Where do the (big) data come from?
… the role of probability sampling

Rod Little

Department of Biostatistics

UNIVERSITY OF MICHIGAN

Outline

- My journey, and the nature of statistics: much more than bean counting
- Statisticians love to toss coins: describe and compare two roles of randomization in

  (a) sample selection (today)

  (b) treatment assignment (next time)
- In my third talk, I'll discuss Maximum Likelihood, a major tool for statistical inference

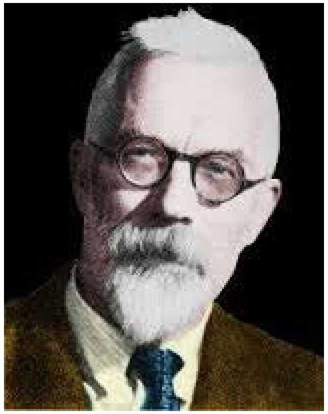
Early days

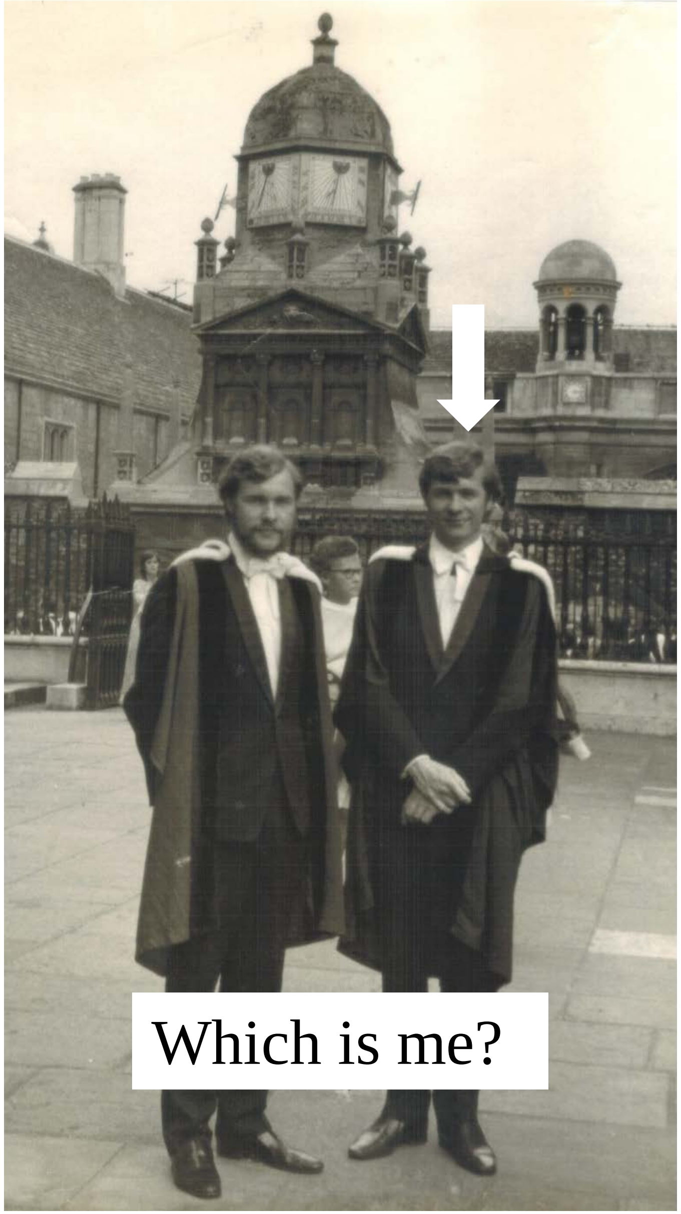
1956-1968: Glasgow Academy

1968-71: BA Mathematics, Gonville and Caius College, Cambridge

(R.A. Fisher's

college)



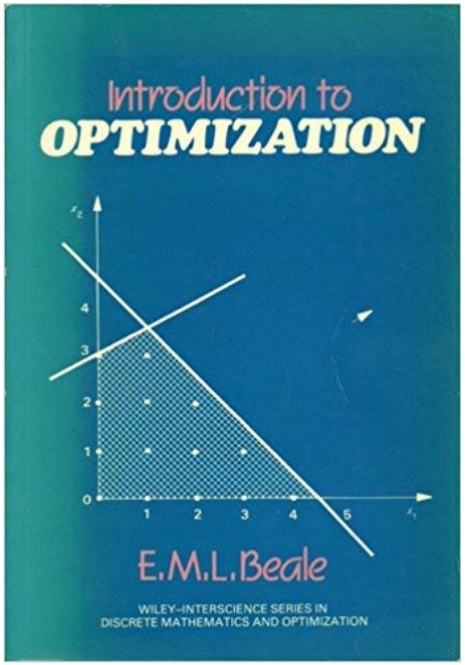


Which is me?

Postgraduate training

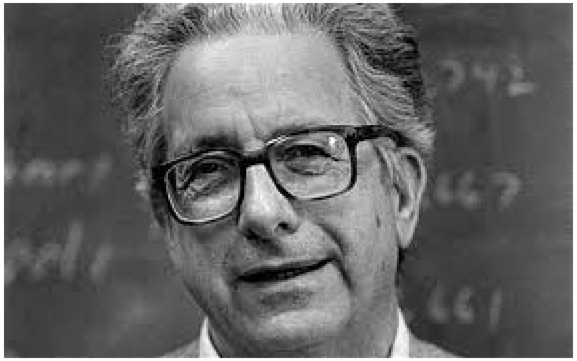
1972-74 MS/PhD in Statistics and Operations Research

D.R. Cox (winner of the 2017 International Prize in Statistics)


Introduction to
OPTIMIZATION
E.M.L.Beale
WILEY–INTERSCIENCE SERIES IN
DISCRETE MATHEMATICS AND OPTIMIZATION

1975-76 Post-doc, Department of Statistics, University of Chicago. My first chair was Paul Meier, of the “Kaplan-Meier curve” in survival analysis. (One of the most cited papers in science)

Paul Meier – from The Telegraph obituary (2011)



"Paul Meier, who died on August 7 aged 87, was a statistician who championed the idea of testing new medical treatments through randomised trials, so helping to lead a revolution in clinical research and **saving, albeit indirectly, millions of lives…**

… The idea of assigning subjects in medical trials solely on the basis of random selection might now seem obvious. But, like many medical innovations, it did not seem so at the time Meier proposed it in the 1950s…. **Many physicians were horrified at the idea that their selection should be random**, together with an equally randomly-selected "control" group of patients who were given the standard treatment or a placebo… At first Meier's arguments met with incomprehension: "When I said 'randomise' in breast cancer trials," he recalled in 2004, "I was looked at with amazement by my medical colleagues: **'Randomise? We know that this treatment is better than that one.' I said, 'Not really!'".**

3. World Fertility Survey (1976-80)

Recruited by Sir Maurice Kendall, Director of the World Fertility Survey (WFS)

