A Likelihood Approach to Statistics: Notes

Alex Balgavy

April-May 2019

1 Introduction

When can we say that there is sufficient evidence? A large issue is with the phrasing of conditional probability. There's a difference between
P(winning | not committed fraud)) and P(committed fraud | winning). The relative probabilities are important.

H_1, H_2: hypotheses
E: evidence/data

$$\begin{aligned}\frac{P(H_1 \mid E)}{P(H_2 \mid E)} &= \frac{P(H_1 \cap E)}{P(H_2 \cap E} \\ &= \frac{P(E \mid H_1)P(H_1)}{P(E \mid H_2)P(H_2)} \\ &= \frac{P(E \mid H_1)}{P(E \mid H_2)} \times \frac{P(H_1)}{P(H_2)}\end{aligned}$$

$$\therefore \underbrace{\frac{P(H_1 \mid E)}{P(H_2 \mid E)}}_{\text{posterior odds}} = \underbrace{\frac{P(E \mid H_1)}{P(E \mid H_2)}}_{\text{likelihood ratio}} \times \underbrace{\frac{P(H_1)}{P(H_2)}}_{\text{prior odds}}$$

Questions:

- When do observations support a hypothesis?
- What does this mean?
- What should I do next? What should I believe?

Evidence are data that **make you change your assessment of the hypotheses of interest**. It doesn't tell you what to believe, but how to change your belief. What to do depends on the risks and consequences.

The **likelihood ratio** is **the extent to which you should change your mind**.

The **evidence** is what determines the likelihood ratio.

1.1 Exercise 3.13

E: driver tests positive on breathalyzer
+: too much alcohol
−: below limit

$$LR = \frac{P(E \mid +)}{P(E \mid -)} = \frac{0.99}{0.10} = 9.9$$

Then:

$$\begin{aligned}
\frac{P(+ \mid E)}{P(- \mid E)} &= LR \times \text{prior odds} = 9.9 \times \overbrace{\frac{P(+)}{P(-)}}^{\text{given in ex.}} \\
&= 9.9 \times \frac{0.1}{9.9} = 1.1 \\
\frac{P(+ \mid E)}{P(- \mid E)} &= \frac{x}{1-x} = 1.1 \\
x &= \frac{11}{21}
\end{aligned}$$

2 Benchmarking

How do you quantify the likelihood ratio? Do a benchmark experiment.

Example with two hypotheses:
H_1: box has all white balls
H_2: box has 50% white, 50% black balls
E: drawing 5 white balls in a row (with replacement)

$$\frac{P(E \mid H_1)}{P(E \mid H_2)} = \frac{1}{\frac{1}{32}} = 32 = LR$$

Then, if some experiment has $LR = 357$, compared to benchmark, it's about as likely as drawing 8-9 white balls in a row. *But* you still can't say whether the situation is H_1 or H_2, because it depends on the prior odds.

3 General LR properties

The likelihood ratio cannot be *wrong*: given the evidence, the LR points a certain way. But it can be *misleading* and point towards a hypothesis that's not true.

Probability theory depends on the available information. Image placing bets for or against something – that's a first indication of the probabilities. **How often can the LR be misleading, and to what degree?**

Example: throw a pin, lands pin up with $\begin{cases} H_1 : p = \frac{1}{2} \\ H_2 : p = \frac{3}{4} \end{cases}$

One throw:

outcome	**LR**
up	$(\frac{1}{2})/(\frac{3}{4}) = \frac{2}{3}$
down	$(\frac{1}{2})/(\frac{1}{4}) = 2$

If H_1 is true, the average LR is

$$\frac{1}{2} \times \frac{2}{3} + \frac{1}{2} \times 2 = \frac{1}{3} + 1 = \frac{4}{3}$$

If H_2 is true, the average LR is

$$\frac{3}{4} \times \frac{2}{3} + \frac{1}{4} \times 2 = 1$$

With two throws, the average LR will be:

- if H_1 true, $(\frac{4}{3})^2$
- if H_2 true, 1

Average LR: $\sum P(\text{outcome}) \times LR_{\text{outcome}}$
If we compute LR for H_1 vs H_2 then:

- If H_1 is true, on average $LR > 1$
- If H_2 is true, on average $LR = 1$
- The following always holds:

$$\frac{P(LR = x \mid H_1)}{P(LR = x \mid H_2)} = x$$

The LR is a sufficient statistic for the two hypotheses, you won't learn more from seeing the evidence. So if we know $LR(E)$, we don't need to know E itself.

