CMS Statistical analysis tool COMBINE

Denys Lontkovskyi

1. Розподіли ймовірності
2. Статистичне тестування гіпотез
3. CMS combine

Література

Викладки та ілюстрації базується на матеріалах з

- Luca Lista. Statistical Methods for Data Analysis With Applications in Particle Physics. (Lecture Notes in Physics)
- Препринт. The CMS Collaboration. The CMS statistical analysis and combination tool: COMBINE (arxiv.org)
- CMS Higgs boson observation statistical model (repository.cern)

Installing

- Docker or podman
  - Docker Desktop: The #1 Containerization Tool for Developers | Docker
  - podman/docs/tutorials/podman-for-windows.md at main · containers/podman (github.com)
  - Requires WSL2 on Windows
  - Requires xqartz and socat packages in MacOS (see next slide)

```
>
docker run -it --rm -e DISPLAY=$DISPLAY -v /tmp/.X11-unix:/tmp/.X11-unix gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1
```

or

```
>
podman run -it gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1
```

Installing on MacOS

- Follow instructions at [Running GUI's with Docker on Mac OS X | by Nils De Moor | Containerizers (cntnr.io)]

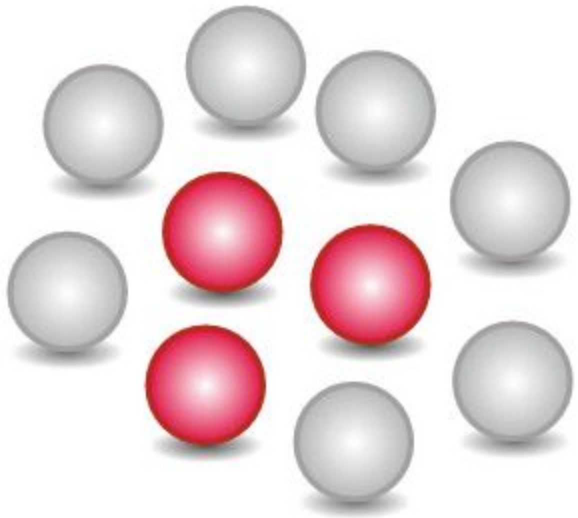
```
> brew install socat

> socat TCP-LISTEN:6000,reuseaddr,fork UNIX-CLIENT:\"$DISPLAY\"

> brew install xquartz

> open -a Xquartz

> ifconfig en0

      en0:

      …

      inet 192.168.0.235 netmask 0xffffff00 broadcast 192.168.199.255
```

- Use IP address of the server in your *docker run* commands, e.g.
- docker run -it --rm -e DISPLAY=192.168.0.235:0 -v /tmp/.X11-unix:/tmp/.X11-unix gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1

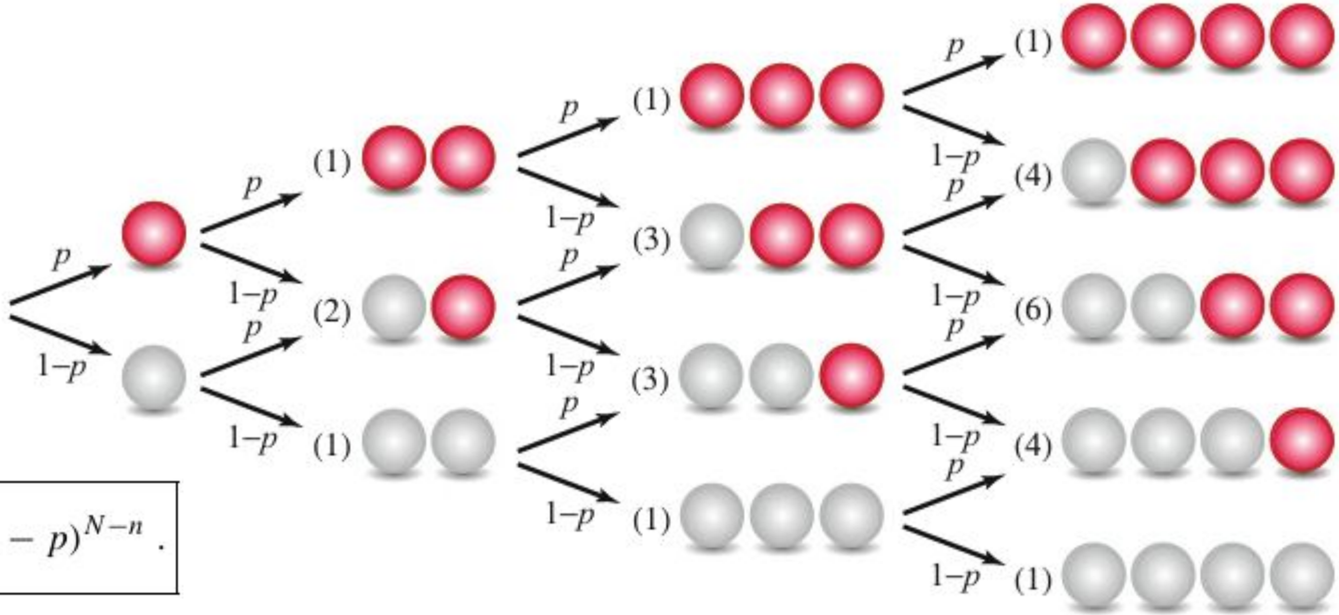
Bernoulli distribution

$$\begin{cases} P(1) = p \,, \\ P(0) = 1 - p \,. \end{cases} \tag{2.55}$$

Fig. 2.1 A set of $R = 3$ red balls plus $W = 7$ white balls considered in a Bernoulli process. The probability to randomly extract a red ball is $p = R/(R + W) = {}^3\!/_{10} = 30\%$



Binomial distribution



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

Fig. 2.2 Binomial process represented as subsequent random extractions of a single red or white ball (Bernoulli process). The tree shows all the possible combinations at each extraction step. Each branching has a corresponding probability equal to p or $1 - p$ for a red or white ball, respectively. The number of paths corresponding to each possible combination is shown in parentheses and is equal to the binomial coefficient in Eq. (2.60)

