Inference

Barbara Brown
National Center for Atmospheric Research
Boulder Colorado USA

bgb@ucar.edu

with contributions from Ian Jolliffe,
Tara Jensen, Tressa Fowler, & Eric Gilleland

May 2017
Berlin, Germany

Introduction

- Statistical inference is needed in many circumstances, not least in forecast verification

  Examples:
  - Agricultural experiments
  - Medical experiments
  - Estimating risks

  Question: *What do these examples have in common with forecast verification?*

- Goals
  - Discuss some of the basic ideas of modern statistical inference
  - Consider how to apply these ideas in verification
- *Emphasis*: interval estimation

Inference – the framework

- We have data that are considered to be a sample from some larger population
- We wish to use the data to make inferences about some population quantities (parameters)

Examples: population mean, variance, correlation, POD, MSE, etc.

Why is inference necessary?

- Forecasts and forecast verification are associated with many kinds of uncertainty
- Statistical inference approaches provide ways to handle some of that uncertainty

*There are some things that you know to be true, and others that you know to be false; yet, despite this extensive knowledge that you have, there remain many things whose truth or falsity is not known to you. We say that you are **uncertain** about them. You are **uncertain**, to varying degrees, about everything in the future; much of the past is hidden from you; and there is a lot of the present about which you do not have full information. **Uncertainty** is everywhere and you cannot escape from it.*

Dennis Lindley, *Understanding Uncertainty* (2006). Wiley-Interscience.

Accounting for uncertainty

- Observational
- Model
  - Model parameters
  - Physics
  - Verification scores

- Sampling
  - Verification statistic is a *realization* of a random process
  - What if the experiment were re-run under identical conditions? Would you get the same answer?

Our population

Age											
20-24											
25-29	F	F	F	F	F	M	M	M	M		
30-34	F	F	F	F	F	F	F	M	M	M	M
35-39	F	F	F	F	F	M	M				
40-44	F	F	F	F	F	M	M				
45-49	F	M	M								
50-54	M	M	M								
55-59											
60-64	F	F	M								
65-69	M										
Count:	1	2	3	4	5	6	7	8	9	10	11

The tutorial age distribution

% male: 44%

Mean age

Overall: 38

For males: 40

For females: 37

What would we expect the results to be if we take samples from this population?

Would our estimates be the same as what's shown at the left?

How much would the samples differ from each other?

Sampling results

	% Male	% Female	Mean Age			Median Age			
			Male	Female	All	Male	Female	All	
Real	44%	56%	40	37	38	39	35	37	N=45
Sample 1	33%	67%	41	43	42	34	42	40	N=12

Random Sampling:
5 samples of 12 people each

Sample 1 results:

- % males too low
- Mean age for males slightly too large
- Mean age for females much too large
- Overall mean is too large
- Medians for females and "All" are too small

Sampling results cont.

	% Male	% Female	Mean Age			Median Age		
			Male	Female	All	Male	Female	All
Real	44%	56%	40	37	38	39	35	37
Sample 1	33%	67%	41	43	42	34	42	40
Sample 2	50%	50%	33	35	34	32	35	32
Sample 3	50%	50%	43	33	38	41	31	36
Sample 4	58%	42%	37	37	37	39	37	38
Sample 5	50%	50%	39	40	40	41	31	36

<u>Summary</u>

- Very different results among samples
- % male almost always over-estimated in this small number of random samples

Types of inference

- Point estimation – simply provide a single number to estimate the parameter, with no indication of the uncertainty associated with it (suggests no uncertainty)
- Interval estimation
  - **One approach**: attach a standard error to a point estimate
  - **Better approach**: construct a ***confidence interval***
- Hypothesis testing
  - May be a good way to address whether any difference in results between two forecasting systems could have arisen by chance.
- **Note**: Confidence intervals and Hypothesis tests are closely related
  - Confidence intervals can be used to show whether there are significant differences between two forecasting systems
  - Confidence intervals provide more information than hypothesis tests (e.g., uncertainty bounds, asymmetries)

Approaches to inference

1. Classical (frequentist) parametric inference
2. Bayesian inference
3. Non-parametric inference
4. Decision theory
5. …

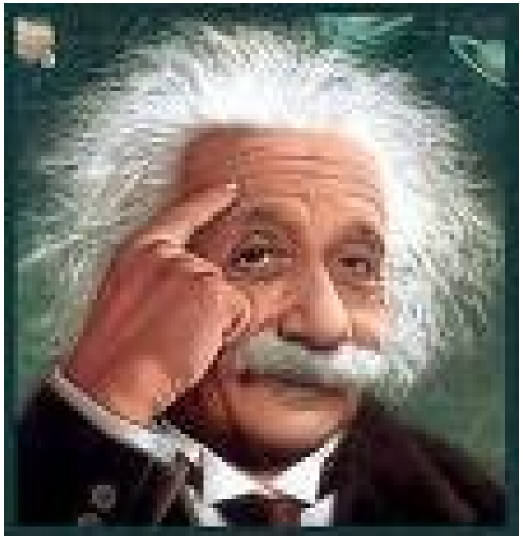
Approaches to inference

1. Classical (frequentist) parametric inference
2. Bayesian inference
3. Non-parametric inference
4. Decision theory
5. …

Focus will be on ***classical*** and ***non-parametric*** confidence intervals (CIs)

Confidence Intervals (CIs)

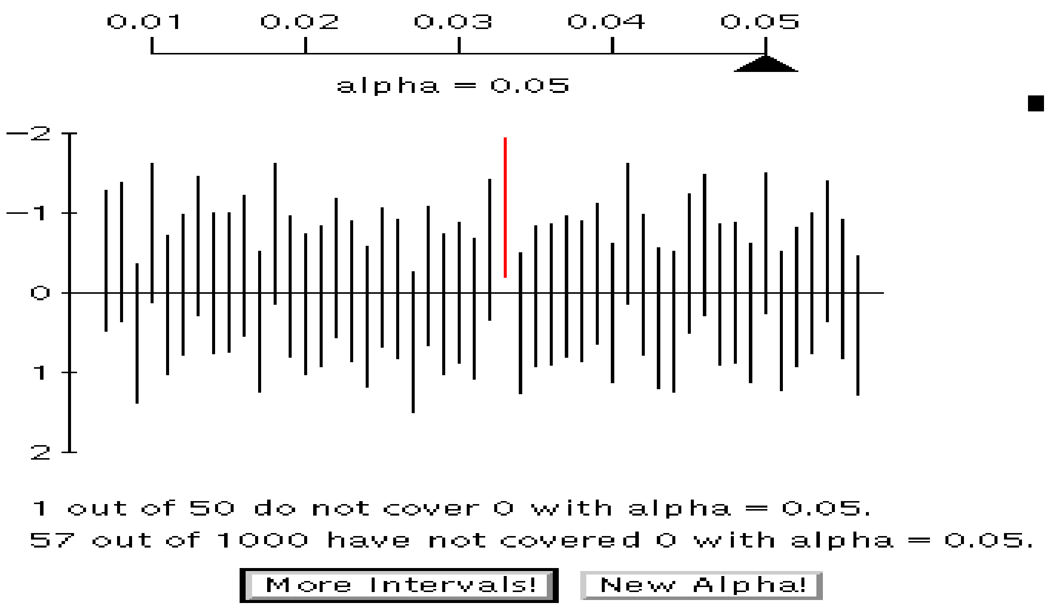
"If we re-run an experiment N times (i.e., create N random samples), and compute a (1-α)100% CI for each one, then *we expect the **true population value** of the parameter to fall inside (1-α)100% of the intervals."*



Confidence intervals can be *parametric* or *non-parametric*…

What is a confidence interval?

