1. Introduction to Statistics; Jan 4, 2010 LM Ch1, and 5.1

1.1: Some info about the class

- The class web site is www.stat.washington.edu/thompson/S341_10/

  For information on exams, homeworks, grading, office-hours, ... see the web page.
- The most useful web page is the schedule www.stat.washington.edu/thompson/S341_10/schedule.shtml.

  It contains links to homeworks, class notes, etc. etc.
- The class is a continuation of Stat340; there is a link to Stat 340 information on the web page.
- The book is Larsen & Marx, 4th ed. (LM).

  It is assumed you have covered the equivalent of LM 2.1-2.7, 3.1-3.6, 4.1-4.5.
- Homeworks due at start of class on Weds; starting Jan 13; see the web page.

  The first homeworks will review probability from 340/394/395.

1.2: What is Statistics?

- Statistics is the science of analysis of data

Science: has theory, methods, principles but no axiomatic basis.

Statistics uses Probability (and its axiomatic basis) as a framework.

Data: No physical laws; data can be anything!

Data derive from experiments, observational studies, sample surveys, ... Data relate to any quantitative or qualitative fact for which we want to understand the basis – in science, society, business, ...

Analysis: Find a unifying structure for data.

Assume a model (a probability distribution) for the data; Find out about the model from the data.

- More limited definition: Devise procedures for extracting information about the model from the data.

1.3 Probability models, parameters and statistical inference

Example: toss a coin for which $P(head) = \theta$ n times, observe k heads.

Model: Number of heads is $Bin(n, \theta)$. Data comes from the *Binomial family*, with index n (fixed), and parameter θ: $\{B(n, \theta); 0 \leq \theta \leq 1\}$.

Statistics vs Probability: In Probability we make statements about the data (e.g. k heads), given some value of θ. In Statistics we make inferences about θ given the data.

There are two basic modes of inference: *In Stat341; we focus on estimation*

Estimation: choose a sensible value for θ, – a function of the data. For example: k/n

Hypothesis testing: We make a statement about the parameter; for example: *The coin in fair:* $\theta = 1/2$. We ask: Do the data enable us to reject this hypothesis? (Stat342 will deal mainly with hypothesis testing.)

1.4 Where does the model come from?

(i) “Almost true” models: Number of heads *is* $Bin(n, \theta)$.

Number of blips on Geiger-counter in time t *is* $\mathcal{P}\mathrm{o}(\lambda t)$.

(ii) Idealized models: e.g. Number of road accidents modeled by Poisson.

Genetic models: assume random mating, no selection,....

(iii) Descriptive models: crop yields, heights, etc. assumed Normal

”density looks like” $1/\pi(1 + (x - \theta)^2))$; Cauchy density.

There is a continuum in the “reality” of models: the three classes merge.

A model is a map; it is NOT Google Earth. Models (like maps) can be more or less detailed; the important thing is that they are useful in understanding the data, and guiding studies/experiments.

2. Estimators and estimates of parameters: Jan 8, 2010, Part 1

2.1: Estimates, Estimators, and statistics

Recall: a *random variable* is any real-valued function on the underlying probability space.

We call the random variables used to model the data, the *data random variables.*

Definition: A *statistic* is any (real-valued) function of the data random variables.

Definition: A statistic used to estimate a parameter is an *estimator.*

Definition: The value taken by the estimator in a particular instance is the *estimate.*

IMPORTANT:

Estimators, statistics, and data random variables are *random variables.*

Estimates, values of a statistics, and the data outcomes are real numbers.

2.2 Estimating parameters in models

Definition: A *model* is a family of (discrete or continuous) probability distributions indexed by some unknown parameter θ.

Example: The family $\{Bin(n, \theta);\ 0 \leq \theta \leq 1\}$.

Note: θ may be a vector; there may be several parameters; e.g. $\{N(\mu, \sigma^2);\ -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$.

Data are assumed to be a realization of a random variable whose distribution is some member of the family.

Example: Number of heads k observed in n tosses of a coin is outcome of random variable $X \sim Bin(n, \theta)$:

$$P(X = k) \ = \ \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Aim is to estimate θ by some function of the data; k.

The *estimator* is that same function of the random variable X.

Important: Always distinguish the *estimate* (a number) from the *estimator* (a random variable).

2.3 Example 1:

We wish to estimate the frequency θ of people of AB blood type in Seattle.

Data: We take a sample of n individuals and type their blood; we find that k of type AB.

Model: The number of type AB in sample size n has $Bin(n, \theta)$ distribution.

Assumptions:?? Random sample; not by family/ by region .... With/without replacement ?

A sensible *estimate* is k/n.

What do we know about the *estimator* X/n when $X \sim Bin(n, \theta)$?

$\mathrm{E}(X) = n\theta$, $\mathrm{var}(X) = n\theta(1-\theta)$. So $\mathrm{E}(X/n) = \theta$, $\mathrm{var}(X/n) = \theta(1-\theta)/n$. (why?)

We can estimate θ by k/n, and can estimate the variance of the estimator X/n by $k(n-k)/n^3$.

2.4 Example 2:

We wish to estimate the rate of accidents on some stretch of highway.

