More on confidence intervals

- because that is what we do!

Morten Medici (mmedici@nbi.ku.dk)

Methods of advanced statistics, March 2018

Brief recap

- slides adapted from R. Barlow

Confidence intervals

- Important part of the statistical reporting of results
- Especially relevant for results which are basically null results.
  - E.g. upper limits on the branching ratio (BR) of a particle decaying in a certain way, testing for new physics:

$$BR < 10^{-20} \text{ @ } 90\% \text{ CL}$$

- Where we have a trade-off between statistical power and size of the interval, e.g.

$$BR < 10^{-19} \text{ @ } 95\% \text{ CL}$$
$$BR < 10^{-20} \text{ @ } 90\% \text{ CL}$$

What is “@ 90% CL”?

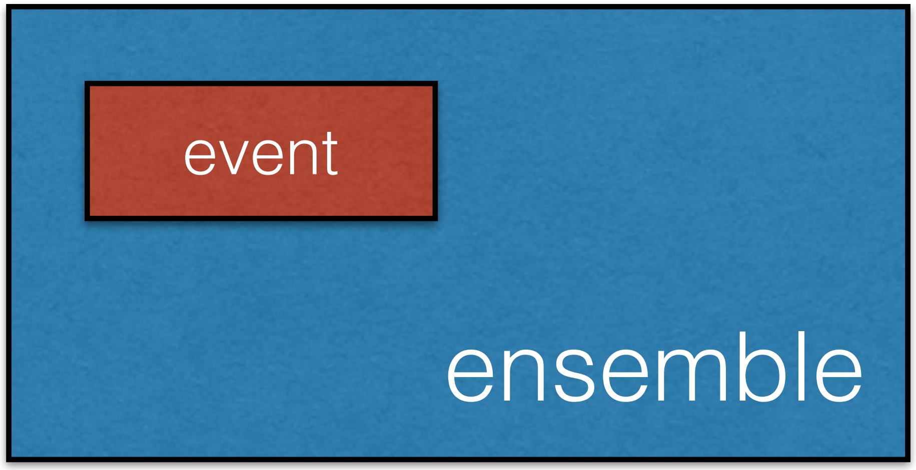
- It is **not** stating “the probability that the result is true”
- Confidence levels are **not** probabilities for results
- However, they are strongly linked to probabilities, so let us take a slight detour

Probability

- The probability of an event to occur is equal to the fraction of experiments where the event occurs compared to all experiments (ensemble), in the limit of a large number of experiments

$$P(\text{event}) = \lim_{N \to \inf} \frac{N_{\text{event}}}{N}$$

- Examples
    - Coin toss: P(tail) = 50%
    - Tau decay: $P(\tau^- \text{ to } \mu^- \nu_\mu \nu_\tau) = 17.4\%$



Depends on ensemble

- The probability is dependent on the event **AND** the ensemble
- Example: '*Nordic study shows that men above 50 with a well-payed job have a 1% risk of getting skin cancer*'
- So a 50-year old danish male has a 99% chance of reaching 51 without getting cancer? **No**
- It all depends on the ensemble you choose
  - Danish males in the study,
  - Danish males
  - Nordic males
  - Male sunbather champions
  - etc…
- Each give a different probability. All values will be valid (if done correctly!)

Probabilities are dependable quantities.. right?

- The probability of the tau lepton decaying to a muon (τ^- to $\mu^-\nu_\mu\nu_\tau$) is 17.4%.
  (*I looked that up in the Particle Data Group (PDG) booklet, so it must be true…*)

- Though in a given analysis that select muons, the fraction of tau leptons that decay to muons might be >17.4%

- If a given analysis is trying to reject muons, the probability might be < 17.4%

- It depends on the ensemble! So does the result in the PDG!

Caveat: When there is no ensemble

- Consider the statement:

“It is likely to be cloudy tomorrow”

or even

“There is a 90% probability for cloudy weather tomorrow”

- There is only one tomorrow. There is no ensemble!
- So P(clouds) is either 0/1=0 or 1/1=1
- Strict frequentists will not be able to arrive at such a statement (could be done with a Bayesian approach)

Getting around the caveat

- Frequentist can instead compile an ensemble of statements, and determine that some of them are true:

  The statement ‘*It will be cloudy tomorrow*’ has a 90% probability of being true

- Translates to defining
  P(clouds) = P(’*It will be cloudy tomorrow*’ is true)

- Where in this case
  P(clouds) = 90%

Still, ensembles matter

- P(cloudy) = 90% can be true at the same time as P(cloudy) = 50% is true

- P(cloudy) = 90% can be true at the same time as P(sun) = 90% is true

- Depending on the ensembles used in the individual studies used to claim those probabilities!

m_τ = 1776.82 +/- 0.16 MeV (at 68% CL)

- 68% of all tau particles have a mass between 1776.66 and 1776.98 MeV? **WRONG**

- The probability of tau-mass being in the range 1776.66-1776.98 MeV is 68%? **WRONG**

- The tau-mass has be measured to be 1776.82 using a technique which give it a 68% probability of being within 0.16 MeV of the true result? **CORRECT**

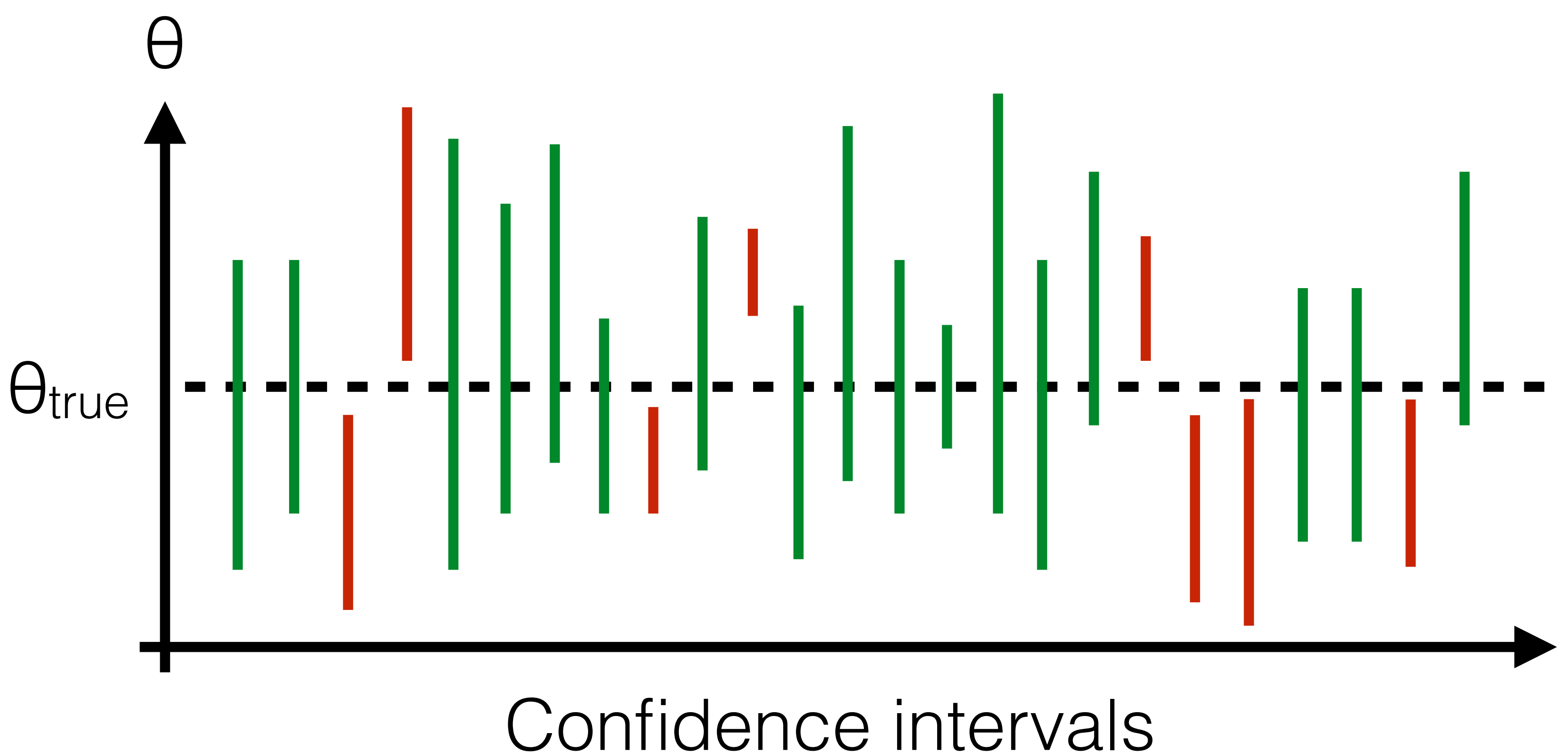
m_τ = 1776.82 +/- 0.16 MeV (at 68% CL)

- Said differently: The statement “the tau-mass is in the range 1776.66-1776.98 MeV” has a 68% probability of being true.

