# *A socio-public health data-based introductory statistics course*

Murray Aitkin

murray.aitkin@unimelb.edu.au

School of Mathematics and Statistics

The University of Melbourne

Australia

## Turbulence in the profession!

- A widespread dissatisfaction with the present curriculum.
- p-values banned by a minor psychology journal.
- ASA statement on p-values recommending their replacement – but by what?
- Q-step support by ESRC and the Nuffield Foundation of the development of new statistics or “quantitative methods" courses for social science graduate and undergraduate students by the social science departments themselves, without the participation of statistics departments.
- Statistics departments sidelined? A warning bell ringing!
- Special issue (November 2015) of The American Statistician on statistics and the undergraduate curriculum.
- The Statistical Society of Australia held a two-day workshop (June 2016) to develop proposals for modernising statistics courses at all levels of school and University.

## Do we need a new introductory stat course?

The editors of the TAS special issue focussed on second- and higher-level courses:

> Likely the first and most important place to start the curriculum conversation is with the courses that follow an introductory statistics course. (N.J. Horton and J.S. Hardin, Special issue editors)

George Cobb, Mount Holyoke College, did not agree with that:

> Mere renovation is too little too late: we need to rethink our undergraduate curriculum from the ground up.

The Special Issue has a curious absence of discussion of content for the first course, apart from issues like bootstrapping replacing parametric inference.

The SSA subgroup which considered the undergraduate curriculum came to a consensus on the first course: it should be data-, models-, and probability-oriented.

# Why do we need a new introductory course? – Cobb

... I have come to the conclusion that our consensus about curriculum needs to be rebuilt from the ground up. Our territory – thinking with and about data – is too valuable to allow old curricular structures to continue to sit contentedly on their aging assets while more vigorous neighbours take advantage of our latest ideas. (p. 267)

... Markov chain Monte Carlo and related methods have led to a widespread use of Bayesian methods for applied work, which use, in turn, has led to a major reversal of an earlier prejudice against what had long been dismissed as an inappropriately subjective approach to data analysis. (p.270)

... If we are truly to rethink our curriculum at a deep level, we ought to start with foundations.

I am convinced we will need an extended period of ferment, experimentation, and settling out to reach a new consensus on content, much as it took us decades to reach the old consensus on the now middle-aged introductory course. (p. 273)

## What's in the traditional intro stats course?

- Population and sample; descriptive statistics – mean and variance – of a sample;
- simple probability; the normal distribution;
- sampling distribution of the sample mean for the normal distribution;
- the Central Limit Theorem and the large-sample normal distribution of the sample mean;
- the z-test for a hypothetical mean with known variance;
- the t-test for a hypothetical mean with unknown variance;
- confidence intervals for the mean and for a population proportion;
- the t-test and confidence intervals for the difference of two means and of two proportions;
- simple linear regression and correlation.

## What's wrong with this? – what is missing?

- A data base. Students see small samples, and may have to collect data themselves, but do not see a realistic large survey or small population data base.
- The research questions – why do we have these data? Who wants to know?
- The importance of the sample design – (not just for sampling distributions).
- An understanding of probability (though sampling distributions are expected to be understood).
- The idea of a probability model.
- Any principles for statistical inference (the Central Limit Theorem is not an inferential principle).

## An ancient syllabus

Apart from the t-test, this intro stat curriculum is pre-1900 (Student's use of the t-test was published in 1908).

How can all these extra topics be fitted into a course which is already overstuffed?

They can't, but space can be made by limiting the range of models and analyses.

## A successful non-standard course

- The data base – a small population of 1296 families in a Child Development Study at UC Berkeley.
- The research questions – what is the effect of mother's (and father's) smoking on their child's development through
  - birthweight;
  - physical and intellectual development at age 10.
- The study design.
- Random sampling from the database.
- The dangers of voluntary response and other non-random sampling methods.
- Sampling binary attributes: the binomial distribution.
- Inference from sample to population: the likelihood function.

## How to handle inference? – frequentist

Need the repeated-sampling distribution of the sample proportion for confidence intervals (point estimates are of no value, since they are always wrong).

Hand-waving needed (statistical theory shows that ...)

Confidence intervals for differences between proportions (more hand-waving) in 2x2 tables (No $X^2$ test ...).

Continuous variables dichotomised at the median or common median.

Not efficient but workable – same theory for differences.

