

# More on confidence intervals

- because that is what we do!

Morten Medici  
Methods of advanced statistics  
March 2017

# Brief recap

- slides borrowed 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:  
$$\text{BR} < 10^{-20} @ 90\% \text{ CL}$$
- Where we have a trade-off between statistical power and size of the interval
  - $$\text{BR} < 10^{-19} @ 95\% \text{ CL}$$
  - $$\text{BR} < 10^{-20} @ 90\% \text{ CL}$$

# What is “@ 90% CL”?

- It is not just 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

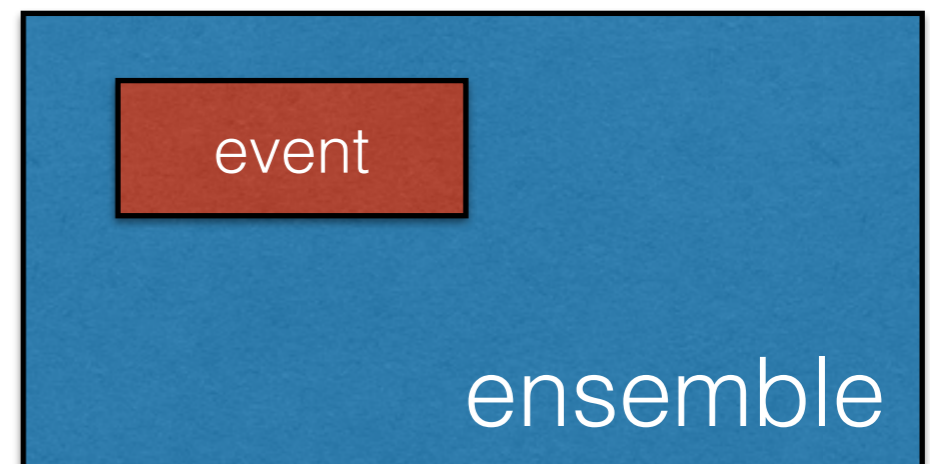
- The probability of an event to occur is equal to the fraction of experiments where the event occurs compared to all experiments, in the limit of a large number of experiments

$$P(\text{event}) = \lim_{N \rightarrow \infty} \frac{N_{\text{event}}}{N}$$

- Examples

- Coin toss:  $P(\text{tail}) = 50\%$

- 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-paid 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 ( $\tau^-$  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 larger than 17.4%
- If a given analysis is trying to reject muons, the probability might be lower than 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  $P(\text{clouds})$  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
$$P(\text{clouds}) = P(\text{'It will be cloudy tomorrow' is true})$$
- Where in this case
$$P(\text{clouds}) = 90\%$$

# Still, ensembles matter

- $P(\text{cloudy}) = 90\%$  can be true at the same time as  $P(\text{cloudy}) = 50\%$  is true
- $P(\text{cloudy}) = 90\%$  can be true at the same time as  $P(\text{sun}) = 90\%$  is true
- Depending on the ensembles used in the individual studies used to claim those probabilities!

$$m_{\tau} = 1776.82 \pm 0.16 \text{ MeV}$$

(at 68% CL)

- 68% of all tau particles have a mass between 1776.66 and 1776.98 MeV?
- The probability of tau-mass being in the range 1776.66-1776.98 MeV is 68%?
- The tau-mass has been measured to be 1776.82 using a technique which gives it a 68% probability of being within 0.16 MeV of the true result?

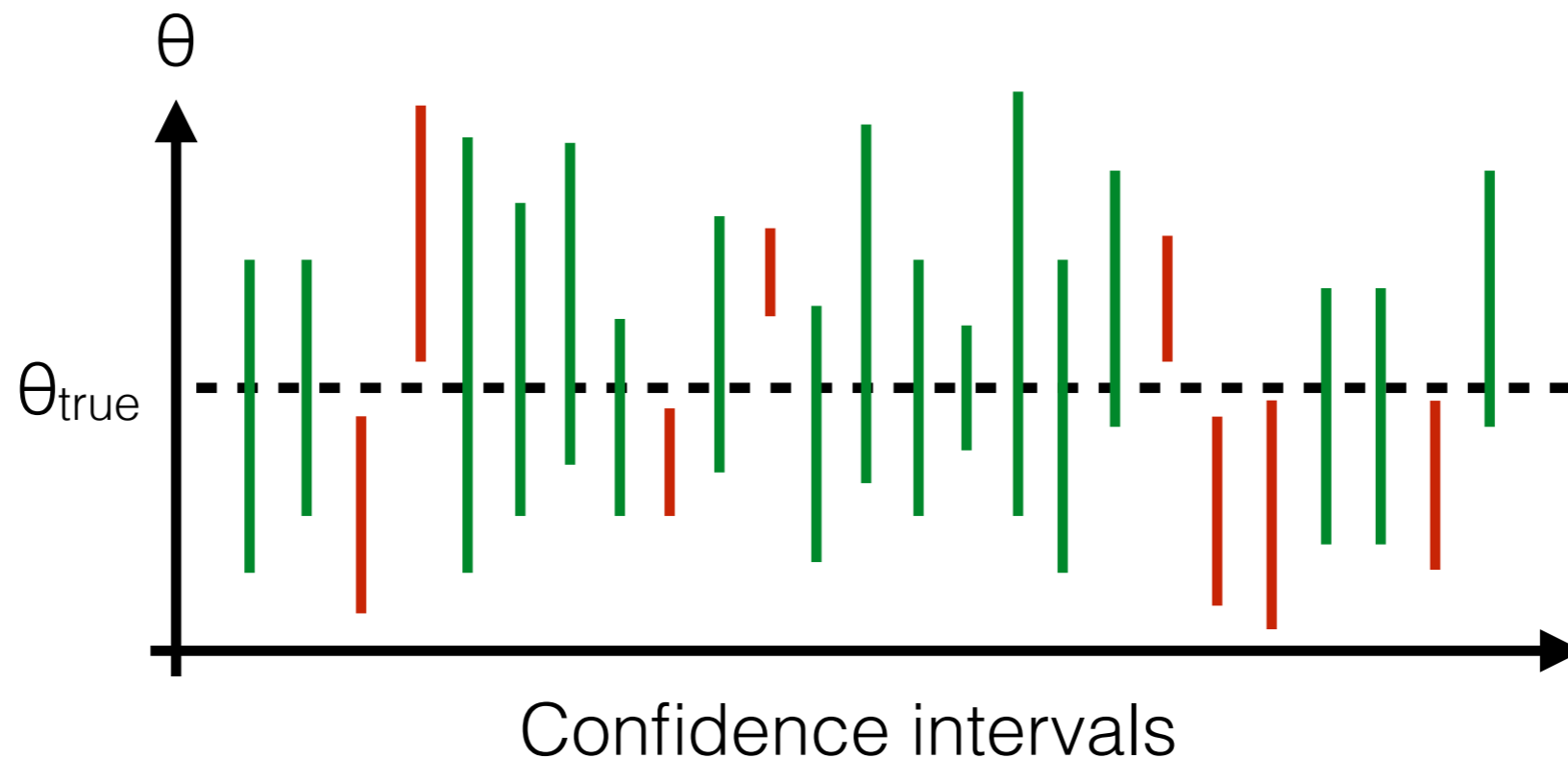
$$m_{\tau} = 1776.82 \pm 0.16 \text{ MeV}$$

(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:  $m_{\tau} = 1776.82 \pm 0.16 \text{ MeV}$  at 68% confidence level

# 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,  $CL = 1 - \alpha$
- Significance  $\alpha$ , is the language of hypothesis testing
- E.g. 'The medicine was effectively improving the 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  $p\text{-value} < \alpha$
- 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

