Повесьте ваши уши на гвоздь внимания !!!!!!

Случайная величина (СВ) и закон ее распределения (з.р.).

Случайная величина обозначается заглавной буквой ***X*** (если случайных величин несколько, то вводят У, Z и т.д.);

значение, которое принимает случайная величина, обозначается малой буквой *x*.

Пишут ***X*** **=** *x*. Это запись означает, что случайная величина приняла некоторое конкретное значение.

Случайной величиной называется числовая функция $X = X(\omega_k)$, заданная на пространстве элементарных исходов случайного эксперимента (т.е. для каждого значения ω_k задается определенное значение *X).*

Следует отметить, что и вероятность является числовой функцией, заданной на пространстве элементарных исходов случайного эксперимента, т.е. $P = P(\omega_k)$

Определение. ***Математическим ожиданием***, *или* ***средним значением***, *M(X) дискретной случайной величины X называется сумма произведений всех ее значений на соответствующие им вероятности*[2]:

$$M(X) = \sum_{i=1}^{n} x_i p_i.$$

Определение. ***Дисперсией*** *D(X) случайной величины X называется математическое ожидание квадрата ее отклонения от математического ожидания*[1]:

$$D(X) = M[X - M(X)]^2,$$

Определение. ***Средним квадратическим отклонением*** (***стандартным отклонением***, или ***стандартом***) σ_x *случайной величины X называется арифметическое значение корня квадратного из ее дисперсии:*

$$\sigma_x = \sqrt{D(X)}.$$

1. *Дисперсия постоянной величины равна нулю:*

$$D(C) = 0.$$

2. *Постоянный множитель можно выносить за знак дисперсии, возведя его при этом в квадрат:*

$$D(kX) = k^2 D(X).$$

3. *Дисперсия случайной величины равна разности между математическим ожиданием квадрата случайной величины и квадратом ее математического ожидания:*

$$D(X) = M(X^2) - [M(X)]^2,$$

4. *Дисперсия алгебраической суммы конечного числа независимых случайных величин равна сумме их дисперсий*[2]*:*

$$D(X \pm Y) = D(X) + D(Y).$$

Математическое ожидание, дисперсия, среднее квадратическое отклонение и другие *числа, призванные в сжатой форме выразить наиболее существенные черты распределения, называются* ***числовыми характеристиками*** *случайной величины.*

Обращаем внимание на то, что сама величина X – *случайная*, а ее *числовые характеристики являются величинами неслучайными, постоянными.* Поэтому их часто называют ***параметрами распределения*** случайной величины.

Существует два типа случайных величин – **дискретные и непрерывные.**

Закон распределения случайной величины – это правило, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Введем универсальный з.р., который подходит как для описания поведения дискретной СВ, так и для описания поведения непрерывной СВ.

Функцией распределения случайной величины называют

$$F(x) = P(X < x)$$

Рассмотрим общие с в о й с т в а функции распределения.

1. *Функция распределения случайной величины есть неотрицательная функция, заключенная между нулем и единицей*:

$$0 \le F(x) \le 1.$$

□ Утверждение следует из того, что функция распределения — это вероятность. ■

2. *Функция распределения случайной величины есть неубывающая функция на всей числовой оси.*

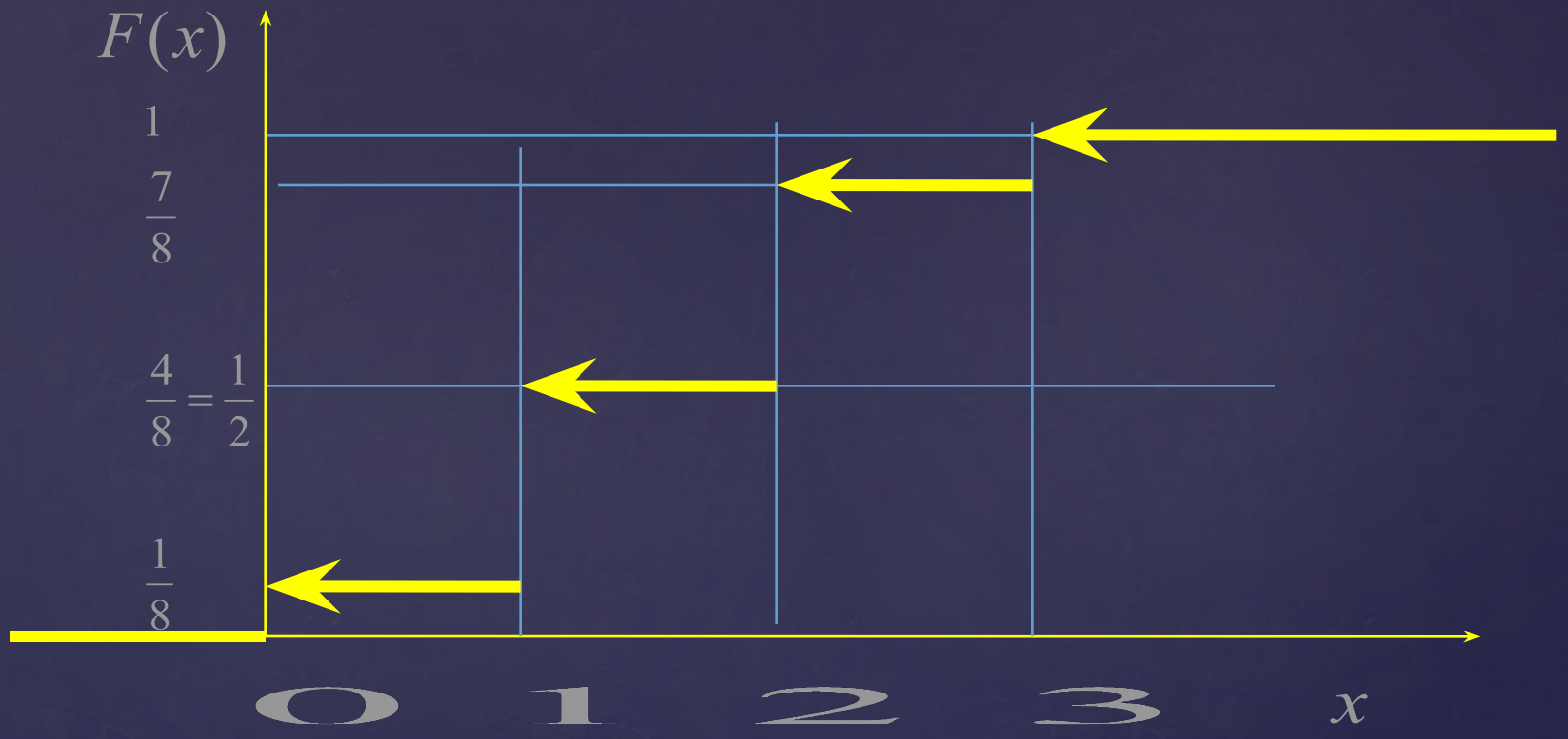
3. *На минус бесконечности функция распределения равна нулю, на плюс бесконечности равна единице*, т.е.

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \qquad F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

4. *Вероятность попадания случайной величины в интервал* $[x_1, x_2)$ (включая x_1) *равна приращению ее функции распределения на этом интервале*, т.е.

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1).$$

Пример графика функции распределения для дискретной случайной величины X – числа выпадений герба при трехкратном бросании правильной монеты.

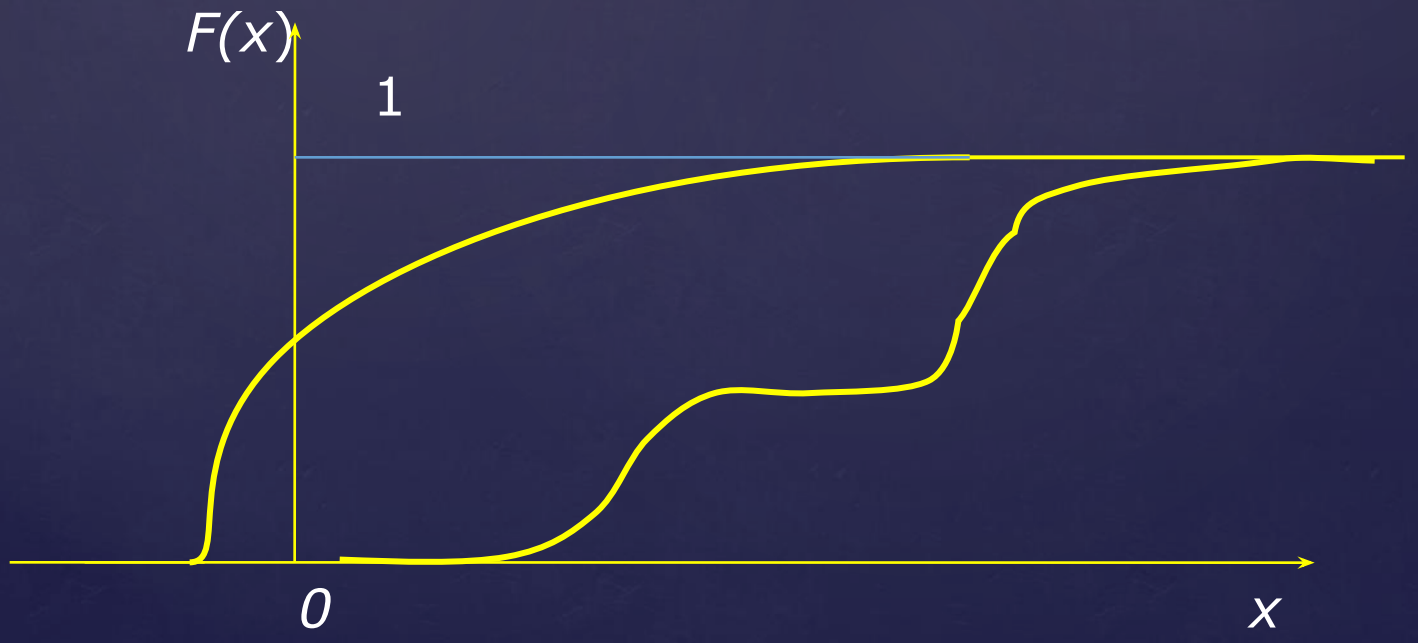


Если случайная величина такова, что ее функция распределения может быть представлена в виде:

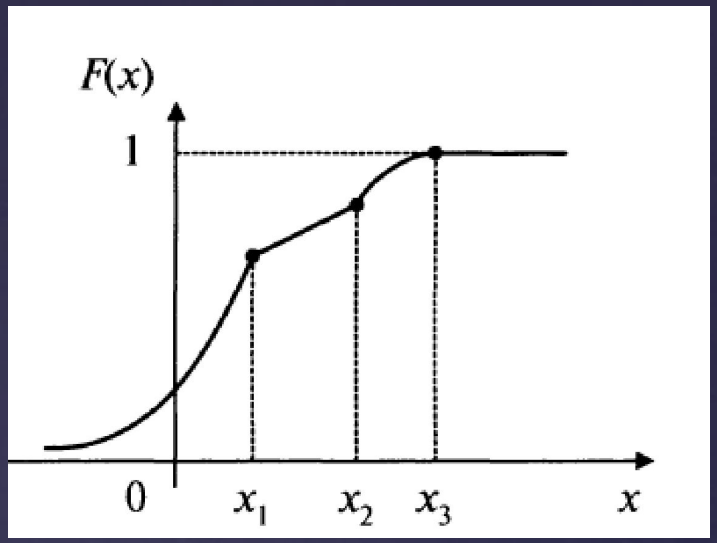
$$F(x) = \int_{-\infty}^{x} f(t) \cdot dt,$$

(здесь t – переменная интегрирования), то мы назовем ее непрерывной случайной величиной.

График функции распределения для непрерывной СВ может выглядеть, например, следующим образом:



Определение. *Случайная величина X называется* ***непрерывной****, если ее функция распределения непрерывна в любой точке и дифференцируема всюду, кроме, быть может, отдельных точек.*



Определение. ***Плотностью вероятности*** (***плотностью распределения*** или просто ***плотностью***) $\varphi(x)$ *непрерывной случайной величины X называется производная ее функции распределения* $\varphi(x) = F'(x)$.

Плотность вероятности $\varphi(x)$, как и функция распределения $F(x)$, является одной из форм закона распределения, но в отличие от функции распределения она существует только для непрерывных случайных величин.

Плотность вероятности иногда называют *дифференциальной функцией* или *дифференциальным законом распределения*.

График плотности вероятности $\varphi(x)$ называется *кривой распределения*.

Отметим с в о й с т в а плотности вероятности непрерывной случайной величины.

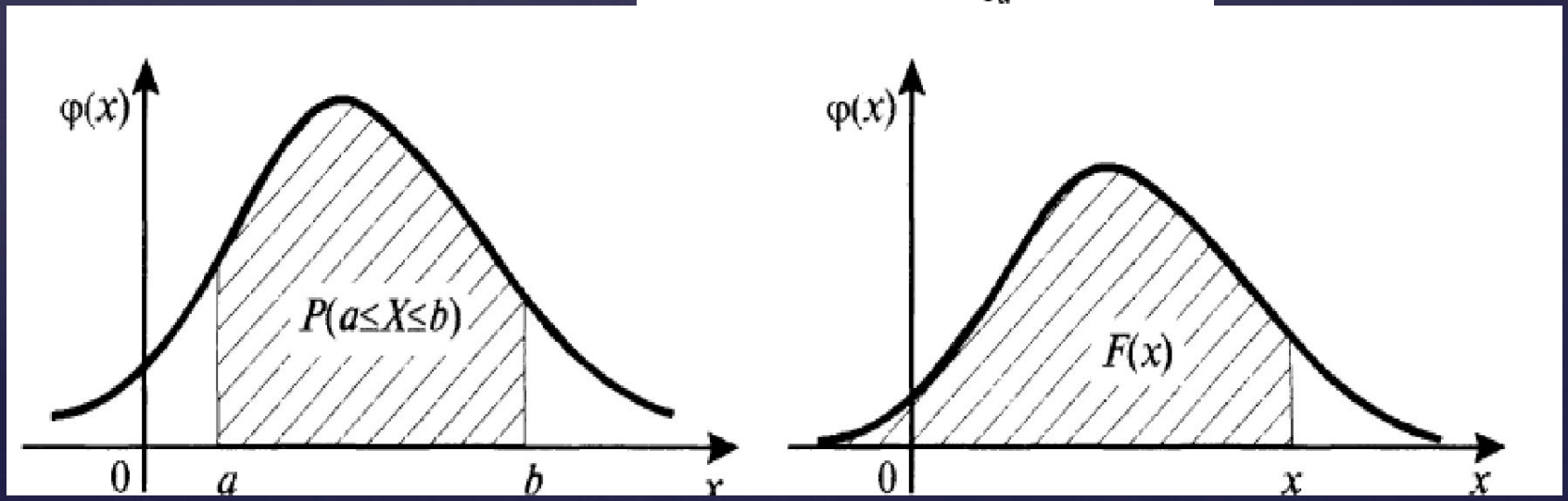
1. *Плотность вероятности — неотрицательная функция*, т.е.

$$\varphi(x) \geq 0.$$

□ $\varphi(x) \geq 0$ как производная монотонно неубывающей функции $F(x)$. ■

2. *Вероятность попадания непрерывной случайной величины в интервал* $[a,b]$ *равна определенному интегралу от ее плотности вероятности в пределах от* a *до* b, т.е.

$$P(a \leq X \leq b) = \int_a^b \varphi(x)dx.$$



3. *Функция распределения непрерывной случайной величины может быть выражена через плотность вероятности по формуле:*

$$F(x) = \int_{-\infty}^{x} \varphi(x)\,dx.$$

4. *Несобственный интеграл в бесконечных пределах от плотности вероятности непрерывной случайной величины равен единице:*

$$\int_{-\infty}^{+\infty} \varphi(x)\,dx = 1.$$

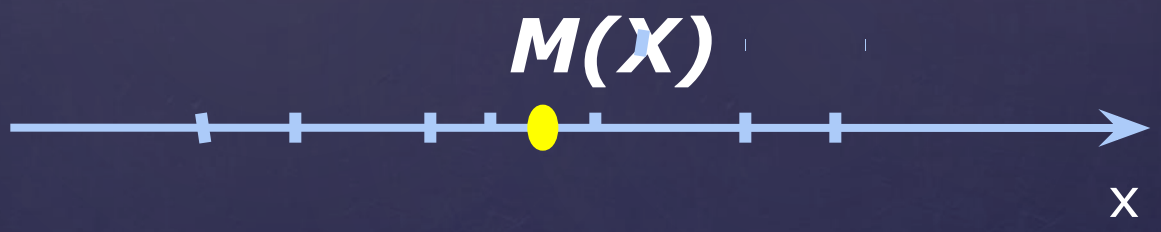
Функцию *f(x)* используют для описания поведения непрерывных случайных величин, ибо она полностью содержит всю информацию, которая нужна для анализа поведения непрерывных случайных величин.

