

# Теория вероятностей

Математическая  
статистика

Владимир Анатольевич Судаков

# Вероятность против статистики

---

- Вероятность занимается прогнозированием вероятности будущих событий, а статистика анализирует частоту прошлых событий.
- Вероятность — это теоретическая часть математики, посвященная следствиям определений, а статистика — это прикладная математика, пытающаяся осмыслить наблюдения из реального мира.

# Литература

- Венцель Е.С. Теория вероятностей
- Пугачев В.С. Теория вероятностей и математическая статистика
- Triola, Mario F. Elementary statistics

# Математическая статистика

---

- Разработка методов регистрации, описания и анализа статистических экспериментальных данных, получаемых в результате наблюдения массовых случайных явлений, составляет предмет специальной науки — математической статистики.
- Задачи математической статистики касаются вопросов обработки наблюдений над массовыми случайными явлениями, но в зависимости от характера решаемого практического вопроса и от объема имеющегося экспериментального материала эти задачи могут принимать ту или иную форму.

# Типичные задачи статистики

- Определить закон распределения случайной величины (системы случайных величин)
- Проверить правдоподобие гипотез
- Найти неизвестные параметры распределения

# Простая статистическая совокупность

- Дана случайная величина  $X$
- Совокупность наблюдаемых значений  $X$ -простой статистический ряд (простая статистическая совокупность)
- Статистическая функция распределения  $X$ :

$$F^*(x) = P^*(X < x)$$

# Визуализация распределений

---

Продажи Apple iPhone стремительно растут, не так ли?

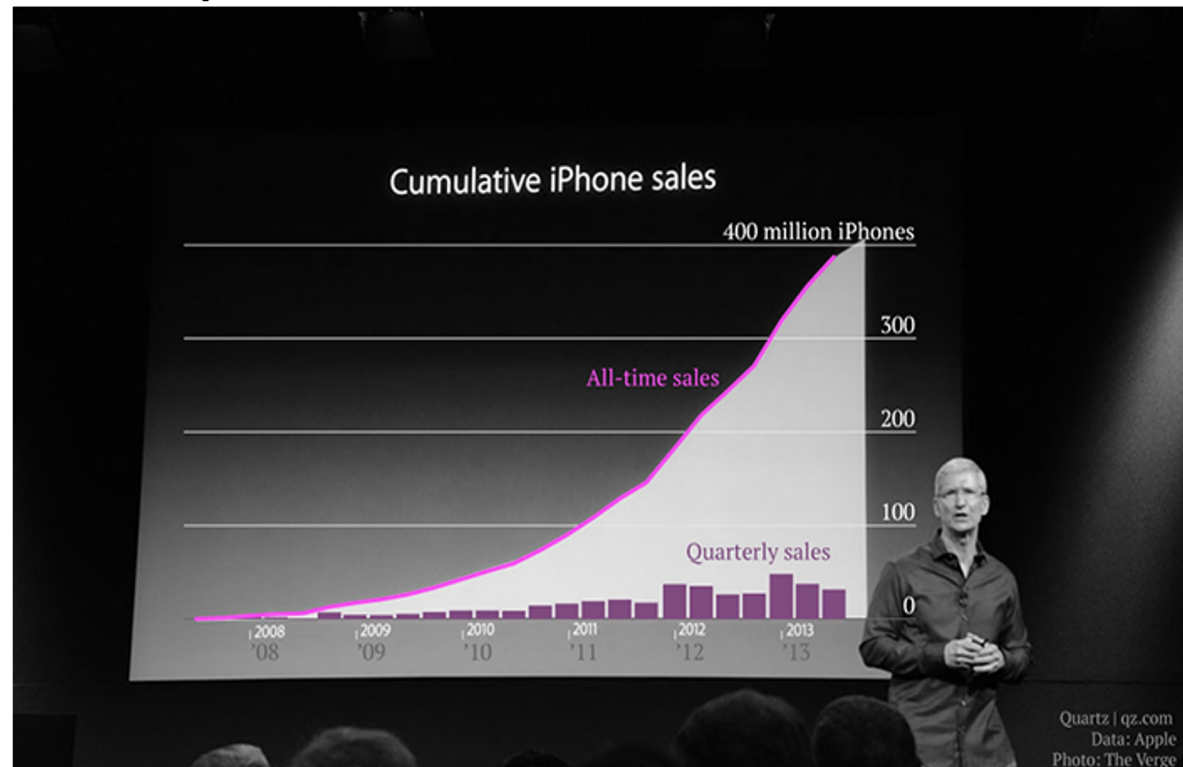


# Насколько взрывным является этот рост на самом деле?

---

Кумулятивные распределения дают ошибочное представление о темпах роста.