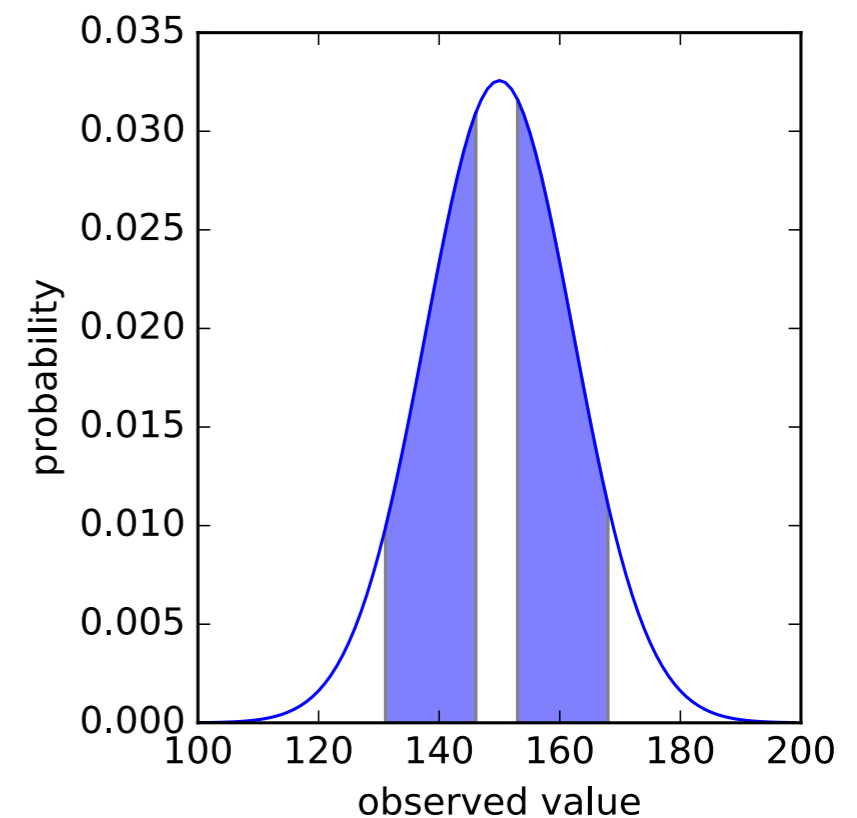
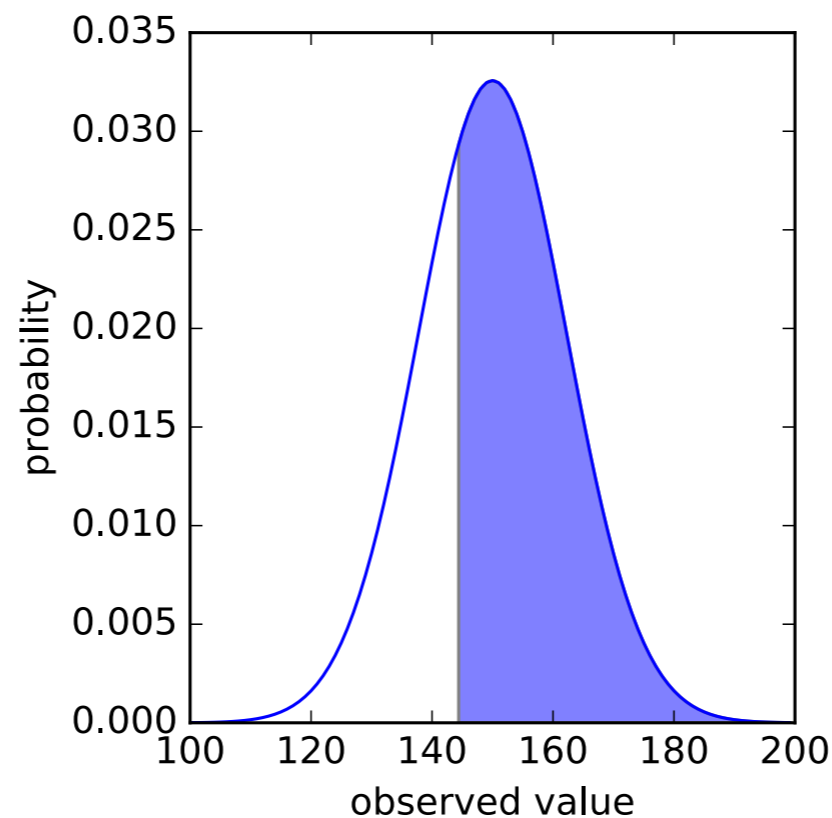
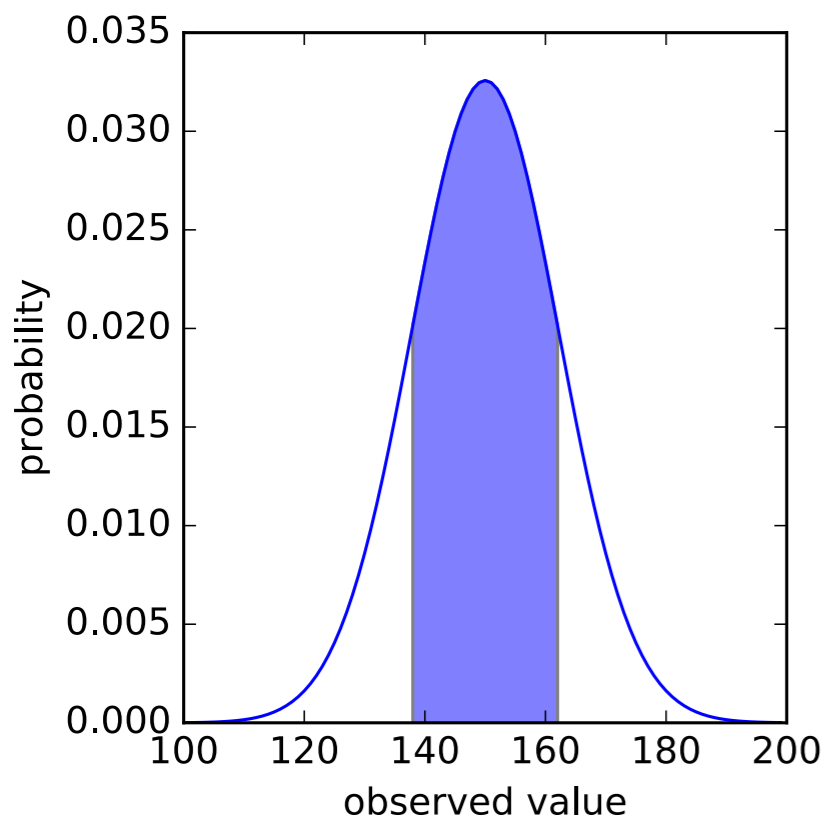
- The frequentist approach can give a statement about the probability of observing a specific value of a parameter given the PDF.
- Use the expression for the PDF to calculate the probability within a given interval

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



# Intervals, intervals, intervals

- You decide which intervals you want to do, though a connected two- or one-sided interval is normally used
- All shaded intervals below hold 68% of all possible outcomes, given a known 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  $\rightarrow$  Data (Experiment)

