Статистические методы и анализ данных (1)

Игорь Бойко

Мифы о статистической погрешности

Рост экономики РФ – на уровне статистической погрешности

ЭКОНОМИСТ
Сергей Хестанов

НЕЗАВИСИМАЯ

ВВП вырос в пределах статистической погрешности

ИЗВЕСТИЯ iz

Радио ТВ

Соответственно, прогнозы 2,9 или 3,8% роста ВВП находятся в пределах статистической погрешности, подчеркивает Денис Палеев.

Европейский банк реконструкции и развития (ЕБРР) ожидает роста ВВП России в 1,2% в 2017 году. Ранее сегодня министр экономического развития России Максим

Рост экономики РФ в 2,3%, язык не поворачивается назвать ростом, считает экономист, директор Института проблем глобализации **Михаил Делягин**. Об этом он заявил 5 мая корреспонденту **ИА REGNUM**, комментируя данные Всемирного банка о рекордном росте экономики РФ, который составил 2,3%.

При этом Делягин подчеркнул, что рост в 2,3% – это действительно рекордный рост за последние шесть лет.

«Это правда. Другое дело, что назвать это ростом язык не поворачивается. Потому что всё, что в пределах 3%, это статистическая погрешность. И у нас, и в США, и во всём мире», – подытожил экономист.

В прошедшем 1997 г. темпы экономического роста были незначительными — увеличение реального ВВП на 0,8% находится в пределах статистической погрешности. Вероятность увеличения ВВП в 1998 г.

Мифы о статистической погрешности

Опрошено 1600 человек в 130 населенных пунктах в 42 областях, краях и республиках России. Статистическая погрешность не превышает 3,4%.

рассказал гендиректор ВЦИОМ **Валерий Федоров**. «Соотношение политических сил установилось и в последний месяц варьируется незначительно», — заметил глава ВЦИОМ.

Рейтинг остальных кандидатов не превышает статистической погрешности, отметил Федоров. Так, поддержка телеведущей **Ксении Собчак** в районе 1%, основателя «Яблока» **Григория Явлинского**, — 0,8%,

В настоящее время рейтинг Пенса находится на уровне статистической погрешности — 3,8% — и он занимает 4-ю строчку среди республиканцев с президентскими амбициями.

Безоговорочным лидером остаётся Трамп — 53,2%,

Когда возникает статистичская погрешность?

- Если мы имеем полную информацию обо всех объектах, то статистическая погрешность равна нулю.
  - Например, если правительство получило отчёты о продукции всех предприятий.
  - Или если на выборах проголосовали 100% граждан.
- Статистическая погрешность возникает, когда свойства генеральной совокупности мы заменяем свойствами выборки (выборочной совокупности).
  - Например, если ВВП страны оценить по ВВП одного города.
  - Или предпочтения избирателей оценить опросом 1600 респондентов

Откуда вообще берётся погрешность??

- Физика считается точной наукой. Откуда же берутся погрешности измерений?
- Флуктуации – это фундаментальное свойство природы (квантовая механика!)
- Некоторые измерения являются статистическими по своей природе. Например, сколько молекул воздуха содержится в 1 $см^3$?
- Ну и многое другое: несовершенство аппаратуры, упрощения, применяемые при измерениях или рассуждениях, ограниченность наших знаний

Каков будет исход измерения?

- Допустим, на коллайдере рождается в среднем 1 Хиггс-бозон в минуту.
- Сколько их родится в течение следующей минуты?
- Может быть 2, а может быть и 0. Или даже 10. «1 в минуту» - это средняя величина, при усреднении большого числа таких минут.
- А какова вероятность, что в следующую минуту родится 0, или 1, или 10 Хиггс-бозонов?

Случайная величина

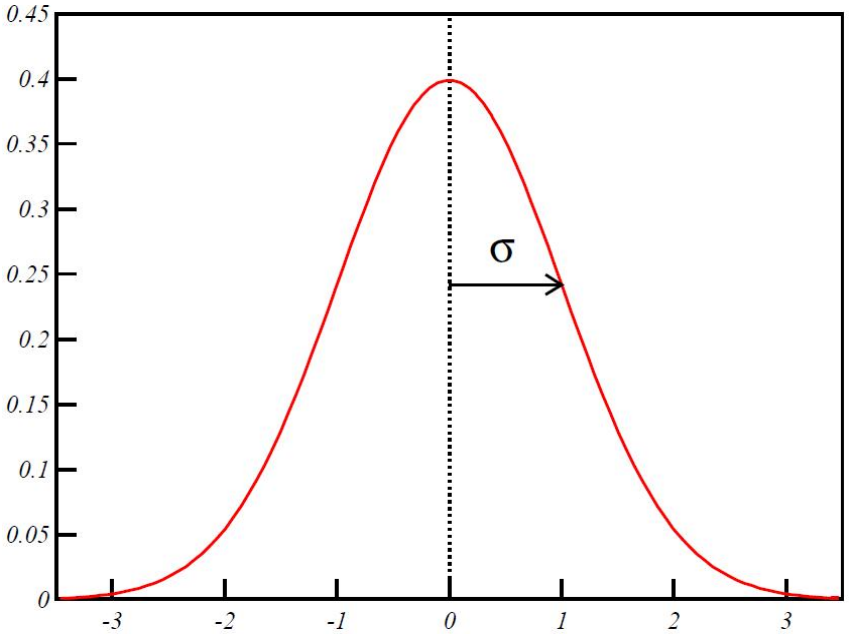
- Случайная величина x – это экспериментально измеряемая переменная, значение которой не может быть предсказано до проведения измерения
- Случайная величина бывает дискретной или непрерывной
- Распределение плотности вероятности (PDF):

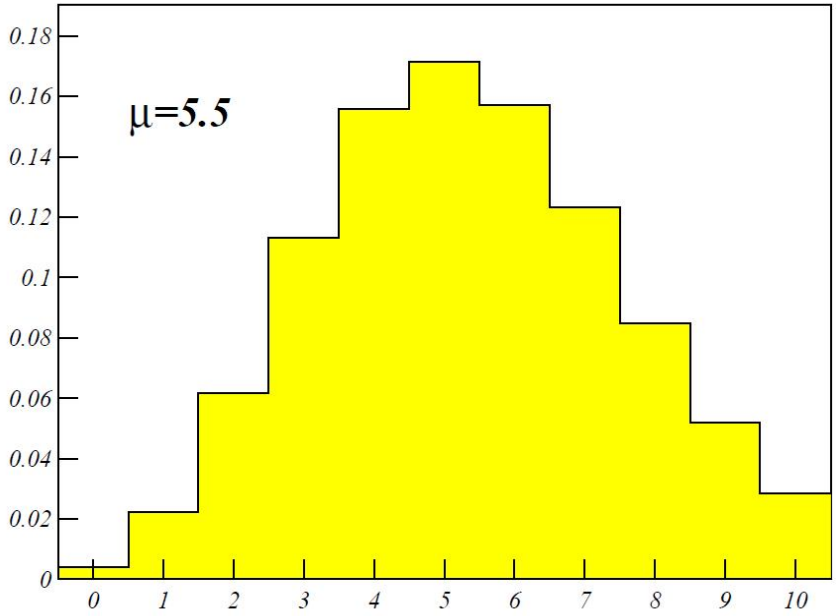
$$p(x < X < x + dx) = f(x)dx$$

$$\int f(x)dx = 1$$

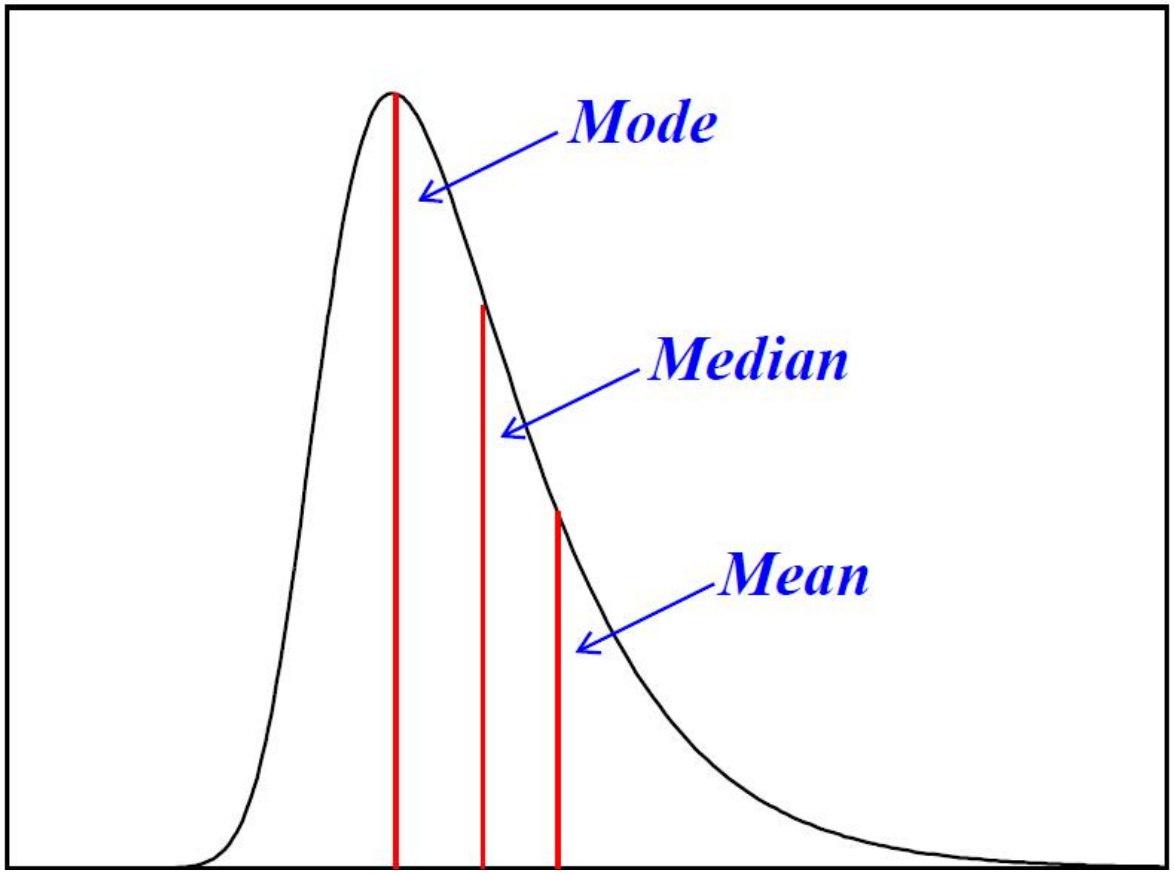
$$\sum_i P_i = 1$$

Примеры плотности вероятности





Характеристики плотности вероятности



- Среднее
- Медиана
- Мода

Среднее и дисперсия

- Функция g(x) от случайной величины – это тоже случайная величина

$$E(g(x)) = \overline{g}_f = \int g(x)\ f(x)dx$$

$$D(g(x)) = E[g(x) - E(g(x))]^2$$

- В частности, среднее и дисперсия самой («исходной») случайной величины x:

$$\mu = E(x) = \overline{x} = \int x\ f(x)dx$$

$$\sigma^2 = D(x) = \overline{(x-\mu)^2} = \int (x-\mu)^2\ f(x)dx = \overline{x^2} - (\overline{x})^2$$

