Занятие 2. Основы статистики: повторение

20 января 2021

Проверка гипотез. Шаг 1

Статистическая гипотеза – предположение о генеральном параметре, тестируемое на основе данных. Нулевая гипотеза тестируется против альтернативы.

Проверка гипотез. Шаг 1

Статистическая гипотеза – предположение о генеральном параметре, тестируемое на основе данных. Нулевая гипотеза тестируется против альтернативы.

Примеры нулевых гипотез

- $E(X) = 5$
- Подбросим монетку. P(орел) = P(решка) = 0.5

Примеры альтернатив

- $E(X) = 3$; $E(X) > 5$; $E(X) < 5$; $E(X) \neq 5$
- P(орел) = 0.7; P(орел) > P(решка); P(орел) < P(решка); P(орел) ≠ P(решка)

Напоминания

1. Статистическая гипотеза формулируется о ГЕНЕРАЛЬНОМ ПАРАМЕТРЕ, а не о его оценке.

Напоминания

1. Статистическая гипотеза формулируется о ГЕНЕРАЛЬНОМ ПАРАМЕТРЕ, а не о его оценке.
2. Чаще всего используются двусторонние альтернативы. Если Вы все же решили воспользоваться односторонней альтернативой, предварительно посмотрите на оценки необходимых параметров. К примеру, если проверяете гипотезу о равенстве средних, сравните средние в двух выборках, чтобы правильно определиться с лево- или правосторонней альтернативой.

Проверка гипотез. Шаг 2

Далее мы формулируем статистику критерия. Что это такое и зачем она нужна?

Проверка гипотез. Шаг 2

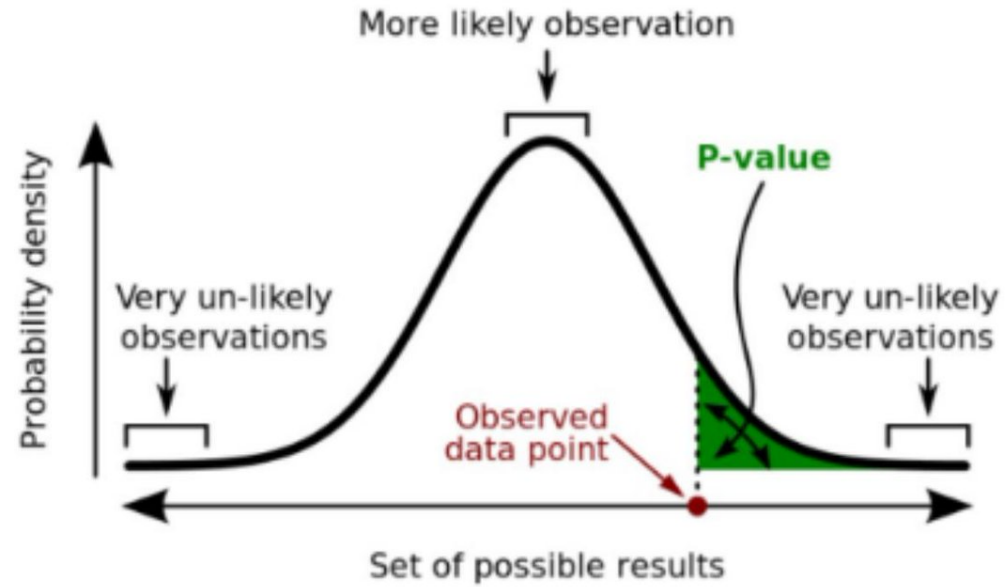
Далее мы формулируем статистику критерия. Что это такое и зачем она нужна?

Ответ

Статистика критерия – функция от выборки, используемая для принятия решения относительно отвержения / неотвержения нулевой гипотезы. К примеру, можно рассмотреть количество выпавших орлов / решек для тестирования, правильная ли монетка.

Проверка гипотез. Шаг 3

Далее мы считаем p-value, или минимальный уровень значимости. Ниже – распределение статистики в условиях верной нулевой гипотезы (H_0).



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Проверка гипотез. Шаг 4

В заключении делаем вывод о H_0.

