Biostat 602 Winter 2016

Lecture Set 1

Review of the Past

Introduction

In scientific research, an investigator often uses one of two types of reasoning, namely the *deductive reasoning* and the *inductive reasoning*.

Deductive Reasoning

- works from the general to specific; typically based on some general laws or rules which are then applied to a specific case.
- We make an assumption about a population and want specifics of a sample
- Suppose the lifetime of a particular bt=rand of car battery has an exponential distribution with a median of 7 years. We want to determine what percentage of these batteries that will last at least 10 years.
- Subject of **Biostat 601**

Inductive Reasoning

- generalizes the conclusion of findings observed from a specific.
- Suppose a particular supplier is providing batteries to a hardware manufacturer and it is intended to estimate the lifetime distribution of this particular brand and substantiate the manufacturer's claim that 90% of the batteries last over 700 hours.
- Typically one would select a random sample of batteries from the batch provided by the supplier and run a life-test on them
- Based on the findings from the sample estimate the distribution of the lifetime and test out the claim
- Subject of statistical inference (**Biostat 602**)

Review of Biostat 601

Probability

Let S be the sample space related to a random experiment. Probability is a set function with range in [0, 1] defined on all subsets of S satisfying:

i. $P(E) \geq 0$, for any event $E \subset S$.

ii. $P(S) = 1$.

iii. If $E_1, E_2, \ldots$ are mutually exclusive, (i.e. $E_i \cap E_j = \phi, \quad i \neq j$), then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \qquad \textit{(countable additivity)}$$

Laws of Probability:

- *Addition Law*

  For any finite set of events $E_1, E_2, \ldots, E_n$,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum\sum_{i<j} P(E_iE_j) + \sum\sum\sum_{i<j<k} P(E_iE_jE_k) + \cdots + (-1)^{n+1} P(E_1E_2\cdots E_n).$$

- *Boole's Inequality*

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i)$$

- *Bonferroni's Inequality*

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geq \sum_{i=1}^{n} P(E_i) - (n-1)$$

- *Law of complementation*
- *Multiplication Law (Conditional Probability)*
- *Law of Independence*
- *Law of Total Probability*

  Suppose $A_1, A_2, \ldots, A_n$ are mutually exclusive and exhaustive events, i.e. $A_i \cap A_j = \phi, \;\; i \neq j$ and $S = \cup_{i=1}^{n} A_i$. Let B be any event in S. Then

$$P(B) = \sum_{i=1}^{n} P(A_i) P(B|A_i).$$

- *Bayes Theorem*

  Suppose $F_1, F_2, \ldots, F_n$ are **mutually exclusive** and **exhaustive** events, i.e. one and only one of them must occur. Suppose for some $j, j = 1, \ldots, n$, we are interested in the conditional probability of F_j given another conditioning event E, i.e. $P(F_j \mid E)$. Bayes' Theorem states that it can be obtained using the *reverse* conditional probability as

$$P(F_j \mid E) = \frac{P(E \mid F_j) P(F_j)}{\displaystyle\sum_{i=1}^{n} P(E \mid F_i) P(F_i)}.$$

Example 1

- The ELISA (Enzyme-Linked Immunosorbent Assay) test is used to detect antibodies in blood and can indicate the presence of the HIV virus.

- Approximately 5% of a population is HIV positive.

- Among those who have HIV virus, 96% test positive with ELISA (Sensitivity).

- Among those who do not have HIV virus, approximately 98% test negative with ELISA (Specificity).

- For a randomly chosen subject from this population if the test is positive, what is the probability that the subject has HIV virus?

Diagnostic Testing Nomenclature

Disease	Test Results +	Test Results −
+	TP	FN
−	FP	TN

In any diagnostic test, there are four quantities which people are interested in:

Sensitivity: Probability of True Positives, i.e. probability of the test result being positive for a diseased individual ($TP/(TP + FN)$)

Specificity: Probability of True Negatives, i.e. probability of the test result showing negative finding for an individual w/o the disease ($TN/(FP + TN)$)

Positive Predictive Value: probability of the individual truly having the disease when the test result is positive ($TP/(TP + FP)$)

Negative Predictive Value: probability of the individual not having the disease when the test result is negative ($TN/(TN + FN)$)

Remarks

- High values of all four quantities are desirable for a diagnostic test.
- In designing the test, care is taken to maintain a reasonably high level of sensitivity and specificity. These two, along with the prevalence of the disease determine the predictive values.
- In our example, sensitivity, specificity are provided. We want to find the positive predictive value.

