Прикладная статистика. Занятие 1

Денис Деркач, Влад Белавин

23 октября 2018 года

Nature description and statistics

Since Pierre-Simon Laplace's times (1749–1827) the universe's fate was deterministic and, in spite of technical difficulties, was considered predictible if the complete equation of state were known.

Challenged by Heisenberg's uncertainty principle (1927), Albert Einstein proclaimed "Gott würfelt nicht" ("God does not play dice"), but so-called hidden variables to bring back determinism through the back door into the quantum world were never found.

In quantum mechanics, particles are represented by wave functions. The size of the wave function gives the probability that the particle will be found in a given position. The rate, at which the wave function varies from point to point, gives the speed of the particle.

Quantum phenomena like particle reactions occur according to certain probabilities. We use probabilistic "Monte Carlo" techniques to simulate event-by-event realisations of quantum probabilities.

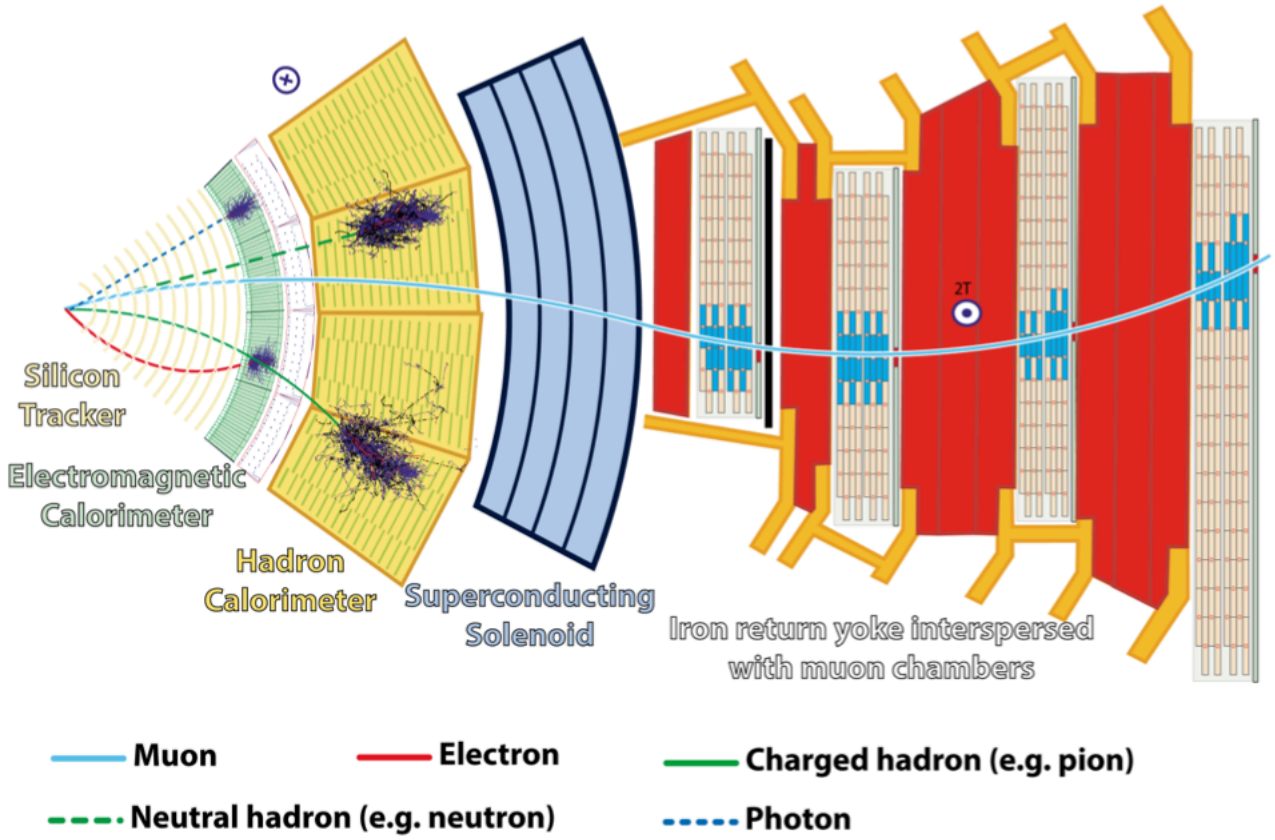

Pierre-Simon de Laplace


Albert Einstein


Werner Heisenberg

Large system fluctuations

In addition to the intrinsic probabilistic character of natural phenomena, the measurement process through the interaction contributes statistical degrees of freedom leading to measurement errors and to genuine systematic effects (eg, detector misalignment), that need to be considered in the statistical analyses.



Probability and statistics are fundamental ingredients & tools in all modern sciences

“Statistics is obvious, so I prefer not to read the literature and just figure it out for myself.”

Noname contributor

Random variable

Generally, we have a set of outcomes of our experiment, we need a function that maps from set of outcomes to real numbers: a random variable. In some example random variables are implicitly used:

Examples of random variables

Experiment	Random variable
Toss two dice	X = sum of the numbers
Toss a coin 25 times	X = number of heads in 25 tosses
Apply different amounts of fertilizer to corn plants	X = yield/acre

Cumulative Density Function

Generally, we are interested in some probabilities connected to the random variable.

We thus introduce the CDF:

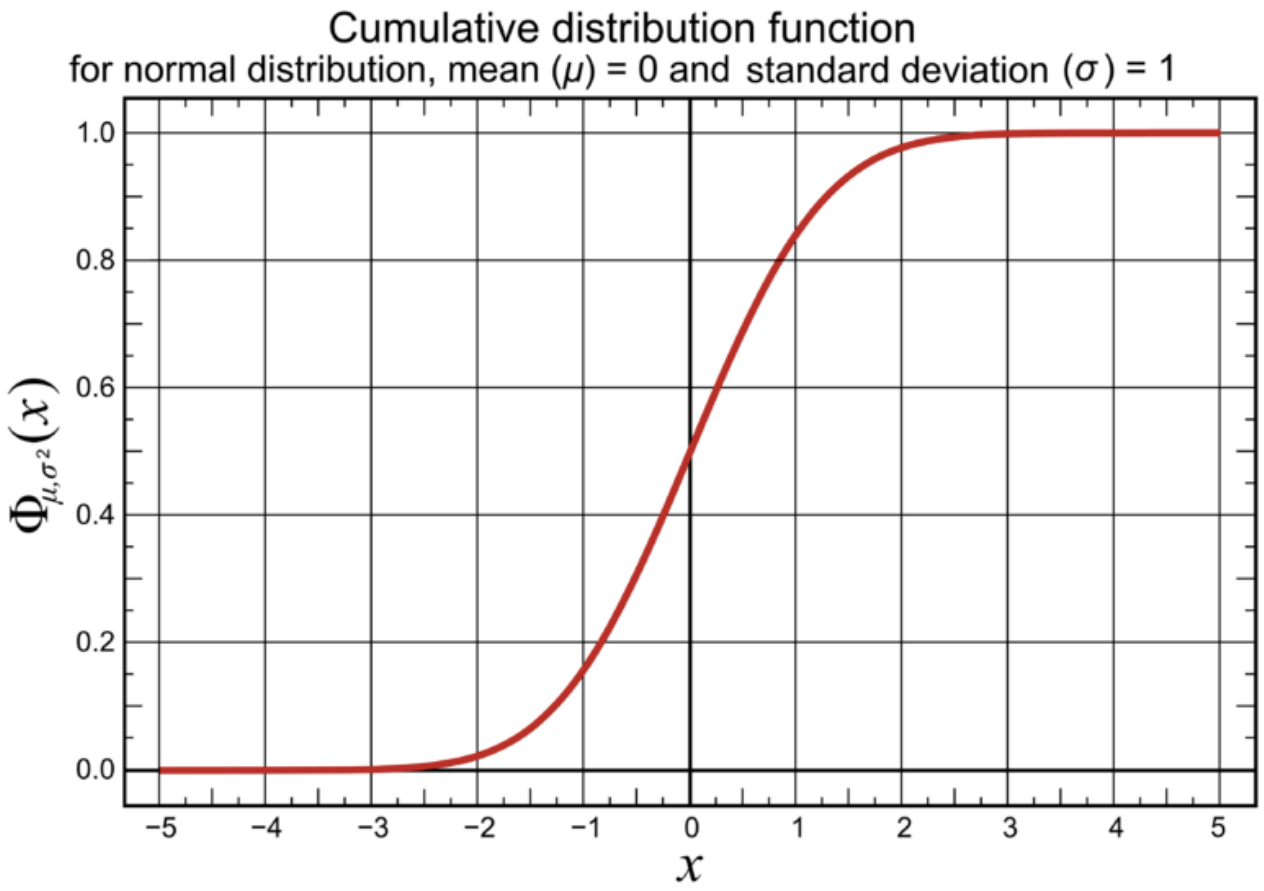
$$F_X(x) = \mathrm{P}(X \leq x)$$

Connected to the probability:

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$$

And probability density functions:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$$



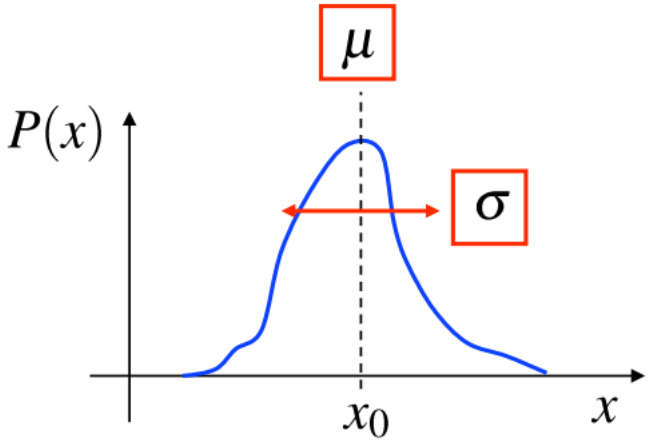
Probability Distributions

Suppose we are trying to measure some quantity with true value the result of a single measurement follows a probability density function (**PDF**) which may or may not be of a known form.

We can define a few important measures of the PDF

Mean: $$\mu = \mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx$$

Variance: $$\mathrm{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\, dx$$



NB more moments exist: Skewness ~ E(X-mu)^3, Kurtosis ~ E(X-mu)^4

Mean and Variance Properties

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx$$

$$\operatorname{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\, dx$$

- E[c] = c
- E[cX] = cE[X]
- E[X+c] = E[X] + c

Var[c] = 0
Var[cX] = c^2Var[X]
Var[X+c] = Var[X]

Experimental Science Measurements

Experimental science concerned with two types of experimental measurement:

- Measurement of a quantity : **parameter estimation**
- Tests of a theory/model : **hypothesis testing**

* Eventually, we want to check the model, either directly or indirectly.

Parameter Estimation

For **parameter estimation** we usually have some data (a set of measurements) from which we want to obtain:

- The best estimate of the true parameter; **“the measured value”**
- The best estimate of how well we have measured the parameter; **”the uncertainty”**

Inferential Statistics

Generally, we have a set of outcomes of our experiment, we need a function that maps from set of outcomes to real numbers: a random variable.