- We add the information about the confidence limit to illustrate this: m_τ = 1776.82 +/- 0.16 MeV at 68% confidence level (CL)

Confidence intervals

- If the experiment is repeated many times, we would get different intervals (ensemble of statements).
- They would be true 68% of the cases, as they would bracket the true value in 68% of the cases.



Confidence/significance

- Confidence level, CL = 1-α
- Significance α, is used when talking the language of hypothesis testing
- A 95% CL result might be stated inversely, e.g.
- '*The medicine was effectively reducing the risk at the 5% level*'
  = If the medicine does nothing, the probability of getting an improvement this size (or better) is 5% (or less)
- Hypothesis testing: Given an observation/measurement the corresponding probability is called the p-value, and the null hypothesis is rejected if p-value < α
- We use this exact approach to construct the intervals

Construction of classic frequentist intervals

- also known as the Neyman construction

Confidence interval - known true value

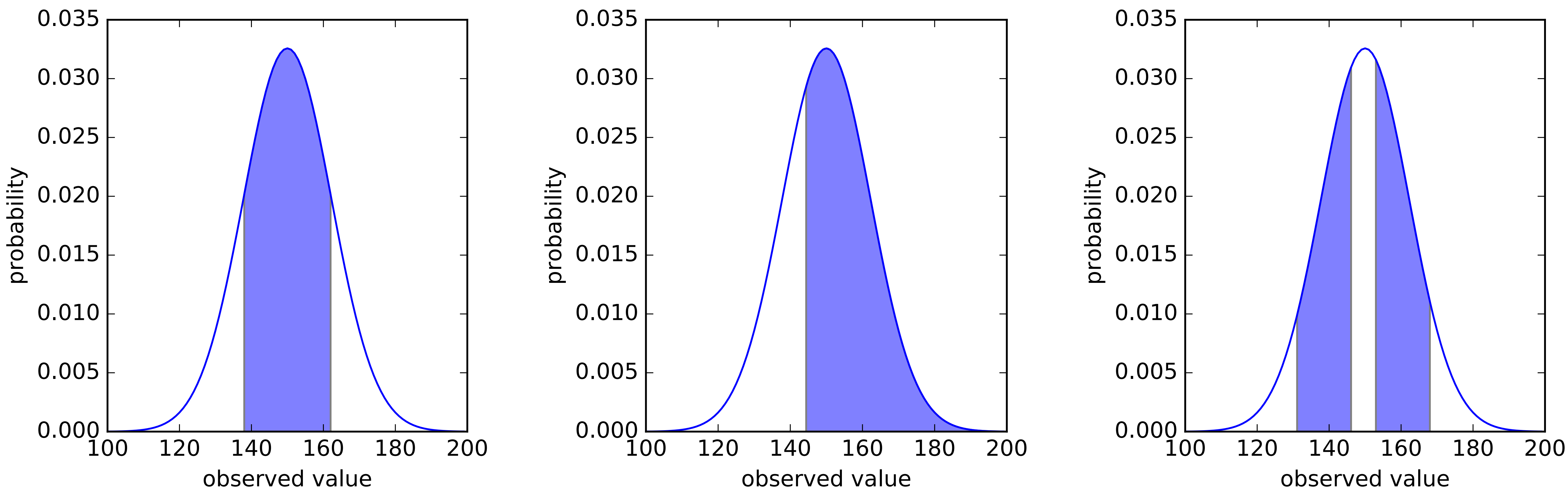
- The frequentist approach can give a statement about the probability of observing a specific value of a parameter given the probability density function (PDF).

- Use the expression for the PDF to calculate the probability of getting *n* within the interval *[a,b]* for a parameter value of θ:

$$P(n \in [a, b] | \theta) = \int_a^b P(n|\theta)$$

Intervals, intervals, intervals

- You decide which intervals you want to do, though a connected two- or one-sided interval is <u>normally used</u>
- All shaded intervals below hold 68% of all possible outcomes of a Gaussian PDF, with mean = 150 and variance = 150



Determine the underlying parameter

- When you know the parameters of a process you can predict the distribution of outcomes

  Hypothesis -> Data (Experiment)

- However, we are often in the situation where we want to infer an estimate of a parameter from data

  Data -> Hypothesis (Statistics)

- That is the real power of confidence intervals (both for frequentist or Bayesian approaches)