The probability of misleading evidence is

$$P(LR_{H_1,H_2}(E) \geq k \mid H_2) \leq \frac{1}{k}$$

regardless of H_1 and H_2.

LR is not additive, but multiplicative:

$$\begin{aligned} LR(E_1, E_2) &= LR(E_1) \times LR(E_2 \mid E_1) \\ &= LR(E_1) \times LR(E_2) \qquad \text{[if independent]} \end{aligned}$$

$$\log(LR(E_1, E_2)) = \log(LR(E_1)) + \log(LR(E_2 \mid E_1))$$

$$\begin{aligned} \log(LR) > 0 &: \quad \text{support } H_1 \\ \log(LR) = 0 &: \quad \text{no evidence either way} \\ \log(LR) < 0 &: \quad \text{support } H_2 \end{aligned}$$

3.1 Exercise 4.1

3.1.1 What's the likelihood ratio in favour of accused's guilt?

$$P(E \mid H_1) = 1$$
$$P(E \mid H_2) = \frac{1}{10000}$$
$$LR = \frac{1}{\frac{1}{10000}} = 10000$$

3.1.2 How can the value be used?

All you can do with the LR is to update the prior odds by multiplying.

3.1.3 What difference would it make if there were a 1% chance of matching results when in reality they are different?

$$P(E \mid H_1) = 1 \quad \text{unchanged}$$
$$P(E \mid H_2) = 0.01$$
$$LR = \frac{1}{0.01} = 100$$

3.1.4 Extra: what if, in 1% of chases, lab mistakenly says there is no match when there really is one?

$$P(E \mid H_1) = 0.99$$
$$P(E \mid H_2) = \frac{1}{10000} \quad \text{unchanged from original}$$
$$LR = \frac{0.99}{\frac{1}{10000}} = 9900$$

4 Assignment 1 Review

Definitions:

$H(1)$: All cards in deck are labelled 1
$H(i)$: All cards in deck are labelled i
H_n : Deck is normal
E : Choosing a card with label 1

How do you derive the result directly?

$$
\begin{aligned}
P(E \mid H(1)) &= 1 \\
P(E \mid \text{not } H(1)) &= \frac{P(E \cap \text{not } H(1)}{\text{not } H(1)} \\
&= \frac{p \times \frac{1}{52}}{1 - \frac{1-p}{52}} && \text{normal deck and choosing 1 out of it} \\
&= \frac{p}{52 - (1-p)} && \text{multiply by 52} \\
&= \frac{p}{51 + p} \\
LR &= \frac{P(E \mid H(1))}{P(E \mid \text{not } H(1))} \\
&= \frac{1}{\frac{p}{51+p}} \\
&= \frac{51+p}{p}
\end{aligned}
$$

5 From Data to Decision

The question now is, "what should I do?"

5.1 With prior probabilities: Bayes rule

Example – nuchal scan of fetus, to assess probability of trisomy 21. Scan produces evidence E. Hypotheses H_1 : trisomy 21, H_2 : no trisomy 21. $P(H_1)$ is given, based on age of mother:

Young mother: $\frac{P(H_1)}{P(H_2)} = \frac{1}{10000}$, action A1 if $LR \geq 40$

Old mother: $\frac{P(H_1)}{P(H_2)} = \frac{1}{5}$, action A1 if $LR \geq \frac{1}{50}$

Then compute posterior odds:

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = LR \times \frac{P(H_1)}{P(H_2)}$$

If $LR = \frac{P(E \mid H_1)}{P(E \mid H_2)}$ large enough, make a decision:

- **A1:** Further testing, or
- **A2:** no action

In The Netherlands, "large enough" means $\geq \frac{1}{250}$.

