Measures of uncertainty, and the P-Value controversy

Roderick Little

UNIVERSITY OF MICHIGAN

Outline

- Widespread concerns about scientific replicability
- Perception that misunderstandings and misuses of hypothesis testing, P-values, contribute to this problem
- American Statistical Association (ASA) "Statement on Statistical Significance and P-Values"
- Review these issues, and discuss alternative approaches for conveying statistical uncertainty– p-values, confidence intervals, Bayesian 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.

Sample		Population	
Mean	$\bar{x}$	Mean	μ
SD	s	SD	σ

- Inference crucially requires that sample is "representative" (e.g. randomly selected) from population (or an assumption that it is)
- Statistical inferences are subject to uncertainty – quantifying uncertainty is an important objective

Tools for assessing uncertainty

- **Hypothesis Testing: basic tool is P-value**
  - **P-value** = Pr("data"|null hypothesis). A low value (e.g. P < 0.05) is interpreted as evidence against the null hypothesis
- **Interval Estimation: basic tool is the Confidence interval** – random interval that includes the true value of a parameter in a given proportion of repeated samples (e.g. 95%)
- **Bayesian methods: basic tool is the Posterior Distribution**
  - More on this later

Hypothesis testing

- Assesses consistency of the data with a particular null value of the parameter
- For example, for inference about a mean
  - Confidence interval: set of values of the mean consistent with the data
  - Hypothesis test: are the data consistent with a particular value of the mean?
- Often the null value corresponds to “no difference” or “no association”

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)

The misinterpretation of p-values: Experiment in McShane and Gall (2017 JASA)

"The study aimed to test how different interventions might affect terminal cancer patients' survival. Subjects were randomly assigned to one of two groups. Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure.

Subjects were then tracked until all had died. **Subjects in Group A lived, on average, 8.2 months** post-diagnosis whereas **subjects in Group B lived, on average, 7.5 months** post-diagnosis **(p = 0.01).** Which statement is the most accurate summary of the results?

McShane and Gill (2017 JASA)

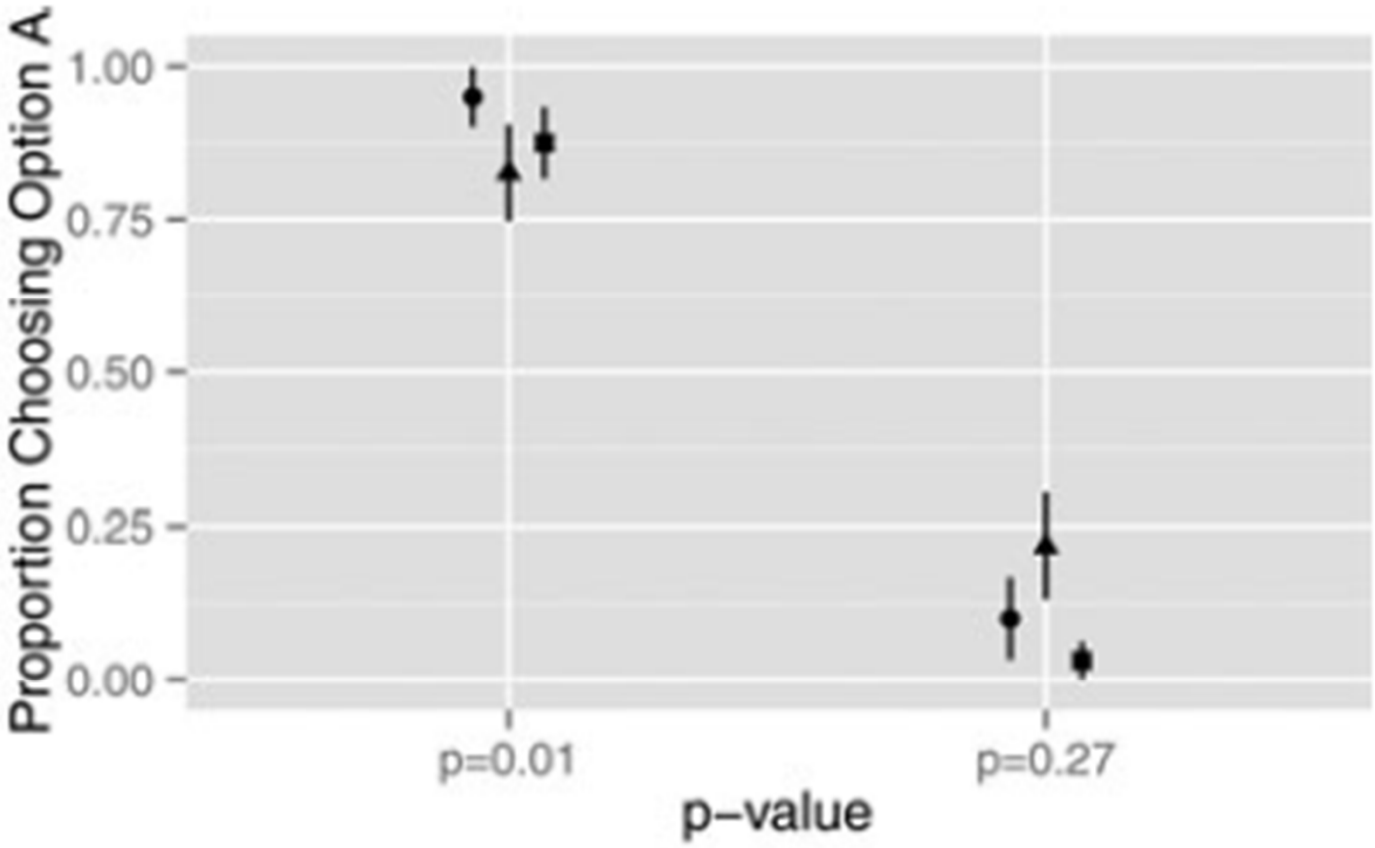
Speaking only of the subjects who took part in this particular study:

