Chapter 1
Statistics, Likelihood and Evidence

Charles A. Rohde

March 19, 2004

Contents

1 Introduction

> "The statistical sciences are concerned with the collection, analysis and interpretation of data that involve uncertainty."

Mathematical Sciences: Some Research Trends. National Academy of Sciences Press 1998.

Most statistics courses focus on the analysis and interpretation of data. In particular statistical inference focuses on the interpretation of data in the context of a probability model.

Why should one study the foundations of statistics? Royall has put it best "Statistics is a mess". In mathematics one counterexample is enough to demonstrate that a stated result or methodology is incorrect or inappropriate. Not so in statistics.

Not only that, there are a variety of definitions of statistical inference e.g.

- " the purpose of inductive reasoning, based on empirical observations, is to improve our understanding of the systems from which these observations are drawn"
  Statistical Methods and Scientific Inference (R.A. Fisher)
- "Inference is usually defined as the process of drawing conclusions from facts, available evidence, and premises. Statistical inference is the term associated with the process of making conclusions on the basis of data that are governed by probability laws".
  Encyclopedia of Biostatistics (L. Fisher)
- " A statistical inference will be defined for the purposes of the present paper to be a statement about statistical populations made from given observations with measured uncertainty."
  Some Problems Connected with Statistical Inference (D.R. Cox)
- " statistics is concerned with decision making under uncertainty"
  Decision Theory (Chernoff and Moses)
- " the problem of inference, or how degrees of belief are altered by data"
  Probability and Statistics from a Bayesian Viewpoint (Lindley)
- "By statistical inference I mean how we find things out - whether with a view to using the new knowledge as a basis for explicit action or not - and how it comes to pass that we often acquire practically identical opinions in the light of evidence"
  The Foundations of Statistical Inference (L.J. Savage)

Consider the simplest problem, that of determining on the basis of an observation x which of two models is true. Assume that the two models are f_0 and f_1. Classical statistics tells us to form the likelihood ratio

$$\frac{f_1(x)}{f_0(x)}$$

and conclude that f_1 is the correct model if this ratio is large. Large being determined by finding the smallest k such that

$$P_0\left(\frac{f_1(X)}{f_0(X)} \geq k\right) \leq \alpha$$

The Neyman-Pearson Lemma tells us that this procedure is optimal in the sense that

- The probability of falsely concluding that f_1 is true when f_0 is true is fixed at α
- The probability of falsely concluding that f_0 is true when f_1 is true is minimized. i.e. this procedure maximizes the power for fixed size.

This result has been extended ad nauseum and has dominated much of statistical practice since its introduction in the early 1930's.

example 1.1 If the most powerful test of H_0 vs H_1 calls for rejection of H_0 it does not mean that the data provide evidence against H_0 in favor of H_1.

Consider a simple situation where Y has one of two pdf's f_0 or f_1 as follows:

	Observed Value of Y		
pdf	$y = 1$	$y = 2$	$y = 3$
f_0	$\frac{19}{20}$	$\frac{1}{20}$	0
f_1	0	$\frac{1}{200}$	$\frac{199}{200}$
$\frac{f_1(y)}{f_0(y)}$	0	$\frac{1}{10}$	∞

The most powerful test of size $\alpha = .05$ rejects H_0 when $y = 2$ or 3 and has power 1. If, however, $y = 2$ is observed even though we reject H_0 the observed data are 10 times more likely under H_0 than under H_1.

Conclusion: Size and power cannot be interpreted as measures of strength of evidence.

example 1.2 Use of hypothesis tests can lead to foolish conclusions.

Suppose that we have Y with one of two pdf's as follows:

	Observed Value of Y			
pdf	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$
f_0	9/10	1/20	1/20	0
f_1	0	99/100	0	1/100

Consider the following two tests

Test 1 rejects if $y = 2$ and has size .05 with power 99/100
Test 2 rejects if $y = 3$ or $y = 4$ and has size .05 with power 1/100

It is clear that before the data are collected Test 1 is better. After the data are collected and we observe $y = 4$ then Test 1 does not reject even though there is perfect evidence against H_0 whereas Test 2 rejects.

Conclusion: Blind evaluation of tests in terms of size and power is silly.

example 1.3 Hypothesis tests are not symmetric.

Suppose that we have Y with one of two pdf's as follows:

	Observed Value of Y			
pdf	$Y=1$	$Y=2$	$Y=3$	$Y=4$
f_0	98/100	1/100	1/100	0
f_1	0	1/100	1/100	98/100
$\frac{f_1(y)}{f_0(y)}$	0	1	1	∞
$\frac{f_0(y)}{f_1(y)}$	∞	1	1	0

- The most powerful test of size $\alpha = .02$ of f_0 vs f_1 rejects if y equals 2,3 or 4 and has power 1.
- The most powerful test of size $\alpha = .02$ of f_1 vs f_0 rejects if y equals 1,2 or 3 and also has power 1.

Thus if y is observed to be 2 or 3 you would reject either f_0 or f_1 depending on which you called the null hypothesis. However, the observed data in this case provide no evidence for or against either of the hypotheses.

2 Statistical Evidence

Let's modify the National Academy of Sciences definition and adopt as a working definition of statistics the following:

Statistics is the discipline concerned with statistical evidence:

- producing it
- interpreting it
- using it

For the moment " Statisical evidence consists of observations in the context of a probability model".

- A model is a collection of probability distributions often indexed by a parameter.
- Observations are conceptualized as having been generated according to one of the distributions in the model.
- Statistics is concerned with using the observations to determine evidence concerning the distribution which generated the observations.

example: Consider a diagnostic test in which a subject has a positive test result. The test has the following probabilistic characteristics:

Disease	Test Result	
Status	$+$	$-$
D	.95	.05
$\overline{D}$	.02	.98

If the observation X represents the test result then we have the following model:

$$H_A: \text{ disease present } \Longrightarrow X \sim P_D \;\; ; \;\; H_B: \text{ disease absent } \Longrightarrow X \sim P_{\overline{D}}$$

On the basis of a positive test result a physician might conclude:

(1) The subject probably does not have the disease.

(2) The subject should be treated for the disease.

(3) The test result is evidence that the subject has the disease.

All of these statements fall within the scope of statistics.

- Which are correct?
- How do we determine which are correct?

To answer (1) we ask when is $P(D|X=1) < \frac{1}{2}$? By Bayes Theorem

$$\begin{aligned} P(D|X=1) &= \frac{P(X=1|D)P(D)}{P(X=1|D)P(D) + P(X=1)|\overline{D})P(\overline{D})} \\ &= \frac{.95\ P(D)}{.95\ P(D) + .02[1 - P(D)]} \\ &= \frac{.95P(D)}{.93P(D) + .02} \end{aligned}$$

It follows that $P(D|X=1)$ increases from 0 to 1 as $P(D)$ increases. Also note that

$$P(D|X=1) < \frac{1}{2} \text{ if and only if } P(D) < .0206$$

Hence conclusion (1) is correct if $P(D)$ is small. It follows that conclusion (1) depends on the probability of the disease before the test and not just on the data provided by the test.