For young mothers, no more tests unless strong evidence for trisomy 21. For old mothers, do further test unless strong evidence against trisomy 21. Result depends not just on LR, but on product of LR with prior odds.

5.2 Without prior probabilities: frequentist approach

Suppose, if H_1 (the "null hypothesis") is true, we take action A1, and if H_2 is true, we take action A2. The options are:

	A1	**A2**
H1	true positive, sensitivity	false positive, type I error
H2	false negative, type II error	true negative, specificity

Often we like to control the probability of a type I error, called α. β is the probability of a type II error ($P(\text{decide A1} \mid H_2)$).

5.2.1 Decision procedure

1. Make probability distributions for evidence E you will gather
2. Define a way to decide – a rejection region R (subset of all possible evidence)
3. If evidence E turns out to be in R, reject H_1 (choose action A2). Otherwise, choose A1.

Such that:

- If H_1 is true, $P(E \in R \mid H_1) = \alpha$ (fixed, often 0.05).
- Preferably $P(E \in R \mid H_2) = 1 - \beta$ as large as possible (this is sometimes called the "power of the test").

5.2.2 Example

The same thumbtack, with $H_1 : p = \frac{1}{4}$, collecting data from 30 trials. If H_1 is true, X successes, observe X = x. How do you choose rejection region R? One way is to select the most unlikely outcomes in R until their joint probability to happen is too large. E.g. $R = \{0, 1, 2, 3, 13, 14, 15 \dots 30\}$, this would give $\alpha = 0.03$.

Well, if there is no alternative, β does not exist. So, take $H_2 : p\frac{3}{4}$.

If there are x successes, then

$$
\begin{aligned}
LR(X) &= \frac{P(X = x \mid p = \frac{1}{4})}{P(X = x \mid p = \frac{3}{4})} \\
&= \frac{\binom{30}{x}(\frac{1}{4})^x(\frac{3}{4})^{30-x}}{\binom{30}{x}(\frac{3}{4})^x(\frac{1}{4})^{30-x}} && [\text{binomial distribution } X \sim B\Big(30, \frac{1}{4}\Big)] \\
&= \frac{\frac{1}{4^x}\frac{3^{30-x}}{4^{30-x}}}{\frac{3^x}{4^x}\frac{1}{4^{30-x}}} \\
&= \frac{\frac{3^{30-x}}{4^{30}}}{\frac{3^x}{4^{30}}} \\
&= \frac{3^{30-x}}{3^x} \\
&= 3^{30} - 2x
\end{aligned}
$$

We also have the property that

$$P(X = x \mid p = \frac{1}{4}) = P(X = 30 - x \mid p = \frac{3}{4})$$

Based on the result, then decide:

- If $x < 15$, $LR > 1$ and supports $H_1 : p = \frac{1}{4}$
- If $x = 15$, $LR = 1$
- If $x > 15$, $LR < 1$ and supports $H_2 : p = \frac{3}{4}$

The table for this binomial distribution with the corresponding LRs is

LR	**Rejection region R**	α	β	$\alpha + \beta$
$LR \leq 729$	$x \geq 12$	0.0506	≈ 0	0.0506
$LR \leq 81$	$x \geq 13$	~ 0.0215	≈ 0	0.0215
$LR \leq 9$	$x \geq 14$	~ 0.0081	0.0002	0.0083
$LR \leq 1$	$x \geq 15$	0.0027	0.0008	0.0035
$LR \leq \frac{1}{9}$	$x \geq 16$	0.0008	0.0027	0.0035
	$x \geq 17$	0.0002	0.0081	0.0083
	etc.			
	$\{0, 1, 2, 3, 13 \ldots 30\}$	~ 0.03	≈ 0	≈ 0.03

5.2.3 Neyman-Pearson lemma

If LR-threshold is used for decision making, eg.

