НЕТОЛОГИЯ

Юстина Иванова

Программист, data scientist

Основные статистические тесты и проверка гипотез.

Спикер

НЕТОЛОГИЯ

Юстина Иванова,
Data scientist по
Компьютерному зрению
в компании ОЦРВ,
Выпускница МГТУ им. Баумана
Магистр по Artificial Intelligence
В University of Southampton

Повторение

Доверительный интервал — интервал, в котором лежит p% данных.

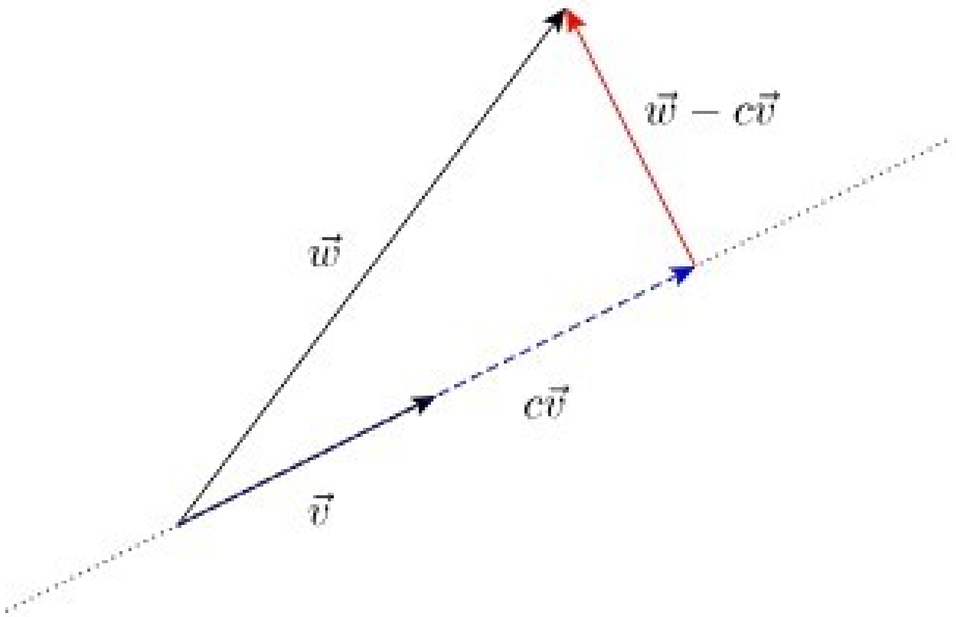
Правило трех сигм.

Виды распределений: дискретное и непрерывное.

Скалярное произведение векторов и проецирование вектора.

Линейная регрессия и классификационный анализ.

Проецирование данных на вектор

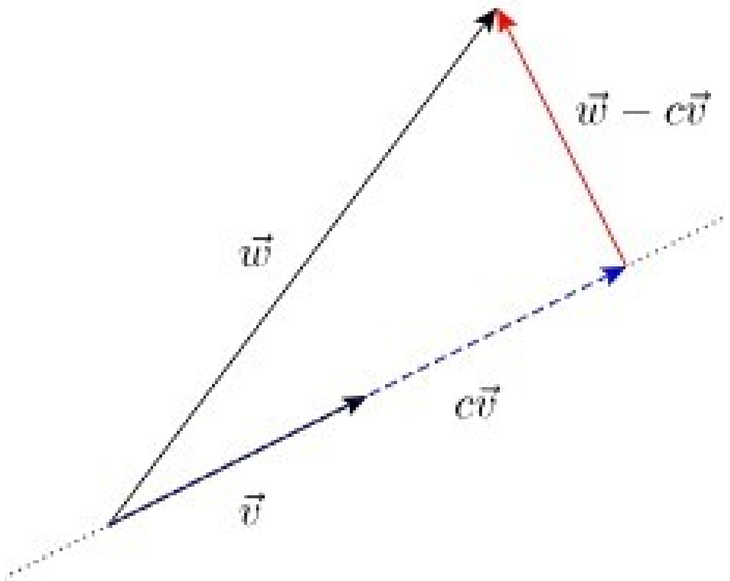


Чтобы посчитать расстояние между точкой и прямой, необходимо знать как проецировать вектор на прямую.

c**v** = np.dot(**w, v**)/np.dot(**v,v**)***v**.

http://qaru.site/questions/15739903/project-vector-w-onto-vector-v-and-draw-perpendicular-line-preparation-for-pca

