Probability Basics

- Experiment: Process by which an observation is made
- Event: Outcome of experiment
- Simple Event: Event composed of one element
- Compound Event: Event composed of multiple elements.
- Sample Space (S): Set of all possible events (outcomes)

We denote the <u>probability</u> of event A as $P(A)$. Probability is a <u>measure</u>.

Probability / Kolmogorov Axioms

controversial in some circles

- For any event E, $P(E) \geq 0$
- $P(S) = 1$
- Let $\{E_i\}$ be a sequence of events such that $E_i \cap E_j = \emptyset$ for $i \neq j$. Then $P(\cup E_i) = \sum P(E_i)$

Any sample space / event can be broken up losslessly

Probability Rules

- Additive Rule: $P(A) + P(B) - P(A \cap B)$
  - $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
  - General form is inclusion-exclusion / sieve principle
- Complement Rule: $P(A^c) = 1 - P(A)$

Sample Point Method

1) Define the experiment
2) Determine elements of S
3) Assign probabilities of elements of S
4) Express event as combination of simple events
5) Apply Kolmogorov's 3rd axiom

This is intractable in general. However, we can simplify it for for simple cases

Counting

<u>Thm</u>: mn rule
If A has m elements & N has n elements, then there are mn pairs containing one element from each set.

<u>Def</u>: Permutation
A <u>permutation</u> is an ordered arrangement of r distinct elements.

Thm:
The number of ways to select an ordered subset of size r from a set n selected w/o replacement is

based on mn rule!

$$(n)_r = P^n_r = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

Order isn't always relevant tho. Sometimes we just want multiple groups (& often only 2! selected or not).

Thm:
The number of ways to assign n distinct objects into k distinct, mutually disjoint groups of size $n_1, n_2, \ldots, n_k$ where $\sum n_i = n$

$$\binom{n}{n_1 n_2 \ldots n_k} = \frac{n!}{n_1! n_2! \ldots n_k!}$$

These are multinomial coefficients

Def: Multinomial Coefficients
Multinomial coefficients b/c they are coefficients in the expansion of the multinomial $(x_1 + x_2 + \ldots + x_k)^n = \sum \binom{n}{n_1 n_2 \ldots n_k} x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}$.

Def: Combination
The number of combinations is the number of ways to select an unordered subset of size r w/o replacement from a set w/ n possible elements

Thm:
The number of combinations is

$$\binom{n}{r} = C^n_r = \frac{(n)_r}{r!} = \frac{n!}{r!(n-r)!}$$ ← Permutations divided by # of repeated ones w/ different order

Note:

$$\binom{n}{r} = \binom{n}{n-r} = \frac{n!}{r!(n-r)!}$$

This is a special case called the binomial coefficient that appears in the expansion of binomial like $(x+y)^n$

Conditional Probability

Def: Conditional Probability
The probability that some event A occurs given some other event B occurs.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ when } P(B) > 0.$$

Essentially, we're limiting our sample space to B.

Def: Independent Events
Two events are independent iff the occurence of one event A does not affect the probability of the other event B.

Similarly, one of the following holds: (They are equivalent)

- $P(A \cap B) = P(A)P(B)$
- $P(A \mid B) = P(A)$
- $P(B \mid A) = P(B)$

Thm: Multiplicative Law

$P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$

This is when we solve the definition of $P(A \mid B)$ solved for intersection.

Example:

Let $S = \{1, 2, 3, 4\}$, $A = \{1, 2\}$, $B = \{1, 3\}$ are independent.

$P(A \cap B) = 1/4 = P(A)P(B)$

Example:

A money throws a red, black, & white balls into corresponding colors. Assuming one ball goes in each box & she is throwing at random, what is the probability of no matches?

Let R, B, W be the events that she matches red, black, & white respectively. We find $1 - P(R \cup B \cup W)$.

$P(R) = P(B) = P(W) = 1/3$

one ball in each box

$P(R \cap B) = P(R)P(B \mid A) = 1/3 \cdot 1/2 = 1/6$

Similarly $P(R \cap W) = P(B \cap W) = 1/6$

$P(R \cap B \cap W) = P(R)P(B \cap W \mid R) = P(R) \cdot P(B \mid R)P(W \mid R \cap B)$

$= 1/3 \cdot 1/2 \cdot 1 = 1/6$

sieve principle

This is helpful to find $P(R \cup B \cup W)$.

$$P(R \cup B \cup W) = P(R) + P(B) + P(W) - P(R \cap B) - P(R \cap W) - P(B \cap W) + P(R \cap B \cap W)$$
$$= 1/3 + 1/3 + \cancel{1/3} - \cancel{1/6} - 1/6 - \cancel{1/6} + \cancel{1/6}$$
$$= 2/3$$

Therefore $1 - P(R \cup B \cup W) = 1 - 2/3 = 1/3$.

Example:

What about the previous example w/ the monkey except exactly one match

$$P(\text{one match}) = P(R \cap B^c \cap W^c) + P(R^c \cap B \cap W^c) + P(R^c \cap B^c \cap W)$$ ← mutually disjoint, so summing is okay

$$= P(R) + P(B) + P(W)$$
$$- 2P(R \cap B) - 2P(R \cap W) - 2P(B \cap W)$$ ← This sum comes from $P(A \cap B^c \cap W^c) = P(R) - P(R \cap B) - P(R \cap W) + P(R \cap B \cap W)$
$$+ 3P(R \cap B \cap W)$$
$$= \cancel{1/3} + \cancel{1/3} + \cancel{1/3} - \cancel{1/3} - \cancel{1/3} - \cancel{1/3} + 1/2$$
$$= 1/2$$

The previous two examples show non-disjoint, non-independent events can be really annoying to deal w/. If we framed things as disjoint events, <u>event composition</u> would be way easier.

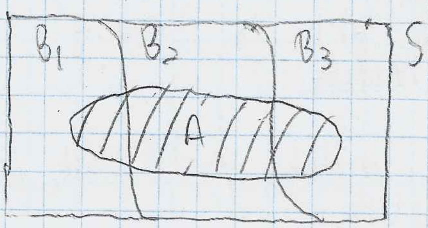
<u>Def:</u> Partition

A <u>partition</u> of S is any collection of sets $\{B_1, \dots, B_k\}$ where

$$S = \bigcup_{i=1}^{k} B_i \quad \& \quad B_i \cap B_j = \emptyset \ \forall i, j \in 1, \dots, k \quad i \neq j$$

With a partition, any set A can be broken down as the union of disjoint sets

$$A = \bigcup_{i=1}^{k} (A \cap B_i) = (A \cap B_1) \cup \dots (A \cap B_k)$$



<u>Thm:</u> Law of Total Probability

Let $\{A_i\}$ be a finite partition of S w/ $P(A_i) > 0 \ \forall i$, then for some event $E \subseteq S$

$$P(E) = \sum_{i=1}^{k} P(E \mid A_i) P(A_i)$$ ← application of union of disjoint intersections above

<u>Thm:</u> Baye's Rule

Let $\{A_i\}$ be a finite partion of S w/ $P(A_i) > 0 \ \forall i$.

$$P(A_j \mid E) = \frac{P(E \mid A_j) P(A_j)}{\sum_{i=1}^{k} P(E \mid A_i) P(A_i)} = \frac{P(A_j \cap E)}{P(E)}$$

You can <u>flip</u> the conditional

top is part of bottom

<u>Def:</u> Sensitivity & Specificity

The <u>sensitivity</u> of a test is the rate at which a test successfully detects a positive case.

The <u>specificity</u> of a test is the rate at which a test successfully ignores a negative case.

Discrete Random Variables & Intro to RVs

Def: Random Variable (RV)
A random variable $X: S \to \mathbb{R}$ are real-valued functions for samples, S.

Def: Support
The support of a RV is the set of all observable real-valued outcomes.

You'd normally call this range but that already exists & is used regularly stats & is used as input to probability function.

Def:
A discrete RV is one w/ a finite or countably infinite support.

Def:
A continuous RV is one w/ an uncountably infinite support.

Example:
Flip a fair coin twice. Let Y be the number of heads.

- Experiment: Flip a fair coin twice & flip 2 heads
- Sample Space: $S = \{TT, TH, HT, HH\}$
- Random Variable (RV): Y = # of heads
- Support: $y \in \{0, 1, 2\}$.

We, studying probability, need to map support to probability.

Def: Probability Function
A probability for a RV Y, typically denoted $p(y)$ or $f(y)$, is a real-valued function where $f(y) = P(Y=y)$ that the value y is observed for Y.

Lowercase = actual value
Uppercase = RV

For a discrete RV, we write
$f: \{y\} \to (0, 1]$ (exclude 0 b/c Y includes only possible samples). Note: $\{y\}$ is the support, that is output of RV.

Def: Probability Distribution
Set of all possible values (support) & their values

Example:
Extending previous example:
- Probability Dist: $p(0) = 0.25$, $p(1) = 0.5$, $p(2) = 0.25$

Thm: For any discrete probability distribution, $0 \le p(y) \le 1 \ \forall y$ & $\sum_y p(y) = 1$.

Examples:

- Is $p(y) = (\frac{1}{2})^{2y} + \frac{1}{24}$ a valid probability function?

  No b/c we don't know the support.

- What about $p(y) = \frac{0.75^{y-1}}{4}$ $y = \{1, 2, 3, \ldots\}$

  This is clearly b/w 0 & 1 ($0 \le p(y) \le 1 \ \forall y$), but does it sum to 1?

$$\sum_y p(y) = \sum_{y=1}^{\infty} 0.25\,(0.75)^{y-1} = \sum_{y=1}^{\infty} \frac{0.25}{0.75}(0.75)^y = \frac{0.25}{0.75}\sum_{y=1}^{\infty} 0.75^y = \frac{0.25}{0.75}\left(\frac{0.75}{0.25}\right) = 1$$

$$\uparrow \quad \frac{0.75}{1-0.75}$$

It does, so $p(y)$ is a valid probability function.

Def: Expected Value

The expected value is in a sense the "average" of a random variable.

For a discrete RV Y the expected value of Y is defined to be

$E[Y] = \sum_y y\,p(y)$.

Example: Binomial Distribution

Let $p(y) = p(1-p)^{1-y}$ for some $p \in \mathbb{R}$ $0 < p < 1$ where $y = \{1, 2, 3, \ldots\}$

Take this to be a discrete probability function (it is). Find its expected value.

$$E[Y] = \sum_{y=1}^{\infty} y\,p(y)$$

$$= \sum_{y=1}^{\infty} y\,p(1-p)^{y-1}$$

$$= p\sum_{y=1}^{\infty} y(1-p)^{y-1} \leftarrow \text{notice this is the power rule wrt } \tfrac{d}{dp}$$

$$= p\sum_{y=1}^{\infty} \frac{d}{dp}\left(-(1-\,)^y\right) \leftarrow \text{Since this is an absolutely convergent function, we can swap the } \Sigma \text{ \& } \tfrac{d}{dp}$$

$$= -p\frac{d}{dp}\sum_{y=1}^{\infty}(1-p)^y$$

$$= -p\frac{d}{dp}\left(\frac{1-p}{p}\right)$$

$$= -p\left(\frac{(p)(-1) - (1-p)(1)}{p^2}\right)$$

$$= -\left(\frac{\cancel{-p} - 1 \cancel{+p}}{p}\right)$$

$$\boxed{E = \frac{1}{p}}$$

Variance

Def: Variance
Variance is the average squared distance from the mean.

For some RV Y, we denote the variance
$V[Y] = E[(Y-\mu)^2]$

If $p(y)$ is a true distribution, $V[Y] = \sigma_Y^2$.

We care about this b/c we want to calculate variance & other fns on Y

Thm:
Let Y be a discrete RV w/ distribution $p(y)$ & let $g_i(Y)$ be a real-valued function of Y for $i=1,\dots,k$

$$\boxed{E\left[\sum_{i=1}^{k} g_i(Y)\right] = \sum_{i=1}^{k} E[g_i(Y)] = \sum_{i=1}^{k} \sum_y g_i(y)\,p(y)}$$

This is to say the expected value E is a linear operator.

Examples:

- $E[c]$ where c is constant $(g_1(Y) = c)$
  $E[c] = \sum_y c\,p(y) = c\sum_y p(y) = c$
- $E[aY+b]$ where a & b are constants
  $E[aY+b] = E[aY] + E[b] = aE[Y] + b$ ← linear function/operator
  $E[aY+b] = \sum_y (ay+b)p(y) = \sum_y ay\,p(y) + \sum_y bp(y) = a\sum_y yp(y) + b\sum_y p(y)$
  $= aE[Y] + b$
- $E[Y^2 - 2E[Y]Y + E^2[Y]]$
  $g_1(Y) = Y^2,\ g_2(Y) = -2E[Y]Y,\ g_3(Y) = E^2[Y]$

  $E[Y^2 - 2E[Y]Y + E^2[Y]]$ ← $V[Y] = E[(Y-E[Y])^2]$ This is just variance !!
  $= E[Y^2] + E[-2E[Y]Y] + E^2[Y]$
  $= E[Y^2] - 2E[Y]E[Y] + E^2[Y]$
  $= E[Y^2] - E^2[Y]$

Binomial Distribution

Def:
A binomial random variable (RV) occurs when we count certain outcomes in a binomial experiment.

