JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Statistics 101

NuSTEC Summer School 2024

What is probability?

- Example: Coin flip
- Probability of throwing heads P(H)
  - Single throw does not tell us much
  - Ratio for lots of throws

$$P(H) = \lim_{N \to \infty} \frac{N_H}{N}$$

  - Frequentist definition
- Probability of coin being fair

$$\text{fair} \equiv (0.5 - \epsilon \leq P(H) \leq 0.5 + \epsilon)$$

  - Frequentist?

$$P(\text{fair}) = \lim_{N \to \infty} \frac{N_{\text{fair}}}{N} = \begin{cases} 1 & \text{if coin is fair} \\ 0 & \text{if coin is not fair} \end{cases}$$

  - No matter how often we throw/measure, coin is fair or not
  - We don’t know which!

Bayesian probability

- Bayesian definition: Probability is interpreted as
  - reasonable expectation
  - representing a state of knowledge
  - quantification of a personal (!) belief
- Allows assigning arbitrary probability of coin being fair

$$P(\text{fair}) \in [0, 1]$$

- If P is subjective, what does it mean exactly?
  - P = 1 → Something is true
  - P = 0 → Something is false
  - P = 0.3 → ???
- Fair bet: Will not lose (or win) money in the long term
  - Considering all information, at what odds would you bet?
    - “Semi frequentist” over all instances when you apply Bayesian stats?

Mathematical probability

- Frequentist vs. Bayesian deals with interpretation
  - Influences what we can consider a random variable
- Mathematical rules are the same
- Probability (Kolmogorov) axioms
  - P are real numbers > 0
  - Total probability = 1
  - P of mutually exclusive events is additive

$$P(E) \in \mathbb{R}, P(E) \geq 0 \qquad \forall E \in F$$

$$P(\Omega) = 1.$$

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

- Conditional probability → Bayes' theorem

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

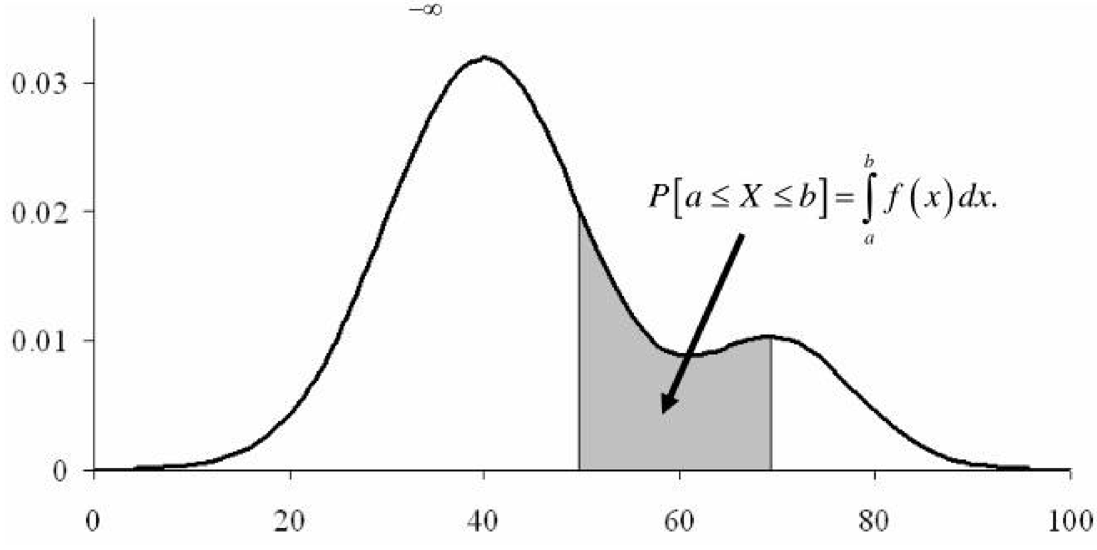
  - Applies to both Bayesian and Frequentist random variables
    - "Fair or not" is not a random variable in Frequentist interpretation!

Probability density

- What if outcome of experiment lies on continuum?
- Define Probability Density Function (PDF)
  f (or sometimes p) so that

Probability density function, $f(x)$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$



- Note that P[X = 40] = 0
  - Infinitely many points on x, P of single point = 0

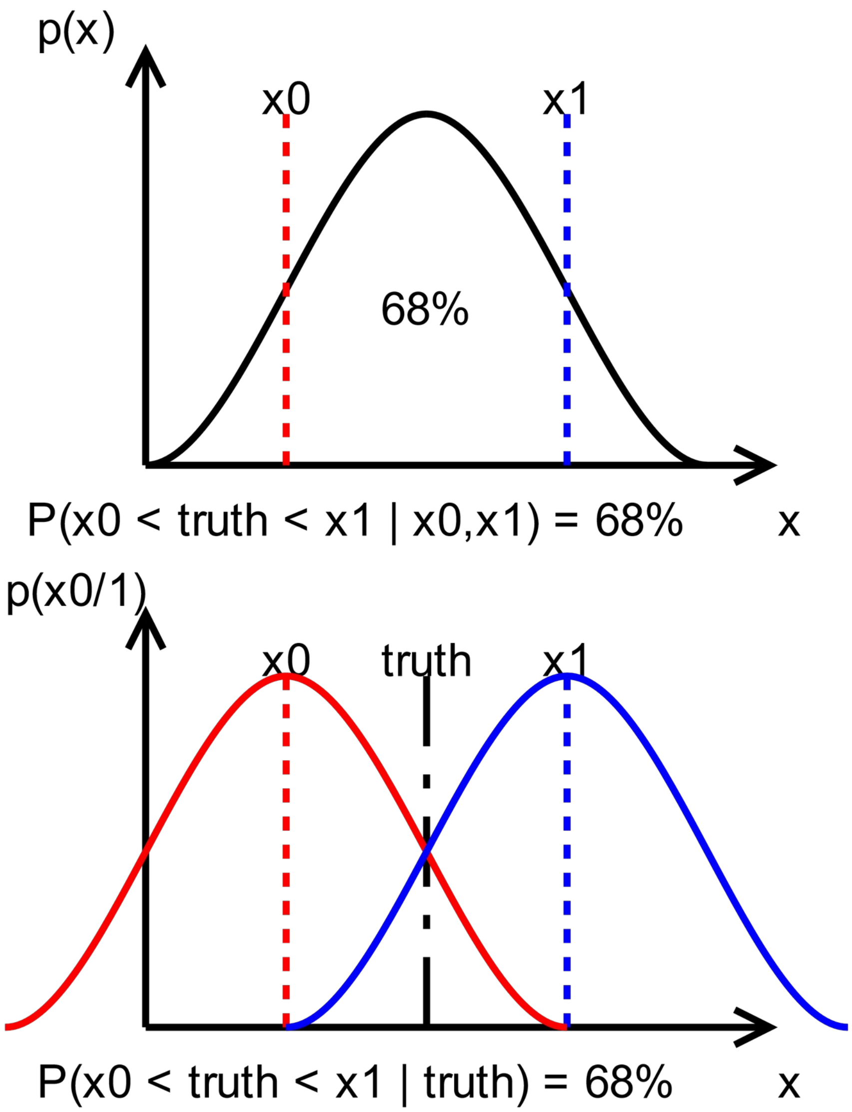
Summary I

- Two different interpretations of probability
  - Frequentist → Probability is frequency limit
  - Bayesian → Probability is certainty or reasonable expectation
- Differ in what can be considered a random variable
- Same rules for calculations with random variables
  - Bayes' theorem works for both!
- Homework: Frequentist probability or not?

P(rolling 1 on D20), P(rolling 1 on next D20 roll),
P(a randomly picked summer day being sunny),
P(yesterday being sunny), P(tomorrow being sunny),
P(people actually doing their homework)

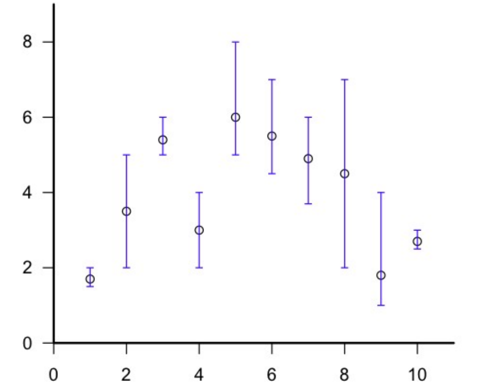
What are error bars?

- Bars in plots can be all sorts of things
  - Bin widths
  - Poisson variance estimate sqrt(N)
  - ???
  - Read the paper
  - Specify in your paper!
- Bayesian → Credible Interval
  - Parameter is random variable
  - Probability of true value within = Credibility Level (often 68%)
- Frequentist → Confidence Interval
  - Parameter is fixed (!) but unknown
  - Interval edges are random variables
  - Probability of CI covering truth = Confidence Level (often 68%)
    - No matter what the true value is
- Not the same!
  - E.g. empty or unphysical confidence intervals are OK! (albeit annoying)
  - Need to be treated differently



Central values

- CI construction is data reduction
  - Condense data into 1 or 2 numbers
  - Many variations: which CL, one- or two-sided,
  - Choice limits future uses of data
  - Which values are “the best”?



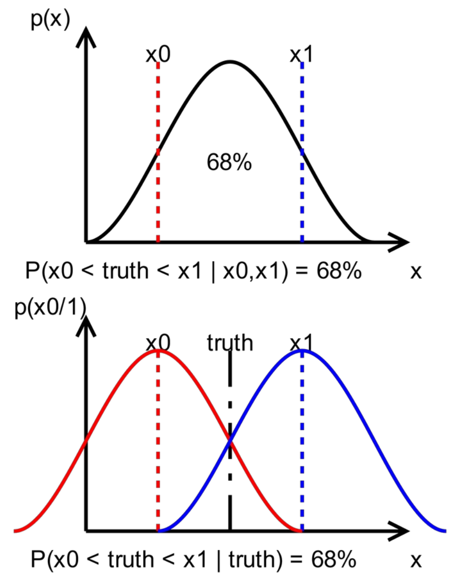
- Add more information by also providing a point estimate
  - E.g. Maximum Likelihood Estimator (MLE), Maximum A-posteriori Probability (MAP), …
  - Does not need to be inside CI! (usually is though)
- Interval w/o central value is useful, e.g. limits
- Central value w/o error bar is useless
  - Is deviation of 10% compatible? 0.1%?
  - Better than nothing? Probably add an error bar in your head

Gaussian approximation

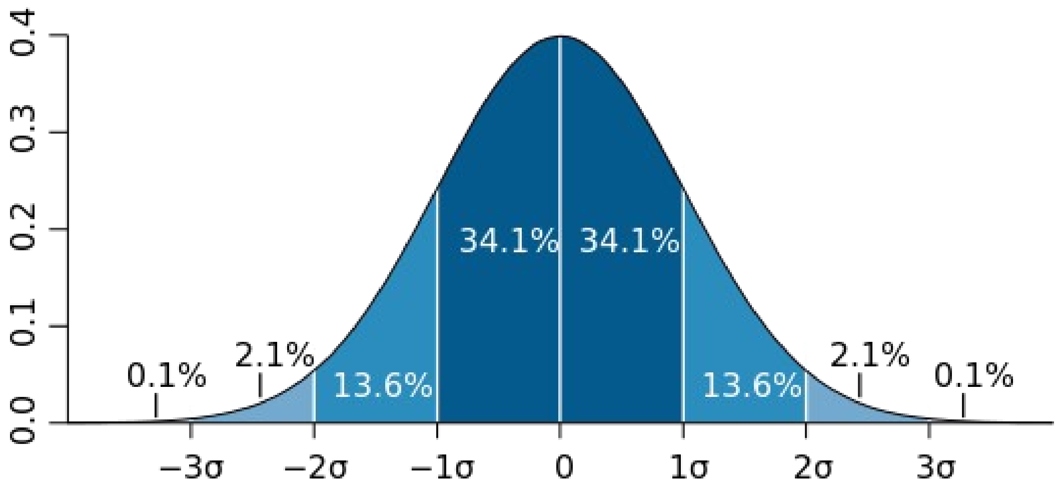
- Even with point estimate, CI only provides single CL
  - What if we want to check for stronger deviations?
  - Would need to specify full probability/likelihood function
- Gaussian approximation
  - Assume a normal distribution of
    - Probability of true value (Bayesian)
    - Point estimates around true value (Frequentist)
  - Completely described by exp. value & variance

$$\mu = \langle x \rangle = \int x\, p(x)\, \mathrm{d}x \qquad \sigma^2(x) = \int (x - \langle x \rangle)^2 p(x)\, \mathrm{d}x$$

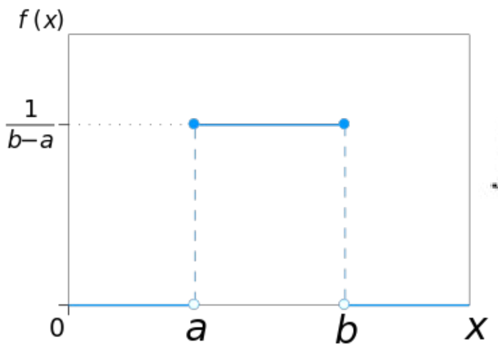
  - Can easily construct CI of any CL:
    - 68%: $\mu \pm \sigma$
    - 95%: $\mu \pm 2\sigma$
    - 99.7%: $\mu \pm 3\sigma$



Why does this work?



