Index of 12-L

Practical Information

Schedule:

- fourth home assignment due today Thursday,
- call for requested topics for final (review) lecture,
- also in last lecture: information about the exam.
- course syllabus for final exam posted at webpage.

Today's lecture

- discussion of issues around statistical inference and in particular statistical testing: strongly recommended to consult the resources at the Media page,
- introduction to sample size calculations (Supplementary notes, Section 1),
  - based on precision, e.g. margin of error for CIs,[1]
  - based on power of a statistical test,[2]
  - both available for many of the simple models/design covered in the course,
  - also include a special situation of our interest: detection of disease (or freedom of disease).

[1] PSLS Chapters 15 & 19; S Sections 7.2-7.3; IPS Sections 6.1 & 8.1.
[2] PSLS Chapter 15; S Section 8.1; IPS Sections 6.4 and 7.1-7.2.

Should *P*-Values be Abolished?

An Internet search on this/related search terms will give a good impression of the strength of the push to eliminate statistical testing from science...
— What is happening?

The origin is in misuse of statistics by its users (to which you will soon belong!).[3] The study and reporting guidelines (Session 8) are part of an attempt to improve the conduct and reporting of scientific studies, including the statistics.

Why statistical testing?

- "since the precise meaning of p-value is hard to grasp, misuse is widespread ...",[4]
- alternatives (Bayesian approach, data science) do not have statistical testing,
- statistical testing tempts users into (false) black-white interpretations, which are then critisized and blamed on the methodology.[5]

Current stand[6]: Statistical testing is often given too much attention in applied data analysis, where one should instead focus on the estimates, their precision (e.g. by confidence intervals) and their interpretation, plus many other decisions prior to the final results.

[3] Inappropriate statistical analysis has historically been well-documented in particular in the health sciences, but due to the increased focus in that area it is nowadays a bigger problem in other disciplines.

[4] Wikipedia page on p-value; as of 25/11/2020.

[5] My personal view, and to some extent agreeing with a statement by Andrew Gelman (one of the big names in statistics): "People want something that they can't really get. They want certainty."; *Nature* 531, 151.

[6] 2019 recommendations (see homepage media links): Don't say "Statistically significant". Accept uncertainty. Be thoughtful, open and modest.

Pitfalls of Statistical Testing

Some points to remember when using statistical tests:[7]

- the test/P-value is only as good as its assumptions. . . ,
- strictly speaking, the theory of statistical tests is for hypotheses determined in advance of data collection, and certainly not inspired by the data. . . ,
- carrying out many tests on the same data may by pure chance cause some of them to be statistically significant (the multiple testing and/or data dredging problems),[8]
- to consider an analysis with $P = 0.049$ a success and discard an analysis when $P = 0.051$, is ridiculous. . . (do not put too strong emphasis on significance at a specific level, and always report P-values instead of just significance yes/no),
- non-significance does not mean proof of no effect:
  — absence of evidence is not evidence of absence![9]
- non-significance may be important in itself (when not caused by insufficient or otherwise poor data),
- statistical significance does not imply practical significance/importance, nor causation.

See also the six principles of ASA's Statement on Statistical Significance and P-Values.

[7] Based on PSLS, Chapter 15, and IPS, Section 6.3.
[8] Computing, say, 10 tests, each with 5% error rate ⇒ simultaneous error rate between 5% and $10 \times 5\% = 50\%$.
[9] Quote usually attributed to Carl Sagan; or to William Cowper (1731-1800).

Misconceptions in Statistical Analysis (in particular, Testing)[10]

1. The P-value is the probability that the test hypothesis is true; ...[11]
2. The P-value for the null hypothesis is the probability that chance alone produced the observed association; ...[12]
3. A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.
4. A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted.
5. A large P-value is evidence in favor of the test hypothesis.
6. A null-hypothesis P-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.
7. Statistical significance indicates a scientifically or substantively important relation has been detected.
8. Lack of statistical significance indicates that the effect size is small.

19. The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size.
20. An effect size outside the 95% confidence interval has been refuted (or excluded) by the data.
21. If two confidence intervals overlap, the difference between two estimates or studies is not significant.
22. An observed 95% confidence interval predicts that 95% of the estimates from future studies will fall inside the observed interval.

[10] Based on Greenland et al. (2016), *European Journal of Epidemiology* 31, 337−350.

[11] ...for example, if a test of the null hypothesis gave $P = 0.01$, the null hypothesis has only a 1% chance of being true; if instead it gave $P = 0.40$, the null hypothesis has a 40% chance of being true.

[12] ...for example, if the P-value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.