Постепенное изменение является производной этой функции, которую трудно визуализировать.





# Тренировка

- Дан ряд углов скольжения самолета в момент сбрасывания бомбы

-20, -60, -10, 30, 60, 70, -10,

-30, -120, -100, -80, 20, 40, -60,

-10, 20, 30, -80, 60, 70

- Построить статистическую функцию распределения
- У кого хорошее решение? Какие ошибки типичны на графике

# Гистограмма

- Если данных много, то простой статистический ряд не удобен
- Разделим наблюдения на разряды и посчитаем частоты попадания:

$$p_i^* = \frac{m_i}{n}$$

- Таблица с интервалами разрядов и  $p_i^*$  называется статистическим рядом

$I_1$	$x_1; x_2$	$x_2; x_3$	$\dots$	$x_i; x_{i+1}$	$\dots$	$x_k; x_{k+1}$
$p_1^*$	$p_1^*$	$p_2^*$	$\dots$	$p_i^*$	$\dots$	$p_k^*$

- Что делать если значение попало на границу интервалов?

Давайте посмотрим решение

<https://colab.research.google.com/drive/1PnR5vCcxVSRN-VdLX86fgH0fYx-LwMJL>

# Построение статистической функции распределения

$$F^*(x_1) = 0$$

$$F^*(x_2) = p_1^*$$

$$F^*(x_3) = p_1^* + p_2^*$$

...

$$F^*(x_k) = \sum_{i=1}^{k-1} p_i^*$$

$$F^*(x_{k+1}) = \sum_{i=1}^k p_i^* = 1$$

# Описательная статистика

---

Описательная статистика предоставляет способы фиксации свойств данного набора данных/выборки.

- Меры центральной тенденции описывают центр распределения данных.
- Меры вариации или изменчивости описывают разброс данных, т.е. насколько далеко измерения лежат от центра.

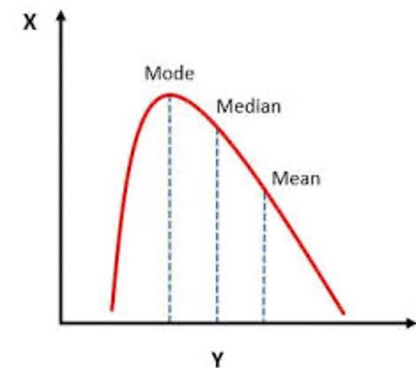
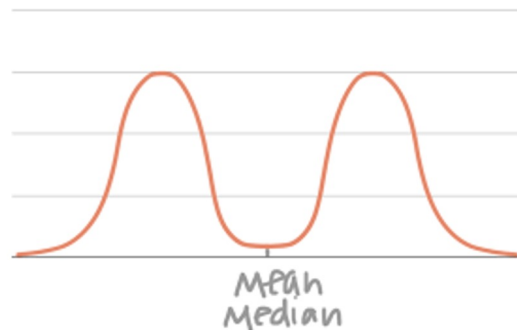
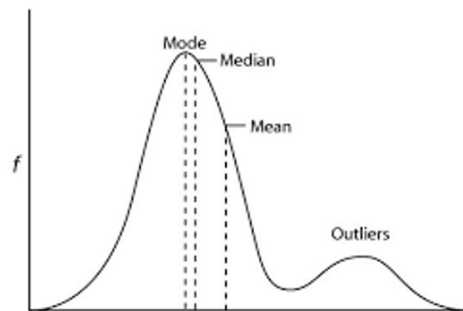
# Мера центральности: среднее значение

---

Чтобы вычислить среднее значение, просуммируйте значения и разделите их на количество наблюдений:

$$\mu_X = \frac{1}{n} \sum_{i=1}^n x_i$$

Среднее значение имеет смысл для симметричных распределений без выбросов.



# Другие меры центральности

---

Медиана представляет собой «серединное» значение.

Среднее геометрическое — это корень  $n$ -й степени из произведения  $n$  значений:

$$\left( \prod_{i=1}^n a_i \right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Среднее геометрическое всегда  $\leq$  среднее арифметическое и более чувствительно к значениям, близким к нулю.

