Week 1

Shengtong Han

Probability

Common notations for sets and elements

`iff` means `if and only if`. For sets: $A, B, C, \cdots$ and the elements $x, y, z, \cdots$,

- $x \in A$ iff x is a member of A.
- $\emptyset$ is the null/empty set without any elements
- A $\subset$ B iff $x \in A$ implies $y \in B$
- A $\cup$ B =$\{$x: x $\in$ A or x $\in$ B$\}$: the union of A and B.
- $AB = \{x : x \in A \; and \; x \in B\}$: the intersection of A and B
- $A \setminus B = \{x : x \in A, \; but \; x \notin B\}$: difference set A less B

What is the probability of **an event** `red three, yellow five`? An event is defined as a set of elementary events.

Axioms for probabilities

Sometimes an event can be a future occurrence, whose probability depends on current knowledge. Formally, suppose there are a pair of events E, H. One is interested in $P(E|H)$: the probability of event E given the hypothesis H. The following axioms hold

- [P1] $P(E|H) \geq 0$ for all E, H
- [P2] $P(H|H) = 1$ for all H
- [P3] $P(E \cup F|H) = P(E|H) + P(F|H)$ when $EFH = \emptyset$
- [P4] $P(E|FH)P(F|H) = P(EF|H)$ $P4$ can be re-written as

$$P(E|FH) = \frac{P(EF|H)}{P(F|H)} when P(F|H) \neq 0$$

The twins are classified according as they had a criminal conviction (C) or not (N) and according as they were monozygotic (M) or dizygotic (D).

\	C	N	Total
M	10	3	13
D	2	15	17
Total	12	18	30

Denote by H the knowledge that an individual has been picked randomly from the population. Then $P(C|H) = \frac{12}{30}$ $P(MC|H) = \frac{10}{30}$ $P(M|CH) = \frac{10}{12}$ (Sample space has changed) Hence $P(M|CH)P(C|H) = P(MC|H)$.

Reference: Fisher, R.A. Statistical methods for research workers, Edinburg: Oliver & Boyd (1925b)

Unconditional probability

Tossing a coin

- Strictly speaking, there is no unconditional probability
- Most statements are made conditional on individual's knowledge or experience
- The probability of `head` is approximately $\frac{1}{2}$ after many tries.

$$P(E) = P(E|\Omega)$$

$$P(E|F) = P(E|F\Omega)$$

where Ω is the whole sample space, consisting of all possible events.

The following axioms hold

- $0 \leq P(E) \leq 1\backslash$
- $P(\Omega) = 1, P(\emptyset) = 0$ \
- $P(\cup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n)$

where E_n are exclusive events.

Odds

Define the odds on E against F given H as the ratio

$$P(E|H)/P(F|H) \; to \; 1$$

or equivalently

$$\frac{P(E|H)}{P(F|H)}$$

with no mention of H. Odds will be used in Bayesian hypothesis testing.

Independence

Two events are said to be `independent` given H if

$$P(EF|H) = P(E|H)P(F|H)$$

It follows immediately from axiom P4 that

$$P(E|FH) = P(E|H)$$

idea: adding extra information of F does not alter the probability of E that given H alone

A sequence of events E_n is said to be `pairwise independent` given H if

$$P(E_m E_n|H) = P(E_m|H)P(E_n|H)$$

for $m \neq n$ and is said to be `mutually independent` given H if for **every finite subset**

$$P(E_{n1}E_{n2}\cdots E_{nk}|H) = P(E_{n1}|H)P(E_{n2}|H)\cdots P(E_{nk}|H)$$

Warning: mutually independent doesn't imply pairwise independent

Bayes' theorem

Derived results of axioms

From P2, P4 and $HH = H$,

$$P(E|H) = P(EH|H)$$

In particular

$$P(E) = P(E\Omega)$$

If given H, E implies F, that is $EH \subset F$ and so $EFH = EH$. $P(E|FH)P(F|H) = P(EF|H)$(*by* P4) = $P(EFH|H) = P(EH|H)$(*because* $EFH = EH$) $= P(E|H)$

Thus $P(E|H) \leq P(F|H)$ if given H, E implies F.