Given a sample value of a measure (statistic), find an interval with a specified level of confidence (e.g., 95%, 99%) of including the corresponding population value of the measure (parameter).



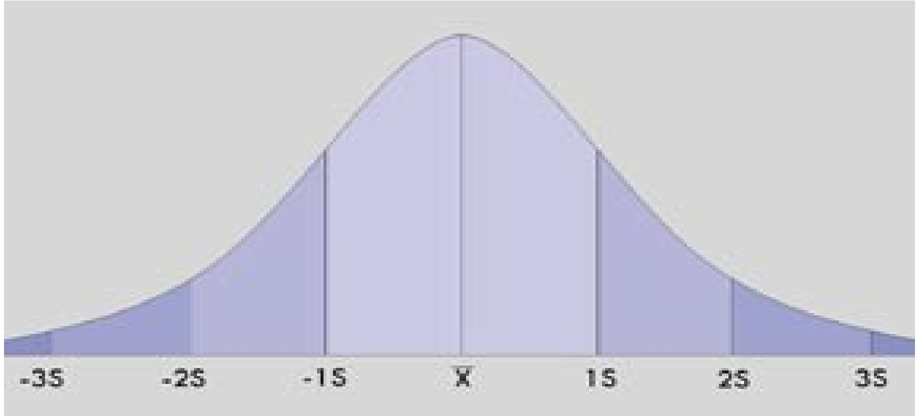
Note:

- The *interval* is random; the *population value* is fixed
- The confidence *level* is the long-run probability that intervals include the parameter, NOT the probability that the parameter is in the interval

http://wise.cgu.edu/portfolio/demo-confidence-interval-creation/

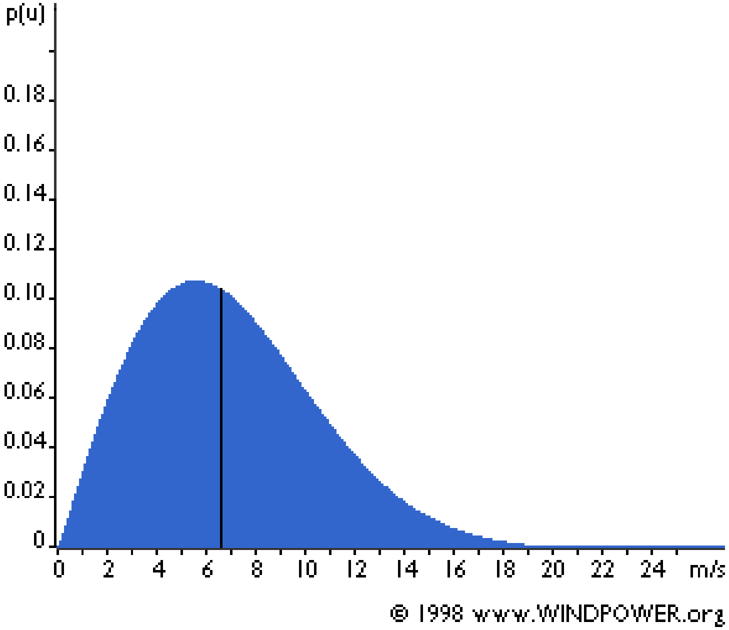
Confidence Intervals (CI's)

- Parametric
  - Assume the observed sample is a realization from a known *population* distribution with possibly unknown parameters (e.g., normal)
  - Normal approximation CI's are most common.
  - Quick and easy



Confidence Intervals (CI’s)

- Nonparametric
  - Assume the distribution of the observed sample is representative of the *population* distribution
  - Bootstrap CI’s are most common
  - Can be computationally intensive, but still easy enough



Normal Approximation CI's

Standard normal variate

Population ("true") parameter

Estimate

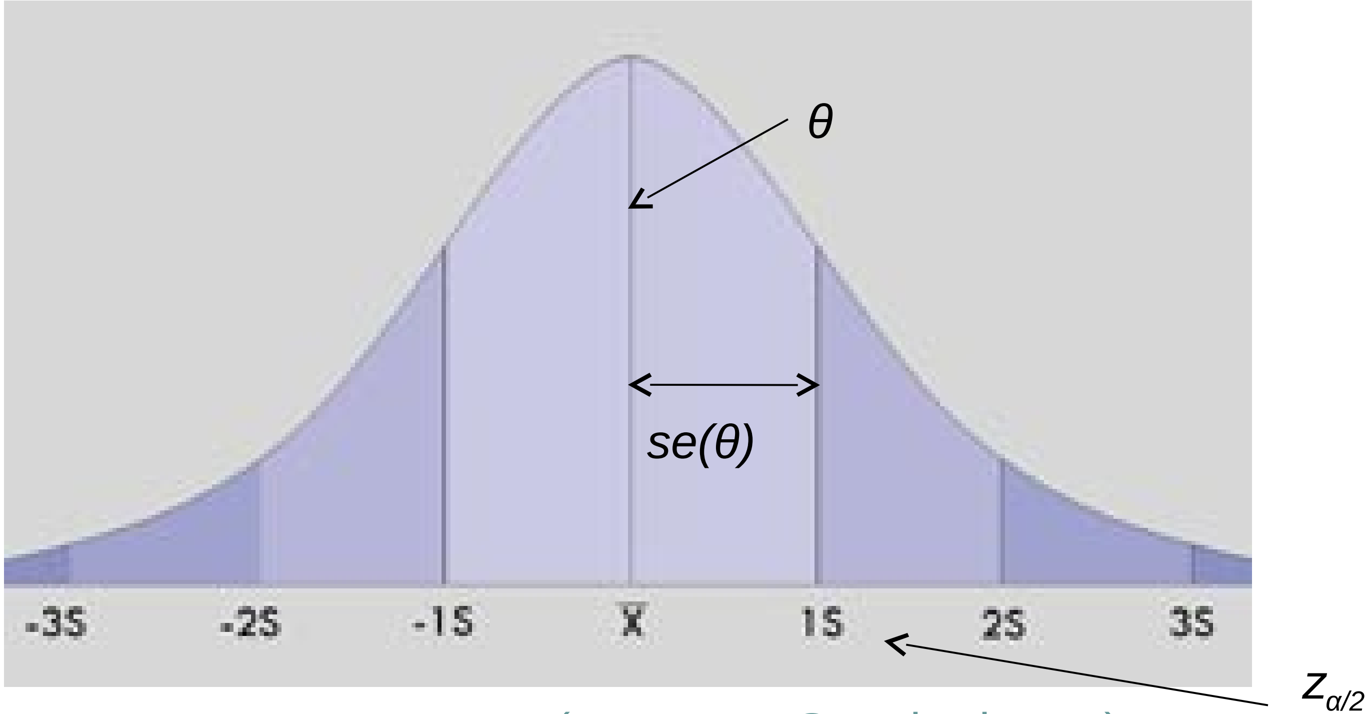
$$\hat{\theta} \pm z_{\alpha/2}\, se(\theta)$$