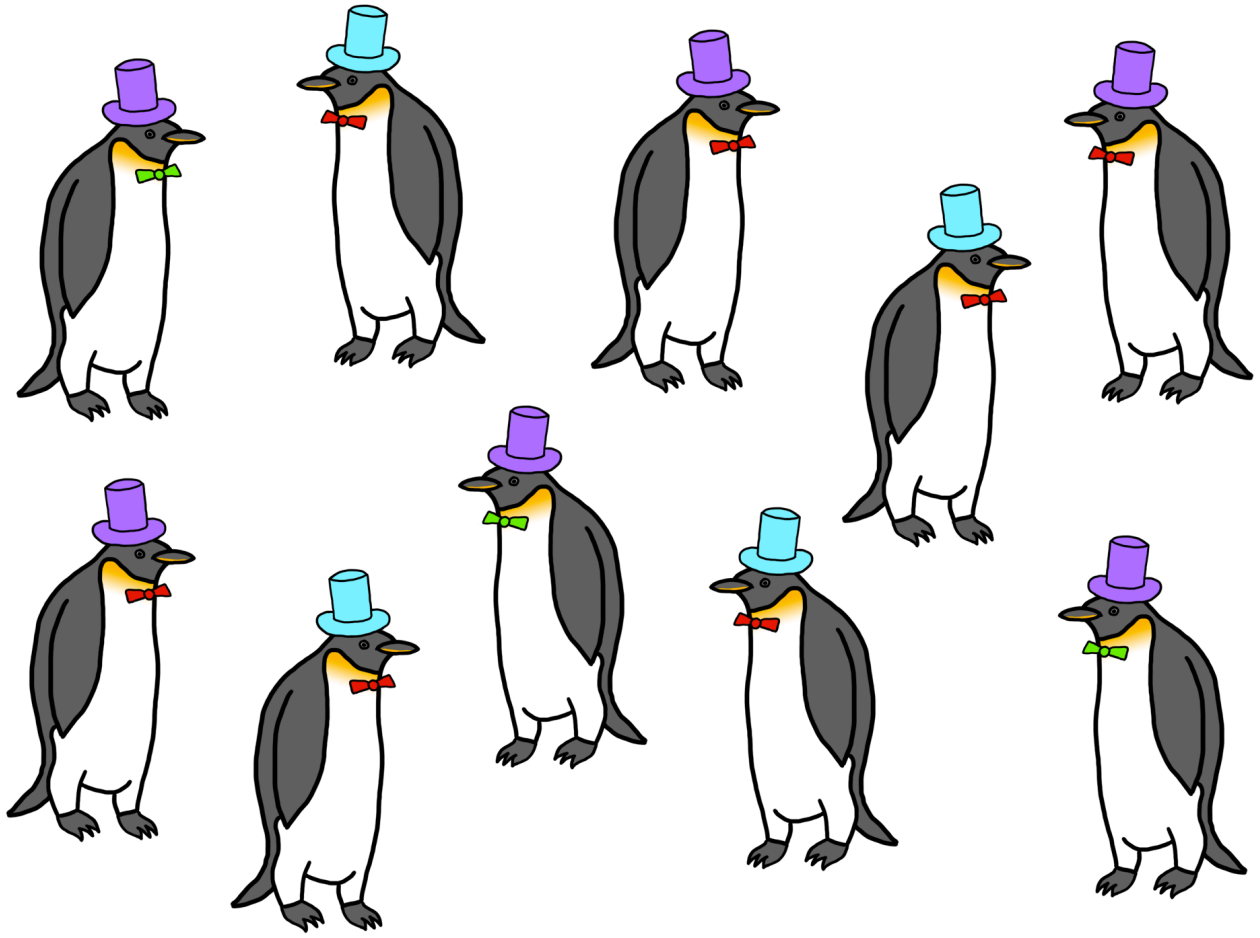
Statistical Inference

The goal of inference is to be able to make a statement about something that is *not observed*, and ideally to be able to characterize any uncertainty you have about that statement. Three main two things are:

- Define the population.
- Define the sampling process.
- Define the model for the population.

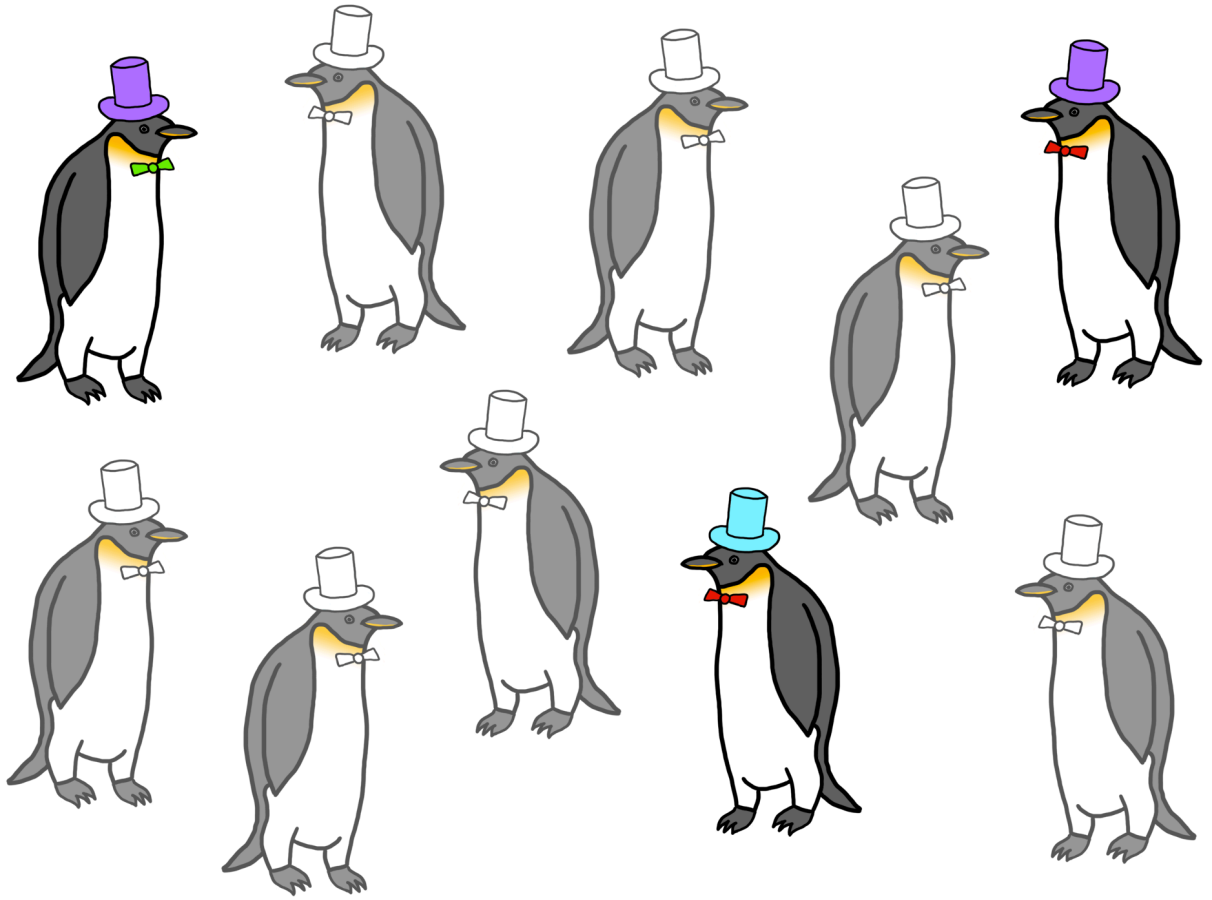
Example of inference

Consider this group of penguins, each wearing either a purple or turquoise hat. There are a total of 10 penguins in this group. We'll call them the *population*.



Example of inference

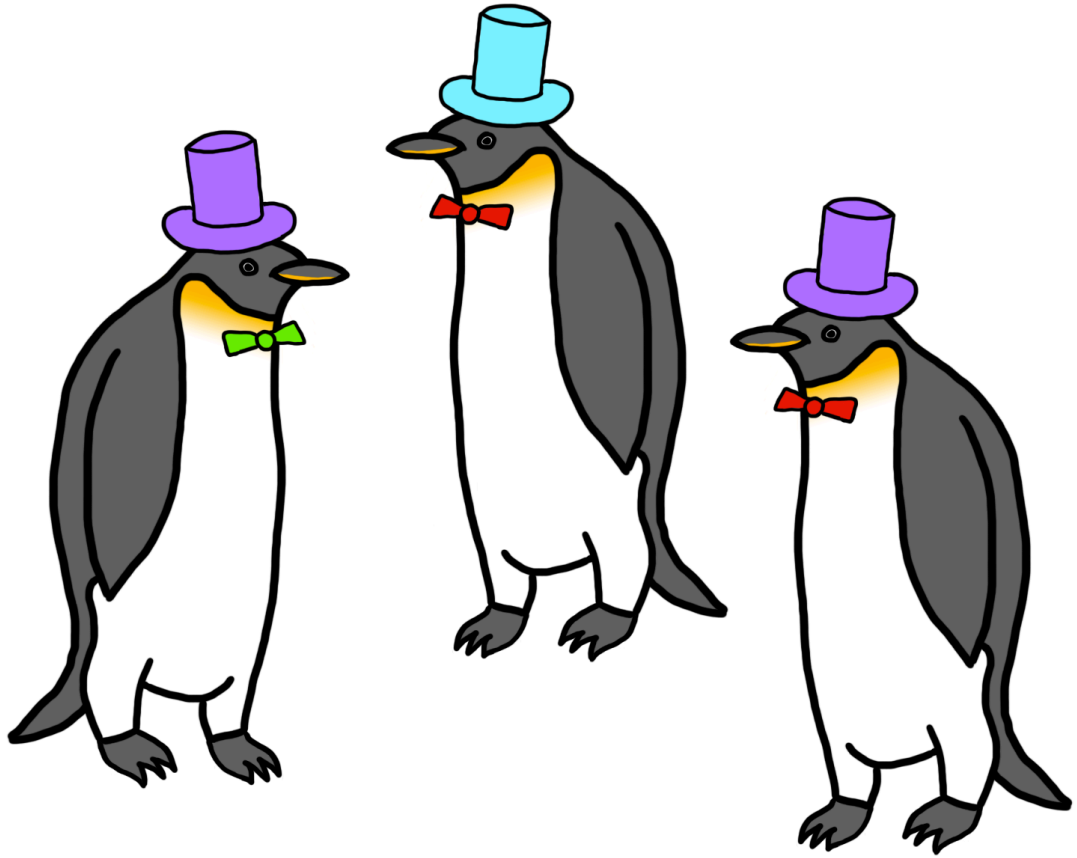
You want to study the population, but due to external constraints, you can only study three birds.



Example of inference

Now, you have three birds. What can you say about their hats?

1/3 is wearing a turquoise hat.



Inference: what can go wrong

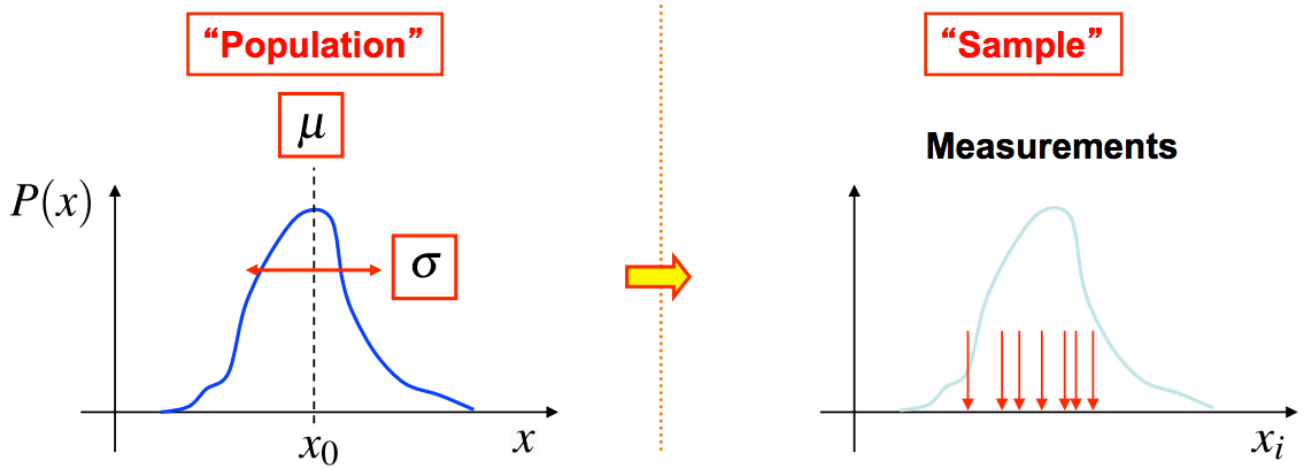
- **selection bias** − violation of one of three stages of inference
- **sampling variability** − sample is not big enough to make precise inference (just increase the sample!)

