Probability - Session 5

Covariance, correlation and the CLT

Elizabeth Williamson
with thanks to Jennifer Rogers

Foundations of Medical Statistics

Session objectives

By the end of this session you should be able to:

- define joint, marginal and conditional distributions for continuous random variables
- explain the concepts of covariance and correlation
- state the Central Limit Theorem and describe its implications
- describe the key properties of the multivariate normal distribution

Outline

Joint probability distributions for continuous random variables

- Previously we introduced the ideas of joint and marginal *discrete* distributions.
- These definitions extend to the case of *continuous* distributions as follows.

Joint density function

- Suppose X and Y are continuous random variables.
- Their **joint density function** $f(x, y)$ is defined such that:

$$P(a \leq X \leq b,\ c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)\, dy\, dx.$$

- The joint density function must satisfy:

$$\begin{aligned} f(x, y) &\geq 0 \text{ for all } x, y \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\ dy\, dx &= 1. \end{aligned}$$

Marginal distributions

- The **marginal** density function of one variable can be obtained from the joint density function:

$$f(x) \;=\; \int_{-\infty}^{\infty} f(x, y)\, dy.$$

The cumulative distribution function (CDF)

- The (joint) **cumulative distribution function** for (X, Y) is defined by:

$$F(x, y) = P(X \leq x \text{ and } Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, dv\, du.$$

Joint distribution

- Can consider joint distributions of mixtures of continuous and discrete random variables.
- Can make corresponding definitions for these joint distributions.
- Details not covered here.

Independence

- Two continuous random variables X and Y are **independent** if and only if:

$$f(x, y) = f(x)f(y) \text{ for all } x, y$$

- Joint density can be factorized into product of marginals.

Outline

Covariance

- Previously we derived the variance of a sum of independent random variables $X + Y$.
- Now suppose X and Y are *not* independent. Then:

$$Var(X+Y) \quad = \quad E\left(\{(X+Y) - E(X+Y)\}^2\right)$$

Covariance

- Previously we derived the variance of a sum of independent random variables $X + Y$.
- Now suppose X and Y are *not* independent. Then:

$$\begin{aligned} Var(X+Y) &= E\left(\{\,(X+Y) - E(X+Y)\,\}^2\right) \\ &= E\left(\{(X+Y) - (E(X) + E(Y))\,\}^2\right) \end{aligned}$$

Covariance

- Previously we derived the variance of a sum of independent random variables $X + Y$.
- Now suppose X and Y are *not* independent. Then:

$$\begin{aligned}
Var(X+Y) &= E\left(\{(X+Y) - E(X+Y)\}^2\right) \\
&= E\left(\{(X+Y) - (E(X) + E(Y))\}^2\right) \\
&= E\left(\{(X - E(X)) + (Y - E(Y))\}^2\right)
\end{aligned}$$

Covariance

- Previously we derived the variance of a sum of independent random variables $X + Y$.
- Now suppose X and Y are *not* independent. Then:

$$
\begin{aligned}
Var(X+Y) &= E\left(\{\,(X+Y) - E(X+Y)\,\}^2\right) \\
&= E\left(\{(X+Y) - (E(X) + E(Y))\,\}^2\right) \\
&= E\left(\{(X - E(X)) + (Y - E(Y))\,\}^2\right) \\
&= E\big((X - E(X))^2 + (Y - E(Y))^2 \\
&\qquad\quad + 2(X - E(X))(Y - E(Y))\big)
\end{aligned}
$$

Covariance

- Previously we derived the variance of a sum of independent random variables $X + Y$.
- Now suppose X and Y are *not* independent. Then:

$$
\begin{aligned}
Var(X+Y) &= E\left(\{(X+Y) - E(X+Y)\}^2\right) \\
&= E\left(\{(X+Y) - (E(X)+E(Y))\}^2\right) \\
&= E\left(\{(X-E(X)) + (Y-E(Y))\}^2\right) \\
&= E\big((X-E(X))^2 + (Y-E(Y))^2 \\
&\qquad +2(X-E(X))(Y-E(Y))\big) \\
&= Var(X) + Var(Y) + 2E((X-E(X))(Y-E(Y))).
\end{aligned}
$$

