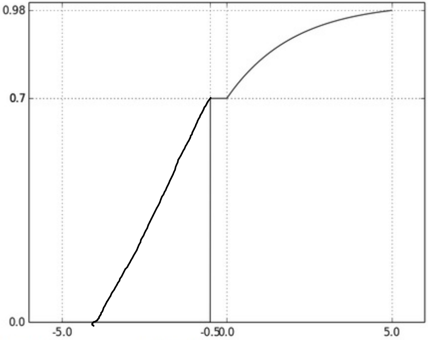
Random Variables, Expectation and Variance

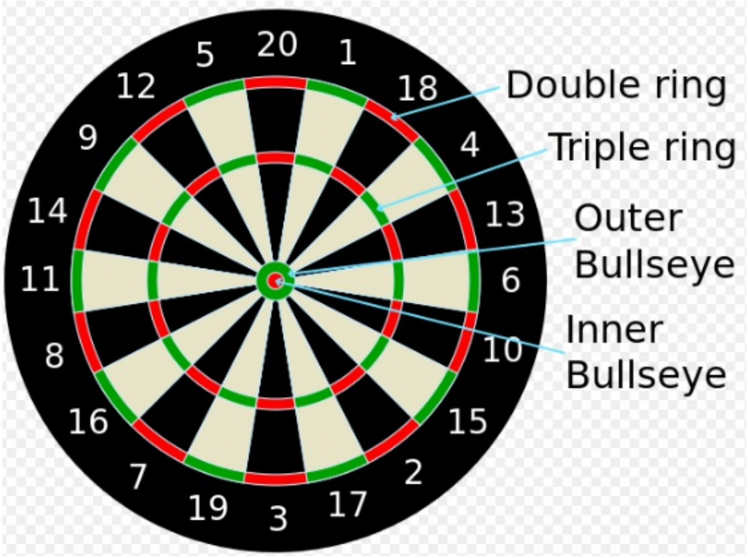
(1 pt)
Below is the CDF of a mixture distribution with **two** components.
One of the components is either a normal or an exponential distribution; the other is either a point mass or a uniform distribution.
All parameters of component distributions are small multiples of 0.5.
λ of exponential components and std of normal components take on value 0.5, 1 or 1.5.
Component weights take on multiples of 0.05 and they need to sum to one.



Identify the component distributions:

- The exponential component has λ of 0.5. Its component weight is 0,3
- Point mass on -0,5 . Its component weight is 0,7

Densities over a 2D space

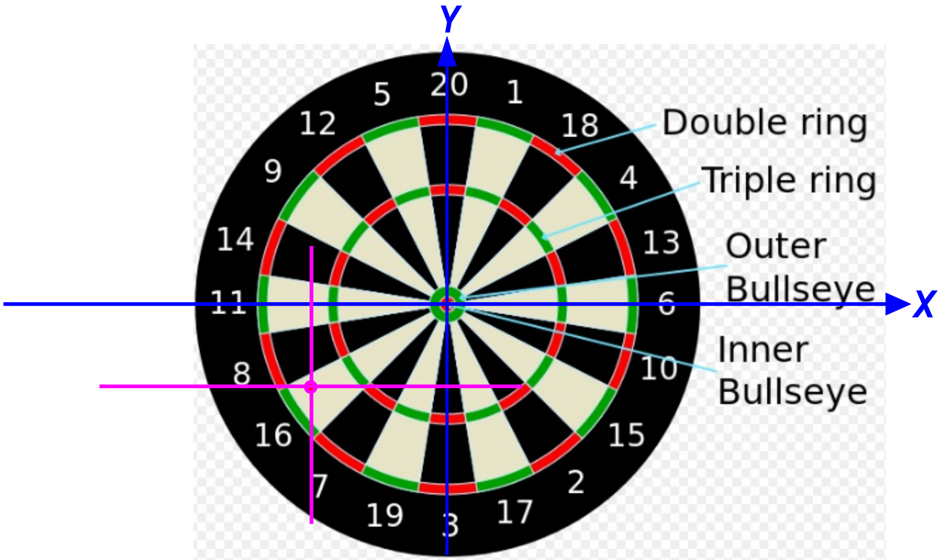


the sample space is the plane

x and y are mappings from the plane to R

Such mappings are called Random Variables

A natural assumption: the probability distribution of the location (x,y) that the dart falls is a density function that is highest at the bullseye and gets lower the further from the bullseye you get.