See properties on next page

For an experiment to be a binomial experiment, the following properties must hold

i) A fixed number of n trials ← when $n=1$, it's a Bernoulli experiment

ii) There are two outcomes ("success" & "failure")

iii) Probability of success is constant. $P(S)=p$ & $P(F)=q=1-p$

iv) The trials are independent

If Y counts successes in a binomial experiment $Y \sim bin(n,p)$ where n is the number of trials & p is the probability of success.

Def:
For discrete RVs, the $p(y)$ or $f(y)$ that gives the probability of an event is the probability mass function (PMF)

Let's derive the PMF of a binomial distribution.

We want y successes & $n-y$ failures. That means that there are $C^n_y = \binom{n}{y}$ elements in the experiment w/ n trials that have y successes

Since we have independence, the probability of y successes (A) & $n-y$ failures (B) for a single element is

$$P(A \cap B) = P(A)P(B) = ppqpq \cdots qp = p^y q^{n-y}$$

Since the events are disjoint we can add $P(A\cap B) = p^y q^{n-y}$ $C^n_y = \binom{n}{y}$ times to get the probability of exactly y successes.

Therefore

$$\boxed{Y \sim bin(n,p) \Rightarrow p(y) = \binom{n}{y} p^y q^{n-y} \quad y=0,1,\dots,n \quad 0 \le p \le 1 \quad q=1-p}$$

We can use this to derive the mean $E[Y]$ & variance $V[Y]$.

$$E[Y] = \sum_{y=0}^{n} y p(y) = \sum_{y=0}^{n} y \binom{n}{y} p^y q^{n-y} = \sum_{y=0}^{n} y \left(\frac{n!}{y!(n-y)!} \right) p^y q^{n-y} \leftarrow \text{Hard! Let's simplify}$$

This is hard! Let's simplify. When $y=0$, the inside is 0 so

$$E[Y] = \sum_{y=1}^{n} y \left(\frac{n!}{y!(n-y)!} \right) p^y q^{y-n} = \sum_{y=1}^{n} \frac{n!}{(y-1)!(n-y)!} p^y q^{n-y} \leftarrow \text{cancel } y$$

$$= np \sum_{y=1}^{n} \frac{n!}{(y-1)!(n-y)!} p^{y-1} q^{n-1}$$

Now let $x = y-1$. This gets us further

$$E[Y] = np \sum_{x=0}^{n-1} \frac{n!}{x!(n-1-x)!} p^x q^{n-1-x}$$

Now, this equation is the PMF / kernel of $Bin(n-1, p)$, summed over its possible inputs. We know the PMF must add to 1, so

$$E[Y] = np \sum_{x=0}^{n-1} \frac{n!}{x!(n-1-x)!} p^x q^{n-1-x} = np \cdot 1 = np$$

$\therefore \boxed{E[Y] = np \text{ for a binomial distribution.}}$

Now we do the variance of the binomial distribution.

$$V[Y] = E[Y^2] - E^2[Y]$$

$E^2[Y]$ is easy to calculate, so we find $E[Y^2]$ independently

$$E[Y^2] = E[Y^2 - Y + Y] = E[Y(Y-1)] + E[Y]$$
$$= \sum_{y=0}^{n} \left(y(y-1) \cdot \binom{n}{y} p^y q^{n-y} \right) + np$$

When $y=0$ or $y=1$, inner $=0$

$$= \sum_{y=2}^{n} \left(y(y-1) \cdot \frac{n!}{y!(n-y)!} p^y q^{n-y} \right) + np$$
$$= n(n-1)p^2 \sum_{y=2}^{n} \left(\frac{n-2!}{(y-2)!(n-y)!} p^{y-2} q^{n-y} \right) + np$$

$\leftarrow$ cancel $y(y-1)$ & pull out $n(n-1)$ from fact & pull out p^2 from p^y

Now we do a similar substitution showing the sum's inside is the binomial kernel, so sum(...) = 1

$$E[Y^2] = n(n-1)p^2 + np$$

Now back to $V[Y]$

$$V[Y] = E[Y^2] - E^2[Y] = n(n-1)p^2 + np - (np)^2 = \cancel{n^2p^2} - np^2 + np - \cancel{n^2p^2}$$
$$= np - np^2$$
$$= np(1-p)$$

$\therefore \boxed{V[Y] = npq \text{ for a binomial distribution.}}$

Geometric

Now, let's consider the bernoulli experiment / trial (binomial w/ $n=1$), when trials is not fixed.

<u>Def:</u>

A <u>geometric RV</u> occurs when we count the number of Bernoulli trials required to reach a success, so the number of trials is not fixed.

Specifically, let Y be a RV which counts the number of trials until the first success occurs. Then

i) Must observe $y-1$ failures first, then 1 success

ii) Only 1 arrangement is possible.

iii) By independence $FF\cdots FS = qq\cdots qp$

Technically a probability function should be defined on all $y \in \mathbb{R}$. To "fix" this we add a "0 otherwise" clause if $y=1,2,\ldots$ doesn't hold.

We say $\boxed{Y \sim geo(p) \Rightarrow p(y) = q^{y-1}p : \text{ where } y=1,2,\ldots,\ 0 \le p \le 1.}$

For a geometric RV $Y \sim geo(p)$

$$\boxed{E[Y] = 1/p \quad \& \quad V[Y] = \frac{q}{p^2}}$$

Proof follows similar form to binomial except w/ geometric PMF

Negative Binomial

What if we swap fixed # of trials w/ fixed # successes?

Bernoulli (1 trial) $\leftrightarrow$ Geometric (1 success)

Binomial (n trials) $\leftrightarrow$ Negative Binomial (r successes)

For geometric, we only wanted 1 success. The negative binomial is a generalization where we want r success.

If we want y trials w/ r'th success on yth trial, that means we want

i) $r-1$ successes in first $y-1$ trials, so there are $\binom{y-1}{r-1}$ ways to do this

This gives us

need at least r trials for r successes

$$\boxed{p(y) = \binom{y-1}{r-1} p^{r-1} q^{y-r} p = \binom{y-1}{r-1} p^r q^{y-r} \quad \text{where } 0 \le p \le 1 \ \& \ y = r, r+1, r+2, \ldots}$$

For a negative binomial RV $Y \sim nb(r,p)$, then

$$\boxed{E[Y] = \frac{r}{p} \quad \& \quad V[Y] = \frac{rq}{p^2}}$$

Hyper Geometric Distribution

What if we sample from a finite population w/o replacement?
This is important when the population you're using is really small.
(e.g. sampling a pop. of 10,000 w/ 5 people sick).

Let us define a new experiment where
the pop. has a finite size N,
the pop. is partioned into two groups A & B, (of interest: A)
group A has r elements,
we sample w/ size n, &
Y counts the number of type A elements in the sample.

there can be multiple groups, just one has to be of interest

- In this case, Y follows a hypergeometric distribution
  $Y \sim hg(N, r, n)$

To work out the PMF of Y, we use the sample point method
- random selection $\Rightarrow$ equiprobability $\Rightarrow p(y) = \frac{|A|}{|S|} = A = \{Y = y\}$
- y type A out of r possible type $A \Rightarrow \binom{r}{y}$ possibilities
- $n-y$ type B out of $N-r$ possible type $B \Rightarrow \binom{N-r}{n-y}$ possibilities
- mn rule $\Rightarrow \binom{r}{y}\binom{N-r}{n-y}$ total ways to get $Y=y$
- # ways to sample n items from $N \Rightarrow \binom{N}{n}$ ways

By all this, $\boxed{p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} \text{ where } Y \sim hg(N,r,n),\ 0 \le y \le n \le N,\ y \le r \le N,\ N = 1, 2, \ldots}$

can't see more than sample size or number possible

For $Y \sim hg(N, r, n)$
$$\boxed{E[Y] = \frac{nr}{N} \quad \& \quad V[Y] = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)}$$

like p like q finite population correction

Note: If $p = r/N$ constant then
$$\lim_{N \to \infty} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{y}} = \binom{n}{y} p^y (1-p)^{n-y}$$

In a sense, hypergeometric is a finite binomial distribution

Poisson Distribution

What if we counted per unit? So what about car crashes per week, not overall, or defects per component.

The fundamental idea is we can make the unit/scale so small that only 0 or 1 events can happen, never multiple.

If we do this then the experimental conditions are
- there are two outcomes (success & failure)
- constant probability,
- independent events,
- unknown n (probably large b/c want tiny samples)

Since as p goes down, n goes up, we define $\lambda = pn$ b/c that stays constant.

The above conditions define a binomial distribution.

$$p(y) = \binom{n}{y} p^y q^{n-y}$$

We take the $\lim_{n\to\infty}$ b/c we need very large sample size to acheive binomial properties.

cancelled $(n-y)!$

$$\lim_{n\to\infty} p(y) = \lim_{n\to\infty} \binom{n}{y} p^y q^{n-y} = \lim_{n\to\infty} \frac{n!}{y!(n-y)!} p^y q^{n-y} = \lim_{n\to\infty} \frac{n(n-1)\cdots(n-y+1)}{y!} p^y q^{n-y}$$

$$= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-y+1)}{y!} \left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y}$$

should've folded earlier

$$= \lim_{n\to\infty} \frac{\lambda^y}{y!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-y} \frac{n(n-1)\cdots(n-y+1)}{n^y}$$

move things w/o n's to left

$$= \frac{\lambda^y}{y!} \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-y} (1)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{y-1}{n}\right)$$

there are y terms in the numerator so we distribute n^{-y} (or $1/n^y$)

$$= \frac{\lambda^y}{y!} \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-y} (1)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{y-1}{n}\right)$$

$$= \frac{\lambda^y}{y!} e^{-\lambda} (1)$$

all limits are 1. ü

Recall, product of limits is limit of products if all limits exist

$$\therefore \lim_{n\to\infty} p(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Thus given $Y \sim Poi(\lambda)$

$$\boxed{\begin{array}{l} p(y) = \dfrac{e^{-\lambda}\lambda^y}{y!} \quad \text{where } \lambda > 0 \text{ \& } y = 0, 1, \ldots \\ E[Y] = \lambda \quad V[Y] = \lambda \end{array}}$$

Example:

A tree has seedlings dispersed randomly w/ a mean 5 seedlings/m^2
Let Y count the number of seedlings found in a single 1 m^2 plot.

Here, $Y \sim Poi(\lambda)$ where $\lambda = 5$ saplings/m^2

a) What is the probability a forester finds no seedlings in a 1 m^2?

$$p(0) = \frac{e^{-\lambda}\lambda^{y}}{y!} = \frac{e^{-5}\cancel{5^0}}{\cancel{0!}} = 0.0067$$

b) What is the mean & standard deviation?

$$\mu = E[Y] = 5$$

$$\sigma = \sqrt{V[Y]} = \sqrt{5} = \underline{2.2361}$$

c) What is the probability at most 3 saplings in 1 m^2?

$$P(Y \leq 3) = p(0) + p(1) + p(2) + p(3) = 0.2650$$

d) Probability of at most 3 given at least one?

$A = Y \leq 3 \Rightarrow P(A) = p(0) + p(1) + p(2) + p(3)$
$B = Y \geq 1 \Rightarrow P(B) = 1 - p(0)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{p(1) + p(2) + p(3)}{1 - p(0)} = 0.2600$$

e) Probability forester inspects 10 1m^2 plots & finds exactly 1 seedling in 3 plots?

This is a binomial where $n = 10$ & $p = p_Y(1) = 0.0337$,
so $Y' \sim bin(10, 0.0337)$, so
$p_{Y'}(3)$ for y' is
$p_{Y'}(3) = \underline{0.0036}$

subscripts are good!

f) Probability forester finds 35 seedlings in 8 m^2 area?

This is a new Poisson distribution $W \sim Poi(8\lambda) = Poi(40)$

$$p_W(35) = \frac{e^{-40}40^{35}}{35!} = \underline{0.0485}$$

don't do $p(35) \cdot 8$ that won't work b/c of independence

In the above example, part (d) is a truncated distribution (changing support). Part (e) is a hierarchical distribution (two different distributions where one depends on other)

Continuous Random Variables

A *continuous random variable* (RV) has an uncountably infinite support (set of possible values).

Def: CDF

Let Y be a RV. ← All distributions (even discrete) have CDFs.

The *cumulative distribution function* (CDF) F is

$F(y) = P(Y \leq y) \quad \forall y \in \mathbb{R}$

That is the probability of getting this value or something lower.

Prop:

For any CDF F,

i) $F(-\infty) = 0$

ii) $F(\infty) = 1$

iii) $F(y_2) \geq F(y_1) \Leftrightarrow y_2 \geq y_1$ ← F is non-decreasing

iv) $\lim_{y \to y_0^+} F(y) = F(y_0)$ ← right continuous

Example: Discrete CDF

Let T be the RV that measures the amount of time to clean up sites w/ Atrazine & Perchlorate contamination. The probability mass function (PMF) is given by

t	$p(t)$
0	0.10
8	0.65
12	0.25

Discrete RV must have a step-function CDF

The CDF F is thus

$$F(t) = \begin{cases} 0 & t < 0 \\ 0.10 & 0 \leq t < 8 \\ 0.75 & 8 \leq t < 12 \\ 1.00 & 12 \leq t \end{cases}$$

A RV is continuous iff its CDF F is continuous on $\mathbb{R}$ & its derivative $\frac{dF(y)}{y}$ has a finite # of discontinuities on any finite interval.