Вероятность попадания непрерывной случайной величины в заданный числовой промежуток определяется формулой:

$$P(a \le X < b) = \int_a^b f(x)dx$$

Числовые характеристики случайной величины - математическое ожидание, дисперсия, стандартное отклонение; их свойства.

Рассмотрим дискретную случайную величину, принимающую некоторые значения на числовой оси:



Определение:

Математическим ожиданием дискретной случайной величины (ДСВ) называется

$$M(X) = \sum_{i=1} x_i \cdot p_i$$

Для случая $n \to \infty$ ряд должен быть сходящимся. Возникают иногда ситуации, когда ряд расходится. Тогда случайная величина не имеет математического ожидания. Такие случай мы рассматривать не будем.

Статистический смысл математического ожидания:

Вычисляя среднее арифметическое всех наблюдаемых значений СВ, получают математическое ожидание СВ в практических задачах.

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n} = x_1 \frac{n_1}{n} + x_2 \frac{n_2}{n} + \ldots + x_k \frac{n_k}{n} =$$

$$x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = EX$$

$$EX = \bar{x} - среднеарифметическое\ для\ ДСВ$$

Определение:

Математическим ожиданием непрерывной случайной величины (НСВ) называется :

$$M(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Математическое ожидание уже не является случайной величиной. Это постоянная величина для данного закона распределения СВ. Она является обобщенной характеристикой данного распределения, указывая то значение, около которого располагаются все возможные значения, принимаемые данной случайной величины.

Рассмотрены свойства математического ожидания.

Математическое ожидание характеризует центр распределения случайной величины и не дает представление о разбросе возможных значений случайной величины, хотя значения случайной величины могут сильно или же не сильно отклоняться от своего теоретического центра (математического ожидания).

Мера разброса возможных значений случайной величины является важной характеристикой поведения случайной величины.

Определение:

Дисперсией случайной величины называется математическое ожидание квадрата отклонения случайной величины от ее теоретического центра:

$$DX = E\left[(X - EX)^2\right] =$$

$$= \begin{cases} \sum_{i=1}^{n\ (\infty)} (x_i - EX)^2 \cdot p_i - для\ \ ДСВ \\ \int_{-\infty}^{\infty} (x - EX)^2 \cdot f(x) \cdot dx - для\ \ НСВ \end{cases}$$

Формула, удобная для вычислений дисперсии:

$$DX = M\left(X^2\right) - \left(M(X)\right)^2$$

Определение:

Стандартным отклонением случайной величины называется

$$\sigma_X = \sqrt{\left(D(X)\right)}$$

Дисперсию можно записать символом как символом ***DX***, так и символом σ^2.

Стандартное отклонение имеет ту же размерность, что и сама случайная величина.

Рассмотрены свойства дисперсии и стандартного отклонения.

Статистический смысл дисперсии:

Вычислили среднее арифметическое на основе данных наблюдений. Далее найдем среднее арифметическое квадратов отклонений от среднего арифметического:

$$\frac{n_1(x_1-\bar{x})^2+n_2(x_2-\bar{x})^2+\ldots+n_k(x_k-\bar{x})^2}{n}=$$

$$=(x_1-\bar{x})^2\frac{n_1}{n}+(x_2-\bar{x})^2\frac{n_2}{n}+\ldots+(x_k-\bar{x})^2\frac{n_k}{n}\approx D(X)$$

$$M(X)\approx\bar{x};\ \frac{n_i}{n}\approx p_i;\ D(X)=\sum_{i=1}^{n}(x_i-M(X))^2\cdot p_i$$

Именно эта формула применяется для практического вычисления дисперсии на основе результатов наблюдений (в действительности знаменатель формулы несколько меняют – вместо n используют (n-1)).

Вычислены математическое ожидание, дисперсия и стандартное отклонение для СВ, распределенной по закону Бернулли (биномиальному закону):

$$DX = npq; \sigma = \sqrt{DX} = \sqrt{npq}$$

В отечественной литературе часто используется другое название для стандартного отклонения σ - среднее квадратическое отклонение.

В коммерческой деятельности стандартное отклонение σ характеризует риск, показывая, насколько неопределённой является ситуация.

Математическое ожидание и стандартное отклонение выражают в сжатой форме наиболее характерные черты закона распределения случайной величины, а именно, его теоретический центр и меру отклонения от этого теоретического центра.

Эти величины для данного распределения являются константами (неслучайными величинами).

Используются и некоторые другие константы распределения, позволяющие выявить особенности данного конкретного распределения. Введем некоторые них.

Определения:

Квантилем уровня *p* **(или *p* - квантилем)** называется такое значение x_p случайной величины, которое является решением уравнения

$$F(x_p) = P(X < x_p) = p,$$

т.е. при котором функция распределения принимает значение, равное *p*.

Модой MoX СВ X называется её наиболее вероятное значение, т.е. это такое значение СВ, для которого вероятность для дискретной СВ или плотность вероятности для непрерывной СВ достигает своего максимума.

Медианой MeX случайной величины называют такое её значение, для которого

$$P(X \le MeX) = P(X \ge MeX) = 0{,}5$$

Медиана – это квантиль уровня 0.5.

§ 12. Наиболее часто используемые законы распределения случайных величин.

Дискретные случайные величины:

Для ДСВ наиболее часто используется биномиальный закон распределения.

Кроме биномиального закона распределения наиболее часто используется ***распределение Пуассона***, которое является следствием (предельным случаем) распределения Бернулли. Оно получено предельным переходом из биномиального закона при выполнении определенных ограничений:

n – велико; p – мало; $\lambda = const = 0(1)$.

Формула Пуассона:

$$P(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Параметр λ называют ***интенсивностью потока событий.***

Формула Пуассона имеет и самостоятельное значение, когда в задаче рассматривается поток событий, имеющий заданную интенсивность.

Для распределения Пуассона $M(X)=\lambda$, $DX=\lambda$.

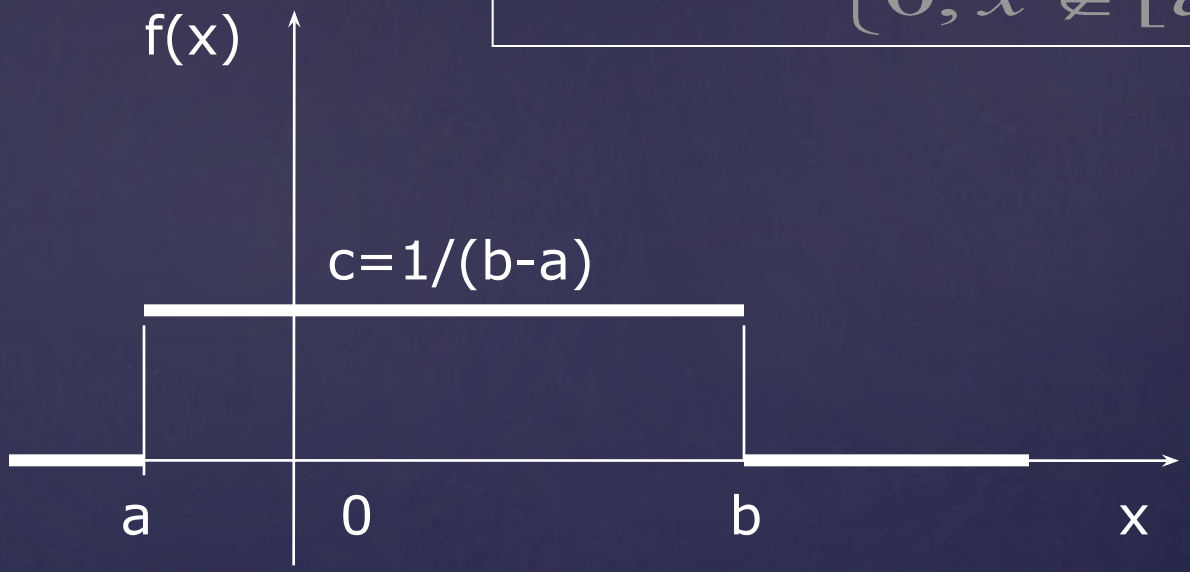
Если нас интересует наступление определенного числа событий А не за единицу времени, а за другой промежуток времени t, отличный от единицы, то формула Пуассона приобретает такой вид:

$$P_t(k) = \frac{(\lambda t)^k \cdot e^{-\lambda t}}{k!}$$

Непрерывные случайные величины:

СВ X имеет равномерный закон распределения на отрезке [a, b], если ее плотность распределения постоянна на этом отрезке и равна нулю вне его:

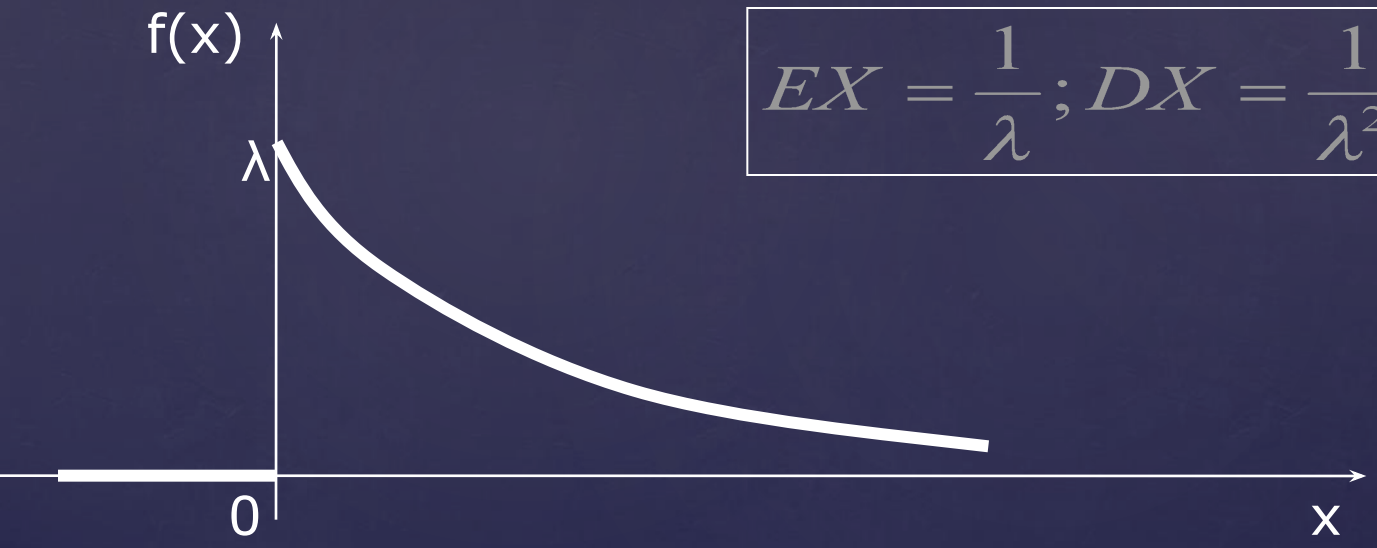
$$f(x)=\begin{cases} c, x \in [a;b] \\ 0, x \notin [a;b] \end{cases}$$



$$EX=\frac{a+b}{2}; DX=\frac{(b-a)^2}{12}; \sigma=\frac{b-a}{2\sqrt{3}}.$$

Непрерывная СВ X имеет показательный (экспоненциальный) закон распределения с параметром λ, если ее плотность распределения имеет вид:

$$f(x)=\begin{cases}\lambda\cdot e^{-\lambda x}, x\geq 0, \lambda>0;\\ 0, x<0.\end{cases}$$



$$EX=\frac{1}{\lambda};\ DX=\frac{1}{\lambda^2};\quad \sigma=\frac{1}{\lambda}.$$

В показательном законе смысл параметра λ тот же самый, что и в законе Пуассона – среднее количество событий за единицу времени.

Между законами распределения Пуассона и показательным существует тесная связь:

Количество событий за любой фиксированный промежуток времени имеет распределение Пуассона, а время ожидания между событиями - показательное распределение.

Поток событий, для описания которого справедливы упомянутые распределения, должен быть подчинен определенным **ограничениям** для того, чтобы его поведение можно было описать такими простыми формулами.

Эти ограничения потока событий таковы:

1. ***Стационарность*** **(интенсивность потока событий λ не зависит от времени);**
2. *Отсутствие последействия* (количество событий, попадающих на данный промежуток времени, не зависит от числа событий, попадающих на другой промежуток времени, не пересекающийся с данным);
3. *Ординарность* (вероятность попадания на малый промежуток времени двух или более событий пренебрежимо мала по сравнению с вероятностью попадания на этот же малый промежуток времени одного события).

