Karlsruhe Institute of Technology

Statistische Methoden der Datenanalyse

Hans Dembinski
IEKP, KIT Karlsruhe

Topics for today

- Confidence limits
- Monte-Carlo and resampling methods
- Testing of hypotheses

Confidence limits

Confidence intervals and limits

Confidence intervals from likelihood ratios (see Thursday's lecture) are always two-sided

What about one-sided limits? Fundamental way of constructing an interval?

Two-sided intervals are not unique

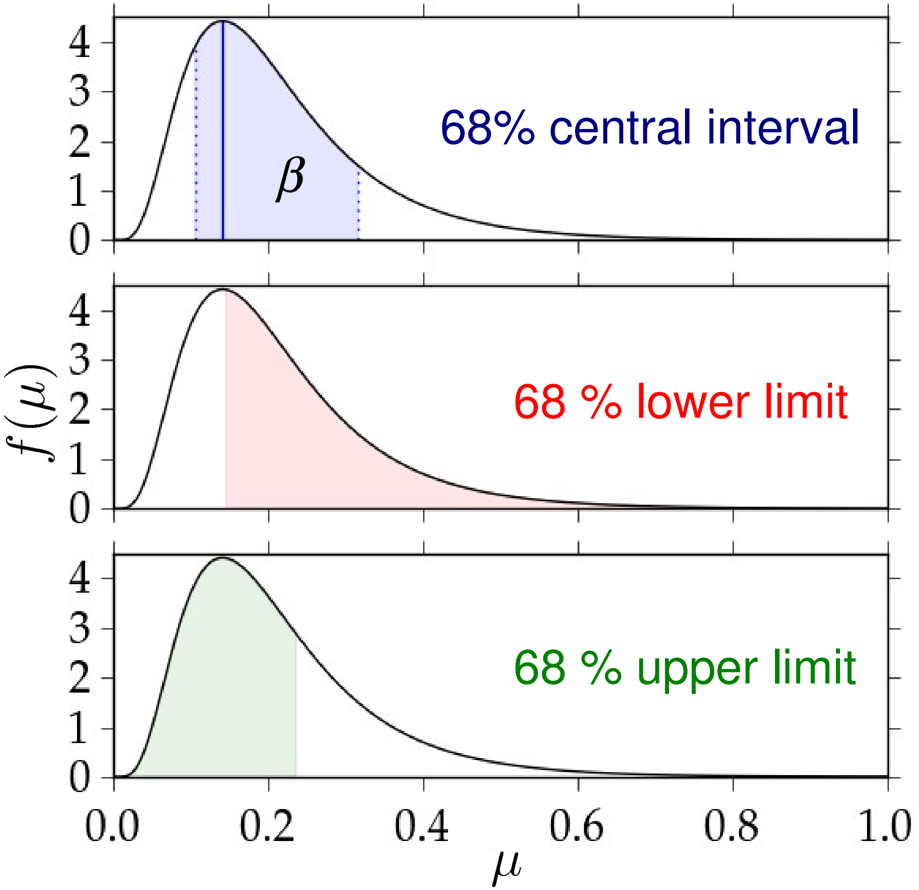
$$\beta = \int_{\mu_l}^{\mu_u} f(\mu)\mathrm{d}\mu$$

Many p_l, p_u give same coverage β

Usual choice: **central interval**

$$\int_{-\infty}^{\mu_l} f(\mu)\mathrm{d}\mu = \int_{\mu_u}^{\infty} f(\mu)\mathrm{d}\mu = (1-\beta)/2$$

No freedom of choice
for upper or lower limit



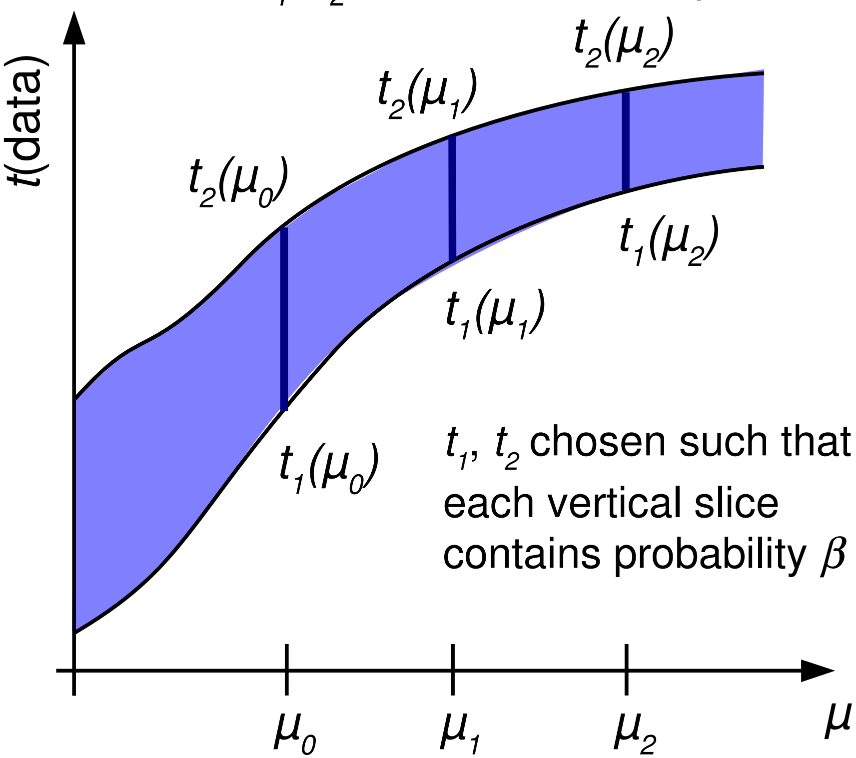
Neyman construction

Let us assume we have a estimator $t(\vec{x})$ of the data $\vec{x}$ with known p.d.f. $f(t|\mu)$

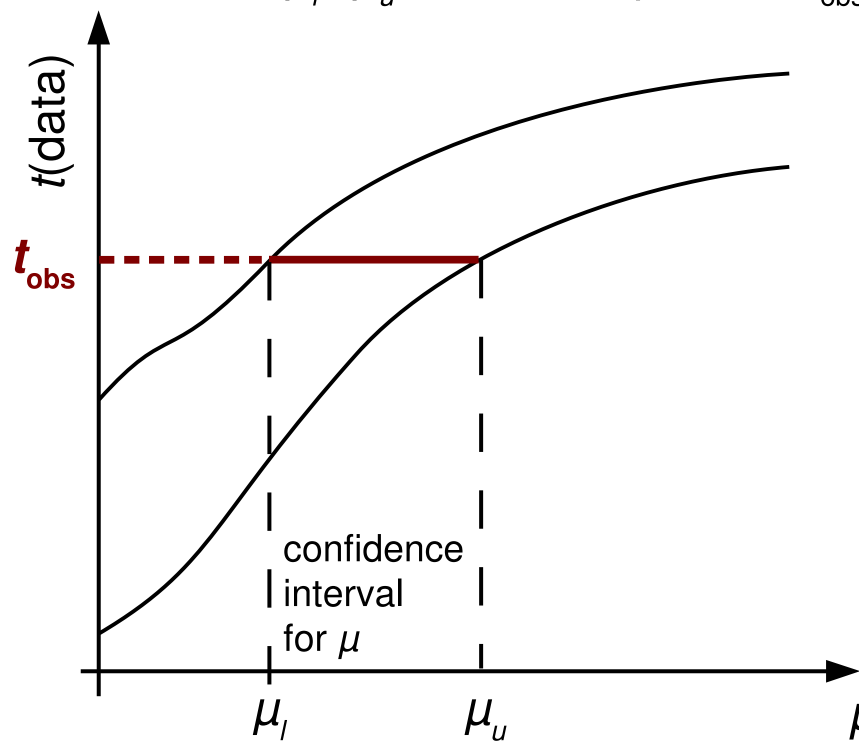
We want to know the confidence interval for μ with coverage $C = \beta$

$$\beta = P[t_1 \le t \le t_2|\mu] = P[t_1(\mu) \le t \le t_2(\mu)] = \int_{t_1}^{t_2} f(t|\mu)\mathrm{d}\mu$$

