КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

ИНСТИТУТ ФУНДАМЕНТАЛЬНОЙ МЕДИЦИНЫ И БИОЛОГИ

Кафедра биохимии и биотехнологии

Н.И.АКБЕРОВА

Выборочные распределения и доверительные интервалы (использование языка R)

Учебно-методическое пособие

Казань – 2014

Выборочные распределения

В этой части рассматриваются характеристики случайной выборки данных, которые могут служить точечными оценками для генеральных параметров.

Рассмотрим реальные данные по продаже недвижимости в города Эймс, штат Айова. Детали каждой сделки с недвижимостью в Эймсе регистрируется в офисе городского налогового инспектора. Мы сосредоточимся продаже всех жилых домов в Эймсе между 2006 г. и 2010 г. Давайте загрузим данные:

```
load(url("http://www.openintro.org/stat/data/ames.RData"))
```

Мы видим довольно много переменных в наборе данных, достаточно, чтобы сделать очень детальный анализ. Ограничимся рассмотрением только двух переменных: площадью дома в квадратных футов (Gr.Liv.Area) и ценой продажи (SalePrice). Создадим две переменные с короткими именами, которые представляют эти две переменные.

```
area <- ames$Gr.Liv.Area
price <- ames$SalePrice
```

Посмотрим на распределение площади в нашей популяции продаж домов путем расчета нескольких суммарных статистик и построим гистограмму.

```
summary(area)
hist(area)
```

Вопрос 1 [выбрать] Что из ниже следующего является ложью?

(a) Распределение площадей домов в городе Эймс унимодально и имеет скос вправо

(b) 50% домов в Эймсе имеют площадь меньше, чем 1500 квадратных футов.

(c) Площадь середины 50% домов колеблется примерно от 1130 квадратных футов до 1740 квадратных футов.

(d) Межквартильный диапазон (IQR)примерно 610 квадратных футов

(e) Наименьший дом имеет площадь 334 квадратных футов и крупнейший - 5642 квадратных футов.

Неизвестное выборочное распределение

В настоящей работе мы имеем доступ к генеральной совокупности, но это редкий случай в реальной жизни. Сбор информации о генеральной совокупности часто является или очень дорогим занятием или вовсе невозможен. В связи с этим чаще всего выборку из совокупности используем для понимания свойств этой совокупности.

Если нас интересует оценка средней жилой площади в Эймсе по выборке, мы можем использовать функцию sample для получения выборки:

```
samp0 <- sample(area, 50)
```

Эта команда осуществляет отбор простой случайной выборки размером 50 из вектора area, которой присваивается имя samp0. Это как пойти в базу данных городского налогового инспектора и случайно выбрать файлы на 50 продаж жилья. Работа с этими 50 файлами будет значительно проще, чем со всеми 2930 продажами жилья.

Теперь когда вы получили выборку, сделайте еще одну выборку и сравните эти две выборки. Являются ли эти выборки одним и тем же? Почему?

Теперь мы готовим выборку:

```
samp1 <- sample(area, 50)
```

Задание Опишите распределение этой выборки. Как оно соотносится с распределением генеральной популяции?

Если мы заинтересованы в оценке средней площади домов в Эймсе, используя выборку, наше лучшее и единственное предположение это выборочное среднее:

```
mean(samp1)
```

В зависимости от 50 домов, которые вы отобрали, ваша оценка может быть немного выше или немного ниже истинного среднего генеральной совокупности, примерно равного 1500 квадратных футов. Однако, в целом, выборочная средняя оказывается довольно хорошей оценкой средней жилой площади, и мы смогли получить ее по выборочным данным менее 3% генеральной совокупности.

Вопрос 2 [выбрать] Предположим, мы взяли еще две выборки, одну размером 100 и одну размером 1000. Какая из двух выборок обеспечит более точную оценку среднего генеральной совокупности?