Sir Maurice was a prominent statistician, noted for the treatise with Alan Stuart “The Advanced Theory of Statistics”

Sir Maurice was proud that WFS conducted **probability surveys** in developing countries, a design that he liked to describe as “scientific”…

Also a poet and joker, see “Hiawatha designs an experiment:” http://www.mscs.mu.edu/~paulb/Pomes/hiawatha.pdf

3. World Fertility Survey (1976-80)

Recruited by Sir Maurice Kendall, Director of the World Fertility Survey (WFS)

Sir Maurice was a prominent statistician, noted for the treatise with Alan Stuart “The Advanced Theory of Statistics”

Sir Maurice was proud that WFS conducted **probability surveys** in developing countries, a design that he liked to describe as “scientific”…

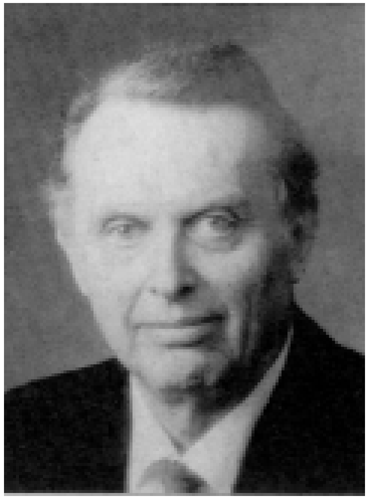
Also a poet and joker, see “Hiawatha designs an experiment:” http://www.mscs.mu.edu/~paulb/Pomes/hiawatha.pdf

5. UCLA Biomathematics (1983-93)

From Wil Dixon's Retirement Party (Oct 1986)

The data of young Sherman Mellinkoff
Had extremes that were knocking his stockings off
He called in Wil Dixon,
Whose trimmed means soon fixed 'em
Dr. M. became Dean, Biomath-took-off (RL)

Sad Chief Wil see his people are needy,
"Stat's too hard!" grunts he, squat in his tepee
Pleads to god of the West Wood,
"Counting beads simply no good."
Gets a vision next day -- BMDP! (Elliot Landaw)

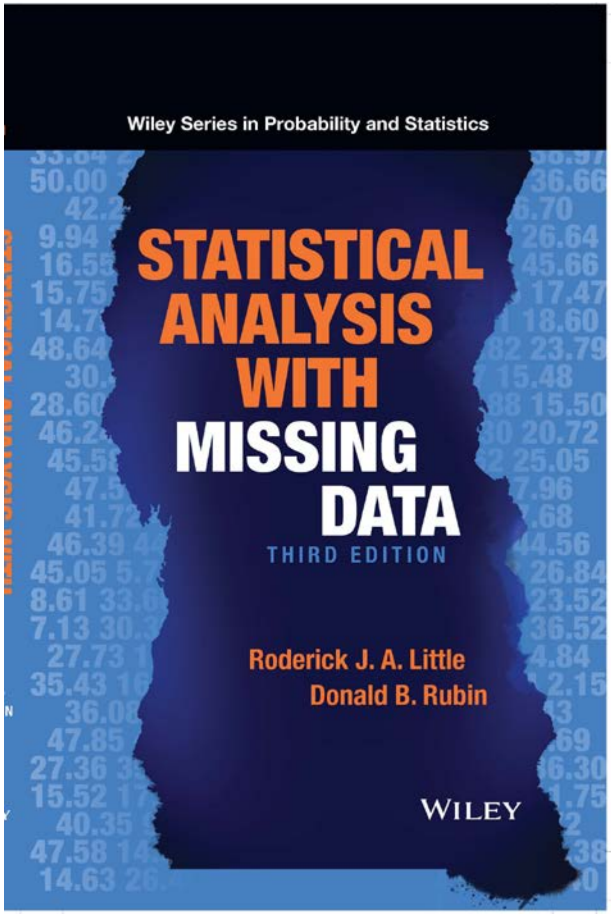


Wilfred Dixon (developer of trimmed means, and BMDP, an important early statistical software program)

6. University of Michigan Biostatistics (1993-present)

- Fine university in a wonderful town
- Great faculty, staff, students
- No earthquakes, hurricanes, floods (just the odd lost tornado)
- Songs and skits…

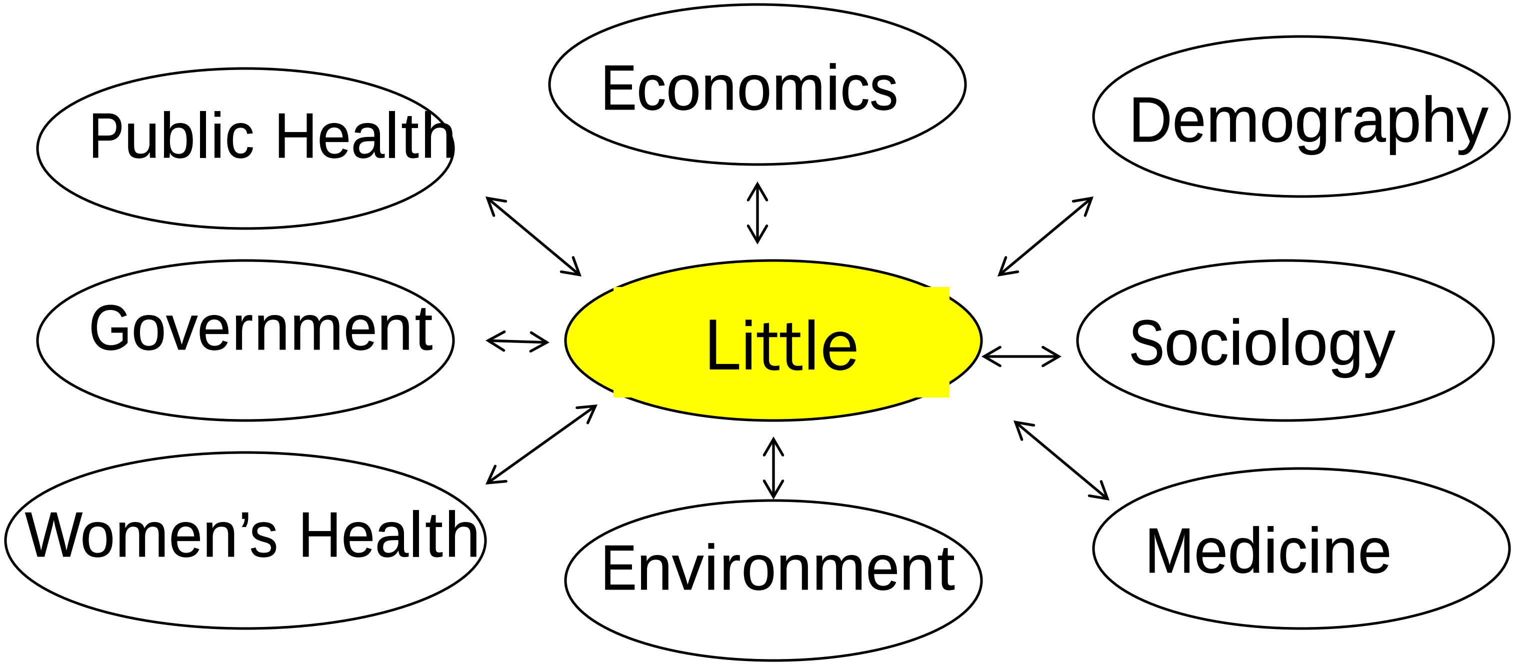
My own work


Wiley Series in Probability and Statistics
STATISTICAL ANALYSIS WITH MISSING DATA
THIRD EDITION
Roderick J. A. Little
Donald B. Rubin
WILEY

