JamaGava 30 сентября 2016 в 10:46

«Правда, чистая правда и статистика» или ‹ вероятности на все случаи жизни»

Занимательные задачки, Анализ и проектирование систем, Алгоритмы, Математика

Tutorial

Статистика приходит к нам на помощь при решении многих задач, например: когда нет возможности построить детерминированную модель, когда слишком много факторов или когда нам необходимо оценить правдоподобие построенной модели с учётом имеющихся данных. Отношение к статистике неоднозначное. Есть мнение, что существует три вида лжи: ложь, наглая ложь и статистика. С другой стороны, многие «пользователи» статистики слишком ей верят, не понимая до конца, как она работает: применяя, например, тест Стьюдента к любым данным без проверки их нормальности. Такая небрежность способна порождать серьёзные ошибки и превращать «поклонников» теста Стьюдента в ненавистников статистики. Попробуем поставить точки над i и разобраться, какие модели случайных величин должны использоваться для описания тех или иных явлений и какая между ними существует генетическая связь.

В первую очередь, данный материал будет интересен студентам, изучающим теорию вероятностей и статистику, хотя и «зрелые» специалисты смогут его использовать в качестве справочника. В одной из следующих работ я покажу пример использования статистики для построения теста оценки значимости показателей биржевых торговых стратегий.

В работе будут рассмотрены дискретные распределения:

1. Бернулли;
2. биномиальное;
3. геометрическое;
4. Паскаля (отрицательное биномиальное);
5. гипергеометрическое;
6. Пуассона,

а также непрерывные распределения:

1. Гаусса (нормальное);
2. хи-квадрат;
3. Стьюдента;
4. Фишера;
5. Коши;
6. экспоненциальное (показательное) и Лапласа (двойное экспоненциальное, двойное показательное);
7. Вейбулла;
8. гамма (Эрланга);
9. бета.

В конце статьи будет задан вопрос для размышлений. Свои размышления по этому поводу я изложу в следующей статье.

Некоторые из приведённых непрерывных распределений являются частными случаями распределения Пирсона.

Дискретные распределения

Дискретные распределения используются для описания событий с недифференцируемыми характеристиками, определёнными в изолированных точках. Проще говоря, для событий, исход которых может быть отнесён к некоторой дискретной категории: успех или неудача, целое число (например, игра в рулетку, в кости), орёл или решка и т.д.

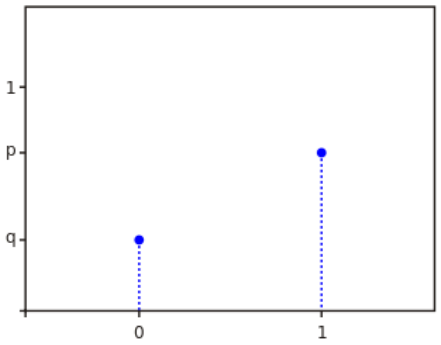
Описывается дискретное распределение вероятностью наступления каждого из возможных исходов события. Как и для любого распределения (в том числе непрерывного) для дискретных событий определены понятия матожидания и дисперсии. Однако, следует понимать, что матожидание для дискретного случайного события — величина в общем случае нереализуемая как исход одиночного случайного события, а скорее как величина, к которой будет стремиться среднее арифметическое исходов событий при увеличении их количества.

В моделировании дискретных случайных событий важную роль играет комбинаторика, так как вероятность исхода события можно определить как отношение количества комбинаций, дающих требуемый исход к общему количеству комбинаций. Например: в корзине лежат 3 белых мяча и 7 чёрных. Когда мы выбираем из корзины 1 мяч, мы можем сделать это 10-ю разными способами (общее количество комбинаций), но только 3 варианта, при которых будет выбран белый мяч (3 комбинации, дающие требуемый исход). Таким образом, вероятность выбрать белый мяч: $\frac{3}{10}$ (распределение Бернулли).

Следует также отличать выборки с возвращением и без возвращения. Например, для описания вероятности выбора двух белых мячей важно определить, будет ли первый мяч возвращён в корзину. Если нет, то мы имеем дело с выборкой без возвращения (гипергеометрическое распределение) и вероятность будет такова: $\frac{3}{10} \times \frac{2}{9}$ — вероятность выбрать белый мяч из начальной выборки умноженная на вероятность снова выбрать белый мяч из оставшихся в корзине. Если же первый мяч возвращается в корзину, то это выборка с возвращением (Биномиальное распределение). В этом случае вероятность выбора двух белых мячей составит $\left(\frac{3}{10}\right)^2$.

наверх

Распределение Бернулли



(взято отсюда)

Если несколько формализовать пример с корзиной следующим образом: пусть исход события может принимать одно из двух значений 0 или 1 с вероятностями q и p соответственно, тогда распределение вероятности получения каждого из предложенных исходов будет называться распределение Бернулли:

$$Bin_{p,q}(x) = \begin{cases} q, x = 0 \\ p, x = 1 \end{cases} \quad (1.1.1)$$

По сложившейся традиции, исход со значением 1 называется «успех», а исход со значением 0 — «неудача». Очевидно, что получение исхода «успех или неудача» наступает с вероятностью $p + q = 1$.

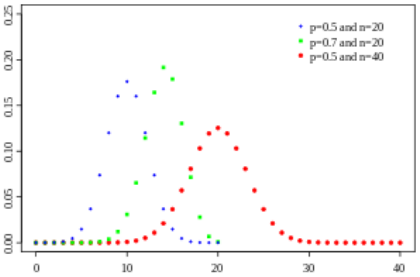
Матожидание и дисперсия распределения Бернулли:

$$E\{Bin_{p,q}\} = p \quad (1.1.2)$$

$$D\{Bin_{p,q}\} = pq = p\,(1-p) \qquad (1.1.3)$$

наверх

Биномиальное распределение



(взято отсюда)

Количество k успехов в n испытаниях, исход которых распределен по Бернулли с вероятностью успеха p (пример с возвращением мячей в корзину), описывается биномиальным распределением:

$$B_{n,p}(k) = C_n^k p^k q^{n-k} \quad (1.2.1)$$

где $C_n^k = \dfrac{n!}{k!(n-k)!}$ — число сочетаний из n по k.

По другому можно сказать, что биномиальное распределение описывает сумму из n независимых случайных величин, умеющих распределение Бернулли с вероятностью успеха n.
Матожидание и дисперсия:

$$E\{B_{n,p}\} = np \quad (1.2.2)$$

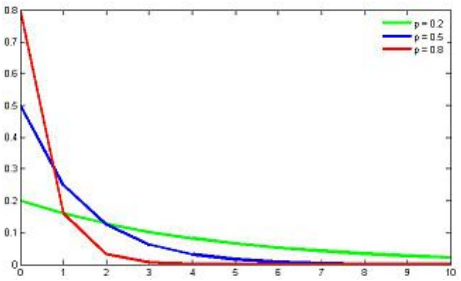
$$D\{B_{n,p}\} = npq \quad (1.2.3)$$

Биномиальное распределение справедливо только для выборки с возвращением, то есть, когда вероятность успеха остаётся постоянной для всей серии испытаний.

Если величины X и Y имеют биномиальные распределения с параметрами $\{n_x, p\}$ и $\{n_y, p\}$ соответственно, то их сумма также будет распределена биномиально с параметрами $\{n_x + n_y, p\}$.

