Занятие 1. Основы статистики: повторение

16 января 2024

Вопрос

Что такое статистическая инференция (statistical inference)?
Какие задачи мы решаем в рамках статистики?

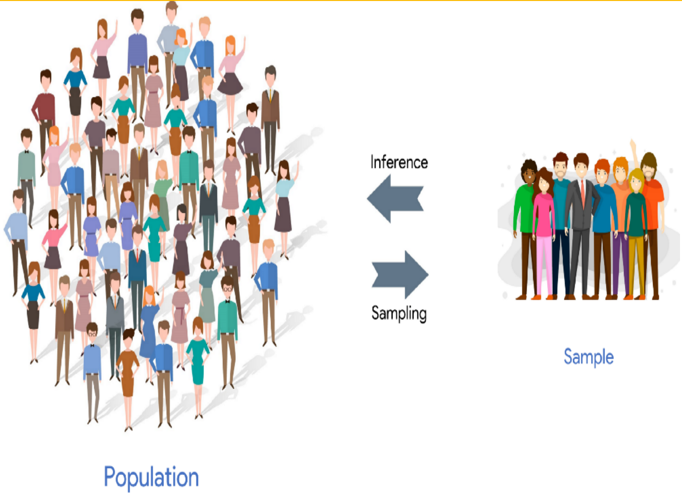
Вопрос

Что такое статистическая инференция (statistical inference)? Какие задачи мы решаем в рамках статистики?

Ответ

Статистическая инференция – перенос результатов с выборки на генеральную совокупность. В соответствии с этим важно разделять генеральные параметры (population parameters) и оценки параметров (sample statistics – estimates of parameters).

Иллюстрация идеи инференции



Вопрос

Каким образом можно осуществить инференцию?

Вопрос

Каким образом можно осуществить инференцию?

Ответ

1. оценивание параметров
   - точечное оценивание (point estimation)
   - интервальное оценивание (interval estimation)
2. проверка гипотез

Вопрос

Приведите примеры генеральных параметров и оценок.

Вопрос

Приведите примеры генеральных параметров и оценок.

Ответ

Параметр	Оценка
Мат. ожидание $E(X)$	Среднее арифметическое
Медиана	Выборочная медиана
Стандартное отклонение: $stdX$	$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

Представление в многомерном пространстве

Представление в многомерном пространстве

$$\begin{pmatrix} EX_1 \\ \dots \\ EX_n \end{pmatrix}$$

Представление в многомерном пространстве

$$\begin{pmatrix} EX_1 \\ \dots \\ EX_n \end{pmatrix}$$

$$\begin{array}{cc} & \begin{array}{cc} X & Y \end{array} \\ \begin{array}{c} X \\ Y \end{array} & \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix} \end{array}$$

Представление в многомерном пространстве

$$\begin{pmatrix} EX_1 \\ \dots \\ EX_n \end{pmatrix}$$

$$\begin{array}{cc} & \begin{matrix} \quad X & \qquad\qquad Y \end{matrix} \\ \begin{matrix} X \\ Y \end{matrix} & \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix} \end{array}$$

$$\begin{array}{cc} & \begin{matrix} \quad X & \qquad\qquad Y \end{matrix} \\ \begin{matrix} X \\ Y \end{matrix} & \begin{pmatrix} 1 & Cor(X,Y) \\ Cor(X,Y) & 1 \end{pmatrix} \end{array}$$

Вопрос

К оценкам с какими свойствами мы стремимся?