Back to AIDS example

- Let H = subject has HIV virus, and Pos = test result is positive.
- It is given that

$$P(H) = 0.05, \quad P(Pos \mid H) = 0.96, \quad P(Pos \mid H^c) = 0.02.$$

- Want to find $P(H \mid Pos)$.
- By definition of conditional probability

$$P(H \mid Pos) = \frac{P(H \cap Pos)}{P(Pos)}$$

- Now

$$P(H \ \cap \ Pos) = P(Pos \mid H)P(H) = 0.96 \times 0.05 = 0.048,$$

and

$$\begin{aligned} P(Pos) &= P(Pos \cap H) + P(Pos \cap H^c) \\ &= P(Pos \mid H)P(H) + P(Pos \mid H^c)P(H^c) \\ &= (0.96)(0.05) + (0.02)(0.95) = 0.067. \end{aligned}$$

- The required probability equals $0.048/0.067 = 0.716$.

Random Variables

A random variable Y is a real-valued function defined on a probability space.

- **Discrete Random Variables:** Probability mass function (pmf), Cumulative Distribution Function (cdf), Calculation of Expectation, Variance from a pmf

- **Continuous Random Variables:** Probability density function (pdf), Cumulative Distribution Function (cdf), Calculation of Expectation, Variance from a given pdf.

- Common Families of Discrete Distributions (Binomial, Poisson, Geometric, Negative Binomial)

- Common Families of Continuous Distributions (Normal, t, chi^2, F, Exponential, Gamma)

Example 2: A point is chosen at random on a line segment of length L. Find the probability that the ratio of the shorter to the longer segment is less than 1/4.

Solution: The given information tantamount to saying that a point randomly picked on the line segment has a length X which is has a *uniform* distribution on $(0, L)$. We are interested in

$$P\left(\frac{\min(x, L-x)}{\max(x, L-x)} \leq 1/4\right).$$

Now note that for $x < L/2$, $\min(x, L-x) = x$ and,

$$\frac{\min(x, L-x)}{\max(x, L-x)} \leq 1/4 \Rightarrow \frac{x}{L-x} \leq 1/4 \Rightarrow x \leq L/5.$$

For $x \geq L/2$, $\min(x, L-x) = L - x$ and,

$$\frac{\min(x, L-x)}{\max(x, L-x)} \leq 1/4 \Rightarrow \frac{L-x}{x} \leq 1/4 \Rightarrow x \geq 4L/5$$

So the required probability equals

$$\begin{aligned}
P[X \leq L/5] + P[X \geq 4L/5] &= \int_0^{\frac{L}{5}} \frac{1}{L}\,dx + \int_{\frac{4L}{5}}^{L} \frac{1}{L}\,dx \\
&= \frac{1}{5} + \frac{1}{5} \\
&= \frac{2}{5}.
\end{aligned}$$

Example 3: Suppose that the travel time from Adam's home to his office is a normally distributed random variable with mean = 40 minutes and standard deviation = 7 minutes.

(a) What proportion of time Adam reaches office within 38 and 45 minutes of leaving home?

Solution: Let X denote Adam's travel time. We need to find $P[38 < X < 45]$. Note

$$\begin{aligned}
P[38 < X < 45] &= P\left[\frac{38-40}{7} < Z < \frac{45-40}{7}\right] \\
&= P\left[Z < \frac{45-40}{7}\right] - P\left[Z < \frac{38-40}{7}\right] \\
&= P[Z < 0.714] - P[Z < -0.286] \\
&= \Phi(0.71) - \Phi(-0.29) = .7611 - .3859 = .3752.
\end{aligned}$$

(b) If Adam wants to be 95% certain that he will not be late for an office appointment at 1 PM, what is the latest time he should leave home?

Solution: This falls under a class of problems involving *inverse transformation*. In these problems, one is interested in finding for a normal random variable X the $100p-th$ percentile x_p. So x_p satisfies the equation $P[X < x_p] = p$; it is the point to the left of which lies $100p\%$ of the distribution. One solves the problem in the following two steps.

