First step in statistics: Getting Good Data

This work is licenced under the Creative Commons Attribution-Non-Commercial-Share Alike 2.0 UK: England & Wales License. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/2.0/uk/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Getting Good Data

Analysis is:

Garbage in → Garbage Out

Getting Good Data

=

- Find a research question that's worth asking.
- Design an experiment that has a chance of answering it.
  - Let the subjects answer the question.
  - Understand your tools.
  - Get enough data and a little more.
- Run a pilot experiment: tune your procedures.
- Collect the data.

- Keep a lab notebook (or equivalent)!

Let the subjects answer the question.

Research is all about doing interesting things and explaining them so that other people can do them too. Research results cannot depend on you, because you won't be there when someone else tries to do it.

Let the subjects answer the question.

Subjects are only too willing to produce the answers you want.
(But if they do, your results are meaningless.)

Clever Hans
(c1904)
Stanley Milgram
(1963)

Let the subjects answer the question.

* Never tell subjects what you expect (until afterwards)
* If practical, design the experiment so it isn't obvious what is being tested.
* Giving subjects the wrong idea is not always helpful: they may act on it and introduce biasses. (Also raises ethical issues.)

Let the subjects answer the question.

You are the easiest person for you to fool.

Physical sciences:

- N-rays
- Polywater
- Cold Fusion

Common thread: proponents never seriously test their techniques.
Proponents try to *prove* their techniques *correct*, rather than trying to understand their limitations.

Let the subjects answer the question.

You are the easiest person for you to fool.

What we have learned:

* Use double-blind experiments (if possible)

* Don't allow yourself unnecessary opportunities to affect the results.

* Minimize changes after you've looked at the data.

* Results only count if someone can follow your recipe and get the same answers.

* Any theorist who isn't worried by adverse experimental results is a fraud.

Understand your tools.

