STA261H1S – Probability and Statistics II – Summer 2013 (web site: UT Portal)

Instructor: Xiaoping Shi SS6026
Office hrs: MW5-6pm

Course Objective: To introduce topics in statistical theory, particularly those in parameter estimation and hypothesis testing, and to prepare you for further study in Statistical Science. Coverage includes chapters 8 and 9 plus a selection of topics from chapters 10-14 of the textbook. Prerequisite: STA 257.

Tutorials

Tutorials begin July 8. Tutorials meet every MW6-7pm at SS2135 same with Lecture room. Assignments will be also posted on the course web site. They will consist of suggested exercises, mostly from the textbook. Bring your solutions to tutorial, along with your questions about these exercises or the related theory and concepts. Expect a quiz on the material as well.

Textbook

Mathematical Statistics and Data Analysis-3rd edition – by John A. Rice

Coverage

Lecture 1: General introduction of probability and statistics and Method of Moments
Lecture 2: Maximum Likelihood Estimation
Lecture 3: Bayesian Inference
Lecture 4: Efficiency and Sufficiency
Lecture 5: Neyman-Pearson Paradigm
Lecture 6: Likelihood Ratio Tests
Lecture 7: Comparing two samples
Lecture 8: Simple linear regression
Lecture 9: Analysis of variance and Matrix algebra
Lecture 10: Multiple linear regression
Lecture 11: Review

Statistics Aid

Your primary source of help with difficulties is your TA in the scheduled tutorial, but additional assistance can be obtained by sending an email to me. Then I would answer common questions in class/ send your email back.

Evaluation

Tutorial quizzes 20% (July 10 & July 31), Midterm Test 30% (July 24) and Final exam 50% (TBD).
Programmable calculators are not permitted on quizzes, test or exam. You must bring your student identification to term tests as well as the final exam.

Missed Midterm Test

There are **no make-up tests**. Should you miss the term test due to illness, you must submit to your lecturer or to SS6018 (Stats office), within one week, completed by yourself and your doctor, **the 'U of T Student Medical Certificate'**, obtainable from your college registrar, the office of the Faculty Registrar (SS 1006), the StatSci Dept. office, or the Koffler health service. The test's weight will then be shifted to the final exam. If proper documentation is not received, your test mark will be zero.

Academic Offences

Academic offences are unacceptable, and harm everyone. Offenders are caught, and **sanctions can be severe** and very with examples including a zero in the course with annotation on the transcript for several years; suspension of a year; even expulsion.

Lecture 1

Xiaoping Shi
Department of Statistics, University of Toronto
xpshi@utstat.toronto.edu
http://www.utstat.utoronto.ca/~ xpshi/sta261-2013s.html

July 3, 2013

- A brief introduction to probability and statistics
- Review
- Estimation of parameters

Probability is a measure or estimation of how likely it is that something will happen or that a statement is true.

Probabilities are given a value between 0 (0% chance or will not happen) and 1 (100% chance or will happen). Tossing a fair coin twice will yield HH with probability 1/4, because the four outcomes HH, HT, TH and TT are possible.

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data.

The difference is that probability starts from the given parameters of a total population to deduce probabilities that pertain to samples. Statistical inference, however, moves in the opposite direction-inferring from samples to the parameters of a total population.