Скалярное произведение



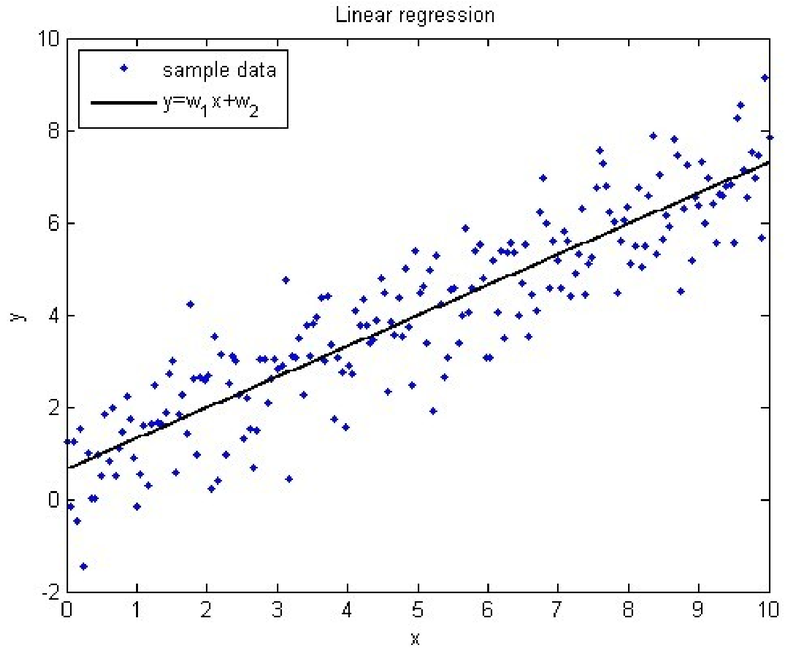
Необходимо для выполнения проецирования данных на вектор.

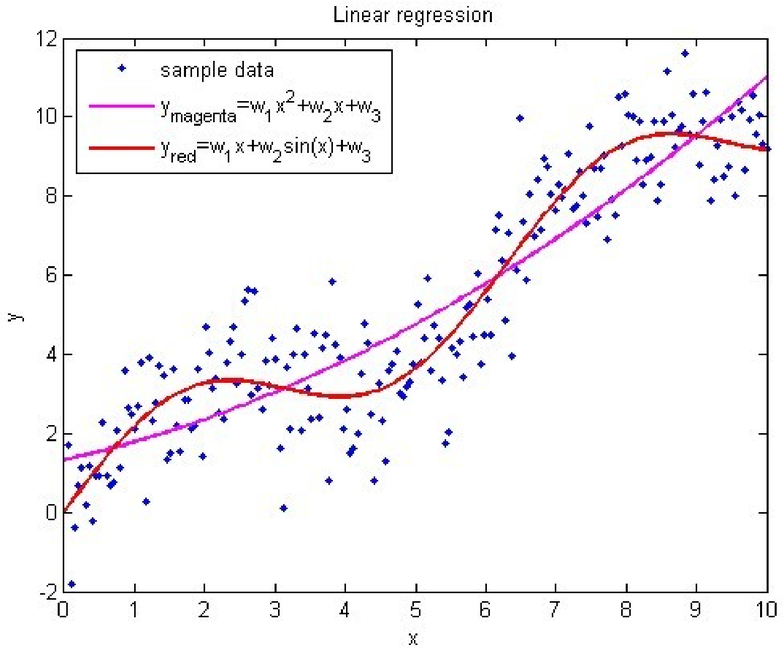
np.dot(w,v) = w1*w2 + v1*v2
Где w = (w1, w2)
v = (v1, v2)

c**v** = np.dot(**w, v**)/np.dot(**v,v**)*****v**.

http://qaru.site/questions/15739903/project-vector-w-onto-vector-v-and-draw-perpendicular-line-preparation-for-pca

Линейная регрессия





Для заданного пространства данных найти уравнение прямой (или кривой), минимизирующую сумму расстояний от точек до нее.

http://www.machinelearning.ru/wiki/index.php?title=%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7