For conclusion (2) assume that we have a table of losses

Model Status	Treament	
	T	$\overline{T}$
D,	$\ell(T,D)$	$\ell(\overline{T},D)$
$\overline{D}$	$\ell(T,\overline{D})$	$\ell(\overline{T},\overline{D})$

where it is natural to assume that

$$\ell(T,D) < \ell(\overline{T},D) \quad , \quad \ell(\overline{T},\overline{D}) < \ell(T,\overline{D})$$

If the patient is treated the expected loss is

$$R_T = \ell(T,D)P(D|X=1) + \ell(T,\overline{D})P(\overline{D}|X=1)$$

while if the pateint is not treated the expected loss is

$$R_{\overline{T}} = \ell(\overline{T},D)P(D|X=1) + \ell(\overline{T},\overline{D})P(\overline{D}|X=1)$$

It is thus better to treat if the expected loss if treated is less than the expected loss if not treated ($R_T < R_{\overline{T}}$) i.e. if

$$\ell(T,D)P(D|X=1) + \ell(T,\overline{D})P(\overline{D}|X=1) \; < \; \ell(\overline{T},D)P(D|X=1) + \ell(\overline{T},\overline{D})P(\overline{D}|X=1)$$

This is equivalent to

$$\frac{P(D|X=1)}{P(\overline{D}|X=1)}\ell(T,D)+\ell(T,\overline{D})<\frac{P(D|X=1)}{P(\overline{D}|X=1)}\ell(\overline{T},D)+\ell(\overline{T},\overline{D})$$

or

$$[\ell(T,\overline{D})-\ell(\overline{T},\overline{D})]<\frac{P(D|X=1)}{P(\overline{D}|X=1)}[\ell(\overline{T},D)-\ell(T,D)]$$

Hence the second conclusion is correct if

$$\frac{P(D|X=1)}{P(\overline{D}|X=1)}>\frac{\ell(T,\overline{D})-\ell(\overline{T},\overline{D})}{\ell(\overline{T},D)-\ell(T,D)}$$

Note that the correctness of the second conclusion depends on the probability of the disease before the test and the loss structure, not just on the data from the test since

$$\frac{P(D|X=1)}{P(\overline{D}|X=1)}=\frac{P(X=1|D)P(D)}{P(X=1|\overline{D})P(\overline{D})}=\left[\frac{P(X=1|D)}{P(X=1|\overline{D})}\right]\left[\frac{P(D)}{P(\overline{D})}\right]$$

For the third conclusion we note that **it is correct** because interpreting a positive test as evidence that the subject does not have the disease is just **wrong** regardless of prior probabilities, possible actions and their consequences.

The three conclusions in the example are anwers to different generic questions:

(1) What should the physician believe?

(2) What should the physician do?

(3) How should the physician interpret the test result as evidence (for D vis-a-vis $\overline{D}$)?

Note that the first two questions are context dependent whereas the third is universal. The first question is Bayesian, the second is decision theoretic and the third we call the evidential interpretation.

The third question is the one we will concentrate on in this course. Saying that a positive test result is evidence that the subject has the disease requires us to determine and investigate "What fundamental principles of statistical reasoning are at work here?"

3 Law of Likelihood

Axiom 3.1 Law of Likelihood: Let hypothesis A imply that a random variable is distributed as p_A and hypothesis B imply that it is distributed as p_B.

- The observation $X = x$ is evidence supporting A over B if and only if

$$\frac{p_A(x)}{p_B(x)} > 1$$

- Moreover the magnitude of the likelihood ratio

$$\frac{p_A(x)}{p_B(x)}$$

  measures the strength of the evidence.

Suppose we take the Law of Likelihood as an axiom. For this axiom to be useful and acceptable in statistics we need to answer several questions:

(1) Does it make intuitive sense?

(2) Is it consistent with probability theory and logic?

(3) Does it work?

4 Intuition

Consider two hypotheses A and B which state that

$$\begin{aligned} A: &\quad \text{under conditions } C,\ X \text{ will happen} \\ B: &\quad \text{under conditions } C,\ X \text{ will not happen} \end{aligned}$$

To investigate these hypotheses we create the conditions (perform an experiment or observational study) and see what happens. An observation of X supports or provides evidence for A vis-a-vis B.

The Law of Likelihood generalizes this statement to

$$\begin{aligned} A: &\quad \text{under conditions } C,\ X \text{ will happen with high probability} \\ B: &\quad \text{under conditions } C,\ X \text{ will happen with low probability} \end{aligned}$$

where high and low refer to **relative probability not absolute probability**.

Note that if probabilities are assigned to A and B then

$$\frac{P(A|X)}{P(B|X)} = \frac{P(X|A)P(A)}{P(X|B)P(B)} = \frac{P_A(X)}{P_B(X)}\frac{P(A)}{P(B)} = \left[\frac{P_A(X)}{P_B(X)}\right]\left[\frac{P(A)}{P(B)}\right]$$

Thus the likelihood ratio is the factor by which the "statistical evidence" changes the probability ratio. In particular a likelihood ratio of say, 5, implies that $X = x$ causes a 5-fold increase in $P(A)/P(B)$ from prior to posterior.

One way to think about the likelihood ratio is that it is similar to a measure of thermal energy: 1 BTU is the energy required to raise the temperature of 1 pound of water at 39.2 degrees F by one degree. This is an attempt to give a meaning to the likelihood ratio in a way such that it always means the same thing, regardless of context. Thus one speaks of the BTUs of one air conditioner relative to another.

Returning to our example

(1) Should this observation lead me to believe that D is present?

(2) Does this observation justify acting as if D were present?

(3) Is this observation evidence that D is present?

Question 3 is the only one that can be answered independently of prior probabilities and consequences. This implies that

"what do the data say?" is independent of beliefs and actions

which in turn implies that the Law of Likelihood provides a numerical measure of the strength of statistical evidence.

In the example a positive test is evidence that the subject has the disease since

$$\text{LR} = \frac{p_D(x)}{p_{\overline{D}}(x)} = \frac{.95}{.02} = 47.5$$

If in fact $\overline{D}$ is true this evidence is misleading. In this particular example we can get misleading evidence of this magnitude with probability at most .021 i.e. if

$$A \Longrightarrow X \sim f_A \text{ and } B \Longrightarrow X \sim f_b$$

then if A is true the probability of misleading evidence satisfies

$$P_A\left(\frac{f_B(X)}{f_A(X)} \geq 47.5\right) \leq \frac{1}{47.5} = .021$$

In a later section we will show that

$$P_A\left\{ \frac{f_B(X)}{f_A(X)} \ \geq \ k \right\} \ \leq \ \frac{1}{k}$$

for any positive k.