Поток событий называется **простейшим** (или **стационарным пуассоновским**), если он одновременно обладает свойствами 1, 2, 3.
Эта модель потока событий обладает свойством, которое называется **характеристическим свойством** или **свойством «отсутствия памяти».**

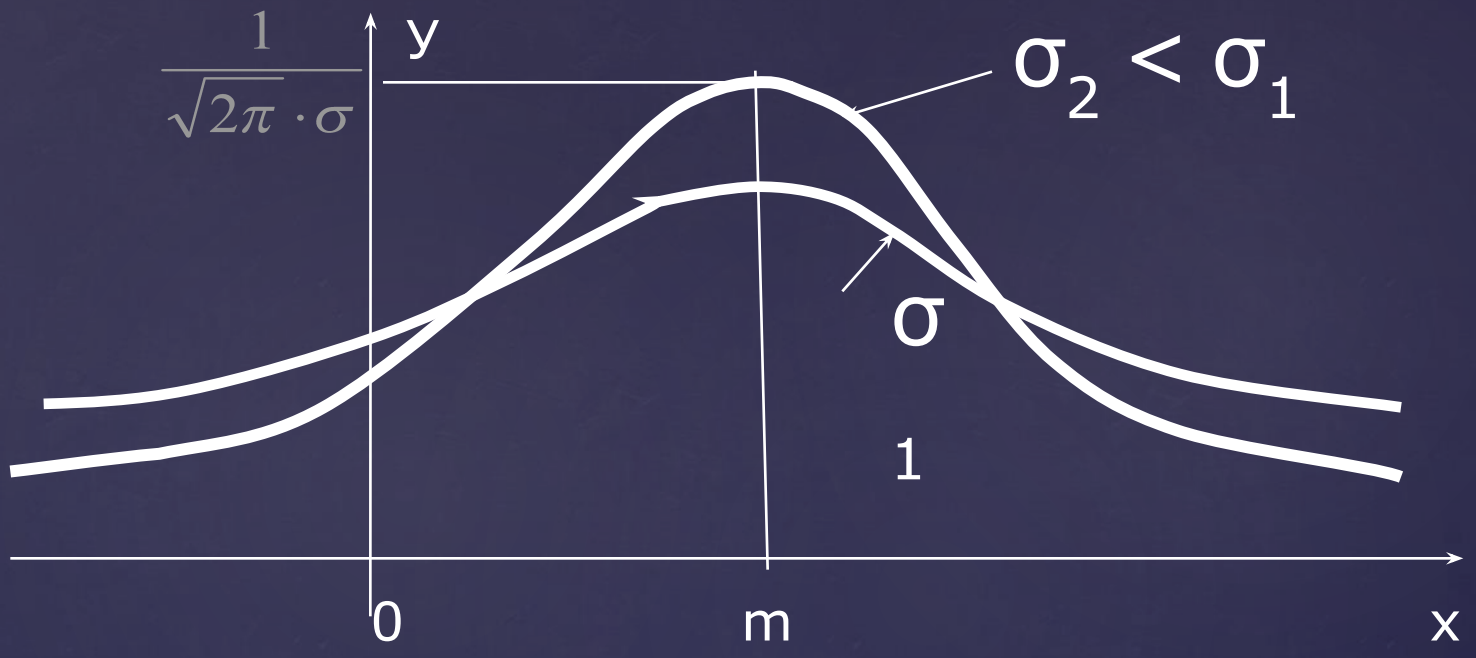
СВ X имеет нормальный закон распределения с параметрами m и σ, если ее плотность распределения имеет вид:

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

Обозначение: $X \sim N(m;\sigma)$

Параметры m и σ имеют определенный смысл. Для выяснения этого смысла следует вычислить математическое ожидание и стандартное отклонение нормально распределенной СВ. Оказывается, что они совпадают с этими параметрами.

График плотности нормального распределения имеет вид:



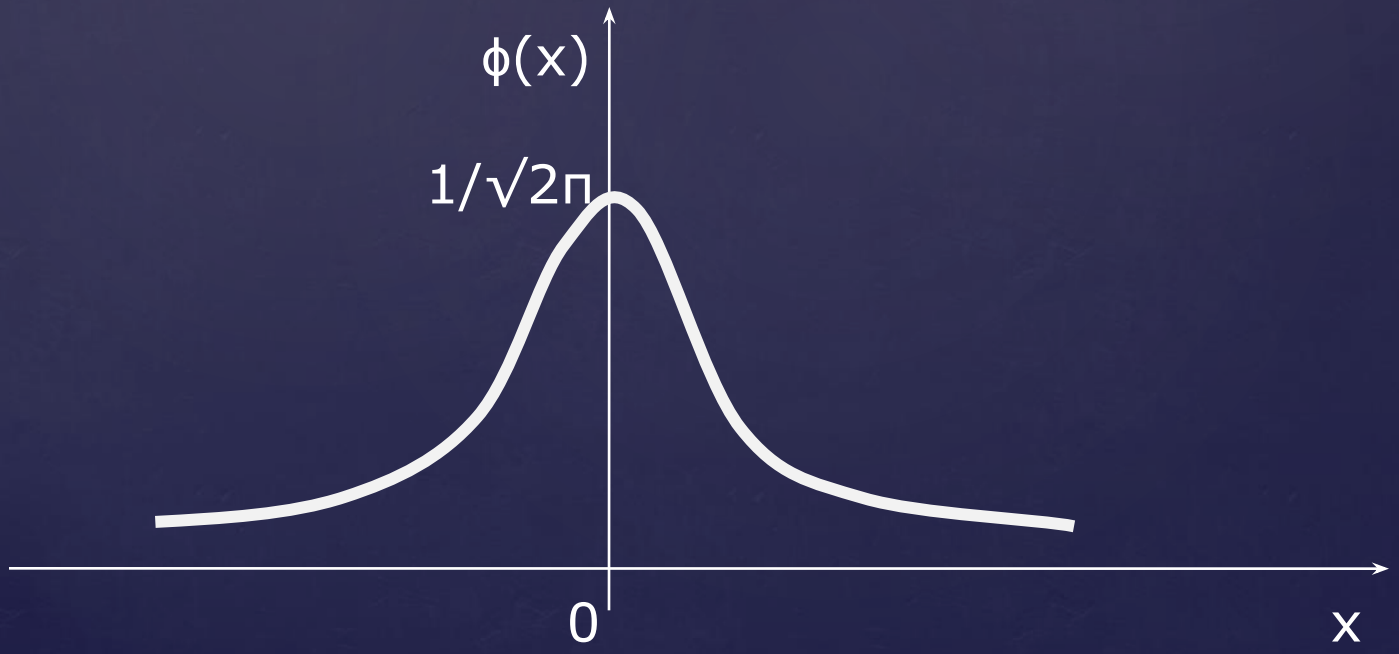
Площадь под кривой сохраняет постоянное значение, равное единице, при любых изменениях σ. Чем больше значение σ, тем более плавно идет кривая плотности.

Стандартным нормальным распределением называется распределение нормальной случайной величины с m=0 и σ=1.

Обозначение: $Z \sim N(0;1)$.

Плотность распределения стандартной нормальной СВ имеет вид:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$



Формула для вычисления вероятности попадания нормально распределенной СВ в заданный интервал:

$$P(x_1 < X < x_2) = P\left(\frac{x_1 - m}{\sigma} < \frac{X - m}{\sigma} < \frac{x_2 - m}{\sigma}\right) =$$

$$= \Phi_0\left(\frac{x_2 - m}{\sigma}\right) - \Phi_0\left(\frac{x_1 - m}{\sigma}\right)$$

Справедлива формула:

$$P(|X - m| < \varepsilon) = 2\Phi_0\left(\frac{\varepsilon}{\sigma}\right)$$

На основе этой формулы может быть получено «**правило трех сигм**»:

$$\varepsilon = 3\sigma \Rightarrow P(|X - m| < \varepsilon) = 2\Phi_0\left(\frac{3\sigma}{\sigma}\right) = 2\Phi_0(3) = 2 \cdot 0.49865 = 0.9973.$$

Если случайная величина распределена нормально, то ее отклонение от математического ожидания практически не превосходит утроенного стандартного отклонения.

Устойчивость некоторых законов распределения.

Если СВ нормально распределена: $X \sim N(m;\sigma)$, то СВ $Y=aX+b$ также подчиняется нормальному закону распределения, причем:

$$M(Y) = aMX + b; \quad \sigma_Y = |a|\sigma_X$$

Закон распределения называется устойчивым, если СВ, равная сумме двух независимых СВ, имеет тот же закон распределения, что и законы распределения суммируемых СВ.

Показано, что если случайная величина Z находится как сумма двух независимых нормально распределенных случайных величин X и Y, то Z также будет нормально распределена, причем

$$Если \quad Z = X + Y, где \quad X \sim N(m_X;\sigma_X) \quad и \quad Y \sim N(m_Y;\sigma_Y),$$

$$то \quad Z \sim N(m_Z;\sigma_Z), \quad где \quad m_Z = m_X + m_Y; \quad \sigma_Z^{\,2} = \sigma_X^{\,2} + \sigma_Y^{\,2}.$$

Предельные теоремы теории вероятностей.

Неравенство Чебышева.

Неравенство Маркова (или лемма Чебышева)

Если случайная величина X принимает только неотрицательные значения и имеет математическое ожидание EX, то для любого положительного числа α справедливо неравенство: $P(X \geq \alpha) \leq \frac{MX}{\alpha}$.

Теорема (*неравенство Чебышева*):

Если случайная величина X имеет математическое ожидание EX и дисперсию DX, то для любого ε > 0 справедливо неравенство:

$$P(|X - MX| \geq \varepsilon) \leq \frac{DX}{\varepsilon^2}.$$

Теорема Чебышева. Закон больших чисел (ЗБЧ).

Введем понятие сходимости по вероятности:

$$\lim_{n\to\infty} P\left(|X_n - a| < \varepsilon\right) = 1 \quad \textit{для сколь угодно малого положительного } \varepsilon;$$

или другая более компактная форма записи:

$$X_n \xrightarrow[n\to\infty]{P} a.$$

Формулировка ЗБЧ в форме Чебышева П.Л. ***(теорема Чебышева):***

Если дисперсии n независимых случайных величин X_1, X_2,…, X_n ограничены сверху одной и той же константой: $DX_i \le C$, $i=1, 2,\ldots, n$,

то для любого сколь угодно малого положительного числа ε

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{m_1 + m_2 + \ldots + m_n}{n}\right| < \varepsilon\right) = 1, \text{где} \quad m_i = MX_i.$$

Или $$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{P} \frac{\sum_{i=1}^{n} m_i}{n}.$$

Следствия ***из теоремы Чебышева***:

Первое следствие: **Теорема Хинчина**

Если независимые случайные величины X_1, X_2,…, X_n имеют одинаковые математические ожидания, равные m, то

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{\text{P}} m.$$

Это соотношение является основой выборочного метода (статистических исследований). Если мы хотим узнать истинное значение какого-то параметра m, нам нужно несколько раз экспериментально получить значения X_i этого параметра и затем на основе этих значений вычислить их среднее арифметическое. Вычисленная величина будет достаточно хорошим приближением истинного значения параметра, причем чем больше включено в расчет экспериментальных значений, тем более точное приближение истинного значения параметра будет получено.

Второе следствие: **Теорема Бернулли**

Пусть проводится *n* независимых испытаний, в каждом из которых событие А может произойти с одной и той же вероятностью *p* (схема Бернулли). При неограниченном возрастании числа опытов *n* частота события А сходится по вероятности к вероятности *p* этого события в отдельном испытании:

$$\frac{k}{n} \xrightarrow[n\to\infty]{P} p$$

Здесь k - количество случаев, когда событие А наблюдалось.

Третье следствие:

ЗБЧ может быть распространен и на зависимые случайные величины (это обобщение принадлежит Маркову А.А.):

Если имеются зависимые случайные величины X_1, X_2, …, X_n и если при

$$при \quad n \to \infty \quad \frac{D\left(\sum_{i=1}^{n} X_i\right)}{n^2} \underset{n\to\infty}{\to} 0, \quad что \quad было \quad выполнено \quad в \quad предыдущей$$

$$теореме, \quad то \quad \frac{\sum_{i=1}^{n} X_i}{n} \underset{n\to\infty}{\xrightarrow{P}} \frac{\sum_{i=1}^{n} m_i}{n}.$$

Смысл и формулировка центральной предельной теоремы (ЦПТ). Интегральная теорема Муавра-Лапласа как следствие ЦПТ.

Эта теорема утверждает, что распределение суммы большого числа независимых и сравнимых по вкладам в сумму случайных величин близко к нормальному закону распределения.

Иначе:

если $Y_n = X_1 + X_2 + \ldots + X_n$, причем

1) Слагаемых много;
2) Слагаемые независимые;
3) Слагаемые сравнимы по вкладам в сумму, т.е. нет слагаемого, которое было бы по вкладу существенно больше остальных,

то ЦПТ утверждает, что СВ Y_n подчиняется нормальному закону распределения.

Именно поэтому нормальный закон распределения так широко применяется в практических задачах, ибо в реальных задачах исследуемые случайные величины часто есть результат сложения многих других случайных величин.

Упрощенная математическая формулировка ЦПТ:

Если X_1 , X_2 ,…, X_n – независимые случайные величины, для каждой из которых существует математическое ожидание $EX_i = m_i$ и дисперсия $DX_i=\sigma_i^{\,2}$, а также выполняется некоторое дополнительное условие , то закон распределения $Y_n = X_1 + X_2 + \ldots + X_n$ при $n \to \infty$ асимптотически приближается к нормальному закону распределения с параметрами

$$m_{Y_n} = \sum_{i=1}^{n} m_i \qquad и \qquad DY_n = \sum_{i=1}^{n} \sigma_i^{\,2}.$$

Что касается упомянутого в формулировке теоремы дополнительного условия, то оно сложно записывается математически, но означает, что вклад каждого слагаемого в сумму ничтожно мал, т.е. слагаемые соразмерны по своим вкладам в сумму.

Из ЦПТ для схемы испытаний Бернулли вытекает как следствие интегральная теорема Муавра – Лапласа.

§17. Многомерная случайная величина и закон ее распределения.

Пусть имеется система случайных величин (СВ), причем эта система может состоять как из дискретных, так и из непрерывных СВ. Будем рассматривать их как координаты случайного вектора.

Определение. n-мерной случайной величиной или случайным вектором называется упорядоченный набор n случайных величин

$$\overrightarrow{X} = (X_1, X_2, \ldots, X_n)$$

Для описания поведения многомерной СВ должен быть введен закон ее распределения:

$$F(x_1, x_2, \ldots, x_n) = (опред.) =$$
$$= P(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n)$$

Эта функция выражает вероятность совместного выполнения неравенств в правой части этого соотношения.

С целью экономии времени изложение выполним для двумерного случая; при этом будем понимать, что все утверждения справедливы и для n>2:

$$F(x, y) = P(X < x, Y < y)$$

Рассмотрены свойства функции F(x,y).

Могут быть получены **частные (маргинальные)** функции распределения на основе функции совместного распределения двух случайных величин:

$$F(x,+\infty) = F_X(x) \qquad F(+\infty, y) = F_Y(y)$$

Для двумерной непрерывной случайной величины (X,Y) функция совместного распределения может быть представлена в виде:

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\cdot du\cdot dv,\ причем\ \ если\ \ продифференцировать$$

$$это\ \ равенство\ \ по\ \ x\ \ и\ \ y,\ \ то\ \ найдем\ \ другую\ \ формулу\ \ связи$$

$$между\ \ этими\ \ функциями:\qquad f(x,y)=\frac{\partial^2 F(x,y)}{\partial x\cdot\partial y}$$

Для функции f(x,y), которая называется плотностью совместного распределения, справедливы те же **свойства**, которые были получены для функции f(x) в одномерном случае.

Зная плотность совместного распределения двух случайных величин, можно найти плотность **частного (маргинального)** распределения одной случайной величины:

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)\cdot dy;\qquad f_Y(y)=\int_{-\infty}^{+\infty} f(x,y)\cdot dx$$

Для независимых случайных величин X и Y независимы события {X<x} и {Y<y}, откуда следует:

$$F(x, y) = F_X(x) \cdot F_Y(y)$$

Для непрерывных СВ из данного соотношения, дифференцируя его по x и y, получим:

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

Для зависимых СВ эти равенства не выполняются:

$$F(x, y) \neq F_X(x) \cdot F_Y(y); \qquad f(x, y) \neq f_X(x) \cdot f_Y(y)$$

§18. Стохастическая зависимость двух случайных величин. Ковариация и коэффициент корреляции.

Если случайные величины зависимы, влияют на поведение друг друга, то следует количественно описать степень их влияния друг на друга.

Определение.

Ковариацией двух СВ X и Y называется математическое ожидание произведения соответствующих центрированных СВ:

$$\operatorname{cov}(X, Y) = E((X - EX) \cdot (Y - EY)) =$$

$$\begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - EX)\cdot(y_j - EY)\cdot p_{ij} - ДСВ;\ p_{ij} \to (x_i; y_j) \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - EX)\cdot(y - EY)\cdot f(x,y)\cdot dxdy - НСВ \end{cases}$$

Рассмотрены свойства ковариации.

Вывод:
ковариация не улавливает сложные виды связей между X и Y. Ковариация отслеживает наличие только **линейной связи** между СВ. При наличии такой линейной связи (стохастической) ковариация отлична от 0.