***Events** are sets.*
Events map outcomes to {True,False}
True = Outcome in set.
False = Outcome not in set.

Examples of Events:

1. ***Dart lands on Inner Bullseye***
2. ***Dart lands on double ring.***
3. ***Dart lands on the "14" section***

***Random variables** are functions (mappings) from Omega to R*
Examples of Random variables:

1. ***X position of dart***
2. ***Y position of dart***
3. ***Distance of dart from middle of target***
4. ***The number associated with the section in which the dart landed***

baseball acronyms

G	Games Played
PA	Plate Appearances
AB	At Bat
R	Runs Scored
H	Hits
D	?
T	?
HR	Home Runs
RBI	Runs Batted In
BB	Bases on Balls (walks)
SO	Strikeouts
BA	?
OBP	On base Percentage
SLG	Slugging Percentage
OPS	OBP+SLG

Player	G	PA	AB	R	H	D	T	HR	RBI	BB	SO	BA	OBP	SLG	OPS
Mike Napoli	17	64	49	13	16	4	0	6	14	15	16	.327	.484	.776	1.260
Josh Donaldson	20	88	72	17	28	8	0	5	15	14	11	.389	.500	.708	1.208
Hunter Pence	21	92	78	19	25	4	0	9	26	13	14	.321	.413	.718	1.131
Matt Carpenter	21	102	88	23	35	11	2	1	11	12	19	.398	.480	.602	1.083
Ryan Zimmerman	21	97	90	22	29	2	0	**11**	16	7	18	.322	.371	.711	1.082
Freddie Freeman	20	85	73	15	26	3	0	6	17	10	15	.356	.435	.644	1.079
Michael Cuddyer	16	67	62	8	26	4	0	3	14	4	10	.419	.448	.629	1.077
Adam Lind	18	60	55	11	16	2	0	7	17	5	11	.291	.350	.709	1.059
Andrew McCutchen	20	83	66	13	22	5	1	3	8	13	11	.333	.470	.576	1.046
Prince Fielder	20	86	79	11	30	7	0	4	14	6	13	.380	.419	.620	1.039
Shin-Soo Choo	18	85	60	15	18	3	0	4	10	21	12	.300	.488	.550	1.038
Paul Goldschmidt	21	92	80	12	27	6	2	4	19	11	19	.338	.424	.613	1.036
Moises Sierra	19	67	63	8	22	12	1	1	9	4	14	.349	.388	.619	1.007
Josmil Pinto	16	62	58	9	21	5	0	3	9	4	12	.362	.403	.603	1.007
Mike Trout	21	95	71	16	21	5	1	3	10	23	21	.296	.474	.521	.995
Yoenis Cespedes	19	80	77	12	26	2	1	6	19	2	19	.338	.363	.623	.986
Matt Holliday	20	92	76	14	28	6	0	2	20	14	13	.368	.457	.526	.983
David Ortiz	20	91	76	18	21	9	0	5	16	13	16	.276	.385	.592	.977
Chase Headley	17	67	56	8	15	2	0	5	9	10	12	.268	.388	.571	.959
Matt Adams	19	72	69	13	21	1	0	7	14	3	21	.304	.333	.623	.957
Joey Votto	20	94	73	12	22	3	0	4	9	20	17	.301	.447	.507	.954
Eric Hosmer	20	87	78	12	27	6	1	2	12	9	18	.346	.414	.526	.939
Wil Myers	20	84	78	10	24	9	0	4	10	6	18	.308	.357	.577	.934
Giancarlo Stanton	20	85	72	12	19	3	0	6	16	11	27	.264	.376	.556	.932
Desmond Jennings	21	83	68	7	19	6	1	3	13	13	16	.279	.398	.529	.927

Outcome space: ***all possible performances of baseball hitters for a month***
Outcome: ***The performance of a particular player***
Random variables: ***measures of performance: G, PA, AB ...***

Events: ***More than 8 home runs,***
OPS higher than 1.0, 1.1, 1.2, ...

Two random variables: X, Y are independent if and only if any event conditioned on X is independent of any event conditioned on Y

A single Random variable X defines
many different events: $X>1$, $X<-3$, $0<X<1$ ...

Two random variables: X, Y are independent if and only if
any event defined using X
is independent of
any event defined using Y

Examples:
$P(X>5 \text{ and } Y<7) = P(X>5)P(Y<7)$
$P(X=2 \text{ and } Y=0) = P(X=2)P(Y=0)$
....

If the set of values the random variables get is infinite,
we need to check and infinite number of conditions!!

If the sets are finite, we can check for independence
using a contingency table.