a) Find t_1, t_2 as a function of μ



b) Find μ_l, μ_u that correspond to t_{obs}

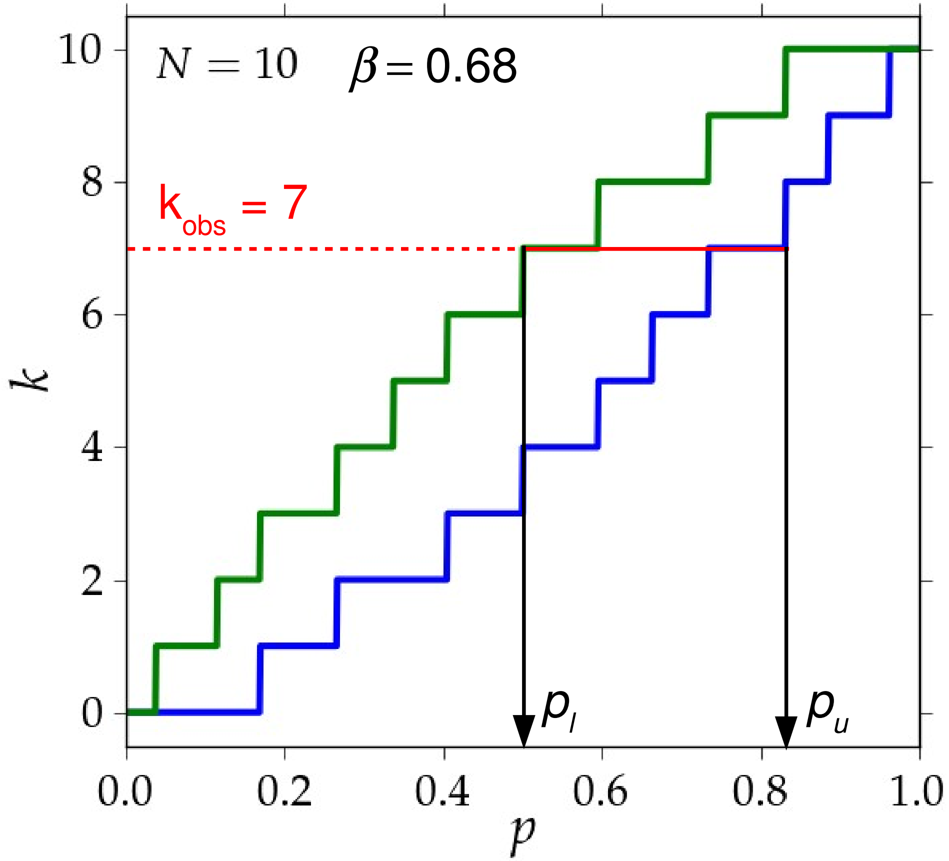


Neyman construction

Neyman construction of confidence intervals for parameter p of binomial distribution

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

Two event classes A, B
Probability $p = P[A] = 1 - P[B]$
$P(k \mid p, N)$ probability of getting k events A out of N total



Neyman-constructed intervals
Note that $p_u > 0$ for $k = 0$
and $p_l < 1$ for $k = N$

Compare with usual method

$$\sigma[p] \approx \frac{\sqrt{V[k]}}{N} = \sqrt{\frac{\frac{k}{N}\left(1-\frac{k}{N}\right)}{N}}$$

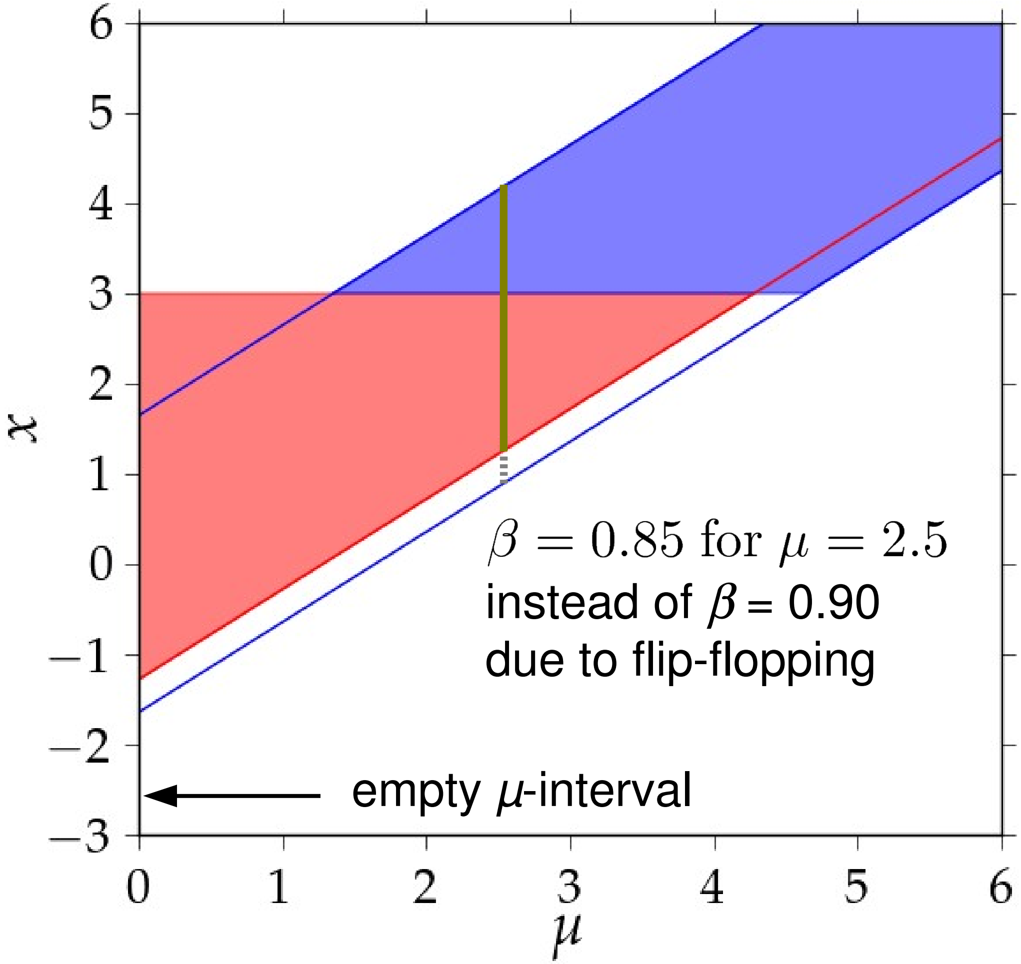
where $\sigma[p] = 0$ for $k = 0$ and $k = N$

Beware:
Discrete distributions $C \geq \beta$
Continuous distributions $C = \beta$

Flip-flopping and empty intervals

Let's regard observation *x* from Normal distribution with $\mu > 0$ (physical constraint)

$\sigma = 1$



Flip-flopping

Two choices for confidence interval, typical approach:

– Give two-sided limit if $x \gg 0$
– Give upper limit if $x \approx 0$

But: Switching method depending on data leads to **under-coverage**

Empty intervals

μ-intervals can be empty for $x \ll \mu$ due to constraint on $\mu > 0$

Solution: Feldman-Cousins limits

Feldman-Cousins limits

Unifies construction of two-sided limits and one-sided limits
Avoids empty intervals

Neyman construction + growth rule
Successively grow *x*-interval at the end with the largest likelihood ratio $L(x|\mu)/L(x|\hat{\mu})$

$$\frac{L(x_1|\mu)}{L(x_1|\hat{\mu})} \qquad \hat{\mu} \qquad \frac{L(x_2|\mu)}{L(x_2|\hat{\mu})}$$

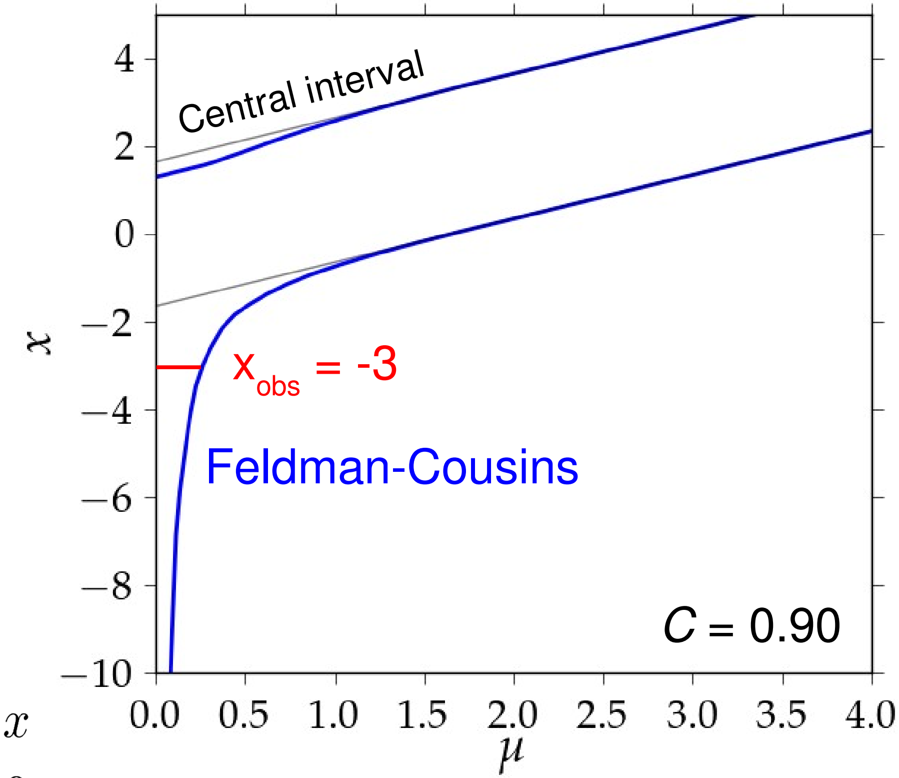
$$x_1, x_1 + \mathrm{d}x \qquad\qquad x_2, x_2 + \mathrm{d}x$$

$\hat{\mu}$ is the maximum likelihood estimate of μ given x under the condition $\hat{\mu} \geq 0$