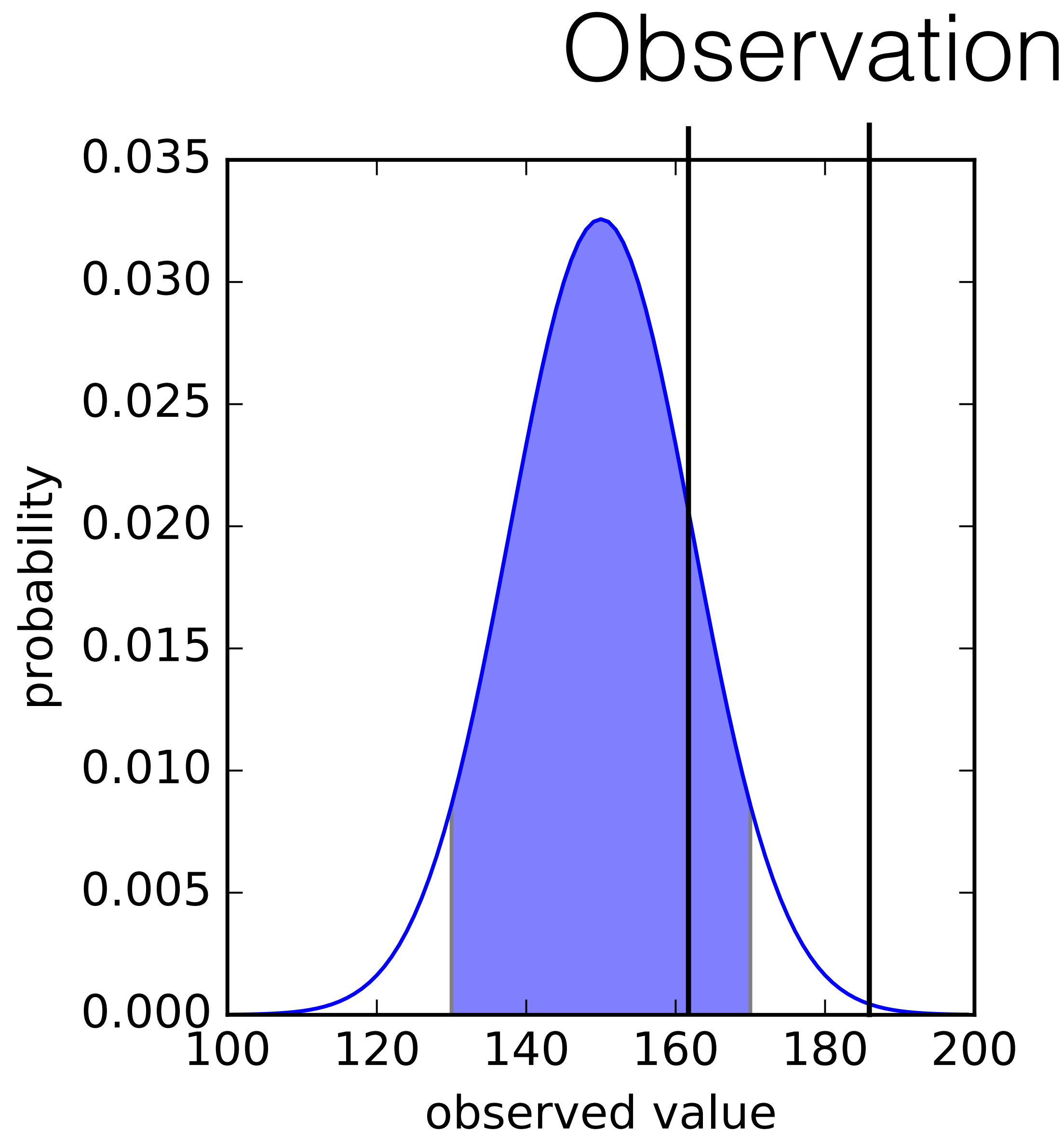
Hypothesis rejection

- An observation of a parameter value that lies outside the 90% confidence interval given a hypothesis (true value) will be rejected at a 90% CL

- However, most often we do not know the true value of the parameter

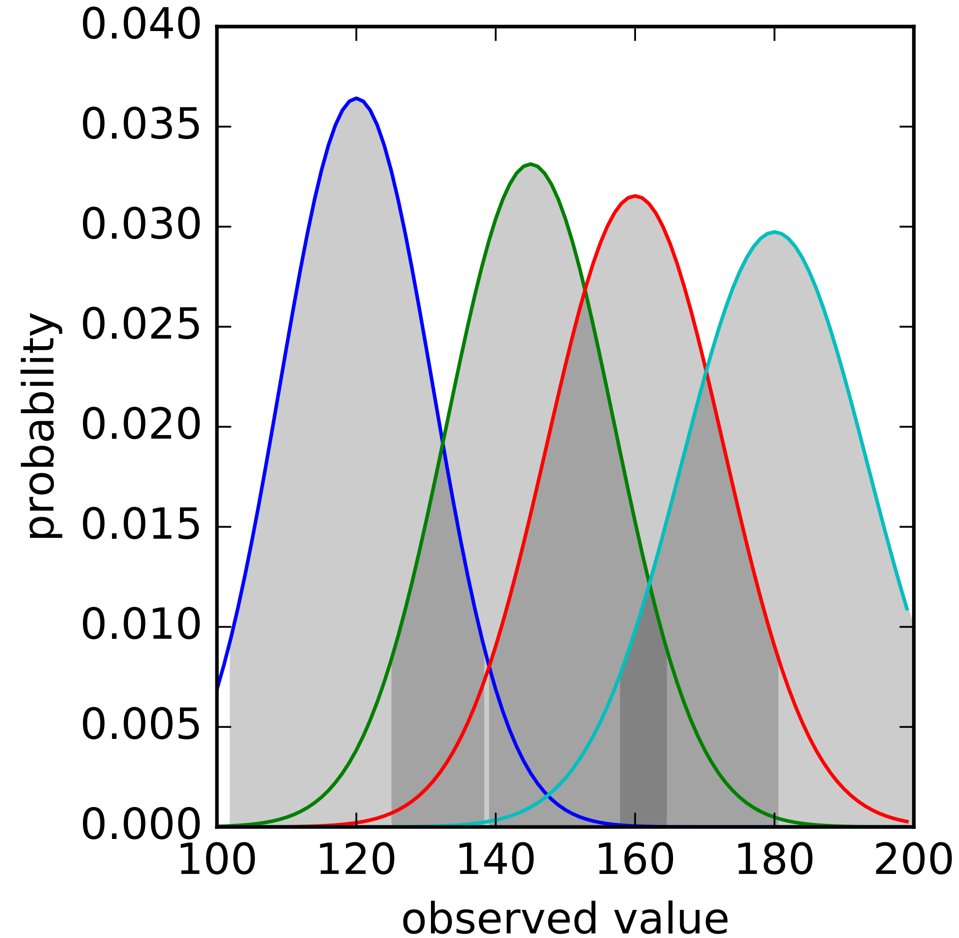
- It could have a different value than we assumed in our hypothesis

- Hence we should look at other hypotheses.



Hypothesis rejection

- Each hypothesis will have an interval within which an observation will confirm (or ’accept’) the hypothesis
- For multiple possible true values of the parameter, these ‘acceptance intervals’ can be determined
- Example figure: 90% central interval for a few true values

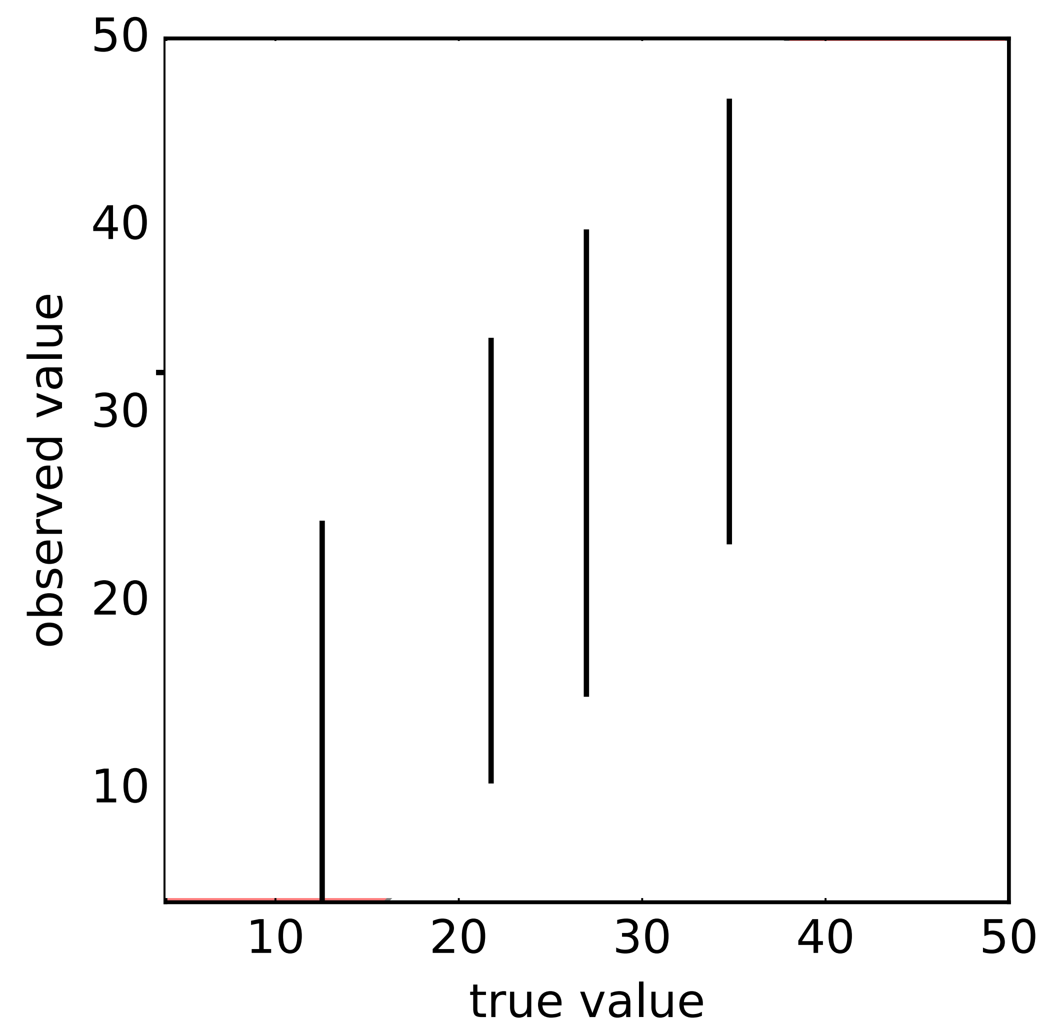


Acceptance belt

- This produce a band ('acceptance belt') that connects the observed value of the parameter to the true value with the correct frequentist interpretation