Геометрические средние имеют смысл с соотношениями:

$1/2$  и  $2/1$  должны в среднем давать 1.

# Какая мера лучше всего?

---

Среднее значение имеет смысл для симметричных распределений без выбросов: например, рост и вес.

Медиана лучше подходит для асимметричных распределений или данных с выбросами: например, богатство и доход.

Билл Гейтс добавляет 250 долларов к среднему доходу на душу населения, но ничего не добавляет к медиане.



# Показатель отклонения: стандартное отклонение

---

Дисперсия представляет собой квадрат сигмы стандартного отклонения.

Мы делим на  $n$  или  $n-1$ ?

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

СКО генеральной совокупности делится на  $n$ , СКО выборки на  $n-1$  (почему – узнаем позже), но для больших  $n$   $n \sim (n-1)$ , так что это не имеет особого значения.

# Интерпретация дисперсии (фондовый рынок)

---

Отношение «сигнал/шум» измерить сложно, поскольку многое из того, что вы видите, — это всего лишь дисперсия.

Рассмотрите возможность измерения относительного «навыка» различных инвесторов фондового рынка.

Ежегодные колебания эффективности фондов таковы, что результаты деятельности инвесторов случайны, а это означает, что реальная разница в навыках незначительна.

# Интерпретация дисперсии (много моделей)

---

Обычно для каждой задачи мы разрабатываем несколько моделей, от очень простых до сложных.

Некоторая разница в производительности будет объяснена простой дисперсией: какие пары обучения/оценки были выбраны, насколько хорошо были оптимизированы параметры и т. д.

Небольшой выигрыш в производительности является аргументом в пользу более простых моделей.

# Методы уменьшения дисперсии

---



Хотя идти на занятия пешком медленнее, чем ехать на автобусе, разница во времени прибытия меньше.

Повторение эксперимента несколько раз уменьшает дисперсию (перекрестная проверка в  $k$ -кратном размере).

То же самое относится и к правильной случайной и детерминированной выборке.

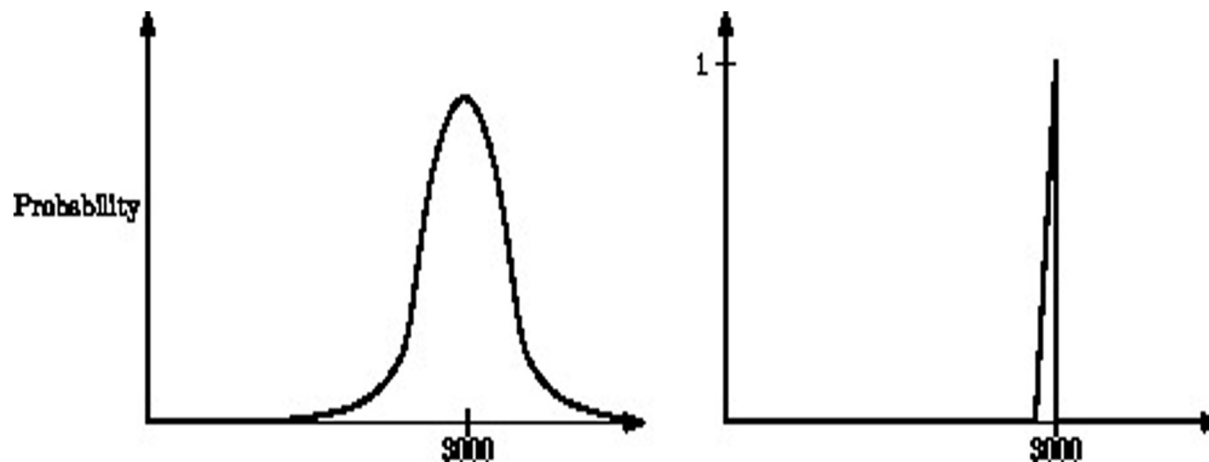
Устранение выбросов (если это оправдано) уменьшает дисперсию.

# Распределение срока службы картриджей принтера

---

Распределения с одинаковым средним значением могут выглядеть очень по-разному.

Но вместе среднее и стандартное отклонение довольно хорошо характеризуют любое распределение.