Определение:

Коэффициентом корреляции двух СВ X и Y называется отношение их ковариации к произведению стандартных отклонений этих величин:

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\cdot\sigma_Y}$$

Рассмотрены свойства коэффициента корреляции.
Значения, принимаемые коэффициентом корреляции:

$$-1\le\rho\le 1$$

Определение.
Случайные величины называются некоррелированными, если их коэффициент корреляции равен нулю. Случайные величины называются коррелированными, если их коэффициент корреляции отличен от нуля.

Было показано, что если случайные величины независимые, то они некоррелированные, а из некоррелированности случайных величин еще не следует их независимость. Из некоррелированности нормальных СВ следует их независимость (в общем случае это не так.)

Коэффициент корреляции характеризует степень линейной зависимости между случайными величинами X и Y в стохастическом смысле и не может отражать более сложных видов зависимостей между случайными величинами.

Графически показана стохастическая линейная связь между случайными величинами при различных значениях коэффициента корреляции.

Введено уравнение линейной регрессии, наилучшим образом описывающим связь между случайными величинами:

$$y = EY + \rho \cdot \frac{\sigma_y}{\sigma_x} \cdot (x - EX)$$

Для вычисления коэффициента корреляции между двумя количественными признаками на практике используется **линейный коэффициент корреляции Пирсона:**

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Введем коэффициент корреляции для изучения тесноты связи между порядковыми случайными величинами.

Если *n* объектов совокупности пронумеровать в соответствии с возрастанием или убыванием изучаемого признака, то говорят, что объекты **ранжированы** по этому признаку. Присвоенный номер называется **рангом.**

Коэффициент ранговой корреляции Спирмена вычисляется по формуле:

$$\rho_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)}, \quad где$$

r_i – ранги по первому признаку; s_i – ранги по второму признаку;

$d_i = (r_i - s_i)$ – разность рангов i-ого объекта; $|\rho_s| \le 1$

В случае совпадения рангов при вычислении коэффициента ранговой корреляции следует брать среднее арифметическое рангов, приходящихся на данные объекты, причем каждому объекту присваивается это среднее арифметическое значение. В формулу вводятся поправки на совпадающие ранги T_a и T_b. Формула приобретает такой вид:

$$\rho_s = \frac{\frac{1}{6}\cdot\left(n^3-n\right)-\sum_{i=1}^{n} d_i^2 - T_a - T_b}{\sqrt{\frac{1}{6}\cdot\left(n^3-n\right)-2T_a}\cdot\sqrt{\frac{1}{6}\cdot\left(n^3-n\right)-2T_b}}$$

$$T_a = \frac{1}{12}\cdot\sum_{k=1}^{m_a}\left(a_k^3-a_k\right), \qquad T_b = \frac{1}{12}\cdot\sum_{k=1}^{m_b}\left(b_k^3-b_k\right), \quad \text{где}$$

a_k, b_k – объемы каждой группы с совпадающими рангами по первому и по второму признаку;

m_a, m_b – количество групп с совпадающими рангами по первому и по второму признаку.

Раздел 2. Элементы математической статистики.

Начнем с нового раздела нумерацию параграфов заново.

§ 1. Случайные выборки. Первичная обработка статистических данных. Вариационные ряды.

Статистика изучает большие массивы информации и устанавливает закономерности, которым подчиняются случайные массовые явления.

Генеральной совокупностью (ГС) называется вся подлежащая изучению какого-либо свойства (говорят, признака) совокупность объектов.

Та часть объектов, которая отобрана для непосредственного изучения какого-либо признака ГС носит название **случайной выборки** (или просто **выборки**).

Объем ГС и объем выборки – это количество элементов в них. Обозначаются , соответственно, N и n.

В дальнейшем будем считать, что объем выборки существенно меньше объема генеральной совокупности. В этом случае получаемые в дальнейшем формулы являются наиболее простыми.

Непрерывная природа изучаемого признака порождает бесконечные ГС.

Для того, чтобы выборка была **репрезентативной** (хорошо представлять элементы ГС), она должна быть отобрана случайно. Случайность отбора элементов в выборку достигается соблюдением принципа равной возможности каждого элемента ГС быть отобранным в выборку.

Нарушение принципов случайного выбора приводит к серьезным ошибкам.

Любое число, полученное на основе выборки, носит название **«выборочная статистика»** (или просто «статистика»).

Пусть получена выборка объема n. Над этим массивом исходных данных выполняется операция ранжирования, т.е. экспериментальные данные выстраиваются в порядке возрастания:

$x_1 < x_2 < x_3 < \ldots < x_k; \quad k \le n;$

причем значение x_i *встречается* n_i *раз*:

$n_1 + n_2 + \ldots + n_k = n;$

вводится терминология:

x_i – *вариант*; n_i – *частота варианта*

(*количество появлений значений* x_i);

$w_i = \frac{n_i}{n}$ – *относительная частота варианта или частость*;

обязательно выполняется $\sum_{i=1}^{k} w_i = 1;$

размах выборки $R = x_{max} - x_{\min} = x_k - x_1.$

Определение.

Вариационным рядом называется ранжированный в порядке возрастания ряд значений (вариантов) с соответствующими им частотами.

Значения x_i	x_1	x_2	...	x_k
Частоты n_i	n_1	n_2	...	n_k
Частости $w_i=n_i/n$	w_1	w_2	... ...	w_k

Данный вариационный ряд носит название дискретного вариационного ряда (его члены принимают отдельные изолированные значения).

Построение дискретного вариационного ряда нецелесообразно, когда число значений в выборке велико или признак имеет непрерывную природу, т.е. может принимать любые значения в пределах некоторого интервала. В этом случае строят интервальный вариационный ряд.

Вид интервального ряда:

Интервалы вариантов	$x_1 - x_2$ *1*	$x_2 - x_3$ *2*	...	$x_{k-1} - x_k$ *k-1*
Частоты n_i *(число вар-тов, попавших в инт-вал)*	n_1	n_2	...	n_{k-1}
Частости $w_i = n_i/n$	w_1	w_2	...	w_{k-1}

В том случае, когда можно предположить, что изучаемый признак в ГС подчиняются нормальному з.р., для вычисления количества интервалов равной длины применяют формулу Стерджесса:

$$m = 1 + 3.3 \cdot \lg n, \quad \text{если} \quad m \in [6;12]$$

Если $m > 12$, то принимают $m = 12$;

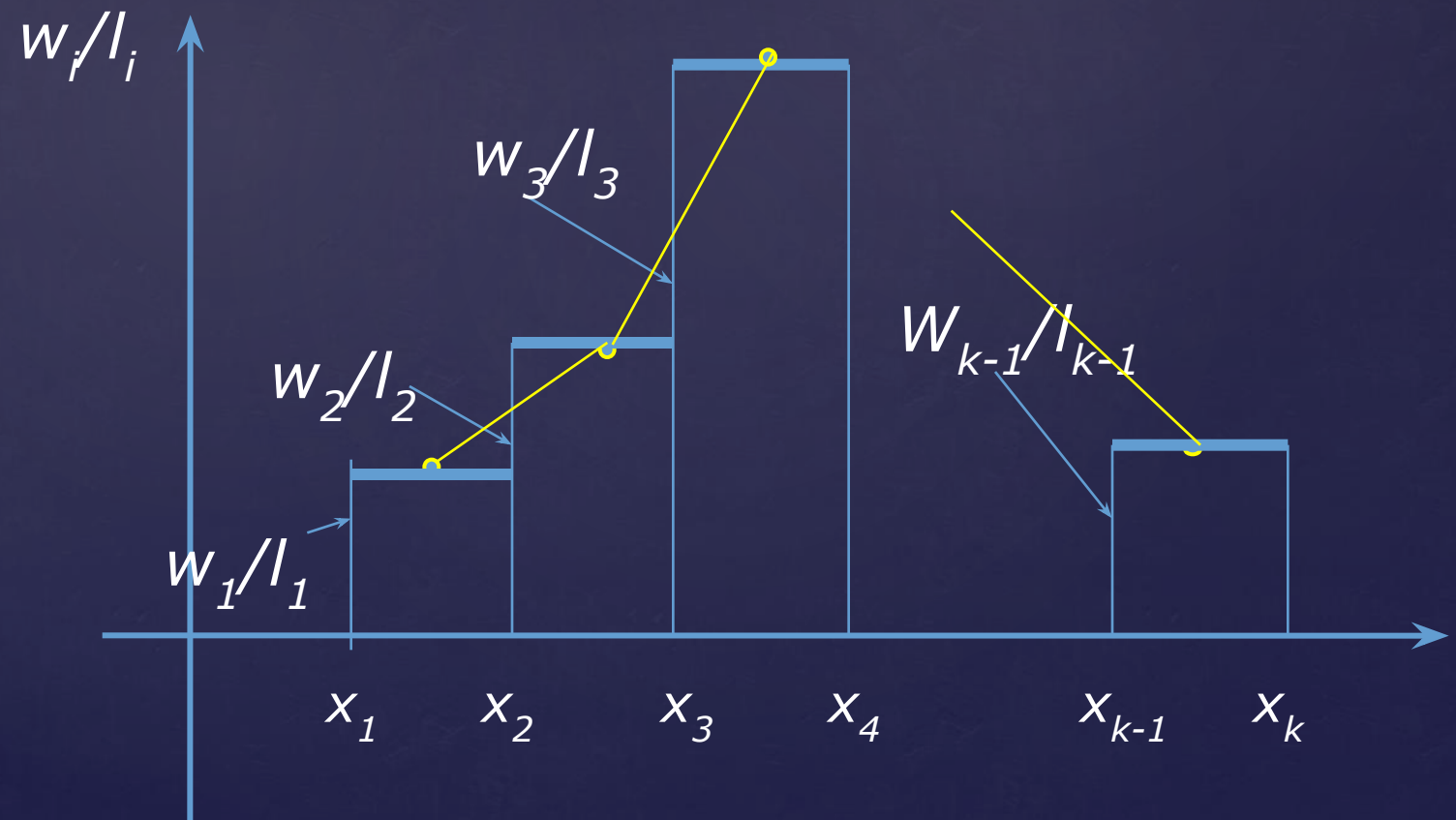
если $m < 6$, то принимают $m = 6$.

Длина отдельного интервала: $l = \frac{R}{m} = \frac{x_{\max} - x_{\min}}{m}$

Существуют различные **приёмы изображения** набора данных, которые дают визуальное представление об основных свойствах экспериментальных данных в целом. Чаще всего для этого используются: полигон, гистограмма, кумулята. Графическое представление вариационных рядов делает картину поведения статистических данных более наглядной.

Полигон распределения частот используется для изображения дискретного вариационного ряда и представляет собой ломаную линию, отрезки которой соединяют точки с координатами (x_i, w_i).

Гистограмма используется для изображения интервальных вариационных рядов и представляет собой ступенчатую фигуру из прямоугольников с основаниями, равными интервалам значений признака l_i ($l_i = x_{i+1} - x_i$) и высотами, равными w_i/l_i.



Эмпирической функцией распределения $F_n(x)$ называется относительная частота того, что случайная величина принимает значение меньше заданного:

$$F_n(x) = W(X<x) = W_x^{нак}$$

Для графического изображения эмпирической функции распределения служит кумулята. Строим ее, соединяя точки $(x_i, W_i^{нак})$.

Следует дополнить вариационные ряды и их графическое изображение некоторыми сводными характеристиками вариационных рядов.

Эти обобщающие показатели в компактном виде характеризуют всю выборку (вариационный ряд) в целом. К таким обобщающим показателям относят:

)Характеристики центральной тенденции - это средние величины, определяющие значения признака, вокруг которого концентрируются все его наблюдаемые значения;

)Характеристики вариации (изменчивости) – это величины, определяющие колебания наблюдаемых значений признака.

В качестве основной характеристики центральной тенденции чаще всего используют среднее арифметическое, вычисленной на основе выборки. Помимо этой величины используют моду и медиану.

Определение:

Медиана – это значение признака, приходящееся на середину ранжированного ряда наблюдений.
Иначе: это то значение варианта, которое делит вариационный ряд на две равные по объему части.

Обозначение:
Теоретическое $MeX;$

Статистическое $\tilde{Me}$ Me

Если число вариант нечетное, т.е. n=2m+1 , то $\tilde{Me} = x_{m+1}$

Если число вариант четное, т.е. n=2m , то $\tilde{Me} = (x_m + x_{m+1})/2$

Определение:

Модой называется значение признака, наиболее часто встречающееся в выборке.

Иначе:

Мода - то значение варианта, которому соответствует наибольшая частота.

Обозначение:

Теоретическое Mo_X;

Статистическое $\tilde{Mo}$

Нам важно знать не только средние значения вариантов, но и отличие значений вариантов от среднего значения. Для отражения изменчивости (вариации) значений признака вводят различные показатели вариации ряда.

Простейшим и весьма приближенным показателем вариации является размах выборки $R = x_{max} - x_{min}$.

Определение.

Выборочной дисперсией вариационного ряда называется среднее арифметическое квадратов отклонений вариантов от их среднего арифметического:

$$S_*^2 = \sum_{i=1}^{n} \frac{\left(x_i - \overline{x}\right)^2}{n} = \sum_{i=1}^{k} \frac{\left(x_i - \overline{x}\right)^2 \cdot n_i}{n} = \sum_{i=1}^{k} \left(x_i - \overline{x}\right)^2 \cdot w_i$$

При вычислении выборочной (или эмпирической) дисперсии формулу несколько меняют. Из некоторых соображений, которые пока для нас с вами скрыты, в знаменателе этой формулы ставят не *n*, а *n-1*, и возникает другая формула для вычисления дисперсии, которую запишем ниже; величину, вычисленную по этой формуле называют «исправленная выборочная дисперсия».

$$S^2 = \sum_{i=1}^{n} \frac{\left(x_i - \overline{x}\right)^2}{n-1} = \sum_{i=1}^{k} \frac{\left(x_i - \overline{x}\right)^2 \cdot n_i}{n-1}$$

Будем всегда выборочную дисперсию вычислять по второй формулу, называя ее просто «выборочная дисперсия». Ясно, что при большом объеме выборки разница между двумя приведенными формулами стирается.

Для меры вариации, выраженной в тех же единицах измерения, что и значение признака, вычисляют выборочное стандартное отклонение:

$$S = \sqrt{S^2} = \sqrt{\sum_{i=1}^{k} \frac{\left(x_i - \bar{x}\right)^2 \cdot n_i}{n-1}}$$

Для сравнения вариаций разных по природе переменных используется относительный показатель вариации:

$$\textit{Коэффициент вариации} \quad V = \frac{S}{\bar{x}} \cdot 100\%$$

Эта величина характеризует, насколько сильно элементы в выборке и, следовательно, в ГС отличаются друг от друга.

§ 2. Точечные оценки параметров генеральной совокупности.

Поставим задачу в общем виде – задачу отыскания хороших (доброкачественных) приближений параметров известных распределений на основе выборки из ГС.

Пусть $x_1, x_2, \ldots, x_n$ - выборка объема n из ГС. Будем рассматривать эту выборку как систему СВ $X_1, X_2, \ldots, X_n$, которая в данном конкретном исследовании приняла именно этот набор числовых значений $x_1, x_2, \ldots, x_n$.

Определение:

Точечной оценкой $\tilde{\theta}_n$ неизвестного параметра θ теоретического закона распределения называют всякую функцию результатов наблюдений над СВ X, значение которой принимают в качестве приближённых значений параметра θ :

$$\tilde{\theta}_n = f(x_1, x_2, \ldots, x_n)$$

Требования, предъявляемые к точечным оценкам
(Иногда говорят : *свойства точечных оценок):*

1.Несмещённость.

