Теория вероятностей.

1. Случайные события.
2. Теоремы сложения и умножения вероятностей.
3. Формула полной вероятности. Формула Байеса.
4. Дискретные и непрерывные случайные величины. Законы распределения и их числовые характеристики.
5. Статистическое распределение выборки. Характеристики вариационного ряда.
6. Точечные и интервальные оценки параметров распределения.
7. Элементы корреляционного анализа. Проверка статистических гипотез.

1. **Случайное событие** — некотороеподмножество множества элементарных исходов случайного эксперимента. Случайное событие может либо произойти или не произойти при осуществлении определенной совокупности условий.

Случайные события обозначаются прописными буквами: A, B, C …

Событие, которое обязательно произойдет при определенной совокупности условий, называется **достоверным** и обозначается символом Ω.

Событие, которое заведомо не произойдет при определенной совокупности условий, называется **невозможным** и обозначается символом ∅.

Вероятность случайного события

$$P(A) = \frac{m}{n},$$

где m - число исходов, благоприятствующих появлению данного события, n – общее число всех равновозможных элементарных исходов. Вероятность – это число, являющееся мерой объективной возможности наступления события.

Вероятность достоверного события $P(\Omega) = 1$, вероятность невозможного события $P(\emptyset) = 0$. Следовательно, вероятность случайного события $0 \leq P(A) \leq 1$.

Для подсчета числа благоприятных исходов и числа равновозможных исходов пользуются комбинаторикой. Существуют два основных правила комбинаторики:

1. Правило сложения.

Если два альтернативных (взаимно исключающих) действия могут быть выполнены nи mспособами, то выполнение одного из них возможно $n + m$ способами.

2. Правило умножения.

Если первое действие можно сделать nспособами, а второе – m способами, то два действия можно сделать $n \times m$ способами.

Перестановками называются комбинации, состоящие из одних и тех же n различных элементов и отличающиеся только порядком их расположения:

$$P(n) = P_n = n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n.$$

Перестановки с повторениями:

$$P_n(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}, \quad \text{где } n_1 + n_2 + \dots + n_k = n.$$

Размещениями называются комбинации, составленные из n различных элементов по m элементов, которые отличаются либо составом элементов, либо их порядком:

$$A_n(m) = A_n^m = \frac{n!}{(n-m)!}.$$

Размещения с повторениями:

$$A_n^m = n^m.$$

Сочетаниями называются комбинации, составленные из n различных элементов по m элементов, которые отличаются составом элементов:

$$C_n^m = \frac{n!}{m! \cdot (n-m)!}.$$

Сочетания с повторениями:

$$C_n^m = C_{n+m-1}^m$$

2. Теоремы сложения и умножения вероятностей.

Два события называются **несовместными**, если появление одного из них исключает появление другого в одном и том же испытании. Иначе, события называются **совместными.**

Вероятность суммы двух несовместных событий равна сумме вероятностей этих событий:

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

Вероятность суммы двух совместных событий равна сумме вероятностей этих событий без вероятности их совместного появления:

$$P(A+B) = P(A) + P(B) - P(AB).$$

Группа событий называется **полной**, если в результате опыта наступает хотя бы одно из этих событий.

Два события называются **противоположными**, если это несовместные события, образующие полную группу. $\overline{A}$ - противоположное событие.

$P(A+\overline{A}) = P(A) + P(\overline{A}) = 1$, следовательно $P(\overline{A}) = 1 - P(A)$.

Произведением двух событий A и B называют событие AB, состоящее в совместном появлении этих событий.

Два события называются **независимыми**, если вероятность одного из них не зависит от появления другого события.

Вероятность совместного появления двух <u>независимых</u> событий равна произведению вероятностей этих событии:

$$P(AB) = P(A)\,P(B).$$

Два события называются зависимыми, если вероятность появления одного из них зависит от появления или не появления другого события.

P(A|B) – это **условная вероятность** события A при условии, что событие B уже произошло.

Вероятность совместного появления двух зависимых событий равна произведению вероятности одного из этих событий на условную вероятность другого:

$$P(\mathrm{AB}) = P(A)P(\mathrm{A}|B) = P(B)P(B|A).$$

3. Формула полной вероятности. Формула Байеса.

