Lecture 3: Probability

Kushal K Dey

Toss a coin

```
TossCoin <- function(n=1){
    return(replicate(n, sample(c("H","T"),1)))
}
TossCoin(1)

[1] "T"

TossCoin(n=2)

[1] "H" "T"

TossCoin(n=10)

 [1] "H" "T" "H" "T" "T" "T" "H" "T" "H" "T"
```

Experimental States

All possible states of an experiment.

1 toss of coin: $\{H, T\}$

Experimental States

All possible states of an experiment.

1 toss of coin: $\{H, T\}$

2 tosses : $\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}$

Experimental States

All possible states of an experiment.

1 toss of coin: $\{H, T\}$

2 tosses : $\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}$

10 tosses: $\{x_1 x_2 \cdots x_n : x_i \in \{H, T\}\}$

Random Variable

A function that associates a unique numerical value with every outcome of an experiment.

Random Variable

A function that associates a unique numerical value with every outcome of an experiment.

For example, associate 1 with H and 0 with T in 1 coin toss experiment.

Random Variable

A function that associates a unique numerical value with every outcome of an experiment.

For example, associate 1 with H and 0 with T in 1 coin toss experiment.

The value of the random variable will vary from trial to trial as the experiment is repeated.

Random Variable

A function that associates a unique numerical value with every outcome of an experiment.

For example, associate 1 with H and 0 with T in 1 coin toss experiment.

The value of the random variable will vary from trial to trial as the experiment is repeated.

Ex: A coin is tossed 2 times. The random variable X is the number of tails that are noted.

state	X
HH	0
HT	1
TH	1
TT	2

Events

Are all the states of an experiment equally likely?

Events

Are all the states of an experiment equally likely?

What if a coin is biased to come up heads 80% times?

Events

Are all the states of an experiment equally likely?

What if a coin is biased to come up heads 80% times?

Events: Sets of outcomes of the experiment

Events

Are all the states of an experiment equally likely?

What if a coin is biased to come up heads 80% times?

Events: Sets of outcomes of the experiment

$X = 1$ is an event, comprises of two outcomes $\{HT, TH\}$

Events

Are all the states of an experiment equally likely?

What if a coin is biased to come up heads 80% times?

Events: Sets of outcomes of the experiment

$X = 1$ is an event, comprises of two outcomes $\{HT, TH\}$]

$X = 2$ is also an event, comprises of $\{HH\}$.

Events

Are all the states of an experiment equally likely?

What if a coin is biased to come up heads 80% times?

Events: Sets of outcomes of the experiment

$X = 1$ is an event, comprises of two outcomes $\{HT, TH\}$]

$X = 2$ is also an event, comprises of $\{HH\}$.

$X = 0$ is also an event, comprises of $\{TT\}$.

Events

Take another example. Let X be a random variable that takes 1 if there is at least one head.

Events

Take another example. Let X be a random variable that takes 1 if there is at least one head.

state	X
HH	1
HT	1
TH	1
TT	0

Then $X = 1$ is an event, comprises of two outcomes $\{HH, HT, TH\}$

Events

Take another example. Let X be a random variable that takes 1 if there is at least one head.

state	X
HH	1
HT	1
TH	1
TT	0

Then $X = 1$ is an event, comprises of two outcomes $\{HH, HT, TH\}$

$X = 0$ is also an event, comprises of $\{TT\}$.

Probability

Take any experiment (say 2 independent tosses of a coin).

Probability

Take any experiment (say 2 independent tosses of a coin).

In one experiment, there may be many states (4 in this case- HH, HT, TH and TT) but remember, you observe only one.

Probability

Take any experiment (say 2 independent tosses of a coin).

In one experiment, there may be many states (4 in this case- HH, HT, TH and TT) but remember, you observe only one.

Repeat the experiment many many times (as many as you can imagine- may be 10^7).

Probability

Take any experiment (say 2 independent tosses of a coin).

In one experiment, there may be many states (4 in this case- HH, HT, TH and TT) but remember, you observe only one.

Repeat the experiment many many times (as many as you can imagine- may be 10^7).

The proportion of times you observe each experimental state after all these many repititions gives you the probability of that state.

Example of Probability

We cite an example of how to measure probability for one coin toss with states H and T.

`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

Example of Probability

We cite an example of how to measure probability for one coin toss with states H and T.

http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html

Cases: Coin fair and unfair!!

Example of Probability

We cite an example of how to measure probability for one coin toss with states H and T.
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

Cases: Coin fair and unfair!!

Question : What is the *probability* of Heads in a coin toss?

Example of Probability

We cite an example of how to measure probability for one coin toss with states H and T.
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

Cases: Coin fair and unfair!!

Question : What is the *probability* of Heads in a coin toss?

Reframe the question: What is the proportion of times you would expect to see Heads if you tossed the coin many many times?

Example of Probability

We cite an example of how to measure probability for one coin toss with states H and T.
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

Cases: Coin fair and unfair!!

Question : What is the *probability* of Heads in a coin toss?

Reframe the question: What is the proportion of times you would expect to see Heads if you tossed the coin many many times?

Probability Table

For a fair toss of coin, we saw

state	Prob
H	$\Pr(\text{H}) = 0.5$
T	$\Pr(\text{T}) = 0.5$

Probability Table

For a fair toss of coin, we saw

state	Prob
H	$\Pr(\mathrm{H}) = 0.5$
T	$\Pr(\mathrm{T}) = 0.5$

Let X be 1 if state is H and 0 if T.

Probability Table

For a fair toss of coin, we saw

state	Prob
H	$\Pr(\text{H}) = 0.5$
T	$\Pr(\text{T}) = 0.5$

Let X be 1 if state is H and 0 if T.

X	Prob
1	$\Pr(\text{X}{=}1) := \Pr(\text{H}) = 0.5$
0	$\Pr(\text{X}{=}0) := \Pr(\text{T}) = 0.5$

Probability of an Event

Probability of an event is the sum of probability of all states making up the event.