$R_t = \{E|LR(E) \leq t\}$ t is threshold, whatever number

Then you get some

$$\begin{cases} \alpha_t & = P(LR(E) \leq t \mid H_1) \\ \beta_t & = P(LR(E) > t \mid H_2) \qquad (< \frac{1}{t}) \end{cases}$$

Suppose I have another procedure with rejection region R and error rates α_R and β_R. If $\alpha_R < \alpha_t$, then $\beta_R > \beta_t$.

So, LR is optimal in the sense that it is impossible to improve upon *both* α_t and β_t at the same time. Therefore, there is no conceptual reason to use a different procedure (though there may be a practical reason).

With $t = 1$, the sum of $\alpha + \beta$ is minimal.

In the example, with $\alpha = 0.05$, we get $5 = 729$.

$R_{729} = \{x \geq 12\}$

That is, if $x \geq 12$, $LR = 729$. Evidence supports H_1, but H_1 is rejected.

Error rates are predictive, they belong to a procedure for decision making:

- If H_1 true: probability α of error.
- If H_2 true: probability β of error.

It's not true that if you decide for H_1, there is a probability α that *you* made an error!

5.2.4 Frequentist vs Bayesian statistics

Frequentist	Bayesian
no priors	priors
predicting data	explaining data
LRs for decision making	LRs for updating odds, *then* decision making

If you don't have priors and no good way to estimate them, it may be better to go with the frequentist approach and accept the errors that come with it.

6 Neyman-Pearson

To recap, the LR "decides" which hypothesis best explains data. Data-driven hypotheses are allowed, but since the posterior odds identity is true, a high LR is compensated by small prior odds.

Procedure:

1. Choose α
2. Choose t, with t such that

$$P(LR < t \mid H_1) = \alpha$$

Choose A_1 if $LR_{H_1,H_2})(E) \geq t$

This means that we choose A_2 while there is evidence for H_1.

6.1 Example (building on the binomial coin from the previous lecture)

$A_1 : \theta = \frac{1}{4}, \quad H_2 : \theta = \frac{3}{4}, \alpha = 0.05$

Choose A_1 if $LR \geq 729$ ($\#successes \geq 12$). Why? Because you insist on a small α.

7 What if only the final decision is given?

What happens if you only get "an expert's opinion" and the final decision they took? You can still figure out the evidential value.

$$\frac{P(E_1 \mid H_1)}{P(E_1 \mid H_2)} = \frac{1-\alpha}{\beta} \qquad \frac{P(E_2 \mid H_1)}{P(E_2 \mid H_2)} = \frac{\alpha}{1-\beta}$$

If β is small, LR increases and you get high evidential value.

8 P-values: what's wrong with them?

8.1 Example: researchers' experiments

Goal is to disprove success probability $p = \frac{1}{2}$. 20 experiments, the result is 14 successes. $\alpha = 0.05$, $H_1 : \ p = \frac{1}{2}$, $H_2 : \ p \neq \frac{1}{2}$.

Compute $P(\geq 14 \cup \leq 6 \mid H_1) = 0.23$. Since this is $> \alpha$, not significant enough so can't reject H_1. But 15 successes would have done it, with probability of 0.0412. So do 20 more trials, with 19 successes. Then $P(\geq 33 \cup \leq 7 \mid H_1) = 0.000422$.

But rejection of H_1 is incorrect here! After 40 experiments, $P(\geq 27 \cup \leq 13 \mid H_1) = 0.05$. The total critical region is:

- $\leq 5 \cup \geq 15$ if 20 experiments
- $\leq 13 \cup \geq 27$ if 40 experiments

Total probability is ≥ 0.05

So what if we do 20 experiments, possibly stopping after 10? Reject H_1 if either:

- After 10 exp. $\geq 9 \cup \leq 1$ successes
- After 20 exp. $\geq 16 \cup \leq 4$ successes

Probability under H_1 is ≤ 0.05.

Then, results are: after 10, 3 successes; after 20, 5 successes. So H_1 is not rejected.

Another researcher only looks after 20 experiments, so for them, 5 successes means reject!

It's strange that we are using probabilities of outcomes we never saw to interpret the evidence. The LR approach doesn't have this problem.

