Class 9: Analysis plan

Miguel F. Morales
Bryna Hazelton

Data

Data cleaning

Calibration

Background removal

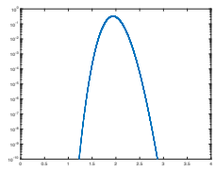
Intermediate integration

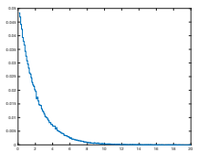
Quality assessment & cleaning pt 2

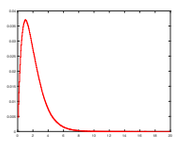
PS transform

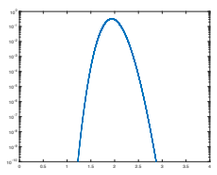
⋮

Error Model









⋮

Worries

Thunder storms

Biasing result

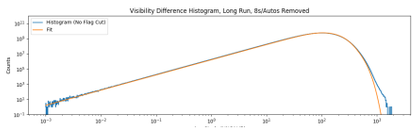
Temperature dep. offset

Signal leakage

⋮

Tests

Jackknife



Correlation

Injection test

⋮

Questions

- Big questions: what you want to know in the end
- Little questions: what should data look like at each step
- Worries: Concerns about the analysis as a whole, but best answered with well understood intermediate products.

Data

Someone else

Step 1

Step 2

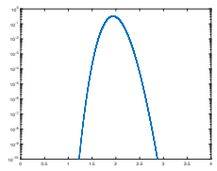
Step 3

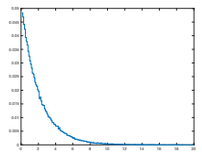
Step 4

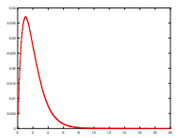
Someone else

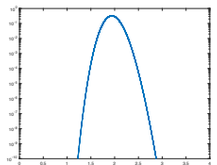
⋮

Error Model









⋮

Worries

Thunder storms

Biasing result

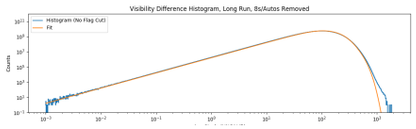
Temperature dep. offset

Signal leakage

⋮

Tests

Jackknife



Correlation

Injection test

⋮

Little steps

Break your analysis problem into a few simple steps.

Examples:

- Filtering data in a specific way (throw outliers on metric A)
- Transforming the data (e.g. averaging, taking a Fourier Transform)
- A calibration step
- Determining or transforming units (e.g. milliVolts to keV)
- Fitting a model (data now in parameter space)

Data

Someone else

Step 1

Step 2

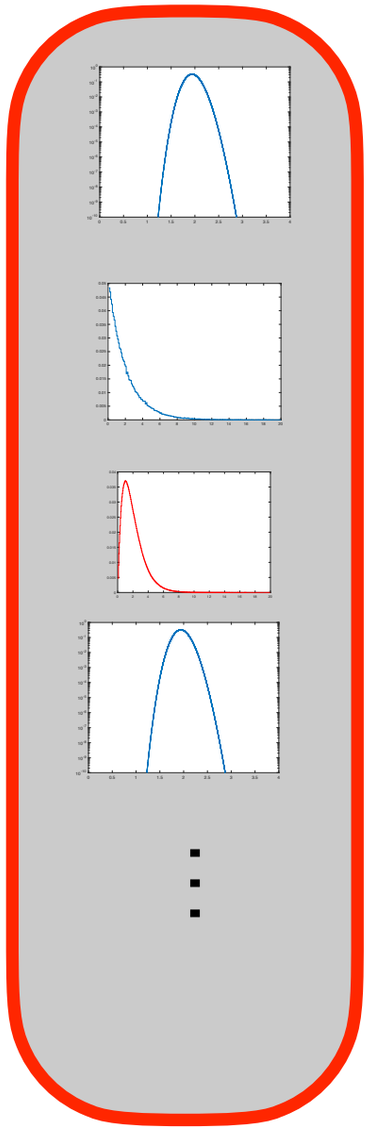
Step 3

Step 4

Someone else

⋮

Error Model



Worries

Thunder storms

Biasing result

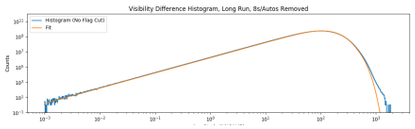
Temperature dep. offset

Signal leakage

⋮

Tests

Jackknife



Correlation

Injection test

⋮

Predict background at each step

Examples:

- Unit or axis change (mV->keV, or uncalibrated to calibrated)

- Filtering or removal of outliers (data quality cuts)

- A mathematical transformation (average, square, Fourier Transform, parameter fit).

Can be placeholders but think about the background and how it changes.