Оценка $\tilde{\theta}_n$ параметра θ называется **несмещённой**, если её математическое ожидание равно оцениваемому параметру:

$$E\tilde{\theta}_n = \theta$$

2. Эффективность.

Оценка $\tilde{\theta}_n$ параметра θ называется **эффективной**, если она имеет наименьшую дисперсию среди всех оценок параметра по выборкам одного и того же объема:

$$D\tilde{\theta}_n \to \min \quad при \quad фиксир. \quad значении \quad n.$$

3. Состоятельность.

Оценка $\tilde{\theta}_n$ параметра θ называется **состоятельной**, если она удовлетворяет ЗБЧ:

$$\tilde{\theta}_n \xrightarrow[n\to\infty]{P} \theta$$

В последнее время стали добавлять еще одно требование к оценкам.

4. Устойчивость.

Смысл этого свойства в том, что при небольших флуктуациях в исходной информации значение оценки не должно существенным образом меняться.

На практике не всегда удается удовлетворить всем требованиям одновременно. *Может оказаться, что для простоты расчетов целесообразно использовать незначительно смещенные оценки или же оценки, обладающие несколько большей дисперсией по сравнению с эффективными оценками.*

Показано, что среднее арифметическое, вычисленное на основе выборки и являющееся точечной оценкой генерального среднего (истинного значения параметра), обладает свойствами 1-4, присущими хорошей оценке.

Показано также, что **выборочная доля *w=k/n*** (иначе: относительная частота появления признака в выборке) является несмещенной и состоятельной оценкой генеральной доли $\mathbf{W_г=K/N}$.

Заметим, что выборочную долю можно трактовать как оценку вероятности в биномиальном законе распределения.

Показано, что выборочная дисперсия, вычисляемая по формуле

$$S^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1},$$

дает несмещенную оценку генеральной дисперсии.

Аналогично, несмещенной точечной оценкой ковариации *cov(X,Y)* является такая оценка:

$$K_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right) - \textit{выборочная ковариация.}$$

В формулах для S^2 и K_{XY} возникает новый параметр **k=n-1**

Он носит название *«число степеней свободы».*Это разность между числом используемых в расчетах отклонений и количеством связей между этими отклонениями.

§ 3. Методы получения точечных оценок параметров генеральной совокупности.

Основное внимание уделим методу, который наиболее часто применяется для этой цели.

1. Метод наибольшего (максимального) правдоподобия.

- это основной метод получения оценок параметров ГС на основе выборки. Метод был предложен американским статистиком Р. Фишером.

Пусть задан известный закон распределения. Ставится задача найти оценку его неизвестного параметра или параметров, если в законе распределения их несколько.

Функцией правдоподобия дискретной СВ X называют функцию аргумента θ (искомого параметра)

$$L(x_1, x_2, \ldots, x_n; \theta) = p(x_1; \theta) \cdot p(x_2; \theta) \cdot \ldots \cdot p(x_n; \theta), \quad где$$

$$x_1, x_2, \ldots, x_n - фиксированные \quad числа.$$

В качестве точечной оценки параметра θ принимают такое его значение $\tilde{\theta}_n$, при котором функция правдоподобия достигает максимума. Оценку $\tilde{\theta}_n$ называют оценкой наибольшего правдоподобия.

Суть подхода заключается в том, чтобы выбрать такое значение оценки параметра, которое обеспечивает наиболее вероятное появление именно данной выборки.

Удобнее рассматривать не саму функцию *L*, а *lnL.*

Методом наибольшего правдоподобия найдена оценка параметра λ в законе распределения Пуассона

$$P(X=k)=\frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Методом наибольшего правдоподобия найдена оценка вероятности успеха в единичном испытании на основе единственной серии испытаний.

Методом наибольшего правдоподобия найдена оценка вероятности успеха в единичном испытании на основе нескольких серий испытаний (биномиальный закон распределения).

Функцией правдоподобия непрерывной СВ X называют функцию аргумента θ (искомого параметра)

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

Здесь x_1, x_2, …, x_n - фиксированные числа.

Методом наибольшего правдоподобия найдена оценка параметра λ показательного з.р.

Методом наибольшего правдоподобия найти оценки параметров m и σ нормального з.р.

По поводу метода наибольшего правдоподобия сделаем **выводы:**

1. Метод наибольшего правдоподобия дает естественные оценки, не противоречащие здравому смыслу.

Усилиями математиков было показано, что в целом эти оценки обладают хорошими свойствам. А именно, они являются состоятельными, эффективными, но иногда слабо смещенными.

2. Метод наибольшего правдоподобия имеет два недостатка:

1) иногда сложно решить уравнение или систему уравнений правдоподобия, которые часто бывают нелинейными.

2) существенное ограничение метода – необходимо точно знать вид закона распределения, что во многих случаях оказывается невозможным.

Существует и другие методы нахождения точечных оценок параметров ГС. Это – **Метод моментов** и

Метод наименьших квадратов.

Суть его заключается в том, что оценка определяется из условия минимизации квадратов отклонений выборочных данных от определяемой оценки.

Следует ввести дополнительные распределения и новые таблицы, созданные на основе этих распределений.

§ 4. Распределения, связанные с нормальным законом распределения.

.Распределение X - квадрат (X^2).

(или распределение Пирсона)

Определение:

Пусть СВ X_1, X_2, …, X_k независимые и каждая из них имеет стандартное нормальное распределение

($X_i \sim N(0;1)$, $i=1, 2,\ldots, n$), тогда случайная величина

$$X^2(k) = X_1^2 + X_2^2 + \ldots + X_k^2$$

имеет распределение хи-квадрат с k степенями свободы.

Значения этого распределения затабулированы.

2. t -распределение *(или распределение Стьюдента)*

Определение:

Пусть СВ Y, X_1, X_2, …, X_k независимые и каждая из них имеет стандартное нормальное распределение

$(Y, X_i \sim N(0;1), \quad i=1, 2,\ldots, k)$,

тогда случайная величина

$$t(k) = \frac{Y}{\sqrt{\frac{1}{k}\left(X_1^2 + X_2^2 + \ldots + X_k^2\right)}} = \frac{Y}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} X_i^2}} = \frac{Y}{\sqrt{\frac{1}{k}\chi^2(k)}}$$

имеет распределение Стьюдента с k степенями свободы.

Значения распределения затабулированы.

§ 5. Интервальные оценки параметров генеральной совокупности.

Наша задача - научиться отыскивать границы интервала, который накроет истинное значение искомого параметра. Для этого будем использовать метод интервального оценивания, который разработал американский статистик Нейман, исходя из идей статистика Фишера. Этот интервал должен накрывать истинное значение параметра θ с большой вероятностью $\gamma = 1-a$, где γ - велико, а a - мало;

γ называется доверительной вероятностью (а также: надежностью, уровнем доверия), a называется уровнем значимости.

Интервал, который мы будем находить, носит название доверительного интервала (иначе: интервальная оценка искомого параметра ГС).

Ставится задача отыскания такого значения ε, для которого выполнено:

$$P\left(\left|\theta-\tilde{\theta}\right|<\varepsilon\right)=P\left(\underbrace{\tilde{\theta}-\varepsilon}_{\tilde{\theta}_1}<\theta<\underbrace{\tilde{\theta}+\varepsilon}_{\tilde{\theta}_2}\right)=\gamma$$

$\tilde{\theta}_1=\tilde{\theta}-\varepsilon;\quad \tilde{\theta}_2=\tilde{\theta}+\varepsilon$ –

– границы доверительного интервала;

$I_\gamma=\left(\tilde{\theta}_1;\tilde{\theta}_2\right)$ – доверительный интервал.

Величина ε называется **«точность оценки»** (или: **«предельная ошибка выборки»**).

Формулы, по которым определяются границы доверительного интервала, зависят от конкретного оцениваемого параметра ГС и конкретной ситуации, поэтому возникает необходимость рассмотреть несколько интересующих нас ситуаций.

1. **Интервальная оценка математического ожидания (или: генерального среднего) нормально распределенной ГС, если известна дисперсия σ^2 для ГС.**

Пусть изучаемый признак X в ГС имеет нормальное распределение с параметрами m и σ независимых СВ. В данной постановке задачи считаем, что σ^2 известна (например, взята из аналогичного предыдущего исследования).

Здесь m – тот неизвестный параметр, для которого мы хотим построить интервальную оценку.

Получено следующее выражение для доверительного интервала:

$$I_\gamma = \left(\overline{x} - \varepsilon; \overline{x} + \varepsilon\right), \quad где \quad \varepsilon = \frac{t_{кр} \cdot \sigma}{\sqrt{n}} = t_{кр} \cdot \sigma_{\overline{X}}$$

(С помощью таблицы функции Φ_0 находим по заданному значению γ $t_{кр}$ - квантиль стандартного нормального з.р. на основе уравнения $\Phi_0(t_{кр}) = \gamma/2$).

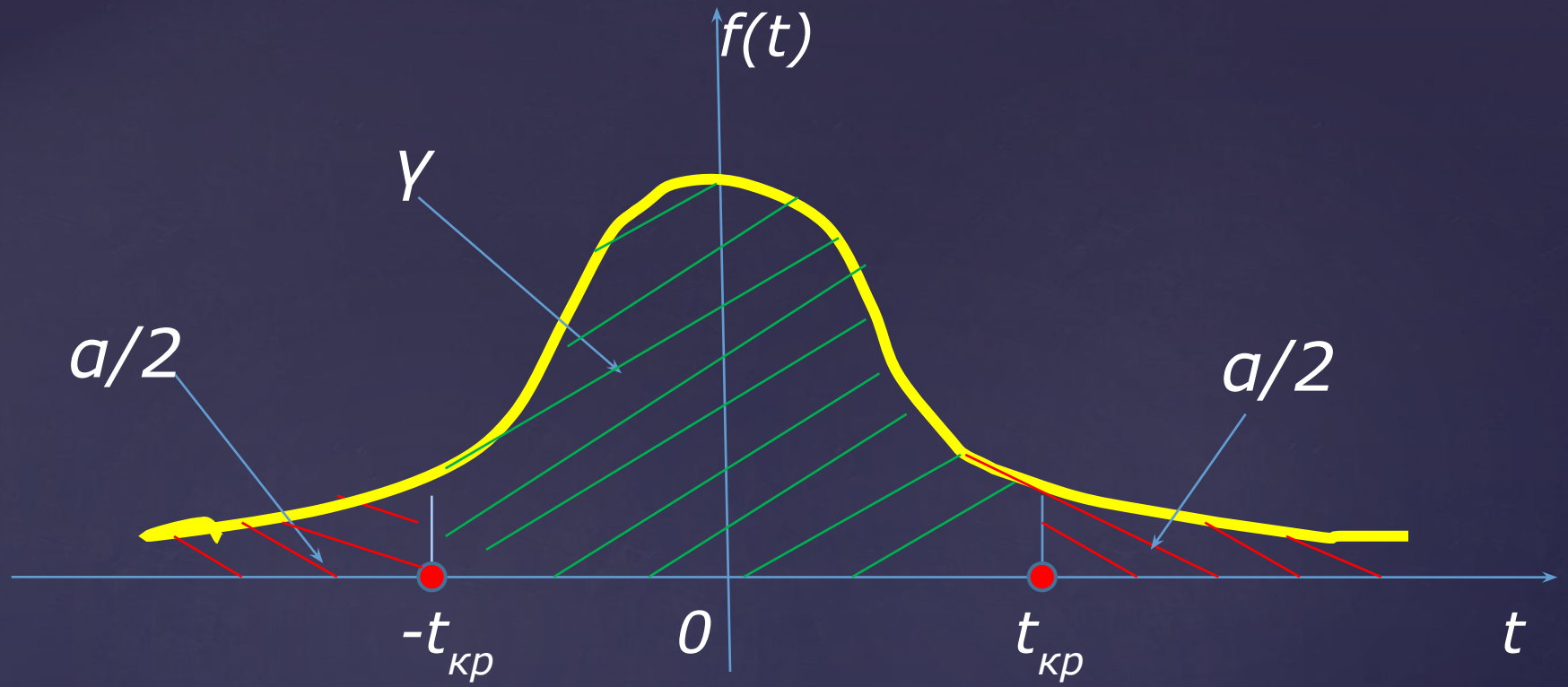
2. Интервальная оценка математического ожидания нормально распределенной ГС, если дисперсия σ^2 для ГС неизвестна.

Теперь вместо неизвестной дисперсии будем использовать ее точечную оценку – выборочную дисперсию

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

$$I_\gamma = \left(\overline{x} - \varepsilon; \overline{x} + \varepsilon\right), \quad где \quad \varepsilon = \frac{t_{кр} \cdot S}{\sqrt{n}} = t_{кр} \cdot S_{\overline{X}}$$

(С помощью таблица **«Критические точки распределения Стьюдента»** по заданным значениям a (двусторонняя критическая область) и $k=n-1$ находим $t_{кр}$ - квантиль распределения Стьюдента).



Замечание:

При $n \leq 30$ (**малые выборки**) следует находить $t_{кр}$ на основе распределения Стьюдента;

При $n > 30$ (**большие выборки**) следует находить $t_{кр}$ на основе стандартного нормального распределения, т.е. на основе функции Лапласа.

Если задана точность оценки ε , то можно найти объем выборки, которая обеспечит эту требуемую точность:

$$n_{\min} = \left(\frac{t_{кр} \cdot S}{\varepsilon} \right)^2; \quad при \quad n \geq n_{\min} \quad эта$$

$$точность \quad будет \quad обеспечена.$$

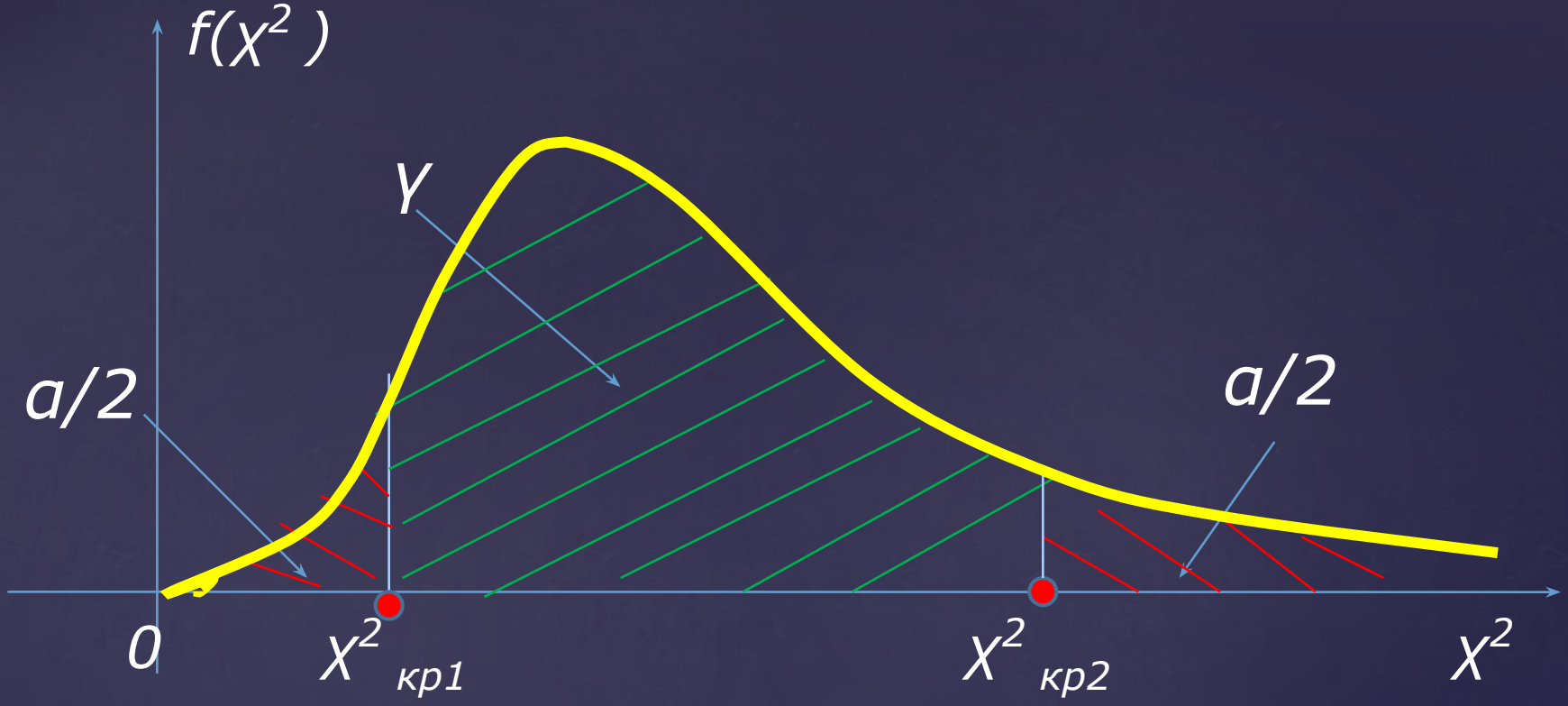
3. Интервальная оценка стандартного отклонения для нормально распределенной ГС.