Example: Normal distribution $\mu > 0$, $\sigma = 1$

$$L(x|\hat{\mu}) = \begin{cases} \frac{1}{\sqrt{2\pi}}, & x \geq 0,\ \hat{\mu} = x \\ \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), & x < 0,\ \hat{\mu} = 0 \end{cases}$$

$$\frac{L(x|\mu)}{L(x|\hat{\mu})} = \begin{cases} \exp(-(x-\mu)^2/2), & x \geq 0 \\ \exp(x\,\mu - \mu^2/2), & x < 0 \end{cases}$$



Feldman-Cousins construction is recommended if you want to report a result close to a physical boundary

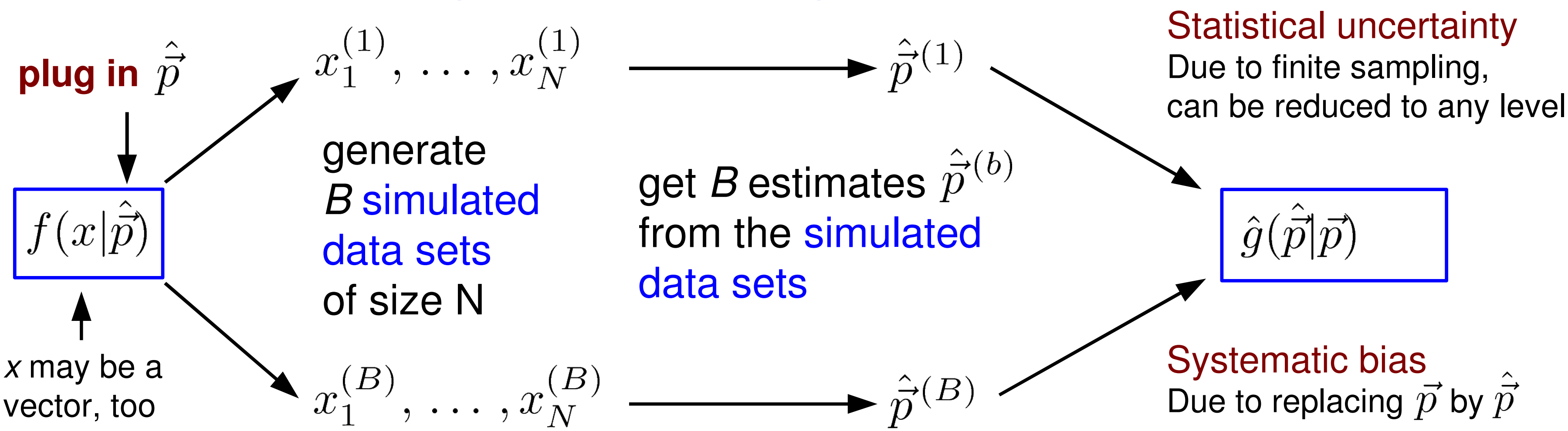
Monte-Carlo and resampling methods

Parametric bootstrap

Let's assume we have the p.d.f. $f(x|\vec{p})$ for an observation *x* given parameters $\vec{p}$ and an estimate $\hat{\vec{p}}$ obtained from *N* observations x_i

We want to know $g(\hat{\vec{p}}|\vec{p})$ or a summary statistic like bias and variance of $\hat{\vec{p}}$

Monte-Carlo method (= parametric bootstrap)



Bias of $\hat{\vec{p}}$

$$\hat{E}[\hat{\vec{p}} - \vec{p}] = \frac{1}{B}\sum_b \hat{\vec{p}}^{(b)} - \hat{\vec{p}}$$

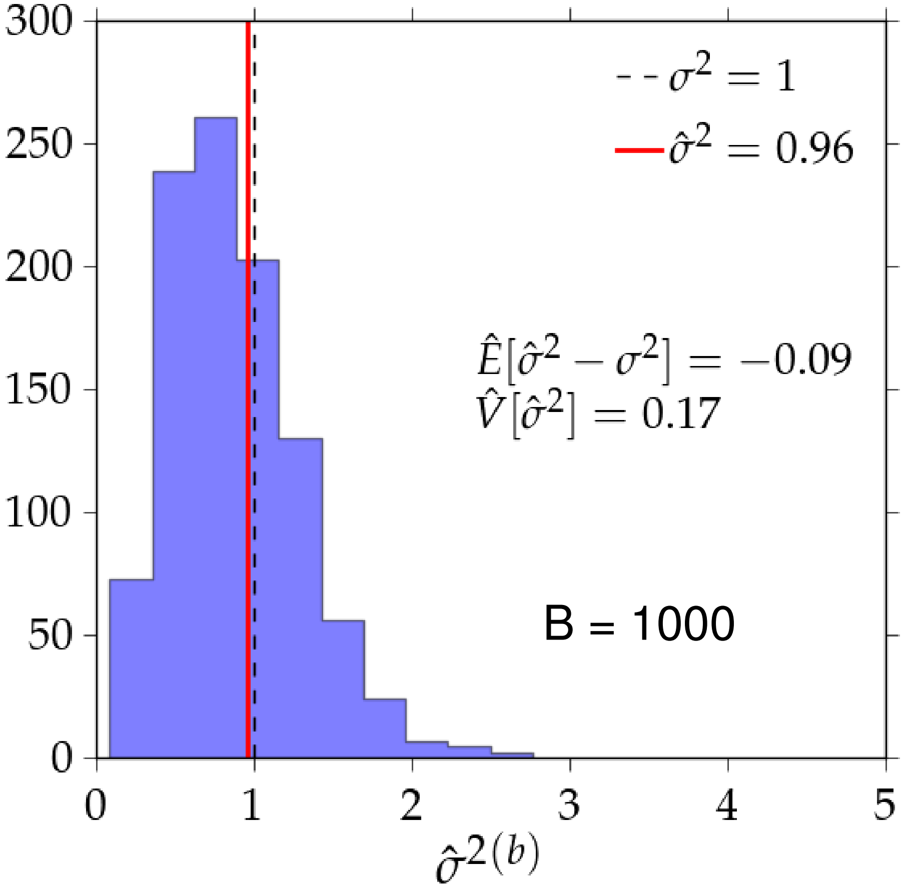
Variance of $\hat{\vec{p}}$

$$\widehat{\mathrm{cov}}[\hat{\vec{p}}]_{ij} = \frac{1}{B-1}\sum_b \hat{p_i}^{(b)}\,\hat{p_j}^{(b)} - \frac{1}{B(B-1)}\left(\sum_b \hat{p_i}^{(b)}\right)^2$$

Parametric bootstrap

Example: Normal distibution $\mu = 0$, $\sigma = 1$, $N = 100$

Study biased estimator $\hat{\sigma}^2 = \frac{1}{N}\sum_i x_i^2 - \frac{1}{N^2}\Big(\sum_j x_j\Big)^2$



Analytical results for normal distribution

$$E[\hat{\sigma}^2 - \sigma^2] = -\frac{\sigma^2}{N} = -0.1$$

$$V[\sigma^2] = 2\sigma^4 \frac{N-1}{N^2} \approx 0.18$$

Parametric bootstrap

+ Bias and variance without analytical effort
+ Works with arbitrarily complex estimators

o Computationally intensive

– Systematic bias can be important if $\hat{\vec{p}}$ is far away from $\vec{p}$

Better performance for large *N*

Random number generation

Scientific programming libraries provide excellent pseudo random number generators

Pseudo random numbers have uniform (flat) distribution, how to get arbitrary $f(\vec{x})$?

a) Transformation method $y = F(x) = \int_{-\infty}^{x} \mathrm{d}x\, f(x) \;\rightarrow\; \boxed{x = F^{-1}(y)}$

↑ (y) follows uniform distribution

Practical only if $F^{-1}(y)$ or a suitable approximation to it is available

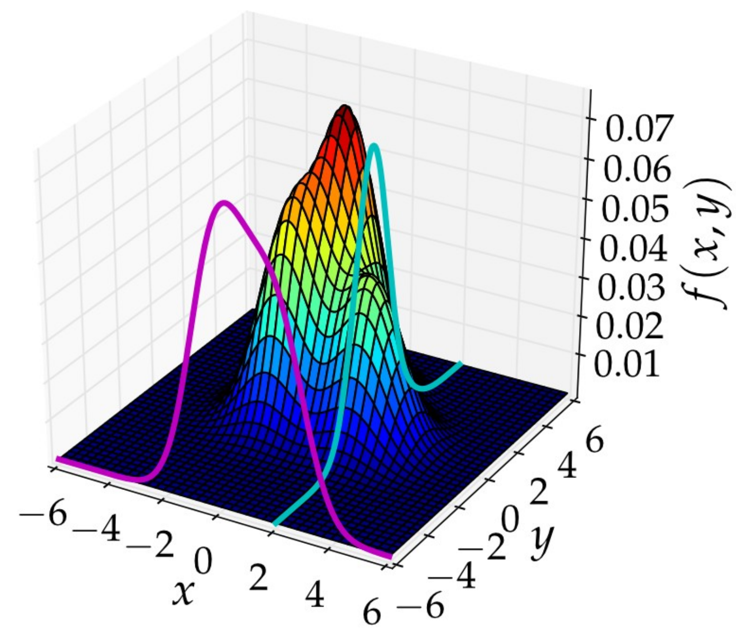
Multivariate case complex, an example in 2d:

Solve in order

$$\int_{-\infty}^{x_0} \mathrm{d}x_0' \int_{-\infty}^{\infty} \mathrm{d}x_1' f(x_0', x_1') = y_0$$

$$\frac{\int_{-\infty}^{x_1} \mathrm{d}x_1' f(x_0, x_1')}{\int_{-\infty}^{\infty} \mathrm{d}x_1' f(x_0, x_1')} = y_1$$

Analog in case of more dimensions



Random number generation

Scientific programming libraries provide excellent pseudo random number generators

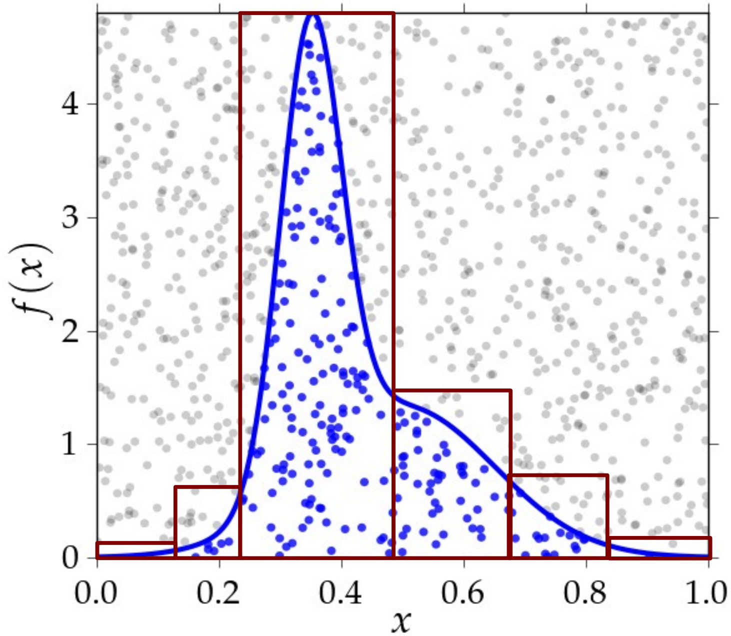
Pseudo random numbers have uniform (flat) distribution, how to get arbitrary $f(\vec{x})$?

b) Accept-Reject method

Construct a (hyper-)rectangle around $f(\vec{x})$ that completely encloses it

Uniformly draw points $(\vec{x}, f')$ from inside the (hyper-)rectangle and accept $\vec{x}$ if $f' < f(\vec{x})$

1d example



+ Very general method
+ Simple to set up
o Need to know max[$f(x)$]
– Inefficient/slow: many points are wasted

Efficiency is greatly improved by sampling from several local boxes

Full bootstrap

What to do if $f(x|\vec{p})$ is unknown?

We could still use the Monte-Carlo method to study an estimator $t(\vec{x})$ of the data $\vec{x}$ if we had an estimate of $f(x)$

Non-parametric maximum-Likelihood estimate of $f(x)$

maximize $\ln \hat{f}(x) = \sum_i \ln \hat{f}(x_i)$ without any further knowledge except $\int_{-\infty}^{\infty} \hat{f}(x) = 1$

non-parametric ML estimate $\hat{f}_B(x) = \frac{1}{N} \sum_i \delta(x - x_i)$

Assumptions: x_i from same $f(x)$, x_i are independent

No proof, but...

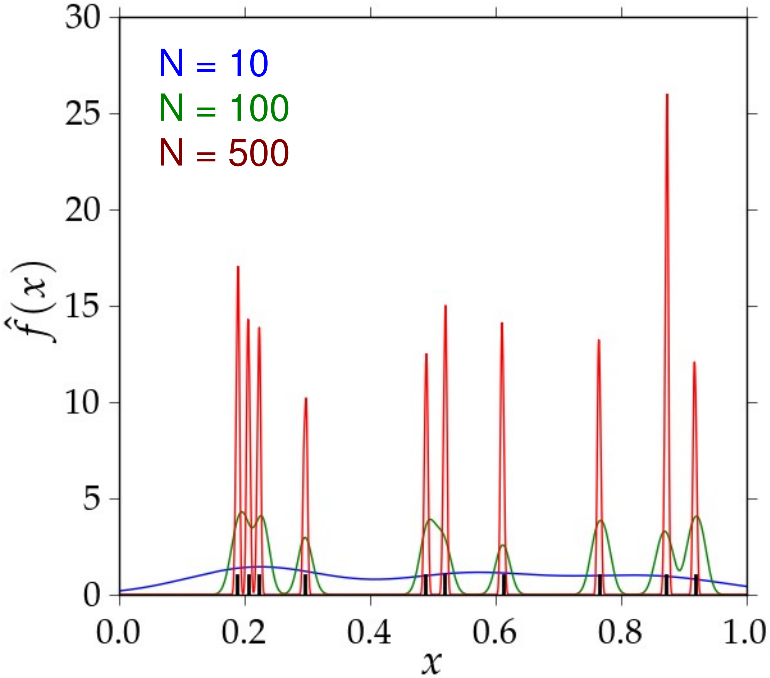
$$f(x|\vec{a}) = \frac{1}{\sum_j a_j} \sum_{k=0}^{K} a_k g(x|\mu_k, \sigma)$$

$$\mu_k = \frac{k}{N-1} \Delta x$$

$g(x|\mu, \sigma)$ Normal p.d.f.

$$\sigma = \frac{1}{N-1} \Delta x$$

converges to $\hat{f}_B(x)$ for $K \to \infty$ (infinite flexibility)



Analytic bootstrap estimates

Plugin principle

Construct bootstrap estimate by replacing true variable in formula by empirical one

$$E_B[x] = \int \mathrm{d}x\, x\, \hat{f}(x) = \int \mathrm{d}x\, x\, \frac{1}{N} \sum_i \delta(x - x_i) = \frac{1}{N} \sum_i x_i$$

sample mean

$$V_B[x] = E_B[x^2] - E_B[x]^2 = \frac{1}{N} \sum_i x_i^2 - \left(\frac{1}{N} \sum_j x_j \right)^2$$

sample variance (biased)

Like any estimator, a bootstrap estimator can be biased if the sample size is small
(Bias can be detected and corrected by a double bootstrap, i.e. bootstrapping the bootstrap)

Two other bootstrap estimates are well known to physicists

Uncertainty of a Poisson count $k \pm \sqrt{k}$: $V[\lambda] = \lambda \rightarrow V_B[\lambda] = k$

Uncertainty of a binomial proportion (e.g. efficiency of a detector):

$$V[k/N] = \frac{p(1-p)}{N} \rightarrow V_B[k/N] = \frac{k/N(1-k/N)}{N}$$

Monte-Carlo bootstrap estimates

Use in Monte-Carlo estimation

Draw random numbers from $\hat{f}(x)$

Pick x_i with equal probability with replacement

Re-examination

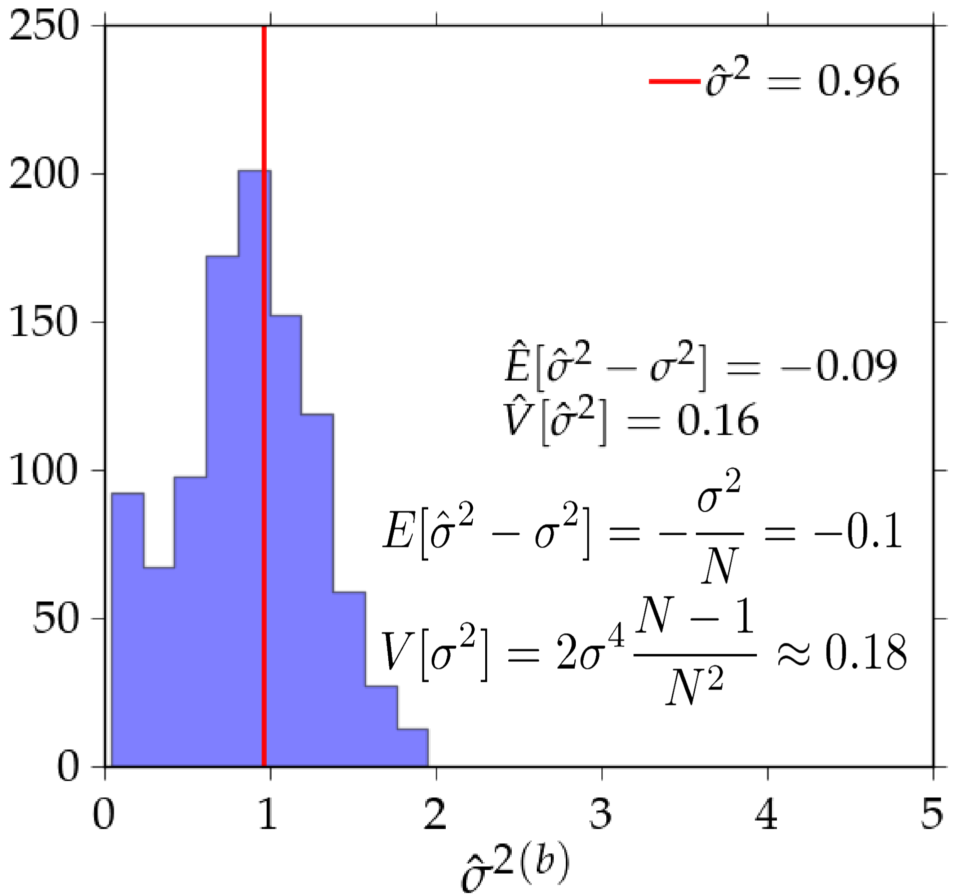
Normal distibution $\mu = 0$, $\sigma = 1$, $N = 100$ and biased estimator