Data

Someone else

Step 1

Step 2

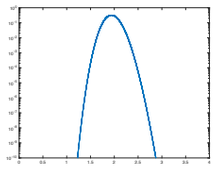
Step 3

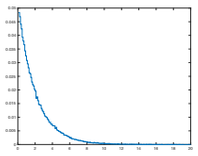
Step 4

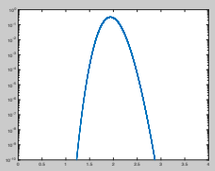
Someone else

⋮

Error Model







⋮

Worries

Thunder storms

Biasing result

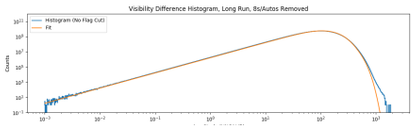
Temperature dep. offset

Signal leakage

⋮

Tests

Jackknife



Correlation

Injection test

⋮

Data

Someone else

Step 1

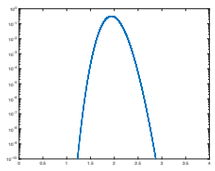
Step 2

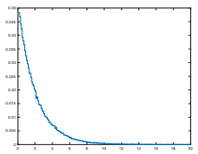
Step 3

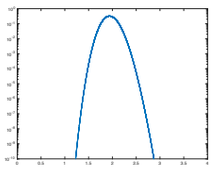
Step 4

Someone else

Error Model







Worries

Tests

Temperature dep. offset

Correlation

Develop a list of worries & tests

Make a list of worries

- Take your time, creative & concrete both useful

Develop a test for each

- jackknife, specific plot, or statistical test

Link to analysis step(s) where test is best performed

- Many questions usually best answered on some intermediate data space and/or internal analysis switch

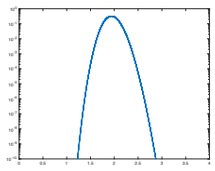
Data

Error Model

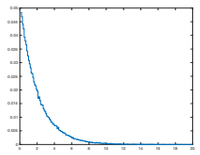
Worries

Tests

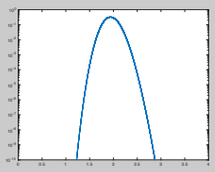
Someone else



Step 1



Step 2



Temperature dep. offset

Correlation

Step 3

Step 4

Someone else

Organize

GitHub issues are great here

- Self-contained analysis steps & code
- Linking of tests to specific analysis & testing code
- Documenting progress on worries (memos also great)
- Checking off of worries (closing issues)
- Ties in with more advanced techniques, e.g. provenance (next week)

Data

Data cleaning

Calibration

Background removal

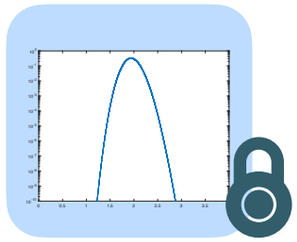
Intermediate integration

Quality assessment & cleaning pt 2

PS transform

⋮

Error Model



⋮

Worries

Thunder storms

Biasing result

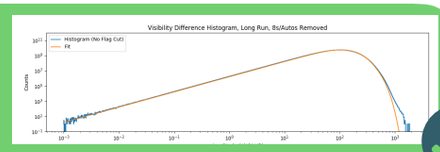
Temperature dep. offset

Signal leakage

⋮

Tests

Jackknife



Correlation

Injection test

⋮

Version controlled software

GitHub issues

DON'T PANIC!

Your assignment

- Design an analysis chain
  - Break into small steps
  - Predict background at each step
- Make a list of worries
  - Develop a jackknife or test for each worry
  - Determine at which step(s) the test is best done

Data

Someone else

Step 1

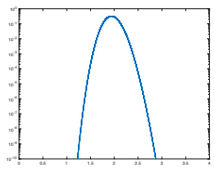
Step 2

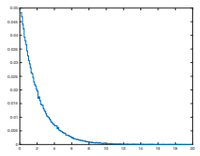
Step 3

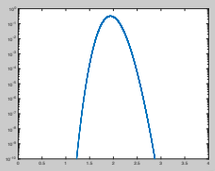
Step 4

Someone else

Error Model







Worries

Temperature dep. offset

Correlation

Tests

Metadata

Metadata

All the information about your data

- When it was taken
- What instrument took it
- How it was calibrated (calibration version, code that did calibration, etc.)
- Prior steps in analysis

Any information you need about the data for a plot

Metadata as a nutrition label for your data

Tagging your data with information

Metadata goals

- **Basic:** you can read the information about your data from the file (nutrition label for everything on a plot)

- **Goal:** you can recreate the analysis as needed (routines run, git hashes, etc.)

- **Advanced:** can recreate full instrument & analysis state (e.g. adds links back into Monitor & Control database; library versions)

How to store metadata

- ***Don't*** store it in the file name
  - Too mutable
  - Not enough space

- Almost all modern file formats have locations for metadata (e.g. headers)
  - FITS, hdf5, AVRO…

Using a standard binary file format (fits, hdf5, avro)

- Accurate (conversion, endian issues, etc.)
- Compact (stores the bits; lossless compression possible)
- Fast (no extra conversion)
- Partial read & write (some of them, important for big data)
- Standard & user defined places to put metadata!

Filenames

- It is useful to have some form of metadata in filename
  - File number, date, etc.
- Too easy to overwrite
- Store all metadata in file headers
  - Copy useful subset into file name for convenience
  - If internal metadata and file name disagree, internal wins