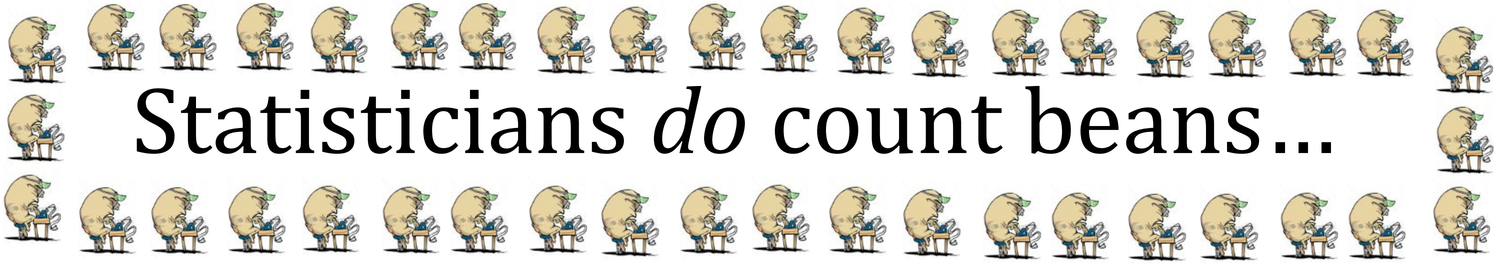
Little,R.J. and Rubin, D.B. (2019) *Statistical Analysis with Missing Data*, 3rd edition. Wiley: New York

Little, R.J. (2012). Calibrated Bayes: an Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder). *Journal of Official Statistics*, 28, 3, 309-372

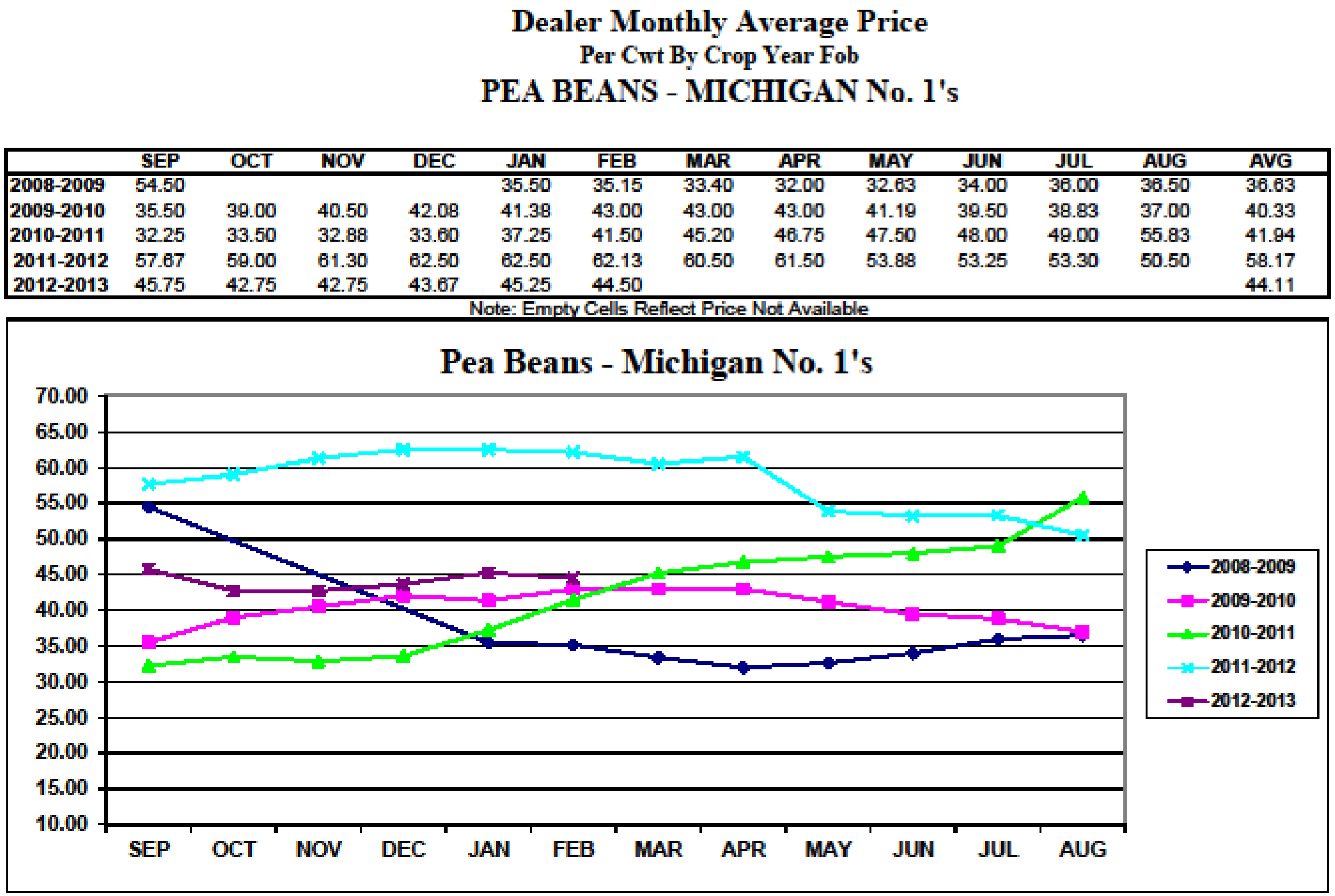
Collaborative Work



The Joy of Stats: So Much More than Bean Counting!



Statisticians *do* count beans…



Dealer Monthly Average Price

Per Cwt By Crop Year Fob

PEA BEANS - MICHIGAN No. 1's

	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	AVG
2008-2009	54.50				35.50	35.15	33.40	32.00	32.63	34.00	36.00	36.50	36.63
2009-2010	35.50	39.00	40.50	42.08	41.38	43.00	43.00	43.00	41.19	39.50	38.83	37.00	40.33
2010-2011	32.25	33.50	32.88	33.60	37.25	41.50	45.20	46.75	47.50	48.00	49.00	55.83	41.94
2011-2012	57.67	59.00	61.30	62.50	62.50	62.13	60.50	61.50	53.88	53.25	53.30	50.50	58.17
2012-2013	45.75	42.75	42.75	43.67	45.25	44.50							44.11

Note: Empty Cells Reflect Price Not Available

2011 Bean Market Summary - USDA Market News - Page 3

It’s what you do with them that matters…

“…now we really do have essentially free and ubiquitous data… so the complementary scarce factor is the ability to understand that data and extract value from it.”

