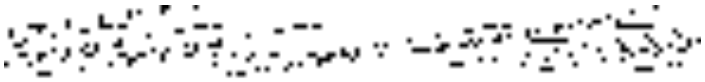
Реферат по курсу математической статистики и теории вероятности

Подготовил: Шевченко Остап
103гр

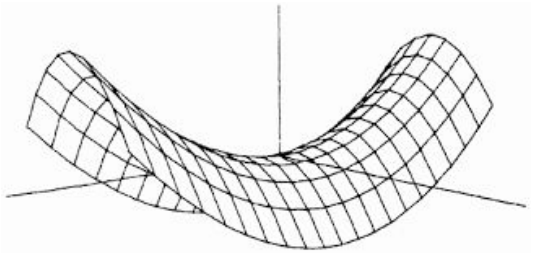
Понятие о совместной функции распределения случайных величин

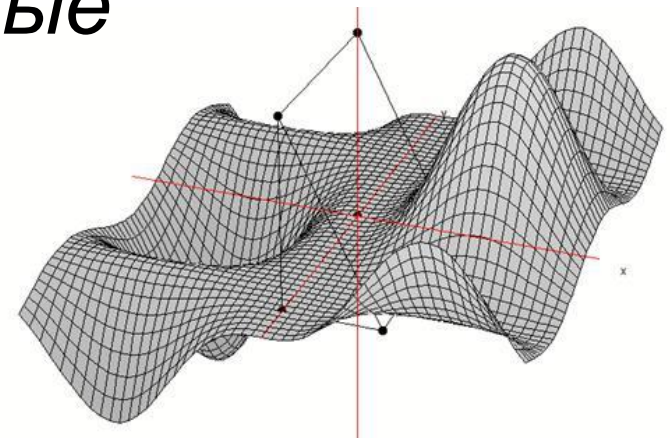
Определения:

Функция нескольких переменных:



где – $x_1, x_2, \ldots, x_n$ - *аргументы* или *независимые переменные*





Функция распределения случайной величины ξ : $F_\xi : \mathbb{R} \to [0,\ 1]$

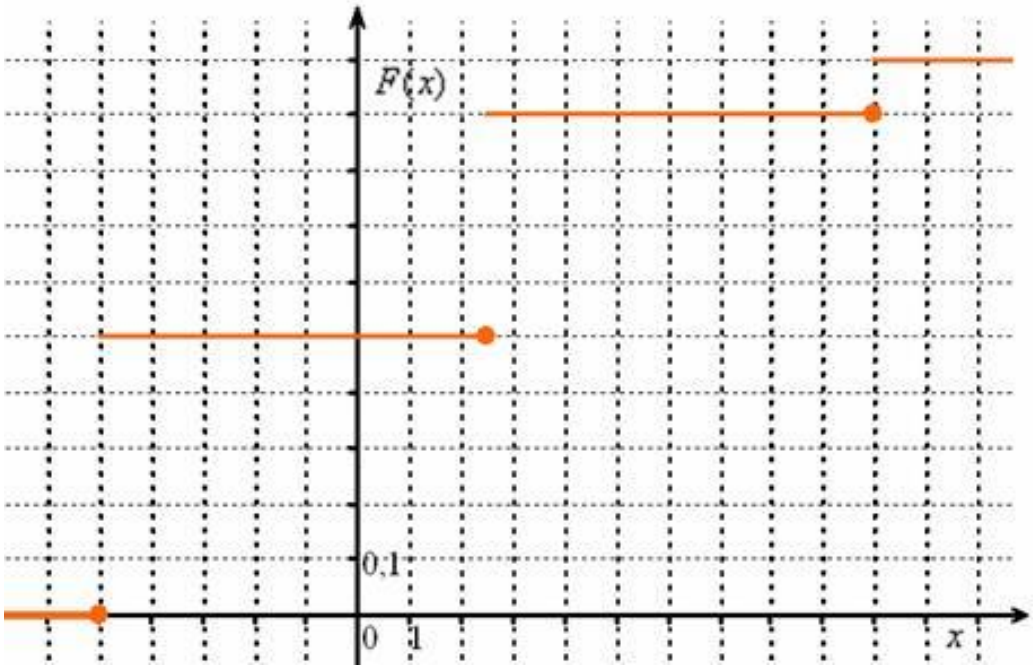
при каждом $x \in \mathbb{R}$

равная вероятности случайной величине ξ принимать значения, меньшие *x*:

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

Построение графика функции распределения случайной величины

x_i	–2	0	3	7
p_i	0,4	0,1	0,3	0,2



Функция совместного распределения случайных величин:

Функция

$$F_{\xi_1,\dots,\xi_n}(x_1,\dots,x_n) = P(\xi_1 < x_1,\dots,\xi_n < x_n)$$

называется *функцией распределения вектора*

$$(\xi_1,\dots,\xi_n)$$

или *функцией совместного распределения* случайных величин

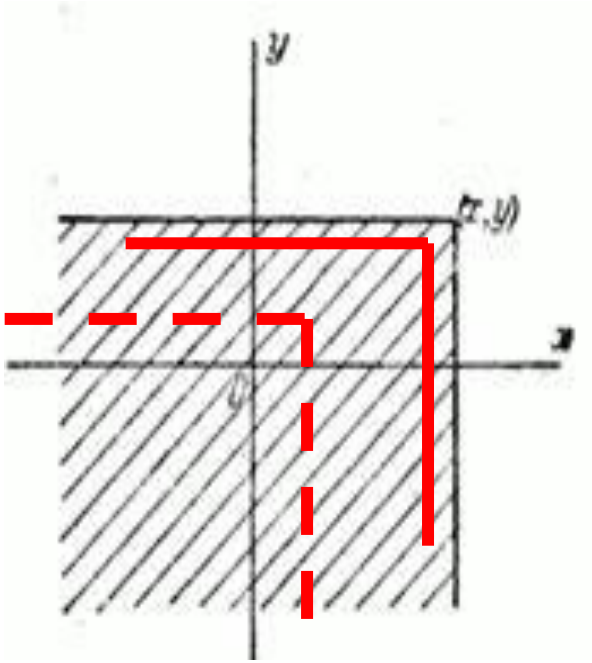
$$(\xi_1,\dots,\xi_n)$$

Свойства функции совместного распределения

Свойство 1: Функция распределения F (x,y) есть неубывающая функция обоих своих аргументов, т. е.:

при $x_2 > x_1$ $F(x_2,y) \geq F(x_1,y)$;

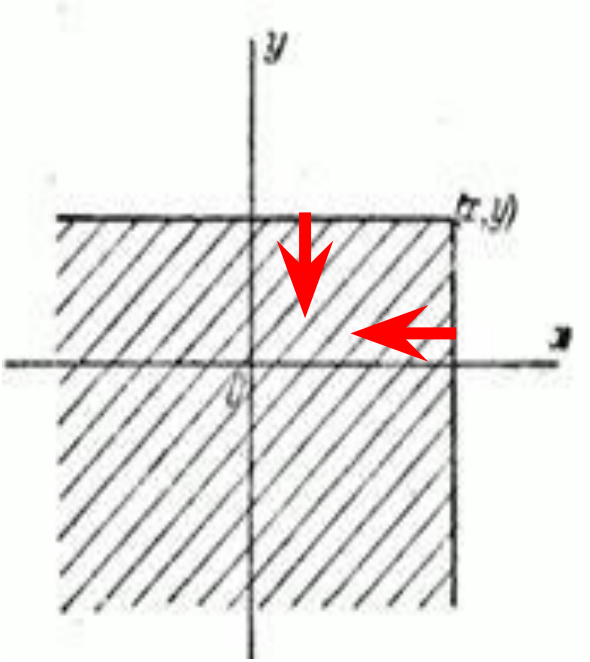
при $y_2 > y_1$ $F(x,y_2) \geq F(x,y_1)$.



Свойства функции совместного распределения

Свойство 2: Повсюду на -□ функция распределения равна нулю:

F(x, -□) = F(-□,y) = F (-□, -□) = 0.



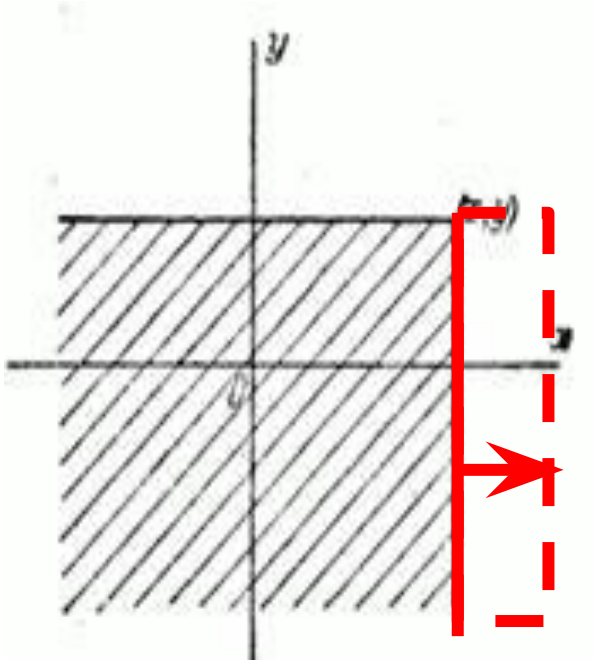
Свойства функции совместного распределения

Свойство 3: При одном из аргументов, равном $+\infty$, функция распределения системы превращается в функцию распределения случайной величины, соответствующей другому аргументу:

$F(x, +\infty) = F_1(x)$;