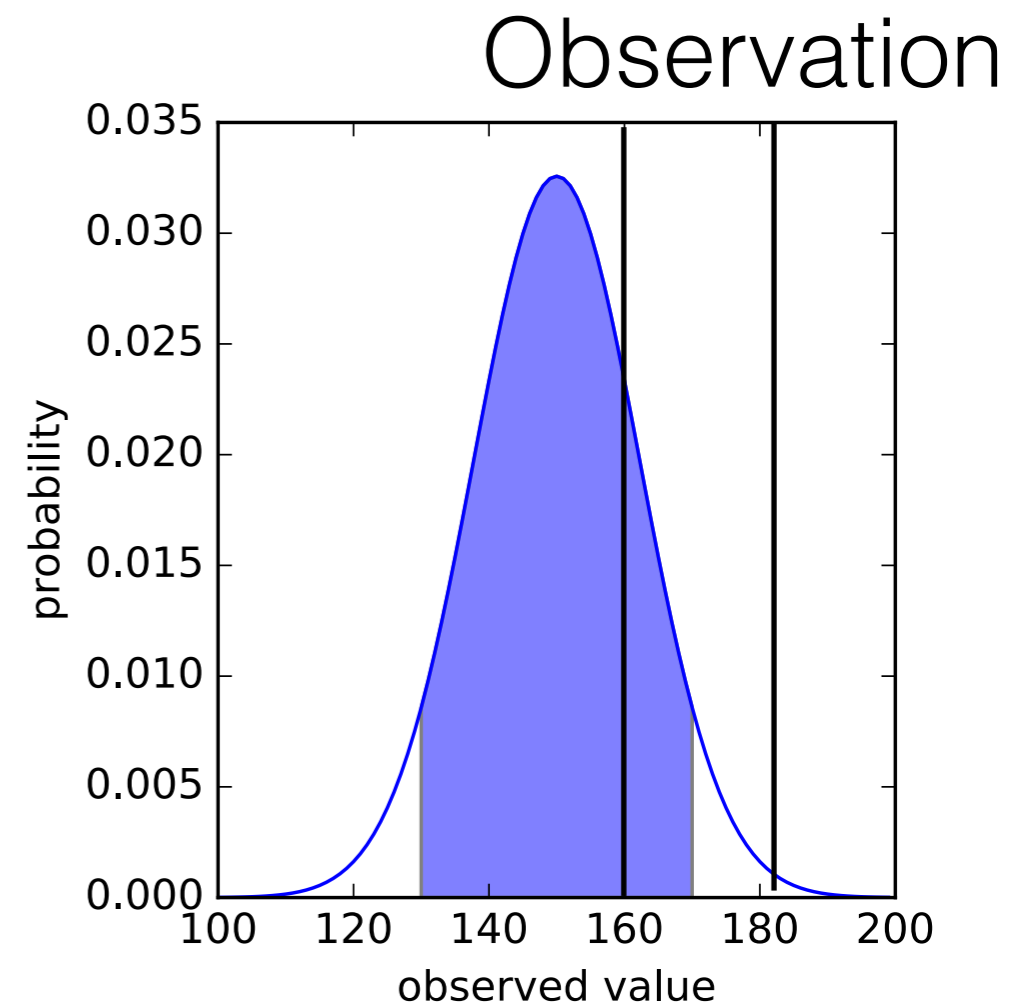
- However, we are often in the situation where we want to infer an estimate of a parameter from data

Data  $\rightarrow$  Hypothesis (Statistics)

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

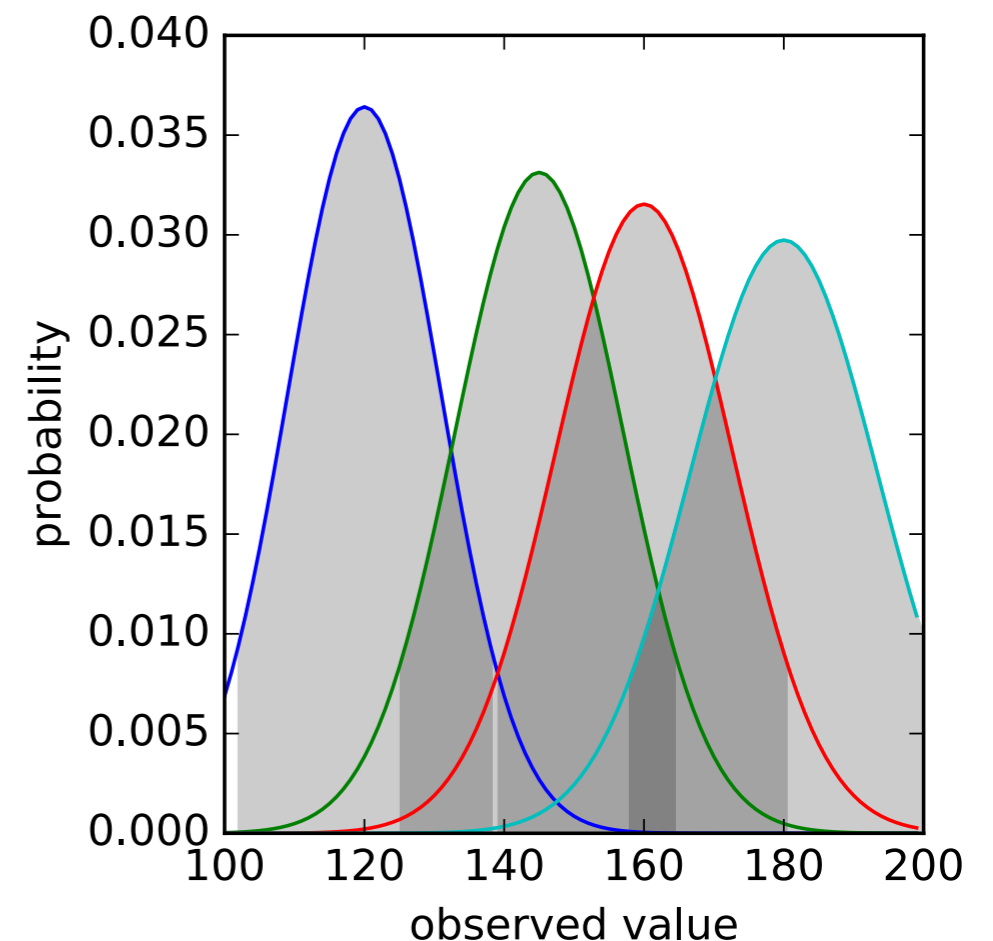
# Hypothesis rejection

- 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, in the situation where we do not know anything about a parameter it could have a different value than what we assumed in the model
- Hence we should look at other hypotheses.