Def:

Let Y be a continuous RV w/ CDF F.

It's *probability density function* (PDF) f is

$f(y) = \frac{dF(y)}{dy} = F'(y)$ ← CDF is integral of PDF

Rem:

i) $f(y) = F'(y) \Rightarrow F(y) = \int_{-\infty}^{y} f(t)\,dt$

ii) $F(y_0) = \int_{-\infty}^{y_0} f(t)\,dt = P(Y \leq y_0)$ (i.e. probability = area)

This is why we say density for continuous RV & mass for discrete RV.

Thm:

For all PDFs f,

i) $f(y) \geq 0 \quad \forall y$

ii) $\int_{-\infty}^{\infty} f(y)\,dy = 1$ ← or really integrate over support

Example:

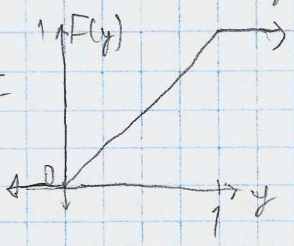
Let F be a PDF for RV Y where

$$F(y) = \begin{cases} 0 & y < 0 \\ y & 0 \leq y < 1 \\ 1 & 1 < y \end{cases}$$

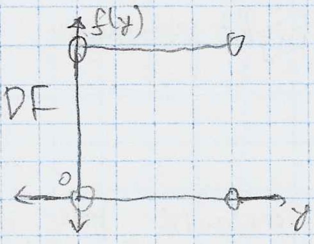
Thus the PMF f is

$$f(y) = \begin{cases} 0 & y < 0 \\ 1 & 0 < y < 1 \\ 0 & 1 < y \end{cases}$$ notice discontinuities!

CDF



PDF



Example:

Let $f(y) = k e^{-y/5}$ for $y \in (0, \infty)$.

What value of k makes f a valid PDF?

$\int_0^\infty f(y)\,dy = \int_0^\infty k e^{-y/5}\,dy = 1$ total prob. must be 1

$= [-5k e^{-y/5}]_0^\infty$

$= 0 - (-5k)$

$= 5k = 1$

$\boxed{k = 1/5}$

What is $F(y)$ given f is PDF?

$F(y) = \int_0^y f(t)\,dt = \int_0^y \frac{1}{5} e^{-t/5}\,dt$

$= [-e^{-t/5}]_0^y$

$\boxed{F(y) = -e^{-y/5} + 1}$

$f(y)$ is valid for $\forall y \in (0, \infty)$ (given) & $F(y)$ is valid $\forall y \in \mathbb{R}$ (req. of CDF)

Example:

Find $F(y)$ if $f(y)$

$$f(y)=\begin{cases} 3y^2 & 0\le y\le 1/2 \\ y+1 & 1/2<y\le 1 \\ 0 & o/w \end{cases}$$

To find $F(y)$ we do piecewise integration

$$F(y)=\int_{-\infty}^{t} f(t)\,dt$$

$$=\begin{cases} 0 \quad \text{It's cumulative} & [-\infty,0) \\ \int_0^y 3t^2\,dt & [0,1/2] \\ \int_{1/2}^{y} t+1\,dt+\int_0^{1/2} 3t^2\,dt & (1/2,1] \\ 1 & (1,\infty) \end{cases}$$

also $\int_{1/2}^{1} t+1+\int_0^{1/2} 3t^2\,dt$

$$=\begin{cases} 0 & (-\infty,0) \\ y^3 & [0,1/2] \\ [1/2t^2+t]_{1/2}^{y}+(1/2)^3 & (1/2,1] \\ 1 & [1,\infty) \end{cases}$$

$$=\begin{cases} 0 \\ y^3 \\ 1/2y^2+y+(1/2)^3+1/2+(1/2)^3=1/2y^2+y+3/4 \\ 1 \end{cases}$$

Thm:

Let Y be continuous RV w/ PDF f & CDF F

$$P(a\le Y\le b)=\int_0^b f(y)\,dy=F(b)-F(a) \quad a\le b$$

by 1st fundamental theorem of calculus

Note: $P(a\le Y\le b)=P(a<Y<b)$ b/c $P(y)=0\ \forall y\in\mathbb{R}$.

Uniform Distribution

Def:

Let $\theta_1,\theta_2\in\mathbb{R}$ where $\theta_1<\theta_2$. This gives us a <u>uniform distribution</u>

$Y\sim u(\theta_1,\theta_2)$

which has a PDF

$$f(y)=\begin{cases} \frac{1}{\theta_2-\theta_1} & \theta_1\le y\le\theta_2 \\ 0 & o/w \end{cases}$$

Thm:

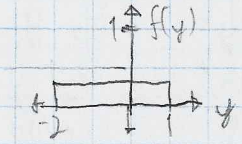
For RV $Y\sim u(\theta_1,\theta_2)$

$$E[Y]=\frac{\theta_1+\theta_2}{2} \qquad V[Y]=\frac{(\theta_2-\theta_1)^2}{12}$$

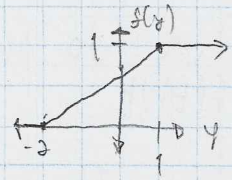
Example:

Consider $Y \sim u(-2, 1)$.

$$\underset{\text{PDF}}{f(y)} = \begin{cases} \frac{1}{1-(-2)} = 1/3 & -2 \le y \le 1 \\ 0 & \text{o/w} \end{cases}$$



$$\underset{\text{CDF}}{F(y)} = \begin{cases} 0 & y < -2 \\ \int_{-2}^{y} \frac{1}{3} dt = y/3 + 2/3 & -2 \le y \le 1 \\ 1 & 1 \le y \end{cases}$$



$P(0 \le Y \le 0.5) = P(-1 \le Y - 0.5) = P(\{-1.75 \le Y \le -1.5\} \cup \{0.25 \le Y \le 0.5\})$

$E[Y] = \frac{-2+1}{2} = -1/2$

$V[Y] = \frac{(1-(-2))^2}{12} = \frac{9}{12} = 3/4$

Normal Distribution

This is the classic "bell curve" (symmetric & unimodal [one peak]) & the input is $\mathbb{R}$.

Def:

Let $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$. This gives us the normal distribution $Y \sim n(\mu, \sigma^2)$

w/ PDF

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} \quad \forall y \in \mathbb{R}$$

We call μ & the σ^2 the mean & variance ← we can't prove b/c $\int f$ isn't possible

$E[Y] = \mu$ & $V[Y] = \sigma^2$

If you simplify (get the kernel), the PDF is e^{-y^2} which we can not integrate. This means we can't show $E[Y] = \mu$ & $V[Y] = \sigma^2$ (easily) nor can we show it is a real probability distribution (easily).

The reason the normal distribution is so popular is that it can be standardized. Standardizing means you can use one particular set of parameters to represent the entire family of distributions. This gives us z-scores.

Def:

Let $Y \sim n(\mu, \sigma^2)$ be a normal distribution. We standardize this into the standard normal distribution Z

$$Z = \frac{Y - \mu}{\sigma^2} \sim n(0, 1)$$

We call Z the z-scores. They can be thought of as unitless values or as σ units.

Standardization made computations easier historically w/ tables & allows for easy comparison of different distributions.

Let's prove that the normal distribution is actually a PDF.

Pf:

We first show $f(y) > 0 \ \forall y$ in the support.
This is clearly true b/c the kernel
$$e^{-y^2} > 0 \quad \forall y$$
& the coefficient to the kernel
$$\frac{1}{\sqrt{2\pi\sigma^2}} > 0 \quad \text{b/c } \sigma^2 > 0.$$

To show that $\int_{\mathbb{R}} f(y) = 1$ is much more difficult. We will do this in steps.
$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy$$
$$= \sqrt{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx}$$
b/c $\forall a \in \mathbb{R}_+ \quad a = \sqrt{aa}$
$$= \sqrt{\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{2\pi\sigma^2} e^{\frac{-(y-\mu)^2}{2\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx\, dy}$$
b/c all products of integrals can be written as a double integral of products (given everything's continuous).

Now we use u-substitution to standardize x & y, that is make $X, Y \sim n(0,1)$, to make the integral simpler. ← other books don't do this

Let $w_1 = \frac{y-\mu}{\sqrt{2}\sigma}$ & $w_2 = \frac{x-\mu}{\sqrt{2}\sigma} \Rightarrow dw_1 = \frac{dy}{\sqrt{2}\sigma}$ & $dw_2 = \frac{dx}{\sqrt{2}\sigma}$

Continuing earlier work
$$= \sqrt{\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{\pi} e^{-w_1^2} e^{-w_2^2} dw_1\, dw_2}$$
$$= \sqrt{\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{\pi} e^{-(w_1^2 + w_2^2)} dw_1\, dw_2}$$

Now we change the variables to convert (w_1, w_2) to polar coordinates (r, θ) where
$$w_1 = r\cos(\theta) \quad \& \quad w_2 = r\sin(\theta)$$

Recall that multivariate change of variables requires the Jacobian of transformation.

$$J = \begin{vmatrix} \frac{dw_1}{dr} & \frac{dw_1}{d\theta} \\ \frac{dw_2}{dr} & \frac{dw_2}{d\theta} \end{vmatrix} = \begin{vmatrix} \cos(\theta)dr & -r\sin(\theta)d\theta \\ \sin(\theta)dr & r\cos(\theta)\,d\theta \end{vmatrix}$$

$$= r\cos^2(\theta)\,dr\,d\theta + r\sin^2(\theta)\,dr\,d\theta$$
$$= r\,dr\,d\theta$$

Thus we have ~~$w_1^2 +$~~

$$w_1^2 + w_2^2 = r^2\cos^2(\theta) + r^2\sin^2(\theta) = r^2$$

Now we actually do our polar coordinate substitution continuing our earlier work

$$\sqrt{\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{\pi} e^{-(w_1^2+w_2^2)}\,dw_1\,dw_2}$$
$$= \sqrt{\int_0^{2\pi}\int_0^{\infty} \frac{1}{\pi} e^{-r^2}\,r\,dr\,d\theta}$$
$$= \sqrt{\int_0^{2\pi}\left[\frac{-1}{2\pi}e^{-r^2}\right]_{r=0}^{r=\infty} d\theta}$$
$$= \sqrt{\int_0^{2\pi}\frac{1}{2\pi}\,d\theta}$$
$$= \sqrt{\frac{2\pi}{2\pi} - \frac{0}{2\pi}}$$
$$= 1$$

Thus the PDF for the normal distribution is valid. $\square$

Gamma Distribution & Friends

The gamma distribution is a common skewed distribution, normally used for lifetimes.

<u>Def:</u>

Let $\alpha, \beta \in \mathbb{R}_+$. The <u>gamma distribution</u> is

$Y \sim \Gamma(\alpha, \beta)$

w/ PDF

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \quad y \in \mathbb{R}_0$$

We call α the <u>shape</u> & β the <u>scale</u>.

this is factorial extended!

$\Gamma(\alpha) = (\alpha - 1)!$

$\alpha \in \mathbb{Z}$ $\alpha \geq 0$

What is the gamma function, $\Gamma(\alpha)$? The <u>gamma function</u> is

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\,dy \qquad \text{Note: } \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

This gamma function <u>normalizes</u> the gamma distribution (makes sure the PDF integrates to 1)

For $Y \sim \Gamma(\alpha, \beta)$

$E[Y] = \alpha\beta \qquad V[Y] = \alpha\beta^2.$

The gamma distribution cannot be standardized.

With the gamma distribution, we'll practice distribution kernel recognition.

Let's prove $E[Y] = \alpha\beta$.

$$E[Y] = \int_{\mathbb{R}} y\, f(y)\,dy$$
$$= \int_0^\infty y\left(\frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}\right)dy$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty y^\alpha e^{-y/\beta}\,dy \leftarrow \text{This is the } \Gamma(\alpha+1,\beta) \text{ kernel!}$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha}\cdot \Gamma(\alpha+1)\beta^{\alpha+1}\int_0^\infty \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} y^\alpha e^{-y/\beta}\,dy \leftarrow \text{this is integral of PDF of } \Gamma(\alpha+1,\beta)\text{, so it is 1}$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha}\cdot \Gamma(\alpha+1)\beta^{\alpha+1}(1)$$
$$= \frac{\alpha\,\cancel{\Gamma(\alpha)}}{\cancel{\Gamma(\alpha)}}\cdot\beta$$
$$= \alpha\beta$$

Def:

The exponential distribution is the simplest special case of the gamma.

$$Y \sim \exp(\beta) \equiv \Gamma(1,\beta)$$

Commonly used for waiting time

w/ PDF similar to $\Gamma(\alpha,\beta)$

$$f(y) = \frac{1}{\beta} e^{-y/\beta} \text{ where } \beta > 0 \text{ \& } y \geq 0.$$

$$E[Y] = \beta \quad \& \quad V[Y] = \beta^2$$

Other books use λ instead of β where β is the mean & λ is the rate. $\beta = 1/\lambda$.