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \frac{1}{N}\sum_j x_j)^2$$

o Same pros/cons as parametric bootstrap

+ Effortless to apply

– Biased for estimators that depend strongly on distribution tails



Other resampling methods

Jackknife – fast approximation to full bootstrap

$$\hat{E}_{\text{jack}}[\hat{p} - p] = \frac{N-1}{N} \sum_j (\hat{p}_{(j)} - \hat{p}) \qquad \hat{V}_{\text{jack}}[\hat{p}] = \frac{N-1}{N} \sum_i \hat{p}^2_{(i)} - \frac{N-1}{N^2} \Big(\sum_j \hat{p}_{(j)} \Big)^2$$

$\hat{p}_{(j)} = t(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_N)$ estimate of *p* without observation x_j

Only needs *N* additional evaluations of $t(\vec{x})$, but less precise

Leave-one-out cross-validation – compare prediction power of models

Can only be used with (*x*,*y*) pairs, *y* = *f(x)*

$$\text{LOOCV} = \sum_i \big(y_i - f_{(i)}(x_i) \big)^2 \propto$$ mean squared error = bias^2 + variance

bias^2 is large if model is not flexible enough, i.e. is missing effects in the data
variance is large if model is too flexible, i.e. “overfitting” the data

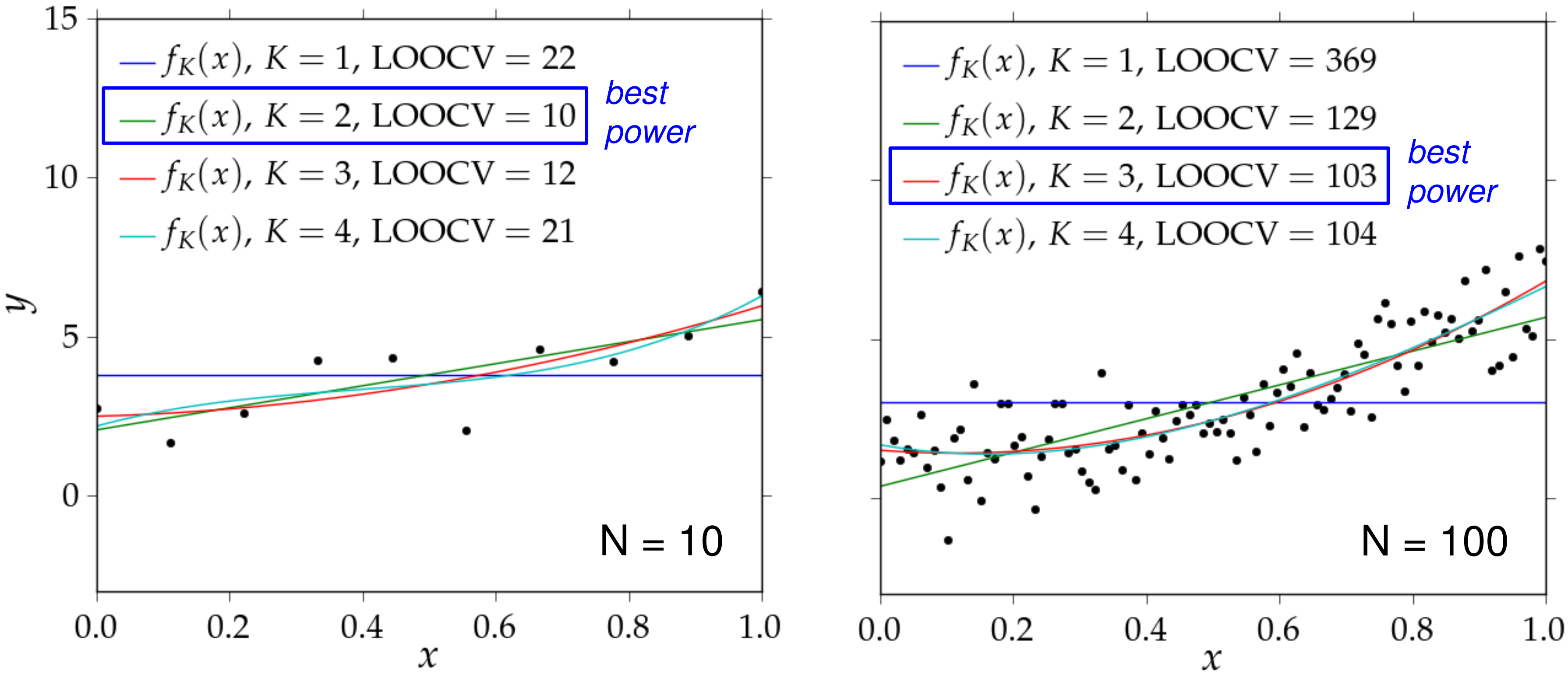
Model with best prediction power has smallest LOOCV value

Maximizing prediction power

Example: fit of a polynomial model

True model $f(x) = 1 + 2x + 3x^2$, $y_{obs} = f(x)$ + Normal fluctuation with $\mu = 0$, $\sigma = 1$

Fitting model $f_K(x) = \sum_{k=0}^{K} p_k\, x^k$, what K to choose if K is unknown?

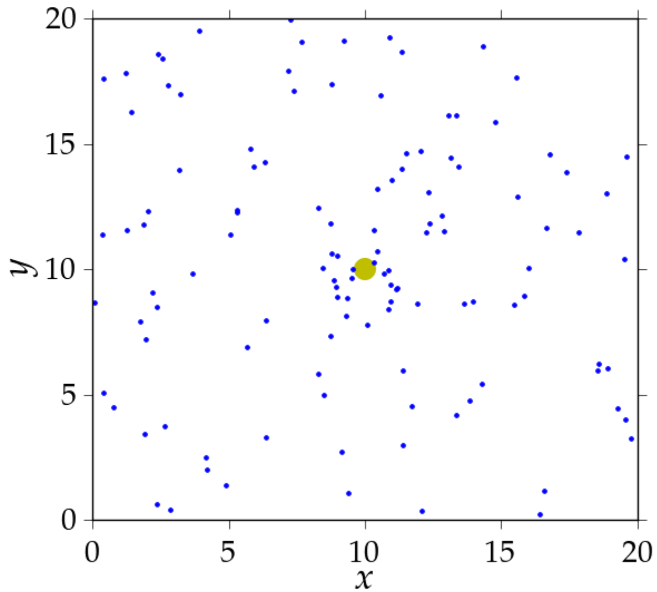


Testing hypotheses

Humor

http://xkcd.com/892

Testing hypotheses



Introductory example

Does a fraction of the sky contain a source of cosmic rays?

Hypothesis H_0 ("background hypothesis")

There is only background and no source

Hypothesis H_1 ("signal hypothesis")

There is background and a source!

Lots of special cases (see literature), most common one for Physicists:

$$f(\vec{x}) = (1-s) f_{\mathrm{B}}(\vec{x}) + s f_{\mathrm{S}}(\vec{x}|\vec{p}_S) \qquad H_0 : s = 0 \qquad H_1 : s > 0$$

→ *Continuous family of hypotheses*

We need a **test statistic** that discriminates between H_0 and H_1

$$-2\ln\lambda = -2\ln\left[\frac{\max L(s=0, \vec{p}_S = 0)}{\max L(s, \vec{p}_S)}\right]$$

likelihood ratio is asymptotically the **most powerful test statistic**

Type I and type II errors

Hypothesis tests are fully characterized by their Type I and Type II errors

Type II error β

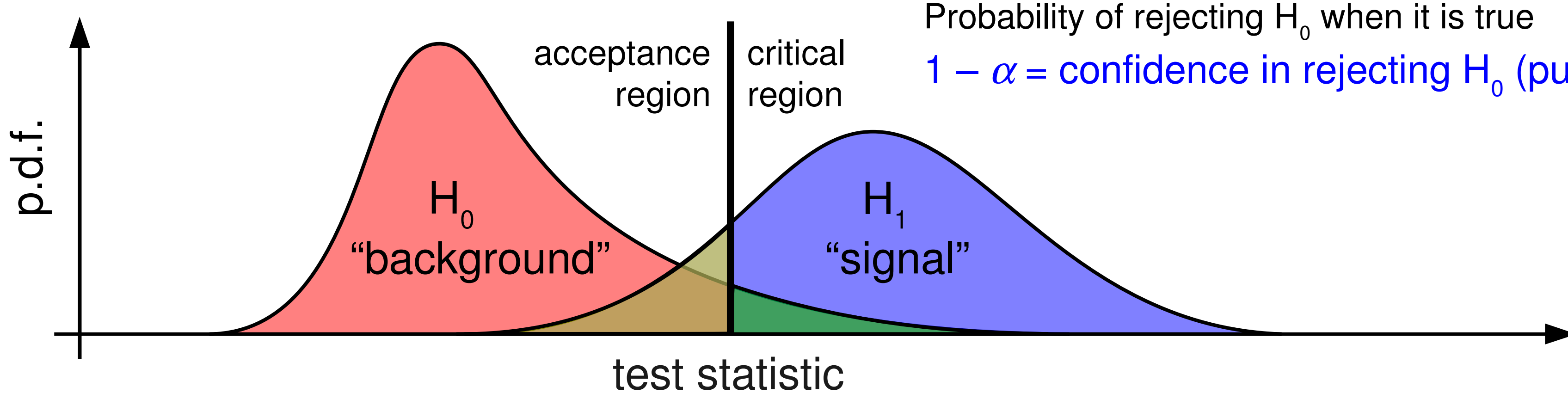
Probability of accepting H_0 when it is false