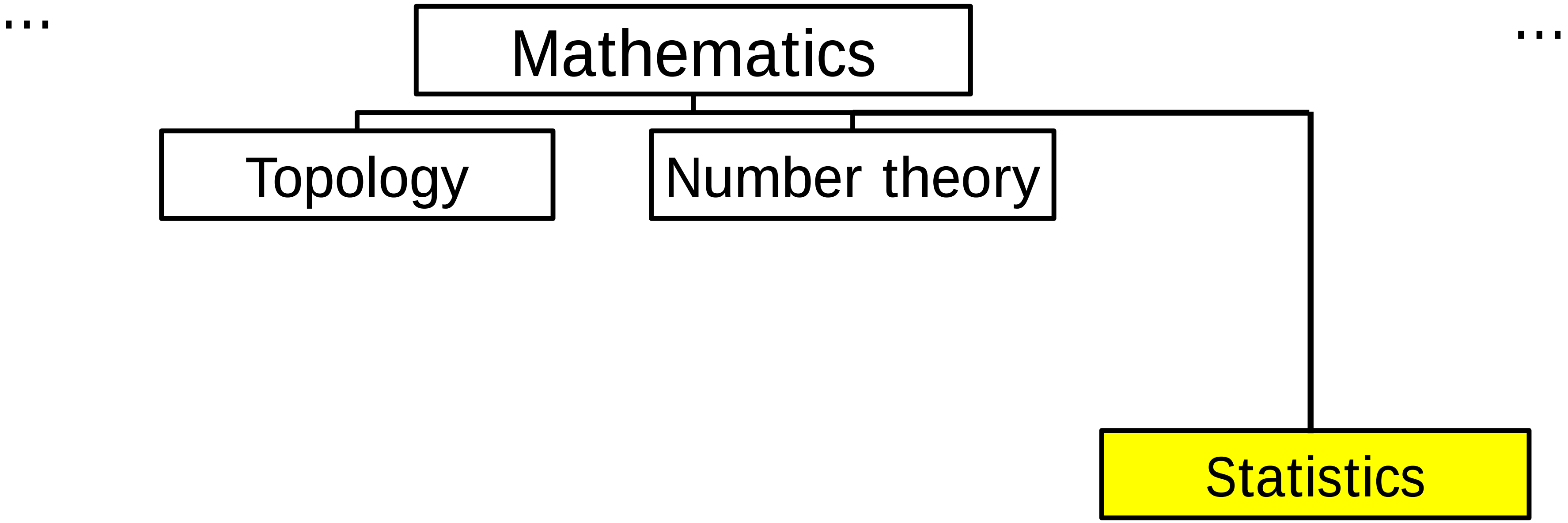
Hal Varian, Chief Economist, Google

Hal Varian

What is statistics?

Not just facts …

or a (rather pedestrian) subfield of math:



…

Mathematics

…

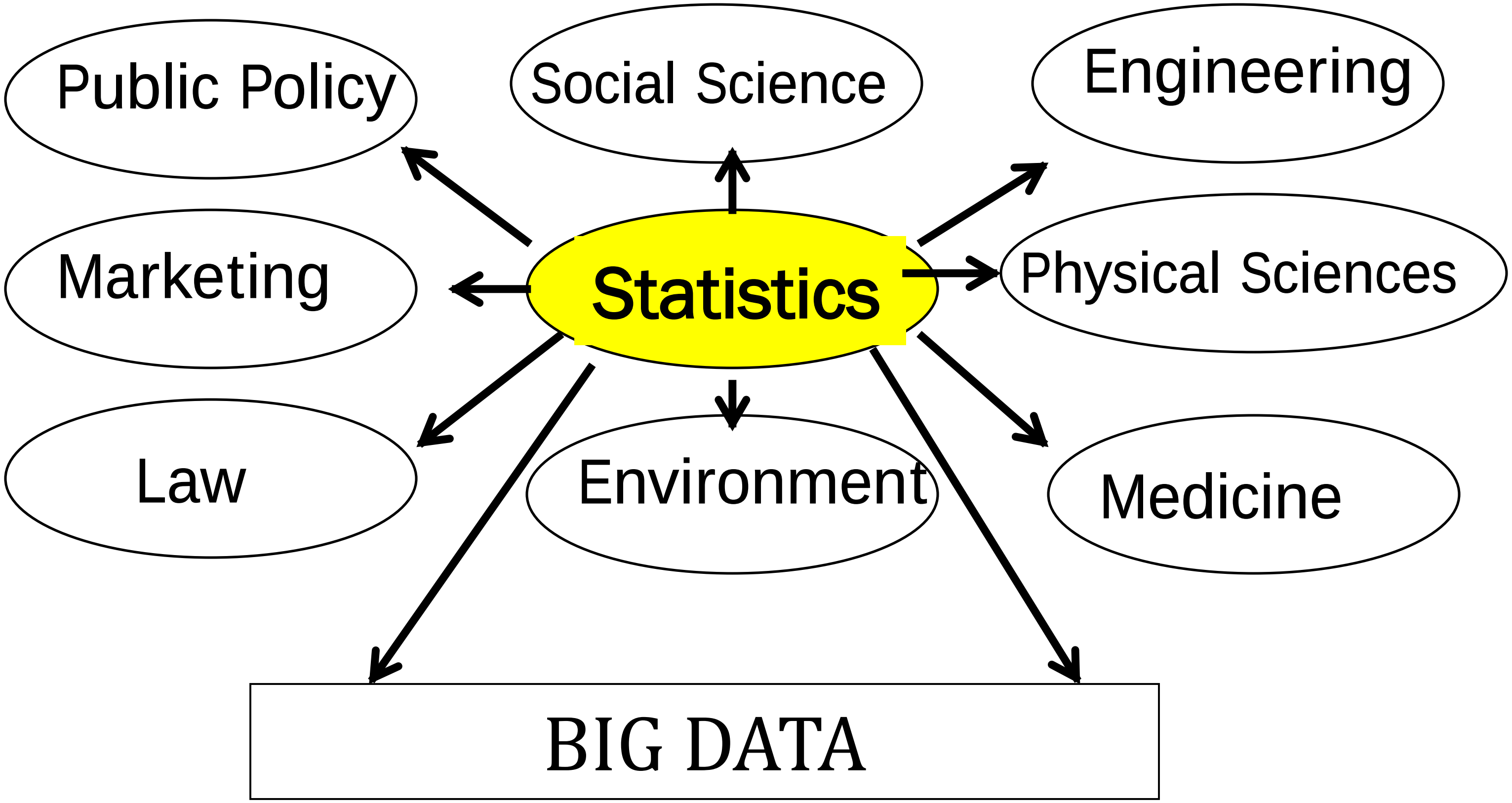
Topology

Number theory

Statistics

What is statistics?