Пусть события $H_1, H_2, \ldots, H_n$ образуют полную группу событий. Тогда для любого события A, которое может произойти при условии наступления одного из событий H_i, имеет место **формула полной вероятности**

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i).$$

Пусть события $H_1, H_2, \ldots, H_n$ образуют полную группу событий. Тогда условная вероятность события H_k ($k = \overline{1,n}$) при условии, что событие A произошло, задается **формулой Байеса**

$$P(H_k|A) = \frac{P(H_k) \cdot P(A|H_k)}{\sum_{i=1}^{n} P(H_i) \cdot P(A|H_i)}.$$

Формула Байеса показывает, как изменяется вероятность гипотезы $P(H_i)$ при реализации события A.

Если производится n независимых испытаний, в каждом из которых вероятность появления события A равна p , а вероятность его не появления равна $q = 1 - p$, то вероятность того, что событие A произойдет m раз определяется **формулой Бернулли**:

$$P_n(m) = C_n^m p^m q^{n-m}, m = 0, 1, 2, \ldots, n.$$

Число m_0, при котором вероятность $P_n(m_0)$ достигает своего максимального значения, называют **наивероятнейшим числом успехов:**

$$np - q \le m_0 \le np + p.$$

4. Дискретные и непрерывные случайные величины. Законы распределения и их числовые характеристики.

Случайной величиной называют величину, которая в результате опыта принимает различные значения, заранее неизвестные.

Дискретная случайная величина принимает отдельные изолированные значения с определенными вероятностями. Законом распределения случайной величины называют соответствие между возможными значениями и их вероятностями.

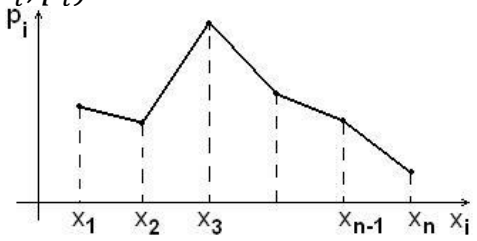
Две случайные величины называются независимыми, если закон распределения одной из них не зависит от того, какие возможные значения приняла другая случайная величина.

Закон распределения дискретной случайной величины можно представить в виде ряда распределения, многоугольника распределения и функции распределения.

Ряд распределения:

X	x_1	x_2	x_3	. . .	x_n
P	p_1	p_2	p_3	. . .	p_n

Многоугольник распределения – это ломаная, вершинами которой являются точки с координатами $(x_i; p_i)$:



Числовыми характеристиками называют параметры, отражающие наиболее существенные черты закона распределения случайной величины.

Математическим ожиданием дискретной случайной величины называется сумма произведений возможных значений случайной величины и их вероятностей:

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

Свойства математического ожидания независимых случайных величин:

1. $M(C) = C$ (С – постоянная),
2. $M(CX) = C \cdot M(X)$,
3. $M(X \cdot Y) = M(X) \cdot M(Y)$,
4. $M(X \pm Y) = M(X) \pm M(Y)$.

Модой дискретной случайной величины называется ее наивероятнейшее значение.

Дисперсией случайной величины называют математическое ожидание квадрата отклонения случайной величины от ее математического ожидания:

$$D(X) = M(X - M(X))^2.$$

Дисперсию можно вычислить по формуле $D(X) = M(X^2) - (M(X))^2$.

Свойства дисперсии:

1. $D(C) = 0$,
2. $D(CX) = C^2 \cdot D(X)$,
3. $D(X + Y) = D(X) + D(Y)$,
4. $D(X - Y) = D(X) + D(Y)$.

Среднее квадратическое отклонение – это $\sigma(X) = \sqrt{D(X)}$. Этот параметр имеет размерность случайной величины и может быть наглядно представлено графически.

Основные распределения дискретной случайной величины.

Биномиальное: Случайная величина *X* представляет собой число появлений события *A* в *n* независимых опытах и принимает целые неотрицательные значения. Вероятность того, что в n испытаниях событие A появится ровно m раз, вычисляется по формуле Бернулли.