A. the average number of post-diagnosis months lived by the subjects who were in Group A was *greater* than that lived by the subjects who were in Group B.
B. the average number of post-diagnosis months lived by the subjects who were in Group A was *less* than that lived by the subjects who were in Group B.
C. The average number of post-diagnosis months lived by the subjects who were in Group A was *no different* than that lived by the subjects who were in Group B.
D. It *cannot be determined* whether the average number of post-diagnosis months lived by the subjects who were in Group A was greater/no different/less than that lived by the subjects who were in Group B.

McShane and Gill (2017 JASA)

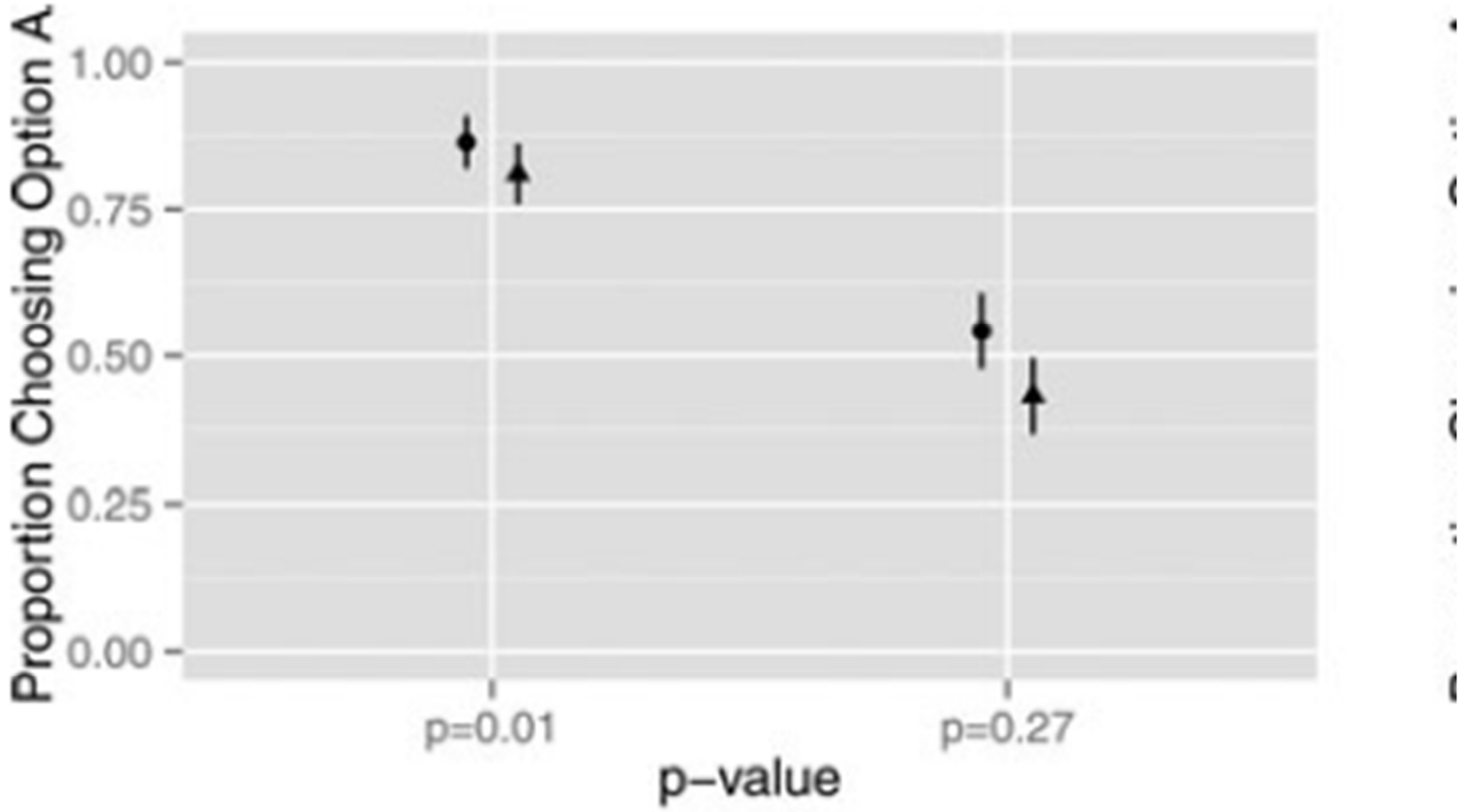
After seeing this question, each subject was asked the same question again but $p = 0.01$ was switched to $p = 0.27$ (or vice versa for the subjects in the condition that presented the $p = 0.27$ version of the question first)"

Proportion choosing A (correct answer): NEJM readers



(b) *NEJM*

Proportion Choosing A: JASA readers



(a) *JASA*

P-Values

P-values can indicate how incompatible the data are with a specified statistical model.

P–values are not:

(a) The probability that the null hypothesis is true

(b) Good measures of the size of an effect:

Smaller deviations from the null can be detected with larger sample sizes, so the P-Value is strongly dependent on sample size

Significance level

- A classical significance test sets a cut off value α, and formally "rejects" the null hypothesis if P-value < α, "accepts" the null hypothesis if P-value > α
- The cut-off α is called the "significance level", "size" or "type 1 error" of the test, and has the property that

$$\Pr(\text{reject Null} | \text{Null true}) = \alpha$$

- The choice of significance level α is somewhat arbitrary; a typical value by convention is 0.05 (but more on this below).
- P = 0.049 is not substantively different from P=0.051, but one "rejects" and the other "accepts" at the 5% level.
- So I think it is better to avoid a cut-off and just report the P-value

Redefining significance

Comparisons with Bayesian hypothesis testing by my ex-colleague Val Johnson suggest that the common "P<.05" significance level is weak evidence against the null, contributing to the lack of replicability of results

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!"

Redefining significance

- A recent 74-author (!) paper (Benjamin…V. Johnson. Redefine Statistical Significance. 2017 *Nature Human Behavior*) argues for changing the threshold from 0.5 to .005, based on comparing P-values with Bayes Factors for a simple null

Let D = data, H = hypothesis.

