**Теория вероятностей и математическая статистика**

**1. Случайный эксперимент и случайные события. σ-алгебра событий.**

**Случайный эксперимент** – это испытание с заранее неизвестным исходом. Примеры: бросание монеты, измерение температуры, подсчет числа клиентов.

**Пространство элементарных исходов (Ω)** – множество всех возможных взаимоисключающих результатов эксперимента.

**Случайное событие** – любое подмножество пространства элементарных исходов. События обозначаются заглавными латинскими буквами (A, B, C).

**σ-алгебра событий –** система подмножеств Ω, замкнутая относительно операций над множествами и удовлетворяющая условиям: - Ω принадлежит σ-алгебре - Если множество A принадлежит σ-алгебре, то и его дополнение принадлежит - Если счетное число множеств принадлежит σ-алгебре, то их объединение тоже принадлежит

Аксиоматическое определение вероятности: вероятность P – это функция, которая каждому событию A из σ-алгебры ставит в соответствие число P(A), удовлетворяющее аксиомам: 1. P(A) ≥ 0 для любого события A 2. P(Ω) = 1 3. Если события A₁, A₂, … попарно несовместны, то P(A₁∪A₂∪…) = P(A₁) + P(A₂) + …

Классическое определение вероятности применяется, когда число элементарных исходов конечно и они равновозможны: P(A) = m/n, где m – число благоприятных исходов, n – общее число исходов.

Геометрическое определение вероятности: P(A) = mes(A)/mes(Ω), где mes – мера множества (длина, площадь, объем).

**2. Условная вероятность и независимость событий. Формулы сложения, полной вероятности и Байеса.**

**Условная вероятность** события A при условии B (P(B) > 0) – это вероятность A с учетом информации о наступлении B: P(A|B) = P(A∩B)/P(B)

События A и B называются **независимыми**, если: P(A∩B) = P(A)·P(B)

**Формула сложения** вероятностей: - Для несовместных событий: P(A∪B) = P(A) + P(B) - Для произвольных событий: P(A∪B) = P(A) + P(B) - P(A∩B)

**Формула полной вероятности** позволяет вычислить вероятность события, если известны условные вероятности при различных гипотезах: P(A) = P(A|H₁)·P(H₁) + P(A|H₂)·P(H₂) + … + P(A|Hₙ)·P(Hₙ) где {H₁, H₂, …, Hₙ} – полная группа несовместных событий (гипотез).

**Формула Байеса** – переоценка вероятностей гипотез после наблюдения события A: P(Hᵢ|A) = [P(A|Hᵢ)·P(Hᵢ)] / P(A) = [P(A|Hᵢ)·P(Hᵢ)] / [Σ P(A|Hⱼ)·P(Hⱼ)]

Эта формула широко применяется в статистике, машинном обучении и теории принятия решений.

**3. Схема Бернулли. Локальная и интегральная предельные теоремы Муавра-Лапласа. Предельная теорема Пуассона.**

**Схема Бернулли** – последовательность n независимых испытаний, в каждом из которых вероятность “успеха” равна p, а “неудачи” – q = 1-p.

Вероятность получения ровно k успехов в n испытаниях: P(X = k) = C(n,k)·pᵏ·q^(n-k), где C(n,k) – число сочетаний из n по k.

**Локальная теорема Муавра-Лапласа**: при больших n вероятность получения ровно k успехов приближенно равна: P(X = k) ≈ φ(x) / √(npq), где x = (k-np)/√(npq), φ(x) = e^(-x²/2)/√(2π) – плотность стандартного нормального распределения.

**Интегральная теорема Муавра-Лапласа**: при больших n вероятность получения от a до b успехов приближенно равна: P(a ≤ X ≤ b) ≈ Φ((b-np)/√(npq)) - Φ((a-np)/√(npq)), где Φ(x) – функция распределения стандартного нормального закона.

**Предельная теорема Пуассона**: если n → ∞, p → 0 так, что np = λ (константа), то: P(X = k) ≈ e^(-λ)·λᵏ/k!

Распределение Пуассона часто используется для моделирования редких событий, например, числа клиентов, прибывающих за единицу времени, числа опечаток в тексте и т.д.