Covariance

So if we do not assume X and Y are independent:

$$Var(X+Y) = Var(X) + Var(Y) + 2E((X-E(X))(Y-E(Y)))$$

We define the covariance between X and Y as:

$$Cov(X,Y) = E\left((X-E(X))(Y-E(Y))\right).$$

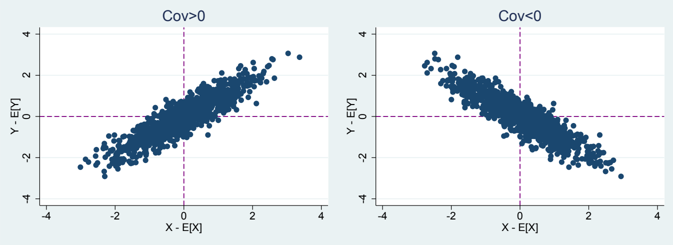
Then:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y),$$

Covariance

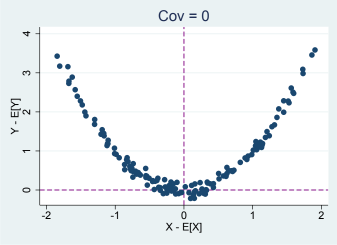
- Covariance measures the magnitude of **linear** association between X and Y.

- Recall $Cov(X, Y) = E((X - E(X))(Y - E(Y)))$.
- $Cov(X, Y) > 0$ when
  - if $X - E(X)$ is positive then $Y - E(Y)$ tends to be positive,
- $Cov(X, Y) < 0$ when
  - if $X - E(X)$ is positive then $Y - E(Y)$ tends to be negative.



More on covariance

- Y and X below appear to be related/associated, yet their covariance is zero.
- This is because covariance measures the magnitude of linear association, and here the association is non-linear, and is such that the linear association is zero.



Some properties of covariance

- $Cov(X, X) = Var(X)$.

Some properties of covariance

- $Cov(X, X) = Var(X)$.
- $Cov(X, Y) = Cov(Y, X)$.

Some properties of covariance

- $Cov(X, X) = Var(X)$.
- $Cov(X, Y) = Cov(Y, X)$.
- $Cov(aX, bY) = abCov(X, Y)$.

Some properties of covariance

- $Cov(X, X) = Var(X)$.
- $Cov(X, Y) = Cov(Y, X)$.
- $Cov(aX, bY) = abCov(X, Y)$.
- $$\begin{aligned} Cov(aR + bS, cX + dY) = {} & acCov(R, X) + adCov(R, Y) \\ & + bcCov(S, X) + bdCov(S, Y). \end{aligned}$$
- $$\begin{aligned} Cov(aX + bY, cX + dY) = {} & acVar(X) + bdVar(Y) \\ & + (ad + bc)Cov(X, Y). \end{aligned}$$
- $Cov(X + Y, X - Y) = Var(X) - Var(Y)$.

Some properties of covariance

- $Cov(X, X) = Var(X)$.
- $Cov(X, Y) = Cov(Y, X)$.
- $Cov(aX, bY) = abCov(X, Y)$.
- $$\begin{aligned} Cov(aR + bS, cX + dY) = {} & acCov(R, X) + adCov(R, Y) \\ & + bcCov(S, X) + bdCov(S, Y). \end{aligned}$$
- $$\begin{aligned} Cov(aX + bY, cX + dY) = {} & acVar(X) + bdVar(Y) \\ & + (ad + bc)Cov(X, Y). \end{aligned}$$
- $Cov(X + Y, X - Y) = Var(X) - Var(Y)$.
- If X and Y are independent, $Cov(X, Y) = 0$ (but **not** vice-versa!).

Proof: $Cov(X, X) = Var(X)$

$$Cov(X, X) = E((X - E(X))(X - E(X)))$$

Proof: $Cov(X, X) = Var(X)$

$$\begin{aligned} Cov(X, X) &= E((X - E(X))(X - E(X))) \\ &= E(X^2 - 2XE(X) + E(X)^2) \end{aligned}$$

Proof: $Cov(X, X) = Var(X)$

$$
\begin{aligned}
Cov(X, X) &= E((X - E(X))(X - E(X))) \\
&= E(X^2 - 2XE(X) + E(X)^2) \\
&= E(X^2) - 2E(X)^2 + E(X)^2
\end{aligned}
$$

Proof: $Cov(X, X) = Var(X)$

$$\begin{aligned} Cov(X, X) &= E((X - E(X))(X - E(X))) \\ &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= Var(X). \end{aligned}$$