# Приближенные вычисления

---

$$m_x^* = M^*[X] = \sum_{i=1}^k \tilde{x}_i p_i^*$$

$$D_x^* = D^*[X] = \sum_{i=1}^k (\tilde{x}_i - m_x^*)^2 p_i^*$$

$$\alpha_s^*[X] = \sum_{i=1}^k \tilde{x}_i^s p_i^*$$

$$\mu_s^*[X] = \sum_{i=1}^k (\tilde{x}_i - m_x^*)^s p_i^*$$

# Выравнивание статистических рядов

---

- Во всяком статистическом распределении неизбежно присутствуют элементы случайности, связанные с тем, что число наблюдений ограничено, что произведены именно те, а не другие опыты, давшие именно те, а не другие результаты.
- Только при очень большом числе наблюдений эти элементы случайности сглаживаются, и случайное явление обнаруживает в полной мере присущую ему закономерность.
- На практике мы почти никогда не имеем дела с таким большим числом наблюдений и вынуждены считаться с тем, что любому статистическому распределению свойственны в большей или меньшей, мере черты случайности.
- Поэтому при обработке статистического материала часто приходится решать вопрос о том, как подобрать для данного статистического ряда теоретическую кривую распределения, выражающую лишь существенные черты статистического материала, но не случайности, связанные с недостаточным объемом экспериментальных данных. Такая задача называется задачей **выравнивания** (сглаживания) **статистических рядов**.
- Задача выравнивания заключается в том, чтобы подобрать теоретическую плавную кривую распределения, с той или иной точки зрения наилучшим образом описывающую данное статистическое распределение.

# Как выравнивать?

---

- Как правило, принципиальный вид теоретической кривой выбирается заранее из соображений, связанных с существом задачи, а в некоторых случаях просто с внешним видом статистического распределения. Аналитическое выражение выбранной кривой распределения зависит от некоторых параметров; задача выравнивания статистического ряда переходит в задачу рационального выбора тех значений параметров, при которых соответствие между статистическим и теоретическим распределениями оказывается наилучшим.

Например,

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

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \text{при } \alpha \leq x \leq \beta \\ 0, & \text{при } x < \alpha \text{ или } x > \beta \end{cases} \quad (2)$$

Что это за законы? Что будем подбирать?



# Требуемые ограничения

---

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

# Метод моментов

---

- Согласно методу моментов, параметры  $a, b, \dots$  выбираются с таким расчетом, чтобы несколько важнейших числовых характеристик (моментов) теоретического распределения были равны соответствующим статистическим характеристикам.
- Например, если теоретическая кривая  $f(x)$  зависит только от двух параметров  $a$  и  $b$ , эти параметры выбираются так, чтобы математическое ожидание и дисперсия теоретического распределения совпадали с соответствующими статистическими характеристиками.
- Если кривая  $f(x)$  зависит от трех параметров, можно подобрать их так, чтобы совпали первые три момента, и т. д.
- При выравнивании статистических рядов может оказаться полезной специально разработанная система **кривых Пирсона**, каждая из которых зависит в общем случае от четырех параметров. При выравнивании эти параметры выбираются с тем расчетом, чтобы сохранить первые четыре момента статистического распределения (математическое ожидание, дисперсию, третий и четвертый моменты).