Probability & Statistics

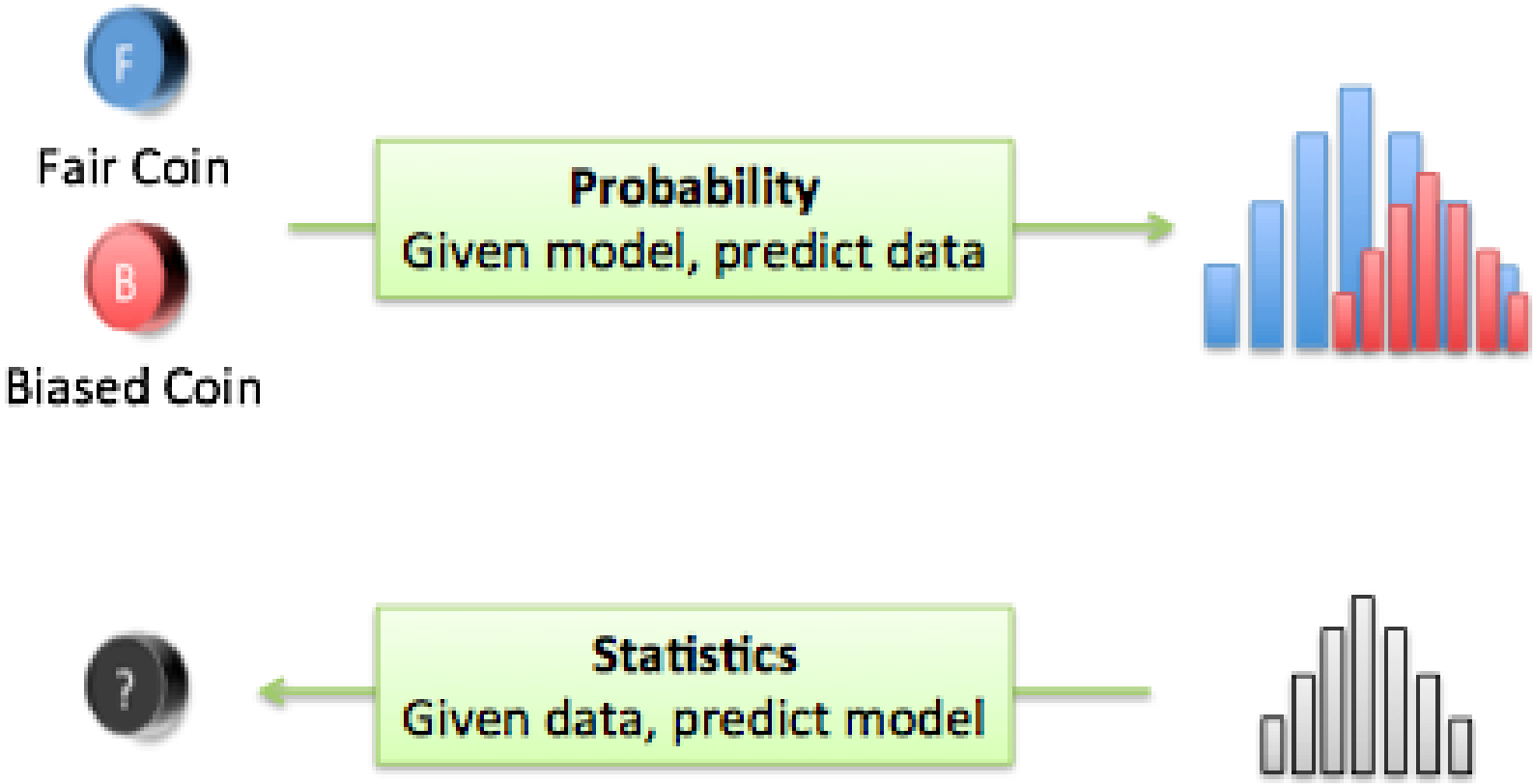

Figure: The difference between prob. and stats.

Probability is straightforward: you have the bear. Measure the foot size, the leg length, and you can deduce the footprints. Oh, Mr. Bubbles weighs 400lbs and has 3-foot legs, and will make tracks like this. More academically: We have a fair coin. After 10 flips, here are the possible outcomes.

Statistics is harder. We measure the footprints and have to guess what animal it could be. A bear? A human? If we get 6 heads and 4 tails, whatre the chances of a fair coin?

Heres how we find the animal with statistics:

Get the tracks: Each piece of data is a point in connect the dots. The more data, the clearer the shape (1 spot in connect-the-dots isnt helpful. One data point makes it hard to find a trend.)

Measure the basic characteristics: Every footprint has a depth, width, and height. Every data set has a mean, median, standard deviation, and so on. These universal, generic descriptions give a rough narrowing: The footprint is 6 inches wide: a small bear, or a large man?

Find the species: There are dozens of possible animals (probability distributions) to consider. We narrow it down with prior knowledge of the system. In the woods? Think horses, not zebras. Dealing with yes/no questions? Consider a binomial distribution.

Look up the specific animal: Once we have the distribution (bears), we look up our generic measurements in a table. A 6-inch wide, 2-inch deep pawprint is most likely a 3-year-old, 400-lbs bear. The lookup table is generated from the probability distribution, i.e. making measurements when the animal is in the zoo.