Determination of Sample Size

Question: how many subjects to take??

- crucial part of study planning: do not start collecting samples without any plan,
- crucial part of any grant proposal submitted for funding, because of its implication for costs and ethics:
  - most funding agencies require a formal justification of the proposed sample size, e.g. by a statistical sample size calculation

  — leading to the totally unrealistic expectation that some "statistical magic" can deliver a definitive number for sample size,
- in practice, main determinants of sample size are logistics and cost (not statistics!).

Plain common sense considerations for choosing a "good" sample size:

- size should be sufficient to detect (statistical significance of) treatment differences of interest,
- avoid "waste" of experimental units,
  - unethical to unnecessarily[13] use individuals (animals, humans) for a study,
  - keep costs as low as possible,
- having replications will reduce the sensitivity to errors.

[13] Individuals would be unnecessary if they no longer contributed any important information, because the conclusion was already clear from a smaller sample.

Statistical Methods to Choose Sample Size

Fact: all formal procedures require pre-decided statistical model involving choices of:

- targeted outcome and scale for its analysis,
- targeted parameters and/or hypotheses of interest,
- assumptions involved in model/analysis,[14]

as well as detailed prior knowledge (estimates or guesses) about the outcomes:

- size of effect of interest, or desired precision for targeted estimate,
- standard deviation of observations (for normal distribution models),

— usually, all this information is not readily available.[15]

Two general statistical approaches for determining sample size:[16]

(1) from desired precision (standard error, size of 95% CI) on selected estimate,

(2) from desired power of test for effect of interest, using always statistical software (e.g., Minitab/Stata/R) or web applications (avoiding hand calculations[17]).

Additionally, specific methods exist for many specialized settings and study types.

[14] Most standard sample size calculations for quantitative outcomes are based on normal distribution assumptions.
[15] Some possible sources of information: *i)* published articles on same research question, *ii)* a pilot study conducted prior to the main study, and *iii)* expert opinion solicited from subject-matter experts.
[16] Approach (2) is far more common in practice, arguably too common; Bland (2009), *BMJ* 339, 1133-1135.
[17] The formulas of IPS and VER, and also Lehr's formula, are not recommended.

Exercise 6.19

Impact of sample size n for study of reading ability of 3rd grade children. Preliminary study has given $s = 12$, so that $\sigma = 12$ is assumed. We also assume an approximate normal distribution for the sample mean $\bar{X}$.

(a) $n = 100$, margin of error for 95% CI (with known σ):

$$z^* \, \sigma/\sqrt{n} = z_{.975} \, \sigma/\sqrt{n} = 1.96 \times 12/\sqrt{100} = 2.35,$$

or more realistically with σ unknown:

$$t^* \, \sigma/\sqrt{n} = t_{.975}(99) \, \sigma/\sqrt{n} = 1.984 \times 12/\sqrt{100} = 2.38,$$

(b) $n = 10$, margin of error for 95% CI:

$$\begin{aligned} z^* \, \sigma/\sqrt{n} \;\; &= 1.96 \times 12/\sqrt{10} = 7.44, \qquad \text{with known } \sigma, \\ t^* \, \sigma/\sqrt{n} = t_{.975}(9) \, \sigma/\sqrt{n} \;\; &= 2.262 \times 12/\sqrt{10} = 8.58, \qquad \text{with unknown } \sigma. \end{aligned}$$

(c) appropriate n for a desired margin of error of $m = 3$ must be between 10 and 100 — we can work our way through trial and error, or use the formula on the next page, for known σ:

$$n = \left(\frac{z^* \, \sigma}{m} \right)^2 = \left(\frac{1.96 \times 12}{3} \right)^2 = 7.84^2 = 61.5,$$

take $n = 62$ to ensure that margin of error ≤ 3.[18]

[18] For unknown σ, we should repeat the calculation with $t^* = t_{.975}(61) = 2.00$, to see whether changing from z^* to t^* affects the required n substantially: it gives $n = 64$.

Controlling Size of Confidence Intervals

How to decrease size of confidence intervals for population means? — based on the formula for 1-sample means and assuming a known σ,

$$\text{margin of error } = z^* \times \sigma/\sqrt{n}.$$

- increase n (more data),
- increase α to decrease $z^* = z_{1-\alpha/2}$, same as decrease $C = 1-\alpha$ (lower certainty/ confidence of interval),
- decrease σ (reduce variation in population, by shifting to another variable or another population, or by variance-reduction techniques such as blocking).

How to adjust sample size to get a desired margin of error (m) for a population mean?