But **data science**:



Statistics reveals the truth …

Revealing the Truth through Statistics

(James Choi, entry to American Statistical Association Public Awareness competition)

Statisticians Impacting Science

20/25 of most-cited mathematicians in science are statisticians (Science Watch 2002)

2 D. L. Donoho Stanford Stat;
3 A.F.M. Smith London Stat
4 E. A. Thompson Washington Biostat;
5 I.M.Johnstone Stanford Stat
6 J. Fan Hong Kong Stat;
7 D.B. Rubin Harvard Stat.
9 A. E. Raftery Washington Stat;
10 A.E. Gelfand U. Conn Stat.
11 S-W Guo Med. Coll. Wisc Biostat;
12 S.L. Zeger JHU Biostat.
13 P.J. Green Bristol Stat; **14** B.P. Carlin Minnesota Biostat
15 J. S. Marron UNC Stat; **16** D.G. Clayton Cambridge Biostat
16 G.O. Roberts Lancaster Stat; 20. X-L Meng Chicago Stat
21. M. P. Wand Harvard Biostat; 22.W.R. Gilks MRC Biostat
23 M. Chris Jones Open U Stat; 25.N. E. Breslow Washington Biostat

Statisticians Impacting Society #1

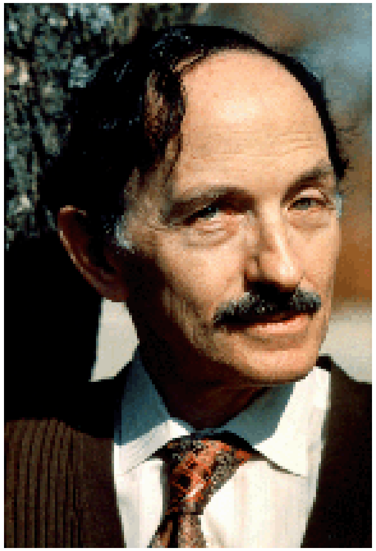
- Sir Ronald Fisher’s experimental designs and analysis of variance have greatly increased the world food supply

Sir Ronald Fisher

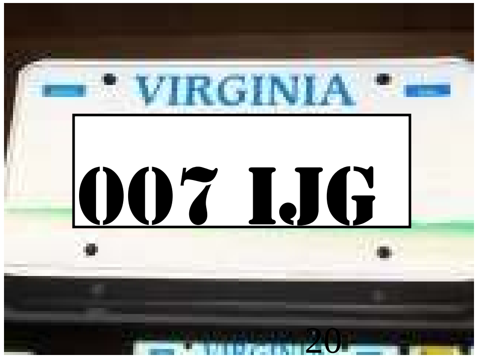
Statisticians Impacting Society #2

Alan Turing sculpture by Stephen Kettle, Bletchley Park. Photo by Jon Callas

Alan Turing and Jack Good's statistical methods helped decode German naval ciphers, arguably reducing the length of World War II by two years or more, saving millions of lives.



I. Jack Good (IJG)



Statisticians Impacting Society #3

- Randomization, a strange but clever idea for
  - – valid answers about populations from surprisingly small sample surveys
  - – randomized clinical trials, pioneered by Sir Bradford Hill, now the gold standard in evidence-based medicine
  - – This is the focus of my first two talks

Statisticians Impacting Society #4

- Official government statisticians … not just bean counters, **guardians of our democracy**.
- http://www.huffingtonpost.com/rod-little/decennial-census_b_3046611.html

Outline

- Statistics: much more than bean counting
- Statisticians love to toss coins: describe and compare two roles of randomization in

    (a) sample selection

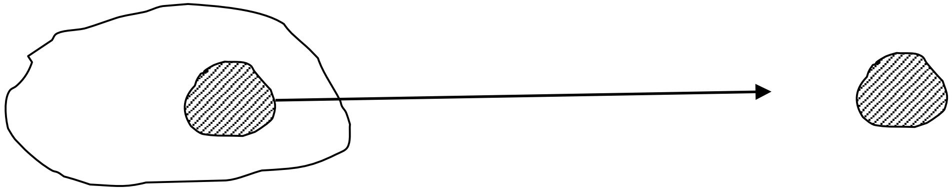
    (b) treatment assignment

What is statistics?

- Statistical Design
- Data collection
- Data description
  - Graphs, tables etc.
- Statistical inference
  - Inferring about a broader population based on a sample
  - Hypothesis testing, confidence intervals, etc.
- This course focuses on the first step … but I’ll say a bit about statistical inference

Inference for a population based on a sample

- Statistical inference: the process of making inferences about parameters of a population based on sample data.
- Usually use Roman symbols for the sample quantities, and Greek symbols for the population quantities



- Population
- Mean μ
- SD σ

- Sample
- Mean $\overline{x}$
- SD s

- Inference crucially requires that sample is randomly selected from population (or an assumption that it is)

The two main tools of classical (frequentist) inference

- **Hypothesis testing:** key output is the P-value
- **Confidence interval:** a random interval that includes the true value of a parameter in a given proportion of repeated samples (e.g. 95%)
- Concepts are related: a 95% confidence interval includes the set of hypotheses that are “consistent with the data” – P > 0.05

Hypothesis testing

- A scientific hypothesis is converted into a null hypothesis about the value or values H_0 of one or more parameters.
- The key output of a hypothesis test is a P-value between 0 and 1 that measures whether the observed data are consistent with the null hypothesis.
  - Small P-Value (say less than 0.05) indicates evidence against the null: either the null hypothesis is false or an unlikely event has occurred. The null hypothesis is "rejected“
  - Large P-Value indicates lack of evidence against the null. The null hypothesis is "accepted", or more precisely, "not rejected".
- *Important:* “Accepting" the null hypothesis does *not* imply that the null hypothesis is true, only that data do not contradict it.

Elements of a hypothesis test

- A scientific hypothesis, e.g. "new treatment is better than old treatment"
- An associated null hypothesis H_0. The null hypothesis is often counter to the scientific hypothesis, e.g. "the average difference in outcomes between treatments is zero".
- An alternative hypothesis H_a: legitimate values of the parameter if H_0 is not true.
- A test statistic T computed from the data, which (a) has a known distribution if the null hypothesis is true and (b) provides information about the truth of the null hypothesis.
- The P-Value for the test is:

$$P = \Pr(\text{test statistic the same or more extreme than } T \mid H_0)$$

- Small P-values are evidence against the null hypothesis

More on P-Value

P-Value = $\Pr(\text{"data"} \mid H_0)$

"data" = "values of T at least as extreme as that observed".