# Пример

---

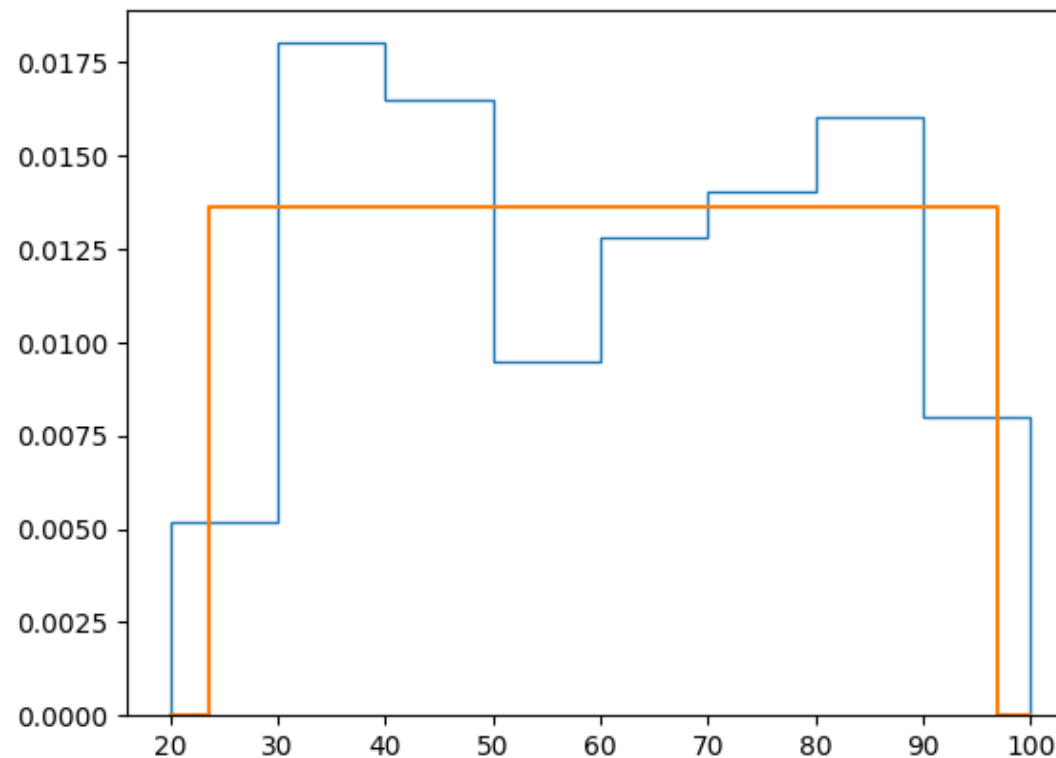
- С целью исследования закона распределения ошибки измерения дальности с помощью радиодальномера произведено 400 измерений дальности. Результаты опытов представлены в виде статистического ряда:

$I_i$ (м)	20; 30	30; 40	40; 50	50; 60	60; 70	70; 80	80; 90	90; 100
$m_i$	21	72	66	38	51	56	64	32
$p_i^*$	0,052	0,180	0,165	0,095	0,128	0,140	0,160	0,080

# Решение

---

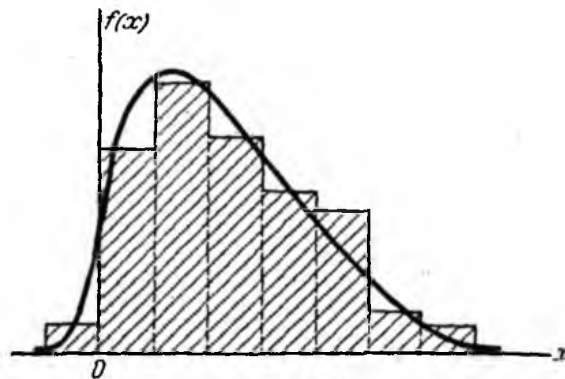
<https://colab.research.google.com/drive/1Ce0rPKw1jdU7mRzOkgNRiKIlIn68XDzR0>



# Критерии согласия

---

- Допустим, что данное статистическое распределение выравнено с помощью некоторой теоретической кривой



- Как бы хорошо, ни была подобрана теоретическая кривая, между нею и статистическим распределением неизбежны некоторые расхождения.
- Вопрос: объясняются ли эти расхождения только случайными обстоятельствами, связанными с ограниченным числом наблюдений, или они являются существенными и связаны с тем, что подобранная нами кривая плохо выравнивает данное статистическое распределение?
- Для ответа служат «критерии согласия».

# Идея метода

---

- Гипотеза ***H***: случайная величина ***X*** подчиняется некоторому определенному закону распределения. Этот закон может быть задан в той или иной форме: например, в виде функции распределения ***F(x)*** или в виде плотности распределения ***f(x)***, или же в виде совокупности вероятностей ***p<sub>i</sub>***, где ***p<sub>i</sub>*** — вероятность того, что величина ***X*** попадет в пределы ***i***-го разряда.
- рассмотрим величину ***U***, характеризующую степень расхождения теоретического и статистического распределений.
- Величина ***U*** может быть выбрана различными способами; например, в качестве ***U*** можно взять сумму квадратов отклонений теоретических вероятностей ***p<sub>i</sub>*** от соответствующих частот ***p<sub>i</sub><sup>\*</sup>*** или же сумму тех же квадратов с некоторыми коэффициентами («весами»), или же максимальное отклонение статистической функции распределения ***F<sup>\*</sup>(x)*** от теоретической ***F(x)*** и т. д. Допустим, что величина ***U*** выбрана тем или иным способом. Очевидно, это есть некоторая ***случайная величина***.

