Математическая статистика

Черновик

Кафедра СМиМ

2019

План

Outline

Математическая статистика

Математическая статистика - наука, разрабатывающая математические методы систематизации и использования статистических (массовых) данных для научных и практических выводов.

Математическая статистика - наука о принятии решений в условиях неопределённости.

Математическая статистика

Задачи

- создание методов сбора, группировки статистических сведений
- анализ статистических данных

Признаки

Для описания изучаемых объектов используются два вида **признаки**

- Значение конкретного признака конкретного объекта можно рассматривать как значение случайной величины
- Количественные - представлены числом
  Рост, масса, прочность, размеры, ...
- Качественные - обозначают порядок или категорию
  место на олимпиаде, цвет, материал стен дома, конструкция крыши

Проблема изучения всех объектов

- Часто невозможно изучать все объекты в силу их большого количества
  Например рост всех людей в определённой стране
- Иногда изучение объекта может быть слишком трудоёмким или затратным
- Изучение некоторых объектов приводит к их разрушению
  например испытания на прочность плит перекрытия
- Поэтому приходится изучать ограниченное число объектов, и по их свойствам делать выводы о всей совокупности

Генеральная совокупность и выборка

- **Выборочная совокупность**, выборка (sample) - совокупность случайно отобранных объектов

  Совокупность всех объектов на основе которых делаются выводы о генеральной совокупности
- **Генеральная совокупность** (statictical population) - совокупность объектов из которых производится выборка.

  Совокупность всех объектов относительно которых делаются выводы
- Объём совокупности - число объектов в этой совокупности

Репрезентативность

- Отбор и изучение только части объектов из генеральной совокупности неизбежно приводит к неточностям в выводах
- Поэтому к выборке предъявляется требование репрезентативности
- **Репрезентативность** - соответствии характеристик выборки характеристикам генеральной совокупности.
- У каждого объекта из генеральной совокупности должен быть один и тот же шанс попасть в выборочною совокупность
- Как правило это достигается *случайным отбором*

Анализ статистических данных

Даже в простых случаях, когда у объекта может быть только один признак анализировать всё множество объектов непосредственно проблематично

Пример. Возраст пассажиров Титаника:

34 47 62 27 22 14 30 26 18 21 46 23 63 47 24 35 21 27 45 55 9 21 48 50 22
22 41 50 24 33 30 18 21 25 39 41 30 45 25 45 60 36 24 27 20 28 10 35 25 36
17 32 18 22 13 18 47 31 60 24 21 29 28 35 32 55 30 24 6 67 49 27 18 2 22 27
25 25 76 29 20 33 43 27 26 16 28 21 18 41 36 18 63 18 1 36 29 12 35 28 17
22 42 24 32 53 43 24 26 26 23 40 10 33 61 28 42 31 22 30 23 60 36 13 24 29
23 42 26 7 26 41 26 48 18 22 27 23 40 15 20 54 36 64 30 37 18 27 40 21 17
40 34 12 61 8 33 6 18 23 0 47 8 25 35 24 33 25 32 17 60 38 42 57 50 30 21 22
21 53 23 40 36 14 21 21 39 20 64 20 18 48 55 45 45 41 22 42 29 1 20 27 24
32 28 19 21 36 21 29 1 30 17 46 26 20 28 40 30 22 23 1 9 2 36 24 30 53 36 26
1 30 29 32 43 24 64 30 1 55 45 18 22 37 55 17 57 19 27 22 26 25 26 33 39 23
12 46 29 21 48 39 19 27 30 32 39 25 18 32 58 16 26 38 24 31 45 25 18 49 0
50 59 30 14 24 31 27 25 22 45 29 21 31 49 44 54 45 22 21 55 5 26 19 24 24
57 21 6 23 51 13 47 29 18 24 48 22 31 30 38 22 17 43 20 23 50 3 37 28 39 38

Анализ статистических данных

Для описания совокупности значений используются подходы похожие на те, что использовались для описания СВ в теории вероятностей

- Числовые характеристики
  - Среднее значение
  - Стандартное отклонение
  - Медиана, квартили
  - Мода
  - ...
- Ряд распределения или эмпирическая функция распределения

