Ch05. Introduction to Probability Theory

Ping Yu

Faculty of Business and Economics
The University of Hong Kong

1. Foundations

2. Random Variables

3. Expectation

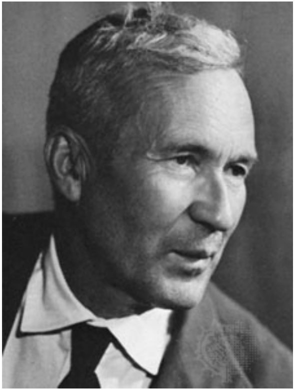
4. Multivariate Random Variables

5. Conditional Distributions and Expectation

6. The Normal and Related Distributions

Foundations

Founder of Modern Probability Theory



Andrey N. Kolmogorov (1903-1987), Russian

- Vladimir Arnold, a student of Kolmogorov, once said: "Kolmogorov – Poincaré – Gauss – Euler – Newton, are only five lives separating us from the source of our science".

Sample Space and Event

- The set Ω of all possible outcomes of an experiment is called the sample space for the experiment.
  - Take the simple example of tossing a coin. There are two outcomes, heads and tails, so we can write $\Omega = \{H, T\}$.
  - If two coins are tossed in sequence, we can write the four outcomes as $\Omega = \{HH, HT, TH, TT\}$.
- An event A is any collection of possible outcomes of an experiment. An event is a subset of Ω, including Ω itself and the null set $\emptyset$.
  - Continuing the two coin example, one event is $A = \{HH, HT\}$, the event that the first coin is heads.

Probability

- A probability function P assigns probabilities (numbers between 0 and 1) to events A in Ω. Probability of an event is the sum of probabilities of the outcomes in the event:

$$P(A) = \frac{\#\text{ of ways (or times) } A \text{ can occur}}{\text{total } \#\text{ of outcomes in } \Omega}.$$

- P satisfies $P(\Omega) = 1$, $P(A^c) = 1 - P(A)$,[1] $P(A) \leq P(B)$ if $A \subseteq B$.
- A probability of 0 means that the event is almost impossible, and a probability of 1 means that the event is almost certain.
- Continuing the two coin example, $P(A) = 1/2$.

[1] This implies $P(\emptyset) = 0$.

Random Variables

Random Variables and CDFs

- A random variable (r.v.) X is a function from a sample space Ω into the real line.
  - The r.v. transforms the abstract elements in Ω to analyzable real values, so is a numerical summary of a random outcome.
  - Notations: we denote r.v.'s by uppercase letters such as X, and use lowercase letters such as x for potential values and realized values.
  - Caution: Be careful about the difference between a random variable and its realized value. The former is a **function** from outcomes to values while the latter is a **value** associated with a specific outcome.
- For a r.v. X we define its cumulative distribution function (cdf) as

$$F(x) = P(X \leq x).$$

  - Notations: Sometimes we write this as $F_X(x)$ to denote that it is the cdf of X.

Discrete Variables

- The r.v. X is discrete if $F(x)$ is a step function.
  - A discrete r.v. can take only finite or countably many values, $x_1, \cdots, x_J$, where J can be ∞
- The probability function for X takes the form of the probability mass function (pmf)

$$P\left(X = x_j\right) = p_j, j = 1, \cdots, J, \tag{1}$$

  where $0 \le p_j \le 1$ and $\sum_{j=1}^J p_j = 1$.
  - $F(x) = \sum_{j=1}^J p_j 1(x_j \le x)$, where $1(\cdot)$ is the indicator function which equals one when the event in the parenthesis is true and zero otherwise.
- A famous discrete r.v. is the Bernoulli (or binary) r.v., where $J = 2$, $x_1 = 0$, $x_2 = 1$, $p_2 = p$ and $p_1 = 1 - p$. [Figure here]
  - The Bernoulli distribution is often used to model sex, employment status, and other dichotomies.

Jacob Bernoulli (1655-1705), Swiss

- Jacob Bernoulli (1655-1705) was one of the many prominent mathematicians in the Bernoulli family.