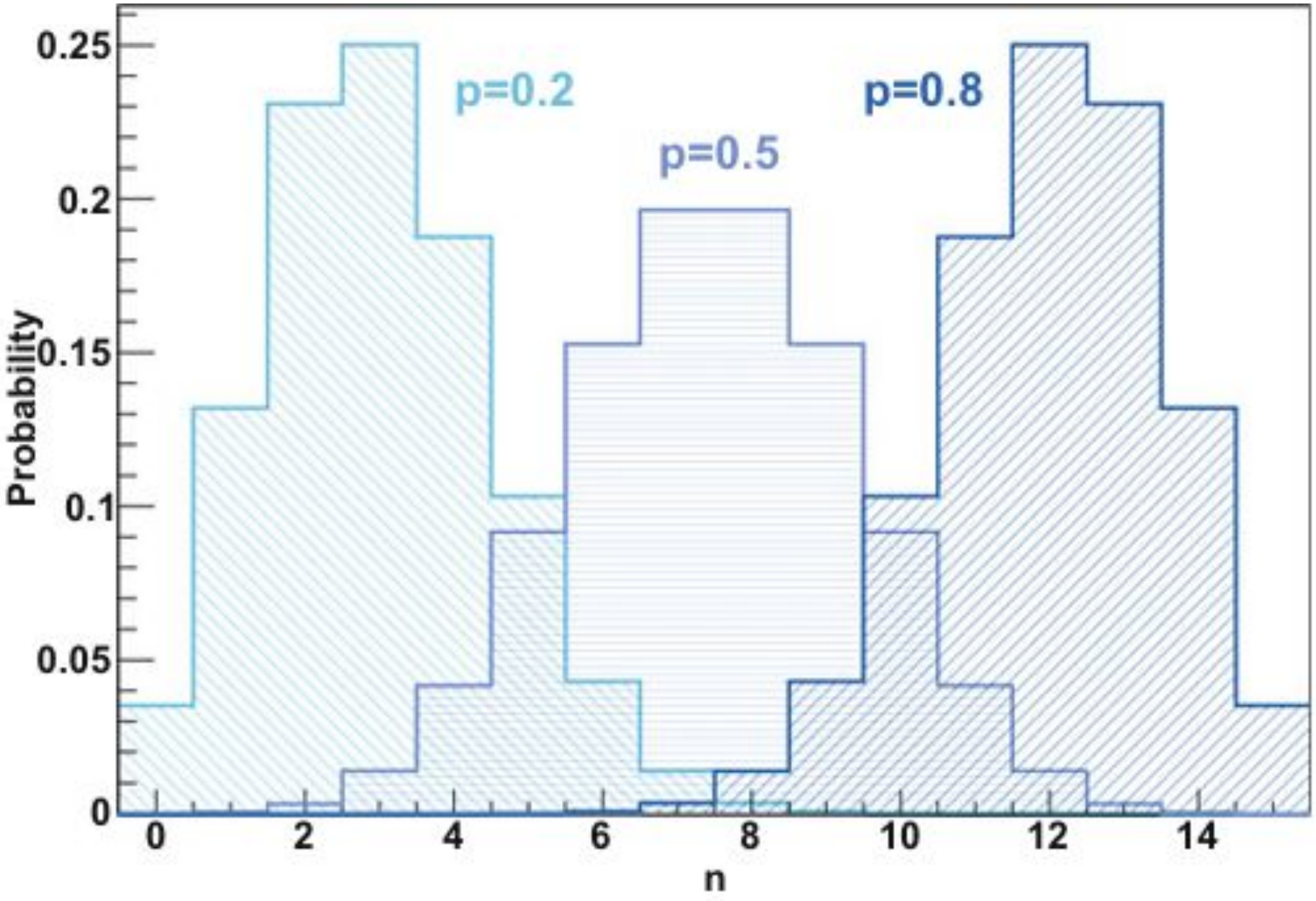


Fig. 2.4 Binomial distributions for $N = 15$ and for $p = 0.2$, 0.5 and 0.8

Poisson distribution is a limit of binomial distribution

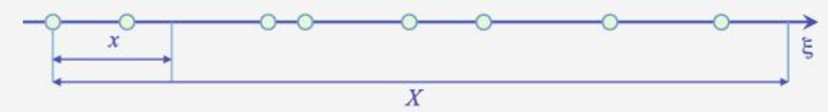
$N \to \infty$

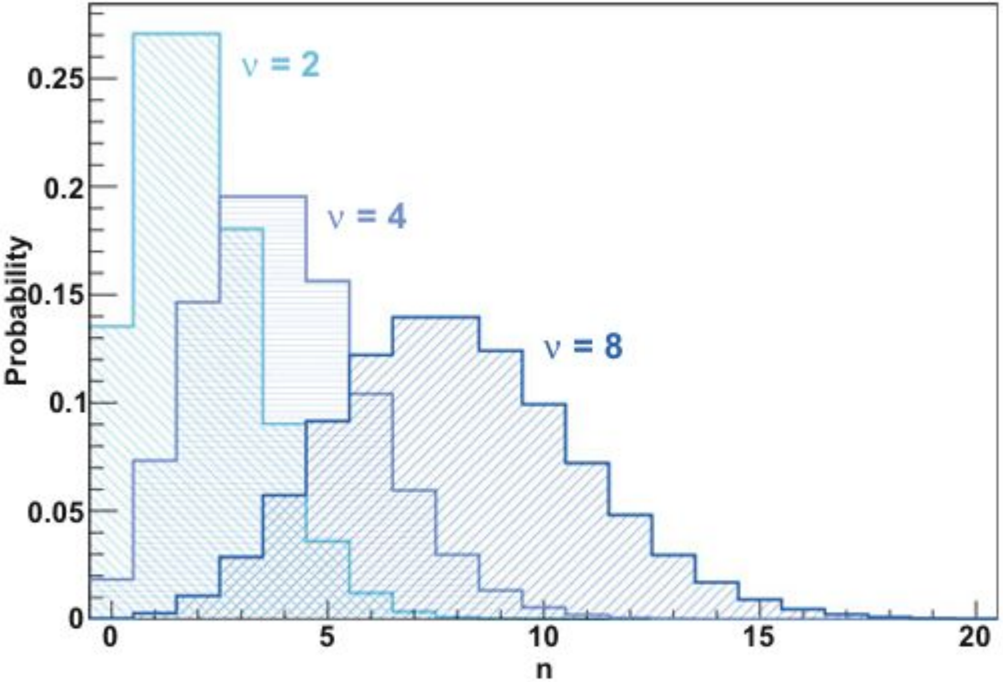
$p \to 0$

$Np \to \nu$

$$\boxed{P(n;\ \nu) = \frac{\nu^n e^{-\nu}}{n!}\ .}$$

Poisson distribution is also obtained from sampling n uniformly distributed points from [0;x) in the limit $N \to \infty$ $X \to \infty$, but fixed density $N/X = r$, $\nu = \langle n \rangle = Nx/X = rx$





Poisson distributions with different value of the rate parameter ν

Large ν limit converges to Gaussian distribution

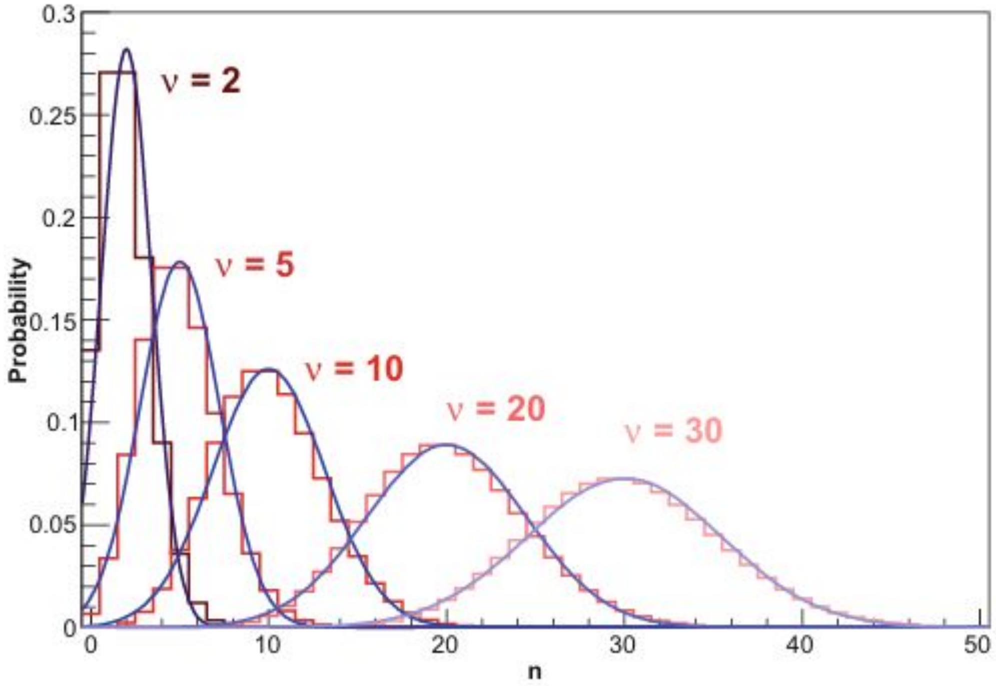