- Where do CL values come from?
- Integrating probability over normal distribution
- Two-sided 1 std = 68% CL only true for normal!
    - E.g. uniform distribution

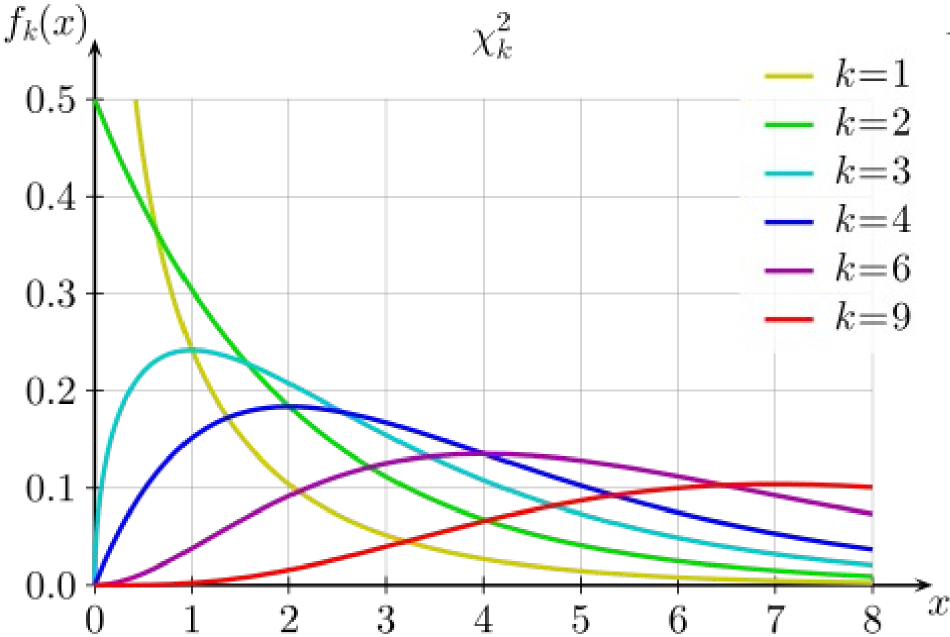


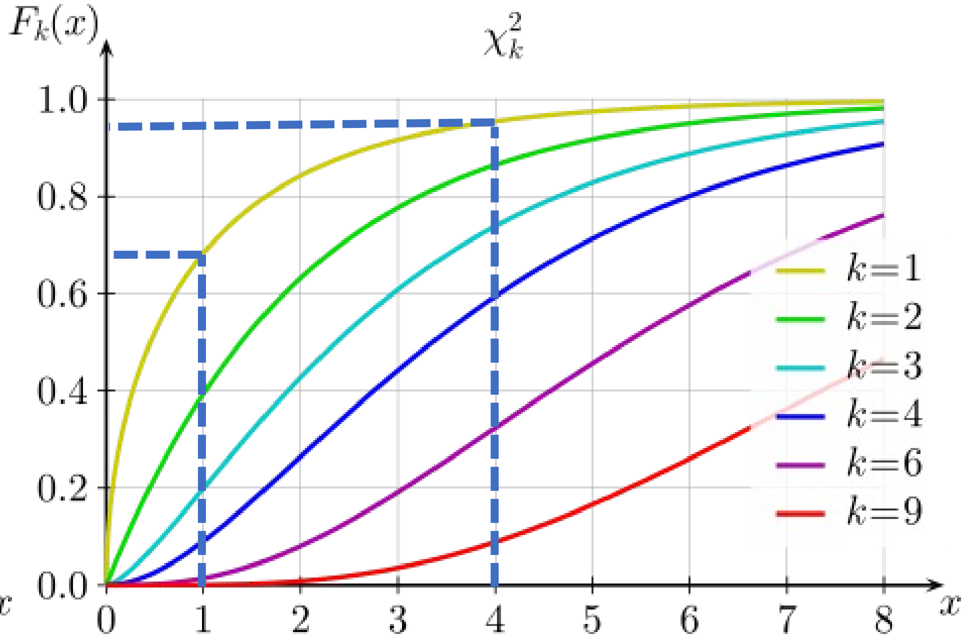
$$\sigma = (b-a)/\sqrt{12}$$

$$\int_{\langle x\rangle-\sigma}^{\langle x\rangle+\sigma} p(x)\,\mathrm{d}x = 2/\sqrt{12} \approx 0.58$$

$$\int_{\langle x\rangle-2\sigma}^{\langle x\rangle+2\sigma} p(x)\,\mathrm{d}x = 1$$

Relation to chi-squared dist.



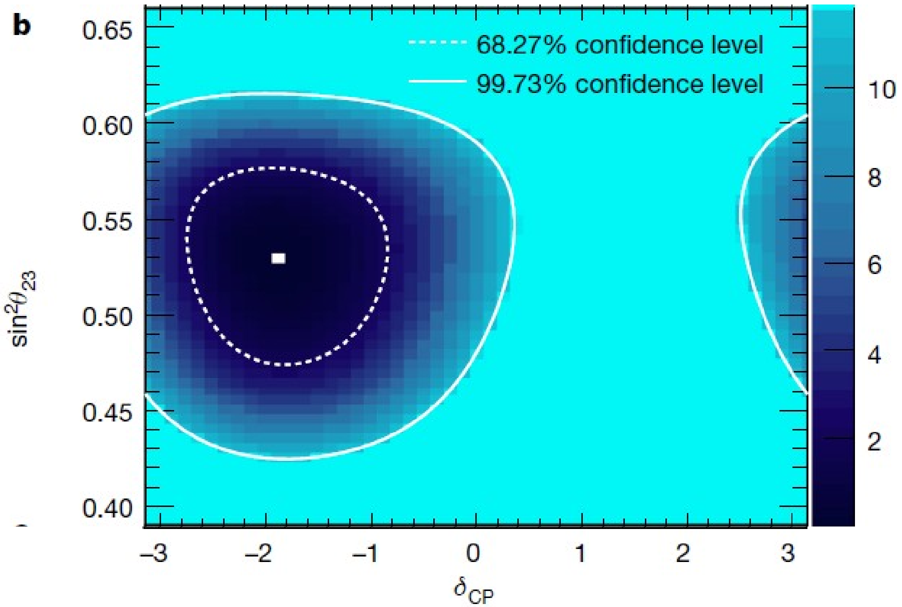


- Squared deviation is Chi-squared distributed with k=1
  - Assuming normal distribution!
  - Integral 0 – 1 = 68%
  - Integral 0 – 4 = 95%
  - Integral 0 – 9 = 99.7%
  - …

$$(x - \langle x \rangle)^2 / \sigma^2 \sim \chi^2_1$$

Multiple parameters

- What if comparing more than one parameter?
  - Multiple XSEC bins, oscillation angles/masses, …
- Confidence/credible
  1D interval → N-dim region

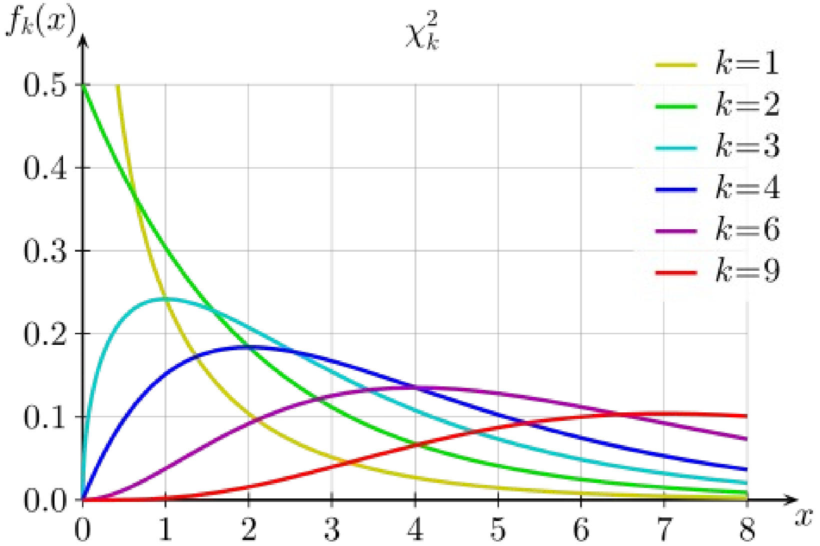


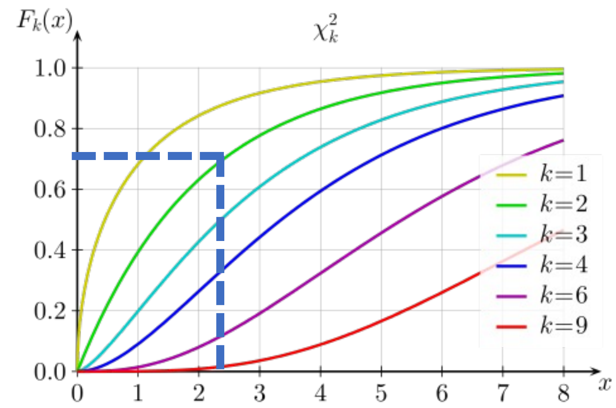
- In principle infinite ways to define confidence/credible region in N dimensions
  - Expand shape around central value until contains desired CL?
    - What shape? Sphere? Box? Heart shape?
- All "correct" as long as CL is right (though not optimal)

Multivariate Gauss

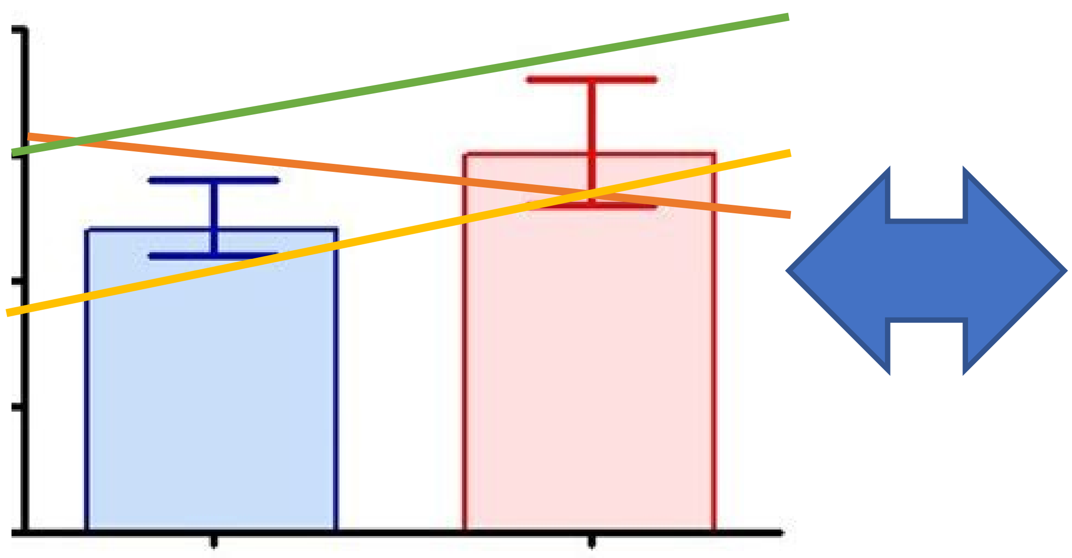
- Uncorrelated N-dim Gaussian approximation
  - Uncertainties on all parameters are independent
  - N mean values, N variances
- Sum of normed squared distances ~ $\chi^2(k=N)$
  - Definition of χ^2!
  - Mode = k – 2, Mean = k
    - That is why "reduced χ^2" should be roughly 1!
  - $\sigma(\chi^2)$ = sqrt(2k)
    - The higher N, the more unlikely a larger sum/N becomes!