5 Strength of evidence

Is a LR of 2 "strong evidence? is 10? is 100? We need benchmarks to relate a numerical scale to verbal descriptions. Consider two urns:

Urn 1: all white balls ; Urn 2: half white balls, half red balls

If we draw balls at random from one of the urns without knowing which urn we are selecting from we have

$$\begin{aligned}
\text{one white ball} &\Longrightarrow \text{LR} = \tfrac{1}{1/2} = 2 \\
\text{two white balls} &\Longrightarrow \text{LR} = \tfrac{1}{1/2^2} = 4 \\
\text{three white balls} &\Longrightarrow \text{LR} = \tfrac{1}{1/2^3} = 8 \\
\text{four white balls} &\Longrightarrow \text{LR} = \tfrac{1}{1/2^4} = 16 \\
\text{five white balls} &\Longrightarrow \text{LR} = \tfrac{1}{1/2^5} = 32
\end{aligned}$$

We call a LR of 8, moderately strong evidence and a LR of 32 strong evidence. Another way of looking at this calibration is to note that

$$2^b = k = \text{LR} \implies b = \frac{\ln(k)}{\ln(2)}$$

where b is the number of white balls drawn. This leads to the following table

LR	2	4	8	10	16	20	32	50	64	100	128	256	512	1000	1028
b	1	2	3	3.3	4	4.3	5	5.6	6	6.6	7	8	9	9.97	10

6 Operational Aspects

Does the Law of Likelihood work? Note that evidence, properly interpreted, can be misleading, but strong misleading evidence cannot occur very often.

Theorem 6.1 Universal Bound on the Probability of Misleading Evidence

$$P_A\left\{\frac{f_B(X)}{f_A(X)} \geq k\right\} \leq \frac{1}{k}$$

Proof: Let

$$S = \left\{x : \frac{f_B(x)}{f_A(x)} \geq k\right\}$$

Then

$$\begin{aligned} 1 &= \int f_B(x)\mu(dx) \\ &= \int_S f_B(x)d\mu(x) + \int_{S^c} f_B(x)\mu(dx) \\ &\geq \int_S k f_A(x)\mu(dx) + \int_{S^c} f_B(x)\mu(dx) \end{aligned}$$

It follows that

$$1 - \int_{S^c} f_B(x)\mu(dx) \geq k \int_S f_A(x)\mu(dx)$$

or

$$\int_S f_B(x)\mu(dx) \geq k \int_S f_A(x)\mu(dx)$$

i.e.

$$\int_S f_A(x)\mu(dx) \leq \frac{1}{k}$$

In fact a much stronger result is true. Consider a sequence of observations $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ such that if A is true then $\mathbf{X}_n \sim f_n$ and when B is true $\mathbf{X}_n \sim g_n$. The likelihood ratio

$$\frac{g_n(\mathbf{x}_n)}{f_n(\mathbf{x}_n)} = z_n$$

is the LR in favor of B after n observations. Then we have the following theorem.

Theorem 6.2 If A is true then

$$P_A(Z_n \geq k \text{ for some } n = 1, 2, \ldots) \leq \frac{1}{k}$$

Proof: Let N be the first n greater than or equal to 1 such that

$$g_n(\mathbf{X}_n) \geq k f_n(\mathbf{X}_n)$$

and define $N = \infty$ if no such n occurs. Let

$$\begin{aligned} S_n &= \{\mathbf{z}_n : N = n\} \\ &= \left\{\mathbf{x}_n : z_j = \frac{g_j(\mathbf{x}_j)}{f_j(\mathbf{x}_j)} < k \ ; \ j = 1, 2, \ldots, n-1 \text{ and } \frac{g_n(\mathbf{x}_n)}{f_n(\mathbf{x}_n)} \geq k\right\} \end{aligned}$$

Then we have

$$\begin{aligned}
\Pr_A(Z_n \geq k \text{ for some } n \geq 1) &= P_A(N < \infty) \\
&= \sum_{n=1}^{\infty} P_A(N = n) \\
&= \sum_{n=1}^{\infty} \int_{S_n} f_n(\mathbf{x}_n)\mu(\mathbf{dx}_n) \\
&\leq \frac{1}{k} \sum_{n=1}^{\infty} \int_{S_n} g_n(\mathbf{x}_n)\mu(\mathbf{dx}_n) \\
&= \frac{1}{k} P_B(N < \infty) \\
&\leq \frac{1}{k}
\end{aligned}$$

Robbins, H. Statistical Methods Related to the Law of the Iterated Logarithm. Annals of Mathematical Statistics 1970 Vol. 41, No. 5, 1397-1409.

Theorem 6.3 If A is true then

$$E_A\left[\frac{f_A(X)}{f_B(X)}\right] > E_A\left[\frac{f_B(X)}{f_A(X)}\right] = 1$$

Proof:

$$\begin{aligned} E_A\left[\frac{f_A(X)}{f_B(X)}\right] &= \int \left[\frac{f_A(x)}{f_B(x)}\right] f_A(x)\mu(dx) \\ &= \int \left[\frac{f_A(x)}{f_B(x)}\right]^2 f_B(x)\mu(dx) \\ &\geq \left\{\int \left[\frac{f_A(x)}{f_B(x)}\right] f_B(x)\mu(dx)\right\}^2 \\ &= 1 \end{aligned}$$

by Jensen's inequality. Note that the inequality is strict if $f_A \neq f_B$.

example 6.1 Suppose we select a card at random from a deck containing the customary 52 cards. We observe that it is the four of clubs. Consider the two hypotheses:

$$H_N: \text{ normal deck of cards } \Longrightarrow P(4\clubsuit) = 1/52$$
$$H_A: \text{ all cards are } 4\clubsuit \Longrightarrow P(4\clubsuit) = 1$$

Thus the observation supports A over N by a factor of 52 which is very strong evidence.

Is this strength of evidence reasonable? Suppose that the deck has a prior probability π of being normal and, if it is not normal, all of the 52 hypotheses defined by

$$\text{all } 4\clubsuit, \text{ all } 2\diamondsuit, \ldots$$

are equally likely i.e.

$$P(H_1) = P(H_2) = \cdots = P(H_{52}) = \frac{1-\pi}{52}$$

Then we have, by Bayes Theorem

$$P(N|4\clubsuit) = \frac{\frac{1}{52}\pi}{\frac{1}{52}\pi + \frac{1-\pi}{52} + 0\frac{1}{52} + \cdots + 0\frac{1}{52}} = \pi$$
$$P(H_1|4\clubsuit) = \frac{\frac{1-\pi}{52}}{\frac{1-\pi}{52} + \frac{1}{52}\pi + 0\frac{1}{52} + \cdots + 0\frac{1}{52}} = 1-\pi$$

It follows that the likelihood ratio is given by

$$\frac{P(H_1|4\clubsuit)}{P(N|4\clubsuit)} = \frac{1-\pi}{\pi} = 52\,\frac{\frac{1-\pi}{52}}{\pi} = 52 \times \text{prior odds}$$

Consider now H_N and H_1 and suppose that we observe the two of diamonds in our draw from the deck. Is the two of diamonds evidence against H_N. i.e. does low probability under H_N make an observation evidence against H_N. The answer to this question is NO.