Examples:

The time you have to wait (in minutes) to check out at the grocery store is

$$Y \sim \exp(10).$$

What's the probability of waiting less than 5 minutes? More than 20?

$$P(Y \leq y) = F(y) = \int_0^y \frac{1}{10} e^{-t/10}\,dt = \left[-e^{-t/10}\right]_0^y = -e^{-y/10} + 1$$

$$P(Y \leq 5) = 1 - e^{-1/2} = 0.3935$$
$$P(Y \geq 20) = 1 - P(Y \leq 20) = 1 - (1 - e^{-2}) = e^{-2} = 0.1353$$

What's the mean time? $E[Y]=10$
Variance? $V[Y]=10^2=100$

Def:
Chi-Squared (χ^2) distributions have 1 variable, the degrees of freedom $v\in\mathbb{Z}_+$
$$Y\sim\chi^2(v)\equiv\Gamma(v/2,2)$$
w/ PDF
$$f(y)=\frac{1}{\Gamma(v/2)\,2^{v/2}}\,y^{v/2-1}e^{-y/2},\quad v\in\mathbb{Z}_+,\ y\geq 0$$
& mean & variance
$$E[Y]=v \quad \& \quad V[Y]=2v$$

Beta Distribution

Unlike the earlier distributions, this has bounded support $[0,1]$. This is useful for modelling probabilities. It can be skew left, skew right, & uniform.

Def:
Let $\alpha,\beta\in\mathbb{R}_+$. The beta distribution is
$$Y\sim beta(\alpha,\beta)$$
w/ PDF
$$f(y)=\frac{1}{B(\alpha,\beta)}y^{\alpha-1}(1-y)^{\beta-1}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$$

The beta function $B(\alpha,\beta)$ again normalizes where
$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Here α controls skew towards 1 & β skew towards 0. They are both shape params.

The mean & variance is
$$E[\alpha]=\frac{\alpha}{\alpha+\beta} \quad \& \quad V[Y]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Moment Generating Functions

i.e. in standard probability theory
Normally, to determine if 2 RV's are equal, we compare their CDFs, that's
$$F(x)=F(y)\iff X\equiv Y$$

How do we determine this for close-ness or cases where we don't know the exact CDF.

To do this, we compare mean, variance, skewness, etc. These are all moments of the distribution

Def:
A moment is a way to characterize the shape of the distribution.
~~In this~~
In this class, we'll talk about the following

i) k^{th} raw moment of Y is $\mu'_k=E[Y^k]$

ii) k^{th} central moment of Y is $\mu_k=E[(Y-\mu)^k]$

iii) ~~k^{th} standardized moment~~ of Y is $\frac{\mu_k}{\sigma_k}$ ← Centered (μ_k) & standardized ($1/\sigma_k$)

Example:

For any RV, find μ_1 & μ_2'

$\mu_1 = E[Y-\mu] = E[Y]-\mu = \mu-\mu = 0$ ← always true

$\mu_2' = E[Y^2] = V[Y] + E^2[Y] = \mu_2 + (\mu_1')^2$

↑

We can write other moments in terms of other moments.
In particular, we can write central moments in terms of raw moments.

Named Moments

i) Mean: $\mu = E[Y]$
Shape Characteristic: Central tendency
Moment: 1st raw moment μ_1'

ii) Variance: $\sigma^2 = V[Y] = E[(Y-\mu)^2]$
Shape Characteristic: Spread
Moment: 2nd central moment μ_2

iii) Skewness: $\gamma = E[(Y-\mu)^3]/\sigma^3$
Shape Characteristic: Symmetry
Moment: 3rd standardized moment μ_3/σ^3

iv) Kurtosis: $\kappa = E[(Y-\mu)^4]/\sigma^4$ (kappa)
Shape Characteristic: tail heaviness
Moment: 4th standardized μ_4/σ^4

Normal distribution has $\kappa=3$.
Excess kurtosis is $\kappa-3$.
This is prob. of outliers/ how "heavy" your tails are.

Applications:

By matching (raw) moments, we can determine equality.
That is $\mu'_{X_k} = \mu'_{Y_k}$ $\forall k \in \mathbb{N} \Leftrightarrow X \equiv Y$.

Estimation using method of moments.

Transformations on RVs.

As you can see from applications, we want an easy way to calculate any moment.

Def: (MGF)

A moment generating function of Y is $m_Y(t) = E[e^{tY}]$ — We take derivatives to find moments!

For our purposes, m_Y must exist around $t=0$, that is $\exists\, \epsilon > 0$ s.t. $m(t) < \infty$ for $|t| \le \epsilon$. ← b/c we want to evaluate @ 0

We care about MGFs b/c they make finding moments easy.

Thm:

If $m(t)$ exists, then for any $k \in \mathbb{N}$

$$\mu_k' = \frac{d^k}{dt^k}(m(t))\Big|_{t=0} = m^{(k)}(0)$$

This gives us an easy way to find raw moments, which makes every other moment easy b/c all moments can be written in terms of raw moments μ_k'

Example:

Find the MGF $m(t)$ for $Y \sim bin(n,p)$ & use it to find $E[Y]$ & $V[Y]$.

$$M(t) = E[e^{tY}]$$
$$= \sum_{y=0}^{n} e^{ty}\binom{n}{y}p^y q^{n-y}$$

bin(n,p)

$$= \sum_{y=0}^{n}\binom{n}{y}(pe^t)^y q^{n-y}$$

$\boxed{M(t) = (pe^t + q)^n}$ by the binomial expansion theorem

Cool! That wasn't too bad.

$$E[Y] = \mu_1' = m^{(1)}(0) = \frac{d}{dt}(pe^t + q) = n(pe^t+q)^{n-1}(pe^t)\Big|_{t=0} = \boxed{np}$$

$$V[Y] = \mu_2' - (\mu_1')^2$$

$$\mu_2' = m^{(2)}(0) = \frac{d^2}{dt^2}(pe^t+q)^n\Big|_{t=0}$$
$$= \frac{d}{dt}\Big[n(pe^t+q)^{n-1}(pe^t)\Big]_{t=0}$$
$$= \Big[n(n-1)(pe^t+q)^{n-2}(pe^t)^2 + n(pe^t+q^{n-1})(pe^t)\Big]_{t=0}$$
$$= n(n-1)(p^2) + np)$$

$$= n(n-1)p^2 + np - (np)^2 = np(1-p) = \boxed{npq}$$

Example:

Identify the distribution for the following RVs.

$m_X(t) = (0.27e^t + 0.73)^3 \iff X \sim bin(3, 0.27)$

$m_Y(t) = \frac{e^{2t} + 2e^t + 1}{4} = \frac{(e^t+1)^2}{2^2} = (\frac{1}{2}e^t + \frac{1}{2})^2 \iff X \sim bin(2, \frac{1}{2})$

& we showed

You've probably learned the form of the binomial distribution's MGF. There's a table in the book for the others.

Example:

Suppose $Y \sim \Gamma(\alpha, \beta)$. Find $m_Y(t)$ & use it to find $E[Y]$ & $V[Y]$.

$$m_Y(t) = E[e^{tY}]$$
$$= \int_0^\infty e^{ty} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}\, dy$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty y^{\alpha-1} e^{-y/\beta + ty}\, dy$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty y^{\alpha-1} e^{-y(1-t\beta)/\beta}\, dy \leftarrow \text{almost kernel of gamma distribution}$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot \frac{(\beta^*)^\alpha}{(\beta^*)^\alpha} \int_0^\infty y^{\alpha-1} e^{-y/\beta^*}\, dy \quad \text{where } \beta^* = \frac{\beta}{1-t\beta}$$
$$= \frac{(\beta^*)^\alpha}{\beta^\alpha} \cancel{\int_0^\infty \frac{1}{\Gamma(\alpha)(\beta^*)^\alpha} y^{\alpha-1} e^{-y/\beta^*}\, dy} \leftarrow \text{kernel of gamma distribution}$$
$$= \frac{(\beta/(1-t\beta))^\alpha}{\beta^\alpha}$$
$$= (1-t\beta)^{-\alpha} \quad \text{for } t < 1/\beta$$

$$E[Y] = \mu_1' = m_Y^{(1)}(0) = \frac{d}{dt}(1-t\beta)^{-\alpha}\Big|_{t=0} =$$
$$= -\alpha(1-t\beta)^{-\alpha-1}(-\beta)\big|_{t=0}$$
$$= \alpha\beta(1-t\beta)^{-\alpha-1}\big|_{t=0}$$
$$= \boxed{\alpha\beta}$$

$$V[Y] = E[Y^2] - E^2[Y] = \mu_2' - (\mu_1')^2$$
$$\mu_2' = m_Y^{(2)}(0) = \frac{d^2}{dt^2}(1-t\beta)^{-\alpha}$$
$$= \frac{d}{dt}\alpha\beta(1-t\beta)^{-\alpha-1}\Big|_{t=0}$$
$$= \alpha\beta(-\alpha-1)(1-t\beta)^{-\alpha-2}(-\beta)\big|_{t=0}$$
$$= \alpha\beta^2(\alpha+1)(1-t\beta)^{-\alpha-2}\big|_{t=0}$$

$= \alpha^2\beta^2 + \alpha\beta^2$
$= \alpha^2\beta^2 + \alpha\beta^2 - \alpha^2\beta^2 = \boxed{\alpha\beta^2}$

Compare CDF & MGF:

CDF $F(y)$ & MGF $m(t)$ have some similarities

i) $F(y)$ gives all probabilities. $m(y)$ gives all moments.

ii) Both uniquely identify a RV

Thm:

If Y is a RV w/ PDF $f(y)$ & $g(Y)$ is a function of Y, then

$$m_{g(Y)}(t) = E[e^{g(Y)t}] = \int_{\mathbb{R}} e^{tg(y)} f(y)\,dy$$

This is particularly useful if g is a linear transformation like $X = a + bY$ for $a, b \in \mathbb{R}$ $b \neq 0$. In that case

$$m_X(t) = e^{at} m_Y(bt) \longleftarrow \text{eat my butt}$$

Example:

If $Y \sim n(\mu, \sigma^2)$ find the distribution $Z = \frac{Y-\mu}{\sigma}$

The MGF of normal distribution is
$m_Y(t) = e^{\mu t + \sigma^2 t/2}$

Note that

$$Z = \frac{-\mu}{\sigma} + \frac{1}{\sigma} Y$$

By the above theorem

$$m_Z = e^{-\mu t/\sigma} e^{\mu t/\sigma + t^2/2} = \underline{e^{t^2/2}}$$

This is the normal MGF w/ $\mu = 0$ & $\sigma = 1$, so $Z \sim n(0,1)$.
Thus, we can standardize $Y \sim n(\mu, \sigma^2)$

Multivariate Probability Distributions (Chp 5)

Often when doing real sampling we have multiple random variables. For example measuring a response variable on n subjects yields n unique RVs. Measuring multiple subjects repeatedly over a week is multiple different RVs where some are independent & some aren't.

In general we break this into

i) Same random variable (RV) on different observational unit (OU).

ii) Different RV on same OU.

For simplicity we start w/ 2 variable or bivariate cases.

Def:

Let Y_1 & Y_2 be discrete RVs.
The joint probability function or joint PMF is

$$p(y_1, y_2) = P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})$$

This joint PMF p has the properties

i) $p(y_1, y_2) \geq 0 \quad \forall y_1 \in Y_1, \ y_2 \in Y_2$

ii) $\sum_{y_1, y_2} p(y_1, y_2) = 1$

Def:

The joint CDF likewise is

$$F(y_1, y_2) = P(\{Y_1 \leq y_1\} \cap \{Y_2 \leq y_2\})$$

Example:

Consider a plant producing cars w/ 3 production lines.
Two cars are selected from the plan randomly.
Let Y_1 & Y_2 count the number of cars coming from line 1 & 2 respectively.

$$S = \left\{ \begin{matrix} (A,A), & (B,A), & (C,A), \\ (A,B), & (B,B), & (C,B), \\ (A,C), & (B,C), & (C,C) \end{matrix} \right\} \leftarrow \text{Sample space}$$

$$(y_1, y_2) \in \left\{ \begin{matrix} (2,0), & (1,1), & (1,0) \\ (1,1), & (0,2), & (0,1), \\ (1,0), & (0,1), & (0,0) \end{matrix} \right\} \leftarrow \text{Support} \ \left(\begin{matrix}\text{should technically} \\ \text{remove duplicates}\end{matrix}\right)$$

Now we have the sample space & support. We find the PMF by counting instances in the (repetitive) support.

		y_2		
		0	1	2
y_1	0	1/9	2/9	1/9
	1	2/9	2/9	0
	2	1/9	0	0

$p(1,2) = 0$

$F(1,2) = P(\{Y_1 \le 1\} \cap \{Y_2 \le 2\}) = \sum_{y_1=0}^{1} \sum_{y_2=0}^{2} = 1/9 + 2/9 + 2/9 + 2/9 + 1/9 + 0 = 8/9$

$F(-2,1) = 0$

Def:

Let Y_1 & Y_2 be continuous RV w/ joint CDF $F(y_1, y_2)$. Note: This might not have a closed form.