8.2 Example: ability to see color

You have 20 colors. Experiment 1: for people that don't see green, reject $p = \frac{1}{2}$ if #successes = {0,1,9,10}. Experiment 2: $H_1 : \ p = \frac{1}{2}$ for all colors, reject H_1 if at least one person gets 0 or 10.

Result: experiment with green has 9 successes, experiment with all others in {1,2,...9}. What then? Reject and don't reject at the same time?

Other experiments should not have an effect on evidential value of an experiment.

8.3 Example: one-tailed vs. two-tailed

Take p to be unknown success probability. $\alpha = 0.05$, 100 experiments. $H : \ p = \frac{1}{2}$, reject H if successes $\leq 59 \cup \geq 61$. $H' : \ p \leq \frac{1}{2}$, reject H' if successes ≥ 59.

Suppose 60 successes. Reject $p \leq \frac{1}{2}$ but not $p = \frac{1}{2}$? Wtf?

8.4 Example: changing alpha

$H:\ p = \frac{1}{2}$, 40 experiments, $\alpha = 0.05$. $P(\geq 29 \cup \leq 11 \mid H) = 0.0003$.

The researcher sees that $\alpha = 0.01$ would also be ok, so they claim to reject H at level $\alpha = 0.01$.

This is wrong! α belongs to the whole experiment, it does not relate to an individual outcome. By changing α from experiment to experiment, it loses the *only* interpretation it has.

9 P-values of LRs

Suppose $LR = 47$. The p-value is then $P(LR \geq 47 \mid H_2)$. The idea is that, if p-value is very small, then the LR of 47 is extreme for H_2. If it's large, then the 47 is 'normal' for H_2.

However, this still has no *evidential* value. LR measures strength of evidence. The p-value tells you how rare such a LR is. However, once you have evidence, it doesn't matter how frequently evidence of that strength occurs.

9.1 Example: genomes

Two people with genomes g1, g2. $H_1 : siblings$, $H_2 : unrelated$.

You can take different types of LRs:

$$LR_{H_1,H_2}(g_1, g_2) = \frac{P(g_1, g_2 \mid H_1)}{P(g_1, g_2 \mid H_2)}$$
$$LR' = \frac{P(g_2 \mid g_1, H_1)}{P(g_2 \mid g_1, H_2)}$$
$$LR'' = \frac{P(g_1 \mid g_2, H_1)}{P(g_1 \mid g_2 H_2)}$$

These are all basically the same. Take notation $p_1 = P(g_1)$, $p_2 = P(g_2)$. $p_1(g_2) = P(g_1$ for a sibling of someone with $g_2)$. $p_2(g_1) = P(g_2$ for a sibling of someone with $g_1)$.

Then we can rewrite the LRs from above:

$$LR_{H_1,H_2}(g_1, g_2) = \frac{p_1 p_2(g_1)}{p_1 p_2} = \frac{p_2(g_1)}{p_2}$$
$$LR' = \frac{p_2(g_1)}{p_2}$$
$$LR'' = \frac{p_2(g_1)}{p_2}$$

The p values will be different though, because depending on the fixed genome, the frequency of how often it occurs will be different. This would lead to different actions, even though the LRs are identical, and thus so is the evidence.

9.2 Example: disease and test results

Take H_1 : disease present, H_2 : disease absent.

Experiment 1:

	+	-
H_1	0.94	0.06
H_2	0.02	0.98

$$LR(+) = \frac{P(+ \mid H_1)}{P(+ \mid H_2)} = \frac{0.94}{0.02} = 47$$
$$LR(-) = \frac{P(- \mid H_1)}{P(- \mid H_2)} = \frac{0.06}{0.98} = \frac{1}{16}$$

Experiment 2 ("0" means experiment is not carried out):

	+	**0**	-
H_1	0.47	0.5	0.03
H_2	0.01	0.5	0.49

$$LR(+) = 47$$
$$LR(-) = \frac{1}{16}$$

Experiment 3 ("*" is negative result or no experiment):

	+	*
H_1	0.47	0.53
H_2	0.01	0.99

$$LR(+) = 47$$

All of the LRs are the same. So essentially, if a "+" is obtained, the evidential value is always the same no matter how it was obtained.