Step 1: Calculate $100p-th$ percentile z_p of Z, that satisfies $P[Z < z_p] = p$.

Step 2: Find x_p using the formula $x_p = \mu + \sigma z_p$.

In our problem, we need to find the 95th percentile of the distribution of X. Using the *qnorm* function in R, $z_{0.95} = 1.645$, and

$$x_{0.95} = 40 + 7(1.645) = 51.515.$$

So, Adam needs to leave his home latest by 12:08 PM.

Multiple Random Variables

- Probability calculations from bivariate distributions
- Bivariate transformations, calculating jacobian, joint to marginal and conditional distribution
- Finding marginal distributions from a hierarchical structure
- Applying Conditional Expectation and Variance formula in Hierarchical Models

$$E(Y) = E\left[E(Y|X)\right], \quad Var(Y) = E\left[Var(Y|X)\right] + Var\left[E(Y|X)\right].$$

- Applying variance and covariance formula for linear combinations

$$Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j),$$

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum\sum\nolimits_{i<j} a_i a_j \; Cov(X_i, X_j).$$

- Chebyshev's Inequality

  If μ and σ are the mean and standard deviation of a random variable X, then for any positive constant k and $\sigma > 0$,

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

- Jensen's Inequality

  For any random variable X, if $g(x)$ is a convex function, then

$$E\left[g(X)\right] \geq g\left(E(X)\right).$$

Example 4: Let X, Y have joint pdf

$$f(x,y) = \begin{cases} cxy & 0 \leq x \leq y < 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

(a) Find c.

(b) Find $P(X+Y \leq 1)$.

(c) Find $E(Y|X=x)$.

Example 5: Suppose X_1, X_2 have the joint pdf

$$f_{X_1,X_2}(x_1, x_2) = 16x_1^3 x_2^3, \quad 0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 1.$$

Consider the transformation to $Y_1 = X_1\sqrt{X_2}$ and $Y_2 = X_2\sqrt{X_1}$. Find the joint density of Y_1 and Y_2. Are they independent?

Example 6: Drugs and HIV

$$
\begin{aligned}
N &= \text{No. of drug injections during specified time period} \\
X_i &= \begin{cases} 1 \text{ if needle is contaminated with HIV} \\ 0 \text{ otherwise} \end{cases} \\
S &= \text{No. of contaminated needles used in time period}
\end{aligned}
$$

$$S|N = n \sim Binomial(n, \theta), \quad N \sim Poisson(\lambda).$$

$$
\begin{aligned}
P(S = s) &= \sum_{n=0}^{\infty} P(S = s|N = n)P(N = n) \\
&= \sum_{n=s}^{\infty} \binom{n}{s} \theta^s (1-\theta)^{n-s} e^{-\lambda} \frac{\lambda^n}{n!} \\
&= e^{-\lambda} (\lambda\theta)^s \sum_{n=s}^{\infty} \binom{n}{s} \frac{\{\lambda(1-\theta)\}^{n-s}}{n!} \\
&= e^{-\lambda} \frac{(\lambda\theta)^s}{s!} \sum_{n=s}^{\infty} \frac{\{\lambda(1-\theta)\}^{n-s}}{(n-s)!} \\
&= e^{-\lambda} \frac{(\lambda\theta)^s}{s!} \sum_{n=0}^{\infty} \frac{\{\lambda(1-\theta)\}^{n}}{n!} \qquad \text{(change of index)} \\
&= e^{-\lambda} \cdot e^{\lambda(1-\theta)} \frac{(\lambda\theta)^s}{s!} \\
&= e^{-\lambda\theta} \frac{(\lambda\theta)^s}{s!}
\end{aligned}
$$

$S \sim Poisson(\lambda\theta)$.