To see this we note that it is equivalent to asking whether low probability under H_N relative to some alternative makes an observation evidence against H_N.

- It is evidence against H_N vis-a-vis H_2.
- It is evidence for H_N vis-a-vis H_1
- But it is not evidence against H_N.

Theorem 6.4 Further Properties of Likelihood Ratios Suppose $A \Longrightarrow X \sim p_A$ and $B \Longrightarrow X \sim p_B$. Assume we observe $x_1, x_2, \ldots, x_n$ which are iid. The likelihood ratio is

$$\mathrm{LR} = \frac{\mathrm{L}_A}{\mathrm{L}_B} = \frac{\prod_{i=1}^n p_A(x_i)}{\prod_{i=1}^n p_B(x_i)} = \prod_{i=1}^n \frac{p_A(x_i)}{p_B(x_i)}$$

and we have

- If A is true then $\mathrm{LR} \xrightarrow{a.s.} \infty$
- If B is true then $\mathrm{LR} \xrightarrow{a.s.} 0$
- Finally if B is true then for $\epsilon > 0$

$$\begin{aligned} P_B\left(\frac{\mathrm{L}_A}{\mathrm{L}_B} > \epsilon\right) &\longrightarrow 0 \\ P_B\left(\frac{\mathrm{L}_A}{\mathrm{L}_B} < \epsilon\right) &\longrightarrow 1 \\ P_B\left(\frac{\mathrm{L}_B}{\mathrm{L}_A} > \frac{1}{\epsilon}\right) &\longrightarrow 1 \end{aligned}$$

7 Testing Statistical Hypotheses

Recall that in the Neyman Pearson theory the best test (in terms of size and power) calls for rejecting H_1 in favor of H_2 if the likelihood ratio is large. Does this mean that observations are evidence supporting H_2 over H_1?

Consider the following situation:

$$n = 30 \text{ iid Bernoulli trials with parameter } \theta$$

and

$$H_1 : \theta = \frac{1}{4} \quad , \quad H_2 : \theta = \frac{3}{4}$$

The uniformly most powerful test of size $\alpha = .05$ rejects H_1 in favor of H_2 if

$$\frac{\mathrm{L}_2}{\mathrm{L}_1} = \frac{\left(\frac{3}{4}\right)^{\sum_{i=1}^{30} x_i} \left(\frac{1}{4}\right)^{30-\sum_{i=1}^{30} x_i}}{\left(\frac{1}{4}\right)^{\sum_{i=1}^{30} x_i} \left(\frac{3}{4}\right)^{30-\sum_{i=1}^{30} x_i}} = 3^{\sum_{i=1}^{30} x_i} \left(\frac{1}{3}\right)^{30-\sum_{i=1}^{30} x_i} = 3^{2s_{30}-30} \geq k$$

where $s_{30} = \sum_{i=1}^{30} x_i$ and k is chosen so that

$$P_{H_1}(S_{30} \geq k) = .05$$

Note that if $k = 12$ we have

$$P_{H_1}(S_{30} \geq 12) = 0.05$$

Hence we reject H_1 in favor of H_2 if $S_{30} = 12, 13, 14, 15, \ldots, 30$. However, if $S_{30} = 12$ then the likelihood ratio is

$$\text{LR} = 3^{2(12)-30} = 3^{-6} = \frac{1}{729}$$

so that we have very strong evidence in favor of H_1 over H_2. Note also that we have the following:

$$\begin{array}{c} \text{if } S_{30} = 13 \implies \text{LR} = \frac{1}{81} \\ \text{if } S_{30} = 14 \implies \text{LR} = \frac{1}{9} \\ \text{if } S_{30} = 15 \implies \text{LR} = 1 \end{array}$$

Thus 15 leads to equal support for H_1 vis-a-vis H_2 even though the Neyman Pearson theory says to reject H_1 in favor of H_2. The P-value for $S_{30} = 15$ is $P_{H_1}(S_{30} \geq 15) = .003$ which would lead, under conventional theory, to the conclusion that H_1 is not tenable.

As another example consider

$$H_1 : \ \theta = \frac{1}{4} \ , \ H_2 : \ \theta = \frac{1}{2} \ , \ H_3 : \ \theta = \frac{3}{4}$$

where X is binomial with $n = 5$ and we observe $X = 0$. Then we have

$$\frac{\mathrm{L}_1}{\mathrm{L}_2} = \frac{\left(\frac{3}{4}\right)^5}{\left(\frac{1}{2}\right)^5} = \left(\frac{3}{2}\right)^5 = 7.6$$

which is pretty strong evidence supporting H_1 over H_2. We also note that

$$\frac{\mathrm{L}_2}{\mathrm{L}_3} = \frac{\left(\frac{1}{2}\right)^5}{\left(\frac{1}{4}\right)^5} = 2^5 = 32$$

which is very strong evidence supporting H_2 over H_3.

Consider now the composite hypothesis

$$H_c : \ \theta = \frac{1}{4} \ \text{ or } \ \theta = \frac{3}{4}$$

Do we have evidence for $\theta = \frac{1}{2}$ vis-a-vis H_c? No.

The Law of Likelihood does not answer this question because H_c does not imply that X has a particular probability model i.e. it makes no prediction of the observations.

example 7.1 Suppose that

$$H_1 : \ \theta = 0.2 \ \ , \ \ H_2 : \ \theta = 0.8$$

and we observe 17 subjects with 9 successes. What do these observations tell us about θ?

The answer is provided by the likelihood function given by

$$\widetilde{\mathrm{L}}(\theta) = \frac{\mathrm{L}(\theta)}{\max[\mathrm{L}(\theta)]} = \frac{\theta^9(1-\theta)^8}{\left(\frac{9}{17}\right)^9\left(\frac{8}{17}\right)^8}$$

in the sense that

$$\frac{\widetilde{\mathrm{L}}(\theta_1)}{\widetilde{\mathrm{L}}(\theta_2)}$$

measures the strength of evidence for θ_1 vs θ_2.

The following graph shows the likelihood function for this situation. For convenience the interval of parameter values where $\tilde{L}(\theta)$ is greater than or equal to 8 and 32 are shown.