Probability of an Event

Probability of an event is the sum of probability of all states making up the event.

For two independent tosses of a fair coin, the probability table on states

Probability of an Event

Probability of an event is the sum of probability of all states making up the event.

For two independent tosses of a fair coin, the probability table on states

state	Prob
HH	Pr(HH) = Pr(H in toss 1) * Pr(H in toss 2) = 0.25
HT	Pr(HT) = Pr(H in toss 1) * Pr(T in toss 2) = 0.25
TH	Pr(TH) = Pr(T in toss 1) * Pr(H in toss 2) = 0.25
TT	Pr(HT) = Pr(T in toss 1) * Pr(T in toss 2) = 0.25

Probability of an Event

Probability of an event is the sum of probability of all states making up the event.

For two independent tosses of a fair coin, the probability table on states

state	Prob
HH	Pr(HH) = Pr(H in toss 1) * Pr(H in toss 2) = 0.25
HT	Pr(HT) = Pr(H in toss 1) * Pr(T in toss 2) = 0.25
TH	Pr(TH) = Pr(T in toss 1) * Pr(H in toss 2) = 0.25
TT	Pr(HT) = Pr(T in toss 1) * Pr(T in toss 2) = 0.25

Probability of an Event

Let X be 1 if we observe at least one H in two coin tosses.

Probability of an Event

Let X be 1 if we observe at least one H in two coin tosses.

Then the probability table for the random variable X.

X	states	Prob
0	TT	Pr(X=0) = Pr(TT) = 0.25
1	{ HH, HT, TH }	Pr(X=1) = Pr(HH) + Pr(HT) + Pr(TH) = 0.7

Note that the sum of the probabilities for any probability table would be 1 (sum across all possible states).

Probability of an Event

Let X be number of Ts observed in two tosses of a coin

Probability of an Event

Let X be number of Ts observed in two tosses of a coin

Then the probability table for the random variable X.

X	states	Prob
0	HH	Pr(X=0) = Pr(HH) = 0.25
1	{ HT, TH }	Pr(X=1) = Pr(HT) + Pr(TH) = 0.50
2	TT	Pr(X=2) = Pr(TT) = 0.25

Note that the sum of the probabilities for any probability table would be 1 (sum across all possible states).

Why do we need Random variables?

When we can define probability in terms of states and events, why do we need random variable X?

Why do we need Random variables?

When we can define probability in terms of states and events, why do we need random variable X?

It is easier to interpret the experiment through numbers and bunch of symbols.
What type of interpreting?

Why do we need Random variables?

When we can define probability in terms of states and events, why do we need random variable X?

It is easier to interpret the experiment through numbers and bunch of symbols.
What type of interpreting?

Measures of central tendency and spread

Probability distribution of random variable

Suppose X is a random variable that takes values $x_1, x_2, \cdots .x_n$ and let us define the probability

$$p(x_i) = Pr(X = x_i)$$

Then we have

$$0 \leq p(x_i) \leq 1$$

$$\sum_{i=1}^{n} p(x_i) = 1$$

Cumulative probability

$$F(x) = Pr(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

Expectation of random variable

If X is a discrete random variable with possible values

$x_1, x_2, x_3, ..., x_n$

$$E(X) := \sum_{i=1}^{n} x_i Pr(X = x_i)$$

For one coin toss, let X be 1 if H and 0 if T.

$$E(X) = 0.Pr(X = 0) + 1.Pr(X = 1) = 0.5$$

If X is number of Ts in 2 independent tosses of a coin

$$E(X) = 0.Pr(X = 0) + 1.Pr(X = 1) + 2.Pr(X = 2) = 0 + 0.5 + 2 * 0.25 = 1$$

If I am asked to choose ONE value of the variable X, this is the value we may want to choose.

Expectation and $\bar{x}$

We saw that if we observe values $x_1, x_2, \cdots, x_n$ with frequencies $f_1, f_2, \cdots, f_n$, then

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i f_i = \frac{1}{n}\sum_{i=1}^{n} x_i \left(\frac{f_i}{n}\right)$$

Now we can argue that as $n \to \infty$, $\frac{f_i}{n} \approx Pr(X = x_i)$, then we get the expectation

$$E(X) = \sum_{i=1}^{n} x_i Pr(X = x_i)$$

Law of Large Numbers

Seems like as n becomes larger, the sample average $\bar{X}_n$ becomes closer and closer to the population average $\mu = E(X)$.

Law of Large Numbers

Seems like as n becomes larger, the sample average $\bar{X}_n$ becomes closer and closer to the population average $\mu = E(X)$.

Can we mathematically formulate it.

$$Pr(\left|\bar{X}_n - \mu\right| > \delta) \to 0$$

for any $\delta > 0$.

Law of Large Numbers

Lets go back to the applet:
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

Law of Large Numbers

Lets go back to the applet:
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

By Law of Large Numbers (LLN), if you toss the coin many times, say 100, you should expect to get closely similar number of heads and tails.

Law of Large Numbers

Lets go back to the applet:
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

By Law of Large Numbers (LLN), if you toss the coin many times, say 100, you should expect to get closely similar number of heads and tails.

If you tossed it many more times than 100, say 10,000, you would expect to get even closer to a 1:1 ratio of heads and tails (closer and closer to 50% heads and 50% tails).

Law of Large Numbers

Lets go back to the applet:
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html`

By Law of Large Numbers (LLN), if you toss the coin many times, say 100, you should expect to get closely similar number of heads and tails.

If you tossed it many more times than 100, say 10,000, you would expect to get even closer to a 1:1 ratio of heads and tails (closer and closer to 50% heads and 50% tails).

The more times you repeat the experiment, the closer you should get to the true probability (which remember is proportion for infinite many trials).