The joint PDF $f(y_1, y_2)$ is a function such that

$$F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f(t_1, t_2)\, dt_2\, dt_1 \quad y_1, y_2 \in \mathbb{R}$$

Likewise (given partial derivatives exist)

$$f(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} F(y_1, y_2)$$

i) $f(y_1, y_2) \ge 0 \quad \forall y_1, y_2$

ii) $\int_{\mathbb{R}} \int_{\mathbb{R}} f(y_1, y_2)\, dy_2\, dy_1 = 1$

Thm:

Let X & Y be RVs w/ joint CDF $F(x,y)$. Then

i) $F(-\infty, -\infty) = F(-\infty, y) = F(x, -\infty) = 0$

ii) $F(\infty, \infty) = 1$

iii) If $x_1 \le x_2$ & $y_1 \le y_2$ then

$$F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \ge 0$$

That is, volume (not area) is non-decreasing.

Example:

Let $f(y_1, y_2) = k$ for $y_1 \in [0,1]$ $y_2 \in [0, 1/2]$. Find k that makes this a valid PDF.

$1 = \int_0^1 \int_0^{1/2} k\, dy_2\, dy_1$

$= \int_0^1 1/2 k\, dy_1$

$= 1/2 k$

$\Rightarrow \underline{k = 2}$

Find $P(Y_1 \le 0.5, Y_2 \le 0.25)$

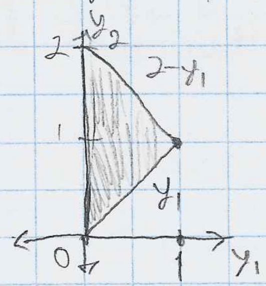
$F(0.5, 0.25) = \int_0^{0.5} \int_0^{0.25} 2\, dy_2\, dy_1 = \int_0^{0.5} 1/2\, dy_1 = 1/4$

Example:

Let $f(y_1, y_2) = 6y_1^2 y_2$ for $0 \le y_1 \le y_2$ & $y_1 + y_2 \le 2$.
Show f is a valid density & find $P(Y_1 + Y_2 < 1)$.

First let's draw our support. We have

$y_2 \ge y_1$ & $y_2 \le 2 - y_1$ & $y_1 \ge 0$



Always do this first for your own understanding

Trivially, $f \ge 0$. Now we must show it integrates to 1.
Let's integrate vertically so we only have 1 integral.

$$\int_0^1 \int_{y_1}^{2-y_1} 6y_1^2 y_2 \, dy_2 \, dy_1$$

$$= \int_0^1 \left[3y_1^2 y_2^2\right]_{y_2 = y_1}^{y_2 = 2-y_1} dy_1$$

$$= \int_0^1 3y_1^2\left((2-y_1)^2 - y_1^2\right) dy_1$$

$$= \int_0^1 3y_1^2 (4 - 4y_1) \, dy_1$$

$$= \int_0^1 12y_1^2 - 12y_1^3 \, dy_1$$

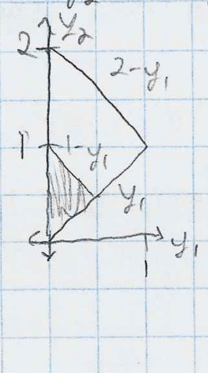
$$= \left[4y_1^3 - 3y_1^4\right]_0^1$$

$$= 1.$$

So f is a valid PDF.

We now find $P(Y_1 + Y_2 < 1)$. This is another inequality so we graph it again.

$y_2 < 1 - y_1$



Always draw & find support first!

We again integrate

$$P(Y_1+Y_2<1)=\int_0^{1/2}\int_{y_1}^{1-y_1} 6y_1^2 y_2\, dy_2\, dy_1 = 1/32$$

Marginal & Conditional Distribution

Useful if we only care about 1 variable

If Y_1 & Y_2 are jointly distributed, can we study one at a time? Yes

Def:

A marginal distribution summarizes the behavior of a single RV across all possible values of the other RV in their joint distribution.

If Y_1 & Y_2 are discrete RVs w/ PMF $p(y_1, y_2)$, then the marginal of Y_1 is

$$p_1(y_1)=\sum_{y_2} p(y_1,y_2).$$

& for y_2

$$p_2(y_2)=\sum_{y_1} p(y_1,y_2)$$

Or p_{Y_1} & p_{Y_2}

If Y_1 & Y_2 are continuous RVs w/ PDF $p(y_1, y_2)$, the the marginal of Y_1 is

$$p_1(y_1)=\int_{\mathbb{R}} p(y_1,y_2)\,dy_2$$

& for y_2

$$p_2(y_2)=\int_{\mathbb{R}} p(y_1,y_2)\,y_1$$

these are called marginal distributions b/c traditionally they are in margin of table (as sum of corresponding row/table)

Example:

Let p be given by

$y_1 \backslash y_2$	0	1	2	$\to p_{Y_1}=p_1$
0	1/9	2/9	1/9	4/9
1	2/9	2/9	0	4/9
2	1/9	0	0	1/9
$p_{Y_2}=p_2$	4/9	4/9	1/9	

That is

y_1	0	1	2
$p_1(y_1)$	4/9	4/9	1/9

y_2	0	1	2
$p(y_2)$	4/9	4/9	1/9

We mentioned earlier that distribution functions can be unique under "special conditions." Those conditions are met here so we can find the RV.

Let $Y' = Y_1 = Y_2$. We don't have a function for p', so let's rearrange the probabilities

$$p'(0)=4/9=(2/3)^2 \qquad p(1)=4/9=2(2/3)(1/3) \qquad p(2)=1/9=(1/3)^2$$

By writing p' in this way, we notice that $Y'=Y_1=Y_2 \sim \text{bin}(2, 1/3)$.

means y_1 & y_2 unique

Example:

Let $f(y_1, y_2) = y_1 + y_2$ for $(y_1, y_2) \in [0,1] \times [0,1]$.
Find the marginals.

We find f_{Y_1} & f_Y

left $[0,1]$

$$f_{Y_1}(y_1) = \int_0^1 y_1 + y_2 \, dy_2 = \left[y_1 y_2 + \tfrac{1}{2} y_2^2 \right]_0^1 = y_1 + \tfrac{1}{2} \qquad y_1 \in [0,1]$$

$$f_{Y_2}(y_2) = \int_0^1 y_1 + y_2 \, dy_1 = \left[\tfrac{1}{2} y_1^2 + y_1 y_2 \right]_0^1 = \tfrac{1}{2} + y_1 \qquad y_2 \in [0,1]$$

right $[0,1]$

What if we care about y_1 only where y_2 is a specific value?
We use a conditional distribution!

Def:

If Y_1 & Y_2 have a joint PMF $p(y_1, y_2)$ & marginals $p_1(y_1), p_2(y_2)$
then the conditional PMF $p_{Y_1|Y_2}$ of $Y_1|Y_2$

intersection(ish)

$$p_{Y_1|Y_2}(y_1|y_2) = P(Y_1 = y_1 \mid Y_2 = y_2) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)}$$

$$= \frac{p(y_1, y_2)}{p_2(y_2)}$$

Example:

Continuing previous example w/ square table of 9ths.

$$p_{Y_1|Y_2}(y_1|y_2) = \frac{p(y_1, y_2)}{p_2(y_2)} =$$

$y_1 \backslash y_2$	0	1	2
0	$\frac{1}{9} / \frac{4}{9}$	$\frac{2}{9} / \frac{4}{9}$	$\frac{1}{9} / \frac{1}{9}$
1	$\frac{2}{9} / \frac{4}{9}$	$\frac{2}{9} / \frac{4}{9}$	$0 / \frac{1}{9}$
2	$\frac{1}{9} / \frac{4}{9}$	$0 / \frac{4}{9}$	$0 / \frac{1}{9}$

=

$y_1 \backslash y_2$	0	1	2
0	$\frac{1}{4}$	$\frac{1}{2}$	1
1	$\frac{1}{2}$	$\frac{1}{2}$	0
2	$\frac{1}{4}$	0	0

Not always columns

Note every column needs to add to 1, since each column represents a different y_2 for $p_{Y_1|Y_2}$.

Def.

If Y_1 & Y_2 are continuous RV w/ joint PDF $f(y_1, y_2)$.
The conditional CDF of Y_1 given $Y_2 = y_2$ is given by

$$F_{Y_1|Y_2}(y_1|y_2) = P(Y_1 \leq Y_2 \mid Y_2 = y_2).$$

If Y_1 & Y_2 have marginals $f_1(y_1)$ & $f_2(y_2)$ then the conditional density of $Y_1|Y_2$ is given by

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)}$$

which is defined $\forall (y_1, y_2)$ s.t. $f_2(y_2) > 0$.

Example:

Suppose $f(y_1, y_2) = \frac{1}{2} y_1 y_2$ for $0 \leq y_1 \leq y_2 \leq y_2$

Find the marginals $f_1(y_1)$ & $f_2(y_2)$

$$f_1(y_1) = \int_{y_1}^{2} \tfrac{1}{2} y_1 y_2 \, dy_2 = \left[\tfrac{1}{4} y_1 y_2^2\right]_{y_1}^{2} = y_1 - \tfrac{1}{4} y_1^3 \quad \text{for } 0 \leq y_1 \leq 2$$

y_1 since $y_2 \geq y_1$

$$f_2(y_2) = \int_0^{y_2} \tfrac{1}{2} y_1 y_2 \, dy_1 = \left[\tfrac{1}{4} y_1^2 y_2\right]_0^{y_2} = \tfrac{1}{4} y_2^3 \quad \text{for } 0 \leq y_2 \leq 2$$

Find $f_{Y_2|Y_1}(y_2|y_1)$

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f(y_1, y_2)}{f(y_1)} = \frac{\frac{1}{2} y_1 y_2}{y_1 - \frac{1}{4} y_1^3} = \frac{2y_2}{4 - y_1^2} \quad \text{for } 0 \leq y_1 \leq y_2 \leq 2 \text{ where } y_1 \neq 2.$$

Find $P(Y_2 \leq 1.5 \mid Y_1 = 1)$.
Since $Y_1 = 1$, we have

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{2y_2}{4-(1)^2} = \tfrac{2}{3} y_2 \quad \text{for } \overset{y_1 = 1}{1} \leq y_2 \leq 2.$$

Now we have conditional PDF, so we find $P(Y_2 \leq 1.5 \mid Y_1 = 1)$ by integration.

$$P(Y_2 \leq 1.5 \mid Y_1 = 1) = \int_1^{1.5} \tfrac{2}{3} y_2 \, dy_2 = \left[\tfrac{1}{3} y_2^2\right]_1^{3/2} = \tfrac{3}{4} - \tfrac{1}{3} = \tfrac{5}{12}$$

Independence of RVs

If A & B are independent events, we write $A \perp B \Rightarrow P(A \cap B) = P(A)P(B)$
We write this in terms of CDFs.

Def:

Let Y_1 & Y_2 be RVs w/ joint CDF $F(y_1, y_2)$ & marginal CDFs $F_1(y_1)$ & $F_2(y_2)$.

Y_1 & Y_2 are <u>independent</u> ($Y_1 \perp Y_2$) iff $F(y_1, y_2) = F(y_1) F(y_2)$.

Thm:

Let Y_1 & Y_2 be RVs.

$$Y_1 \perp\!\!\!\perp Y_2 \iff \underbrace{p(y_1, y_2) = p_1(y_1)\, p_2(y_2)}_{\text{PMF, discrete}} \iff \underbrace{f(y_1, y_2) = f_1(y_1)\, f_2(y_2)}_{\text{PDF, continuous}}$$

Example:

Let $f(y_1, y_2) = 6y_1^2 y_2 \quad 0 \le y_1 \le y_2 \quad y_1 + y_2 \le 2$. Show Y_1 & Y_2 are dependent.

Since the supports of Y_1 & Y_2 depend on each other, we can't apply the above theorem. from diagram of support

$$f_1(y_1) = \int_{y_1}^{2-y_1} 6y_1^2 y_2 \, dy_2 = \left[3y_1^2 y_2^2\right]_{y_1}^{2-y_1} = 12y_1^2 - 12y_1^3 \quad \text{for } 0 \le y_1 \le 1$$

$$f_2(y_2) = \begin{cases} \int_0^{y_2} 6y_1^2 y_2 \, dy_1 & 0 \le y_2 \le 1 \\ \int_0^{2-y_2} 6y_1^2 y_2 \, dy_1 & 1 \le y_2 \le 2 \end{cases}$$

$$= \begin{cases} 2y_2^4 & 0 \le y_2 \le 1 \\ 2y_2(2-y_2)^3 & 1 \le y_2 \le 2 \end{cases}$$

Clearly $f(y_1, y_2) \ne f_1(y_1) f_2(y_2)$ b/c f_2 is piecewise. Thus Y_1 & Y_2 are dependent.

The process above (finding all PDFs & their product) doesn't scale to 100s+ of RVs. We need an easier way.