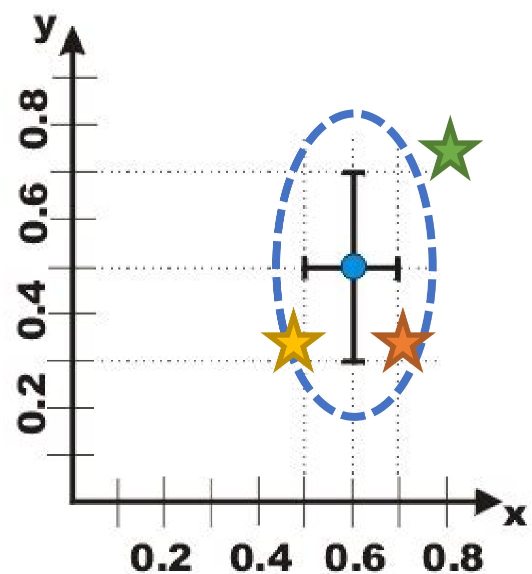
$$\sum_i^N (x_i - \langle x_i \rangle)^2 / \sigma_i^2 \sim \chi_N^2$$





- Look up value of $\chi^2(k)$ distribution where integral = CL
- Compare to sum of squared deviations
- Error bar → error ellipse in N-dim parameter space
  - Sphere in normalised parameter space





Correlations

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

- What about correlated variables?
- Assume multivariate normal distribution
  - Fully described by N mean values, N variances, and N*(N-1)/2 covariances → N mean values & N*N covariance matrix

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = (\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k])^{\mathbf{T}}, \quad \Sigma_{i,j} := \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j]$$

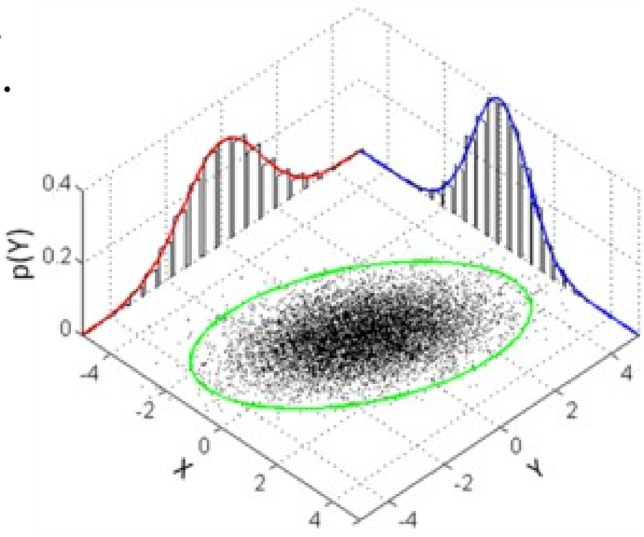
- Mahalanobis distance (“the chi-squared”)
  - Generalisation of sum of squared deviations for correlated variables

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

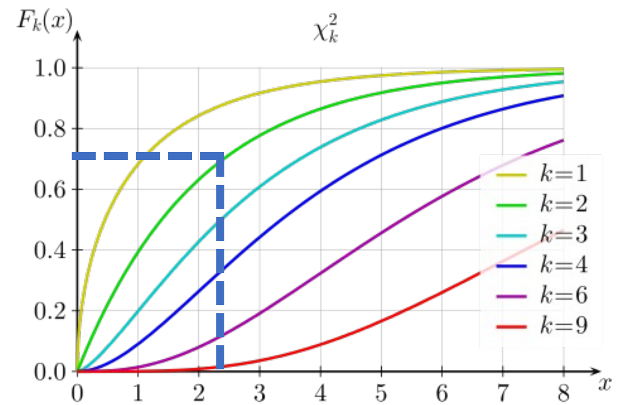
- Chi2 distributed k=N

$$D_M^2 \sim \chi_N^2$$

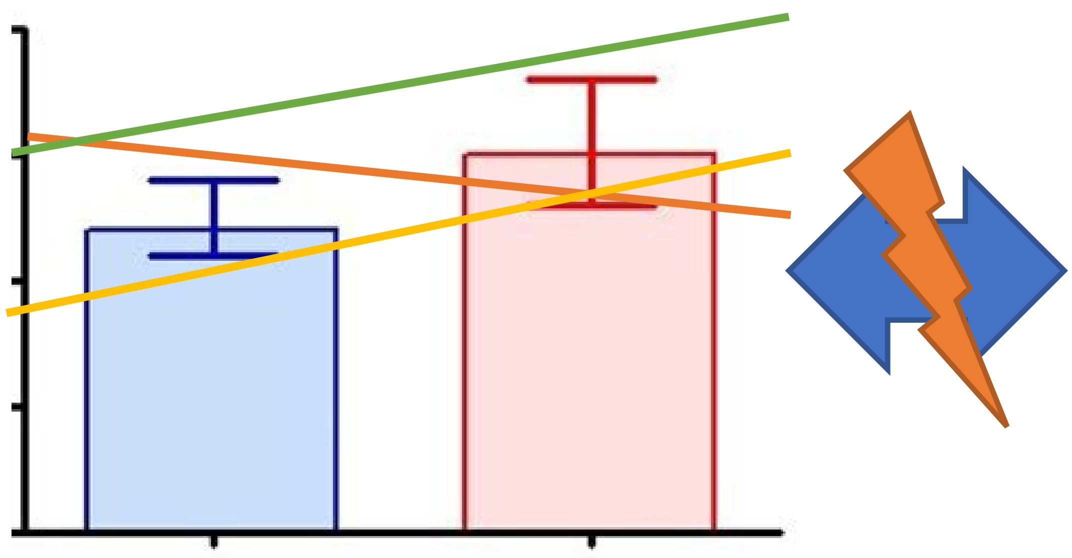
- Construct CI with chi2 quantile
  - N-dim ellipse in parameter space

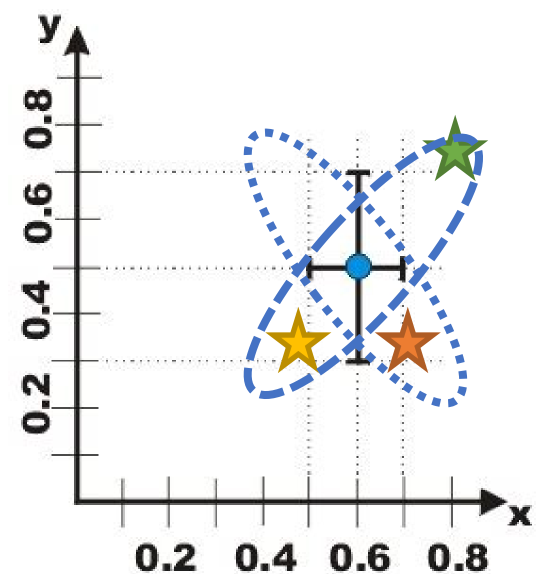


- Look up value of chi2(k) distribution where integral = CL



- Compare to squared Mahalanobis distance
- Error ellipse → rotated error ellipse
  - Can no longer judge fit by looking at plot! → Give us the χ^2!





Summary II

- Error bars are often confidence/credible intervals
  - Former cover truth in CL instances of repeated experiments
  - Latter cover CL of posterior probability
- Can convey more information by Gaussian approx.
  - Assumptions about shape of uncertainty!
  - Construct CI with arbitrary CL w/o reanalysing the data
    - Mahalanobis distance & quantiles of chi2 distribution
  - Need full covariance if uncertainties are correlated!
- Homework

Repeat 20 times: “I will provide a full covariance matrix and always quote ‘the chi-squared’ in plots when dealing with correlated data points.”

Constructing a CI

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

- Bayesian straight forward(ish)
  - Assume a prior (a can of worms for another workshop)
  - Update prior with data and Bayes' theorem
  - "Cut out" area of posterior with desired credible level
    - Many ways to do this, e.g. central, shortest possible, one-sided, …
    - Why not provide full (parametrised) posterior?
- Frequentist more complicated
  - Devise rule for including or excluding a point in parameter space based on the data
  - Rule must accept true parameter in CL of cases, no matter what the real parameter actually is
    - Necessary for correct CI, but not the only desirable property
    - Consider ignoring data, randomly accepting or rejecting full parameter space

Bayesian inference

- Example: Coin flip
  - Parameter: P(H)
  - Prior: p(P(H))
  - Experiment:
    - Throw N times
    - Get N(H) times heads
    - Binomial distribution
- Apply Bayes' theorem
  - Prior probability of N(H) can be seen as normalisation const.
  - Posterior p'(P(H)) can get complicated
    - Using posterior as prior for next measurement?
    - No closed form of p(x)?

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)} = \frac{p(x \mid \theta)\, p(\theta)}{\int_\theta p(x, \theta)\, d\theta} = \frac{p(x \mid \theta)\, p(\theta)}{\int_\theta p(x \mid \theta)\, p(\theta)\, d\theta}$$

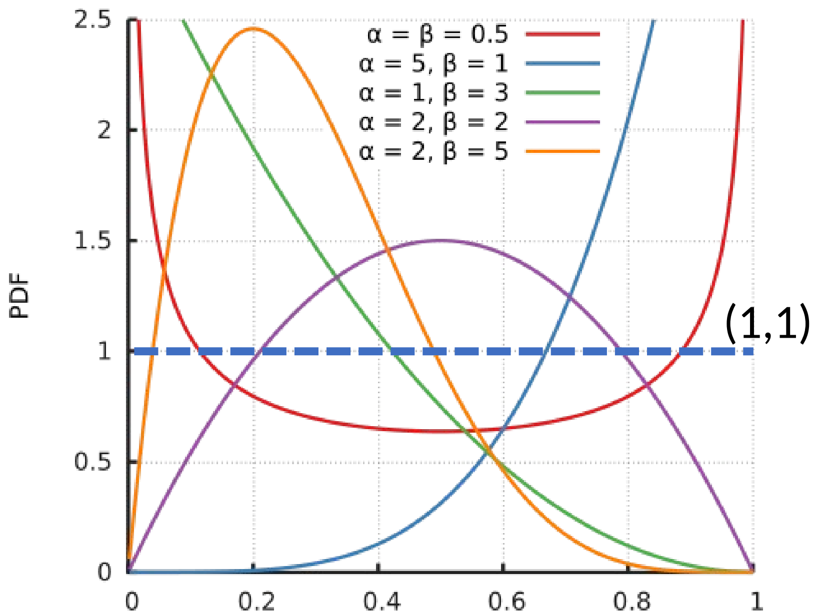
Conjugate priors

- Special class of prior probability functions
  - Posterior probability is same class of function as prior
  - p(P(H)) = f(a,b,c,…)
  - p’(P(H)) = f(a’,b’,c’,…)
  - Rules to update hyperparameters a,b,c,… → a’,b’,c’,… with data
- Update rules much simpler than calculating Bayes’!
- Depends on data process!
  - E.g. for Binomial process (coin toss) → Beta distribution