Fig. 2.6 Poisson distributions with different value of the parameter ν compared with Gaussian distributions with $\mu = \nu$ and $\sigma = \sqrt{\nu}$

X^2 distribution

- Sum of independent normally distributed random variables

$$\chi^2 = \sum_{j=1}^{k} z_i^2 \, .$$

$$\chi^2 = \sum_{j=1}^{k} \frac{(x_j - \mu_j)^2}{\sigma_j^2} \, .$$

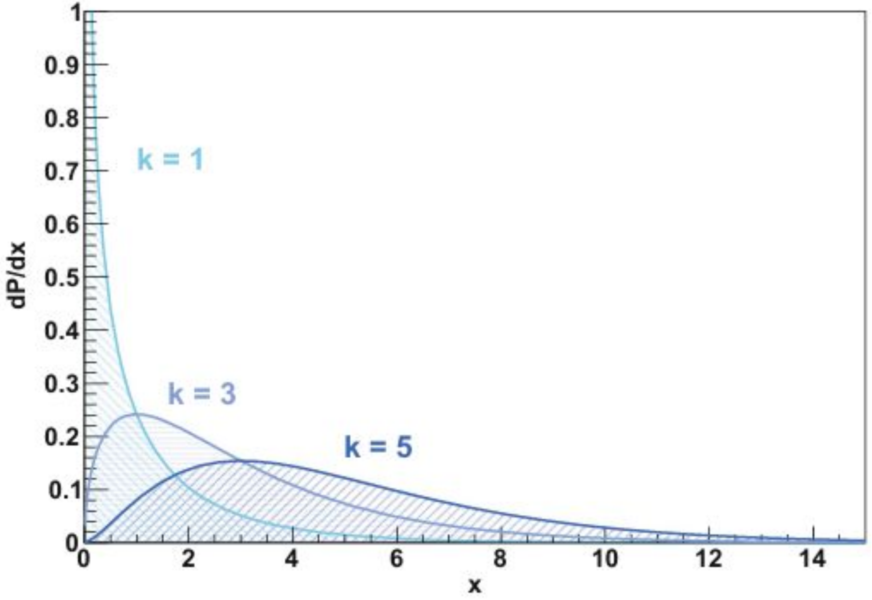


Fig. 3.3 χ^2 distributions with different numbers of degrees of freedom k

Likelihood function

- The joint probability density function of the random variables x_1, x_2, … x_n

$$L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m) = \frac{\mathrm{d}P(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)}{\mathrm{d}x_1 \cdots \mathrm{d}x_n} .$$