Is a (1-α)100% Normal CI for Θ, where

- Θ is the statistic of interest (e.g., the forecast mean)
- se(Θ) is the standard error for the statistic
- z_v is the v-th quantile of the standard normal distribution where v= α/2.
- A typical value of α is 0.05 so (1-α)100% is referred to as the 95th percentile Normal CI

Normal Approximation CI's

$$\hat{\theta} \pm z_{\alpha/2} se(\theta)$$



(note: se = Standard error)

Normal Approximation CI's

- Normal approximation is appropriate for numerous verification measures

  Examples: *Mean error, Correlation, ACC, BASER, POD, FAR, CSI*

- Alternative CI estimates are available for other types of variables

  Examples: forecast/observation *variance*, *GSS, HSS, FBIAS*

- All approaches expect the sample values to be *independent* and *identically distributed* (iid)

Application of Normal Approximation CI's

- **Independence assumption** (i.e., "iid") – temporal and spatial
  - Should check the validity of the independence assumption
  - Relatively simple methods are available to account for first-order temporal correlation
    - More difficult to account for spatial correlation (an advanced topic…)
- **Normal distribution assumption**
- Should check validity of the normal distribution (e.g., qq-plots, Kolmagorov-Smirnov test, χ^2 test)

Normal CI Example

		Observed		
orecast		Yes	No	Total
	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

POD (Hit Rate)= 0.55
FAR= 0.72

What are appropriate CI's for these two statistics?

CIs for POD and FAR

- Like several other verification measures POD and FAR represent the *proportion of times that something occurs or something doesn't occur*
  - **POD**: *The proportion of hits that were forecast*
  - **FAR**: *The proportion of forecasts that weren't associated with an event occurrence*
  - Denote these proportions by p_1 and p_2.
- CIs can be found for the underlying probability of
  - A correct forecast, given that the event occurred
  - A non-event given that the forecast was of an event
  - Call these probabilities θ_1 and θ_2.
- Statistical analogy:
  - Find a confidence interval for the 'probability of success' in a *binomial distribution*
  - Various approaches can be used

Binomial CIs

- Distributions of p_1 and p_2 can be approximated by Gaussian distributions with
  - Means θ_1 and θ_2 and
  - Variances $p_1(1-p_1)/n_1$ and $p_2(1-p_2)/n_2$

    [n's are the 'numbers of trials' (number of observed Yes for POD and number of forecasted Yes for FAR)]
- The intervals have endpoints

$$p_1 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}} \quad \text{and} \quad p_2 \pm z_{\alpha/2}\sqrt{\frac{p_2(1-p_2)}{n_2}}$$

where $z_{\alpha/2} = 1.96$ for a 95% interval

- Other approximations for binomial CIs are available which may be somewhat better than this simple one in some cases

Normal CI Example

		Observed		
		Yes	No	Total
orecast	Yes	28	72	100
	No	23	2680	2703
	Total	51	2752	2803

POD (Hit Rate)= 0.55 ≈ (0.41, 0.69)

FAR= 0.72 ≈ (0.63, 0.81)

95% normal approximation CI shown in red

Note: These CIs are symmetric

(Nonparametric) Bootstrap CI's

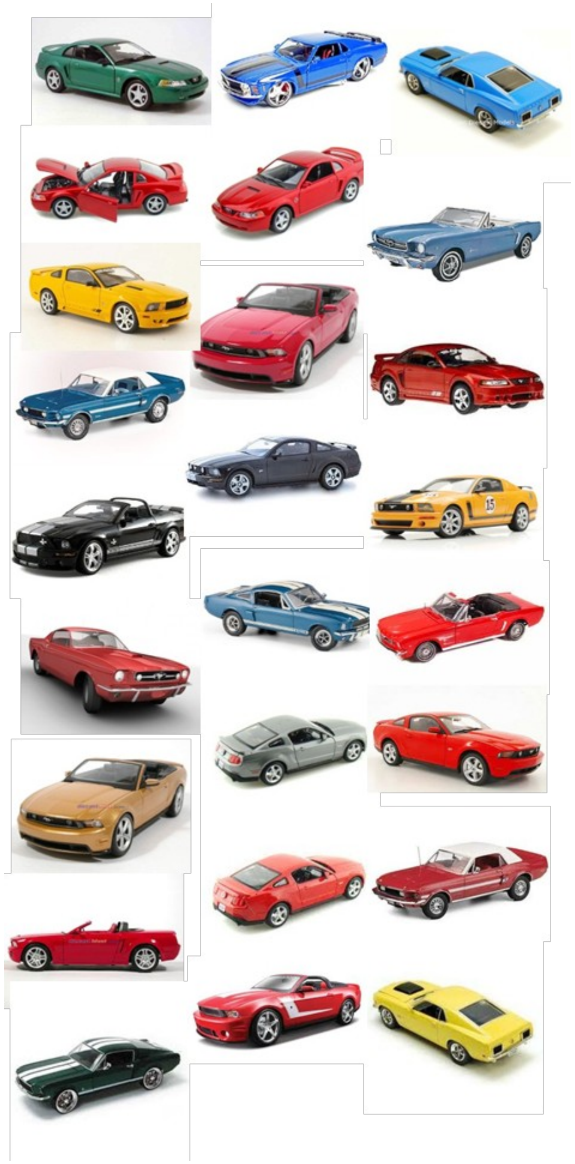
IID Bootstrap Algorithm

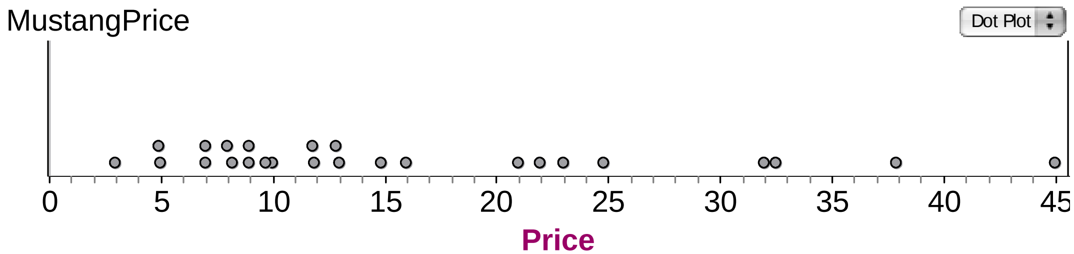
1. Resample *with replacement* from the sample,

   $x_1, x_2, \ldots, x_n$

2. Calculate the verification statistic(s) of interest from the resample in step 1.
3. Repeat steps 1 and 2 many times, say B times, to obtain a sample of the verification statistic(s) θ_B .
4. Estimate (1-α)100% CI's from the sample in step 3.