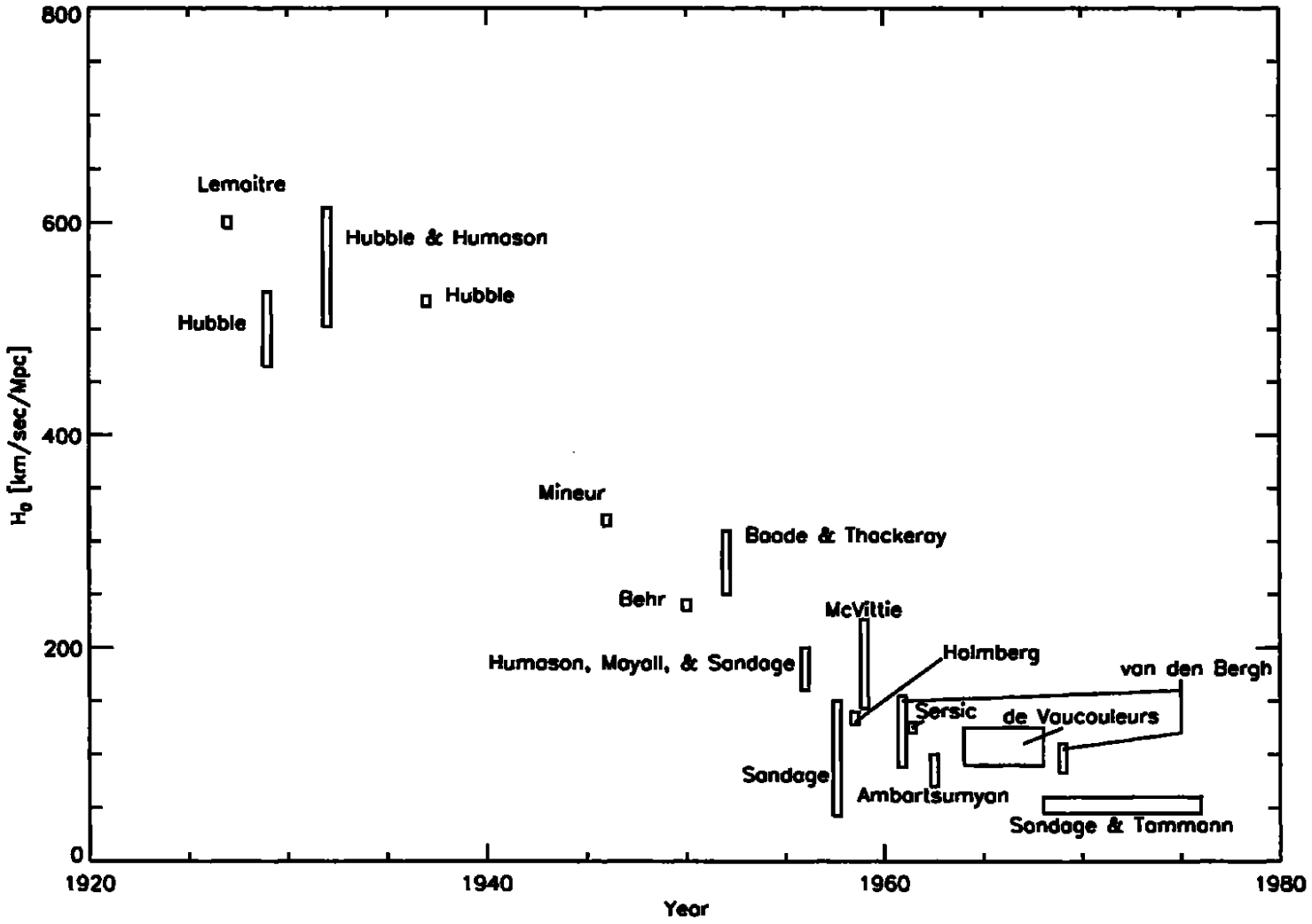

FIG. 1—Published values of the Hubble constant from Lemaître (1927) to the hardening of the battle lines. Rectangle dimensions are intended to suggest a range of values/uncertainties or a range of dates. Except where the errors listed below are larger, all uncertainties were claimed to be of order 10% or less (occasionally much less). A straight-line fit to the numbers from 1927 to 1965 or so would have suggested that the Hubble constant might have become negative within a decade or two (discovered by astronomy graduate students at Caltech in the 1960s and undoubtedly by many others). This did not actually happen. The numerical values represented are Lemaître 600, Hubble 465, 513, 535; Hubble and Humason 526; Mineur 320; Behr 240; Baade and Thackeray 280±30; Hubble, Mayall, and Sandage 180±20; Sandage 75 (+75,−40); Holmberg 134±6; McVittie 143–227; Sersic 125±5; van den Bergh 100 (+20, −12), 120 (+25,−20); Ambartsumyan 70–100; de Vaucouleurs 125, 100±10, 100±10; van den Bergh 95 (+15,−12); Sandage and Tammann 45–60.

Getting Good Data

=

- Find a research question that's worth asking.
- Design an experiment that has a chance of answering it.
  - Let the subjects answer the question.
- Run a pilot experiment: tune your procedures.
- Collect the data.

- Keep a lab notebook (or equivalent)!

Pilot Experiment

Why?

* You can make changes in the midst of a pilot experiment.
* You might think of a better procedure.
* You'll learn what all those buttons and knobs do.

Things you might find out:

The experiment takes too long.

Subjects cannot actually say the necessary sentences.

A survey question was vague or misinterpreted.

The corpus you wanted to use has HTML tags in it.

Word frequencies you get from Google are strange.

Pilot Experiment

Find out if you need to do data selection.

If so, establish rules for accepting/rejecting data.

If

* The rules aren't clear and easy to apply, or
* You cannot explain the rules to a colleague and expect to get the same results, or
* You end up rejecting a lot of data,

then

Improve the rules or re-design the experiment.

Getting Good Data

=

- Find a research question that's worth asking.
- Design an experiment that has a chance of answering it.
  - Let the subjects answer the question.
- Run a pilot experiment: tune your procedures.
- Collect the data.

- Keep a lab notebook (or equivalent)!

Collect the data

* This can require ingenuity when things break
* If things change, ask:
    * “Is there any plausible way this could affect the results?”
    * “How will I describe the change in the published paper?”
* If there is, you may need to start again from scratch

UNIVERSITY OF OXFORD

Keep a Lab notebook (or equivalent)

Communicate with your future self.

* Papers can take a while.

Reviewers can ask for details.

One less opportunity to fool yourself by conveniently misremembering.

Keep a Lab notebook (or equivalent)
What's the equivalent?

Everything is done by writing scripts (little programs).
Scripts are stored in Bazaar version control system.
Scripts have at least a little documentation inside.
Scripts should say where their data comes from and where it goes.

Data files have headers that describe where they came from.
(Intermediate results, too!)

Most scripts, when run, add a line to a LOG file.

The goal: *know out where each result came from,
*be able to understand it a year later,
* be able to fix a mistake and re-compute easily.

Conclusion:

You can design an experiment that takes reliable data.

We know how to do it.

Sometimes, it is too expensive/hard, though. In that case, do what you can and recognize that your results may be due (at least partially) to luck.

Lots of linguistics is built on insufficient / slightly dubious data. Don't be surprised if a few things turn out to be wrong.