В эксперименте всегда измеряется *выборка* распределения случайной величины

- Таким образом, следует различать:
  - (истинные) параметры распределения $(\mu, \sigma, \ldots)$
  - (измеренные) характеристики выборки $(\overline{x}, S, \ldots)$
  - оценки параметров распределения по измеренным характеристикам выборки $(\hat{\mu}, \hat{\sigma}, \ldots)$

Характеристики выборки

$$\overline{\mathrm{x}} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\mathrm{S}^2 = \frac{1}{n}\sum_{i}^{n} (x_i - \overline{x})^2$$

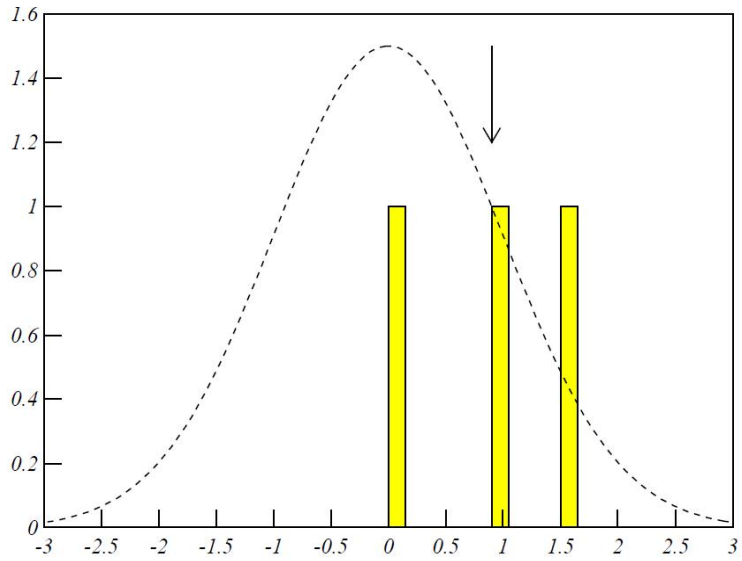
- Оценка параметров распределения по характеристикам выборки:

$$\hat{\mu} = \overline{\mathrm{x}} \pm \frac{\sigma}{\sqrt{n}}$$

$$\hat{\sigma}^2 = \frac{n}{n-1}\mathrm{S}^2 \qquad (!!!)$$

- Для нормального распределения: $D(\hat{\sigma}^2) = \frac{2\sigma^4}{n-1}$

Откуда берется единичка в формуле $\hat{\sigma}^2 = \frac{n}{n-1} S^2$



- Дисперсию следовало бы вычислять вокруг истинного среднего μ. Но оно неизвестно! В эксперименте мы вынуждены использовать среднее выборки $\hat{\mu}$ которое обеспечивает разброс меньше (или равный) истинной дисперсии

Два самых важных распределения

- Распределение Пуассона
  - Количество независимых событий, которые случаются за фиксированный промежуток времени
- Нормальное (гауссово) распределение
  - Результаты измерений как правило имеют нормальное распределение вокруг истинного значения

Распределение Пуассона

- Бозоны Хиггса рождаются (почти) независимо друг от друга.
- При небольшом траффике машины проезжают по дороге независимо друг от друга.
- Поэтому количество бозонов или машин в течение 1 минуты подчиняется распределению Пуассона

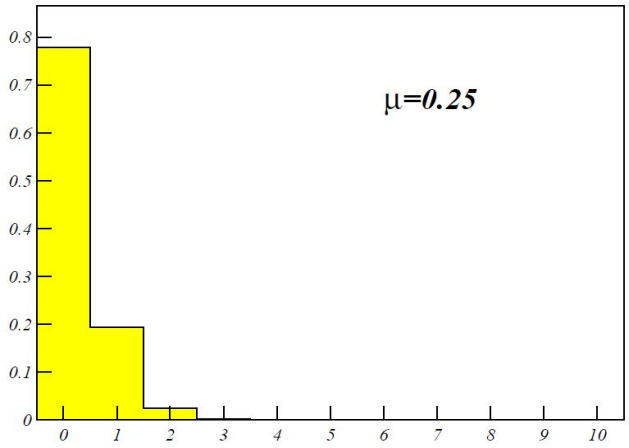
$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

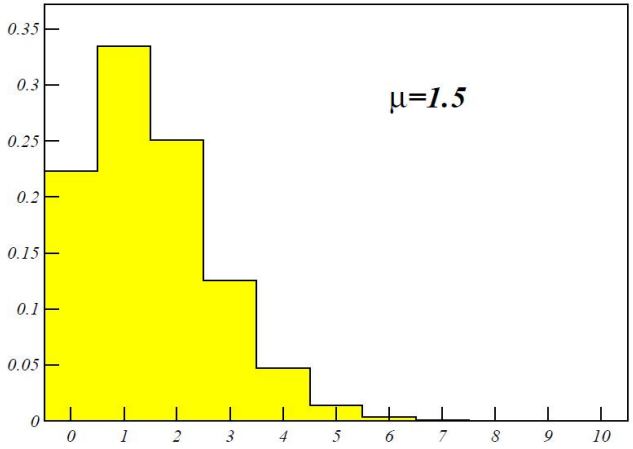
Распределение Пуассона

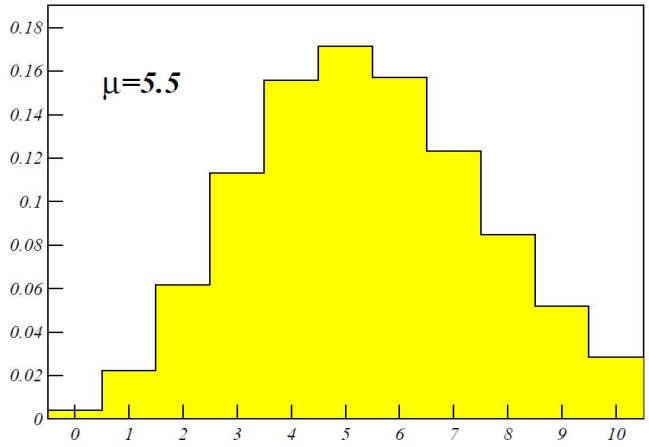
$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

