Прикладная статистика

Слайды к лекции 7

С. А. Спирин

28 февраля 2023

1. Точечное оценивание

Общая теория и терминология

Математическая статистика

Проверка гипотез

Оценка параметров

– точечное оценивание
– интервальное оценивание

Во всех задачах исходным материалом является **выборка**, которая в математической статистике рассматривается как набор независимых одинаково распределённых случайных величин, а на практике представляет собой набор чисел (выборка из генеральной совокупности или результаты экспериментов).

Вместо «выборка» часто говорят «наблюдения».
Выборка может быть как-то структурирована, например состоять из пар чисел, или делиться на две подвыборки и т. д.

Точечное оценивание

Предполагается, что выборка происходит из распределения, зависящего от параметра θ.

Например:

– из экспоненциального распределения с неизвестным средним θ;
– из нормального распределения с неизвестным средним θ;
– из нормального распределения с известным средним (например, 0), но неизвестной дисперсией θ;
– …

Результатом является **оценка** θ' — функция от наблюдений, значение которой принимается за предполагаемое значение параметра θ.

В мат. статистике оценка рассматривается как функция от случайных величин, поэтому сама является случайной величиной с распределением, зависящим от θ.

Точечное оценивание

(*терминология*)

•Оценка θ' называется **состоятельной**, если θ' сходится к θ с ростом числа наблюдений.

•Оценка θ' называется **несмещённой**, если математическое ожидание θ' равно θ.

•**Эффективностью** оценки θ' (неформально) называется то, насколько хорошо θ' приближает θ. Мерой эффективности обычно служит среднее квадратичное отклонение θ' от θ: $E(\theta' - \theta)^2$.
Для несмещённой оценки эта мера равна дисперсии θ'.

*В мат. статистике оценка называется **эффективной**, если она несмещённая и её дисперсия **при любом θ** наименьшая среди всех оценок.*

Оценка среднего значения

Эффективной (в любом смысле) оценкой среднего значения генеральной совокупности почти всегда служит среднее по выборке.

В математической статистике среднему значению генеральной совокупности соответствует матожидание случайной величины

Бывают экзотические случаи, когда среднего по генеральной совокупности просто нет, например, если измерять расстояние от центра мишени при равномерном распределении угла стрельбы в интервале (0°, 90°)

Но и, например, для равномерного распределения на интервале с известной левой, но неизвестной правой границей оценивать матожидание как среднее по выборке не всегда хорошо. Действительно, разумно предполагать, что оценка правой границы должна быть такой, чтобы оценка среднего оказалась посредине между левой границей и оценкой правой границы. Но вполне может получиться, что часть наблюдений вылезет за оценённую так правую границу.

Оценка матожидания

Эффективной оценкой математического ожидания большинства распределений является среднее по выборке.

Для выборки из нормального распределения среднее по выборке тоже распределено нормально с тем же матожиданием и с дисперсией $D(\theta') = D/n$, где D — дисперсия исходного распределения, а n — число наблюдений.

Для выборки из другого распределения (с конечной дисперсией) среднее по выборке для больших n распределено почти нормально (распределение среднего по выборке стремится к нормальному с ростом n)

Квадратный корень из дисперсии среднего по выборке принято называть **стандартной ошибкой.**

Стандартное отклонение и стандартная ошибка

Стандартное отклонение = корень квадратный из дисперсии исходного распределения

(то есть среднее квадратичное отклонение от среднего значения по генеральной совокупности).

Ст. отклонение не зависит от выборки, это свойство генеральной совокупности.