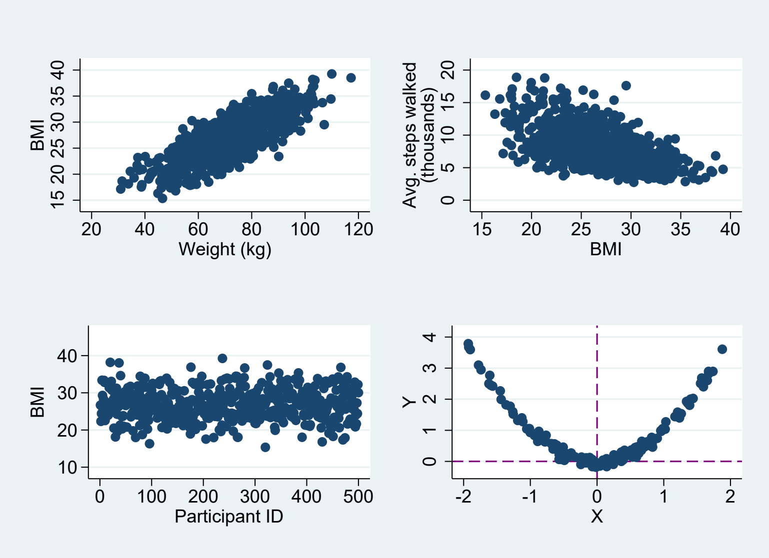
Correlation

- $Cov(X, Y)$ depends on the scale/magnitude of variability of X and Y.
- Correlation is a standardized version of covariance, which lies between -1 and 1:

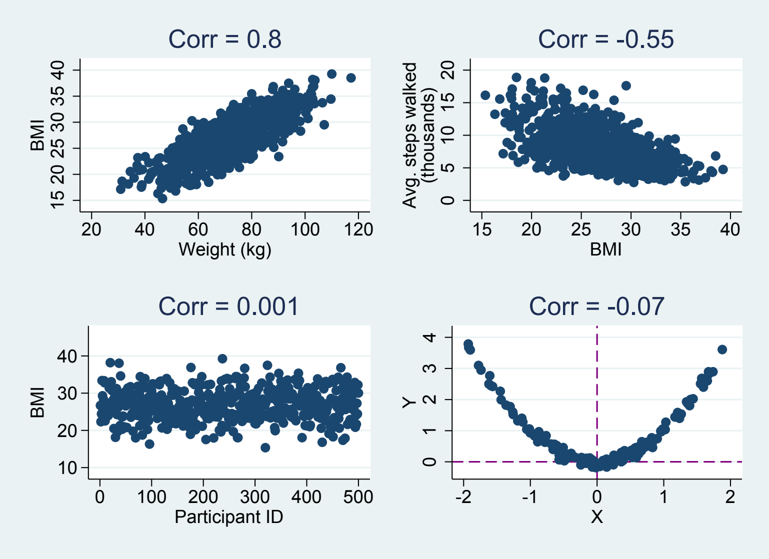
$$Corr(X, Y) = \frac{Cov(X, Y)}{SD(X)SD(Y)} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

- $Corr(X, Y) = 1$ or -1:
  - means X and Y are perfectly correlated.
  - does **not** necessarily mean X and Y are equal.
  - It means $Y = aX + b$ for some constants a and b.
  - e.g. $Y = 2X$ have correlation 1, but are not equal.
- Possible for two variables to be dependent (associated) but have zero correlation.

Correlation - some examples



Correlation - some examples



Outline

Sampling distribution of the mean

- Suppose that:
  - we draw a random sample of size n from a population $(X_1, X_2, ..., X_n)$
  - the $X's$ are independent and identically distributed (i.i.d.)
  - $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

- Define the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- We know that $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$ (Practical 4).
- We often want to make inferences about the mean (Inference)
- To do this, it is useful to know distribution of $\bar{X}_n$ in repeated sampling

The Central Limit Theorem (CLT)

- The CLT is a hugely important theorem in statistics.
- It plays a central role in large sample theory.

- CLT:

  "If you draw a sample of size n from a population and calculate the sample mean, $\bar{X}_n$, the sampling distribution of $\bar{X}_n$ tends to a normal distribution as $n \to \infty$".

- CLT tells us that the distribution of $\bar{X}_n$ is normal, when n is sufficiently large, irrespective of what distribution the individual X_is follow.
- This holds even if the X_is have a discrete distribution.