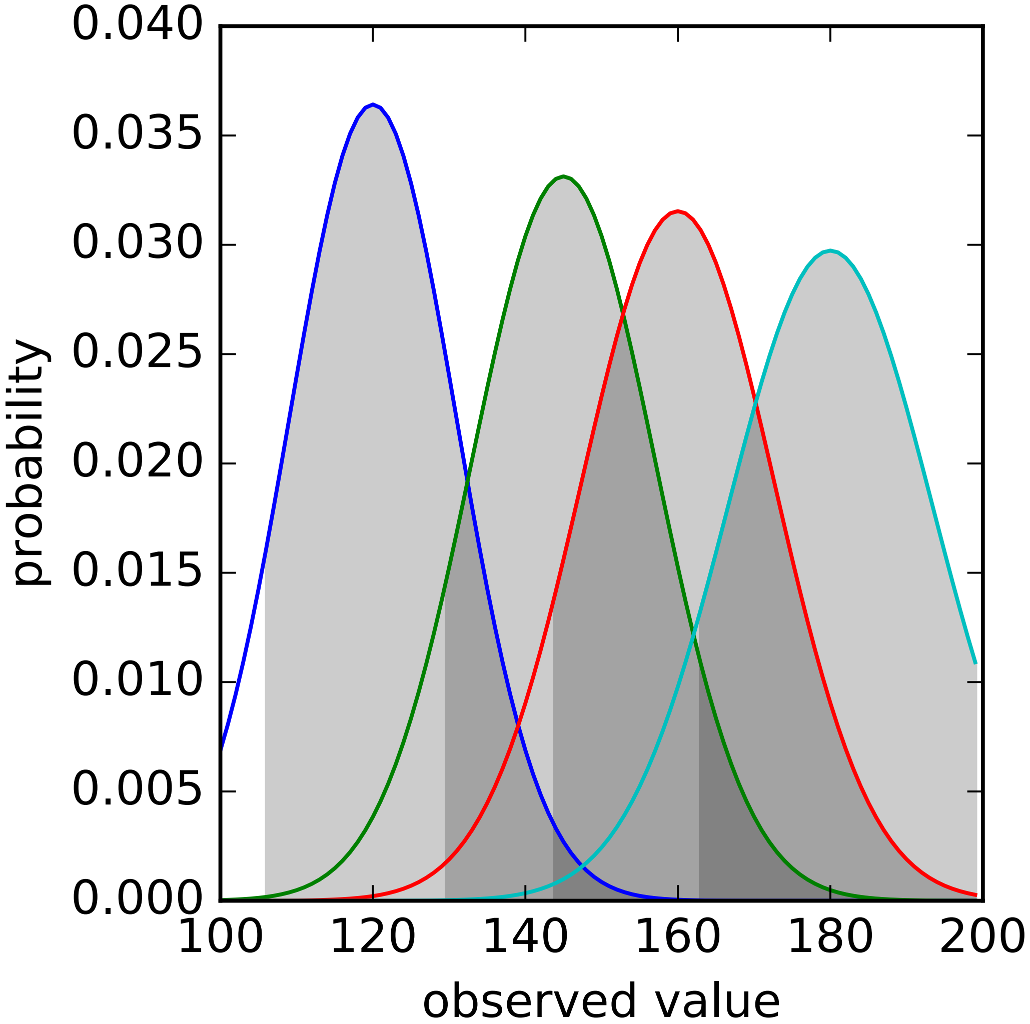
- For a given observation, the interval on the true parameter can be determined at a given CL

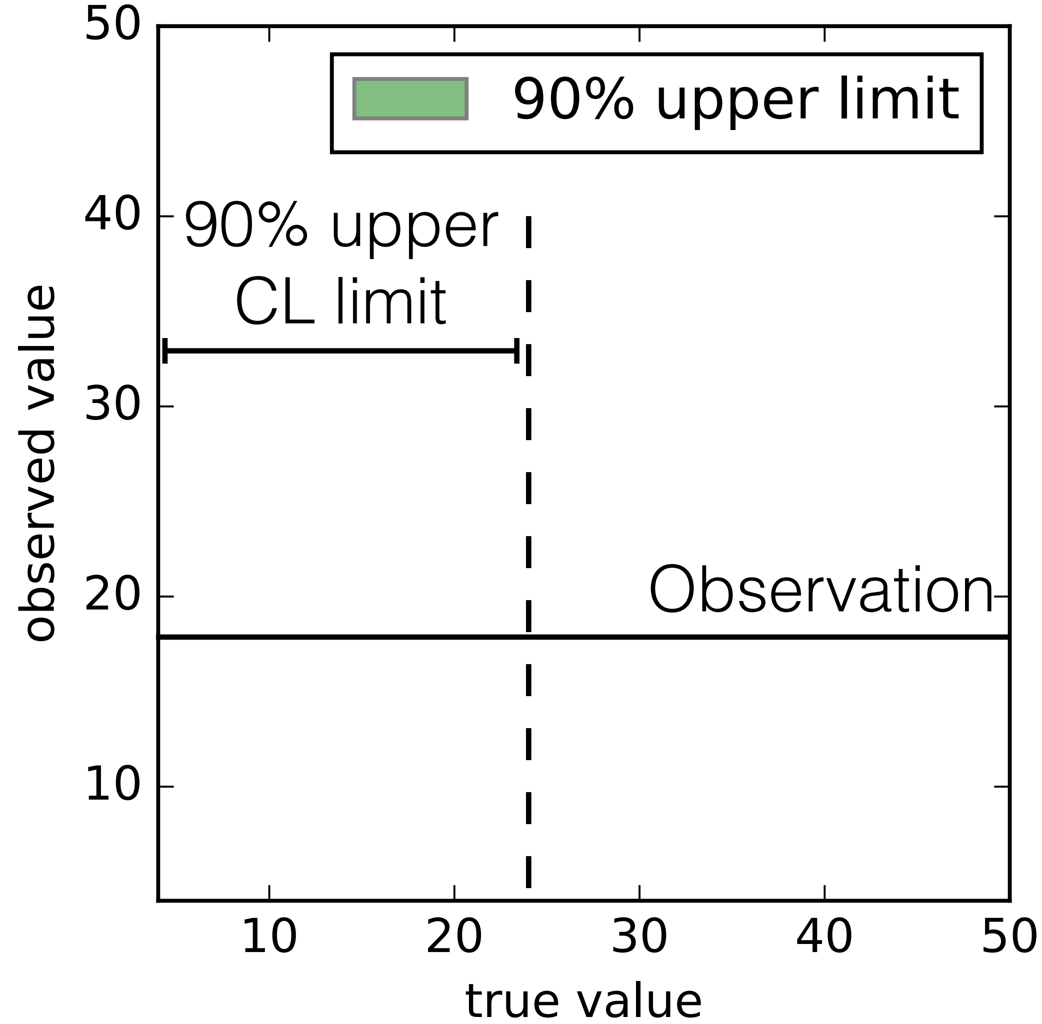
- By construction, this method gives confidence intervals which contain the true value with an exact known probability.
(90% in the example to the right)



Acceptance belt

- Similarly can we produce the acceptance belt for a 90% upper limit





Exercise 1

- Assume a measurement of θ, that is distributed with a Gaussian from the true value θ_{true} with variance equal to one. Do the following:

1. Plot the 68% central limit acceptance belt for values of θ_{true} between zero and ten, analytically or numerically (steps of 0.1)