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

  - Update rules

$$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$$



Markov Chain Monte Carlo

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

- Analytical solutions work well for “small” problems
- “Big” problems (lots of parameters, complicated likelihoods) easier to solve with numerical methods
- Markov chain Monte Carlo (MCMC)
  - Randomly throw parameter values
  - Accept according to rules based on likelihood and prior
  - Result is sample of parameters from posterior distribution
- Use sample to fill histograms, fit functions, find central value, etc…
- Lots of complicating details…

Frequentist inference

- Significance test
  - Hypothesis = proposition about stochastic process of data
    - E.g. binomial distribution with P(H) = 0.49
    - Not a random variable!
  - Test statistic = function of data, measuring “extremeness”
    - Example: standard deviation (z-score), Mahalanobis distance
  - p-value = probability of getting “more extreme” result assuming the hypothesis is correct
- When p-value < significance = α = 1 – CL
  - Exclude hypothesis from CI
  - By construction will happen randomly at rate alpha
- Critical value: Test statistic where p-value = α
- Critical region: Area of test statistic p-value <= α

Gaussian example

- Example: Speed control
  - Measure a vehicles speed once

$$\hat{v} = 55\,\mathrm{km/h}$$

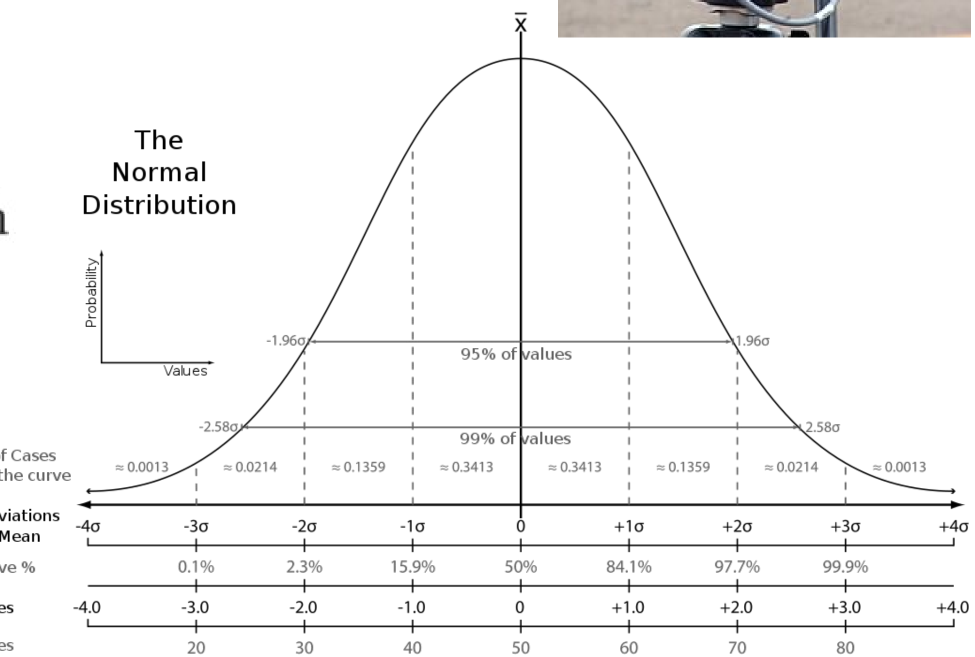
  - Know value is normal distributed around truth

$$\sigma = 2\,\mathrm{km/h}$$

- Choose z-score (“sigmas”) as test statistic
  - 99% CL in Z = +/- 2.58
- Naïve approach:

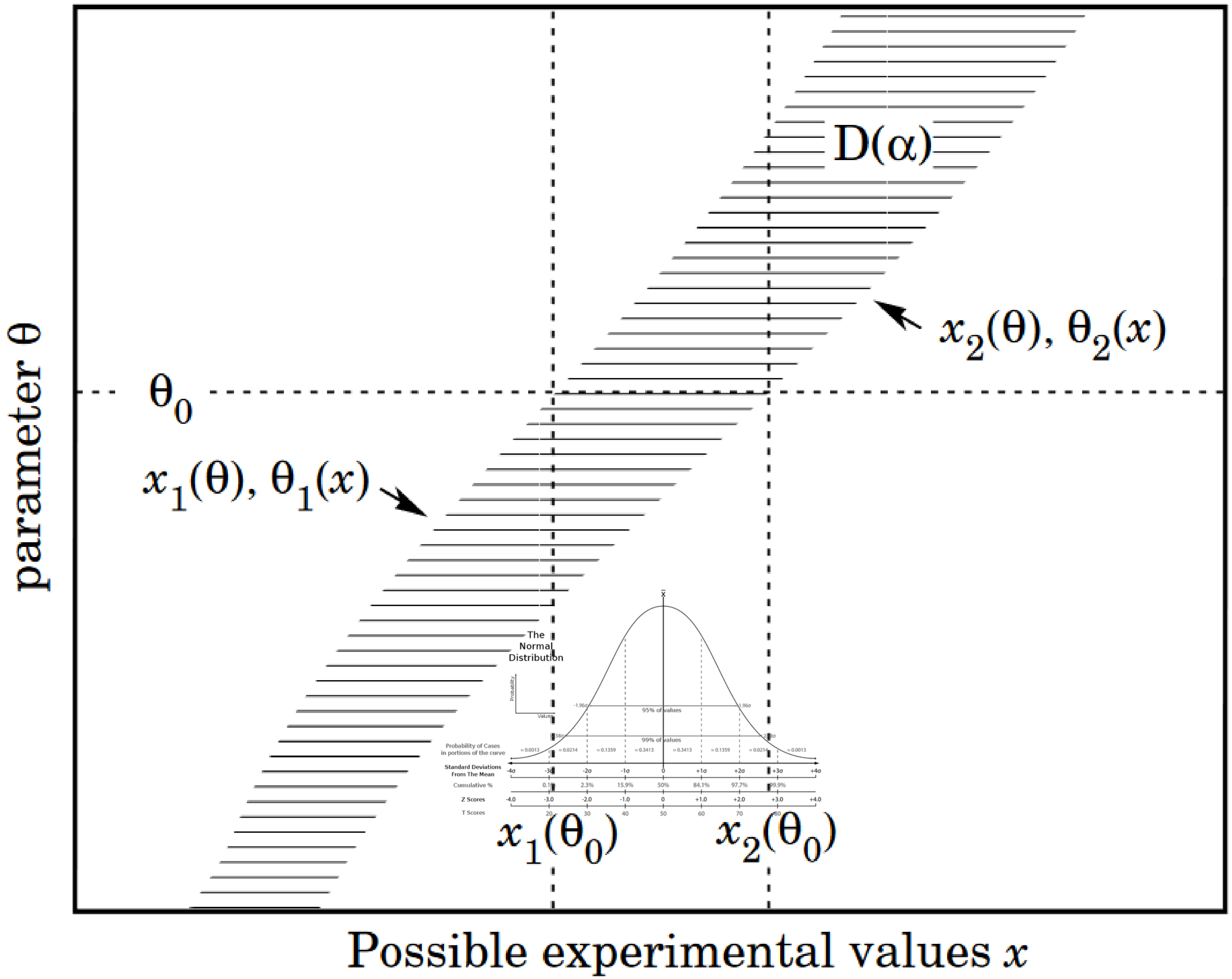
$$v = \hat{v} \pm 2.58\sigma = (55 \pm 5.16)\,\mathrm{km/h}$$

  - How does this work?
  - v is not a random variable!
  - Bayesian propaganda?