- In case of independent observations, likelihood is a product of individual PDFs for each observation

$$\boxed{L(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta}) = \prod_{i=1}^{N} p(\vec{x}_i; \vec{\theta}) .}$$

- It is more convenient to work with sums rather than products, therefore

$$-\log L(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta}) = -\sum_{i=1}^{N} \log p(\vec{x}_i; \vec{\theta}) .$$

Neyman Confidence Intervals

Two steps:

1. **Construction of a confidence belt**
2. Inversion of the confidence belt

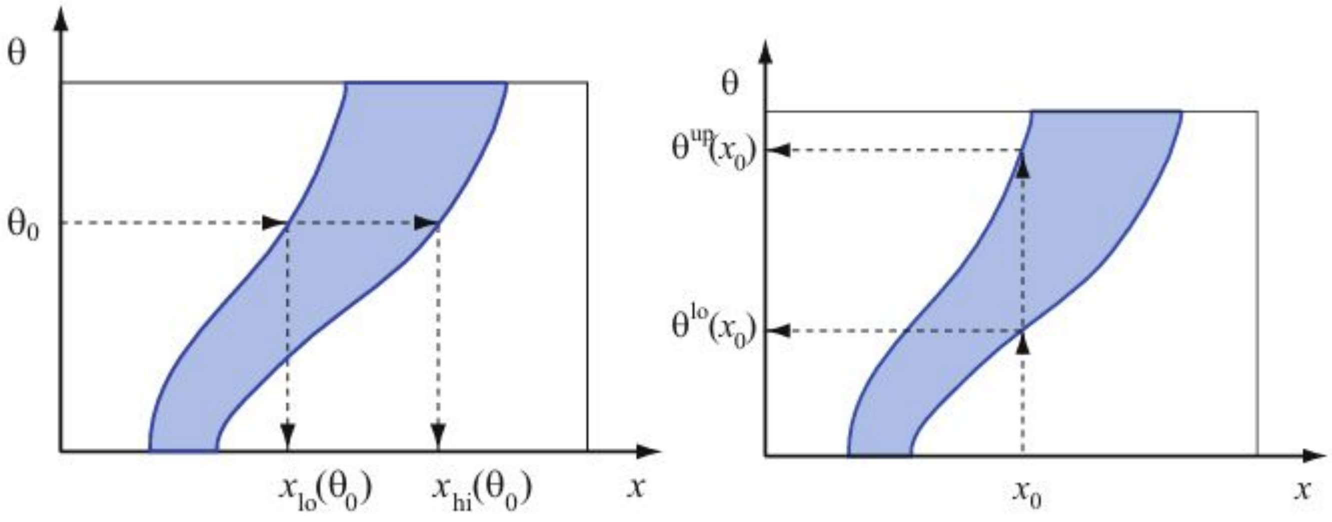


Fig. 8.1 Illustration of Neyman belt construction (left) and inversion (right)

$$1 - \alpha = \int_{x^{\,\mathrm{lo}}(\theta_0)}^{x^{\,\mathrm{up}}(\theta_0)} f(x \mid \theta_0)\, \mathrm{d}x \ .$$

Neyman Confidence Intervals contd.

Two steps:

1. Construction of a confidence belt
2. **Inversion of the confidence belt**

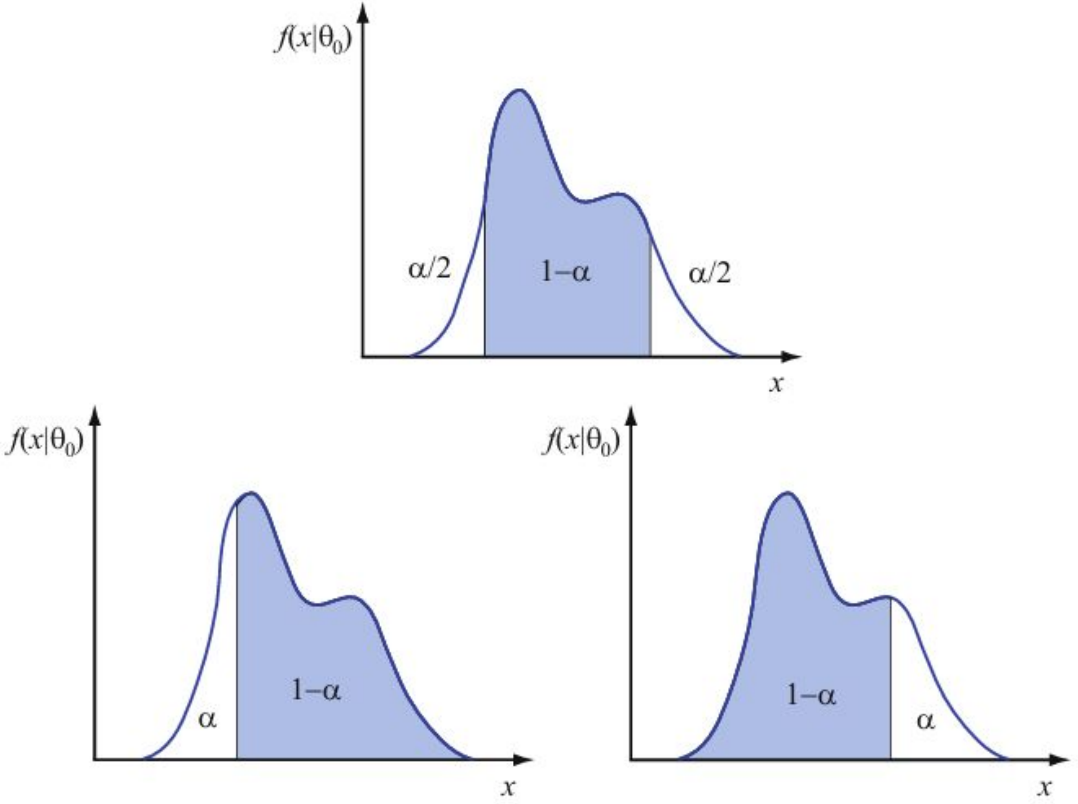


Fig. 8.2 Three possible choices of ordering rule: central interval (top) and fully asymmetric intervals (bottom left, right)