(a) Объем выборки 50
(b) Объем выборки 100
(c) Объем выборки 1000

Не удивительно, что каждый раз, когда мы делаем еще одну случайную выборку, мы получаем различные выборочные средние. Так мы можем получить представление о том, какую изменчивость следует ожидать при оценке среднего генеральной совокупности. Распределение выборочных средних, называется выборочным распределением, может помочь нам понять эту изменчивость. Поскольку у нас есть сейчас доступ к генеральной совокупности, мы можем построить распределение выборки для выборочных средних, повторив описанные выше шаги много раз. Мы сгенерируем 5000

выборок и вычислим выборочное среднее для каждой из них:

```
sample_means50 <- rep(NA, 5000)
for (i in 1:5000) {
    samp <- sample(area, 50)
    sample_means50[i] <- mean(samp)
}
hist(sample_means50)
```

Если вы хотите настроить ширину столбцов гистограммы, чтобы показать распределение более подробно, вы можете сделать это, изменив аргумент breaks:

```
hist(sample_means50, breaks = 25)
```

Таким образом, используя цикл ***for*** (*for loop)*применяем R для того, чтобы взять 5000 выборок размером 50 из генеральной совокупности, рассчитать среднее каждой выборки, и хранить каждый результат в векторе, называемом sample_means50. Далее рассмотрим в деталях, как этот код работает.

> ***Задание***. Опишите выборочное распределение (распределение средних выборок, что вы только что создали), и убедитесь, что правильно отметили его центр.

Вставка про R: for loop

Давайте попробуем осмыслить этот блок кода. Вы только что запустили ваш первый цикл. Следует отметить, что цикл является краеугольным камнем компьютерного программирования. Идея цикла **for** - итерации: он позволяет выполнить код столько раз, сколько вы хотите без необходимости вводить его на каждой итерации. В приведенном выше случае, мы хотим повторять две строки кода в фигурных скобках, которые отбирают случайную выборку размером 50

из area и сохраняют выборочные средние в вектор sample_means50. Без цикла **for** сделать это было бы нелегко:

```
sample_means50 <- rep(NA, 5000)
samp <- sample(area, 50)
sample_means50[1] <- mean(samp)
samp <- sample(area, 50)
sample_means50[2] <- mean(samp)
samp <- sample(area, 50)
sample_means50[3] <- mean(samp)
samp <- sample(area, 50)
sample_means50[4] <- mean(samp)


и.т.д.
```

С циклом **for** эти тысячи строк кода сжимаются в несколько строк. Мы добавили одну дополнительную строку в коде ниже, который печатает переменную i в каждой итерации цикла. Выполните этот код:

```
sample_means50 <- rep(NA, 5000)
for (i in 1:5000) {
    samp <- sample(area, 50)
    sample_means50[i] <- mean(samp)
    print(i)
}
```

Рассмотрим этот код построчно, чтобы выяснить, что он делает. В первой строке мы инициализировали вектор. В этом случае мы создали пустой вектор из 5000 NAs, называемый sample_means50. Этот вектор будет хранить значения, сгенерированные в **for** цикле. NA означает «не доступны», и в этом случае они используются как заполнители, пока мы не заполним значения фактическими выборочными средними. NA также часто используется в R для обозначения отсутствующих данных.

Вторая строка вызывает сам цикл **for.** Синтаксис можно свободно читать как,

"для каждого элемента **i** от 1 до 5000, выполните следующие строки кода". Вы можете считать i счетчиком, который отслеживает, на каком шаге вы находитесь. Точнее, цикл будет выполняться один раз, когда i=1, один раз при i=2, и так далее вплоть до i=5000.

Тело **for** цикла является частью в фигурных скобках, и этот код выполняется для каждого значения i. В каждом цикле мы берем случайную выборку размером 50 из area, считаем ее среднее и сохраняем его в качестве i -го элемента sample_means50.

Для того чтобы отобразить, что это происходит на самом деле, мы попросили R печатать i на каждой итерации. Эта строка кода не является обязательной и используется только для отображения того, что происходит в то время, как **for** цикл работает.

Цикл **for** позволяет нам не только выполнить код 5000 раз, но аккуратно упаковать результаты, элемент за элементом, в пустой вектор, который мы инициализировали с самого начала.

Задание Чтобы убедиться, что вы понимаете, что вы сделали в этом цикле, попробуйте запустить уменьшенную версию. Инициализировать вектор из 100 NAs с именем sample_means_small. Запустите цикл, который берет выборку размером 50 из area и сохраняет ее среднее в sample_means_small. Напечатайте вывод на экран вашего компьютера (напечатайте sample_means_small в консоль и нажмите Enter).

Вопрос 3 [выбрать] Сколько элементов в этом объекте под названием sample_means_small?

(a) 0

(b) 30
(c) 50
(d) 100
(e) 5000

Вопрос 4 [выбрать] Что из нижеперечисленного верно об элементах выборочных распределений, созданных вами?

