# Categorical data

Датасет содержит информацию о статистике одобрения заявок на научные проекты в Австралийском университете.

df <- read.csv("grants.csv")

str(df)

Прокомментируем некоторые поля из датасета:

npersons – количество людей в проекте;

years\_in\_uni – количество лет, которое провел в университете руководитель проекта;

oldest\_age – самый старый сотрудник;

field – область исследований

status – была ли поддержана заявка, т.к. данная переменная номинативная, то сделаем из неё фактор:

df$status <- as.factor(df$status)

levels(df$status) <- c("Not funded", "Funded") # levels позволяет просматривать уровни фактора. Установим для фактора уровни "Not funded", "Funded" (не поддержана/поддержана)

Тоже самое можно сделать, выполнив:

df$status <- factor(df$status, labels = c("Not funded", "Funded"))

# 1d Table

Выводим таблицу по поддержанным и не поддержанным заявкам:

t1 <- table(df$status)

t1

dim(t1) #размерность таблицы (2 уровня)

# 2d Table

Построим таблицу сопряженности по 2-м переменным, показывающую количество поддержанных/не поддержанных заявок по областям:

t2 <- table(df$status, df$field)

t2

Добавляем имена

t2 <- table(status = df$status, field = df$field)

t2

dim(t2)

Выводит таблицу сопряженности в пропорциях:

prop.table(t2)

100% по строке

prop.table(t2, 1)

100% по столбцу (в нашем примере более показательно) :

prop.table(t2, 2)

# 3d Table

Добавим ещё одну переменную и получим разбиение на 2 таблицы, одна по поддержанным заявкам, а вторая – по не поддержанным:

t3 <- table(Years = df$years\_in\_uni, Field = df$field, Status = df$status)

t3

dim(t3)

# plots (Графики)

Столбчатая диаграмма:

barplot(t1)

barplot(t2)

Добавляем легенду:

barplot(t2, legend.text = TRUE, args.legend = list(x = "topright"))

Добавим параметр beside, в результате на каждую область будет выведено по 2 столбика (поддержанные и не поддержанные заявки):

barplot(t2, legend.text = TRUE, args.legend = list(x = "topright"), beside = TRUE)

Мозаичный график:

mosaicplot(t2)

##########################

**Нулевая гипотеза** – это*о*сновное проверяемое предположение, которое обычно формулируется как отсутствие различий, отсутствие влияния фактора, отсутствие эффекта, равенство нулю значений выборочных характеристик и т.п.

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

Чем меньше p-уровень значимости, тем больше оснований отклонить нулевую гипотезу (p<0.05 – отклоняем нулевую гипотезу).

Обычно нулевая гипотеза отклоняется, и различия считаются статистически достоверными, если p < 0,05. Однако часто в статистике используется более жесткий критерий достоверности различий, например, при условии, что p < 0,01. Значение p-уровня значимости, которое выбирается, в качестве порога обозначается буквой *α* (альфа). Например, если исследователь решил, что *α*=0,05, то и нулевая гипотеза будет отклоняться при условии, что p < 0,05.

# Binomial Test (Биномиальный тест)

Проверяем соответствует распределение теоретическому биномиальному

Немного теории: БТ – это двусторонний тест для оценки нулевой гипотезы о том, что вероятность успеха в эксперименте Бернулли равна заданной.

**Пример:** Частота нарушений сна в общей популяции достигает 30%. Было обследовано 156 пациентов, страдающих головной болью напряжения. Установлено, что у 67 из них имели место нарушения сна. Являются ли нарушения сна характерными для больных с головной болью напряжения по сравнению с общей популяцией?

Таким образом, по условию вероятность появления симптома в общей популяции 0.3, количество испытаний 156, количество исходов 67. В результате расчета p=0.001.

**Вывод:** Нулевая гипотеза отвергается с вероятностью ошибки меньше 0,1%, значит частота появления нарушений сна в обследованной выборке пациентов с головной болью напряжения достоверно выше, чем в общей популяции.

binom.test(x = 5, n = 20, p = 0.5) #подбрасываем монетку 20 раз, количество исходов = 5 (например, количество орлов), 0,5 - априорная вероятность успеха; p-value здесь трактуем, как: какова вероятность, что в нормальной монетке орел выпадет 5 и менее раз

Для таблиц:

binom.test(t1) #проверяем гипотезу, что заявки чаще отвергаются

# Chi-Square (Хи-квадрат)

Критерий согласия Пирсона или критерий согласия хи-квадрат — непараметрический метод, который позволяет оценить значимость различий между фактическим (выявленным в результате исследования) количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы. Выражаясь проще, метод позволяет оценить статистическую значимость различий двух или нескольких относительных показателей (частот, долей). Является наиболее часто употребляемым критерием для проверки гипотезы о принадлежности наблюдаемой выборки x1,x2,...,xn объёмом n некоторому теоретическому закону распределения. Критерий хи-квадрат для анализа таблиц сопряжённости был разработан и предложен в 1900 году основателем математической статистики английским учёным Карлом Пирсоном.

Нулевая гипотеза хи квадрат

Нулевая гипотеза заключается в том, что частоты согласованы, то есть фактические данные не противоречат ожидаемым. Альтернативная гипотеза – отклонения в частотах выходят за рамки случайных колебаний, расхождения статистически значимы.

t1

chisq.test(t1) # проверяем гипотезу о том, насколько эмпирическое распределение удачных/неудачных заявок отличается от случайного, в результате приходим к выводу, что наше распределение заявок не соответствует равномерному распределению (p<0.05)

chi <- chisq.test(t1)

chi$exp #ожидаемые значения

chi$obs #наблюдаемые значения

t2

chisq.test(t2)

# Fisher's Exact Test (Тест Фишера)

Тест обычно используется, чтобы исследовать значимость взаимосвязи между двумя переменными в факторной таблице размерности 2×2 (таблице сопряжённости признаков). Величина вероятности p теста вычисляется, как если бы значения на границах таблицы известны. С большими выборками может использоваться тест хи-квадрат. Однако этот тест не является подходящим, когда математическое ожидание значений в любой из ячеек таблицы с заданными границами оказывается ниже 10: вычисленное выборочное распределение испытуемой статистической величины только приблизительно равно теоретическому распределению хи-квадрат, и приближение неадекватно в этих условиях (которые возникают, когда размеры выборки малы, или данные очень неравноценно распределены среди ячеек таблицы). Тест Фишера, как следует из его названия, является точным и может поэтому использоваться независимо от особенностей выборки. Тест становится трудновычислимым для больших выборок или хорошо уравновешенных таблиц, но, к счастью, именно для этих условий хорошо применим критерий Пирсона.

fisher.test(t2)