Estimating the Mean and Variance

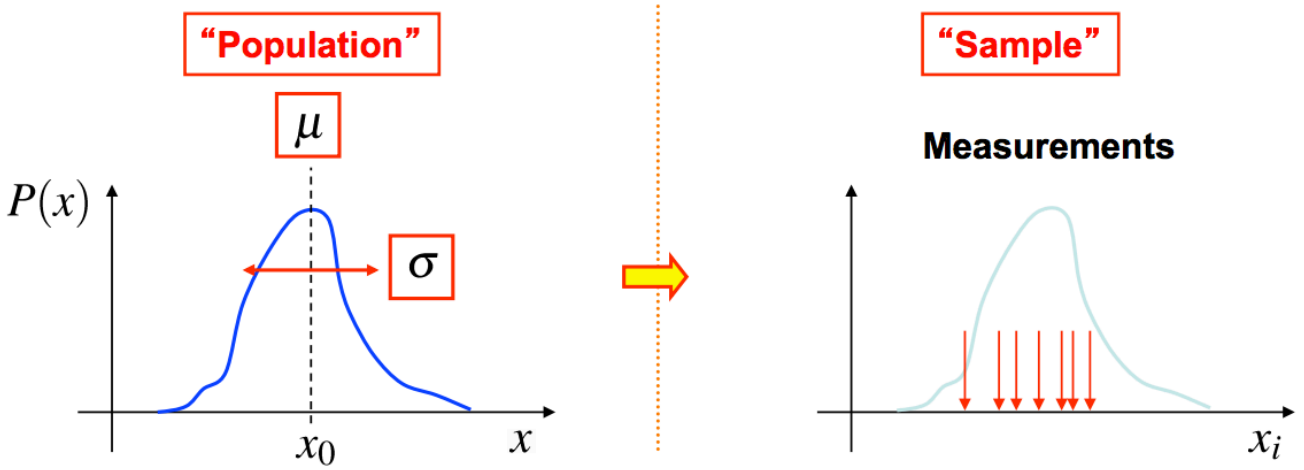
In general do not know the PDF – instead have a number of measurements distributed according to the PDF.

Unless one has a infinite number of measurements cannot fully reconstruct the PDF (not a particularly useful thing to do anyway).

But can obtain unbiased estimates of the mean and variance.



Estimating the Mean and Variance

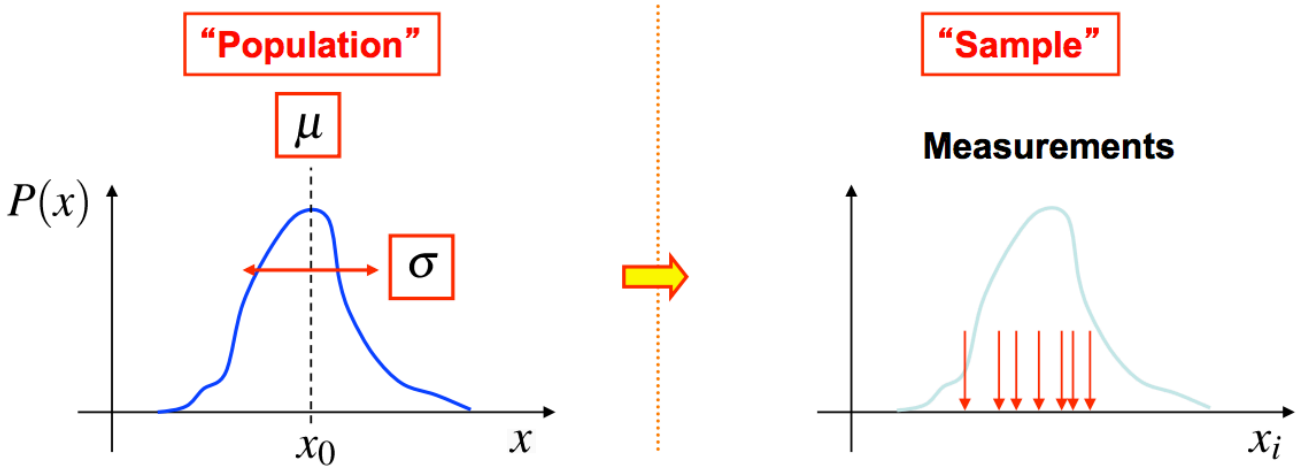


$$\mu = \mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx$$

$$\mathrm{Var}(X) = \sigma^2 = \int (x - \mu)^2 f(x)\, dx$$

For independent identically distributed random variables (iid)

Estimating the Mean and Variance



$$\mu = \mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\,dx$$

$$\mathrm{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\,dx$$

$$\overline{X} = \frac{\sum X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

NB while S^2 is unbiased estimator of sigma^2, S is biased estimator of sigma (bias = sigma/4n)

Three Types of “Errors”

- **Statistical** Uncertainties:
  - Random fluctuations
    - e.g. shot noise, measuring small currents, how many electrons arrive in a fixed time
    - Tossing a coin N times, how many heads
- **Systematic** Uncertainties:
  - Biases
    - e.g. energy calibration wrong
    - Thermal expansion of measuring devices
    - Imperfect theoretical predications
- Blunders, i.e. **errors**:
  - Mistakes
    - Forgot to include a particular background in analysis
    - Bugs in analysis code

Uncertainty of mean

What is the standard error (i.e. square root of the variance) on the sample mean?

$$Var(\bar{x}) \equiv \sigma_{\bar{x}}^2 \quad = \quad \langle (\bar{x} - \mu)^2 \rangle = \frac{\sigma^2}{n}$$

Hence the uncertainty on the mean is smaller than the uncertainty on a single measurement

Note: this is general result – doesn't rely on distribution

Uncertainty of variance

What is the standard error (i.e. square root of the variance) on the sample variance?

$$Var(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$$

where $\mu_j = E(X-\mu)^j$

NB1: this is general result – doesn't rely on distribution
NB2: constructing an unbiased estimate is somewhat long expression (see for example https://modelingwithdata.org/pdfs/moments.pdf)

Errors and Confidence

- Within ±1σ: "1σ Confidence Level",
  or "68.27% Confidence level"

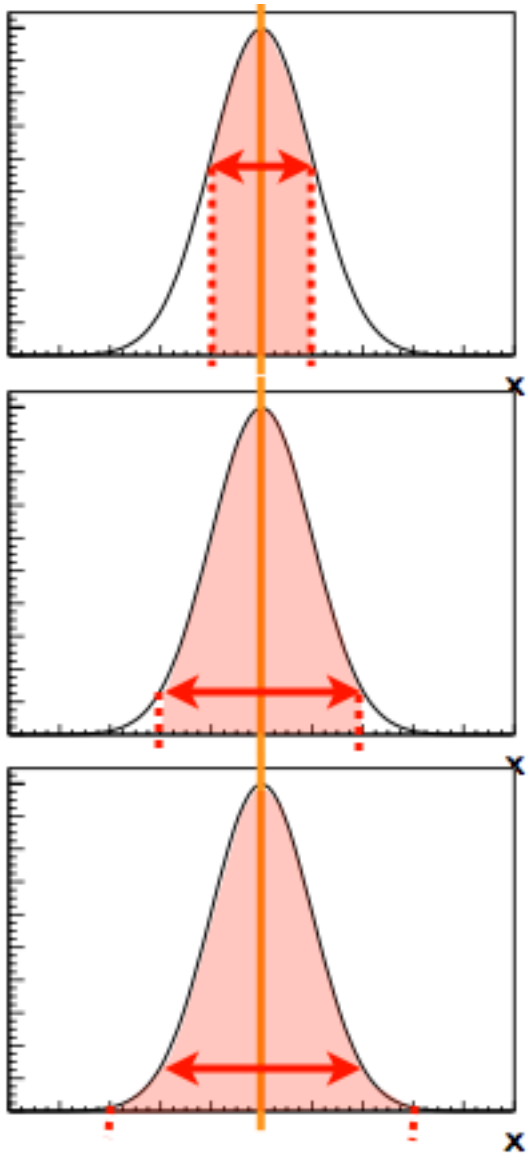
$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 68.27\%$$

- Within ±2σ: "2σ CL" or "95.45% CL"

$$\int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 95.45\%$$

- Within ±3σ: "3σ" or "99.73% CL"

$$\int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 99.73\%$$



Talking to engineers

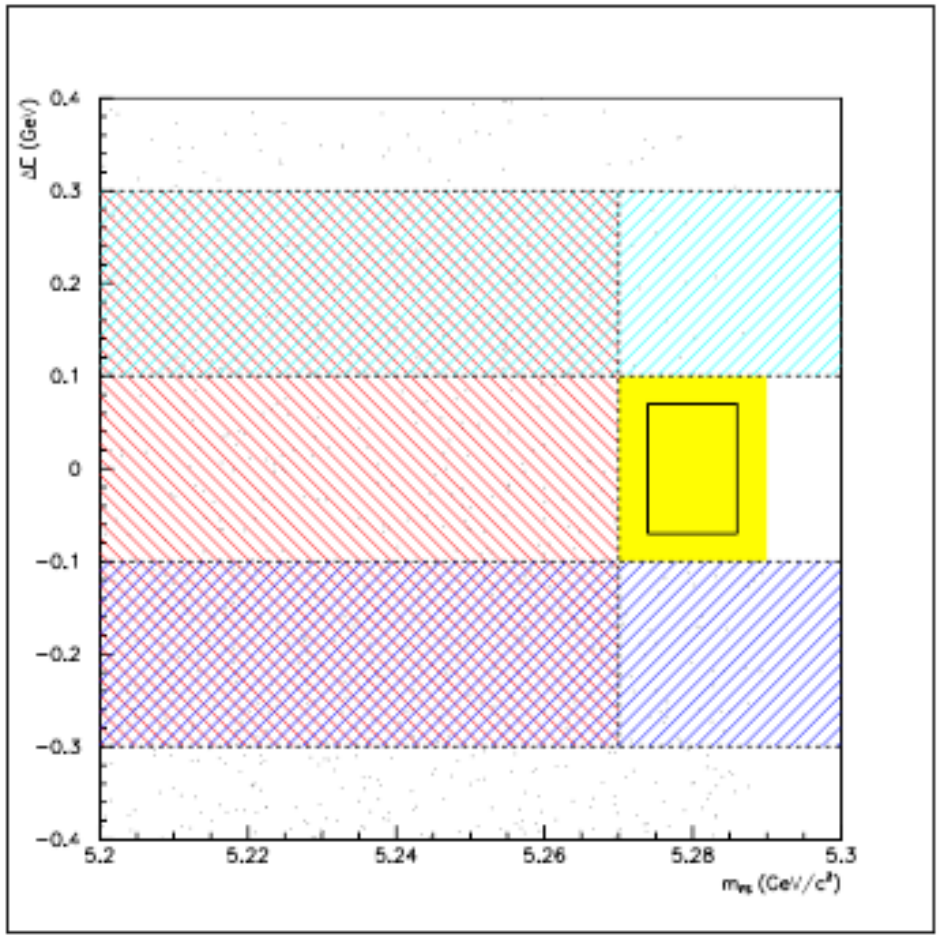
- Physicists quote their errors as 1σ (Gaussian) confidence intervals.
- The probability that a result is outside the quoted error is 32%. About 1/3 of measurements should be outside the error bars. Results outside error bars are OK - it just shouldn't happen too often. And it shouldn't be too far: *P(outside μ±2σ) ~5%, P(outside μ±3σ) ~0.3%*)
- Engineers *guarantee* that the actual value is within mean ± tolerance.

Cool Hand Luke (1967)

"What we've got here is...failure to communicate.

Some men you just can't reach."