Make additional predictions: Once we know the animal, we can predict future behavior and other traits. Statistics helps us get information about the origin of the data, from the data itself.

The necessary steps for inferential statistics:

(1) Specify the questions to be answered and identify the population of interest.
(2) Decide how to select the sample from the population.
(3) Analyze the sample information and make an inference about the population.

Probability:Axioms & Rules

- $0 \leq P(A) \leq 1$ for any event A.
- $P(S) = 1$ and $P(\emptyset) = 0$.
- Let $A_1, A_2, \ldots, A_m$ be mutually exclusive events. Then

$$P(A_1 \cup A_2 \cup \cdots \cup A_m) = P(A_1) + P(A_2) + \cdots + P(A_m)$$

  - this must hold even for a countably infinite collection of mutually exclusive events
- Complement Rule: $P(\overline{A}) = 1 - P(A)$.
  - $\overline{A}$: complement of A ie. not A

- Inclusion-exclusion (addition theorem) rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- If A and B are mutually exclusive:

$$P(A \cup B) = P(A) + P(B)$$

- Useful complement result:

$$P(\overline{A \cup B}) = P(\overline{A} \cap \overline{B})$$

Conditional probability

- Is there a subset of the population that can be *recognized* that has a different probability?
- Let $P(B) > 0$
- The *conditional* probability of A given that B occurs is:

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

  - If $P(B) = 0$, $P(A|B)$ is not defined.
- Multiplication theorem:

$$P(A \cap B) = P(A|B)P(B)$$

Independence of Events

- Two events A and B with $P(B) > 0$ are *independent* events if

$$P(A|B) = P(A)$$

  - knowledge that B occurred does not alter the probability that A occurs
- Independence of A and B:

$$P(A \cap B) = P(A)P(B)$$

Discrete Random Variables

- *Random variable*: realized value determined by chance through a random mechanism.
- Y: a *discrete* random variable
- Y can assume values $y_1, y_2, \ldots$
  - finite or at most countably infinite
- *Probability (density or mass) function* f is:

$$f(y) = P(Y = y)$$

- Properties of f:
  - $f(y) \geq 0$
  - $f(y) > 0$ for possible distinct values
  - summing f over possible values yields 1

$$\sum_{\text{all } y} f(y) = 1$$

- *Expected* value or *mean* of Y:

$$\begin{aligned} \mu &= E[Y] \\ &= \sum_{\text{all } y} yP(Y=y) \\ &= \sum_{\text{all } y} yf(y) \end{aligned}$$

  - measure of center
- $H(Y)$: any function of Y
- Expected value of $H(Y)$:

$$\begin{aligned} E[H(Y)] &= \sum_{\text{all } y} H(y)P(Y=y) \\ &= \sum_{\text{all } y} H(y)f(y) \end{aligned}$$

- Variance of Y: $\sigma^2\{Y\}$ (or $Var(Y)$)

$$\begin{aligned}\sigma^2\{Y\} &= E[(Y - E[Y])^2] \\ &= E[Y^2] - E[Y]^2\end{aligned}$$

- *Standard deviation* of Y: $\sigma\{Y\}$
  - measure of spread
- a & c constants:

$$E[a + cY] = a + cE[Y]$$

$$Var(a + cY) = c^2 Var(Y)$$

Binomial Distribution

- Fixed number of trials n are carried out.
- Two possible outcomes:success or failure for each trial
- Probability of success: p
  - constant across trials
- All trials are independent.
- Y: number of successes

- The probability mass function of Y:

$$\begin{aligned} f(y) &= P(Y = y) \\ &= \binom{n}{y} p^y (1-p)^{n-y} \end{aligned}$$

  - for $y = 0, 1, \ldots, n$.
- $E[Y] = np$.
- $\sigma^2\{Y\} = np(1-p)$.

Continuous Random Variables

- Y: a continuous random variable
- Y can assume any value over entire intervals of real numbers.
- Model for weights, lengths, etc.
- *Probability density function* $f(y)$ gives the probability per unit length for values very close to y.