Chebyshev Inequality

General Chebyshev Inequality: For any random variable X, for any fixed $\epsilon > 0$ and a function $g(X) \geq 0$,

$$Pr[g(X) \geq \epsilon] \leq \frac{E(g(X))}{\epsilon}$$

Chebyshev Inequality:

$$Pr[\left|\bar{X}_n - \mu\right| \geq \epsilon] \leq \frac{E(\bar{X}_n - \mu)^2}{\epsilon^2}$$

Variance

Variance of a random variable is a non-negative number which gives an idea of how widely spread the values of the random variable are likely to be; the larger the variance, the more scattered the observations on average.

Variance

Variance of a random variable is a non-negative number which gives an idea of how widely spread the values of the random variable are likely to be; the larger the variance, the more scattered the observations on average.

Define it as

$$
\begin{aligned}
Var(X) &:= \sum_{i=1}^{n} (x_i - E(X))^2 Pr(X = x_i) && (1) \\
&= E(X - E(X))^2 && (2) \\
&= E(X^2) - E(X)^2 && (3) \\
&= \sum_{i=1}^{n} x_i^2 Pr(X = x_i) - \left(\sum_{i=1}^{n} x_i Pr(X = x_i) \right)^2 && (4) \\
& && (5)
\end{aligned}
$$

Our first Probability Distribution

Let X be a random variable taking values 0 and 1 for T and H in a coin toss experiment.

Our first Probability Distribution

Let X be a random variable taking values 0 and 1 for T and H in a coin toss experiment.

$$
\begin{aligned}
X &= 1 \qquad \textit{prob} \;\; p & (6)\\
&= 0 \qquad \textit{prob} \;\; (1-p) & (7)\\
& & (8)
\end{aligned}
$$

where p is the probability of head (0.5 in fair coin toss case).

Our first Probability Distribution

Let X be a random variable taking values 0 and 1 for T and H in a coin toss experiment.

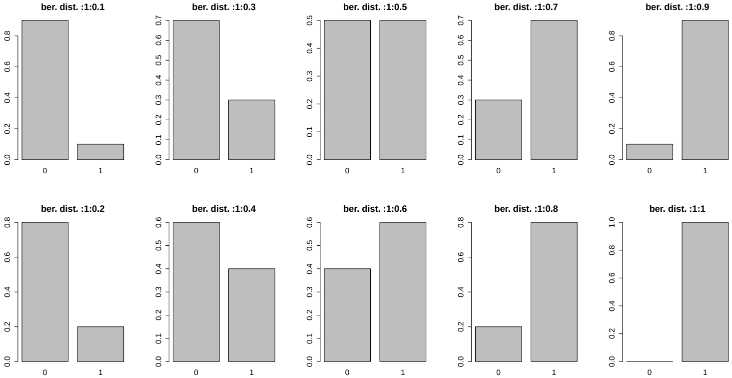
$$\begin{aligned} X &= 1 \qquad prob \;\; p & (6) \\ &= 0 \qquad prob \;\; (1-p) & (7) \\ & & (8) \end{aligned}$$

where p is the probability of head (0.5 in fair coin toss case).

This distribution is called a **Bernoulli distribution** and we term $X \sim Ber(p)$

$$E(X) = p \qquad var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$$

Bernoulli distribution graph

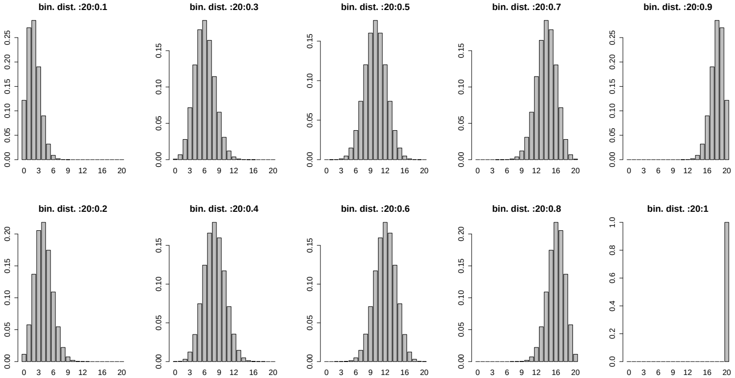


Binomial Distribution

Here X denotes number of heads observed in n tosses. Let probability of observing head in one toss be p

$$
\begin{aligned}
Pr(X = 0) &= (1-p)^n && (9)\\
Pr(X = 1) &= np(1-p)^{n-1} && (10)\\
Pr(X = 2) &= \binom{n}{k} p^2 (1-p)^{n-2} && (11)\\
&....... && (12)\\
Pr(X = x) &= \binom{n}{k} p^x (1-p)^{n-x} && (13)\\
&........ && (14)\\
Pr(X = n) &= p^n && (15)\\
& && (16)
\end{aligned}
$$

Binomial Distribution graph (n=20)



Moment generating function

The moment generating function (mgf) is given by

$$
\begin{aligned}
mgf(t) &:= E(e^{tX}) && (17)\\
&= E\left(\left\{\sum_{k=1}^{\infty}\frac{(tX)^k}{k!}\right\} Pr(X=k)\right) && (18)\\
&= \sum_{k=1}^{K}\frac{t^k}{k!}E\left(X^k\right) && (19)\\
&= \sum_{k=1}^{K}\frac{t^k}{k!}\mu_k && (20)
\end{aligned}
$$

μ_k is called the kth moment of the distribution of X.

Moment generating function

The moment generating function (mgf) is given by

$$
\begin{aligned}
mgf(t) &:= E(e^{tX}) && (17)\\
&= E\left(\left\{\sum_{k=1}^{\infty}\frac{(tX)^k}{k!}\right\}Pr(X=k)\right) && (18)\\
&= \sum_{k=1}^{K}\frac{t^k}{k!}E\left(X^k\right) && (19)\\
&= \sum_{k=1}^{K}\frac{t^k}{k!}\mu_k && (20)
\end{aligned}
$$

μ_k is called the kth moment of the distribution of X.