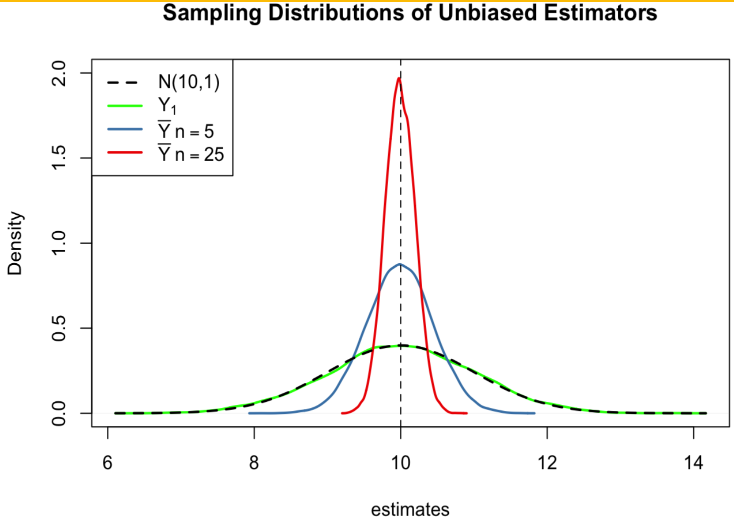
Вопрос

К оценкам с какими свойствами мы стремимся?

Ответ

1. несмещенные (unbiased): $E(\hat{\theta}) = \theta$
2. эффективные (efficient): минимальная вариация
3. состоятельные (consistent): при увеличении выборки $\hat{\theta}$ сходится по вероятности к θ

Несмещенные оценки с разными вариациями



Вопрос

Каковы ограничения точечного оценивания?

Вопрос

Каковы ограничения точечного оценивания?

Ответ

Точечные оценки – это конкретные значения. В этом случае у нас нет информации относительно степени уверенности, насколько мы близки к истинному параметру.

Доверительные интервалы

Какой общий принцип построения доверительных интервалов?

Доверительные интервалы

Какой общий принцип построения доверительных интервалов?

Общий вид доверительного интервала

$$\left[\hat{\theta} - Z_c \times se(\hat{\theta}); \hat{\theta} + Z_c \times se(\hat{\theta})\right], \text{ где}$$

$\hat{\theta}$ – оценка параметра,

$se(\hat{\theta})$ – стандартная ошибка оценки параметра,

Z_c – критическая точка (в частном случае, рассчитанная по стандартному нормальному распределению).

Для того, чтобы вспомнить нормальное распределение и лучше представить критическую точку, перейдем по ссылке

Границы доверительного интервала

1. Давайте вспомним, как связано стандартное и произвольное нормальное распределение:

$$Z = \frac{Y - EY}{stdY}$$

2. Если нам нужно построить 95%-ый ДИ, то найдем такие точки на графике функции плотности Z, между которыми располагаются 95% вероятности (симметрично относительно центра): $Z_{0.025}; Z_{0.975}$

$$-Z_{0.975} \leq \frac{Y - EY}{stdY} \leq Z_{0.975}$$

$$Y - Z_{0.975} \times stdY \leq EY \leq Y + Z_{0.975} \times stdY$$

Доверительный интервал для среднего

Интересующий нас параметр – EX

Пусть $X \sim N(a, \sigma^2)$

Какое распределение в таком случае имеет $\bar{x}$?

Доверительный интервал для среднего

Интересующий нас параметр – EX

Пусть $X \sim N(a, \sigma^2)$

Какое распределение в таком случае имеет $\bar{x}$?

Найдем $E(\bar{x})$: $E(\dfrac{x_1 + x_2 + ... + x_n}{n}) = \dfrac{na}{n} = a$

Доверительный интервал для среднего

Интересующий нас параметр – EX

Пусть $X \sim N(a, \sigma^2)$

Какое распределение в таком случае имеет $\bar{x}$?

Найдем $E(\bar{x})$: $E(\frac{x_1 + x_2 + ... + x_n}{n}) = \frac{na}{n} = a$

Найдем $Var(\bar{x})$: $Var(\bar{x}) = Var(\frac{x_1 + x_2 + ... + x_n}{n}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

Таким образом, $\bar{x} \sim N(a, \frac{\sigma^2}{n})$

Доверительный интервал для среднего:

$$\left[\overline{x} - Z_c \times \frac{\hat{\sigma}}{\sqrt{n}}; \overline{x} + Z_c \times \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

Доверительный интервал для среднего

Пример

Социологическое исследование, выборка которого включала 64 домашних хозяйств, показало, что в среднем в домашней библиотеке жителей страны А находится 36 книг. Выборочная оценка стандартного отклонения составляет 11 книг. Построим 95%-ый доверительный интервал для среднего числа книг домашней библиотеке семьи, проживающей в стране А.