Cut-and-Count analysis

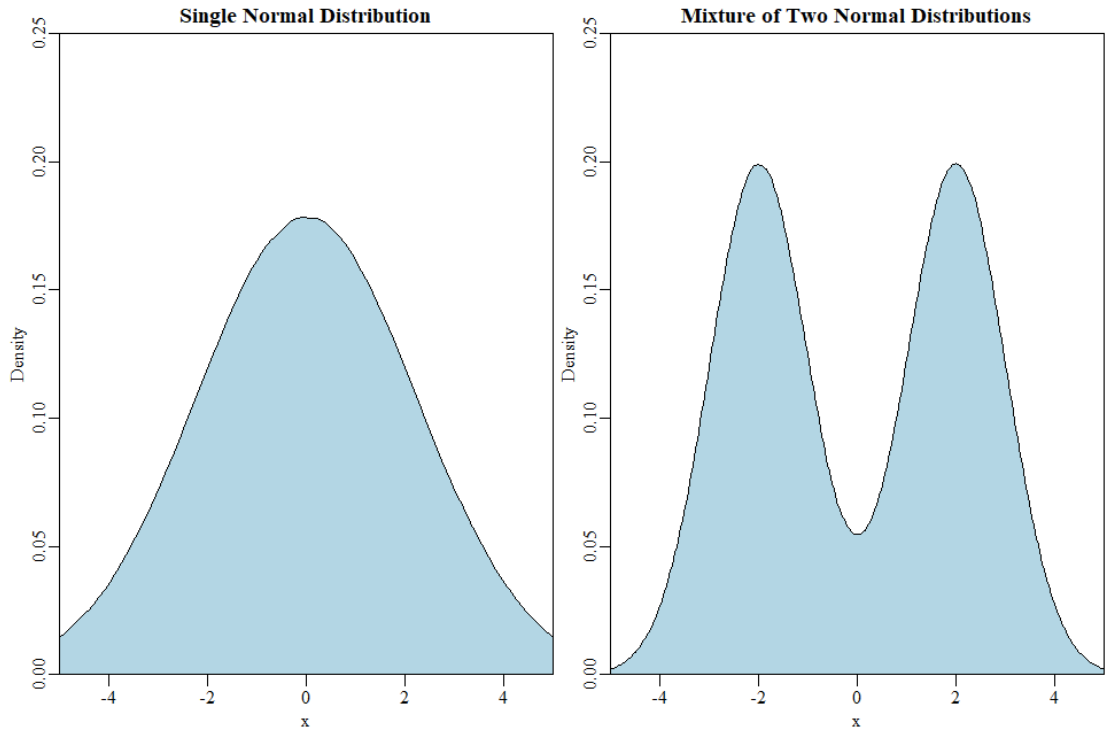


We thus can already perform an analysis, using these simple estimators.
Imagine zero-background experiment and a theory that predicts an average value.

Dangers

Mean is extremely sensitive to outliers

But not to the shape of distributions



Statistical distributions

Measurement results typically follow some “distribution”, ie, the data do not appear at fixed values, but are “spread out” in a characteristic way.

- Which type of distribution it follows depends on the particular case. It is important to know the occurring distributions to be able to pick the correct one when interpreting the data (example: *Poisson* vs. *Compound Poisson*)
- ...and it is important to know their characteristics to extract the correct information

Binomial distribution

How often (likely) is $\boldsymbol{k}\times\mathbf{head}$ and $(\boldsymbol{N}-\boldsymbol{k})\times\mathbf{tail}$?

- Each coin: $P(\mathrm{head}) = p, P(\mathrm{tail}) = 1-p$
- Pick $\boldsymbol{k}$ particular coins → the probability of all having **head** is:

$$P(k\times\mathrm{head}) = P(\mathrm{head})\cdot P(\mathrm{head})\cdot\cdots\cdot P(\mathrm{head}) = P(\mathrm{head})^k = p^k$$

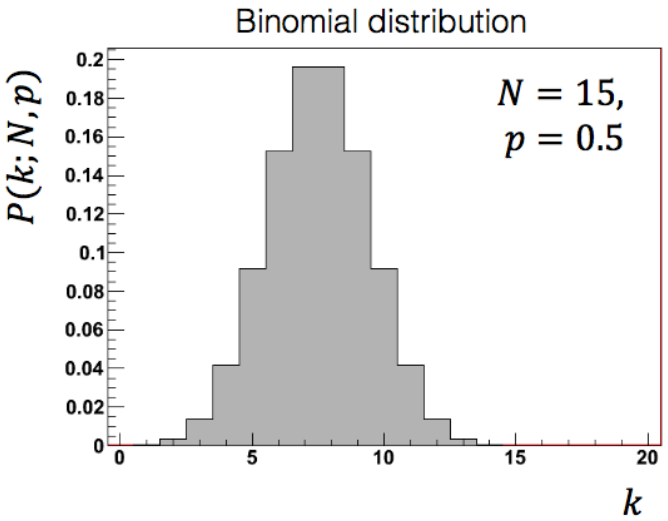
- Multiply this by the probability that all remaining $\boldsymbol{N}$–$\boldsymbol{k}$ coins land on **tail**:

$$P(\mathrm{head})^k \cdot P(\mathrm{tail})^{N-k} = p^k\ (1-p)^{N-k}$$

- This was for a particular choice of $\boldsymbol{k}$ coins
- Now include all $\binom{N}{k}$ permutations for *any* $\boldsymbol{k}$ coins

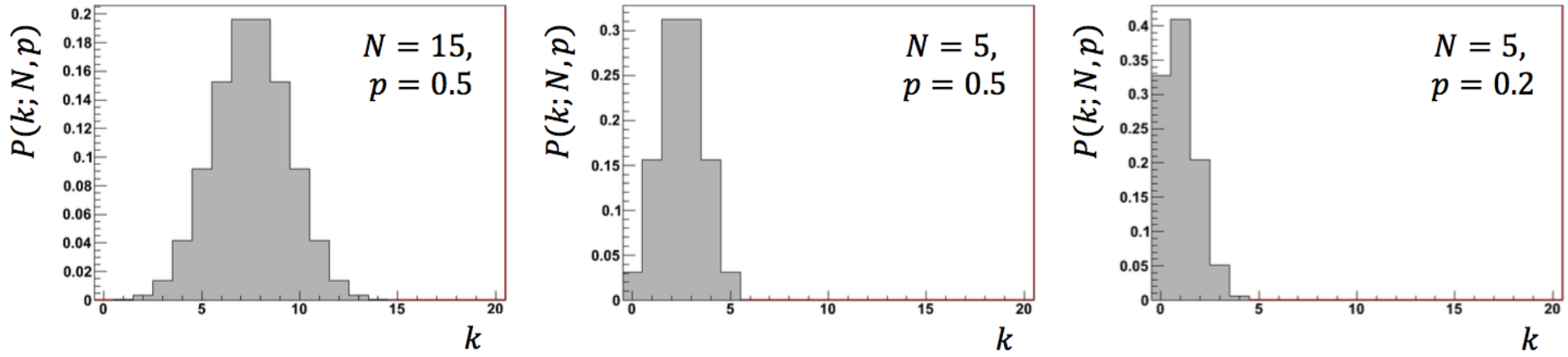
$$P(k;N,p) = p^k\ (1-p)^{N-k}\binom{N}{k}$$

where $\binom{N}{k} = \frac{k!}{k!(N-k)!}$ is the binomial coefficient



Binomial distribution

Example binomial distributions:



- *Expectation value*: sum over all possible outcomes and *average* (i.e.: weighted average)

$$\mathrm{E}[k] = \sum_k kP(k;N,p) = Np$$

- *Variance:*

$$V[k] = Np(1-p)$$

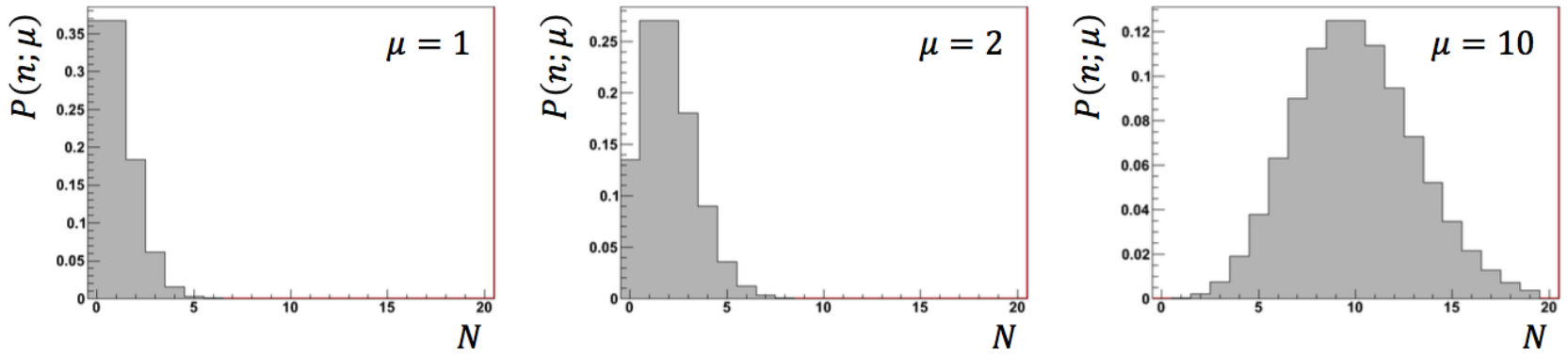
Poisson distribution

Recall: for individual events each with two possible outcomes → Binomial distribution

How about: number of counts in radioactive decay experiment during given time interval $\boldsymbol{\Delta t}$?

- Events happen "randomly" but there is no such 2nd outcome; $\boldsymbol{\Delta t}$ is continuous, no discrete num. of trials
- $\boldsymbol{\mu}$: average number of counts in $\boldsymbol{\Delta t}$. What is the probability of observing $\boldsymbol{N}$ counts?
- Limit of Binomial distribution for large number of trials and small p: $N \to \infty$ & $p \to 0$ so that $Np \to \mu$.

→ Poisson distribution: $P(N;\mu) = \frac{\mu^N}{N!} e^{-\mu}$



Expectation value: $E[N] = \sum_N N \cdot P(N;\mu) = \mu$, Variance: $V[N] = \mu$

Poisson is good approximation for Binomial distribution for $N \gg \mu\ (= Np)$

Gaussian distribution

In limit of large $\boldsymbol{\mu}$ a Poisson distribution approaches a symmetric Gaussian distribution

- This is the case not only for the Poisson distributions, but for almost any sufficiently large sum of samples with different sub-properties (mean & variance) → **Central Limit Theorem** (will discuss later)
- Gaussian distribution is of utter use, and luckily has simple properties

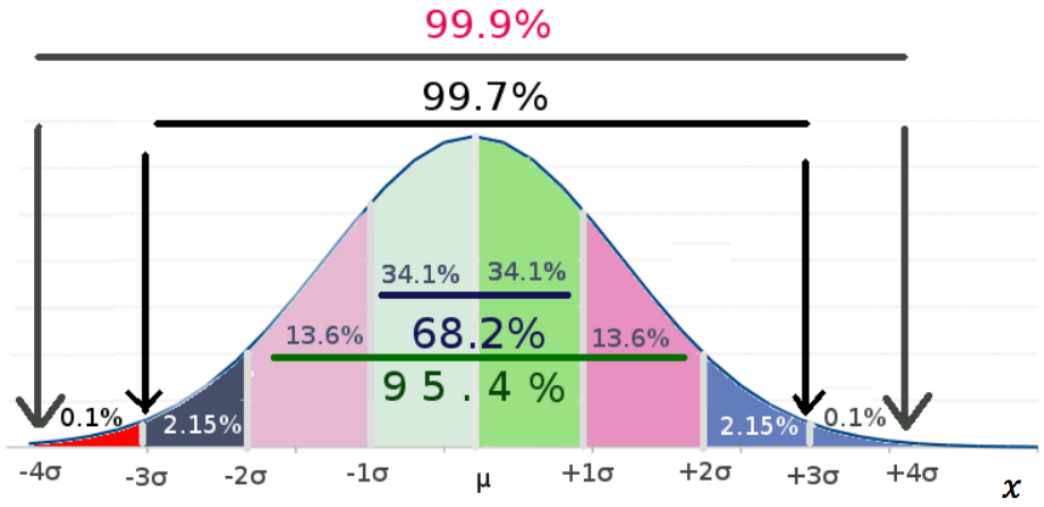
→ **Gauss distribution:** $P(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Symmetric distribution:

- Expectation value: $E[x] = \mu$
- Variance: $V[x] = \sigma^2$
- Probability content:

$$\int_{-\sigma}^{+\sigma} P(x;\mu,\sigma)dx = 68.2\%$$

$$\int_{-2\sigma}^{+2\sigma} P(x;\mu,\sigma)dx = 95.4\%$$



Gaussian distribution

In limit of large $\boldsymbol{\mu}$ a Poisson distribution approaches a symmetric Gaussian distribution

- This is the case not only for the Poisson distributions, but for almost any sufficiently large sum of samples with different sub-properties (mean & variance) → **Central Limit Theorem** (will discuss later)
- Gaussian distribution is of utter use, and luckily has simple properties

→ **Gauss distribution:** $P(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