Thm: Factorization Theorem

Let Y_1, Y_2 be RVs w/ joint PDF $f(y_1, y_2)$ w/ support $(y_1, y_2) \in [a, b] \times [c, d]$. ← Ensures support doesn't depend on RV (e.g. ~~no~~ $0 \le y_1 \le y_2 \le$)

Y_1 & Y_2 are independent iff $f(y_1, y_2)$ can be factored into

$$f(y_1, y_2) = g(y_1)\, h(y_2)$$

for some non-negative, univariate functions g & h.

We say supports are seperable

This no longer requires finding marginals But it only applies when support is a box.

Example:

Let Y_1 & Y_2 be RVs w/ joint PDF

$f(y_1, y_2) = 3e^{-(y_1/2 + 6y_2)}$ $\quad y_1 \geq 0$ & $y_2 \geq 0$

Since the supports are seperable, we use the factorization theorem.

$f(y_1, y_2) = 3e^{-(y_1/2 + 6y_2)} = (3e^{-y_1/2})(e^{-6y_2}) = g(y_1)h(y_2)$ where

$g(y_1) = 3e^{-y_1/2}$ & $h(y_2) = e^{-6y_2}$

Thus Y_1 & Y_2 are independent $Y_1 \perp\!\!\!\perp Y_2$.

We could pick different g & h!

What distribution do Y_1 & Y_2 follow? If we choose

$g(y_1) = \frac{1}{2}e^{-y_1/2}$ & $h(y_2) = 6e^{-6y_2}$

so

$Y_1 \sim exp(2)$ & $Y_2 \sim exp(\frac{1}{6})$

Expectations

Thm:

Let Y_1 & Y_2 be RVs. Let c_i be constants & g_i be real-valued functions of Y_1 & Y_2.

discrete or continuous!

Then E is a linear operation

$$E_{Y_1,Y_2}\left[\sum_{i=1}^{k} c_i g_i(Y_1, Y_2)\right] = \sum_{i=1}^{k} c_i E_{Y_1,Y_2}[g_i(Y_1, Y_2)]$$

If Y_1 & Y_2 are both discrete, we get

$$= \sum_{i=1}^{k} c_i\left(\sum_{y_1}\sum_{y_2} g_i(y_1, y_2)p(y_1, y_2)\right)$$

If Y_1 & Y_2 are continuous, we get

$$= \sum_{i=1}^{k} c_i \int_{\mathbb{R}}\int_{\mathbb{R}} g_i(y_1, y_2) f(y_1, y_2)\, dy_2\, dy_1$$

Rem

$E_{Y_1,Y_2}[Y_1] = E_{Y_1}[Y_1]$, which is why we can normally elide the subscript

Example:

Consider $f(y_1, y_2) = \frac{2}{3}y_1 e^{-y_2/3}$ for $y_1 \in [0,1]$ $y_2 \in [0,\infty)$ ← seperable supports

How do you find $E_{Y_1,Y_2}[Y_1Y_2]$.

i) Directly integrate $E_{Y_1,Y_2}[Y_1Y_2] = \int_0^\infty \int_0^1 y_1 y_2 \frac{2}{3} y_1 e^{-y_2/3}\, dy_1\, dy_2$ — we do simpler integral first b/c y_1 y_2 seperable.

$dy_1 \Rightarrow$ power rule

$dy_2 \Rightarrow$ integration by parts

ii) By factorization theorem, Y_1 & Y_2 are independent, so separate integration into

$$\left(\int_0^\infty y_2 \cdot \tfrac{1}{3} e^{-y_2/3}\, dy_2\right)\left(\int_0^1 y_1 \cdot 2y_1\, dy_1\right)$$

The first integral looks like $E[Y]$ for exp(3) & the 2nd beta(2,1)

Method ii is way easier so now we find

$$E_{Y_1,Y_2}[Y_1 Y_2] = \left(\int_0^\infty y_2 \cdot \tfrac{1}{3} e^{-y_2/3}\, dy_2\right)\left(\int_0^1 y_1 \cdot 2y_1\, dy_1\right)$$

$= E[\exp(3)]\ E[\text{beta}(2,1)]$
$= (3)\cdot(2/3)$
$= 2$

Example:

What about the same Y_1 & Y_2. Finding $E[Y_1^{Y_2}]$.
We can only use method i b/c the expected value is no longer seperable.

even tho Y_1 & Y_2 are independent!

Thm:

Let Y_1 & Y_2 be independent RVs. (i.e. $f(y_1, y_2) = f_1(y_1) f_2(y_2)$)
Let $g(Y_1)$ & $h(Y_2)$ be univariate functions.
If all the following expectations exist, then

$$E_{Y_1,Y_2}[g(Y_1)\, h(Y_2)] = E_{Y_1,Y_2}[g(Y_1)]\, E_{Y_1,Y_2}[h(Y_2)]$$

$$= E_{Y_1}[g(Y_1)]\, E_{Y_2}[h(Y_2)]$$

Pf:

$E_{Y_1,Y_2} = [g(Y_1)\, h(Y_2)] = \int_{\mathbb{R}}\int_{\mathbb{R}} g(y_1)\, h(y_2)\, f(y_1, y_2)\, dy_1\, dy_2$
$= \int_{\mathbb{R}}\int_{\mathbb{R}} g(y_1)\, h(y_2)\, f_1(y_1)\, f_2(y_2)\, dy_1\, dy_2$
$= \left(\int_{\mathbb{R}} g(y_1)\, f_1(y_1)\, dy_1\right)\left(\int_{\mathbb{R}} h(y_2)\, f_2(y_2)\, dy_2\right)$
$= E[Y_1]\, E[Y_2]$

Example:

Let Y_1 & Y_2 be independent & $Y_1 \sim N(\mu, \sigma^2)$ $Y_2 \sim N(\mu, \sigma^2)$.

Find the moment generating function (MGF) of $Y_1 + Y_2$

$m_{Y_1+Y_2}(t) = E_{Y_1,Y_2}[e^{t(Y_1+Y_2)}] = E_{Y_1,Y_2}[e^{tY_1} e^{tY_2}]$ ← univariate seperable fns

$= E_{Y_1}[e^{tY_1}]\, E_{Y_2}[e^{tY_2}]$ & Y_1 & Y_2 independent

$= m_{Y_1}(t)\, m_{Y_2}(t)$

$= \left(e^{\mu t + \sigma^2 t^2/2}\right)\left(e^{\mu t + \sigma^2 t^2/2}\right) = \boxed{e^{2\mu t + \sigma^2 t^2}}$

Find the moment generating function (MGF) of $\bar{Y} = \frac{Y_1 + Y_2}{2}$

$$m_{\bar{Y}}(t) = m_{\frac{Y_1+Y_2}{2}}(t) = m_{Y_1+Y_2}(t/2) = e^{\mu t + \sigma^2 t^2/4}$$

Looking at the MGFs, we can find the distributions of $Y_1 + Y_2$ & $\bar{Y}$.

$$m_{Y_1+Y_2}(t) = e^{2\mu t + \sigma^2 t^2} \Rightarrow Y_1 + Y_2 \sim N(2\mu, 2\sigma^2)$$

$$m_{\bar{Y}}(t) = e^{\mu t + \sigma^2 t^2/4} \Rightarrow \bar{Y} \sim N(\mu, \sigma^2/2)$$

$$m_Y(t) = e^{\mu t + \sigma^2 t^2/2} \Leftrightarrow Y \sim N(\mu, \sigma^2)$$

Now we can see the nice properties of the normal distribution!
It is "closed" under addition of normal distributions, which is unusual.

Covariance

Note: $X \not\perp\!\!\!\perp Y$ means X & Y dependent

If we have 2 RVs Y_1 & Y_2 are dependent $\not\perp\!\!\!\perp$, then changes in Y_1 may result in changes in Y_2 & vice versa. How do we quantify this?

We use moments of a joint distribution $f(y_1, y_2)$! Here we'll talk about the moment called covariance.

Def:
Let Y_1 & Y_2 be RVs where $E[Y_1] = \mu_1$ & $E[Y_2] = \mu_2$.
The covariance of Y_1 & Y_2 measures the linear relationship b/w 2 RVs & is calculated

$$COV[Y_1, Y_2] = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

We can simplify this to

$$Cov[Y_1, Y_2] = E[Y_1 Y_2] - E[Y_1]E[Y_2]$$

Thm:
If Y_1 & Y_2 are independent, then

$$COV[Y_1, Y_2] = E[Y_1 Y_2] - E[Y_1]E[Y_2] = 0$$ ← Converse does not hold in general

b/c $E[Y_1 Y_2] = E[Y_1]E[Y_2]$.

Remark:
$$COV[Y_1, Y_1] = E[Y_1^2] - E^2[Y_1] = V[Y_1]$$

Example:
Let $f(y_1, y_2) = 4y_1 y_2$ for $y_1 \in [0,1]$ & $y_2 \in [0,1]$. Find $COV[Y_1, Y_2]$.

Since f is seperable & so are the supports, Y_1 & Y_2 are independent $Y_1 \perp\!\!\!\perp Y_2$
$\therefore COV[Y_1, Y_2] = 0$.

Example:
Let $f(y_1, y_2) = 6y_1^2 y_2$ for $0 \le y_1 \le y_2$ & $y_1 + y_2 \le 2$. Find $COV[Y_1, Y_2]$.

$$E[Y_1 Y_2] = \int_0^1 \int_{y_1}^{2-y_1} y_1 y_2 \, 6y_1^2 y_2 \, dy_2 \, dy_1$$ ← bounds determined by (way) earlier graph

$$= \int_0^1 \int_{y_1}^{2-y_1} 6y_1^3 y_2^2 \, dy_2 \, dy_1$$

$$= \int_0^1 [2y_1^3 y_2^3]_{y_1}^{2-y_1} \, dy_1$$

$$= \int_0^1 2y_1^3(2-y_1)^3 - 2y_1^6 \, dy_1$$

$$= \int_0^1 2y_1^3(-y_1^3 + 6y_1^2 - 12y_1 + 8) - 2y_1^6 \, dy_1$$

$$= \int_0^1 -4y_1^6 + 12y_1^5 - 24y_1^4 + 16y_1^3 \, dy_1$$

$$= \left[-\tfrac{4}{7} y_1^7 + 2y_1^6 - \tfrac{24}{5} y_1^5 + 4y_1^4 \right]_0^1$$

$$= -\tfrac{4}{7} + 2 - \tfrac{24}{5} + 4$$

$$= -\tfrac{20}{35} + \tfrac{70}{35} - \tfrac{168}{35} + \tfrac{140}{35}$$

$$= -\tfrac{168}{35} + \tfrac{190}{35}$$

$$= \tfrac{22}{35}$$

$$E[Y_1] = \int_0^1 \int_{y_1}^{2-y_2} y_1 \, 6y_1^2 y_2 \, dy_2 \, dy_1 = \tfrac{3}{5}$$

$$E[Y_2] = \int_0^1 \int_{y_1}^{2-y_2} y_2 \, 6y_1^2 y_2 \, dy_2 \, dy_1 = \tfrac{16}{15}$$

$$\text{COV}[Y_1, Y_2] = E[Y_1 Y_2] - E[Y_1]E[Y_2] = \tfrac{22}{35} - \left(\tfrac{3}{5}\right)\left(\tfrac{16}{15}\right) = -\tfrac{2}{175}$$

B/c COV is negative, as the values of either RV increase, the other value decreases.

Can't say variables aren't very related b/c covariance is near 0 or any sort of scale.

<u>Cons of Covariance:</u>

Covariance is unbounded & depends on unit of measurements.

To address these cons of measuring <u>linear</u> dependence, we will use correlation coefficients, specifically Pearson's <u>Product Moment</u> Correlation Coefficient. moment of $(Y_1 - \mu_1)(Y_2 - \mu_2)$:)

<u>Def:</u>

The <u>correlation</u> b/w Y_1 & Y_2 is defined as

$$\rho_{1,2} = \frac{\text{COV}[Y_1, Y_2]}{\sigma_1 \sigma_2} = \frac{\sigma_{1,2}}{\sigma_1 \sigma_2}$$ ← covariance b/w Y_1 & Y_2

Here we're <u>normalizing</u> the coefficient by making it unitless (i.e. not depend on units) & bounded b/w $-1 \leq \rho_{1,2} \leq 1$.

Example:

Using $f(y_1, y_2) = 6y_1^2 y_2$ for $0 \le y_1 \le y_2$ & $y_1 + y_2 \le 2$ again, we find the correlation coefficient $\rho_{1,2}$.