Стандартная ошибка = корень квадратный из дисперсии среднего по выборке

Ст. ошибка равна ст. отклонению, деленному на квадратный корень из размера выборки

Несмещённая оценка дисперсии

При **известном** матожидании μ несмещённой (и эффективной) оценкой дисперсии является

$D' = \sum(x_i - \mu)^2/n$

что неудивительно, ведь дисперсия — это среднее значение величины $Y = (X - \mu)^2$

Несмещённая оценка дисперсии

При **неизвестном** матожидании сначала оцениваем его средним по выборке: $\bar{x} = \sum x_i / n$,
после этого оцениваем дисперсию как $s^2 = \sum (x_i - \bar{x})^2/(n - 1)$

Если вместо $(n - 1)$ поставить n, получится смещённая оценка: её матожидание будет несколько меньше реального значения дисперсии σ^2. Это происходит потому, что x_i всегда в среднем ближе к $\bar{x}$, чем к μ.

Стандартное отклонение и стандартная ошибка оцениваются из этих оценок дисперсии как квадратный корень из s^2 и квадратный корень из s^2/n, соответственно.

Эти оценки — смещённые, но достаточно эффективные, ими все пользуются.

Обычно именно эти оценки называются стандартным отклонением и стандартной ошибкой выборки.

Оценка частоты

Пусть имеем бернуллиевские испытания с неизвестной вероятностью успеха θ.
Как по результатам испытаний (число успехов/число неуспехов) оценить θ?

Оценка частоты

Пусть имеем бернуллиевские испытания с неизвестной вероятностью успеха θ.
Как по результатам испытаний (число успехов/число неуспехов) оценить θ?

Введём случайную величину, принимающую значение 1 в случае успеха и 0 в случае неуспеха.
Тогда θ — это математическое ожидание этой случайной величины.
Поэтому самая разумная оценка θ — среднее значение по проведённым испытаниям, то есть **отношение числа успехов к числу испытаний**.
(*Что, конечно, совсем не удивительно :)*)

Оценка по максимальному правдоподобию

Что делать, если нужно оценить какую-нибудь не вполне стандартную величину?

Есть универсальный подход: принцип максимального правдоподобия.

Этот принцип гласит: из конкурирующих гипотез выбираем ту, что дала бы максимальную вероятность того, что мы наблюдаем.

Оценка по максимальному правдоподобию: дискретный случай

Пусть сначала наши наблюдения $X_1, X_2, \ldots, X_n$ происходят из дискретного распределения.

Это значит, что каждое X может принимать лишь значения из какого-то дискретного множества, и каждое такое значение имеет ненулевую вероятность $P_\theta(X)$

Вероятность каждого значения зависит от θ, которого мы не знаем и хотим оценить.

Оценка по максимальному правдоподобию: дискретный случай

Посчитаем (зависящую от θ) вероятность наших наблюдений:

$P(X_1, X_2, \ldots, X_n \mid \theta) = \prod_i P_\theta(X_i)$

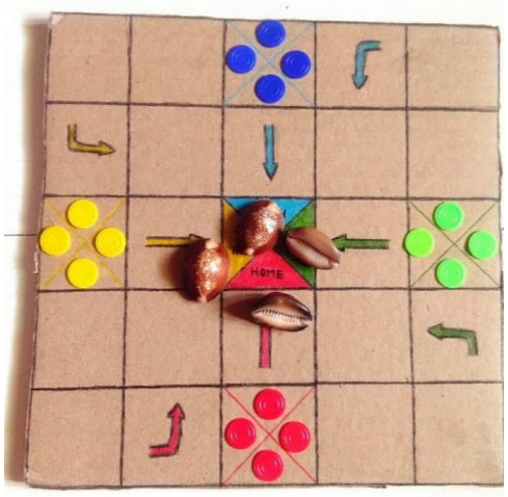
Получилась функция от θ.

Если мы можем найти такое $\theta = \theta'$, при котором данная вероятность достигает максимума, то это значение и будет максимально правдоподобной оценкой θ.

Правдоподобием (likelihood) называется вероятность сделанных наблюдений.

Пример 1

У нас есть раковина каури, мы хотим узнать, с какой вероятностью она падает выпуклостью вверх.



Индийское лудо, или тхайайам

Неизвестный параметр θ : вероятность падения выпуклостью вверх.