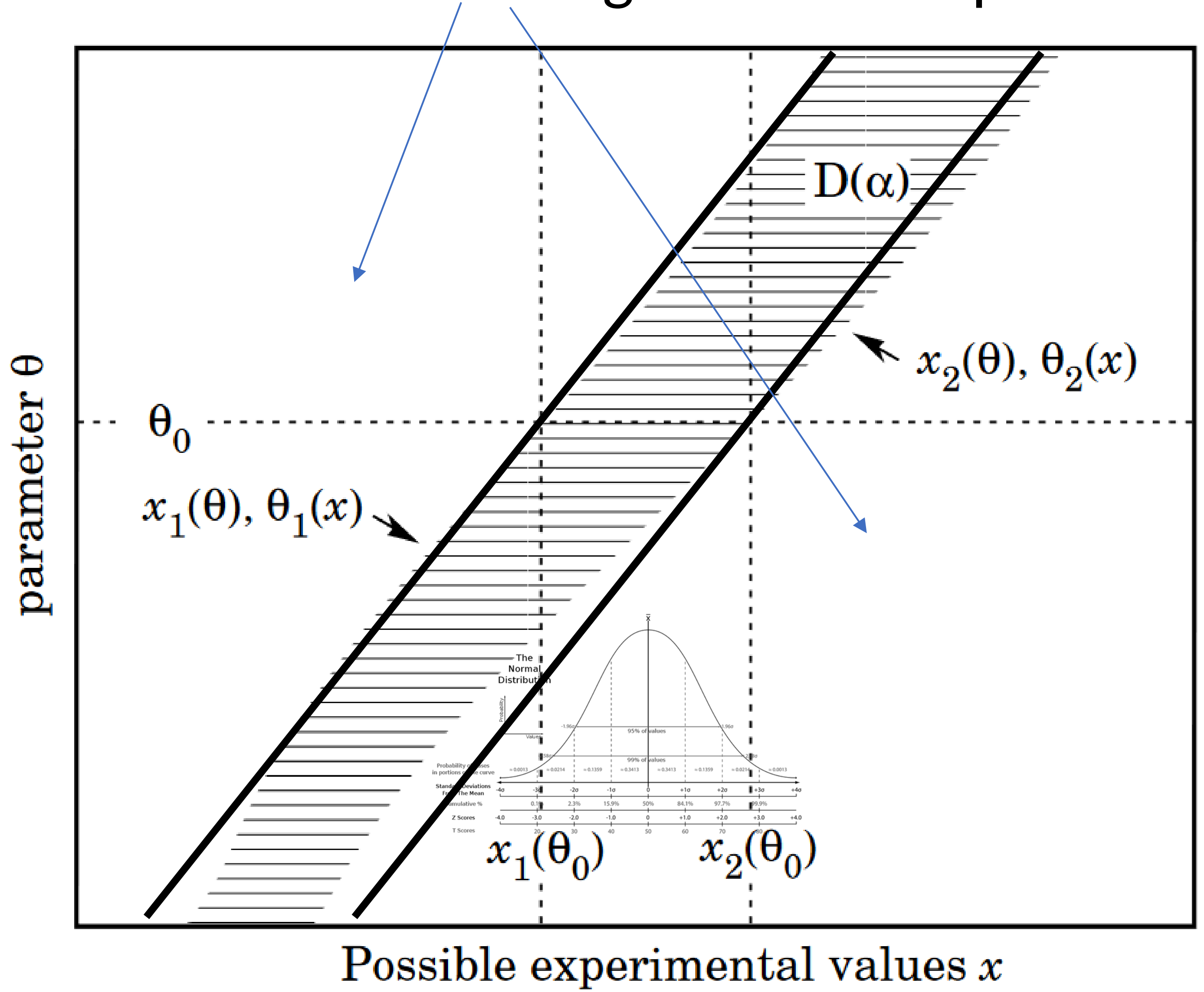
Gaussian example

- Need to define a critical region for each possible value

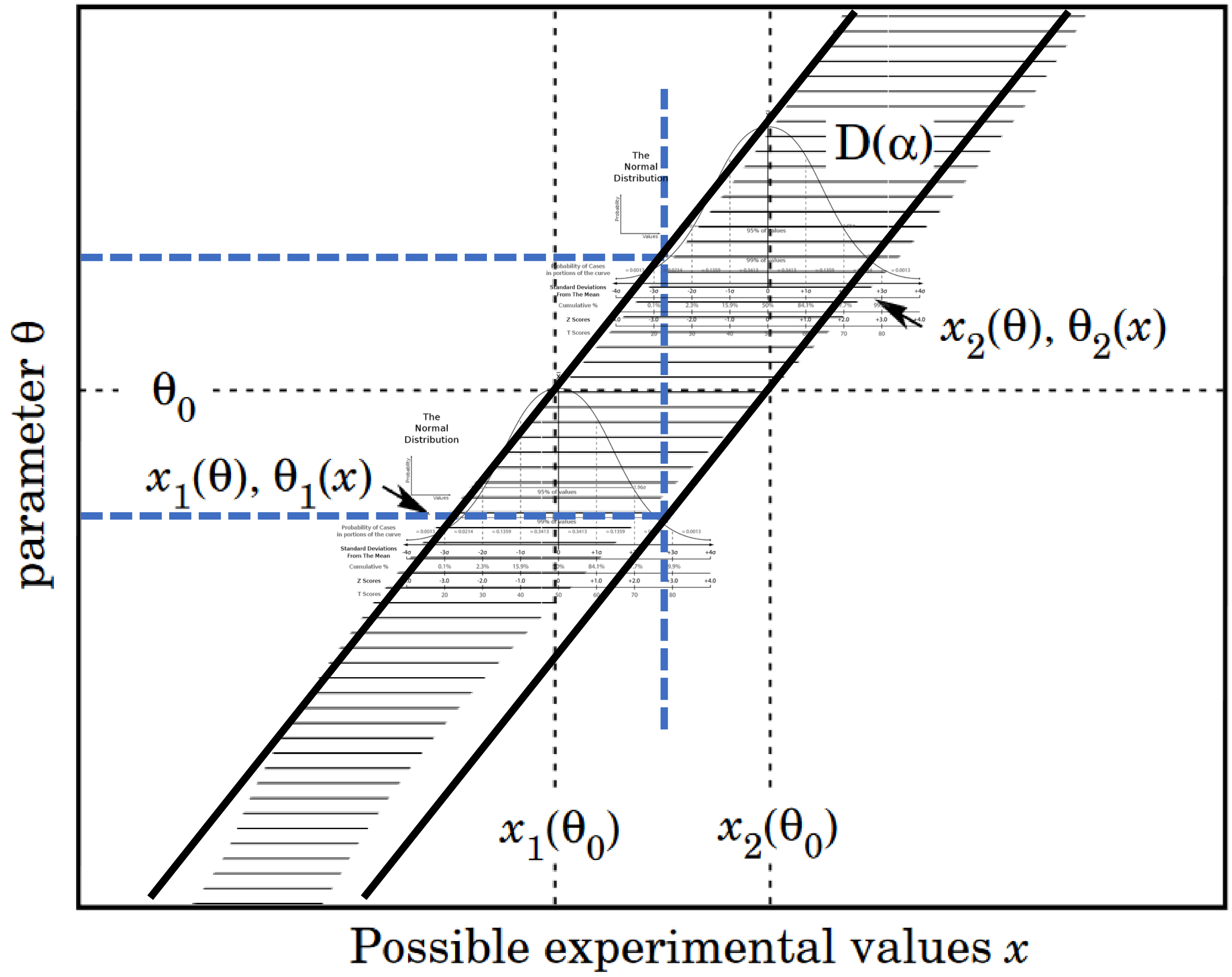


Gaussian example

- Need to define a critical region for each possible value

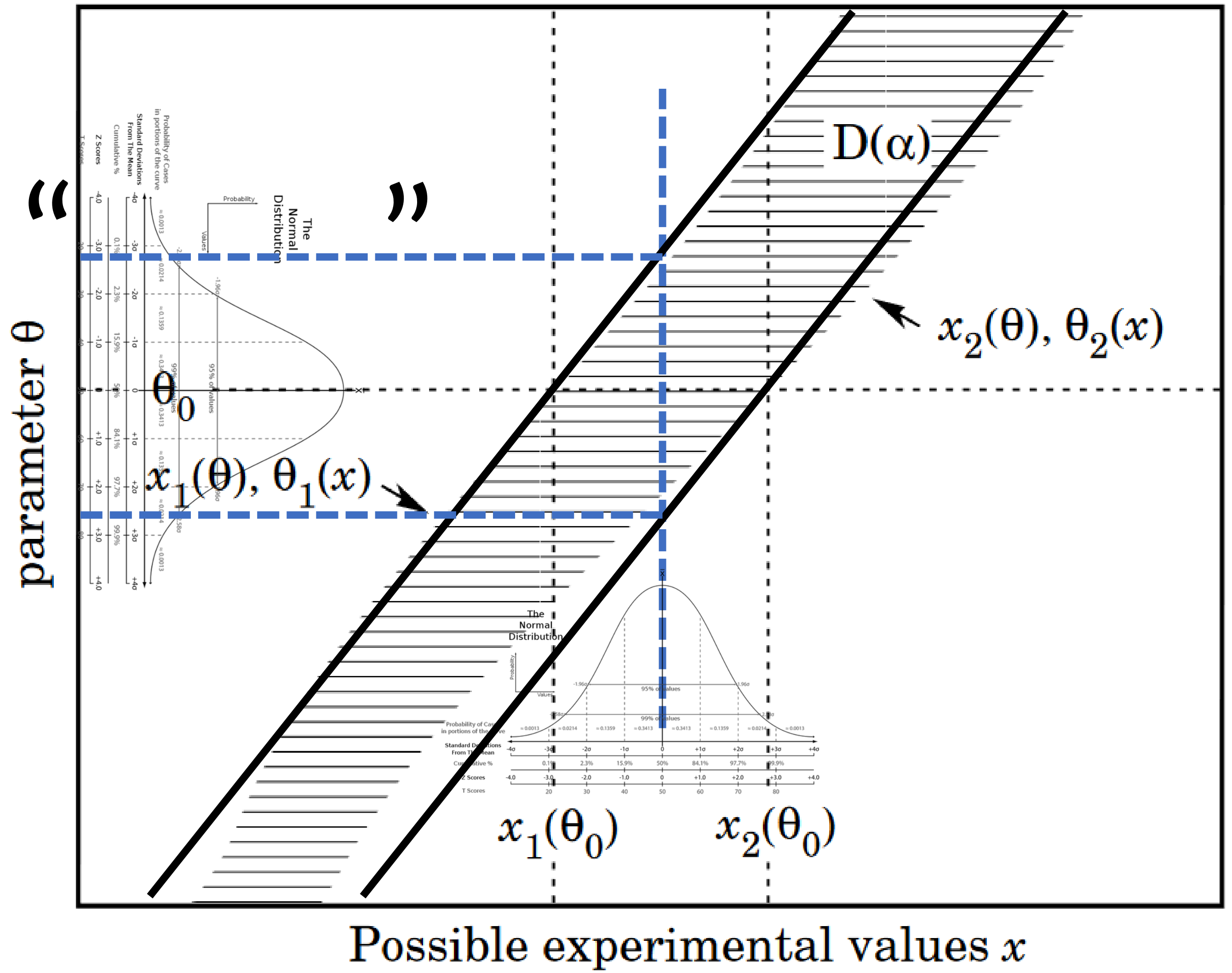


- Check for which parameters the result is in critical region



Gaussian example

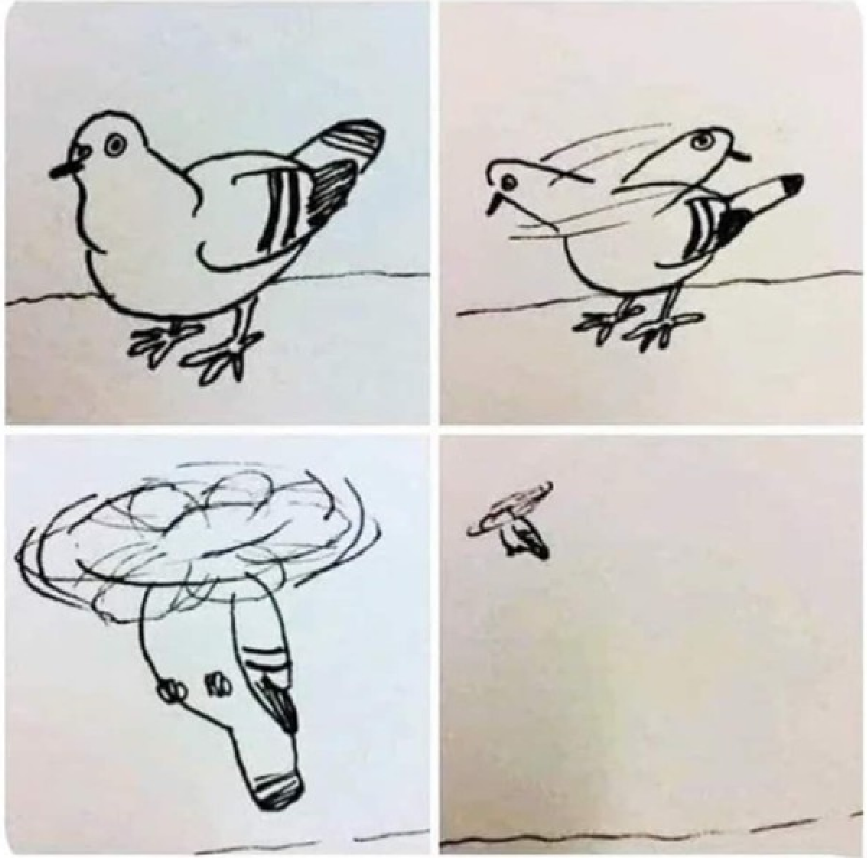
- Accepted region happens to be same as naïve solution



Gaussian example

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

- Lots of symmetry in this Gaussian problem
  - PDF symmetric around mean
  - Parameter corresponds directly to data
  - Data variance is constant
- When treating estimated data PDF as PDF for parameter we get the correct answers by accident!
- Can do the wrong things for the wrong reasons and still get a correct answer



Poisson example

- Example: Speeding rate
  - How many speeders are there per year
  - Count number of speeding tickets N = 25
- Decide: sufficiently Gaussian, use z-score again
- Naïve solution:

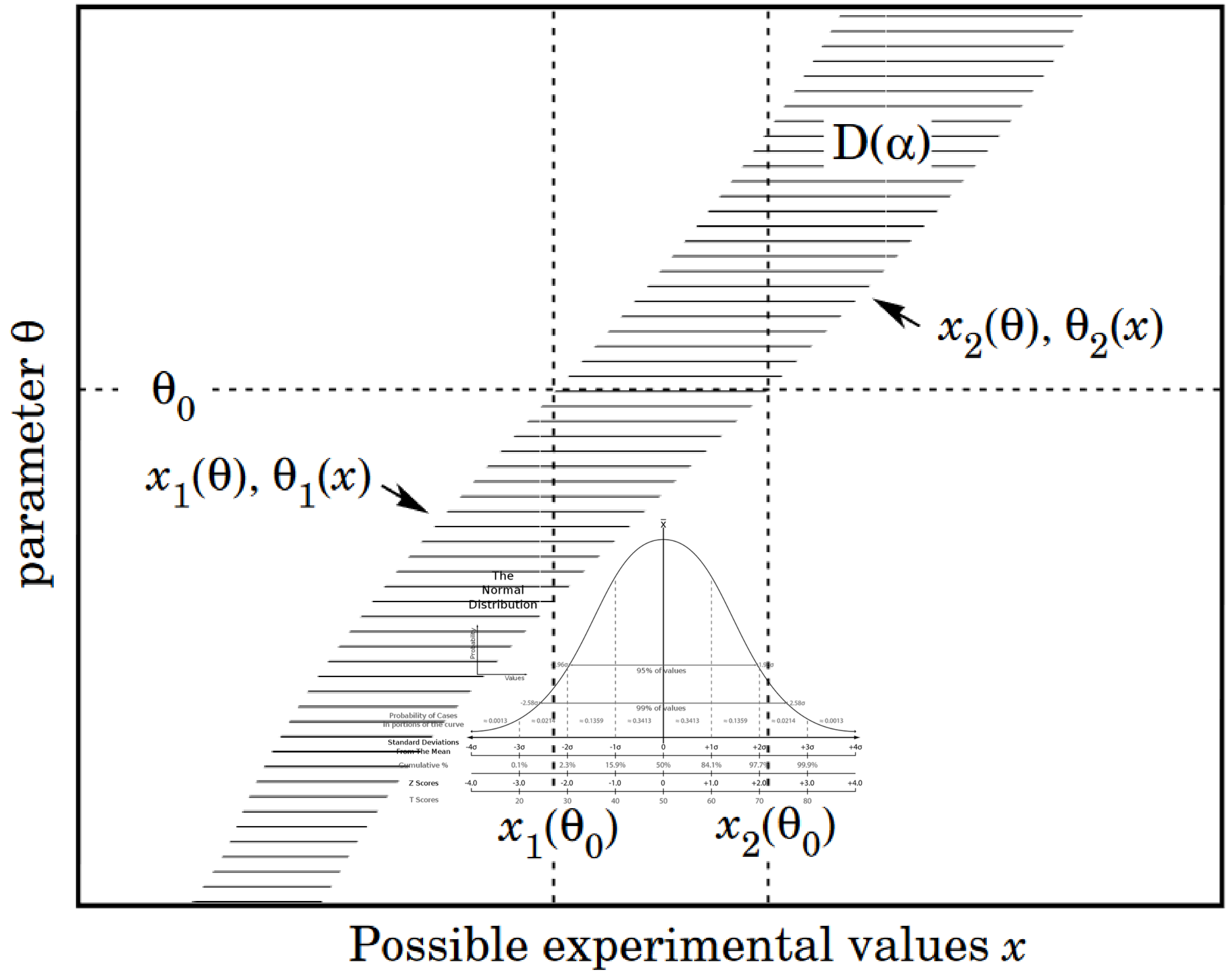
$$\hat{\lambda} = 25$$

$$\hat{\sigma} = \sqrt{\hat{\lambda}} = 5$$

$$\lambda = 25 \pm 5$$

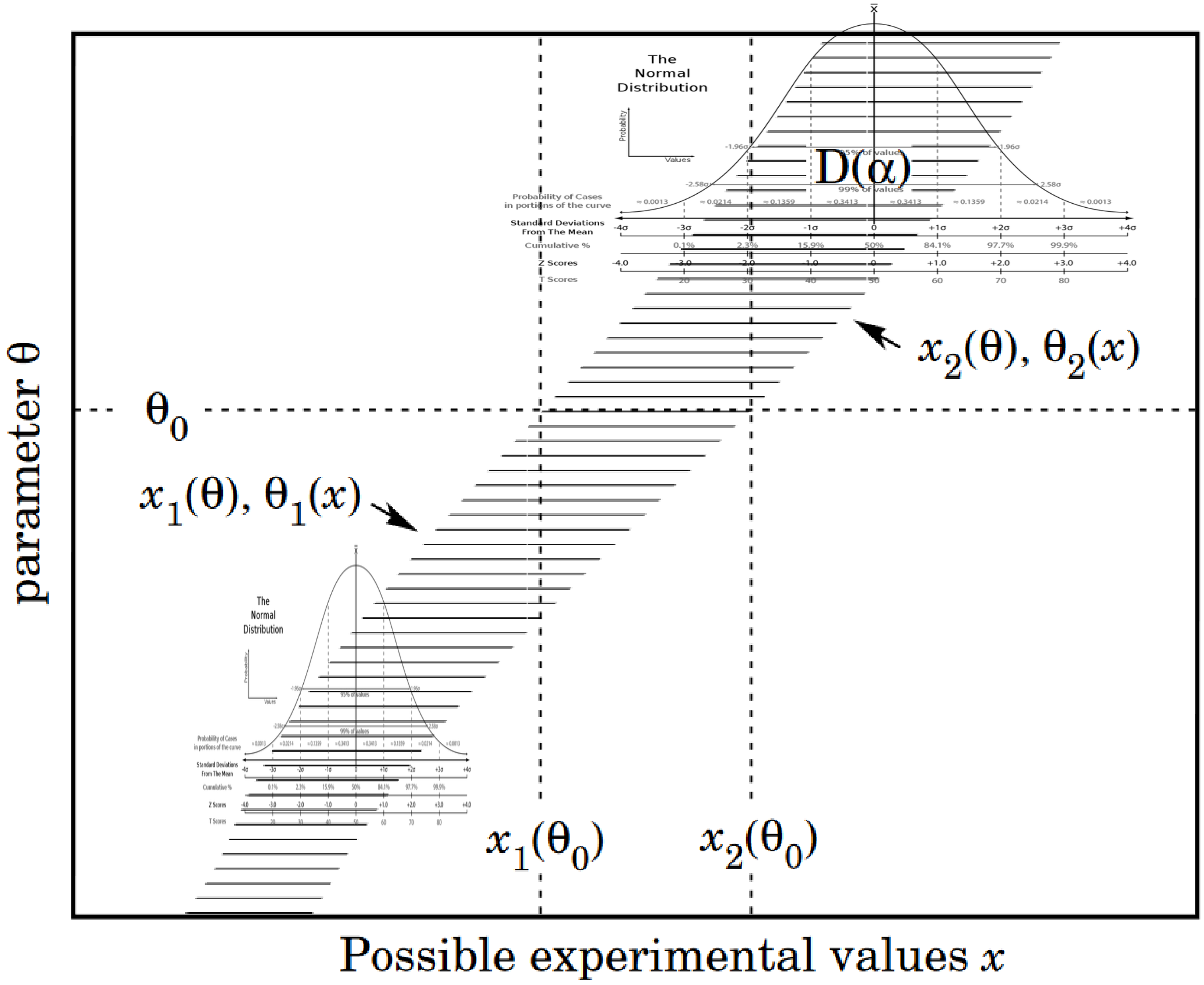
- How does it hold up this time?

- Need to define a critical region for each possible value



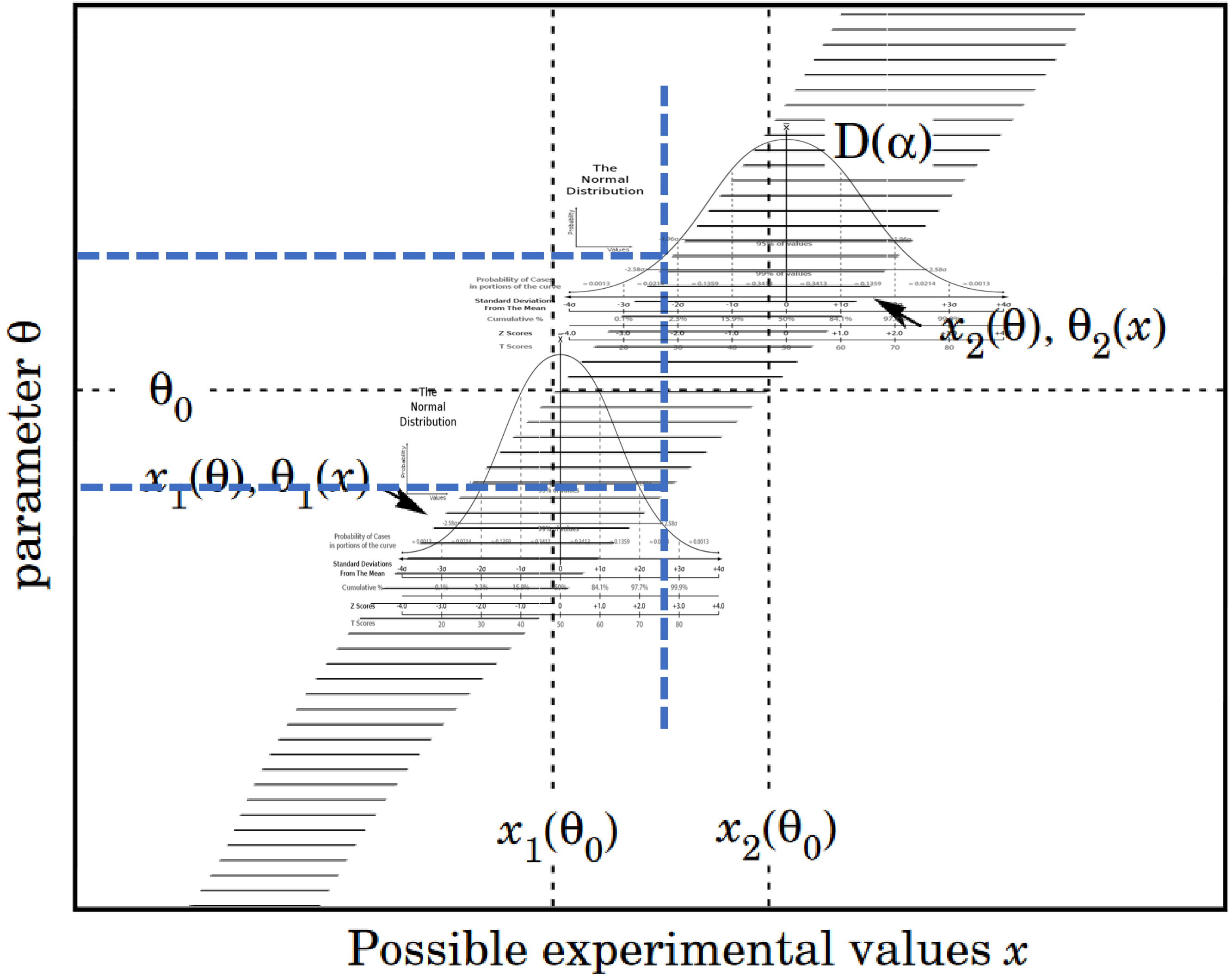
Poisson example

- Expected data variance depends on parameter value!



