**Лекция №6 Сравнение долей. Построение доверительных интервалов.**

**Цель лекции:**

* Научиться строить доверительные интервалы для средних значений и их разности
* Выполнить построение доверительных интервалов для доли и разности долей
* Научиться сравнивать доли
* Применить биномиальное распределение при работе с выборками маленьких объемов.

**Материал прошлого урока:**

На прошлом вебинаре мы разбирали тестирование гипотез. Мы научились сравнивать средние арифметические генеральной совокупности, используя Z и t критерии. Мы изучили общий алгоритм тестирования гипотез, который применим и к другим критериям. Сегодня мы продолжаем знакомство с тестированием гипотез, но уже для долей, а также рассмотрим построение доверительных интервалов для среднего и для доли.

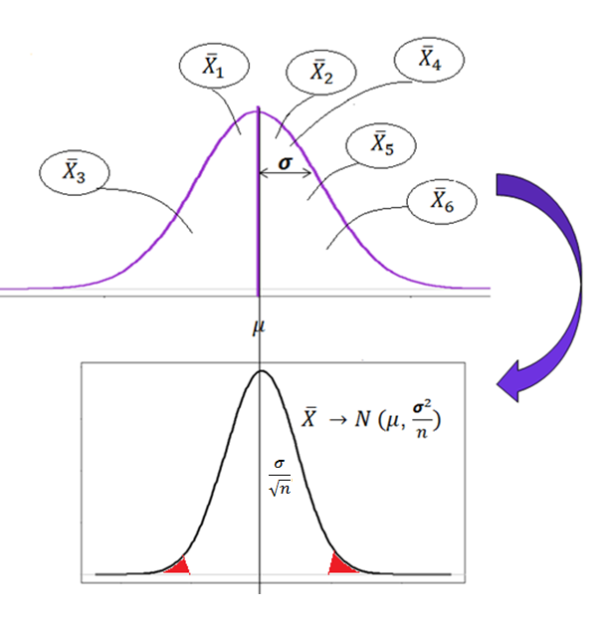
**План урока:**

1. Доверительный интервал для средних арифметических
2. Интервальная оценка для разности средних арифметических
3. Доверительный интервал для доли
4. Маленькие объемы выборок
5. Сравнение долей
6. Интервал для разности долей

**Доверительный интервал для средних арифметических**

Мы научились сравнивать средние арифметические. Т.е. статистика дает нам метод, который позволяет понять, а обнаружены ли статистически значимые различия или нет между средними значениями. Еще один статистический метод – это построение доверительных интервалов. Данный метод позволяет давать интервальную оценку для среднего генеральной совокупности. Ведь, напоминаю, в реальной жизни у нас нет доступа ко всей генеральной совокупности, поэтому мы берем выборку и по ней уже можем рассчитать точечную оценку истинного среднего арифметического. Но это не очень надежно. Потому что наша средняя выборочная это случайная величина, мы можем получить, как слишком высокую точечную оценку, так и слишком низкую точечную оценку истинного среднего арифметического.

Мы помним, что средняя выборочная стремится к нормальному распределению с средним арифметическим, равным среднему арифметическому генеральной совокупности, откуда была взята выборка, и с дисперсией, равной дисперсии генеральной совокупности, деленой на объем выборки . Давайте взглянем на рисунок ниже:

****

Мы видим на этом распределении красные хвостики, которые показывают долю очень маленьких средних арифметических и долю очень больших средних арифметических. Для 95% интервала эти хвостики будут площадью по 0.025 или 2,5 % значений средних по выборкам будут очень маленькими и 2,5% будут очень большими. Т.е. при построении 95% интервала для этих средних, которые попадают в эти области, интервалы не захватят истинного среднего арифметического. Мы с вами подошли к идее доверительного интервала. Давайте поясним это.