Квантиль

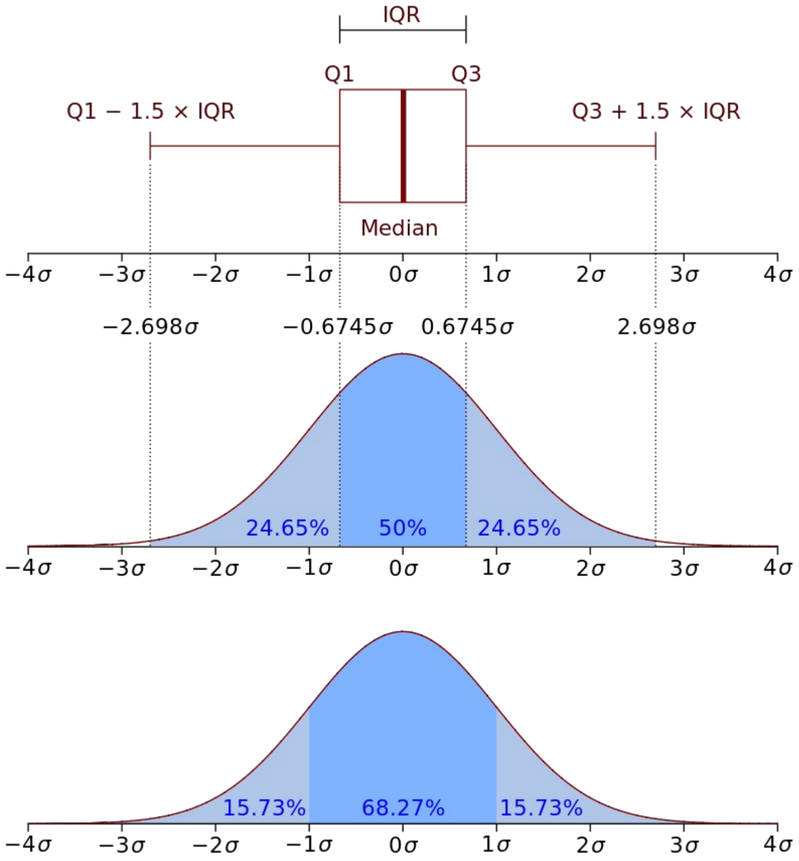
0,25-квантиль называется первым (или нижним) квартúлем;

0,5-квантиль называется медианой или вторым квартúлем;

0,75-квантиль называется третьим (или верхним) квартúлем.

Интерквартúльным размахом называется разность между третьим и первым квартилями, то есть x_{0,75}-x_{0,25}. Интерквартильный размах является характеристикой разброса распределения величины и является аналогом дисперсии. Вместе, медиана и интерквартильный размах могут быть использованы вместо математического ожидания и дисперсии в случае распределений с большими выбросами, либо при невозможности вычисления последних.

Значение квартиля

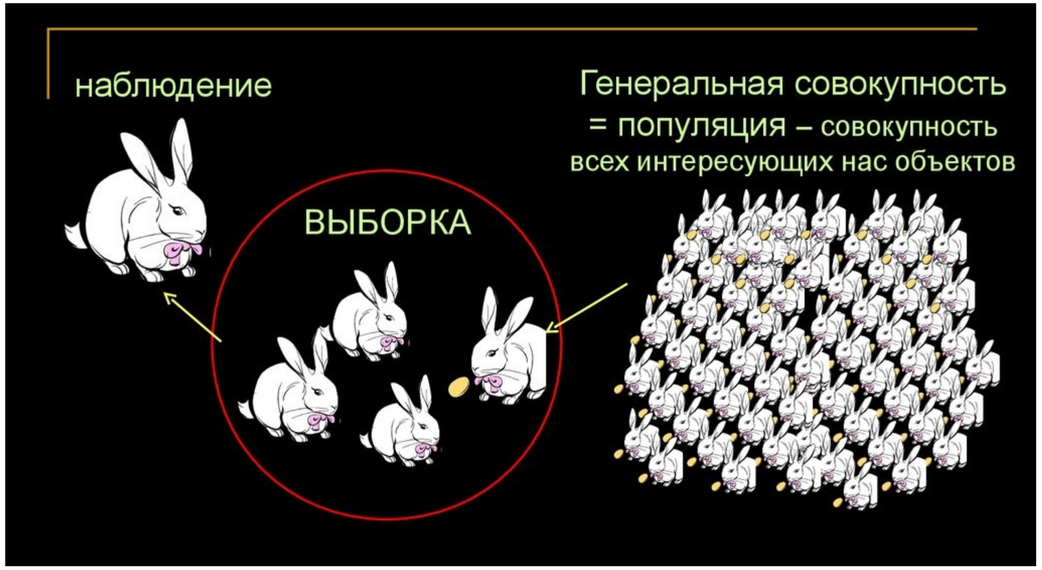


Предоставляют важную информацию о структуре **вариационного** (колонку таблицы) ряда признака. Вместе с медианой они делят вариационный ряд на 4 равные части. Квартилей две, их обозначают символами Q, верхняя и нижняя квартиль. 25% значений меньше, чем нижняя квартиль, 75% значений меньше, чем верхняя квартиль.

https://en.wikipedia.org/wiki/Quartile

Генеральная совокупность и выборка.