Why do we care about $mgf(t)$?

The problem

Where is the connection to the last 4 classes?

How to connect to previous classes

Suppose I want to find out what proportion of Males/Females study in UChicago.

How to connect to previous classes

Suppose I want to find out what proportion of Males/Females study in UChicago.

I stand at the center of the Quad for a day and mark the gender of each person passing by.

How to connect to previous classes

Suppose I want to find out what proportion of Males/Females study in UChicago.

I stand at the center of the Quad for a day and mark the gender of each person passing by.

Each record I take of a person is one run of the experiment. The states of the space for one run is $\{M, F\}$.

How to connect to previous classes

Say, if I observe M, I am noting down 1 and if F, I am noting down 0.

How to connect to previous classes

Say, if I observe M, I am noting down 1 and if F, I am noting down 0.

For the first record I note down, define

$$
\begin{aligned}
X &= 1 \;\; if \;\; M && (21)\\
&= 0 \;\; if \;\; F && (22)\\
& && (23)
\end{aligned}
$$

X is a random variable.

How to connect to previous classes

Say, if I observe M, I am noting down 1 and if F, I am noting down 0.

For the first record I note down, define

$$\begin{aligned} X &= 1 \ \ if \ \ M && (21) \\ &= 0 \ \ if \ \ F && (22) \\ & && (23) \end{aligned}$$

X is a random variable.

Assume the actual proportion of males in UChicago is p. We do not know p (otherwise we would not have sampled). Then

$$X \sim Ber(p)$$

How to connect to previous classes

Say, if I observe M, I am noting down 1 and if F, I am noting down 0.

For the first record I note down, define

$$\begin{aligned} X &= 1 \ \ if \ \ M \quad (21) \\ &= 0 \ \ if \ \ F \quad (22) \\ & \quad (23) \end{aligned}$$

X is a random variable.

Assume the actual proportion of males in UChicago is p. We do not know p (otherwise we would not have sampled). Then

$$X \sim Ber(p)$$

How to connect to previous classes

Continue taking records. Suppose I have taken 100 independent records.

How to connect to previous classes

Continue taking records. Suppose I have taken 100 independent records.

My state space now is

$$S = \{(x_1 x_2 \cdots x_n) : x_i \in \{M, F\}\}$$

How to connect to previous classes

Continue taking records. Suppose I have taken 100 independent records.

My state space now is

$$S = \{(x_1 x_2 \cdots x_n) : x_i \in \{M, F\}\}$$

Now I can define n random variables $X_1, X_2, \cdots, X_n$, so that

$$\begin{aligned} X_1 &= 1 \;\; if \;\; (M x_2 x_3 \cdots x_n) && (24) \\ &= 0 \;\; if \;\; (F x_2 x_3 \cdots x_n) && (25) \\ & && (26) \end{aligned}$$

and so on.

How to connect to previous classes

Continue taking records. Suppose I have taken 100 independent records.

My state space now is

$$S = \{(x_1 x_2 \cdots x_n) : x_i \in \{M, F\}\}$$

Now I can define n random variables $X_1, X_2, \cdots, X_n$, so that

$$\begin{aligned} X_1 &= 1 \;\; if \;\; (M x_2 x_3 \cdots x_n) && (24) \\ &= 0 \;\; if \;\; (F x_2 x_3 \cdots x_n) && (25) \\ & && (26) \end{aligned}$$

and so on.

How to connect to previous classes

Continue taking records. Suppose I have taken 100 independent records.

My state space now is

$$S = \{(x_1 x_2 \cdots x_n) : x_i \in \{M, F\}\}$$

Now I can define n random variables $X_1, X_2, \cdots, X_n$, so that

$$\begin{aligned} X_1 &= 1 \;\; if \;\; (M x_2 x_3 \cdots x_n) && (24) \\ &= 0 \;\; if \;\; (F x_2 x_3 \cdots x_n) && (25) \\ & && (26) \end{aligned}$$

and so on.

How to connect to previous classes

Look at $\frac{Y}{n} = \bar{X}_n$. Use that as an approximation for p.

How to connect to previous classes

Look at $\frac{Y}{n} = \bar{X}_n$. Use that as an approximation for p.

The more sample records you choose, more is n and by Law of Large numbers, the closer your approximation will be to the true value p.

How to connect to previous classes

Look at $\frac{Y}{n} = \bar{X}_n$. Use that as an approximation for p.

The more sample records you choose, more is n and by Law of Large numbers, the closer your approximation will be to the true value p.

In an ideal world, if you had managed to record all the students in Uchicago, you will be able to determine p exactly.

How to connect to previous classes

Look at $\frac{Y}{n} = \bar{X}_n$. Use that as an approximation for p.

The more sample records you choose, more is n and by Law of Large numbers, the closer your approximation will be to the true value p.

In an ideal world, if you had managed to record all the students in Uchicago, you will be able to determine p exactly.

The point of sampling is if by just recording 100/1000 students can give you a somewhat close approximation to actual parameter p, why bother recording all students and waste your time and energy and resources?

How to connect to previous classes

Look at $\frac{Y}{n} = \bar{X}_n$. Use that as an approximation for p.

The more sample records you choose, more is n and by Law of Large numbers, the closer your approximation will be to the true value p.

In an ideal world, if you had managed to record all the students in Uchicago, you will be able to determine p exactly.

The point of sampling is if by just recording 100/1000 students can give you a somewhat close approximation to actual parameter p, why bother recording all students and waste your time and energy and resources?

That is the main reason why we do Statistics!!

A fun matching problem

http://www.rossmanchance.com/applets/randomBabies/RandomBabies.html

Joint Distribution

Consider two independent tosses of a coin.

