



Критерии согласия

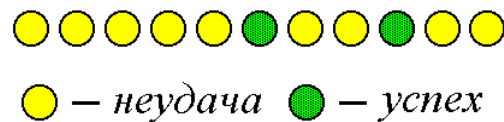
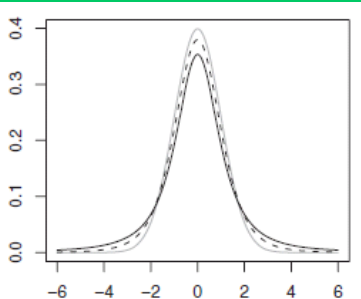
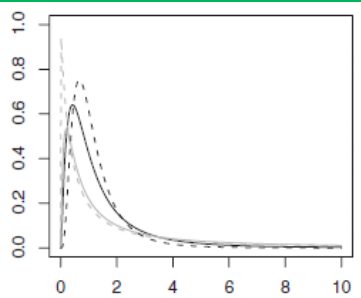
He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may be cast.

Leonardo da Vinci

Статистические модели

Статистическая модель — это некоторое семейство вероятностных распределений, приближающее неизвестное распределение интересующего исследователя признака (признаков) во всей генеральной совокупности объектов.

Как правило, функции распределения в семействе зависят от одного или нескольких неизвестных параметров. Например, в **модели Бернулли** распределения зависят от параметра p , где $0 < p < 1$ — вероятность «успеха» (выпадения несимметричной монеты «гербом» вверх).



● — неудача ● — успех

All models are wrong.

Some models are better than others.

The correct model can never be known with certainty.

The simpler the model, the better it is.

Из книги М. J. Crawley «Statistical Computing. An Introduction to Data Analysis using S-PLUS», Wiley, Chichester, 2002 (с. 17).



Важнейшие понятия статистики

- Статистическая гипотеза
- Критерий
- Статистика критерия
- Уровень значимости
- Фактический уровень
значимости



Многие вещи нам непонятны не потому, что наши понятия слабы; но потому, что сии вещи не входят в круг наших понятий.

Козьма Прутков



Определения основных понятий теории проверки гипотез

Статистическая гипотеза (H) — предположение о том, что наблюдения можно рассматривать как выборку из определённого семейства вероятностных распределений. Гипотеза называется *простой*, если семейство состоит из единственного распределения. В противном случае гипотеза называется *сложной*. В последнем случае функции распределения в модели зависят от одного или нескольких неизвестных параметров.

Критерий ($test$) — правило, позволяющее принять решение о том, верна или нет проверяемая статистическая гипотеза.

Статистика критерия (S) — некоторая функция от наблюдений (скажем, сумма или максимум), очень большие (или очень малые) значения которой противоречат проверяемой гипотезе.

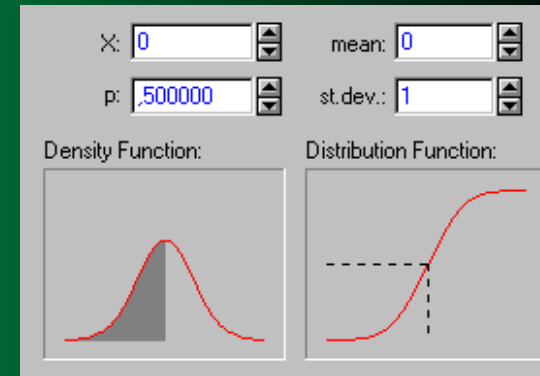
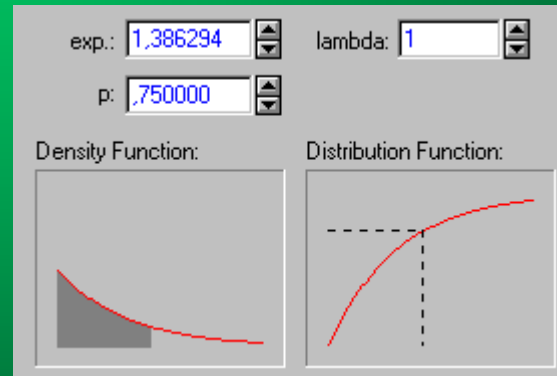
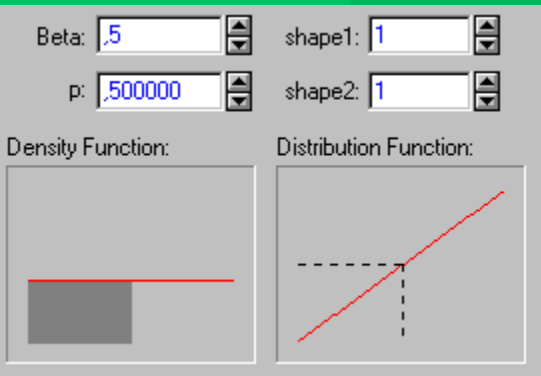
Уровень значимости критерия (α) — малая вероятность, с которой допускается ошибочное отвержение верной гипотезы из-за случайности. На его основе рассчитывают критическое значение статистики критерия. Как правило, на практике берут $\alpha = 0,05$.