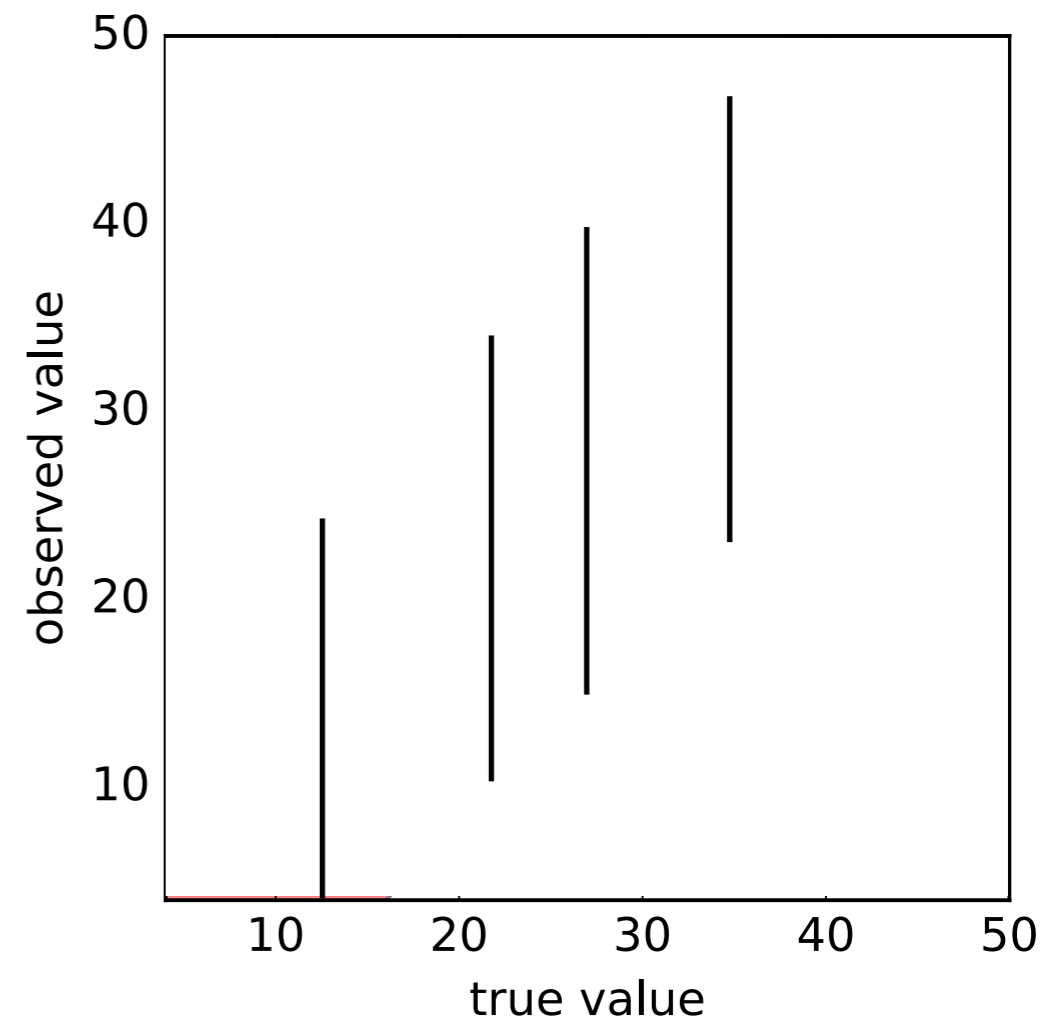
# Hypothesis rejection

- Each hypothesis will have an interval within which an observation will accept the hypothesis
- For multiple possible true values of the parameter, the acceptance intervals can be determined
- Example on the right for 68% central interval for a few true values



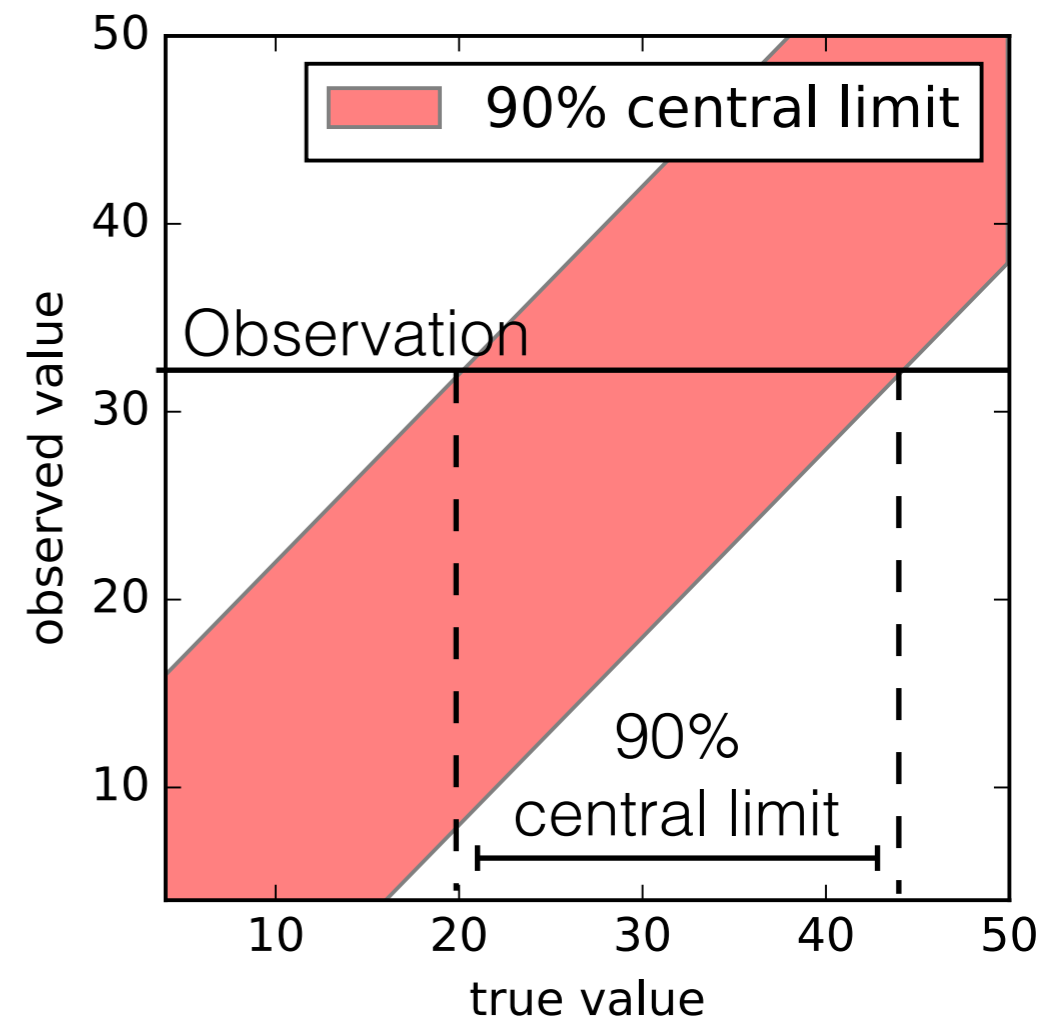
# Acceptance belt

- This produce a band 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 the probability of the confidence interval from this method to contain the true value is exactly XX%, in the example to the right 90%