Во-первых, доверительный интервал можно представить таким образом:

(1-α) = 0.95 , т.е. α =0.05

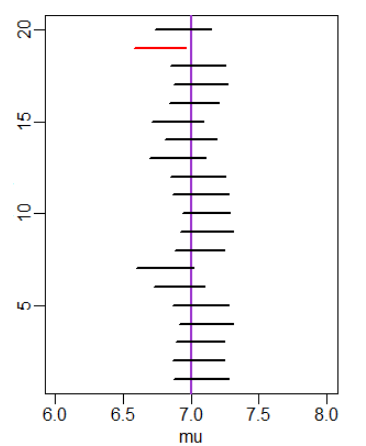
Так что же все-таки означает 95% доверительный интервал? Это означает, что если мы возьмем, к примеру, 100 случайных выборок одинакового объема, посчитаем по ним средние арифметические и построим 95% доверительный интервал для каждого среднего, то получится, что приблизительно 95 интервалов из 100 захватят истинное среднее арифметическое генеральной совокупности. Т.е. 95% интервал означает, что с вероятностью 95% интервал захватит истинное генеральной совокупности. Т.к. это доля интервалов, захвативших среднее генеральной совокупности, а доля – это и есть вероятность. И обратите внимание, что это не то же самое, как, если бы мы сказали, что с вероятностью 95% истинное генеральной совокупности попадет в доверительный интервал.

На рисунке ниже изображены 20 интервалов. Красным показан один интервал, который не захватывает истинного среднего. 1 из 20 , составляет 5% от 20, т.е. 95% интервалов (19 штук) захватили истинное среднее арифметическое генеральной совокупности.

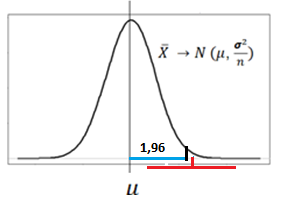
Общая формула для нахождения интервала для среднего арифметического

\* при известной генеральной совокупности, и

\* , если неизвестна, тогда вычисляем ее по выборке, используя формулу для несмещенного стандартного отклонения, или функция в Python std (x, ddf =1).

Для α = 5% = 1.96. Что показывает 1.96? Это показывает в скольких стандартных ошибках ( так называется стандартное отклонение для среднего выборочного) лежит среднее по выборке от истинного среднего. Т.е., если среднее выборочное будет лежать дальше, чем на расстоянии 1,96 стандартных ошибок ( 1.96\* ), иными словами, попадать в красные области, а интервал при этом же интервал рассчитывается как \* , то он у нас просто не дотянется до истинного среднего арифметического.

Посмотрите на рисунок ниже. Красный интервал – это интервал для среднего, которое попало в те самые 2,5% очень высоких значений. Голубым цветом обозначена длина, равная 1.96 стандартных ошибок. Если этот голубой отрезок отложить от красной вертикальной линии, то он не дотянется до истинного среднего.



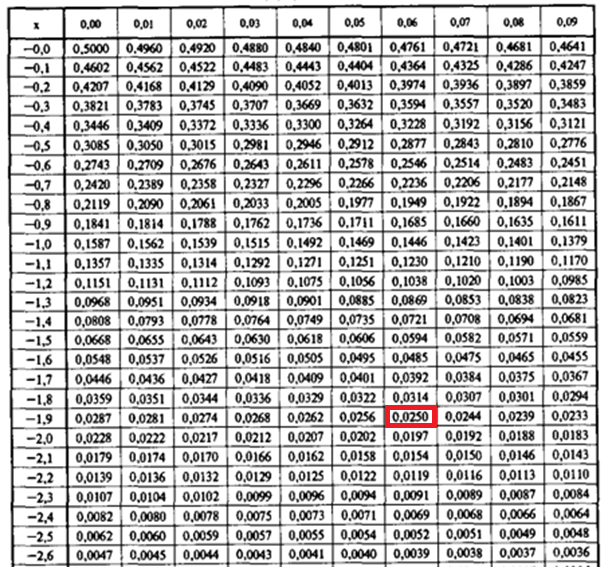
Давайте рассмотрим следующую задачу:

Известно, что генеральная совокупность распределена нормально со средним квадратичным 5. Найти доверительный интервал для оценки среднего арифметического с надежностью 0,95, если выборочная средняя М =24.15, а объем выборки 100.