Фактический уровень значимости критерия (p -level или p -value) — вероятность, что статистика критерия примет значение не меньше наблюдаемого в предположении, что проверяемая гипотеза верна.



Три распределения

- Равномерное на отрезке $[0, 1]$
- Показательное с параметром $\lambda = 1$
- Стандартное нормальное $N(0, 1)$



Эти распределения связаны следующей задачей:

Рассмотрим случайную величину ξ , имеющую плотность распределения $p(x)$ с носителем $A = \{x : p(x) > 0\}$. Задача состоит в максимизации энтропии $H(\xi) = - \int p(x) \log p(x) dx$ при выполнении одного из следующих наборов условий:

- $A = (0, 1)$,
- $A = (0, +\infty)$ и $\mathbf{M}\xi = 1$,
- $A = (-\infty, +\infty)$, $\mathbf{M}\xi = 0$ и $\mathbf{D}\xi = 1$.

Энтропия
— это мера
беспорядка,
непредска-
зуемости

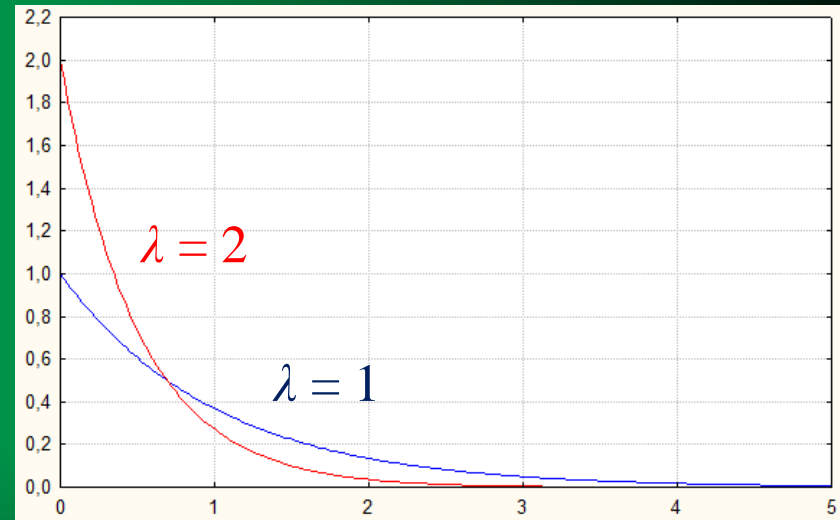


Показательная и нормальная статистические модели

Пусть случайная величина τ имеет показательное (экспоненциальное) распределение с параметром $\lambda > 0$. Тогда она имеет плотность

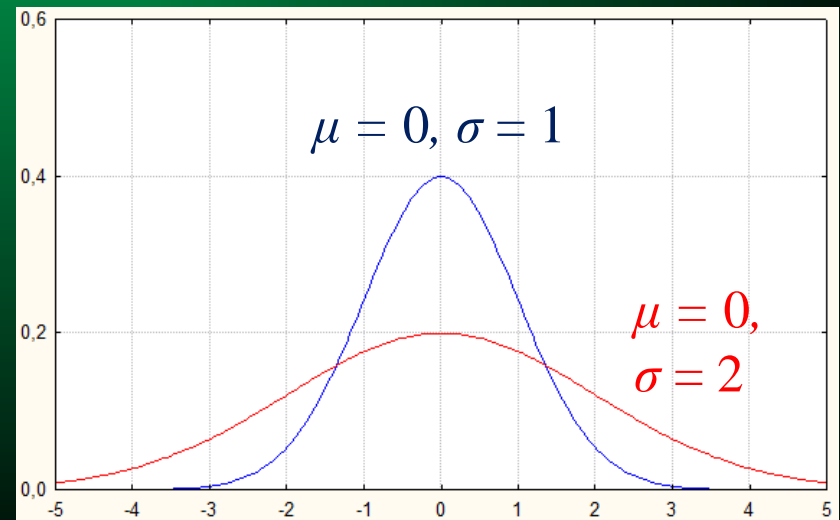
$$f_{\lambda}(x) = \lambda e^{-\lambda x} I_{\{x>0\}},$$

где $I_{\{x>0\}}$ — индикатор множества $\{x>0\}$.