In particular, replace H with Ω, if E implies F then

$$P(E|F)P(F) = P(E)$$
$$P(E) \leq P(F)$$

Let H_n be a sequence of exclusive and exhaustive events and E be any event.

$$P(E) = \sum_n P(E|H_n)P(H_n)$$

generalized addition law

By P4

$$P(H_n|E)P(E) = P(EH_n) = P(H_n)P(E|H_n)$$

`Thus`

$P(H_n|E) = \frac{P(H_n)P(E|H_n)}{P(E)} \propto P(H_n)P(E|H_n)$(*provided* $P(E) \neq 0$)

Let H_n be a sequence of exclusive and exhaustive events, then

$$P(H_n|E) = \frac{P(H_n)P(E|H_n)}{\sum_m P(H_m)P(E|H_m)}$$

`By repeatedly applying P4,`

$$P(H_1H_2\cdots H_n) = P(H_1)P(H_2|H_1)\cdots P(H_n|H_1H_2\cdots H_{n-1})$$

where $H_1, H_2, \cdots, H_n$ are any events.

An Example: The Biology of twins

Background: Twins can be either Monozygotic(M) (from the same egg) or dizygotic (D). Monozygotic twins are identical and thus look very similar and more important are of the same sex. While dizygotic twins can be of opposite sex, but assuming are equally probable. Denote by the sexes of a pair of twins GG, BB, GB (GB is indistinguishable from BG).

$$P(GG|M) = P(BB|M) = \frac{1}{2}, P(GB|M) = 0$$

$$P(GG|D) = P(BB|M) = \frac{1}{4}, P(GB|D) = \frac{1}{2}$$

`By Bayes's theorem`

$P(GG) = P(GG|M)P(M) + P(GG|D)P(D) = \frac{1}{2}P(M) + \frac{1}{4}(1 - P(M))$

thus

$$P(M) = 4P(GG) - 1$$

The sex distribution of all twins can be used to estimate the proportion of monozygotic twins in the whole population.

Random variables

Discrete random variables

Let Ω be a set of all elementary events, suppose with each elementary event $\omega \in \Omega$, there is a function $\tilde{m}$ mapping Ω to the set of all integers. This function is said to be a **random variable** or an r.v.

For example, in the experiment of tossing a red die and a blue die, ω could be a event of `red three, yellow two` and $\tilde{m}$ could be 5.

There are two ways to describe the distribution of a random variable

- Probability density function (PDF): $p(m) = P\{\omega, \tilde{m}(\omega) = m\}$

the probability of random variable $\tilde{m}$ taking value of m.

- Cumulative distribution function (CDF), defined by $F(m) = P(\tilde{m} \leq m) = \sum_{k \leq m} p(k)$

Because PDF has properties $p(m) \geq 0, \sum_m p(m) = 1$ leading to

$F(m) \leq F(m')$ if $m \leq m'$,
$lim_{m \to -\infty} F(m) = 0, lim_{m \to \infty} F(m) = 1$ #### Binomial distribution In a sequence of n independent trials, each of which results in a success with probability π or failure $(1 - \pi)$, the number of successes X is said to be following a Binomial distribution with parameter π, denoted as $X \sim B(n, \pi)$

and $p(k) = \binom{n}{k}\pi^k(1 - \pi)^{n-k}$, $0 \leq k \leq n$

Show that: if X, Y are independent and $X \sim B(m, \pi), Y \sim B(n, \pi)$, then $X + Y \sim B(m + n, \pi)$.

Continuous random variables

In contrast to discrete random variable, continuous random variable takes a real number $\tilde{x}(\omega)$ for each elementary event $\omega \in \Omega$. Its cumulative distribution function is

$F(x) = P(\tilde{x} \leq x) = P(\{\omega : \tilde{x}(\omega) \leq x\})$ Similarly,

$F(x) \leq F(x')$ if $x \leq x'$, $lim_{x \to -\infty} F(x) = 0, lim_{x \to \infty} F(x) = 1$