наверх

Геометрическое распределение



(взято отсюда)

Представим ситуацию, что мы вытягиваем мячи из корзины и возвращаем обратно до тех пор, пока не будет вытянут белый шар. Количество таких операций описывается геометрическим распределением. Иными словами: геометрическое распределение описывает количество испытаний n до первого успеха при вероятности наступления успеха в каждом испытании p. Если n подразумевается номер испытания, в котором наступил успех, то геометрическое распределение будет описываться следующей формулой:

$$Geom_p(n) = q^{n-1}p \quad (1.3.1)$$

Матожидание и дисперсия геометрического распределения:

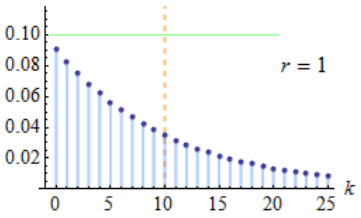
$$E\{Geom_p\} = \frac{1}{p} \quad (1.3.2)$$

$$D\{Geom_p\} = \frac{q}{p^2} \quad (1.3.3)$$

Геометрическое распределение генетически связано с экспоненциальным распределением, которое описывает непрерывную случайную величину: время до наступления события, при постоянной интенсивности событий. Геометрическое распределение также является частным случаем отрицательного биномиального распределения.

наверх

Распределение Паскаля (отрицательное биномиальное рспределение)



(взято отсюда)

Распределение Паскаля является обобщением геометрического распределения: описывает распределение количества неудач k в независимых испытаниях, исход которых распределен по Бернулли с вероятностью успеха p до наступления r успехов в сумме. При $r = 1$, мы получим геометрическое распределение для величины $k + 1$.

$$NB_{r,p}(k) = C^k_{k+r-1}p^r q^k \quad (1.4.1)$$

где $C^k_n = \frac{n!}{k!(n-k)!}$ — число сочетаний из n по k.

Матожидание и дисперсия отрицательного биномиального распределения:

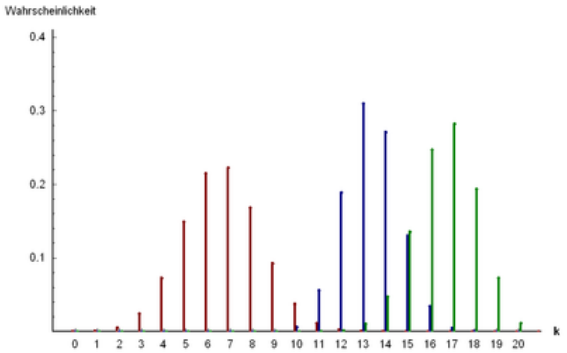
$$E\{NB_{r,p}\} = \frac{rq}{p} \quad (1.4.2)$$

$$D\{NB_{r,p}\} = \frac{rq}{p^2} \quad (1.4.3)$$

Сумма независимых случайных величин, распределённых по Паскалю, также распределена по Паскалю: пусть X имеет распределение $NB_{r_x,p}$, а Y — $NB_{r_y,p}$. Пусть также X и Y независимы, тогда их сумма будет иметь распределение $NB_{r_x+r_y,p}$

наверх

Гипергеометрическое распределение



(взято отсюда)

До сих пор мы рассматривали примеры выборок с возвращением, то есть, вероятность исхода не менялась от испытания к испытанию.

Теперь рассмотрим ситуацию без возвращения и опишем вероятность количества успешных выборок из совокупности с заранее известным количеством успехов и и неудач (заранее известное количество белых и чёрных мячей в корзине, козырных карт в колоде, бракованных деталей в партии и т.д.).

Пусть общая совокупность содержит N объектов, из них D помечены как «1», а $N - D$ как «0». Будем считать выбор объекта с меткой «1», как успех, а с меткой «0» как неудачу. Проведём n испытаний, причём выбранные объектв больше не будут участвовать в дальнейших испытаниях. Вероятность наступления k успехов будет подчиняться гипергеометрическому распределению:

$$HG_{N,D,n}(k) = \frac{C_D^k C_{N-D}^{n-k}}{C_N^n} \quad (1.5.1)$$

где $C_n^k = \frac{n!}{k!(n-k)!}$ — число сочетаний из n по k.

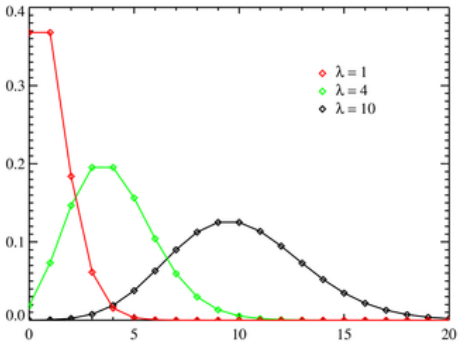
Матожидание и дисперсия:

$$E\{HG_{N,D,n}\} = \frac{nD}{N} \quad (1.5.2)$$

$$D\{HG_{N,D,n}\} = n\frac{D}{N}\frac{N-D}{N}\frac{N-n}{N-1} \quad (1.5.3)$$

наверх

Распределение Пуассона



(взято отсюда)

Распределение Пуассона значительно отличается от рассмотренных выше распределений своей «предметной» областью: теперь рассматривается не вероятность наступления того или иного исхода испытания, а интенсивность событий, то есть среднее количество событий в единицу времени.

Распределение Пуассона описывает вероятность наступления k независимых событий за время t при средней интенсивности событий λ:

$$P_{\lambda,t}(k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad (1.6.1)$$

Матожидание и дисперсия распределения Пуассона:

$$E\{P_{\lambda,t}\} = \lambda t \quad (1.6.2)$$

$$D\{P_{\lambda,t}\} = \lambda t \quad (1.6.3)$$

Дисперсия и матожидание распределения Пуассона тождественно равны.

Распределение Пуассона в сочетании с экспоненциальным распределением, описывающим интервалы времени между

наступлениями независимых событий, составляют математическую основу теории надёжности.

наверх

Непрерывные распределения

Непрерывные распределения, в отличие от дискретных, описываются функциями плотности (распределения) вероятности $f(x)$, определёнными, в общем случае, на некоторых интервалах.

Если известна плотность вероятности для величины x: $f(x)$ и определено преобразование $y = g(x)$, то плотность вероятности для y может быть получена автоматически:

$$f_y(y) = f\left(g^{-1}(y)\right) \left| \frac{dg^{-1}}{dy}(y) \right| \quad (2.0.1)$$

при условии однозначности и дифференцируемости $g^{-1}(x)$.

Плотность вероятности $h(z)$ суммы случайных величин x и y ($z = x + y$) с распределениями $f(x)$ и $g(y)$ описывается свёрткой f и g:

$$h(z) = \int f(t)g(z-t)dt = (f * g)(z) \quad (2.0.2)$$

Если распределение суммы случайных величин принадлежит к тому же распределению, что и слагаемые, такое распределение называется бесконечно делимым. Примеры бесконечно делимых распределений: нормальное, хи-квадрат, гамма, распределение Коши.