Uncomfortable with hand-waving – Bayes is easier

## How to handle inference? – Bayesian

Likelihood function conveys the data information.

How to turn this into a probability statement?

Prior distribution on $p$.

Finite population of size $N$, so $p$ must be one of the values $0/N, 1/N, ..., N/N$.

If no prior preference for one of these over another, all have equal prior probability $1/(N+1)$.

Bayes's theorem provides the solution.

How to demonstrate or justify it? – the screening test – a widely used and useful example.

(New intro Bayesian book: Statistical Rethinking: a Bayesian course with examples in R and Stan. Richard McElreath, Chapman and Hall 2016.)

## The screening test

- A condition $C$ is uncommon – present in 2% of the population.
- For those people who have the condition, the screening test gives a true positive result 95% of the time.
- For those people who do not have the condition, the screening test gives a false positive result 10% of the time.
- We test a population of 1000 people in a small town. The true positive and false positive rates apply to this town population.

What do we conclude from the results of the screening test?

# Venn diagram – contingency table

We write + if the test result is positive, and - if the test result is negative.

We write yes if the person tested has the condition, and no if the person tested does not have the condition.

The test results for the town are:

| | condition present | |
|---|---|---|
| test | yes | no | Total |
| + | 19 | 98 | 117 |
| - | 1 | 882 | 883 |
| Total | 20 | 980 | 1000 |

What to conclude?

# How effective is the screening test?

| | condition present | |
|---|---|---|
| test | yes | no | Total |
| + | 19 | 98 | 117 |
| - | 1 | 882 | 883 |
| Total | 20 | 980 | 1000 |

- There are 20 (2%) cases and 980 (98%) non-cases.
- Of the 20 cases, 19 (95%) are correctly identified – true +.
- Of the 980 non-cases, 98 (10%) are incorrectly identified – false +.
- Of the 117 + tests, 19/117 (16%) are from people who had the condition.
- Of the 883 - tests, 1/883 (0.11%) are from people who had the condition.

## Conclusion

- If your test was negative, you can be reasssured that you are very unlikely to have the condition.
- If your test was positive, the probability that you have the condition is only 16%.

So what was the point of the screening test?

Many students find this shocking.

The true positive rate is 95%!

So surely almost everyone testing positive must have the condition!

The fallacy of the transposed conditional.

Bayes's theorem from a 2x2 table, without algebra.

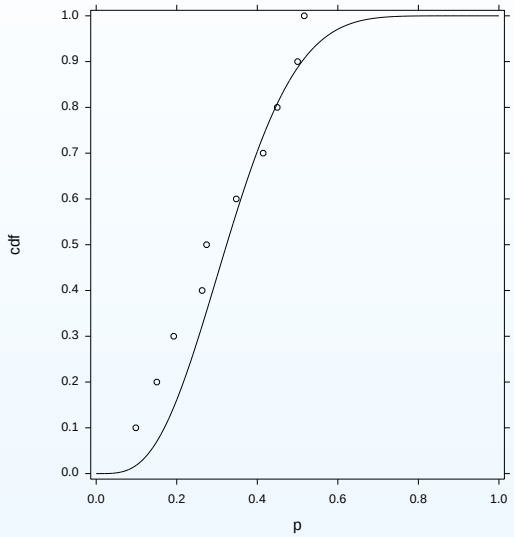
## The role of posterior simulation

- Bayesian analysis has been greatly enriched, but also simplified, by the ability to generate large numbers of posterior draws of model parameters through MCMC,
- and to combine these in any way, with or without observed data,
- to give posterior distributions of any functions of data and parameters.

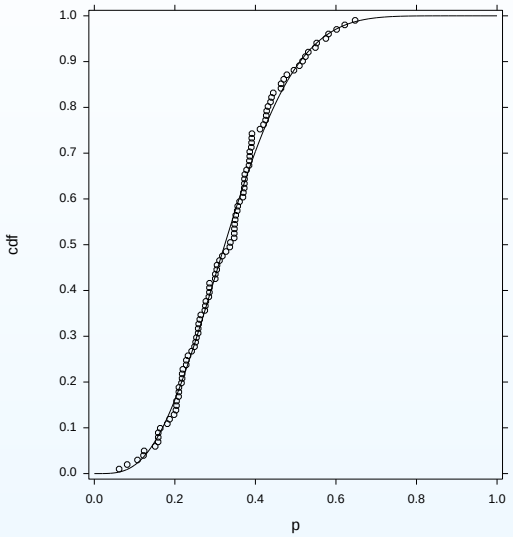
This idea is new to students, and is illustrated here with successively larger numbers of random draws from the Beta (4,8) distribution; we show the empirical cdf of the draws and the true cdf (solid curve).