$$E(n) = \mu \qquad D(n) = \mu$$

$$\hat{\mu} = n \pm \sqrt{n}$$

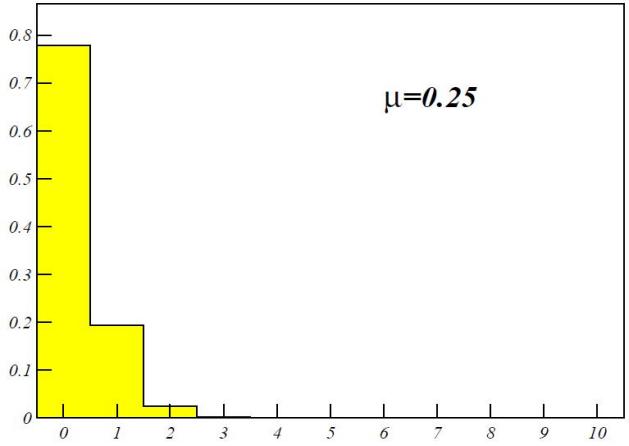


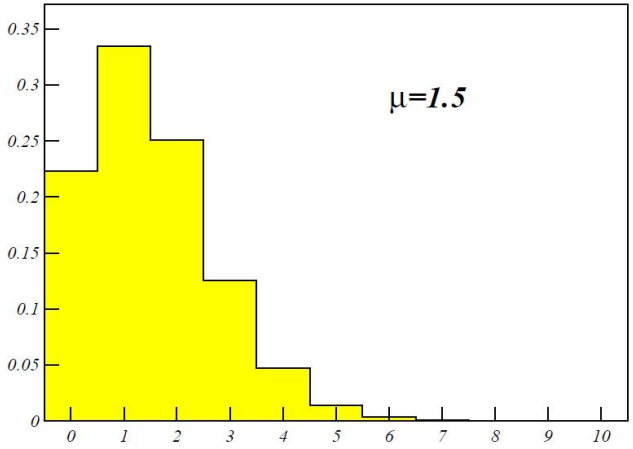


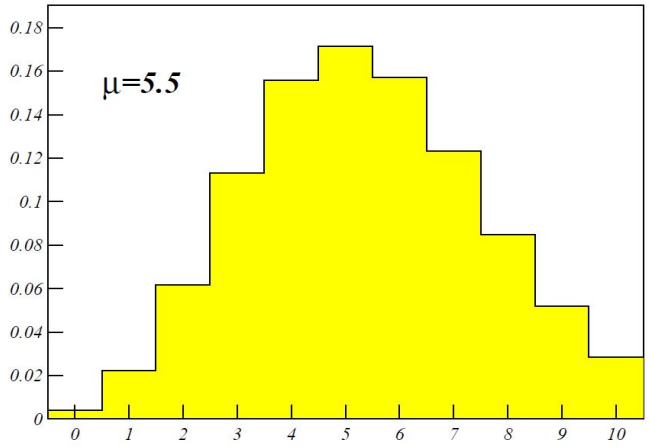


Распределение Пуассона

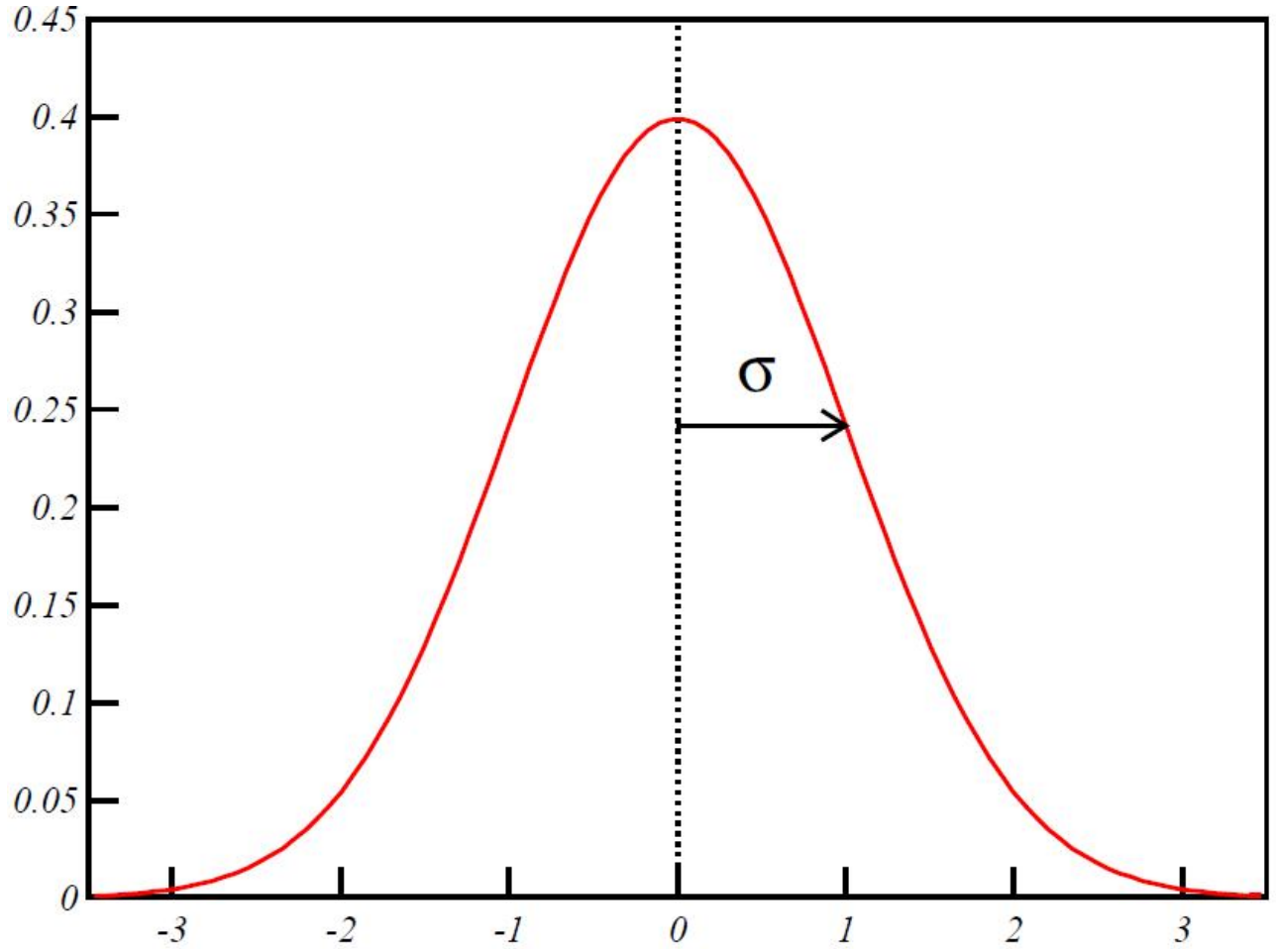
- Если $\mu<<1$, то $P(0)\sim 1-\mu$, $P(1)\sim\mu$, $P(2)\sim 0$
- Если $\mu>>1$, переходит в распределение Гаусса
  - $\langle n\rangle=\mu$, $\sigma=\sqrt{\mu}$







Нормальное распределение



Нормальное распределение

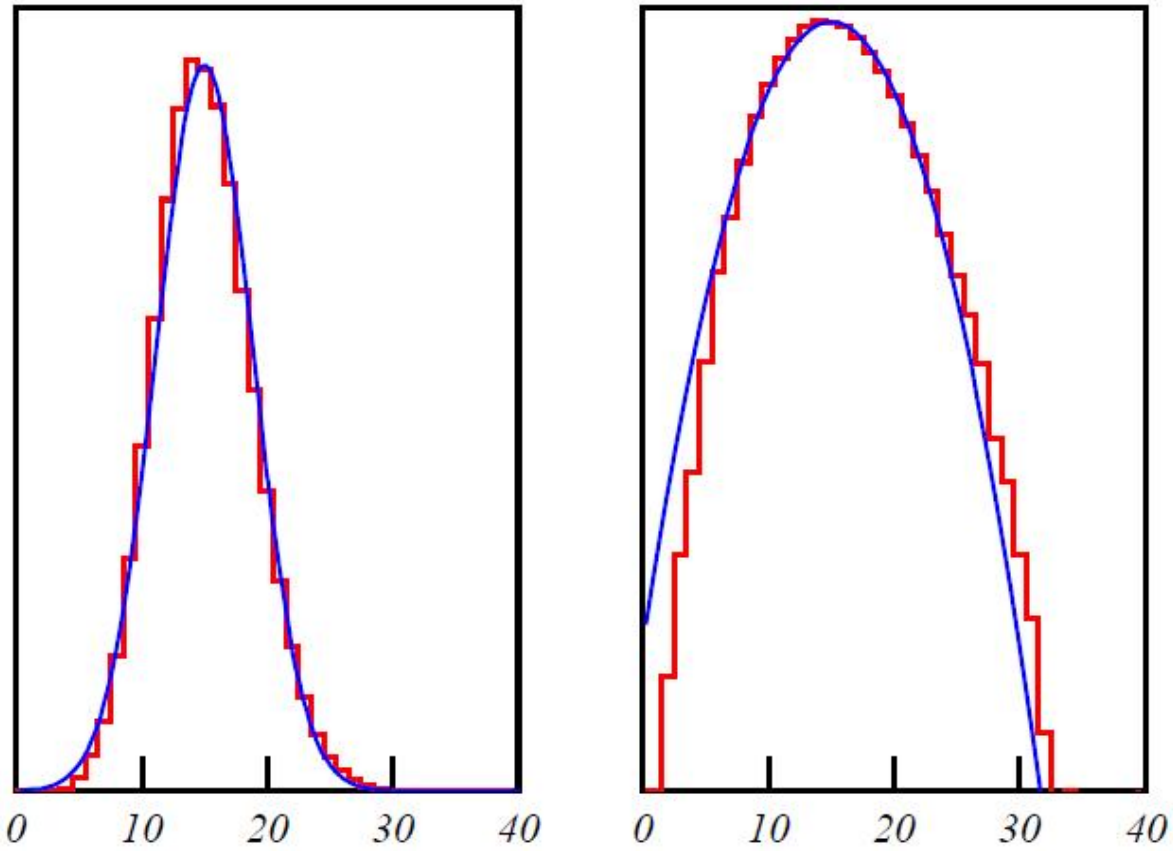
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2} = N(\mu, \sigma^2)$$

$$E(x) = \mu \qquad\qquad D(x) = \sigma^2$$

- Стандартное нормальное распределение: $\mu = 0,\ \sigma = 1$

$$N(0,1) = g(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}$$

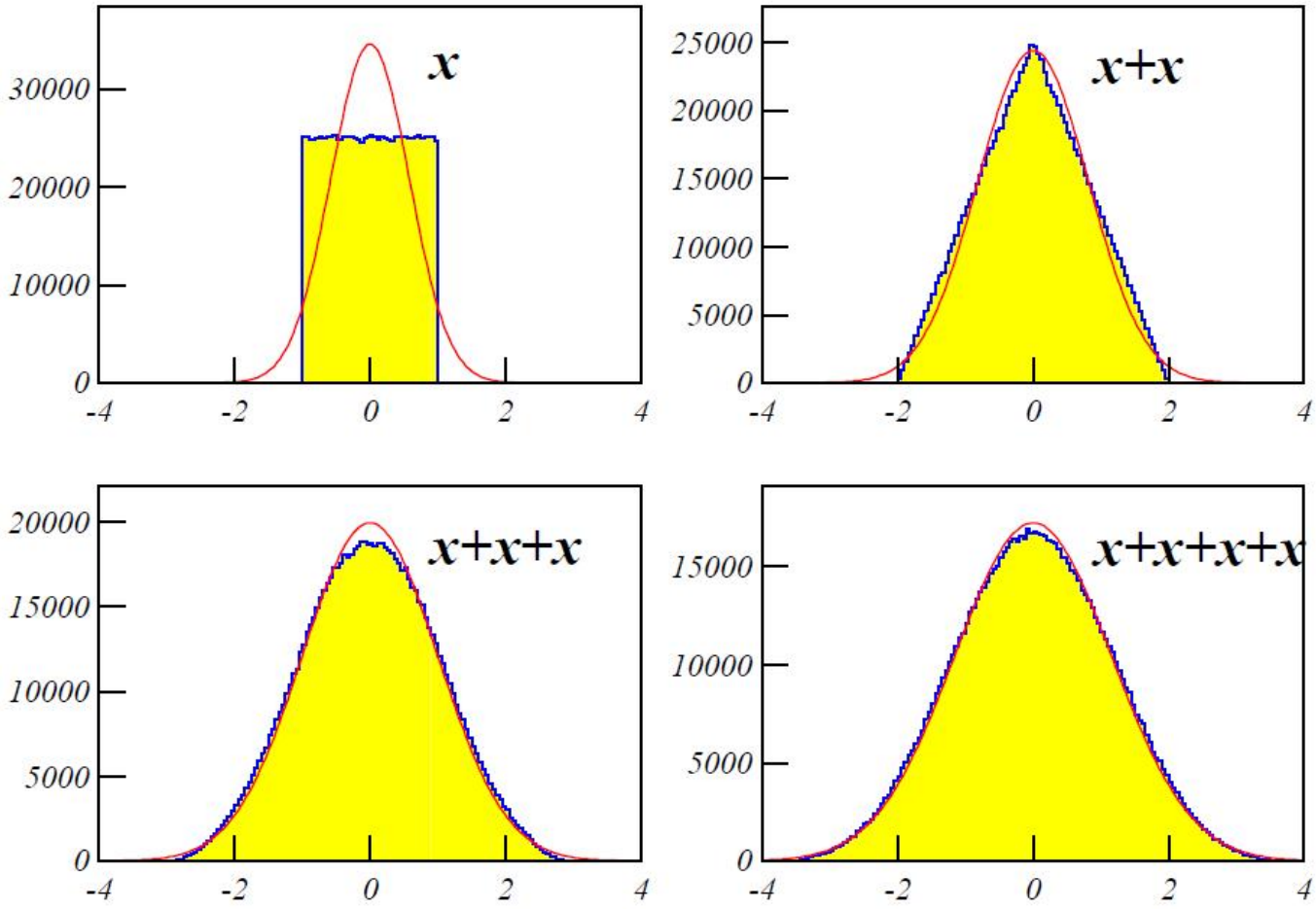
Сравнение равпределений Пуассона (μ=15) и Гаусса



Центральная предельная теорема

- Сумма большого количества слагаемых имеет нормальное распределение в пределе бесконечного числа слагаемых.
  - Внимание! Эта формулировка неправильная (предельно упрощенная), но в целом передает смысл.
- Никогда не путайте сумму распределений $f(x)+g(x)+\ldots$ и распределение суммы $f(x+y+\ldots)$!!!
- Как правило, экспериментальная погрешность складывается из множества случайных факторов. Поэтому ошибка измерений как правило хорошо описывается распределением Гаусса.