Параметры биномиального распределения $M(X) = np, D(X) = npq, \sigma(X) = \sqrt{npq}$.

Пуассоновское: Случайная величина *X* принимает целые неотрицательные значения 0, 1, 2, 3 ...,n, где n достаточно большое число. Вероятность появления события в одном опыте p является достаточно малым числом.

Однако, $\lambda = np$ограничено. Тогда вероятность того, что в n испытаниях событие A появится ровно m раз, вычисляется по формуле Пуассона

$$P_n(m) = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Параметры распределения Пуассона $M(X) = \lambda, D(X) = \lambda, \sigma(X) = \sqrt{\lambda}$.

Геометрическое: Пусть производятся независимые испытания, в каждом из которых событие наступает с вероятностью p. Испытания заканчиваются, как только появится событие A. Вероятность появления события A не менее чем в m опытах, вычисляется по формуле

$$P(m) = q^{m-1}p.$$

Параметры геометрического распределения $M(X) = \frac{1}{p}, D(X) = \frac{q}{p^2}, \sigma(X) = \frac{\sqrt{q}}{p^2}$.

Гипергеометрическое: Пусть имеется конечная совокупность, состоящая из N элементов, среди которых M обладают определенным свойством. Случайным образом из общей совокупности выбирается группа из n элементов. Вероятность того, что в данной выборке окажется ровно m элементов, обладающих указанным свойством, вычисляется по формуле

$$P(m) = \frac{C_M^m \cdot C_{N-M}^{n-m}}{C_N^n}.$$

Параметры гипергеометрического распределения $M(X) = \frac{M \cdot n}{N}, D(X) = \frac{M \cdot n \cdot (N-M)(N-n)}{N^2(N-1)}$.

Непрерывная случайная величина принимает все значения из некоторого конечного или бесконечного промежутка.

Закон распределения непрерывной случайной величины можно представить в виде функции распределения и плотности распределения.

Функция распределения – это универсальная форма закона распределения, так как она характеризует не только непрерывную случайную величину, но и дискретную тоже.

Функцией распределения случайной величины X называется функция F(x), которая для любого x $\in R$ равна вероятности события (X<x): $F(x) = P(X < x)$.

Свойства функции распределения:

1. F(x) – неубывающая функция на R,
2. $0 \leq F(x) \leq 1$,
3. $F(-\infty) = 0$,
4. $F(+\infty) = 1$.

Плотностью распределения вероятностей непрерывной случайной величины X называется производная от функции распределения: $f(x) = F'(x)$.

Свойства плотности распределения:

1. $f(x) \geq 0$,
2. $\int_{-\infty}^{\infty} f(x)dx = 1$,
3. $P(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx$,
4. $F(x) = \int_{-\infty}^{x} f(t)dt$.

Математическое ожидание непрерывной случайной величины на интервале $[a; b]$равно: M(x) = $\int_a^b xf(x)dx$, дисперсия равна D(X) =$\int_a^b (x - M(X))^2 f(x)dx$. Дисперсию также можно вычислить по формуле: $D(x) = \int_a^b (x)^2 f(x)dx - M(X)^2$

Модой непрерывной случайной величины называется точка локального максимума функции плотности $f(x)$.

Корень уравнения $F(x) = 1/2$ называется медианой случайной величины.

Основные законы распределения непрерывной случайной величины:

Равномерное распределение задается плотностью распределения

$$f(x) = \begin{cases} \frac{1}{b-a}, \text{при } x \in [a, b], \\ 0, \text{при } x \notin [a, b]. \end{cases}$$

Параметры равномерного распределения $M(X) = \frac{a+b}{2}, D(X) = \frac{(b-a)^2}{12}, \sigma(X) = \frac{(b-a)}{2\sqrt{3}}$.

Показательное распределение задается функцией плотности

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

Параметры показательного распределения M(X) = $\frac{1}{\lambda}$,D(X) = $\frac{1}{\lambda^{2,}} \sigma(X) = \frac{1}{\lambda}$.