Data: In n months we observe a total of k accidents.

Model: Accidents occur as a Poisson Process rate λ per month.

That is, the number of accidents has a $\mathcal{P}\mathrm{o}(\lambda n)$ distribution (why?).

Assumptions: independent? constant rate? (day/night; weather??).

A sensible *estimate* is k/n.

What do we know about the *estimator* X/n when $X \sim \mathcal{P}\mathrm{o}(n\lambda)$?

$\mathrm{E}(X) = n\lambda$, $\mathrm{var}(X) = n\lambda$. So $\mathrm{E}(X/n) = \lambda$, $\mathrm{var}(X/n) = \lambda/n$. (why?)

We can estimate λ by k/n, and can estimate the variance of the estimator X/n by $(k/n)/n = k/n^2$.

3. Estimators based on n-samples: Jan 8,2010, part 2.

3.1 Data random variables as an n-sample

Definition of *n-sample*: Often our data come as a set of n outcomes from repeated experiments/sampling. Such data are said to be an *n-sample.*

Definition of *i.i.d*: The underlying data random variables $X_1, ..., X_n$ are *independent and identically distributed* (i.i.d).

Definition of *parameter space*: Each of the data outcomes $x_1, ..., x_n$ is an independent realization from the family of distributions $\{f_X(x;\theta); \theta \in \Theta\}$, indexed by parameter θ in *parameter space* Θ.

Note: we will use $f_X(x;\theta)$ to denote the pmf or pdf (discrete of continuous) of random variable X, indexed by the parameter θ. X is a single data random variable; each of the i.i.d. X_i, $i = 1, ..., n$, has this same pmf/pdf $f_X(x;\theta)$.

3.2 An n-sample of Bernoulli or Binomial outcomes

In the example of 2.3 above; each of the outcomes is a Bernoulli outcome: $X_i = 1$ if person i has blood type AB, $X_i = 0$ if not. $P(X_i = 1) = \theta$. $P(X_i = 0) = (1-\theta)$. $\Theta = \{\theta; 0 \leq \theta \leq 1\} \ = \ [0, 1]$.
The statistic $T(X_1, ..., X_n) = \sum_1^n X_i$ is Binomial $Bin(n, \theta)$, and we would use the estimator T/n for θ just as before.

We could also have an n-sample of Binomial outcomes. For example, each of n technicians samples and types r people (in a very large population), and reports the number of bloodtype AB in his sample.
Then the i.i.d data random variables are $X_i \ \sim Bin(r, \theta)$.
The statistic $T(X_1, ..., X_n) = \sum_1^n X_i$ is Binomial $Bin(rn, \theta)$ (why?).
We would use the estimator T/rn for θ; why?. $\mathrm{E}(T/rn) = \theta$. $\mathrm{var}(T/rn) = \theta/rn$.

3.3 An n-sample of Poisson outcomes

In example 2.4 above, instead of just reporting the total accidents in n months, we could report the number x_i in each month i, $i = 1, ..., n$. The data are then an n-sample from the $\mathcal{P}\mathrm{o}(\lambda)$ distribution; that is X_i are i.i.d and $X_i \ \sim \ \mathcal{P}\mathrm{o}(\lambda)$, $i = 1, ..., n$. The parameter space is $\{\lambda; 0 \leq \lambda < \infty\} \ = \ [0, \infty)$.
The statistic $T(X_1, ..., X_n) \ = \ \sum_1^n X_i$ is Poisson $\mathcal{P}\mathrm{o}(n\lambda)$, and we could use the estimator T/n for θ just as before.

How do we know this would be the best thing to do?? –as yet, we don't.

3.4 An n-sample from a Uniform distribution

Consider a n-sample from a Uniform distribution, $U(0, \theta)$.
We wish to estimate θ. The i.i.d data random variables $X_i \ \sim \ U(0, \theta)$, $i = 1, ..., n$.
Now $\mathrm{E}(X_i) \ = \ \theta/2$ so $\mathrm{E}(\sum_{i=1}^n X_i) \ = \ n\theta/2$, so we could use the estimator $T \ = \ 2\sum X_i/n$.
However, this might not be sensible; suppose there is some X_i larger than the value of T ??
Maybe we should use an estimator based on $W = \max(X_i)$ (why?).
The CDF $F_X(x) = P(X \leq x) = x/\theta$ for $0 \leq x \leq \theta$.
So the CDF of W is $F_W(w) = P(W \leq w) = P(\text{all } X_i \text{ are } \leq w) \ = \ (w/\theta)^n$ (using independence).
So the pdf is $f_W(w) = nw^{n-1}/\theta^n$ on $0 < w < \theta$ and 0 otherwise.
So now $\mathrm{E}(W) \ = \ \int_0^\theta nw^n/\theta^n \ dw \ = \ (n/\theta^n)[w^n/(n+1)]_0^\theta \ = \ n\theta/(n+1)$.
Thus perhaps the estimator $(n+1)W/n \ = (n+1)\max(X_i)/n$ is better than $T \ = \ 2\sum_{i=1}^n X_i$ – we shall see that it is!!