Continuous Random Variables

- The r.v. X is continuous if $F(x)$ is continuous in x.
  - In this case $P(X = x) = 0$ for all $x \in \mathbb{R}$ so the representation (1) is unavailable.
- We instead represent the relative probabilities by the probability density function (pdf)
$$f(x) = \frac{d}{dx}F(x).$$
  - A function $f(x)$ is a pdf iff $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.
- By the fundamental theorem of calculus,
$$F(x) = \int_{-\infty}^{x} f(u)du$$
and
$$P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(u)du.$$

Examples of Continuous R.V.s

- A famous continuous r.v. is the standard normal r.v. [Figure here]
- The standard normal density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), -\infty < x < \infty,$$

  which is a symmetric, bell-shaped distribution with a single peak
  - Notations: write $X \sim N(0,1)$, and denote the standard normal cdf by $\Phi(x)$. $\Phi(x)$ has no closed-form solution. [Figure here]
- Another famous continuous r.v. is the standard uniform r.v. whose pdf is

$$f(x) = 1(0 \le x \le 1),$$

  i.e., X can occur only on $[0,1]$ and occurs **uniformly**. We denote $X \sim U[0,1]$. [Figure here]
  - A generalization is $X \sim U[a,b]$, $a < b$.

Carl F. Gauss (1777-1855), Göttingen

- The normal distribution was invented by Carl F. Gauss (1777-1855), so it is also known as the Gaussian distribution.

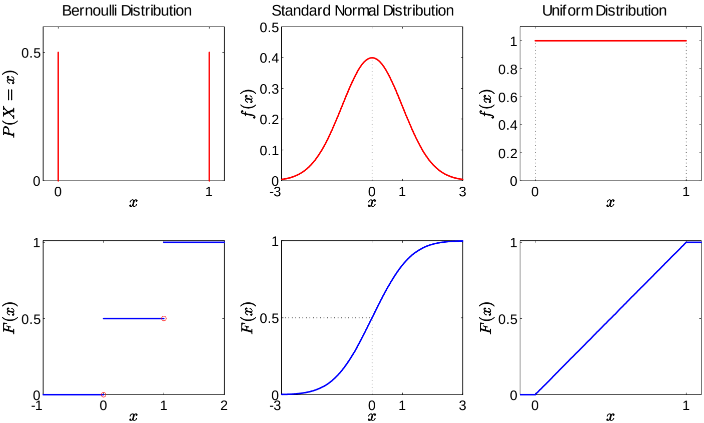


Figure: PMF, PDF and CDF: $p = 0.5$ in the Bernoulli Distribution

Expectation

Expectation

- For any real function g, we define the mean or expectation $E\left[g(X)\right]$ as follows: if X is discrete,

  $$E\left[g(X)\right] = \sum_{j=1}^{J} g(x_j)p_j,$$

  and if X is continuous

  $$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

  - The mean is a **weighted average** of all possible values of $g(X)$ with the weights determined by its pmf or pdf.

- Since $E\left[a + bX\right] = a + b \cdot E\left[X\right]$, we say that expectation is a **linear operator**.

Moments

- For $m > 0$, we define the mth moment of X as $E[X^m]$, the mth central moment as $E[(X - E[X])^m]$, the mth absolute moment of X as $E[|X|^m]$, and the mth absolute central moment as $E[|X - E[X]|^m]$.
- Two special moments are the first moment - the mean $\mu = E[X]$, which is a measure of central tendency, and the second central moment - the variance $\sigma^2 = E\left[(X-\mu)^2\right] = E\left[X^2\right] - \mu^2$, which is a measure of variability or dispersion.
  - We call $\sigma = \sqrt{\sigma^2}$ the standard deviation of X.
  - The definition of variance implies $E\left[X^2\right] = \sigma^2 + \mu^2$, the second moment is the variance plus the first moment squared.
  - We also write $\sigma^2 = Var(X)$, which allows the convenient expression $Var(a + bX) = b^2 Var(X)$.
- The standard normal density has all moments finite, e.g., $\mu = 0$ and $\sigma = 1$.

Standardizing a Random Variable

- For a r.v. X, the standardized r.v. is defined as

$$Z = \frac{X-\mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma}.$$

- Let $a = -\mu/\sigma$ anb $b = 1/\sigma$, we have

$$E[Z] = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\mu = 0,$$

and

$$Var(Z) = \frac{Var(X)}{\sigma^2} = 1.$$

- This transformation is frequently used in statistical inference.