$1 - \beta$ = power of the test (efficiency)

Type I error α

Probability of rejecting H_0 when it is true

$1 - \alpha$ = confidence in rejecting H_0 (purity)



Desired confidence $1\text{-}\alpha$ defines the critical region, so tests are compared by their power $1\text{-}\beta$

In our case, $1\text{-}\beta$ cannot be calculated, since H_1 is not fully determined

Fortunately, only H_0 is needed to determine the critical region

> Test must be completely defined **before** seeing the data
> Confidence of rejecting H_0 **is not** confidence in choosing H_1 ($\alpha \neq \beta$)

Critical region

How to determine critical region for given confidence 1-α?

$$f(\vec{x}) = (1-s) f_{\mathrm{B}}(\vec{x}) + s f_{\mathrm{S}}(\vec{x}|\vec{p}_S) \qquad H_0:\ s=0 \qquad H_1:\ s>0$$

$$-2\ln\lambda = -2\ln\left[\frac{\max L(s=0, \vec{p}_S = 0)}{\max L(s, \vec{p}_S)}\right]$$

is asymptotically distributed as $\chi^2(r)$

r = number of parameters fixed by H_0 but left free by H_1

Asymptotic properties are nice, but test usually used with small data sets...

Recommended: Monte-Carlo-based determination of critical region

$f_B(\vec{x})$

generate B simulated data sets of size N

$x_1^{(1)}, \ldots, x_N^{(1)} \longrightarrow -2\ln\lambda^{(1)}$

$x_1^{(B)}, \ldots, x_N^{(B)} \longrightarrow -2\ln\lambda^{(B)}$

get B values of $-2\ln\lambda$ from the simulated data sets

1-α quantile of $-2\ln\lambda$ distribution defines critical region

Critical region – example

In our example

Background: 2d uniform distribution
Signal: 2d normal distribution

$$f(x,y) = (1-s) f_B(x,y) + s f_S(x,y)$$

$$f_B(x,y) = \frac{1}{\Delta x \Delta y}$$

$$f_S(x,y) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x^2+y^2)\right]$$

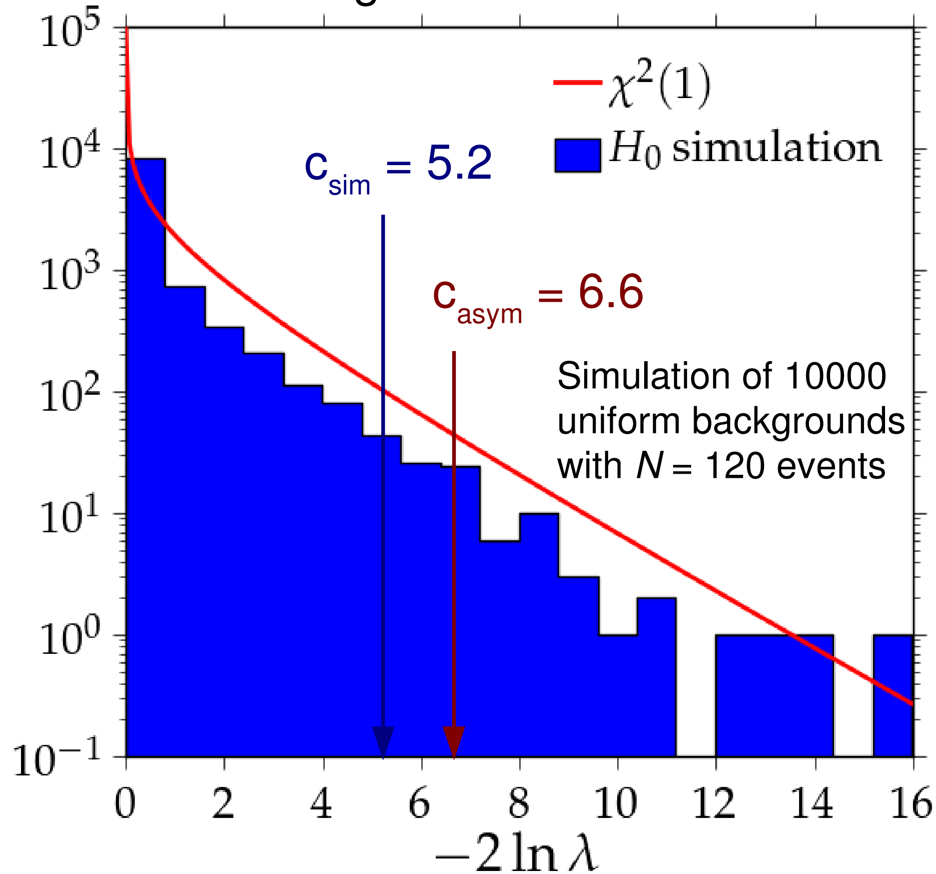
$$H_0 : \lambda_S = 0 \qquad H_1 : \lambda_S > 0$$

s is free in H_1, but fixed in H_0
no other free parameters

→ asymptotic distribution of $-2\ln\lambda$ is $\chi^2(1)$

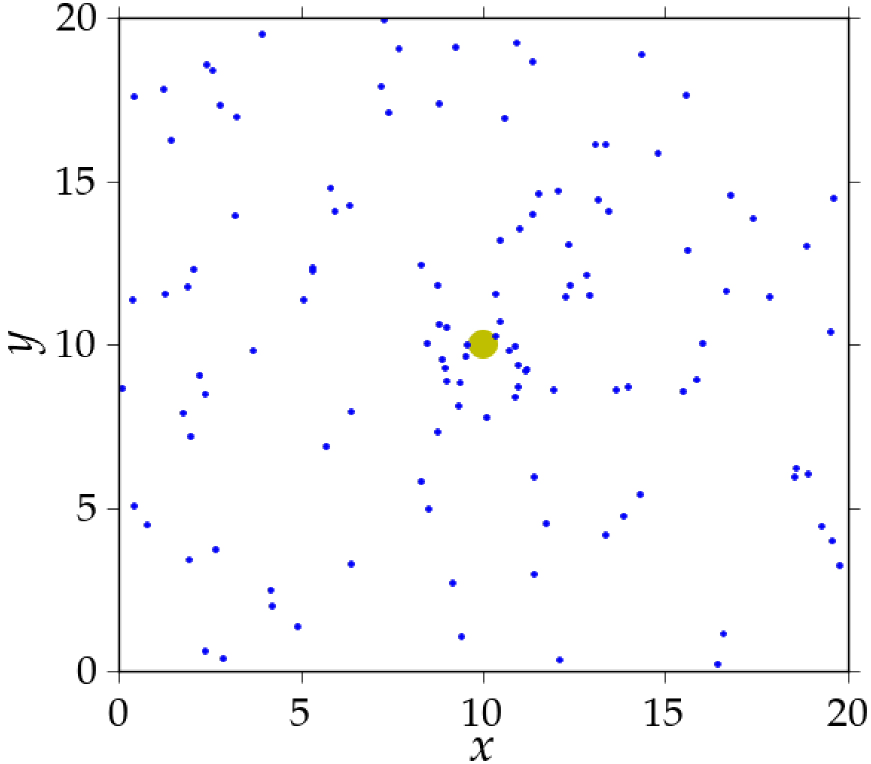
Reject H_0 if in real data set $-2\ln\lambda > c$