$F(+\infty, y) = F_2(y)$,

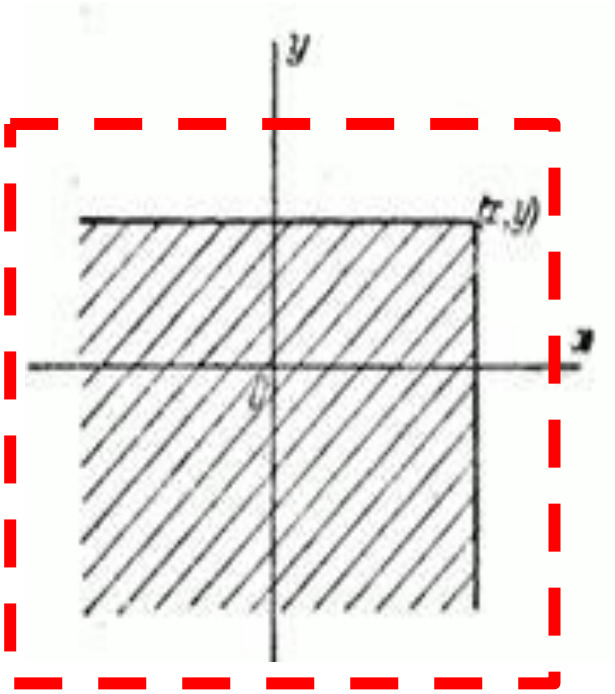
где $F_1(x)$, $F_2(y)$ - соответственно функции распределения случайных величин X и Y.



Свойства функции совместного распределения

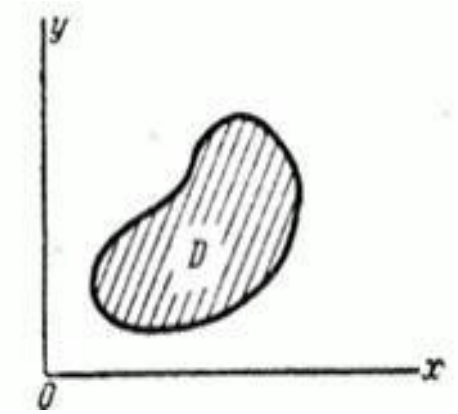
Свойство 4. Если оба аргумента равны $+\infty$, функция распределения системы равна единице:

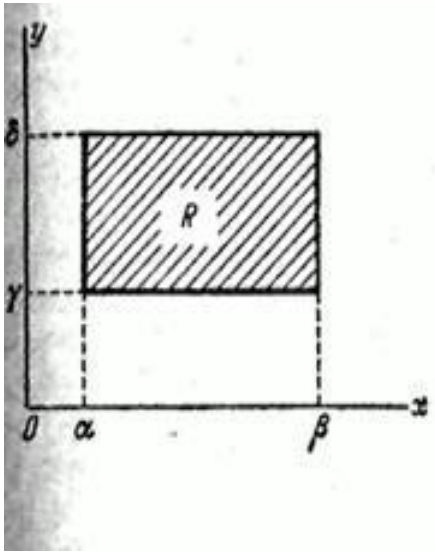
$F(+\infty, +\infty) = 1.$

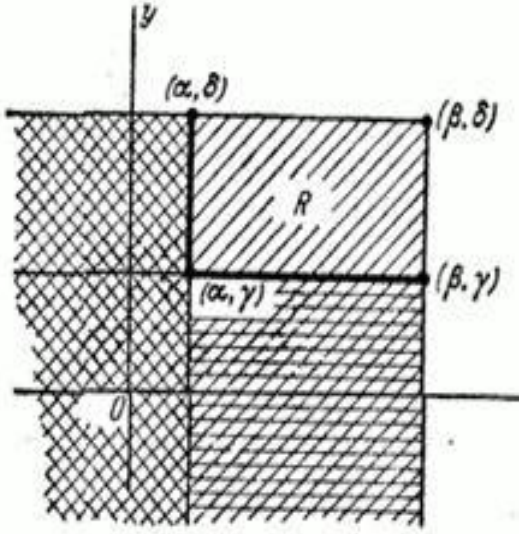


Свойства функции совместного распределения

Для системы двух случайных величин актуальным является вопрос о вероятности попадания случайной точки (X, Y) в пределы заданной области D на плоскости xOy:





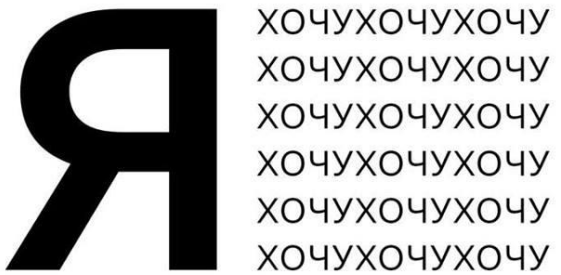


$$P((X, Y) \subset R) = F(\beta, \delta) - F(\alpha, \delta) - F(\beta, \gamma) + F(\alpha, \gamma)$$

Доверительные интервалы для параметра a в случае выборки из нормального распределения N (a,σ^2):

а) при известном σ^2;

б) при неизвестном σ^2



Определения:

- *Генеральная совокупность* - совокупность всех объектов (единиц), относительно которых предполагается делать выводы при изучении конкретной задачи. Генеральная совокупность состоит из всех объектов, которые имеют качества, свойства, интересующие исследователя.
- *Выборка* или *выборочная совокупность* — часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).