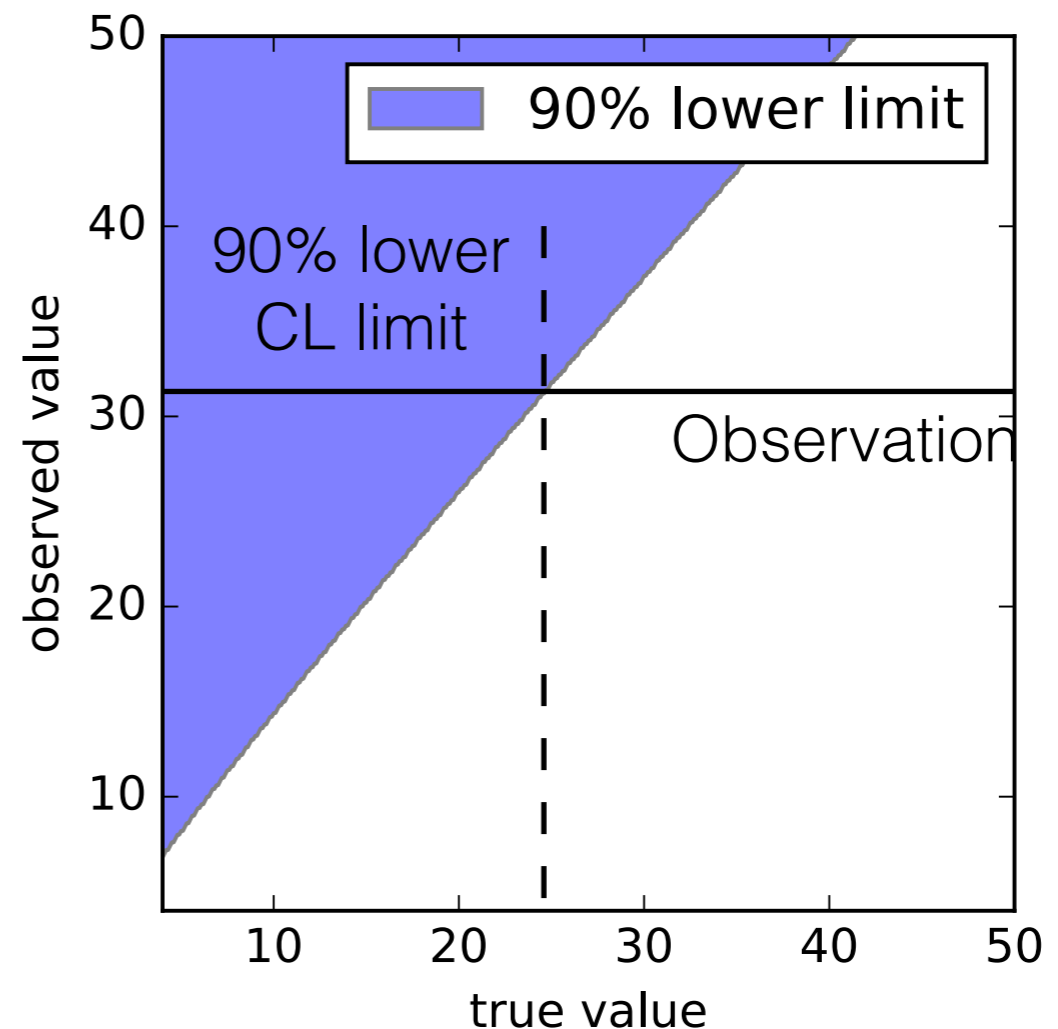
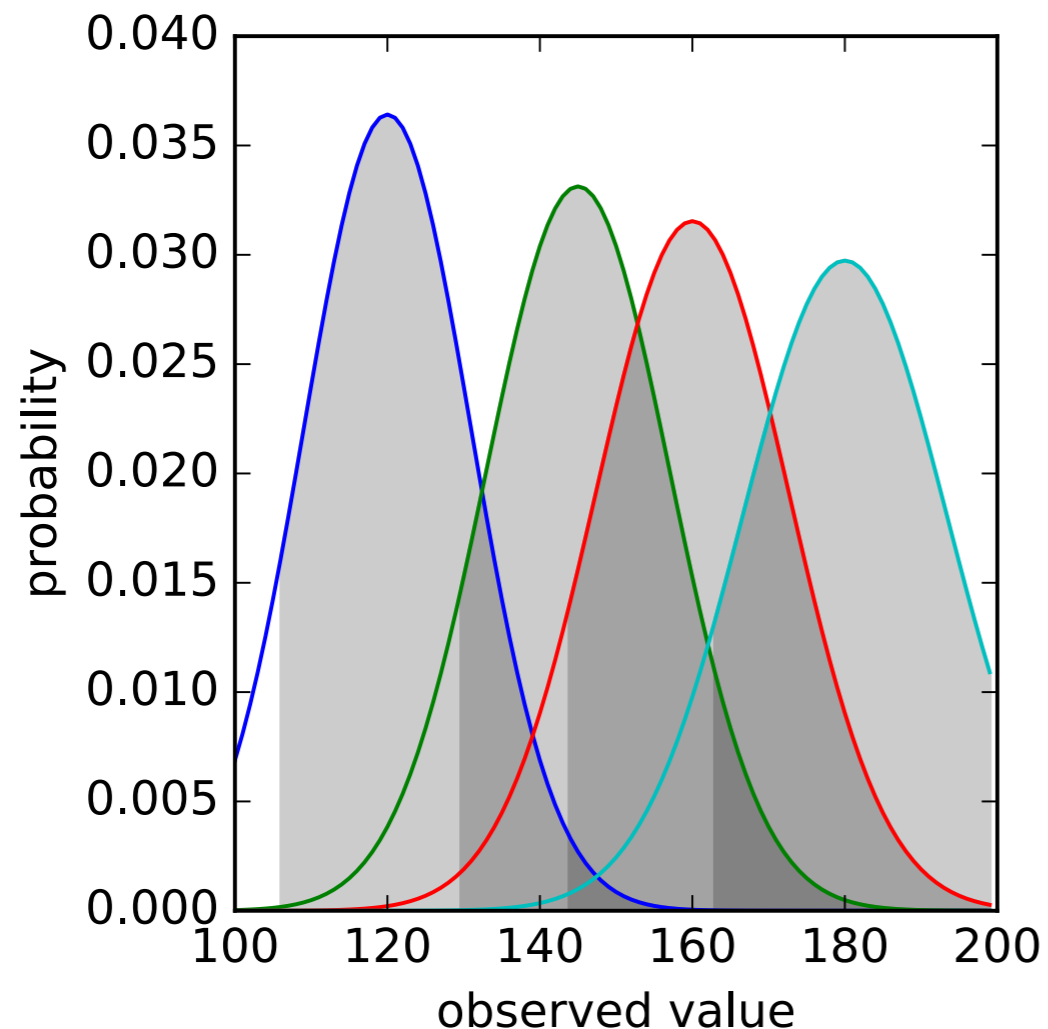
# Acceptance belt

- This produce a band 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 the probability of the confidence interval from this method to contain the true value is exactly XX%, in the example to the right 90%



# Acceptance belt

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



# Exercise 1

- Assume a Gaussian measurement of  $\theta$  with variance equal to one
- Plot the 68% central limit acceptance band for values of  $\theta$  between zero and ten, in steps of 0.1 or less
- Assume you observe  $n_{\text{obs}} = 8$  events. What is the corresponding 68% central limit?
- Extra: Repeat the exercise with a 68% upper and lower limit. Repeat at a 90% CL and 95% CL and compare the value of  $n_{\text{obs}}$  required to set a lower limit above 0



# Exercise 1

- Resulting limits on n

$n_{\text{obs}}=8$	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  $\theta$  can be non-integer)

- To make a 68% upper limit:

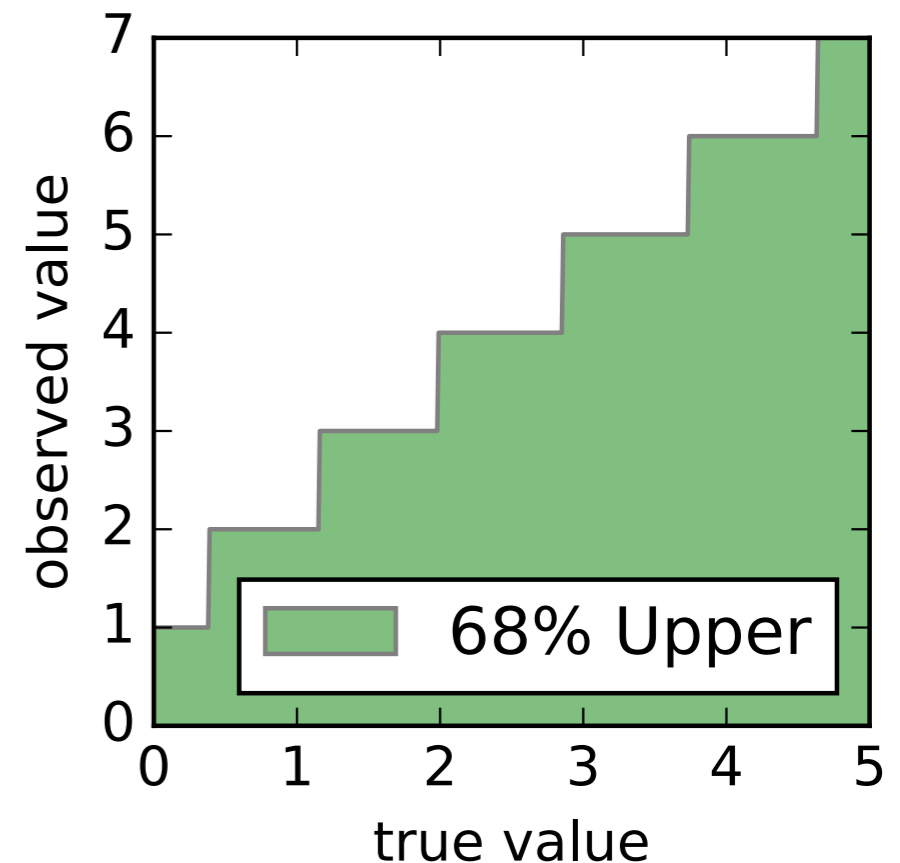
$$P(n|\theta) = e^{-\theta} \frac{\theta^n}{n!}$$