For continuous random variable, usually there is a function $p(x)$, such that

$F(x) = \int_{-\infty}^{x} p(\delta)d\delta$

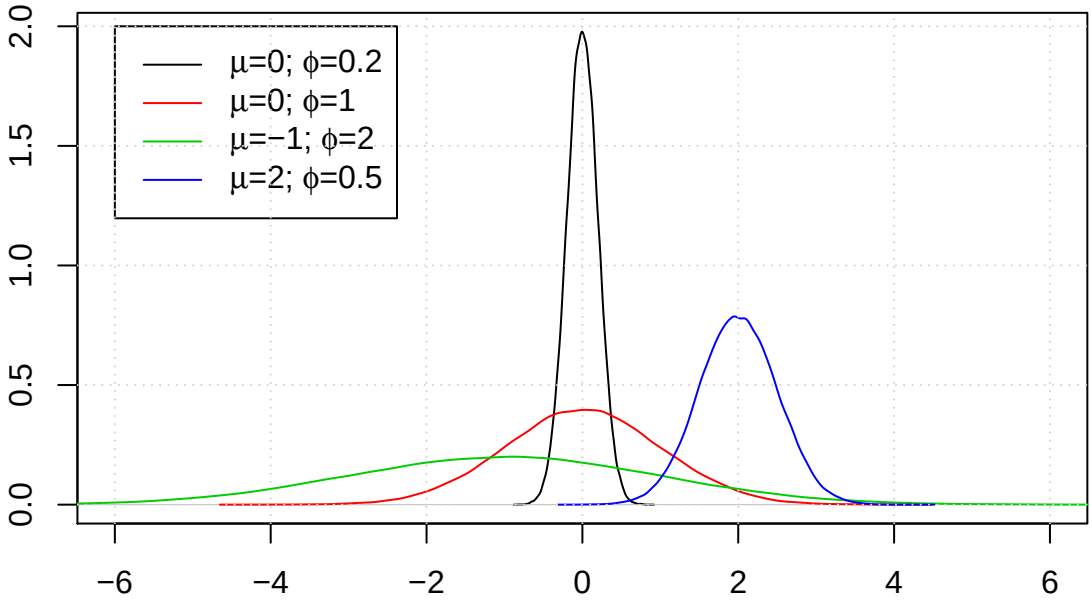
where $p(x)$ is called probability density function (PDF). It is hard to interpret $p(x)$, and for small δx,

$p(x)\delta x \cong P(x < \tilde{x} \leq x + \delta x) = P(\{\omega : x < \tilde{x}(\omega) \leq x + \delta x\})$

The normal distribution

The most important continuous distribution is normal distribution or Gaussian distribution. The standard normal distribution has PDF, i.e. $z \sim N(0,1)$ $p(z) = (2\pi)^{-\frac{1}{2}} exp(-\frac{z^2}{2})$ For standard normal distribution

- 68% area between -1 and 1
- 95 % area between -2 and 2
- 99.7% area between -3 and 3



More generally, if x follows a normal distribution with mean μ and variance ϕ, i.e. $x \sim N(\mu, \phi)$, the PDF is

$p(x) = (2\pi\phi)^{-\frac{1}{2}} exp\{-\frac{(x-\mu)^2}{2\phi}\}$

Normal distribution is important mainly because of the **Central Limit Theorem** which says if a random variable can be expressed as a sum of large number of independent components, no one of which dominates others, then this sum will be approximately normally distributed.

Two discrete random variables

The sequence $p(m,n)$ is said to be a bivariate density function or bivariate PDF or joint PDF of random variables m, n if $p(m,n) = P(\{\omega : \tilde{m}(\omega) = m, \tilde{n}(\omega) = n\})$ Clearly, $p(m,n) \geq 0$, $\sum_m \sum_n p(m,n) = 1$

- Joint distribution function of (m,n) is $F(m,n) = \sum_{k \leq m} \sum_{l \leq n} p(k,l)$
- The marginal density of m is $p(m) = \sum_n p(m,n)$
- The conditional density of n given m is

$p(n|m) = P(\tilde{n} = n | \tilde{m} = m) = \frac{p(m,n)}{p(m)}$ if $p(m) \neq 0$

- Conditional distribution function

$F(n|m) = P(\tilde{n} \leq n | \tilde{m} = m) = \sum_{k \leq n} p(k|m)$

Two continuous random variables

The joint distribution function of two continuous random variables is defined as

$F(x,y) = P(\tilde{x} \leq x, \tilde{y} \leq y)$

Marginal distribution functions are

$F(x, +\infty)$, $F(+\infty, y)$

If $p(x, y)$ is the joint density function, then

$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p(\eta, \delta) d\eta d\delta$

If we know joint distribution function $F(x, y)$, then

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

For joint density function $p(x, y)$, it holds

$p(x, y) \geq 0, \int \int p(x, y) dx dy = 1$

The marginal density is

$p(x) = \int p(x, y) dy$

The conditional density is, when $p(x) \neq 0$

$p(y|x) = \frac{p(x,y)}{p(x)}$

and the conditional distribution function is

$F(y|x) = \int_{-\infty}^{y} p(\eta|x) d\eta$

Bayes' theorem for random variables

It holds for the conditional density, $p(y|x)$

$p(y|x) \geq 0$, $\int p(y|x) dy = 1$ $p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(y)p(x|y)}{p(x)}$

So

$p(y|x) \propto p(y)p(x|y)$ Bayes' formula

The constant of proportionality is * $\frac{1}{p(x)} = \frac{1}{\int p(y)p(x|y)dy}$, for continuous r.v. * $\frac{1}{p(x)} = \frac{1}{\sum_y p(y)p(x|y)}$ for discrete r.v.