Пусть изучаемый признак X в ГС имеет нормальное распределение: $X \sim N(m,\sigma)$, причем параметры распределения неизвестны.

Для случая малых объемов выборки ($n \leq 30$):

$$Доверительный \quad интервал \quad для \quad \sigma \quad имеет \quad вид:$$

$$I_\gamma = \left(S \cdot \sqrt{\frac{(n-1)}{\chi^2_{кр_2}}}; \quad S \cdot \sqrt{\frac{(n-1)}{\chi^2_{кр_1}}} \right)$$



Очевидно, что значения $X^2_{кр1}$ и $X^2_{кр2}$ определяются неоднозначно при одном и том же значении заштрихованной площади, равной γ . Границы красных зон выбираем так, чтобы вероятности попадания в них были бы одинаковыми, равными $a/2$.

Для случая больших объемов выборки (*n>30):*

$$I_\gamma = \left(S \cdot \frac{\sqrt{2(n-1)}}{\sqrt{2 \cdot n - 3} + t_{кр}}; \quad S \cdot \frac{\sqrt{2(n-1)}}{\sqrt{2 \cdot n - 3} - t_{кр}} \right),$$

где $t_{кр}$ *находим из табл. решения уравнения:* $\Phi_0(t_{кр}) = \frac{\gamma}{2}.$

4. Интервальная оценка истинного значения вероятности биномиального закона распределения (генеральной доли).

Рассмотрим два случая:

А. Случай умеренно больших выборок
(n>30 до нескольких сотен, например, до 200).

Далее в формуле $t_{кр}$ - квантиль стандартного нормального з.р. на основе уравнения $\Phi_0(t_{кр}) = \gamma /2.$

$$p_{1,2} = \frac{\left(w + \frac{t_{кр}^{\,2}}{2n}\right) \pm t_{кр} \cdot \sqrt{\frac{w(1-w)}{n} + \frac{t_{кр}^{\,2}}{4n^2}}}{\left(w + \frac{t_{кр}^{\,2}}{n}\right)} =$$

$$= \frac{n}{\left(n + t_{кр}^{\,2}\right)} \cdot \left(w + \frac{t_{кр}^{\,2}}{2n} \pm t_{кр} \cdot \sqrt{\frac{w(1-w)}{n} + \frac{t_{кр}^{\,2}}{4n^2}}\right) \rightarrow$$

$\rightarrow$ доверит. интервал для p находится следующим образом: $p_1 < p < p_2$,

где p_1, p_2 – меньший и больший корни этого уравнения; иначе: $I_\gamma = (p_1; \quad p_2)$.

Б. Случай больших выборок

(порядка сотен и более ; например, от 200 и более).

Формулы для вычисления границ доверительного интервала существенно упрощаются при таких больших объемах выборок.

$$I_\gamma = \left(w - \varepsilon; \quad w + \varepsilon \right), где \quad \varepsilon = t_{кр} \cdot \sqrt{\frac{w(1-w)}{n}} = t_{кр} \cdot S_w$$

При больших объемах выборок n возникает простая формула для ε, на основе которой при заданном ε можно вычислить соответствующее n:

$$n_{\min} = \frac{t_{кр}^2 \cdot w \cdot (1-w)}{\varepsilon^2}.$$

В. Случай выборок малого объема (*n≤30*)

В этом случае для вычисления S_w используется формула

$$S_w = \sqrt{\frac{w \cdot (1 - w)}{n - 1}}$$

Доверительный интервал определяется по формуле предыдущего пункта; $t_{кр}$ находится по распределению Стьюдента по *к=n-1.*

Замечание:

В литературе часто приводят упрощенный способ вычисления доверительного интервала, рассматривая только большие и малые выборки. В этом случае выделяют два пункта при вычислении доверительного интервала:

1) Большая выборка (*n* более 30) - вычисление ведут по пункту Б.
2) Малая выборка (*n* меньше или равно 30) – вычисление ведут по пункту В.

и конкурирующая гипотезы. Критерий. Критические области и область принятия нулевой гипотезы.

Гипотеза – утверждение, которое надо либо доказать, подтвердить, исходя из разумных предположений, либо опровергнуть.

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известного распределения.

Нулевой (основной) называют выдвинутую гипотезу H_0. **Конкурирующей** (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

Статистическим критерием или просто критерием называют случайную величину К, которая служит для проверки нулевой гипотезы H_0.

Областью принятия гипотезы (областью допустимых значений критерия) называют совокупность значений критерия, при которых нулевую гипотезу принимают.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.
Это такие значения критерия, которые не характерны для данного распределения, т.е. возникающие с малой вероятностью.

Основной принцип проверки статистической гипотезы можно сформулировать так: если наблюдаемое значение критерия принадлежит области принятия гипотезы, то принимают нулевую гипотезу; если наблюдаемое значение критерия принадлежит критической области, то нулевую гипотезу отвергают и принимают альтернативную гипотезу;

Гипотеза называется параметрической, если речь идет об утверждении, связанном с каким-то конкретным параметром. В противном случае она называется непараметрической.

Гипотеза называется простой, если речь идет о том, что неизвестный параметр принимает какое-то конкретное значение. Если речь идет о многих значениях параметра, то она называется сложной.

Процедура проверки простой параметрической гипотезы выглядит так:

1. Формируют нулевую гипотезу H_0 и альтернативную гипотезу H_1 на основе выборочных данных.

2. Конструируют, исходя из логики задачи, СВ на основе результатов выборки (критерий); *распределение критерия в случае истинности гипотезы H_0 известно.*

3. Вся область возможных значений критерия разбивается на два подмножества. Одно подмножество – это совокупность естественных (правдоподобных), т.е. наиболее вероятных для данного распределения значений. В это подмножество критерий попадает с высокой вероятностью γ. Эту вероятность мы задаем сами. Она носит название «д**оверительная вероятность**» (уровень доверия) (γ = 0.90; 0.95; 0.99). Другое подмножество – это область редко возникающих для данного з.р. значений (неправдоподобных значений).

Вероятность попадания в эту область мала и равна $\alpha = 1-\gamma$.

α носит название «уровень значимости»($\alpha = 0.10;0.05;0.01$).

4. Вычисляют значение критерия $K_{набл}$ на основе выборочных значений изучаемого признака. Если $K_{набл}$ попадает в область правдоподобных значений, то с вероятностью γ утверждают, что гипотеза H_0 не противоречит экспериментальным данным, а поэтому принимают основную гипотезу.
   Если значения $K_{набл}$ попадает в область неправдоподобных для данного з.р. значений, то отвергают гипотезу H_0 и принимают альтернативную гипотезу H_1.

5. Если при проверке гипотезы H_0 эта гипотеза принимается, то этот факт не означает, что высказанное нами утверждение является единственно верным. Просто оно не противоречит имеющимся выборочным данным. Возможно, что и другое утверждение также не будет противоречить выборочным данным.

6. Если наблюдаемое значение критерия $K_{набл}$ попадает в область неестественных значений и мы, следовательно, отвергаем гипотезу H_0 и принимаем гипотезу H_1, то не можем ли мы при этом совершить ошибку - отвергнуть верную гипотезу H_0 и принять ложную гипотезу H_1? Да, можем, но вероятность этой ошибки мала; она равна величине α .

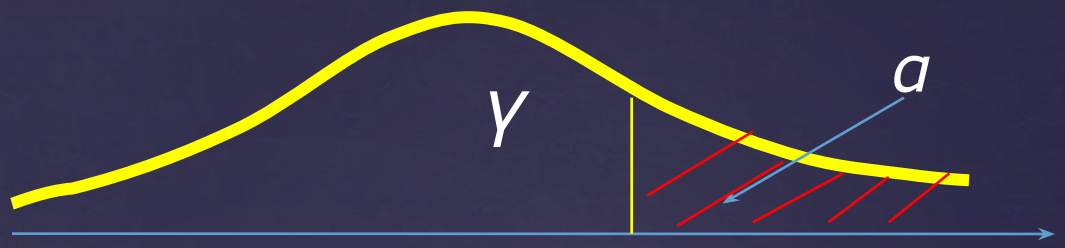
Типы альтернативных гипотез (для исходной простой параметрической гипотезы $\mathbf{H_0 : \theta = \theta_0}$)

1. $H_1: \boldsymbol{\theta \neq \theta_0}$

$\gamma + \alpha = 1$

Двусторонняя критическая область

2. $H_1: \theta > \theta_0$



Правосторонняя критическая область

3. $H_1: \theta < \theta_0$

Левосторонняя критическая область

§ 7. Проверка гипотезы о числовом значении математического ожидания m (генеральной средней $\overline{x_1}$) нормально распределенной ГС.

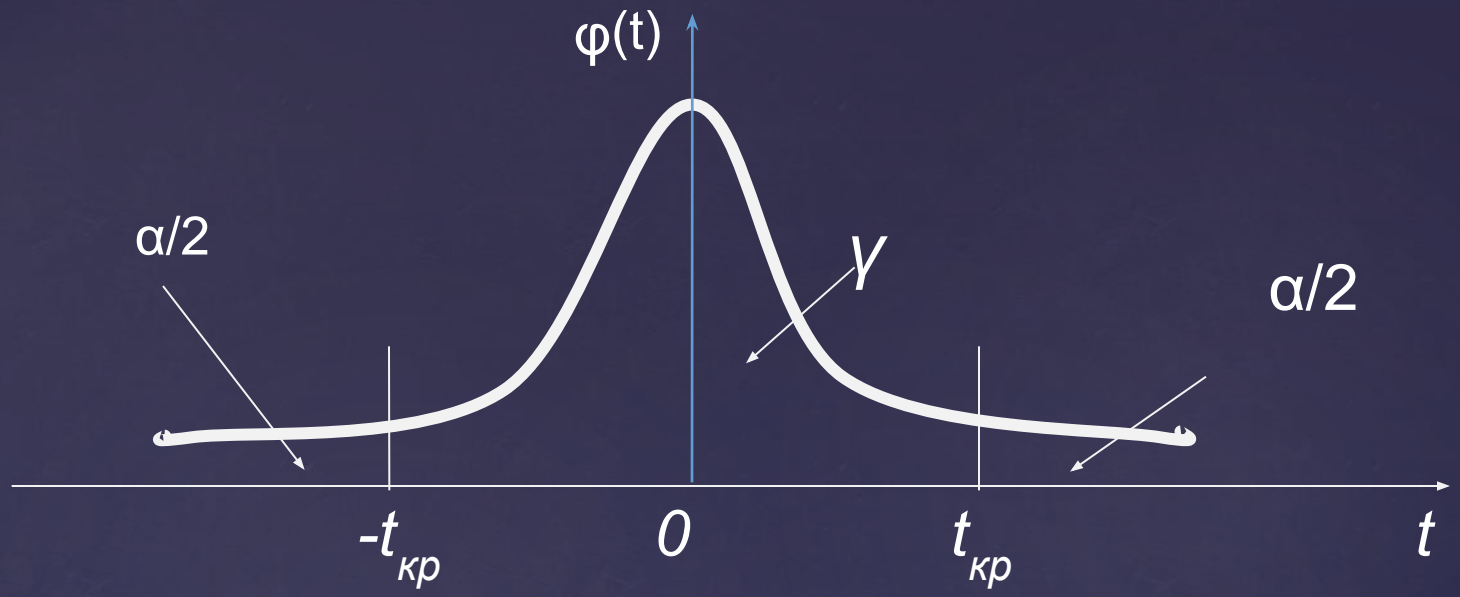
1. *Дисперсия ГС известна* (или n>30)

Считаем, что в ГС изучаемый признак X распределен нормально, причем мат. ожидание неизвестно, но есть основание полагать, что оно равно какому-то определенному значению m_0.

В этом пункте считаем, что дисперсия σ^2 в ГС известна либо из предшествующего опыта, либо же вычислена на основе данного опыта, но по выборке большого объема (по большой выборке можно получить весьма хорошее приближение для истинной дисперсии в ГС на основе рассчитанной по выборке выборочной дисперсии).

Поставим задачу следующим образом:

$H_0: \ m = m_0$



При конкурирующей гипотезе $H_1: \; m \neq m_0$ следует вводить двустороннюю критическую область.

Из условия $P(|t| < t_{кр}) = \gamma = 2\Phi_0(t_{кр})$ с помощь таблиц функции Лапласа находим значение $t_{кр}$.

Здесь введен критерий

$$t = \frac{\left(\overline{X} - m_0\right)\sqrt{n}}{\sigma}$$

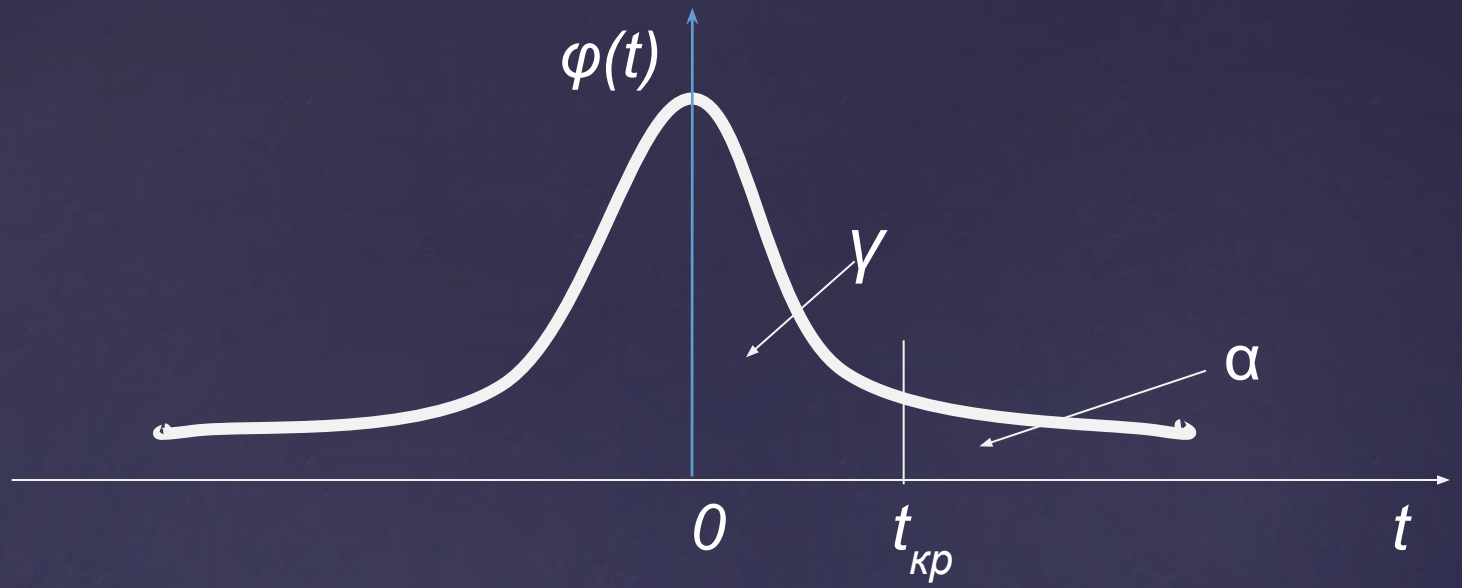
Если окажется, что вычисленное на основе экспериментальных данных значение $t_{набл}$ таково, что $|t_{набл}| < t_{кр}$, то нет оснований отвергнуть гипотезу H_0;

если $|t_{набл}| \geq t_{кр}$, то отвергаем нулевую гипотезу как противоречащую экспериментальным данным и принимаем альтернативную гипотезу H_1.