Bayes' rule converts $\Pr(D|H)$ into $\Pr(H|D)$, and is a simple consequence of basic rules of probability:

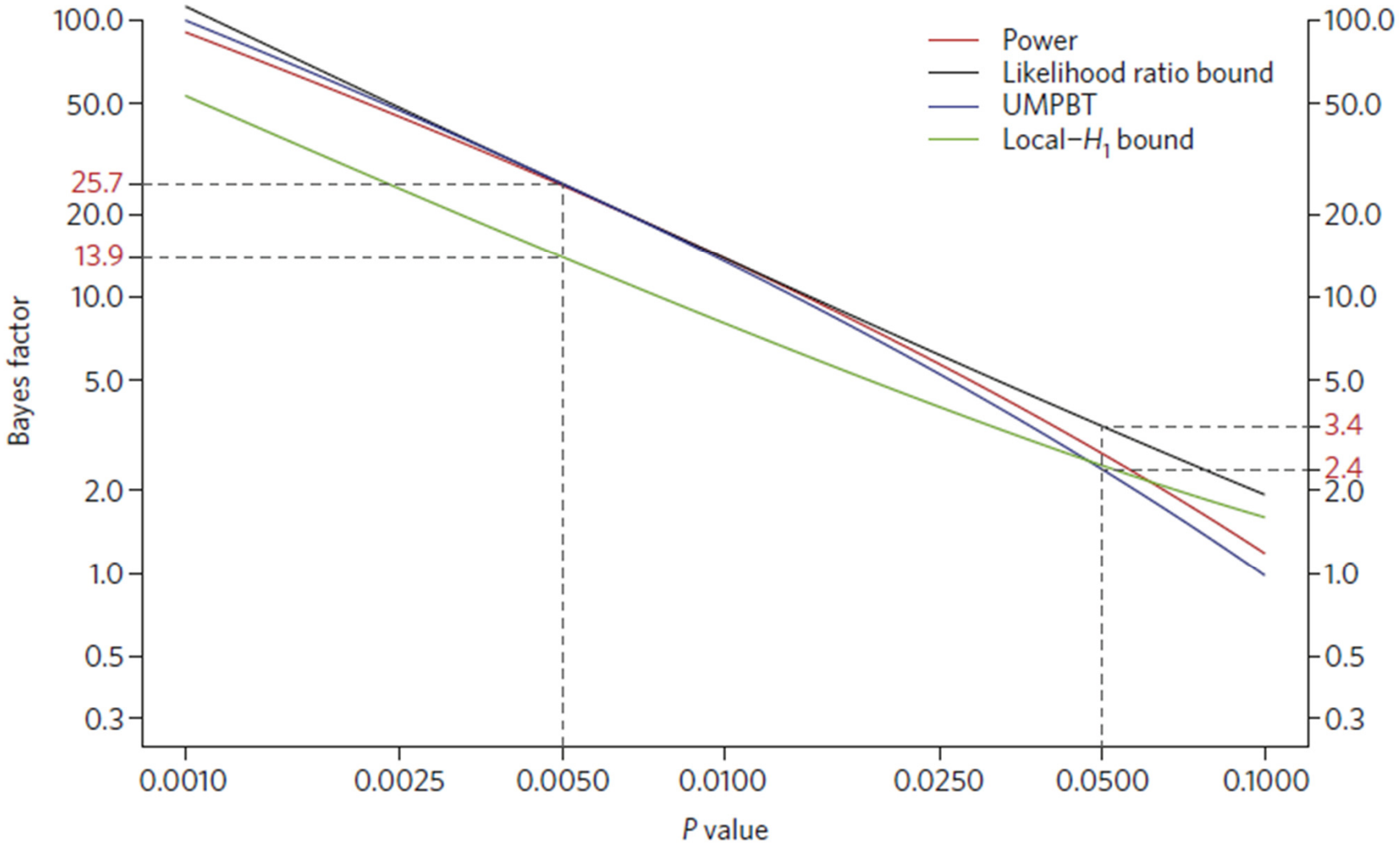
$$\Pr(H, D) = \Pr(D) \times \Pr(H \mid D) = \Pr(H) \times \Pr(D \mid H)$$

$$\Pr(H \mid D) = \Pr(H) \times \Pr(D \mid H) / \Pr(D)$$

$$\text{Hence,} \quad \frac{\Pr(H \mid D)}{\Pr(H' \mid D)} = \frac{\Pr(H)}{\Pr(H')} \times \frac{\Pr(D \mid H)}{\Pr(D \mid H')}$$