Mustang example





$$n = 25,\ \bar{x} = 15.98,\ s = 11.11$$

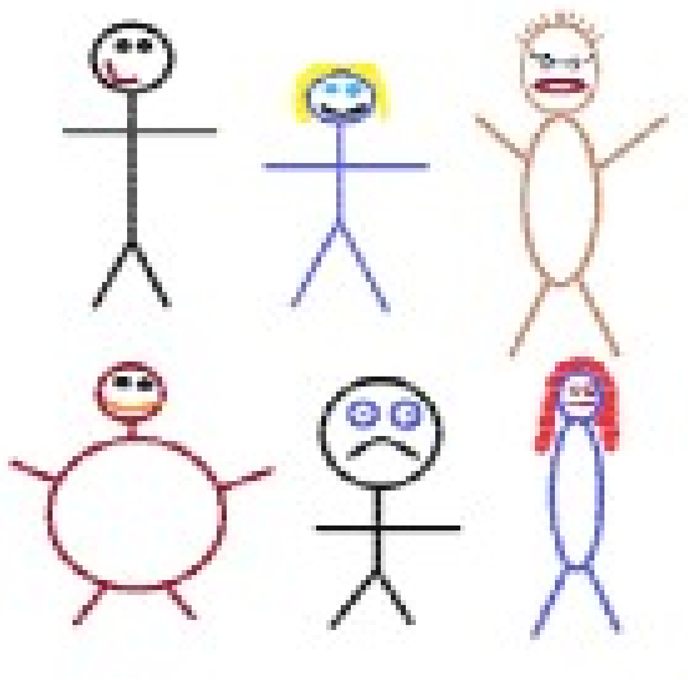
Our best estimate of the average price of used Mustangs is \$15,980

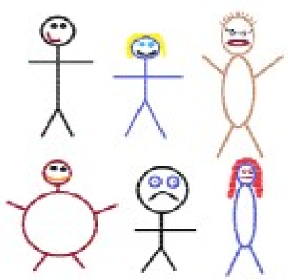
How do we estimate the confidence interval for Mustang prices?

Original Sample

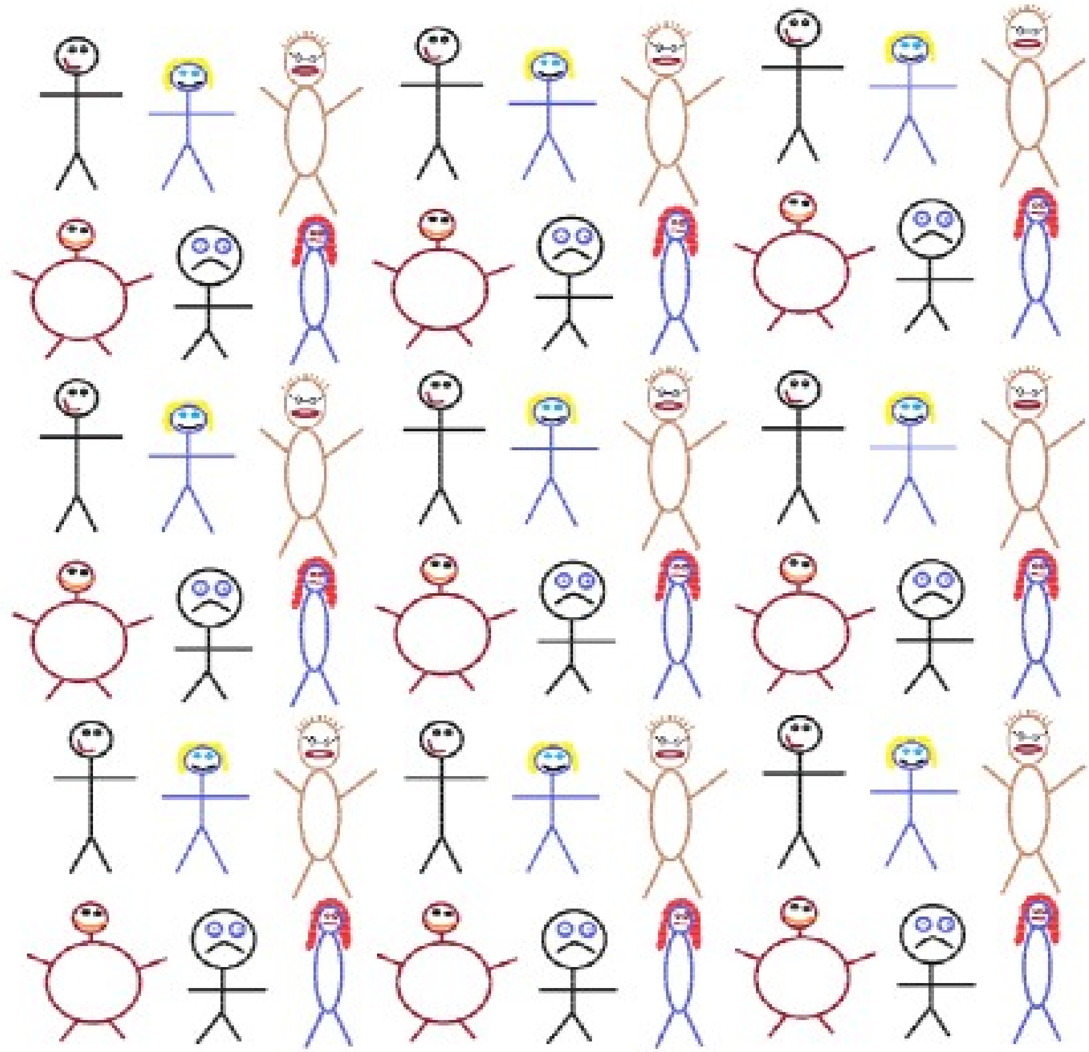
Bootstrap Sample

Suppose we have a random sample of 6 people:



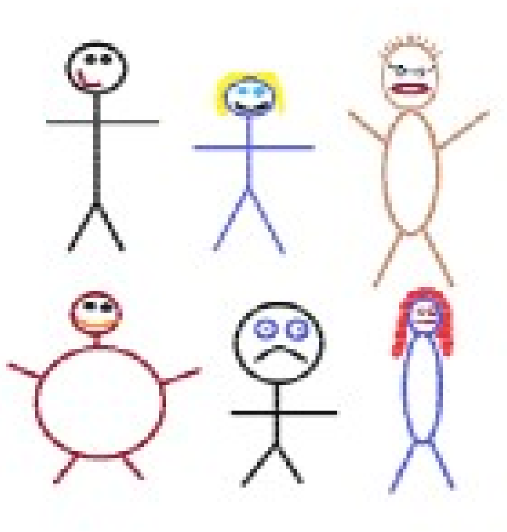


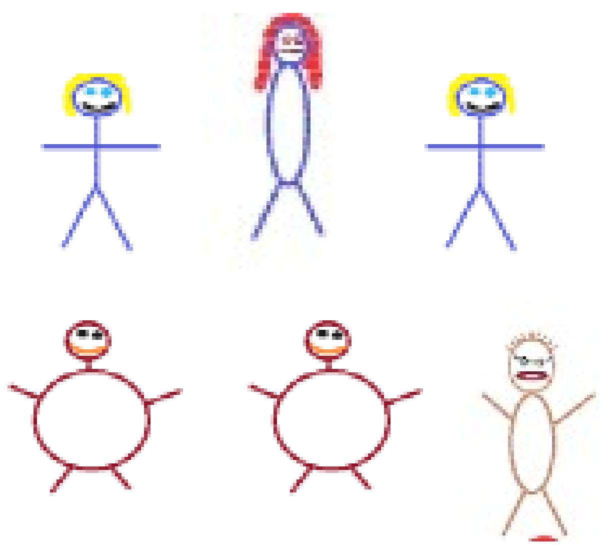
Original Sample



A simulated “population” to sample from

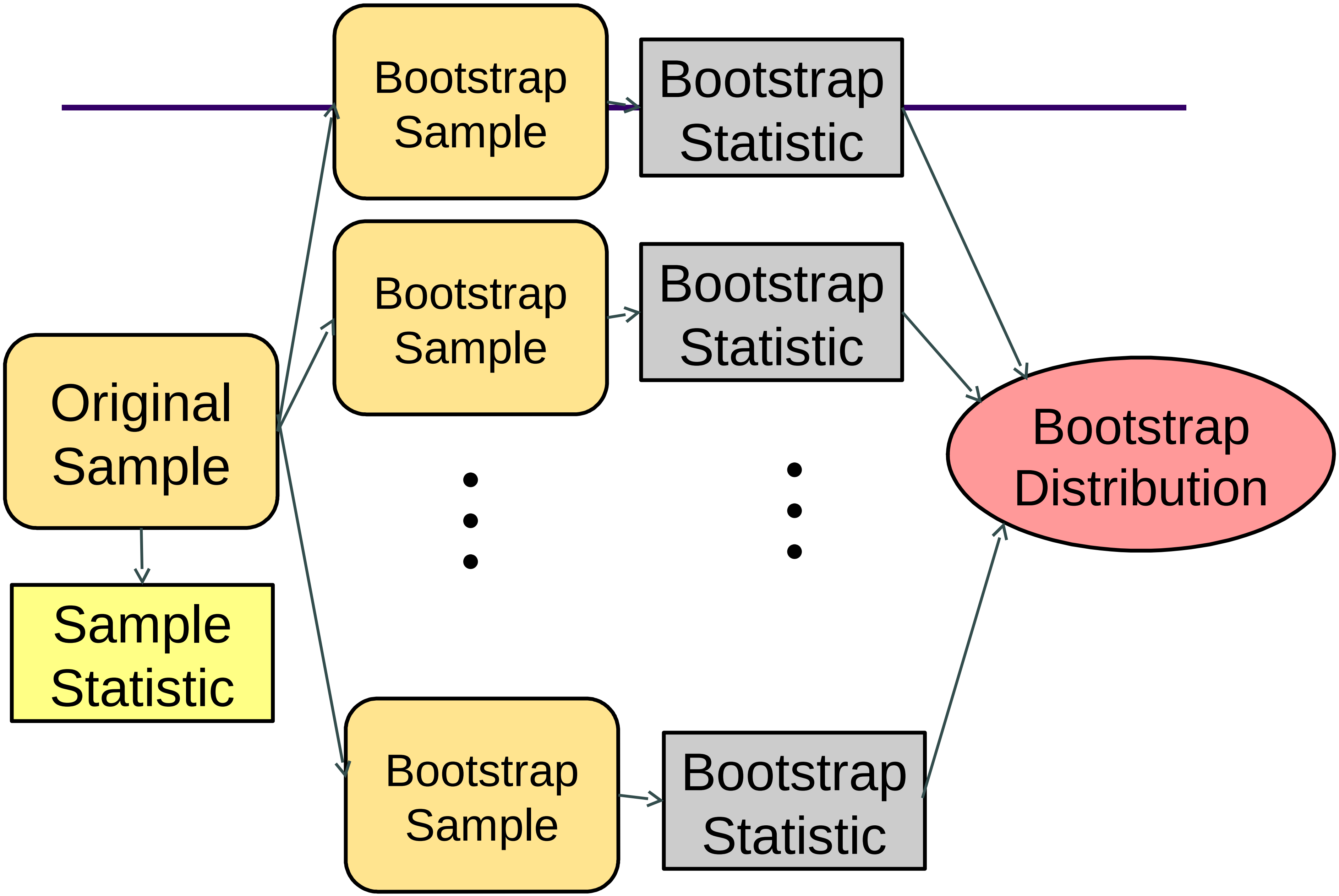
Bootstrap Sample: Sample with replacement from the original sample, using the same sample size.