Random Samples

- Basic objective in statistical inference is to estimate population parameters of interest, such as mean, median, sd, prevalence, odds.
- Inference on the population parameters is based on the corresponding measure derived from a sample. For example, the prevalence of a chronic condition in a certain population can be estimated on the basis of the proportion of individuals having this condition in a *random sample* drawn from the population.
- A random sample is a collection of random variables.
- A collection of random variables $X_1, X_2, \ldots, X_n$ is called a **random sample** of size n from a population with pdf/pmf $f(x)$ if
  1. $X_1, X_2, \ldots, X_n$ are mutually independent;
  2. The marginal pdf or pmf of X_i is the same as $f(x)$.
- Alternatively, we say $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, expressed as

$$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} f(x)$$

- The joint pdf or pmf of $X_1, X_2, \ldots X_n$ (also called the *likelihood function*) is

$$f(x_1, \ldots, x_n) = f(x_1) \times f(x_2) \times \ldots \times f(x_n) = \prod_{i=1}^{n} f(x_i)$$

Properties of sample mean and variance

Result: Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

(a) $E(\overline{X}) = \mu$.

(b) $Var(\overline{X}) = \sigma^2/n$.

(c) $E(S^2) = \sigma^2$.

(d) $Var(S^2) = \left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)/n$, where μ_4 is the fourth central moment of the population.

Properties of sample mean and variance from Normal population

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, and let

$$\overline{X} = \left(\sum_{i=1}^{n} X_i\right) \Big/ n \quad \text{and} \quad S^2 = \left\{\sum_{i=1}^{n}(X_i - \overline{X})^2\right\} \Big/ (n-1).$$

- **Result 1:** $\overline{X}$ and S^2 are independent random variables.
- **Result 2:**
$$\overline{X} \sim N(\mu, \sigma^2/n).$$
- **Result 3:**
$$(n-1)S^2/\sigma^2 \sim \chi^2(n-1).$$
- **Result 4:**
$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

- **Result 5:** Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu_X, \sigma_X^2)$ population, and $Y_1, \ldots, Y_m$ is a random sample from an independent $N(\mu_Y, \sigma_Y^2)$ population. Then

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.$$

- **Result 6:** Suppose $X_1, \ldots, X_n$ is a random sample from an arbitrary distribution F. Define $\overline{X}$ and S^2 as above. Then $\overline{X}$ and S^2 are *independently* distributed *if and only if* F is normal.

Order Statistics

Consider a continuous population. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d with cdf and pdf $F_Y(y)$, $f_Y(y)$, respectively. The ordered observations

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

are called order statistics. For example, the *minimum* is $Y_{(1)}$ and the *maximum* is $Y_{(n)}$. We are interested in finding the distribution of an arbitrary $Y_{(i)}$, as well as the joint distributions of sets of $Y_{(i)}$'s and $Y_{(j)}$'s.

I. Distribution of $Y_{(r)}$

Marginal pdf of the r-th order statistic is

$$f_{Y_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!} F(y)^{r-1}[1 - F(y)]^{n-r} f(y)$$

II. Joint distribution of $Y_{(r)}, Y_{(s)}, \quad r < s$

Joint pdf of any pair of order statistics Y_r, Y_s is given by

$$\begin{aligned} f_{Y_{(r)},Y_{(s)}}(u,v) &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F_Y(u)^{r-1} \\ &\quad \times \left[F_Y(v) - F_Y(u)\right]^{s-r-1} (1 - F_Y(v))^{n-s} f_Y(u) f_Y(v) \end{aligned}$$

III. Joint distribution of first r order statistics, $r < n$

Joint pdf of $Y_{(1)}, \ldots, Y_{(r)}$ from a sample of size n is

$$f_{Y_{(1)},\ldots,Y_{(r)}}(u_1, \ldots, u_r) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f_Y(u_i) \Big(1 - F(u_r)\Big)^{n-r}, \quad u_1 < u_2 < \ldots < u_r.$$

Large Sample Theory

Convergence of a sequence of random variables

A sequence of random variables $\{X_n\}$ is said to converge, as $n \longrightarrow \infty$,

(i) almost surely (or with probability 1) to a random variable X (Notation: $X_n \xrightarrow{a.s.} X$) if for any $\epsilon > 0$

$$P\left[\lim_{n\to\infty} |X_n - X| > \epsilon\right] = 0.$$

(ii) in probability to a random variable X (Notation: $X_n \xrightarrow{P} X$) if for any $\epsilon > 0$

$$\lim_{n\to\infty} P\Bigg[|X_n - X| > \epsilon\Bigg] = 0.$$

(iii) in distribution to a random variable X (Notation: $X_n \xrightarrow{d} X$) if

$$\lim_{n\to\infty} P(X_n \le x) = \lim_{n\to\infty} F_{X_n}(x) = F_X(x) = P(X \le x)$$

at all continuity points of $F_X(x)$.