Measures consistency of data with H_0

P-Value is *not* $\Pr(H_0 \mid \text{data})$

That is, is not the probability that H_0 is true given the data

(Latter is computed in Bayesian hypothesis testing)

Strength of evidence against null

As measures of statistical evidence, we can informally divide P-Values into intervals, as follows:

- P < 0.01: strong evidence against null (but some argue for P < 0.005)
- 0.01 < P < 0.05: weak evidence against null
- 0.05 < P < 0.1: at best marginal evidence against null
- P > 0.1: data consistent with null, different values of P above 0.1 (e.g. 0.2, 0.7) have little impact on conclusions

Smaller deviations from the null can be detected with larger sample sizes, so the P-Value is strongly dependent on sample size – it is <u>not</u> a good measure of the size of the effect.

Ex: Confidence interval for population mean

- A confidence interval (CI) is a measure of uncertainty for a sample estimate (e.g. $\overline{x}$) of a population quantity (e.g.μ)
- range of values computed from the sample within which the population quantity is likely to lie
- A 95% confidence interval for a population mean μ (if the sample size *n* is large, say greater than 30) is

$$C_{.95}(\mu) = \overline{x} \pm 1.96\left(s / \sqrt{n}\right)$$

- Works because of the *Central Limit Theorem*, which shows that means have a normal distribution in repeated samples
- Random sample from population is a key assumption
- Confidence interval interpretation: in 95% of repeated samples, this random interval covers the true mean

Problems with P-Values

- P-value is <u>not</u> the probability that the null hypothesis is true, $p(H_0|data)$; it measures consistency of the data with the null hypothesis, $p(data|H_0)$
- P-value is poor measure of the size of an effect –
  - mixes estimate of effect and its uncertainty
  - size of P-value has no clinical meaning
  - P value is strongly determined by sample size – since nothing is <u>exactly</u> zero, anything is significant with a large enough data … so P-Values have limited use for big data!
  - The more important question is the size of the effect, not whether it differs from zero

Problems with P-Values

“Hypothesis testing, as performed in the applied sciences, is criticized. Then assumptions that the author believes should be axiomatic in all statistical analyses are listed. These assumptions render many hypothesis tests superfluous. The author argues that the image of statisticians will not improve until the nexus between hypothesis testing and statistics is broken.”

M. NESTER, An Applied Statistician's Creed *Applied Statistics* (1996) 45,*No.4,pp.* 401-410

See also the “ASA Statement on P-Values” posted on the course site

Problems with P-Values

- The conventional cut-off for statistical significance – $P < 0.05$ – is weak evidence – when translated to the Bayes factor for reasonable choices of alternative, it is too weak to establish effects.
- Some (e.g. Val Johnson) advocate the more stringent cut-off $P < 0.005$. Hence my limerick:

“In statistics, one thing do we cherish

P .05 we publish else perish

Val says, that’s so out-of-date

Our studies don’t replicate

P .005, then null is rubbish!

Confidence intervals

- A confidence interval -- estimate with associated measure of uncertainty
- Confidence interval property – in hypothetical repeated samples, the 95% interval includes the true value of the parameter at least 95% of the time. Here 95% is the "nominal coverage" of the CI
  - Example: 95% CI for population mean in a normal sample of size *n* with mean $\bar{x}$, sd *s* is

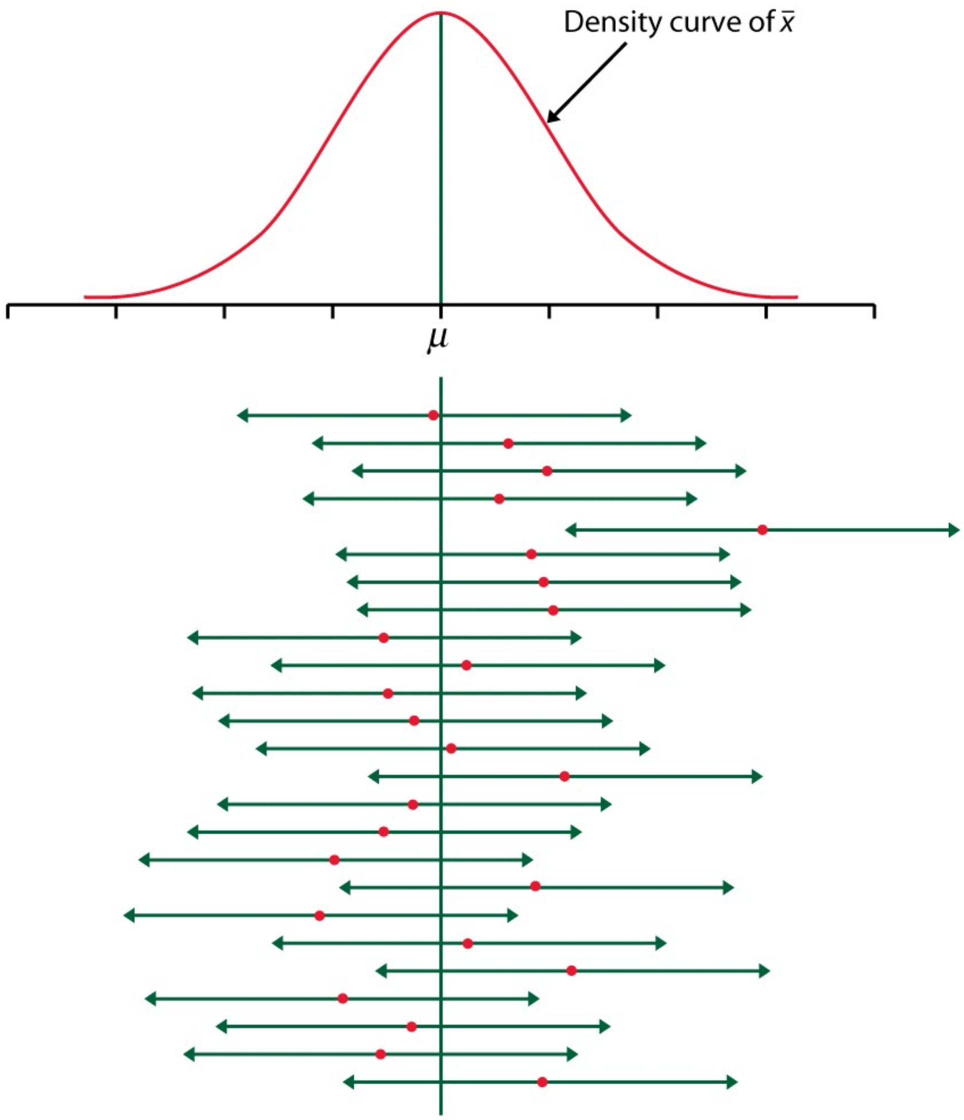
$$\bar{x} \pm t_{.975} s / \sqrt{n}$$

where $t_{.975}$ is the 97.5th percentile of the *t* distribution with n – 1 degrees of freedom. In particular

$t_{.975} = 1.96$ if $n > 50$, $t_{.975} = 2.447$ if $n = 7$.