Функция распределения случайной величины X -

$$F_X(t) = P(X<t)\ t \in \mathbb{R}$$

Математическое ожидание - мера среднего значения случайной величины в теории вероятностей (задается интегралом Лебега — Стилтьеса) –

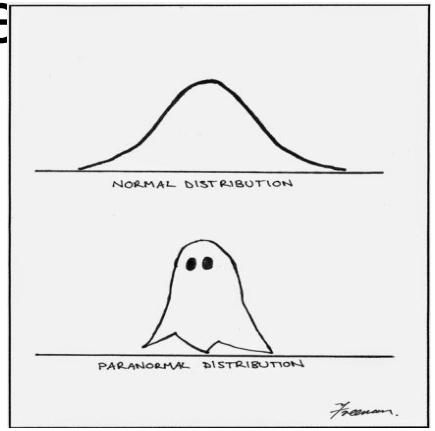
$$M[X] = \int_{-\infty}^{\infty} x\, dF_X(x)$$

Дисперсия (D[X], σ^2)- мера разброса значений случайной величины относительно её математического ожидания -

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

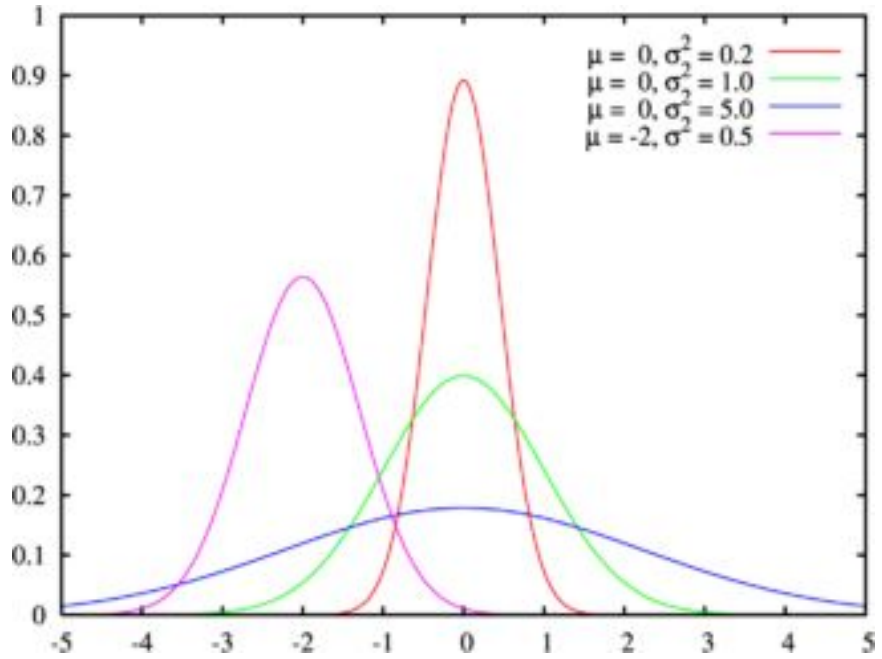
Закон распределения – это некоторая функция, полностью описывающая случайную величину с вероятностной точки зрения.

Нормальное распределение (распределение Гаусса) – семейство распределения вероятностей, которое играет важнейшую роль во многих областях знаний и зависит от двух параметров – *смещения* (коэффициент сдвига μ) и *масштаба* (коэффициент масштаба $\sigma > 0$). σ, μ – веществе



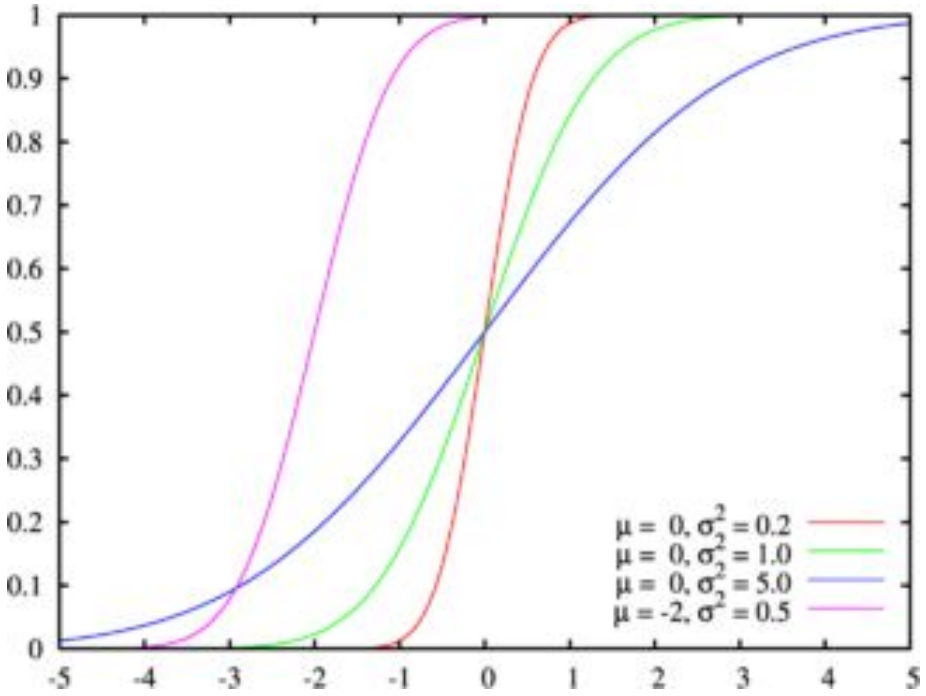
Плотность вероятности нормального распределения

$$p(x;\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad x \in (-\infty;+\infty)$$



Функция нормального распределения

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$



Доверительный интервал - это интервал, построенный с помощью случайной выборки из распределения с неизвестным параметром, такой, что он содержит данный параметр с заданной вероятностью.

