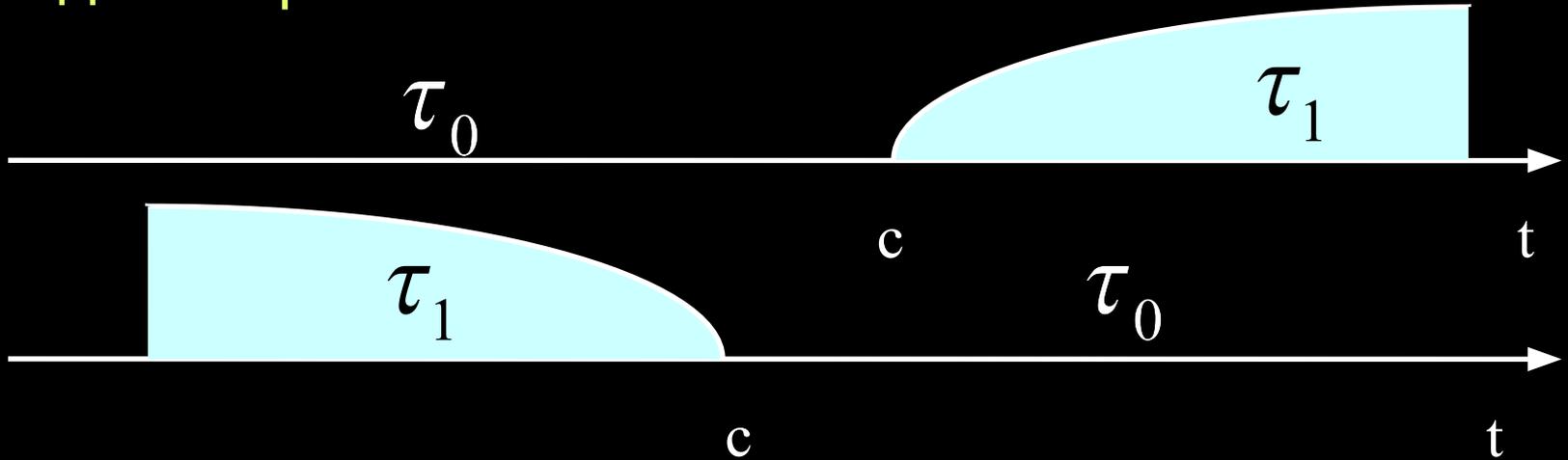
- принять или отклонить нулевую
гипотезу H_0

Каждому критерию отвечает разбиение области значений *статистики критерия* на две непересекающихся части:

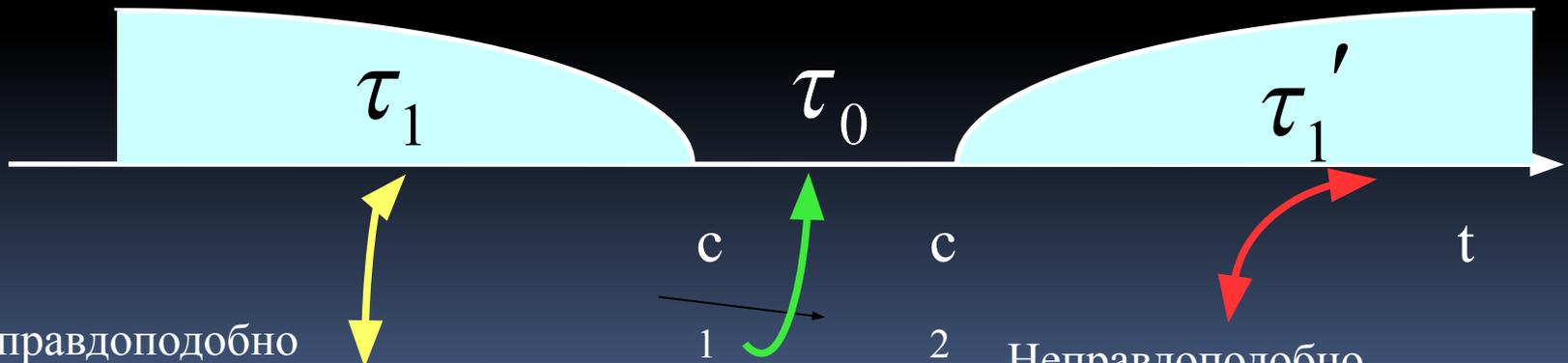
- *критическую область τ_1*
- *область принятия гипотезы τ_0*

Критические области

Односторонние



Двусторонняя



Неправдоподобно
маленькие значения

1 2
Приемлемые значения

Неправдоподобно
большие значения

Если значение статистики критерия попадает в область принятия гипотезы τ_0 , то принимается *нулевая* гипотеза, в противном случае она отвергается (принимается *альтернативная* гипотеза)

Задать статистический критерий

значит:

- задать статистику критерия
- задать критическую область

В ходе проверки гипотезы H_0 можно прийти к правильному выводу, либо совершить **два рода ошибок**:

- ошибку первого рода -- **отклонить H_0** , когда она верна
- ошибку второго рода -- **принять H_0** , когда она не верна.