Бросим раковину много раз и посчитаем число падений выпуклостью вверх.

Пусть из n бросков k раз раковина упала выпуклостью вверх.

Вероятность такого исхода зависит от параметра θ, она равна
$P(\theta) = C_n^k \, \theta^k \, (1-\theta)^{n-k}$

Найдём такое θ , при котором $P(\theta)$ достигает максимума.

Для этого решим уравнение $P'(\theta) = 0$.

$P'(\theta)$ с точностью до не зависящего от θ множителя равна
$(k(1-\theta) - (n-k)\,\theta)\,\theta^{k-1}\,(1-\theta)^{n-k-1}$, откуда получаем уравнение
$k - n\theta = 0$
Его решение: $\theta = k/n$

Пример 2

В коробке лежат два внешне одинаковых кубика с цифрами 1, …, 6 на гранях.
Известно, что для одного кубика вероятность упасть любой гранью вверх равна 1/6, а для другого вероятность упасть вверх гранью 6 равна 1/2, а прочие — 1/10.

Вы взяли наугад один кубик, бросили его три раза и все три раза выпали шестёрки.
Какой кубик вы взяли?

Пример 2

В коробке лежат два внешне одинаковых кубика с цифрами 1, …, 6 на гранях.
Известно, что для одного кубика вероятность упасть любой гранью вверх равна 1/6, а для другого вероятность упасть вверх гранью 6 равна 1/2, а прочие — 1/10.

Вы взяли наугад один кубик, бросили его три раза и все три раза выпали шестёрки.
Какой кубик вы взяли?

Разумеется, точный ответ «неизвестно».
Но вероятнее, что это второй кубик.

Пример 2

В коробке лежат два внешне одинаковых кубика с цифрами 1, …, 6 на гранях.
Известно, что для одного кубика вероятность упасть любой гранью вверх равна 1/6, а для другого вероятность упасть вверх гранью 6 равна 1/2, а прочие — 1/10.

Вы взяли наугад один кубик, бросили его три раза и все три раза выпали шестёрки.
Какой кубик вы взяли?

Разумеется, точный ответ «неизвестно».
Но вероятнее, что это второй кубик.

В данном случае можно даже посчитать вероятность, это делается по формуле Байеса:
$P(A|B) = P(B|A)\cdot P(A)/P(B)$ A — взят второй кубик, B — три раза выпала шестёрка
$P(A) = 1/2$
$P(B|A) = (1/2)^3 = 1/8$
$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 1/8 \cdot 1/2 + (1/6)^3 \cdot 1/2 \approx 0{,}0648$
Получаем $P(A|B) \approx 0{,}0625/0{,}0648 \approx 0{,}964$
Мы смогли посчитать вероятность только потому, что заранее знали распределение параметра

Пример 2 (варианты)

При тех же условиях шестёрка выпала два раза из трёх.

Посчитаем вероятность такого исхода при условии «взят первый кубик» и её же при условии «взят второй кубик».
Первая вероятность равна $3 \cdot (1/6)^2 \cdot (5/6) \approx 0{,}07$, а вторая $3 \cdot (1/2)^3 = 0{,}375$
Поэтому по принципу ML мы должны выбрать вариант «второй кубик»

Если выпала одна шестёрка из трёх, то первая вероятность равна $3 \cdot (1/6) \cdot (5/6)^2 \approx 0{,}347$, а вторая снова 0,375. Значит, и при этом условии по принципу ML мы должны выбрать второй кубик.

Пример 3

Вам дали кубик и попросили оценить вероятность выпадения шестёрки.
Предположим, вам разрешили бросить кубик только три раза, и шестёрка ни разу не выпала.
В этом случае по максимальному правдоподобию следовало бы оценить эту вероятность как нулевую (что очевидно нелепо).

Выход — в применении байесовой статистики и выборе подходящего «приора» (априорного распределения параметра)

Другой выход — в использовании не точечной оценки, а доверительного интервала для параметра.