Пусть $x_1,\ldots,x_n$ – выборка из некоторого распределения с плотностью $p(x;\theta) = p(x_1,\ldots,x_n;\theta)$, зависящей от параметра θ, который может изменяться в интервале $\theta_0 < \theta < \theta_1$.

Пусть $y(x_1,\ldots,x_n)$ – некоторая статистика и $F(x;\theta) = P\{\eta \le x\}$ – функция распределения случайной величины $\eta = y(x_1,\ldots,x_n)$, когда выборка $x_1,\ldots,x_n$ имеет распределение с плотностью $p(x_1,\ldots,x_n;\theta)$.

Предположим, что $F(x;\theta)$ есть убывающая функция от параметра θ.

Обозначим $x_\gamma(\theta)$ квантиль распределения $F(x;\theta)$, тогда $x_\gamma(\theta)$ - есть возрастающая функция от θ.

Зафиксируем близкое к нулю положительное число α (например, 0.05 или 0.01). Пусть $\alpha = \alpha_1 + \alpha_2$. При каждом θ неравенства

$$x_{1-\alpha_2}(\theta) \le \eta \le x_{\alpha 1}(\theta) \qquad (1)$$

выполняются с вероятностью 1-α, близкой к единице.

Перепишем неравенства (1) в другом виде:

$$x_{\alpha_1}^{-1}(\eta) \le \theta \le x_{1-\alpha_2}^{-1}(\eta) \qquad (2)$$

Обозначим $x_{\alpha_1}^{-1}(\eta) = \underline{\theta}(\eta) \qquad x_{1-\alpha_2}^{-1}(\eta) = \bar{\theta}(\eta)$

и запишем (2) в следующем виде:

$$P_0\left\{\underline{\theta}(\eta) \le \theta \le \bar{\theta}(\eta)\right\} = 1-\alpha$$

Интервал $\underline{\theta}(\eta) \le \theta \le \bar{\theta}(\eta)$ называется доверительным интервалом для параметра θ, а вероятность 1-α – доверительной вероятностью.

Доверительный интервал для математического ожидания (μ) в случае нормальной генеральной совокупности и известной дисперсии

$$\bar{x} - \frac{z\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z\sigma}{\sqrt{n}}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Вывод полученного выражения

0

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

1

$$P(|Z| \geq z) = \alpha$$

$$P(|Z| \geq z) = 1 - P(-z < Z < z)$$

2

$$1 - \alpha = P(-z < Z < z) = P\left(-z < \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} < z\right)$$

3

$$1 - \alpha = P\left(-z < \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} < z\right) = P\left(-\frac{z\sigma}{\sqrt{n}} < \bar{X} - \mu < \frac{z\sigma}{\sqrt{n}}\right) = P\left(\bar{X} - \frac{z\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z\sigma}{\sqrt{n}}\right)$$

$$\bar{x} - \frac{z\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z\sigma}{\sqrt{n}}$$

Доверительный интервал для математического ожидания (μ) в случае нормальной генеральной совокупности и неизвестной дисперсии

$$\bar{x} - \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}} < \mu < \bar{x} + \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}}$$

Вывод полученного выражения

1

$$U = \frac{\bar{X} - \mu}{\left(\hat{\sigma} / \sqrt{n}\right)} \qquad U = \frac{\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right)}{\sqrt{\frac{1}{n-1}\left(\frac{(n-1)\hat{\sigma}^2}{\sigma^2}\right)}}$$

Теперь нужно найти такое значение t, что P(|U|≥t)=α. Его обычно обозначают:

2

$$t_{\alpha/2}$$

$$P(U \geq t_{\alpha/2}) = \alpha/2$$

3

$$1-\alpha = P(-t_{\alpha/2} < U < t_{\alpha/2}) = P(-t_{\alpha/2} < \frac{\bar{X} - \mu}{\hat{\sigma} / \sqrt{n}} < t_{\alpha/2}) =$$

$$P(-\frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}} < \bar{X} - \mu < \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}}) = P(\bar{X} - \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}})$$

$$\bar{x} - \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}} < \mu < \bar{x} + \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}}$$

Творческое задание. Анализ статьи «Inflammation, Aspirin, and the Risk of Cardiovascular Disease in Apparently Healthy Men»

The NEW ENGLAND JOURNAL of MEDICINE

Что изучалось

Увеличивает ли воспалительный процесс риск возникновения тромботических заболеваний; снижает ли приём аспирина этот риск.

Методика

Авторы измерили уровень плазменного С-реактивного белка, маркер системного воспаления, у 543 здоровых мужчин, у которых впоследствии развился инфаркт миокарда, инсульт или венозный тромбоз, и у 543 участников исследования, которые не сообщили о сосудистых заболеваниях в течение последующего периода, превышающего восемь лет. Участники были рандомизированы для приёма аспирина или плацебо в начале исследования.