Per experiment, $P(LR \geq 47 \mid H_2) =$

1. 0.02
2. 0.01
3. 0.01

These are not all the same!

The p-value relates to the *entire* procedure, that's why it's not the same. The LR relates to an individual outcome, so it's always the same.

10 Why confidence intervals are similarly fucked

Recall testing H1 vs H2: define rejection region R, s.t. if sampled data are in R, you "reject H1" (take some action). Otherwise, do not reject.

P-values define R in terms of what might happen if H1 is true, s.t. total probability for data to be in R is α. The point is that you can't interpret data in R as evidence against H1.

Neyman-Pearson: define R using LR threshold t. $R\{E|LR(E) \leq t\}$ gives you optimality.

Why p-values suck (recap):

- do not measure strength of evidence in E against H1
- they are ambiguous (several ways of defining them)
- the probability α is a property of the procedure that you do (how data are gathered), not of the obtained data

10.1 Confidence intervals

Say we have a model (e.g. a Binomial distribution) that generates the data and has unknown param θ that we want to estimate. Example: θ mean height of people, model $N(\theta, \sigma^2)$.

A CI of $1 - \alpha$ consists of two functions on data that can be obtained, θ_{min} and θ_{max}, such that if θ is true value of the param of interest, it lies between $\theta_{min}(E)$ and $\theta_{max}(E)$ with probability $1 - \alpha$ if we repeat sampling of E.

10.1.1 Commonly encountered 95% CI

For data from $N(\theta, \sigma^2)$, if I sample n points $x_1, \ldots, x_n$, estimate θ by

$$\hat{\theta} = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

And take as 95% CI $\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$. 1.96 is the z-score for the CI.

Why? It gives the smallest 95% CI for such data.

10.1.2 Binomial data (2 possible outcomes)

Data $x_1, \ldots, x_n$, interested in "success prob" p of $p = P(x = 1)$.

With success probability p, in n points $x_1, \ldots, x_n$, there are k successes (ones) with prob

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

(Binomial distribution probability).

A 95% CI can b e computed with this, but a good approximation is

$$\theta_{min,max} = \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} = \frac{k}{n}$.

The CI at level $1-\alpha$ contains exactly the values that would not lead to rejection with significance level α (i.e. p-value $\geq \alpha$).

10.2 Problems with CIs

CIs suffer from the same problems as p-values:

- α is a property of the procedure, not of any realized outcome
- ambiguity: lots of choices possible

10.2.1 Example

Want to estimate θ, gather data x.

$$P(x|\theta) = \begin{cases} \frac{1}{2} & x = \theta \\ \frac{1}{2} & x = \theta + 1 \end{cases}$$

Gather two points x_1, x_2. CI defined as $\left[\theta_{min} = \min(x_1, x_2), \quad \theta_{max} = \max(x_1, x_2)\right]$

This is a 75% CI.

But if data are $x_1 = 28$, $x_2 = 29$, then CI is $[28, 29]$ and definitely contains θ. If $x_1 = x_2 = 30$, then CI is $[30]$, θ could be 29 or 30. If the values for θ are equally likely, 50% chance to contain θ.

10.2.2 Example

With n points from $N(\mu, \sigma^2)$ normal dist, 95% CI is $\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$.

Let $n = 1$ or $n = 100$ with probability $\frac{1}{2}$. What's a good 95% CI?

1. If $n = 1$, 95% CI is $x_1 \pm 1.96\sigma$. If $n = 100$, 95% CI is $\overline{x} \pm 1.96\frac{\sigma}{10}$. But you can do better.
2. If $n = 1$, $x_1 \pm 1.62\sigma$ (91% CI). If $n = 100$, $\overline{x} \pm 2.72\frac{\sigma}{10}$ (99% CI). Overall, this is also 95% CI.