- Если p-value мало, значит наблюдаемое значение статистики ближе к хвостам распределения («экстремальным» значениям), следовательно, на основании имеющихся данных мы отвергаем нулевую гипотезу в пользу альтернативы.

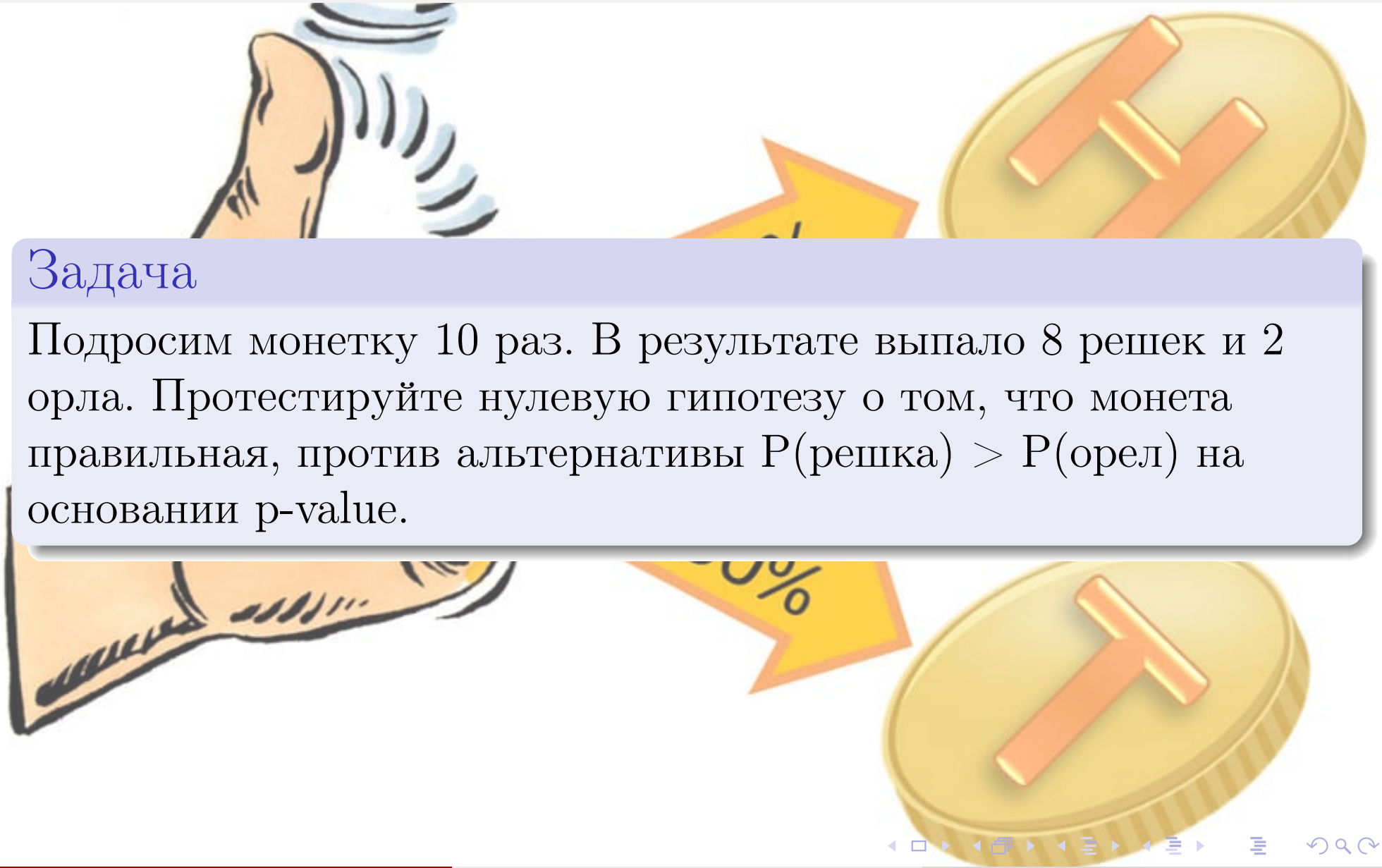
Проверка гипотез. Шаг 4

В заключении делаем вывод о H_0.