Какие наши рассуждения? Я всегда рекомендую начинать прямо с вопроса задачи. В задаче просят найти доверительный интервал для среднего арифметического. Т.е. сначала запишем формулу для нахождения доверительного интервала, чтобы она уже была перед глазами, а потом будем думать, что и откуда находить. Поскольку известно среднее квадратичное генеральной совокупности, то используем формулу с z критерием:

Еще раз посмотрим, как получить z критерий в данном случае для 2.5%

По таблице для z значений мы находим этот критерий для 0.025 – это доля, которая в z-распределении лежит ниже этого квантиля. Найдя 0.025 в таблице, мы смотрим на горизонтальный и вертикальный первые столбцы. Видим, что по вертикали это -1,9 , а по горизонтали 0,06. Получается z = 1.96

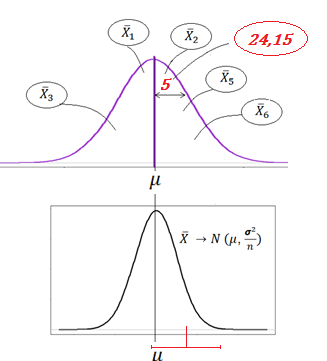


Получаем

[ 23,17; 25,13 ]

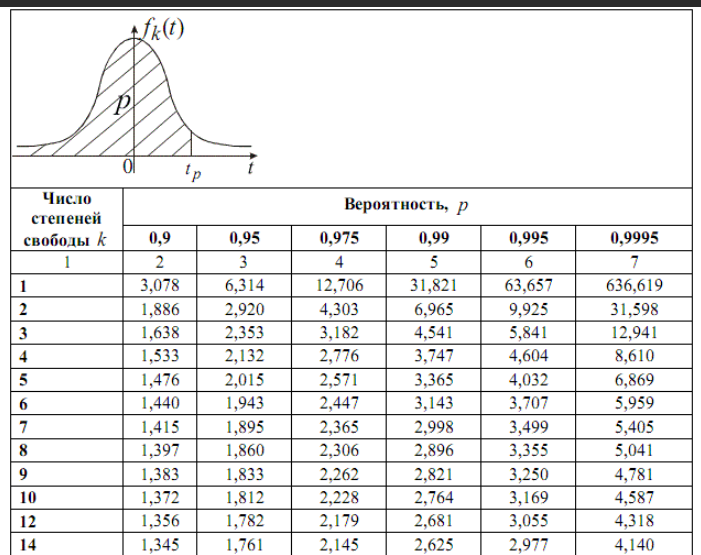
Механически здесь все легко считается, давайте еще раз рассмотрим смысл выполненных действий.

По задачи у нас дана некоторая генеральная совокупность. Смотрим на рисунок ниже. Стандартное отклонение совокупности известно 5 . Мы можем брать множество выборок и считать по ним средние арифметические. Эти средние образовали распределение, обозначенное черной кривой на этом же рисунке. Но наша выборка дала среднее 24.15. Мы не знаем – это больше или меньше истинного среднего. Просто схематично можем отметить справа или слева от истинного среднего. И затем мы построили к нему интервал. Мы не знаем наверняка, захватит этот интервал истинное среднее арифметическое или нет. Чтобы знать наверняка, нам придется измерить всю генеральную совокупность и измерить среднее истинное по ней (но тогда не понадобится интервал). А с интервальной оценкой мы можем сказать, что с вероятностью 95% этот интервал захватит истинное генеральной совокупности.

**

А теперь представим, что нам неизвестна сигма генеральной совокупности. И объем выборки 10. Мы помним с прошлого занятия, что распределение Стьюдента зависит от степеней свободы (n - 1). Мы строим тот же 95% интервал. Т.е.

По таблице ниже берем параметры 0.975 для p и 9 степеней свободы. Получаем t=2.262

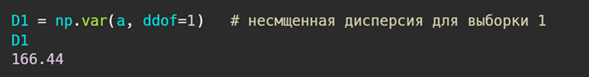
**

Давайте рассмотрим пример в Python. У нас будет дана выборка – рост десяти человек.