- 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 getting something that is not 2,3,4,5,6 is now less than the 32% originally intended

n	P(n   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  $\theta = 4.3$  and  $\theta = 4.5$ , the same 5 values of  $n$  would have to be included in the 68% central limit
- This will be the case over a range of values of  $\theta$ , so the confidence belt will change in steps
- Multiple true values will cover the same range of observed values

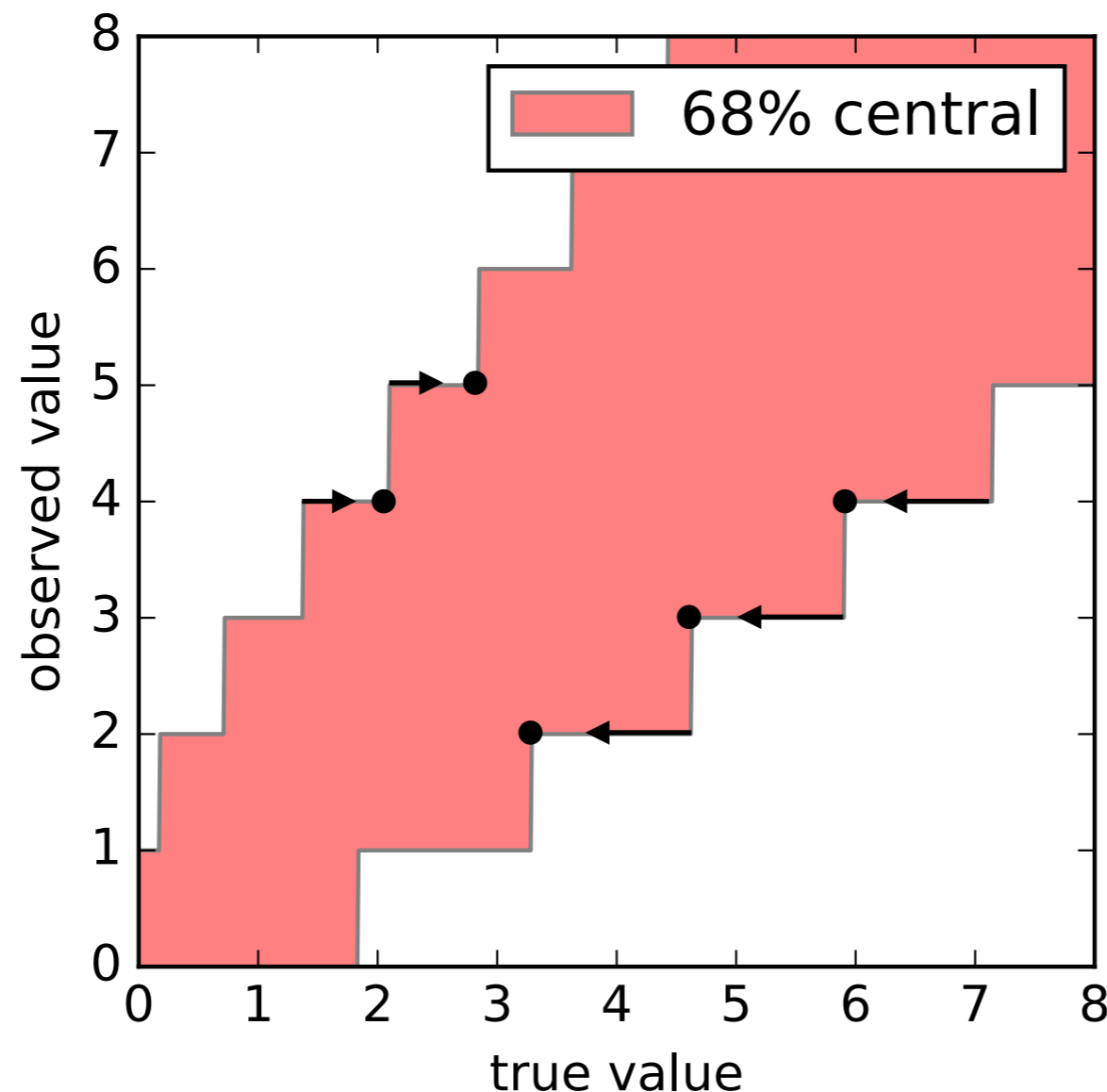


# Coverage

- A frequentist test may have a coverage greater than the confidence level = over-coverage
- Though it should never undercover (by construction)

# Complication: Discrete observations

- Use smallest true value of  $\theta$  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?
- Extra: Determine the coverage across numerous values of  $\theta$

# Exercise 2

- Results

<b>n<sub>obs</sub>=8</b>	<b>lower</b>	<b>upper</b>	<b>central</b>
<b>68 %</b>			
<b>90 %</b>	4.7	13.0	4.0-14.4
<b>95 %</b>			



# Upper limits

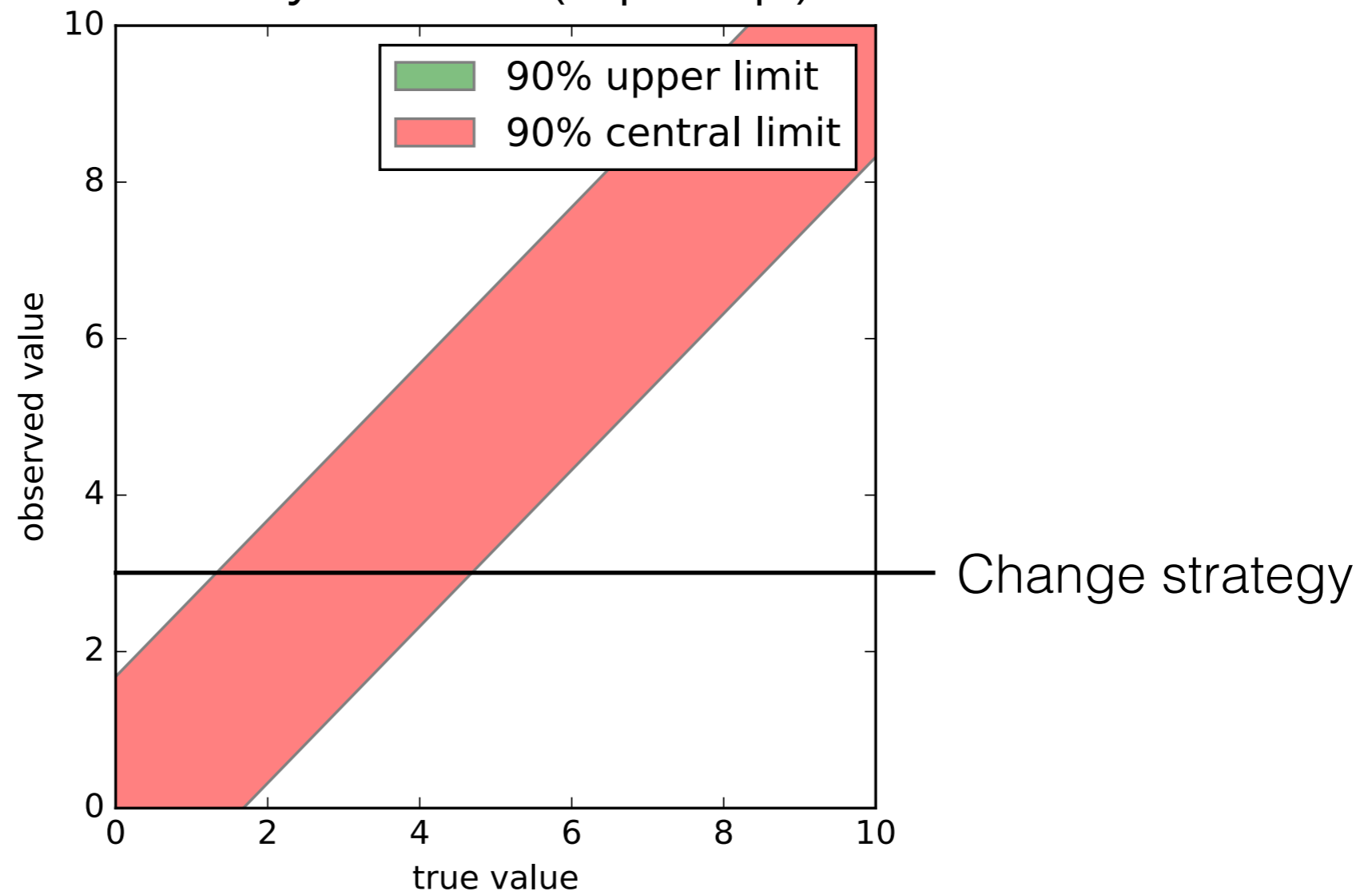
- Consider the case of observing  $n_{\text{obs}}$  events
- We assume poisson uncertainty on the number of observed events given a true number of events  $n$
- The number of events are expected to be small, so after our observation we will be reporting an 90% upper limit on  $n$ .
- Example, if zero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