2. From the plot, determine the 68% central limit on θ_{true} resulting from an observation of $\theta_{obs} = 8$.

3. <u>Extra</u>: Repeat the exercise with a 68% upper and lower limit. Repeat at a 90% CL and 95% CL and compare the value of θ_{obs} required to set a lower limit above 0

Exercise 1

- Resulting limits on n

$n_{obs}=8$	lower	upper	central
68 %	7.5	8.4	7.0-9.0
90 %	6.7	9.3	6.3-9.7
95 %	6.3	9.6	6.0-10.0

Complications for classic frequentist intervals

Complication: Discrete observations

- If we use e.g. the poisson formula as a PDF, we can only count integer values (even though θ can be non-integer)

$$P(n|\theta) = e^{-\theta}\frac{\theta^n}{n!}$$

- To make a 68% lower limit:
  - Include 0,1,2,3,4 to get 57.0%
  - Include 0,1,2,3,4,5 to get 73.6%
- Be conservative and include 5, even though it corresponds to 'too much' probability
- Actually, the probability of getting something above 5 is less than the 32% originally intended

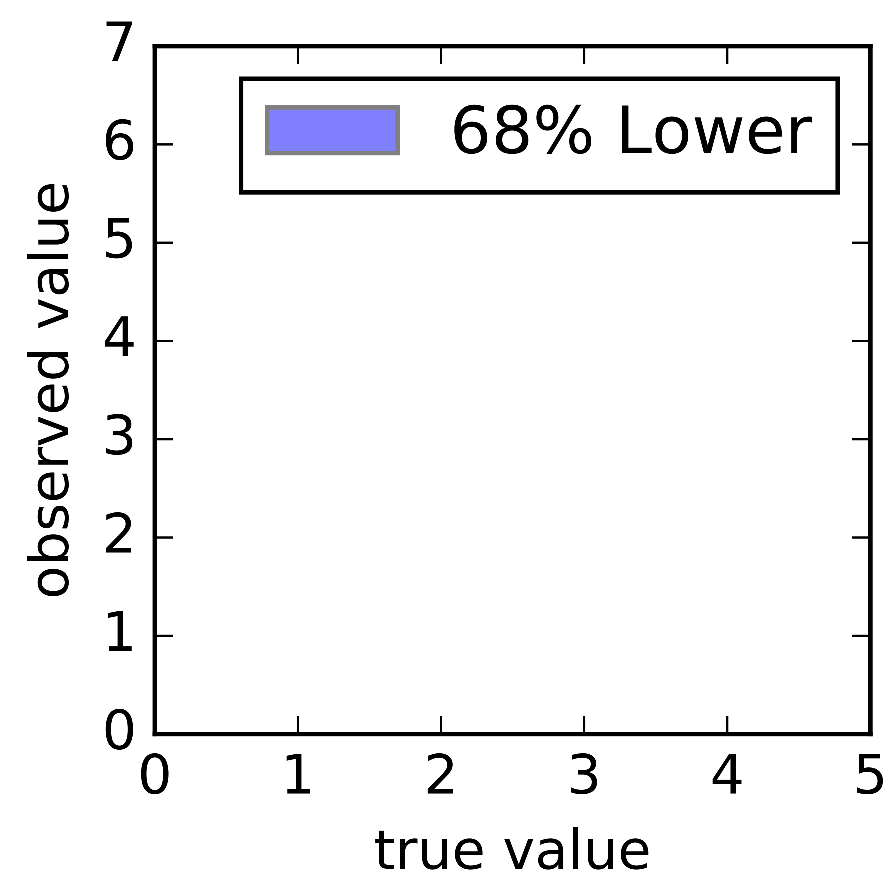
n	P(n I 4.3)
0	1.4 %
1	5.8 %
2	12.5 %
3	18.0 %
4	19.3 %
5	16.6 %
6	11.9 %

Complication: Discrete observations

- For both a poisson with θ = 4.3 and θ = 4.5, the same 5 values of n would have to be included in the 68% lower limit

- This will be the case over a range of values of θ, so the confidence belt will change in steps

- Multiple true values will cover the same range of observed values

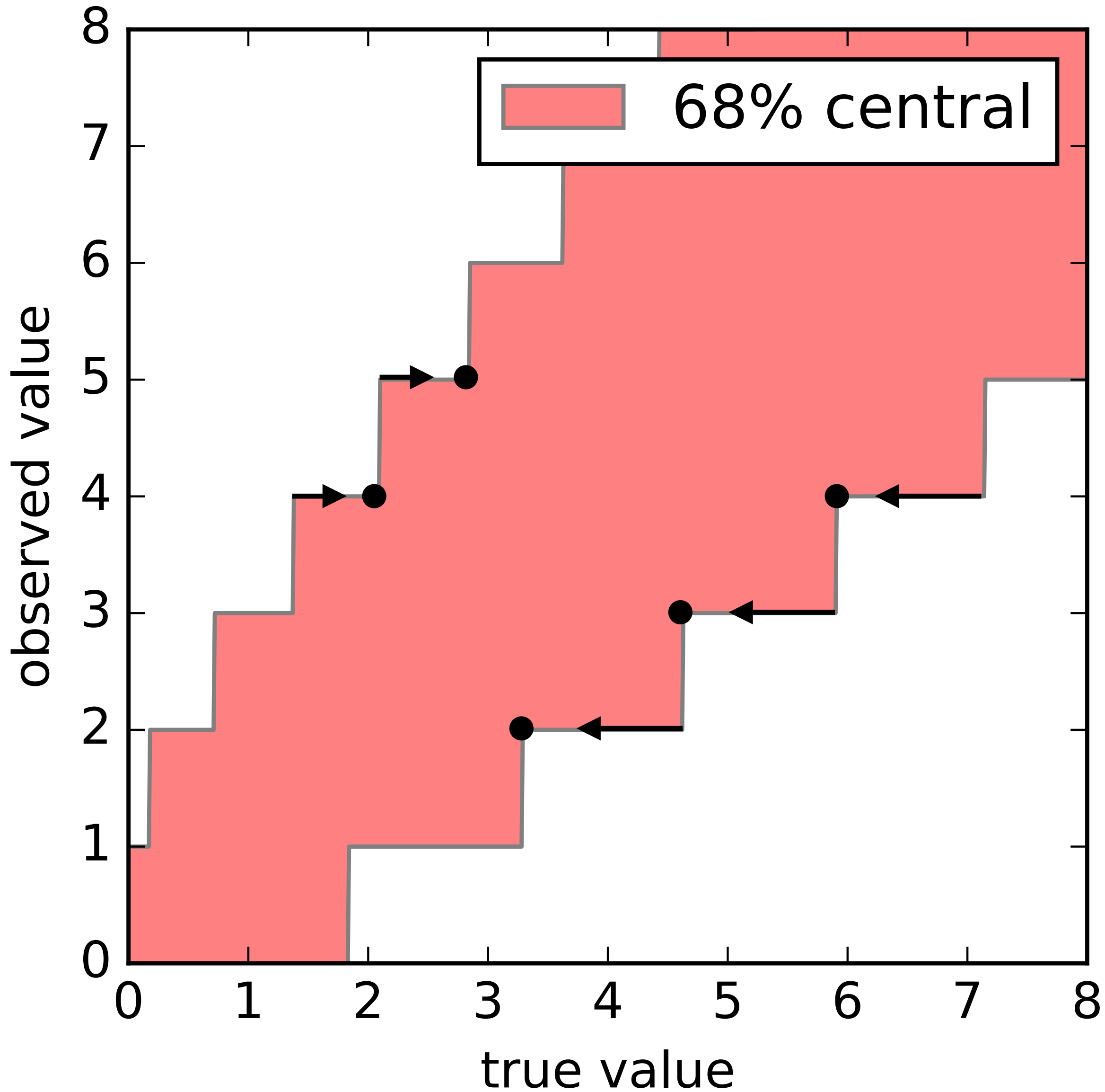


Coverage

- A frequentist test may have a coverage greater than the confidence level = over-coverage
- Though it should never undercover (by construction)
- If it undercovers, the analyser did something wrong!