That is, posterior odds = prior odds × Bayes factor

Strength of evidence against null



More on significance level

- *Regardless of the threshold, it is a bad idea to publish only statistically significant results,* since this leads to *publication bias*
  - for interpretation, we need to know about negative studies too!
  - journals should report results from methodologically sound studies that address important questions, whether or not results are significant

From ASA P-Value Statement

- *"P*-values can indicate how incompatible the data are with a specified statistical model.
- *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
- Proper inference requires full reporting and transparency
- A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis."

Full reporting and transparency

- Bad practice: Carry out many statistical tests and only report significant ones. Transparency here is to report all the tests carried out, whether or not significant.
- 20 independent tests: one will be significant even at 5% level even if all effects are null
- Question is whether interest is in controlling type 1 error of each individual test, or over all the tests in the experiment.
- If latter, one simple (if crude) approach is the Bonferroni correction: divide the significance level by number of tests made; e.g. if 10 tests and sig level .05, test at .05/10 = .005 level
- Related: in genetics with many genes tested, significance level is chosen to be very low.

P-Value is not the effect size

- P-value is poor measure of the size of an effect –
  - size of P-value has no clinical meaning
  - mixes estimate of effect and its uncertainty
  - strongly determined by sample size – since nothing is exactly zero, anything is significant with a large enough data … and we are entering the era of big data!
  - One-sided or two sided alternative – not always clear
  - 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."

