Concepts of Experiments at Future Colliders II

PD Dr. Oliver Kortner

06.06.2025

Fundamentals of statistical treatment of experimental data

Introductory example: beam energy measurement

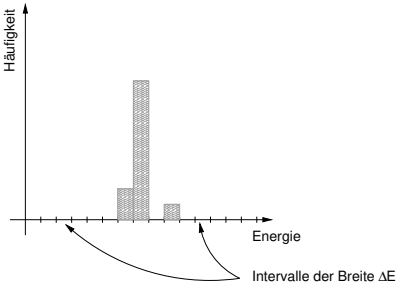
Example: Measurement of the energy of a monoenergetic particle beam.

Notations

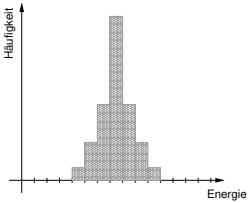
E_S: actual beam energy.
N: number of measurements of beam energy.
E_k: result of the k-th measurement of beam energy.

Frequency Distributions

N small

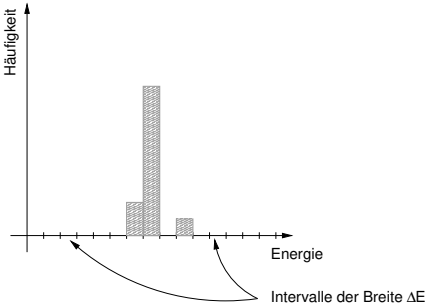


N large

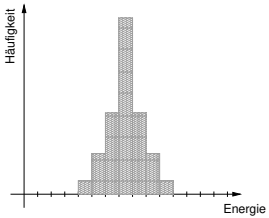


Introductory example: beam energy measurement

N small



N large

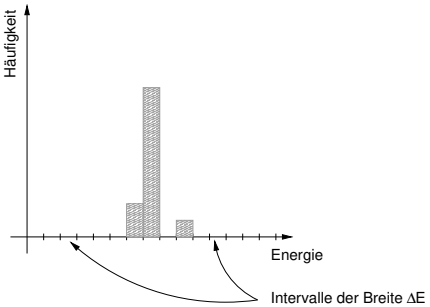


- When N is large, repeating the N measurements yields (nearly) the same frequency distribution.
- In the limit $N \to \infty$, the frequency distribution converges to the probability distribution for the outcome of the measurement.

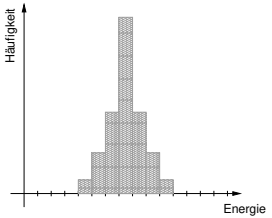
Recapitulation of the previous lecture

Introductory example: beam energy measurement

N small



N large



- The probability of measuring E_k when the beam energy is E_S depends on the value of E_S and the measurement method. If one knows the probability function $p(E_k; E_S)$, one can determine E_S from the measurement of the frequency distribution.
- In practice, $p(E_k; E_S)$ is only partially known, and one tries to infer $p(E_k; E_S)$ from the measured frequency distribution, which provides an estimate of E_S. In statistics, methods are employed to infer the underlying probability distributions from frequency distributions.

Recapitulation of the previous lecture

Probability distributions

- A physical measurement is a random process.
- A measured quantity x, which represents the outcome of a random process, is called a random variable or random quantity.
- Any function of x is also a random variable.
- If the random variable can only take discrete values, there is a probability for the occurrence of each of these values, which is the probability function.
- For random variables with continuous range of values, the probability density $p(x)$ replaces the probability function. Let Ω be a measurable set of possible values of x, whose measure is greater than zero. Then

$$\int_{\Omega} p(x)dx$$

is the probability of observing a value $x \in \Omega$.

Axiomatic Definition of Probability

The mathematical field of probability theory is based on Kolmogorov's Axioms.

Kolmogorov's Axioms

Let Σ denote a set of events.

1. For every event $A \in \Sigma$, the probability of the occurrence of A is a real number $p(A) \in [0, 1]$.
2. The certain event $S \in \Sigma$ has probability $p(S) = 1$.
3. The probability of the union of countably many incompatible events is equal to the sum of the probabilities of the individual events. Here, events A_k are incompatible if they are pairwise disjoint, i.e., $A_k \cap A_\ell = \emptyset$ for all $k \neq \ell$.

