Lecture 4

Ciprian Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

August 12, 2016

Table of contents

Outline

1. Define random vectors
2. Independent events and variables
3. IID random variables
4. Covariance and correlation
5. Standard error of the mean
6. Unbiasedness of the sample variance

Random vectors

- Random vectors are random variables collected into a vector
  - If X and Y are random variables (X, Y) is a random vector
  - If $X_1, \ldots, X_n$ are random variables $(X_1, \ldots, X_n)$ is a random vector
  - The "columns" of most common data structures are realizations of a random vector. Each column is a realization of one random variable
- Joint density $f(x, y)$ satisfies $f(x, y) \geq 0$ for every x and y and $\int\int f(x, y)dxdy = 1$
- For discrete random variables $\sum_x \sum_y f(x, y) = \sum_x \sum_y P(X = x, Y = y) = 1$
- In this lecture we focus on **independent** random variables where $f(x, y) = f(x)g(y)$

Independent events

- Two events A and B are **independent** if

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

- Two random variables, X and Y are independent if for any two sets A and B

$$P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$$

- If A is independent of B then
  - A^c is independent of B
  - A is independent of B^c
  - A^c is independent of B^c

Example

- What is the probability of getting two consecutive heads?
- $A = \{\text{Head on flip 1}\} \quad P(A) = .5$
- $B = \{\text{Head on flip 2}\} \quad P(B) = .5$
- $A \cap B = \{\text{Head on flips 1 and 2}\}$
- $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

Example

- Volume 309 of *Science* reports on a physician who was on trial for expert testimony in a criminal trial
- Based on an estimated prevalence of sudden infant death syndrome of 1 out of $8,543$, Dr Meadow testified that that the probability of a mother having two children with SIDS was $\left(\frac{1}{8{,}543}\right)^2$
- The mother on trial was convicted of murder
- What was Dr Meadow's mistake(s)?

Example: continued

- For the purposes of this class, the principal mistake was to *assume* that the probabilities of having SIDs within a family are independent
- That is, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$
- Biological processes that have a believed genetic or familiar environmental component, of course, tend to be dependent within families
- In addition, the estimated prevalence was obtained from an *unpublished* report on single cases; hence having no information about recurrence of SIDs within families

Useful fact

- We will use the following fact extensively in this class:

  If a collection of random variables $X_1, X_2, \ldots, X_n$ are independent, then their joint densities or mass functions is the product of their individual densities or mass functions

  That is, if f_i is the density for the random variable X_i we have that

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

- The X_i variables do not need to have the same distribution

Simulating independent discrete variables

Independent Bernoulli draws

```
x<-rbinom(10,1,prob=0.3)
bernm<-seq(0,1,by=0.1)
x<-rbinom(3*length(bernm),1,prob=bernm)
mx=matrix(x,ncol=length(bernm),byrow=TRUE)
```

Independent Poisson draws

```
x<-rpois(10000,20)
poism<-c(1,2.5,5,7.5,10,1000)
x<-rpois(24,poism)
```

Simulating independent Normal variables

Independent Normal draws

```
x<-rnorm(1000,mean=2,sd=9)
normm<-1:100
sdm<-normm/3
x<-rnorm(150*length(normm),mean=normm,sd=sdm)
mx=matrix(x,ncol=length(normm),byrow=TRUE)
```

Checking results

```
dim(mx)
colMeans(mx)
```

IID random variables

- In the instance where $f_1 = f_2 = \ldots = f_n$ we say that the X_i are **iid** for *independent* and *identically distributed*
- iid random variables are the default model for random samples
- Many of the important theories of statistics are founded on assuming that variables are iid

Example

- Suppose that we flip a biased coin 4 times with success probability p and we obtain (1,0,1,1)
- What is the joint probability mass function of the collection of outcomes?
- Therefore