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

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

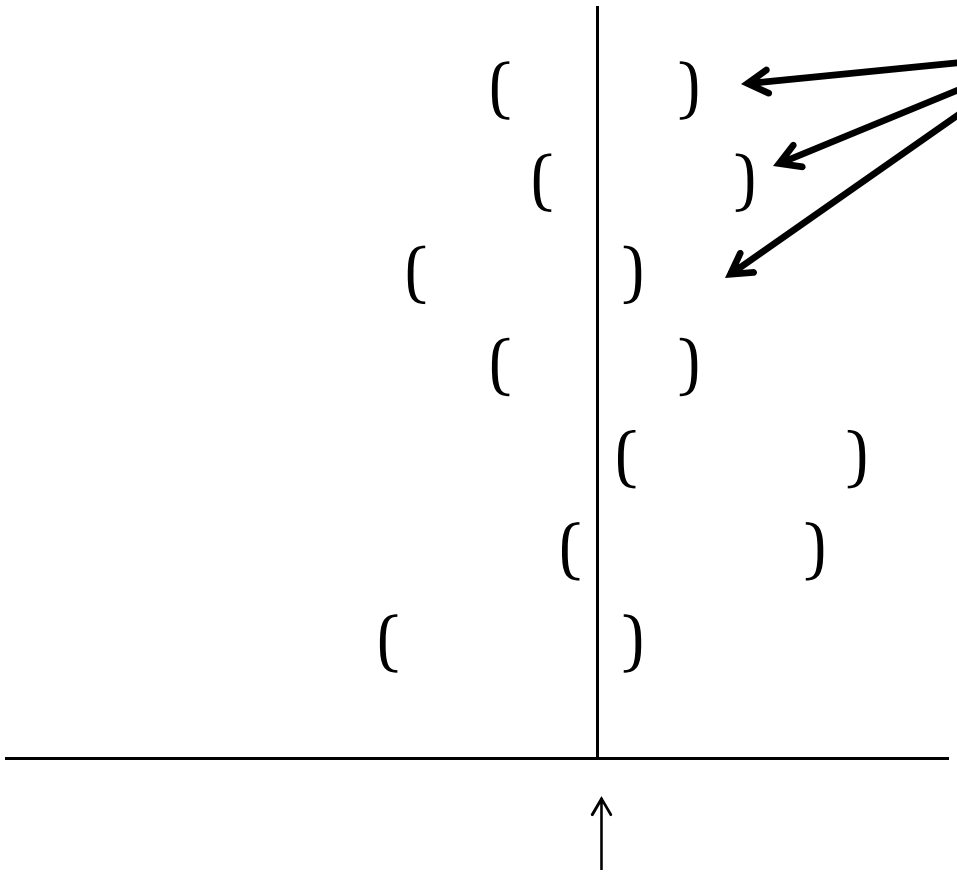
$$\bar{x} \pm t_{.975} s / \sqrt{n}$$

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

Confidence Intervals



Hypothetical repeated samples

.
.
.

Confidence interval property: (at least) 95% of these random intervals include the true value

Unknown true value of parameter θ

Confidence Intervals: better for inference than P-values

- Estimate has clinical meaning – closer to the science. Good measurement is the heart of statistics
- Width of interval captures uncertainty
- Confidence interval summarizes the evidence in a natural way

Study A: small trial

	Success	Failure
Treatment 1	10 (50%)	10 (50%)
Treatment 2	15 (75%)	5 (25%)

- Null Hypothesis H_0: Outcome independent of treatment, or treatments equally effective
- Chi-squared test of equality of proportions: $P = 0.102$
- P = Pr(Tables with treatment differences as or more extreme than that observed | H_0)
- Conclusion: "accept" H_0 at 5% level

Study B: large trial

	Success	Failure
Treatment 1	500 (50%)	500 (50%)
Treatment 2	550 (55%)	450 (45%)

- Null Hypothesis H_0: Outcome independent of treatment, or treatments equally effective
- Test of equality of proportions: $P = 0.025$
- Conclusion: Reject H_0 at 5% level

Examples

- Study A: 95% CI for Diff = (-4.6%, 54.6%)

  Wide, consistent with no difference, but large differences also possible

P-Value = .102. Not significant (NS), but doesn't mean there is no effect – NS does not mean null hypothesis is true!

- Study B: 95% CI for Diff = (0.6%, 9.4%)

  Narrow, not consistent with no difference, but large difference is unlikely

P-Value = .025. Statistically significant, but evidence is that effect is not clinically significant!

Can warfarin be continued during dental extraction? Results of a randomized controlled trial

- I. L. Evans, M. S. Sayers, A. J. Gibbons, G. Price, H. Snooks, A. W. Sugar. *Brit. J. Oral & Maxillofacial Surgery* (2002) **40**, 248–252