**4. Случайные величины (СВ). Свойства функции распределения (ФР).**

**Случайная величина** – это функция, которая каждому элементарному исходу ставит в соответствие число.

Функция распределения (ФР) случайной величины X – это функция F(x) = P(X < x), определяющая вероятность того, что X примет значение, меньшее x.

**Свойства функции распределения:** 1. F(x) – неубывающая функция: если x₁ < x₂, то F(x₁) ≤ F(x₂) 2. F(-∞) = 0, F(+∞) = 1 3. F(x) непрерывна справа: F(x) = F(x+0) 4. P(a ≤ X < b) = F(b) - F(a) 5. P(X = a) = F(a) - F(a-0)

Дискретные случайные величины принимают конечное или счетное множество значений. Они задаются рядом распределения – таблицей значений и соответствующих вероятностей.

Непрерывные случайные величины имеют абсолютно непрерывную функцию распределения и могут быть описаны плотностью распределения f(x) = F’(x).

Свойства плотности распределения: 1. f(x) ≥ 0 для всех x 2. ∫f(x)dx = 1 (интеграл по всей числовой оси) 3. F(x) = ∫f(t)dt (от -∞ до x) 4. P(a ≤ X ≤ b) = ∫f(x)dx (от a до b)

**5. Многомерные СВ – определение. ФР – определение и свойства. Непрерывные и дискретные многомерные СВ.**

**Многомерная случайная величина** (X₁, X₂, …, Xₙ) – это набор случайных величин, определенных на одном вероятностном пространстве.

**Функция распределения многомерной СВ**: F(x₁, x₂, …, xₙ) = P(X₁ < x₁, X₂ < x₂, …, Xₙ < xₙ)

**Свойства многомерной ФР**: 1. F(x₁, x₂, …, xₙ) неубывает по каждому аргументу 2. F(-∞, …, -∞) = 0, F(+∞, …, +∞) = 1 3. F непрерывна справа по каждому аргументу 4. Вероятность попадания в прямоугольник равна сумме значений F в вершинах с соответствующими знаками

**Дискретные многомерные СВ** принимают счетное множество значений (x₁, x₂, …, xₙ) с вероятностями P(X₁ = x₁, X₂ = x₂, …, Xₙ = xₙ).

**Непрерывные многомерные СВ** имеют плотность распределения f(x₁, x₂, …, xₙ), такую что: F(x₁, x₂, …, xₙ) = ∫…∫f(t₁, t₂, …, tₙ)dt₁dt₂…dtₙ (интегрирование по всем tᵢ от -∞ до xᵢ)

Условные распределения позволяют найти распределение одних компонент при фиксированных значениях других: P(X₁ = x₁ | X₂ = x₂) = P(X₁ = x₁, X₂ = x₂) / P(X₂ = x₂)

Независимость случайных величин: X₁, X₂, …, Xₙ независимы, если: F(x₁, x₂, …, xₙ) = F₁(x₁)·F₂(x₂)·…·Fₙ(xₙ) где Fᵢ – функция распределения Xᵢ.

**6. Математическое ожидание, дисперсия. Моменты высших порядков.**

**Математическое ожидание** (среднее значение) случайной величины X – мера ее центральной тенденции:

Для дискретной СВ: E(X) = Σ xᵢ·P(X = xᵢ) Для непрерывной СВ: E(X) = ∫x·f(x)dx

Свойства математического ожидания: 1. E(c) = c, где c – константа 2. E(cX) = c·E(X) 3. E(X + Y) = E(X) + E(Y) 4. Если X и Y независимы, то E(X·Y) = E(X)·E(Y)

**Дисперсия** – мера разброса значений случайной величины вокруг среднего: D(X) = E((X - E(X))²) = E(X²) - (E(X))²

Свойства дисперсии: 1. D(c) = 0, где c – константа 2. D(cX) = c²·D(X) 3. Если X и Y независимы, то D(X + Y) = D(X) + D(Y)