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1) = p(1-p)pp = p^3(1-p)$$

Example

- Suppose that we flip a biased coin with success probability p n times
- What is the joint probability mass function of the collection of outcomes?
- These are independent random variables X_1, X_2, X_3, X_4
- $P(X_1 = 1) = p$ and $P(X_1 = 0) = 1 - p$. Equivalently $P(X_1 = x_1) = p^{x_1}(1-p)^{1-x_1}$, where x_1 is the outcome of the first coin flip
- In general, the pmf $P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$
- Therefore

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$P(X = x_1, \ldots, x_n) = P(x_1) \; P(x_2) \cdots P(x_n)$

Standard normal

- The standard normal density is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$$

- Mean 0, variance 1
- Suppose that one draws n independent samples, $X_1, \ldots, X_n$ from a distribution with the pdf given above
- What is the joint density of the vector $(X_1, \ldots, X_n)$?

Standard normal

- Let's suppose that $n = 2$ (one makes two independent draws from a standard normal)
- What is $P(X_1 \geq 1.5, X_2 \geq 1)$?

R code

```
probt=(1-pnorm(1.5))*(1-pnorm(1))
nsim=100000
x1=rnorm(nsim)
x2=rnorm(nsim)
probs=mean((x1>1.5) & (x2>1))
probt # display theoretical value
probs # display simulated value
abs(probs-probt)/probt
```

Monte Carlo methods are incredibly powerful for evaluating probabilities.

Product of independent variables

- Assume that X and Y are independent random variables
- Show that $E[XY] = E[X]E[Y]$

Covariance

- The **covariance** between two random variables X and Y is defined as

$$\mathrm{Cov}(X,Y) = E\{(X - E[X])(Y - E[Y])\} = E[XY] - E[X]E[Y]$$

- Prove this result
- If X and Y are independent then $\mathrm{Cov}(X,Y) = 0$
- The following are useful facts about covariance
  1. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
  2. $\mathrm{Cov}(X,Y)$ can be negative or positive
  3. $|\mathrm{Cov}(X,Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$

Correlation

- The **correlation** between X and Y is

$$\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$$

1. $-1 \leq \mathrm{Cor}(X,Y) \leq 1$ $\Rightarrow$ $|cov| \leq \sqrt{var(x) \cdot var(y)}$
2. $\mathrm{Cor}(X,Y) = \pm 1$ if and only if $X = a + bY$ for some constants a and b
3. $\mathrm{Cor}(X,Y)$ is unitless
4. X and Y are **uncorrelated** if $\mathrm{Cor}(X,Y) = 0$
5. X and Y are more positively correlated, the closer $\mathrm{Cor}(X,Y)$ is to 1
6. X and Y are more negatively correlated, the closer $\mathrm{Cor}(X,Y)$ is to -1

Zero correlation does not imply independence

- Consider X a normal random variable with mean zero and variance one
- Show that $\mathrm{Cor}(X, X^2) = 0$
- Show that $P(X > 1, X^2 > 1) \neq P(X > 1)P(X^2 > 1)$

Zero correlation among the entries of a random normal vector implies independence

Intra-class correlation

- Consider the case when one is interested in measuring the systolic blood pressure (SBP) in a population
- Take two measurements (replication study). For example, in two consecutive days
- Denote by W_{ij} the measurement for subject $i = 1, \ldots, n$ on day $j = 1, 2$
- $\mathrm{Cor}(W_{i1}, W_{i2})$ is the intra-class correlation (ICC) coefficient
- An estimator of this coefficient is

$$\widehat{\mathrm{Cor}}(W_{i1}, W_{i2}) = \widehat{\mathrm{Cov}}(W_{i1}, W_{i2}) / \widehat{\mathrm{Var}}(W_{i1})$$