- *SUMMARY*. **A randomized controlled trial was set up to investigate whether patients who were taking warfarin … require cessation of their anticoagulation drugs before dental extractions.**
- **Of 109 patients who completed the trial, 52 were allocated to the control group (warfarin stopped 2 days before extraction) and 57 patients were allocated to the intervention group (warfarin continued).**
- **The incidence of bleeding complications in the intervention group was higher (15/57, 26%) than in the control group (7/52, 14%)**
- **but this difference was not significant… we found no evidence of an increase in clinically important bleeding. As there are risks associated with stopping warfarin, the practice of routinely discontinuing it before dental extractions should be reconsidered.**

Clinical vs statistical significance

- "Incidence of bleeding complications in the intervention group was higher (15/57, 26%) than in the control group (7/52, 14%) but this difference was not significant ...we found no evidence of an increase in clinically important bleeding."
  - Is 26% vs 14% *clinically* significant? 95% confidence interval for difference in proportions = (0, 0.28)
  - Study seems underpowered (sample size too small)
  - a common problem in clinical trials

Some objections to CIs

- Confidence intervals are peculiar objects: the interval is random, but the parameter is fixed
- For some basic problems there is no CI procedure that gives exactly the nominal coverage
  - Behrens-Fisher problem: comparing means of two normal samples with unknown means and variances, not assumed to be equal.
- Basing inference on sampling distribution violates the likelihood principle – experiments leading to the same likelihood function should have the same inference

A related problem with CIs

- What should be included in the set of hypothetical repetitions -- the *reference set* -- is not always clear
  - and different choices give different confidence intervals