Генеральная совокупность — множество всех объектов, обладающих изучаемым признаком.



На основе свойств выборки делаем заключение о свойствах генеральной совокупности.

https://www.mindomo.com/es/mindmap/mind-map-ee84f2f780eb465fa8b31f9b7c59bf1b

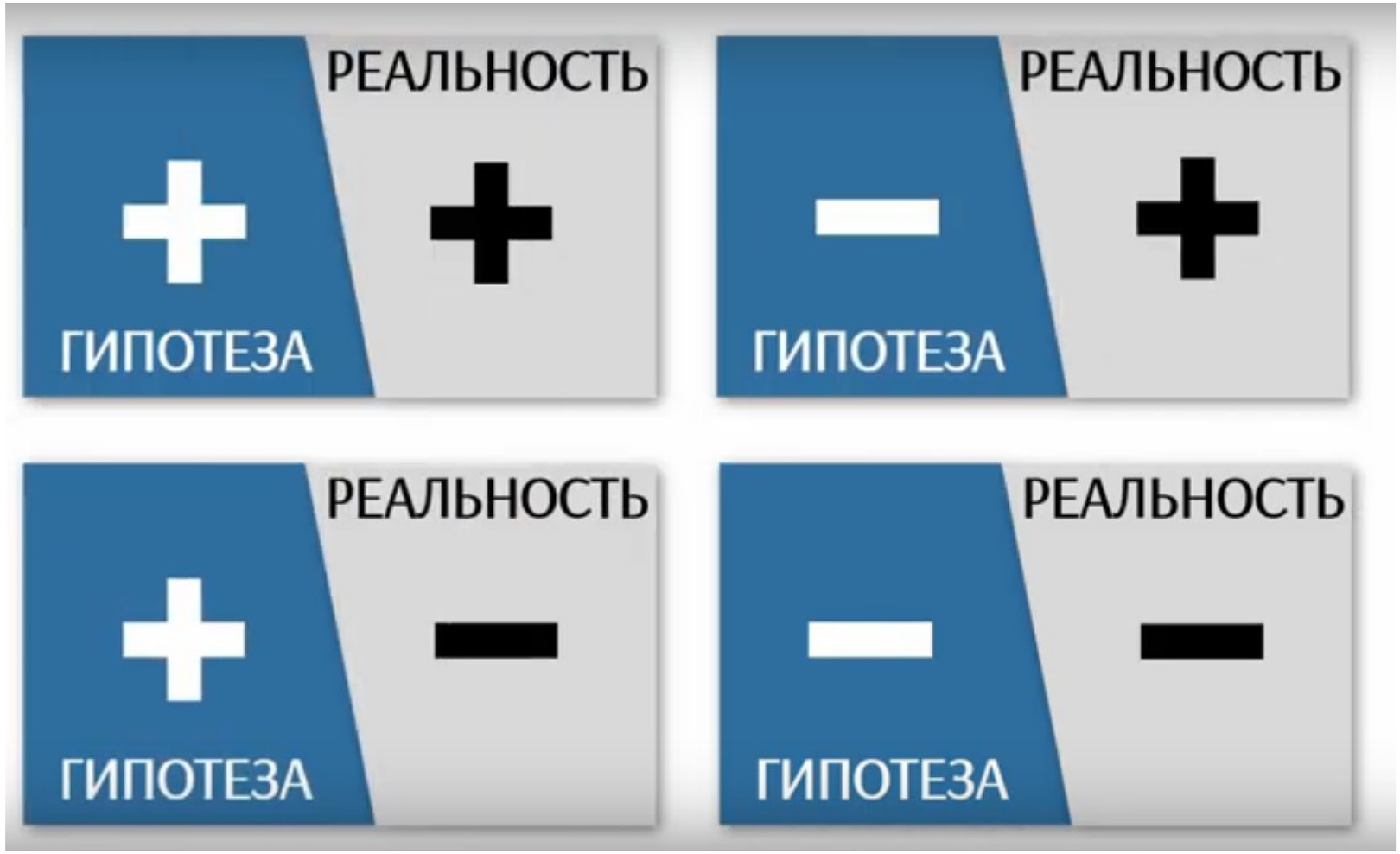
Статистические гипотезы о данных

Выборочная совокупность — множество всех объектов, отобранных случайно из генеральной совокупности для изучения.