Skewness and Kurtosis

- Skewness: measure of asymmetry of a distribution, $E\left[Z^3\right]=\frac{E\left[(X-\mu)^3\right]}{\sigma^3}$.

  - If X has a symmetric distribution about μ, then its skewness is zero, e.g., the standard normal distribution.

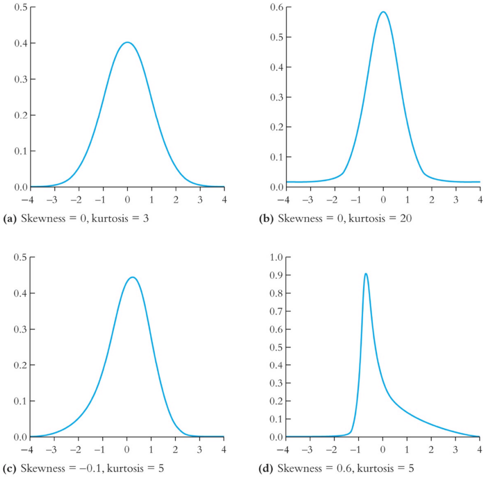
  - If X has a long right tail, then its skewness is positive, and X is called positive- (or right-) skewed.

  - If X has a long left tail, then its skewness is negative, and X is called negative- (or left-) skewed.

- Kurtosis: measure of the heavy-tailedness of a distribution, $E\left[Z^4\right]=\frac{E\left[(X-\mu)^4\right]}{\sigma^4}$.

  - The normal distribution has kurtosis 3: $E\left[Z^4\right]=3$.

  - If the kurtosis of a distribution is greater than 3, then it is called heavy-tailed or leptokurtoic.



(a) Skewness = 0, kurtosis = 3

(b) Skewness = 0, kurtosis = 20

(c) Skewness = −0.1, kurtosis = 5

(d) Skewness = 0.6, kurtosis = 5

Figure: Four Distributions with Different Skewness and Kurtosis and the Same Mean 0 and Variance 1

Multivariate Random Variables

Bivariate Random Variables

- A pair of bivariate r.v.'s (X, Y) is a function from the sample space into $\mathbb{R}^2$.
- The joint cdf of (X, Y) is

$$F(x,y) = P(X \leq x, Y \leq y).$$

- If F is continuous, the joint pdf is

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y).$$

- For any set $A \subset \mathbb{R}^2$,

$$P((X, Y) \in A) = \int\int_A f(x,y)dxdy.$$

- The discrete case can be parallely discussed. [Exercise]

- For any function $g(x,y)$,

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy.$$

Marginal Distributions

- The marginal distribution of X is

$$F_X(x) = P(X \leq x) = \lim_{y \to \infty} F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x,y)dydx,$$

- So by the fundamental theorem of calculus, the marginal density of X is

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x,y)dy.$$

- Similarly, the marginal density of Y is

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

Independence of Two R.V.s

- The r.v.'s X and Y are defined to be independent if $f(x,y) = f_X(x) f_Y(y)$.
- If X and Y are independent, then

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy \\ &= E[g(X)]E[h(Y)]. \end{aligned} \tag{2}$$

- For example, if X and Y are independent then

$$E[XY] = E[X]E[Y].$$

Covariance and Correlation

- The covariance between X and Y is
$$Cov(X,Y)=\sigma_{XY}=E\left[(X-E[X])(Y-E[Y])\right]=E[XY]-E[X]E[Y],$$
whose unit is the unit of X times the unit of Y.
- The correlation between X and Y is
$$Corr(X,Y)=\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$
- The Cauchy-Schwarz Inequality implies that
$$|\rho_{XY}|\le 1.$$
- The correlation is a measure of **linear** dependence, free of units of measurement. [Figure here]
- If X and Y are independent, then $\sigma_{XY}=0$ and $\rho_{XY}=0$. The reverse, however, is not true. For example, if $E[X]=0$ and $E\left[X^3\right]=0$, then $Cov(X,X^2)=0$. [Figure here]
- $Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$.
  - If X and Y are independent, then $Var(X+Y)=Var(X)+Var(Y)$ - the variance of the sum is the sum of the variances.
- $Var(X)=Cov(X,X)=E\left[(X-\mu_X)^2\right]=E\left[X^2\right]-\mu_X^2$ is the covariance of X with itself.