With 10,000 draws the empirical cdf is very smooth, and overlaps the true cdf almost exactly.

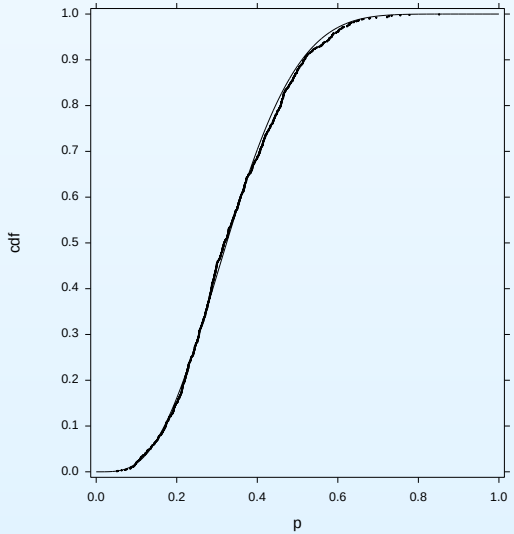
# Simulations from Beta(4,8)



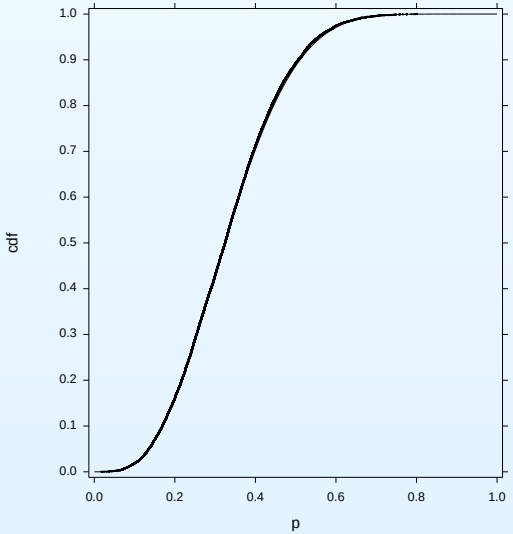

(a) $n = 10$




(b) $n$=100



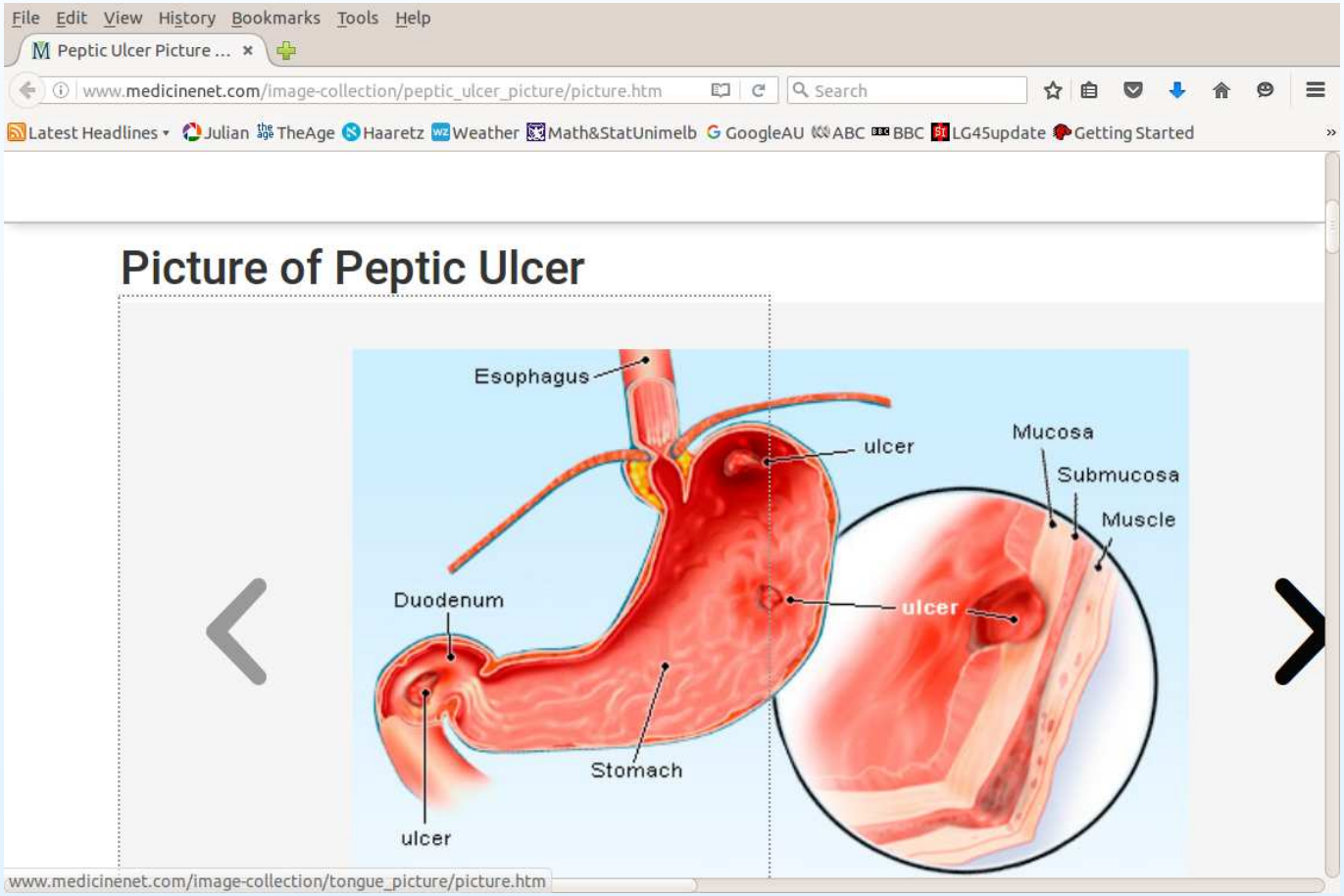

(c) $n$=1,000




(d) $n$=10,000

# RCT of Depepsen for the treatment of duodenal ulcers

Posterior simulations provide a simple procedure for credible intervals for the difference in proportions responding in a randomised clinical trial:



# RCT of Depepsen

- A study carried out at the Royal North Shore hospital in Sydney by Professor D.W. Piper and co-workers.
- Depepsen (a trade name for sodium amylosulphate) had been found effective in the treatment of gastric (stomach) ulcers.