Critical region for 99 % confidence

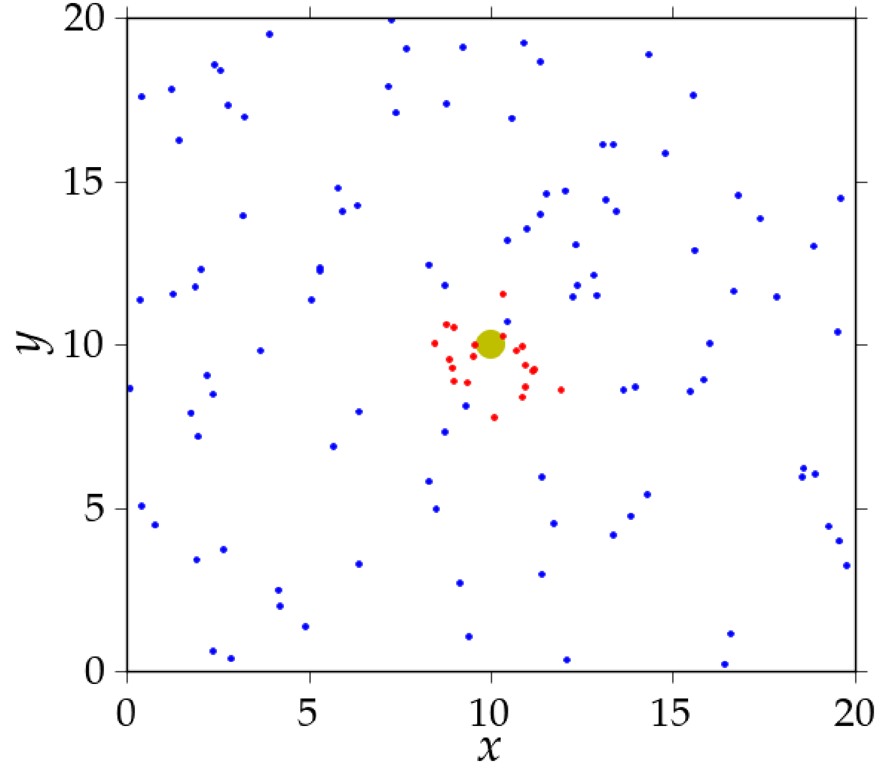


Hypothesis test – example

$N = 120$



uncovered: $N_B = 100$, $N_S = 20$



Test statistic in real data	$-2\ln\lambda = 38.3 > c$

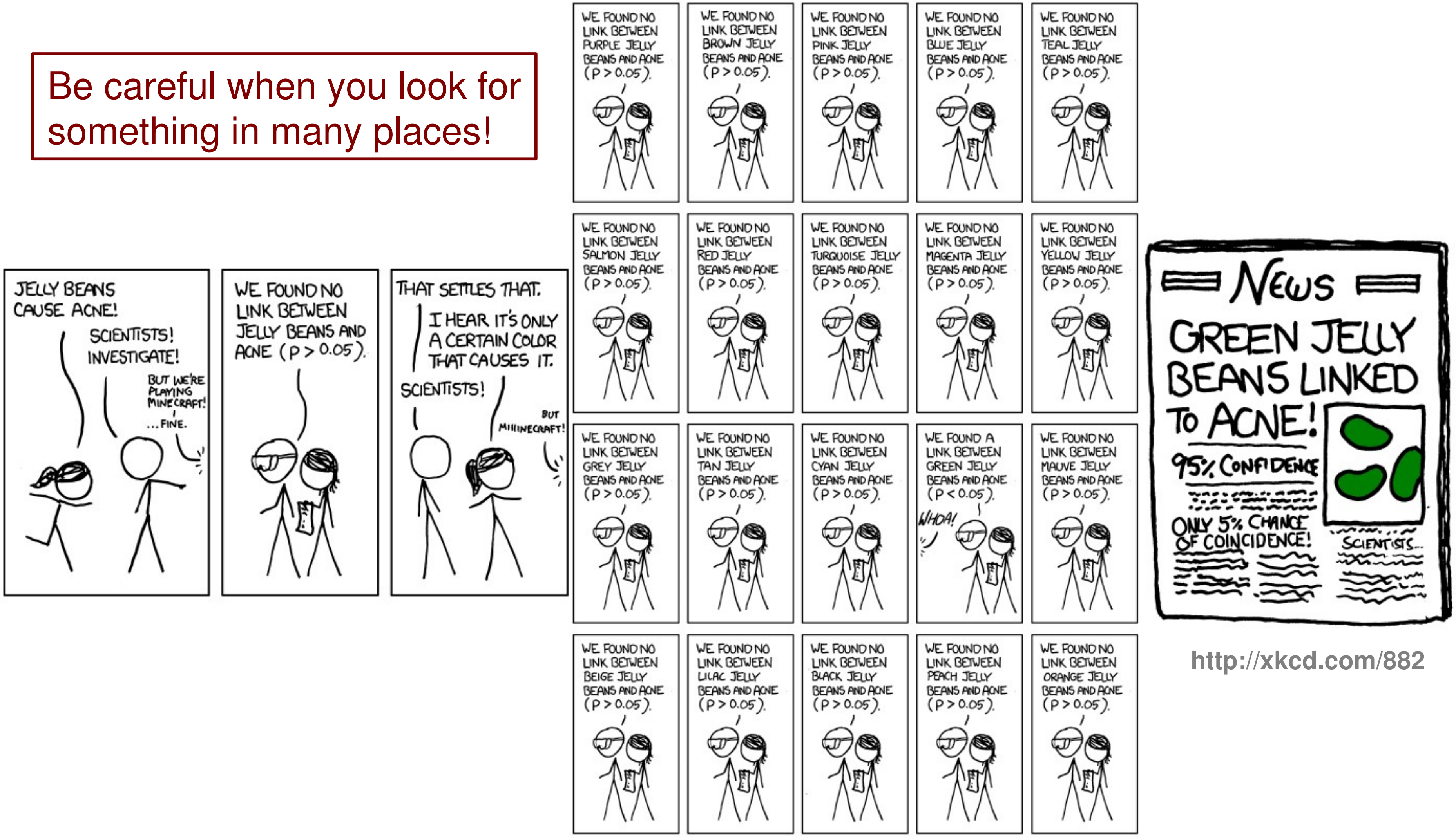
⇨ We reject the background-only hypothesis H_0 with a confidence of at least 99 %

Confidence does **not** increase even if $-2\ln\lambda \gg c$! (property of test, not of data)

Trial factors

a.k.a. Look-Elsewhere-Effect

Be careful when you look for something in many places!



http://xkcd.com/882

Trial factors – example

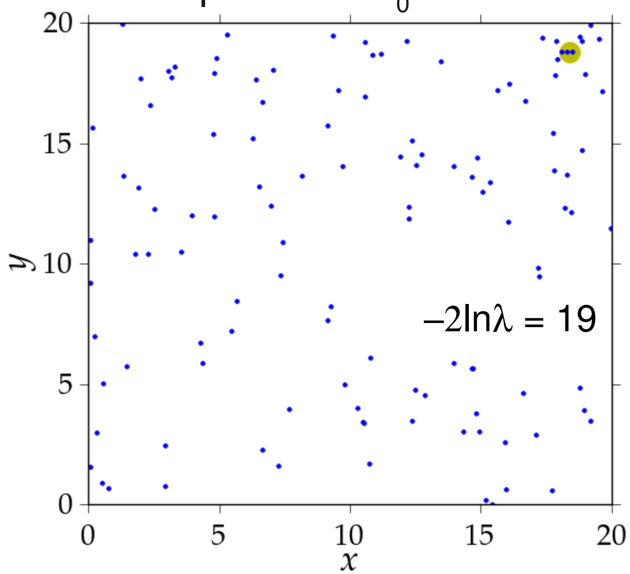
Our example revisited with signal location unknown

$$f_S(x, y|\mu_x, \mu_y) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}\left((x-\underline{\mu_x})^2 + (y-\underline{\mu_y})^2\right)\right]$$

source position free in signal model

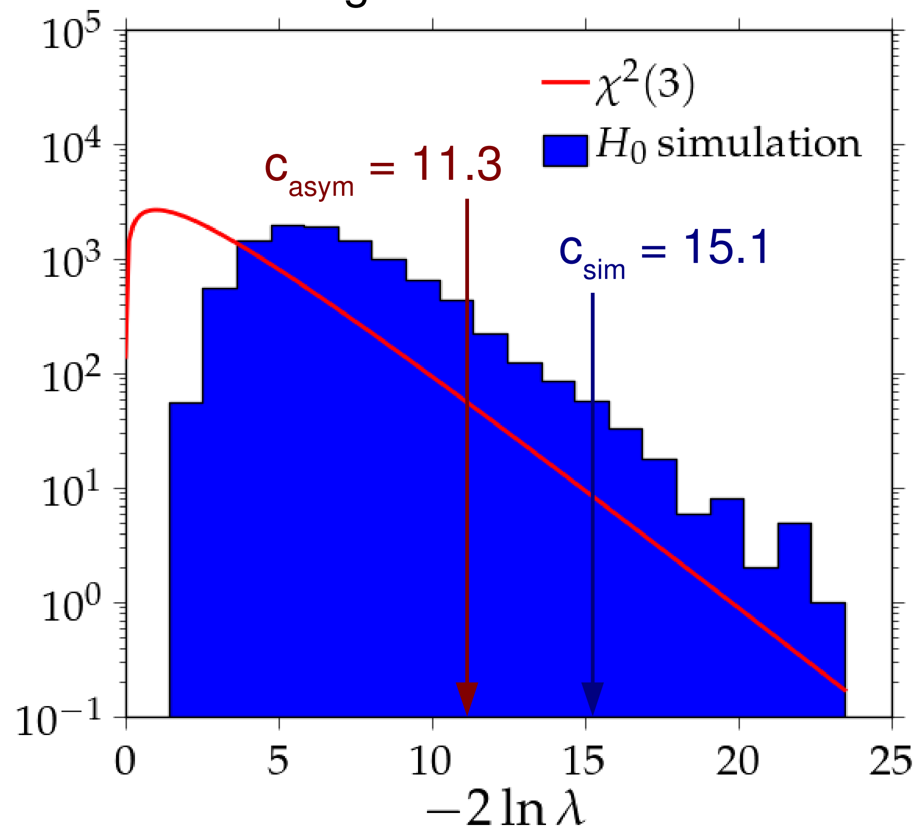
Samples from uniform distribution tend to form clusters
→ Clusters appear like sources
→ large $-2\ln\lambda$ more frequent

Example from H_0 simulation



Natural form of apophenia!