Иллюстрация ЦПТ

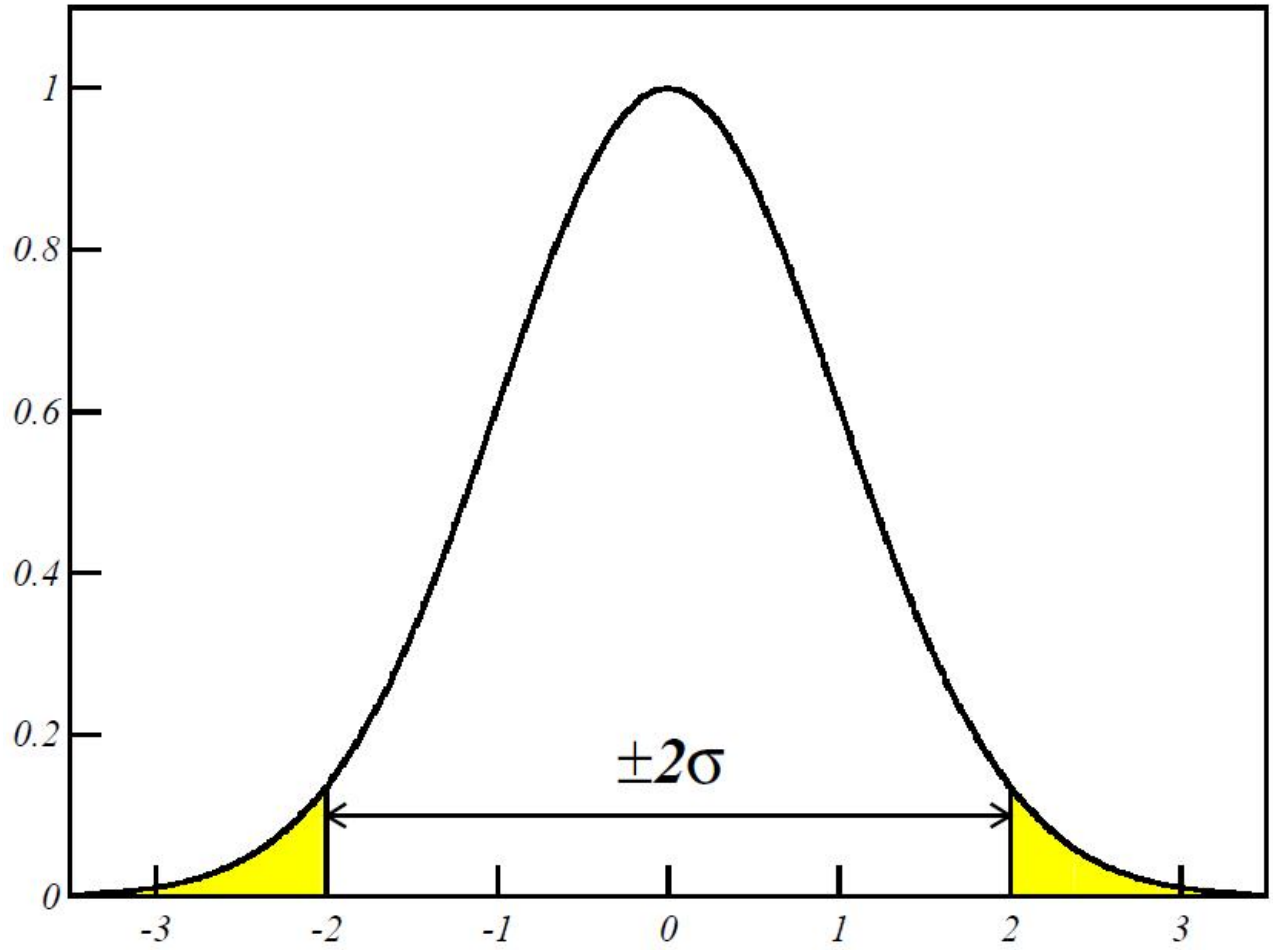


Квантиль

- Квантиль уровня α – это значение, которое с вероятностью α не будет превышено случайной величиной.
- Пример: если ровно 90% людей имеют собственность меньше 1M$, то мы говорим, что 1M$ это 90-процентный квантиль распределения богатства
- Для нормального распределения удобно использовать 2-сторонний квантиль: расстояние от среднего, которое не превышается с вероятностью α:

$$\alpha = P(|x-\mu| < k\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-k\sigma}^{\mu+k\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

Пример 2-стороннего квантиля



Квантили нормального распределения

Квантиль $\|x-\mu\|<k\sigma$	Уровень α
1 σ	68.27%
2 σ	95.42%
3 σ	99.73%
1.645 σ	90%
1.960 σ	95%
2.57 σ	99%
3.29 σ	99.9%

“Столько-то сигма”

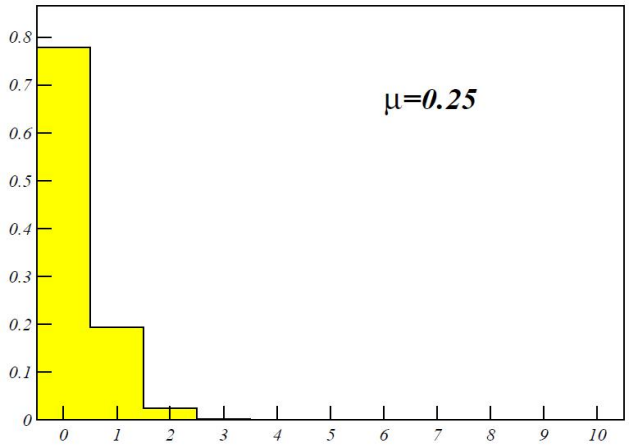
- Если измерения имеют нормальное распределение, то результат измерения отклоняется от истинного значения:
  - в пределах $\pm 1\sigma$ – с вероятностью 68%
  - в пределах $\pm 2\sigma$ – с вероятностью 95%
  - в пределах $\pm 3\sigma$ – с вероятностью 99.7%
- Даже для негауссовых распределений вероятность часто переводят в квантили нормального распределения. Например, мы говорим: «эффект составляет 2.5σ». Это означает, что наблюдается явление, имеющее вероятность 1.2%

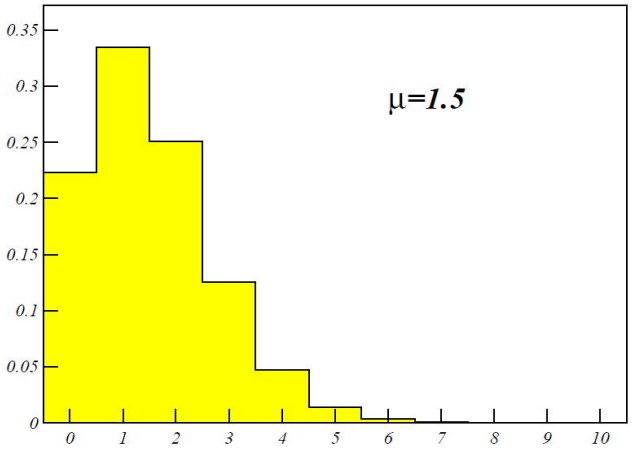
Вернёмся к распределению Пуассона…

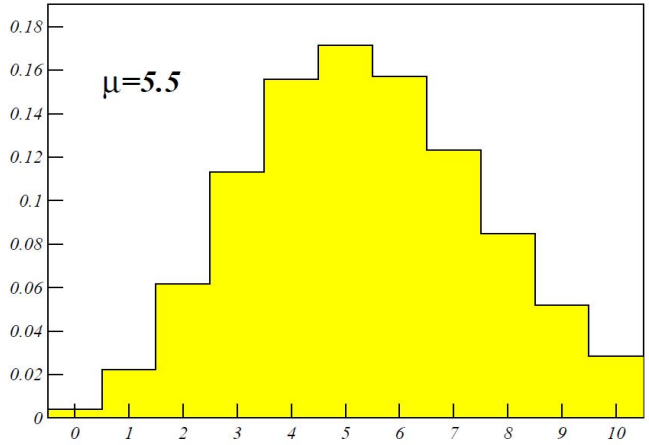
$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

$$E(n) = \mu \qquad D(n) = \mu$$

$$\hat{\mu} = n \pm \sqrt{n}$$







А если события не независимы?

- По определению, распределение Пуассона – это число независимых событий.
- Если в городе 1000 машин, значит за 1 минуту по дороге не сможет проехать 1001. Тот факт, что случились предыдущие 1000 событий – влияет на вероятность 1001-го.
- Количество столкновений в коллайдере конечно, значит Хиггс-бозонов не может быть больше, чем столкновений.
- Более практичный пример. Ваш экспериментальный анализ отбирает 50% всех событий. Допустим, всего было 10 событий. Сколько из них будет отобрано в результате анализа?
  - Это число подчиняется биномиальному распределению

Биномиальное распределение

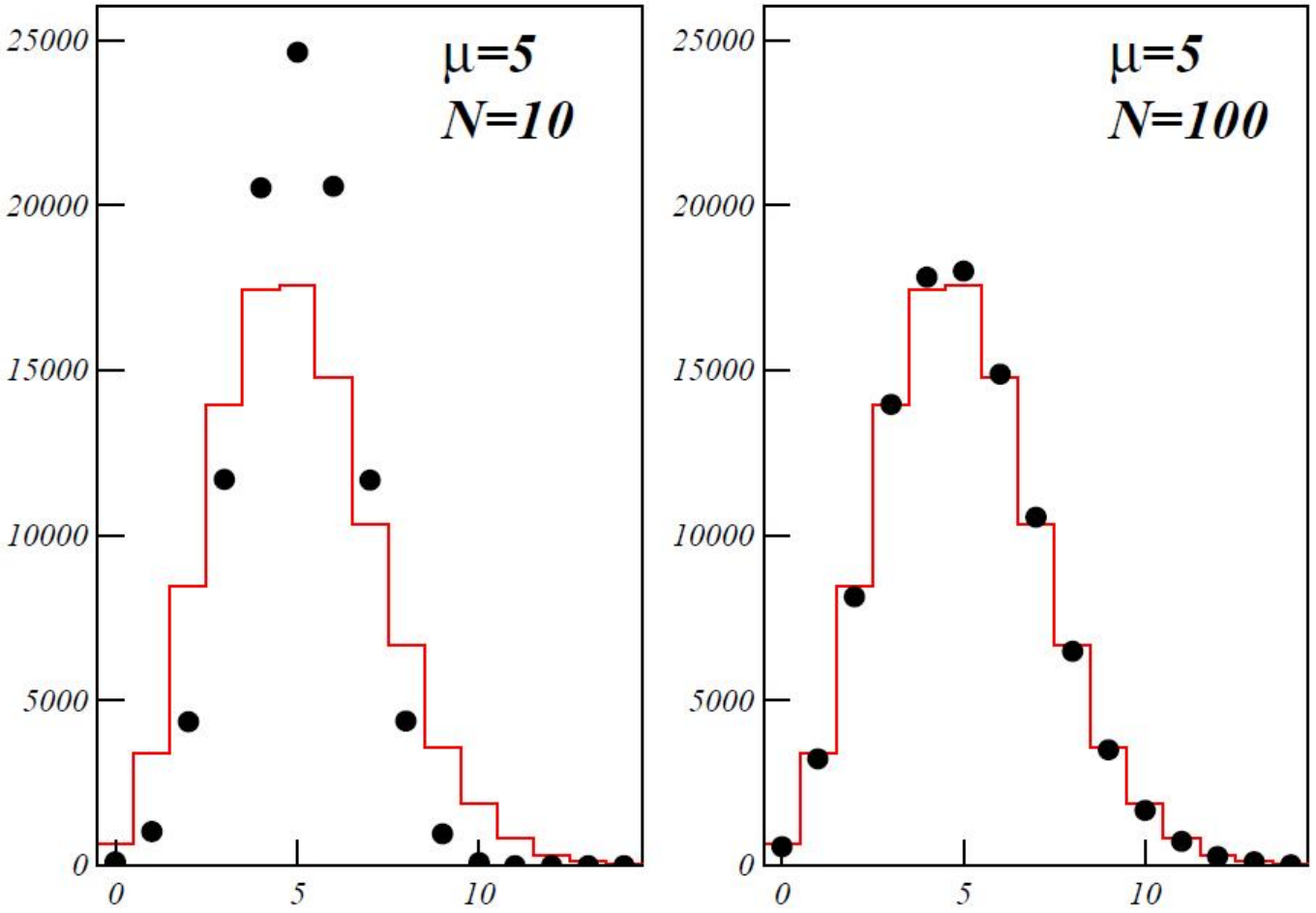
- Пусть в эксперименте возможны 2 исхода, вероятности p (благоприятный) и 1-p=q
- После N испытаний вероятность наблюдения n благоприятных исходов:

$$P(n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$\mu = E(n) = Np \qquad \sigma^2 = Npq = \mu(1-p)$$

