STAT 830

Notes on an introduction to inference

Definition: A **model** is a family $\{P_\theta; \theta \in \Theta\}$ of possible distributions for some random variable X. (Our data set is X, so X will generally be a big vector or matrix or even more complicated object.)

We will assume throughout this course that the true distribution P of X is in fact some P_{θ_0} for some $\theta_0 \in \Theta$. We call θ_0 the true value of the parameter. Notice that this assumption will be wrong; we hope it is not wrong in an important way. If we are very worried that it is wrong we enlarge our model putting in more distributions and making Θ bigger. Lots of work has been done to think about what happens when the true *data generating mechanism* (the true distribution P of X) is not in the model; this problem is called *model mis-specification*.

Our goal is to observe the value of X and then guess θ_0 or some property of θ_0. We will consider the following classic mathematical versions of this:

1. Point estimation: we must compute an estimate $\hat\theta = \hat\theta(X)$ which lies in Θ (or something close to Θ).

2. Point estimation of a function of θ: we must compute an estimate $\hat\phi = \hat\phi(X)$ of $\phi = g(\theta)$.

3. Interval (or set) estimation. We must compute a set $C = C(X)$ in Θ which we think will contain θ_0.

4. Hypothesis testing: We must choose between $\theta_0 \in \Theta_0$ and $\theta_0 \not\in \Theta_0$ where $\Theta_0 \subset \Theta$.

5. Prediction: we must guess the value of an observable random variable Y whose distribution depends on θ_0. Typically Y is the value of the variable X in a repetition of the experiment.

There are several schools of statistical thinking. Some of the main schools of thought can be summarized roughly as follows:

- **Neyman Pearson**: A statistical procedure is evaluated by its long run frequency performance. Imagine repeating the data collection exercise many times, independently. The quality of a procedure is measured by its average performance when the true distribution of X values is P_{θ_0}.

For instance, estimates are studied by computing their sampling properties such as mean, variance, bias, and mean squared error.

Definition: If $\hat{\phi}$ is an estimator of some parameter ϕ then the bias, variance and mean squared error are the following functions of the unknown distribution F of the data.

Notice that I switch from P to F. Sometimes we talk about all the probabilities using the symbol P and sometimes we talk about cumulative distribution functions and use F. Sometimes we use a subscript θ for a point in the parameter space.

Bias:

$$\text{bias}_{\hat{\phi}}(F) = \text{E}_F(\hat{\phi}) - \phi(F).$$

Variance:

$$\text{bias}_{\hat{\phi}}(F) = \text{Var}_F(\hat{\phi}).$$

Mean Squared Error:

$$\text{MSE}_{\hat{\phi}}(F) = \text{E}\left[\left\{\hat{\phi} - \phi(F)\right\}^2\right].$$

Several features of these definitions deserve discussion. First, each distribution F in the model $\mathcal{F}$ must have some value for the parameter ϕ. We denote this value $\phi(F)$ in the definitions above. In parametric models the distribution F is indexed by the parameter θ and we write $\phi(\theta)$ instead of $\phi(F)$. Second, the subscripts F on E and Var remind us that while the model has many possible distributions when we come to compute probabilities and moments we have to use some particular distribution. Third, notice that the subscript F, indicating which distribution goes into computing the means and variances is the same as the one going into ϕ. Fourth, you need to know the following decomposition of MSE:

$$\text{MSE} = \text{bias}^2 + \text{Variance}.$$

Finally, the idea is that good estimators have small biases, small variances and small mean squared errors. They are being judged on the basis of their long-run or average or expected performance NOT on the basis of how well they will work with today's data. This is the Neyman-Pearson approach to inference – ask the question "how well does my statistical procedure work on average?"

Confidence sets or intervals are also to be judged on the basis of their average performance. A confidence set is a random subset $C(X)$ of Θ or Φ (where Φ is the set of possible values of some parameter ϕ). The set has *level* β if

$$P_F(\phi(F) \in C(X)) \geq \beta$$

for all $F \in \mathcal{F}$. It is absolutely crucial to note that the only thing random in this formula is the set $C(X)$, NOT, $\phi(F)$. That means that the probability describes the average behaviour of the procedure used to compute the set $C(X)$ NOT the behaviour on today's data set.

Several details should be mentioned. First if we replace $\geq \beta$ by $\equiv \beta$ then the set is *exact*. (WARNING: I recently (June 2020) read a paper which used *exact* to mean the inequality was correct for every parameter value.) Second the random set $C(X)$ is usually just a random interval $[L(X), U(X)]$ – all the values of ϕ between these two random limits. Third in practice the desired property is more stringent that we can achieve. Generally we can only replace $\geq \beta$ with the assertion that the probability is approximately β or approximately some number $\geq \beta$.

Example: You all know that for samples of size n from the $N(\mu, \sigma^2)$ distribution the interval

$$\bar{X} \pm t_{n-1,\alpha/2}s/\sqrt{n} \text{ or } L = \bar{X} - t_{n-1,\alpha/2}s/\sqrt{n} \text{ to } U = \bar{X} + t_{n-1,\alpha/2}s/\sqrt{n}$$

is an exact level $1-\alpha$ confidence interval for μ. (As usual $t_{\nu,\alpha}$ is the upper α critical point for a Student's t distribution on ν degrees of freedom.

There are more features to discuss in a confidence interval beyond its coverage probability $P_F(\phi \in C(X))$. For instance the probability it does not include a given wrong value of ϕ should be high. The set should be as small as possible since that corresponds to a precise estimate of ϕ.

Hypothesis tests are judged on the basis of error rates. For problems when a hypothesis is true we ask how often we conclude the hypothesis is true. The probability we incorrectly conclude the hypothesis is wrong is an error rate. Note particularly that we just ask what fraction of data sets the procedure works for, NOT, whether or not it appears likely to work with today's data.

- **Bayes**: Treat θ as random just like X. Compute conditional law of unknown quantities given known quantities. In particular ask how a procedure will work on the data we actually got – no averaging over data we might have got.

  For point estimation the Bayesian would study the distribution of the estimation error $\hat{\phi}(X) - \phi(F)$ *given* the data X. Now only F is random – X is known and treated as a fixed deterministic object. The Bayesian then chooses $\hat{\phi}(X)$ to make the estimation error as small as possible – as measured by some feature of its distribution give X; this distribution is called a *posterior* distribution since it applies *after* the data are observed.

  For confidence sets the Bayesian, too, would work out a set $C(X)$ of values of ϕ which s/he considers likely to contain the true value but now the Bayesian wants

  $$P(\phi \in C(X)|X)$$

  to be large while making $C(X)$ as small as possible. Typically the Bayesian insists that

  $$P(\phi \in C(X)|X) = \beta$$

  for some given β. The Bayesian asks only about today's data X as s/he observed it and not about other data which might have been observed but was not.

  For hypothesis testing the Bayesian naturally computes the probability, given X that each hypothesis is correct.

- **Likelihood**: Try to combine previous 2 by looking only at actual data while trying to avoid treating θ as random.

  I will try, later in the course, to describe this school of inference.

We use the Neyman Pearson approach to evaluate the quality of likelihood and other methods in this course – and even to study the behaviour of Bayesian methods.