(a) Каждый элемент представляет собой среднюю площадь в квадратных футах простой случайной выборки из 50 домов.

(b) Каждый элемент представляет собой площадь дома в квадратных футах.

(c) Каждый элемент представляет истинное математическое ожидание (среднее) метража домов.

Размер выборки и выборочное распределение

Вернемся к причине, для чего мы использовали **for** цикл: для вычисления

```
hist(sample_means50)
```

Выборочное распределение, которое мы вычислили, многое говорит об оценке средней жилой площади в домах Эймса. Поскольку среднее выборки есть несмещенная оценка, выборочное распределение имеет центр в истинной средней жилой площади генеральной совокупности, и размах распределения показывает, какова изменчивость по выборочным данным 50 сделок по продаже жилья.

Чтобы получить представление об эффекте, который имеет объем выборки на наше выборочное распределение, давайте построим еще два выборочных распределения: одно основано на выборке из 10 и другое, основанное на выборке из 100:

```
sample_means10 <- rep(NA, 5000)
sample_means100 <- rep(NA, 5000)
for (i in 1:5000) {
    samp <- sample(area, 10)
    sample_means10[i] <- mean(samp)
    samp <- sample(area, 100)
    sample_means100[i] <- mean(samp)
}
```

Здесь мы можем использовать один цикл для создания двух распределений, добавляя дополнительные строки внутри фигурных скобок. Не нужно беспокоится о том, что samp используется для названия двух разных объектов. Во второй команде **for** цикла, среднее samp сохраняется в соответствующем месте в вектор sample_means10. Поскольку средние сохранены, мы можем переписывать объект samp для новой выборки, на этот раз объемом 100. В общем, вы создаете объект, используя имя, которое уже используется, старый объект будет заменен на новый, т.е. R будет писать поверх существующего объекта новый.

Чтобы увидеть эффект различных объемов выборок на выборочное распределение, постройте графики трех распределений поверх друг друга. Для этого введите:

```
par(mfrow = c(3, 1))
xlimits = range(sample_means10)
hist(sample_means10, breaks = 20, xlim = xlimits)
hist(sample_means50, breaks = 20, xlim = xlimits)
hist(sample_means100, breaks = 20, xlim = xlimits)
```

Первая команда указывает, что вы хотели бы разделить область построения графиков на 3 ряда и 1 колонку. Возможно, вам придется растянуть окно для размещения дополнительных графиков. Чтобы вернуться к настройкам по умолчанию (одновременно один график) выполните следующую команду:

```
par(mfrow = c(1, 1))
```

Аргумент breaks определяет количество «бункеров», столбцов, используемых при построении гистограммы. Аргумент xlim задает диапазон оси x

гистограммы, и, установив его равным xlimits каждой гистограммы, мы будем уверены, что все три гистограммы будут построены с теми же ограничениями по оси x.

Вопрос 5 [выбрать] Интуиция подсказывает, что по мере увеличения объема выборки, центр выборочного распределения становится все более надежной оценкой для истинного среднего совокупности. Кроме того, с увеличением размера выборки, изменчивость распределения выборки_____.

(a) уменьшается

(b) увеличивается

(c) остается неизменной

До сих пор мы работали с оценкой средней жилой площади в домах Эймса. Теперь попытайтесь оценить среднюю стоимость дома.

Задание Возьмите случайную выборку объемом 50 из price. При использовании этой выборки, что является лучшей точечной оценкой среднего по генеральной совокупности?

Задание Поскольку у вас есть доступ к генеральной совокупности, смоделируйте выборочное распределение для $\bar{x}_{price}$, взяв 5000 выборок из генеральной совокупности объемом 50 и вычислите 5000 выборочных средних. Сохраните их в векторе с именем sample_means50. Постройте график, а затем опишите форму этого выборочного распределения. На основе полученного выборочного распределения, какова средняя цена дома генеральной совокупности?

Задание Измените размер выборки с 50 на 150, затем вычислите выборочное распределение, используя для этого тот же метод, что и выше, сохраните эти полученные значения в новый вектор с названием sample_means150. Опишите форму этого выборочного распределения и сравните его с распределением для объема выборок 50. На основе полученного выборочного распределения, что бы вы сказали о средней цене продажи домов в Эймсе?

Вопрос 6 [выбрать] Что из ниже перечисленного является ложным?