- При $\mu << N$ переходит в распределение Пуассона

Распределения Пуассона и биномиальное



Применение биномиального распределения

- При отборе событий
- Пусть зарегистрировано N событий; после применения отбора их осталось n
- Число n имеет биномиальное распределение
- Эффективность отбора: $\varepsilon = n/N$
- Статистическая погрешность эффективности:

$$\sigma(\varepsilon) = \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}} = \frac{\sqrt{n(1-\varepsilon)}}{N}$$

Оценка погрешности эффективности «в уме»

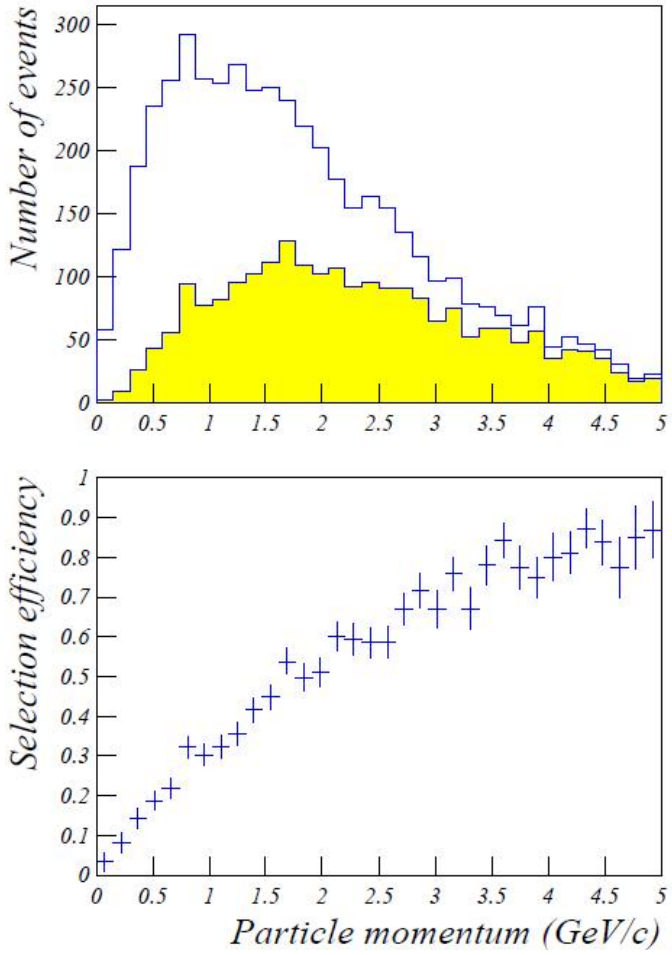
- Пусть отбор прошло n событий из общего числа N
- Если $n \ll N$, тогда $\sigma(n) \sim \sqrt{n}$
- Если $n \sim N$, тогда $\sigma(n) \sim \sqrt{(N-n)}$
  - результат симметричен относительно перестановки прошедших/непрошедших
- Если $n \sim N/2$, тогда $\sigma(n) \sim \sqrt{n}/\sqrt{2}$

Опросы общественного мнения

- Всем известны опросы общественного мнения: ответы 1600 респондентов, «статистическая погрешность не превышает 2.5%».
  - Если ответ выбрали 2.5%, то «число ответов не превышает статистической погрешности».
- По всей видимости, число 2.5% получено как $1/\sqrt{N} = 1/\sqrt{1600}$ (хотя при чём тут в данном случае 1600??)
- Из дисперсии биномиального распределения $\sigma^2=Npq$ мы легко оценим погрешность опроса.
- ***Максимальная*** погрешность достигается при ε=50%=800/1600: 50.00% ± 1.25%
- Если ответ выбрало 2.5% (40 респондентов), тогда ε =2.5% ± 0.4%, вполне надёжная оценка!
  - Кстати, если ε = 97.5%, погрешность такая же: ε = 97.5 ± 0.4%,

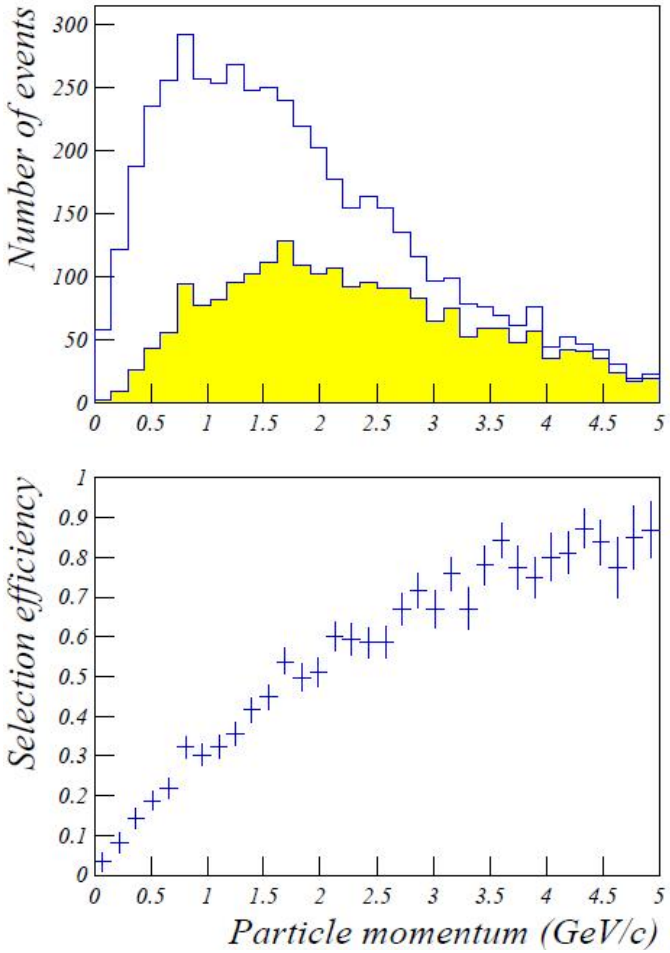
Эффективность отбора

- ε=(прошедшие отбор)/(все события)



Эффективность отбора

- **ε=(прошедшие отбор)/(все события)**
- Погрешность оценки эффективности отбора подчиняется дисперсии биномиального распределения
- Эффективность вычисляется не из анализируемых данных, а из независимого источника (например, моделирование Монте-Карло)



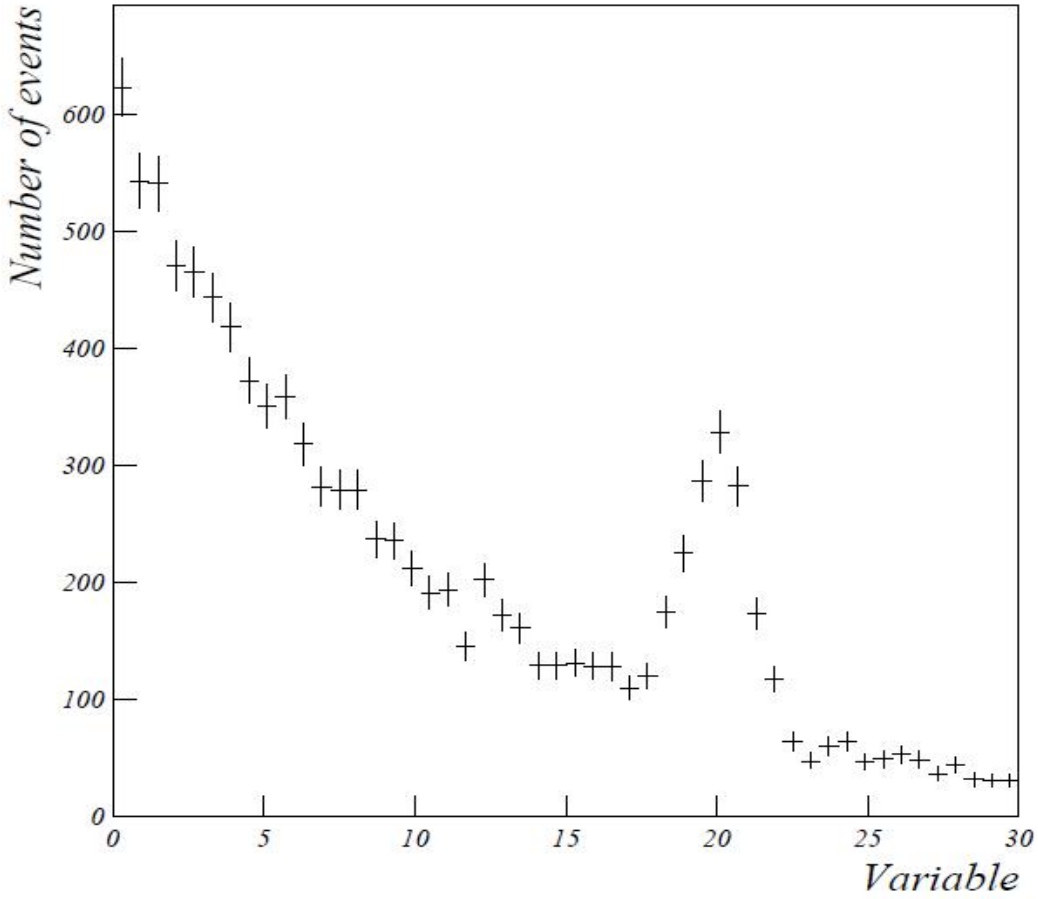
Погрешность числа событий

- Допустим, за 1 минуту родилось 100 Хиггс-бозонов. С какой погрешностью мы измерили поток Хиггс-бозонов?
- Согласно распределению Пуассона, $\hat{\mu} = n \pm \sqrt{n}$
- Таким образом, результат: 100±10 в минуту.
- А что если фон составляет 900 событий в минуту? Всего мы зарегистрировали 1000 событий, они тоже подчиняются распределению Пуассона: 1000±32.
- Если фон (900) известен идеально (без погрешности), тогда измеренный поток Хиггс-бозонов составляет 100±32.