Let X_1 be the number of Ts observed in the 2 tosses.

Let X_2 be 1 if at least one T observed, 0 otherwise.

Joint Distribution

Consider two independent tosses of a coin.

Let X_1 be the number of Ts observed in the 2 tosses.

Let X_2 be 1 if at least one T observed, 0 otherwise.

The table

states	X_1	X_2	Prob
HH	0	0	0.25
HT	1	1	0.25
TH	1	1	0.25
TT	2	1	0.25

Joint Probability Table

$X1 \| X2$	0	1
0	0.25	0
1	0	0.50
2	0	0.25

Joint Probability Table

$X1 \vert X2$	0	1
0	0.25	0
1	0	0.50
2	0	0.25

$X1 \vert X2$	0	1	Total
0	0.25	0	0.25
1	0	0.50	0.50
2	0	0.25	0.25
Total	0.25	0.75	1

Joint Probability Table

Marginal distribution of $X1$

$X1 \mid X2$	0	1	Total
0	0.25	0	0.25
1	0	0.50	0.50
2	0	0.25	0.25
Total	0.25	0.75	1

Joint Probability Table

Marginal distribution of $X1$

$X1 \mid X2$	0	1	Total
0	0.25	0	0.25
1	0	0.50	0.50
2	0	0.25	0.25
Total	0.25	0.75	1

Marginal distribution of $X2$

Joint Probability Table

Marginal distribution of $X1$

$X1 \mid X2$	0	1	Total
0	0.25	0	0.25
1	0	0.50	0.50
2	0	0.25	0.25
Total	0.25	0.75	1

Marginal distribution of $X2$

$X1 \mid X2$	0	1	Total
0	0.25	0	0.25
1	0	0.50	0.50
2	0	0.25	0.25
Total	0.25	0.75	1

Joint and Marginal Probabilities

For two random variables X and Y, the joint probability

$$p(x, y) = Pr(X = x, Y = y)$$

where x and y are possible relaizations of the random variables X and Y.

Joint and Marginal Probabilities

For two random variables X and Y, the joint probability

$$p(x, y) = Pr(X = x, Y = y)$$

where x and y are possible relaizations of the random variables X and Y.

From what we observed in the last slide, marginal distribution of X

$$p_X(x) = Pr(X = x) = \sum_y p(x, y)$$

where the $\sum_y$ is over all realizations of Y.

Joint and Marginal Probabilities

For two random variables X and Y, the joint probability

$$p(x, y) = Pr(X = x, Y = y)$$

where x and y are possible relaizations of the random variables X and Y.

From what we observed in the last slide, marginal distribution of X

$$p_X(x) = Pr(X = x) = \sum_y p(x, y)$$

where the $\sum_y$ is over all realizations of Y.

Similarly for Y

$$p_Y(y) = Pr(Y = y) = \sum_x p(x, y)$$

Conditional Probabilities

we define a conditional probability as

$$Pr(X = x | Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

Intuition: Given we have observed realization y of Y, what is the probability distribution of X.

Conditional Probabilities

we define a conditional probability as

$$Pr(X = x | Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

Intuition: Given we have observed realization y of Y, what is the probability distribution of X.

The distribution of $X1$ conditional on $X2 = 0$ and $X2 = 1$ in the previous example.

Conditional Probabilities

we define a conditional probability as

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

Intuition: Given we have observed realization y of Y, what is the probability distribution of X.

The distribution of $X1$ conditional on $X2 = 0$ and $X2 = 1$ in the previous example.

$X1 \vert X2$	0	1
0	1	0
1	0	0.67
2	0	0.33
Total	1	1

Conditional Probabilities

we define a conditional probability as

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

Intuition: Given we have observed realization y of Y, what is the probability distribution of X.

The distribution of $X2$ conditional on $X1 = 0$, $X1 = 1$ and $X1 = 2$ in the previous example.

Conditional Probabilities

we define a conditional probability as

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

Intuition: Given we have observed realization y of Y, what is the probability distribution of X.

The distribution of $X2$ conditional on $X1 = 0$, $X1 = 1$ and $X1 = 2$ in the previous example.

$X1 \| X2$	0	1	Total
0	1	0	1
1	0	1	1
2	0	1	1

Sum of random variables

Consider two independent tosses of a coin.

Let X_1 be the number of Ts observed in the 2 tosses.

Let X_2 be 1 if at least one T observed, 0 otherwise.

Sum of random variables

Consider two independent tosses of a coin.

Let X_1 be the number of Ts observed in the 2 tosses.

Let X_2 be 1 if at least one T observed, 0 otherwise.

The table

states	X_1	X_2	$Y = X_1 + X_2$	Prob
HH	0	0	0	0.25
HT	1	1	2	0.25
TH	1	1	2	0.25
TT	2	1	3	0.25

Sum of random variables

$$
\begin{aligned}
E(Y) &= E(X_1 + X_2) && (27)\\
&= 0 \times 0.25 + 2 \times 0.50 + 3 \times 0.25 && (28)\\
&= 1.75 && (29)\\
& && (30)
\end{aligned}
$$

Sum of random variables

$$
\begin{align}
E(Y) &= E(X_1 + X_2) \tag{27}\\
&= 0 \times 0.25 + 2 \times 0.50 + 3 \times 0.25 \tag{28}\\
&= 1.75 \tag{29}\\
\tag{30}
\end{align}
$$

$$E(X_1) = 1 \qquad E(X_2) = 0.75$$

Sum of random variables

$$
\begin{aligned}
E(Y) &= E(X_1 + X_2) && (27) \\
&= 0 \times 0.25 + 2 \times 0.50 + 3 \times 0.25 && (28) \\
&= 1.75 && (29) \\
& && (30)
\end{aligned}
$$

$$E(X_1) = 1 \qquad E(X_2) = 0.75$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

Sum of random variables

$$
\begin{align}
E(Y) &= E(X_1 + X_2) \tag{27} \\
&= 0 \times 0.25 + 2 \times 0.50 + 3 \times 0.25 \tag{28} \\
&= 1.75 \tag{29} \\
& \tag{30}
\end{align}
$$

$$E(X_1) = 1 \qquad E(X_2) = 0.75$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

Is that always true?