Joint distribution of two dependent random variables

	X=1	X=2	X=10	P(Y=y)
Y=-1	1/12	2/12	1/12	4/12= 1/3
Y=+1	2/12	1/12	5/12	8/12= 2/3
P(X=x)	3/12= 1/4	3/12= 1/4	6/12= 1/2	

Marginals

2/12

$P(X=2 \wedge Y=-1)$

$\overset{?}{\neq} P(X=2)P(Y=-1)$

1/4 · 1/3

Joint distribution of two independent random variables

	X=1	X=2	X=10	P(Y=y)
Y=-1	1/12	1/12	2/12	4/12= 1/3
Y=+1	2/12	2/12	4/12	8/12= 2/3
P(X=x)	3/12= 1/4	3/12= 1/4	6/12= 1/2	

Marginals

Expected Value

- Suppose X is a discrete random variable $P(X = a_i) = p_i$
  - The expected value of X is $E(X) = \sum_{i=1}^{n} p_i a_i$
- Suppose X is a continuous random variable with density f
  - The expected value of X is $E(X) = \int_{-\infty}^{+\infty} f(x) x dx$
- $E(X)$ is a property of the distribution, it is not a random variable.
- The average is a random variable:
  - $Average(x_1, x_2, \ldots, x_n) \doteq \frac{1}{n} \sum_{i=1}^{n} x_i$
- When n is large, the average tends to be close to the mean.

Expected Value

- Suppose X is a discrete random variable $P(X = a_i) = p_i$
  - The expected value of X is $E(X) = \sum_{i=1}^{n} p_i a_i$
- Suppose X is a continuous random variable with density f
  - The expected value of X is $E(X) = \int_{-\infty}^{+\infty} f(x)xdx$
- $E(X)$ is a property of the distribution, it is not a random variable.
- The average is a random variable:
  - $Average(x_1, x_2, \dots, x_n) \doteq \frac{1}{n}\sum_{i=1}^{n} x_i$
- When n is large, the average tends to be close to the mean.

Example - Binary random variables:

Let $X_1, X_2, \ldots, X_{100}$

Be independent binary random variables: $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$

Let $S = \frac{1}{100}\sum_{i=1}^{100} X_i$ $\quad$ S is the average, S is/is-not a random variable?

$E(X_i) = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2}$, $\quad E(X_i)$ is/is-not a random variable?

What is $E(S)$? $= E(X) = \frac{1}{2}$

Expectation for a random variable with n possible values

The random variable X takes on the value x_i with probability p_i for $i = 1,2,\ldots,n$

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

Expectation is a linear operator. X is a random variable, a,b are constants

$$E(aX + b) = \sum_{i=1}^{n} p_i(ax_i + b) = a\sum_{i=1}^{n} p_i x_i + b\sum_{i=1}^{n} p_i = aE(x) + b$$

Linearity of Expectation: the expectation of a sum is equal to the sum of the expectations.
X,Y are random variables.

$$E(X + Y) = \sum_{i=1}^{n} p_i(x_i + y_i) = \sum_{i=1}^{n} p_i x_i + \sum_{i=1}^{n} p_i y_i = E(x) + E(y)$$

Rules for expected value:

1. If a, b are constants and X is a random variable then

$$E(aX+b) = aE(X)+b$$

2. If X, Y are random variables (dependent or independent)

$$E(X+Y) = E(X)+E(Y)$$

—> what is $E(aX+bY+c)=?$ $aE(x)-bE(y)+c$

3. If the distribution of the RV X is a mixture of two distributions:

$P = \mu P_1 + (1-\mu)P_2$ then

$$E_P(X) = \mu E_{P_1}(X) + (1-\mu)E_{P_2}(X)$$

Suppose $x_1, x_2, \ldots, x_n$ are random variables that all have the same mean: $E(x_i) = a$
What is the expected value of the average? $y = \frac{1}{n}\sum_{i=1}^{n} x_i$?

$$E(y) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}(na) = a$$

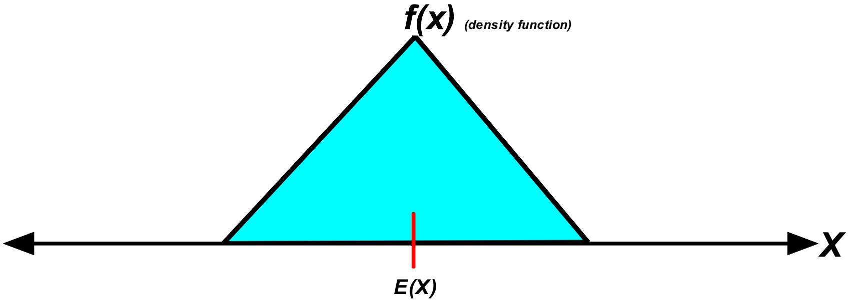
The mean is the center of mass of the distribution

If the distribution is symmetric around zero, then the mean is zero.
If the distribution is symmetric around a, then the mean is a.