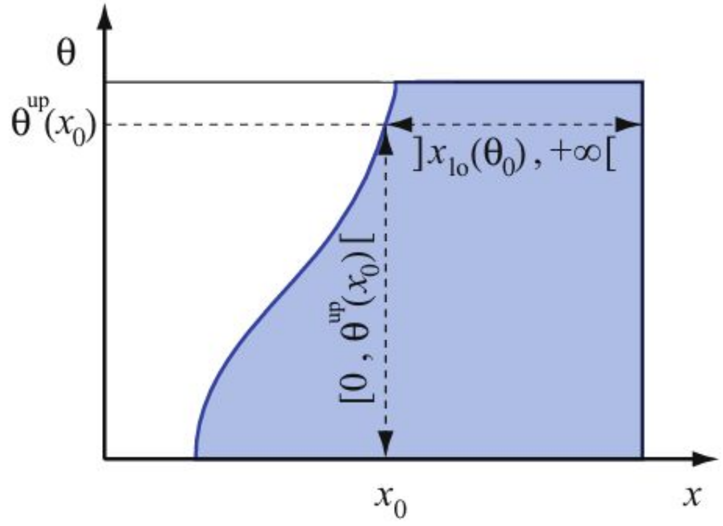


Fig. 12.3 Illustration of Neyman belt construction for upper limits determination

$$P(n;\ s) = \frac{e^{-s}s^n}{n!}\,. \qquad p = P(0;\ s) = e^{-s}, \quad p = e^{-s} > \alpha \quad s < -\log\alpha = s^{\mathrm{up}}\,.$$

$$s < 3.00 \text{ at } 95\% \text{ CL}\,,$$

$$s < 2.30 \text{ at } 90\% \text{ CL}\,.$$

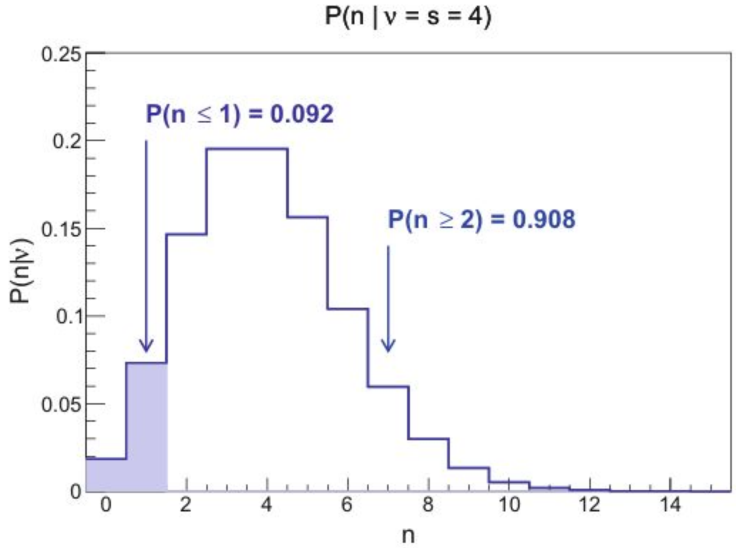
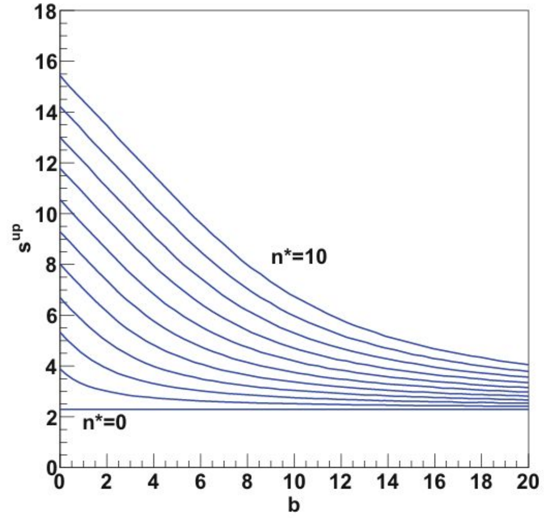


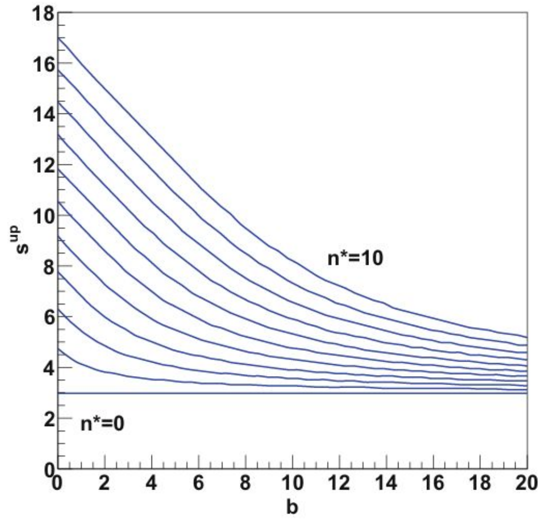
Fig. 12.4 Poisson distribution of the total number of counts $n = s + b$ for $s = 4$ and $b = 0$. The white bins show the smallest possible fully asymmetric confidence interval, $\{2, 3, 4, \cdots\}$ in this case, that gives at least the required coverage of 90%

$$\boxed{\alpha = e^{-s^{\mathrm{up}}}\frac{\sum_{m=0}^{n^\star}(s^{\mathrm{up}}+b)^m/m!}{\sum_{m=0}^{n^\star}b^m/m!}}\,,$$

Upper limits in the presence of negligible background

	$1-\alpha = 90\%$	$1-\alpha = 95\%$
$n^\star$	s^{up}	s^{up}
0	2.30	3.00
1	3.89	4.74
2	5.32	6.30
3	6.68	7.75
4	7.99	9.15
5	9.27	10.51
6	10.53	11.84
7	11.77	13.15
8	12.99	14.43
9	14.21	15.71
10	15.41	19.96