Среднеквадратическое отклонение: σ(X) = √D(X)

**Моменты высших порядков:** - Начальный момент порядка k: μₖ = E(Xᵏ) - Центральный момент порядка k: νₖ = E((X - E(X))ᵏ)

Коэффициент асимметрии (показывает несимметричность распределения): γ₁ = ν₃/σ³

Коэффициент эксцесса (показывает “остроту” пика распределения): γ₂ = ν₄/σ⁴ - 3

7**. Моменты многомерных СВ. Ковариация и коэффициент корреляции.**

Для многомерных случайных величин важны совместные моменты, характеризующие взаимосвязь компонент.

**Ковариация** – мера линейной зависимости между двумя случайными величинами: cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(X·Y) - E(X)·E(Y)

Свойства ковариации: 1. cov(X,X) = D(X) 2. cov(X,Y) = cov(Y,X) 3. cov(aX + b, cY + d) = a·c·cov(X,Y) 4. Если X и Y независимы, то cov(X,Y) = 0 (обратное неверно)

**Коэффициент корреляции** – нормированная ковариация: ρ(X,Y) = cov(X,Y) / (σ(X)·σ(Y))

Свойства коэффициента корреляции: 1. -1 ≤ ρ(X,Y) ≤ 1 2. |ρ(X,Y)| = 1 тогда и только тогда, когда между X и Y существует линейная зависимость 3. Если X и Y независимы, то ρ(X,Y) = 0 (обратное неверно)

Корреляционная матрица для многомерной СВ (X₁, X₂, …, Xₙ): R = [ρ(Xᵢ,Xⱼ)]

Эта матрица симметрична и положительно определена, на диагонали стоят единицы.

**8. Характеристические функции (ХФ). ХФ основных распределений.**

**Характеристическая функция** случайной величины X – это математическое ожидание e^(itX): φₓ(t) = E(e^(itX))

**Для дискретной СВ**: φₓ(t) = Σ e^(itxᵢ)·P(X = xᵢ) Для непрерывной СВ: φₓ(t) = ∫e^(itx)·f(x)dx

Свойства характеристических функций: 1. φₓ(0) = 1 2. |φₓ(t)| ≤ 1 3. Если X и Y независимы, то φₓ₊ᵧ(t) = φₓ(t)·φᵧ(t) 4. φₐₓ₊ᵦ(t) = e^(itb)·φₓ(at) 5. Характеристическая функция однозначно определяет распределение

**ХФ основных распределений:** - Нормальное N(μ,σ²): φ(t) = exp(iμt - σ²t²/2) - Пуассона с параметром λ: φ(t) = exp(λ(e^(it) - 1)) - Равномерное на [a,b]: φ(t) = (e^(itb) - e^(ita))/(it(b-a)) - Показательное с параметром λ: φ(t) = λ/(λ-it)

Характеристические функции используются для: - Получения моментов случайной величины - Изучения суммы независимых случайных величин - Доказательства предельных теорем

**9. Неравенство Чебышева и закон больших чисел. Центральная предельная теорема.**

**Неравенство Чебышева** – оценка вероятности отклонения случайной величины от ее математического ожидания: P(|X - E(X)| ≥ ε) ≤ D(X)/ε²

Это неравенство показывает, что вероятность значительного отклонения от среднего мала для случайных величин с небольшой дисперсией.

**Закон больших чисел (ЗБЧ)** – теорема, утверждающая, что среднее арифметическое большого числа независимых случайных величин стабилизируется вокруг их математического ожидания:

**Теорема Чебышева (ЗБЧ):** Если X₁, X₂, … – независимые случайные величины с одинаковым математическим ожиданием μ и ограниченными дисперсиями, то для любого ε > 0: P(|(X₁ + X₂ + … + Xₙ)/n - μ| < ε) → 1 при n → ∞

**Центральная предельная теорема (ЦПТ)** описывает предельное распределение нормированной суммы независимых случайных величин:

Если X₁, X₂, … – независимые случайные величины с одинаковым распределением, E(Xᵢ) = μ, D(Xᵢ) = σ², то: (X₁ + X₂ + … + Xₙ - nμ)/(σ√n) → N(0,1) при n → ∞

То есть, распределение нормированной суммы приближается к стандартному нормальному распределению при большом числе слагаемых.