Sum of random variables

$$
\begin{aligned}
E(Y) &= E(X_1 + X_2) && (27)\\
&= 0 \times 0.25 + 2 \times 0.50 + 3 \times 0.25 && (28)\\
&= 1.75 && (29)\\
& && (30)
\end{aligned}
$$

$$E(X_1) = 1 \qquad E(X_2) = 0.75$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

Is that always true? YES!!! Lets derive it.

Similar Results

- $E(c) = c$ where c is constant
- $E(aX) = aE(X)$
- $E(X + Y) = E(X) + E(Y)$
- $E(aX + bY) = aE(X) + bE(Y)$ where a, b are constants.

Expected Value of a Linear Combination

Show that $E(aX + bY) = aE(X) + bE(Y)$.

$$E(aX + bY) = \sum_x \sum_y (ax + by)p(x, y)$$

Expected Value of a Linear Combination

Show that $E(aX + bY) = aE(X) + bE(Y)$.

$$E(aX + bY) = \sum_x \sum_y (ax + by)p(x, y)$$

$$= \sum_x \sum_y axp(x, y) + \sum_x \sum_y byp(x, y)$$

$$= \sum_x \sum_y axp(x, y) + \sum_y \sum_x byp(x, y)$$

Expected Value of a Linear Combination

Show that $E(aX + bY) = aE(X) + bE(Y)$.

$$E(aX + bY) = \sum_x \sum_y (ax + by)p(x, y)$$

$$= \sum_x \sum_y axp(x, y) + \sum_x \sum_y byp(x, y)$$

$$= \sum_x \sum_y axp(x, y) + \sum_y \sum_x byp(x, y)$$

$$= \sum_x ax \underbrace{\left[\sum_y p(x, y)\right]} + \sum_y by \underbrace{\left[\sum_x p(x, y)\right]} \quad *$$

* See HW0 for more details about the "double sum."

Expected Value of a Linear Combination

Show that $E(aX + bY) = aE(X) + bE(Y)$.

$$E(aX + bY) = \sum_x \sum_y (ax + by) p(x, y)$$

$$= \sum_x \sum_y axp(x, y) + \sum_x \sum_y byp(x, y)$$

$$= \sum_x \sum_y axp(x, y) + \sum_y \sum_x byp(x, y)$$

$$= \sum_x ax \underbrace{\left[\sum_y p(x, y) \right]}_{=p_X(x)} + \sum_y by \underbrace{\left[\sum_x p(x, y) \right]} \quad *$$

* See HW0 for more details about the "double sum."

Expected Value of a Linear Combination

Show that $E(aX + bY) = aE(X) + bE(Y)$.

$$E(aX+bY) = \sum_x \sum_y (ax+by)p(x,y)$$

$$= \sum_x \sum_y axp(x,y) + \sum_x \sum_y byp(x,y)$$

$$= \sum_x \sum_y axp(x,y) + \sum_y \sum_x byp(x,y)$$

$$= \sum_x ax \underbrace{\left[\sum_y p(x,y)\right]}_{=p_X(x)} + \sum_y by \underbrace{\left[\sum_x p(x,y)\right]}_{=p_Y(y)} \quad *$$

* See HW0 for more details about the "double sum."

Expected Value of a Linear Combination

Show that $E(aX + bY) = aE(X) + bE(Y)$.

$$
\begin{aligned}
E(aX + bY) &= \sum_x \sum_y (ax + by)p(x, y) \\
&= \sum_x \sum_y axp(x, y) + \sum_x \sum_y byp(x, y) \\
&= \sum_x \sum_y axp(x, y) + \sum_y \sum_x byp(x, y) \\
&= \sum_x ax \underbrace{\left[\sum_y p(x, y)\right]}_{=p_X(x)} + \sum_y by \underbrace{\left[\sum_x p(x, y)\right]}_{=p_Y(y)} \quad * \\
&= a \sum_x xp_X(x) + b \sum_y yp_Y(y)
\end{aligned}
$$

* See HW0 for more details about the "double sum."

Expected Value of a Linear Combination

Show that $E(aX + bY) = aE(X) + bE(Y)$.

$$
\begin{aligned}
E(aX + bY) &= \sum_x \sum_y (ax + by) p(x, y) \\
&= \sum_x \sum_y axp(x, y) + \sum_x \sum_y byp(x, y) \\
&= \sum_x \sum_y axp(x, y) + \sum_y \sum_x byp(x, y) \\
&= \sum_x ax \underbrace{\left[\sum_y p(x, y) \right]}_{= p_X(x)} + \sum_y by \underbrace{\left[\sum_x p(x, y) \right]}_{= p_Y(y)} \quad * \\
&= a \sum_x x p_X(x) + b \sum_y y p_Y(y) \qquad = aE(X) + bE(Y)
\end{aligned}
$$

* See HW0 for more details about the "double sum."

Independence of two random variables

- The variable X is **independent** of variable Y if its chances are not affected by the variable Y,

$$Pr[X = x|Y = y] = Pr[X = x] \quad \forall\, x \quad each \;\; y$$

- The following 3 definitions of independence are equivalent.
  - $Pr[X = x|Y = y] = Pr[X = x]$
  - $Pr[Y = y|X = x] = Pr[Y = y]$
  - $Pr[X = x, Y = y] = p_{XY}(x, y) = Pr[X = x]Pr[Y = y] = p_X(x)p_Y(y)$

Independence for Random Variables

$$p(x, y) = p_X(x)p_Y(y).$$

Note: Independence implies that

$$p(y|X = x) = \frac{p(x, y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y).$$

Y \ X	0	1	2	$p(y)$
0	2/6	1/6	0	3/6
1	0	1/6	2/6	3/6
$p(x)$	2/6	2/6	2/6	

Consider the pair $(x, y,) = (0, 0)$.

$$p(0, 0) = 2/6 \neq p_X(0)p_Y(0) = (1/3)(1/2) = 1/6.$$

X and Y are not independent.