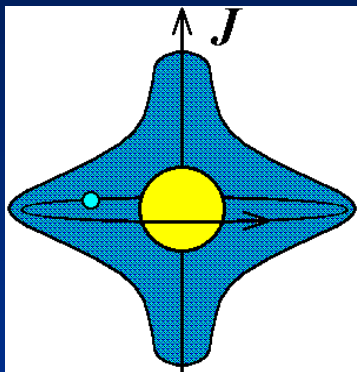


Пусть случайная величина Z имеет распределение $N(0, 1)$. Тогда случайная величина $X = \mu + \sigma Z$ распределена по закону $N(\mu, \sigma^2)$, имеющему плотность

$$f_{\mu, \sigma}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Гипотеза Лапласа о кометах



В 1812 г. Лаплас исследовал такую проблему: образовались ли кометы в общем с планетами «волчке» или же они — всего лишь «гости», захваченные притяжением Солнца. В последнем случае углы между нормальными к плоскостям орбит комет и **моментом импульса J** должны не концентрироваться вблизи 0 , а быть равномерно распределёнными на отрезке $[0, \pi/2]$.

Проведя статистическую обработку известных к тому времени астрономических данных, Лаплас пришёл к выводу, что гипотеза равномерности может быть принята. Согласно современным представлениям большинство комет и астероидов сосредоточены в так называемом **«поясе Оорта»**. В результате вращения Нашей Галактики некоторые из них покидают пояс и захватываются притяжением Солнца.





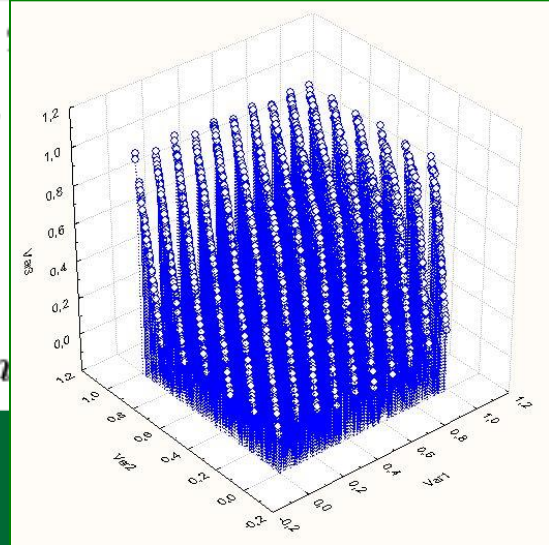
Датчики случайных чисел

➤ Мультипликативные датчики

Задаются стартовое число k_0 , множитель m .
Далее последовательно вычисляются y_1, y_2, \dots

$$\begin{cases} k_n = (m \cdot k_{n-1}) \bmod d, \\ y_n = k_n / d. \end{cases}$$

Здесь запись " $a \bmod b$ " обозначает остаток от деления a на b .