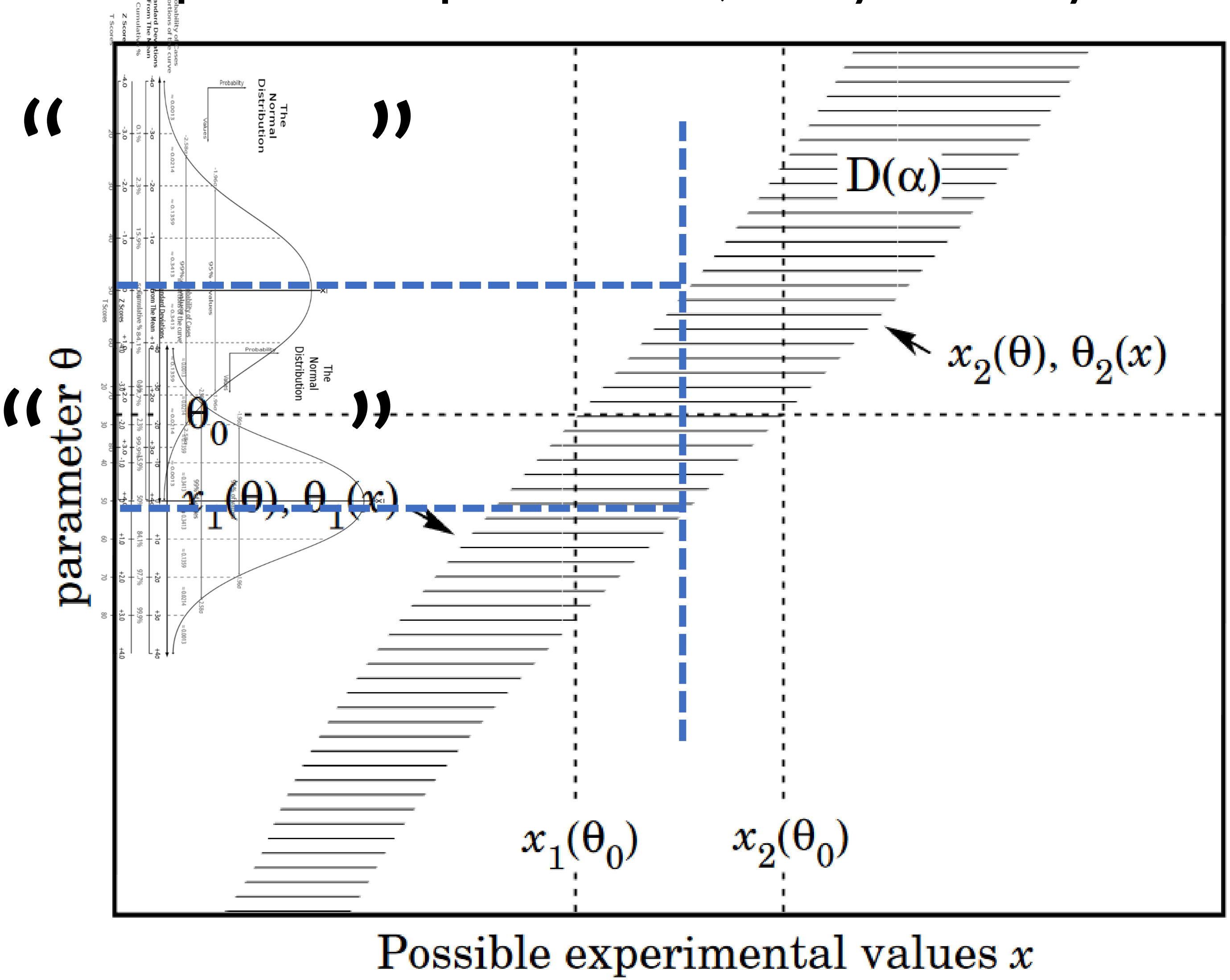
Poisson example

- Check for which parameters the result is in critical region



Poisson example

- Width depends on parameter, so symmetry broken!



Poisson example

- To find CI limits, solve

$$x \pm \sqrt{x} = 25$$

$$x = \frac{51 \mp \sqrt{101}}{2} = 25.5 \mp 5.02$$

- CI is shifted by 0.5 compared to naïve solution
  - It is also slightly wider
- Point estimate is still valid though!

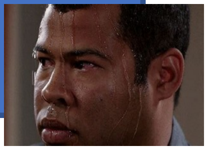
$$\lambda = 25^{+5.5}_{-4.5}$$

- Even better: use Poisson likelihood instead of z-score
- The variance of data at the parameter point estimate is not the same as the variance of the parameter!
  - The test statistic has to be calculated for each parameter point

Summary III

- Bayesian CI are constructed by updating a prior using Bayes' theorem
  - Conjugate priors can be used to make updating very easy
  - MCMC can be used if no closed form is available
  - Many ways to "cut" CI with right CL out of posterior
- Frequentist CI by defining a critical data region for each possible parameter value *a priori* (Neyman construction)
  - All values with measured data in crit. region are excluded
  - Often data is not used directly → test statistic as intermediary
  - Many ways to define critical region
- Homework

How does the speeding result change when looking for 99% CL lower limit, i.e. one sided z-score test?



JOHANNES GUTENBERG UNIVERSITÄT MAINZ JG|U

- More desirable properties for CI construction
  - Use data efficiently
    - Use all the information
  - Maximise test "power" = probability to exclude an untrue parameter from CI
    - Aim for short CI

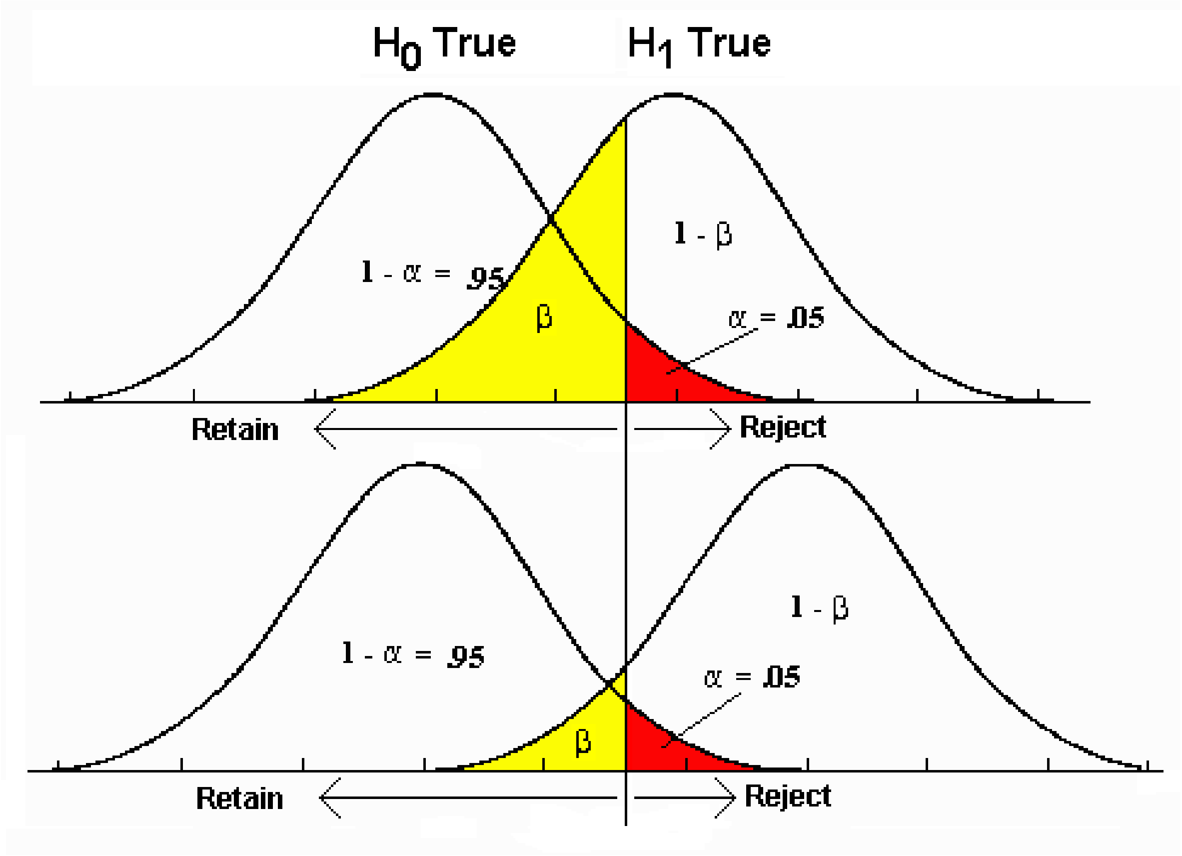
		Actual scenario	
		H_0 true	H_0 is false
Decision	Accept H_0	**Correct Decision**	**Type-II error: β** Accepting H_0 when it is false
	Reject H_0	**Type-I error: α** Rejecting H_0 when it is true	**Correct Decision** **1-β** **Power of test**

- To quantify power, need alternative hypothesis to evaluate probability → Hypothesis tests

Hypothesis tests

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

- H0 = hypothesis we want to test
- H1 = alternative hypothesis needed to evaluate β



- Select test statistic that for given α minimises β

Likelihood ratios

- Simple hypothesis = w/o any free parameters
  - Includes hypothesis where you fixed all parameters!
- Likelihood ratio proven to be most powerful test statistic (Neyman–Pearson lemma)

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x),$$

$$\Lambda(x) \equiv \frac{\mathcal{L}(\theta_0 \mid x)}{\mathcal{L}(\theta_1 \mid x)}$$

  - Alternative expression as log likelihood ratio
    - Easier to compute with (additions instead of multiplications)

$$\lambda = \ln \Lambda$$

- Need to determine critical value of ratio for every possible parameter point we want to exclude
  - E.g. MC simulation of distribution assuming H0
  - Reject H0 at low (!) likelihood ratios

Likelihood ratios

- Simple hypothesis = w/o any free parameters
  - Includes hypothesis where you fixed all parameters!
- Likelihood ratio proven to be most powerful test statistic (Neyman–Pearson lemma)

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x),$$

$$\Lambda(x) \equiv \frac{\mathcal{L}(\theta_0 \mid x)}{\mathcal{L}(\theta_1 \mid x)}$$

  - Alternative expression as log likelihood ratio
    - Easier to compute with (additions instead of multiplications)

$$\lambda = \ln \Lambda$$

No standard nomenclature!

- Need to determine critical value of ratio for every possible parameter point we want to exclude
  - E.g. MC simulation of distribution assuming H0
  - Reject H0 at low (!) likelihood ratios

Composite hypotheses

- Composite hypothesis = with free parameters

$$H_{0/1}(\boldsymbol{\theta}_{0/1})$$

- Can use ratio of maximum likelihoods as test statistic
  - Compare best fit point of H0 with best fit point of H1

$$\lambda = \ln \frac{\sup_{\boldsymbol{\theta}_0} \mathcal{L}(H_0(\boldsymbol{\theta}_0))}{\sup_{\boldsymbol{\theta}_1} \mathcal{L}(H_1(\boldsymbol{\theta}_1))}$$

- Not guaranteed to be most powerful test
  - But it is under certain circumstances (Karlin–Rubin theorem)
  - Usually considered to be pretty good
- Finding critical value becomes very hard in general case
  - What values to assume for free parameters in MC?

Wilks’ theorem

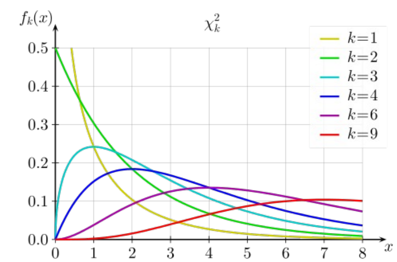
- Under some (generic) conditions
  - H0 is subset of H1
  - “Data is good enough”, e.g. MLE are norm. distr.
  - If point on H0 is true
- Likelihood ratio is χ^2 distributed

$$H_0(\boldsymbol{\theta}_0) \subset H_1(\boldsymbol{\theta}_1)$$

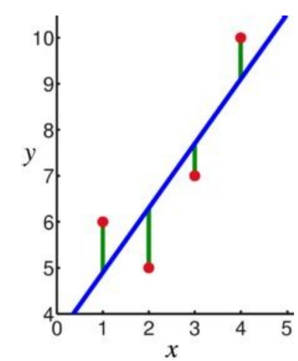
$$\lambda = \ln \frac{\sup_{\boldsymbol{\theta}_0} \mathcal{L}(H_0(\boldsymbol{\theta}_0))}{\sup_{\boldsymbol{\theta}_1} \mathcal{L}(H_1(\boldsymbol{\theta}_1))}$$

nparam

$$-2\lambda \sim \chi^2(k = ||\boldsymbol{\theta}_1|| - ||\boldsymbol{\theta}_0||)$$



- Can use χ^2 quantiles to calculate critical values
  - No need for time consuming MC
  - Reason why ndof = N(data points) – N(fit parameters) for GOF
    - H0 = Expectation value calculated from fit parameters
    - H1 = Expectation values of all data free
  - Bad “chi-squared” → we say the data does not fit
    - Actually a hypothesis test!