Complication: Discrete observations

- Use smallest true value of θ for upper limit and largest true value for lower (which correspond to the correct CL)



report the center-most points

other points will overcover

Exercise 2

- Same as exercise 1, produce a 90% central limit acceptance band assuming a **poisson** PDF, between true values of 0 and 15 in steps of 0.1 or less.

- Assume you measure n = 8 events, which confidence interval do you report?

- <u>Extra</u>: Determine the coverage across numerous values of θ

Exercise 2

- Results

$n_{obs}=8$	lower	upper	central
68 %			
90 %	4.7	13.0	4.0-14.4
95 %			

Upper limits

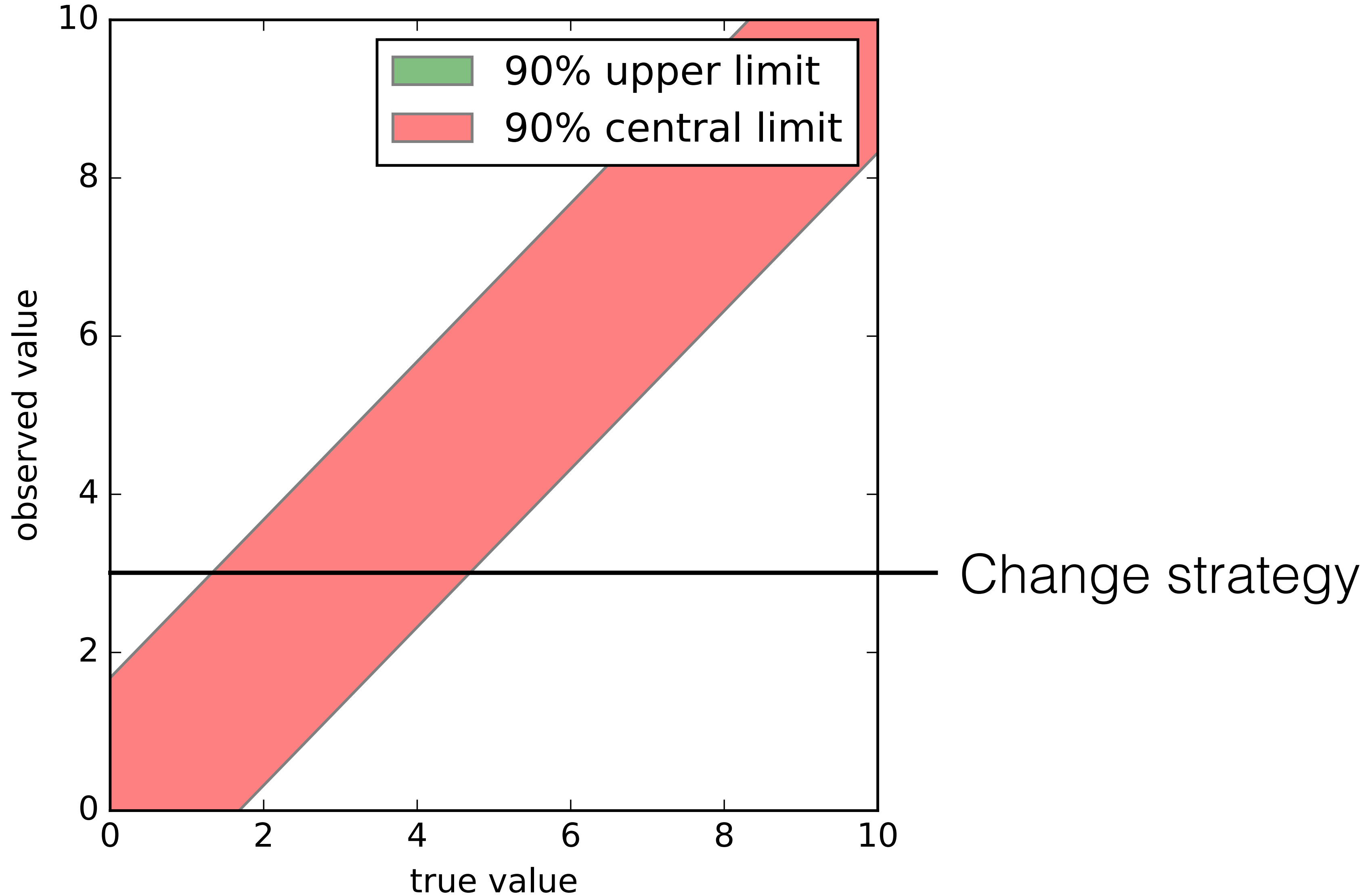
- Consider the case of observing n_{obs} events
- We assume poisson uncertainty on the number of observed events given a true number of events n
- The number of events are expected to be small, so after our observation we will be reporting a 90% upper limit on n.
- Example, if zero events are observed (n_{obs} =0), a 90% upper limit of 2.3 can be set.

The hunt for discoveries

- If the signal s is expected to be small, it would be sensational if the number of observed events is significantly above 0
- In that case we could be inclined to calculate a central limit instead, to illustrate a discovery
- So depending on the number of observed events we will quote either an upper limit or a central limit