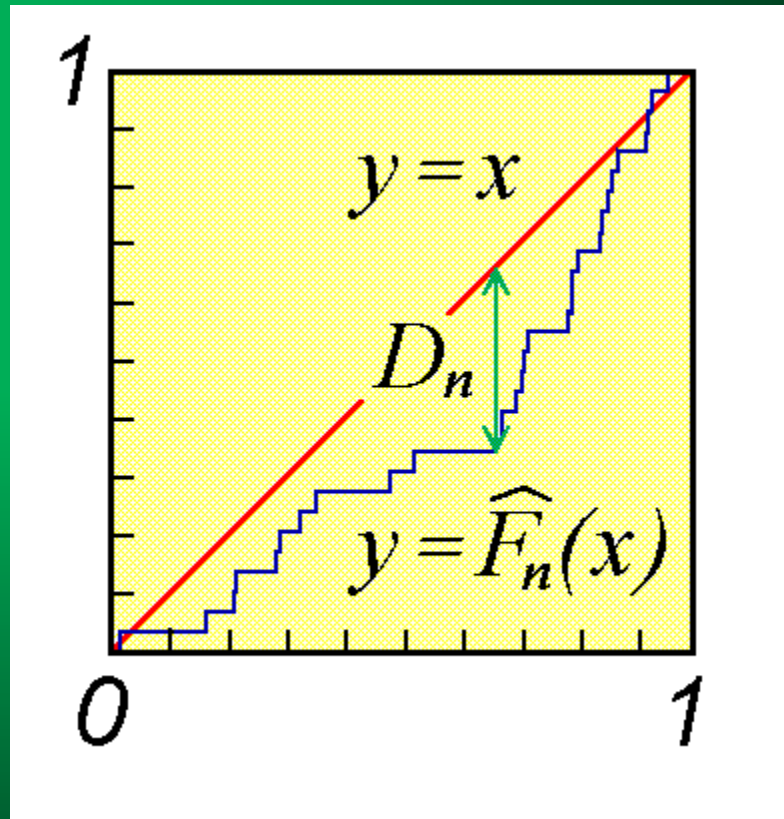


➤ Плохой датчик

Известным примером такого датчика служит **RANDU** ($d = 2^{31}$, $m = 2^{16} + 3 = 65539$), вошедший в **SSP** — библиотеку научных программ для **IBM-360**. Оказалось, что все точки с координатами $(y_{3n-2}, y_{3n-1}, y_{3n})$ располагаются **в точности** на одной из **15** плоскостей с уравнениями $9y_{3n-2} - 6y_{3n-1} + y_{3n} = k$, где $k = -5, \dots, 9$, вместо того, чтобы равномерно плотно заполнять единичный куб.

Критерий Колмогорова

В случае проверки гипотезы равномерности на $[0, 1]$ теоретической функцией распределения служит диагональ единичного квадрата



Эмпирическая функция распределения

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x\}}$$

Данный критерий базируется на величине максимального отклонения D_n эмпирической функции распределения от теоретической функции распределения


Практическое задание 1

1) Моделируйте две независимые выборки x_1 и x_2 размера $n = 100$: элементы первой выборки должны иметь равномерное распределение на отрезке $[0, 1]$, элементы второй — функцию распределения $F(x) = x^2$ на отрезке $[0, 1]$ (используйте команду `runif` и метод обратной функции для моделирования выборки x_2)

2) Постройте графики эмпирических функций распределения каждой из выборок вместе прямой $y = x$ с помощью команд `plot(ecdf(...))` и `abline`

(`plot` — график, `ecdf` — empirical cumulative (накопленная) distribution function, команда `abline` добавляет на график прямую линию с уравнением $y = a + bx$)

3) Изучите `help` функции `ks.test`, реализующей критерий Колмогорова (`ks` — сокращение для английских переводов фамилий Колмогоров и Смирнов) и проверьте каждую из моделированных выборок на равномерность критерием Колмогорова (здесь важно правильно задать аргумент `y`)



Отсутствие памяти (характеристическое свойство показательного распределения)



Допустим, что времена X_i работы прибора до i -го отказа ($i = 1, \dots, n$) образуют выборку из некоторого непрерывного закона. Нас интересует гипотеза

$$H_0: \mathbf{P}(X_1 \geq x + y \mid X_1 \geq x) = \mathbf{P}(X_1 \geq y) \quad \text{для всех } x, y \geq 0.$$

(Здесь $\mathbf{P}(A \mid B)$ — условная вероятность события A при условии события B .) Гипотеза H_0 означает, что вероятность отсутствия поломок за дополнительный период времени y при условии, что прибор уже проработал в течение периода времени x , равна вероятности того, что новый (еще не работавший) прибор прослужит начальный период времени y . Утверждение о том, что это верно для всех $x, y \geq 0$ эквивалентно утверждению о том, что работающие приборы любого “возраста” не лучше и не хуже, чем новые.

Работает ли прибор ещё на фазе «стабильного функционирования» или уже на фазе «старения»?

Практическое задание 2

В файле Condit.txt содержатся данные о количествах лётных часов между последовательными отказами установки для кондиционирования воздуха на самолете типа «Боинг-720».