# Идея метода (2)

---

- Закон распределения случайной величины  $U$  зависит от закона распределения случайной величины  $X$ , над которой производились опыты, и от числа опытов  $n$ . Если гипотеза  $H$  верна, то закон распределения величины  $U$  определяется законом распределения величины  $X$  (функцией  $F(x)$ ) и числом  $n$ .
- Допустим, что этот закон распределения нам известен. В результате данной серии опытов обнаружено, что выбранная нами мера расхождения  $U$  приняла некоторое значение  $u$ .
- Можно ли объяснить это случайными причинами или же это расхождение слишком велико и указывает на наличие существенной разницы между теоретическим и статистическим распределениями и, следовательно, на непригодность гипотезы  $H$ ?

## Идея метода (3)

---

- Предположим, что гипотеза ***H*** верна, и вычислим в этом предположении вероятность того, что за счет случайных причин, связанных с недостаточным объемом опытного материала, мера расхождения ***u*** окажется не меньше, чем наблюденное нами в опыте значение. Вычислим вероятность события:

$$U \geq u$$

- Если эта вероятность весьма мала, то гипотезу следует *отвергнуть* как мало правдоподобную
- Если же эта вероятность значительна, следует признать, что *экспериментальные данные не противоречат гипотезе ***H****.



# Как следует выбирать $U$ ?

---

- При некоторых способах ее выбора закон распределения величины  $U$  обладает весьма простыми свойствами при достаточно большом  $n$  практически не зависит от функции  $F(x)$ .

# Критерий Хи-квадрат Пирсона

---

- Произведено  $n$  независимых опытов, в каждом из которых случайная величина  $X$  приняла определенное значение. Результаты опытов сведены в  $k$  разрядов и оформлены в виде статистического ряда:

$I_l$	$x_1; x_2$	$x_2; x_3$	$\dots$	$x_k; x_{k+1}$
$p_l^*$	$p_1^*$	$p_2^*$	$\dots$	$p_k^*$

- Зная теоретический закон распределения, можно найти теоретические вероятности попадания случайной величины в каждый из разрядов:

$$p_1, p_2, \dots, p_k$$

- В качестве меры возьмем

$$U = \sum_{i=1}^k c_i (p_i^* - p_i)^2$$

# Критерий Хи-квадрат Пирсона (2)

---

- Коэффициенты  $c_i$  («веса» разрядов) вводятся потому, что в общем случае отклонения, относящиеся к различным разрядам, нельзя считать равноправными по значимости.
- Одно и то же по абсолютной величине отклонение  $p_i^* - p_i$  может быть мало значительным, если сама вероятность  $p_i$  велика, и очень заметным, если она мала. Поэтому естественно «веса»  $c_i$  взять обратно пропорциональными вероятностям разрядов  $p_i$ .

- **Если положить**

$$c_i = \frac{n}{p_i}$$

то при больших  $n$  закон распределения величины  $U$  обладает весьма простыми свойствами:

он практически не зависит от функции распределения  $F(x)$  и от числа опытов  $n$ , а зависит только от числа разрядов  $k$ ,

Этот закон при увеличении  $n$  приближается к «распределению  $\chi^2$ »

# Мера расхождения

---

$$\chi^2 = n \sum_{i=1}^k \frac{(p_i^* - p_i)^2}{p_i}$$

$$U = \chi^2 = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}$$

1) Распределением  $\chi^2$  с  $r$  степенями свободы называется распределение суммы квадратов  $r$  независимых случайных величин, каждая из которых подчинена нормальному закону с математическим ожиданием, равным нулю, и дисперсией, равной единице. Это распределение характеризуется плотностью

$$k_r(u) = \begin{cases} \frac{1}{2^{\frac{r}{2}} \Gamma\left(\frac{r}{2}\right)} u^{\frac{r}{2}-1} e^{-\frac{u}{2}} & \text{при } u > 0, \\ 0 & \text{при } u < 0, \end{cases}$$

где  $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$  — известная гамма-функция.