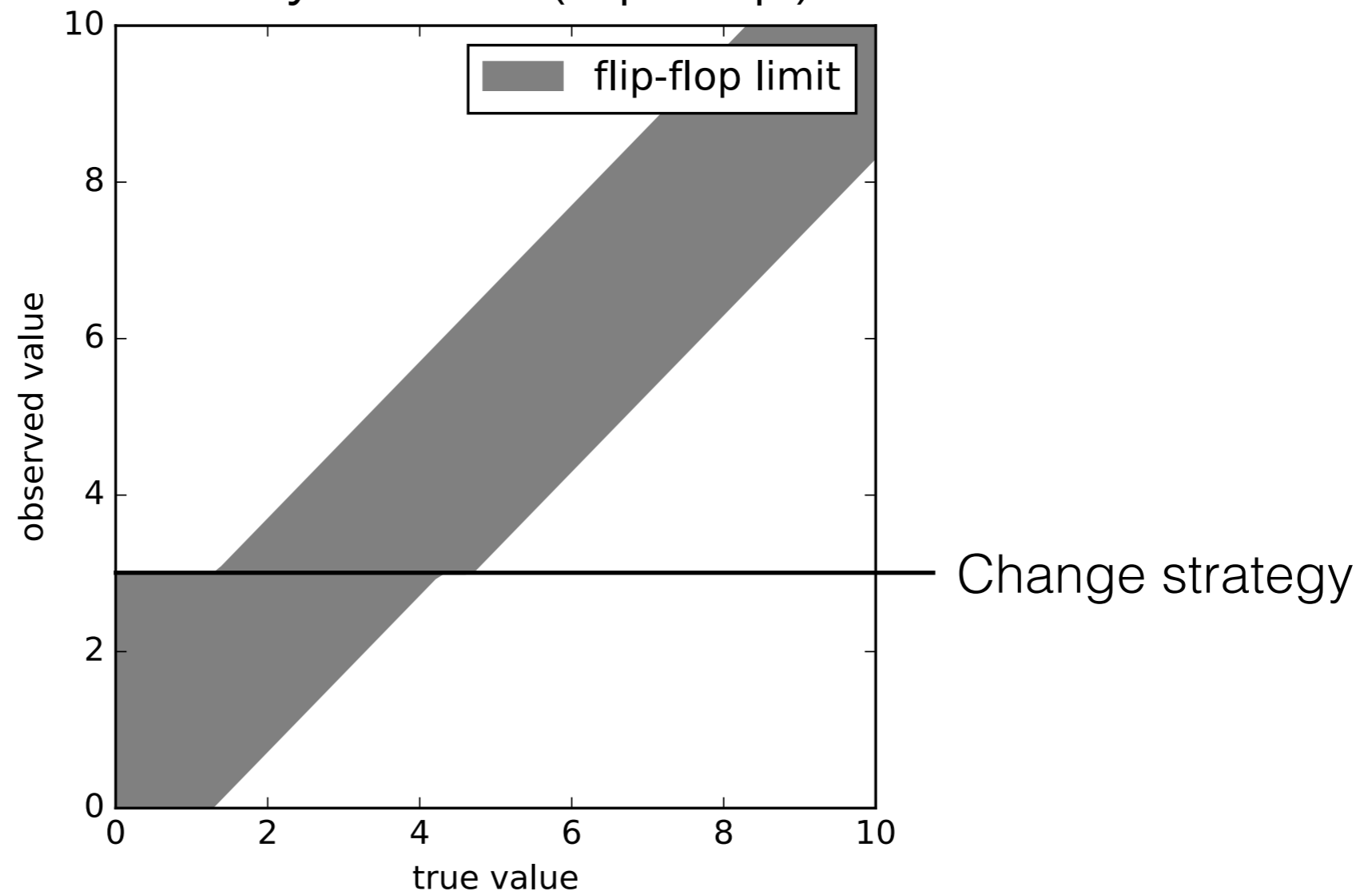
# Complication: Choosing strategy later

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



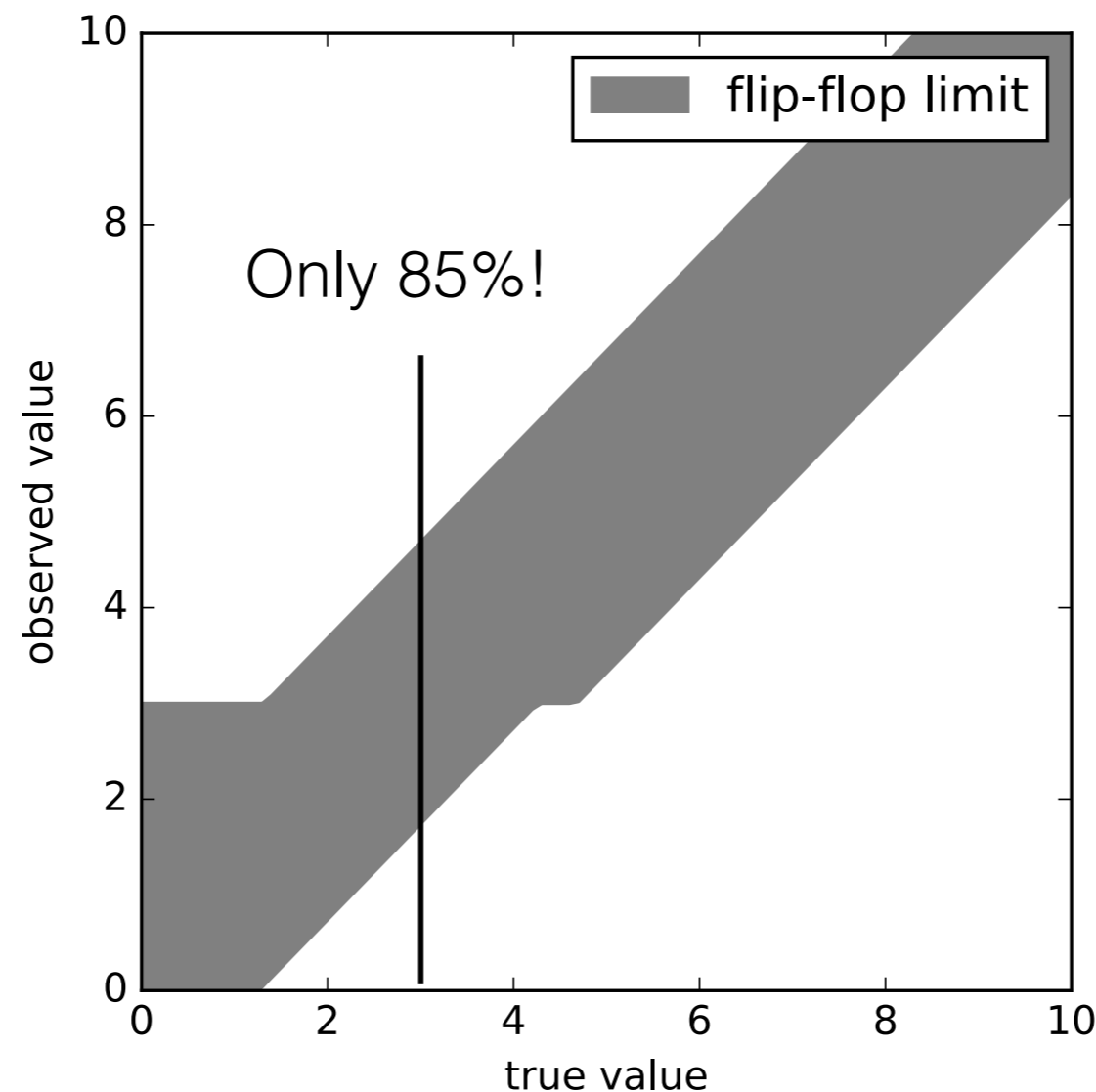
# Complication: Choosing strategy later

- Assume gaussian PDF with  $\sigma = 1$ , with the strategy of changing from 90% upper limits to 90% central limit if the observation is  $3\sigma$  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  $n_{\text{obs}} = 4$  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 positive



# 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:

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

- With the likelihood value of observing  $n$  given a true value  $\theta$ , or the best fit value of the parameter  $\theta_{\text{best}}$  given the dataset and any constraints on  $\theta$
- Completely rethink the construction of acceptance intervals for the acceptance belt: For a given true value  $\theta$ , include values of  $n$  to the interval from highest rank  $R(\theta)$  to lower, until the desired probability is reached

# Approach

- Determine the PDF for your hypothesis, which will provide the likelihood used
- For each true value  $\theta$ :
  1. Determine for all possible outcomes  $n$ :
    - A. The value  $\theta_{\text{best}}$  that maximises the likelihood  $L$
    - B. Calculate the rank  $R(\theta)$
  2. Construct the acceptance interval by including the values of  $n$ , that has the highest rank 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  $\theta_{\text{best}} = n$
- Let us produce the acceptance interval for one true value  $\theta = 1$
- Then we should do this for multiple values of  $\theta$

# Approach - Example

- Assume a Poisson measurement with true value  $\theta = 1$