Roughly "estimate +/- two se's" for moderate size *n*

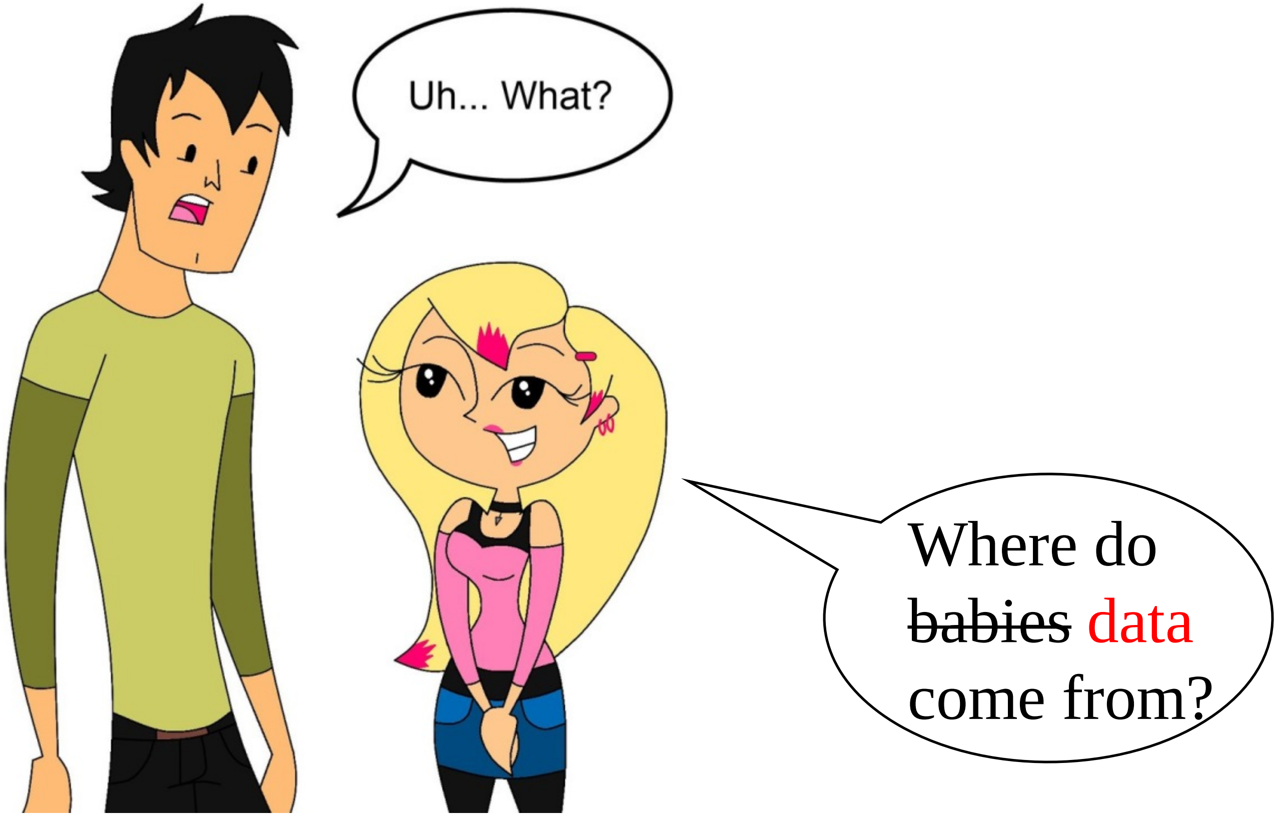
CI interpretation: “in 95% of repeated samples, this random interval covers the true mean”



Confidence Intervals: better than P-values

- Estimate has clinical meaning – closer to the science. Good measurement is the heart of statistics
- width of interval captures uncertainty in a natural way
- Confidence interval summarizes the evidence in a natural way
- But confidence intervals are peculiar objects: the interval is random, but the parameter is fixed
- *Bayesian statistics* provide credibility intervals where interval is fixed, parameter is random

Where do data come from?



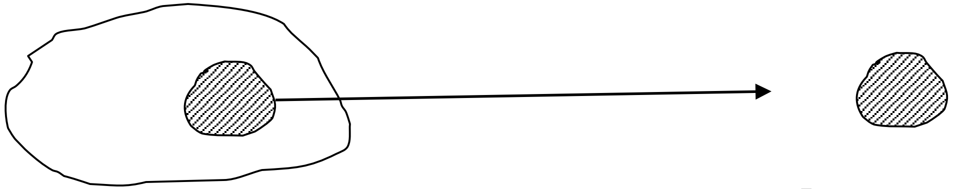
Source: themightypen.deviantart.com

Some quotes about data

- Data always have errors --
- “If a statistic is interesting, it’s probably wrong!”
  - Sir Claus Moser, UK Central Statistics Office
- For “found” big data, not collected with any particular objective or design, beware of the GIGO principle:
- “garbage in, garbage out”
- “It ain’t so much the things we know that get us into trouble. It’s the things we know that just ain’t so” -- Artemis Ward

Root mean squared error of an estimate

- Statistical inference: the process of making inferences about parameters of a population based on sample data



population quantity θ

sample estimate $\hat{\theta}$

- Statistical measures of quality of the estimate include:

Bias: $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Variance: $Var(\hat{\theta})$, Standard Error: $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$

Mean Squared Error: $MSE(\hat{\theta}) = B^2(\hat{\theta}) + Var(\hat{\theta})$

Root Mean Squared Error $RMSE(\hat{\theta}) = \sqrt{B^2(\hat{\theta}) + Var(\hat{\theta})}$

Precision and accuracy

- An estimate is precise if it has low uncertainty -- small standard error, narrow confidence interval
- An estimate is accurate if it is precise and close to the true value – small bias and standard error, small RMSE, narrow confidence interval around true value
- E.g. suppose a true target proportion is **T=0.4**
- Estimate = 0.5, confidence interval = (0.1, 0.9) [-----T------]
    - low precision and accuracy, but no evidence of bias
- Estimate = 0.5, Confidence interval (0.47, 0.53) T [---]
    - high precision, low accuracy (biased)
- Estimate = 0.42, Confidence interval = (0.39, 0.45) [-T-]
    - high precision, high accuracy (no evidence of bias)

Big data: precise but potentially inaccurate?