Перед **рандомизацией** в период с августа 1982 года по декабрь 1984 года потенциальным участникам было предложено предоставлять образцы опытной линии крови в течение 16-недельного периода, в течение которого *всем участникам был дан аспирин*, и никто не получал плацебо. Из 22 071 участников 14 916 (68%) предоставили образцы «опытной» плазмы.

Контроль был выбран случайным образом среди участников исследования, которые соответствовали критериям соответствия *возраста* (±1 год), *статусу курения* (курение в настоящее время, курили в прошлом или никогда не курили), а также *продолжительность времени*, прошедшего после после рандомизации (через 6-месячные интервалы). Используя эти методы, авторы оценили **543 пациента и 543 контроля**.

Базовые характеристики участников исследования

CHARACTERISTIC	CARDIOVASCULAR DISEASE DURING FOLLOW-UP*				
	NONE (N=543)	ANY (N=543)	MYOCARDIAL INFARCTION (N=246)	STROKE (N=196)	VENOUS THROMBOSIS (N=101)
Age (yr)	59±9.1	59±9.2	58±8.6	62±9.1	57±9.4
Smoking status (%)					
Never smoked	44	44	45	42	50
Smoked in the past	41	41	40	40	44
Currently a smoker	15	15	15	18	6
Diabetes (%)	4	7	5	12	2
Body-mass index†	25±2.8	26±3.2	26±3.3	25±3.2	26±2.9
History of high plasma cholesterol (%)	9	13	17	10	7
History of hypertension (%)	16	29	27	35	20
Parental history of coronary artery disease (%)	10	13	17	11	8

*Plus–minus values are means ±SD.

†The body-mass index is the weight in kilograms divided by the square of the height in meters.

Статистика

Для пациентов из контрольной группы были рассчитаны средние или доли для базовых факторов риска. Значение любой разницы в средних было проверено с использованием ***t*-критерия Стьюдента**, а значение любых различий в долях было проверено с использованием статистики χ^2. Поскольку значения С-реактивного белка искажены, вычислялись средние концентрации, и значение любых различий в средних значениях между пациентами и контрольной группой оценивали с использованием рангового **теста Уилкоксона** *(будет рассмотрен далее)*. Геометрические средние концентрации С-реактивного белка также вычислялись после **логарифмирования**, что приводило к почти нормальному распределению. Авторы использовали **тест для тренда**, чтобы оценить любое соотношение возрастающих значений С-реактивного белка с риском будущего сосудистого заболевания после деления образца на квартили, определяемые распределением контрольных значений. Авторы получили скорректированные оценки с использованием условных моделей **логистической регрессии**, которые учитывали сопоставимые переменные и контролировали назначение случайного лечения, индекс массы тела, диабет, историю гипертонии и родительскую историю болезни коронарной артерии. Аналогичные модели использовались для корректировки измеренных концентраций общей массы и холестерина, ЛПВП, триглицеридов, липопротеинов, антигена t-PA, фибриногена, D-димера и гомоцистеина. Чтобы оценить, повлиял ли аспирин на эти отношения, анализы были повторены для всех случаев инфаркта миокарда, произошедшего 25 января 1988 года или до этого, — даты, когда рандомизированное назначение аспирина прекращалось.

Концентрация плазменной концентрации С-реактивного белка в базовой линии у участников исследования, у которых не проявилось сосудистых заболеваний во время наблюдения (контроль) и у тех, у кого произошел инфаркт миокарда, инсульт или венозный тромбоз (пациенты)