Recapitulation of the previous lecture

Characteristics of probability distributions

Remark. In this section, we consider probability densities. Probability functions of discrete variables are also covered if one considers the δ-distribution as a probability density.

Nomenclature. D: Range of values of a random variable $x = (x_1, ..., x_n)$.
$p(x)$: Probability density of x.
(D is the domain of p.)

Definitions

The expectation value of x, $E(x)$ (also $< x >$), is defined as

$$E(x) := \int_D x \cdot p(x) dx.$$

The covariance matrix $cov(x_k, x_l)$ is defined as

$$cov(x_k, x_l) := < (x_k - < x_k >) \cdot (x_l - < x_l >) > .$$

The diagonal element $cov(x_k, x_k)$ is called the variance of x_k, $Var(x_k)$, and $\sqrt{Var(x_k)}$ is the standard deviation $\sigma(x_k)$.

Expectation value of a function of a random variable

- A function $f(x)$ is also a random variable.

$$< f >= \int_D f(x)p(x)dx.$$

- If $f(x) = f(x - < x > + < x >)$ is significantly different from 0 only for small values of $|x - < x > |$, one can approximate $f(x)$ by

$$f(< x >) + \left.\frac{df}{dx}\right|_{<x>} \cdot (x - < x >)$$

Then

$$\begin{aligned} < f > &\approx \left\langle f(< x >) + \left.\frac{df}{dx}\right|_{<x>} \cdot (x - < x >) \right\rangle \\ &= < f(x) > + \left\langle \left.\frac{df}{dx}\right|_{<x>} \cdot (x - < x >) \right\rangle \\ &= f(< x >) + \left.\frac{df}{dx}\right|_{<x>} \cdot (< x > - < x >) = f(< x >). \end{aligned}$$

Variance of a function of a random variable

Special Case: $f(x) \in \mathbb{R}$.

$$
\begin{aligned}
Var(f) &= \left\langle (f - < f >)^2 \right\rangle = \langle [f - f(< x >)] \rangle \\
&\approx \left\langle \left[\sum_{k=1}^{n} \left. \frac{df}{dx_k} \right|_{<x>} \cdot (x_k - < x_k >) \right]^2 \right\rangle \\
&= \left\langle \left[\sum_{k,\ell=1}^{n} \left. \frac{df}{dx_k} \right|_{<x>} \left. \frac{df}{dx_\ell} \right|_{<x>} \cdot (x_k - < x_k >) \cdot (x_\ell - < x_\ell >) \right] \right\rangle \\
&= \sum_{k,\ell=1}^{n} \left. \frac{df}{dx_k} \right|_{<x>} \left. \frac{df}{dx_\ell} \right|_{<x>} \cdot \langle (x_k - < x_k >) \cdot (x_\ell - < x_\ell >) \rangle \\
&= \sum_{k,\ell=1}^{n} \left. \frac{df}{dx_k} \right|_{<x>} \left. \frac{df}{dx_\ell} \right|_{<x>} \cdot cov(x_k, x_\ell),
\end{aligned}
$$

which is the well-known error propagation formula.

Examples of important probability distributions

The binomial distribution

- The binomial distribution gives the probability of observing n_k events out of a total of N events when ν_k events are expected:

$$p(n_k;\nu_k) = \binom{N}{n_k}\left(\frac{\nu_k}{N}\right)^{n_k}\left(1-\frac{\nu_k}{N}\right)^{N-n_k}.$$

- With $p := \frac{\nu_k}{N}$, one obtains from

$$\begin{aligned}
0 &= \frac{d}{dp}1 = \frac{d}{dp}\sum_{n_k=0}^{N}\binom{N}{n_k}p^{n_k}(1-p)^{N-n_k} \\
&= \sum_{n_k=0}^{N}\binom{N}{n_k}\left[n_k p^{n_k-1}(1-p)^{N-n_k} - (N-n_k)p^{n_k}(1-p)^{N-n_k-1}\right] \\
&= \frac{1}{p} < n_k > - \frac{1}{1-p} < N-n_k > = \left(\frac{1}{p}+\frac{1}{1-p}\right) < n_k > + \frac{N}{1-p} \\
&= \frac{1}{p(1-p)} < n_k > + \frac{N}{1-p} \Leftrightarrow < n_k > = N\cdot p = N\cdot\frac{\nu_k}{N} = \nu_k.
\end{aligned}$$

- Using the same calculation trick, one obtains $Var(n_k) = \nu_k(1-\frac{\nu_k}{N})$.