- Properties of f:
  - defined for all real numbers.
  - $f(y) \geq 0$.
  - region under graph of f & above the y axis has area 1.

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

- For real numbers $a \leq b$, $P(a \leq Y \leq b)$ is given by area bounded by:
  - the graph of f &
  - the lines $y = a$, $y = b$ &
  - the y axis.

$$P(a \leq Y \leq b) = \int_a^b f(y)dy$$

- $P(Y = a) = 0$ for this idealization.
  - zero chance of hitting any specific value
- $E[Y]$ given by

$$E[Y] = \int_{-\infty}^{\infty} yf(y)dy$$

Normal distribution

- Bell-shaped density that can model many continuous variables of interest.
- The density has the form:

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

  - for $-\infty < y < \infty$
  - where μ is the mean
  - and σ is the standard deviation

- Linear transformation: $Y' = a + cY$
  - a & c constants
- Y' again normal
  - mean: $a + c\mu$
  - variance: $c^2\sigma^2$
- Z: standardized version of Y.

$$Z = \frac{Y - \mu}{\sigma}$$

  - signed number of standard deviations Y from μ
  - Z has a $N(0, 1)$.
- $P(Z \leq z\{A\}) = A$ tabulated in Table B.1.

68-95-99.7 Rule

- In any normal distribution (correct to given number of decimal places):
  - 68% (68.3%) of the population values fall within σ of the mean μ.
  - 95% (95.4%) of the population values fall within 2σ of the mean μ.
  - 99.7% (99.74%) of the population values fall within 3σ of the mean μ.
- Normal distribution often used to describe unexplained random error in regression analysis.

Linear combinations of r.v.

- $Y_1, Y_2, \ldots, Y_n$: r.v.'s
- $a_1, \ldots, a_n$: constants

$$E[\sum_{i=1}^{n} a_i Y_i] = \sum_{i=1}^{n} a_i E[Y_i]$$

- If $Y_1, \ldots, Y_n$ are independent

$$Var(\sum_{i=1}^{n} a_i Y_i) = \sum_{i=1}^{n} a_i^2 Var(Y_i)$$

- If $Y_1, \ldots, Y_n$ independent and normally distributed:
  - then $\sum_{i=1}^{n} a_i Y_i$ normally distributed

Center & spread for distribution of averages

- $Y_1, \ldots, Y_n$: n independent r.v.'s with mean μ and standard deviation σ.
- $\overline{Y} = (Y_1 + Y_2 + \cdots + Y_n)/n$.
- $\mu_{\overline{Y}}$: mean of the sample mean
- $\sigma_{\overline{Y}}$: standard deviation of the sample mean
  - called the *standard error* of the sample mean

$$\mu_{\overline{Y}} = \mu$$

$$\sigma^2_{\overline{Y}} = Var(\overline{Y}) = \frac{\sigma^2}{n}$$

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

- $Y_1, \ldots, Y_n$: n independent *normal* r.v.'s with mean μ and standard deviation σ:
  - $\overline{Y}$ normally distributed
  - mean μ
  - standard deviation $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem

- $Y_1, \ldots, Y_n$: n independent r.v.'s with mean μ and finite standard deviation σ.
- $\overline{Y} = (Y_1 + Y_2 + \cdots + Y_n)/n$.
- For n sufficiently large, the sampling distribution of $\overline{Y}$ becomes approximately normal with:
  - mean $\mu_{\overline{Y}} = \mu$
  - standard deviation $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$

Precise meaning

- Standardize $\overline{Y}$, to produce Z.
- As $n \to \infty$, the distribution of Z becomes standard normal ie. $N(0, 1)$ where

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

- For many statistical purposes, sufficient accuracy is typically achieved with $n \geq 30$.

- Introduction
- Method of moment
- Method of maximum of likelihood

Example 1: Original Data-Alpha Decay of Americium-241 (Berkson, 1966). The experimenters recorded **10,220** times between successive emissions. The first two columns of the following table display the counts, n, that were observed in 1207 intervals, each of length **10** sec. Our aim is to estimate the mean emissions rate per sec (or per 10 sec).