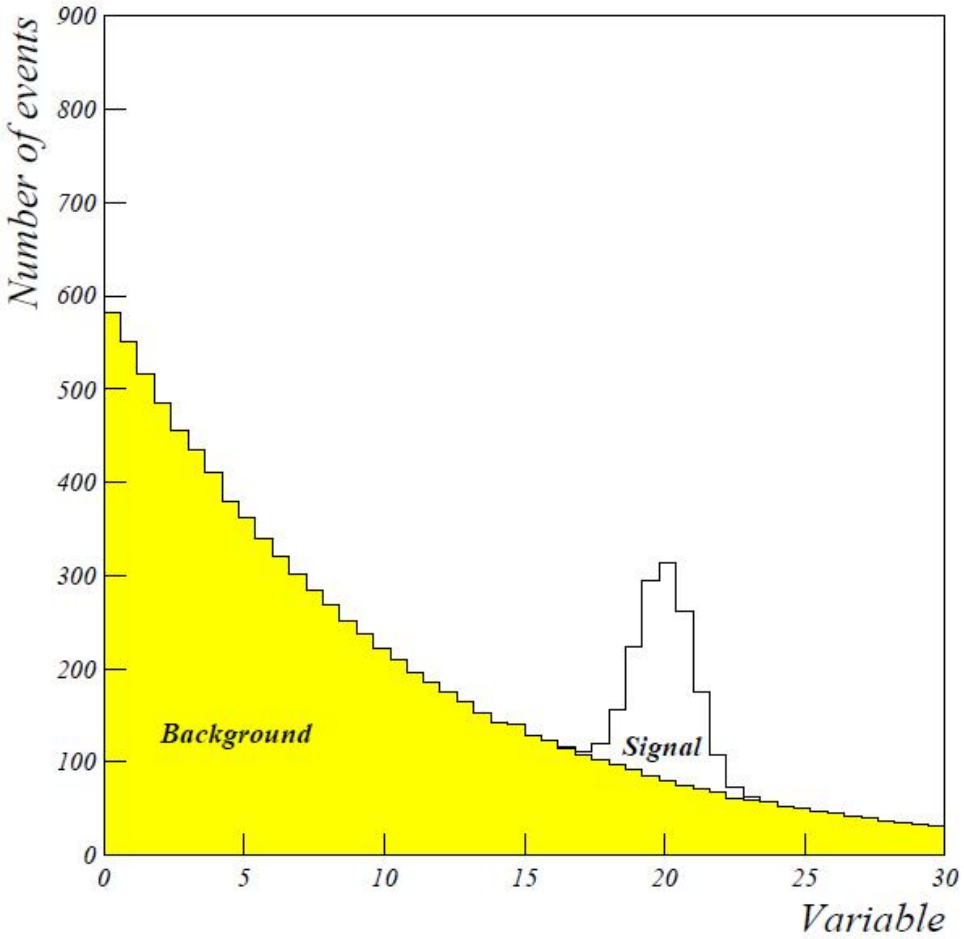
Как повысить точность?

- Итак, даже если фон известен идеально, само его наличие ухудшает точность измерений с 100±10 до 100±32.
- Можно ли ослабить влияние фона? И что делать, если сам фон неизвестен, или известен плохо?
- Как правило, решение существует, но для этого нам понадобится Монте-Карло
  - Метод Монте-Карло – это метод математического моделирования, позволяющий генерировать такие же события, как и в экспериментальных данных.
  - Смысл метода в том, что (в отличие от реальных данных) для каждого моделированного события известно, является ли это событие сигналом или фоном.

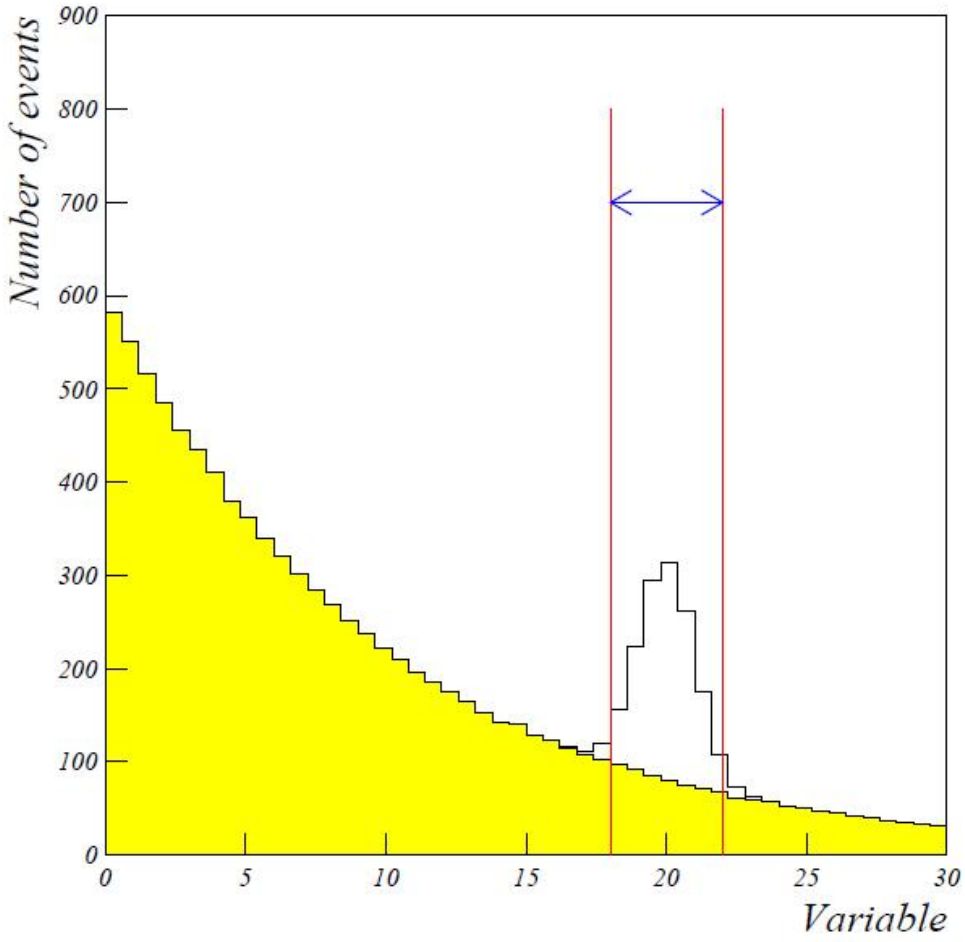
Измеренное распределение



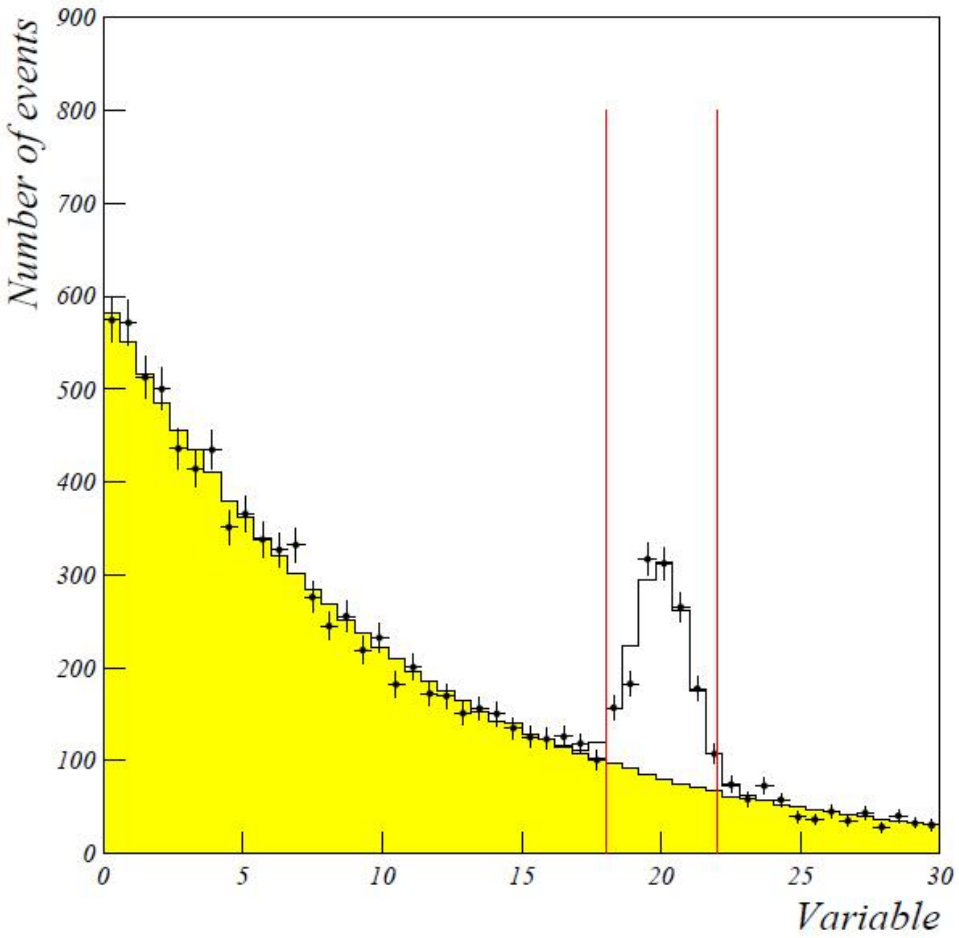
А что нам скажет Монте-Карло?