Плотность вероятности $h(z)$ произведения случайных величин x и y ($z = xy$) с распределениями $f(x)$ и $g(y)$ может быть вычислена следующим образом:

$$h(z) = \int f(t)g(z/t)dt \quad (2.0.3)$$

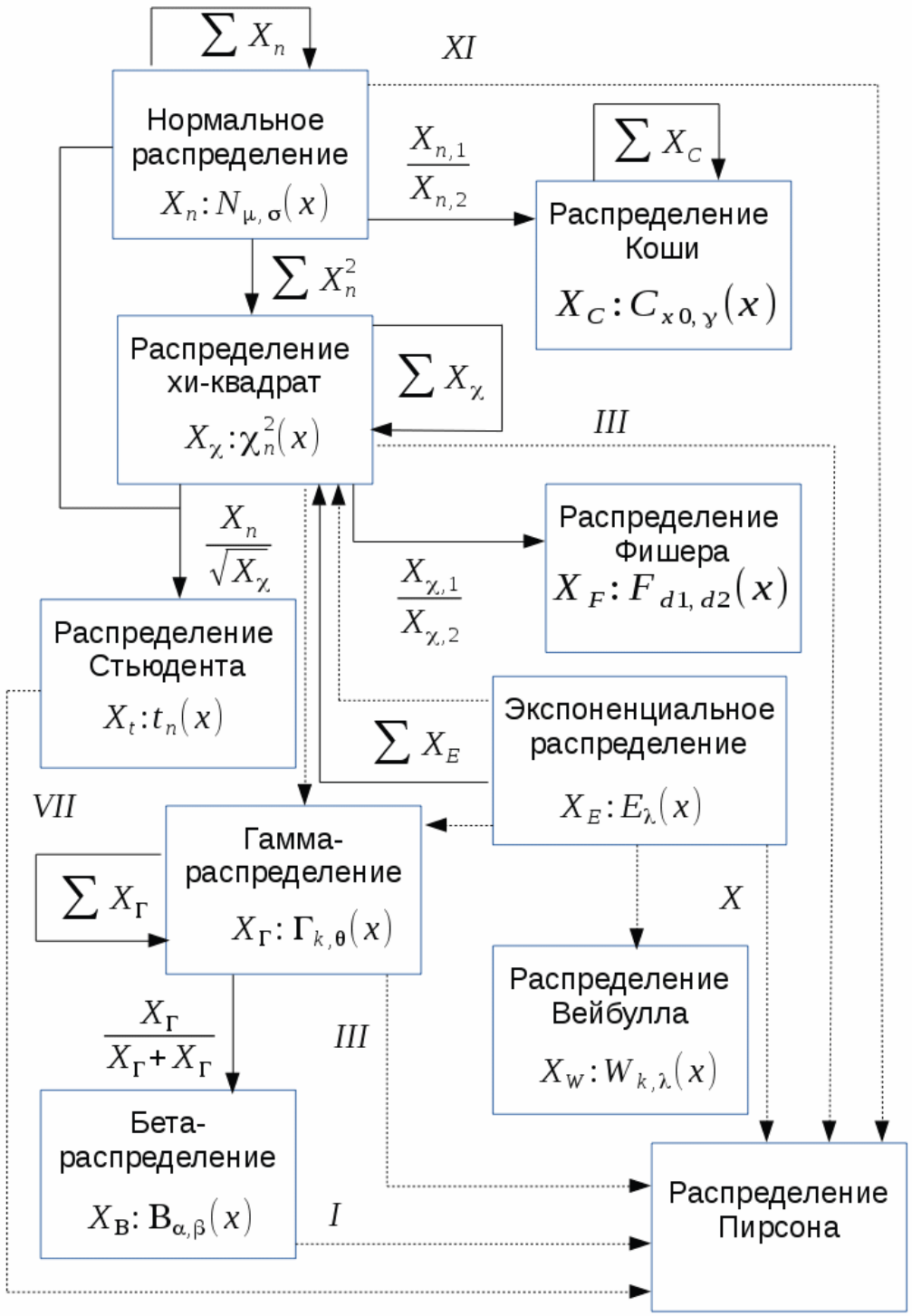
Некоторые из приведённых ниже распределений являются частными случаями распределения Пирсона, которое, в свою очередь, является решением уравнения:

$$\frac{df}{dx}(x) = \frac{a_0 + a_1 x}{b_0 + 2b_1 x + b_2 x^2} f(x) \quad (2.0.4)$$

где a_i и b_i — параметры распределения. Известны 12 типов распределения Пирсона, в зависимости от значений параметров.

Распределения, которые будут рассмотрены в этом разделе, имеют тесные взаимосвязи друг с другом. Эти связи выражаются в том, что некоторые распределения являются частными случаями других распределений, либо описывают преобразования случайных величин, имеющих другие распределения.

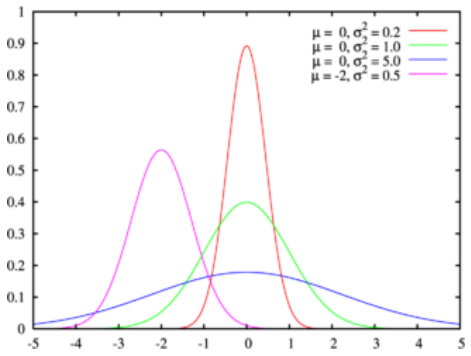
На приведённой ниже схеме отражены взаимосвязи между некоторыми из непрерывных распределений, которые будут рассмотрены в настоящей работе. На схеме сплошными стрелками показано преобразование случайных величин (начало стрелки указывает на изначальное распределение, конец стрелки — на результирующее), а пунктирными — отношение обобщения (начало стрелки указывает на распределение, являющееся частным случаем того, на которое указывает конец стрелки). Для частных случаев распределения Пирсона над пунктирными стрелками указан соответствующий тип распределения Пирсона.



Предложенный ниже обзор распределений охватывает многие случаи, которые встречаются в анализе данных и моделировании процессов, хотя, конечно, и не содержит абсолютно все известные науке распределения.

наверх

Нормальное распределение (распределение Гаусса)



(взято отсюда)

Плотность вероятности нормального распределения $N_{\mu,\sigma}(x)$ с параметрами μ и σ описывается функцией Гаусса:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.1.1)$$

Если $\sigma = 1$ и $\mu = 0$, то такое распределение называется стандартным.

Матожидание и дисперсия нормального распределения:

$$E\{N_{\mu,\sigma}\} = \mu \quad (2.1.2)$$

$$D\{N_{\mu,\sigma}\} = \sigma^2 \quad (2.1.3)$$

Область определения нормального распределения — множество дествительных чисел.

Нормальное распределение является распределение Пирсона типа XI.

Сумма квадратов независимых нормальных величин имеет распределение хи-квадрат, а отношение независимых Гауссовых величин распределено по Коши.

Нормальное распределение является бесконечно делимым: сумма нормально распределенных величин x и y с параметрами $\{\mu_x, \sigma_x\}$ и $\{\mu_y, \sigma_y\}$ соответственно также имеет нормальное распределение с параметрами $\{\mu_{x+y}, \sigma_{x+y}\}$, где $\mu_{x+y} = \mu_x + \mu_y$ и $\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y$.

Нормальное распределение хорошо моделирует величины, описывающие природные явления, шумы термодинамической природы и погрешности измерений.

Кроме того, согласно центральной предельной теореме, сумма большого количества независимых слагаемых одного порядка сходится к нормальному распределению, независимо от распределений слагаемых. Благодаря этому свойству, нормальное распределение популярно в статистическом анализе, многие статистические тесты рассчитаны на нормально распределенные данные.