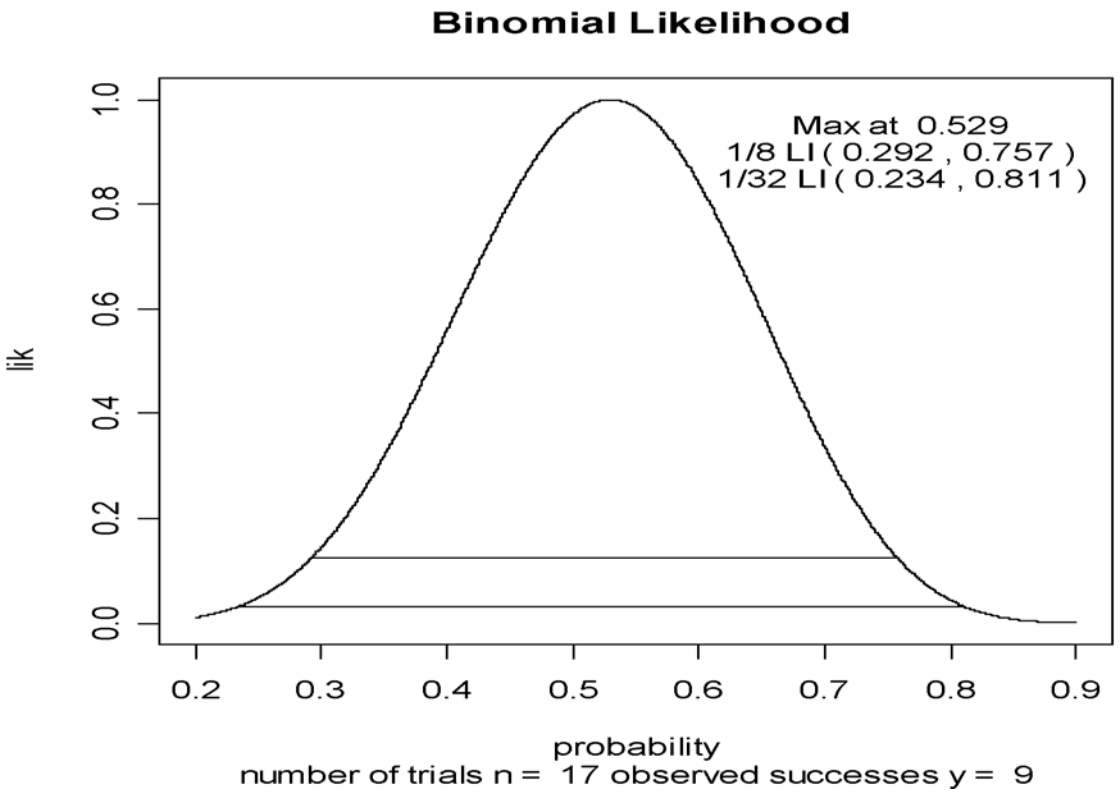

Figure 1:

The graph shows the strength of evidence in the observations. Do we have

- Evidence that $\theta > 0.20$?
- Evidence that $H_2 : \ \theta > 0.20$ vs $H_1 : \ \theta \leq 0.20$?

The answer is provided by looking at the likelihood function. Can answer yes in the sense that

- Values near $\theta = 0.5$ are much better supported than 0.20 or any smaller value.
- Note however, that $\theta = 0.20$ is better supported than $\theta = 0.9$ or any larger value.

Now look at some different evidence: we observe 20,500 successes in 100,000 trials. In this case the likelihood is given by

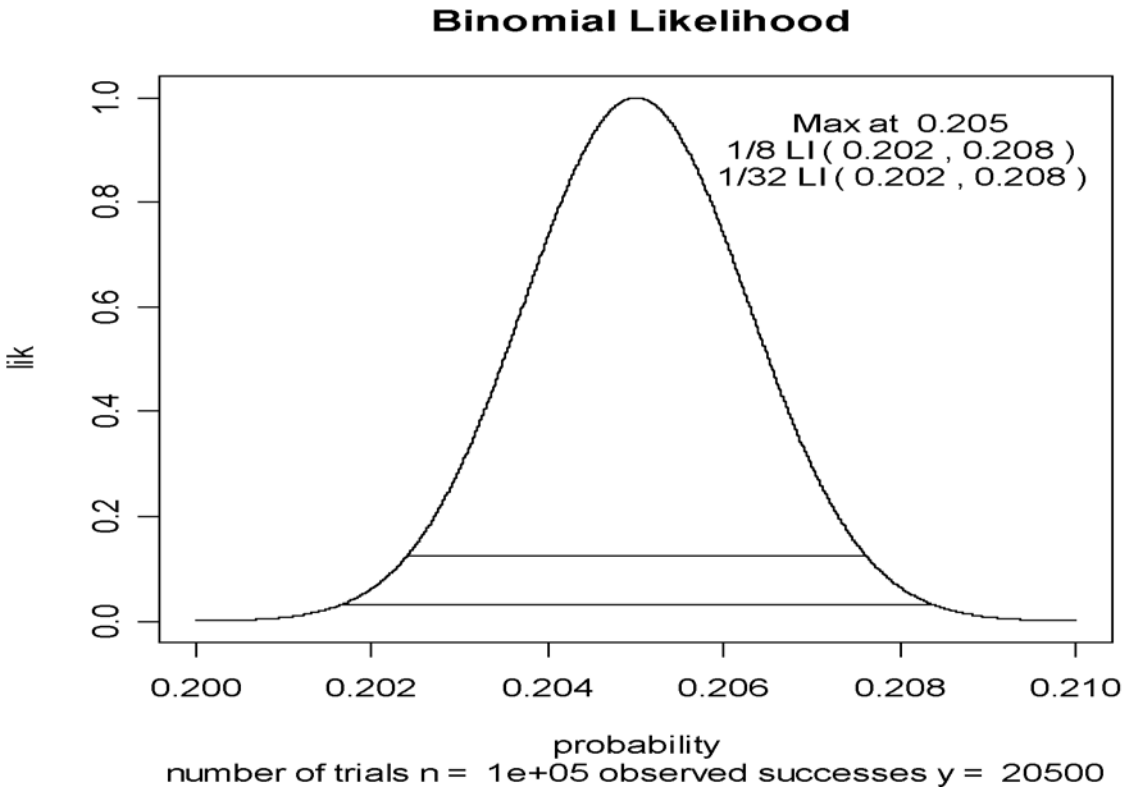

Figure 2:

In this case we have

- Strong evidence that θ is greater than 0.2 but it is also strong evidence that θ is not much larger than 0.2
- Looking at the likelihood function tells the complete story
- The likelihood function shows the evidence in the observations about the parameter.

8 Irrelevance of the Sample Space

Consider the following three models and hypotheses

(1) $A \Longrightarrow X \sim f_A\,, B \Longrightarrow X \sim f_B$ where

	Sample Space	
Model	1	0
f_A	.95	.05
f_B	.02	.98

(2) $A \Longrightarrow X \sim f_A\,, B \Longrightarrow X \sim f_B$ where

	Sample Space		
Model	1	2	3
f_A	.95	.02	.03
f_B	.02	.22	.76

(3) $A \Longrightarrow X \sim f_A\,, B \Longrightarrow X \sim f_B$ where

	Sample Space			
Model	1	6	7	2000
f_A	.95	.02	.01	.02
f_B	.02	.95	.00	.03

Is the evidence in the observation $X = 1$ different in the three situations?

- It doesn't matter (for interpreting the observation $X = 1$ as evidence for A vs B) what values X might have taken if it had not been 1
- or how the probability that $X \neq 1$ is spread out over the other values.

example 8.1 Suppose that we perform 20 independent Bernoulli trials in Afghanistan to learn something about a parameter θ.