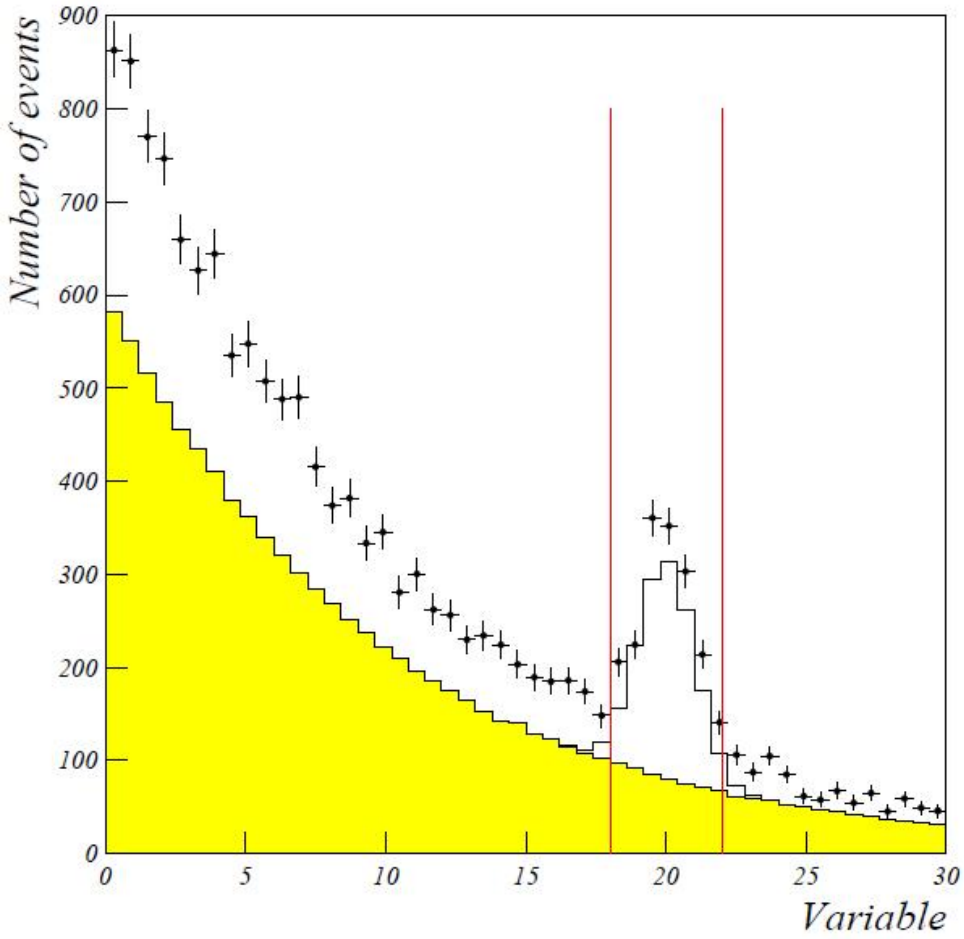
Выберем область, где много сигнала и мало фона…



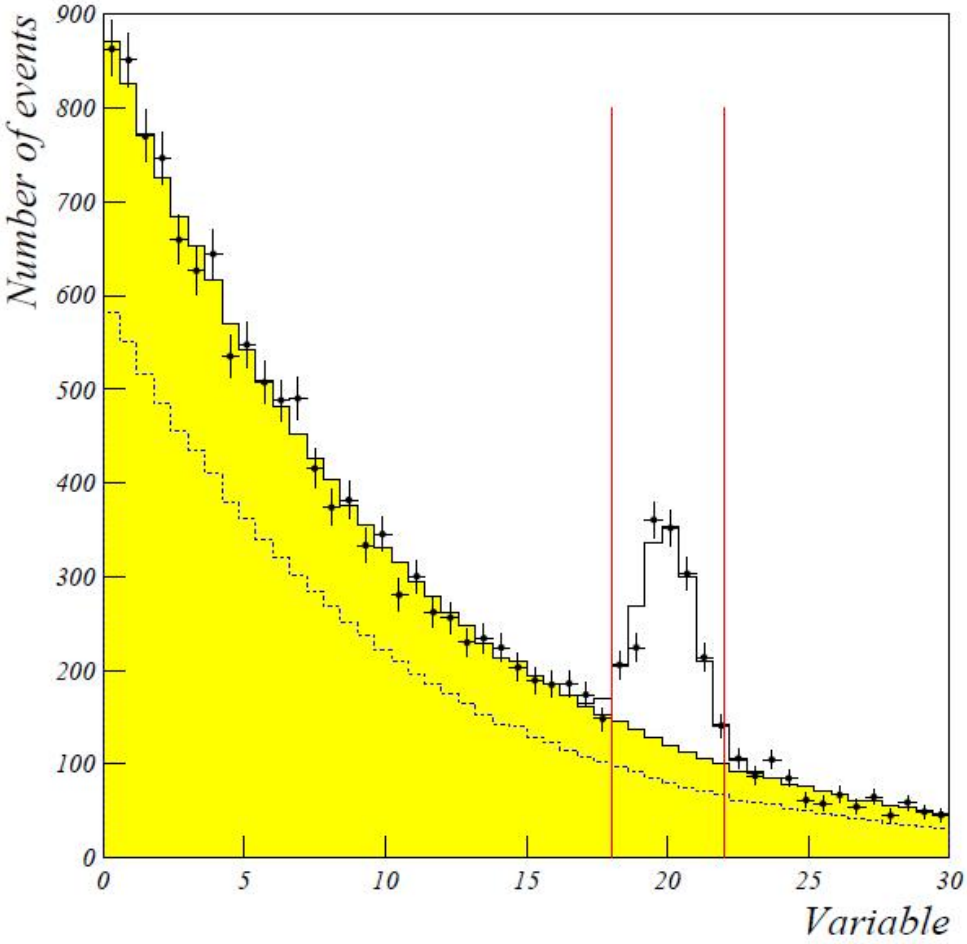
Из данных в этом окошке вычтем смоделированный фон



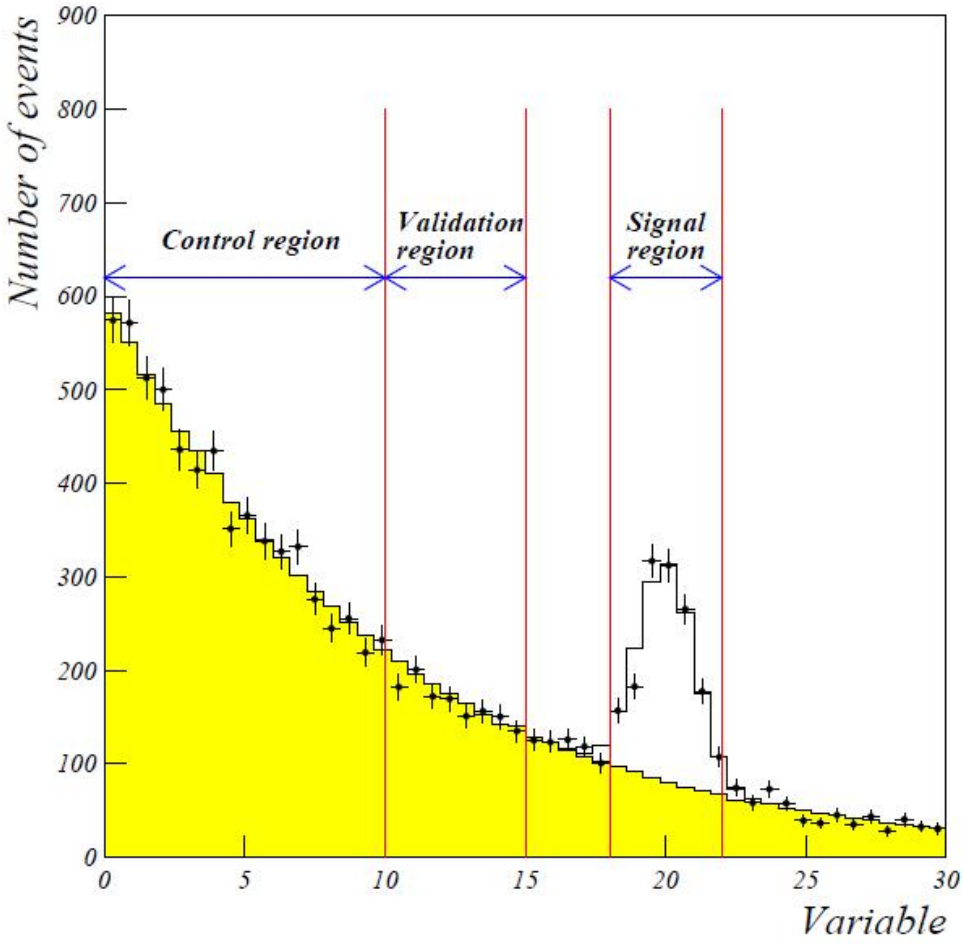
А что делать, если фон смоделирован плохо?



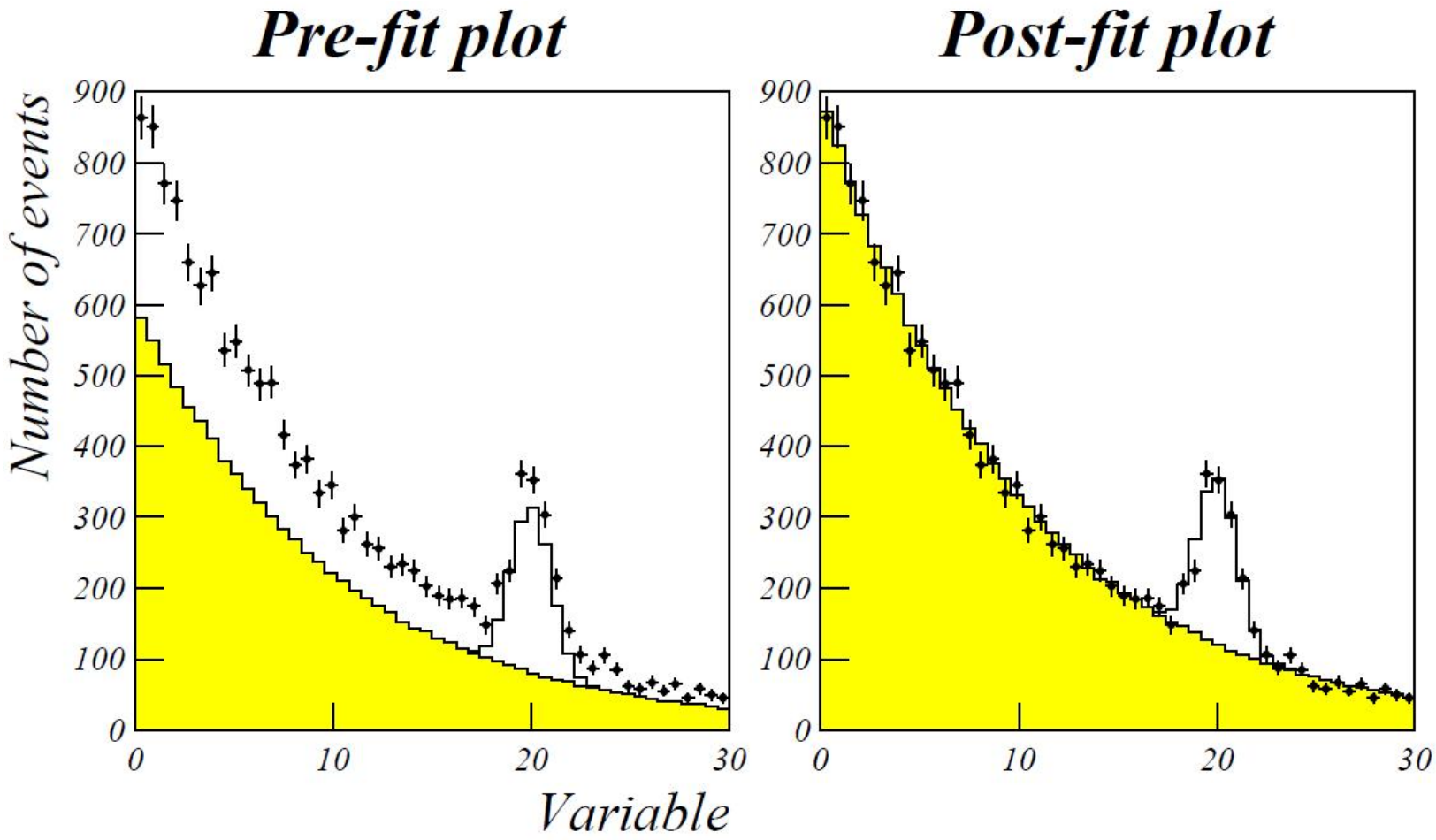
Подгоним (отфитируем) фон под реальные данные



