

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and Brian Caffo. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

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 simply random variables collected into a vector
 - ▶ For example if X and Y are random variables (X, Y) is a random vector
- Joint density $f(x, y)$ satisfies $f > 0$ and $\int \int f(x, y) dx dy = 1$
- For discrete random variables $\sum \sum f(x, y) = 1$
- 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?

$$A = \{\text{Head on flip 1}\} \quad P(A) = .5$$

$$B = \{\text{Head on flip 2}\} \quad P(B) = .5$$

$$A \cap B = \{\text{Head on flips 1 and 2}\}$$

$$P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$$

Useful fact

We will use the following fact extensively in this class:

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

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

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

IID random variables

- In the instance where $f_1 = f_2 = \dots = f_n$ 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 with success probability p n times, what is the joint density of the collection of outcomes?

These random variables are iid with densities $p^{x_i}(1-p)^{1-x_i}$

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

Correlation

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

$$\text{Cov}(X, Y) = E[(X - \mu_x)(y - \mu_y)] = E[XY] - E[X]E[Y].$$

The following are useful facts about covariance.

- i.* $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
- ii.* $\text{Cov}(X, Y)$ can be negative or positive.
- iii.* $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\text{Var}(y)}$

Correlation

The **correlation** between X and Y is

$$\text{Cor}(X, Y) = \text{Cov}(X, Y) / \sqrt{\text{Var}(X)\text{Var}(y)}.$$

- i.* $-1 \leq \text{Cor}(X, Y) \leq 1$.
- ii.* $\text{Cor}(X, Y) = \pm 1$ if and only if $X = a + bY$ for some constants a and b .
- iii.* $\text{Cor}(X, Y)$ is unitless.
- iv.* X and Y are **uncorrelated** if $\text{Cor}(X, Y) = 0$.
- v.* X and Y are more positively correlated, the closer $\text{Cor}(X, Y)$ is to 1.
- vi.* X and Y are more negatively correlated, the closer $\text{Cor}(X, Y)$ is to -1 .

Some useful results

Let $\{X_i\}_{i=1}^n$ be a collection of random variables

- When the $\{X_i\}$ are uncorrelated

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

- Otherwise

$$\text{Var} \left(\sum_{i=1}^n a_i X_i + b \right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i}^n a_i a_j \text{Cov}(X_i, X_j).$$

- If the X_i are iid with variance σ^2 then $\text{Var}(\bar{X}) = \sigma^2/n$ and $E[S^2] = \sigma^2$

Example proof

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

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

The sample mean

Suppose X_i are iid with variance σ^2

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

Some comments

- When X_i are independent with a common variance
$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$
- σ/\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}

The sample variance

- The **sample variance** is defined as

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

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

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

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

The sample variance is unbiased

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

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 σ/\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 = $1151.281/\sqrt{495} = 5.062$