Why number 2? Expected width of intervals:

1. $\frac{1}{2}(2 \times 1.96\sigma) + \frac{1}{2}(2 \times 1.96\frac{\sigma}{10} = 1.96\sigma + 0.196\sigma = 2.156\sigma$
2. $\frac{1}{2}(2 \times 1.62\sigma) + \frac{1}{2}(2 * 2.72\frac{\sigma}{10} = 1.62\sigma + 0.272\sigma = 1.892\sigma$

10.2.3 Example

Heart/lung problems with newborns. Conventional medical treatment not very adequate, survival rate not precisely known but 20%. New, promising treatment ECMO, survival rate estimated possibly around 80%. Study ECMO vs CMT. How large should it be?

Take n patients, number of recoveries is x.

Say, test $\theta_{ECMO} = 0.8$ vs $\theta_{ECMO} = 0.2$. If x recoveries, LR is

$$\begin{aligned}\frac{P(x \mid \theta = 0.8)}{P(x \mid \theta = 0.2)} &= \frac{\binom{n}{x}(0.8)^x(0.2)^{n-x}}{\binom{n}{x}(0.2)^x(0.8)^{n-x}} \\ &= \frac{4^x}{4^{n-x}} = 4^{2x-n} \\ &= 2^{4x-n}\end{aligned}$$

If I want $LR \geq 32 (\geq 2^5)$, I need $4x - 2n \geq 5$. So $x \geq \frac{2n+5}{4}$.

We can compute probability to get sufficiently strong evidence in favor of the true hypothesis, or probability of strongly misleading evidence, or probability of not obtaining strong evidence.

Now suppose 13 out of 17 recoveries. What does this say about θ_{ECMO}?

We could CI that shit, but CIs have problems.

The best $\theta_{ECMO} = \frac{13}{17} = 0.76$. How much better than $\theta = 0.5$?

$$\frac{P(13 \text{ out of } 17 \mid \theta = 0.8)}{P(13 \text{ out of } 17 \mid \theta = 0.5)} = 11.5$$

A likelihood interval. E.g. $\frac{1}{32}$ LI is all values θ such that LR for $\theta = \frac{13}{17}$ vs $\theta = \theta_0$ is at most 32.

11 Notes from Ioannidis

These are notes from the class when discussing the article "Why Most Published Research Findings Are False" by Ioannidis.

S : significant result ($p < 0.05$).

$$\frac{P(\overline{H_0} \mid S)}{P(H_0 \mid S)} = \frac{P(S \mid \overline{H_0})}{P(S \mid H_0)} \times \underbrace{\frac{P(\overline{H_0})}{P(H_0)}}_{R \text{ in the article}}$$

$$\frac{P(S \mid \overline{H_0})}{P(S \mid H_0)} = \frac{1-\beta}{\alpha}$$

So $\frac{1-\beta}{\alpha} \times R > 1$ for H_0 to be false.

In notation of Ioannidis, $(1-\beta)R > \alpha$.

Odds $\frac{P\overline{H_0}}{P(H_0)} = R$ are equivalent to

$$P(\overline{H_0}) = \frac{R}{R+1} \qquad P(H_0) = \frac{1}{R+1}$$

.

Total number of research questions c is then:

$$\begin{cases} c\frac{R}{R+1} & \text{if } \overline{H_0} \text{ true} \longrightarrow \begin{cases} S(\overline{H_0} \text{ true}) = (1-\beta)c\frac{R}{R+1} \\ \overline{S}(\overline{H_0} \text{ true}) = \beta c\frac{R}{R+1} \end{cases} \\ c\frac{1}{R+1} & \text{if } H_0 \text{ true} \longrightarrow \begin{cases} S(H_0 \text{ true}) = \alpha c\frac{1}{R+1} \\ \overline{S}(H_0 \text{ true}) = (1-\alpha)c(\frac{1}{R+1}) \end{cases} \end{cases}$$

Ioannidis: $\beta = 0.2$, $\alpha = 0.05$. So

$$LR(S) = \frac{1-0.2}{0.05} = \frac{0.8}{0.05} = 16$$

11.1 Bias

Bias is when you get more significant findings than warranted by the data. E.g. you try to 'clean up the data'. But then your original error rates don't apply anymore.