History of the Correlation Coefficient

Karl Pearson (1857-1936), UCL

- Karl Pearson (1857-1936) is the inventor of the correlation coefficient, so the correlation coefficient is also called the Pearson correlation coefficient.

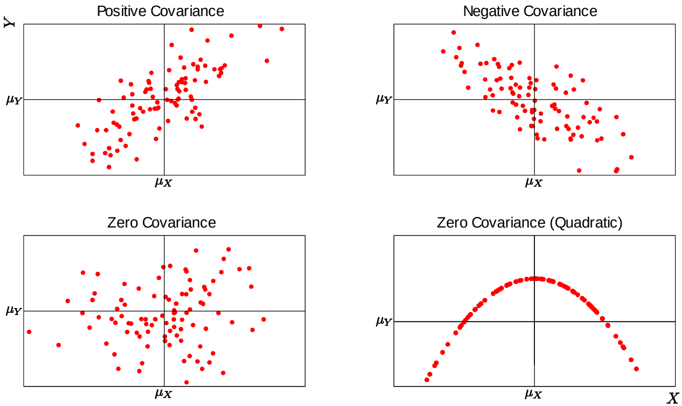


Figure: Positive, Negative an Zero Covariance

Conditional Distributions and Expectation

Conditional Distributions and Expectation

- The conditional density of Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

  if $f_X(x) > 0$.

- The conditional mean or conditional expectation is the function

$$m(x) = E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy,$$

  which is the mean of Y for the (slice of) individuals with $X = x$.

- The conditional mean $m(x)$ is a function, meaning that when X equals x, then the expected value of Y is $m(x)$.

- The conditional variance of Y given $X = x$ is

$$\begin{aligned}\sigma^2(x) &= Var(Y|X = x) = E\left[(Y - m(x))^2 \middle| X = x\right] \\ &= E\left[Y^2 \middle| X = x\right] - m(x)^2.\end{aligned}$$

- If Y and X are independent, then $E[Y|X = x] = E[Y]$ and $Var(Y|X = x) = Var(Y)$. (why?)

The Normal and Related Distributions

The Normal Distribution

- If X follows the normal distribution, then it has the density

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

- The mean and variance of the distribution are μ and σ^2, and it is conventional to write $X \sim N(\mu, \sigma^2)$.
  - The distribution of X is determined **only** by its mean and variance.
- The normal distribution is often used to model human heights and weights, test scores, and county unemployment rates.
- In some cases, normality can be achieved through transformations (e.g. log(wage) would follow a normal distribution).
- If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ follows $N(0,1)$.
- If X and Y are jointly normally distributed, then they are independent iff $Cov(X, Y) = 0$.
- Any linear combination of independent normal random variables has a normal distribution.

History of the χ^2 Distribution

Friedrich R. Helmert (1843-1917), University of Berlin

- The χ^2 distribution was first described by Friedrich Robert Helmert in papers of 1875-6, and was independently rediscovered by Karl Pearson in 1900.

χ^2 Distribution

- If $Z_1, \cdots, Z_r$ are independently distributed random variables such that $Z_i \sim N(0,1)$, $i = 1, \cdots, r$, then
$$X = \sum_{i=1}^{r} Z_i^2$$
follows the χ^2 (chi-square) distribution with r degrees of freedom (df), denoted as $X \sim \chi_r^2$.
- $E[X] = rE\left[Z_i^2\right] = r$ and $Var(X) = rVar\left(Z_i^2\right) = 2r$.
  - $Var\left(Z_i^2\right) = E\left[Z_i^4\right] - E\left[Z_i^2\right]^2 = 3 - 1 = 2$.