An example

In some cases we may have one continuous r.v and another is discrete r.v. All definitions and formulae still apply.

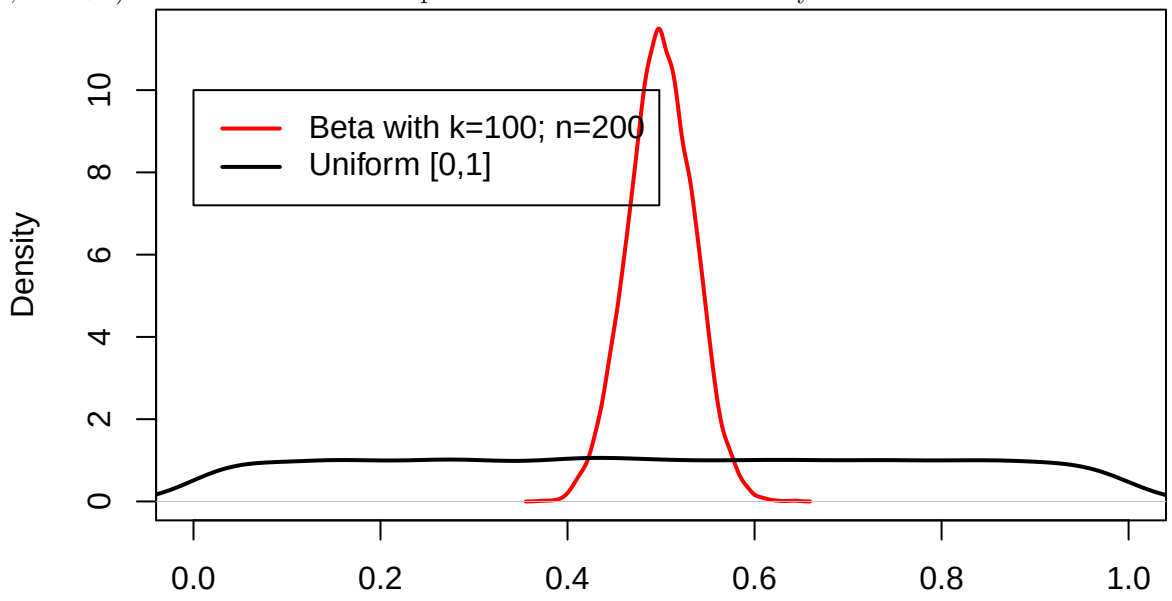
Suppose k is the number of successes in n Bernoulli trials, i.e. $k \sim B(n, \pi)$. But π is unknown, and assumed to be uniformly distributed within interval $[0, 1]$.

$p(k|\pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$, $k = 0, 1, \cdots, n \setminus p(\pi) = 1$, $0 \leq \pi \leq 1$

so

$p(\pi|k) \propto p(\pi)p(k|\pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} \propto \pi^k (1 - \pi)^{n-k}$

Thus $\pi|k \sim Be(k+1, n-k+1)$. This beta distribution represents the beliefs about π after you have observed k successes in n trials.



Two random variables are said to be $\{independent\}$ *if* $p(x, y) = p(x)p(y)$

for all values of x *and* y*. It applies to both continuous and discrete cases. Basic idea:* ***Knowing*** x ***doesn't affect the distribution of*** y*.*

Mean and variances

Expectation

For discrete random variable m*, if* $\sum |m|p(m) < \infty$*, the* ***mean or expectation*** *of a discrete random variable is defined as*

$E(m) = \sum mp(m)$

It can be viewed as a long term average. m *can be generalized to a function of* $g(m)$*, and its expectation is* $E(g(m)) = \sum g(m)p(m)$*. In a similar way, for continuous random variable* x*,* $E(x) = \int xp(x)dx$*,* $E(g(x)) = \int g(x)p(x)dx$

Properties

For any two continuous random variables x, y

$E(ax + by + c) = \int\int (ax + by + c)p(x, y)dxdy = a\int\int xp(x, y)dxdy + b\int\int yp(x, y)dxdy + c\int\int p(x, y)dxdy = a\int xp(x)dx + b\int yp(y)dy + c \;\; (by\ marginal\ PDF) = aE(x) + bE(y) + c$