- Если p-value мало, значит наблюдаемое значение статистики ближе к хвостам распределения («экстремальным» значениям), следовательно, на основании имеющихся данных мы отвергаем нулевую гипотезу в пользу альтернативы.
- И наоборот, если p-value достаточно велико, значит наблюдаемое значение статистики ближе к центру распределения (характерным значениям), следовательно, на основании имеющихся данных мы не можем отвергать нулевую гипотезу в пользу альтернативы.

Тестирование гипотез: practice makes perfect

Тестирование гипотез: practice makes perfect



Задача

Подросим монетку 10 раз. В результате выпало 8 решек и 2 орла. Протестируйте нулевую гипотезу о том, что монета правильная, против альтернативы P(решка) $>$ P(орел) на основании p-value.

Классическая линейная регрессия

Вопрос

Что дает оценивание регрессии по сравнению с корреляцией?

Классическая линейная регрессия

Вопрос

Что дает оценивание регрессии по сравнению с корреляцией?

Ответ

Корреляция	Регрессионная модель
Сила связи между X и Y	Эффект изменения X на Y (предсказание Y по X)
Линейная взаимосвязь	Можем моделировать нелинейность *
Нет зависимой и независимой переменных	Выбор отклика имеет значение

* N.B.: модели линейные с точки зрения коэффициентов, но при этом могут учитывать нелинейную связь X и Y

Классическая линейная регрессия

Вопрос

Запишите спецификацию линейной регрессии в общем виде.

Классическая линейная регрессия

Вопрос

Запишите спецификацию линейной регрессии в общем виде.

Ответ

$y_i = b_0 + b_1 x_{1i} ... + b_k x_{ki} + e_i$,
где y_i – зависимая переменная (отклик),
b_0 – константа (intercept),
b_1, ..., b_k – коэффициенты при предикторах,
x_i – независимая переменная (предиктор),
e_i – ошибка.

Классическая линейная регрессия

Вопрос

Запишите спецификацию линейной регрессии в общем виде.

Ответ

$y_i = b_0 + b_1 x_{1i} ... + b_k x_{ki} + e_i$,
где y_i – зависимая переменная (отклик),
b_0 – константа (intercept),
b_1, ..., b_k – коэффициенты при предикторах,
x_i – независимая переменная (предиктор),
e_i – ошибка.

$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{1i} ... + \hat{b}_k x_{ki}$ – это предсказанное значение зависимой переменной;
$\hat{e}_i = y_i - \hat{y}_i$, где $\hat{e}_i$ – это остаток (оценка ошибки).

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Объясните основной принцип этого метода.

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Объясните основной принцип этого метода.

Ответ

В соответствии с МНК выбираем такие оценки коэффициентов, при которых линия предсказания наиболее близка к наблюдениям. Математически происходит минимизация суммы квадратов остатков:

$$\min \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x_{1i} ... + \hat{b}_k x_{ki}))^2$$

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Объясните основной принцип этого метода.