Classical measurement error and ICC

- If X_i is the true long term SBP of subject i then the classical measurement error model assumes

$$W_{ij} = X_i + U_{ij}$$

where U_{ij} are the measurement errors

- U_{ij} are mutually independent and independent of X_i, $i = 1, \ldots, n$
- It can be shown that

$$\mathrm{Cor}(W_{i1}, W_{i2}) = \frac{\mathrm{Var}(X_i)}{\mathrm{Var}(X_i) + \mathrm{Var}(U_{ij})} = \frac{\mathrm{Var}(X_i)}{\mathrm{Var}(W_{i1})}$$

- ICC is sometimes called the reliability of the replication study

Classical ME and ICC in R

```
# Simulate true SBP
X<-rnorm(200,130,10)
# Simulate contamination
U<-matrix(rnorm(400,m=0,sd=10),ncol=2)
# Obtain contaminated variables
W<-X+U
cor(W)
```

10/10+10

- True ICC (reliability) was 0.5 (how do I know that?)
- Things are a bit more complex with more than 2 replicates

Some useful results

- Let $\{X_i\}_{i=1}^n$ be a collection of random variables
  - When the $\{X_i\}$ are uncorrelated

$$\mathrm{Var}\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X_i)$$

  - Otherwise $Var(X_1+X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1,X_2)$

$$\mathrm{Var}\left(\sum_{i=1}^n a_i X_i + b\right)$$

$$= \sum_{i=1}^n a_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$$

  - If the X_i are iid with variance σ^2 then $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$

$Var\left(\frac{1}{n}(X_1+\cdots+X_n)\right) = \frac{1}{n^2} Var(X_1+\cdots+X_n) = \frac{1}{n^2} n \cdot Var(X)$

Location change variables

- Consider a random variable X with pdf $f_X(x)$ and cdf $F_X(x)$
- If $Y = X + b$ what are its pdf $f_Y(y)$ and cdf $F_Y(y)$?
- What is $E[Y]$?
- What is $\mathrm{Var}(Y)$?

Example proof

Prove that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$

$$\begin{aligned}
& \mathrm{Var}(X+Y) \\
= \; & E[(X+Y)(X+Y)] - E[X+Y]^2 \\
= \; & E[X^2 + 2XY + Y^2] - (\mu_x + \mu_y)^2 \\
= \; & E[X^2 + 2XY + Y^2] - \mu_x^2 - 2\mu_x\mu_y - \mu_y^2 \\
= \; & (E[X^2] - \mu_x^2) + (E[Y^2] - \mu_y^2) + 2(E[XY] - \mu_x\mu_y) \\
= \; & \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)
\end{aligned}$$

Result

- A commonly used subcase from these properties is that *if a collection of random variables* $\{X_i\}$ *are uncorrelated*, then the variance of the sum is the sum of the variances

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

- Therefore, it is sums of variances that tend to be useful, not sums of standard deviations; that is, the standard deviation of the sum of bunch of independent random variables is the square root of the sum of the variances, not the sum of the standard deviations

A pitfall

- Calculating the SD of a sum of independent variables

$$\mathrm{SD}(\sum_i X_i) = \sqrt{\mathrm{Var}(\sum_i X_i)} = \sqrt{\sum_i \mathrm{Var}(X_i)}$$

- Sum of standard deviations

$$\sum_i \mathrm{SD}(X_i) = \sum_i \sqrt{\mathrm{Var}(X_i)}$$

- In general $\mathrm{SD}(\sum_i X_i) < \sum_i \mathrm{SD}(X_i)$
- When X_i are independent with variance 1 then $\mathrm{SD}(\sum_i X_i) = \sqrt{n}$ and $\sum_i \mathrm{SD}(X_i) = n$
- When $n = 100$ the difference is of one order of magnitude!