Transition to the Poisson distribution

If $\nu_k \gtrsim 10$, $\nu_k \ll N$, and N are large, one can approximate it by the Poission distribution. The approximation is a results of the Stirling formula:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \; f\ddot{u}r \; n \to \infty.$$

$$\begin{aligned}
p(n_k;\nu_k) &= \frac{N!}{n_k!(N-n_k)!} p^{n_k}(1-p)^{N-n_k} \\
&\approx \frac{1}{n_k!} p^{n_k} \left(\frac{N}{e}\right)^N \sqrt{2\pi N} \frac{1}{\left(\frac{N-n_k}{e}\right)^{N-n_k} \sqrt{2\pi(N-n_k)}} (1-p)^{N-n_k} \\
&= \frac{1}{n_k} p^{n_k} e^{-n_k} \underbrace{\sqrt{\frac{N}{N-n_k}}}_{\to 1 \;\; f. \;\; N\to\infty} \frac{N^N}{(N-n_k)^{N-n_k}} (1-p)^{N-n_k} \\
&\approx \frac{1}{n_k!} e^{-n_k} p^{n_k} N^{n_k} N^{N-n_k} (1-p)^{N-n_k} \frac{1}{(N-n_k)^{N-n_k}} \\
&= \frac{\nu_k}{n_k!} e^{-n_k} \frac{(N-\nu_k)^{N-n_k}}{(N-n_k)^{N-n_k}} \approx \frac{\nu_k^{n_k}}{n_k!} e^{-\nu_k} \text{ (Poisson distribution).}
\end{aligned}$$

Properties of the Poisson distribution

Poisson distribution

$$p(n_k;\nu_k) = \frac{\nu_k^{n_k}}{n_k!} e^{-\nu_k}.$$

Normalization

$$\sum_{n_k=0}^{\infty} p(n_k;\nu_k) = e^{-\nu_k} \sum_{n_k=0}^{\infty} \frac{\nu_k^{n_k}}{n_k!} = e^{-\nu_k} \cdot e^{\nu_k} = 1.$$

Expectation value: ν_k, resulting from $0 = \frac{d}{d\nu_k} \sum_{n_k=0}^{\infty} p(n_k;\nu_k)$.

Variance: ν_k, resulting from $0 = \frac{d^2}{d\nu_k^2} \sum_{n_k=0}^{\infty} p(n_k;\nu_k)$.

Poisson distribution for $\nu_k \to \infty$

When ν_k becomes large, the probability of the occurrence of small values of n_k is small. Then n_k can be considered large, and for $n_k!$ in the Poisson distribution, Stirling's approximation can be used:

$$
\begin{aligned}
\frac{\nu_k^{n_k}}{n_k!}e^{-\nu_k} &\approx \frac{\nu_k^{n_k}}{n_k^{n_k}}\frac{1}{\sqrt{2\pi n_k}}e^{n_k-\nu_k} \\
&\approx \frac{1}{\sqrt{2\pi\nu_k}}\exp\left(\ln\frac{\nu_k^{n_k}}{n_k^{n_k}}\right)\exp(n_k-\nu_k) \\
&= \frac{1}{\sqrt{2\pi\nu_k}}\exp\left(n_k\ln\frac{\nu_k}{\nu_k+n_k-\nu_k}\right)\exp(n_k-\nu_k) \\
&= \frac{1}{\sqrt{2\pi\nu_k}}\exp\left(n_k\ln\frac{1}{1-\frac{n_k-\nu_k}{\nu_k}}\right)\exp(n_k-\nu_k) \\
&\approx \frac{1}{\sqrt{2\pi\nu_k}}\exp\underbrace{\left[n_k\cdot\left(-\frac{n_k-\nu_k}{\nu_k}-\frac{1}{2}\frac{(n_k-\nu_k)^2}{\nu_k^2}\right)\right]}_{\approx-(n_k-\nu_k)-\frac{(n_k-\nu_k)^2}{2\nu_k}}\exp(n_k-\nu_k) \\
&\approx \frac{1}{\sqrt{2\pi\nu_k}}e^{-\frac{(n_k-\nu_k)^2}{2\nu_k}}.
\end{aligned}
$$

The normal distribution

Normal distribution of a one-dimensional random variable $x \in \mathbb{R}$

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- $< x >= \mu$, $Var(x) = \sigma^2$.
- The Poisson distribution approaches a normal distribution in the limit $\nu_k \to \infty$ with the expected value ν_k and the variance ν_k.

