Программа экзамена

1. Критерии согласия (Пирсона и Колмогорова)
2. Таблицы сопряженности
3. Одновыборочный критерий Стьюдента. Условия его применимости. Парный критерий Стьюдента. Непараметрические парные критерии (знаков и Уилкоксона)
4. Двухвыборочные критерии (Смирнова, Уилкоксона, Стьюдента, Фишера, Z-тест)
5. Дисперсионный анализ (ANOVA): постановка задачи, схема решения, условия применимости
6. Проблема множественного тестирования при проверке статистических гипотез. FWER и FDR.
7. Теория точечного оценивания. Состоятельность, несмещённость и эффективность оценок. Принцип максимального правдоподобия при получении оценок параметров.
8. Байесовы оценки: по максимуму апостериорной вероятности, по среднему и медиане апостериорного распределения. Связь с псевдоотсчётами.
9. Доверительные интервалы и их интерпретация. Получение симметричного и одностороннего доверительного интервала из точечной оценки.
10. Коэффициент корреляции Пирсона, его интерпретация. Связь скореллированности и независимости. Проверка гипотезы о нескореллированности. Корреляция Спирмена.
11. Регрессия. Выбросы и влиятельные значения. Коэффициент детерминации. Разложение суммы квадратов. Формулировки статистических задач, связанных с линейной регрессией.

Регрессия

*x*1, …, *xn* — условия

*y*1, …, *yn* — наблюдения

*y* = *F*(*x*)+ ε — **модель**  
здесь *F*(*x*) — некоторая функция, а ε — «ошибка измерения» (случайная величина)  
(например, если *F*(*x*) = *bx* + *a*, то это «линейная регрессия»)

*ŷi* = *F*(*xi*) — предсказанные значения.

*ri* = *ŷi* – *yi* — невязки.

Функция *F* может быть подобрана из разных соображений. Чаще всего стараются минимизировать сумму квадратов невязок.

Пусть *x*1, …, *xn*; *y*1, …, *yn* — две выборки чисел одинаковой длины.

Гипотеза состоит в том, что значения *y* зависят от значений *x* линейно с точностью до ошибки измерения, которая (ошибка) нормально распределена со средним 0 :  
*y* = *bx* + *a* + ε  
ε ~ N(0; σ2)  
(σ неизвестно, но предполагается постоянным)  
Наша задача — оценить *a* и *b*.

Для этого минимизируем сумму квадратов невязок:  
*Q*(*a, b*) = ∑(*bxi* + *a* – *yi*)2

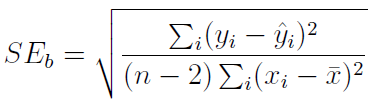
решение:

*b̂ =* ∑ (*xi –x̅*) (*yi –y̅*)/ ∑ (*xi –x̅*)2

*â* = *y̅* – *b̂ x̅*

При линейной регрессии *b̂ = r sy/sx*

где *r*  — коэффициент корреляции между *y*1, …, *yn* и *x*1, …, *xn*  *,*а *sy* и *sx*  — стандартные отклонения выборок *y*1, …, *yn* и *x*1, …, *xn*

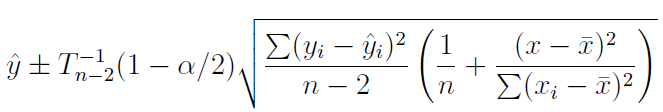
Стандартная ошибка для *b* :

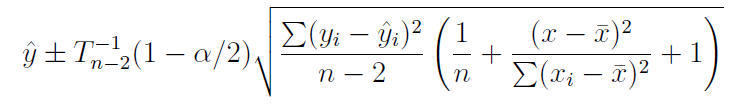
Пусть мы определили параметры линейной регрессии по наблюдениям   
*x*1, …, *xn*; *y*1, …, *yn* — и пусть нам хочется предсказать значение *y* для какого-либо нового *x.*

На самом деле тут **две разные**  задачи:

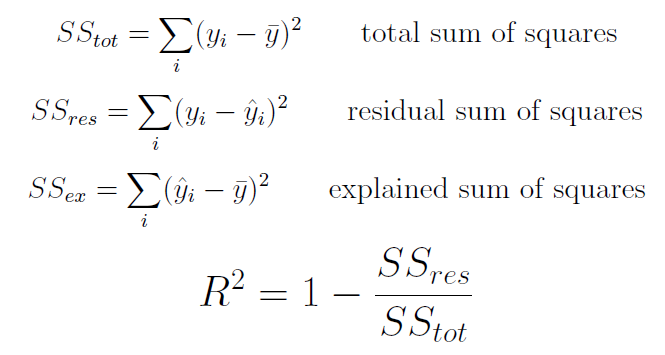
1. Оценить **среднее** значение *y* при таком *x*
2. Оценить границы доверительного интервала значений *y*

Интуитивно понятно, что точечная оценка для среднего значения *y* — это   
*ŷ* = *â + b̂ x*

Но вот границы доверительного интервала для **самого** *y* и для его среднего разные. Для среднего это 

А для самого *y* :

Коэффициент детерминации



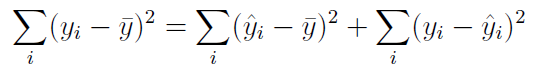
десь *ŷi* = *F*(*xi*) — предсказанные моделью значения, *y̅* — среднее значение *yi*

Величина *R*2 называется «коэффициент детерминации»

*R*2 = 1 означает идеальное соответствие модели наблюдениям (все *ŷi* = *yi*)

Близкое к 0 или тем более отрицательное значение *R*2 означает, что модель ничего не объяснила (остаточная сумма квадратов практически равна полной сумме квадратов)

Разложение суммы квадратов



Если модель линейная: *ŷi* = *bxi* + *a*и **оба коэффициента** получены методом наименьших квадратов (!) ,   
то полная сумма квадратов равна сумме объяснённой суммы квадратов и остаточной суммы квадратов.

*Упражнение: докажите это*

Поэтому *R*2 в этом случае равен доле объяснённой суммы квадратов в полной сумме квадратов (или, что то же самое, доле объяснённой дисперсии в полной дисперсии — дисперсия получается из суммы квадратов делением на *n*):

*R*2 = *SSex*/*SStot*

Кроме того, *R*2 в этом случае равен квадрату коэффициента корреляции между *x*1, …, *xn* и *y*1, …, *yn* .

Как следствие, в этом случае 0 ≤ *R*2 ≤ 1

Выбросы и влиятельные значения

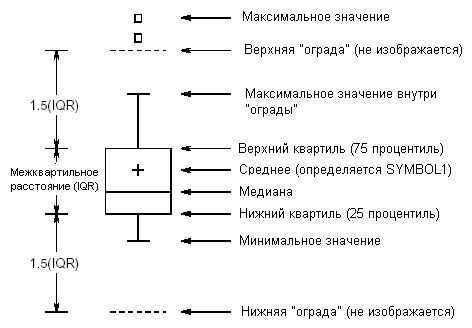
В регрессии выбросом (outlier) называется пара x, y с большой (по модулю) невязкой.

Точнее: обычно выбросами называются те пары, для которых невязки выбиваются из нормального распределения, в то время как для большинства пар они хорошо соответствуют гипотезе нормальности.

Влиятельным (influential) значением называется пара x, y, сильно влияющая на параметры регрессии, то есть такая, после выкидывания которой параметры регрессии существенно меняются.