The sample mean

Suppose X_i are iid with variance σ^2

$$\begin{aligned}
\mathrm{Var}(\bar{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\
&= \frac{1}{n^2} \times n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Variance of sample and mean

Suppose X_i are iid $N(\mu, \sigma^2)$

- The density of X_i is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

- If $X_i \sim N(0,1)$ then $\mu + \sigma X_i \sim N(\mu, \sigma)$
- $E[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2$
- $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$

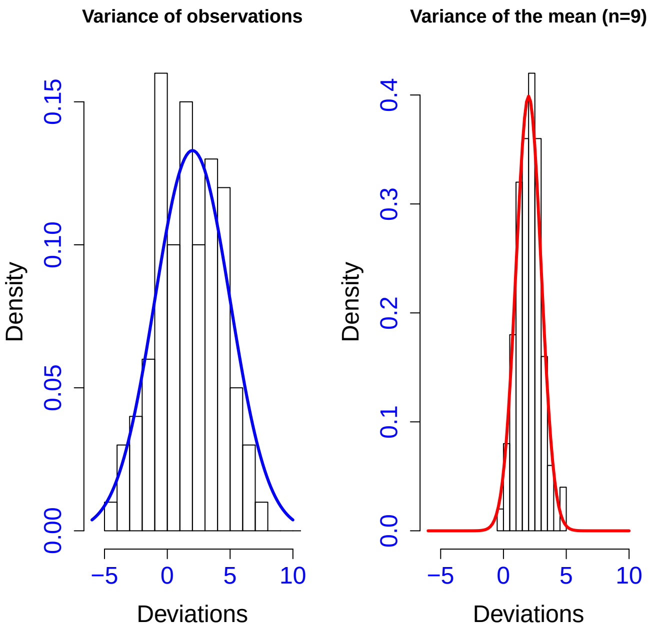
Theoretical and sampling distributions

Distribution of $X \sim N(2, 9)$

```
par(mfrow = c(1, 2))
x=seq(-6,10,length=101)
y=dnorm(x,m=2,sd=3)
ry=rnorm(100,m=2,sd=3)
hist(ry,probability=TRUE,xlim=c(-6,10))
lines(x,y,type="l",col="blue",lwd=3)
```

Distribution of $\bar{X}_9 \sim N(2, 1)$

```
ym=dnorm(x,m=2,sd=1)
rym<-rep(0,100)
for (i in 1:100)
  {rym[i]<-mean(rnorm(9,m=2,sd=3))}
hist(rym,probability=TRUE,xlim=c(-6,10))
lines(x,ym,type="l",col="red",lwd=3)
```


Variance of observations
Density
0.00
0.05
0.10
0.15
−5
0
5
10
Deviations
Variance of the mean (n=9)
Density
0.0
0.1
0.2
0.3
0.4
−5
0
5
10
Deviations

Some comments

- When X_i are iid $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- $\sigma/\sqrt{n}$ is called **the standard error** of the sample mean
- The standard error of the sample mean is the standard deviation of the distribution of the sample mean
- σ is the standard deviation of the distribution of a single observation
- Easy way to remember, the sample mean has to be less variable than a single observation, therefore its standard deviation is divided by a $\sqrt{n}$

Mixture of distributions

- Consider the case of mixture of two normals
- $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$
- $f(x) = \pi f_1(x) + (1 - \pi) f_2(x)$
- If $X \sim f(x)$ calculate $E[X]$ and $\mathrm{Var}(X)$

Mean of the mixture of distributions

$$
\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x f(x) dx \\
&= \int_{-\infty}^{\infty} x\{\pi f_1(x) + (1-\pi) f_2(x)\} dx \\
&= \pi \int_{-\infty}^{\infty} x f_1(x) dx + (1-\pi) \int_{-\infty}^{\infty} x f_2(x) dx \\
&= \pi \mu_1 + (1-\pi)\mu_2
\end{aligned}
$$

The expected value of a mixture distribution is the weighted mean of the individual distribution means, where the weights are equal to the proportion of each population.