Covariance between Random Variables

The covariance between random variables X and Y is defined as

$$\begin{aligned}
cov(X, Y) &= E(X - E(X)(Y - E(Y))) && (31)\\
&= E(XY - E(X)Y - XE(Y) + E(X)E(Y)) && (32)\\
&= E(XY) - E(E(X)Y) - E(XE(Y)) + E(E(X)E(Y)) && (33)\\
&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) && (34)\\
&= E(XY) - E(X)E(Y) && (35)\\
&= \sum_{x,y} xyPr(X = x, Y = y) - E(X)E(Y) && (36)\\
& && (37)
\end{aligned}$$

Covariance between Random Variables

The covariance between random variables X and Y is defined as

$$
\begin{aligned}
cov(X,Y) &= E(X - E(X)(Y - E(Y))) && (31)\\
&= E(XY - E(X)Y - XE(Y) + E(X)E(Y)) && (32)\\
&= E(XY) - E(E(X)Y) - E(XE(Y)) + E(E(X)E(Y)) && (33)\\
&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) && (34)\\
&= E(XY) - E(X)E(Y) && (35)\\
&= \sum_{x,y} xyPr(X = x, Y = y) - E(X)E(Y) && (36)\\
& && (37)
\end{aligned}
$$

When X and Y are independent, we also have $cov(X,Y) = 0$

Covariance between Random Variables

The covariance between random variables X and Y is defined as

$$
\begin{aligned}
cov(X,Y) &= E(X - E(X)(Y - E(Y))) && (31)\\
&= E(XY - E(X)Y - XE(Y) + E(X)E(Y)) && (32)\\
&= E(XY) - E(E(X)Y) - E(XE(Y)) + E(E(X)E(Y)) && (33)\\
&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) && (34)\\
&= E(XY) - E(X)E(Y) && (35)\\
&= \sum_{x,y} xyPr(X = x, Y = y) - E(X)E(Y) && (36)\\
& && (37)
\end{aligned}
$$

When X and Y are independent, we also have $cov(X,Y) = 0$

Example to disprove the converse

Let X be a random variable

$$X = 1 \;\; prob \;\; 0.5 \qquad\qquad = -1 \;\; prob \;\; 0.5 \tag{38}$$

Let Y be another random variable

$$\begin{aligned} Y &= 0 \;\; if \; X = 1 && (39)\\ &= +1 \;\; if \; X = -1 \; prob \, 0.5 && (40)\\ &= -1 \;\; if \; X = -1 \; prob \, 0.5 && (41)\\ & && (42) \end{aligned}$$

Then check that $E(XY) = E(X) = E(Y) = 0$ and so $cov(X, Y) = 0$

Show X and Y are not independent

Correlation between Random Variables

Correlation between two random variables X and Y is given by

$$cor(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$

It can be shown that (check Cauchy Schwartz inequality)

$$-1 \leq cor(X, Y) \leq +1$$

Correlation (and covariance) is a measure of *linear* dependency. Correaltion of -1 means negative linear dependency, $+1$ means positive linear dependency, 0 means no linear dependency.

Recap

- States of an experiment, Events, Random Variables
- Probability (Joint, Marginal and Conditional)
- Expectation, Variance, Moments, MGF
- Law of Large Numbers, Chebyshev Inequality
- Independence and Covariance
- Sums of random variables
- Connect with Statistical Data !!

Is there something I very conveniently skipped?

Continuous Random Variables

Position yourself again at the middle of the Quad. Suppose now you record the height of each person crossing the Quad.

Continuous Random Variables

Position yourself again at the middle of the Quad. Suppose now you record the height of each person crossing the Quad.

In this case each person whose height is measured is a run of the experiment. The state is the height (in ft/inches). Since the state itself is numeric, we can directly define a random variable X same as the state.

Continuous Random Variables

Position yourself again at the middle of the Quad. Suppose now you record the height of each person crossing the Quad.

In this case each person whose height is measured is a run of the experiment. The state is the height (in ft/inches). Since the state itself is numeric, we can directly define a random variable X same as the state.

Now we record the states for n individuals. Then the state space

$$S = \{(h_1 h_2 \cdots h_n)\}$$

and define n random variables $X_1, X_2, \cdots, X_n$ such that

$$X_1 = h_1 \quad X_2 = h_2 \quad \cdots X_n = h_n$$

Continuous Random Variables

If we had recorded the heights of all students, the assume the histogram would have been approximated by a normal bell-shaped curve with center μ and variance σ^2 very accurately. Then

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma)$$

Continuous Random Variables

If we had recorded the heights of all students, the assume the histogram would have been approximated by a normal bell-shaped curve with center μ and variance σ^2 very accurately.Then

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma)$$

You can then build the histogram of the n realizations. Then we fit a curve as a fit to the histogram.

Continuous Random Variables

If we had recorded the heights of all students, the assume the histogram would have been approximated by a normal bell-shaped curve with center μ and variance σ^2 very accurately. Then

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma)$$

You can then build the histogram of the n realizations. Then we fit a curve as a fit to the histogram.

Continuous Random Variables

As $n \to \infty$, the fitted curve would get closer and closer to the normal bell-shaped curve with the above mean and variance parameters.

Continuous Random Variables

As $n \to \infty$, the fitted curve would get closer and closer to the normal bell-shaped curve with the above mean and variance parameters.