На бесконечной делимости нормального распределении основан z-тест. Этот тест используется для проверки равенства матожидания выборки нормально распределённых величин некоторому значению. Значение дисперсии должно быть **известно**. Если значение дисперсии неизвестно и рассчитывается на основании анализируемой выборки, то применяется t-тест, основанный на распределении Стьюдента.

Пусть у нас имеется выборка объёмом n независимых нормально распределенных величин X_i из генеральной совокупности со стандартным отклонением σ выдвинем гипотезу, что $\bar{X} = \mu$. Тогда величина $z = \frac{\bar{X} - \mu}{\sigma\sqrt{n}}$ будет иметь стандартное нормальное распределение. Сравнивая полученное значение z с квантилями стандартного распределения можно принимать или отклонять гипотезу с требуемым уровнем значимости.

Благодаря широкой распространённости распределения Гаусса, многие, не очень хорошо знающие статистику исследователи забывают проверять данные на нормальность, либо оценивают график плотности распределения «на глазок», слепо полагая, что имеют дело с Гауссовыми данными. Соответственно, смело применяя тесты, предназначенные для нормального распределения и получая совершенно некорректные результаты. Наверное, отсюда и пошла молва про статистику как самый страшный вид лжи.

Рассмотрим пример: нам надо измерить сопротивления набора резистров некоторого номинала. Сопротивление имеет

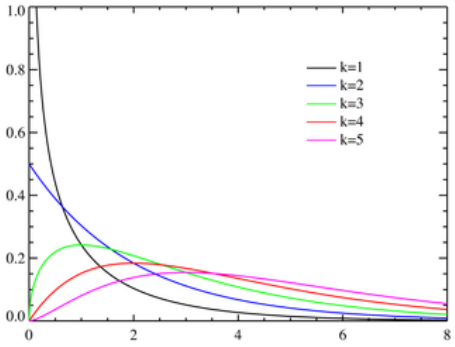
физическую природу, логично предположить, что распределение отклонений сопротивления от номинала будет нормальным. Меряем, получаем колоколообразную функцию плотности вероятности для измеренных значений с модой в окрестности номинала резистров. Это нормальное распределение? Если да, то будем искать бракованные резистры используя тест Стьюдента, либо z-тест, если нам заранее известна дисперсия распределения. Думаю, что многие именно так и поступят.

Но давайте внимательнее посмотрим на технологию измерения сопротивления: сопротивление определяется как отношение приложенного напряжения к протекающему току. Ток и напряжение мы измеряли приборами, которые, в свою очередь, имеют нормально распределенные погрешности. То есть, измеренные значения тока и напряжения — это **нормально распределенные случайные величины** с матожиданиями, соответствующими истинным значениям измеряемых величин. А это значит, что полученные значения сопротивления распределены по Коши, а не по Гауссу.

Распределение Коши лишь напоминает внешне нормальное распределение, но имеет более тяжёлые хвосты. А значит предложенные тесты неуместны. Надо строить тест на основании распределения Коши или вычислить квадрат сопротивления, который в данном случае будет иметь распределение Фишера с параметрами (1, 1).

к схеме
наверх

Распределение хи-квадрат



(взято отсюда)

Распределение χ^2 описывает сумму n квадратов случайных величин X_i, каждая из которых распределена по стандартному нормальному закону $N_{0,1}$:

$$\chi^2_k(x) = \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad (2.2.1)$$

где n — число степеней свободы, $x = \sum_{i=1}^{n} X_i^2$.

Матожидание и дисперсия распределения χ^2:

$$E\{\chi^2_n\} = n \quad (2.2.2)$$

$$D\{\chi^2_n\} = 2n \quad (2.2.3)$$

Область определения — множество неотрицательных натуральных чисел. χ^2 является бесконечно делимым распределением. Если x и y — распределены по χ^2 и имеют n_x и n_y степеней свободы соответственно, то их сумма также будет распределена по χ^2 и иметь $n_x + n_y$ степеней свободы.

χ^2 является частным случаем гамма-распределения (а следовательно, распределением Пирсона типа III) и обобщением экспоненциального распределения. Отношение величин, распределенных по χ^2 распределено по Фишеру.

На распределении χ^2 основан критерий согласия Пирсона. с помощью этого критерия можно проверять достоверность принадлежности выборки случайной величины некоторому теоретическому распределению.

Предположим, что у нас имеется выборка некоторой случайной величины X_i. На основании этой выборки рассчитаем вероятности P_k попадания значений X в n интервалов ($k = 1 : n$). Пусть также есть предположение об аналитическом

выражении распределения, в соответствие с которым, вероятности попадания в выбранные интервалы должны составлять Q_k. Тогда величины $D_k = P_k - Q_k$ будут распределены по нормальному закону.

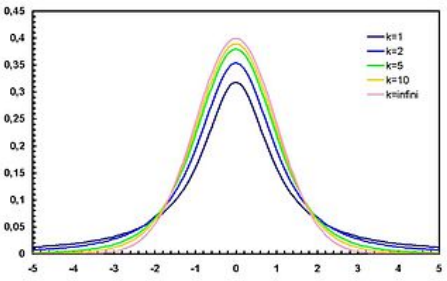
Приведем D_k к стандартному нормальному распределению: $d_k = \frac{D_k - m}{S}$,
где $m = \frac{1}{n}\sum_{i=1}^{n} D_i$ и $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} D_i^2}$.

Полученные величины d_i имеют нормальное распределение с параметрами (0, 1), а следовательно, сумма их квадратов распределена по χ^2 с $n-1$ степенью свободы. Снижение степени свободы связано с дополнительным ограничением на сумму вероятностей попадания значений в интервалы: она должна быть равна 1.

Сравнивая значение $z = \sum_{i=1}^{n} d_i^2$ с квантилями распределения χ^2 можно принять или отклонить гипотезу о теоретическом распределении данных с требуемым уровнем значимости.

к схеме
наверх

Распределение Стьюдента (t-распределение)



(взято отсюда)

Распределение Стьюдента используется для проведения t-теста: теста на равенство матожидания выборки стандартно нормально распределённых случайных величин некоторому значению, либо равенства матожиданий двух нормальных выборок с одинаковой дисперсией (равенство дисперсий необходимо проверять f-тестом). Распределение Стьюдента описывает отношение нормально распределённой случайной величины к величине, распределённой по хи-квадрат.

T-тест является аналогом z-теста для случая, когда дисперсия или стандартное отклонение выборки неизвестно и должно быть оценено на основании самой выборки.

Рассмотрим пример проверки равенства матожидания нормальной выборки некоторому значению: пусть нам дана выборка X_i нормальных величин объёмом n из некоторой генеральной совокупности, выдвинем и проверим гипотезу о том, что матожидание этой совокупности равно m.

Рассчитаем величину $S = \frac{\sum_{i=1}^{n} X_i^2}{n-1}$. Эта величина будет иметь распределение хи-квадрат. Тогда величина $t = \frac{\frac{\sum_{i=1}^{n} X_i - m}{n-1}}{\frac{S}{\sqrt{n}}}$
будет иметь распределение Стьюдента $T_{n-1}(x)$ с $n-1$ степенью свободы, где:

$$T_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}} \quad (2.3.1)$$

где $\Gamma(x)$ — гамма-функция Эйлера.

