OVERVIEW OF PRINCIPLES OF STATISTICS

F. James
CERN, CH-1211 Geneva 23, Switzerland

Abstract
A summary of the basic principles of statistics. Both the Bayesian and Frequentist points of view are exposed.

1 The Problems that Statistics is supposed to Solve.

Statistical problems can be grouped into five classes:

- *Point Estimation:* Find the "best" value for a parameter.
- *Interval Estimation:* Find a range within which the true value should lie, with a given confidence.
- *Hypothesis Testing:* Compare two hypotheses. Find which one is better supported by the data.
- *Goodness-of-Fit Testing:* Find how well one hypothesis is supported by the data.
- *Decision Making:* Make the best decision, based on data.

In the Frequentist methodology, this separation is especially important, and books on Statistics are often organized into chapters with just these titles. The reason for this importance is that often the same problem can be formulated in different ways so that it fits into different classes, but the fundamental question being asked is different in each class, so the resulting solution must be expected to be different. The lesson is: Make sure you know what question you want to ask, and then choose the appropriate methods for *that* question. And be aware that seemingly unimportant differences in the way a problem is posed can make large differences in the answer. The secret to getting the right answer is to understand the question.

In the Bayesian methodology, this separation is much less important, and Bayesian treatments tend not to be organized in this way. Bayes' Theorem is the concept which unifies Bayesian inference, since the methods for solving problems in all classes are based on the same theorem.

2 Probability

All statistical methods are based on calculations of *probability*.

In Mathematics, probability is an abstract (undefined) concept which obeys certain rules. We will need a specific operational definition. There are basically two such definitions we could use:

- *Frequentist probability* is defined as the *limiting frequency* of a particular outcome in a large number of identical experiments.
- *Bayesian probability* is defined as the *degree of belief* in a particular outcome of a single experiment.

2.1 Frequentist Probability

This probability of an event A is defined as the number of times A occurs, divided by the total number of trials, in the limit of a large number of identical trials:

$$P(A) = \lim_{N\to\infty} \frac{N(A)}{N}$$

where A occurs $N(A)$ times in N trials. Frequentist probability is used in most scientific work, because it is *objective.* It can (in principle) be determined to any desired accuracy and is the same for all observers. It is the probability of Quantum Mechanics.

Just like the definition of electric charge [1], the definition of frequentist probability is a conceptual definition which communicates clearly its meaning and can in principle be used to evaluate it, but in practice one seldom has to resort to such a primitive procedure and go experimentally to a limit (in the case of the electric field, it is even physically impossible to go to the limit because charge is quantised, but this only illustrates that the definition is more conceptual than practical).

However, even though one does not usually have to repeat experiments in order to evaluate probabilities, the definition does imply a serious limitation: It can only be applied to phenomena that are in principle exactly repeatable. This implies also that the phenomena must be random, that is: identical situations can give rise to different results, something we are accustomed to in Quantum Mechanics. There is great debate about whether macroscopic phenomena like coin-tossing are random or not; in principle coin-tossing is classical mechanics and the initial conditions determine the outcome, so it is not random. But such phenomena are usually treated as random; it is sufficient that the phenomenon behaves as though it were random: initial conditions which are experimentally indistinguishable yield results which are unpredictably different.

2.2 Bayesian Probability

This kind of probability is more general, since it can apply also to unrepeatable phenomena (for example, the probability that it will rain **tomorrow**). However, it depends not only on the phenomenon itself, but also on the state of knowledge and beliefs of the observer. Therefore, Bayesian P(A) will in general change with time. The probability that it will rain at 12:00 on Friday will change as we get closer to that date until it becomes either zero or one on Friday at 12:00.

We cannot verify if the Bayesian probability P(A) is "correct" by observing the frequency with which A occurs, since this is not the way probability is defined. The operational definition is based on "the coherent bet" (de Finetti [2]). O'Hagan [3] gives two different definitions, one of which is based on a comparison with the belief in the outcome of a process for which the frequentist probability is known.

There has been considerable effort (in particular, by Jeffreys) to develop an objective Bayesianism, but this is generally considered to be not entirely successful. Nearly all modern definitions of Bayesian probability are subjective, so we will consider here mainly subjective Bayesianism.

3 Fundamental Underlying Concepts

The **hypothesis** is what we want to test, verify, measure, decide.