$\sigma_{1,2} = COV[Y_1, Y_2] = -2/125$ ← from earlier

$\sigma_1^2 = V[Y_1] = 1/25$ ← could find doing $E[Y_1^2] - E^2[Y_1]$

$\sigma_2^2 = V[Y_2] = 14/225$ ← could also find

We now find $\rho_{1,2}$

$$\rho_{1,2} = \frac{\sigma_{1,2}}{\sigma_1 \sigma_2} = \frac{-2/125}{\sqrt{1/25}\sqrt{14/225}} = -0.229$$

More Advanced Combinations of RVs

Def:

Let $Y_1, \ldots, Y_n$ be RVs & $a_1, \ldots, a_n$ be scalars. Define W_1 as

$$W_1 = \sum_{j=1}^{n} a_j Y_j$$

Thm:

Let $Y_1, \ldots, Y_n$ be RVs where $E[Y_j] = \mu_j$ } may not be

Let $X_1, \ldots, X_m$ be RVs where $E[X_i] = \eta_i$ } independent

For $a_1, \ldots, a_n$ & $b_1, \ldots, b_m$ scalars we get

$V[Y+X] = V[Y] + V[X] + 2\,COV[X,Y]$

$V[Y-X] = V[Y] + V[X] - 2\,COV[X,Y]$

i) $E[W_1] = \sum_{j=1} a_j \underset{\text{or } E[Y_j]}{\mu_j}$ ← expectation E is linear

ii) $V[W_1] = \sum_{j=1} a_j^2 V[Y_j] + 2\sum_{k<l}\sum a_k a_l COV[Y_k, Y_l]$

where $W_1 = \sum_{j=1}^{n} a_j Y_j$ & $W_2 = \sum_{i=1}^{m} a_i X_i$.

iii) $COV[W_1, W_2] = \sum_{j=1}^{n} \sum_{i=1}^{m} a_j b_i COV[Y_j, X_i]$ ← Whoops

If you've used matrices, like I have, this form of the theorem is helpful.

Let $a^T = (a_1, \ldots, a_n)$ & $b^T = (b_1, \ldots, b_m)$.
Let $Y^T = (Y_1, \ldots, Y_n)$ & $X^T = (X_1, \ldots, X_m)$.
Let $\mu^T = (\mu_1, \ldots, \mu_n)$ & $\eta^T = (\eta_1, \ldots, \eta_m)$.
Then

i) $E[W_1] = E[a^T Y] = a^T \mu$

ii) $V[W_1] = a^T\, COV[Y]\, a$

iii) $COV[W_1, W_2] = a^T\, COV[Y, X]\, b$

Example:

A study by fivethirtyeight.com tracked 29 voters on their opinion of political candidates over many months.
W/in a month let Y_1 count the number supporting B & Y_2 supporting W.

Then $\hat{p}_1 = Y_1/29$ & $\hat{p}_2 = Y_2/29$

Since $\hat{p}_1$ & $\hat{p}_2$ depend on RVs, they are random variables.

From previous experience we know

$E[\hat{p}_1] = 0.1 \quad V[\hat{p}_1] = 0.003 \quad COV[\hat{p}_1, \hat{p}_2] = -0.002$

$E[\hat{p}_2] = 0.48 \quad V[\hat{p}_2] = 0.012$

$\hat{p}_1$ & $\hat{p}_2$ are dependent b/c if a subject supports B then they don't support W, since the probability must sum to 1.

What is $E[\hat{p}_2 - \hat{p}_1]$? That is how far apart is their popularity on average. Likewise $V[\hat{p}_2 - \hat{p}_1]$

38 percentage point lead for candidate 2

$$E[\hat{p}_2 - \hat{p}_1] = E[\hat{p}_2] - E[\hat{p}_1] = 0.48 - 0.1 = 0.38$$

$$V[\hat{p}_2 - \hat{p}_1] = V[\hat{p}_2] + V[\hat{p}_1] - 2COV[\hat{p}_2, \hat{p}_1] = 0.012 + 0.003 + 0.004 = 0.019$$

Give a 95% confidence interval for $\hat{p}_2 - \hat{p}_1$.
We'll assume the values are normally distributed, so we can use a z-score. We calculate this for mean

$$(\hat{p}_2 - \hat{p}_1) \pm Z_{\alpha/2}\sqrt{V[\hat{p}_2 - \hat{p}_1]}$$

Recall a confidence interval is Point Estimate ± (Critical value) * (Standard Error)

$$= 0.38 \pm 1.96\sqrt{0.019}$$
$$= 0.38 \pm 0.27$$
$$= (0.11, 0.65)$$

We are thus 95% confident that the true difference (W − B) in the proportion of voters who support these candidates in a particular month is b/w 0.11 & 0.65. Since the interval is entirely above 0, there is evidence to suggest W is the more strongly supported candidate

Def:
Let Y_1 & Y_2 be RVs. The conditional expectation of function $g(Y_1)$ given $Y_2 = y_2$ is defined as

$$E[g(Y_1) \mid Y_2 = y_2] = \int_{-\infty}^{\infty} g(y_1) f(y_1 \mid y_2)\,dy_1 \quad \text{when } Y_1, Y_2 \text{ continuous}$$

$$E[g(Y_1) \mid Y_2 = y_2] = \sum_{y_1} g(y_1) f(y_1 \mid y_2) \quad \text{when } Y_1, Y_2 \text{ discrete}$$

This $E[g(Y_1) \mid Y_2 = y_2]$ still depends on a random variable Y_2 so it itself is another random variable. Thus we can find its expected value which leads to iterated expectations

Iterated Expectations

Thm:

Let Y_1 & Y_2 be RVs.

$$E[Y_1] = E_{Y_2}[E_{Y_1|Y_2}[Y_1|Y_2]] = E[E[Y_1|Y_2]]$$

This means we can find a mean w/o the joint distribution or marginal for Y_1!

Since variance $V[Y_1 | Y_2 = y_2]$ is just an expected value, we get

$$V[Y_1|Y_2 = y_2] = E[Y_1^2|Y_2 = y_2] - E^2[Y_1|Y_2 = y_2]$$

which gives us the following theorem.

Thm:

Let Y_1 & Y_2 be RVs. Then

$$V[Y_1] = E[V[Y_1|Y_2]] + V[E[Y_1|Y_2]]$$

Example:

Wild Things throws strikes w/ probability p. We want to count the number of strikes Wild Things throws out in 6 pitches. If p is known, then $Y|p \sim bin(6,p)$

However, we don't know p as it varies. We know from historical data $p \sim beta(0.25, 0.75)$

We use iterated expectations to find mean $E[Y]$

$$\begin{aligned} E[Y] &= E[E[Y|p]] \\ &= E[6p] \\ &= 6E[p] \\ &= 6\left(\frac{0.25}{0.25+0.75}\right) \\ &= 1.5 \end{aligned}$$

We use iterated expectations to find variance $V[Y]$

$$\begin{aligned} V[Y] &= E[V[Y|p]] + V[E[Y|p]] \\ &= E[6pq] + V[6p] \\ &= 6E[p(1-p)] + 6^2V[p] \\ &= 6E[p-p^2] + 36V[p] \\ &= 6E[p] - 6E[p^2] + 36V[p] \end{aligned}$$

We don't know $E[p^2]$ normally, but we know it is part of $V[p]$ so we rearrange $V[p] = E[p^2] - E^2[p]$ to give us $E[p^2] = V[p] + E^2[p]$

$$\begin{aligned} V[Y] & \\ &= 6E[p] - 6(V[p] + E^2[p]) + 36V[p] \\ &= 6(0.25) - 6(0.09375 + 0.25^2) + 36(0.09375) \\ &= 3.9375 \end{aligned}$$

Note: These concepts can be extended to more than two RVs trivially, especially using linear algebra & matrices.

Functions of Random Variables

So far, we've just been taking linear functions of random variables. How do we do that? Why do that?

Often, we want to talk about distributions of statistics, like the distribution of sample means (linear), variance (non-linear), range (non-linear), etc. Here's a list.

1) Mean: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
2) Proportion: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \in [0,1]$
3) Variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
4) Range: $R = \max(Y_i) - \min(Y_i) = Y_{(n)} - Y_{(1)}$ ← $Y_{(i)}$ is i-th Y arranged in sorted order
5) Z-Statistic: $\frac{\bar{Y}-\mu}{\sigma^2/\sqrt{n}}$ } everything but $\bar{Y}$ is constant so this is linear
6) T-Statistic: $\frac{\bar{Y}-\mu}{S^2/\sqrt{n}}$ } no longer linear b/c S^2 not constant
7) $T_{2\text{-sample}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$
8) $MSE_{MLR} = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

All of the above are statistics (as RVs) & are all functions of the given random variable. <u>This is true for all statistics.</u>

If we can't find the probability distribution of a statistic, then we can't do hypothesis testing or confidence intervals b/c we would have no way to quantify the likelihood of getting said statistic value.

Before we can start describing properties of functions of random variables, we need to formalize them.

Let $Y_1, \ldots, Y_n$ be a sequence of n random variables.
Let $W_1 = h(Y_1, \ldots, Y_n)$ & $W_2 = g(Y_1, \ldots, Y_n)$ where g & h are functions of RVs. We denote $Y_1, \ldots, Y_n$ as $\vec{Y}$.

We need to find the marginal distributions & joint distributions for W_1 & W_2. That is

$f_{W_1}(w_1)$ & $f_{W_2}(w_2)$ — Marginal Distribution
$f_{W_1,W_2}(w_1, w_2)$ — Joint Distribution

Then we use the earlier methods we've learned to find our desired quantities

Let's go over 3 techniques. We'll go more in-depth later.

Technique 1: Method of Distribution Functions
Let $W = h(\vec{Y})$ be a RV where joint distribution of $\vec{Y}$ is $f_{\vec{Y}}(\vec{Y})$.

1) Define a region of the support of $f_{\vec{Y}}$ where $W \le w$ for some w. (set of n numbers)

2) Find $F_W(w) = P(W \le w)$ by integrating $f_{\vec{Y}}(\vec{Y})$ over this region.

3) Differentiate $F_W(w)$ wrt w to get $\frac{d}{dw} F_W(w) = f_W(w)$.

Technique 2: (Univariate) Method of Transformations
Let $W = h(Y)$ be a strictly increasing or decreasing function over the support of $f_Y(y)$.

1) Find $y = h^{-1}(w)$ which exists b/c h is strictly monotonic.

2) $f_W(w) = f_Y(h^{-1}(w)) \left| \frac{d}{dw} h^{-1}(w) \right|$

This is limited b/c it only works for one-variable & the function must be monotonic.

Technique 3: Method of Moment Generating Functions
Let $W = h(\vec{Y})$.

W is the RV that we care about, the function of the RV.

1) Find the moment generating function of W

$m_W(t) = E[e^{tW}] = E[e^{th(\vec{Y})}]$ if it exists

2) Identify the moment generating functions

When should we use these techniques?
1) Always works but hard
2) Only works for invertible, univariate functions but kinda easy
3) Easiest but sometimes you can't identify the MGF.

Method of Distribution Functions

Here we'll be applying the technique 1 discussed above.

Example: 1 RV
Let $W = 2Y - 1$ where $Y \sim \beta(2,2)$. Here w/ W we are modelling the difference b/w two proportions. (You can see this w/ support & other properties)

1) What is the support of W?

Since Y is a beta distribution, the support of Y is $[0,1]$. Plugging this into W we get a support $[-1,1]$.

2) What is the CDF F_W for W?

$$F_W(w) = \begin{cases} 0 & w < -1 \\ ? & -1 < w < 1 \\ 1 & 1 < w \end{cases}$$

Let's look at what happens for $-1 < w < 1$.
Using the method of distribution functions we know

$$F_W(w) = P_W(W \le w)$$
$$= P_Y(2Y - 1 \le w) \leftarrow \text{Rewrite in terms of the known } Y$$
$$= P_Y\left(Y \le \frac{w+1}{2}\right) \leftarrow \text{Find the "region" for } Y \text{ that corresponds to that of } W$$

Now we integrate to find P_Y

$$F_W(w) = P_Y\left(Y \le \frac{w+1}{2}\right)$$
$$= \int_0^{\frac{w+1}{2}} \frac{\Gamma(2+2)}{\Gamma(2)\Gamma(2)} y^{2-1}(1-y)^{2-1}\,dy$$
$$= \int_0^{\frac{w+1}{2}} 6y(1-y)\,dy$$
$$= \int_0^{\frac{w+1}{2}} 6y - 6y^2\,dy$$
$$= \left[3y^2 - 2y^3\right]_0^{\frac{w+1}{2}}$$
$$= \left(3\left(\tfrac{w+1}{2}\right)^2 - 2\left(\tfrac{w+1}{2}\right)^3\right)$$
$$= 1/4(3w^2 + 6w + 3 - w^3 - 3w^2 - 3w - 1)$$
$$= 1/4(-w^3 + 3w + 2)$$

Thus the CDF F_W for W is

$$F_W = \begin{cases} 0 & (-\infty, -1) \\ 1/4(-w^3 + 3w + 2) & [-1, 1] \\ 1 & (1, \infty) \end{cases}$$

3) What is the PDF f_W for W?

$$f_W(w) = \frac{d}{dw} F_W(w) = 1/4(-3w^2 + 3) = 3/4(1+w)(1-w) \quad -1 \le w \le 1$$

This kinda looks like a beta. It's actually the generalized beta, which deals w/ any support.

Examples 2 RVs

Let $W = Y_2 - Y_1$ be the simple difference of RVs Y_1 & Y_2. This is common for comparisons in statistics. Suppose the joint distribution f_{Y_1, Y_2} for Y_1 & Y_2 is

$$f_{Y_1, Y_2}(y_1, y_2) = e^{-y_2} \quad \text{for } 0 \le y_1 \le y_2 < \infty$$

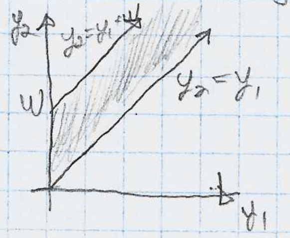
1) What is the support of W? Since $y_2 \ge y_1$, the min is zero & there is no max. The support is $[0, \infty)$.