(a) Изменчивость выборочного распределения с меньшим размером выборок (sample_means50) меньше, чем изменчивость выборочного распределения с большим размером выборок (sample_means150).

(b) Средние значения обоих выборочных распределений примерно одинаковы

(c) Оба выборочных распределения симметричны

Доверительные интервалы

Выборки из г. Эймс, штат Айова

Если у вас есть доступ к данным целой совокупности, говоря о размере каждого дома в Эймс, штат Айова, можно прямо ответить на вопросы: "Каков размер типичного дома в Эймсе?" и "Насколько значительно отличаются в размерах дома?". Если у вас есть доступ только к выборке из совокупности, как это обычно бывает, задача усложняется. Если вы знаете только размеры нескольких десятков домов, то каким будет самое лучшее предположение для типичного размера? Такая ситуация требует, чтобы вы использовали выборку для того, чтобы сделать логический вывод о том, как выглядит генеральная совокупность.

Начнем с простой случайной выборки объемом 60 из генеральной совокупности всех домов. Следует отметить, что набор данных содержит информацию о многих переменных, но сначала мы сфокусируемся на размере дома, представленного переменной Gr.Liv.Area.

```
load(url("http://www.openintro.org/stat/data/ames.RData"))
```

```
population <- ames$Gr.Liv.Area
samp <- sample(population, 60)
```

Упражнение Опишите распределение вашей выборки. Каков "типичный" размер в пределах вашей выборки? Как именно вы интерпретируете "типичное" среднее?

Вопрос 1 [TRUE / FALSE] Мое распределение должно быть похоже на распределения других людей, которые также собирают случайные выборки из этой генеральной совокупности, но это, скорее всего, не совсем то же, так как это случайная выборка.

- TRUE
- FALSE

Интервальные оценки

Один из наиболее распространенных способов описать типичное или центральное значение распределения является использование среднего. В этом случае мы можем вычислить выборочное среднее с помощью:

```
sample_mean <- mean(samp)
```

Вернемся к вопросу, какой можно сделать вывод о генеральной совокупности на основе этой выборки? Основываясь только на этой одной выборке, лучшая оценка домов со средней площадью, проданных в Эймсе, будет выборочное среднее, которое обычно обозначается как $\bar{x}$ (sample_mean). Это хорошая точечная оценка, но было бы полезно также сообщить, насколько уверены мы в этой оценке. Это можно сделать с помощью доверительного интервала.

Мы можем вычислить 95% доверительный интервал для среднего, добавляя 1,96 стандартных ошибок к точечной оценке и вычитая 1,96 стандартных ошибок из точечной оценки:

```
se <- sd(samp)/sqrt(60)
lower <- sample_mean - 1.96 * se
upper <- sample_mean + 1.96 * se
c(lower, upper)
```

Это важный вывод, который мы только что сделали: хотя мы не знаем, как в полной мере выглядит генеральная совокупность, мы в 95% уверены, что истинный средний размер дома в Эймсе лежит между значениями нижней lower и верхней upper границ доверительного интервала. Есть несколько условий, которые должны быть выполнены, чтобы построенный интервал был в силе.

Вопрос 2 [выбрать]

Доверительный интервал работает, когда выборка нормально распределена и среднее имеет стандартную ошибку, равную стандартное отклонение *s*/корень из *n*.

Что из нижеперечисленного не является необходимым условием ?

- выборка случайная.
- размер выборки, 60, меньше 10% всех домов.
- распределение выборки должно быть близким к нормальному.

Доверительные уровни

Вопрос 3 [выбрать] Что означает «95% доверие»?

- 95% времени истинная средняя площадь домов в Эймсе, штат Айова, будет в этом интервале.
- 95% от случайных выборок размера 60 даст доверительные интервалы, содержащие истинную среднюю площадь домов в Эймсе, штат Айова.
- 95% домов в Эймсе имеют площадь в этом интервале.
- 95% увереность, что выборочное среднее находится в этом интервале.

В нашем случае у нас есть редкая возможность знать истинное математическое ожидание (среднее), так как у нас есть данные по генеральной совокупности. Это значение может быть вычислено с помощью следующей команды:

```
mean(population)
```

Задание Содержится ли истинный средний размер дома в Эймсе в вашем доверительном интервале?

Вопрос 4 [выбрать] Какая часть 95% доверительного интервала будет содержать среднее генеральной совокупности?