Illustration: The Central Limit Theorem

- Suppose we are interested in estimating the mean of X
- Suppose the population distribution of X is:

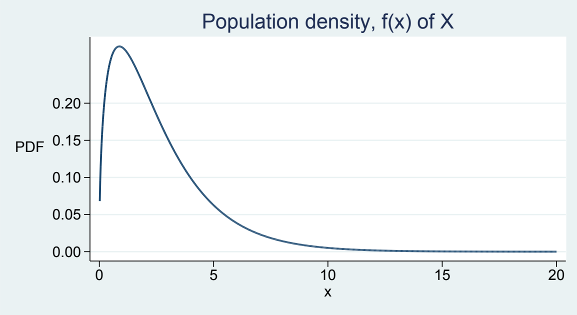
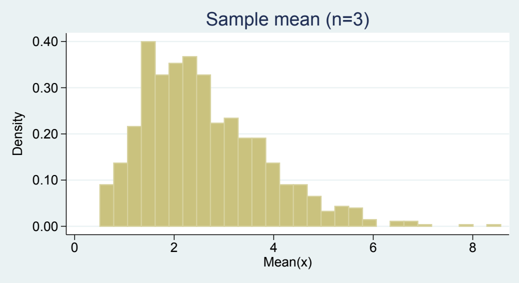


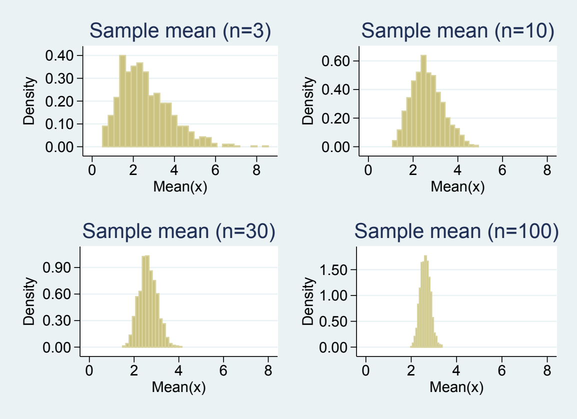
Illustration: The Central Limit Theorem

- Suppose we repeatedly take a random sample of size n from the population, and calculate the mean $\bar{X} = \sum_{i=1}^{n} X_i / n$.

- Here is an example of a histogram of these $\bar{X}$s, with $n = 3$



Distribution of $\bar{X}_n$ for different n



As n increases,

- the distribution of $\bar{X}_n$ appears to become more symmetric, less skewed, and shaped more like a normal distribution.

The CLT: Distribution of $\bar{X}_n$

If samples are i.i.d, with $E(X) = \mu$ and $Var(X) = \sigma^2$, the Central Limit Theorem says that:

- As n increases,

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- An alternative formulation, as n increases,

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2),$$

- Note: this theorem does **not** assume normality of the original distribution of X.
- What becomes normal is the distribution (in repeated sampling) of the sample mean $\bar{X}_n$.

An application of the CLT: normal approximation to the Binomial

There are many applications of the CLT. For example,

- Suppose we have $X \sim Bin(n, \pi)$ with large n.
- Difficult to calculate $P(X = x)$ unless x is small or close to n
- Solution:
  - Think of X as the sum of n Bernoulli trials
  - Apply the CLT.
- The CLT says that for large n,

$$X \sim N(n\pi, n\pi(1 - \pi))$$

- **Caution**:

  this only works well if $n > 20$ and $n\pi > 5$ and $n(1 - \pi) > 5$.

An application of the CLT: normal approximation to the Poisson

- Suppose $X \sim Po(\mu)$, over an interval t. Can we use the CLT to approximate the Poisson by the normal?
- Consider dividing the time t into n equal length intervals. Then the number of events in the ith interval is Poisson:

$$X_i \sim Po\left(\frac{\mu}{n}\right),$$

  with $E(X_i) = \mu/n$ and $Var(X) = \mu/n$.
- The original X can then be viewed as the sum of the X_i, and we can apply the CLT:

$$X = \sum_{i=1}^{n} X_i \sim N\left(\frac{n\mu}{n}, \frac{n\mu}{n}\right) = N(\mu, \mu),$$

- **Caution**: this approximation works well when $\mu > 10$.

Continuity corrections

- If we approximate a discrete distribution by the normal, we should usually use a *continuity correction*.
- i.e. if we want to find $P(X = 15)$, and use a normal approximation, we should calculate $P(14.5 \leq X \leq 15.5)$.

Outline

Two random variables

Suppose we have a two random variables X_1 and X_2, with

$$\begin{aligned} E(X_1) &= \mu_1, \qquad Var(X_1) = \sigma_1^2, \\ E(X_2) &= \mu_2, \qquad Var(X_2) = \sigma_2^2, \end{aligned}$$

and $Corr(X_1, X_2) = \rho$, so $Cov(X_1, X_2) = \rho\sigma_1\sigma_2 = \sigma_{12}$.