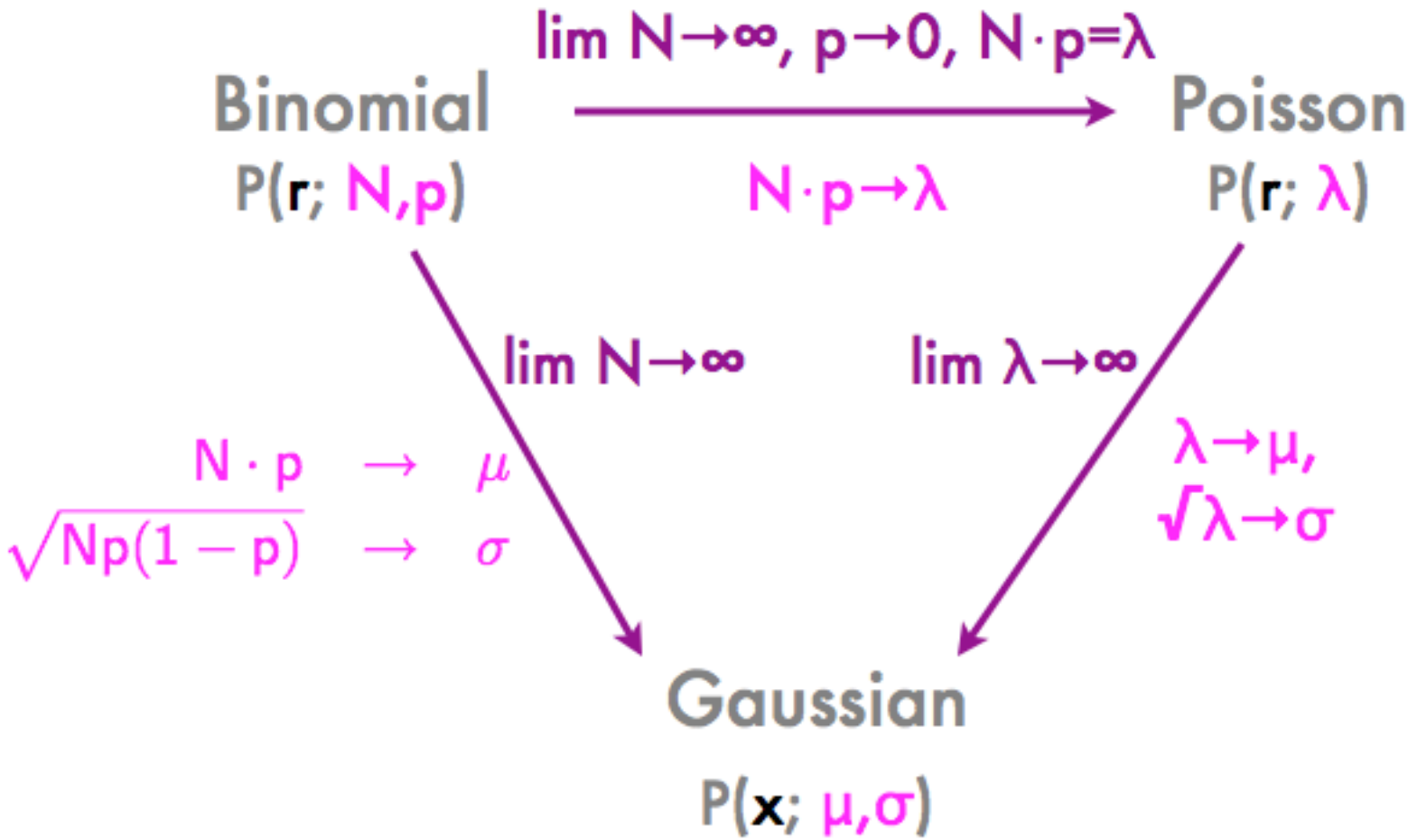
Symmetric distribution:

- Expectation value: $E[x] = \mu$
- Variance: $V[x] = \sigma^2$

Poisson distribution:

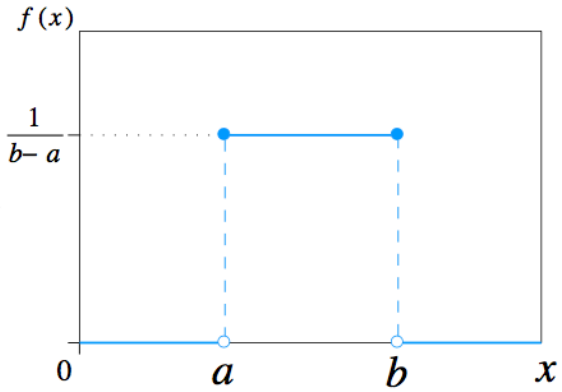
- Expectation value: $E[x] = \mu$
- Variance: $V[x] = \mu$

→ For large μ, the standard deviation (σ) of the expected event counts is $\sqrt{\mu}$!

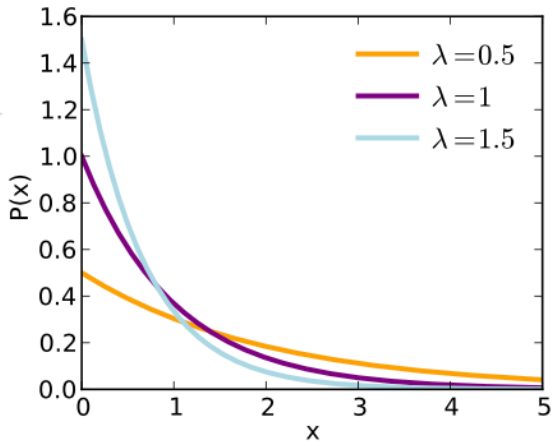

lim N→∞, p→0, N·p=λ
Binomial
P(r; N,p)
Poisson
P(r; λ)
N·p→λ
lim N→∞
lim λ→∞
N · p → μ
√Np(1 − p) → σ
λ→μ,
√λ→σ
Gaussian
P(x; μ,σ)

Some other distributions

Uniform ("flat") distribution



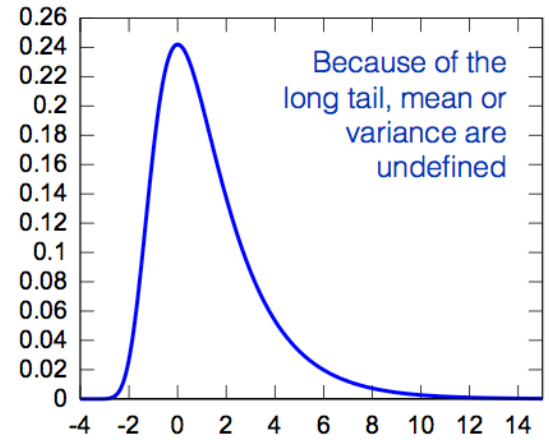
Exponential distribution



Chi-squared (χ^2) distribution

- Sum of squares of Gaussian distributed variables; used to derive goodness of a fit to describe data

Landau distribution



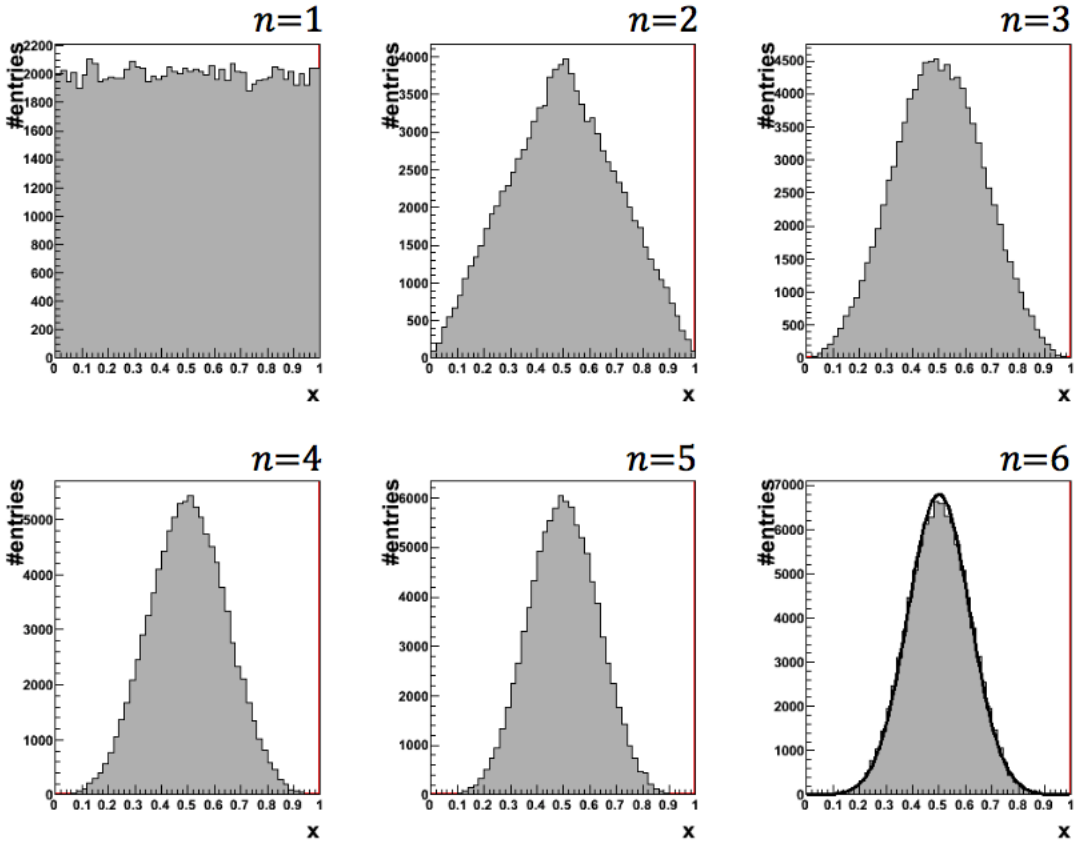
- Fluctuation of energy loss by ionization of charged particle in thin matter (eg, charge deposition in silicon detector)

Many more, see http://pdg.lbl.gov/2015/reviews/rpp2015-rev-probability.pdf for definitions and properties.

Central limit theorem (CLT)

CLT: the sum of $\boldsymbol{n}$ independent samples $\boldsymbol{x_i}$ $(i = 1, \ldots, n)$ drawn from any PDF $\boldsymbol{D(x_i)}$ with well defined expectation value and variance is Gaussian distributed in the limit $n \to \infty$

$$\boldsymbol{D}\text{: } E_D[x_i] = \mu;\ V_D[x_i] = \sigma_D^2\text{, and: } y = \frac{1}{n}\sum_{i=1}^{n} x_i \implies E_{\text{Gauss}}[y] = \mu;\ V_{\text{Gauss}}[y] = \frac{\sigma_D^2}{n}$$

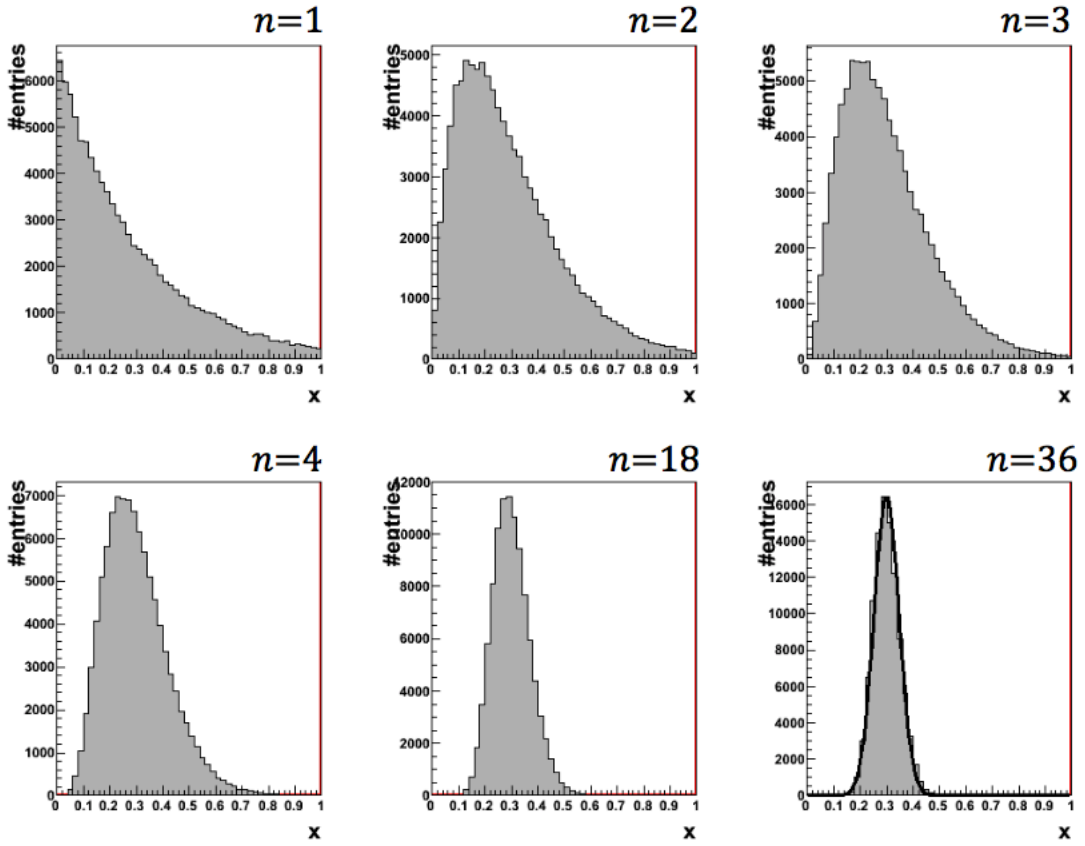


Example: summing up uniformly distributed ensembles within [0,1]