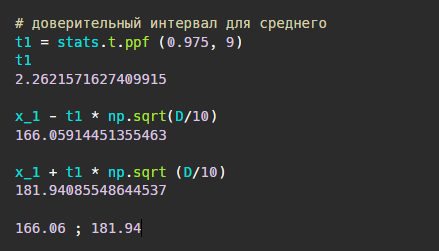


Найдем среднее по выборке

 Теперь найдем несмещенную дисперсию для выборки



Осталось определить критерий Стьюдента и посчитать интервал



**Интервальная оценка для разности средних арифметических**

Предположим, что мы исследуем рост двух разных популяций и хотим оценить различие в росте между двумя средними арифметическими этих популяций. Чтобы получить истинную разницу в росте, нам необходимо измерить всех в одной и другой популяции, найти истинные средние арифметические и и произвести элементарное действие - . Но в реальной жизни мы работаем с выборками. Пусть у нас будет 2 выборки по 10 человек. Если мы измерим по ним средние арифметические, то получим и их разность – будет оценкой истинной разницы. Мы помним, что если мы возьмем другие 2 выборки и посчитаем по ним среднее арифметическое, а затем вычтем одно значение из другого, то получим другое значение . Поэтому в этой задаче также необходимо предоставить интервальную оценку. Но дисперсии в каждой группе различны. Поэтому для начала найдем общую дисперсию, как среднюю дисперсию обеих групп.

, где – дисперсии обеих групп

А вместо стандартной ошибки среднего найдем стандартную ошибку разности средних:

, где и – объемы выборок.

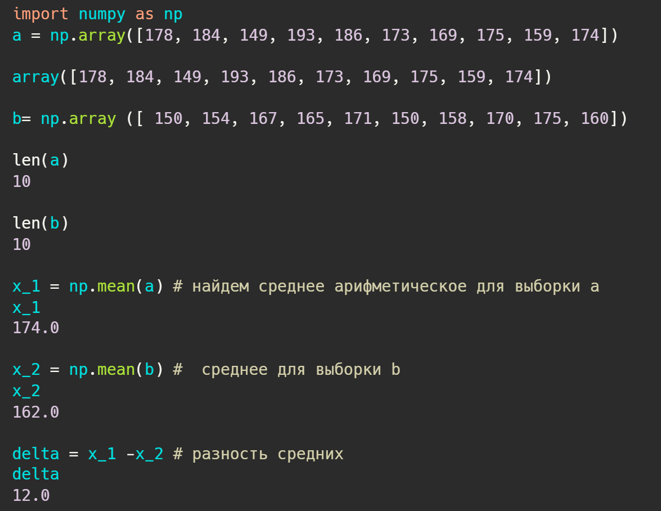
И таким образом, формула для нахождения доверительного интервала для разности средних арифметических будет:

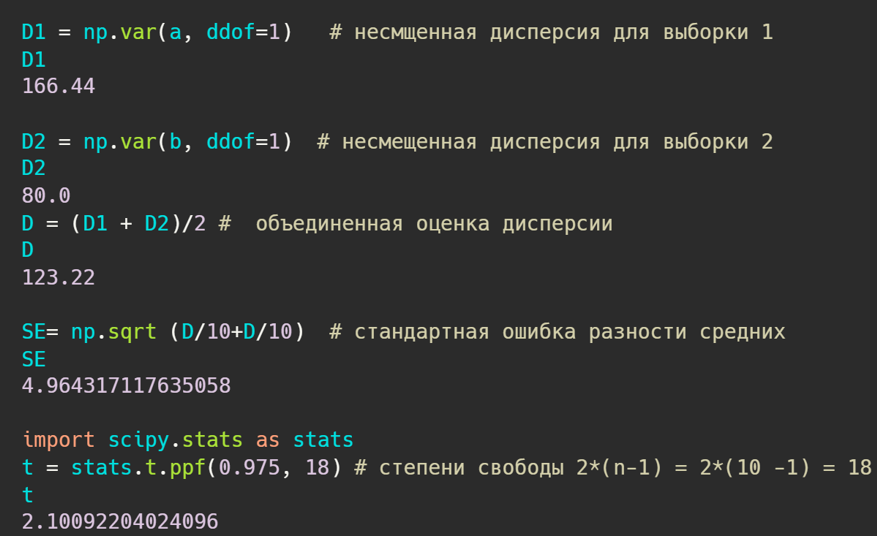
, где

В данном случае для критерия Стьюдента степени свободы

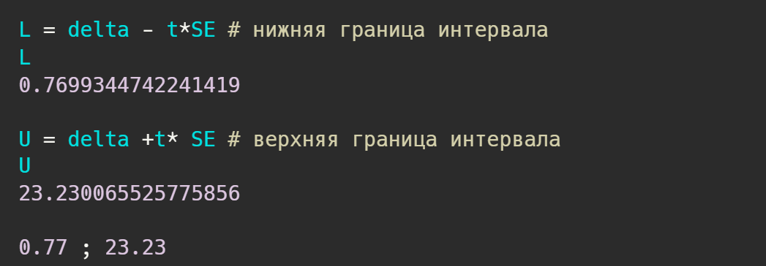
Давайте теперь все эти вычисления выполним в Python.

Создадим выборку a и выборку b , найдем их средние и и их разность



Находим несмещенные дисперсии для каждой выборки и затем стандартную ошибку для разности средних и критерий Стьюдента.

Теперь осталось посчитать верхнюю и нижнюю границы доверительного интервала L и U соответственно. И интервал получится 0.77; 23.23



**Доверительный интервал для доли**

Чтобы сравнивать два средних арифметических, мы можем использовать Z и t критерии при соблюдении определенных условий. Это параметрические критерии, используемые при работе с количественными случайными величинами (далее СВ). Но в статистике также есть качественные случайные величины (мальчики/ девочки, слоны/ жирафы, зеленые / красные). Давайте сегодня сфокусируемся именно на таких случайных величинах.

Работая с количественной СВ, например, рост, мы можем найти среднее арифметическое, стандартное отклонение. Но и работая с долями, мы также можем воспользоваться данными характеристиками описательной статистики, но для этого нам надо посчитать, сколько объектов относится к тому или иному классу. И посчитать их долю p в генеральной совокупности. Эта же величина будет являться вероятностью того, что мы возьмем объект определенного класса. Итак, вместо среднего арифметического мы используем вероятность (долю) p. Следующий важный параметр – стандартное отклонение, которое определяется по формуле:

σ =

Предположим, что в какой – то сфере работают 20% женщин и 80% мужчин.

Давайте посчитаем стандартное отклонение для p = 0.2

σ = = 0.4

А что будет, если p = 0.5?

σ = = 0.5,

для p = 0 , т.е. в коллективе вообще нет женщин, одни мужчины

σ = = 0

для p = 1, когда в коллективе все женщины и совсем нет мужчин:

σ = = 0

Мы видим, что самое большое стандартное отклонение при вероятности p = 0.5.

**А какая же тогда стандартная ошибка для доли?**

Мы помним, что среднее выборочное имеет стандартную ошибку , где n- объем выборки. Стандартная ошибка для доли находится аналогичным образом и получаем .

На самом деле вероятность p сродни среднему арифметическому. Представим, что у нас есть разделение на зеленые и красные. Зеленых объектов 2, а красных 8. Зеленые обозначим 1, а красные 0.

Найдем среднее для этой маленькой генеральной совокупности.

, т.е*.*

Из центральной предельной теоремы следует, что при достаточно большом объеме выборки выборочная оценка стремится к нормальному распределению с параметрами p и стандартной ошибкой SE = . Этим приближением мы пользуемся, когда больше 5. Это утверждение нарушается при маленьких объемах выборки и близких к 0 или 1. Поэтому для построения доверительных интервалов для доли, рассчитанной по маленькой выборке, мы рассмотрим другой пример позже. А при соблюдении выше указанных условий мы, будем пользоваться критерием Z.