Examples: H: The data are consistent with the Standard Model.
H: The mass of the proton is m_p (unknown)
H: Aspirin is effective in preventing heart disease

A **Random Variable** is data which can take on different values, unpredictable except in probability: $P(\text{data}|\text{hypothesis})$ is assumed known, provided any unknowns in the hypothesis are given some assumed values.

Example: for a Poisson process, N is a random variable taking on positive integer values, and P is the probability of observing N events when the expected rate is μ:

$$P(N|\mu) = \frac{e^{-\mu}\mu^N}{N!}$$

A **Nuisance parameter** is an unknown whose value does not interest us, but is unfortunately necessary for the calculation of $P(\text{data}|\text{hypothesis})$.

The **Likelihood Function** $\mathcal{L}$ is $P(\text{data}|\text{hypothesis})$ evaluated at the observed data, and considered as a function of the (unknowns in the) hypothesis.

3.1 Bayes' Theorem

We first need to define conditional probability: $P(A|B)$ means the probability that A is true, given that B is true. For example $P(\text{symptom}|\text{illness})$ such as $P(\text{headache}|\text{influenza})$ is the probability of the patient having a headache if she has influenza.

Bayes' Theorem says that the probability of both A and B being true simultaneously can be written: $P(A|B)P(B) = P(B|A)P(A)$ which implies:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

which can be written:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\text{not}B)P(\text{not}B)}$$

This theorem therefore tells us how to invert conditional probability to obtain $P(A|B)$ when we know $P(B|A)$.

Example of Bayes' Theorem

Suppose we have a test for influenza, such that if a person has flu, the probability of a positive result is 90%, and is only 1% if he doesn't have it:

$$P(T^+ \,|\text{flu}) = 0.9 \;\; \text{[10\% false negatives]}$$

$$P(T^+ \,|\text{not flu}) = 0.01 \;\; \text{[1\% false positives]}$$

Now patient P tests positive. What is the probability that he has the flu? The answer by Bayes' Theorem:

$$P(\text{flu}|T^+) = \frac{P(T^+ \,|\text{flu})P(\text{flu})}{P(T^+ \,|\text{flu})P(\text{flu}) + P(T^+ \,|\text{not flu})P(\text{not flu})}$$

So the answer depends on the **Prior Probability** of the person having flu, that is:

- for Frequentists, the frequency of occurence of flu in the general population.
- for Bayesians, the prior belief that the person has the flu, before we know the outcome of any tests.

If we are in the winter in Durham, perhaps $P(\text{flu})$ is 1% . On the other hand, we may be in another country where it is a very rare disease with $P(\text{flu}) = 10^{-6}$

If we apply the same diagnostic test in each of these two places, we would get the following probabilities:

	flu = 1%	flu $= 10^{-6}$
$P(\text{flu}\|T^+)$	0.48	10^{-4}
$P(\text{flu}\|T^-)$	0.001	10^{-7}

So this test would be useful for diagnosing the flu in Durham, but in another place where it was a rare disease it would always lead to the conclusion that the person probably does not have the flu even if the test is positive.

Note that, as long as all the probabilities are meaningful in the context of a given methodology, Bayes' Theorem can be used as well by Frequentists as by Bayesians. The use of Bayes' Theorem does not imply that a method is Bayesian, however the inverse is true: all Bayesian methods make use (at least implicitly) of Bayes' Theorem.

4 Point Estimation - Frequentist

Common notation: for all estimation (sections 4 – 6), we are estimating a parameter x using some data, and it is assumed that we know $P(data|x)$, which can be thought of as the Monte Carlo for the experiment, for any assumed value of x.

An **Estimator** is a function of the data which will be used to estimate (measure) the unknown parameter x. The problem is to find that function which gives estimates of x closest to the true value assumed for x. This can be done because we know $P(\text{data}|\text{true value of } x)$ and because the estimate is a function of the data. The general procedure would therefore be to take a lot of trial estimator functions, and for each one calculate the expected distribution of estimates about the assumed true value of x. [All this can be done without any experimental data.] Then the best (most efficient) estimator is the one which gives estimates grouped closest to the true value (having a distribution centred on the true value and as narrow as possible).

Fortunately, we don't have to do all that work, because it turns out that under very general conditions, it can be shown that the best estimator will be the one which maximizes the **Likelihood** $\mathcal{L}(x)$. This is the justification for the well-known method of Maximum Likelihood.