ЦПТ объясняет, почему нормальное распределение так часто встречается в природе: многие наблюдаемые величины являются результатом сложения большого числа независимых факторов.

**10. Основные понятия математической статистики. Выборка, вариационный ряд, эмпирическая ФР.**

**Математическая статистика** – раздел математики, изучающий методы сбора, обработки и анализа данных.

**Выборка** – набор значений {x₁, x₂, …, xₙ}, полученных в результате наблюдений над случайной величиной X.

Вариационный ряд – выборка, упорядоченная по возрастанию: x₍₁₎ ≤ x₍₂₎ ≤ … ≤ x₍ₙ₎

**Эмпирическая функция распределения (ЭФР)** – оценка функции распределения, построенная по выборке: Fₙ(x) = (число элементов выборки ≤ x) / n

**Свойства ЭФР:** 1. Fₙ(x) – неубывающая функция, принимающая значения от 0 до 1 2. Fₙ(x) – ступенчатая функция с скачками в точках выборки 3. При n → ∞, Fₙ(x) → F(x) по вероятности (закон больших чисел)

**Гистограмма и полигон частот** – графические представления выборки.

**Выборочные характеристики** – оценки параметров распределения: - Выборочное среднее: x̄ = (x₁ + x₂ + … + xₙ)/n - **Выборочная дисперсия:** s² = Σ(xᵢ - x̄)²/(n-1) - Выборочная медиана: med = x₍ₙ₊₁₎₍₂₎, если n нечетное, или (x₍ₙ/₂₎ + x₍ₙ/₂₊₁₎)/2, если n четное

**11. Классификация оценок. Эффективность оценок. Метод моментов. Функция правдоподобия и оценки максимального правдоподобия.**

Точечная оценка параметра θ – функция от выборки θ̂ = θ̂(x₁, x₂, …, xₙ).

Свойства оценок: 1. Несмещенность: E(θ̂) = θ 2. Состоятельность: θ̂ → θ по вероятности при n → ∞ 3. **Эффективность**: минимальная дисперсия среди всех несмещенных оценок

**Метод моментов** – подход к построению оценок, основанный на приравнивании выборочных моментов к теоретическим.

Для одного параметра θ: E(X) = g(θ) → 1/n·Σxᵢ = g(θ̂) → находим θ̂

**Функция правдоподобия** – вероятность (плотность вероятности) получения наблюдаемой выборки как функция от параметров:

Для дискретной СВ: L(θ) = P(X₁ = x₁, X₂ = x₂, …, Xₙ = xₙ) Для непрерывной СВ: L(θ) = f(x₁, x₂, …, xₙ; θ)

Если наблюдения независимы: L(θ) = f(x₁; θ)·f(x₂; θ)·…·f(xₙ; θ)

**Метод максимального правдоподобия (ММП):** оценка θ̂ находится из условия максимума функции правдоподобия L(θ) или логарифмической функции правдоподобия ln L(θ).

Оценки ММП обладают свойствами: 1. Состоятельность 2. Асимптотическая нормальность 3. Асимптотическая эффективность

**12. Проверка статистических гипотез. Уровень значимости и мощность критерия. Ошибки 1-го и 2-го рода.**

**Статистическая гипотеза** – предположение о свойствах распределения случайной величины.

Нулевая гипотеза H₀ – проверяемая гипотеза. Альтернативная гипотеза H₁ – гипотеза, принимаемая при отклонении H₀.

**Критерий проверки** – правило, по которому принимается решение о принятии или отклонении гипотезы.

**Критическая область** – множество значений статистики, при которых H₀ отклоняется.

**Ошибки при проверке гипотез:** - Ошибка 1-го рода (ложное отклонение): отклонение H₀, когда она верна. Вероятность α = P(отклонить H₀ | H₀ верна). - Ошибка 2-го рода (ложное принятие): принятие H₀, когда она неверна. Вероятность β = P(принять H₀ | H₁ верна).