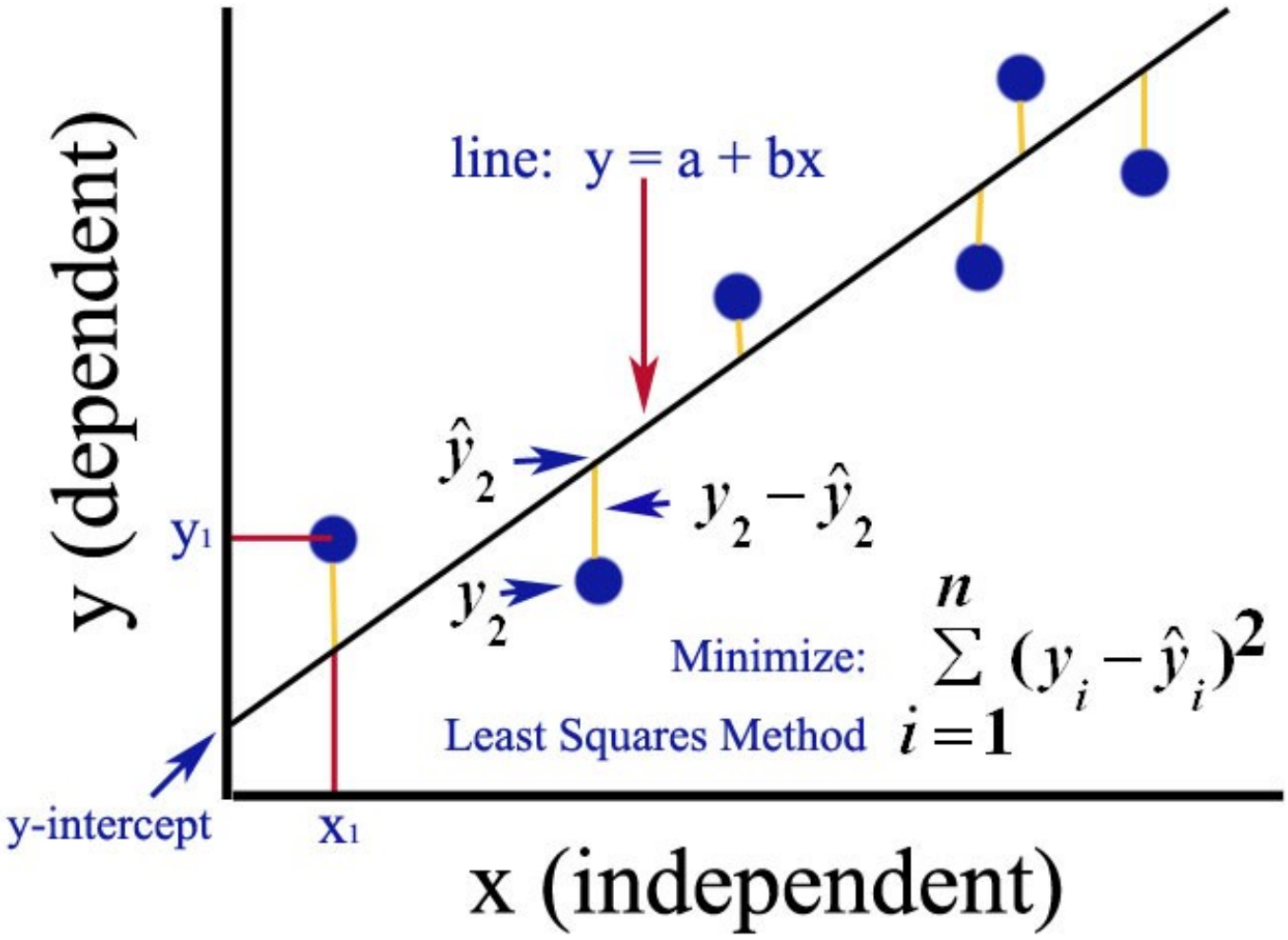
Ответ

В соответствии с МНК выбираем такие оценки коэффициентов, при которых линия предсказания наиболее близка к наблюдениям. Математически происходит минимизация суммы квадратов остатков:

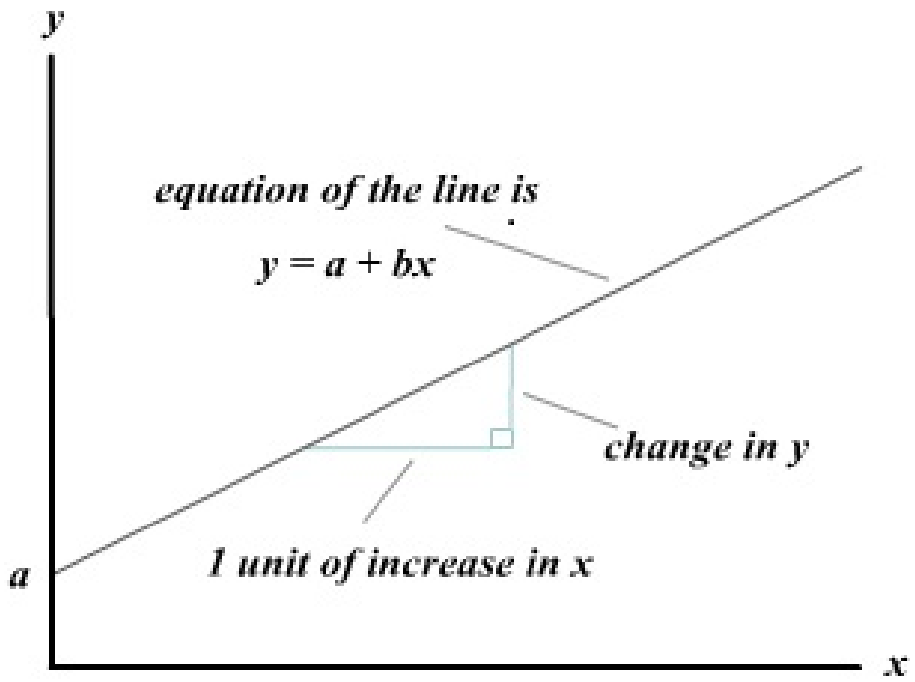
$$\min \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x_{1i} ... + \hat{b}_k x_{ki}))^2$$

Или можем переписать это в таком виде: $\min \sum_{n}^{n} (y_i - \hat{y}_i)^2$

Иллюстрация принципа МНК



Интерпретация оценок коэффициентов



$\hat{b}_0$ (также обозначается как a) – среднее значение отклика при условии равенства предикторов 0.
$\hat{b}_1$ – на сколько в среднем изменяется отклик при увеличении предиктора на единицу измерения при прочих равных.

Желаемые свойства МНК-оценок

Вопрос

Каким требованиям должны соответствовать ошибки в регрессии для получения несмещенных и наиболее эффективных оценок среди класса линейных оценок (BLUE)?

Желаемые свойства МНК-оценок

Вопрос

Каким требованиям должны соответствовать ошибки в регрессии для получения несмещенных и наиболее эффективных оценок среди класса линейных оценок (BLUE)?

Допущения

- Корректная спецификация регрессионной модели

Желаемые свойства МНК-оценок

Вопрос

Каким требованиям должны соответствовать ошибки в регрессии для получения несмещенных и наиболее эффективных оценок среди класса линейных оценок (BLUE)?

Допущения

- Корректная спецификация регрессионной модели
- Нет строгой мультиколлинеарности
- $E(e_i|x_i) = 0$

Желаемые свойства МНК-оценок

Вопрос

Каким требованиям должны соответствовать ошибки в регрессии для получения несмещенных и наиболее эффективных оценок среди класса линейных оценок (BLUE)?

Допущения

- Корректная спецификация регрессионной модели
- Нет строгой мультиколлинеарности
- $E(e_i|x_i) = 0$
- $Var(e_i|x_i) = const$ (гомоскедастичность)

Желаемые свойства МНК-оценок

Вопрос

Каким требованиям должны соответствовать ошибки в регрессии для получения несмещенных и наиболее эффективных оценок среди класса линейных оценок (BLUE)?

Допущения

- Корректная спецификация регрессионной модели
- Нет строгой мультиколлинеарности
- $E(e_i|x_i) = 0$
- $Var(e_i|x_i) = const$ (гомоскедастичность)
- $Cov(e_i, e_j) = 0$ (отсутствие автокорреляции)

Желаемые свойства МНК-оценок

Вопрос

Каким требованиям должны соответствовать ошибки в регрессии для получения несмещенных и наиболее эффективных оценок среди класса линейных оценок (BLUE)?

Допущения

- Корректная спецификация регрессионной модели
- Нет строгой мультиколлинеарности
- $E(e_i|x_i) = 0$
- $Var(e_i|x_i) = const$ (гомоскедастичность)
- $Cov(e_i, e_j) = 0$ (отсутствие автокорреляции)
- $Cov(e_i, x_i) = 0$ (экзогенность)

Ссылки

Источники картинок:

- https://examvictor.com/probability-basic-concepts/
- https://www.econometrics-with-r.org/
- https://elliptigon.com/statistical-significance-and-inference/
- https://bookdown.org/sbikienga/Intro_to_stat_book/