In matrix notation, we can write $\mathbf{X} = (X_1, X_2)^T$, i.e.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

which has expectation and variance (covariance matrix)

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad Var(\mathbf{X}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Example: systolic and diastolic blood pressure

Suppose we are interested in the relationship between systolic blood pressure (SBP) and diastolic blood pressure (DBP).

- SBP has a mean of 130, and standard deviation of 15
- DBP has a mean of 90, and standard deviation of 10.
- The correlation between SBP and DBP is 0.75.

Then if $\mathbf{X} = (SBP, DBP)^T$,

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} 130 \\ 90 \end{pmatrix}, \qquad Var(\mathbf{X}) = \boldsymbol{\Sigma} = \begin{pmatrix} 225 & 112.5 \\ 112.5 & 100 \end{pmatrix}$$

The bivariate normal PDF

We say that $\mathbf{X}$ follows a bivariate normal distribution, or $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if

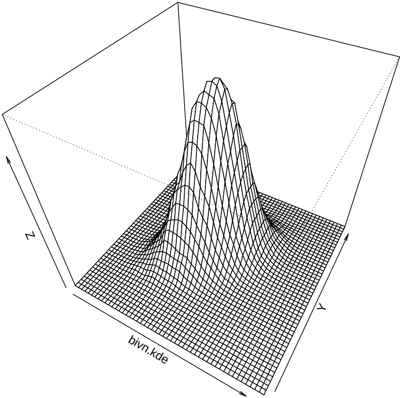
$$f(x_1, x_2) = \frac{exp\left(\frac{-z}{2(1-\rho^2)}\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

with

$$z = \frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}$$

The bivariate normal PDF

A bivariate normal probability density function looks like this:



Conditional and marginal distributions

If $\mathbf{X} = (X_1, X_2)^T$ follows a bivariate normal distribution:

- The marginal distributions of the two variables are normal

$$X_1 \sim N(\mu_1, \sigma_1^2), \qquad X_1 \sim N(\mu_2, \sigma_2^2)$$

- The conditional distribution of X_1 given X_2 is normal, with

$$E(X_1|X_2) = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(X_2 - \mu_2),$$

and

$$Var(X_1|X_2) = \sigma_1^2(1 - \rho^2)$$

(And X_2 given X_1 is similarly normal)

Example: systolic and diastolic blood pressure

Suppose systolic and diastolic blood pressure follow a bivariate normal distribution.

What is the conditional distribution of systolic blood pressure given diastolic blood pressure?

- The conditional expectation is:

$$E(SBP|DBP) = 130 + \frac{0.75 \times 15}{10}(DBP - 90)$$

- E.g. Among people with diastolic blood pressure of 95

$$E(SBP|DBP = 95) = 136.$$

- The conditional variance is equal to:

$$Var(SBP|DBP) \quad = \quad 15^2(1 - 0.75^2) = 98.4.$$

- The conditional standard deviation (9.92) is less than the marginal standard deviation (15).

The multivariate normal

Suppose we have n random variables, $\mathbf{X} = (X_1, .., X_n)^T$

$\mathbf{X}$ follows a multivariate normal (MVN) distribution if its joint density function is

$$f(\mathbf{x}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp(-(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})/2)$$

- $\boldsymbol{\mu}$ is a vector of means: $E(\mathbf{X}) = \boldsymbol{\mu}$.
- Σ is an $n \times n$ covariance matrix: $Var(\mathbf{X}) = \Sigma$.

For a MVN distribution:

- All marginal distributions are also normal
- All conditional distributions are also normal (multivariate if more than one variable).

Outline

Summary

- We have looked at how joint, marginal and conditional distributions are defined for continuous random variables.
- We have described the definitions and meanings of covariance and correlation.
- Covariance and correlation are measures of *linear* association only.
- The Central Limit Theorem is of critical importance for inferential methods, and is also useful in some settings for approximating distributions by the normal.
- Introduced the multivariate normal distribution, and some of its properties.

Overview of Probability

- Session 1
  - Probability as a concept
  - Axioms of probability
  - Conditional probabilities and independence
- Session 2
  - Bayes theorem
  - Random variables
  - Expectation and variance of random variables
- Session 3
  - Discrete probability distributions
  - Combinatorics
  - Binomial and Poisson distributions
- Session 4
  - Continuous probability distributions and density functions
  - Continuous distributions, including the normal
- Session 5
  - Joint distributions
  - Covariance and correlation
  - The Central Limit Theorem