# Как определить число степеней свободы

---

- Распределение  $\chi^2$  зависит от параметра  $r$ , называемого числом «степеней свободы» распределения. Число «степеней свободы»  $r$  равно числу разрядов  $k$  минус число независимых условий («связей»), наложенных на частоты  $p_i$ .
- **Примеры условий:**

$$\sum_{i=1}^k p_i^* = 1$$

$$\sum_{i=1}^k \tilde{x}_i p_i^* = m_x$$

$$\sum_{i=1}^k (\tilde{x}_i - m_x^*)^2 p_i^* = D_x$$

# Как посчитать критерий

---

<https://colab.research.google.com/drive/1EuF6rmUqZt8EQEThsVeiSbNZ8G7TKcx>

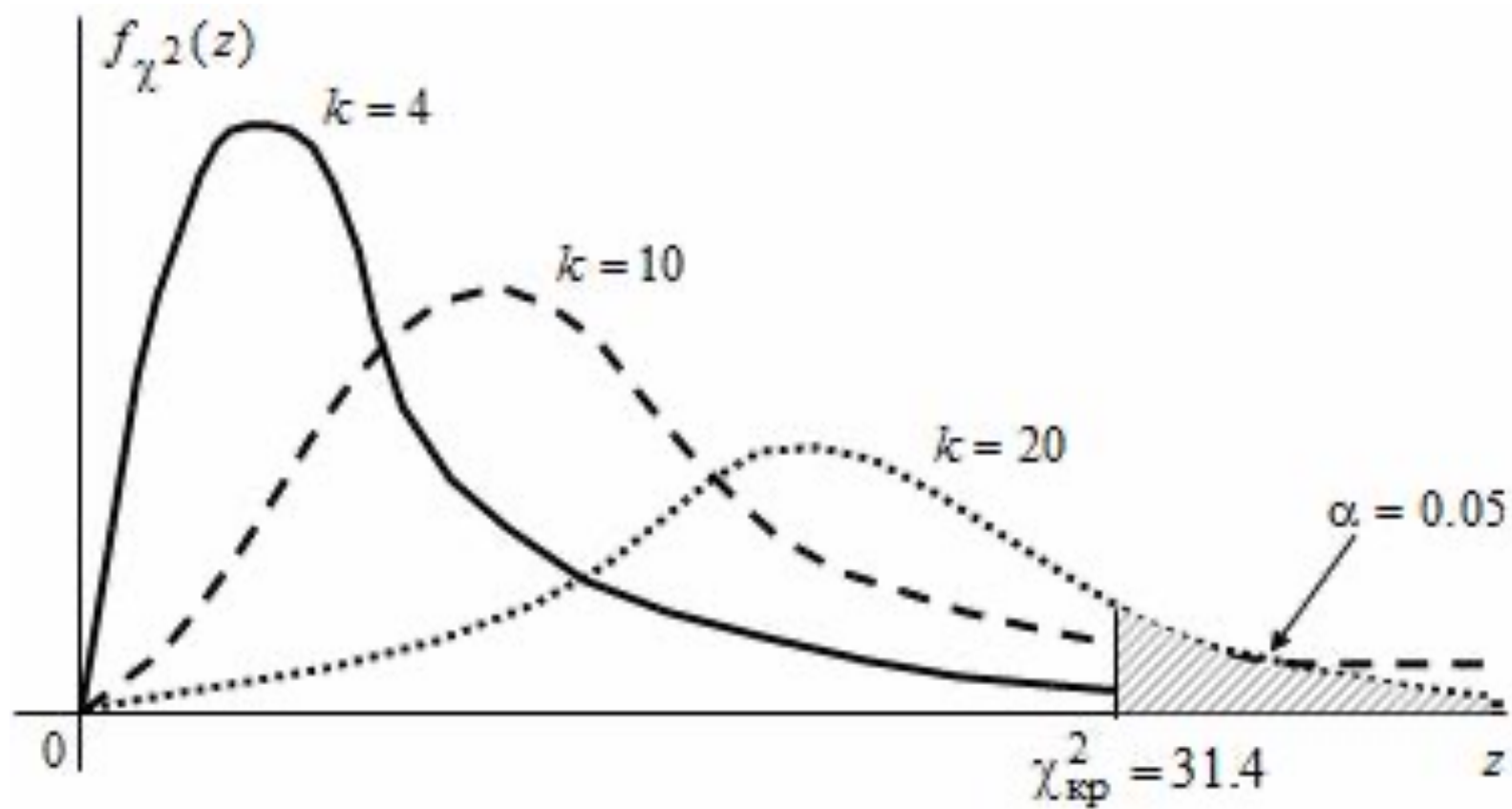
# Смысл p-value

---

- Распределение  $\chi^2$  дает возможность оценить степень согласованности теоретического и статистического распределений,
- Будем исходить из того, что величина  $X$  действительно распределена по закону  $F(x)$ .
- Тогда вероятность  $p$  (**p-value**), есть вероятность того, что за счет чисто случайных причин мера расхождения теоретического и статистического распределений будет не меньше, чем фактически наблюдаемое в данной серии опытов значение  $\chi^2$ .
- Если эта вероятность,  $p$  весьма мала (настолько мала, что событие с такой вероятностью можно считать практически невозможным), то результат опыта следует считать противоречащим гипотезе  $H$  о том, что закон распределения величины  $X$  есть  $F(x)$ .
- Эту гипотезу следует отбросить как неправдоподобную. Напротив, если вероятность  $p$  сравнительно велика, можно признать расхождения между теоретическим и статистическим распределениями несущественными и отнести их за счет случайных причин. Гипотезу  $H$  о том, что величина  $X$  распределена по закону  $F(x)$ , можно считать правдоподобной или, по крайней мере, не противоречащей опытными данным.

# Смысл p-value

---





# На сколько должно быть мало $p$ -value?

---

- Вопрос неопределенный; он не может быть решен из математических соображений, так же как и вопрос о том, насколько мала должна быть вероятность события для того, чтобы считать его практически невозможным.