- Generally speaking, as sample size *n* increases:
- Precision increases, but bias stays constant (or may even increase)
- With small sample sizes, maximizing precision is important
- With large sample sizes, minimizing bias is important

Big Data

- “Big Data” – large data sets, often not collected for a specific research objective with a statistical design – e.g. internet data, administrative data
- Large implies high statistical precision, but high potential for bias (that is, estimates may be inaccurate – and get the wrong answer)

Two roles of randomization to avoid bias

- Random selection of participants
  - Ensures unbiased selection
  - Enhances external validity – inference for population based on sample
- Random allocation of treatments/factors
  - Ensures unbiased assignment
  - Enhances internal validity – valid treatment effect for individuals in the sample
  - Absent in observational studies
- Ideally we would like both, but this is very rarely achieved

Two roles of randomization to avoid bias

- Random selection of participants
  - Ensures unbiased selection
  - Enhances external validity – inference for population based on sample
- Random allocation of treatments/factors
  - Ensures unbiased assignment
  - Enhances internal validity – valid treatment effect for individuals in the sample
  - Absent in observational studies
- Ideally we would like both, but this is very rarely achieved

Properties of a good sampling scheme

- "representative" of the population (... whatever that means)
- demonstrably free of selection bias
- repeatable (at least in theory)
- efficient: lowest cost for given level of precision
- measurable precision: e.g., can quantify how close the sample mean is to the population mean it is estimating.
- Only probability (or random) sampling designs have these properties. Probability samples are characterized by the following two properties:
  - every sample has a known (maybe zero) probability of selection
  - every element (individual) in the population has a (known) positive probability of selection.

Probability sampling defined

- Probability samples are characterized by the following two properties:
  - every sample has a known (maybe zero) probability of selection
  - every element (individual) in the population has a (known) positive probability of selection
  - Examples to follow
- “scientific” sampling – allows statements about uncertainty for estimates of population quantities
  - In practice, frame errors and nonresponse can reduce effectiveness

Non-random sampling methods

- Examples of non-random sampling methods are:
  - Convenience sampling: Sample readily accessible individuals
  - Purposive or judgmental sampling (???)
  - Self-selected samples – e.g. open-access internet surveys
  - Quota sampling
  - Snowball sampling
- These methods are less scientific and less trustworthy than probability sampling, since they are subject to hidden biases.
- Note: “big data” are usually not based on random samples

Non-probability samples

- Examples include:
- Convenience sampling
- Quota sampling
- Open-access internet surveys
- Opt-in internet surveys
- Snowball sampling

Probability sampling and “big data”

Xiao-Li Meng’s (2018) “Law of Large Populations” (LLP) suggests that the value of probability sampling increases with the population size, and is surprisingly high:

“Estimates obtained from the Cooperative Congressional Election Study (CCES) of the 2016 US presidential election suggest a $\rho_{R,X} \approx -0.005$ for self-reporting to vote for Donald Trump. Because of LLP, this seemingly minuscule data defect correlation implies that the simple sample proportion of the self-reported voting preference for Trump from 1% of the US eligible voters, that is, **$n \approx 2{,}300{,}000$**, has the **same mean squared error** as the corresponding sample proportion from a **genuine simple random sample of size $n \approx 400$**, a 99.98% reduction of sample size (and hence our confidence).”

Simple Random Sampling

- The most familiar form of probability sampling (but there are others)
- With and without replacement
- Simple random sampling without replacement corresponds to selecting *n* balls out of a well-mixed urn containing *N* balls (like some lotteries). For this method:
  - All possible samples of size *n* have an equal probability of being selected.
  - All samples of size not equal to *n* have zero probability of selection
  - every individual has probability *n*/*N* of selection

SRS example

- Example. Suppose the urn contains $N = 5$ balls, labeled {A B C D E}; this is our population. We select a simple random sample of $n = 2$ balls. There are 10 possible samples of size 2, namely:
- AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
- Since all these samples have the same chance of being selected,
  - Pr(any size 2 sample selected) = 0.1
  - Pr(any other sample selected) = 0
  - Pr(any particular ball is included) = 0.4

Neyman’s famous paper

ON THE TWO DIFFERENT ASPECTS OF THE REPRESENTATIVE METHOD: THE METHOD OF STRATIFIED SAMPLING AND THE METHOD OF PURPOSIVE SELECTION.

By JERZY NEYMAN

(Biometric Laboratory, Nencki Institute, Soc. Sci. Lit. Varsoviensis, Warsaw).

[Read before the Royal Statistical Society, June 19th, 1934, the PRESIDENT, the RT. Hon. LORD MESTON of Agra and Dunottar, K.C.S.I., LL.D., in the Chair.]

Probability Sampling versus “Purposive Sampling”

- Initially, probability sampling was equated with its basic form, simple random sampling (SRS)
  - Every sample of size *n* has *equal* chance of being selected, hence an equal probability of selection method (*epsem*)
  - Samples of size other than *n* have no chance of being selected
  - With and without replacement

“Purposive Sampling”

- “Non-probability sampling” – but hard to define a negative.
- Units are picked so that sample matches distribution of a characteristic known for the population.
- E.g. if we know distribution of age and gender in population, choose sample cases to match this distribution.
- A common form is *quota sampling*: interviewers are given a quota for each age group and gender and interview individuals until this quota is met

The Controversy

- Under simple random sampling, distribution of a known characteristic in the sample can deviate considerably from its (known) distribution in the population, purely by chance
- This “lack of representativeness” led some to prefer purposively picking the sample to match the population distribution

Neyman’s “Resolution”

- Neyman (1934) showed that we can get the best of both worlds by stratified sampling:
  - Create strata by the classifying population according to the known characteristics
  - Select a simple random sample of known size n_j from population of size N_j in stratum j
- If f_j=n_j/N_j = const., results in epsem sample, retains probabilistic selection, and sample matches distribution of strata in population
- Also one can vary f_j and weight sample cases by 1/ f_j: Neyman’s optimal allocation

Footnote to Neyman’s paper

- Neyman’s paper is also famous for introducing (in English) the idea of confidence intervals– intervals with at least the nominal coverage in repeated samples.
- Ushered in the era of Neyman and Pearson significance testing
- Fisher notably fought with Neyman over this idea, calling it a “confidence trick”

More Complex Designs

- Neyman’s paper helped to set the stage for extensions to cluster sampling, multistage sampling, greatly extending the practical feasibility and utility of probability sampling in practice
- E.g. simple random sampling of people in the US is not feasible – we do not have a complete list of everyone in the population from which to sample
- Work of Mahalanobis, Hansen, Cochran, Kish, ….

Checking "Representativeness"

- One way of assessing representativeness is to compare distributions of known variables for the sample and the population
  - e.g. target population = U.S. Civilians
  - compare sample distribution of age, race, and sex with the population distribution from the nearest census.
  - should be done if possible, but are of limited value: really need to compare variables closely associated with the variables of interest

Summary

- Random sampling: a scientific way of achieving representativeness (on average)
- Big data that are not a random sample of the population may yield biased answers – proceed with caution

Next time

- Random selection of participants
  - Ensures unbiased selection
  - Enhances external validity – inference for population based on sample
- Random allocation of treatments/factors
  - Ensures unbiased assignment
  - Enhances internal validity – valid treatment effect for individuals in the sample
  - Absent in observational studies
- Ideally we would like both, but this is very rarely achieved

Next time

- Clinical trials for comparing treatments
- A little homework: read the (a) Cameron and Pauling and (b) Creagan et al. articles on Vitamin C as a treatment of advanced cancer, in the course site
- Why do they give some different conclusions? What are strengths and weaknesses
- We’ll be discussing these next time.