Iacta alea est. (The die is cast.) –Julius Caesar

Problems 21-30 on Basic Statistics Objects (STAT2802 Statistical Models Tutorial notes for the week of 08-OCT-2012)

A datum is formalized as a realization/observation/measurement of a random variable. A datum is a point on the random variable's sample space. This formalization was non-trivial before it first appeared, but it is now felt as a trivial initiation of any statistical modeling. A character of the reality behind this formalization is that datum repeats itself many times into a sample of data. In some real cases, the statistician has some control over this repetition so that they can design the repetition to yield a sample of data holding an optimal structure of information. In this case, the statisticians are effectively performing and studying *statistical experiments*.

In some other real cases, the repetition is done by Nature before the sample of data is observed by the statistician. In this case, the information contained within the sample of data may not be optimally structured to answer the statistician's question. Some part of the information may even severely mask other parts; moreover, the statistician has limited flexibility to perform controlled experiments to unmask them. In this case, the statistician is studying *observational data*.

Statistical inference is directed towards properties of the states of our world. These properties shape the behavior of a random variable on its sample space in a fundamental way. When the sample space is still too large and complex for complete comprehension, such as one often in an observational study, the sample space is called a *population*, carrying the taste of practice. Thus, although the DNAs among human beings are almost identical, the height of an individual human being varies according to a distribution. The individual human being (as a developmental result of his DNA) can be regarded as a random variable observable by his/her height; the set of all possible values of this random variable (its sample space) is the *population* of heights (word play of "the heights of a population of human individuals"). Keep in mind that all these terminologies can be streamlined mathematically as a random variable with its domain and its range and also as compositions of random variables. Also keep in mind that at the target of a statistical model is the population distribution (i.e., the distribution on the random variable's sample space); the statistician aims to learn this population distribution.

The two major procedures of statistical inference are estimation and hypothesis testing. Both use a statistic as the vehicle for formal manipulation. The statistic is called an *estimator* in the estimation settings and a *testing statistic* in the hypothesis testing settings. A statistic is based on a sample of data, which is more extensive than a single datum. The value of a statistic is definitively evaluable from those of the data: data observed → statistic evaluated. Mathematically, the statistic is a function whose domain is $S^n = S \times S \times \cdots \times S$, where S is the sample space (population), $\times$ is the Cartesian product between sets, and n is the number of datum in the particular sample of data. Thus the statistic is a transformation of an n-dimensional random vector. The transformation not only transforms the n-tuple realization, but also transforms the joint distribution of this n-tuple, to the range of the statistic. The distribution of the statistic (on its range) is specifically called the *sampling distribution* of that statistic.

The statistician starts with making scientifically informed hypotheses about the population distribution. Thus, for measurements of lengths, there is good reason to assume the shape of the distribution is a normal one. The statistician leaves the *parameters* of the population distribution to be determined by the estimation procedure. We say that the population distribution belongs to a family *indexed* by the parameters. This is the beginning of *parameter estimation*. Next, the statistician constructs a statistic called a *parameter estimator* so that the range of the statistic is the same as the set of all possible values of the parameter. This is a sensible requirement. Next, the estimator is constructed with the aim to hold as lean (*sufficient* and *efficient*), correct (*unbiased* and *consistent*), and *complete* as possible the information of the parameter so that this information can be extracted by evaluating the statistic on the data.

Notations

Random variable: $X: \Omega \to S(\in \mathbb{R})$

Random Sample of size n: $(X_1, X_2, \dots, X_n): \Omega^n \to S^n(\text{usu.} \subset \mathbb{R}^n)$

Family of distributions of X indexed(parameterized) by θ: $\mathbb{P}_\theta^X: \mathcal{B}(S) \to [0,1]$ ($\mathcal{B}(S)$ denotes the set of all events in S)

Parameter and Parameter space: $\theta \in \Theta$

Statistic: $T: S^n \to \mathbb{R}$ (or $\mathbb{R}^m$)

Estimator: $\hat{\theta}: S^n \to \Theta$ ($\text{usu.} \subset \mathbb{R}$ or $\mathbb{R}^m$).

Problems 21-30

21. Let $\hat{\theta}: S^n \to \Theta$ be an estimator for θ. State the mathematical definitions of its bias $Bias(\hat{\theta})$, variance $\mathbb{V}(\hat{\theta})$, and mean square error $MSE(\hat{\theta})$. Then prove the identity: $MSE(\hat{\theta}) = \mathbb{V}(\hat{\theta}) + Bias(\hat{\theta})^2$. Then construct, respectively, an unbiased estimator for the population mean $\mathbb{E}(X)$ and population variance $\mathbb{V}(X)$.

22. Let $\hat{\theta}: S^n \to \Theta$ be an estimator for θ. State the mathematical definition of $\hat{\theta}$ being a consistent estimator. Construct a consistent estimator for the 3rd population moment $\mathbb{E}(X^3)$, is it also unbiased?

23. A size-n random sample $\{X_i\}_{i=1}^n$ is obtained from a Bernoulli trial (e.g. by tossing an unfair coin n times or by observing a stable natural yes-no phenomenon n times). The statistician wants to infer the quantity $\theta(1-\theta)$ where $\theta = \mathbb{P}(X = 1)$. First, explicitly state S, Θ, and $\mathbb{P}_\theta^X$, then construct an unbiased estimator $U: S^n \to \Theta$ and derive $\mathbb{V}(U)$.

24. Same as 23, except that now the estimation target is $\theta(1-\theta)^2$.

25. Same Bernoulli trail as in 23, except that now the estimation target is θ and that the statistician is considering using the number of consecutive heads(1s) between two tails (0s), which is a random variable $Z \sim Geometric(\theta)$ with p.m.f $p_Z(k) = (1-\theta)^k\theta, k = 0,1,2, \dots$, as the basis for constructing an estimator for θ. Is an unbiased estimator possible?

26. State and prove the fundamental result "Poisson approximation to Binomial" which gives the relationship between the two families of distributions.

27. Let $X_1, \dots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Show that its variance $\mathbb{V}(S^2) = \frac{2\sigma^4}{n-1}$. *Hint: Chi-sq distribution with degree of freedom n-1*.

28. The continuous random variables X, Y, and Z defined in the range $(0 \le X, Y, Z < \infty)$ have the joint probability distribution with the density function $f(X = x, Y = y, Z = z) = (xyz)^{-1/2} \cdot g(x + y + z)$. Derive the marginal distributions of the random variables (i) $U = X + Y + Z$; (ii) $V = Y/X$; and (iii) $W = Z/(X + Y)$.

29. Let X_1, X_2 be a size-2 random sample from the population density $f_X(x; \theta) = \frac{1}{\theta}e^{-x/\theta}$, for $0 \le x \le \infty$ and θ being a parameter. Derive the sampling distributions of the statistics $U = X_1 + X_2$ and $V = \frac{X_1}{X_1 + X_2}$, and hence prove that $U \perp\!\!\!\perp V$. Find the mean and variance of V.

30. Let X_1, X_2, X_3 be a size-3 random sample from $N(m, \sigma^2)$. Derive the joint sampling distribution of (i) $U = X_1 - X_3$; (ii) $V = X_2 - X_3$; and (iii) $W = X_1 + X_2 + X_3 - 3m$.