- You know the language so that you will observe

$$X \sim \text{ binomial}(20, \theta)$$

- I know only the word for six and nothing else so I will observe

$$Y = 6I(X = 6) \Longrightarrow \frac{Y}{6} \sim \text{ Bernoulli}\left(1, \binom{20}{6}\theta^6(1-\theta)^{14}\right)$$

Suppose that an assistant reports 6 successes. For you the LR for comparing θ_1 to θ_2 is

$$\frac{\binom{20}{6}\theta_1^6(1-\theta_1)^{14}}{\binom{20}{6}\theta_2^6(1-\theta_2)^{14}} = \frac{\mathrm{P}(X = 6, \theta_1)}{\mathrm{P}(X = 6, \theta_2)}$$

For me, the same likelihood is obtained so that

$$\frac{\mathrm{P}(X = 6, \theta_1)}{\mathrm{P}(X = 6, \theta_2)} = \frac{\mathrm{P}(Y = 6, \theta_1)}{\mathrm{P}(Y = 6, \theta_2)}$$

Thus, according to the Law of Likelihood, we have the same evidence for comparing θ_1 to θ_2

Your sample space is $\{0, 1, 2, \ldots, 20\}$ mine is $\{6, \text{not } 6\}$. However, if we test $H_1 : \theta = .5$ vs $H_2 : \theta = .2$ your P-value is

$$P(X \leq 6; \theta = .5) = 0.06$$

while mine is

$$P(Y = 6; \theta = .5) = 0.04$$

Thus we have the following conclusions:

- P-values violate "irrelevance of the sample space" for determining evidence.
- The P-value (as a measure of the strength of evidence) is wrong because it says that the evidence is stronger for me than you but it is not.
- Evidence is independent of the sample space, it depends only on the observed point and its probability under various hypotheses.

Still a third possibility in the above scenario is that we observe the number of trials Z until we obtain 6 successes. In this case

$$\begin{aligned} P(Z=z) &= P(5 \text{ successes in } z-1 \text{ trails})P(\text{next trial a success}) \\ &= \binom{z-1}{5}\theta^5(1-\theta)^{z-1-5}\theta \\ &= \binom{z-1}{5}\theta^6(1-\theta)^{z-6} \end{aligned}$$

If $Z = 20$ we obtain

$$P(Z=20) = \binom{19}{5}\theta^6(1-\theta)^{14}$$

which leads to a LR for comparing θ_1 to θ_2 of

$$\frac{\theta_1^6(1-\theta_1)^{14}}{\theta_2^6(1-\theta_2)^{14}}$$

which implies the "irrelevance of the stopping rule".

It follows that $X = 6, Y = 6$ and $Z = 6$ are equivalent as evidence about θ because the strength of evidence for θ_1 vs θ_2 as given by the Law of Likelihood is the same. That is these three instances of statistical evidence all generate the same likelihood function given by

$$\mathrm{L}(\theta) \propto \theta^6(1-\theta)^{14}$$

In general: If $X \sim f_X(x;\theta)$ for $x \in \mathcal{X}$ and $\theta \in \Theta$ the observation $X = x_0$ implies that the LF satisfies

$$\mathrm{L}_X(\theta, x_0) = c(x_0) f_X(x_0;\theta) \propto f_X(x_0;\theta)$$

- Before the observation is recorded $f_X(x;\theta)$ represents uncertainty about what value of X will be observed.
- After the observation of $X = x_0$

  $$f_X(x_0;\theta) \propto \mathrm{L}_X(\theta; x_0)$$

  represents the evidence about θ in x_0.

The Law of Likelihood gives the likelihood function its meaning in the sense that

$$\frac{\mathrm{L}(\theta_1; x_0)}{\mathrm{L}(\theta_2; x_0)}$$

is the strength of the evidence in x_0 for comparing θ_1 to θ_2.

9 Likelihood Principle

Suppose that

$$Y \sim f_Y(y;\theta)\ ;\ y \in \mathcal{Y}\ \ \theta \in \Theta$$

If $Y = y_0$ is observed then the likelihood function satisfies

$$\mathrm{L}(\theta; y_0) \propto f_Y(y_0;\theta)$$

If it happens that

$$f_Y(y_0;\theta) \propto f_X(x_0;\theta)$$

then $X = x_0$ and $Y = y_0$ determine the same likelihood function. Thus for any pair of values θ_1 and θ_2 the evidence in $X = x_0$ and $Y = y_0$ is the same. This fact is called the Likelihood Principle.

Likelihood Principle: If $X = x_0$ and $Y = y_0$ determine the same likelihood function then $X = x_0$ and $Y = y_0$ are equivalent as evidence about θ.

Key concepts:

- Two instances of statistical evidence are equivalent if and only if they generate the same likelihood function (LF).
- This implies that the likelihood function is the mathematical representation of the concept of statistical evidence.

In the example of the previous section frequentist statistics says that $6/20$ is an unbiased estimate if it is $X/20$ but not if it is $6/Z$ because

$$E\left(\frac{X}{20}\right) = \theta \text{ but } E\left(\frac{6}{Z}\right) \geq \frac{6}{E(Z)} = \frac{6}{6 + 6\frac{1-\theta}{\theta}} = \theta$$

Actually the inequality is strict since $\frac{1}{X}$ is strictly convex unless Z is degenerate (which it isn't in this case).

Similarly we have that

- s.e. $\left(\frac{X}{20}\right) \neq$ s.e. $\left(\frac{6}{Z}\right)$
- $P_\theta(X \leq 6) \neq P_\theta(Z \geq 20)$
- The best 95% confidence interval for θ is different in both cases.

The only sensible conclusion is that standard inference methods are all defective (for the purpose of representing or interpreting the data as evidence about θ). The reasons are that

- Standard methods assert that the evidence about θ in $X = 6$ is different from that in $Z = 20$
- Standard methods thus violate the likelihood principle.

10 Misconceptions and Misinterpretations

(1) "The likelihood principle says that conclusions based on $X = 6$ should be the same as those based on $Z = 20$. i.e. everyone faced with the same likelihood should give the same point estimate, (P-value, etc.)."

Conclusions and decisions are not likelihood concepts i.e. they are not evidence.

(2) "The likelihood principle is a rule for data reduction."

Consider the following quote from Casella and Berger:

> "The Likelihood Principle specifies how the likelihood function should be used as a data reduction device.
>
> LIKELIHOOD PRINCIPLE: If x and y are two sample points such that $\mathrm{L}(\theta|x)$ is propotional to $\mathrm{L}(\theta|y)$, that is, there exists a constant $C(x, y)$ such that
>
> $$\mathrm{L}(\theta|x) = C(x, y)\mathrm{L}(\theta|y) \text{ for all } \theta$$
>
> then the conclusions drawn from x and y should be identical."

Again this is incorrect since the likelihood principle appies to evidence not conclusions.