Давайте построим 95% доверительный интервал для истинного p – доли женщин, работающих в определенной сфере. Предположим, есть выборка объемом 100, и по выборке получена доля 0.2, тогда интервал находим как

0.2 1.96\* = 0.2 0.08

[0.1216; 0.2784] или в процентах [12,16% ; 27,84%]

**Маленькие объемы выборок**

Вышеприведенный способ вычисления доверительного интервала для доли, как было сказано, применим, если выполняются условия > 5 и > 5. Если это условие не выполняется, то z критерий не используется, а используют биномиальное распределение.

Предположим у нас есть выборка объемом 5. Это пациенты, у которых исследуется наличие побочного эффекта нового препарата. Только у 1 пациента из 5 был зафиксирован побочный эффект. Т.е. вероятность события .

, что меньше 5. Уже это условие не выполняется, поэтому для построения 95% доверительного интервала будем использовать биномиальное распределение. Из 5 пациентов теоретически побочный эффект может наступить у 0, 1, 2, 3, 4, 5 пациентов. Если он наступит у 0 из 5, то доля пациентов с побочным эффектом будет 0, если у 1 из 5, то доля будет , для 2 из 5 равна и т.д. Занесем эти данные в таблицу ниже в столбец «Доля в выборке». А также посчитаем соответствующие вероятности того, что событие наступит 0 раз из 5, 1 раз из 5, 2 раза из 5 и т.д. Используем формулу Бернулли. Мы ее изучали:

где p – вероятность наступления события, q – противоположная вероятность (1 – p).

Накопленная вероятность будет рассчитываться, как сумма текущей и предыдущих вероятностей.

| Значение СВ | Доля в выборке | Вероятность | Накопленная вероятность |
| --- | --- | --- | --- |
| 0 | 0 | 0.32768 | 0.32768 |
| 1 | 0.2 | 0.4096 | 0.73728 |
| 2 | 0.4 | 0.2048 | 0.94208 |
| 3 | 0.6 | 0.0512 | 0.99328 |
| 4 | 0.8 | 0.0064 | 0.99968 |
| 5 | 1 | 0.00032 | 1 |

Доля по выборке получилась 0.2 ( у 1-го из 5 был зафиксирован побочный эффект ), если мы сложим вероятности, соответствующие долям в выборке выше и ниже от 0.2, т.е. для 0 и для 0.4, то получим следующее выражение

0.32768 + 0.4096+ 0.2048 = 0.94208

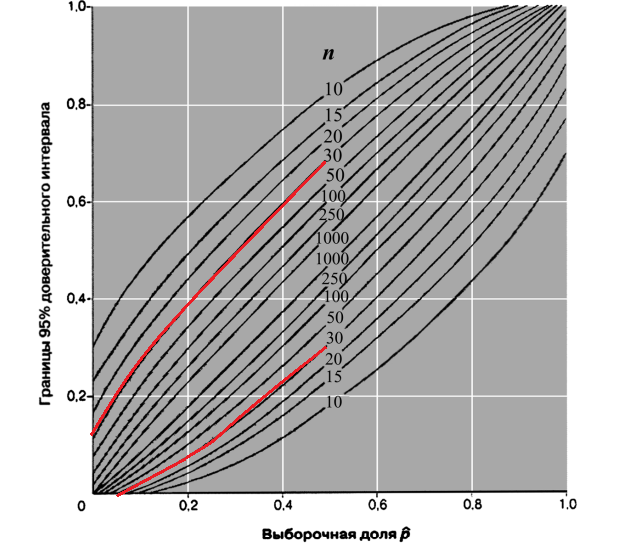
Т.е. 94% интервал составляет 0 – 0.4. Обратите внимание, что мы не можем построить более точный, 95% интервал.

В книге С. Гланца приведен графический метод определения доверительного интервала для выборок, когда условия применимости z –критерия не выполняются.

В книге приводится пример выборки уже из 30 пациентов, ни у одного пациента не обнаружено побочных эффектов, т.е.