Полученное значение можно сравнивать с квантилями распределения Стьюдента и принимать либо отклонять гипотезу о равенстве маотожидания значению m с требуемым уровнем значимости.

Матожидание и дисперсия распределения Стьюдента:

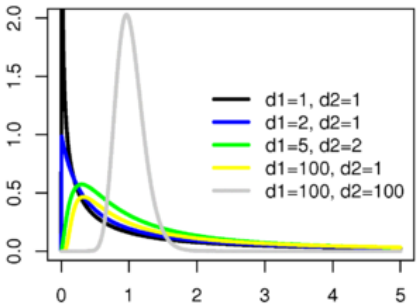
$$E\{T_n\} = 0 \quad (2.3.2)$$

$$D\{T_n\} = \frac{n}{n-2} \quad (2.3.3)$$

при $n > 2$.

к схеме
наверх

Распределение Фишера



(взято отсюда)

Пусть X и Y независимые случайные величины, имеющие распределение хи-квадрат со степенями свободы n_x и n_y соответственно. Тогда величина $F = \frac{n_y X}{n_x Y}$ будет иметь распределение Фишера со степенями свободы (n_x, n_y), а величина F^{-1} — распределение Фишера со степенями свободы (n_y, n_x).
Распределение Фишера определено для действительных неотрицательных аргументов и имеет плотность вероятности:

$$F_{n_1,n_2}(x) = \frac{\sqrt{\frac{(n_1 x)^{n_1} n_2^{n_2}}{(n_1 x + n_2)^{n_1+n_2}}}}{x B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \quad (2.4.1)$$

Матожидание и дисперсия распределения Фишера:

$$E\{F_{n_1,n_2}\} = \frac{n_2}{n_2 - 2} \quad (2.4.2)$$

$$D\{F_{n_1,n_2}\} = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} \quad (2.4.3)$$

Матожидание определено для $n_2 > 2$, а дисперсия — для $n_2 > 4$.

На распределении Фишера основан ряд статистических тестов, таких как оценка значимости параметров регрессии, тест на гетероскедастичность и тест на равенство дисперсий нормальных выборок (f-тест, следует отличать от **точного** теста Фишера).

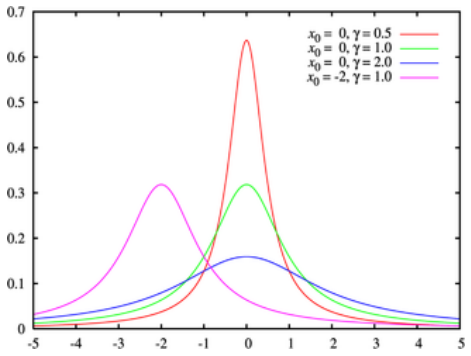
F-тест: пусть имеются две независимые выборки x_i и y_j нормально распределенных данных объёмами n_x и n_y соответственно. Выдвинем гипотезу о равенстве дисперсий выборок и проверим её статистически.

Рассчитаем величину $F = \frac{n_y \sum_{i=1}^{n_x} x_i^2}{n_x \sum_{j=1}^{n_y} y_j^2}$. Она будет иметь распределение Фишера со степенями свободы $(n_x - 1, n_y - 1)$.

Сравнивая значение F с квантилями соответствующего распределения Фишера, мы можем принять или отклонить гипотезу о равенстве дисперсий выборок с требуемым уровнем значимости.

к схеме
наверх

Распределение Коши



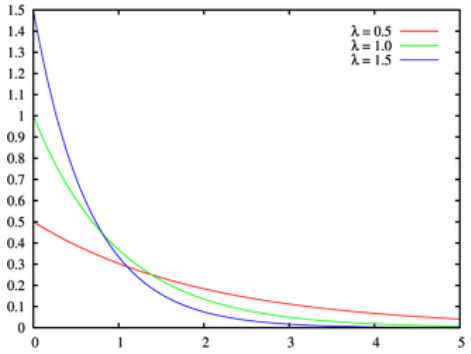
(взято отсюда)
Распределение Коши описывает отношение двух нормально распределенных случайных величин. В отличие от других распределений, для распределения Коши не определены матожидание и дисперсия. Для описания распределения используются коэффициенты сдвига x_0 и масштаба γ.

$$C_{x_0,\gamma}(x) = \frac{1}{\pi\gamma\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)} \quad (2.5.1)$$

Распределение Коши является бесконечно делимым: сумма независимых случайных величин, распределённых по Коши, также распределена по Коши.

к схеме
наверх

Экспоненциальное (показательное) распределение и распределение Лапласа (двойное экспоненциальное, двойное показательное)



(взято отсюда)

Экспоненциальное распределение описывает интервалы времени между независимыми событиями, происходящими со средней интенсивностью λ. Количество наступлений такого события за некоторый отрезок времени описывается дискретным распределением Пуассона. Экспоненциальное распределение вместе с распределением Пуассона составляют математическую основу теории надёжности.

Кроме теории надёжности, экспоненциальное распределение применяется в описании социальных явлений, в экономике, в теории массового обслуживания, в транспортной логистике — везде, где необходимо моделировать поток событий.

Экспоненциальное распределение является частным случаем распределения хи-квадрат (для n=2), а следовательно, и гамма-распределения. Так-как экспоненциально распределённая величина является величиной хи-квадрат с 2-мя степенями свободы, то она может быть интерпретирована как сумма квадратов двух независимых нормально распределенных величин.

Кроме того, экспоненциальное распределение является честным случаем распределения Вейбулла.

Дискретный вариант экспоненциального распределения — это геометрическое распределение.

Плотность вероятности экспоненциально распределения:

$$E_\lambda(x) = \lambda e^{-\lambda x} \quad (2.6.1)$$

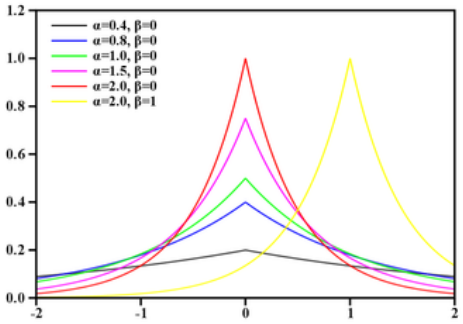
определена для неотрицательных действительных значений x.

Матожидание и дисперсия экспоненциального распределения:

$$E\{E_\lambda\} = \frac{1}{\lambda} \quad (2.6.2)$$

$$E\{E_\lambda\} = \frac{1}{\lambda^2} \quad (2.6.3)$$

двойным показательным.



(взято отсюда)

Для большего обобщения, вводится параметр сдвига, смещающий центр «соединения» левой и правой частей распределения вдоль оси абсцисс. В отличие от экспоненциального, распределение Лапласа, определено на всей действительной числовой оси.

$$L_{\alpha,\beta}(x) = \frac{\alpha}{2} e^{-\alpha|x-\beta|} \quad (2.6.4)$$

где α — параметр масштаба, а β — параметр сдвига.