**Уровень значимости α** – допустимая вероятность ошибки 1-го рода, обычно принимают α = 0.05 или 0.01.

**Мощность критерия 1-β** – вероятность отклонения H₀, когда верна H₁.

**P-значение** – вероятность получить значение статистики не менее экстремальное, чем наблюдаемое, при условии, что H₀ верна.

**Алгоритм проверки гипотез:** 1. Формулировка H₀ и H₁ 2. Выбор уровня значимости α 3. Выбор статистики критерия и построение критической области 4. Вычисление наблюдаемого значения статистики критерия 5. Принятие решения: если статистика попадает в критическую область, H₀ отклоняется

**13. Критерий отношения правдоподобия. Критерий согласия Пирсона.**

**Критерий отношения правдоподобия** позволяет проверять гипотезы путем сравнения значений функции правдоподобия.

Отношение правдоподобия: λ = L(θ₀)/L(θ̂), где L(θ₀) - значение при нулевой гипотезе, L(θ̂) - максимальное значение. Чем меньше λ, тем менее правдоподобна нулевая гипотеза.

Статистика -2ln(λ) при больших выборках имеет распределение χ² с числом степеней свободы, равным разности размерностей параметрических пространств.

**Критерий согласия Пирсона (χ²)** проверяет соответствие эмпирического распределения теоретическому: 1. Разбить область значений на интервалы 2. Подсчитать эмпирические и теоретические частоты 3. Вычислить χ² = Σ(nᵢ - n’ᵢ)²/n’ᵢ 4. Сравнить с критическим значением χ²ₐ,ₖ₋ᵣ₋₁

Рекомендуется, чтобы теоретическая частота в каждом интервале была не менее 5.

**14. Доверительные интервалы. Метод построения доверительных интервалов.**

**Доверительный интервал** с уровнем доверия 1-α для параметра θ - это интервал (θ̂₁, θ̂₂), такой что P(θ̂₁ < θ < θ̂₂) = 1-α.

Метод построения: 1. Найти статистику T(X₁,…,Xₙ,θ) с известным распределением 2. Определить значения c₁ и c₂ такие, что P(c₁ < T < c₂) = 1-α 3. Преобразовать неравенство к виду θ̂₁ < θ < θ̂₂

Примеры: - Для среднего нормального распределения с известной σ²: (x̄ - z₁₋ₐ/₂·σ/√n, x̄ + z₁₋ₐ/₂·σ/√n) - С неизвестной σ²: (x̄ - t₁₋ₐ/₂,ₙ₋₁·s/√n, x̄ + t₁₋ₐ/₂,ₙ₋₁·s/√n) - Для дисперсии: ((n-1)s²/χ²ₐ/₂,ₙ₋₁, (n-1)s²/χ²₁₋ₐ/₂,ₙ₋₁)

Длина интервала уменьшается с ростом объема выборки и увеличивается с повышением уровня доверия.

**15. Регрессионный анализ. Линейная регрессия. Метод наименьших квадратов.**

**Регрессионный анализ** исследует зависимость между переменными.

**Линейная модель**: Y = β₀ + β₁X₁ + … + βₚXₚ + ε

**Метод наименьших квадратов (МНК)** минимизирует сумму квадратов отклонений: Q = Σ(yᵢ - (β₀ + β₁x₁ᵢ + … + βₚxₚᵢ))² → min

Для простой линейной регрессии: β̂₁ = Σ(xᵢ - x̄)(yᵢ - ȳ) / Σ(xᵢ - x̄)² β̂₀ = ȳ - β̂₁x̄

Качество модели оценивается с помощью: - Коэффициента детерминации R² = 1 - SSres/SStot - F-статистики для проверки значимости регрессии - t-статистик для проверки значимости коэффициентов

Остатки eᵢ = yᵢ - ŷᵢ должны быть независимы, иметь нормальное распределение с нулевым средним и постоянной дисперсией.