Оценка по максимальному правдоподобию: пример из биоинформатики

Выравнивание:

```
Seq. 1: TTTATATCGTGTGTACATATAAATATGTACACACGGCTTTTAGGTAGAATAT
Seq. 2: TTCATATTATGAGTACGTTTAAATGTGTACACACAGCTTTTAAGTAGAGTCT
```

Мы видим результат эволюции (замены в гомологичных последовательностях) и хотим оценить число мутаций. Неизвестный параметр θ равен числу мутаций.

Число мутаций $\neq$ числу замен (мутации могли происходить в одном месте не раз)

Стандартный подход: оценить наиболее правдоподобное число мутаций.

Этапы:

- Выбор вероятностной модели эволюции
- Подсчёт логарифма правдоподобия $\log L(n)$ (логарифма вероятности наблюдать данную пару последовательностей) для каждого числа мутаций на пути от первой ко второй
- Выбор числа мутаций, максимизирующего $\log L(n)$

Оценка по максимальному правдоподобию: непрерывный случай

Пусть теперь наблюдения $X_1, X_2, \ldots, X_n$ происходят из непрерывного распределения.
Это значит, что каждое X может принимать любое действительное значение (или любое значение из какого-то интервала: $(0,\infty)$ или (a, b)),
вероятность каждого конкретного значения равно нулю и для любого числа определена плотность вероятности в окрестности этого числа:
$p(x) = \lim_{\varepsilon \to 0} P(x - \varepsilon < X < x + \varepsilon)/2\varepsilon$

Тогда нужно считать плотность вероятности в окрестности наших наблюдений:

$$p_\theta(X_1, X_2, \ldots, X_n) = \prod_i p_\theta(X_i)$$

(Такую плотность тоже нередко называют правдоподобием, хотя строго говоря, это плотность правдоподобия)

Оценкой максимального правдоподобия в непрерывном случае называется значение θ, максимизирующее плотность правдоподобия.

Часто от правдоподобия (или его плотности) берут логарифм: положение максимума не меняется, а вычисления удобнее.

Оценка по максимальному правдоподобию: упражнения

- Докажите, что максимально правдоподобная оценка неизвестного среднего нормального распределения совпадает со средним по выборке
- То же, для экспоненциального распределения
- Докажите, что максимально правдоподобная оценка неизвестной дисперсии нормального распределения (при известном среднем) совпадает со средним квадратичным отклонением от среднего по выборке
- Распределение равномерное на отрезке [0, θ], где θ неизвестно. Чему равна максимально правдоподобная оценка θ?

Равномерное распределение,

оценка границ при известной длине интервала

Известно, что выборка $X_1, X_2, \ldots, X_n$ происходит из равномерного распределения на отрезке $[\theta, \theta + 1]$, где θ неизвестно. Чему равна максимально правдоподобная оценка θ?

Ответ: её в общем случае нет :(
Любое θ такое, что $\theta < \min(X_1, X_2, \ldots, X_n)$ и $\max(X_1, X_2, \ldots, X_n) < \theta + 1$, даст одно и то же значение плотности правдоподобия (равное 1).

2. Байесова статистика

Оценки по максимуму апостериорной вероятности

Байесовские оценки

Предположим, у нас есть какое-то априорное знание о параметре θ, выраженное в виде вероятностей разных его значений.
Тогда стоит оценивать θ исходя из формулы Байеса:

$$P(\theta \mid X_1, X_2, \ldots, X_n) = P(X_1, X_2, \ldots, X_n \mid \theta) \cdot P(\theta) / P(X_1, X_2, \ldots, X_n)$$

Пусть сначала всё дискретно
(включая априорное распределение параметра θ, например, параметр задаётся выбором из конечного числа объектов).