(3) "The likelihood principle applies only when you have a parametric model that you know to be true."

This is incorrect. No model is true, the best we can do is state the evidence based on the model assumptions and investigate the consequences of model inadequacy.

11 Relation to Bayes and Decision Theory

Consider the model

$$X \sim f_X(x;\theta) \;\; ; \;\; X \in \mathcal{X} \;\; ; \;\; \theta \in \Theta$$

and the effect of statistical evidence on the state of uncertainty about θ. We assume that before the evidence is obtained our uncertainty about θ is expressed as a pdf $g(\theta)$.

After the evidence $X = x$ is obtained Bayes theorem implies that our state of uncertainty is given by

$$\begin{aligned} g(\theta|x) &= \frac{f(x;\theta)g(\theta)}{\int_\Theta f(x;\theta)g(\theta)\mu(d\theta)} \\ &= \frac{c\mathrm{L}(\theta;x)g(\theta)}{\int_\theta c\mathrm{L}(\theta;x)g(\theta)\mu(d\theta)} \\ &= \frac{\mathrm{L}(\theta;x)g(\theta)}{\int_\theta \mathrm{L}(\theta;x)g(\theta)\mu(d\theta)} \end{aligned}$$

Thus evidence effects the state of uncertainty only through the likelihood function.

It follows that two instances of statistical evidence that generate the same likelihood function have the same effect on the state of uncertainty about θ.

It also follows that the likelihood principle applies to areas of statistics answering the question - What do I (or should I) believe?

With regard to decision theory (What should I do?), suppose we have a model and a prior which generates $g(\theta|x)$, an action space $\mathcal{A}$ and a loss function defined by

$L(a, \theta) =$ loss if we take action a and θ is the true value of the parameter

The expected loss (risk) if we take action a after observing $X = x$ is

$$R(a|x) = \int_\Theta L(a, \theta) g(\theta|x) \mu(d\theta)$$

Decision theory chooses a to minimize $R(a|x)$.

Again the data enter the analysis only through the likelihood function.

Thus two instances of statistical evidence that produce the same likelihood function should produce the same action in a given problem.

12 Evidence and Uncertainty

The Law of Likelihood implies that evidence has a different mathematical form (likelihood) than uncertainty (defined by a probability distribution).

12.1 Example 1

To illustrate suppose we are interested in the probability of success θ (heads) when I toss a pnickel (a coin formed by joining a penny and a nickle). To learn about θ we toss a nickel 20 times and it is a success 9 times.

Model : $Y \sim \text{Bernoulli}(\theta)$, $X \sim \text{binomial}(20, 9, 1/2)$ for $0 \leq \theta \leq 1$

Observation : $X = 9$

To represent the evidence about θ we look at the likelihood function which is

$$\text{L}(\theta) = P_{X,Y}(X = 9; \theta) = \binom{20}{9} \left(\frac{1}{2}\right)^{20}$$

which is a constant as a function of θ. Normalizing by the maximum the likelihood function is

$$\text{L}(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

This likelihood function represents the absence of evidence about θ.

This same function can also be a probability density function defined by

$$f_\Theta(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

As a pdf, $f_\Theta(\theta)$ represents a particular state of knowledge about θ i.e.

$$P\left(\Theta > \frac{1}{2}\right) = \frac{1}{2} \quad , \quad P\left(\Theta < \frac{1}{8}\right) = \frac{1}{8}$$

and so on.

Two Questions:

(i) Does this pdf represent ignorance (the absence of knowledge about θ?

(ii) Does it represent what $X = 9$ tells us about θ?

Suppose we look at another parameter $\phi = \theta^2$, the probability of heads on two tosses of the pnickel. Note that ϕ is equivalent to θ since it is a one to one function on this parameter space.

What is the evidence about ϕ in the observation $X = 9$? The evidence about $\theta = \sqrt{\phi}$ is given by

$$L_1(\theta) = c$$

and the evidence about ϕ is given by

$$L_2(\phi) = L_1(\sqrt{\phi}) = L_1(\theta) = c$$

Thus there is no evidence (or neutral evidence) about ϕ. The observation $X = 9$ tells us nothing about θ and it tells us nothing about $\theta^2 = \phi$ (or any other function of θ).

What is the uncertainty about ϕ? Since $\phi = \theta^2$ we have that the pdf for ϕ is given by

$$f_\Phi(\phi) = f_\Theta(\theta(\phi))\frac{d\theta}{d\phi} = \begin{cases} \frac{1}{2\sqrt{\phi}} & \text{if } 0 \le \phi \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

We now note that

$$P\left(\Phi < \frac{1}{2}\right) = \int_0^{\frac{1}{2}} \frac{1}{2\sqrt{\phi}} d\phi = \phi^{\frac{1}{2}}\Big|_0^{\frac{1}{2}} = \frac{1}{\sqrt{2}}$$

which is equal to the probability that θ is less than $\frac{1}{\sqrt{2}}$.

If ignorance about a probability θ is represented by the uniform (0,1) pdf then we are not ignorant of the probability $\Phi = \theta^2$ because its distribution is not uniform.

- If the uniform (or any other) probability distribution represents ignorance, then we can be ignorant of only one function of θ.
- But ignorance (the absence of knowledge) about θ is equivalent to the ignorance about $\phi = \theta^2$ (or any other function of θ.
- It follows that a probability distribution represents a particular, specific state of knowledge. It cannot and does not represent ignorance.

A likelihood function, on the other hand, represents evidence. It does not represent uncertainty.

To make this explicit $\mathrm{L}(\theta)$ looks like a pdf. Since the scale is arbitrary why not scale so that the area under the likelihood function is one, and treat it as a pdf? (This is the integrated likelihood approach) Why not measure the "evidence" for $H_0 : \theta \leq \frac{1}{2}$ by

$$\int_o^{\frac{1}{2}} \mathrm{L}_1^*(\theta) d\theta = \frac{1}{2}$$

where $\mathrm{L}_1^*(\theta)$ is the rescaled version of L?

This approach would lead to inconsistencies i.e.

$$\theta \leq \frac{1}{2} \iff \phi \leq \frac{1}{4}$$

But the rescaled likelihood for ϕ would have evidence for the hypothesis $H_0 : \ \phi \leq \frac{1}{4}$ given by

$$\int_o^{\frac{1}{4}} \mathrm{L}_2^*(\theta) d\theta = \frac{1}{4}$$

and we have two equivalent hypotheses, the same data and yet we assign different evidence to them.

It follows that a likelihood function cannot be integrated. When you change variables (reparametrize) in a likelihood function the new likelihood function is obtained by simple substitution i.e.

$$\mathrm{L}_2(\phi) = \mathrm{L}(\theta(\phi))$$

There is no Jacobian as there is when dealing with pdfs. This is essential to maintain consistency with the Law of Likelihood.