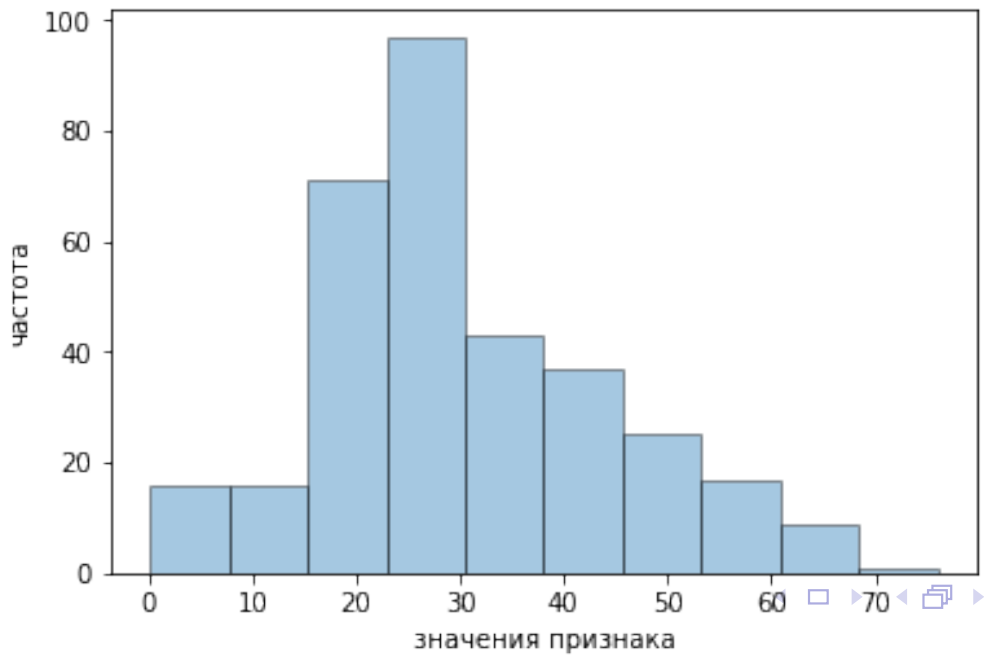
Ряд распределения

- Для построения **интервального ряда распределения** диапазон значений выборки разбивается на равные участки
- Далее для каждого интервала подсчитывается число попавших в него значений - **абсолютная частота**
- Вместо абсолютной частоты иногда используют **относительную частоту** - долю объектов попавших в диапазон

от	0	7.753	15.336	22.919	30.502	38.085	45.668	53.251	60.834	68.417
до	7.753	15.336	22.919	30.502	38.085	45.668	53.251	60.834	68.417	76.0
	16	16	71	97	43	37	25	17	9	1

Гистограмма

- Для графического представления ряда распределения используется *гистограмма*
- Высота столбца определяется количеством (**абсолютной частотой**) значений из выборки попавших в интервал, определяемый шириной столбца
- Вместо частот могут быть использованы **относительные частоты** или проценты - доля объектов выборки попавшая в интервал



Outline

Числовых характеристики статистического распределения

- Выборочные - вычисленные по данным выборки
- Генеральные - вычисляются по данным генеральной совокупности, или (чаще всего) *оцениваются* с помощью аналогичных характеристик выборки

Оценки

Числовые характеристики выборки не описывают в точности генеральную совокупность, поэтому они являются **оценками** характеристик генеральной совокупности

Оценки делаться на две категории

- Точечные
  Представляется одним числом. Например - среднее выборочное. Может заметно отличатся от значения оцениваемой величины.
- Интервальные
  Представляют собой интервал, в которое в заданной вероятностью попадает оцениваемое значение

Интервальная оценка

- θ - оцениваемая величина (характеризует генеральную совокупность)
- θ^* - статистическая характеристика найденная по выборке

Вероятность того, что оцениваемая величина будет отличатся θ от своей оценки θ^* не более чем на δ с ***надёжностью*** γ

$$P(|\theta - \theta^*| < \delta) = \gamma$$

- δ - половина ширины интервала
- $(\theta^* - \delta, \theta^* + \delta)$ - доверительный интервал
- γ - вероятность с которой величина θ попадёт в доверительный интервал

Оценка дисперсии генеральной совокупности

- SD - стандартное отклонение генеральной совокупности
- sd - стандартное отклонение выборочной совокупности

$$sd = \sqrt{\frac{\sum n_i (x_i - \bar{x})^2}{n - 1}}$$

x_i - значения признака из выборки (в случае интервального ряда

берётся середина i-го интервала),
$\bar{x}$ - среднее значение признака в выборке,
n_i - число значений признака x_i (в случае интервального ряда - число значений попавших в i-й интервал)