1) Импортируйте файл Condit.txt (см. текст домашнего задания из темы 2), ради краткости запишите 1-й столбец таблицы в вектор x

2) Постройте для x диаграмму размахов, найдите «выбросы»

3) Постройте для x гистограмму, чтобы получить представление о форме плотности распределения исследуемого признака

4) Для проверки выборки x на согласие с показательным законом преобразуйте x в выборки y и z согласно следующим формулам:

$$y_k = x_1 + \dots + x_k, \quad z_k = y_k / y_n, \quad k = 1, \dots, n,$$

где n — размер выборки x , вычисляемый командой $n = \text{length}(x)$.

Для выполнения преобразования используйте функцию `cumsum` (накопленные суммы) или *команду цикла*: `for (k in 1:n) {...}`

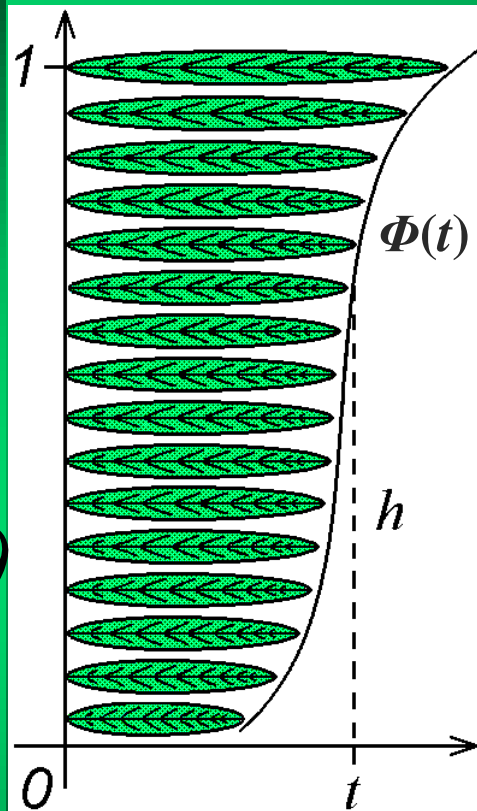
5) Замените z_n на пропуск (NA), постройте график ЭФР для z и проверьте выборку z на равномерность критерием Колмогорова

Практическое задание 3

- 1) Замените на пропуски возможные «выбросы» в выборке `x` командой `ifelse` и сохраните результат замены в самом векторе `x`
- 2) Исклучите возможные «выбросы» командой `na.omit` (`omit` — опустить, отбросить) и сохраните результат замены в векторе `x`
- 3) Выполните ещё раз пункты 2-5 из практического задания 2 для выборки `x` без «выбросов» и сформулируйте окончательный вывод

Рекомендация. Вызывайте прошлые команды с помощью клавиши `↑` или воспользуйтесь списком всех выполненных команд на вкладке `History`, находящейся в правом верхнем окне Rstudio: если дважды кликнуть по команде из списка, то она будет скопирована в командную строку RStudio

Нормальный закон



“Я до сих пор живо помню, как однажды, когда я был еще ребенком, мой отец привел меня на край города, где на берегу стояли ивы, и велел мне сорвать наугад сотню ивовых листочков. После отбора листьев с поврежденными кончиками у нас осталось 89 целых листиков. Вернувшись домой, мы расположили их в ряд по росту, как солдат. Затем мой отец через кончики листьев провел кривую и сказал: “Это и есть кривая Кетле.*) Глядя на нее, ты видишь, что посредственности всегда составляют большинство и лишь немногие поднимаются выше или так и остаются внизу”.

(Ван дер Варден)

Математики уверены в том, что нормального закона подчиняется большинство явлений. Установлено, что в природе и обществе считают его

Л. А. Ж. Кетле (1796–1874)
— бельгийский социолог.

Липпман (согласно Пуанкаре)





Визуальная проверка согласия с вероятностным законом

Проблема. Пусть элементы выборки X_1, \dots, X_n имеют функцию распределения $F((x - \mu)/\sigma)$, где F известна, а параметры сдвига μ и масштаба $\sigma > 0$ — нет. Как их оценить?

Согласно закону больших чисел эмпирическая функция распределения $\hat{F}_n(x)$ служит естественным приближением к теоретической функции распределения $F((x - \mu)/\sigma)$. Среди функций этого двухпараметрического семейства следовало бы выбрать такую функцию $F((x - \hat{\mu})/\hat{\sigma})$, чтобы она «меньше всего» отличалась от $\hat{F}_n(x)$, и взять соответствующие $\hat{\mu}$ и $\hat{\sigma}$ в качестве искомых оценок. Однако, в общем случае из-за нелинейности F это сделать затруднительно. Идея метода оценивания, приведенного ниже, состоит в «распрямлении» графика $F((x - \mu)/\sigma)$ и последующей подгонки прямой, сглаживающей соответствующее «облако» точек плоскости.