Central limit theorem (CLT)

CLT: the sum of $\boldsymbol{n}$ independent samples $\boldsymbol{x_i}$ $(i = 1, \ldots, n)$ drawn from any PDF $\boldsymbol{D(x_i)}$ with well defined expectation value and variance is Gaussian distributed in the limit $\boldsymbol{n \to \infty}$

$$\boldsymbol{D}\text{: } E_D[x_i] = \mu;\ V_D[x_i] = \sigma_D^2\text{, and: } y = \frac{1}{n}\sum_{i=1}^{n} x_i \implies E_{\text{Gauss}}[y] = \mu;\ V_{\text{Gauss}}[y] = \frac{\sigma_D^2}{n}$$



Example: summing up exponential distributions

Central Gaussian limit works even if $\boldsymbol{D}$ *doesn't look Gaussian at all*

- Many different random processes contribute to the over-all measurement errors.

- The CLT ensures that errors due to such random processes are well described by a Gaussian.

- So, if a parameter has a true value of m, the probability to measure x is
  P(x) = m + gauss(x-m; mu, error) = gauss(x; m+mu ,error)

- For truly random processes, mu=0. (If mu is not zero, we deal with "systematic errors", which have to be treated separately). So

P(x) = gauss(x; m, error).

Multidimensional random variables

What if a measurement consists of two variables?

Let:

$\boldsymbol{A}$ = measurement $\boldsymbol{x}$ in $[\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{dx}]$

$\boldsymbol{B}$ = measurement $\boldsymbol{y}$ in $[\boldsymbol{y}, \boldsymbol{y} + \boldsymbol{dy}]$

Joint probability: $P(A \cap B) = p_{xy}(x, y)dxdy$

(where $p_{xy}(x, y)$ is joint PDF)

If the two variables are independent:

$$P(A \cap B) = P(A) \cdot P(B)$$

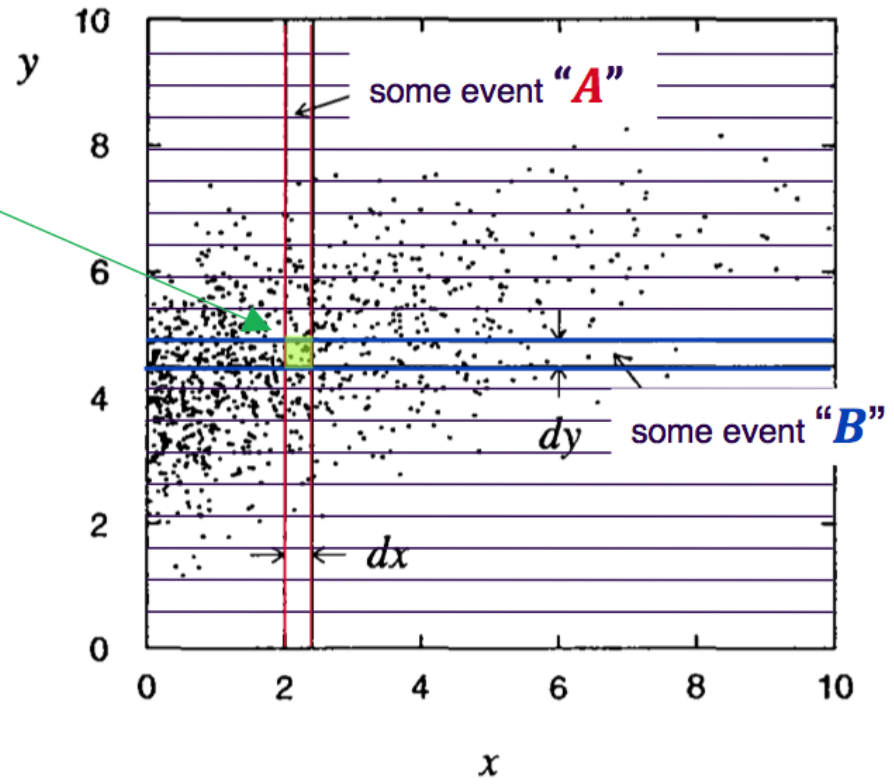
$$p_{xy}(x, y) = p_x(x) \cdot p_y\,(y)$$

Marginal PDF: if one is not interested in dependence on $\boldsymbol{y}$ (or cannot measure it),

→ integrate out ("marginalise") $\boldsymbol{y}$, ie, project onto $\boldsymbol{x}$

→ resulting one-dimensional PDF: $p_x(x) = \int p_{xy}(x, y)dy$

From: Glen Cowan, Statistical data analysis



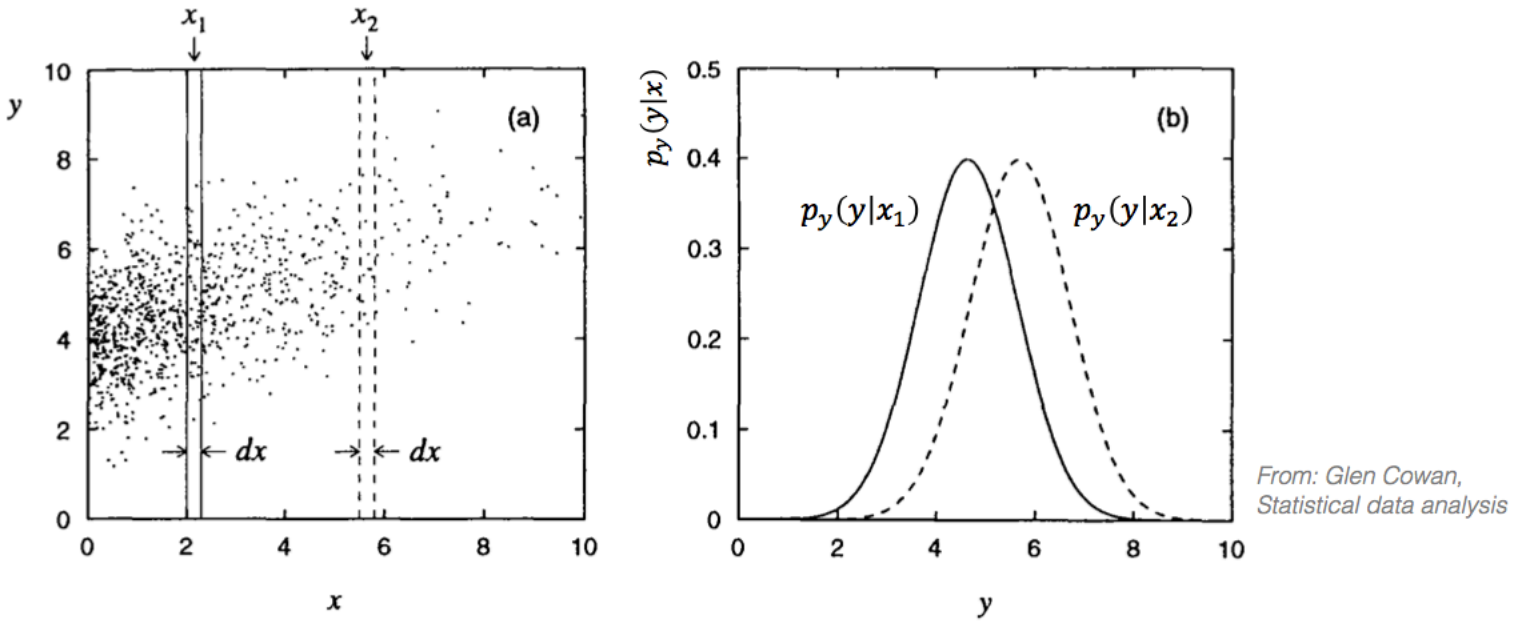
Conditioning versus marginalisation

Conditional probability $\boldsymbol{P(A|B)}$: [read: $P(A|B)$ = "probability of A *given* B"]

$$P(A \cap B) = P(A|B) \cdot P(B) \quad \Leftrightarrow \quad \boldsymbol{P(A|B)} = \frac{P(A \cap B)}{P(B)} = \frac{p_{xy}(x,y)dxdy}{p_y(y)dx}$$

Rather than integrating over the whole y region (marginalisation),
look at one-dimensional (1D) slices of the two-dimensional (2D) PDF $p_{xy}(x,y)$:

$$p_y(y|x_1) = p_{xy}(x = \text{const} = x_1, y)$$



From: Glen Cowan, Statistical data analysis

Covariance and correlation

Recall, for 1D PDF $\boldsymbol{p_x(x)}$ we had: $E[x] = \mu_x;\ \ V[x] = \sigma_x^2$

For a 2D PDF $\boldsymbol{p_{xy}(x,y)}$, one correspondingly has: μ_x, μ_y, σ_x, σ_y

How do $\boldsymbol{x}$ and $\boldsymbol{y}$ co-vary ? $\rightarrow$ $\mathbf{C}_{xy} = \mathbf{covariance}_{xy} = E\big[(x-\mu_x)(y-\mu_y)\big] = E[xy] - \mu_x\mu_y$

From this define the scale / dimension invariant *correlation coefficient*:

$$\rho_{xy} = \frac{\mathbf{C}_{xy}}{\sigma_x\sigma_y}, \quad \text{where } \rho_{xy} \subset [-1,+1]$$