Нулевая гипотеза (H0)— гипотеза о сходстве

Альтернативная гипотеза, конкурирующая, (H1)— гипотеза о различиях

Статистические гипотезы о данных



https://www.youtube.com/watch?v=4eyEp_NTXAU

Пример: тест на наличие болезни.

	Disease present	Disease absent
Positive	**a** True positive	**b** False positive
Negative	**c** False negative	**d** True negative

https://www.nejm.org/doi/full/10.1056/NEJM199908193410823

Статистические гипотезы о данных

Ошибка 1 рода:
Вероятность отвергнуть гипотезу,
Но в действительности она верна
Alpha — вероятность ошибки.
Критически значимый уровень
alpha = 0.05

Ошибка 2 рода:
Вероятность принять гипотезу,
Но в действительности она неверна
beta — вероятность ошибки.
Мощность исследования = 1-beta.

https://www.youtube.com/watch?v=4eyEp_NTXAU

Статистически значимость

СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ (ЗНАЧЕНИЕ P) – РАСЧЕТНАЯ ВЕРОЯТНОСТЬ ОШИБКИ ПЕРВОГО РОДА, КОТОРАЯ РАССЧИТЫВАЕТСЯ С ПОМОЩЬЮ РАЗЛИЧНЫХ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ

$P < 0,05$

Подсчитывается с помощью разных критериев.

A/B тесты

A/B тестирование — это мощный маркетинговый инструмент для повышения эффективности работы вашего интернет-ресурса.

Ниже на картинках приведены примеры распределения значений показателя в сегментах.

https://www.dataquest.io/blog/a-b-testing-the-definitive-guide-to-improving-your-product/

Пример A/B теста: Wallmonkeys

Компания WallMonkeys решила оптимизировать веб-сайт на клики и конверсию.

https://www.crazyegg.com/blog/ab-testing-examples/

Пример A/B теста: WallMonkeys

1 тест: 27% кликов.

2 тест: 550% кликов

https://www.crazyegg.com/case-studies

Виды статистических критериев

Критерии согласия -
проверка на согласие подразумевает проверку предположения о том, что исследуемая случайная величина подчиняется предполагаемому закону.

Параметрические критерии -
группа статистических критериев,
которые включают в расчет параметры вероятностного распределения признака (средние и дисперсии).

Непараметрические критерии -
группа статистических критериев,
которые не включают в расчёт параметры вероятностного распределения и основаны на оперировании частотами или рангами.

Параметрическая — непараметрическая гипотеза

Параметрические критерии - группа статистических критериев, которые включают в расчет параметры вероятностного распределения признака (средние и дисперсии).

t-критерий Стьюдента
Критерий Фишера
Критерий отношения правдоподобия
Критерий Романовского

https://medium.com/statistics-experiments/%D1%87%D1%82%D0%BE-%D0%B1%D1%83%D0%B4%D0%B5%D1%82-%D0%B5%D1%81%D0%BB%D0%B8-%D0%B8%D1%81%D0%BF%D0%BE%D0%BB%D1%8C%D0%B7%D0%BE%D0%B2%D0%B0%D1%82%D1%8C-%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9-%D0%BA%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9-%D0%BD%D0%B0-%D0%BD%D0%B5%D0%BD%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE-%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%BD%D0%BE%D0%B9-%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B5-94be5d2afaa7

Параметрическая — непараметрическая гипотеза

Непараметрические критерии
Группа статистических критериев, которые не включают в расчёт параметры вероятностного распределения и основаны на оперировании частотами или рангами.

Q-критерий Розенбаума
U-критерий Манна — Уитни
Критерий Уилкоксона
Критерий Пирсона
Критерий Колмогорова — Смирнова

https://medium.com/statistics-experiments/%D1%87%D1%82%D0%BE-%D0%B1%D1%83%D0%B4%D0%B5%D1%82-%D0%B5%D1%81%D0%BB%D0%B8-%D0%B8%D1%81%D0%BF%D0%BE%D0%BB%D1%8C%D0%B7%D0%BE%D0%B2%D0%B0%D1%82%D1%8C-%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9-%D0%BA%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9-%D0%BD%D0%B0-%D0%BD%D0%B5%D0%BD%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE-%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%BD%D0%BE%D0%B9-%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B5-94be5d2afaa7

НЕТОЛОГИЯ

Распределение Стьюдента

Мы хотим сгенерировать нормальное распределение, но по некоторым причинам не можем вычислить среднеквадратичное отклонение (например, выборка маленькая). Мы можем найти выборочное среднее и выборочную дисперсию по выборке.