Тогда справа в числителе — вероятность наблюдений при заданном θ, умноженная на **априорную** вероятность данного θ, в знаменателе — **полная** вероятность наблюдений, которую в дискретном случае можно посчитать как сумму $P(X_1, X_2, \ldots, X_n \mid \theta_k) \cdot P(\theta_k)$ по всем возможным значениям θ_k параметра θ.

Слева получается **апостериорная** (то есть посчитанная на основе наблюдений) вероятность каждого значения θ.

Байесовские оценки

Апостериорная вероятность:
$P(\theta \mid X_1, X_2, \ldots, X_n) = P(X_1, X_2, \ldots, X_n \mid \theta) \cdot P(\theta) / P(X_1, X_2, \ldots, X_n)$

В дискретном случае единственной байесовской оценкой является MAP (maximum of a posterior probability) — в качестве оценки берётся значение θ, максимизирующее апостериорную вероятность

Если все априорные вероятности $P(\theta)$ равны между собой, то MAP-оценка превращается в ML-оценку
(ML = maximum likelihood = максимальное правдоподобие)

Байесовская оценка в дискретном случае: пример

В коробке лежит множество внешне неразличимых кубиков. 9/10 из них «нормальные» (вероятность каждой грани 1/6), а десятая часть падает шестёркой вверх в половине случаев. Мы взяли один кубик и бросили его три раза, два раза выпала шестёрка. Оцените, какой кубик мы взяли, по максимуму апостериорной вероятности.

Решение.
Знаменатель в формуле Байеса от θ не зависит, поэтому нам надо сравнить выражения $P(X_1, X_2, X_3 \mid \theta) \cdot P(\theta)$ для двух значений θ.
Мы помним, что для «нормального» кубика вероятность наблюдения $\approx 0{,}07$, а для «перекошенного» = 0,375 (поэтому по ML мы оценивали кубик как перекошенный).
Теперь мы должны умножить первую вероятность на 9/10, а вторую на 1/10 — получаем, что по MAP при таком приоре мы оцениваем кубик как нормальный.

Если бы все три раза выпала шестёрка, мы бы и по MAP оценили кубик как перекошенный (почему)?

Байесовские оценки: непрерывный случай

В формуле для апостериорной вероятности:

$P(\theta \mid X_1, X_2, \ldots, X_n) = P(X_1, X_2, \ldots, X_n \mid \theta) \cdot P(\theta) / P(X_1, X_2, \ldots, X_n)$

надо заменить вероятности на плотности вероятности везде, где распределение непрерывно.

Потом можно применять MAP, а если непрерывно распределение θ, то можно и другие приёмы:

- posterior mean — в качестве оценки берётся апостериорное матожидание θ
- posterior median — в качестве оценки берётся апостериорная медиана θ

(в дискретном случае медиана и матожидание могут не попасть в множество допустимых значений θ)

Чаще всего применяют MAP (его обычно легче всего считать)

Байесовские оценки: непрерывный случай

В формуле для апостериорной вероятности:

$P(\theta \mid X_1, X_2, \ldots, X_n) = P(X_1, X_2, \ldots, X_n \mid \theta) \cdot P(\theta) / P(X_1, X_2, \ldots, X_n)$

надо заменить вероятности на плотности вероятности везде, где распределение непрерывно.

Потом можно применять MAP, а если непрерывно распределение θ, то можно и другие приёмы:

- posterior mean — в качестве оценки берётся апостериорное матожидание θ
- posterior median — в качестве оценки берётся апостериорная медиана θ

Если непрерывно распределение параметра θ , возникает «засада».

Предположим, что нам удобнее оценивать не θ, а θ^2 или другую взаимно однозначную функцию от θ.

Но если $\chi = \theta^2$, то плотность для χ **не равна** плотности для θ в соответствующей точке!
(Она равна $p(\theta)/2\theta$— почему?)

Поэтому максимум у неё может быть в другом месте!