## Дополнительные вопросы

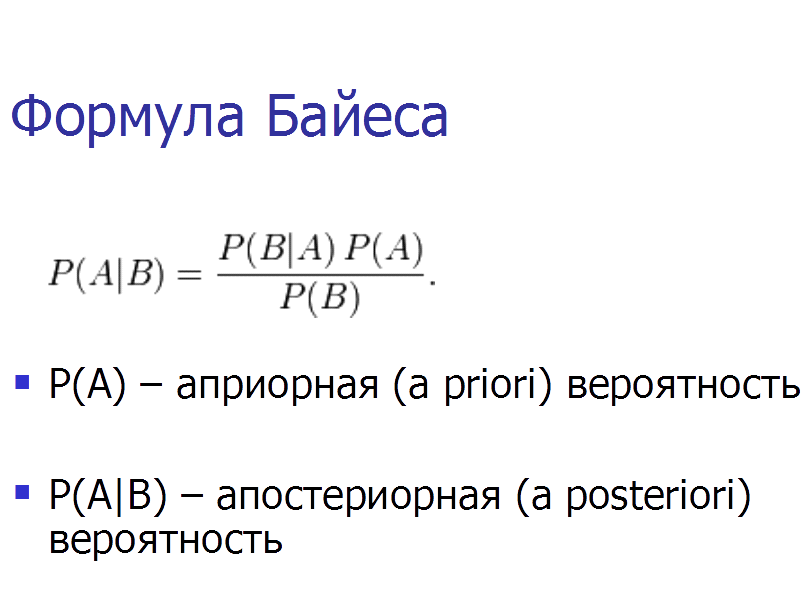
1. **Ящики с усами (box plots), какие элементы что могут означать?**



1. **Гистограмма, график оценки плотности распределения, скрипичная диаграмма**

Скрипичная диаграмма (violin plot) — как бы гибрид ящика с усами и графика оценки плотности

1. **Что такое p-значение?**
2. **Как распределено p-значение при нулевой гипотезе?**
3. **Формула Байеса**

****

1. **Отношение правдоподобия**
2. **Стандартное отклонение и стандартная ошибка**
3. **Выбор между односторонней и двусторонней альтернативами**
4. **Сведение сравнения выборок к таблице сопряжённости**
5. **При одних условиях наблюдалось *m* успехов, а при других — *n* успехов. Как по числам *m* и *n* понять, достоверно ли влияние условий на число успехов?**
6. **Биномиальное распределение (формула, где возникает, матожидание и дисперсия)**

Биномиально распределённая величина = число успехов в *n* независимых испытаниях; параметр *p*=вероятность успеха в одном испытании

1. **Распределение Пуассона (формула, где возникает, матожидание и дисперсия)**

Случайная величина, распределённая по Пуассону = число (достаточно) редких событий за (достаточно большой) промежуток времени или в (достаточно большой) области пространства.   
Имеет один параметр: λ — среднее число событий.

Вероятность наблюдать ровно k событий

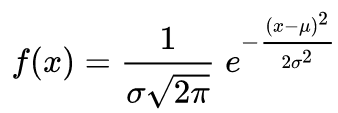
1. **Экспоненциальное распределение (формула для функции распределения, формула для плотности распределения, где возникает, матожидание и дисперсия)**

Функция распределения 

Плотность распределения

1. **Нормальное распределение (формула для плотности распределения, в каких ситуациях возникает, при каких значениях *x* функция распределения *F*(*x*) принимает значения 0,05 и 0,025)**

Функция распределения - не имеет формулы

Плотность распределения 

Смысл параметров: μ — мат. ожидание, σ2 — дисперсия.   
σ часто называют стандартным отклонением

Нормальное распределение возникает везде, где величина представляет собой сумму большого количества элементов, вносящих приблизительно одинаковый вклад (без сильного доминирования небольшого числа из них)

Например:

* длина тела животных одной популяции, как правило, распределена нормально
* ошибки измерений в большинстве экспериментов распределены нормально
* количество крупинок в 1 кг сахарного песка распределено нормально
* число выпадений «орла» при бросании монеты 1 млн. раз распределено нормально

«Правило трёх сигм»: вероятность удалиться от среднего (в заранее заданную сторону)  
 более чем на три стандартных отклонения – около одной тысячной

1. **Перестановочные тесты**
2. **Простая альтернатива и оптимальный критерий Неймана – Пирсона**
3. **Генерация псевдослучайных чисел по заданной функции распределения**
4. **Генерация псевдослучайных чисел по заданной плотности распределения**