Example: Independence in 2x2 Contingency Table

Treatment	Outcome S	Outcome F
A	170	2
B	162	9

$$H_0 : \pi_A = \pi_B;\ \ H_a : \pi_A > \pi_B$$

Alternative tests

Pearson chi-squared (C)	P=0.016
Yates continuity corrected (Y)	P=0.032
Fisher exact test (F)	P=0.030
Bayes $\Pr(\pi_A < \pi_B \mid data)$	Pr=0.013

Independence in 2x2 tables

- Choice of test doesn't matter in large samples, but it does in small/moderate samples
- Fisher test is conservative when one margin is fixed in repeated sampling (as is common in many practical designs), but exact if both margins are fixed
- Should the reference set condition on second margin or not? It's debatable (Yates 1984, Little 1989)
- Frequentist theory is ambiguous, and frequentists disagree about which is the right test

A CI is not a probability interval

Most people interpret a confidence interval as a probability interval: a fixed interval that includes the unknown parameter with 95% probability. That is, the interval is fixed, the parameter is random. Unfortunately, confidence intervals have some properties that are in conflict with this idea:

For example, an interval A that includes an interval B on a particular data set may have lower confidence coverage!

Bayes turns confidence interval into probability intervals, and P(D|H) (as in P-values) into P(H|D) (what we really want)…

Example: Inference for a mean with bound on precision

A normal sample with $n = 7,\ \bar{y} = 1,\ s = 1$ yields

$$PI^{BRP}_{.05}(s=1) = CI^{F}_{.05}(s=1) = \bar{y} \pm 2.447\left(1/\sqrt{n}\right) = 1 \pm 0.92 \qquad (1)$$

Experimenter E tells us that true sd $\sigma = 1.5$

$$PI^{BRP}_{.05}(\sigma=1.5) = CI^{F}_{.05}(\sigma=1.5) = 1 \pm 1.96\left(1.5/\sqrt{7}\right) = 1 \pm 1.11 \qquad (2)$$

E: oops there's more variance! In fact $\sigma > 1.5$!

$$PI^{BRP}_{.05}(\sigma > 1.5) = 1 \pm 1.45 \qquad (3)$$

What does a frequentist do? Pick your poison:
(1) is an exact 95% CI but is clearly the wrong inference!
(2) is an anti-conservative 95% CI (though it contains (1)!)
(3) is correctly wider than (2), but it's Bayes, not a 95% CI, and depends on the choice of prior

Pr(D|H) or Pr(H|D)?

- Pr(D|H) is easier, but Pr(H|D) is what we really care about
- Classical or frequentist statistics (the stuff you learnt in a basic statistics course) stops at Pr(D|H):
  - P-value = Pr(D|H), *not* Pr(H|D)
  - Confidence intervals: proportion of intervals in repeated sampling that include a fixed parameter, *not* Pr(fixed interval includes parameter)
- Bayesian statistics tries for Pr(H|D)

Pr(D|H) or Pr(H|D)?

- Bayesians boldly (rashly?) seek Pr(H|D)
- Getting from Pr(D|H) to Pr(H|D) is called the *inverse probability* problem: *Bayes' rule* is the link…

Bayes' rule

- Bayes' rule converts $\Pr(D|H)$ into $\Pr(H|D)$, and is a simple consequence of basic rules of probability:

$$\Pr(H, D) = \Pr(D) \times \Pr(H \mid D) = \Pr(H) \times \Pr(D \mid H)$$

$$\Pr(H \mid D) = \Pr(H) \times \Pr(D \mid H) / \Pr(D)$$

$$\text{Hence,} \quad \frac{\Pr(H \mid D)}{\Pr(H' \mid D)} = \frac{\Pr(H)}{\Pr(H')} \times \frac{\Pr(D \mid H)}{\Pr(D \mid H')}$$

That is, posterior odds = prior odds × Bayes factor

- Bayes rule also converts confidence interval statements into posterior distributions for parameters θ:

$$p(\theta \mid D) \propto p(\theta) \times p(D \mid \theta)$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

