Where do the (big) data come from?

2. Statistical Inference, and Probability sampling

Rod Little

Department of Biostatistics

University of Michigan
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 $\mu$
  - SD $\sigma$

- Sample
  - Mean $\bar{x}$
  - 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 $H_0$ about the value or values 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 = \( \Pr(\text{"data"}| H_0) \)
"data" = "values of \( T \) at least as extreme as that observed".
Measures consistency of data with \( H_0 \)
P-Value is not \( \Pr(H_0 | \text{data}) \)
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 not a good measure of the size of the effect.
Problems with P-Values

• P-value is not the probability that the null hypothesis is true, \( p(H_0|\text{data}) \); it measures consistency of the data with the null hypothesis, \( p(\text{data}|H_0) \)

• 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 exactly 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

• 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 $n$
Ex: Confidence interval for population mean

• A confidence interval (CI) is a measure of uncertainty for a sample estimate (e.g. $\bar{x}$) of a population quantity (e.g. $\mu$)

• range of values computed from the sample within which the population quantity is likely to lie

• A 95% confidence interval for a population mean $\mu$ (if the sample size $n$ is large, say greater than 30) is

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

• 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
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
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

\[
\text{population quantity } \theta \quad \text{sample estimate } \hat{\theta}
\]

- Statistical measures of quality of the estimate include:

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

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

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

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

Where do data come from?
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)

Where do data come from?
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

Where do big data come from?
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

Where do big data come from?
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
  – Opt-in sampling for internet surveys
  – 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
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 sample size, and is surprisingly high:

Meng’s “data-defect correlation” $\rho_{R,X}$ is the correlation between inclusion in a sample (R) and a survey outcome (X). For random sampling $\rho_{R,X} = 0$, non-zero values indicate selection bias.

A sample of 2,300,000 self-reporting voters from the Cooperative Congressional Election Study (CCES) of the 2016 US presidential election had a miniscule $\rho_{R,X} \approx -0.005$ for voting for Donald Trump. This (seemingly minuscule) data-defect correlation implies that for proportion voting for Trump, this sample of $n \approx 2,300,000$, has the same mean squared error as the sample proportion from a genuine simple random sample of size $n(srs) \approx ???$
n = 2.3 million from CCES has equivalent RMSE to a simple random sample size n(srs) of

230,000
23,000
4,000
400
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

Where do big data come from?
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,
  - $\Pr(\text{any size 2 sample selected}) = 0.1$
  - $\Pr(\text{any other sample selected}) = 0$
  - $\Pr(\text{any particular ball is included}) = 0.4$
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.

Where do big data come from?
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 $f_j = n_j/N_j = \text{const.}$, 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
Tomorrow’s topic

• 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
For Wednesday

• Clinical trials for comparing treatments
• A little homework before tomorrow: 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