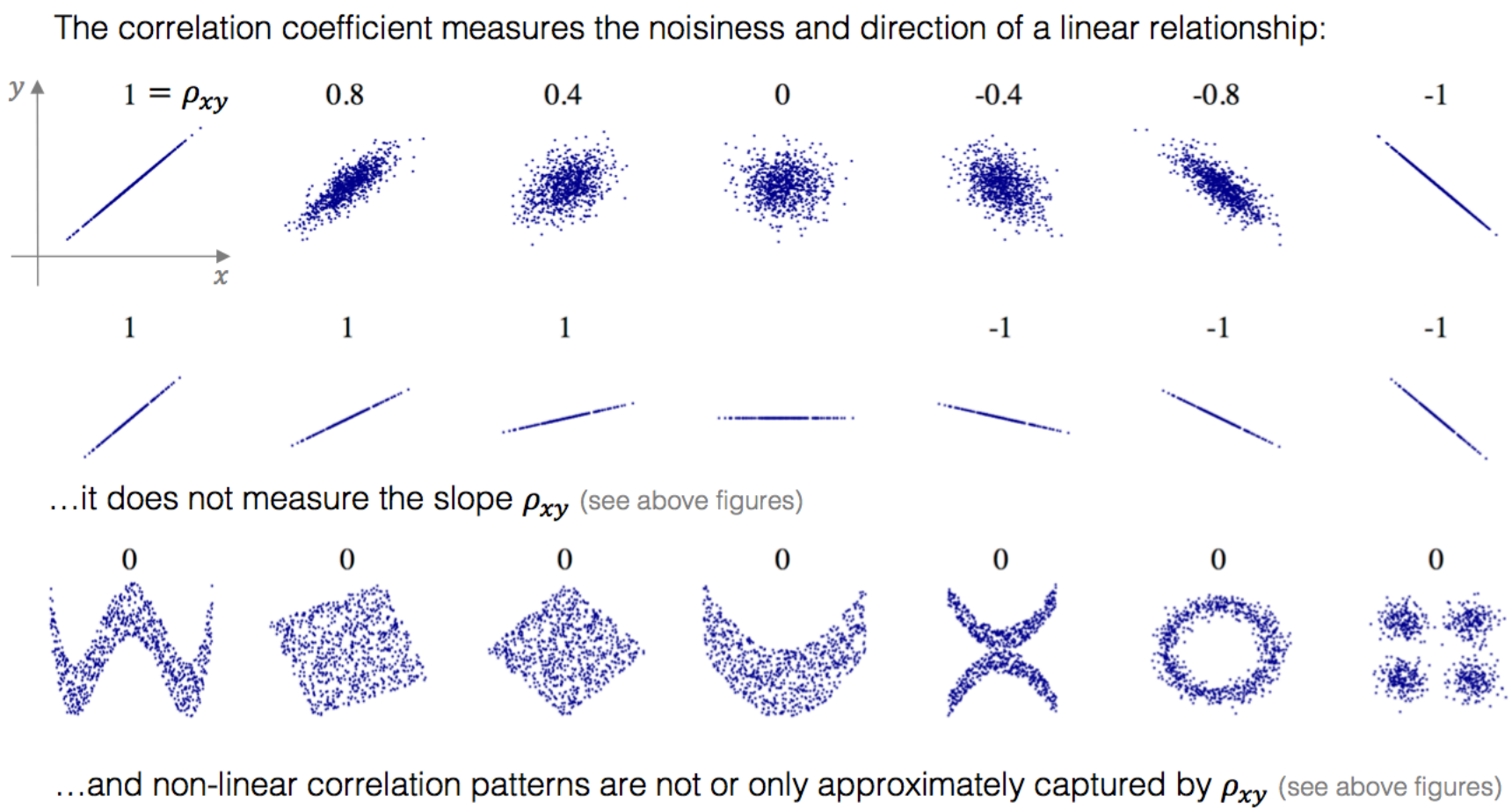
- If $\boldsymbol{x}$, $\boldsymbol{y}$ are independent: $\rho_{xy} = 0$, ie, they are *uncorrelated* (or they *factorise*)

  Proof: $E[xy] = \iint xy \cdot p_{xy}\,(x,y)dxdy = \iint xy \cdot p_x(x)p_y(y)dxdy = \int x \cdot p_x(x)dx \cdot \int y \cdot p_y(y)dy = \mu_x\mu_y$

- Note that the contrary is not always true: non-linear correlations can lead to $\rho_{xy} = 0$,
  $\rightarrow$ see next page

Correlations

Figure from: https://en.wikipedia.org/wiki/Correlation_and_dependence



Correlations

Non-linear correlation can be captured by the "*mutual information*" quantity $\boldsymbol{I_{xy}}$:

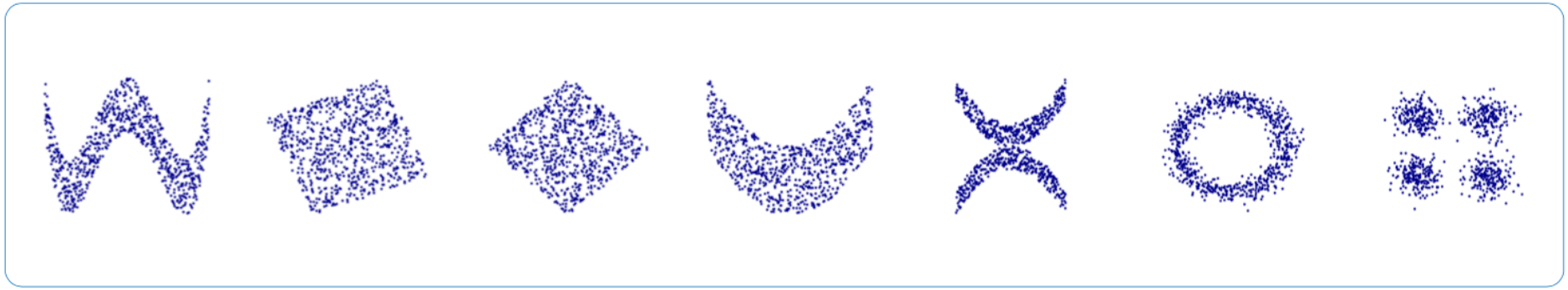
$$I_{xy} = \iint p_{xy}(x,y) \cdot \ln\left(\frac{p_{xy}(x,y)}{p_x(x)p_y(y)}\right) dxdy$$

Measure of mutual dependence between two variables:
"How much information is shared among them"

where I_{xy} =0 only if $\boldsymbol{x}, \boldsymbol{y}$ are fully statistically independent

Proof: if independent, then $p_{xy}(x,y) = p_x(x)p_y(y) \Rightarrow \ln(\dots) = 0$

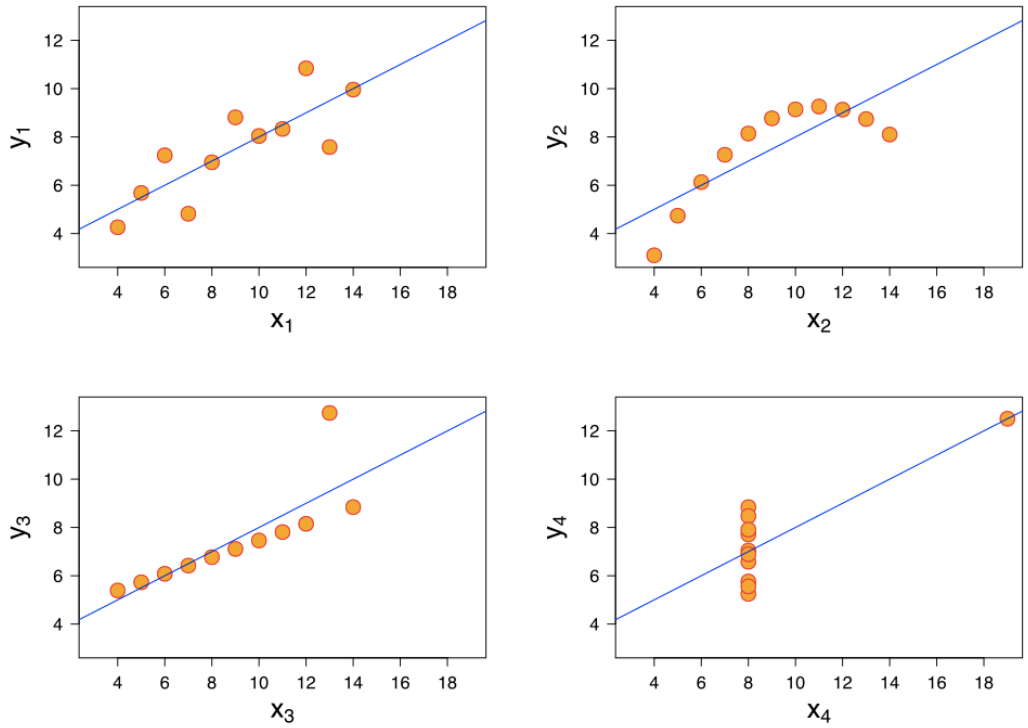
NB: $I_{xy} = H_x - H_x(y) = H_y - H_y(x)$,
where $H_x = -\int p_x(x) \cdot \ln\big(p_x(x)\big)\, dx$ is *entropy*, $H_x(y)$ is *conditional entropy*



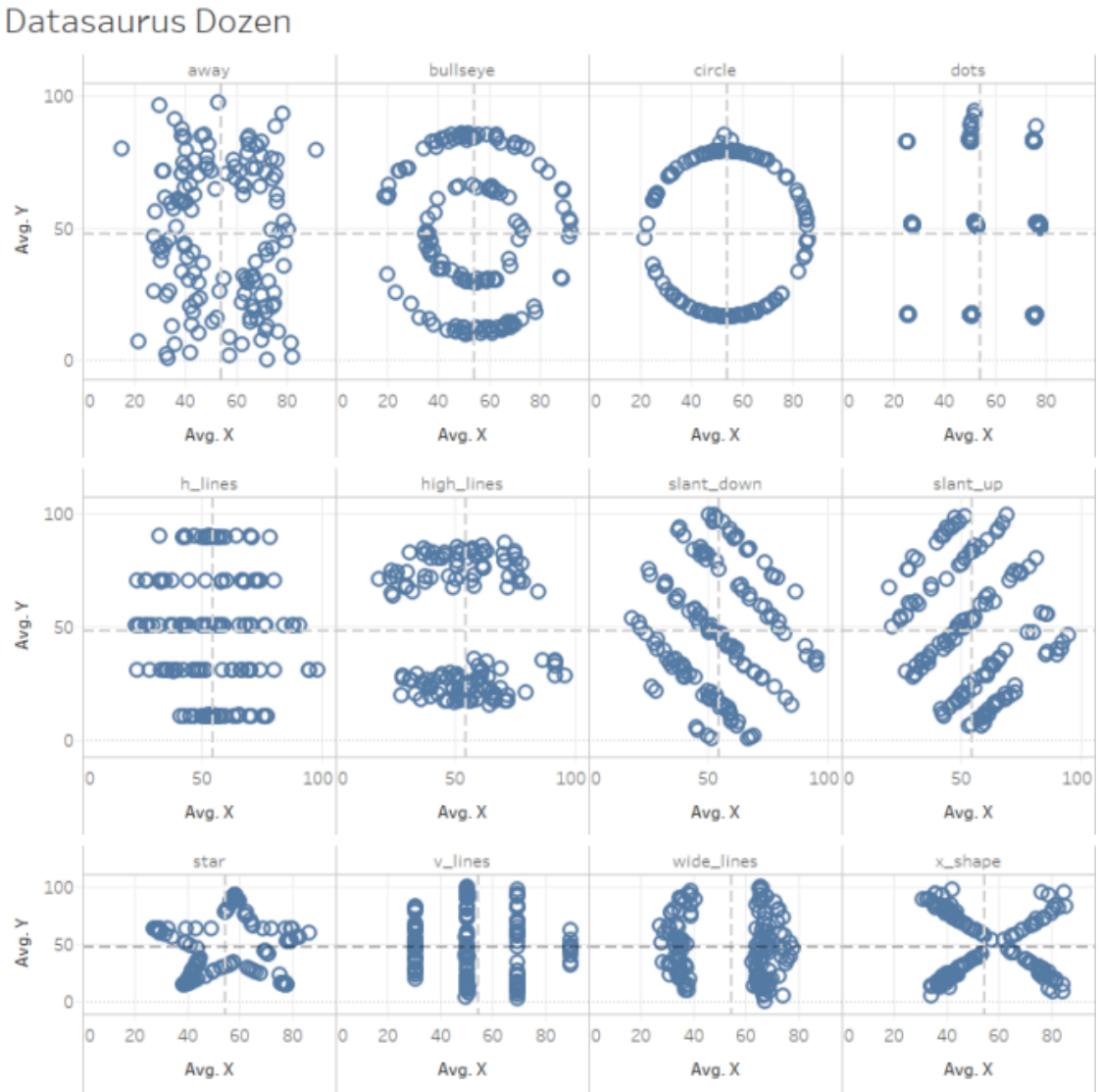
Anscombe’s quartet

For all four datasets:

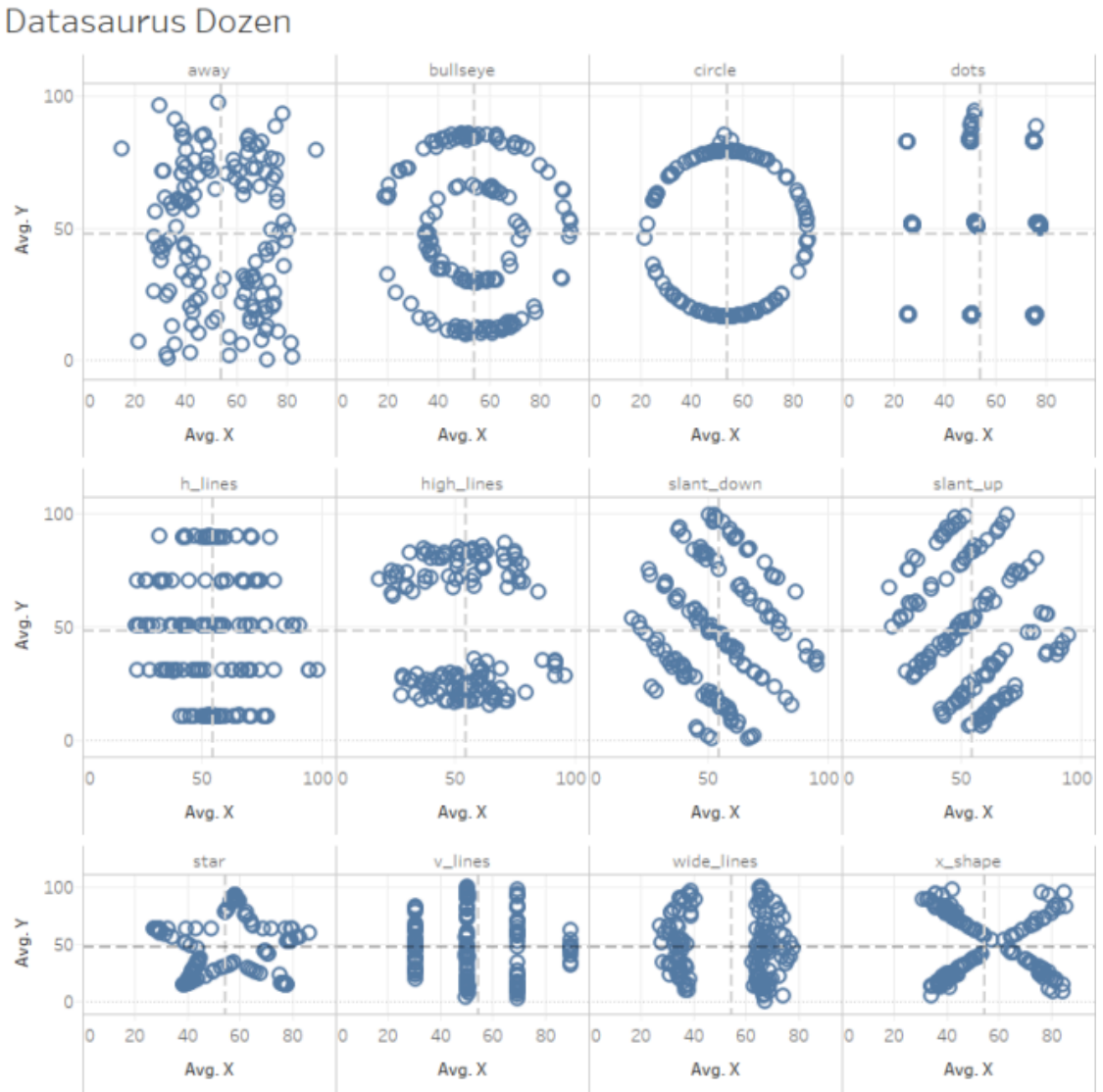
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

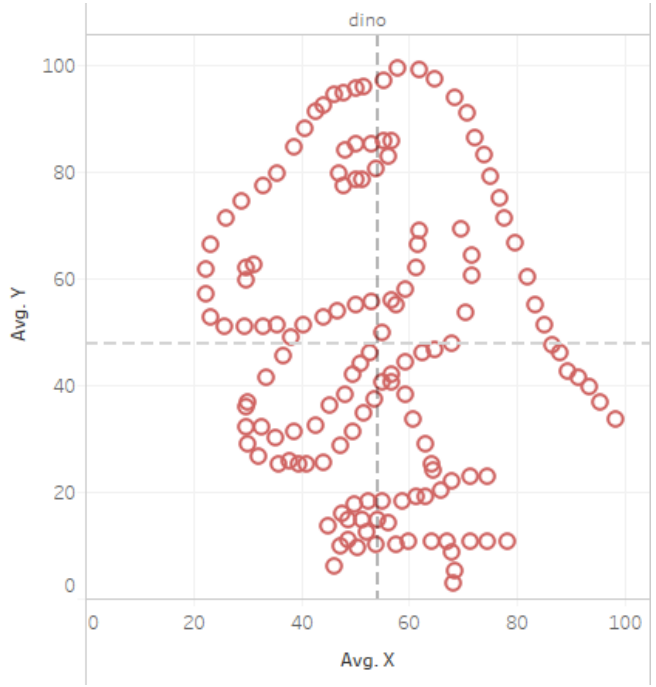


Modern Anscombe’s quartet



Modern Anscombe’s quartet

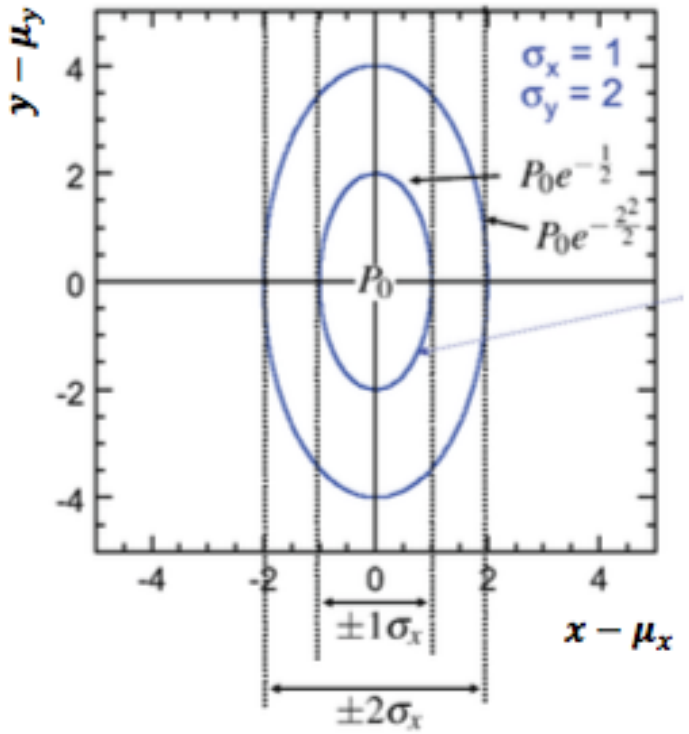




2D Gaussians

Two variable $\boldsymbol{x}, \boldsymbol{y}$ are independent: $[\, p_{xy}(x,y) = p_x(x) \cdot p_y(y) \,]$

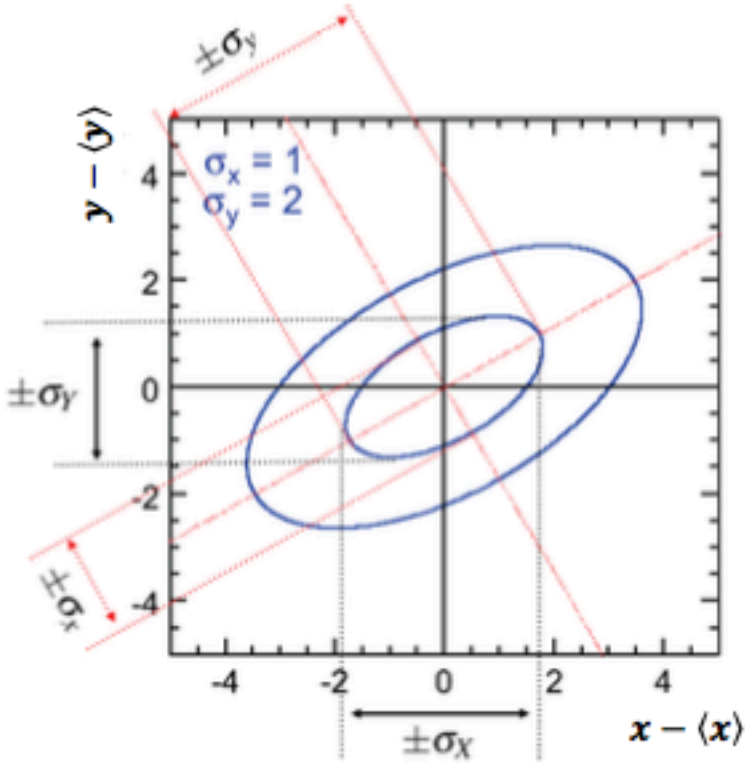
$$p_{xy}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}$$



2D Gaussians

Two variable $\boldsymbol{x}, \boldsymbol{y}$ are *not* independent: $[\, p_{xy}(x,y) \neq p_x(x) \cdot p_y(y) \,]$

$$p_{\vec{x}}(\vec{x}) = \frac{1}{2\pi\sqrt{\det(C)}} \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})\right)$$



where:

$$C = \begin{pmatrix} \langle x^2 \rangle - \langle x \rangle^2 & \langle xy \rangle - \langle x \rangle \langle y \rangle \\ \langle xy \rangle - \langle x \rangle \langle y \rangle & \langle y^2 \rangle - \langle y \rangle^2 \end{pmatrix}$$

is the (symmetric) *covariance matrix*

Corresponding correlation matrix elements:

$$\rho_{ij} = \rho_{ji} = \frac{C_{ij}}{\sqrt{C_{ii} \cdot C_{jj}}}$$

Avoiding False Discoveries

• Imagine you have 100 independent measurements. What is the probability that at least one of them would show a 3-sigma effect?

• Let's assume that we are not looking for excesses (that would be a one-tailed Gaussian), but for deviations either way, so the 2-tailed limits apply. Then:

P(at least one 3sigma effect) = 1 - P(all within 3sigma) = 1 - $(0.9973)^{100}$ = 1.0 - 0.76 = **0.24**

P(at least one 3sigma effect in 1 measurement) = 0.27%
P(at least one 3sigma effect in 10 measurements) = 2.7%
P(at least one 3sigma effect in 100 measurements) = 24%
P(at least one 3sigma effect in 1000 measurements) = 93%

Avoiding False Discoveries

• Imagine you have 100 independent measurements. What is the probability that at least one of them would show a 3-sigma effect?

• Let's assume that we are not looking for excesses (that would be a one-tailed Gaussian), but for deviations either way, so the 2-tailed limits apply. Then:

P(at least one 3sigma effect) = 1 - P(all within 3sigma) = 1 - $(0.9973)^{100}$ = 1.0 - 0.76 = 0.24

P(at least one 3sigma effect in 1 measurement) = 0.27%
P(at least one 3sigma effect in 10 measurements) = 2.7%
P(at least one 3sigma effect in 100 measurements) = 24%
P(at least one 3sigma effect in 1000 measurements) = 93%

The important, and somewhat brain-twisting, conclusion from this section is that the significance of a any deviation in a measurement from the expected value, does not only depend on the measurement itself, but on how many other measurements you made, how many other places you looked. REMEMBER LOOK ELSEWHERE EFFECT.

General Considerations

•This sort of consideration is very important in “data mining”, when looking for significant effects, say, in the general census. Since we look at 100s of different data items (age, income, height, spending on food, ...) we’re bound to find effects, and correlations) that look significant at first glance, but aren’t.

•**The correct way** of dealing with it is:

- Take these effects from data mining as hints what might be interesting.
- Then take a new data sample, look only at the small number of measurement where you suspect an effect, and see if a significant effect persists in the new data.
- Because you looked at far fewer places the 2nd time, the same n-sigma effect would have much higher significance in the new data.
- One way to do this in practice is to split your data sample and use one part for data mining, and the other part for seeking significant evidence for or against a small number of hypotheses.

Blind analyses

In order to avoid false discovery, one should use blind analysis technique, i.e. not optimising in the final sample.

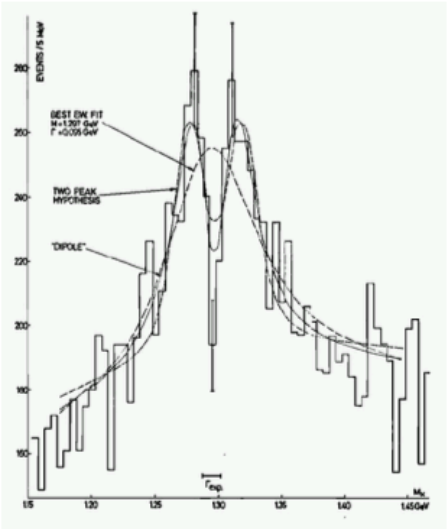


Fig. 3: The "discovery" of the split A_2 in the missing mass spectrum of the process $\pi^- + p \to p + MM^-$.