- 1%
- 5%
- 99%
- 95%

Используя R, мы собираемся воссоздать много выборок, чтобы узнать больше о том, как выборочные средние и доверительные интервалы изменяются от выборки к выборке. Здесь пригодятся циклы.

Используйте следующую схему:

- Получите случайную выборку.
- Рассчитайте выборочные среднее и стандартные отклонения.
- Используйте эти выборочные параметры для определения доверительного интервала.
- Повторите шаги (1)-(3) 50 раз.

Но прежде, чем сделать все это, следует сначала создать пустые векторы, где мы можем сохранить средние и стандартные отклонения, которые будут рассчитаны из каждой выборки, в это время будем хранить необходимый размер выборки как n.

```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60
```

Теперь мы готовы к циклам, где мы вычислим средние величины и стандартные отклонения 50 случайных выборок.

```
for(i in 1:50){
  samp <- sample(population, n) # извлечение выборки объемом n = 60 из ген.совокупности.
  samp_mean[i] <- mean(samp)    # сохраняет выб.среднее в i-том элементе samp_mean samp_sd[i]
  <- sd(samp)                   # сохраняет выб. sd в i-том элементе samp_sd
}
```

Наконец, мы строим доверительные интервалы:

```
lower <- samp_mean - 1.96 * samp_sd/sqrt(n)
upper <- samp_mean + 1.96 * samp_sd/sqrt(n)
```

Нижние границы этих 50-ти доверительных интервалов хранятся в lower, и верхние границы находятся в upper. Давайте рассмотрим первый интервал.

```
c(lower[1], upper[1])
```

Чтобы визуализировать рассчитанные доверительные интервалы и показать среднее генеральной совокупности применяем:

```
plot_ci(lower, upper, mean(population))
```

Задание Доля доверительных интервалов, которые включают истинное среднее по совокупности, точно равна уровню доверия? Если нет, объясните почему.

Вопрос 5 [выбрать] Каким должно быть критическое значение для 99% доверительного уровня?

(a) 0.01
(b) 0.99
(c) 1.96
(d) 2.33
(e) 2.58

Задание Рассчитайте 50 доверительных интервалов на уровне 99%. Вам не нужно получать новые выборки, просто рассчитайте новые интервалы, основанные на выборочных средних и стандартных отклонениях, уже собранных вами. Используя функцию plot ci, нарисуйте график всех интервалов и рассчитайте долю интервалов, которые включают истинное среднее по совокупности.

Вопрос 6 [TRUE / FALSE] Мы ожидаем, что 99% интервалов будут содержать истинное среднее генеральной совокупности.

- TRUE
- FALSE

Контрольные задания

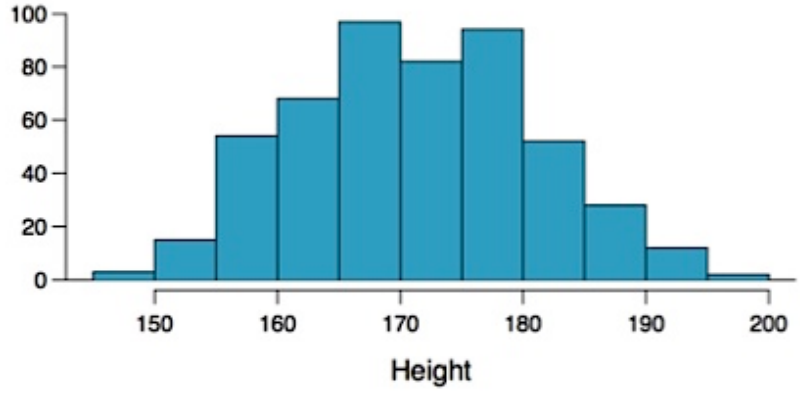
Задание 1

Предположим, что мы изучаем, сколько шоколада (в граммах в неделю) потребляется студентами Coursera. После опроса 500 студентов, мы рассчитываем в среднем 175 граммов в неделю со стандартным отклонением 195 граммов в неделю. Что из перечисленного ниже не обязательно верно?

- Выборочное распределение будет скошено вправо.
- $\bar{x}=175$, $s=195$
- $\mu=175$, $\sigma=195$
- Точечной оценкой генерального стандартного отклонения является 195.