Тем самым MAP-оценка может зависеть от «способа измерения» параметра (оцениваем дисперсию — получаем одно, оцениваем стандартное отклонение — другое)

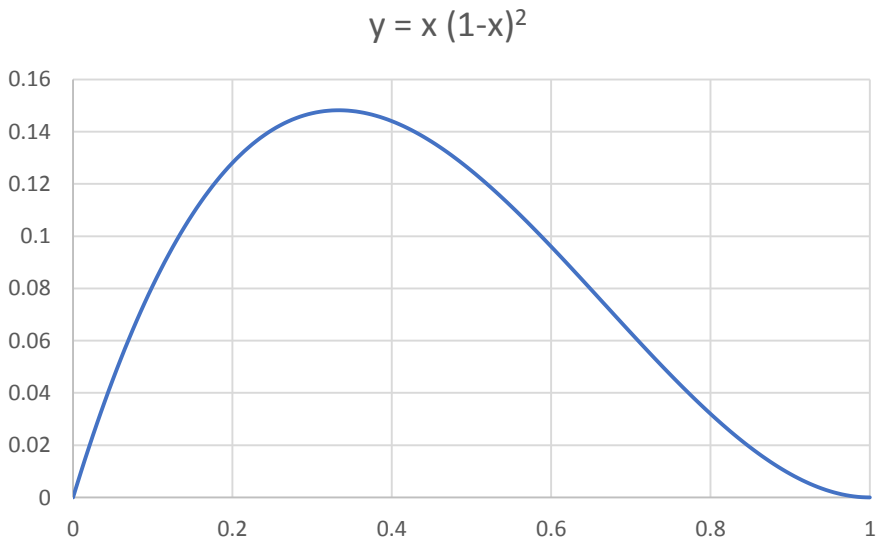
Для матожидания ещё хуже: матожидание квадрата **никогда** не равно квадрату матожидания.

(Правда, с медианой ничего не случится).

Байесовская оценка: пример

Для данной раковины каури хотим узнать, с какой вероятностью она падает выпуклостью вверх.
Мы знаем, что типичное значение для таких раковин — $\theta = 1/3$.

Придумаем подходящее априорное распределение для θ. Из соображений удобства обычно используют бета-распределение с плотностью $p(x) \sim x^a(1-x)^b$ $(0<x<1)$.
Максимум такой плотности находится в точке $a/(a+b)$, поэтому в нашем случае годится $a = 1$, $b = 2$



Байесовская оценка: пример

Для данной раковины каури хотим узнать, с какой вероятностью она падает выпуклостью вверх.
Мы знаем, что типичное значение для таких раковин — $\theta = 1/3$.

Придумаем подходящее априорное распределение для θ. Из соображений удобства обычно используют бета-распределение с плотностью $p(x) \sim x^a(1-x)^b$.
Максимум такой плотности находится в точке $\alpha/(\alpha+\beta)$, поэтому в нашем случае годится $a = 1$, $b = 2$
Для получения MAP оценки нам нужно найти максимум функции

$$P(X_1, X_2, \ldots, X_n \mid \theta) \cdot p(\theta)$$

то есть:

$$C_n^k \theta^k (1-\theta)^{n-k} p(\theta) \sim \theta^k (1-\theta)^{n-k} \theta^a (1-\theta)^b \sim \theta^{k+a} (1-\theta)^{n-k+b}$$

Тем самым вычисления будут те же, что для ML оценки, но не для k успехов в n испытаниях, а для $k + a$ успехов в $n + a + b$ испытаниях.

Байесовская оценка: пример

Для данной раковины каури хотим узнать, с какой вероятностью она падает выпуклостью вверх.
Мы знаем, что типичное значение для таких раковин — $\theta = 1/3$.