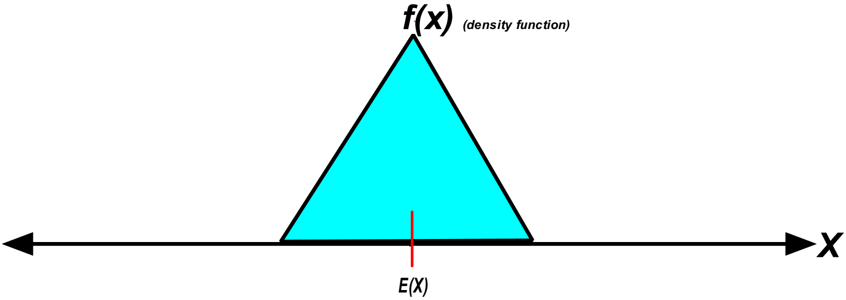
1. If a,b are constants and X is a random variable then

$$E\left(aX+b\right)=aE(X)+b$$

E(X) corresponds to the location. ***If we subtract the mean we have a distribution centered at zero: E(X-E(X))=E(X)-E(X)=0***



The mean corresponds to the location of the "center" of the distribution.
How do we measure the "width" of the distribution?



Measuring the width of the distribution

Lets use $\mu \doteq E(X)$

We already know that $E(X-\mu)=0$

To find the width we could use $E\left(|X-\mu|\right)$

But it is much more convenient to use:

$$Var(X) \doteq E\left((X-\mu)^2\right)$$

Using the rules for expected value (remember that μ is a constant)

$$Var(X) \doteq E\left((X-\mu)^2\right) = E\left(X^2 - 2\mu X + \mu^2\right)$$

$$= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - E(X)^2$$

Properties of the variance

$$Var(aX) = E[(aX)^2] - (E[aX]\,)^2 = a^2E[X^2] - a^2(E[X])^2 = a^2Var(X)$$

$$Var(X+b) = E[(X+b)^2] - (E[X+b]\,)^2 = E[X^2 + 2Xb + b^2] - (E[X]+b)^2 =$$
$$(E[X^2] + 2E[X]b + b^2) - (E[X]^2 + 2E[X]b + b^2) = E[X^2] - E[X]^2 = Var(X)$$

Properties of the variance

1. If a, b are constants and X is a random variable then

$$Var(aX + b) = a^2 Var(X)$$

2. If X, Y are Independent Random Variables, then

$$Var(X + Y) = Var(X) + Var(Y)$$

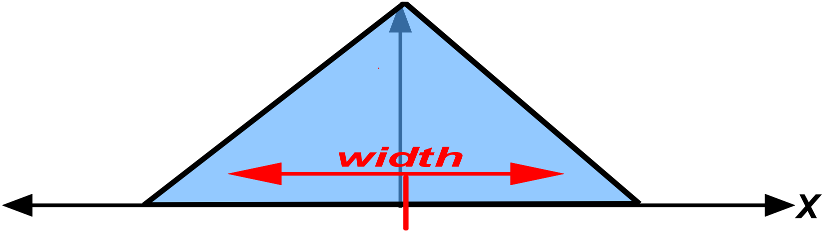
3. If the distribution of the RV X is a mixture of two distributions:

$P = \mu P_1 + (1 - \mu) P_2$ then..... (nothing)

The standard Deviation

Recall that $Var(aX) = a^2 Var(X)$

But the width has to increase by a factor of a when we multiply the random variable X by a not a^2



Solution: Use the square-root of the variance

Width=Standard Deviation=std=$\sigma(X) = \sqrt{Var(X)}$

The std scales linearly with a : $\sigma(aX) = a\sigma(X)$

Excercise:

Suppose X, Y are independent random variables.

$$E(X) = 5, Var(X) = 2$$
$$E(Y) = -1, Var(Y) = 1$$

Calculate:

1. $E[X+Y] =$ $E(X) + E(Y) = 4$
2. $Var[X+Y] =$ $Var(X) + Var(Y) = 3$
3. $E[2X - Y + 5]$= $2E(X) - E(Y) + 5 = 21$
4. $Var[2X - Y + 5] =$ $2^2 Var(X) + (-1)^2 Var(Y)$

$= 8 + 1 = 9$