Normal distribution of a d-dimensional random variable $x \in \mathbb{R}^d$

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{\det(\Sigma)} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma (x-\mu)\right).$$

$$\Sigma \in \mathbb{R}^{d\times d},\ \mu \in \mathbb{R}^d.$$

- $< x >= \mu$.
- $cov(x_k, x_l) = \Sigma_{k,l}$.

Properties of the one-dimensional normal distribution

$w_n :=$ Probability of observing a value $x \in [\mu - n\sigma, \mu + n\sigma]$.

n	w_n
1	0.6827
2	0.9545
3	0.9973
4	$1 - 6.3 \cdot 10^{-5}$
5	$1 - 5.7 \cdot 10^{-7}$

w_n	n
0.900	1.645
0.950	1.960
0.990	2.576
0.999	3.290

Concept of stochastic convergence

(t_n) is a sequence of random variables and T is also a random variable. We say t_n converges stochastically to T if for every $p \in [0, 1[$ and $\epsilon > 0$, there exists an N such that the probability P that $|t_n - T| > \epsilon$ is less than p for all $n > N$:

$$P(|t_n - T| > \epsilon) < p \ (n > N).$$

In other words: The probability of observing a value t_n different from T vanishes as $n \to \infty$.

Law of large numbers. Central limit theorem

The law of large numbers

(x_n) is a sequence of independent random variables, each following the same distribution function. μ denotes the expected value of x_n. Then the arithmetic mean

$$\frac{1}{N}\sum_{n=1}^{N} x_n$$

converges stochastically to μ.

The central limit theorem

(x_n) is a sequence of identically distributed random variables with mean μ and standard deviation σ. Then as $N \to \infty$, the standardized random variable

$$Z_N := \frac{\sum_{n=1}^{N} x_n - N\mu}{\sigma\sqrt{N}}$$

converges pointwise to a normal distribution with mean 0 and standard deviation 1.

Point estimation

Let α be a parameter of a probability distribution. The goal of point estimation is to find the best estimate (the best measurement in the terminology of physicists) of α.

x: Random variable corresponding to the experimental measurements.
$p(x;\alpha)$: Probability density for the measurement of x as a function of the parameter α.

x and α can be multidimensional.

Definition. A point estimator $\mathcal{E}_\alpha$ is a function of x used to estimate the value of the parameter α. Let $\hat{\alpha}$ denote this estimate. Thus, $\hat{\alpha} = \mathcal{E}_\alpha(x)$.

Goal is to find a function $\mathcal{E}_\alpha$ such that $\hat{\alpha}$ is as close as possible to the true value of α.

Since $\hat{\alpha}$ is a function of random variables, $\hat{\alpha}$ itself is a random variable.

$$p(\hat{\alpha}) = \int_D \mathcal{E}_\alpha(x) p(x;\alpha) dx,$$

where α denotes the true value of the parameter.

Quality criteria for point estimators

Consistency

n: Number of measurements used for the point estimation.
$\hat{\alpha}_n$: Corresponding estimate.
α_0: True value of α.
$\mathcal{E}_\alpha$ is called a consistent point estimator if $\hat{\alpha}_n$ converges stochastically to α_0. This means that the probability of estimating a value different from α_0 goes to 0 as $n \to \infty$.