Hypothesis testing

According to Neyman-Pearson lemma, the likelihood ratio

$$\lambda(\vec{x}\,) = \frac{L(\vec{x} \mid H_1)}{L(\vec{x} \mid H_0)} \,.$$

Achieves

Largest **signal selection efficiency**

1 − β for fixed **background misidentification probability α**

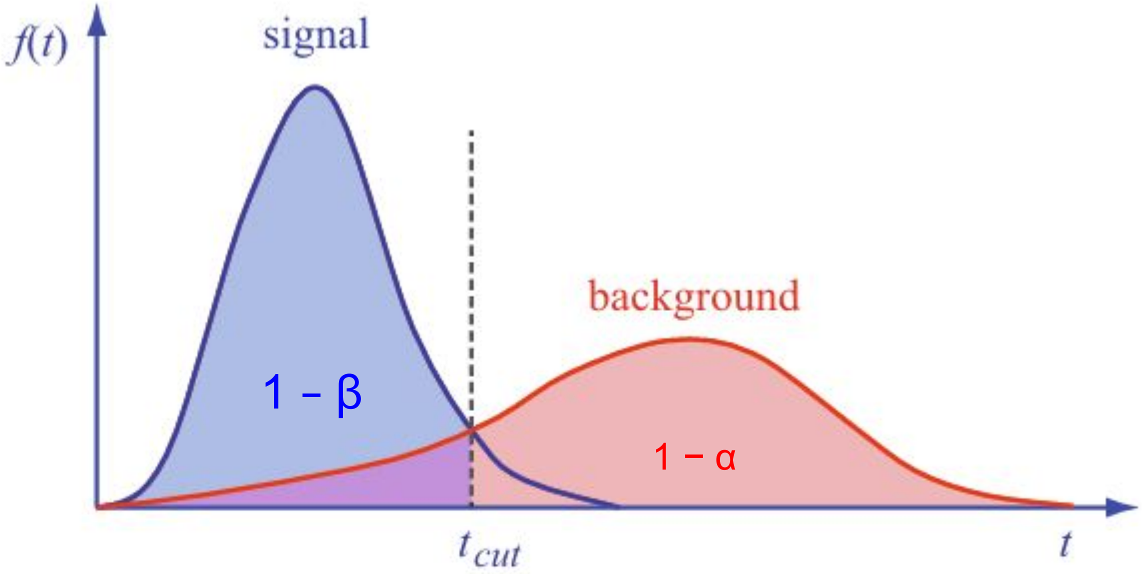
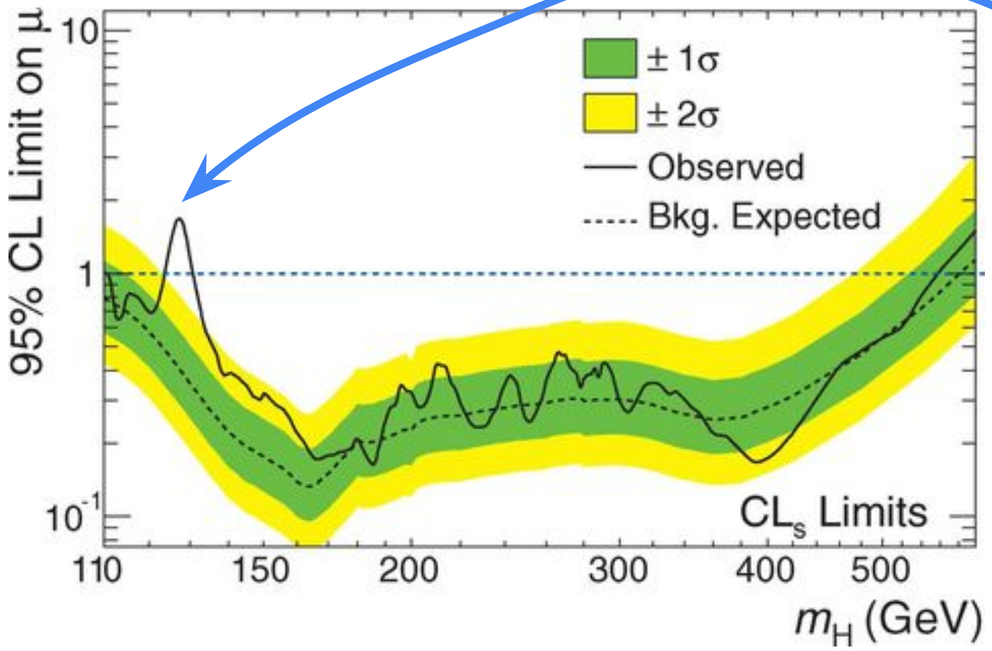
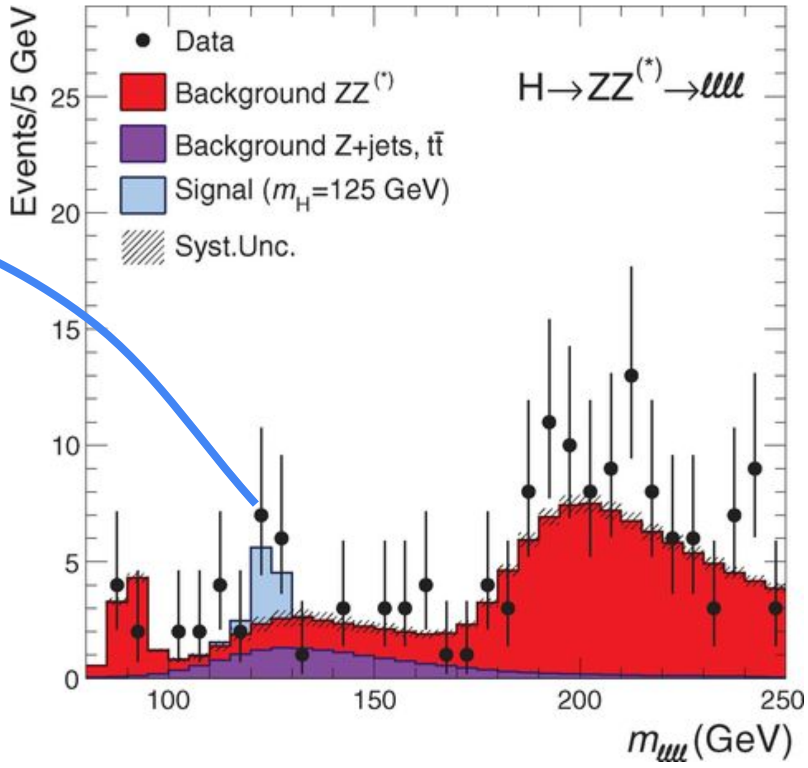