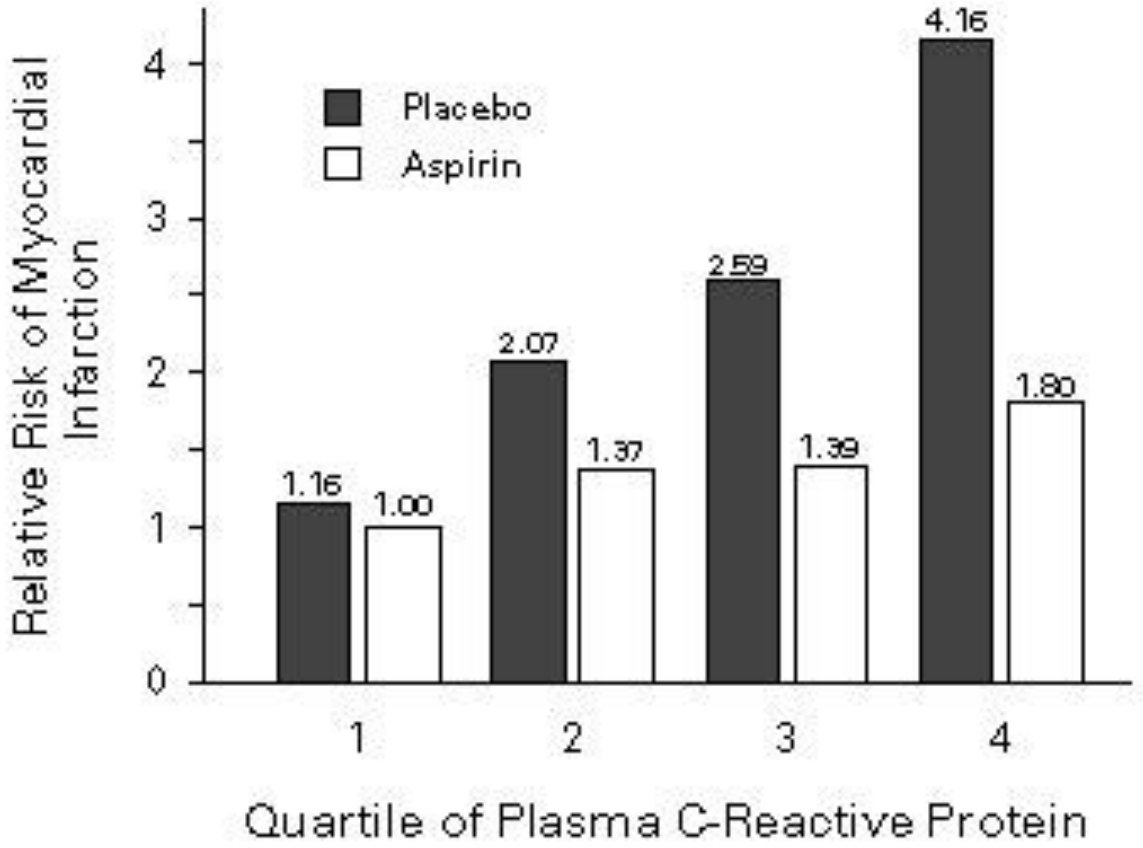
Cardiovascular Disease during Follow-up	Plasma C-Reactive Protein			
	Geometric mean	P value	Median	P value
	mg/liter		mg/liter	
None (n=543)	1.10	—	1.13	—
Any vascular event (n=543)	1.37	<0.001	1.40	<0.001
Myocardial infarction (n=246)	1.48	<0.001	1.51	<0.001
Any stroke (n=196)	1.30	0.03	1.36	0.03
Ischemic stroke (n=154)	1.36	0.01	1.38	0.02
Venous thrombosis (n=101)	1.24	0.22	1.26	0.34

Относительный риск будущего инфаркта миокарда, инсульта и венозного тромбоза в соответствии с концентрацией плазмы С-реактивного белка в базовой линии

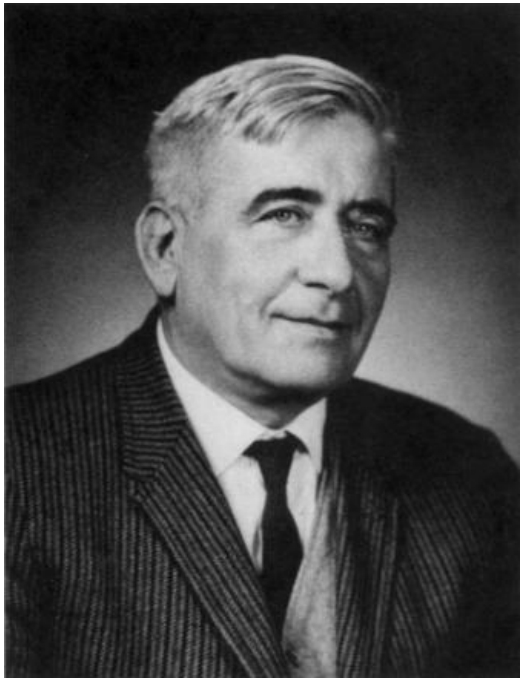
Vascular Event*	Quartile of C-Reactive Protein Concentration (mg/liter)				P for Trend
	≤0.55	0.56–1.14	1.15–2.10	≥2.11	
Myocardial infarction (total cohort)					
Relative risk	1.0	1.7	2.6	2.9	<0.001
95% CI	—	1.1–2.9	1.6–4.3	1.8–4.6	
P value	—	0.03	<0.001	<0.001	
Myocardial infarction (nonsmokers)					
Relative risk	1.0	1.7	2.5	2.8	<0.001
95% CI	—	1.0–2.8	1.5–4.1	1.7–4.7	
P value	—	0.06	<0.001	<0.001	
Ischemic stroke					
Relative risk	1.0	1.7	1.9	1.9	0.03
95% CI	—	0.9–2.9	1.1–3.2	1.1–3.3	
P value	—	0.07	0.02	0.02	
Venous thrombosis					
Relative risk	1.0	1.1	1.2	1.3	0.38
95% CI	—	0.6–2.0	0.7–2.3	0.7–2.4	
P value	—	0.78	0.51	0.42	

*CI denotes confidence interval.