Доверительный интервал для среднего

Пример

Социологическое исследование, выборка которого включала 64 домашних хозяйств, показало, что в среднем в домашней библиотеке жителей страны A находится 36 книг. Выборочная оценка стандартного отклонения составляет 11 книг. Построим 95%-ый доверительный интервал для среднего числа книг домашней библиотеке семьи, проживающей в стране A.

$$se = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{11}{\sqrt{64}} = 1.375$$

$$36 - 1.375 \times 1.96 \leq EX \leq 36 + 1.375 \times 1.96$$

$$33.305 \leq EX \leq 38.695$$

Интерпретация доверительного интервала

Процедура многократного сэмплинга

1. Извлекаем много случайных выборок из заданного распределения одного и того же размера
2. Строим по каждой такой выборке доверительный интервал заданного уровня доверия (возьмем, к примеру, 95%-ый доверительный интервал)
   Интерактивную демонстрацию многократного сэмплинга можно посмотреть здесь.

В таком случае мы с 95% уверенностью можем говорить, что доверительный интервал накрывает истинное значение параметра. Или иными словами, 95% доверительных интервалов будут включать истинное значение параметра.

True/False?

Определим, является ли утверждение верным или ложным.

С увеличением стандартного отклонения при прочих равных условиях длина доверительного интервала увеличивается.

True/False?

Определим, является ли утверждение верным или ложным.

С увеличением стандартного отклонения при прочих равных условиях длина доверительного интервала увеличивается.
ВЕРНОЕ

True/False?

Определим, является ли утверждение верным или ложным.

С увеличением стандартного отклонения при прочих равных условиях длина доверительного интервала увеличивается.
ВЕРНОЕ

С увеличением уровня доверия при прочих равных условиях длина доверительного интервала увеличивается.

True/False?

Определим, является ли утверждение верным или ложным.

С увеличением стандартного отклонения при прочих равных условиях длина доверительного интервала увеличивается. ВЕРНОЕ

С увеличением уровня доверия при прочих равных условиях длина доверительного интервала увеличивается. ВЕРНОЕ

True/False?

Определим, является ли утверждение верным или ложным.

С увеличением стандартного отклонения при прочих равных условиях длина доверительного интервала увеличивается. ВЕРНОЕ

С увеличением уровня доверия при прочих равных условиях длина доверительного интервала увеличивается. ВЕРНОЕ

С увеличением размера выборки при прочих равных условиях длина доверительного интервала увеличивается.

True/False?

Определим, является ли утверждение верным или ложным.

С увеличением стандартного отклонения при прочих равных условиях длина доверительного интервала увеличивается. ВЕРНОЕ

С увеличением уровня доверия при прочих равных условиях длина доверительного интервала увеличивается. ВЕРНОЕ

С увеличением размера выборки при прочих равных условиях длина доверительного интервала увеличивается. ЛОЖНОЕ

True/False?

Определим, является ли утверждение верным или ложным.

В рамках многократного сэмплинга мы извлекаем выборки независимым образом.

True/False?

Определим, является ли утверждение верным или ложным.

В рамках многократного сэмплинга мы извлекаем выборки независимым образом. ВЕРНОЕ

Был построен 95% доверительный интервал для среднего: [5;15]. Мы можем проинтерпретировать его следующим образом: С уверенностью 0.95 оценка среднего лежит в интервале [5;15].

True/False?

Определим, является ли утверждение верным или ложным.

В рамках многократного сэмплинга мы извлекаем выборки независимым образом. ВЕРНОЕ

Был построен 95% доверительный интервал для среднего: [5;15]. Мы можем проинтерпретировать его следующим образом: С уверенностью 0.95 оценка среднего лежит в интервале [5;15]. ЛОЖНОЕ

Доверительный интервал для доли

Пример

Известно, что среди случайно отобранных 200 студентов 40 имеют научные публикации. Построим 99%-ый доверительный интервал для доли студентов, имеющих научные публикации.