Продолжение.

Диаграмма квантилей (Q-Q plot)

Для простоты допустим, что F непрерывна и строго монотонна. Тогда для нее определена обратная функция F^{-1} . Посмотрим, во что переходит график функции $y = F((x - \mu)/\sigma)$ при преобразовании $(x, y) \rightarrow (x, F^{-1}(y))$:

$$(x, F((x - \mu)/\sigma)) \rightarrow (x, F^{-1}(F((x - \mu)/\sigma))) = (x, (x - \mu)/\sigma).$$

Значит, график переходит в прямую $y = (x - \mu)/\sigma$.

Отсюда вытекает следующий способ оценивания μ и σ : преобразуем график эмпирической функции распределения $y = \hat{F}_n(x)$ в $y = F^{-1}(\hat{F}_n(x))$ и подберем «на глаз» наиболее тесно прилегающую к нему прямую $y = (x - \hat{\mu})/\hat{\sigma}$.

Для реализации этого способа получения оценок нет необходимости строить целиком график $y = F^{-1}(\hat{F}_n(x))$. Достаточно отметить только точки $(x_{(i)}, F^{-1}(i/n))$, отвечающие скачкам функции $\hat{F}_n(x)$, и подогнать прямую к этому «облаку» точек.

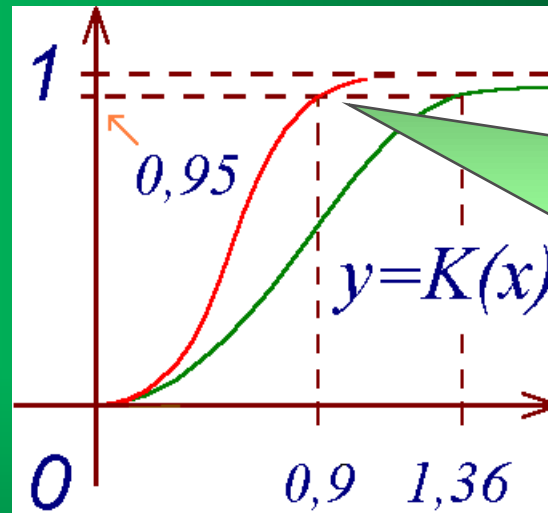
Практическое задание 4

В файле `Employees.txt` содержатся данные о заработной плате 100 сотрудников некоторой фирмы.

- 1) Импортируйте файл `Employees.txt`, ради краткости запишите признак `SALARY` в вектор `x`
- 2) Постройте для выборки `x` диаграмму размахов, гистограмму и выясните, есть ли резко выделяющиеся наблюдения («выбросы»)
- 3) Выполните визуальную проверку на согласие с нормальным законом, построив диаграмму квантилей (Q-Q plot — Quantile-Quantile Plot) с помощью функций `qqnorm` и `qqline`

Парадокс критерия Колмогорова

Подстановка оценок \bar{X} и S вместо неизвестных параметров μ и σ очень сильно меняет вид предельного закона для статистики критерия Колмогорова



Коррекция (поправка) предложена Лильефорсом (Lilliefors)

Важно отметить, что предельное распределение статистики Колмогорова при проверке (сложной) гипотезы нормальности отличается от $K(x)$. Дело в том, что для вычисления оценок \bar{X} и S используются те же самые X_1, \dots, X_n , что и при построении эмпирической функции распределения. Поэтому эмпирическая функция $\hat{F}_n(x)$ и $\Phi((x - \bar{X})/S)$ оказываются ближе друг к другу, чем в случае проверки простой гипотезы. При этом критическими становятся существенно меньшие значения статистики Колмогорова.

Ошибка в книге Байена

В книге R. H. Baayen «Analyzing Linguistic Data», Cambridge University Press, 2008, на с. 73 допущена **статистическая ошибка**: при проверке нормальности логарифма частоты встречаемости 985 голландских слов с приставкой ver (см. файл Prefix-ver.txt) вместо неизвестных параметров были подставлены их оценки без применения (!) поправки Лильефорса.