Нормальное распределение задается функцией плотности

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

Параметры нормального распределения $M(X) = a, D(X) = \sigma^2$.

5. Статистическое распределение выборки. Характеристики вариационного ряда.

Для изучения случайной величины X производят ряд независимых опытов. В каждом из этих опытов случайная величина X принимает то или иное значение. Пусть она приняла n_1 раз значение x_1, n_2 раз - значение x_2, ... n_kраз – значение x_k.

Значения x_1, x_2, ... x_k называются вариантами случайной величины X. Числа n_i, показывающие, сколько раз встречается варианта x_i в выборке, называются частотами, а $\omega_i = \frac{n_i}{n}$ называются относительными частотами. Число $\sum n_i = n$ - это объем выборки.

Последовательность вариант, упорядоченных по возрастанию, называется вариационным рядом.

Перечень вариант и соответствующих им частот (и/или относительных частот) называется статистическим рядом или статистическим распределением выборки.

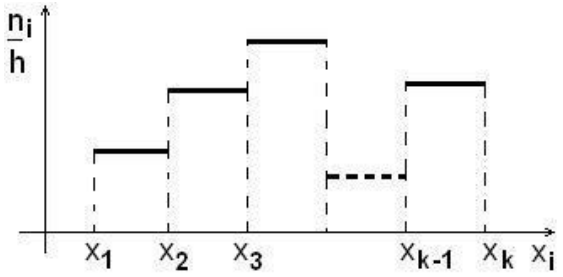
Статистический ряд может быть представлен в виде таблицы

x_i	x_1	x_2	x_3	...	x_k

n_i	n_1	n_2	n_3	...	n_k

Или графически в виде полигона частот - ломаной, соединяющей точки (x_i; n_i). В случае непрерывного признака распределения целесообразно строить гистограмму частот.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высоты равны $\frac{n_i}{h}$ (плотность частоты). Площадь гистограммы равна объему выборки n.



Числовые характеристики выборки:

1. выборочная средняя $\bar{x}_{\text{в}} = \frac{1}{\text{n}}\sum_{i=1}^{k} x_i \cdot n_i$

2. выборочная дисперсия $\text{D}_{\text{в}} = \frac{1}{\text{n}}\sum_{i=1}^{k}(x_i - \bar{x}_{\text{в}})^2 \cdot n_i$

3. выборочное среднее квадратическое отклонение $\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$

4. исправленная выборочная дисперсия $S^2 = \frac{n}{n-1} D_{\text{в}}$

5. размах вариации $\text{R} = \text{x}_{\text{max}} - \text{x}_{\text{min}}$

6. мода вариационного ряда $\widetilde{Mo}$ - варианта, имеющая наибольшую частоту

7. медиана вариационного ряда $\widetilde{Me}$ - значение варианты, приходящейся на середину вариационного ряда

6. Точечные и интервальные оценки параметров распределения.

Пусть изучается случайная величина X с математическим ожиданием M(X) и дисперсией D(X), оба эти параметра неизвестны.

Статистика, используемая в качестве приближенного значения неизвестного параметра, называется ее **точечной оценкой**. Качество оценки определяют, проверяя, обладает ли она свойствами несмещенности, состоятельности и эффективности.

Несмещенность оценки означает отсутствие систематических погрешностей в наблюдаемых данных, для этого ее математическое ожидание должно быть равно оцениваемому параметру.

Состоятельность оценки заключается в том, что с ростом числа наблюдений дисперсия стремится к нулю.

Для исследуемого параметра оценка **эффективна**, если имеет минимальную дисперсию среди всех возможных оценок, построенных по данной выборке.

Выборочное среднее $\bar{x}_{\text{в}}$ является несмещенной и состоятельной оценкой математического ожидания M(X). Исправленная выборочная дисперсия S^2 является несмещенной и состоятельной оценкой дисперсии D(X).

Интервальные оценки являются более полными и надежными по сравнению с точечными, они применяются как для больших, так и для малых выборок. **Интервальной оценкой** служит симметричный относительно точечной оценки интервал. Если дана интервальная оценка $(a; b)$, то точечная оценка определяется по интервальной как $\frac{b+a}{2}$. Точность интервальной оценки $(a; b)$ определяется как $\delta = \frac{b-a}{2}$.