Fig. 10.1 Distribution of a test statistic *t* according to two possible PDFs for the signal (blue) and background (red) hypotheses under test

Brazilian plot

Typical illustration of upper limits in HEP papers





Combine tool

[Combine (cms-analysis.github.io)](https://cms-analysis.github.io)

The package, originally designed to perform searches for a Higgs boson and the combined analysis of those searches, has evolved to become the statistical analysis tool presently used in the majority of measurements and searches performed by the CMS Collaboration

Counting experiment

- docker run [--platform linux/amd64] -it gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1
- combine data/tutorials/CAT23001/datacard-1-counting-experiment.txt

- LHC-style: `--LHCmode LHC-limits`. The test statistic is defined using a ratio of profile likelihoods,

$$\tilde{q}_{\mathrm{LHC}}(\mu) = \begin{cases} -2\ln\left(\dfrac{\mathcal{L}(\mu,\hat{\hat{\nu}}(\mu))}{\mathcal{L}(\hat{\mu},\hat{\nu})}\right) & \text{if } 0 \leq \hat{\mu} \leq \mu, \\ -2\ln\left(\dfrac{\mathcal{L}(\mu,\hat{\hat{\nu}}(\mu))}{\mathcal{L}(0,\hat{\hat{\nu}}(0))}\right) & \text{if } \hat{\mu} < 0, \\ 0 & \text{if } \hat{\mu} > \mu, \end{cases} \tag{24}$$

Wilk’s theorem

- Wilk’s theorem ensures that in the limit of the test statistic follows χ^2 distribution with 1 degree of freedom

$$\chi_r^2 = -2 \log \frac{\sup_{\vec{\theta} \in \Theta_0} \prod_{i=1}^{N} L(\vec{x}_i \,;\, \vec{\theta}\,)}{\sup_{\vec{\theta} \in \Theta_1} \prod_{i=1}^{N} L(\vec{x}_i \,;\, \vec{\theta}\,)} .$$

- This property is applied in combine AsymptoticLimit option

Applying Wilk's theorem to searches test statistic

$$L(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta}) = \frac{e^{-\nu(\vec{\theta})}\nu(\vec{\theta})^N}{N!} \prod_{i=1}^{N} f(\vec{x}_i; \vec{\theta}),$$

$$f(\vec{x}; \vec{\theta}) = \frac{\mu s}{\mu s + b} f_s(\vec{x}; \vec{\theta}) + \frac{b}{\mu s + b} f_b(\vec{x}; \vec{\theta}).$$

$$L_{s+b}(\vec{x}_1, \cdots, \vec{x}_N; \mu, \vec{\theta}) = \frac{e^{-(\mu s(\vec{\theta}) + b(\vec{\theta}))}}{N!} \prod_{i=1}^{N} (\mu s f_s(\vec{x}_i; \vec{\theta}) + b f_b(\vec{x}_i; \vec{\theta})).$$

$$L_b(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta}) = \frac{e^{-b(\vec{\theta})}}{N!} \prod_{i=1}^{N} b f_b(\vec{x}_i; \vec{\theta}). \tag{10.33}$$

Applying Wilk's theorem to searches test statistic

$$\lambda(\mu, \vec{\theta}) = \frac{L_{s+b}(\vec{x}_1, \cdots, \vec{x}_N; \mu, \vec{\theta})}{L_b(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta})} =$$

$$= \frac{e^{-(\mu s(\vec{\theta})+b(\vec{\theta}))}}{e^{-b(\vec{\theta})}} \prod_{i=1}^{N} \frac{\mu s f_s(\vec{x}_i; \vec{\theta}) + b f_b(\vec{x}_i; \vec{\theta})}{b f_b(\vec{x}_i; \vec{\theta})} =$$

$$= e^{-\mu s(\vec{\theta})} \prod_{i=1}^{N} \left(\frac{\mu s f_s(\vec{x}_i; \vec{\theta})}{b f_b(\vec{x}_i; \vec{\theta})} + 1 \right) .$$

$$-\log \lambda(\mu, \vec{\theta}) = \mu s(\vec{\theta}) - \sum_{i=1}^{N} \log \left(\frac{\mu s f_s(\vec{x}_i; \vec{\theta})}{b f_b(\vec{x}_i; \vec{\theta})} + 1 \right) .$$

Applying Wilk's theorem to searches test statistic

- In case of counting experiment f_s and f_b terms are dropped and the expression simplifies to

$$\lambda(\vec{\theta}\,) = \frac{e^{-(\mu s(\vec{\theta}\,)+b(\vec{\theta}\,))}(\mu s(\vec{\theta}\,)+b(\vec{\theta}\,))^N}{N!}\frac{N!}{e^{-b(\vec{\theta}\,)}b(\vec{\theta}\,)^N} =$$

$$= e^{-\mu s(\vec{\theta}\,)}\left(\frac{\mu s(\vec{\theta}\,)}{b(\vec{\theta}\,)}+1\right)^N .$$

$$-\log\lambda(\vec{\theta}\,) = \mu s(\vec{\theta}\,) - N\log\left(\frac{\mu s(\vec{\theta}\,)}{b(\vec{\theta}\,)}+1\right) ,$$

