Concepts of Experiments at Future Colliders II

PD Dr. Oliver Kortner

07.06.2024

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:

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

- 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 $Var(n_k) = \nu_k(1-\frac{\nu_k}{N})$.

Recapitulation of the previous lecture

Transition to the Poisson distribution

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

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

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

Recapitulation of the previous lecture

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 $0 = \frac{d^2}{d\nu_k^2} \sum_{n_k=0}^{\infty} p(n_k;\nu_k)$.

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

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

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

Recapitulation of the previous lecture

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

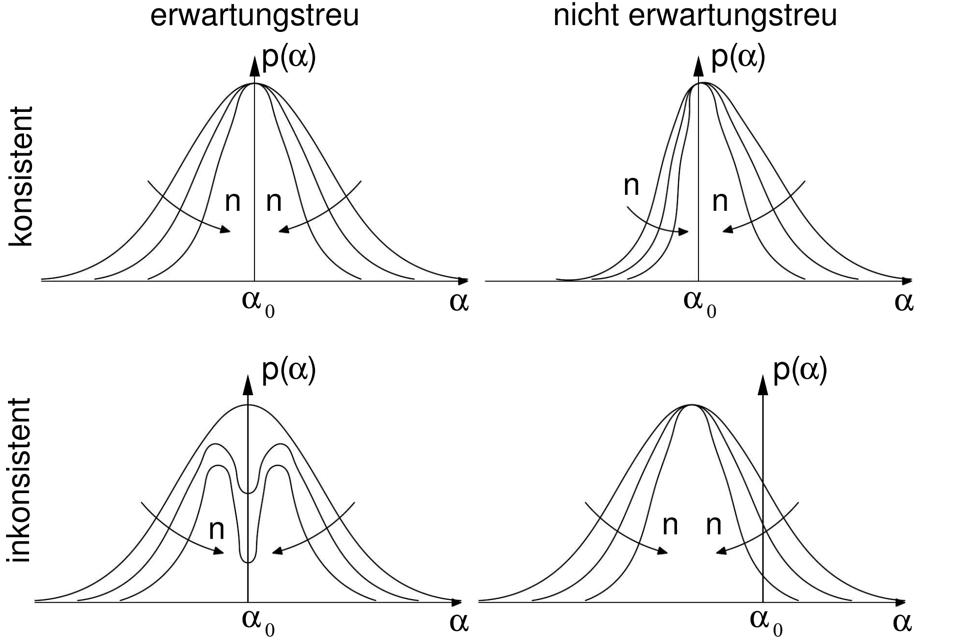
$$b_n(\hat{\alpha}) := E(\hat{\alpha}_n - \alpha_0) = E(\hat{\alpha}_n) - \alpha_0.$$

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 $x = (x_1, \dots, x_n)$, we have

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

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

Recapitulation of the previous lecture

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.

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

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.

Interval estimation

Goal: Determination of an interval which contains the true value of a parameter with a given probability.

Limit case of the normal distribution

Let us assume the variable $x \in \mathbb{R}$ is normally distributed, i.e.

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

If μ and σ are known, then

$$p(a < x < b) = \int_a^b N(x;\mu,\sigma)dx =: \beta.$$

If μ is unknown, one can calculate $p(\mu + c < x < \mu + d)$:

$$\begin{aligned} \beta = p(\mu + c < x < \mu + d) &= \int_{\mu+c}^{\mu+d} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \int_c^d \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{y^2}{\sigma^2}} dy \\ &= p(c - x < -\mu < d - x) = p(x - d < \mu < x - c). \end{aligned}$$

Interval estimation with the normal distribution

$$\beta = p(\mu + c < x < \mu + d) = \int_{\mu+c}^{\mu+d} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \int_{c}^{d} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{y^2}{\sigma^2}} dy$$

$$= p(c - x < -\mu < d - x) = p(x - d < \mu < x - c).$$

That is, if x has been measured, the probability that the desired value of μ lies between $x - d$ and $x - c$ is equal to β.

- If x is a parameter $\hat{\alpha}$ from a point estimation conducted using the method of maximum likelihood or the method of least squares, then $\hat{\alpha}$ is asymptotically normally distributed, and the above formulas can be applied for interval estimation.
- The intervals $[a, b]$ or $[x - d, x - c]$ are called confidence intervals. β is the confidence level corresponding to the confidence level.

Generalization to the multidimensional case

$$Q(x;\mu,\Sigma) := (x-\mu)^t\Sigma^{-1}(x-\mu),\ \ x,\mu \in \mathrm{|R}.$$

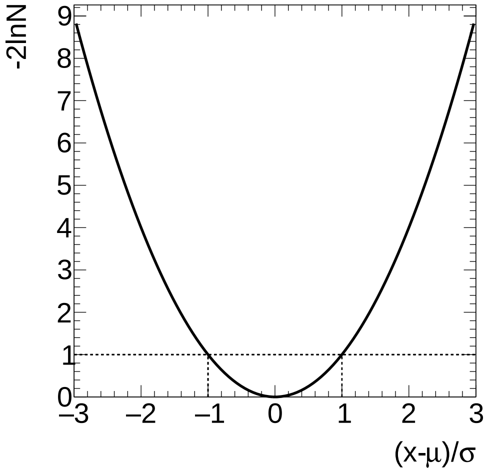
$$p(Q) = \frac{1}{(2\pi)^{d/2}} \cdot \frac{1}{\sqrt{det(\Sigma)}} \exp\left(-\frac{1}{2}Q(x;\mu,\Sigma)\right).$$

In multiple dimensions, the confidence interval becomes a confidence region corresponding to the confidence level β:

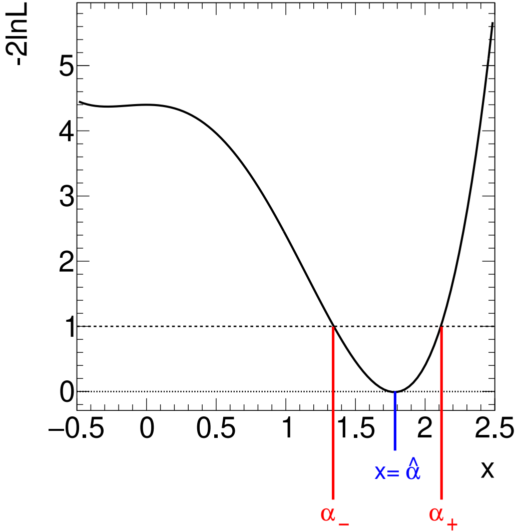
$$p(Q(x;\mu,\Sigma) < K_\beta^2) = \beta.$$

Likelihood-based confidence intervals

$$-2\ln N(x = \mu \pm \sigma; \mu, \sigma) - [-2\ln N(x = \mu; \mu, \sigma] = 1.$$



Generalization



Confidence Interval: $[\alpha_{-}, \alpha+]$.

Hypothesis testing

Goal, to determine which hypothesis (for a probability distribution) describes the recorded data point distributions (data).

Nomenclature. H_0: null hypothesis.
H_1: alternative hypothesis.

Simple and Composite Hypotheses

- When the hypotheses H_0 and H_1 are given completely without free parameters, the hypotheses are called simple hypotheses.
- If a hypothesis contains at least one free parameter, it is referred to as a composite hypothesis.

Procedure

For hypothesis testing, W must be chosen such that

$$p(\text{data} \in W | H_0) = \alpha$$

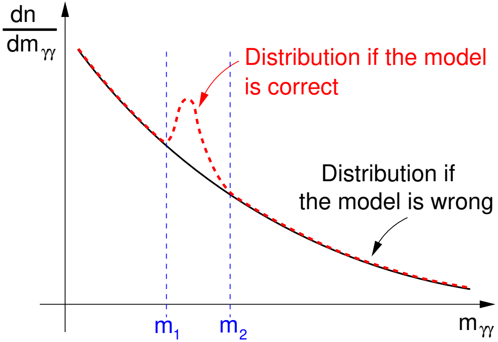
with a small value of α and simultaneously

$$p(\text{data} \in D \setminus W | H_1) = \beta$$

with the smallest possible β.

Introductory example of hypothesis testing

A theoretical model predicts the existence of a particle with mass M, the production cross-section, and the partial width for decay into a photon pair. To confirm or refute this model, one must examine the distribution of $m_{\gamma\gamma}$.

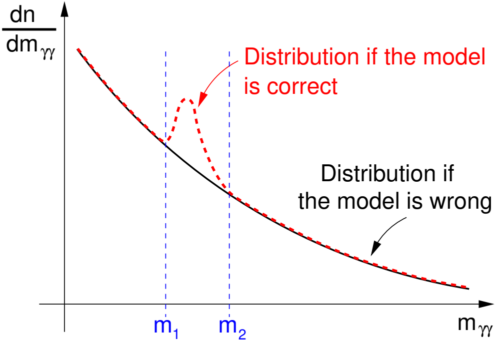


In the interval $[m_1, m_2]$, one is sensitive to the model's prediction. There are two hypotheses, namely that the theory is correct or incorrect.

- H_0: Null hypothesis: TTheory is incorrect.“
- H_1: Alternative hypothesis: TTheory is correct.“

With a sufficiently large amount of data, the probability that the measured $m_{\gamma\gamma}$ distribution looks like H_0 is small if the theory is correct. At the same time, the probability that the measured mass distribution looks like H_1 is large.

Introductory example of hypothesis testing



n: Number of events measured in the interval $[m_1, m_2]$.
One must now choose a threshold value N such that

$$p(n > N|H_0) = \alpha$$

with a small value of α and

$$p(n \leq N|H_1) = \beta$$

is as small as possible if the theory, i.e., H_1, is correct.

n: Number of events measured in the interval $[m_1, m_2]$.
One must now choose a threshold value N such that

$$p(n > N|H_0) = \alpha$$

with a small value of α and

$$p(n \leq N|H_1) = \beta$$

is as small as possible if the theory, i.e., H_1, is correct.

Experimental Practice

- $\alpha = 5.7 \cdot 10^{-7}$, which corresponds to 5$\sigma$ of a normal distribution, to claim the discovery of a particle.
- With a value of $\alpha = 0.3\%$, which corresponds to 3σ of a normal distribution, one says there is evidence for the existence of a new particle.

Type I and type II errors

The confidence level α is defined as the probability that $x \in W$ if the null hypothesis H_0 is correct:

$$p(x \in W | H_0) = \alpha.$$

The probability β represents the likelihood of incorrectly rejecting the alternative hypothesis H_1:

$$p(x \in D \setminus W | H_1) = \beta.$$

Approach	H_0 correct	H_1 correct
$x \notin W \Rightarrow H_0$ is considered correct	Good acceptance, since $p(x \in D \setminus W | H_0) = 1 - \alpha$ is large	Contamination Type II error $p(x \in D \setminus W | H_1) = \beta$.
$x \in W \Rightarrow H_0$ is rejected, H_1 is considered correct	Wrong decision Type I error $p(x \in W | H_0) = \alpha$ is small	Rejecting H_0 good, since $p(x \in W | H_1) = 1 - \beta$ is large.