(iv) in p-th mean to a random variable X (Notation: $X_n \xrightarrow{L_p} X$) if

$$\lim_{n\to\infty} E\Bigg[|X_n - X|^p\Bigg] = 0.$$

Example 7: Suppose $X_1, X_2, \dots X_n$ be a random sample from a *lomax* distribution with parameter σ having pdf

$$f_X(x) = \frac{1}{\sigma\left(1 + \frac{x}{\sigma}\right)^2}, \qquad x > 0, \sigma > 0.$$

(a) Let $X_{(1)}$ be the minimum based on the random sample. Show that $nX_{(1)} \xrightarrow{d} Exp(\sigma)$ as $n \longrightarrow \infty$.

(b) Show that $X_{(1)} \xrightarrow{P} 0$ as $n \longrightarrow \infty$.

<u>Proof:</u>

Example 8: Suppose $X_1, X_2, \dots X_n$ be a random sample from a *lomax* distribution with parameter σ having pdf

$$f_X(x) = \frac{1}{\sigma\left(1 + \frac{x}{\sigma}\right)^2}, \qquad x > 0, \sigma > 0.$$

(a) Let $X_{(1)}$ be the minimum based on the random sample. Show that $nX_{(1)} \xrightarrow{d} Exp(\sigma)$ as $n \longrightarrow \infty$.

(b) Show that $X_{(1)} \xrightarrow{P} 0$ as $n \longrightarrow \infty$.

<u>Proof:</u>

Slutsky's theorem

If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} b$, and $Z_n \xrightarrow{P} a$, where a and b are constants, then

$$Z_n X_n + Y_n \xrightarrow{d} aX + b.$$

Weak law of large numbers

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E(Y_i) = m$ and $\mathrm{V}(Y_i) = \sigma^2$. Then $\overline{Y}_n = (Y_1 + \cdots + Y_n)/n \xrightarrow{P} m$

Strong law of large numbers

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of i.i.d. random variables with $E(Y_i) = m < \infty$. Then the Strong Law of Large Numbers states that $\overline{Y}_n \xrightarrow{a.s.} m$. In other words,

$$P\left\{\lim_{n\to\infty} \overline{Y}_n = m\right\} = 1.$$

Example 9: Let $X_n \sim F(n, n)$, a F distribution with n and n degrees of freedom. Show that as $n \longrightarrow \infty$,

$$X \xrightarrow{P} 1, \quad X \xrightarrow{a.s.} 1.$$

Central Limit Theorem (Laplace)

Let Y_i for $i = 1, 2, \ldots, n$, be i.i.d. each with finite mean $\mu < \infty$ and finite variance $\sigma^2 < \infty$. Then, the *Central Limit Theorem* states that

$$Z_n = \frac{(\overline{Y}_n - \mu)}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1).$$

This implies $\lim_{n\to\infty} P(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-x^2/2)dx.$

Example 10: Let $X_n \sim gamma(n, \beta)$.

(a) Show that $\frac{X_n}{n} \xrightarrow{P} \beta$.

(b) What is the limiting distribution of suitably scaled and centered X_n/n?

Delta Method

Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g^{(1)}(\cdot)$ exists, continuous, and $g^{(1)}(\theta) \neq 0$. Then

$$\sqrt{n}\left[g(Y_n) - g(\theta)\right] \xrightarrow{d} N\left\{0, \sigma^2 \left[g^{(1)}(\theta)\right]^2\right\}$$

Example 11: Let $X_n \sim gamma(n, \beta)$. Define $Y_n = X_n/n$.

(a) Obtain the limiting distribution of $\sqrt{n}(Y_n - \beta)$.

(b) Obtain the limiting distribution of $\sqrt{n}(\log(Y_n) - \log(\beta))$.

(c) What is the limiting distribution of (scaled and centered) Y_n^{-1}?

Example 12: Let $X_1, X_2, \ldots, X_n$ be a random sample from $Bernoulli(p)$. Consider the transformation function $g(x) = x(1-x)$. Find the large-sample distribution of suitably scaled and centered random variable $g(\overline{X}_n)$.