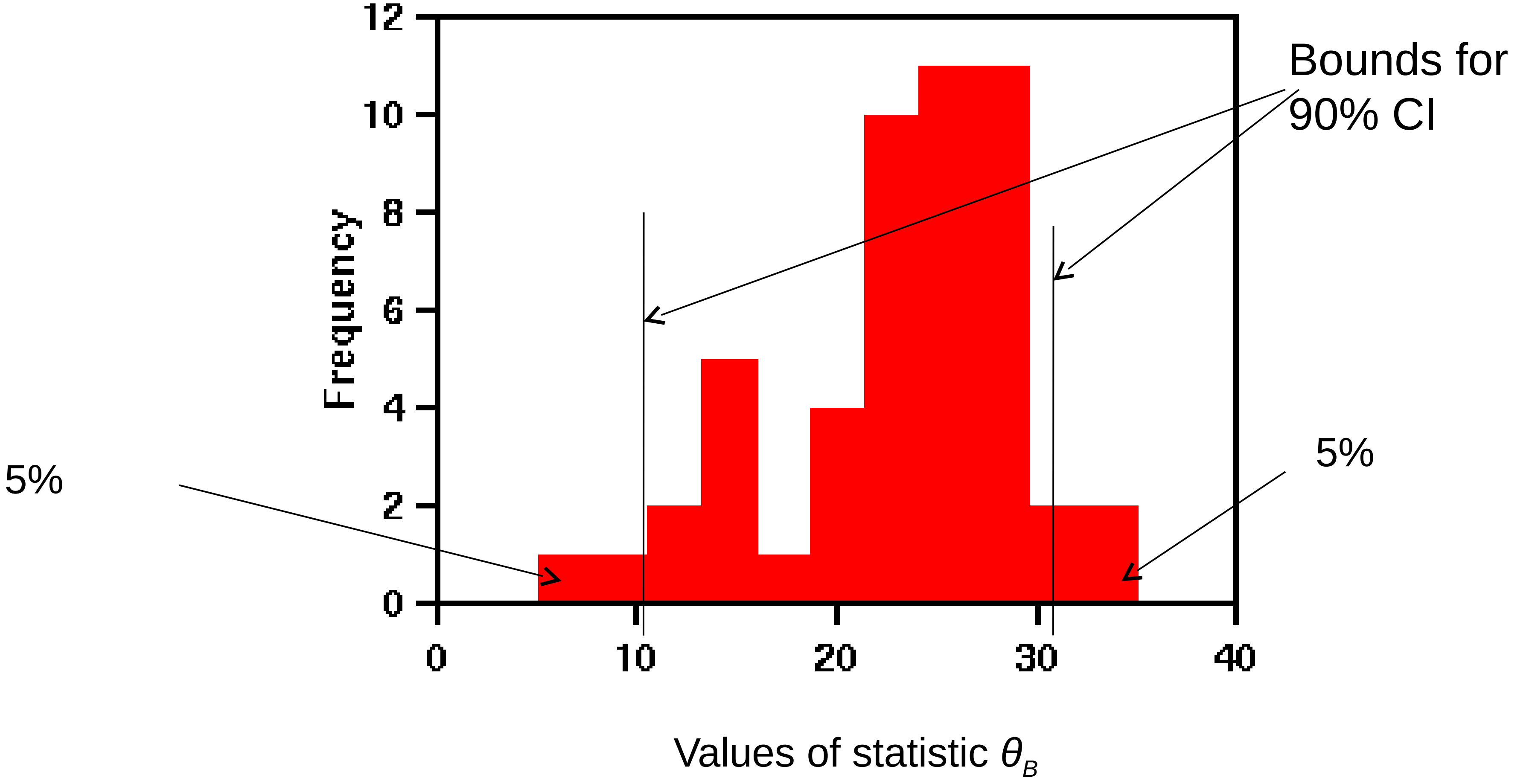


Original
Sample

Bootstrap Sample


Bootstrap Sample
Bootstrap Statistic
Bootstrap Sample
Bootstrap Statistic
Original Sample
Sample Statistic
Bootstrap Distribution
Bootstrap Sample
Bootstrap Statistic

Bootstrap Distribution: Empirical Distribution (Histogram) of statistic calculated on repeated samples



Bootstrap CI's

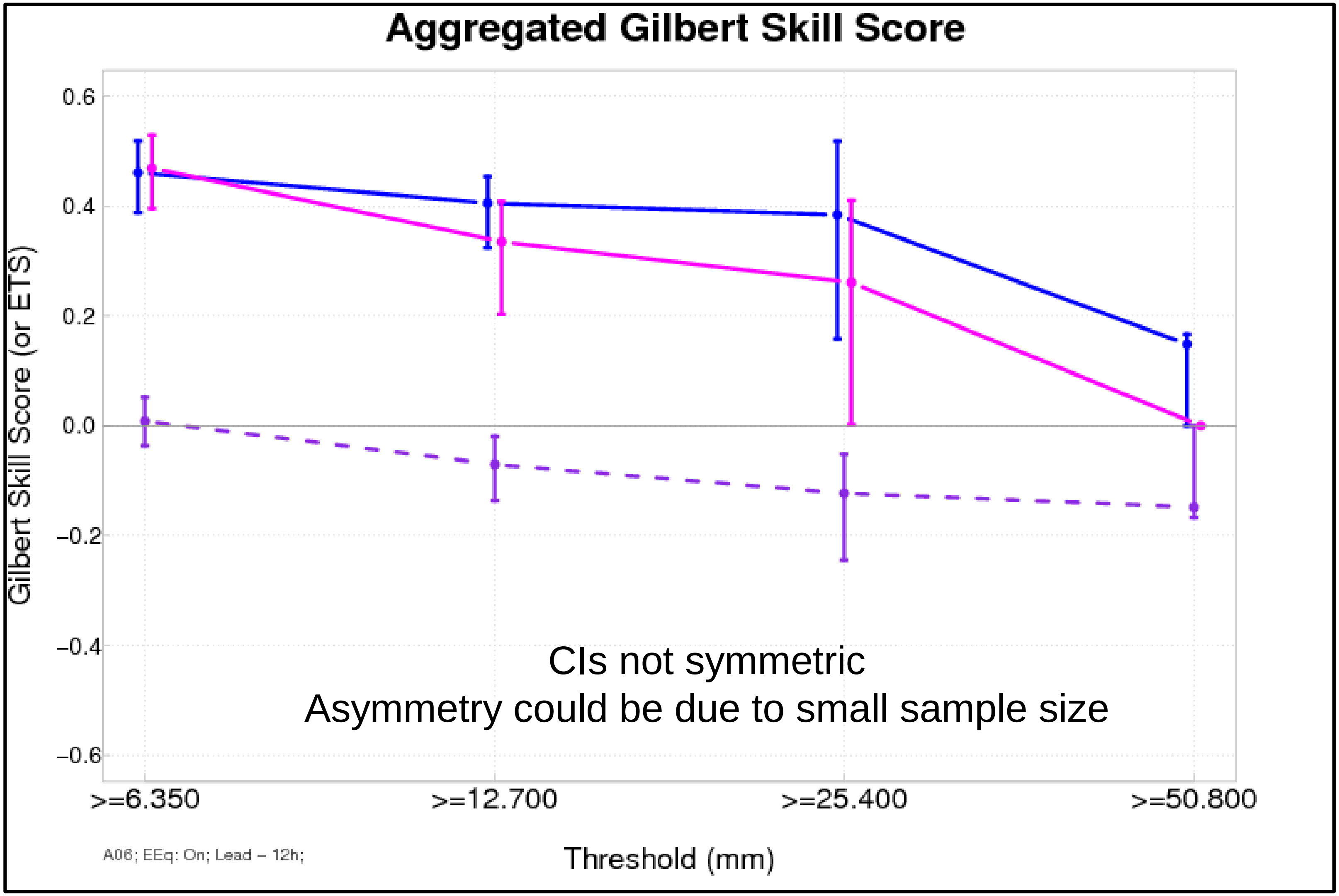
IID Bootstrap Algorithm: Types of CI's

1. Percentile Method CI's
2. Bias-corrected and adjusted (BCa)[1]
3. ABC
4. Basic bootstrap CI's
5. Normal approximation
6. Bootstrap-t