- It was used in an RCT for duodenal ulcers, together with best current treatment (bed rest, antacids, light diet, sedatives), and compared with placebo with current best treatment.
- The criterion for success was complete healing of the ulcer within a period of 8 weeks after the beginning of treatment.
- 18 patients were randomised to Depepsen, and 17 to placebo.

## Outcome

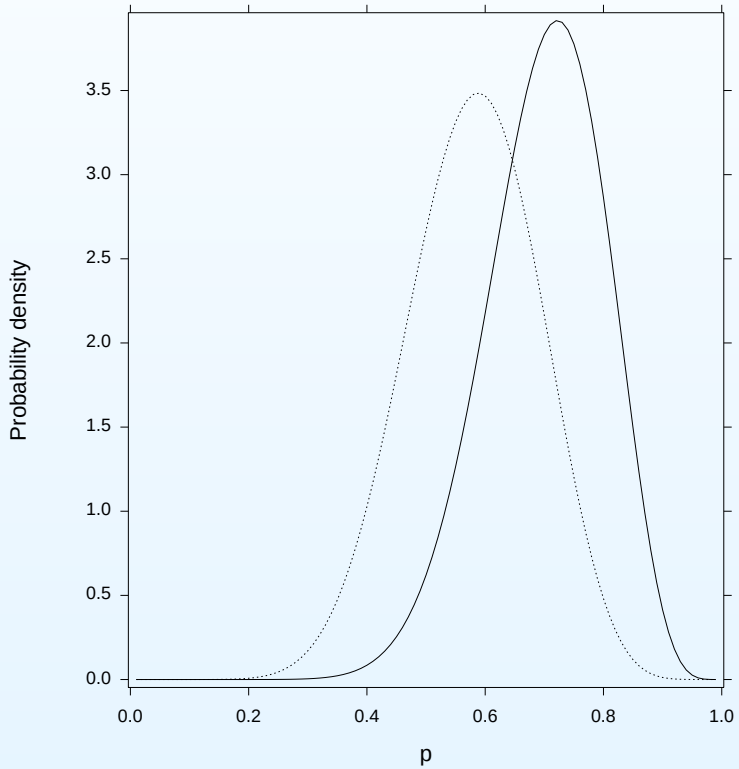
| | Depepsen | Placebo | Total |
|---|---|---|---|
| Healed | 13 | 10 | 23 |
| Not healed | 5 | 7 | 12 |
| Total | 18 | 17 | 35 |

A slightly higher proportion of Depepsen patients recovered in 8 weeks: 0.72 vs 0.59. What can we say about the true value of $p_D - p_0$?