Матожидание и дисперсия:

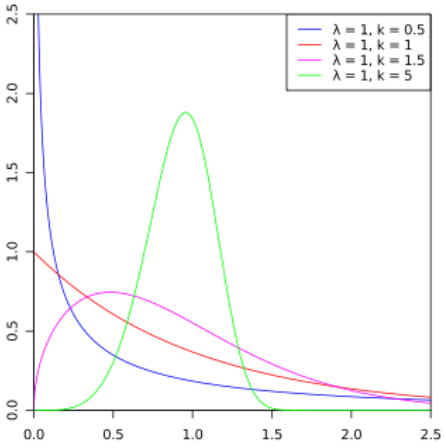
$$E\{L_{\alpha,\beta}\} = \beta \quad (2.6.5)$$

$$D\{L_{\alpha,\beta}\} = \frac{2}{\alpha^2} \quad (2.6.6)$$

Благодаря более тяжёлым хвостам, чем у нормального распределения, распределение Лапласа используется для моделирования некоторых видов погрешностей измерения в энергетике, а также находит применение в физике, экономике, финансовой статистике, телекоммуникации и т.д.

к схеме
наверх

Распределение Вейбулла



(взято отсюда)

Распределение Вейбулла описывается функцией плотности вероятности следующего вида:

$$W_{k,\lambda}(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} \quad (2.7.1)$$

где λ (λ > 0)- интенсивность событий (аналогично параметру экспоненциального распределения), а k — показатель нестационарности ($k > 0$). При $k = 1$ распределение Вейбулла вырождается в экспоненциальное распределение, а в остальных случаях описывает поток независимых событий с нестационарной интенсивностью. При $k > 1$ моделируется поток событий с растущей со временем интенсивностью, а при $k < 1$ — со снижающейся. Область определения функции распределения плотности вероятностей: неотрицательные действительные числа.

Таким образом, распределение Вейбулла — обобщение экспоненциального распределения на случай нестационарной интенсивности событий. Используется в теории надёжности, моделировании процессов в технике, в прогнозировании погоды, в описании процесса измельчения и т.д.

Матожидание и дисперсия распределения Вейбулла:

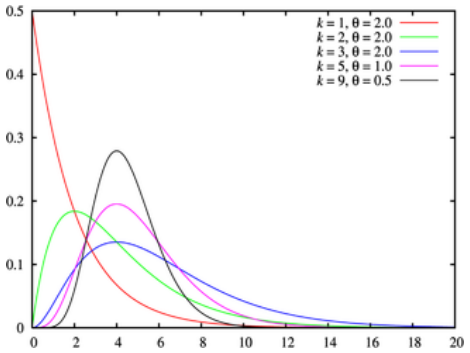
$$E\{W_{k,\lambda}\} = \lambda\Gamma\left(1 + \frac{1}{k}\right) \quad (2.7.2)$$

$$D\{W_{k,\lambda}\} = \lambda^2\left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma\left(1 + \frac{1}{k}\right)^2\right) \quad (2.7.3)$$

где $\Gamma(x)$ — гамма-функция Эйлера.

к схеме
наверх

Гамма-распределение (распределение Эрланга)



(взято отсюда)

Гамма-распределение является обобщением рапспределения хи-квадрат и, соответственно, экспоненциального распределения. Суммы квадратов нормально распределённых величин, а также суммы величин распределённых по хи-квадрат и по

экспоненциальному распределению будут иметь гамма-распределение.

Гамма-распределение является распределением Пирсона III рода. Область определения гамма-распределения — натуральные неотрицательные числа.

Гамма-распределение определяется двумя неотрицательными параметрами k — число степеней свободы (при целом значении степеней свободы, гамма-распределение называется распределением Эрланга) и коэффициент масштаба θ.

Гамма-распределение является бесконечно делимым: если величины X и Y имеют распределения $G_{kx,\theta}$ и $G_{ky,\theta}$ соответсвенно, то величина $X + Y$ будет иметь распределение $G_{kx+ky,\theta}$

$$G_{k,\theta}(x) = x^{k-1}\frac{e^{-\frac{x}{\theta}}}{\Gamma(k)\theta^k} \quad (2.8.1)$$

где $\Gamma(x)$ — гамма-функция Эйлера.

Матожидание и дисперсия:

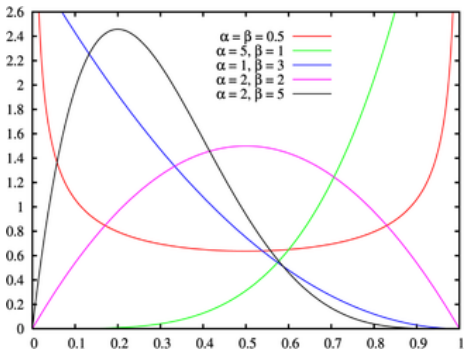
$$E\{G_{k,\theta}\} = k\theta \quad (2.8.2)$$

$$D\{G_{k,\theta}\} = k\theta^2 \quad (2.8.3)$$

Гамма распределение широко применяется для моделирования сложных потоков событий, сумм временных интервалов между событиями, в экономике, теории массового обслуживания, в логистике, описывает продолжительность жизни в медицине. Является своеобразным аналогом дискретного отрицательного биномиального распределения.

к схеме
наверх

Бета-распределение



(взято отсюда)

Бета-распределение описывает долю суммы двух слагаемых, приходящуюся на каждое из них, если слагаемые являются случайными величинами, имеющими гамма-распределение. То есть, если величины X_1 и X_2 имеют гамма-распределение, величины $\frac{X_1}{X_1 + X_2}$ и $\frac{X_2}{X_1 + X_2}$ будут иметь бета-распределение.

Очевидно, что область определения бета-распределения $[0, 1]$. Бета-распределение является распределение Пирсона I типа.

$$B_{\alpha,\beta} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad (2.9.1)$$

где параметры α и β — положительные натуральные числа, $B(x, y)$ — бета-функция Эйлера.

Матожидание и дисперсия:

$$E\{B_{\alpha,\beta}\} = \frac{\alpha}{\alpha+\beta} \quad (2.9.2)$$

$$D\{B_{\alpha,\beta}\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (2.9.3)$$

к схеме
наверх

Вместо заключения

Мы рассмотрели 15 распределений вероятности, которые, на мой взгляд, охватывают большинство наиболее популярных приложений статистики.

Напоследок, небольшое домашнее задание: для оценки надёжности биржевых торговых систем используется такой показатель как профит-фактор. Профит-фактор рассчитывается как отношение суммарного дохода к суммарному убытку. Очевидно, что для системы, приносящей доход, профит-фактор больше единицы, и чем его значение выше, тем система надёжнее.

Вопрос: какое распределение имеет значение профит-фактора?

Свои размышления по этому поводу я изложу в следующей статье.

P.S. Если Вы захотите сослатья на нумерованные формулы из этой статьи, то можете использовать такую сслыку: *ссылка_на_статью*#x_y_z, где (x.y.z)- номер формулы, на которую Вы ссылаетесь.