A second test that can be used is the KOLMOGOROV-SMIRNOV ONE-SAMPLE TEST. Its first argument is the observed vector of values; its second argument is the name of the density function that we want to compare our observed vector with. As we are considering a normal distribution here, this second argument is pnorm. The remaining arguments are the corresponding parameters, in this case, the mean and standard deviation which we estimate from the (log-transformed) frequency vector:

```
> ks.test(ver$Frequency, "pnorm",  
+ mean(ver$Frequency), sd(ver$Frequency))  
      One-sample Kolmogorov-Smirnov test  
data:  ver$Frequency  
D = 0.1493, p-value < 2.2e-16  
alternative hypothesis: two.sided
```

Критерий омега-квадрат

Критерий базируется на величине (интегрального) среднеквадратичного отклонения эмпирической функции распределения от теоретической функции распределения

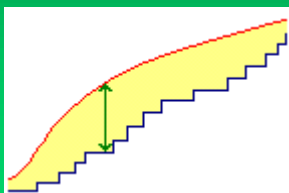
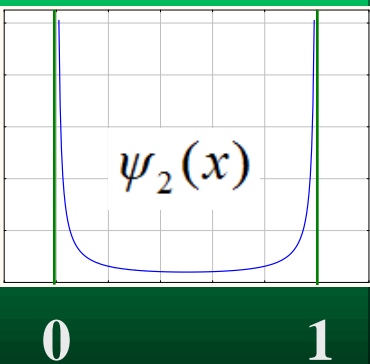


График
весовой
функции



$$\omega^2(\psi) = \int_{-\infty}^{+\infty} (\hat{F}_n(x) - F(x))^2 \psi(F(x)) dF(x)$$

где $\psi(x)$ — заданная на отрезке $[0, 1]$ весовая функция.

Рассмотрим два варианта:

$\psi_1(x) \equiv 1$ — критерий Крамера — фон Мизеса,

$\psi_2(x) = 1/[x(1-x)]$ — критерий Андерсона — Дарлинга.

Первый хорошо улавливает расхождение между эмпирической и теоретической функциями в области «типичных значений». Второй способен обнаружить различие «на хвостах» распределения, которому придается дополнительный вес благодаря функции ψ_2

Асимметрия и эксцесс

Простые критерии, которые несколько больше, чем критерий Колмогорова, учитывают поведение «хвостов» распределения, основаны на *центральных выборочных моментах*

$$M_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, \quad k = 1, 2, \dots$$

При помощи величин M_2 , M_3 и M_4 вычисляются *выборочные коэффициенты асимметрии G_1 и эксцесса G_2* :

$$G_1 = M_3/M_2^{3/2}, \quad G_2 = M_4/M_2^2 - 3.$$

Эти случайные величины можно использовать в качестве оценок для (независящих от сдвига и масштаба) *теоретических коэффициентов асимметрии $\gamma_1 = \mu_3/\mu_2^{3/2}$ и эксцесса $\gamma_2 = \mu_4/\mu_2^2 - 3$* , где $\mu_k = \mathbf{M}(X_1 - \mathbf{M}X_1)^k$ — *центральные теоретические моменты*. (Для нормального закона $\gamma_1 = \gamma_2 = 0$.)

К. Пирсон в 1930 г. показал, что выборочный коэффициент асимметрии довольно быстро сходится к теоретическому, а выборочный коэффициент эксцесса сходится очень медленно.

Критерии Гири и Дэвида – Хартли – Пирсона

3.2.2.9. Критерий среднего абсолютного отклонения (критерий Гири)

Гири [261–263] рассмотрел критерий нормальности распределения случайных величин, основанный на статистике

$$d = \frac{1}{ns} \sum_{i=1}^n |x_i - \bar{x}|,$$

«Тяжёлые
хвосты»

являющейся отношением среднего абсолютного отклонения $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$ к выборочному стандартному отклонению

$$s = \left\{ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right\}^{\frac{1}{2}}.$$

3.2.2.10. Критерий Дэвида–Хартли–Пирсона

В [264] предложен критерий нормальности распределения вероятностей случайной величины, основанный на распределении отношения размаха к стандартному отклонению.

Статистика критерия имеет вид

$$U = \frac{R}{s},$$

«Лёгкие
хвосты»

где $R = x_{\max} - x_{\min}$ (или $(x_n - x_1)$ для упорядоченного по возрастанию ряда выборочных значений) — размах выборки; s — стандартное отклонение.