- fix m, α and σ at suitable values,
- solve for n in above formula for the margin of error:

$$m \geq \frac{z^* \times \sigma}{\sqrt{n}} \quad \text{or} \quad n \geq \left(\frac{z^* \times \sigma}{m}\right)^2.$$

- formula guarantees margin of error $\leq m$, provided assumptions for confidence interval met.

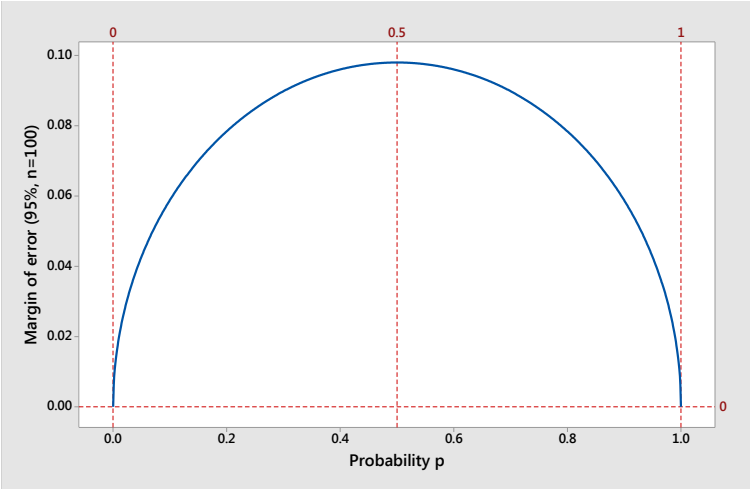
If σ cannot be assumed known (most realistically), the calculation should be redone with $t^* = t_{.975}(n-1)$ to assess any changes in the required n.

Controlling Margin of Error for a Single Proportion

Margin of error: (based on the normal approximation[19])

$$\text{margin of error} = z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad z^* = z_{1-\alpha/2},$$

- ways to reduce margin of error: increase n or α,
- margin of error largest at $p = 0.5$ (see graph).



How to adjust sample size to get a desired margin of error (m) for a proportion?

- fix m, α and p (guessed) at suitable values,
- solve for n in above formula for margin of error:

$$m \geq z^* \times \sqrt{\frac{p(1-p)}{n}} \quad \text{or} \quad n \geq p(1-p)(z^*/m)^2,$$

- formula guarantees margin of error $\leq m$, provided assumptions for confidence interval met,
- using $p = 0.5$ is conservative ($\Rightarrow$ maybe (much) too large n).

[19] The Minitab menu (Sample Size for Estimation) uses a more exact calculation.

Sample Size based on Estimation Precision

General approach for normal distribution models:

- assume estimated/guessed/known standard deviation σ,
- assume (mean) parameter of interest μ and estimate $\hat{\mu}$, with standard error $\text{SE}(\hat{\mu}) = \sigma \times c(n)$, where $c(n)$ is a known constant (depending on number of obs. n),
- approximate 95% CI[20]: $\hat{\mu} \pm 2\,\sigma\,c(n)$,

Compute n to achieve desired margin of error (M) by solving for n in the equation:

$$M(\text{desired value}) \geq 2\,\sigma\,c(n).$$

Example: blood pressure measured on patients before and after an intervention,

- design: two paired samples, use the differences D = after − before,
- model: i.i.d. sample of size n from $\text{N}(\mu, \sigma)$, with guessed value $\sigma = 10$ (mm Hg),
- μ = population mean, $\hat{\mu} = \bar{D}$ (sample mean), $c(n) = 1/\sqrt{n}$,
- assume, the desired margin of error of CI for μ is $M = 3$,[21]
- solve: $3 \geq 2 \times 10/\sqrt{n} \Rightarrow n \geq (2 \times 10/3)^2 = 44.4 \approx 45$,
- conclusion: with $n = 45$ (Minitab: 46) patients, a 95% CI for the difference would have a margin of error of 3 (if the sample standard deviation equalled σ)[22].

[20] Approximation requires either σ known, or σ unknown and n so large that $t^* = t_{.975}(\text{df}) \approx z^* = z_{.975} \approx 2$, say $n \geq 40$.
[21] If $M = 3$, then an observed sample mean of 3 would be (just) significant at the 5% level.
[22] Some software (e.g. Stata) allows to also include variability in the estimate s, further complicating the procedure.

Sampling to Detect Disease

Question of interest: does a (rare) disease exist in a population of individuals (say animals)?