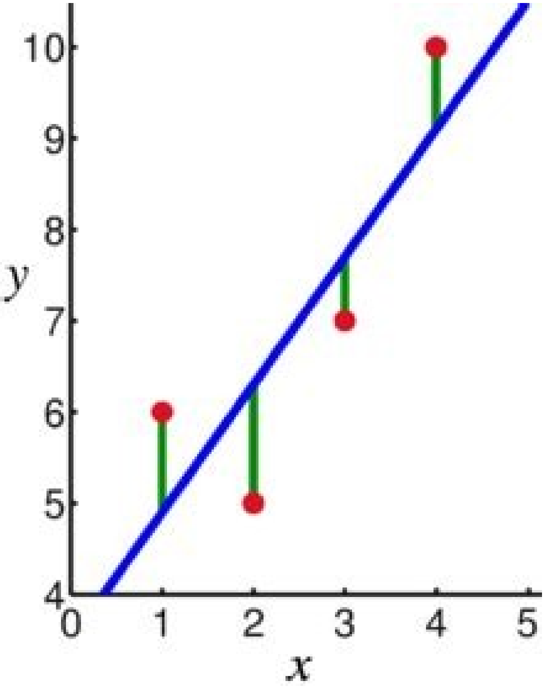


Used for CI construction

- When constructing CI have (at least) two choices for H1
  - H1 = most general hypothesis possible (ndof = ndata)
  - H1 = H0 but with all parameters free (ndof = nparam)
- In both cases ndof(H0) < ndof(H1)
  - (some) parameters fixed, where we want to check whether they are inside the CI or not
- Example: Linear fit to 4 points
  - H0: y = ax + b; a=1, b=4; ndof = 0
  - H1a: y = ax + b; a, b; ndof = 2
  - H1b: y1, y2, y3, y4; ndof = 4
- Case a: Will always find accepted region

$$a = \hat{a}, b = \hat{b} \rightarrow -2\lambda = 0$$

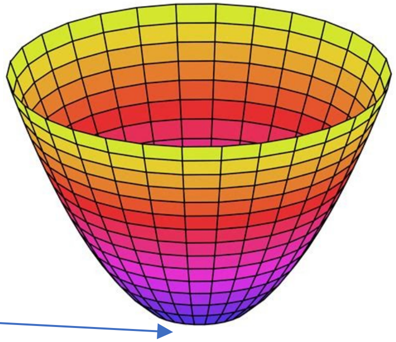
- Case b: CI might be empty (if fit is bad)



Normal approximation again

- Can provide more than just CI again (e.g. what fitters do)
  - Scan log-likelihood surface

$$-2\ln\mathcal{L}_0(\boldsymbol{\theta})\ [+\ 2\ln\mathcal{L}_1]$$



  - Find best fit (max likelihood, MLE) point

$$\hat{\boldsymbol{\theta}}$$

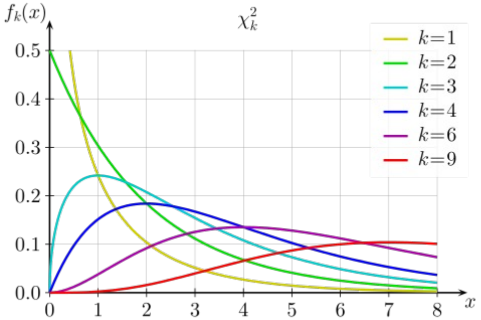
  - Use curvature around MLE (or other technique?) to approximate surface as quadratic function
    - In case of Gaussian uncertainties w/ fixed variance, this is exact!
  - Return covariance matrix S and MLE so that

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T S^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \approx -2\ln\mathcal{L}_0(\boldsymbol{\theta}) - (-2\ln\mathcal{L}_0(\hat{\boldsymbol{\theta}}))$$

    - RHS is chi-square distributed with ndof = nparam
    - LHS looks like Mahalanobis distance!

(Case "a")

$$D_M^2 \sim \chi_N^2$$



  - Can use MLE and covariance to construct CI
    - As if it described a PDF of the parameters

Summary IV

- Defining an alternative H1 and Type II error helps us decide on a test statistic
  - Minimise P(Type II) for a given P(Type I)
- Likelihood ratios are usually a very good choice
- When $H_0(\boldsymbol{\theta}_0) \subset H_1(\boldsymbol{\theta}_1)$ (and other requirements)

$$-2\lambda \sim \chi^2(k = ||\boldsymbol{\theta}_1|| - ||\boldsymbol{\theta}_0||)$$

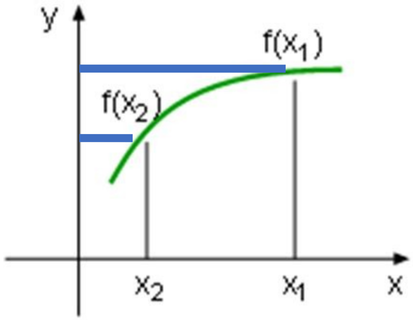
- Likelihood surface (function of parameters) often approximated as quadratic function → covariance matrix
  - Gaussian approximation → symmetry → easy CI construction
- Homework

Show that the a two-sided z-score and likelihood ratio tests are equivalent for normally distributed data.

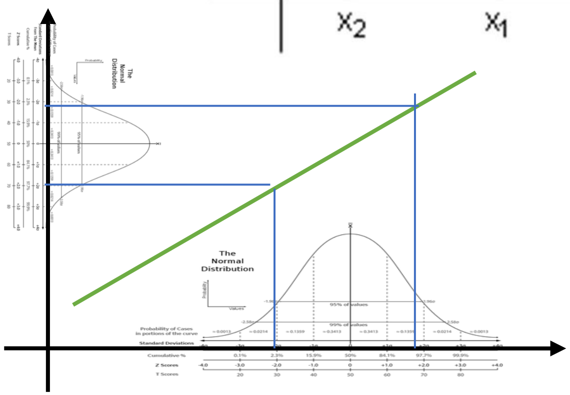
Propagation of uncertainty

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

- Have uncertainty of parameters covered
- How to propagate to uncertainty of prediction?
- Monotone 1D functions → easy
  - f(CI edges)



- Linear function → even better
  - Gaussian approx. → new Gaussian



$$\hat{y} = f(\hat{x}), \quad \sigma_y = \frac{\mathrm{d}y}{\mathrm{d}x}\sigma_\theta = a\sigma_x$$

- N-dim linear combination
  - N-dim Gauss → M-dim new Gauss

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{b}$$
$$\hat{\boldsymbol{y}} = A\hat{\boldsymbol{x}} + \boldsymbol{b}$$
$$\Sigma^y = A\Sigma^x A^T$$

$$\Sigma^x = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \cdots \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \Sigma_{11}^x & \Sigma_{12}^x & \Sigma_{13}^x & \cdots \\ \Sigma_{12}^x & \Sigma_{22}^x & \Sigma_{23}^x & \cdots \\ \Sigma_{13}^x & \Sigma_{23}^x & \Sigma_{33}^x & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

If M > N, new covariance will be degenerate!

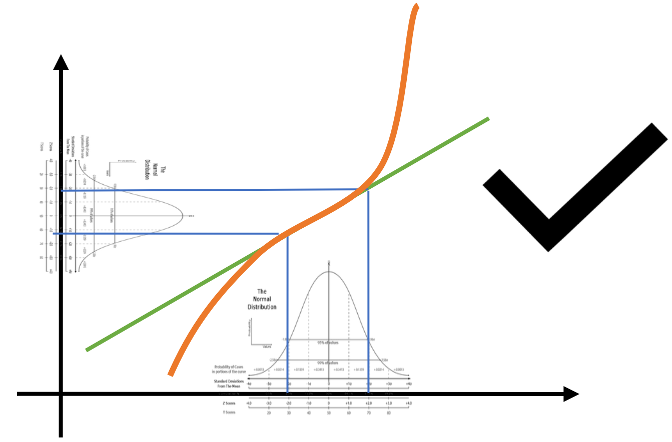
Local linear approximation

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

- Non-linear function, but “straight” on scale of uncert.
  - Approximate as linear (Taylor expansion)

$$y = A \cdot (x - \hat{x}) + b$$

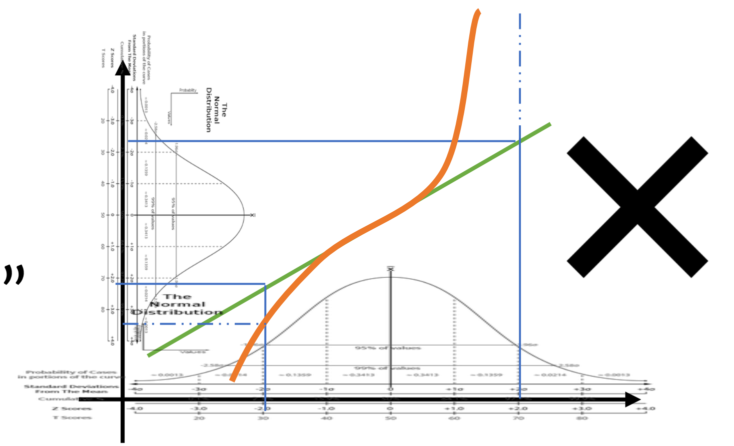
$$b = y(\hat{x}), \quad A_{ij} = \left.\frac{\partial y_i}{\partial x_j}\right|_{\hat{x}}$$



  - Rest stays same as in linear case

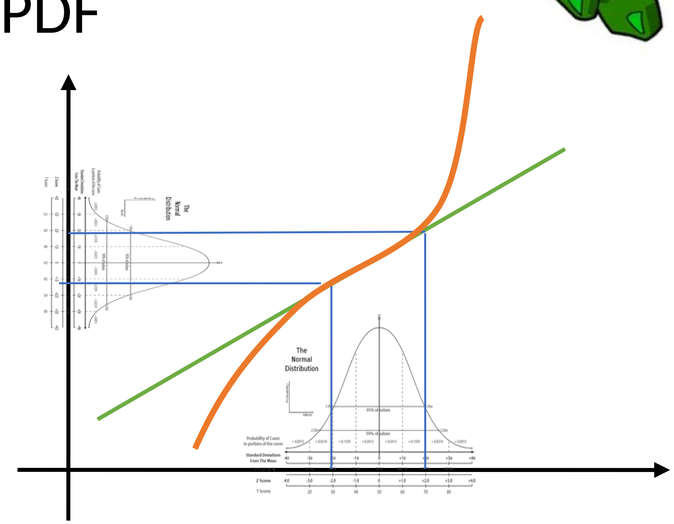
$$\Sigma^y = A \Sigma^x A^T$$

$$\sigma^2_{y_k} = \Sigma^y_{kk} = \sum_{i,j} \frac{\partial y_k}{\partial x_i} \frac{\partial y_k}{\partial x_j} \Sigma^x_{ij}$$



- No one ever checks “straightness”
  - Coverage tests are important!

- Function difficult to differentiate?
  - Throw parameters
  - Calculate function for each throw
  - Extract uncertainty on prediction from sample
- Always possible in Bayesian statistics
  - Parameter uncertainty is probability distribution
- Frequentist? Harder to justify
  - Uncertainty describes likelihood, not a PDF
    - But only ratios matter!
  - Does the right thing in linear case
  - Distorts (relative) likelihoods when not linear
    - What we want: $\mathcal{L}(y) = \mathcal{L}(x(y))$
    - What we get: $\frac{\mathrm{d}x}{\mathrm{d}y}(y)\mathcal{L}(x(y))$



Summary V

- Propagation of uncertainty works as expected in Ideal Linear Normal Land
  - Just use "regular" error propagation
    - Analytical or MC
  - Works in Frequentist and Bayesian

- When function is not linear enough
  - Bayesian MC method still works
  - No simple solution for Frequentists (I am aware of)
    - At least not in the general N-dimensional case
    - when monotone, can calculate $\mathcal{L}(x(y))$ or translate CI edges directly
- Homework

We measured the sides of a cube to be (2 +- 0.3) mm.
What are the uncertainty and CI on the volume?
At what significance have we shown that volume > 0?

Summary Summary

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

- Statistics can be hard
- Understanding will lead to better physics results
  - Blindly following “rules” can lead to mistakes
  - Understanding comes with taking this seriously over and over
  - Question what you are doing until you know it makes sense!
- Frequentist probabilities are strictly defined, “objective”
  - Though talking about parameters/uncertainties is a pain
- Bayesian probability definition is softer, “subjective”
  - Much easier to think/talk about
- Further reading
  - Wikipedia
  - PDG Particle Data Booklet
  - Cowan (1998) – Statistical Data Analysis
  - Bohm – Introduction to Statistics and Data Analysis for Physicist (free PDF!)
  - Papers/books referenced in the above
    - Can be dense, conventions/lingo differs between stats and physics