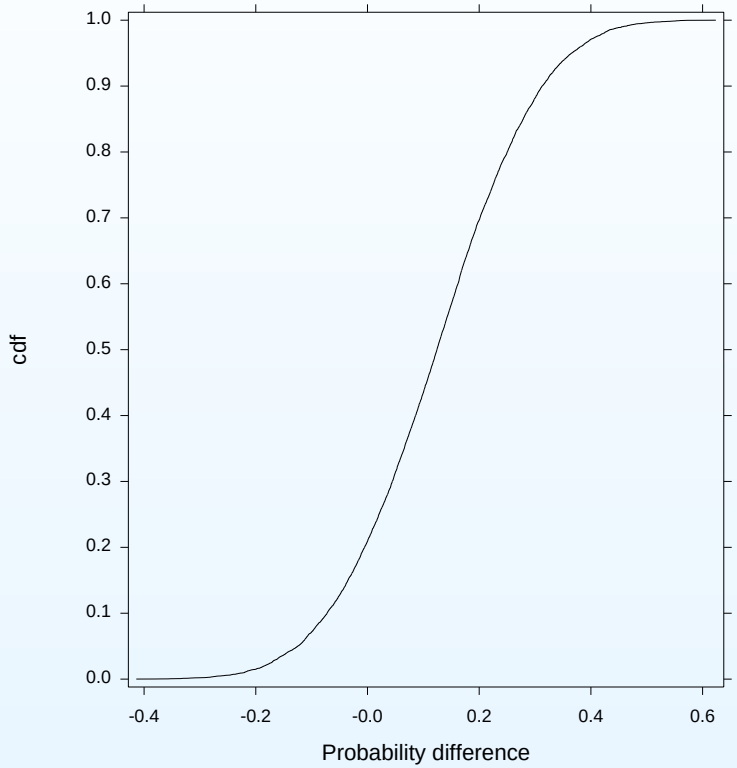
With uniform priors on $p_D$ and $p_0$, the posterior distributions are Beta (14,6) and Beta (11,8).

We make 10,000 independent random draws $p_D^{[m]}$ and $p_0^{[m]}$ from these posteriors, and form the differences $\delta^{[m]} = p_D^{[m]} - p_0^{[m]}$. Their cdf follows.

# Posterior densities placebo (dotted) and Depepsen (solid)



# Cdfs of 10,000 differences $p_D - p_0$



## Credible intervals

- The median difference is 0.12 (close to the difference 0.13 in sample proportions);
- the lower and upper 2.5% points of the cdf are $-0.17$ and 0.41, so
- the central 95% credible interval for $p_D - p_0$ is $[-0.17, 0.41]$ which includes zero, the no difference value. (The asymptotic central 95% confidence interval for $p_D - p_0$ is $[-0.19, 0.45]$.)
- The difference in recovery proportions in the two populations could plausibly be as much as 0.41 in favour of Depepsen, or as much as 0.17 in favour of placebo, or zero.

The trial is so small that the small difference in sample proportions is a poor indicator of the difference in the population proportions, which could be zero – a critical issue for recommending the Depepsen treatment.

## Future treatments

Soon after this trial, a different drug treatment for duodenal ulcers – cimetidine (trade name Tagamet) – was found to be effective, and trials of Depepsen for the treatment of duodenal ulcers were abandoned.

In the last ten years, these drug treatments, which were based on reducing acidity in the stomach, have been replaced by an entirely different treatment with antibiotics –

It was discovered by Dr Barry J. Marshall and Dr J. Robin Warren of Perth, Western Australia, that most ulcers develop from a stomach infection by the Helicobacter pylori bacterium, which responds rapidly to antibiotic drug treatment.

They were awarded the 2005 Nobel Prize for Medicine and Physiology.

☺ Thank you! ☺