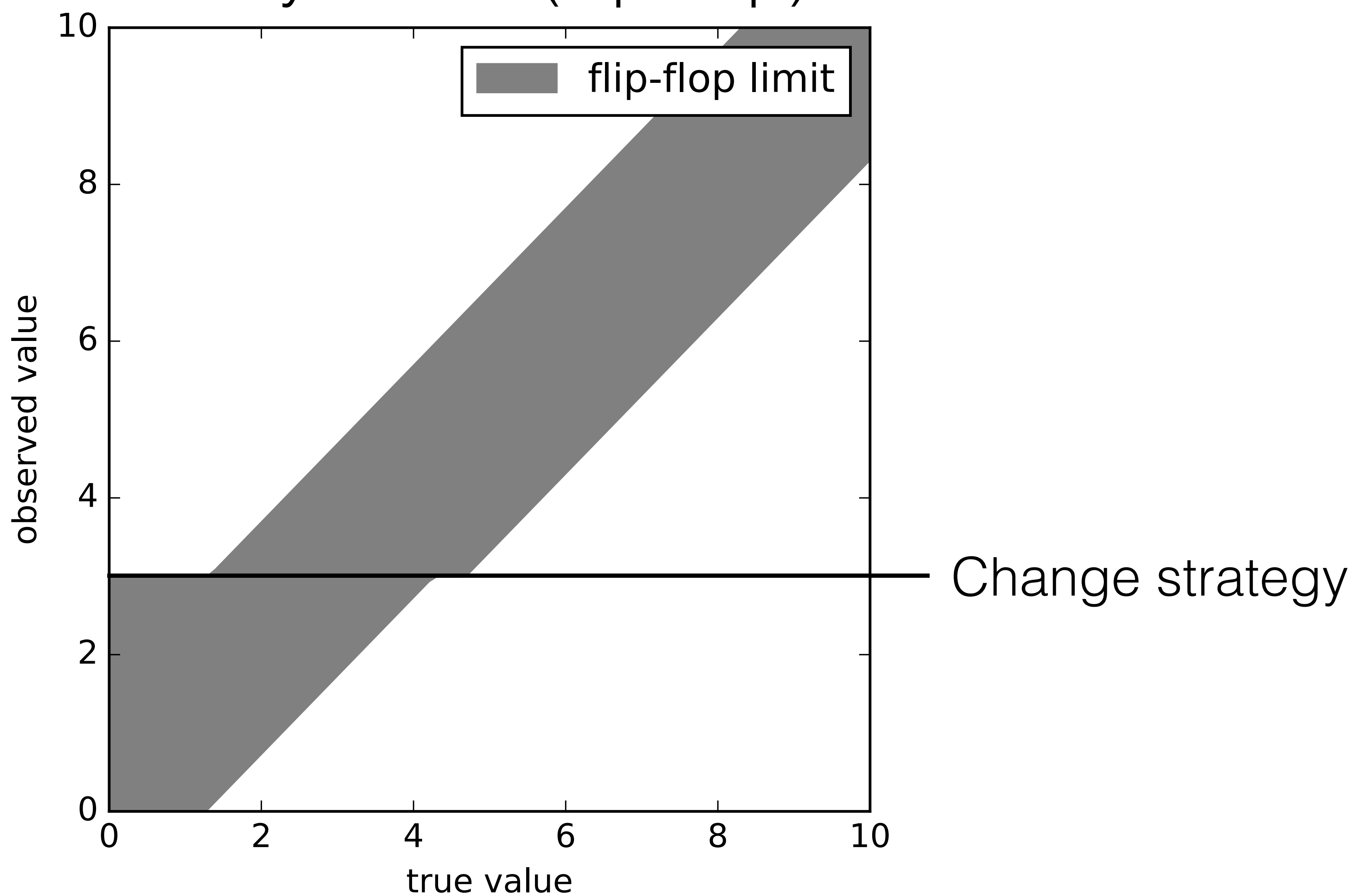
Complication: Choosing strategy later

- Assume gaussian PDF with σ = 1, with the strategy of changing from 90% upper limits to 90% central limit if the observation is 3σ away from 0 (flip-flop)



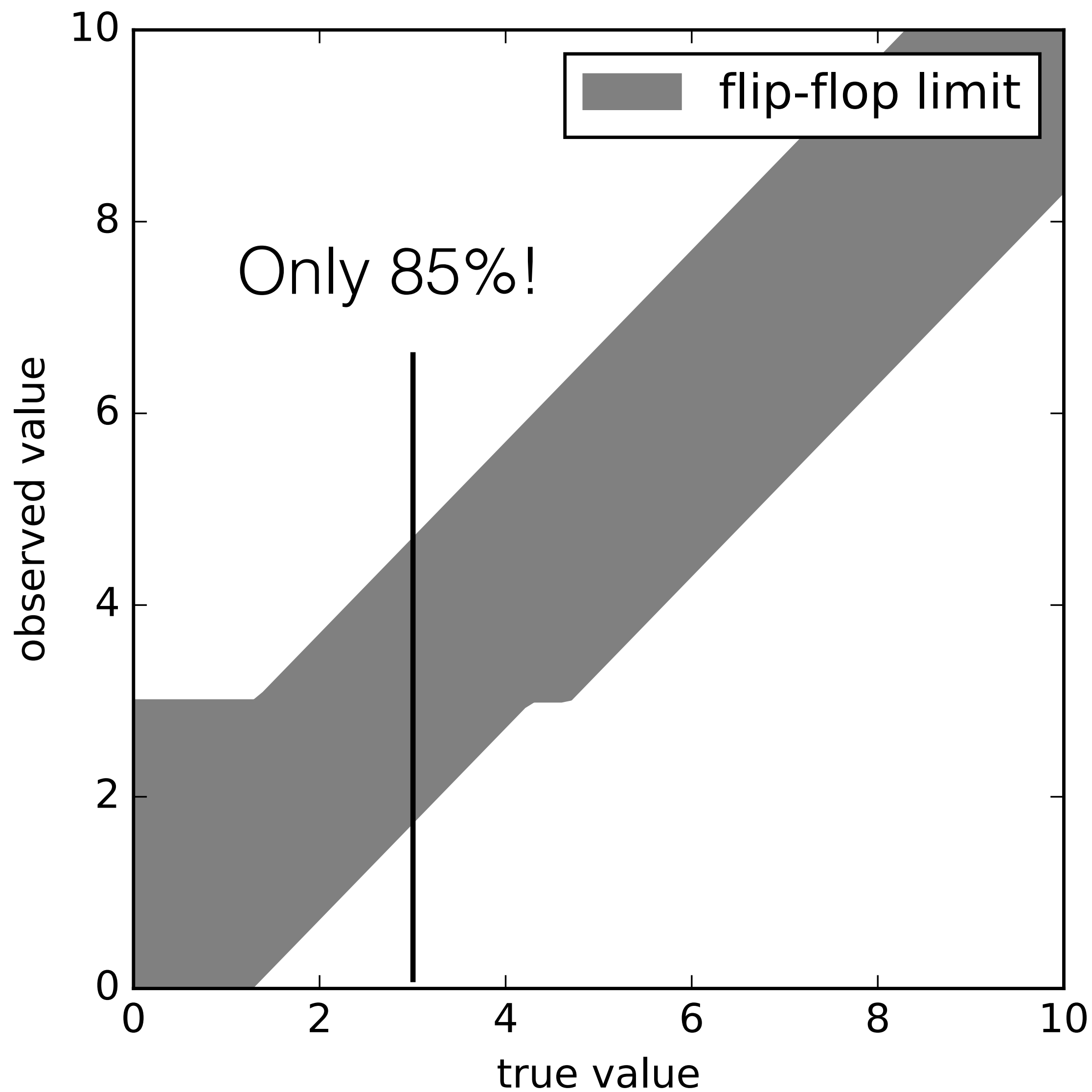
Complication: Choosing strategy later

- Assume gaussian PDF with σ = 1, with the strategy of changing from 90% upper limits to 90% central limit if the observation is 3σ away from 0 (flip-flop)



Complication: Choosing strategy later

- Problem: Part of the range only has 85% coverage, not the 90% that we designed the method for



Complication: Choosing strategy later? No!

- In order for the coverage to be meaningful, the type of limit must be decided ahead of time

- Only way to get around the issue: Stick to the ideal approach:

    1. Choose strategy (upper/lower or central limit)

    2. Examine data

    3. Quote result

Signal+background

- Consider the case of measuring a number of events $n = n_s + n_b$
- With n_s and n_b corresponding to the number of signal and background events, respectively
- Both signal and background are given by gaussian distributions with mean s and b, and variance equal to one
- The signal is expected to be small, so after our observation we will be reporting a 90% upper limit on s.

Complication: Constrained parameters

- Since we are counting events, the number cannot be negative
- Assume the background mean is known, $b = 7$
- For $n_{obs} = 4$ we can determine that $N = s+b \sim 5.3$ (at 90% CL)
- Hence we can conclude that $s < -1.7$ (at 90% CL)
- Or can we? The number of events should be zero or above

Complication: Constrained parameters

- Do we claim $s < -1.7$ (at 90% CL)?
- Answer: The interval will only cover the right result 90% of the time, this is one of those 10%-cases
- Answer: We should publish this result to avoid biasing the reported numbers
- Answer: This is clearly unphysical, we can not publish a result based on a broken approach, we should use a statistical method that fixes this

Feldman-Cousins Method

- also known as the “Unified Approach” (mainly by G. Feldman and R. Cousins)

See paper: G J Feldman and R D Cousins, *Unified approach to the classical statistical analysis of small signals*, Phys Rev D, 1998 vol. 57 (7) pp. 3873-3889.