Note that the support (in $X = 9$) for $\theta = \frac{1}{2}$ vs $\theta = \frac{1}{4}$ is given by

$$\frac{\mathrm{L}_1\left(\frac{1}{2}\right)}{\mathrm{L}_1\left(\frac{1}{4}\right)} = 1$$

The support for the equivalent hypotheses $\phi_1 = \theta_1^2 = \frac{1}{4}$ vs $\phi_2 = \theta_2^2 = \frac{1}{16}$ is given by

$$\frac{\mathrm{L}_2\left(\frac{1}{4}\right)}{\mathrm{L}_2\left(\frac{1}{16}\right)} = 1$$

which would be destroyed if

$$\mathrm{L}_2(\phi) = \mathrm{L}_1(\theta(\phi))\frac{d\theta}{d\phi}$$

In this example we have seen that

no evidence for θ_1 vs θ_2 is equivalent to no evidence for ϕ_1 vs ϕ_2

but

equal pdfs at θ_1 and θ_2 is not equivalent to equal pdfs at ϕ_1 and ϕ_2

i.e.

$$f_\Theta(\theta_1) = f_\Theta(\theta_2)$$

and yet

$$f_\Phi(\phi_1) = f_\Theta(\theta(\phi_1)) \left[\frac{d\theta(\phi)}{d\phi}\right|_{\phi=\phi_1} \neq f_\Phi(\phi_2) = f_\Theta(\theta(\phi_2)) \left[\frac{d\theta(\phi)}{d\phi}\right|_{\phi=\phi_2}$$

Thus we have the following implications:

- The likelihood function represents evidence. It does represent uncertainty.
- A pdf represents uncertainty. It does not represent evidence.

It follows that no state of uncertainty (no pdf) represents ignorance (absence of evidence).

In summary:

- $\mathrm{L}(\theta)$ represents evidence about θ and evidence about a function of θ, $\phi(\theta)$ is represented by
$$\mathrm{L}_2(\phi) = \mathrm{L}(\theta(\phi))$$
- $f(\theta)$ represents a state of knowledge or uncertainty about θ. The state of knowledge about a function of θ, $\phi(\theta)$ is represented by the pdf
$$g(\phi) = f(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right|$$

Thus the curve
$$\frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x-\theta)^2 \right\}$$
has a different meaning and different mathematical properties (as a function of θ) depending on whether it is a pdf $f(\theta)$ or a likelihood function $\mathrm{L}(\theta)$. In the first case it represents uncertainty. In the second case it represents evidence.

12.2 Example 2

Suppose that X and Y are jointly distributed random variables. We are interested in the value of Y in a realization (x, y). We can only observe the value of X.

- The joint density $f_{X,Y}(\,\cdot\,,\,\cdot\,)$ represents the uncertainty about the pair (X, Y)
- The marginal density

$$f_Y(\,\cdot\,) = \int f_{X,Y}(x,\,\cdot\,)d\mu(x)$$

  represents the uncertainty about Y.

- After X is observed the conditional density of Y given X

$$f_{Y|X}(\,\cdot\,|x) = \frac{f_{X,Y}(x,\,\cdot\,)}{f_X(x)}$$

  represents the uncertainty about Y.

- After $X = x$ is observed $f_{X|Y}(x|y)$ is the likelihood function. It represents the evidence in the observation $X = x$ about y.

As a specific example suppose that (X, Y) is bivariate normal i.e.

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_Y^2 \end{bmatrix} \right)$$

where we assume that all parameters are known.

We are interested in the value of y of Y (realized but not observed). Before making any observation, uncertainty about y is described by a $\text{N}(\mu_y, \sigma_y^2)$ probability distribution i.e. by

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left\{ -\frac{1}{2\sigma_y^2}(y - \mu_y)^2 \right\}$$

If we observe $X = x$ what is our knowledge (or belief) in the state of uncertainty about y? It is described by the conditional distribution of Y given $X = x$ i.e. by a normal distribution with mean equal to $\mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x)$ and variance equal to $\sigma_y^2(1 - \rho_{xy}^2)$ where ρ_{xy} is the correlation between X and Y.

The density is thus given by

$$f_{Y|X}(y|x) == \frac{1}{\sqrt{2\pi\sigma_y^2(1-\rho_{xy}^2}} \exp\left\{-\frac{1}{2\sigma_y^2(1-\rho_{xy}^2)}\left[y-\mu_y-\frac{\sigma_{xy}}{\sigma_x^2}(x-\mu_x)\right]^2\right\}$$

The observation $X = x$ is evidence about y. How do we represent and interpret that evidence?

We use the likelihood function given by

$$\mathrm{L}(y) \propto f_{X|Y}(x|y) = \mathrm{N}\left(\mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y-\mu_y)\ ,\ \sigma_x^2(1-\rho_{xy}^2)\right)$$

Each value of y determines a probability distribution for X so the observation $X = x$ generates a likelihood function for y.

This likelihood function is given by

$$\begin{aligned}
\mathrm{L}(y) &\propto \exp\left\{-\frac{1}{2\sigma_x^2(1-\rho_{xy}^2)}\left[x-\mu_x-\frac{\sigma_{xy}}{\sigma_y^2}(y-\mu_y)\right]^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma_x^2(1-\rho_{xy}^2)}\left(\frac{\sigma_{xy}}{\sigma_y^2}\right)^2\left[\frac{\sigma_y^2}{\sigma_{xy}}(x-\mu_x)-(y-\mu_y)\right]^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma_y^2\frac{(1-\rho_{xy}^2)}{\rho_{xy}^2}}\left(\frac{\sigma_{xy}}{\sigma_y^2}\right)^2\left[y-\mu_y-\frac{\sigma_y^2}{\sigma_{xy}}(x-\mu_x)\right]^2\right\}
\end{aligned}$$

where we have assumed that $\sigma_{xy} \neq 0$.

The ratio $\mathrm{L}(y_1)/\mathrm{L}(y_2)$ measures the support or evidence for y_1 vs y_2 given that $X = x$ was observed.

To see the difference between a likelihood function and a pdf in this case consider the situation where $\sigma_{xy} = 0$. In this case

$$\mathrm{L}(y) \propto \text{constant i.e.} \quad \frac{\mathrm{L}(y_1)}{\mathrm{L}(y_2)} = 1 \text{ for all } y_1, y_2$$

so that there is no evidence in $X = x$ about the value of y.

On the other hand, the conditional distribution of Y given $X = x$ is in this case

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left\{-\frac{1}{2\sigma_y^2}(y - \mu_y)^2\right\}$$

and it represents the uncertainty about Y, but not the evidence in $X = x$ about y.

Evidence acts on uncertainty. **Evidence is, literally, that which alters the state of uncertainty**.

13 Summary

- The Law of Likelihood answers the question; "When do observations provide evidence for one hypothesis vis a vis another"
- The Law of Likelihood is intuitive, consistent with probability theory and using the likelihood function defines statistical evidence by the likelihood ratio.
- Examples show that the customary methods of modern frequentist statistics do not measure evidence.
- Evidence is measured not by probabilities but by likelihood ratios.