Доверительный интервал для МО

- Рассмотрим выборку объёмом n
- Пусть исследуемый признак распределён по нормальному закону с известным стандартным отклонением σ для генеральной совокупности
- Требуется получить интервальную оценку математического ожидания $M(X)$ признака в генеральной совокупности с надёжностью γ, если известно его среднее выборочное значение $\bar{x}$

$$P(\bar{x} - \delta < M(X) < \bar{x} + \delta) = \gamma$$

$\delta = z \cdot \frac{\sigma}{\sqrt{n}}$
$\gamma = 2\Phi(z)$

$\Phi(z)$ - значение функции распределения стандартного нормального распределения

Доверительный интервал для МО

- Рассмотрим выборку объёмом n
- Пусть исследуемый признак распределён по нормальному закону с **не**известным стандартным отклонением для генеральной совокупности
- Вместо стандартного ген. совокупности отклонения будем использовать аналогичную величину для выборки - *sd*
- Требуется получить интервальную оценку математического ожидания $M(X)$ признака в генеральной совокупности с надёжностью γ, если известно его среднее выборочное значение $\bar{x}$

$$P(\bar{x} - \delta < M(X) < \bar{x} + \delta) = \gamma$$

$$\delta = t_\gamma \cdot \frac{\sigma}{\sqrt{n}};\ \gamma = 2\int_0^{t_\gamma} S(t, n)dt$$

$S(t, n)$ - значение функции распределения Стьюдента с параметром n

Таблица распределения Стьюдента

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02
df								
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602

t-table.pdf

df = n - 1

Доверительный интервал для МО

Пример

Оценить м.о. генеральной совокупности распределённой по нормальному закону распределения при помощи доверительного интервала с надёжностью 0.95
Известно среднее выборочное - 20.2, выборочное с.к.о. - 0.8

Доверительный интервал для МО

Пример

Оценить м.о. генеральной совокупности распределённой по нормальному закону распределения при помощи доверительного интервала с надёжностью 0.95
Известно среднее выборочное - 20.2, выборочное с.к.о. - 0.8

1. Определим параметр df распределения Стьюдента: df = n - 1
2. Зная значение функции веростности - 0.95 по таблице распределения определим соответствующее значение t_γ
3. Примечание: в таблице (в строке two-tails) вместо γ приводится величина $1 - \gamma$
4. $t_\gamma = 2.131$

Доверительный интервал для МО

Пример

Оценить м.о. генеральной совокупности распределённой по нормальному закону распределения при помощи доверительного интервала с надёжностью 0.95
Известно среднее выборочное - 20.2, выборочное с.к.о. - 0.8

1. Определим параметр df распределения Стьюдента: df = n - 1
2. Зная значение функции веростности - 0.95 по таблице распределения определим соответствующее значение t_γ
3. Примечание: в таблице (в строке two-tails) вместо γ приводится величина $1 - \gamma$
4. $t_\gamma = 2.131$
5. Доверительный интервал $(19.774, 20.686)$

Outline

Статистические гипотезы

Статистическая гипотеза - предположение о виде распределения и о свойствах случайной величины (или нескольких), которое можно опровергнуть или подтвердить применением статистических методов.

Примеры статистических гипотез:

- Математическое ожидание равно 10
- Признак имеет нормальное распределение
- Математические ожидания двух генеральных совокупностей равны

Статистические гипотезы

Зачем?

- В выборках всегда присутствует элемент случайности, поэтому их характеристики не могут в точности соответствовать характеристикам генеральной совокупности
- Нельзя безоговорочно доверять даже средним значениям репрезентативных выборок
- Поэтому все характеристики выборок (числовые характеристики, распределения) можно считать *предположениями* о генеральной совокупности
- А значит появляется необходимость оценивать надёжность этих предположений - статистическая проверка гипотез

Статистические гипотезы

На практике обычно рассматривают две гипотезы - основную и противоположную ей.

- Нулевая гипотеза H_0 - основная гипотеза
- Альтернативная гипотеза H_1
- В результате проверки, нулевая гипотеза либо принимается либо отвергается
- Во втором случае следует автоматическое принятие альтернативной гипотезы

Например, если основная гипотеза формулируется так $H_0 : M(X) = 10$, то альтернативная будет такой $H_1 : M(X) \neq 10$

Статистические критерии

- Для проверки гипотез разного вида и разных условий их применения разработаны статистические критерии
- **Статистический критерий** - строгое математическое правило, по которому принимается или отвергается та или иная статистическая гипотеза с известным уровнем значимости
- Уровень значимости α - вероятность отвергнуть правильную гипотезу
- Обычно уровень значимости выбирается равным 0.05 или меньше