Критерий Шапиро – Уилка

3.2.2.1. Критерий Шапиро–Уилка

Критерий Шапиро–Уилка [240] основан на отношении оптимальной линейной несмещенной оценки дисперсии (см. раздел 2.1.2.1.6.6) к ее обычной оценке методом максимального правдоподобия (см. раздел 2.1.2.1.1). Статистика критерия имеет вид

$$W = \frac{1}{s^2} \left[\sum_{i=1}^k a_{n-i+1} (x_{n-i+1} - x_i) \right]^2, \quad \text{где} \quad s^2 = \sum_{i=1}^n (x_i - \bar{x})^2; \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Здесь $x_1 \leq x_2 \leq \dots \leq x_n$ — упорядоченная выборка.

Изучение мощности критерия Шапиро–Уилка [243] показало, что это — один из наиболее эффективных критериев проверки нормальности распределения случайных величин. Для больших n таблицы коэффициентов a_{n-i+1} становятся неудобными, поэтому была предложена модификация критерия Шапиро–Уилка — критерий Шапиро–Франча [244].

Его статистика имеет вид

$$W' = \frac{1}{s^2} \left[\sum_{i=1}^k c_{n-i+1} (x_{n-i+1} - x_i) \right]^2, \quad \text{где} \quad c_{n-i+1} = \frac{m_{n-i+1}}{\left(\sum_{i=1}^n m_{i,n}^2 \right)^{\frac{1}{2}}}$$

и $m_{i,n}$ — математическое ожидание i -й порядковой статистики из стандартного нормального распределения.

П. Ройстон (1995) для $n \leq 5000$.

Практическое задание 5

В практическом задании 4 был импортирован файл `Employees.txt` и признак `SALARY` из него был ради краткости записан в вектор `x`.

1) Для проверки нормальности выборки `x` с помощью разных критериев установите в Rstudio пакет `nortest`: в правом нижнем окне на вкладке `Packages` нажмите кнопку `Install Packages`, введите имя пакета и нажмите `Install` (для установки пакета нужен Internet), затем активизируйте пакет `nortest`, поставив перед ним «галку» в списке пакетов на вкладке `Packages`

2) Проверьте согласие с нормальным законом с помощью функций:

- а) `lillie.test` — критерий Колмогорова с поправкой Лиллиефорса,
- б) `cvm.test` — критерий Крамера — фон Мизеса,
- в) `ad.test` — критерий Андерсона — Дарлинга,
- г) `sf.test` — критерий Шапиро — Франчия

3) Импортируйте файл `Prefix-ver.txt` и проверьте нормальность распределения признака `LogFrequency` с помощью диаграммы квантилей и критериев из пункта 2




Главное в теме

- Если фактический уровень значимости критерия (p -value) меньше 0,05, то гипотеза равномерности, показательности или нормальности отвергается, иначе — принимается (но нельзя считать это правило пригодным во всех возможных случаях)
- Критерий Андерсона — Дарлинга лучше, чем критерий Колмогорова, улавливает расхождение между эмпирической и теоретической функциями в области «хвостов» распределения
- Критерий Шапиро — Уилка обычно оказывается самым эффективным для проверки нормальности
- При проверке гипотезы нормальности критерием Колмогорова подстановка вместо неизвестных параметров оценок максимального правдоподобия приводит к парадоксальному изменению предельного распределения

Those who
ignore
statistics
are condemned
to reinvent it.

Brad Efron



Первое — это понять правило,
второе — научиться его применять.
Первое достигается разумом и сразу,
второе — опытом и постепенно.

Артур Шопенгауэр

Домашнее задание

В файле Uniform.txt содержится выборка размера $n = 100$, моделированная из равномерного распределения на $[0, 1]$.

1) На её основе получите методом обратной функции выборку из *логистического закона*, имеющего функцию распределения

$$F(x) = 1 / (1 + e^{-x})$$

2) Проверьте гипотезу нормальности для моделированной выборки критериями Колмогорова, Крамера — Мизеса, Андерсона — Дарлинга, Шапиро — Уилка

3) Визуально проверьте согласие с нормальностью с помощью диаграммы квантилей

4) Объясните, почему критерии не отвергают гипотезу нормальности распределения для выборки с логистическим распределением элементов. (Объяснение, что нормальное и логистическое распределения похожи, не является достаточным.)