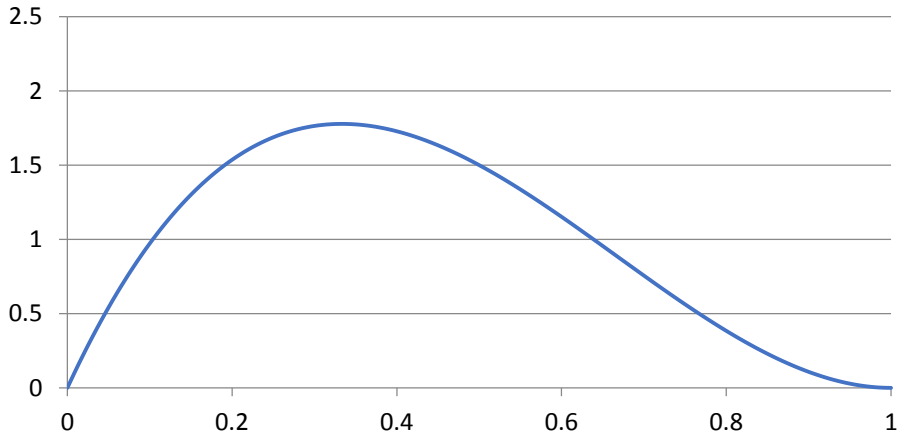
Придумаем подходящее априорное распределение для θ. Из соображений удобства обычно используют бета-распределение с плотностью $p(x) \sim x^a(1–x)^b$.
Максимум такой плотности находится в точке $a/(a+b)$, поэтому в нашем случае годится $a = 1$, $b = 2$
Для получения MAP оценки нам нужно найти максимум функции
$P(X_1, X_2, \ldots, X_n \mid \theta) \cdot p(\theta)$, то есть:

$$C_n^k \theta^k (1–\theta)^{n–k} p(\theta) \sim \theta^k (1–\theta)^{n–k} \theta^a(1–\theta)^b \sim \theta^{k+a} (1–\theta)^{n–k+b}$$

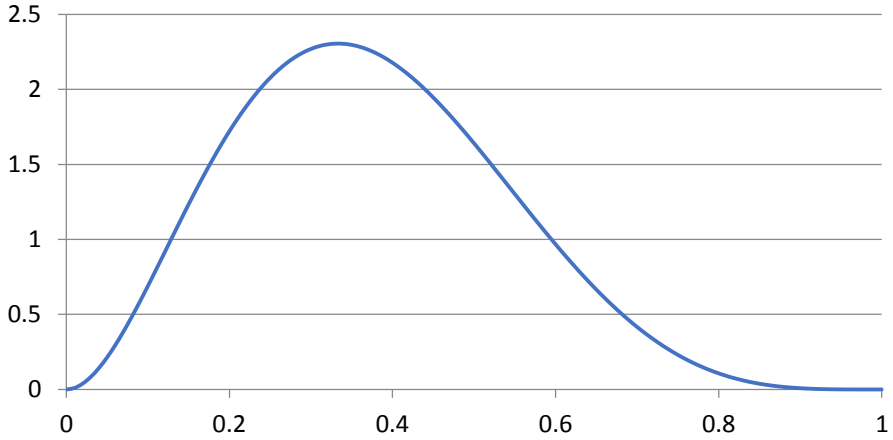
Тем самым вычисления будут те же, что для ML оценки, но не для *k* успехов в *n* испытаниях, а для *k* + *a* успехов в *n* + *a* + *b* испытаниях.
Мы «притворяемся», что сделали не n, а $n + 3$ броска, и раковина упала выпуклостью вверх не в k, а в $k + 1$ случае. Оценка: $\theta = (k+1)/(n+3)$.
Это называется «псевдоотсчёты» (pseudocounts). Априорное знание позволяет нам как бы увеличить число испытаний.