Области: сигнальная, контрольная, проверочная



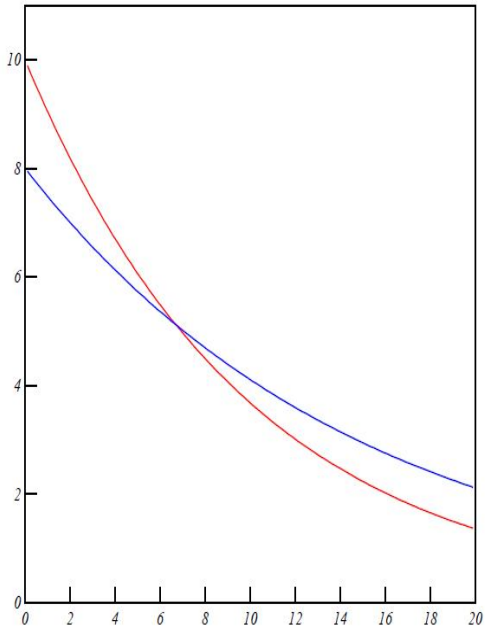
Пре-фит и пост-фит



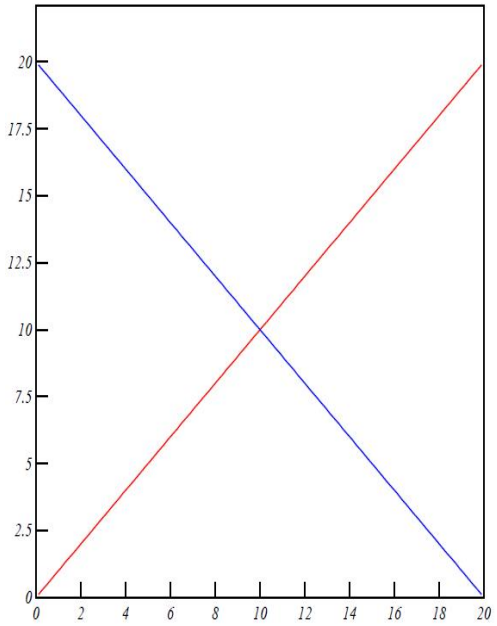
Что нам даёт Монте-Карло

- Итак, моделирование методом Монте-Карло позволяет нам:
  - Определить область с максимальным соотношением сигнал/фон
  - Предсказать количество фоновых событий, чтобы вычесть их из данных
  - При помощи data-driven анализа измерить фон прямо из реальных данных, если предсказание теории имеет плохую точность
  - Определить эффективность отбора сигнала (вспоминаем биномиальное распределение)

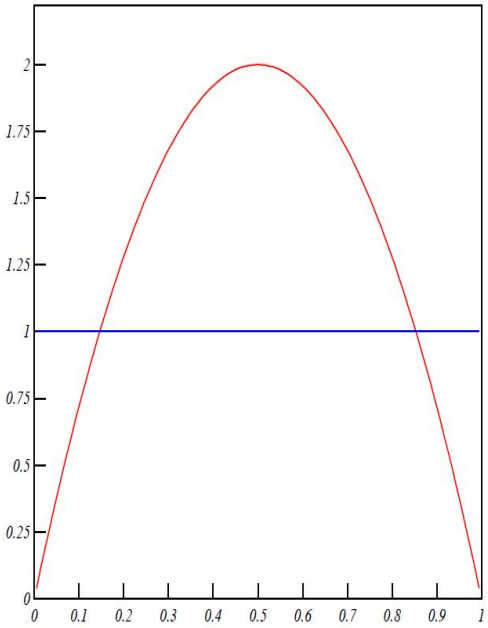
Примеры разделения сигнала и фона



Кривые распада двух нуклидов



Угол распада частиц со спином +1/2 и -1/2



Угол распада частиц со спином 0 и 1

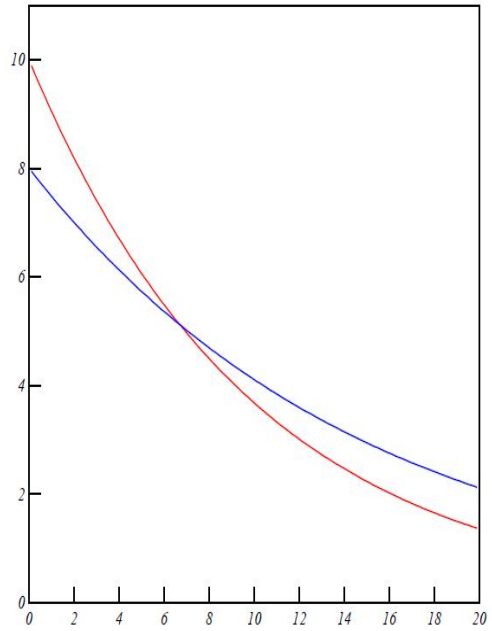
Разделительная способность

- Чем сильнее отличаются распределения сигнала и фона, тем с большей точностью можно измерить сигнал, даже не зная величину фона
- А какова числовая мера понятия «распределения отличаются»?
- Это «разделительная способность» (Separation Power)

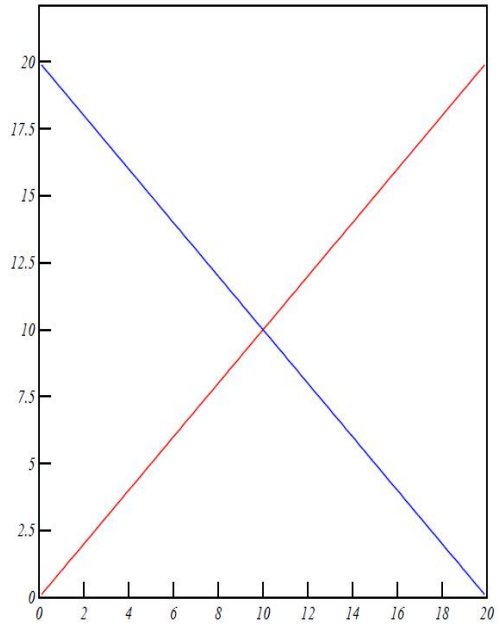
$$SP = \frac{1}{2}\sum \frac{(s_i - b_i)^2}{(s_i + b_i)} \qquad s_i = S_i / \sum S_i \quad b_i = B_i / \sum B_i$$

- SP=0: сигнал и фон идентичны по форме, измерение невозможно (разве только вычесть весь фон, если он известен).
- SP=1.0: идеальное разделение, сигнал и фон нигде не пересекаются. Сигнал измеряется с точностью $N \pm \sqrt{N}$
- Если сигнал и фон примерно равны, то сигнал можно измерить с точностью $N \pm \sqrt{(N/SP)}$

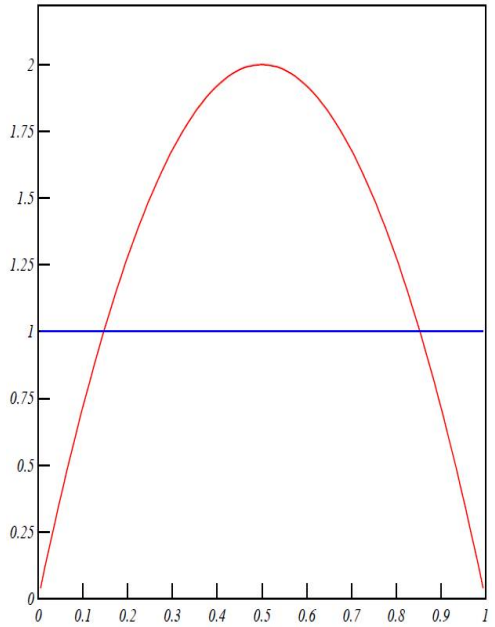
Примеры разделительной способности



SP=0.008

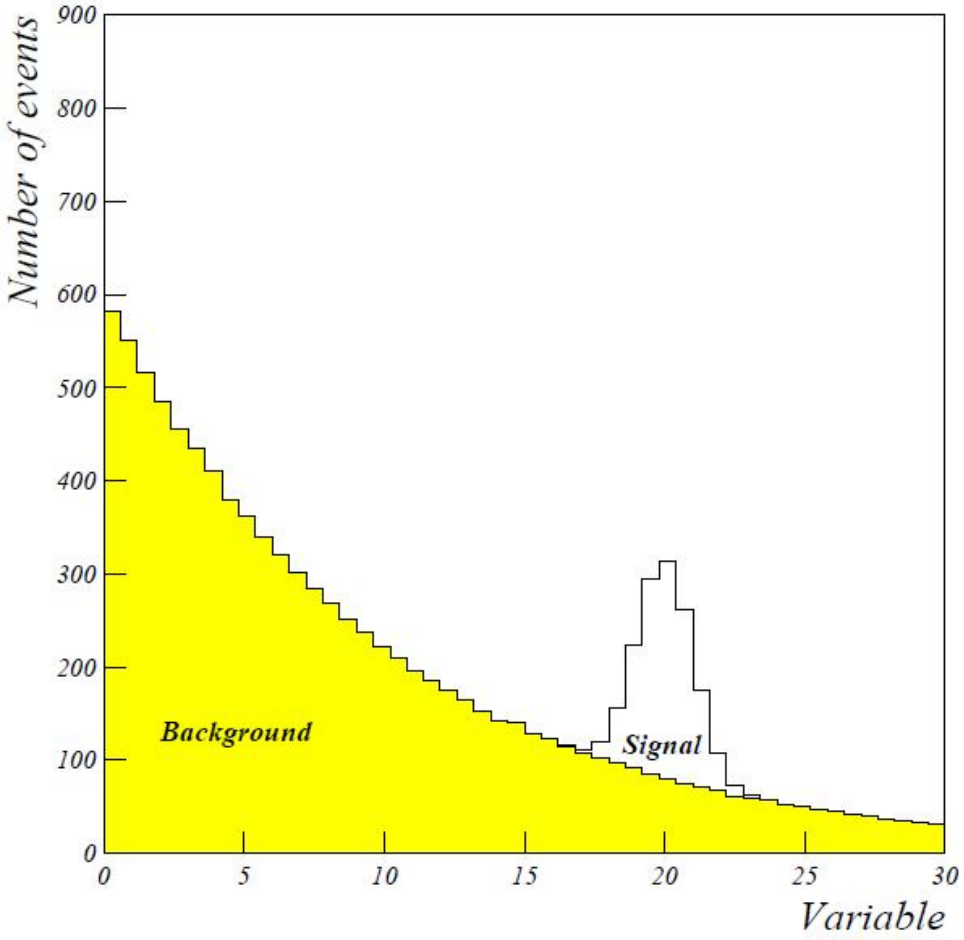


SP=0.3333



SP = 0.07

«Наш» пример: пик над фоном



- SP = 0.86
- Сигнал можно измерить почти с такой же точностью, как если бы фона не было
- Не удивительно: когда сигнал это узкий пик, влияние фона всегда минимально