30 \* 0 < 5 и 30 \* ( 1 – 0 ) > 5. Первое условие не выполняется, поэтому используем графический метод, приведенный ниже.

Чтобы найти интервал, по горизонтальной оси находим долю , в нашем случае это 0, а затем смотрим, где пересекают ось y линии, соответствующие заданному объему выборки. Обратите внимание, что ось y заканчивается на 0. По оси y и будет интервал. Самая нижняя граница по оси y = 0. И для объема выборки 30 интервал будет от 0 до 0.13 или в процентах от 0% до 13%.



**Сравнение долей**

Мы научились на прошлом вебинаре сравнивать средние арифметические, давайте рассмотрим, как проводится тест гипотезы, где надо сравнить доли.

Для сравнения долей используют z критерий, который находят в случае сравнения долей по формуле:

где – объемы выборок, а – число интересующих объектов в выборке.

Величина Z – это величина, которая следует нормальному стандартному распределению, но выше вычисленная величина z немного отклоняется от нормального распределения, поэтому при сравнении долей нам нужно воспользоваться поправкой Йейтса на непрерывность. Данная поправка сокращает отклонения от нормальности и z критерий будем в итоге находить по формуле:

Как в любом тестировании гипотез теперь расчетное значение сравниваем с табличным значением z и делаем вывод.

**Пример решения задачи.** Есть две группы студентов объемом 56 и 61, которые сдают тест на знание иностранного языка. Максимальное число баллов за тест 120. Высокая оценка считается выше 100 баллов. В первой группе высокую оценку получили 7 студентов, а во второй 22. Есть ли статистически значимые различия в долях студентов, сдавших тест на высокий балл.

Решение:

Для двустороннего теста табличный критерий . На рисунке ниже изображены красные хвостики по 2,5% с обеих сторон, им соответствует как раз и все что находится за этими табличными значениями, попадает в область принятия гипотезы , т. е. расчетный критерий попадает в область принятия альтернативной гипотезы. Во второй группе частота сдача тест на высокий бал выше, чем в первой.

На нашем курсе мы рассмотрели сравнение двух долей, но бывают случаи, когда признак может принимать более 2х значений. Использование Z критерия в этом случае не подойдет. Надо использовать критерий , где строятся таблицы сопряженности, а если в одной из ячеек таблицы значение будет меньше 5, то надо будет воспользоваться критерием Фишера. Эти методы мы не будем рассматривать, потому что существует множество различных тестов, которые просто невозможно охватить в одном курсе. Но самые частые случаи мы рассматриваем здесь. Может случиться так, что некоторыми методами вы не воспользуетесь, а какие-то, напротив, будут очень часто применяться в Вашей сфере деятельности. Но будьте уверены, что для Вас ни один новый тест теперь не составит труда, потому что все тесты имеют общую идею и сводятся к сравнению табличного значения с расчетным. Но современные технологии облегчают задачи расчетов и готовые функции выдают значения pvalue, которое мы уже научились интерпретировать на прошлом вебинаре.

**Интервал для разности долей**

Рассмотрим еще одну задачу. На одном сайте из 153 посетителей 75 оформили заказ, а на другом сайте из 120 заказ оформили 50 человек. Давайте оценим с помощью доверительного интервала разность долей покупателей, совершивших покупку.

Интервал находим по формуле

\* , где – разность долей

– стандартная ошибка разности долей

Z - критерий стремится к нормальному распределению и при работе с долями, поэтому используем его и в этой задаче. Но этот критерий работает хорошо, если мы имеем дело с выборками больших объемов.

0.417 = 0.073

Чтобы найти стандартную ошибку разности долей, надо сначала найти общую долю .

Стандартная ошибка разности долей:

Осталось вычислить интервал:

0.0731.96\*0.06

[ -0.045; 0.009]

Сегодня мы научились строить интервалы для различных ситуаций, а также сравнивать доли с помощью z критерия. На следующем вебинаре мы продолжим тестирование гипотез и поговорим уже о непараметрических тестах, к которым прибегают, когда условия применимости параметрических тестов нарушаются.