8. Элементы корреляционного анализа. Проверка статистических гипотез.

Корреляционной зависимостью называют такую зависимость, при которой изменение одной из случайных величин влечет изменение средних значений другой случайной величины.

Основные задачи корреляционного анализа:

1. определение формы корреляционной связи,
2. оценка тесноты связи,
3. проверка значимости корреляционной связи.

При изучении зависимости между двумя величинами чаще других используется линейная регрессия. Уравнение парной линейной регрессии имеет вид$y = \alpha + \beta x + \delta$, где β - коэффициент регрессии. Уравнение регрессии всегда дополняется показателем тесноты связи изучаемых величин. Для линейной регрессии в качестве такого показателя выступает коэффициент корреляции r_{xy}. Коэффициент корреляции принадлежит промежутку [-1; 1] и его знак совпадает со знаком коэффициента регрессии.

Коэффициент корреляции можно найти по следующим формулам:

$$r_{\text{ху}} = \frac{\mu_{xy}}{\sigma_x \sigma_y}, \text{где } \mu_{xy} = \overline{XY} - \overline{X} \cdot \overline{Y}.$$

$$\overline{X} = M(X) = \frac{\sum_i x_i}{n}, \overline{Y} = M(Y) = \frac{\sum_i y_i}{n}, \overline{XY} = \frac{\sum_i x_i \cdot y_i}{n},$$

$$\sigma_x = \sqrt{D_{\text{в}}(X)}, \sigma_y = \sqrt{D_{\text{в}}(Y)}, D_{\text{в}}(X) = \frac{1}{\text{n}} \sum_{i=1}^{k} (x_i - \bar{x}_{\text{в}})^2 \cdot n_i \,, D_{\text{в}}(Y)$$

$$= \frac{1}{\text{n}} \sum_{i=1}^{k} (y_i - \bar{y}_{\text{в}})^2 \cdot n_i.$$

Статистической гипотезойH_0 называют гипотезу о виде неизвестного распределения или о параметрах известного распределения. Конкурирующей

(альтернативной) называют гипотезу H_1, которая противоречит основной гипотезе.

Правило проверки статистических параметрических гипотез.

1. Формулируют H_0 и H_1.
2. Назначают уровень значимости α.
3. Выбирают статистику критерия *K* для проверки H_0.
4. Определяют $К_{набл}$ по выборке при условии, что H_0 верна.
5. В зависимости от H_1 определяют критическую область (левую, правую или двухстороннюю).
6. По таблице определяют значение $К_{кр}$.
7. Если $К_{набл} > К_{кр}$, то гипотезу H_0 отвергают, если же $К_{набл} < К_{кр}$, то гипотеза H_0 не противоречит наблюдаемым данным.

Проверка гипотезы о виде распределения случайной величины осуществляется с помощью специально подобранной случайной величины – критерия согласия. ***Критерием согласия*** называют статистический критерий проверки гипотезы о предполагаемом законе распределения.Критерий согласия Пирсона χ^2устанавливает, при уровне значимости α, согласуется ли гипотеза с опытными данными.

$$\chi^2_{\text{набл}} = \sum_{i=1}^{k} \frac{\left(n_i - n_i^T\right)^2}{n_i^T},$$

где n_i – эмпирические частоты, $\sum_{i=1}^{k} n_i = n$; n_i^T – теоретические частоты, $\sum_{i=1}^{k} n_i^T = n$.

Случайная величина *V* распределена по закону χ^2 со степенями свободы $k = s - r - 1$, где S – число интервалов выборки, r – число параметров предполагаемого распределения. Из уравнения $P\left(\chi^2_{\text{набл}} - \chi^2_{\text{кр}}\right) = \alpha$ по соответствующим таблицам определяем $\chi^2_{\text{кр}}$. Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$, то гипотезу H_0о том, что измеряемая величина распределена нормально, следует отбросить как несостоятельную. Если же $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$, то гипотеза о нормальном распределении этой выборки не противоречит наблюдаемым данным.