Приоры с одинаковым средним, но разной «уверенностью»

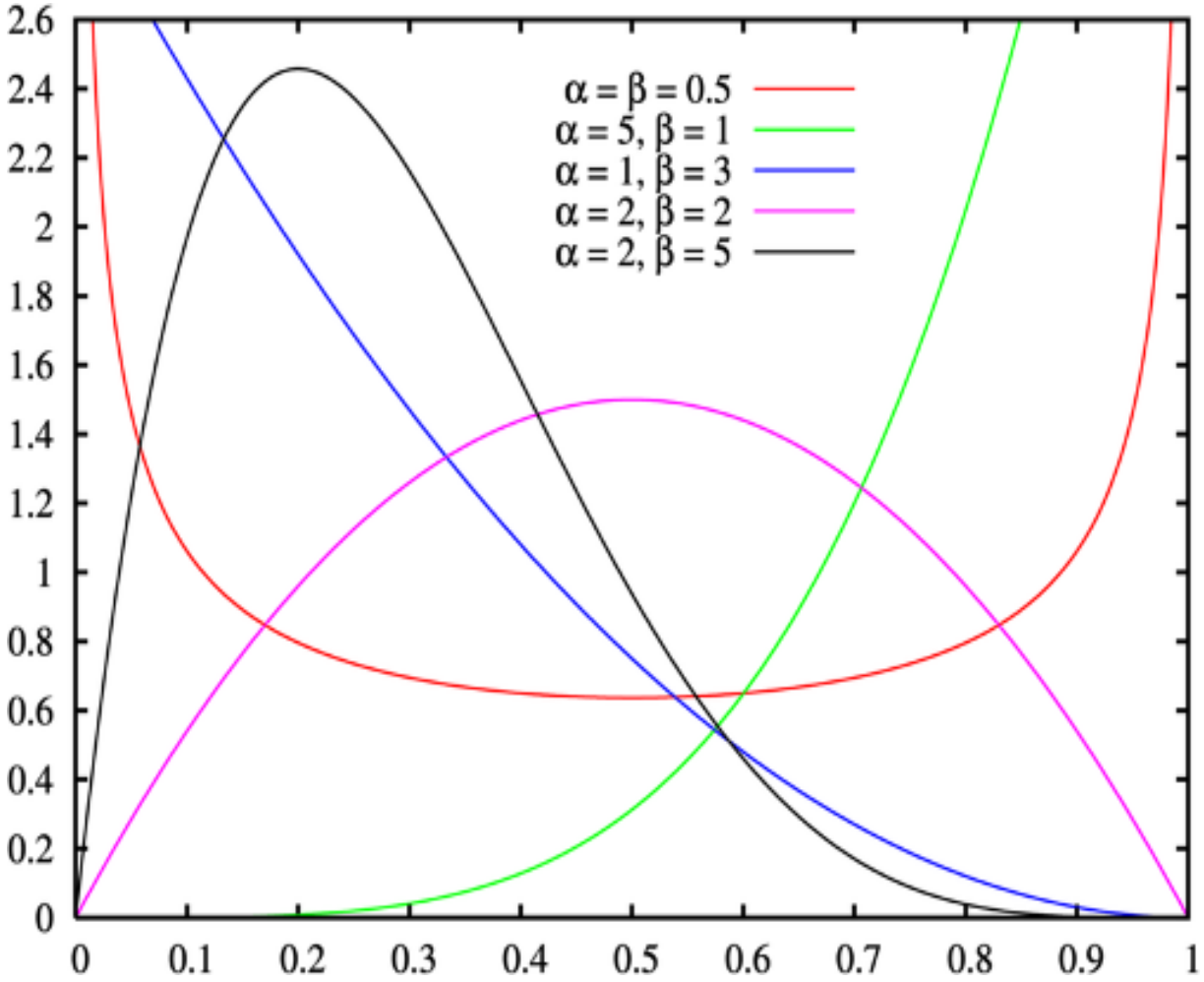
$y = 12x(1-x)^2$



$y = 110x^2(1-x)^4$



Бета-распределение



$p(x) = C(\alpha, \beta)x^{\alpha-1}(1-x)^{\beta-1}$ $\alpha > 0, \beta > 0$