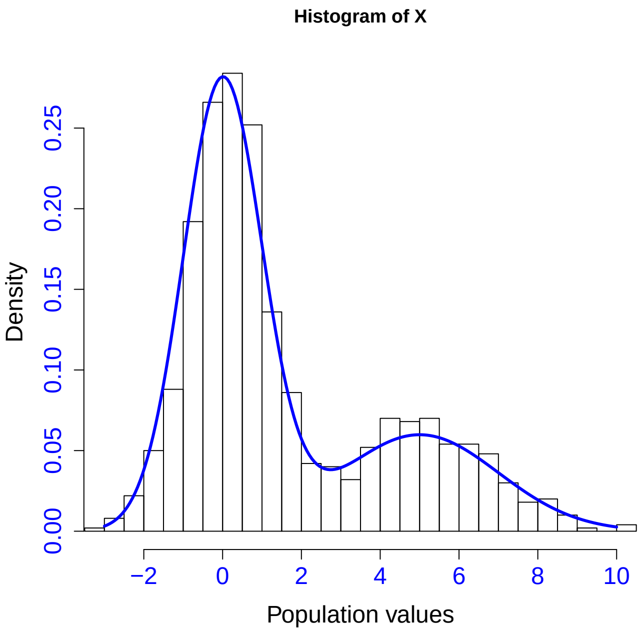
Mixture of distributions: R code

Density of a mixture

```
x=seq(-3,10,length=201)
```

Simulating a mixture of distributions

```
X1<-rnorm(1000)
X2<-rnorm(1000,m=5,sd=2)
U<-rbinom(1000,1,p=.7)
X=U*X1+(1-U)*X2
hist(X,breaks=30,probability=TRUE)
lines(x,dx,type="l",col="blue",lwd=3)
```


Histogram of X
Density
0.00
0.05
0.10
0.15
0.20
0.25
−2
0
2
4
6
8
10
Population values

The sample variance

- The **sample variance** is defined as

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$$

- The sample variance is an estimator of σ^2
- The numerator has a version that's quicker for calculation

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2$$

- The sample variance is (nearly) the mean of the squared deviations from the mean

The sample variance is unbiased

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] &= \sum_{i=1}^{n} E\left[X_i^2\right] - nE\left[\bar{X}_n^2\right] \\
&= \sum_{i=1}^{n}\left\{\mathrm{Var}(X_i) + \mu^2\right\} - n\left\{\mathrm{Var}(\bar{X}_n) + \mu^2\right\} \\
&= \sum_{i=1}^{n}\left\{\sigma^2 + \mu^2\right\} - n\left\{\sigma^2/n + \mu^2\right\} \\
&= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2
\end{aligned}
$$

The sample variance is unbiased

- Is this estimator of the variance unbiased?

$$S_B^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}$$

$\sim (n-1)\sigma^2$

- If it is biased then what is its bias?
- Bias of an estimator is $E[U(X)] - \theta$
- Note that

$$S_B^2 = \frac{n-1}{n} S^2$$

- Show that $\mathrm{Var}(S_B^2) < \mathrm{Var}(S^2)$

Hoping to avoid some confusion

- Suppose X_i are iid with mean μ and variance σ^2
- S^2 estimates σ^2
- The calculation of S^2 involves dividing by $n-1$
- $S/\sqrt{n}$ estimates $\sigma/\sqrt{n}$ the standard error of the mean
- $S/\sqrt{n}$ is called the sample standard error (of the mean)

Example

- In a study of 495 organo-lead workers, the following summaries were obtained for TBV in cm^3
- mean $= 1151.281$
- sum of squared observations $= 662361978$
- sample sd $=$ $\sqrt{(662361978 - 495 \times 1151.281^2)/494} = 112.6215$
- estimated se of the mean $= 112.6215/\sqrt{495} = 5.062$

$$\sum_n (x_i - \bar{X})^2$$

$$= \sum_n X_i^2 - n\bar{X}^2$$

Minimizing sums of squares

Consider the following measure of deviation

- $D(a) = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$
- This is the average of square distances from the sample observations to a point a
- What is the minimum $D(a)$ and where is it attained?