Critical region for 99 % confidence



Goodness-of-fit tests

Goodness-of-fit (GOF) test = Lesser form of Hypothesis test

Test of H_0 with against all possible other hypotheses: H_1 completely unspecified

→ Power 1-β unknown

Components of a GOF test

Test statistic t and c.d.f. $F(t) = \int_t^{-\infty} \mathrm{d}t' \, f(t|H_0)$ to convert t into P-value

Small P-values indicate “bad fit” of model to data

P-value $= P(\mathrm{data}|H_0)$ **is not** $P(H_0|\mathrm{data})$

Large P-values are not evidence in favor of H_0!

Some GOF test statistics are independent of H_0 (distribution-free) $f(t|H_0) = f(t)$

e.g. Pearson's Chi-square test and Smirnov-Cramér-von Mises' test

(for data pairs and binned data) (for unbinned data)

For combined tests calculate P-value from Monte-Carlo simulations of H_0

Pearson's Chi-square test

Idea: sum up squares of normalized residuals of data points around model

$$t = \left(\vec{y} - \vec{f}(\vec{x})\right)^T \tilde{V}^{-1} \left(\vec{y} - \vec{f}(\vec{x})\right) = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2 = \sum_{i=1}^{N} z_i^2$$

if y_i are uncorrelated

z_i have normal distribution with $\mu = 0$, $\sigma = 1$ independent of H_0

If y_i are correlated, one can find transformation to decorrelate them and get same result

$$E[t] = \sum_{i=1}^{N} E[z_i^2] = N \qquad V[t] = \sum_{i=1}^{N} V[z_i^2] = 2N \qquad f(t) = \frac{\frac{1}{2}\left(\frac{t}{2}\right)^{N/2-1} e^{-t/2}}{\Gamma\left(\frac{N}{2}\right)}$$

If *f(x)* has *k* free parameters fitted to the y_i, replace *N* by $N - k$

No formal proof here, but intuition:
Due to fit of *f(x)*, z_i are no longer independent → *k* "degrees of freedom" lost

Smirnov-Cramér-von Mises test

$$t = \int_{-\infty}^{\infty} \mathrm{d}x \, [\hat{F}(x) - F(x)]^2 f(x)$$

is independent of *f(x)* (= H_0)

$$\hat{F}(x) = \int_{-\infty}^{x} \mathrm{d}x' \, \hat{f}(x) = \frac{1}{N} \int_{-\infty}^{x} \mathrm{d}x' \sum_i \delta(x - x_i) = \frac{1}{N} \sum_i H(x - x_i)$$

H(x) Heaviside step function

Proof: insert substitution $y = F(x)$ $\longrightarrow$ $t = \int_{-\infty}^{\infty} \mathrm{d}y \, [\hat{F}(y) - y]^2$

$$E[t] = \frac{1}{6N} \qquad V[t] = \frac{4N - 3}{180N^3}$$

no $f(t)$ in closed form → tables

Based on asymptotic distribution, reached for $N \geq 3$

Confidence level 1-α	Critical value of $N\,t$
0.90	0.347
0.95	0.461
0.99	0.743
0.999	1.168

Backup

General formula for any distribution

$$V[\hat{\sigma}^2] = \frac{1}{N}\left(\mu_4 - \frac{N-3}{N-1}\sigma^4\right)$$

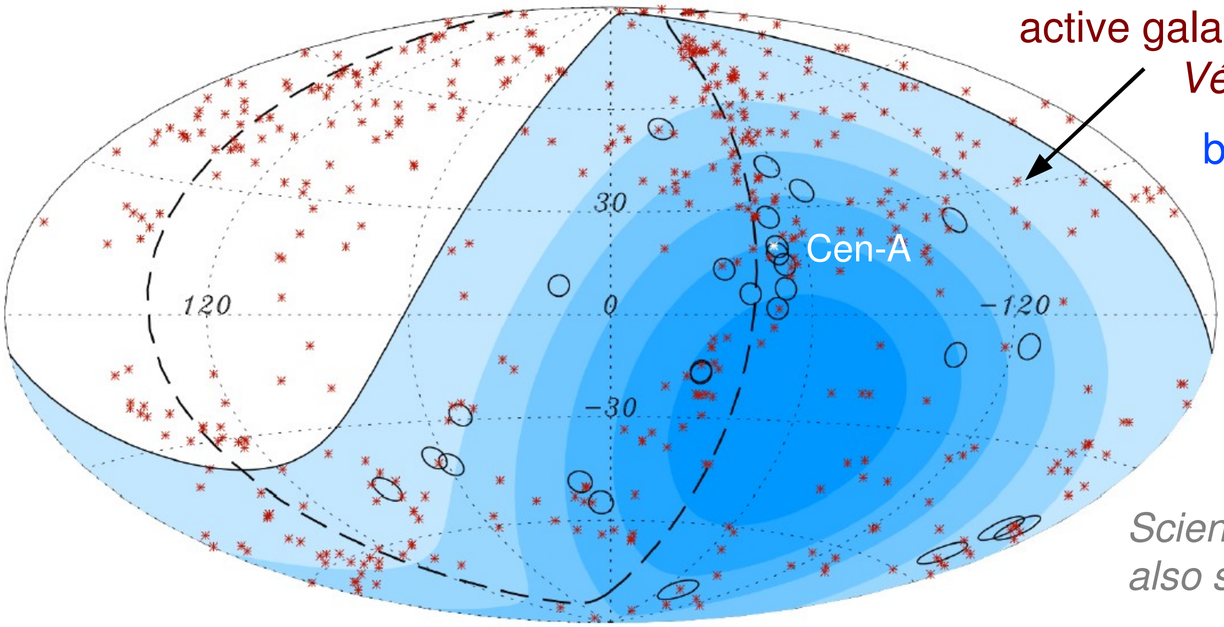
with $\mu_4 = \frac{1}{N}\sum_i (x_i - \mu)^4$

Hypothesis probability after seeing data

Testing hypotheses

Recent example: structures in cosmic ray sky found by the Pierre Auger Observatory

sky map of CR arrival directions in galactic coordinates



active galaxies closer than 75 Mpc from *Véron-Cetty & Véron* catalogue

blue color indicates exposure

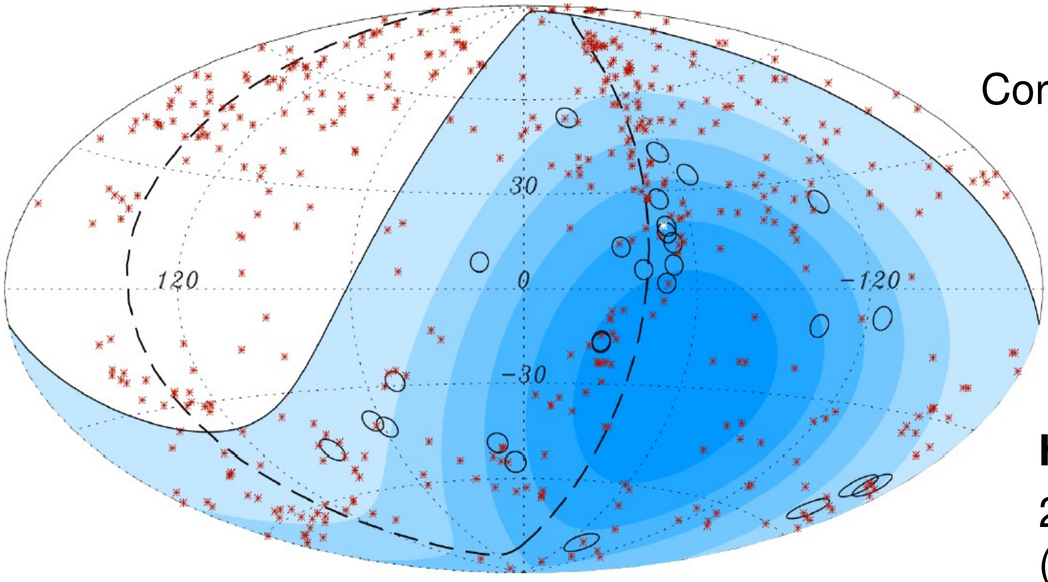
◯ Auger data

27 events with $E > 10^{19.7}$ eV
(data up to September 2007)

Science 2007
also see Astropart. Phys. 2010

UHECR sky seems anisotropic, let's reject the hypothesis H_0 [CRs are isotropic]!
With what confidence can we do it? **or** What is the probability to be mistaken?

Testing hypotheses



Correlation = angle(cosmic ray, AGN) < 3.2°

Test statistic
Number of correlating events *k* out of *N*
(*k* follows binomial distribution)

H_0 prediction (isotropy)
21 % of cosmic rays correlate → p = 0.21
(AGN coverage of the sky)

N = 13 (first 14 events were used to define test)
k = 8

H_1 prediction (anisotropy)
p >= 0.21