Относительный риск первого инфаркта миокарда, связанного с концентрацией плазмы С-реактивного белка в базовой линии, стратифицированной в соответствии с рандомизированным назначением на аспирин или плацебо-терапию



Разбор статистической методики ***U*-критерий Манна — Уитни**



Представление данных

Выборка 1 (объём n_1): $x_{11}, x_{21}, \ldots, ;$

Выборка 2 (объём n_2): $x_{12}, x_{22}, \ldots, .$

Наблюдения из двух выборок объёма n_1 и n_2 объединяются и упорядочиваются, например, по возрастанию. Затем наблюдениям присваиваются ***ранги*.**

Выборка **первая** (объём n_1)

Наблюдение $x_{11}, x_{21}, \ldots,$

Ранг $r_{11}, r_{21}, \ldots,$

Сумма рангов в первой выборк

$$R_1 = \sum_{i=1}^{n_1} r_{i1}$$

Представление данных

Выборка **вторая** (объём n_2)

Наблюдение $x_{12}, x_{22}, \ldots,$

Ранг $r_{12}, r_{22}, \ldots,$

Сумма рангов во второй выборке

$$R_2 = \sum_{i=1}^{n_2} r_{i2}$$

Общее число наблюдений $N = n_1 + n_2$.

Статистическая модель

Все наблюдения независимы.

Наблюдения, входящих в одну выборку, относятся к одной совокупности.

Гипотезы

H_0: совокупности одинаково распределены;

H_1: нулевая гипотеза неверна

Критериальная статистика

Малые выборки

Вычисляются

$$U_1 = n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1) - R_1$$

$$U_2 = n_1 n_2 + \frac{1}{2} n_2 (n_2 + 1) - R_2$$

и берётся $U = \max(U_1, U_2)$

Критериальная статистика

Большие выборки

В том случае, когда объём меньшей выборки больше 20 или объём большей выборки превышает 40, то *U* распределение Манна — Уитни приближается к нормальному.

Пусть

$$\mu_U = \frac{n_1 n_2}{2} \qquad \sigma_U = \sqrt{\frac{n_1 n_2 (N+1)}{12}} \qquad z = \frac{|U - \mu_U|}{\sigma_U}$$

Критериальная статистика

В том случае, если совпадающие ранги существуют, то

$$\sigma_U = \sqrt{\frac{n_1 n_2 \left(N^3 - N - \sum_j \left(t_j^3 - t_j\right)\right)}{12\left(N^2 - N\right)}}$$

где j — число связок, t_j — число элементов в связке

Поправка Йейтса

$$z = \frac{|U - \mu_U| - 0{,}5}{\sigma_U}$$

Отсутствие поправки на непрерывность приводит к увеличению значения статистики и, соответственно, уменьшению величины достигнутого уровня значимости. Это приводит к более частому отклонению нулевой гипотезы и принятию гипотезы H_1.

Результаты статьи

В статье были сравнены концентрации С-реактивного белка у двух групп мужчин (по 543 человека в каждой в соответствии, стало быть, указанного выше «рецепта» применения данного критерия). Точно проследить использование данного критерия не представляется возможным по данной статье, так как авторы не приводят первичные данные для 1086 участников.

Концентрации С-реактивных белков плазмы в «эксперименте» были выше среди мужчин, у которых был инфаркт миокарда (1,51 против 1,13 мг/л, $P < 0,001$) или ишемический инсульт (1,38 против 1,13 мг/л, $P = 0,02$), но не венозный тромбоз (1,26 против 1,13 мг на литр, $P = 0,34$), чем у мужчин без сосудистых событий. У мужчин в квартилях с самыми высокими значениями концентрации С-реактивного белка риск возникновения инфаркта миокарда в три (относительный риск, 2,9, $P < 0,001$) и риск возникновения ишемического инсульта (относительный риск 1,9; $P = 0,02$) в два раза превышал таковой у мужчин в наименьшей квартили. Риски были стабильными в течение длительного периода времени, их значения не были подвергнуты влиянию курению и не зависели от других факторов риска, связанных и не связанных с липидами. Использование аспирина было связано со значительным снижением риска инфаркта миокарда (снижение на 55,7%, $P = = 0,02$) среди мужчин в самом высоком квартиле, но с небольшими незначительными

Результаты статьи

Экспериментальная концентрации С-реактивного белка в плазме предсказывает риск будущего инфаркта миокарда и инсульта. Более того, снижение, связанное с использованием аспирина в риске развития первого инфаркта миокарда, по-видимому, напрямую связано с уровнем С-реактивного белка, повышая вероятность того, что противовоспалительные агенты могут иметь клинические преимущества в профилактике сердечно-сосудистых заболеваний.

Список использованной литературы:

- *Ивашёв-Мусатов О. С.* Теория вероятностей и математическая статистика: Учеб. пособие. — 2-е изд., перераб. и доп. — М.: ФИМА, 2003. — 224 с.
- *Гланц С.* Медико-биологическая статистика. Пер . с англ. — М., Практика, 1998. — 459 с.
- *Кочнева Л.Ф., Липкина З.С., Новосельцева В. И.* Теория вероятностей и математическая статистика (Часть III): Учеб. пособие - федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Московский государственный университет путей сообщения», Москва, 2012. – 44с.
- *Ridker P. M. et al.* Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men //New England journal of medicine. — 1997. — V. 336. — N. 14. — Pp. 973-979.
- *Яровая Е. Б.* Лекции курса основ теории вероятностей и математической статистики, прочитанные в МГУ имени М. В. Ломоносова на факультете фундаментальной медицины с 10.02.2017 по 18.05.2018.