Теги: теория вероятностей, статистика, распределение Бернулли, биноминальное распределение, геометрическое рапределение, гипергеометрическое распределение, отрицательное биноминальное, распределение Паскаля, распределение Пуассона, нормальное распределение, распределение Гаусса, z-распределение, распределение Стьюдента, t-распределение, распределение хи квадрат, распределение Фишера, f-распределение, распределение Коши, экспоненциальное распределение, показательное распределение, распределение Лапласса, двойное показательное, двойное экспоненциальное, распределение Вейбулла, гамма распределение, распределение Эрланга, бета распределение, распределение Пирсона, случайная величина, плотность вероятности, статистический тест, тест Стьюдента, t-тест, t-критерий, тест Фишера, f-тест, f-критерий, z-тест, z-критерий Фишера, критерий согласия, критерий Пирсона, критерий хи квадрат, статитический анализ, моделирование, профит фактор

Хабы: Занимательные задачки, Анализ и проектирование систем, Алгоритмы, Математика

Комментарии 29

GeMir 30 сентября 2016 в 11:03 +1

Неплохо бы статью перед публикацией показать корректору, который почистит текст от остатков «(3/10) x (2/9)» и «Г(x)» (раз уж в большинстве случаев используется красивый TeX). Нумерация формул кажется лишней, а сама статья скорее похожа на цитату из скрипта лекции по вероятностному исчислению. Справочников по теме предостаточно и как студенты, так и «зрелые специалисты», думается мне, первым делом обратятся к ним, а не к поиску по Хабру.

JamaGava 30 сентября 2016 в 13:45 0

Спасибо за комментарий. По поводу формул — согласен, буду дорабатывать.
А что касается тематики и нумерации формул: эта статья «база» для дальнейших более IT-шных статей по анализу данных (чтобы ставить ссылки на конкретные формулы, собранные вместе).

H_0 **kxx** 30 сентября 2016 в 14:44 +1

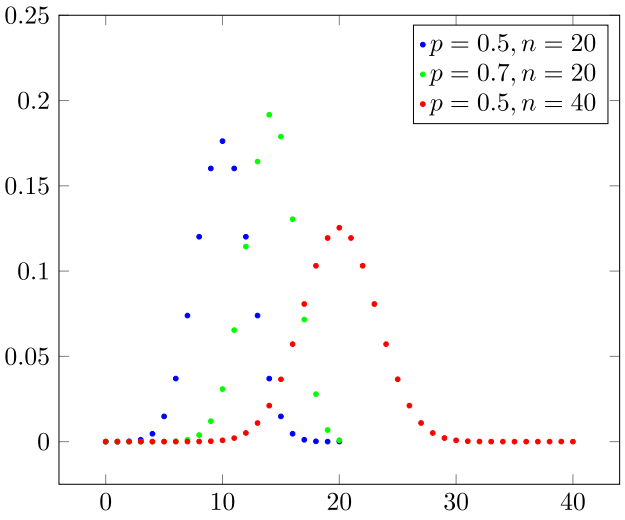
И есть небольшая неточность: биномиальный коэффициент записывается обычно либо как C_n^k, либо как $\binom{n}{k}$, но не как C_k^n.

JamaGava 30 сентября 2016 в 23:30 0

Это действительно ляп с моей стороны… Спасибо.
Уже исправлено.

parpalak 1 октября 2016 в 12:37 +1

Картинки, кстати, тоже можно сделать красивыми:



Вот исходник:

```
\begin{tikzpicture}[mark=*,mark size=1,only marks]
\begin{axis}[
    yticklabel style={/pgf/number format/fixed},
    ymax=0.25,
    width=10cm,
    declare function={binom(\k, \n, \p)=(\p^(\k))*((1-\p)^(\n-\k))*factorial(\n)/(factorial(\n-\k)*factorial(\k));}
]
\addlegendimage{blue}
\addlegendimage{green}
\addlegendimage{red}
\foreach \n/\p/\c in {20/0.5/blue, 20/0.7/green, 40/0.5/red} {
    \foreach \k in {0,...,\n} {
        \edef\temp{\noexpand
            \addplot[\c] coordinates {(\k,{binom(\k, \n, \p)})};
        }\temp
    }
    \edef\temp{\noexpand
        \addlegendentry{$p=\p, n=\n$}
    }\temp
}
\end{axis}
\end{tikzpicture}
```

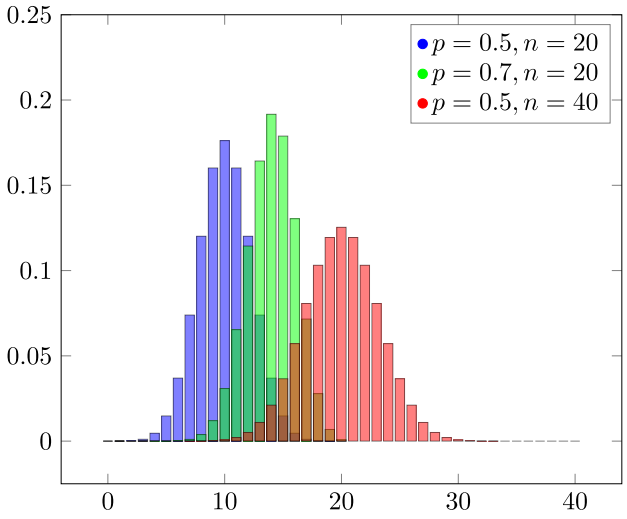
tomzarubin 2 октября 2016 в 14:12 +2

Непонятно почему вы считаете свой график лучше представленных выше. Точки тут неуместны— чтобы понять о чём график(а потом и форму распределения), надо напрячься.

parpalak 2 октября 2016 в 14:38 +2

Согласен, конечно. Мы тут обсуждали оформление, а не содержание.

В тексте была откуда-то скопированная png-картинка. Моя svg-картинка сделана в латехе с пакетом tikz. Способ не без проблем: человеку без опыта тяжело сразу готовить такие картинки. Зато исходник можно править на лету, не перерисовывая картинку. Я вот за 5 минут поменял точки на столбцы, и диаграмма стала понятнее:



tomzarubin 4 октября 2016 в 17:04 +1

Постойте, так про оформление и речь. Пример выше, конечно, чище и понятней.

JamaGava 4 октября 2016 в 20:05 0

Такие диаграммы красивее. Я взял материал со страниц википедии (в тексте приведены ссылки). Если для кого-нибудь действительно совсем несложно переделать графики — Вы можете улучшить Вики.
Сам бы занялся, но не владею технологией построения настолько красивых диаграмм :)
(Хотя, уже захотелось освоить)

parpalak 4 октября 2016 в 21:11 +1

Переделать — не сложно, но и не просто :)

Хотите освоить — посмотрите введение к официальной документации tikz. В нем последовательным усложнением строятся полноценные примеры графиков и диагамм. Это хорошая отправная точка.

JamaGava 4 октября 2016 в 23:13 0

Спасибо

tomzarubin 6 октября 2016 в 09:10 +1

Я не в претензии)
Кстати, в R есть прекрасные библиотеки. Тот же «классический уже» ggplot2 или seaborn + несколько библиотек интерактивных графиков.
Ну и R, по ощущениям, очень прекрасен для статистики и EDA. Самое главное— побороть непонимание синтаксиса самого R.

temas 30 сентября 2016 в 12:47 0

Интересно, но ожидал больше примеров. (В дискретных распределениях примеры были)

JamaGava 30 сентября 2016 в 13:43 0

Большой пример скоро будет в отдельной статье

Lelushak 30 сентября 2016 в 18:48 –1

Зачем захламлять теги отдельным названием каждого распределения? У них другое предназначение.

ratatosk 30 сентября 2016 в 19:03 +1

Вот очень хорошая вводная статья про виды распределений: Common Probability Distributions: The Data Scientist's Crib Sheet.

JamaGava 1 октября 2016 в 14:52 0

Название действительно провокационное, для привлечения внимания. В самой статье я пишу, что предложенные распределения используются в «наиболее часто встречающихся задачах».