Задание 2

Исследователи собрали различные антропометрические показатели 507 физически активных людей. На гистограмме показано выборочное распределение роста в см. Будем считать, что эти 507 человек представляют случайную выборку, тогда выборочное среднее является точечной оценкой среднего роста всех физически активных людей. Какую меру следует использовать для количественной оценки изменчивости такой оценки? Вычислите эту величину, используя данные из примера, и выберите лучший ответ.



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

- Стандартная ошибка = 0.019
- Стандартное отклонение = 0.019
- Стандартная ошибка = 0.417
- Стандартное отклонение = 0.417
- Среднеквадратичная ошибка= 0.105

Задание 3

Что описывают стандартные отклонения σ и s и стандартная ошибка SE?

- *σ*: изменчивость выборочных данных; *SE*: изменчивость точечных оценок на основании различных выборок одного размера из одной и той же; *s*: изменчивость в генеральной совокупности
- *SE* : изменчивость выборочных данных; *σ*: изменчивость точечных оценок на основании различных выборок одного размера из одной и той же; *s*: изменчивость в генеральной совокупности
- *σ*: изменчивость выборочных данных; *s*: изменчивость точечных оценок на основании различных выборок одного размера из одной и той же; *SE*: изменчивость в генеральной совокупности
- *s*: изменчивость выборочных данных; *SE*: изменчивость точечных оценок на основании различных выборок одного размера из одной и той же; *σ*: изменчивость в генеральной совокупности

Задание 4

Что из нижеперечисленного неверно о центральной предельной теореме (ЦПТ)?

- ЦПТ утверждает, что центр выборочного распределения выборки будет истинным параметром генеральной совокупности
- По мере увеличения размера выборки, выборочное распределение среднего, скорее всего, будет почти нормальным, независимо от формы исходного распределения в генеральной совокупности

- Если взять больше выборок (одного размера) из исходной совокупности, выборочное распределение средних, скорее всего, будет близко к нормальному
- Если распределение совокупности является нормальным, выборочное распределение среднего также будет почти нормальным, независимо от размера выборки

Задание 5

Чтобы получить оценку потребительских расходов в США после Дня Благодарения, в произвольном порядке было опрошено 436 взрослых американцев. Среднее их ежедневных расходов шестидневного периода после Дня Благодарения составляет $ 84,71. 95% -ный доверительный интервал на основе этой выборке ($ 80,31, $ 89,11). Что из нижеперечисленного верно?

I.Мы на 95% уверены, что в среднем расходы 436 взрослых американцев из данной выборки составляют от $ 80,31 до $ 89,11.

II. Если мы собрали много случайных выборок одинакового размера и рассчитали доверительный интервал для ежедневных расходов для каждой выборки, то мы ожидаем, что 95% интервалы содержат истинные параметры населения.

III. Мы на 95% уверены, что в среднем расходы всех взрослых американцев составляет от $ 80,31 до $ 89,11

- I,II
- I,III
- II,III
- I,II,III

Задание 6

Некое исследование показывает, что средний студент тратит 2 часа в неделю, общаясь с другими людьми в Интернете. Вы считаете, что это недооценка, и решили собрать свою собственную выборку для проверки гипотезы. Вы случайно образом опросили 60 студентов из вашего общежития и выяснили, что в среднем они провели 3,5 часа в неделю, общаясь с друзьями онлайн. Что из нижеперечисленного является правильным набором гипотез?

- o $H_o: \mu = 2, H_A: \mu > 2$
- o $H_o: \mu = 2, H_A: \mu < 2$
- o $H_o: \overline{x} = 2, H_A: \overline{x} < 2$
- o $H_o: \overline{x} = 2, H_A: \overline{x} > 2$

Задание 7

Предположим, что мы собрали выборку размером n = 100 из некоторой популяции и используем данные для расчета 95% доверительный интервал для истинного среднего. Теперь предположим, что мы собираемся увеличить размер выборки до n = 300. Если все остальное останется неизменным, что можно ожидать в результате увеличения размера выборки?

I. Стандартная ошибка будет уменьшаться.
II. Ширина 95% доверительного интервала будет увеличиваться.
III. Предел погрешности уменьшится.

- o I,III
- o I,II
- o II,III
- o I,II,III
- o Нет ответа

Задание 8

Односторонние альтернативные гипотезы сформулированы в терминах:

- o < or >
- o ≤ or ≥
- o ≈ or =
- o ≠

Задание 9

Вы используете двусторонний тест для истинной средней μ с нулевой гипотезой H_0: μ = 100, выбрали уровень значимости α = 0,05. Значение p-value, вычисленное из данных, составляет 0,12, и, следовательно, вы не в состоянии отвергнуть нулевую гипотезу. Предположим, что после того как ваш анализ был завершен и опубликован, эксперт сообщил, что истинное значение μ является 104. Как бы Вы охарактеризовали результат вашего анализа?