Originally,

$$P(S \mid H_0) = \alpha$$
$$P(S \mid \overline{H_0}) = 1 - \beta$$

Now,

$$P(S \mid H_0) = \alpha + (1-\alpha)u$$
$$P(S \mid \overline{H_0}) = (1-\beta) + \beta u$$

where u is the probability of data becoming significant when they are not.

Now, with bias, LR of S for $\overline{H_0}$ vs H_0 becomes

$$\frac{P(S \mid \overline{H_0})}{P(S \mid H_0)} = \frac{1-\beta+\beta u}{\alpha+(1-\alpha)u}$$

.

$PPV = P(\overline{H_0} \mid S)$. Plot y-axis odds $\frac{P(\overline{H_0} \mid S)}{P(H_0 \mid S)}$, x-axis u.

Suppose several teams:

- all the same research question
- all the same α and β
- result is published as soon as at least 1 team finds statistically significant result

S : at least one team has $p < 0.05$.

$$\frac{P(S \mid \overline{H_0})}{P(S \mid H_0)} = \frac{1-\beta^n}{1-(1-\alpha)^n}$$

As n goes to infinity, the result tends towards 1.

Corollaries:

- smaller studies = less likely for findings to be true
- smaller effect sizes = less likely for findings to be true
- greater number and less selection of tested relationships = less likely for findings to be true
- greater flexibility = less likely for findings to be true

12 The Paradox of the Ravens (Hempel)

H: all ravens are black. Equivalent to saying "all not black things are not ravens". So observation of non-black should be evidence for H.

Suppose two vases, one with only ravens (R, amount n_R), and one with only non-ravens (NR, amount n_{NR}). P_R is probability of black in raven vase, P_{NR} is probability of black in non-ravens.

X is a draw from R. $H_A : P_R = 1$, $H_B : P_R = p < 1$.

Evidence is that X is black.

$$\begin{aligned} LR_{A,B}(E) &= \frac{P(Xisblack \mid A)}{P(Xisblack \mid B)} \\ &= \frac{1}{p} > 1 \end{aligned}$$

So this is evidence that all ravens are black.

Y is a draw from NR. Evidence is that Y is white.

$$\begin{aligned} LR_{A,B}(E') &= \frac{P(Yiswhite \mid A)}{P(Yiswhite \mid B)} \\ &= 1 \quad \text{A,B do not affect NR, just R} \end{aligned}$$

That's not evidence for H. It's neutral.

12.1 But there's a big but

What if we do this:

1. Mix all the things
2. Choose non-black object from the mix
3. Suppose this non-black object came from NR
4. Claim this is evidence for H

Z is outcome, R or NR.

$$
\begin{aligned}
LR_{A,B}(Z = NR) &= \frac{P(Z = NR \mid A)}{P(Z = NR \mid B)} \\
&= \frac{1}{P(Z = NR \mid B)} \\
P(Z = NR) &= \frac{\text{num of non-black objects in NR}}{\text{total num of non-black objects}} \\
&= \frac{n_{NR}(1 - P_{NR}}{n_{NR}(1 - P_{NR}) + n_R(1 - P_R} \\
\therefore LR_{A,B}(Z = NR) &= \frac{1}{P(Z = NR} \\
&= \frac{n_{NR}(1 - P_{NR}) + n_R(1 - P_R}{n_{NR}(1 - P_{NR}} \\
&= 1 + \frac{n_R(1 - P_R)}{n_{NR}(1 - P_{NR}} \\
&> 1, \text{ if } P_R \text{ is not 1 (assumed)}
\end{aligned}
$$

So this *is* evidence for H! (though not very strong evidence)

So, two ways of sampling, which one you use *is* definitely relevant. If you select something that you *know* is not a raven, and see that it's not black, that' snot evidence. If you randomly select something that's not black, and see that it's not a raven, it *is* evidence.