The sample mean $\bar{X}_n$ and sample standard deviation s_n^2 will also get closer and closer to μ and σ respectively as $n \to \infty$.

Continuous Random Variables

As $n \to \infty$, the fitted curve would get closer and closer to the normal bell-shaped curve with the above mean and variance parameters.

The sample mean $\bar{X}_n$ and sample standard deviation s_n^2 will also get closer and closer to μ and σ respectively as $n \to \infty$.

But again, we save our time and energy by looking at only 100/1000 samples and then approximate the normal model by approximately determining the population or model parameters μ and σ from the n data points and corresponding histogram.

Continuous Probabilities

We would want to define $Pr(X = x)$ for a continuous random variable X, but unfortunately it is 0. We will soon see why

Continuous Probabilities

We would want to define $Pr(X = x)$ for a continuous random variable X, but unfortunately it is 0. We will soon see why

So how do we proceed to build something like a probability table for discrete variable?
We define **cumulative density function** (cdf)

$$F(x) = Pr(X \leq x)$$

If we differentiate this function, we get **probability density function** (pdf)

$$\frac{d}{dx}F(x) = f(x)$$

$$f(x) \geq 0 \qquad \int f(x)dx = 1$$

Continuous Probabilities

For normal distribution

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

Cumulative distribution

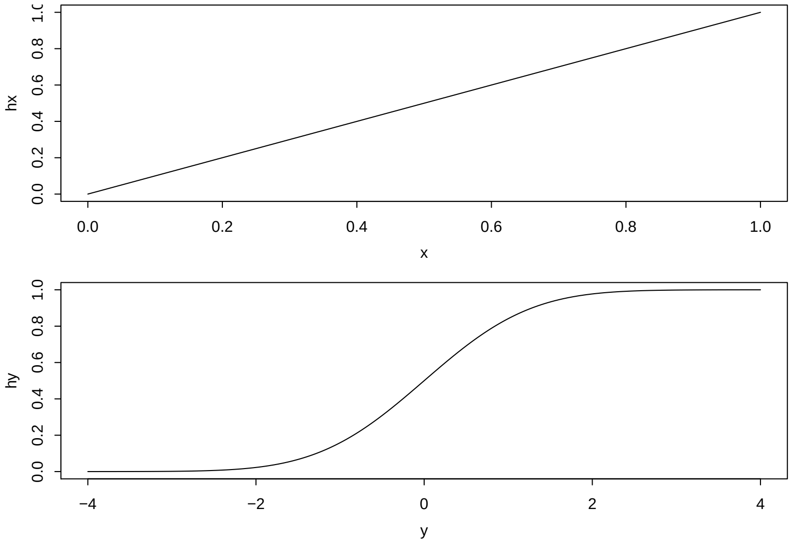
$$\Phi(x) := \int_{-\infty}^{x} \phi(x)dx$$

Uniform distribution

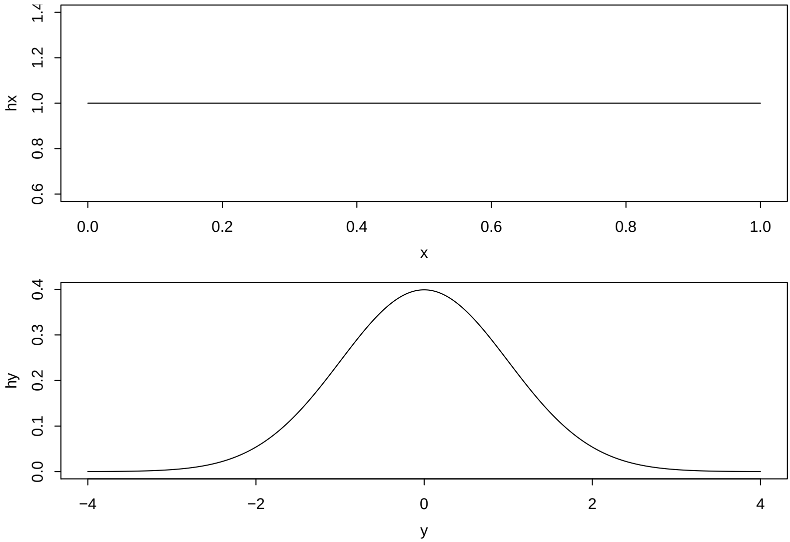
$$F(x) = Pr(X \leq x) = x$$

which implies $f(x) = 1$ for all x.

cumulative density graphs (Uniform and Normal)



probability density graphs (Uniform and Normal)



Properties of Continous Random Variable

We define the expectation analogous to the discrete random variable.

$$E(X) := \int xf(x)dx$$

$$var(X) := E(X - E(X))^2 = \int x^2 f(x)dx - E^2(X)$$

$$MGF(t) := E(e^{tX}) = \int e^{tx} f(x)dx$$

For normal distribution, one can easily show that

$$E(X) = \mu \quad var(X) = \sigma^2 \quad MGF(t) = exp(\mu t + \frac{\sigma^2 t^2}{2})$$

what about uniform? HW

Properties of Continous Random Variable

For two continuous random variables X and Y, we can define the joint cumulative density

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

Then we define joint density

$$f(x, y) = \frac{d^2}{dxdy} F(x, y) = f_{XY}(x, y)$$

Define marginal

$$\int_X f_{XY}(x, y) = f_Y(y) \qquad \int_Y f_{XY}(x, y) = f_X(x)$$

Conditional

$$f_{X|Y}(x|y) := \frac{f_{XY}(x, y)}{f_Y(y)}$$

Properties of Continous Random Variable

We say X and Y are independent when

$$f_{X|Y}(x|y) = f_X(x)$$

or

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

As for discrete variables, independence implies uncorrelated but converse is not true.
Similarly results related to sums and linear transform of two or more variables stay true even for continuous variables.