Approach

- Introduce ranking principle based on the following likelihood ratio, or rank:

$$R(n) = \frac{L(n|\theta)}{L(n|\theta_{\mathrm{best}})}$$

- With the likelihood value of observing n given a true value θ, or the best fit value of the parameter θ_{best} given the dataset and any constraints on θ

- Completely rethink the construction of acceptance intervals for the acceptance belt: For a given true value θ, include values of n to the interval from highest rank R(n) to lower, until the desired confidence is reached

Approach

- Determine the PDF for your hypothesis, which will provide the likelihood used

- For each true value θ:

  1. Determine for all possible outcomes n:

     A. The value θ_{best} that maximises the likelihood L

     B. Calculate the rank R(n)

  2. Construct the acceptance interval by including the values of n, that has the highest rank R(n) to lower until the desired confidence is reached

Approach - Example

- Assume a Poisson measurement, so $L(n|\theta) = \text{Poisson}(n|\theta)$
- For a Poisson the ML estimator is $\theta_{best} = n$

1. We determine the acceptance interval for one true value (e.g. $\theta = 1$)
2. Repeat 1. for multiple values of θ

Approach - Example

- Assume a Poisson measurement with true value θ = 1
- ‘rank’ indicates in which order the values of n are included for a 90% interval

n	$P(n \vert \theta=1)$	θ_{best}	$P(n \vert \theta_{best})$	R(n)	rank
0	0.368	0	1	0.368	3
1	0.368	1	0.368	1	1
2	0.184	2	0.271	0.680	2
3	0.061	3	0.224	0.274	
4	0.015	4	0.195	0.079	
5	0.003	5	0.175	0.017	

Example: Constrained Gaussian

- Consider again the case of measuring a number of events
  $n = n_s + n_b$

- Where again both the signal and background are given by Gaussian distributions with mean s and b, and variance equal to one

- Assume the background mean is known, b = 3

- So if we observe n = 1, which effectively corresponds to $n_s = n - b = -2$

Example: Constrained Gaussian

- However, when determining the 90% confidence interval on s, we have to require that, s > 0

- So we incorporate this in the definition of s_{best}:

$$s_{\mathrm{best}} = \begin{cases} n - b & \text{if } n > b \\ 0 & \text{otherwise} \end{cases}$$

- And use that when we calculate R(s)

- For each signal true value the acceptance interval [α,β] is determined such that

$$90\% = \int_{\alpha}^{\beta} P(n|s) \qquad \text{and} \qquad R(\alpha) = R(\beta)$$

Example: Constrained Gaussian

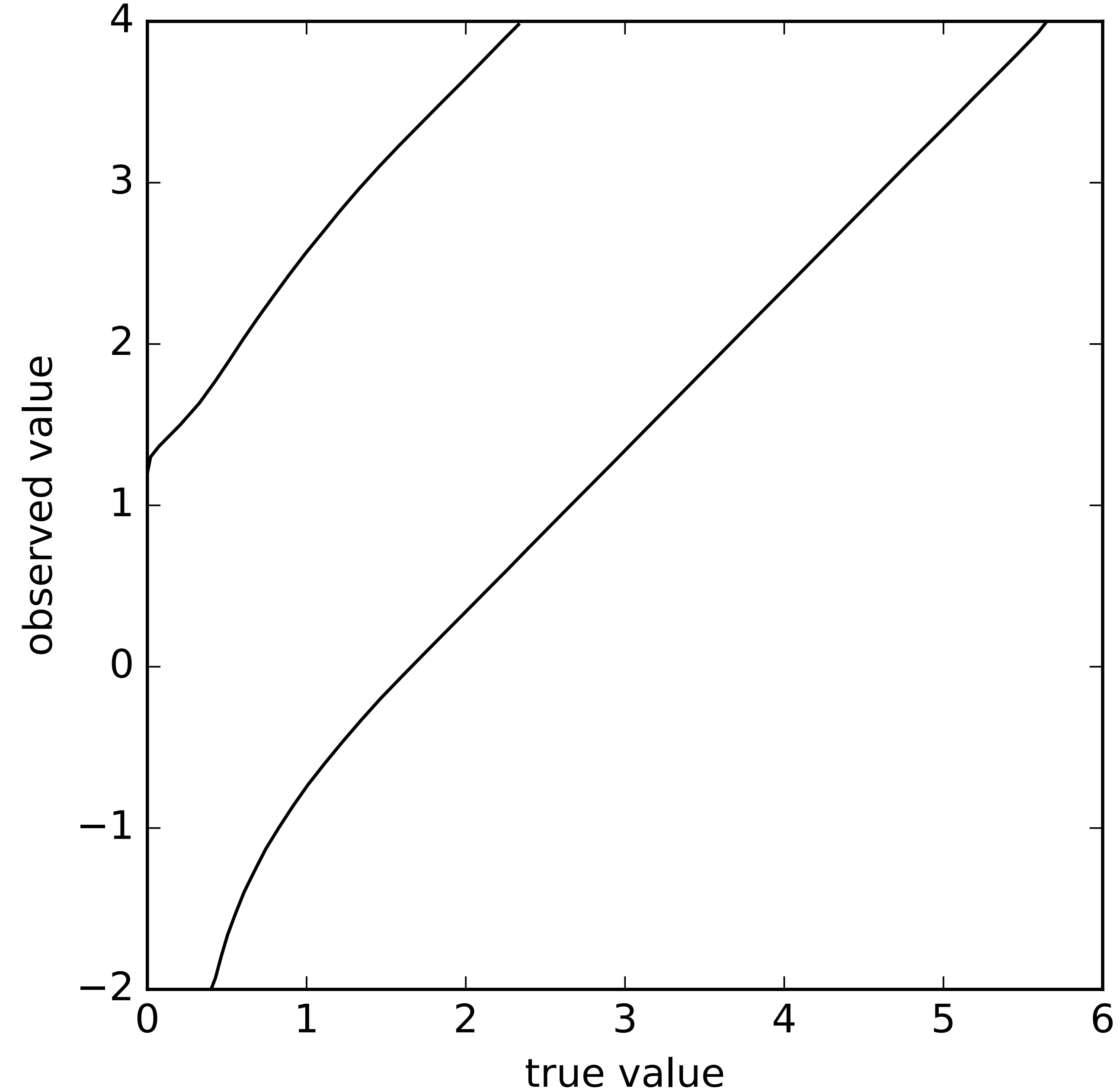
- Shown is the 90% confidence belt when applying the FC for a known background of b = 3

- It automatically transitions between an upper limit and a central limit

- Decides for you whether an upper limit or central limit is appropriate to quote based on the observation

- If we observe $n = n_s + n_b = 2$ the measured number of signal events is effectively $n_s = -1$

- The corresponding 90% interval is then $s < 0.81$ (at 90% CL)



Argument against (0)

- <u>Argument</u>: It is more cumbersome to implement!
- Yes. But, if your problem does not offer any other way around you will have to use it
- Just because it is right, does not mean that it is easy

Argument against (1)

- Argument: Takes power away from analysers!
- Yes. But that exactly why this method should be used. Such that your results are statistically sound (if applied correctly!)
- You are welcome to choose the CL, but once chosen, this method invalidates the conventional approach of having to make a choice

- Experiment 1 (spent time/money removing backgrounds):
  - $b = 0$, $n_{obs} = 1$
  - Feldman-Cousins limit: $s < 2.44$ (at 90% CL)

- Experiment 2 (less optimised):
  - $b = 10$, $n_{obs} = 1$
  - Feldman-Cousins limit: $s < 0.75$ (at 90% CL)

- Argument: This is unfair to the hardworking group!

- But experiment 2 needs to get extremely lucky to get zero events, and lucky experiments will always quote better limits (though averaging out luck, experiment 1 will be better off)

Exercise 3

- For a measurement of n which is distributed by a Poisson distribution from the true value n_S.

1. Determine Feldman-Cousins 90 % acceptance belt

2. Suppose you observe n = 10 events what is the 90% confidence interval on n_S, what if you observe n = 1?

3. Compare to the central limit using the Neyman method

Exercise 3 - extra

- Similarly to the previous exercise, now assume there is a known background component. So we have a Poisson measurement of

$$n = n_s + n_b\text{, with a known background of } n_b = 4$$

- Include the constraint: $n_{best} = 0$ for $n_{obs} < 0$

1. Determine Feldman-Cousins 90 % acceptance belt
2. Suppose you observe n = 10 events what is the 90% confidence interval on n_s, what if you observe n = 1?
3. Compare to the central limit using the Neyman method
4. <u>Extra</u>: Determine the coverage across the considered values of n
5. <u>Extra Extra</u>: Do the calculations for 68% and 95% and various values of n_{obs} .

Exercise 3