- 'rank' indicates in which order the values of  $n$  are included for a 90% interval

<b>n</b>	<b><math>P(n \theta=1)</math></b>	<b><math>\theta_{\text{best}}</math></b>	<b><math>P(n \theta_{\text{best}})</math></b>	<b>R</b>	<b>rank</b>
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  $n = 0$ , it corresponds to  
 $n_s = n - b = -3$

# 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  $s_{\text{best}}$ :

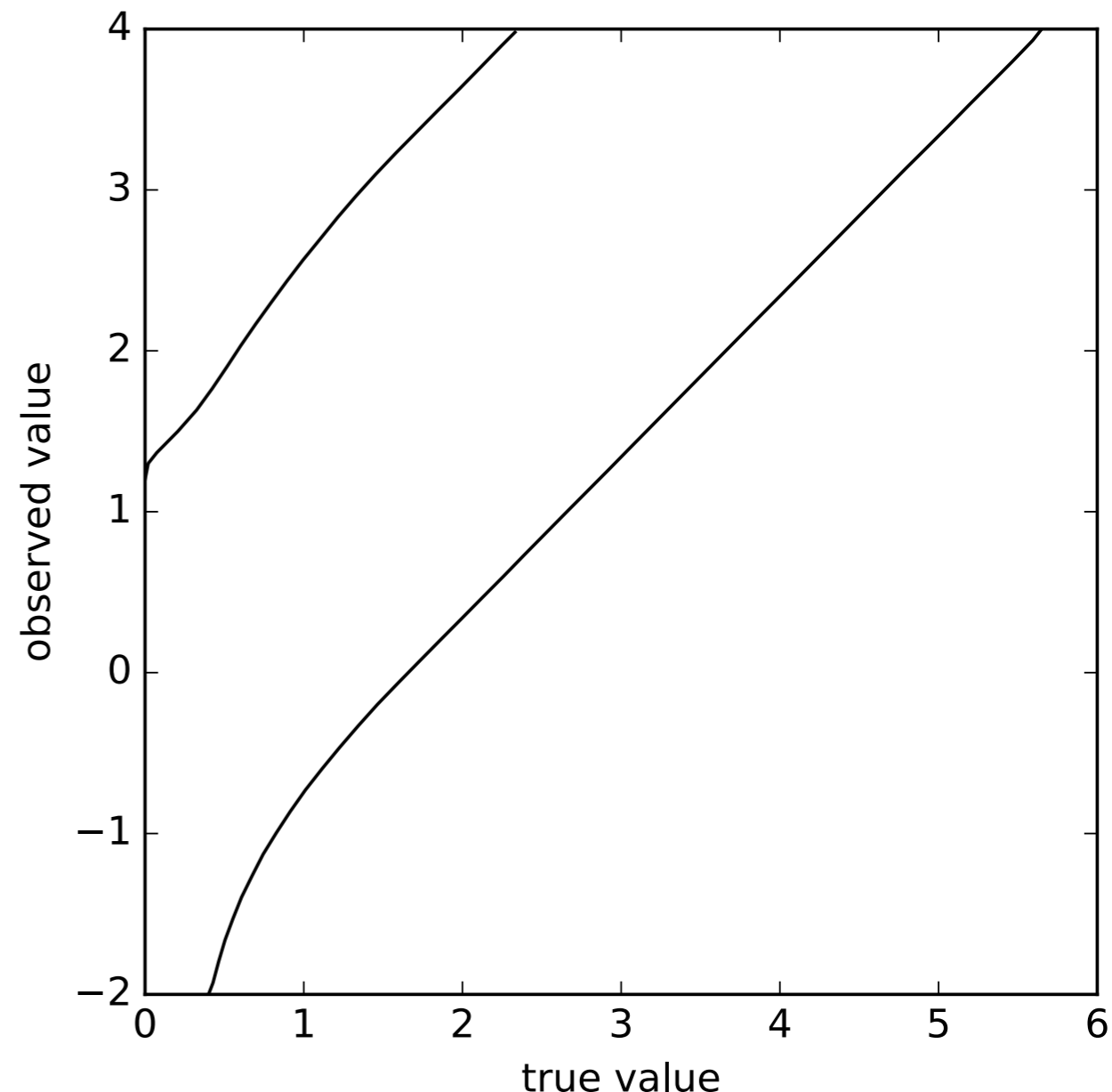
$$s_{\text{best}} = \begin{cases} n - b & \text{if } n > b \\ 0 & \text{otherwise} \end{cases}$$

- And use that when we calculate  $R(s)$



# Example: Constrained Gaussian

- Shown is the 90% confidence belt when applying the FC
- 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)

- Argument: 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

# Arguments against (2)

- Experiment 1:
  - $b = 0, n_{\text{obs}} = 1$
  - Feldman-Cousins limit:  $s < 2.44$  (at 90% CL)
- Experiment 2:
  - $b = 10, n_{\text{obs}} = 1$
  - 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

- Determine Feldman-Cousins 90 % acceptance belt for  $n_s$  in a Poisson measurement of  $n = n_s + n_b$ , with a known background of  $n_b = 4$   
Include the constraint:  $n_{\text{best}} = 0$  for  $n_{\text{obs}} < 0$   
Use steps in  $n_s$  of 0.1 or less
- Suppose you observe  $n = 10$  events what is the 90% confidence interval on  $n_s$ , what if you observe  $n = 1$ ?
- Extra: Determine the coverage across the considered values of  $n$
- Extra Extra: Do the calculations for 68% and 95% as well and compare limits for various numbers of  $n_{\text{obs}}$   
Compare to the central limit using the Neyman method

# Exercise 3