Note that the definition of the "narrowest distribution" of estimates requires specifying a norm for the width; the usual criterion, whereby the width is defined as the variance, leads to the Maximum Likelihood solution, since this is (asymptotically) the minimum-variance estimator.

An important and well-known property of the Maximum-likelihood estimate is that it is metric-independent: If the hat represents the Maximum-likelihood estimate, then $\hat{f}(x) = f(\hat{x})$.

5 Point Estimation - Bayesian

For parameter estimation, we can rewrite Bayes' Theorem:

$$P(\text{hyp}|\text{data}) = \frac{P(\text{data}|\text{hyp})P(\text{hyp})}{P(\text{data})}$$

and if the hypothesis concerns the value of x:

$$P(x|\text{data}) = \frac{P(\text{data}|x)P(x)}{P(\text{data})}$$

which is a **probability density function** in the unknown x. Since it is a **pdf**, it must be normalized: $\int_x P(x|\text{data}) = 1$, which determines $P(\text{data})$, considered now as a normalization constant.

Assigning names to the different factors, we get:

$$\text{Posterior pdf}(x) \propto \mathcal{L}(x) \times \text{Prior pdf}(x)$$

The Bayesian point estimate is usually taken as the maximum value of the Posterior *pdf*.

If the Prior *pdf* is taken to be the uniform distribution in x, then the maximum of the Posterior will occur at the maximum of $\mathcal{L}(x)$, which means that in practice the Bayesian point estimate is often the same as the Frequentist point estimate, although following a very different reasoning!

Note that the choice of a uniform Prior is not well justified in Bayesian theory (for example, it seldom corresponds to anyone's actual prior belief about x), so the best Bayesian solution is not necessarily the Maximum Likelihood.

Note also that the choice of the maximum of the posterior density has the unfortunate property of being dependent on the metric chosen for x. In particular, consider the "natural" metric, that function $f(x)$ in which the *pdf* $P[f(x)]$ is uniform between zero and one: in this metric P has no maximum. This problem is easily solved by choosing the point estimate corresponding to the median P (50th percentile) instead of the mode (maximum), but then it will not in general coincide with the Maximum Likelihood.

6 Interval Estimation - Bayesian

Here the goal is to find an interval which will contain the true value with a given probability, say 90%. Since the Posterior Probability distribution is known from Bayes' Theorem (see above), we have only to find an interval such that the integral under the Posterior *pdf* is equal to 0.9 . As this interval is not unique, the usual convention is to choose the interval with the largest values of the posterior *pdf*.

There are three arbitrary choices to be made in Bayesian estimation, and the most common choices are:

1. The uniform prior.
2. The point estimate as the maximum of the posterior *pdf*.
3. The interval estimate as the interval containing the largest values of the posterior *pdf*.

Note that *all these choices* produce metric-dependent results (they give a different answer under change of variables), but the first two happen to cancel to yield the metric-independent frequentist result.

A metric-independent solution is easily found for the third case, the most obvious possibility being the *central intervals*, defined such that there is equal probability above and below the confidence interval. However, this would have the unfortunate consequence that a Bayesian result could never be given as an upper limit: Even if no events are observed, the central Bayesian interval would always be two-sided with an upper and a lower limit.

7 Interval Estimation - Frequentist

Assuming as usual that we know $P(data|x)$, the goal is to find two functions of the data $F_1(\text{data}|x)$ and $F_2(\text{data}|x)$ such that, for any (true) value of x,

$$P(F_1 < x < F_2) = 0.9$$

Then the 90% interval is defined by $F_1(\text{observed data})$ and $F_2(\text{observed data})$. If we could find such functions, this would assure that:
If the experiment were repeated many times, and the data were treated using the functions F_1 and F_2 to define the interval, then the interval would contain the true value in 90% of the cases. This property is called *coverage*.