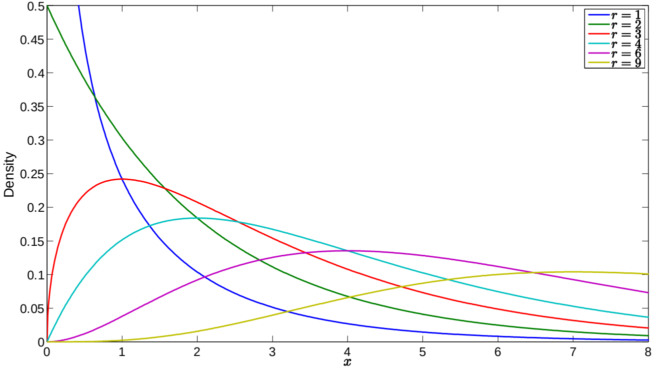


Figure: Chi-Square Density for $r = 1, 2, 3, 4, 6, 9$

History of the t-Distribution

William S. Gosset (1876-1937)

- The t-distribution is named after Gosset (1908), “The probable error of a mean”. At the time, Gosset worked at Guiness Brewery, which prohibited its employees from publishing in order to prevent the possible loss of trade secrets. To circumvent this barrier, Gosset published under the pseudonym “Student”. Consequently, this famous distribution is known as the student's t rather than Gosset's t! The name “t” was popularized by R.A. Fisher (we will discuss him later).

t-Distribution

- If Z is a standard normal variable,

$$Z \sim N(0,1),$$

and variable X has a χ^2 (chi-square) distribution with r degrees of freedom (df),

$$X \sim \chi_r^2,$$

independent of Z, then

$$\frac{Z}{\sqrt{X/r}} = \frac{\text{standard normal variable}}{\sqrt{\text{independent chi-square variable}/df}} \sim t_r,$$

a t-distribution with r degrees of freedom.

- $E[X] = 0$ if $r > 1$ and $Var(X) = \frac{r}{r-2}$ if $r > 2$.
- As $r \to \infty$, $t_r \to N(0,1)$.

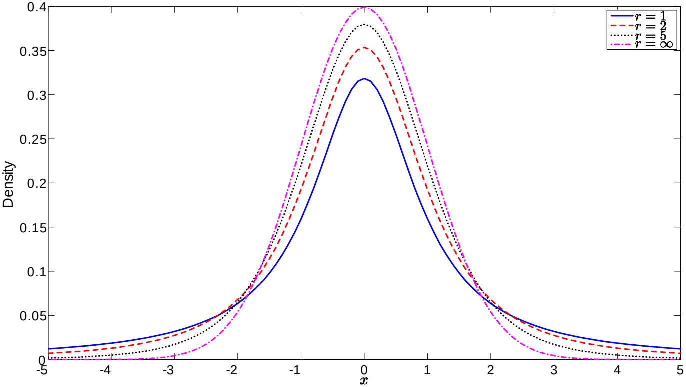


Figure: Density of the t Distribution for $r = 1, 2, 5, \infty$

History of the F-Distribution

Ronald A. Fisher (1890-1962), UCL

- Ronald A. Fisher (1890-1962) is one iconic founder of modern statistical theory. The name of F-distribution was coined by G.W. Snedecor, in honor of R.A. Fisher. The p-value discussed in the next chapter is also credited to him.

F-Distribution

- If X_1 follows a χ^2 distribution with q degrees of freedom,

$$X_1 \sim \chi^2_q,$$

and X_2 follows a χ^2 distribution with r degrees of freedom,

$$X_2 \sim \chi^2_r,$$

independent of X_1, then

$$\frac{X_1/q}{X_2/r} = \frac{\text{chi-square variable}/df}{\text{independent chi-square variable}/df} \sim F_{q,r},$$

an F-distribution with degrees of freedom q and r.

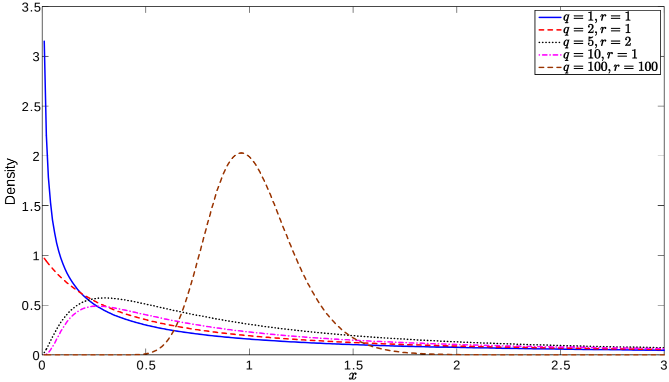


Figure: Density of the F Distribution for $(q, r) = (1, 1), (2, 1), (5, 2), (10, 1), (100, 100)$