More representative but also much more Compute-intensive

[1]See Gilleland 2010 for more information about alternative methods

Bootstrap CI Example

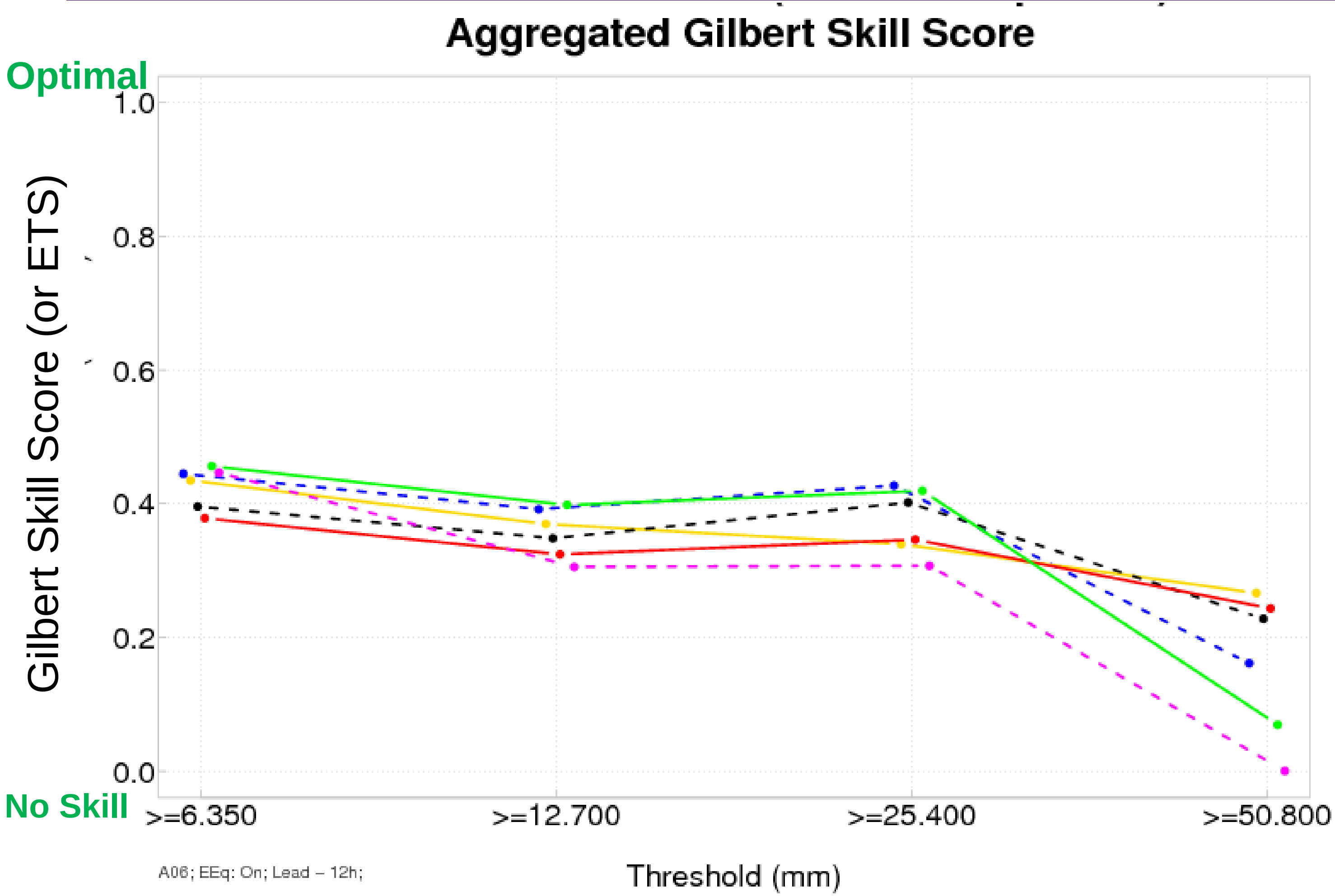


Pairwise comparisons

Pairwise comparisons are often advantageous when comparing performance for two forecasting systems

- Reduced variance associated with the comparison statistic (for normal distribution approaches)
- More “efficient” testing procedure
- More “powerful” comparisons

6 hours accumulated precipitation evaluation



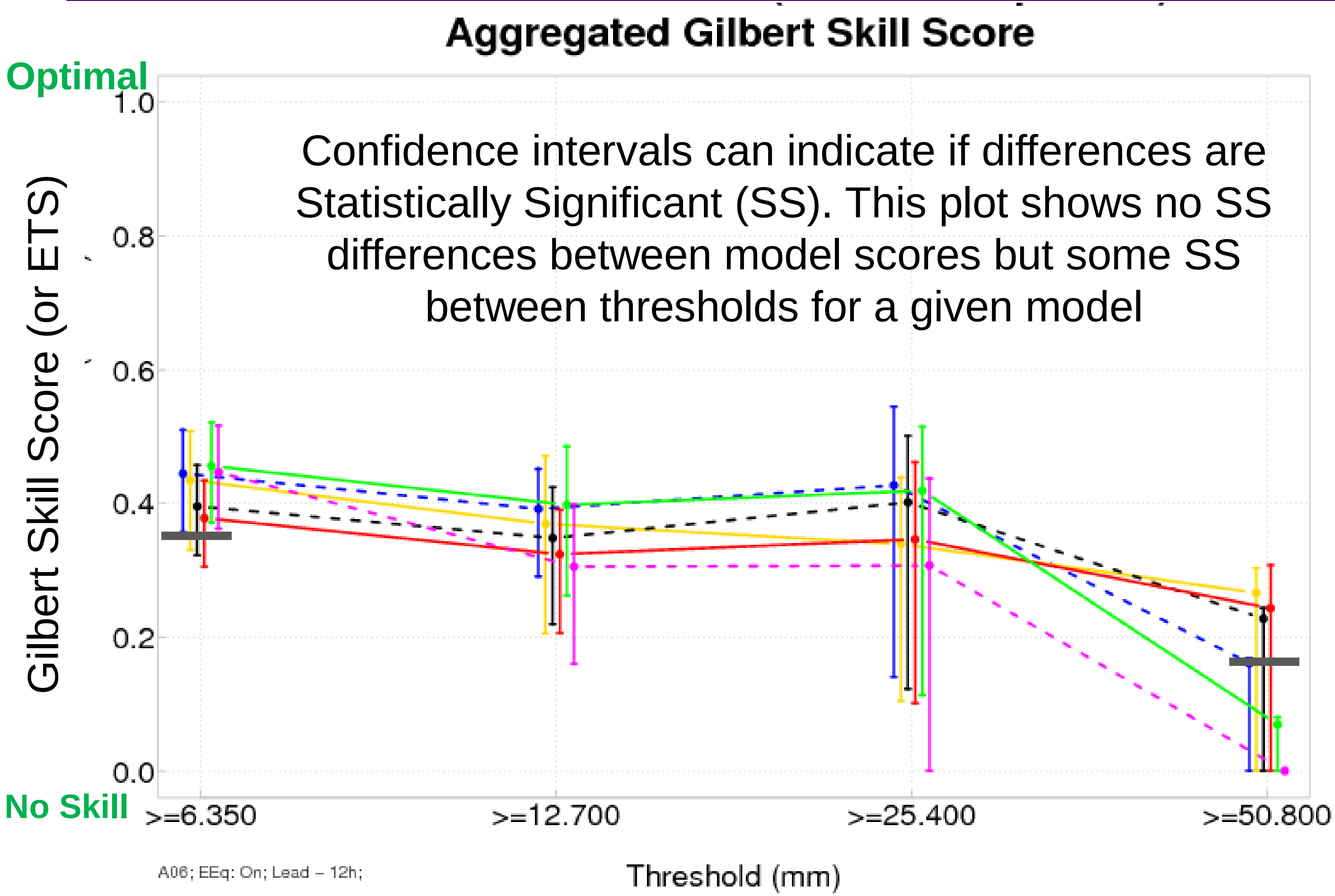
Aggregated GSS :