Доверительный интервал для доли

Пример

Известно, что среди случайно отобранных 200 студентов 40 имеют научные публикации. Построим 99%-ый доверительный интервал для доли студентов, имеющих научные публикации.

$$\hat{p} = \frac{40}{200} = 0.2$$

$$se = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = \frac{\sqrt{0.2 \times 0.8}}{\sqrt{200}} = 0.028$$

$$0.2 - 0.028 \times 2.58 \leq p \leq 0.2 + 0.028 \times 2.58$$

$$0.128 \leq p \leq 0.272$$

Проверка гипотез. Шаг 1

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

Проверка гипотез. Шаг 1

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

Примеры нулевых гипотез

- $E(X) = 5$
- Подбросим монетку. $\mathrm{P}(\text{орел}) = \mathrm{P}(\text{решка}) = 0.5$

Примеры альтернатив

- $E(X) = 3$; $E(X) > 5$; $E(X) < 5$; $E(X) \neq 5$
- $\mathrm{P}(\text{орел}) = 0.7$; $\mathrm{P}(\text{орел}) > \mathrm{P}(\text{решка})$; $\mathrm{P}(\text{орел}) < \mathrm{P}(\text{решка})$; $\mathrm{P}(\text{орел}) \neq \mathrm{P}(\text{решка})$

Напоминания

1. Статистическая гипотеза формулируется об истинном параметре, а не о его оценке.

Напоминания

1. Статистическая гипотеза формулируется об истинном 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

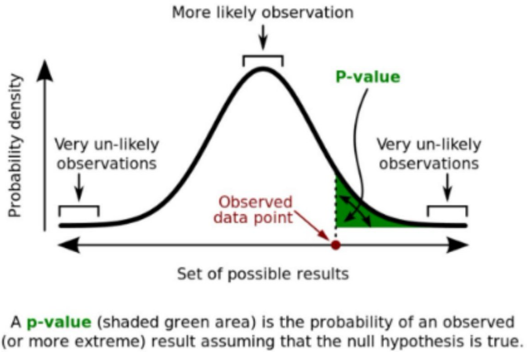
Далее мы формулируем статистику критерия. Что это такое и зачем она нужна?

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).



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.

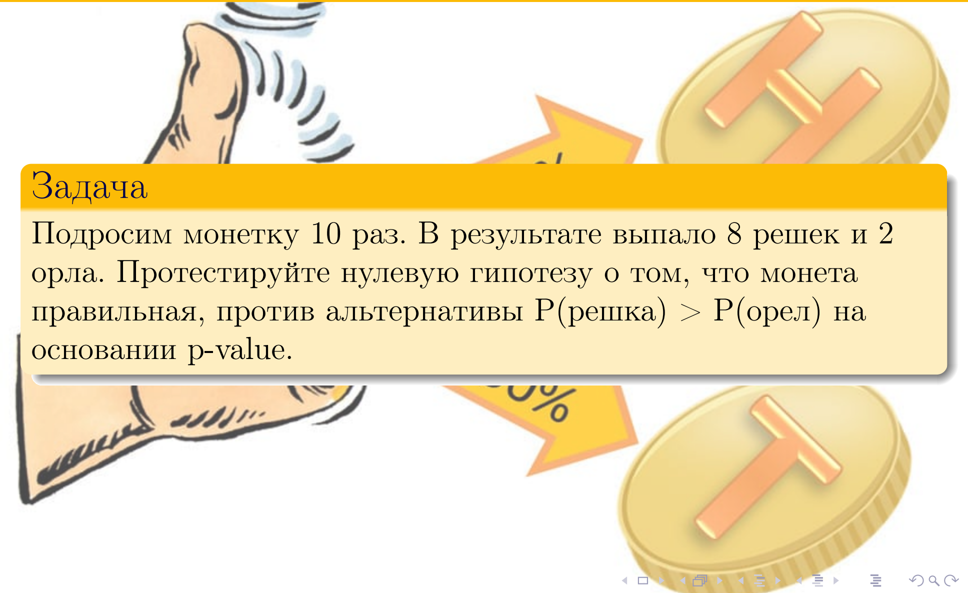
Проверка гипотез. Шаг 4

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.