Example datacard for counting experiment

```
imax 1
jmax 2
kmax 3
# A single channel - ch1 - in which 0 events are observed in data
bin               ch1
observation       0
# ----------
bin               ch1   ch1   ch1
process           ppX   WW    tt
process           0     1     2
rate              1.47  0.64  0.22
# ----------
lumi     lnN      1.11  1.11   1.11
xs       lnN      1.20    -      -
nWW      gmN 4      -   0.16     -
```

combine data/tutorials/CAT23001/datacard-1-counting-experiment.txt

Example datacard for template shapes analysis

```
imax 1
jmax 1
kmax 4
# --------------
shapes * * template-analysis-datacard-input.root $PROCESS $PROCESS_$SYSTEMATIC
# --------------
bin          ch1
observation 85
# ------------------------------
bin             ch1        ch1
process         signal     background
process         0          1
rate            24         100
# --------------------------------
lumi     lnN    1.1       1.0
bgnorm   lnN    -         1.3
alpha   shape   -          1   # uncertainty in the background template.
sigma   shape   0.5        -   # uncertainty in the signal template.
```

combine data/tutorials/CAT23001/datacard-2-template-analysis.txt

BONUS: CMS higgs observation statistical analysis

- Exit container and check container name
  - Ctrl+D
  - docker container ls --all # find the name of your container

```
→ hep_lectures docker container ls --all
CONTAINER ID   IMAGE                                                       COMMAND                  CREATED              STATUS                       PORTS     NAMES
efdaa9c5aa32   gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1   "/bin/bash -l -c /bi…"   About a minute ago   Exited (0) About a minute ago          compassionate_keldysh
a735f8717912   gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1   "/bin/bash -l -c /bi…"   3 hours ago          Exited (127) 5 minutes ago             optimistic_kepler
7d8a506d58ec   gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1   "/bin/bash -l -c /bi…"   3 hours ago          Exited (130) 3 hours ago               dazzling_jennings
```

- Copy datacards into container
  - wget https://repository.cern/records/c2948-e8875/files/cms-h-observation-public-v1.0.tar.gz\?download\=1
  - mv cms-h-observation-public-v1.0.tar.gz\?download=1 cms-h-observation-public-v1.0.tar.gz
  - docker container cp cms-h-observation-public-v1.0.tar.gz \ <container_name>:/code/HiggsAnalysis/CombinedLimit
  - docker container restart <container_name>
  - docker container attach <container_name>
  - tar zxvf cms-h-observation-public-v1.0.tar.gz
- Podman has equivalent commands

Домашнє завдання

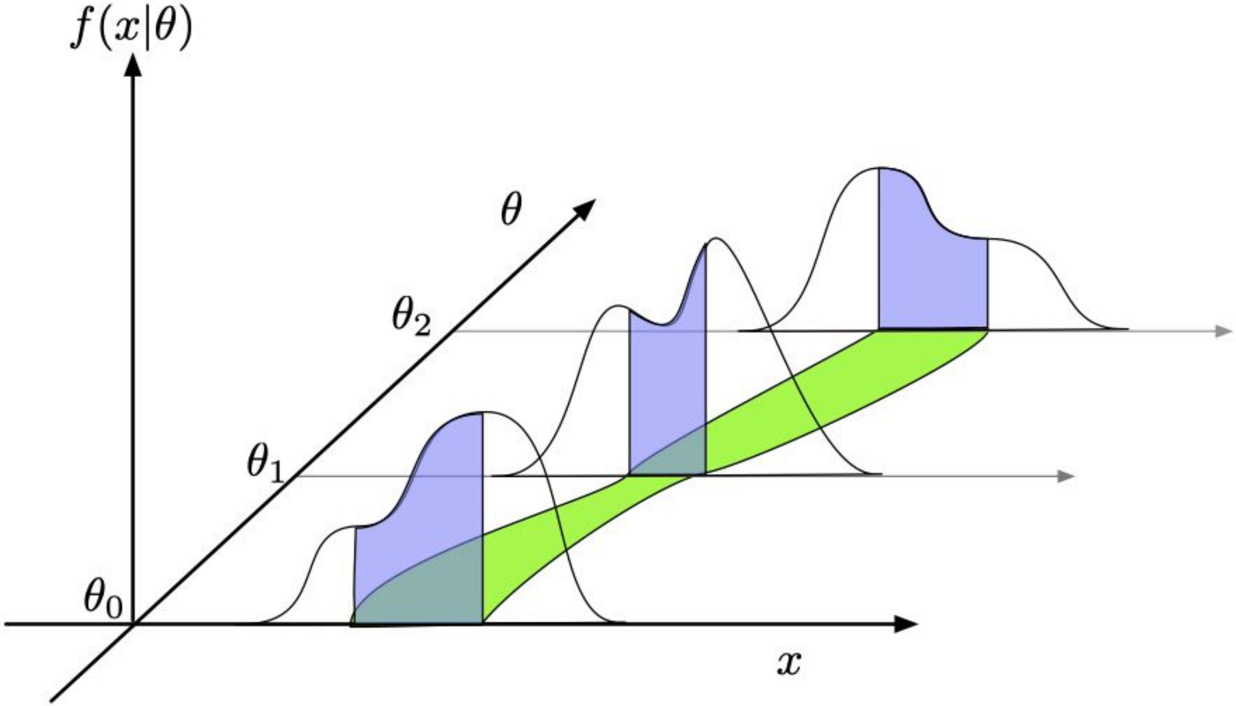
- Створити власну карту для counting experiment
    - Одине джерело фонових подій + сигнал

- Порівняти результати
    - LEP, TEVATRON, LHC upper limits
    - Для різних рівнів фонових подій

BACKUP

NEYMAN CONSTRUCTION EXAMPLE

This makes a **confidence belt** for θ



[Neyman construction — Statistics and Data Science (theoryandpractice.org)]

A RESTATEMENT OF THE CONSTRUCTION

For every point θ, if it were true, the data would fall in its acceptance region with probability $1-\alpha$

If the data fell in that region, the point θ would be in the interval

So the interval $[\theta_-, \theta_+]$ covers the true value with probability $1-\alpha$