Что касается «дз» это тизер готовящейся к публикации работы. Идея такова: мы не можем знать всё обо всех стратегиях, но может поступить по аналогии с тем, как строятся другие стат.тесты, а именно, построить распределение профитфактора для системы, торгующей случайно. Тогда значение профитфактора реальной системы должно быть таковым, чтобы случайное достижения такого значения являлось маловероятным.

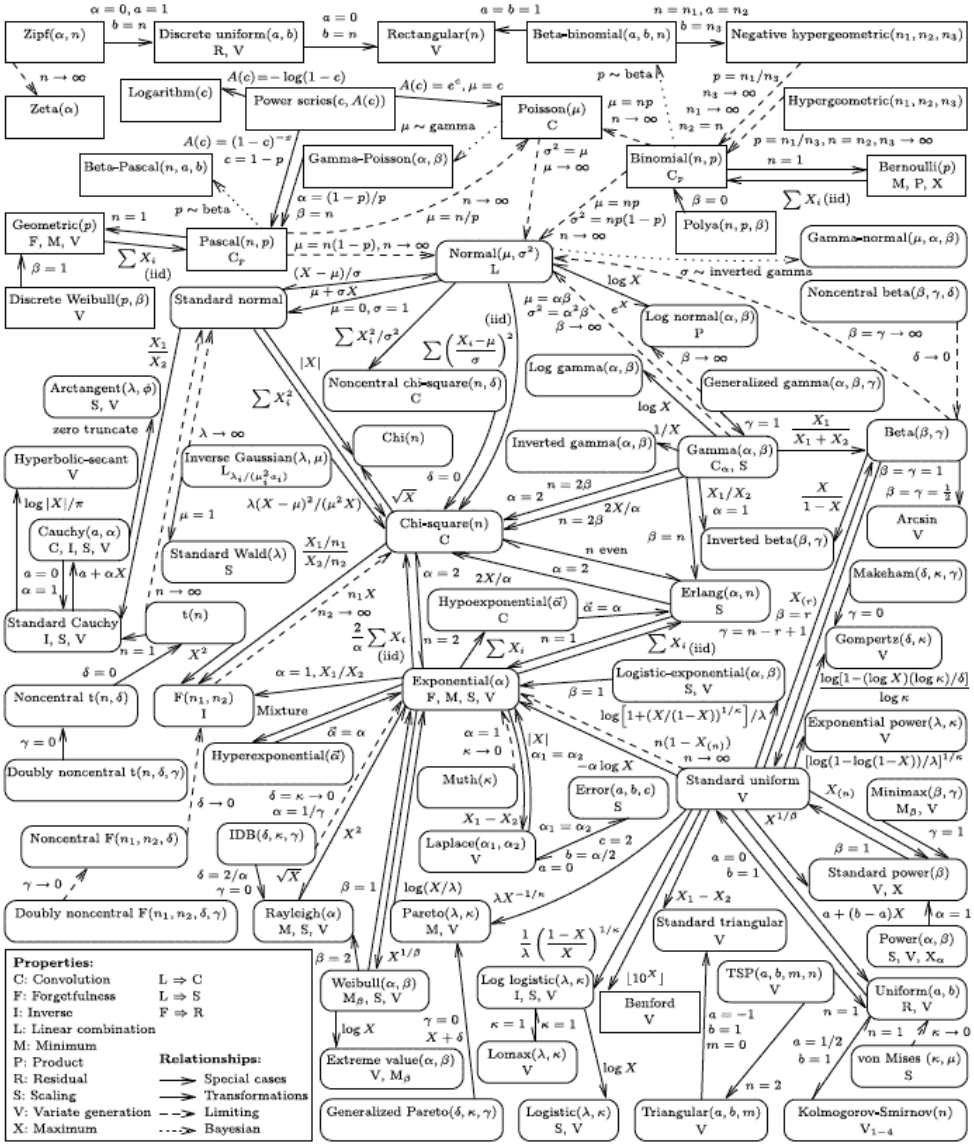
smxfem 1 октября 2016 в 14:59 +1

Название статьи провокационное. Получается, что сингулярные распределения — бесполезная альтернатива и создание этих распределений не продиктовано непригодностью описанных тут распределений к некоторым задачам — «случаям жизни»?

Домашнее задание смахивает на задачку с подвохом. Требуется смоделировать множество биржевых систем (тут требуются знания предметной области на уровне бога) и напрямую получить распределения аналитически или откуда-то взять статистику по биржевым системам, подобрать для нескольких дающих надежду распределений оптимальные параметры (оптимизационная задача) и выбрать среди полученных конкретных распределений наиболее точные?

ShashkovS 3 октября 2016 в 00:37 +4

Я просто оставлю это здесь :)



Ну, и с подробностями: http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

JamaGava 3 октября 2016 в 01:19 +1

Вот это действительно: «Вау!» :)
Спасибо!

JamaGava 3 октября 2016 в 01:38 0

У меня вообще подозрение, что почти все непрерывные распределения должны укладываться в форму $f(x) = const \times (P(x))^a\, e^{-b(Q(x))^c}$, где $P(x)$ и $Q(x)$ — полиномы. Либо являться подстановкой функции (например, нецелой степени либо логарифма) от x вместо аргумента в эту формулу.

Кто-нибудь встречал подобные обобщение?

smxfem 8 октября 2016 в 10:33 +2

Есть обобщенное нормальное и обобщенное гиперболическое распределения, которые содержат многие распределения. Но обобщенный вид затрудняет оценку параметров (не критично), а полу-гуманитарные критерии применимости распределений к реальным задачам слабо распространены на такие виды. Попробуйте убедить, предположим, медиков, что какое-то обобщенное распределение лучше подходит для описания некоторой случайной величины, когда 100 лет уже её описывают в обширных исследованиях, которые считаются фундаментальными, более простым частным распределением. Хотя, если, после оценки, вклад всех параметров в распределение будет существенным, то никто не отвертится и придется принять обобщение. Знакомый психолог публиковала в зарубежном журнале статью, где математических претензий не было, а отфутболивали именно из-за отсутствия обоснования применимости использованных распределений, пока не были найдены необходимые подтверждающие ссылки.

H_0 **kxx** 4 октября 2016 в 22:23 +1

Граф впечатляет. Даже не подозревал о некоторых распределениях.

Leo5700 5 октября 2016 в 01:27 0

Ближайшие полчаса путешествую по графу)

JamaGava 5 октября 2016 в 09:35 0

Студентам буду на зачёт задание давать: воспроизвести граф по памяти :)

nickolaym 6 октября 2016 в 02:00 +1

Для плотности f(x) и y=g(x), — формула плотности fy(y) справедлива, только если $g^{-1}(y)$ однозначна. То есть, g(x) биективна. А дифференцируемость g(x) не только недостаточное, но и необязательное условие: легко придумать кучу примеров как с разрывными отображениями
— например, пила вдоль линии y = int(x)-frac(x)
так и дифференцируемые отображения, для которых обратные не дифференцируемы
— например, с седловой точкой, $y = x^3$ (при том, что оно биективно)

JamaGava 6 октября 2016 в 11:40 0

Согласен с Вашим, замечанием: дифференцируемой должна быть обратная к g(x) (опечатка).

Andreylvanoff 17 октября 2016 в 21:16 +1

Вот интересное оформление справки по теме.

Figure 6A.15: Distributional Choices