Формулируем H_0 и H_1 на статистическом языке и определяем статистику критерия

Формулируем H_0 и H_1 на статистическом языке и определяем статистику критерия

$H_0 : P(O) = P(P) = 0.5$
$H_1 : P(P) > P(O)$

Формулируем H_0 и H_1 на статистическом языке и определяем статистику критерия

$H_0 : P(O) = P(P) = 0.5$
$H_1 : P(P) > P(O)$

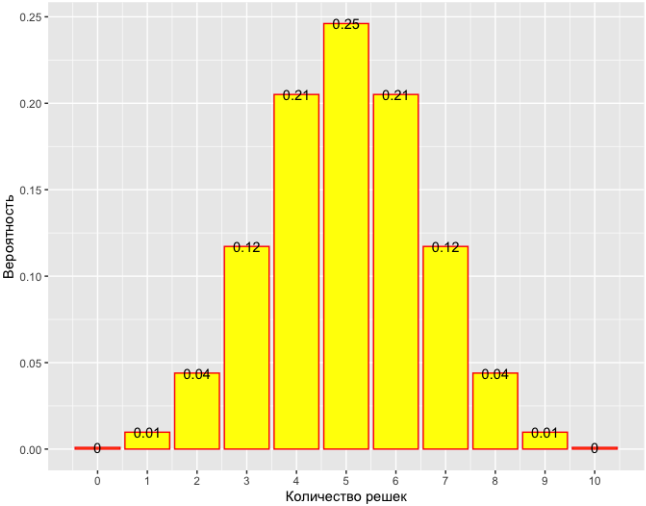
В качестве статистики возьмем случайную величину – количество выпавших решек. Такая статистика имеет биномиальное распределение. Параметры: n = 10; p = 0.5.

Как рассчитать вероятность того, что биномиальная сл. в. принимает определенное значение k:

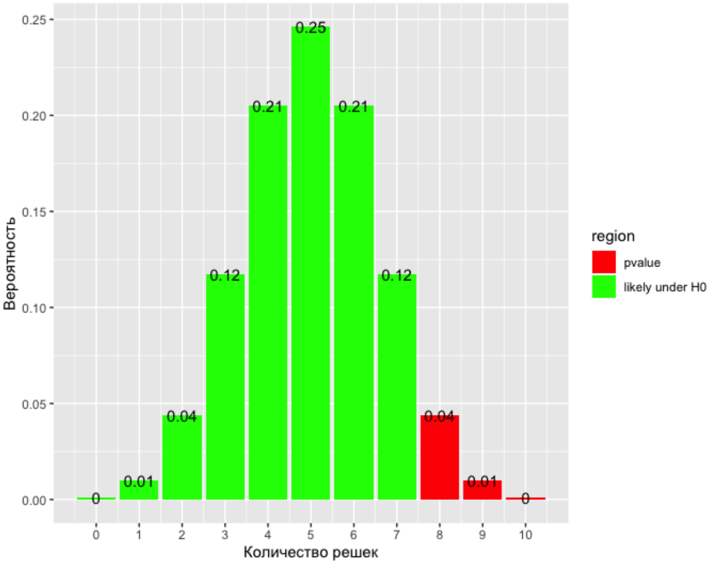
$$P(S = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{(n-k)}$$

Пример: $P(S = 8) = \frac{10!}{8!(10-8)!} \times 0.5^8 \times 0.5^2 \approx 0.044$

Строим распределение статистики критерия при верной H_0



Обозначим p-value на графике



Вывод на основе p-value

Рассчитаем p-value

$P(S = 8) + P(S = 9) + P(S = 10) \approx 0.055$

Сделаем вывод

Значение p-value достаточно большое (превышает конвенциональное значение 0.05) и, следовательно, можно сделать вывод о том, что в данном случае нет оснований отвергнуть нулевую гипотезу (монета правильная). Однако обратите внимание на то, что выборка маленькая.