Outline

Гипотеза о равенстве МО двух генеральных совокупностей

- Из двух генеральных совокупностей (X и Y) сделаны выборки объёмом n и m соответственно
- Известны дисперсии генеральных совокупностей $D(X)$ и $D(Y)$
- Средние выборочные значения ($\bar{x}$ и $\bar{y}$) этих выборок близки
- Однако эти значения могли получится близкими случайно, в силу того, что сделаны из выборок
- Требуется определить, равны ли математические ожидания генеральных совокупностей

Гипотеза о равенстве МО двух генеральных совокупностей

1. Сформулируем гипотезы:
   $H_0 : M(X) = M(Y)$
   $H_1 : M(X) \neq M(Y)$
2. Вычислим значение (**значение критерия**) Z[1], которое будем использовать для принятия решения

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/m}}$$

3. Определим критические значения z используя заранее заданный уровень значимости ($\alpha = 0.05$) из выражения $P(Z < z_k) = \Phi(z_k) = \alpha/2$

[1] это значение вычисляется использованием случайных величин $\bar{x}$ и $\bar{y}$ и само является сдучайной величиной

Гипотеза о равенстве МО двух генеральных совокупностей

1. Примем нулевую гипотезу если

$$z < -z_k \text{или} z > +z_k$$

2. Иначе, отклоним нулевую гипотезу и примем альтернативную

Проверка статистических гипотез

- значение критерия Z, по которому принимается решение о принятии или отклонении гипотезы является случайной величиной
- Для разных статистических критериев (например для сравнения дисперсий) используется своя формула для вычисления значения, но везде вычисленное значение - случайная величина
- Распределение случайной величины тоже изменяется от критерия к критерию

Проверка статистических гипотез

p-value

- Программные средства (и модули языков программирования) имеющие средства проверки статистических гипотез могут вместо *значения критерия* выдавать p-value[2]
- **p-value** вычисляется с помощью *значения критерия*
- Правила принятия и отклонения нулевой гипотезы:
  $p - value > \alpha$ - нулевая гипотеза принимается
  $p - value < \alpha$ - нулевая гипотеза отклоняется
  где α - уровень значимости[3]

[2] Подробнее о смысле p-value: https://habr.com/en/company/stepic/blog/250527/

[3] обычно выбирается равным 0.05, 0.01, ...

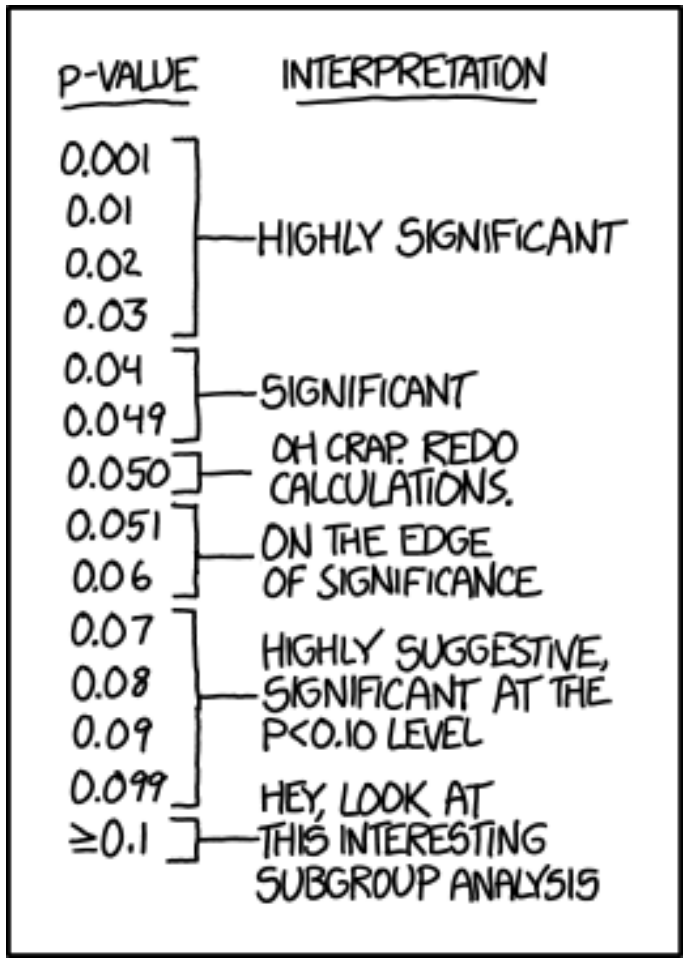
Проверка статистических гипотез

p-value — это вероятность получить такие или более выраженные различия при условии, что в генеральной совокупности никаких различий на самом деле нет.