Emissions (n)	Observed Counts	Expected Counts
0	1	0.27
1	4	2.30
2	13	9.63
3	28	26.95
4	56	56.54
5	105	94.90
6	126	132.73
7	146	159.12
8	164	166.92
9	161	155.64
10	123	130.62
11	101	99.65
12	74	69.69
13	53	44.99
14	23	26.97
15	15	15.09
16	9	7.91
17	3	3.91
19	1	0.80
	1207	1207

The Poisson distribution is frequently used as a model for radioactive decay based on three assumptions:

(1) the underlying rate at which the events occur is constant in space or time;
(2) events in disjoint intervals of space or time occur independently;
(3) there are no multiple events.

The probability function of Poisson is

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Since λ is equal to

$$E(X) = \sum_{k=0}^{\infty} kP(X = k),$$

based on the second column, we have an estimated $P(X = k)$, which is Observed Counts$(k)/1207$.

Then the estimator of λ,

$$\hat{\lambda} = \sum_{k=0}^{19} k \times \textit{Observed Counts}(k)/1207 = 8.367$$

which is 0.8367 per sec, which is close to 0.8392 obtained by Berkson (1966) by recording **10,220** experimenters.
The estimated probability function of $P(X = k)$ is

$$\frac{\hat{\lambda}^k e^{-\hat{\lambda}}}{k!},$$

shown in the third column "Expected Counts" based on the estimated value 0.8392 of λ.

Normal distribution $N(\mu, \sigma^2)$

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \tag{2.1}$$

The use of the normal distribution as a model is usually justified using the central limit theorem (CLT), which says the sum of a large number of independent random variables (r.v.'s) is approximately normally distributed.

There are two parameters μ and σ^2 to be estimated.

Gamma distribution $G(\alpha, \lambda)$

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \quad 0 \leq x < \infty. \tag{2.2}$$

Gamma distribution fits to environmental data such as rain fall.

There are two parameters α and λ to be estimated.

We will take the following basic approach to the study of parameter estimation:

(1) Given population distribution, we want to find the estimate of parameter θ in the distribution.

(2) We have the observed data regarded as realizations of r.v.'s $X_1, X_2, \cdots, X_n$ whose joint distribution depends on unknown θ.

(3) An estimate of θ will be a function of $X_1, X_2, \cdots, X_n$, $g(X_1, X_2, \cdots, X_n)$, and tis sampling distribution will be investigated to assess the variability, eg. unbias and efficience by expectation and variance of the estimate.

Two methods will be introduced:

(1) The method of moments

(2) The method of maximum likelihood

The kth moment

$$\mu_k = E(X^k) \tag{2.3}$$

may be related to the unknown parameters to be estimated.

The kth sample moment

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{2.4}$$

given iid sample $X_1, X_2, \cdots, X_n$.

Let $\hat{\mu}_k = \mu_k$, we call this as the method of moment.

Note that we should choose the lowest possible order moment

Example 2: Consider $X_1, \cdots, X_n$ is iid Poisson(λ) with $E(X_i) = Var(X_i) = \lambda$.

Our aim is to show that we should choose the lowest possible order moment in the method of moment.

Two estimators of λ by the method of moment are $\hat{\lambda}_1 = \sum_{i=1}^n X_i/n$ and $\hat{\lambda}_2 = \sum_{i=1}^n (X_i - \sum_{i=1}^n X_i/n)^2/n$.

We use simulations to show that $\hat{\lambda}_1$ is better by applying the lower order moment.
For $n = 20$ and $n = 100$, we take $\lambda = 5$ and $\lambda = 10$. Simulate 1000 times and we obtain $\hat{\lambda}_{1i}, \hat{\lambda}_{2i}$ for $i = 1, \cdots, 1000$. We calculate the sample means and sample variances of them as presented in the following Table.

n	λ	$\hat{E}(\hat{\lambda}_1)$	$\hat{V}ar(\hat{\lambda}_1)$	$\hat{E}(\hat{\lambda}_2)$	$\hat{V}ar(\hat{\lambda}_2)$
20	5	5.0025	0.2656	4.7813	2.5178
	10	10.0406	0.5163	9.5677	9.8858
100	5	4.9864	0.0500	4.9440	0.5278
	10	9.9916	0.0964	9.8545	2.0582

The Table indicates $\hat{\lambda}_1$ is more accurate than $\hat{\lambda}_2$, why?