J. Neyman [4] showed how to construct such functions in the most general case, thereby solving the problem of how to find confidence intervals which have a given coverage. Since coverage is the most important property of confidence intervals, this was a very important milestone in frequentist statistics. Some comments:

1. Coverage alone does not determine the confidence intervals uniquely. There is another degree of freedom remaining, and this can be resolved in various ways, the most common being:
   - *Central intervals* are central *in the data*, not in the parameter, so they can as well produce upper limits as two-sided limits, and they have the nice property of being unbiased, but also the not-so-nice property that the interval can be empty (for example, an upper limit could be zero for a parameter that must be positive).
   - *Feldman-Cousins intervals* are the closest to central (least biased) among all intervals which are guaranteed to be non-empty. This is currently considered to be the state-of-the-art. The authors point out that these intervals are just standard frequentist intervals using the likelihood-ratio ordering based on the Neyman-Pearson criteria as given in Kendall and Stuart [5], however the paper by Feldman and Cousins [6] gives the best unified treatment.
   - *Ciampolillo intervals*[7] are the most biased but have the nice property that when no events are observed, the upper limit is independent of the expected background.

All the above have exact frequentist coverage when the data is continuous. For discrete data there is an additional problem that exact coverage is not always possible, so we have to accept some over-coverage.

2. The Neyman procedure in general, and in particular all of the three examples above are fully metric-independent in both the data and the parameter spaces.
3. The probability statement that defines the coverage of frequentist intervals appears to be a statement about the probability of the true value falling inside the confidence interval, but it is in fact the probability of the (random) confidence interval covering the (fixed but unknown) true value. That means that coverage is not a property of one confidence interval, it is a property of the ensemble of confidence intervals you could have obtained as results to your experiment. This somewhat unintuitive property causes considerable misunderstanding.

8 Hypothesis Testing - Frequentist

Compare two hypotheses to see which one better explains (predicts) the data. The two hypotheses are conventionally denoted: H_0 the null hypothesis; and H_1 the alternative hypothesis. If the hypotheses are **simple hypotheses**, they are completely specified so we know $P(\text{data}|H_0)$ and $P(\text{data}|H_1)$.

If W is the space of all possible data, the problem is to find a **Critical Region** (in which we reject H_0) $\omega \in W$ such that

$$P(\text{data} \in \omega|H_0) = \alpha$$

is as small as possible, and at the same time,

$$P(\text{data} \in W - \omega|H_1) = \beta$$

is also as small as possible.

α is the probability of rejecting H_0 when it is true. This is the error of the first kind, or **loss**. $1 - \alpha$ is the **acceptance** of the test. Some books interchange the definitions of α and $1 - \alpha$.

β is the probability of accepting H_0 when H_1 is true. This is the error of the second kind, or **contamination**. $1 - \beta$ is the **power** of the test.

When the two hypotheses are *simple hypotheses*, then it can be shown that the *most powerful test* is the Neyman-Pearson Test [8], which consists in taking as the critical region that region with the largest values of λ_0/λ_1, where λ_i is the likelihood under hypothesis H_i.

When a hypothesis contains unknown parameters, it is said to be not completely specified and is called a *composite hypothesis.* This important case is much more complicated than that of simple hypotheses, and the theory is less satisfactory, general results holding only asymptotically and under certain conditions. In practice, Monte Carlo calculations are required in order to calculate α and β exactly for composite hypotheses.

9 Hypothesis Testing - Bayesian

Recall that according to Bayes' Theorem:

$$P(\text{hyp}|\text{data}) = \frac{P(\text{data}|\text{hyp})P(\text{hyp})}{P(\text{data})}$$

The normalization factor $P(\text{data})$ can be determined for the case of parameter estimation, where all the possible values of the parameter are known, but in hypothesis testing it doesn't work, since we cannot enumerate all possible hypotheses. However it can be used to find the **ratio of probabilities** for two hypotheses, since the normalizations cancel:

$$R = \frac{P(H_0|\text{data})}{P(H_1|\text{data})} = \frac{\mathcal{L}(H_0)P(H_0)}{\mathcal{L}(H_1)P(H_1)}$$

10 Goodness-of-Fit Testing (GOF)

Here we are testing only one hypothesis H_0. The alternative is everything else, unspecified.

The Frequentist method for GOF is the same as for hypothesis testing, except that now only H_0 and α are known. We cannot know the **power** of the test since there is no alternative hypothesis (we don't know what we are trying to exclude). We can only say that if the data fall in the critical region, they fail the test (incompatible with the hypothesis H_0).

The most important GOF test is the Pearson Chisquared Test [9]. Indeed it is without a doubt the most often used statistical method in history. One can estimate that in the reconstruction of HEP data alone, it is probably invoked thousands of times per second in computers around the world.