Пусть $x_1, ...x_n$ — выборка размером n

Выборочное среднее $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Выборочная дисперсия $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Распределение Стьюдента

Случайная величина t имеет распределение Стьюдента с n^{-1} степенями свободы, где n — размер выборки.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Данный критерий был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны (руководство Гиннесса считало таковой использование статистического аппарата в своей работе), статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент).

Проверка гипотез методом Стьюдента

Владимир Савельев «Статистика и котики»

Тесты для проверки гипотез: тест Фишера

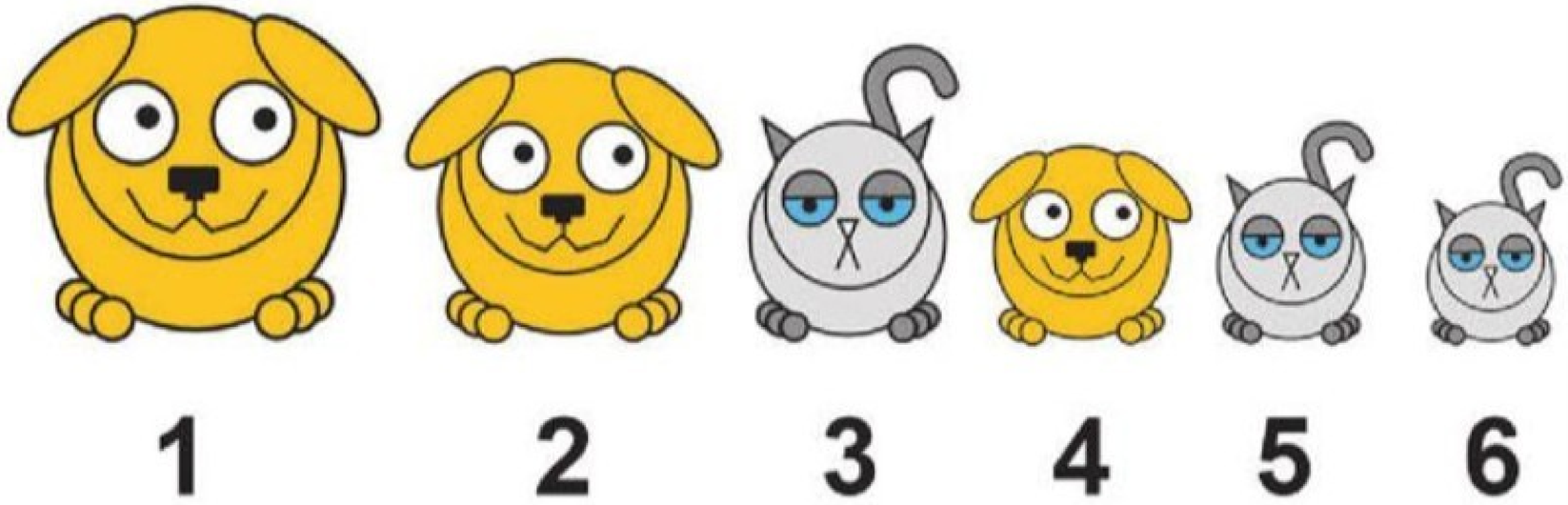
Метод Стьюдента чувствителен к выбросам => параметрический метод: мы можем посмотреть, являются ли песики более разнообразными по размеру, чем котики, или же нет. Для этого мы можем воспользоваться F-критерием равенства дисперсий Фишера,который укажет нам, насколько различаются между собой эти показатели.

Владимир Савельев «Статистика и котики»

Проверка гипотез методом Манна-Уитни

Метод Стьюдента чувствителен к выбросам => параметрический метод:

Чтобы рассчитать критерий Манна-Уитни, необходимо выстроить всех песиков и котиков в один ряд, от самого мелкого к самому крупному, и назначить им ранги. Самому большому зверьку достанется первый ранг, а самому маленькому–последний.После этого мы снова делим их на две группы и считаем суммы рангов отдельно для песиков и для котиков. Общая логика такова: чем сильнее будут различаться эти суммы, тем больше различаются песики и котики.



Владимир Савельев «Статистика и котики»

Вопросы?

Контакты спикера:
yustiks@gmail.com