Так как статистика критерия $T(X) \in \tau$ есть случайная величина со своим законом распределения, то попадание её в ту или иную область характеризуется соответствующими вероятностями:

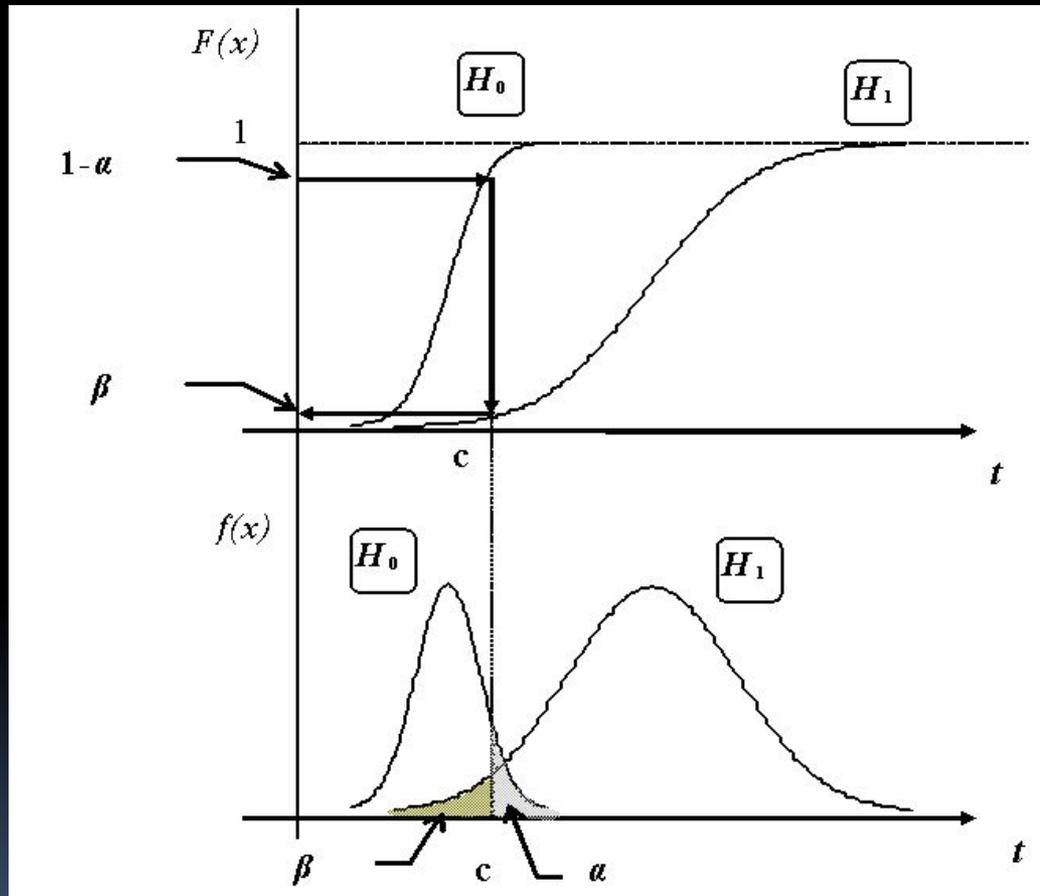
- вероятностью ошибки первого рода α
- вероятностью ошибки второго рода β

Ошибку первого рода α ещё называют **уровнем значимости критерия**.

Часто пользуются понятием **мощности критерия W** -- вероятности попадания в критическую область при условии справедливости альтернативной гипотезы

$$W = 1 - \beta$$

Распределение статистики критерия для нулевой и альтернативной гипотез (односторонний критерий)



Пять шагов проверки гипотезы

- 1 шаг – выдвигается основная гипотеза H_0
- 2 шаг – задается уровень значимости α
- 3 шаг – задается статистика критерия $T(X)$ с известным законом распределения

- 4 шаг – из таблиц распределения статистики критерия находятся квантили, соответствующие границам критической области
- 5 шаг – для данной выборки рассчитывается значение статистики критерия

Если значение статистики критерия попадает в область принятия гипотезы, то нулевая гипотеза принимается на уровне значимости α .

В противном случае принимается альтернативная гипотеза (отвергается нулевая гипотеза)

Пример: На основании сделанного прогноза средняя дебиторская задолженность одготипных предприятий региона должна составить $\alpha_0 = 120$ ден, ед. выборочная проверка 10 предприятий установила, что средняя задолженность $\bar{x} = 135$ ден.ед. $s = 20$ ден.ед.

На уровне значимости $\alpha = 0,05$ выяснить можно ли принять данный прогноз.

Решение:

Для проверки нулевой гипотезы $H_0: \alpha_0 = 120$

при альтернативной $H_1: \alpha_1 = 135$ построим статистику

$$t = \frac{\bar{X} - \alpha_0}{S} \sqrt{n-1} = \frac{135 - 120}{20} \sqrt{10-1} = 2,25,$$

$\alpha_1 > \alpha_0$ строим правостороннюю критическую область

$$t_{кр}(2\alpha, n-1) = St^{-1}(2\alpha, n-1)$$

$$t_{кр}(2 \cdot 0,05, n-1) = t_{кр}(0,1; 9) = 1,83$$

$t_{набл} > t_{кр}$ H_0 отвергаем, т.е на 5% уровне значимости сделанный прогноз должен быть отвергнут.