For the Pearson test, the *test statistic* is the sum of the squares of deviations between data points and the hypothesis, with each deviation divided by the standard deviation of the data. Pearson showed that under the null hypothesis, this statistic is distributed asymptotically as a known function (now usually called the Chisquared Function) with N degrees of freedom if there are N data points, independently of the hypothesis being fitted. Tests for which the expected values of the test statistic do not depend on the hypothesis are called *distribution-free*.

There are many other tests which have been found to work well for particular problems. For physicists, the most important is probably the Kolmogorov-Smirnov test for compatibility of one-dimensional distributions (unbinned).

There is no way to do Bayesian hypothesis testing without an alternative hypothesis. Goodness-of-fit testing is therefore the domain of Frequentist statistics.

11 Decision Theory

For decision-making we need to introduce a new concept, the **loss** incurred in making the wrong decision, or more generally the losses incurred in taking different decisions as a function of which hypothesis is true. Sometimes the negative loss *(utility)* is used.

Simplest possible example: Decide whether to bring an umbrella to work.

The loss function may be:	Loss (umbrella if rain)	= 1
	Loss (umbrella if no rain)	= 1
	Loss (no umbrella if no rain)	= 0
	Loss (no umbrella if rain)	= 5

In order to make a decision, we need, in addition to the loss function, a **decision rule**. The most obvious and most common rule is to *minimize the expected loss.* Let $P(\text{rain})$ be the (Bayesian) probability that it will rain. Then we can write:

$$\begin{aligned} \text{Expected loss}|\text{umbrella} &= 1 \times P(\text{rain}) + 1 \times P(\text{no rain}) = 1 \\ \text{Expected loss}|\text{no umbrella} &= 5 \times P(\text{rain}) + 0 \times P(\text{no rain}) = 5 \times P(\text{rain}) \end{aligned}$$

The expected loss depends on the probability of rain, and with this loss function it is minimized if you take an umbrella to work whenever the probability of rain is more than 1/5.

An example of a different *decision rule* is the *minimax rule* which consists in minimizing the maximum loss. This rule does not require knowing the (Bayesian) probability of rain and is therefore a non-Bayesian decision rule. The minimax decision in the present case would be to carry the umbrella always, since the maximum loss is then only one point.

It can be shown that for any non-Bayesian decision rule, there is always a Bayesian rule which is as good or better (in the sense that it leads to no more loss than the non-Bayesian rule).

Since the *loss function* is in general subjective, and in view of the result that no decision rule can be better than a Bayesian decision rule, it is natural and reasonable to treat the whole decision process within the domain of Bayesian statistics.

References

[1] Wolfgang Panofsky and Melba Phillips, *Classical Electricity and Magnetism*, Addison-Wesley 1955, Section 1-2.

[2] Bruno de Finetti, *Annales de l'Institut Henri Poincaré* 7 (1937) 1-68. English Translation reprinted in *Breakthroughs in Statistics, Vol. 1*, Kotz and Johnson eds., Springer 1992.

[3] Anthony O'Hagan, *Kendall's Advanced Theory Of Statistics,* Vol. 2B (1994), Chapter 4.

[4] J. Neyman, *Phil. Trans. R. Soc. Ser. A* 236 (1937) 333, reprinted in *A Selection of Early Statistical Papers on J. Neyman*, Univ. of Cal. Press, Berkeley, 1967.

[5] *Kendall's Advanced Theory of Statistics*: In the Fifth Edition (1991) the authors are Stuart and Ord, and this material is at the beginning of chapter 23 in Volume 2. In the Sixth Edition (1999) the authors are Stuart, Ord and Arnold, and this material appears at the beginning of chapter 22 in Volume 2A.

[6] G. J. Feldman and R. D. Cousins, *Phys Rev* D57 (1998) 3873

[7] S. Ciampolillo, Il Nuovo Cimento 111 (1998) 1415

[8] J. Neyman and E. S. Pearson, *Phil. Trans. R. Soc. Ser. A* 231 (1933) 289-337, reprinted in *Breakthroughs in Statistics, Vol. 1*, Kotz and Johnson eds., Springer 1992.

[9] Karl Pearson, *Phil. Mag. Ser. 5* (1900) 157-175, reprinted in *Breakthroughs in Statistics, Vol. 2*, Kotz and Johnson eds., Springer 1992.