All of the scores are similar at low thresholds

Scores seem to be much different at larger thresholds

A06 - 12hr Lead Time

6 hours accumulated precipitation evaluation



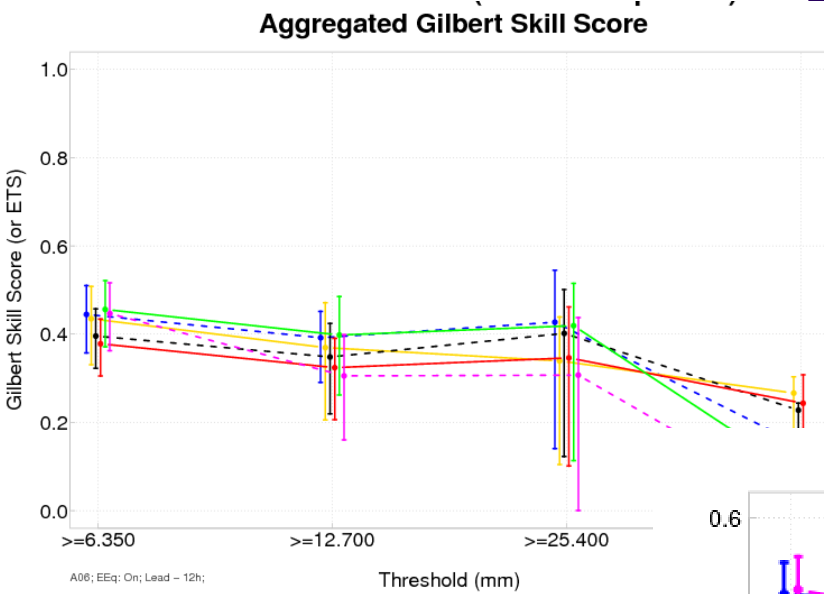
Aggregated GSS :

Overlapping confidence intervals indicate no significant difference because of large sample uncertainty

Statistical significance indicated when CIs don't overlap

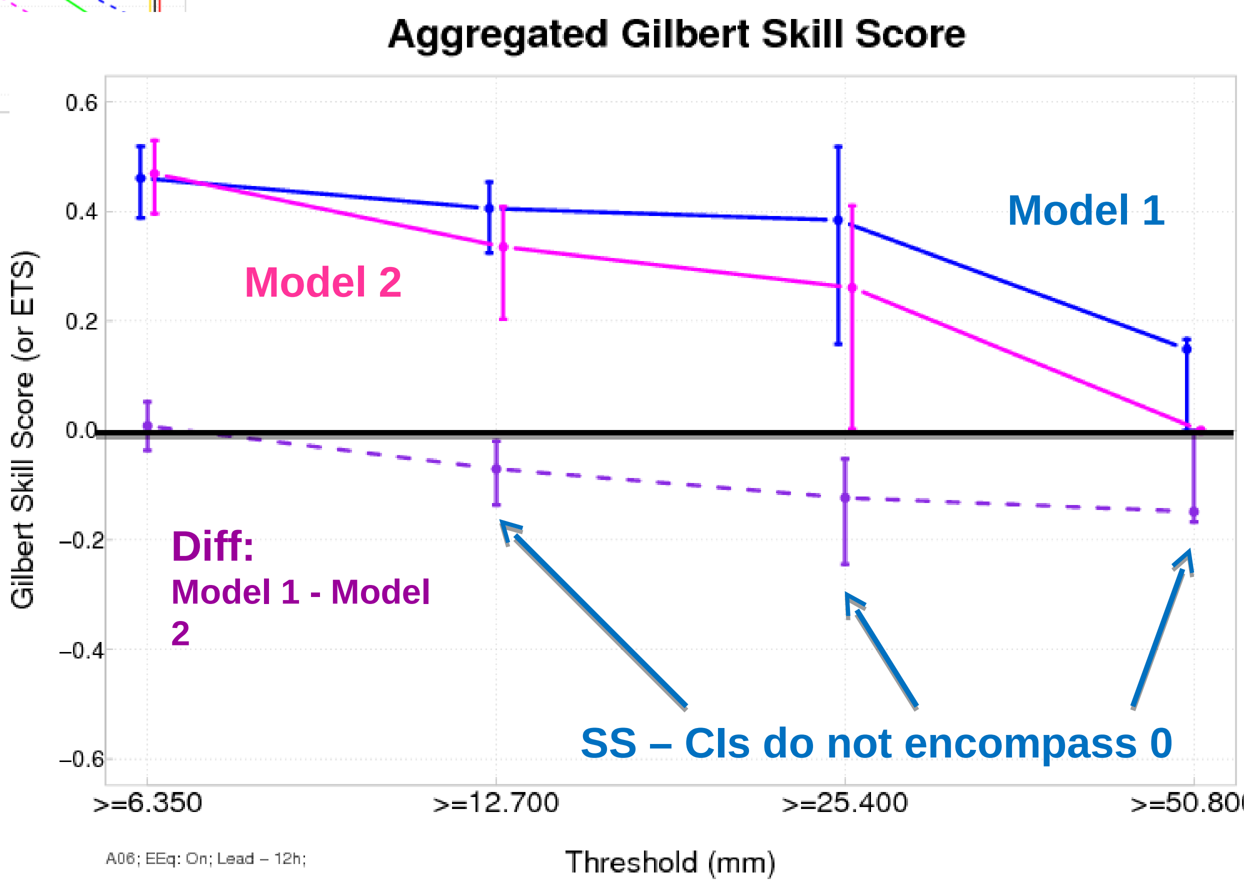
A06 - 12hr Lead Time

Two ways to examine scores



CI about Actual Scores

may be difficult to differentiate model performance differences



CI about Pairwise Differences

may allow for differentiation of model performance

CI application considerations

Normal approximation

- Quick
- Generally pretty accurate
- Only valid for certain measures

Bootstrap approach

- Speed depends on number of points
  - Using grids can be expensive (quicker with points)
- Speed depends on number of resamples
  - Recommended #: 1000
  - If that’s too many: determine where solutions converge to pick the value

Reminders and other considerations

- Normal approaches only work for some verification measures
  - Need to evaluate appropriateness of normal approx for verification statistics
- For all CIs:
  - Need to consider non-independence and ways to account for it
- Multiplicity (computing lots of confidence intervals) makes the error rate much larger than indicated by α
- CIs provide a meaningful and useful way to compare forecast performance

References and further reading

- Garthwaite PH, Jolliffe IT & Jones B (2002). *Statistical Inference*, 2nd edition. Oxford University Press.
- Gilleland, E., 2010: Confidence intervals for forecast verification. NCAR Technical Note NCAR/TN-479+STR, 71pp. Available at:http://nldr.library.ucar.edu/collections/technotes/asset-000-000-000-846.pdf
- Jolliffe IT (2007). Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 637-650.
- Jolliffe and Stephenson (2011): *Forecast verification: A practitioner's guide*, 2nd Edition, Wiley & sons
- JWGFVR (2009): Recommendation on verification of precipitation forecasts. WMO/TD report, no.1485 WWRP 2009-1
- Nurmi (2003): Recommendations on the verification of local weather forecasts. ECMWF Technical Memorandum, no. 430
- Wilks (2011): Statistical methods in the atmospheric sciences, Ch. 7. Academic Press
- http://www.cawcr.gov.au/projects/verification/