Unbiasedness

The bias of an estimate $\hat{\alpha}$ is defined as

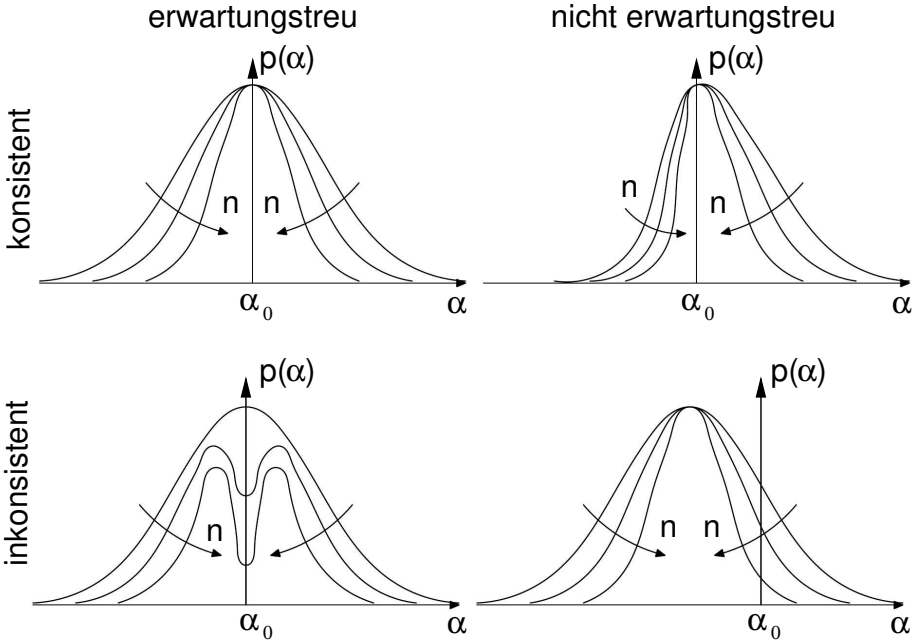
$$b_n(\hat{\alpha}) := E(\hat{\alpha}_n - \alpha_0) = E(\hat{\alpha}_n) - \alpha_0.$$

The point estimator is unbiased if

$$b_n(\hat{\alpha}) = 0, \text{ or } E(\hat{\alpha}_n) = \alpha_0$$

for all n.

Illustration of Consistency and Unbiasedness



Further quality criteria for point estimators

Efficiency

Let V_{min} be the minimum possible variance among all point estimators of a real-valued parameter. The efficiency of a particular point estimator is given by the ratio $\frac{V_{min}}{Var(\hat{\alpha})}$, where $Var(\hat{\alpha})$ is the variance of $\hat{\alpha}$ for that point estimator.

Sufficiency

Any function of data x is called a statistic. A sufficient statistic for α is a function of the data that contains all the information about α.

Point estimators used in high energy physics

Maximum likelihood method

$p(x;\alpha)$: Probability of obtaining the measured values x given a parameter α.

- Substituting the measured values x into the function $p(x;\alpha)$ yields a statistic of x, which is called the likelihood or the likelihood function $L(x;\alpha)$.
- The term likelihood is used to indicate the relationship with the probability density $p(x;\alpha)$ while making it clear that L is not a probability function.

Let $f(x_k;\alpha)$ be the probability density for the outcome of a single measurement x_k. With n independent measurements $x = (x_1, \ldots, x_n)$, we have

$$L(x_1, \ldots, x_n; \alpha) = \prod_{k=1}^{n} f(x_k; \alpha).$$

In the method of maximum likelihood, the estimate for α is taken as the value of α that maximizes $L(x;\alpha)$.

Asymptotic behavior of maximum likelihood

$n \to \infty$

- The point estimator is consistent.
- The point estimator is efficient.
- $\hat{\alpha}$ is normally distributed.
- Due to consistency, the point estimator is asymptotically unbiased.

Finite n

To determine the behavior of the point estimator with limited data size n, experimental practice uses ensembles of randomly generated simulated data to which the point estimator is applied.

Method of least squares

n measurements $x_1, \ldots, x_n$.

$E(x_k; \alpha)$: Expectation value of x_k given α (ttheoretical predictionffor the value of x_k).

$V = (cov(x_k, x_\ell))$: Covariance matrix. In general, V is also a function of α.

$$Q^2 := \sum_{k,\ell=1}^{n} \left[x_k - E(x_k; \alpha)\right] V_{k\ell}^{-1}(\alpha) \left[x_\ell - E(x_\ell; \alpha)\right].$$

In the method of least squares, the estimate for α is chosen as the value for which Q^2 is minimized.

Remark. If $V_{k\ell}(\alpha)$ is unbounded, we may obtain nonsensical results for α. For example, if $V_{k\ell}(\alpha) \to \infty$ as $\alpha \to \alpha_{\text{non-sense}}$ and $x_k - E(x_k; \alpha)$ remains bounded, the minimization yields $\alpha_{\text{non-sense}}$. In practice, Q^2 is often minimized iteratively. One starts with an estimate for V and varies V during the minimization of Q^2. Then, V is recalculated for the obtained estimate of α, and the minimization is repeated with V fixed until $\hat{\alpha}$ no longer changes significantly.