`More generally`

$E[ag(x) + bh(y) + c] = aE(g(x)) + bE(h(y)) + c$

If random variables x, y *are independent* $E(xy) = \int\int xyp(x, y)dxdy = \int\int xyp(x)p(y)dxdy \;\; (by\ independence) = (\int xp(x)dx)(\int yp(y)dy) = (Ex)(Ey)\;\; more\ generally Eg(x)h(y) = (Eg(x))(Eh(y))$

Variance, precision and standard deviation

`To measure how spread out a distribution is, variance is introduced by`

$var(x) = E(x - Ex)^2$

- *when the distribution is little spread out,* $(x - Ex)^2$ *is small with high probability and thus* $var(x)$ *is small*
- *when the distribution is very spread out,* $var(x)$ *is large*

- *the reciprocal of variance is called* ***precision***
- *the positive square root is called* ***standard deviation*** $var(x) = E(x - Ex)^2 = Ex^2 - (Ex)^2$

Examples

- *Suppose* $k \sim B(n, \pi)$*, then* $Ek = \sum_{k=0}^{n} k\binom{n}{k}\pi^k(1-\pi)^{n-k} = n\pi$*,* $Ek(k-1) = n(n-1)\pi^2$*, and* $var(k) = n\pi(1-\pi)$*.(proof that)*
- *If* $x \sim N(\mu, \phi)$*, then* $Ex = \mu, var(x) = \phi$ *(proof that)*

covariance and correlation

Define covariance of x, y *as* $cov(x, y) = E(x - Ex)(y - Ey) = Exy - (Ex)(Ey)$ *note that*

$$\begin{aligned} var(x+y) &= E[x + y - E(x+y)]^2 \\ &= E[(x - Ex) + (y - Ey)]^2 \\ &= E(x - Ex)^2 + E(y - Ey)^2 + 2E(x - Ex)(y - Ey) \\ &= var(x) + var(y) + 2cov(x, y). \end{aligned}$$

More generally, $var(ax + by + c) = a^2var(x) + b^2var(y) + 2abcov(x, y) \geq 0$ *for any constants* a, b, c*. Treat this expression as a quadratic of* a *and it can not have two unequal real roots because it is always nonnegative. So* $cov(x, y) \leq (var(x))(var(y))$*. Define the correlation coefficient* $\rho(x, y)$ *as*

$$\rho(x, y) = \frac{cov(x, y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

Clearly $-1 \leq \rho(x, y) \leq 1$*.*

- $\rho(x, y) = 1$ *if and only if* $ax + by + c = 0$ *with probability 1 for constants* a, b, c *and* a, b *have opposite signs*
- $\rho(x, y) = 1$ *if and only if* $ax + by + c = 0$ *with probability 1 for constants* a, b, c *and* a, b *have same signs*
- *If* x, y *are independent* $cov(x, y) = E(xy) - (Ex)(Ey) = 0$*. But the converse is not true, that is* $\rho(x, y) = 0 \nRightarrow x, y$ *are independent.*

Approximations

In some cases, it is useful to use approximations to the mean and variance of a function of a random variable. Suppose $z = g(x)$*, where* g *is a smooth function and* x *is not too far from its expectation. By Taylor's theorem,*

$$z \cong g(Ex) + (x - Ex)g'(Ex)$$

- *taking expectations in both sides, a fair approximation to the expectation of* z *is* $Ez = g(Ex)$*.*

- *taking variance in both sides, a reasonable approximation to the variance of* z *is* $var(z) = var(x)[g'(Ex)]^2$*.*

An example

Suppose $x \sim B(n, \pi)$, $E(x) = n\pi$. *Let* $z = g(x)$, *where* $g(x) = sin^{-1}\sqrt{x/\pi}$. $g'(x) = \frac{1}{2n\sqrt{\frac{x}{n}(1-\frac{x}{n})}}$. *Therefore*

$E(z) \approxeq sin^{-1}\sqrt{\pi}, var(z) \approxeq \frac{1}{4n}$

Conditional expectations

Define conditional expectation of y *given* x *by*

$E(y|x) = \int yp(y|x)dy$.

More generally

$E(g(y)|x) \int g(y)p(y|x)dy$.

Define a conditional variance as

$var(y|x) = E[(y - E(y|x))^2|x] = E(y^2|x) - [E(y|x)]^2$.

Medians and modes

Define ***median*** *as any value* x_0 *such that*

$P(x \leq x_0) \leq \frac{1}{2}$ *and* $P(x \geq x_0) \geq \frac{1}{2}$

In continuous case, there is usually a unique median such that

$P(x \geq x_0) = P(x \leq x_0) = \frac{1}{2}$

Mode is defined as a value at which the PDF achieves the maximum.
An empirical relation among them is $mean - mode = 3(mean - median)$