**16. Однофакторный дисперсионный анализ (ANOVA).**

**Дисперсионный анализ** проверяет гипотезы о равенстве средних значений в нескольких группах.

Модель: Yᵢⱼ = μ + αᵢ + εᵢⱼ

Гипотеза H₀: α₁ = α₂ = … = αₖ = 0 (все средние равны)

Принцип: разложение общей изменчивости (SST) на межгрупповую (SSB) и внутригрупповую (SSW): SST = SSB + SSW

F-статистика: F = (SSB/(k-1)) / (SSW/(n-k)) = MSB/MSW

При верности H₀ статистика имеет F-распределение с k-1 и n-k степенями свободы.

Предположения метода: 1. Независимость наблюдений 2. Нормальность распределения в каждой группе 3. Однородность дисперсий (гомоскедастичность)

**17. Анализ временных рядов. Основные компоненты. Методы сглаживания.**

**Временной ряд** - последовательность наблюдений, упорядоченная во времени.

Компоненты: 1. Тренд (T) - долговременная тенденция 2. Сезонность (S) - регулярные колебания 3. Цикличность (C) - колебания с переменным периодом 4. Случайная компонента (ε)

Модели: аддитивные (xt = Tt + St + εt) или мультипликативные (xt = Tt × St × εt)

**Методы сглаживания**: 1. Скользящее среднее: x̄t = (xt-m + … + xt + … + xt+m)/(2m+1) 2. Экспоненциальное сглаживание: St = αxt + (1-α)St-1 3. Метод Хольта-Уинтерса - учитывает тренд и сезонность

**Стационарность** - свойство, при котором статистические характеристики ряда не меняются со временем.

Преобразования нестационарных рядов: - Взятие разностей: Δxt = xt - xt-1 - Логарифмирование - Удаление тренда и сезонности

**18. Модели авторегрессии и скользящего среднего (ARMA, ARIMA).**

**Модель авторегрессии AR(p)**: xt = φ₁xt-1 + … + φₚxt-p + εt - Текущее значение зависит от p предыдущих значений

**Модель скользящего среднего MA(q)**: xt = εt + θ₁εt-1 + … + θqεt-q - Текущее значение зависит от q предыдущих значений случайной составляющей

**Смешанная модель ARMA(p,q)**: xt = φ₁xt-1 + … + φₚxt-p + εt + θ₁εt-1 + … + θqεt-q

**Интегрированная модель ARIMA(p,d,q)**: - Для нестационарных рядов - d - порядок дифференцирования для достижения стационарности

**Методология Бокса-Дженкинса**: 1. Идентификация модели (параметры p, d, q) 2. Оценка параметров 3. Диагностика модели 4. Прогнозирование

**Сезонные модели SARIMA** учитывают сезонные колебания.

**Прогнозирование**: - Точечный прогноз: x̂t+h = E[xt+h | xt, xt-1, …] - Интервальный прогноз: x̂t+h ± z₁₋ₐ/₂ σ̂h

**19. Непараметрические методы статистики. Ранговые критерии.**

**Непараметрические методы** не требуют предположений о виде распределения данных.

Основные методы: 1. **Критерий знаков** - проверка гипотезы о медиане 2. **Критерий Вилкоксона** (знаковых рангов) - для зависимых выборок 3. **Критерий Манна-Уитни** (U-тест) - для двух независимых выборок 4. **Критерий Краскела-Уоллиса** - для нескольких независимых выборок 5. **Критерий Фридмана** - для нескольких зависимых выборок 6. **Коэффициент ранговой корреляции Спирмена** - мера зависимости между переменными 7. **Критерий Колмогорова-Смирнова** - проверка соответствия распределений

Преимущества: - Не требуют нормальности распределения - Применимы к порядковым данным - Устойчивы к выбросам

Недостатки: - Меньшая мощность при выполнении параметрических предпосылок - Потеря информации при переходе к рангам