При иной конкурирующей гипотезе, например,

$$H_1: m > m_0$$

следует формировать правостороннюю критическую область.



Если $t_{набл} < t_{кр}$, то принимается гипотеза H_0;

если $t_{набл} \geq t_{кр}$, то отвергаем нулевую гипотезу и принимаем альтернативную гипотезу H_1.

2. Дисперсия ГС неизвестна

Вычисляем выборочную дисперсию S^2 для аппроксимации значения генеральной дисперсии σ^2 . Формулы полностью сохраняются, только вместо σ используем S и $t_{кр}$ определяем по таблице критических точек распределения Стьюдента для критической области по заданному уровню значимости a и по числу степеней свободы $k=n-1$.

Здесь вводится критерий

$$t = \frac{\left(\overline{X} - m_0\right)\sqrt{n}}{S}$$

3. *Связь между двусторонней критической областью и доверительным интервалом*

Отыскивая двустороннюю критическую область мы проделывали совершенно такие же преобразования как и при нахождении доверительного интервала для математического ожидания.

$$P\left(\left|\overline{X}-m_0\right|<\varepsilon_{кр}\right)=\gamma \rightarrow P\left(|t|<t_{кр}\right)=\gamma, \quad где \quad t=\frac{\left(\overline{X}-m_0\right)\sqrt{n}}{S}; \quad t_{кр}=\frac{\varepsilon\sqrt{n}}{S}$$

$$или \quad P\left(\overline{X}-\frac{t_{кр}\cdot S}{\sqrt{n}}<m_0<\overline{X}+\frac{t_{кр}\cdot S}{\sqrt{n}}\right)=\gamma$$

Область принятия нулевой гипотезы и доверительный интервал совпадают.

Можно сделать следующий ***вывод:***

Если предполагаемое в основной гипотезе числовое значение m_0 неизвестного параметра попадает в доверительный интервал этого параметра, отвечающего заданному уровню доверия γ, то гипотезу H_0 принимаем, в противном случае ее отклоняем и принимаем альтернативную гипотезу H_1.

§ 8. Проверка гипотезы о числовом значении вероятности p биномиального закона распределения (о числовом значении генеральной доли W)

Требуется при заданном уровне доверия γ проверить нулевую гипотезу $H_0: p = p_0$

Альтернативная гипотеза может быть трех видов

$$H_1: p \neq p_0 \quad (p < p_0;\ p > p_0)$$

Здесь мы будем рассматривать только случай умеренно больших (от 30 до нескольких сотен) и больших (более нескольких сотен) выборок, т.е. n>30.

Используется критерий $t = \dfrac{w - p_0}{\sqrt{\dfrac{p_0 \cdot q_0}{n}}}$ $\quad P(|t| < t_{кр}) = 2\Phi_0(t_{кр}) = \gamma$

математических ожиданий (генеральных средних) двух нормально распределенных ГС.

Пусть имеются две нормально распределенные ГС, причем в первой совокупности изучаемый признак $X \sim N(m_1;\sigma_1)$, во второй совокупности изучаемый признак $Y \sim N(m_2;\sigma_2)$.

Предположим, что m_1 и m_2 неизвестны, а σ_1 и σ_2 известны (значения стандартных отклонений взяты либо из предшествующего опыта, либо при больших выборках получены на основе этих же выборок, поскольку хорошо аппроксимируют значения стандартных отклонений в ГС).

Проверим гипотезу

$$H_0: m_1 = m_2$$

$$H_1: : m_1 \neq m_2 \quad (m_1 < m_2 \text{ или } m_1 > m_2)$$

Подчеркнем: мы в данной формуле берем значения σ_1 и σ_2 либо из предыдущего опыта (и тогда нет ограничений на величины объемов выборок), либо получаем на основе выборок из данного опыта, но при этом полагаем, что выборки большие, т. е. $n_1>30$, $n_2>30$. Используется такой критерий:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Далее в конкретных примерах в зависимости от конкурирующих гипотез выстраивают критические области, вычисляют наблюдаемое значение критерия и смотрят, попадает ли это значение в область правдоподобных значений критерия при справедливости нулевой гипотезы или же, напротив, в область неправдоподобных значений критерия. И в зависимости от этого принимают или же отвергают нулевую гипотезу, т.е. реализуют обычный алгоритм проверки гипотезы.

§10. Проверка гипотезы о равенстве вероятностей биномиального закона распределения (о равенстве долей признака) двух генеральных совокупностей.

Рассмотрим две ГС.

Из первой ГС делается случайная выборка объемом n_1, и на основе этой выборки выясняется, сколько объектов выборки обладает изучаемым признаком – этих объектов k_1. Из второй ГС делается случайная выборка объемом n_2; количество объектов выборки, обладающих изучаемым признаком, - k_2.

Выборочные доли признака равны соответственно

$w_1 = k_1 / n_1$; $w_2 = k_2 / n_2$.

В данном пункте мы ограничимся лишь случаем, когда выборки достаточно большие : $n_1 > 30$, $n_2 > 30$.

Сформулируем задачу:

Имеются две ГС, вероятности проявления признака (генеральные доли) в которых равны соответственно p_1 и p_2. Необходимо проверить нулевую гипотезу о равенстве вероятностей (генеральных долей):

$$H_0: \quad p_1 = p_2; \quad (иначе: \quad W_{Г_1} = W_{Г_2})$$

$$H_1: \quad p_1 \neq p_2; \quad (иначе: \quad W_{Г_1} \neq W_{Г_2})$$

Могут быть поставлены и другие альтернативные гипотезы: $p_1 > p_2$; $p_1 < p_2$.

В качестве критерия используется случайная величина:

$$t = \frac{w_1 - w_2}{\sqrt{\tilde{p} \cdot (1 - \tilde{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

В качестве неизвестного значения вероятности p, входящего в выражение критерия t, берут ее наилучшую оценку:

$$\tilde{p} = \frac{k_1 + k_2}{n_1 + n_2}.$$

$t_{кр}$ находится на основе функции Лапласа.

§11. Проверка гипотезы о значимости выборочного коэффициента корреляции Пирсона.

Рассматривается двумерная нормально распределенная генеральная совокупность (X,Y), т.е. случайные величины X и Y в ней распределены нормально. Из этой совокупности извлечена выборка объемом n пар (x_i , y_i) и по ней вычислен выборочный коэффициент корреляции Пирсона, который оказался отличным от нуля.

На основе выборочных данных мы бы хотели получить обоснованный вывод о наличии связи между изучаемыми признаками во всей ГС.

Всегда проверяется нулевая гипотеза об отсутствии линейной корреляционной связи в ГС, а альтернатива заключается в предположении о том, что этот коэффициент в ГС отличен от нуля:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Если нулевая гипотеза отвергается, то это означает, что выборочный коэффициент корреляции значимо отличается от нуля, и, следовательно, в ГС признаки X и Y связаны линейной зависимостью. Если же принимается нулевая гипотеза, то выборочный коэффициент корреляции незначим, и, следовательно, признаки X и Y в ГС не связаны линейной зависимостью.

В качестве критерия проверки нулевой гипотезы используется случайная величина

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

Показано, что эта СВ при справедливости нулевой гипотезы имеет распределение Стьюдента с $k=n-2$ степенями свободы.

Ясно также, что при больших объемах выборки ($n>30$) можно вместо распределения Стьюдента использовать стандартный нормальный з.р.

Поскольку конкурирующая гипотеза имеет вид $\rho \neq 0$, то следует строить двустороннюю критическую область.

Определив, куда попадает вычисленное значение $t_{набл}$, делаем вывод о справедливости нулевой или же альтернативной гипотезы:

Если $|t_{набл}| < t_{кр}$, то принимается гипотеза H_0,

Если $|t_{набл}| \geq t_{кр}$, то принимается гипотеза H_1.

Проверка гипотезы о значимости выборочного коэффициента корреляции Спирмена

При проверке коэффициента корреляции Спирмена поступают совершенно аналогично тому, как мы поступали, работая с коэффициентом Пирсона.

$H_0: \quad \rho_{S_\Gamma} = 0$

$H_1: \quad \rho_{S_\Gamma} \neq 0;$

через ρ_{S_Γ} обозначаем ранговый коэффициент корреляции во всей ГС;

через ρ_S обозначаем коэффициент корреляции Спирмена, вычисленный по выборке (иначе: выборочный ранговый коэффициент корреляции)

Если объем выборки совсем маленький ($n<9$), то для выяснения значимости коэффициента корреляции нужны специальные таблицы, которые приводятся в специальных руководствах (этот случай мы рассматривать не будем).

Если объем выборки $n \geq 9$, то при справедливости гипотезы H_0 критерий

$$t = \frac{\rho_S \cdot \sqrt{n-2}}{\sqrt{1-\rho_S{}^2}}$$

имеет распределение Стьюдента с $k=n-2$ степенями свободы.

$t_{кр}$ находим по таблице критических точек распределения Стьюдента по значениям a и k для двусторонней критической области. Вычисляем $t_{набл}$ на основе приведенной выше формулы.

Если $|t_{набл}| < t_{кр}$, то принимается гипотеза H_0,

Если $|t_{набл}| \geq t_{кр}$, то принимается гипотеза H_1, т.е. коэффициент корреляции является значимым и в ГС между качественными признаками имеется корреляционная связь.

При объеме выборки больше 30 следует вместо распределения Стьюдента перейти с стандартному нормальному з.р.

Критерий знаков не связан с заданием каких-то конкретных значений параметров распределения, и поэтому на основе этого критерия формулируются так называемые **непараметрические статистические гипотезы.**

Это самый простой критерий непараметрической статистики. Простота критерия объясняется двумя причинами:

- Не делается предположение о том, что ГС имеет нормальное распределение или какое-то другое распределение. Единственное предположение – распределение должно быть **непрерывным.**

§12. Критерий знаков.

- Критерий знаков использует только знаки различий между двумя числами, а не их количественную меру. Поэтому иногда его называют **«ранговый критерий проверки гипотез».**

Пусть имеются две выборки **одинакового объема** n, и эти выборки проранжированы:

$$x_1<x_2<\ldots<x_n \quad \text{и} \quad y_1<y_2<\ldots<y_n$$

Введем разность $r_i=x_i-y_i$.

Исследуем знаки разностей r_i и найдем число положительных разностей (это для нас успех), т.е. найдем ***число успехов***, которое обозначим величиной ***k***.

В случае справедливости нулевой гипотезы о том, что выборки извлечены из совпадающих генеральных совокупностей (или из одной и той же ГС), положительные и отрицательные разности r_i будут появляться с одинаковой вероятностью.

Задание гипотезы H_0 возможно в и других форматах, например, $P(x-y>0)=P(x-y<0)=½$;
или проверить, равны ли друг другу генеральные средние

$$\bar{x}_Г = \bar{y}_Г$$

Если разность r_i окажется равной нулю, то ее исключают из рассмотрения.

При справедливости гипотезы H_0 *k – дискретная случайная величина, распределенная по биномиальному з.р.* с параметрами *n* и *p=1/2*, причем *n* - число отличных от нуля разностей:

$$P_n(k) = C_n^k \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{n-k} = \left(\frac{1}{2}\right)^n \cdot C_n^k, \quad где$$

$$Ek = np = \frac{n}{2}; \quad Dk = npq = \frac{n}{4}.$$

Критическая область строится в зависимости от альтернативной гипотезы, а вид альтернативной гипотезы связан с данными конкретной рассматриваемой задачи.

Алгоритм реализации критерия знаков таков:

1. Рассматривают серию из n испытаний и подсчитывают число положительных и отрицательных разностей r_i, нулевые разности исключаются из рассмотрения, выясняют число положительных разностей (число успехов k).

2. Для получения выводов используется критерий следующего вида:

$$W(n,k) = \frac{1}{2^n} \cdot \sum_{i=o}^{k} C_n^i$$

Понятно, что $W(n,0) \approx 0$, а $W(n,n)=1$.

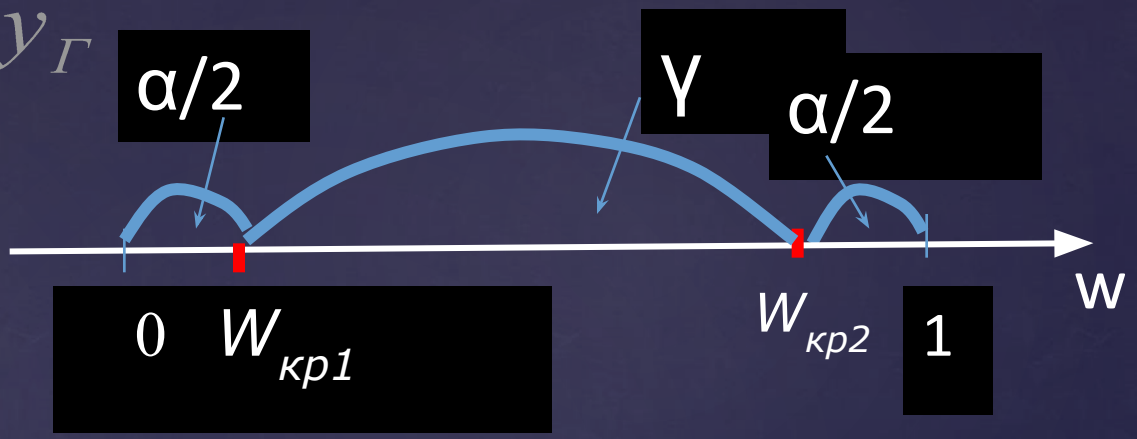
3. На основе свойств биномиальных коэффициентов для облегчения вычислений можно использовать равенство

W(n;k) = 1 – W(n; n-k-1).