- o Была допущена ошибка второго типа, потому что вы не смогли отвергнуть нулевую гипотезу, хотя μ действительно не равна 100
- o Была допущена ошибка первого типа, потому что вы не смогли отвергнуть нулевую гипотезу, хотя μ действительно не равна 100
- o Вы не сделали ошибок ни первого, ни второго типа

Задание 10

Статистик изучает уровни давления крови итальянцев в возрасте 75-80. Ниже приведены некоторые сведения об этом исследовании.

(1) Данные были собраны путем опроса, проведенного по электронной почте, при этом никакие меры не были приняты, чтобы получить информацию от тех, кто не ответил на начальном этапе обследования.

(2) Выборка наблюдений составляют около 4% населения.

(3) Размер выборки 2047.

(4)Распределение выборки имеет хорошо видимый, хотя и не очень экстремальный, перекос.

Исследователь собирается использовать центральную предельную теорему (ЦПТ) в основной части своего анализа. Какой аспект исследования, скорее всего, помешает использовать ЦПТ?

- (2), потому что данные собраны для маленькой части населения
- (3), так как размер выборки слишком мал по сравнению со всеми итальянцами в возрасте 75-80
- (1), потому что выборка может быть не случайной, и, следовательно, наблюдения могут быть не независимыми
- (4), потому что есть некоторый перекос в распределении выборки

Задание 11

Распределение SAT баллов (Академический Оценочный Тест - стандартизованный тест для приема в высшие учебные заведения в США) имеет среднее значение 1500 и стандартное отклонение 300. Вы хотите оценить средний балл SAT студентов первого года обучения в вашем колледже. Сколько студентов вы должны включить в выборку, чтобы статистическая погрешность 95%-ного доверительного интервала составляла 25?

- 392
- 554
- 553
- 393
- 13,83

Задание 12

В каком случае выше риск отвергнуть нулевую гипотезу, которая на самом деле справедлива: при большем или меньшем уровне значимости?

- меньший
- больший

Задание 13

Этикетка на пакете чипсов говорит, что одна унция (28 грамм) порции картофельных чипсов имеет 130 калорий и содержит 10 граммов жира, в т.ч. 3 грамма насыщенных жиров. Случайная выборка из 35 пакетов имела среднее 134 калорий со стандартным отклонением 17 калорий. Попробуйте оценить на 10% уровне значимости, обеспечивают ли эти данные убедительные доказательства, что на этикетке не обеспечивается точное измерение калорий в пакетах чипсов. Что из нижеперечисленного является правильным?

- p-value составляет примерно 16%, что означает, что мы должны отвергнуть нулевую гипотезу и можем утверждать, что эти данные дают убедительные доказательства, что этикетка не обеспечивает точное измерение калорий в пакетах чипсов.

- p-value составляет примерно 8%, что означает, что мы не можем отвергнуть нулевую гипотезу, и эти данные не дают убедительных доказательств, что этикетка не обеспечивает точное измерение калорий в пакетах чипсов.
- p-value составляет примерно 8%, что означает, что мы должны отвергнуть нулевую гипотезу и можем утверждать, что эти данные дают убедительные доказательства, что этикетка не обеспечивает точное измерение калорий в пакетах чипсов
- p-value составляет примерно 16%, что означает, что мы не можем отвергнуть нулевую гипотезу, и эти данные не дают убедительных доказательств, что этикетка не обеспечивает точное измерение калорий в пакетах чипсов.

Задание 14

Компания предлагает онлайн-курсы по скорости чтения и утверждает, что студенты, которые учатся на этих курсах, демонстрируют увеличение в 5 раз (500%) количества слов, что они могут прочитать в минуту без потери понимания. Случайная выборка из 100 студентов показала среднее 415% и стандартное отклонение 220%. Рассчитайте 95% доверительный интервал для среднего увеличения количества слов, которые студенты могут читать в минуту без потери понимания. Выберите самый точный ответ.

- (411.37, 418.63)
- (412.09, 417.91)
- (378.7, 451.3)
- (371.88, 458.12)