Chapter 2
Statistical Concepts

Paul C. Taylor
University of Hertfordshire

27th June 2005

2.1 Introduction

This will be a very brief skim over what Ad wrote in Chapter 2 because:

- there isn’t enough time to include everything;
- you can get the details from the book.

2.2 Probability

2.2.1 Random Experiments

A trial where the result is randomly determined, e.g.,

- rolling a die,
- tossing a coin,
- planting a field of turnips and seeing what the yield is.

We refer to the possible results as outcomes or *events*.

2.2.3 Frequency Definition of Probability

The probability of event A is written $P(A)$ and is

the long-run proportion of trials for which A occurred.

Mathematically, let m be the number of trials and let $m(A)$ be the number of trials for which A occurred, then

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

2.2.4 Subjective Belief

For example

I think there is a 90% probability of extra-terrestrial life.

does not fit into the frequency definition (because the result isn't random, . . .).

In this case the probability represents subjective *amount of belief*.

2.2.7 Random Variables

We can map the outcomes from our experiment to the real numbers.

E.g., the number of heads gained from tossing a coin twice is a (discrete) random variable.

2.2.8 Probability Distribution

This tells us the probability of a discrete random variable taking any particular value.

For continuous random variables it tells us the chance of the random variable falling in any particular interval of values.

2.2.10 Expectation

We often summarise a probabilty distribution by its expectation and variance.

Expectation is the long-run average (mean) value of the random variable, is the same way that the probabilities are the long-run proportions.

Usually denoted as μ.

Variance is the expectation of $(X - \mu)^2$, and it measures the how much the values taken by the random variable vary around μ.

Usually denoted as σ^2.

2.2.15 Some Named Discrete Distributions

- Binomial
- Multinomial
- Poisson

2.2.16 Some Named Continuous Distributions

- Uniform
- Normal (Gaussian)
- Exponential (also called negative exponential)
- Beta

2.3 Sampling and Sampling Distributions

Statisticians often view data collection as the carrying out a sequence of random experiments, each of which generates one or more data values.

We cannot keep going forever, so our data constitute a (random) sample of possible results.

Sampling distributions are probability distributions where the random variable of interest is some summary calculated form a sample.

Why do we want to know this? See, for example, bias/variance tradeoff in Section 2.5

2.4 Inference

2.4.2 Likelihood

The data are made up of results of random experiments and the probabilty distributions for the data are based on some vector of unknown parameters (constants), θ.

We want to estimate the unknown parameters.

The likelihood is

$$L(t) = P(\text{getting the data we collected if } \theta = t) \ .$$

There is a standard notational abuse: we usually write $L(\theta)$ instead of $L(t)$.

Our estimate of θ is denoted as $\hat{\theta}$.

We choose so that $L(\hat{\theta})$ maximises the likelihood function.

We know a lot about the asymptotic sampling properties of $\hat{\theta}$.

2.4.3 Bayesian Inference

Chapter 4 covers this topic, but here's the basic idea.

We think of $\boldsymbol{\theta}$ as the value taken by a random variable and the probability of this happening is $p(\boldsymbol{\theta})$.

So, $p(\boldsymbol{\theta})$ is our subjective probability of the parameters taking the value $\boldsymbol{\theta}$. This is called the *prior probability*.

We want to update $p(\boldsymbol{\theta})$ in the light of the data that we have collected, to give the *posterior probability*, which we will write as $p(\boldsymbol{\theta}|\text{data})$.

It turns out that

$$p(\boldsymbol{\theta}|\text{data}) \propto L(\boldsymbol{\theta})p(\boldsymbol{\theta}) \ ,$$

or

$$p(\boldsymbol{\theta}|\text{data}) = \frac{L(\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \ .$$

Note the appeal of this idea in the context of automated data collection.

2.5 Prediction and Prediction Error

2.5.1 Prediction Error in Regression

If we are trying to estimate a function, say $f(\boldsymbol{x})$, and we estimate it by, say $\widehat{f}(\boldsymbol{x})$ then it is natural to look at the expectation of

$$\left(f(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x})\right)^2$$

to judge how good the prediction is; we call this a *mean square error*.

It turns out that we can break this mean square error down as

$$\text{bias}^2 + \text{variance} \ .$$

(We can do similar things for classification, i.e., unsupervised learning, problems.)

This is important because:

- we can reduce bias by fitting more complicated/flexible models;
- we can reduce variance by using bigger data sets;
- using more complicated/flexible models tends to increase the variance.

In data-mining there is a focus on complicated/flexible models (to reduce bias) applied to massive data sets, which reduces the variance (compensating for the variance drawback of flexible models).

See the book for more details.

2.6 Resampling

The sort of models that are being used these days are mathematically complicated and do not lend themselves to determining properties of sampling distributions by mathematical approaches.

An altenative attack is *computer simulation*.

Resampling uses the data collected (sampled) as a source of outcomes from our random experiment. The data are (pseudo-)randomly resampled repeatedly to detemine sampling properties, such as the bias.

Techniques discussed in this section of the book include: *cross-validation*; *bootstrapping*; *bagging*; *arcing*.