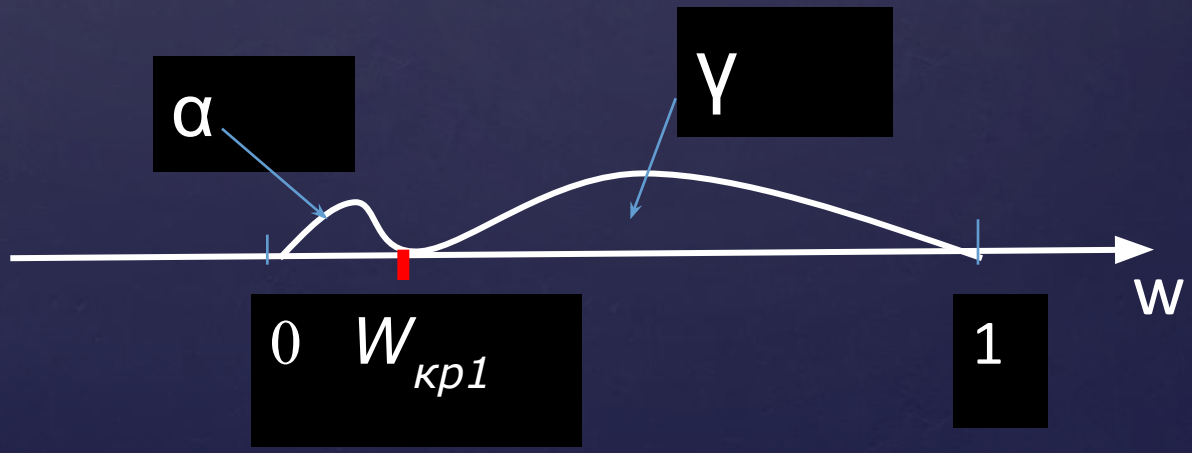
Это равенство удобно использовать, когда $k>n/2$.

4. Критические области для значений критерия связаны с видом альтернативной гипотезы H_1:

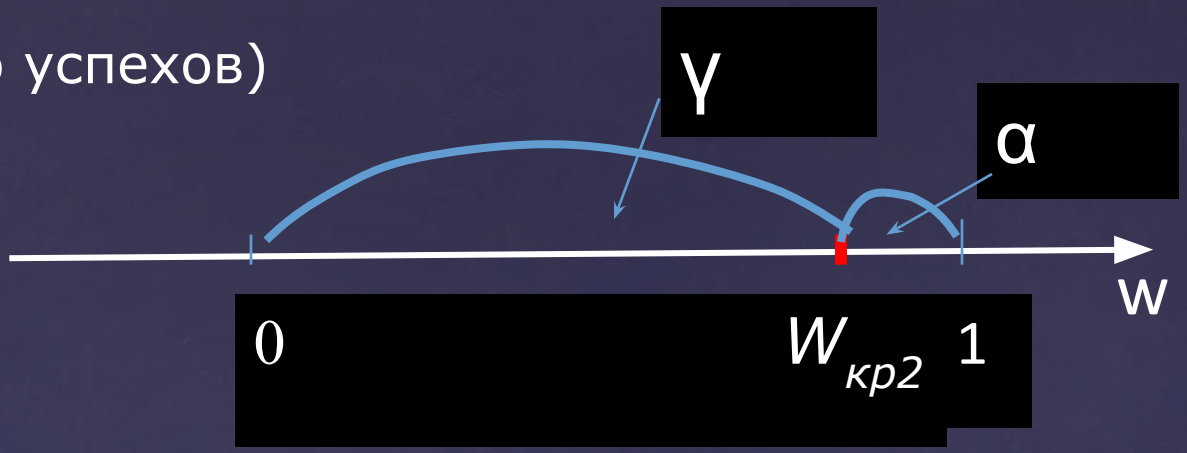
а) $H_1: \quad \bar{x}_Г \neq \bar{y}_Г$



б) $H_1: \quad \bar{x}_Г < \bar{y}_Г$ → положительных разностей мало (мало успехов)



в) $H_1: \quad \bar{x}_Г > \bar{y}_Г$ → положительных разностей много
(много успехов)



4. Вычисление критерия *W(n;k)* проводят при малых выборках (*n≤30*). При больших выборках (*n>30*) биномиальный з.р. переходит в нормальный з.р. , поэтому при *n>30* обычно вводят иной критерий, ибо вычисления по нему существенно упрощаются. Этот критерий t при справедливости гипотезы H_0 имеет стандартный нормальный з.р.:

$$t = \frac{w - p_0}{\sqrt{\frac{p_0 q_0}{n}}}, \quad где \quad w = \frac{k}{n}, \quad p_0 = \frac{1}{2}.$$

§ 13. Шкалы измерений признаков.

Ранее были рассмотрены признаки, измеряемые в количественных шкалах - в этом случае для выяснения тесноты связи между признаками был использован коэффициент корреляции Пирсона, а также признаки, измеренные в шкале порядков - был использован коэффициент корреляции Спирмена.

До сих пор не рассматривались ситуации, когда возникает необходимость изучить связи таких признаков, как профессия, и, допустим, политические убеждения, или уровень образования и политические убеждения, и тому подобное.

Возникает новое понятие номинальных признаков и номинальных (неметрических) шкал измерений.

В этом случае объекты группируются по различным классам так, чтобы внутри класса они были идентичны по измеряемому свойству. Следует научится выявлять наличие или же отсутствие связи между номинальными признаками и научиться количественно оценивать тесноту связи между ними, если она будет выявлена.

Предположим, что признаки статистически независимы, тогда введем две гипотезы:

H_0: признаки независимы

H_1: признаки зависимы

Рассмотрен конкретный пример, в котором для простоты ограничились лишь двумя признаками:

Признак А имеет **r**=2 уровня.

Признак В имеет **s**=3 уровня.

§ 14. Связь номинальных признаков (таблицы сопряженности)

B(s)

A \ B	B_1	B_2	B_3	Итого
A_1	42 n_{11}	66 n_{12}	28 n_{13}	$n_{1\bullet}$ =136
A_2	8 n_{21}	14 n_{22}	42 n_{23}	$n_{2\bullet}$ =64
итого	$n_{\bullet 1}$=50	$n_{\bullet 2}$=80	$n_{\bullet 3}$=70	n=200

A (r)

Возникла таблица 2×3.

Она называется таблицей сопряженности признаков А и В.

Введем обозначения:

i - номер строки ($i=1,2,\ldots,r$)

j- номер столбца ($j=1,2,\ldots,s$)

n_{ij} - частота события $A_i \cap B_j$ – это количество объектов, обладающих комбинацией уровней A_i и B_j признаков A и B.

Через • будем обозначать суммирование по соответствующему признаку, тогда

Определение.

Величины
называются **ожидаемыми** или **теоретическими частотами**
(имеется в виду ожидаемыми при выполнении гипотезы H_0)

При выполнении гипотезы H_0 ожидаемые частоты не должны сильно отличаться от наблюдаемых частот n_{ij} .

Если равенства (*) примерно выполняются, то гипотезу H_0 можно признать справедливой.

Если же равенства (*) плохо выполняются, то гипотезу H_0 отвергаем, т.е. отвергаем утверждение о независимости признаков и признаем справедливой альтернативную гипотезу H_1: признаки зависимые.

Сопоставим наблюдаемые ***H*** и теоретические частоты ***T:***
Мера согласия опытных данных с теоретической моделью:

Суммы берется по всем ячейкам таблицы сопряженности.

Для ответа на вопрос, что такое большое значение случайной величины X^2, надо знать распределение этой СВ. Ответ на этот вопрос дает следующая теорема:

Теорема (К. Пирсон, Р. Фишер):

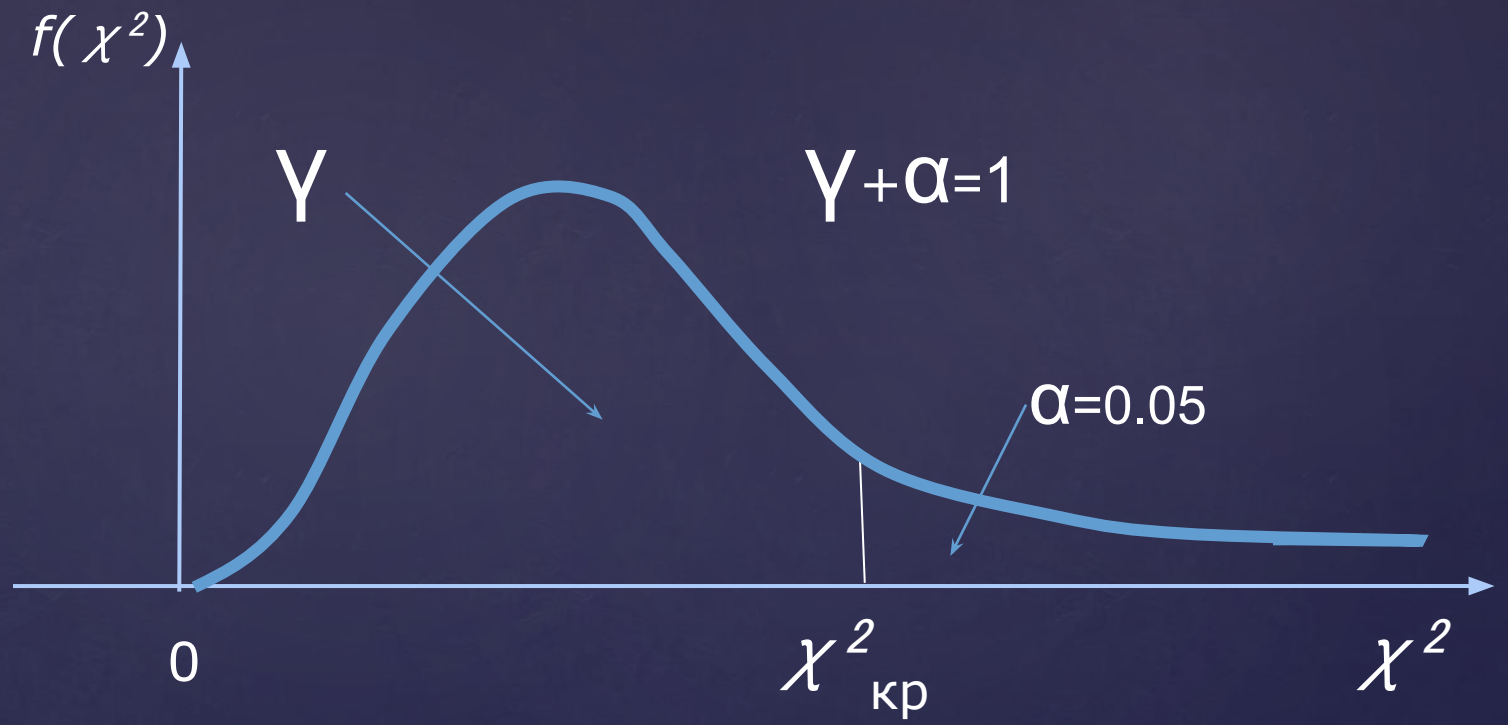
Если справедлива гипотеза H_0, на основе которой рассчитаны теоретические частоты T, то при неограниченном росте числа наблюдений n распределение СВ X^2 стремится к распределению χ - квадрат (χ^2).

Число степеней свободы этого распределения равно разности между числом событий и числом связей между n_{ij}, заложенных в таблице сопряженности.

Число степеней свободы:

Как было сказано, распределение χ^2 является предельным для СВ X^2 , поэтому использовать его как приближение для реальных распределений X^2 можно только при большом числе наблюдений n . Считается достаточным для возможности заменить распределение СВ X^2 распределением СВ χ^2 выполнение следующего **ограничения:** для каждой ячейки теоретические частоты должны быть не меньше 5:

Значения X^2 считаются настолько большими, если они превосходят критические значения распределения χ^2, соответствующие выбранному уровню значимости.



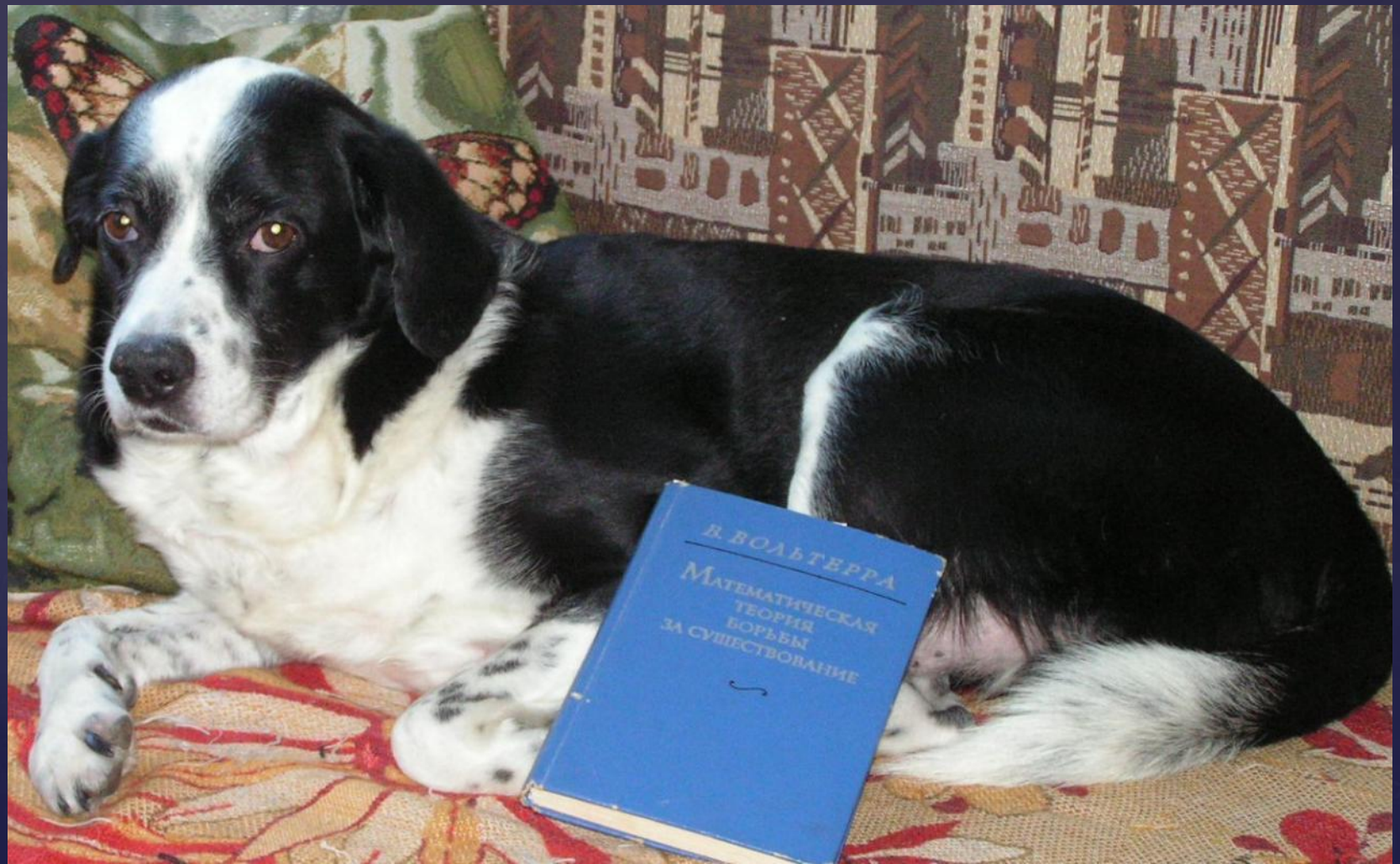
Здесь всегда по смыслу рассматривается правосторонняя критическая область, т.к. если нулевая гипотеза неверна, то X^2 принимает большое значение и, следовательно, χ^2 также принимает большое значение.

Коэффициенты для вычисления тесноты связи между номинальными признаками:

.Коэффициент «фи»

.Коэффициент взаимной сопряженности Пирсона

Благодарю за внимание! Желаю удачи в написании итоговой контрольной работы !!!!!!!!!!!



Благодарю за внимание!