Чем меньше p-value тем с большей надёжностью можно отклонять нулевую гипотезу

Проверка статистических гипотез

p-value

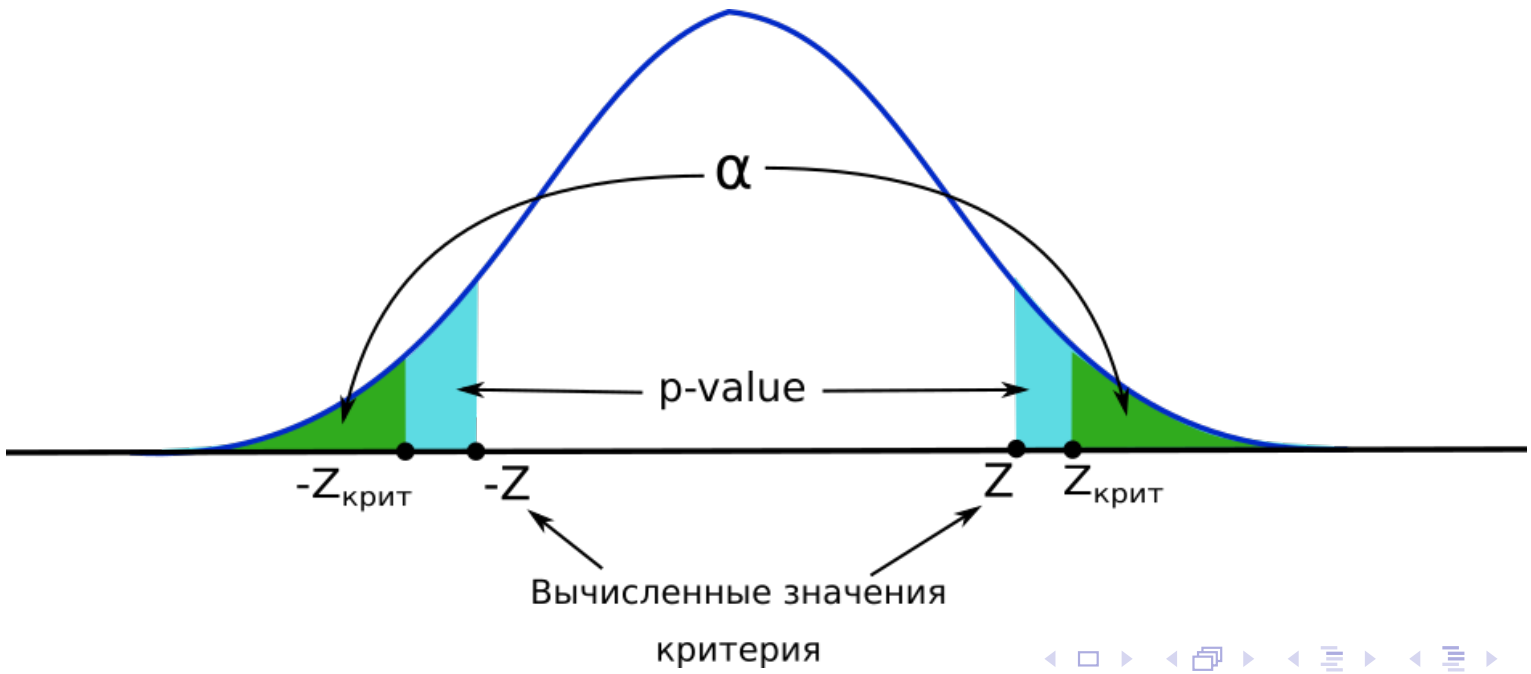


https://xkcd.com/1478/

Проверка статистических гипотез

p-value и значение критерия

- значение критерия - случайная величина
- p-value веростность
- Для проверки гипотезы на равенство чего-либо (математических ожиданий между собой, коф.корреляции нулю и т.п.) p-value и значение критерия связываются следующим образом



Outline

Outline

Источники

- Теория вероятностей и математическая статистика. Гмурман В.Е. biblio-online.ru/book/teoriya-veroyatnostey-i-matematicheskaya-statistika-431095
- Руководство к решению задач по теории вероятностей и математической статистике. В. Е. Гмурман. — 11-е изд., Издательство Юрайт, 2019. — 406 с www.biblio-online.ru/book/02E0C1D3-4EEA-43AA-AA6B-5E25C4991D0

Ссылки

Материалы курса

github.com/VetrovSV/ST