- in order to know for certain, we need to test until a positive animal is found (or all animals have been tested),
- in classical statistics, we cannot make probability statements about the proportion p of diseased animals in the population, because it is a parameter (and constant),
- but we can compute a $(1-\alpha)$ confidence interval for p, based on having observed a number n of (all negative) animals $\Rightarrow$ a desired upper bound for the (one-sided) CI can be converted into a required sample size.

Two basic situations for the sampling ($\sim$ binomial/hypergeometric distributions; 4L–10):

- sampling with replacement, or very large population: $n = \ln(\alpha)/\ln(1-p_{\text{min}})$,
- sampling without replacement from a finite population of size N with at most D infected animals (where $p_{\text{min}} = D/N$): $n = (1-\alpha^{1/D}) \times (N-(D-1)/2)$,

— both allow the statement that $p \leq p_{\text{min}}$ with confidence $(1-\alpha)$ after observing n negative animals.

Some extensions: imperfect tests $\rightarrow$ software calculation[23]; prior distribution for $p \rightarrow$ Bayesian methods; also the larger area freedom of disease modelling.

[23] One of the standard calculators is Freecalc, `https://epitools.ausvet.com.au/freecalctwo`.

Errors of Type I–II and Power

Errors of type I and II:

- type I error: to reject H_0, when H_0 in reality true,[24]
- type II error: to not reject H_0, when H_0 in reality false,
- power of statistical tests involves type II errors (below),
- schematic:

Conclusion	Truth about population	
from sample	H_0 true	H_a true (H_0 false)
reject H_0	type I error	no error
not reject H_0	no error	type II error

Power of a statistical test:

- involves a specific alternative, e.g. H_a: $\mu = 0.84$ in the laboratory analysis example (5L–13/17) with H_0: $\mu = 0.86$,
- definition: power = probability that the statistical test will reject H_0, when the specific alternative H_a is true = 1− type II error,
- important for planning of experiments: what chance of a significant result?
- difficult to calculate in complex models (lots of formulas and software exist).

[24] The definition of statistical tests involves only type I errors, which are controlled by the significance level (α).

Sample Size based on Power

Requirements for sample size determination based on power:

- statistical model / design and corresponding software,
- size of effect[25] desired to be detected,
- standard deviation of model (for normal distribution data/models),
- desired value of power (0.8, or 80%, commonly used),
- significance level α of test employed (usually $\alpha = 0.05$, or 5%),
- statistical software offering sample size calculations for the specific setup.

Blood pressure example continued:

- same model and setup as before (known $\sigma = 10$); assume interest in a true population mean (difference) of 3 units,
- computation of power for $n = 45$ and significance level 0.05:
  - two-sided alternative: power = 0.52 (unknown σ: 0.50),
  - one-sided alternative: power = 0.64 (unknown σ: 0.63),
- computation of necessary sample size to achieve power = 0.8 at $\alpha = 0.05$:
  - two-sided alternative: required $n = 88$ (unknown σ: 90),
  - one-sided alternative: required $n = 69$ (unknown σ: 71).

[25] In this context (not generally), effect size is the difference between H_0 and H_a values.

Sample Size for ANOVA

Based on desired precision (e.g., for the size of CI):

- CI could be for group mean, a difference between group means or a more general contrast,
- requires known (guessed) σ and desired margin of error (m),
- general approach based on approximate 95% CI for a mean (or contrast) parameter μ, based on its estimate $\hat{\mu}$:

$$\hat{\mu} \pm 2\,\mathrm{SE}(\hat{\mu})$$

— determine $\mathrm{SE}(\hat{\mu})$ as a function of n, and solve with respect to n the equation:

$$m \geq 2\,\mathrm{SE}(\hat{\mu}).$$

Based on power for F-test:

- requires known (guessed) σ and group means μ_i, plus significance level,
- calculation available for one-way ANOVA in Minitab/Stata/R, and more generally on websites. [26] [27]

[26] One recommended website for sample size calculations is at: `http://homepage.stat.uiowa.edu/~rlenth/Power/`.
[27] An updated, comprehensive and diverse list of links to statistical websites, including for sample size calculations, is at: `http://statpages.org`.

Sample Size Complexities

How to deal with more complex situations (in practice) than the simplest designs?

- more complex situations require more information to be modelled/taken into account . . . which may not be available (e.g. in multifactorial studies), and may involve extra assumptions that are not easily understood and/or difficult to assess,
- in practice, it is often preferable to restrict to a simplified setup that incorporates the primary objective(s) of the study/data,[28]
- sample size requirements from separate outcomes/analyses are usually conflicting, so need to be balanced with regard to study objectives.