- На практике, если  $p$  оказывается меньшим чем 0,1, рекомендуется проверить эксперимент, если возможно — повторить его и в случае, если заметные расхождения снова появятся, пытаться искать более подходящий для описания статистических данных закон распределения.
- С помощью критерия  $\chi^2$  (или любого другого критерия согласия) можно только в некоторых случаях **опровергнуть** выбранную гипотезу  $H$  и отбросить ее как явно несогласную с опытными данными.
- Если же вероятность  $p$  велика, то этот факт сам по себе ни в коем случае не может считаться доказательством справедливости гипотезы  $H$ , а указывает только на то, что гипотеза не противоречит опытными данным.

# Критерий Колмогорова

---

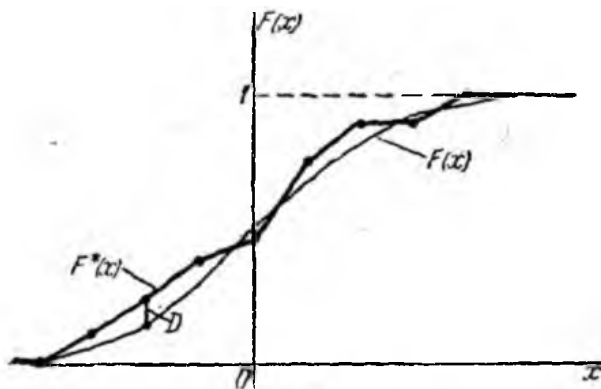
$$D = \max |F^*(x) - F(x)|$$

- Какова бы ни была функция распределения  $F(x)$  непрерывной случайной величины  $X$ , при неограниченном возрастании числа независимых наблюдений  $n$  вероятность неравенства

$$D\sqrt{n} \geq \lambda$$

- стремится к пределу

$$P(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 \lambda^2}$$



# Критерий Колмогорова. Когда применим?

---

- можно применять только в случае, когда гипотетическое распределение  $F(x)$  полностью известно заранее из каких-либо теоретических соображений, т. е. когда известен не только вид функции распределения  $F(x)$ , но и все входящие в нее параметры.
- Такой случай сравнительно редко встречается на практике. Обычно из теоретических соображений известен только общий вид функции  $F(x)$ , а входящие в нее числовые параметры определяются по данному статистическому материалу.
- При применении критерия  $\chi^2$  это обстоятельство учитывается соответствующим уменьшением числа степеней свободы распределения  $\chi^2$ . Критерий А. Н. Колмогорова такого согласования не предусматривает. Если все же применять этот критерий в тех случаях, когда параметры теоретического распределения выбираются по статистическим данным, критерий дает заведомо завышенные значения вероятности  $P(X)$ ; поэтому мы в ряде случаев рискуем принять как правдоподобную гипотезу, в действительности плохо согласующуюся с опытными данными.