A simple application of Bayes: Screening Tests

- A friend is diagnosed by a screening test (D = result of test, + or -) to have an extremely rare form of cancer (H = has cancer). Only one out of a million people in his age group have the cancer.
- Naturally he is very upset as the test is pretty accurate:

  Sensitivity: Pr(+| has cancer)=0.99, implying

  Pr(-| has cancer)=0.01 (False negative)

  Specificity: P(-|no cancer)=0.999, implying

  Pr(+| no cancer)=0.001 (False positive)

False Positive

- The probability that matters is the positive predictive value, which by Bayes Rule is

$$\Pr(\text{has cancer}|+)=\frac{\Pr(+|\text{has cancer})\Pr(\text{has cancer})}{\Pr(+)}$$

$$=\frac{(0.99)(1/1000000)}{(0.99)(1/1000000)+(0.001)(999999/1000000)}$$

$$=0.001 \quad (!)$$

False Positive

Very likely, the friend does not have cancer.

Bayesian statistics treats all unknowns (including fixed quantities) as random

- Frequentist statistics does not allow probability statements about fixed quantities – such as the true value of a parameter. Probability is the limit of the frequency of events in repeated sampling
- **Bayes uses probability statements to express uncertainty about all unknowns, whether "fixed" or "random"**
- In this sense any unknown is treated as a random variable, until its value is known.
- This idea greatly extends the reach of probabilistic statements.

History of Bayes

- Much maligned in the last century, Bayesian statistics has since experienced a dramatic revival
- See for example “The theory that would not die” by Sharon McGrayne

Bayes and the University of Michigan

- Arthur Bailey: BS from U of M Actuarial mathematics in 1928, affirmed Bayesian roots of "credibility theory" for setting workers' compensation insurance rates
- Allen Mayerson, actuarial professor at U-M, wrote about Bailey's seminal role
- Howard Raiffa: enrolled in actuarial mathematics at U of M, got his Ph.D. in 1952. With Robert Schlaifer wrote a highly influential book on Bayesian decision theory.

Bayes at U Michigan

- Leonard Jimmie Savage (mathematics PhD at U of M and professor at Chicago and later U of M) became a leader of the Bayesian revival
- In 1969 Bill Ericson (U of M Statistics Department) wrote the seminal paper on Bayes for sample surveys

LJ Savage

Calibrated Bayes

"… frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events."

Don Rubin (1984 Annals of Statistics)

Factoring in scientific plausibility

- Bayesian hypothesis testing formally allows prior scientific plausibility to modify the assessment of evidence, through the choice of prior distribution:

H = "Homeopathy works", $\bar{H}$ = "Homeopathy doesn't work".

$$\frac{\Pr(H \mid \text{data})}{\Pr(\bar{H} \mid \text{data})} = \frac{\Pr(H)}{\Pr(\bar{H})} \times \frac{\Pr(\text{data} \mid H)}{\Pr(\text{data} \mid \bar{H})}$$

$$\text{Posterior odds} = \text{Prior odds} \times \text{Bayes factor}$$

- For example, I'd give theories like homeopathy based on dubious science "skeptical priors".

What's bad about Bayes?

- "OK for gambling, but too subjective for science"
  - But frequentist methods can also make strong assumptions
  - Bayes makes assumptions in a model explicit, subject to criticism
  - Bayesian methods differ greatly in degree of subjective, e.g. in choice of model or prior
- Requires a high degree of model specification
  - Bad models yield bad answers
  - Need to pay attention to developing a good statistical model
- Too much work, computationally tractable
  - But computation is now feasible, using monte-carlo simulation methods

Summary

- Hypothesis testing and associated P-Values are widely viewed as flawed for assessing evidence
- Confidence intervals are better ways of assessing evidence, but they have some problems.
- Bayesian methods provide direct answers to the questions we really want to answer – what's the probability that a hypothesis is correct, or that an interval contains the parameter of interest
- So Bayesian methods are an alternative or complement to frequentist methods
- … although we would like our Bayesian methods to have good frequentist properties (to be *well calibrated*).