Adjustments for specific study/data features (selection, others exist):

- finite population correction: less samples are required when sampling from finite populations, and an (ad-hoc) correction factor exists[29],
- clustering: if animals are sampled from multiple herds, animals from the same herd can usually not be assumed independent, which in some instances leads to larger required sample sizes, and an (ad-hoc) correction factor exists[30],
- sample size for reference intervals $\rightarrow$ specialized methods/literature.

[28] For example, maybe restrict to two (primary) groups even if multiple groups are planned, or maybe restrict to a single (primary) time point among repeated measures.

[29] If n is the infinite population sample size and N is the population size, use instead $n^* = 1/[(1/n) + (1/N)]$.

[30] If n is the infinite population sample size, m is the number of animals per herd and ρ is the within-herd correlation, use instead $n^* = n[1 + \rho(m-1)]$.

Sample Size Misconceptions

Common misconceptions[31] in sample size calculations:

- use of standard effect sizes (general definitions of "small", "medium" and "large" effects, relative to std. dev.): effects of interest should be determined exclusively from the context of your study,
- retrospective power calculation: after a study has been carried out and using its estimated values:
  - power/sample size calculations aid in planning of new studies, not in interpreting results of data analysis,
  - confidence intervals give the best information about the unknown parameters from a study,
  - if H_0 was not rejected, the conclusion may sometimes be strengthened by an equivalence test (instead of arguing from the study's power), see next slide.

Extra (technical) misconceptions from the Greenland (2006) paper:

24. If you accept the null hypothesis because the null *P*-value exceeds 0.05 and the power of your test is 90%, the chance you are in error (the chance that your finding is a false negative) is 10%.
25. If the null P value exceeds 0.05 and the power of this test is 90% at an alternative, the results support the null over the alternative.

[31] Largely based on Lenth (2001), *The American Statistician* 55, 187−193; also discussed in Greenland et al. (2016).

Equivalence Testing

An equivalence test[32] is for making a statement that effects of two "treatments" (say) differ at most by a small amount (say δ),[33]

- for $H_0 : \theta = 0$ (θ being a difference in means, or other parameters), not rejecting the H_0 is a weak and non-quantitative conclusion,
- a CI for the difference θ contains useful information,
- a non-equivalence hypothesis $H_0{}^{(ne)} : |\theta| \geq \delta$ can be tested against the $H_a{}^{(ne)} : |\theta| < \delta$, as follows (at a 5% significance level):
  - compute a 90% CI (not 95% CI) for θ,
  - reject $H_0{}^{(ne)}$, if the interval $(-\delta, \delta)$ entirely includes the CI.

Example: Radon detectors (Exercises 6.95/7.64),

- Data: $X_1, \ldots, X_{12} \sim N(\mu, \sigma)$; $n = 12$, $\bar{X} = 104.13$, $s = 9.40$, targeted mean $= 105$,
- deviation: $\theta = \mu - 105$, say we set $\delta = 3 \sim$ acceptable deviation from target,
- 90% CI for θ: $\bar{X} - 105 \pm t^* \, s/\sqrt{n} = 104.1 - 105 \pm 1.796 \cdot 9.4/\sqrt{12} = (-5.74, 4.00)$,
- Conclusion: 90% CI is not inside equivalence interval $(-3, 3) \Rightarrow$ cannot claim equivalence (because the deviation to the target value could be larger than 3).

[32] A non-inferiority test is a similar construct, only with both hypotheses expressed in terms of θ instead of $|\theta|$, so as to involve only one direction for the comparison, e.g. that one treatment is at most δ worse the other, or better.

[33] One difficulty with the setup is the choice of δ, which should be determined entirely from the study context, e.g. as the difference between treatments that would make a biological/clinical difference.

Summary Notes

Statistical testing has come under strong criticism, in part due to numerous and consistent errors in its application (use, reporting and interpretation) $\Rightarrow$ care is needed with the use of statistical tests to avoid damaging comments by readers/reviewers.

Statistical sample size calculation – two main approaches:

- based on estimation precision: determine sample size to achieve a desired precision for an estimate of interest,
  - requirements: desired precision (e.g. margin of error for CI),
  - implementation: hand calculation formulas (+ Minitab), typically derived from the SE of the estimate of interest (formulas exist for standard settings),
- based on power of statistical tests:
  - test H_0 against one- or two-sided H_a (standard setup),
  - type I and type II error of statistical testing,
  - power against specific alternative hypothesis H_a,
  - requirements: targeted effect (e.g. mean difference) to detect, desired power level, test settings (incl. significance level, type of H_a),
  - implementation: statistical software/web applications.
- both approaches also require a standard deviation for quantitative (normal distribution) data.