2) Find the region of support of f_{Y_1,Y_2} where $W \le w$. Draw a graph of the support of f_{Y_1,Y_2} & indicate where $W \le w$.

$$W \le w \Rightarrow Y_2 - Y_1 \le w \Rightarrow Y_2 \le Y_1 + w$$

Thus our region is the space below $Y_2 = Y_1 + w$. Likewise, since $y_2 \ge y_1$ we are bounded below by $y_2 = y_1$.

Let's draw a graph.



3) Find the CDF $F_W(w)$ for W

$$F_W(w) = P_W(W \le w)$$

$$= P_{Y_1,Y_2}(Y_2 - Y_1 \le w)$$ similar to above

$$= P_{Y_1,Y_2}(Y_2 \le Y_1 + w)$$

$$= \begin{cases} 0 & (-\infty, 0) \\ \int_0^\infty \int_{y_1}^{y_1+w} e^{-y_2}\, dy_2\, dy_1 & [0, \infty) \end{cases}$$

$$= \begin{cases} 0 & '' \\ \int_0^\infty [-e^{-y_2}]_{y_1}^{y_1+w}\, dy_1 & '' \end{cases}$$

$$= \begin{cases} 0 & '' \\ \int_0^\infty -e^{-y_1-w} + e^{-y_1}\, dy_1 & '' \end{cases}$$

$$= \begin{cases} 0 & '' \\ [e^{-y_1-w} - e^{-y_1}]_0^\infty & '' \end{cases}$$

$$= \begin{cases} 0 & '' \\ -e^{-w} + 1 & '' \end{cases}$$

Look at this CDF. Is it familiar? It kinda looks like an exponential to me. Let's find the PDF to confirm

4) Find the PDF $f_W(w)$ for W

$$f_W(w) = \frac{d}{dw} F_W(w) = e^{-w} \qquad w \in [0, \infty)$$

This is the exponential distribution! Specifically $W \sim exp(1)$.

Method of (Univariate) Transformations

Recall:

Let $W = f(Y)$ be a strictly increasing or strictly decreasing (i.e. monotonic) function w/ a support of $f_Y(y)$.

The reason this needs a univariate, monotonic function $h(Y)$ so its derivative is either always positive or always negative.

This method has only 2 steps once you have your (univariate) monotonic function $h(Y)$.

1) Find $y = h^{-1}(w)$ which exists b/c h strictly monotonic.
2) Find $f_W(w) = f_Y(h^{-1}(w)) \left| \frac{d}{dw} h^{-1}(w) \right|$.

Example:

Suppose $f_Y(y) = 1/5 \, e^{-y/5}$ for $y \geq 0$ $(Y \sim \exp(5))$.
Let $W = 2Y$.

Can you use the method of (univariate) transformations here?
Yes. $W = h(Y) = 2Y$ is strictly increasing over support $(0, \infty)$.

Find $h^{-1}(w)$
$$h^{-1}(w) = \tfrac{1}{2} w$$

Find $\frac{d}{dw} h^{-1}(w)$
$$\frac{d}{dw} h^{-1}(w) = 1/2$$

Now we substitute to find
$$f_W(w) = f_Y(h^{-1}(y)) \left| \frac{d}{dw} h^{-1}(y) \right|$$
$$= f_Y(\tfrac{1}{2}w) \, \tfrac{1}{2}$$
$$= 1/10 \, e^{-w/10}$$

Note that $W \sim \exp(10)$

Note that $Y \sim \Gamma(1, 5) = \exp(5)$ & $W \sim \Gamma(1, 10) = \exp(10)$.
Unsurprisingly scaling the RV by c 'scales' the scale factor β by c.

Why does this work?

First, our region $W \leq w \Rightarrow h(Y) \leq w$. If we assume $h(Y)$ monotonically increases the
$h(Y) \leq w \Rightarrow Y \leq h^{-1}(w)$
Note if $h(Y)$ is monotonically decreasing then $h(Y) \leq w \Rightarrow Y \geq h^{-1}(w)$.

Thus our CDF is
$F_W(w) = P(W \leq w) = P(Y \leq h^{-1}(w)) = F_Y(h^{-1}(w))$.
Then our PDF is
$f_W(w) = f(h^{-1}(w)) \left| \frac{d}{dw} h^{-1}(w) \right|$ ← absolute value gets rid of sign difference for monotonically increasing/decreasing functions.

This is a special case of the method of distribution functions.

Example:
Let Y be a RV w/ PDF $f_Y(y) = 2y$ for $0 \leq y \leq 1$.
Let $W = 1 - Y$. Find the distribution of W.

$h(Y) = 1 - Y$ is strictly decreasing on $[0,1]$ so we can use method of (univariate) transformations.

$h^{-1}(w) = 1 - w$ & $\frac{d}{dw} h^{-1}(w) = -1$

Now we substitute to find f
$f_W(w) = f_Y(h^{-1}(w)) \left| \frac{d}{dw} h^{-1}(w) \right| = f_Y(1-w) \cdot -1 = -2(w-1) = 2(1-w)$. $0 \leq w \leq 1$

Note that $Y \sim beta(2,1)$ & $W \sim beta(1,2)$. This should make sense.

Method of Moment Generating Functions

Recall the steps of this method is
1) Find the MGF of $h(Y)$
2) Identify it

However, notice that this MGF method doesn't provide PDFs, which is why identification is crucial.

Example:
Let Y_1 count how many pitches until wild thing throws his 1st pitch, Y_2 how many pitches after 1st strike until 2nd, & Y_3 until 3rd.

What is the distribution of total number of pitches before the 3rd strike? Let $W = Y_1 + Y_2 + Y_3$. W is what we want

Note $Y_i \sim geo(p)$.

We find the MGF of W
$M_W(t) = E[e^{tW}] = E[e^{t(Y_1+Y_2+Y_3)}]$
$= E[e^{tY_1}]\, E[e^{tY_2}]\, E[e^{tY_3}]$
$= \left(\frac{pe^t}{1-qe^t}\right)$
Since MGFs are unique, we know $W \sim nb(3, p)$.

Conceptually, this makes sense as the negative binomial is essentially defined as a sum of geometrics (which have the same p.

We can simplify our method of MGFs for a common case

Thm: 6.2
If $Y_1, \ldots, Y_n$ are independent RVs & $W = \sum_{i=1}^{n} Y_i$, then the MGF is
$$M_W(t) = \prod_{i=1}^{n} M_{Y_i}(t)$$

Thm: 6.3
Let $Y_1, \ldots, Y_n$ be independent RVs w/ $Y_i \sim n(\mu_i, \sigma_i^2)$ & let $a_1, \ldots, a_n$ be constants.
Let $W = \sum_{i=1}^{n} a_i Y_i$. Then $W \sim n\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$.

Here we can see combining independent normal random variables still results in normal RVs.

Pf: 6.3
Since $Y_i \sim n(\mu_i, \sigma_i^2)$, we have MGF $M_{Y_i}(t) = e^{\mu_i t + \sigma_i^2 t^2/2}$

We know by the definition of MGFs
$$M_{a_i Y_i}(t) = e^{\mu_i a_i t + \sigma_i^2 a_i^2 t^2/2}$$

From theorem 6.2 we get
$$M_W(t) = \prod_{i=1}^{n} e^{\mu_i a_i t + \sigma_i^2 a_i^2 t^2/2}$$
$$= e^{\sum_{i=1}^{n} \mu_i a_i t + \sigma_i^2 a_i^2 t^2/2}$$
$$= e^{\sum_{i=1}^{n} (\mu_i a_i) t + \sum_{i=1}^{n} (\sigma_i^2 a_i^2) t^2/2}$$

Since MGFs are unique
$$W \sim n\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right). \quad \square$$

Example:
Let $Z \sim n(0,1)$. What is the distribution of $W = Z^2$.

$$M_W(t) = E[e^{tZ^2}] = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad \text{(PDF of normal)}$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2(1/2 - t)} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/(2(1-2t)^{-1})} dz \quad \text{(try to make it a normal kernel)}$$

$$= \sqrt{(1-2t)^{-1}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-2t)^{-1}}} e^{-z^2/(2(1-2t)^{-1})} \, dz$$

(make look more like normal kernel)

$$= \sqrt{(1-2t)^{-1}} \quad \text{if} \quad 1-2t > 0$$

$$= (1-2t)^{-1/2} \quad \text{for} \quad t < 1/2$$

Recognize that this is the MGF of $\chi^2(1)$. Thus $W \sim \chi^2(1)$.

χ^2 comes from squaring the normal distribution

Thm:

Consider independent RVs $Z_1, \ldots, Z_n$ where $Z_i \sim n(0,1)$.
Let $W = \sum_{i=1}^{n} Z_i^2$. Then

$$W = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

Pf:

From our previous theorem $Z_i^2 \sim \chi^2(1)$. Thus $M_{Z_i^2}(t) = (1-2t)^{-1/2}$.

By theorem 6.2

$$M_W(t) = \prod_{i=1}^{n} (1-2t)^{-1/2} = (1-2t)^{-n/2}$$

Since MGFs are unique we know $W \sim \chi^2(N)$.

This can be used for all RVs but is easy for independent RVs & useful if you recognize the MGF.

Method of Jacobians

The method of Jacobians is like the method of (univariate) transformations for multivariate functions. Let's motivate this by talking about bivariate functions.

Example: Bivariate

Suppose $W_1 = h_1(Y_1, Y_2)$ & $W_2 = h_2(Y_1, Y_2)$ are RVs where Y_1 & Y_2 are also RVs.

We still need to assume h_1^{-1} & h_2^{-1} exist, but this time we need partial derivatives.

Let's diverge to cover what a Jacobian is.

Def:

A Jacobian is a matrix of partial derivatives.

Let's return to the example.

Example: Bivariate cont.

We form a Jacobian matrix J of inverses, where rows are functions & columns are derivatives. In particular

$$J = \begin{bmatrix} \frac{dh_1^{-1}}{dw_1} & \frac{dh_1^{-1}}{dw_2} \\ \frac{dh_2^{-1}}{dw_1} & \frac{dh_2^{-1}}{dw_2} \end{bmatrix}$$

Consider in particular
$W_1 = h_1(Y_1, Y_2) = Y_1 + 2Y_2$ & $W_2 = h_2(Y_1, Y_2) = 3Y_1 - Y_2$

We find Y_1 & Y_2 in terms of W_1 & W_2 to give us h_1^{-1} & h_2^{-1} respectively.

Using Elimination
$$W_1 + 2W_2 = 7Y_1 \Rightarrow Y_1 = \frac{W_1 + 2W_2}{7} = h_1^{-1}(W_1, W_2),$$
$$3W_1 - W_2 = 7Y_2 \Rightarrow Y_2 = \frac{3W_1 - W_2}{7} = h_2^{-1}(W_1, W_2).$$

Thus we get Jacobian matrix J by doing simple derivatives
$$J = \begin{bmatrix} 1/7 & 2/7 \\ 3/7 & -1/7 \end{bmatrix}$$

Now we have our Jacobian & we need to fit it into
$$f_W(w) = f_Y(h^{-1}(w)) \left| \frac{d}{dw} h^{-1}(w) \right|.$$

Our Jacobian is almost the analog of $\frac{dh^{-1}(w)}{dw}$. However, we need to do the determinant to find the "scale" of J & to make it scalar.

Now w/ the Jacobian handled, the rest of the process is fairly simple. Let's go to the general case. (Note: We use a bivariate f_W, f_Y, & h^{-1} in this particular case & a multivariate ones in the general case).

The <u>method of Jacobians</u> is thus as follows:

For a set of random variables $\vec{Y} = \{Y_1, \ldots, Y_n\}$, suppose
- $W_1 = h_1(\vec{Y}), \ldots, W_n = h_n(\vec{Y})$,
- Each h_i has an inverse h_i^{-1} st $Y_i = h_i^{-1}(\vec{W})$,
- Each h_i^{-1} has a continuous partial derivative wrt each W_i,
- The determinant of the Jacobian is non-zero.

Then the joint density of $\vec{W}$ is given by
$$f_{\vec{W}}(w_1, \ldots, w_n) = f_{\vec{Y}}(h_1^{-1}(\vec{w}), \ldots, h_n^{-1}(\vec{w})) \, |\det(J)|$$

Example:

Suppose $W_1 = Y_1 + 2Y_2$ & $W_2 = 3Y_1 - Y_2$. Find the joint distribution of W_1 & W_2 if $f_{Y_1,Y_2}(y_1,y_2) = \frac{y_1+y_2}{15}$ for $(y_1, y_2) \in [0,2]\times[0,3]$.

Recall from our previous work

$h_1^{-1}(w_1,w_2) = \frac{w_1+2w_2}{7}$, $h_2^{-1}(w_1,w_2)=\frac{3w_1-w_2}{7}$, & $J = \begin{bmatrix} 1/7 & 2/7 \\ 3/7 & -1/7\end{bmatrix}$

$\Rightarrow \det(J) = -1/7$

Now we plug in

$$f_{W_1,W_2}(w_1,w_2) = f_{Y_1,Y_2}(h_1^{-1}(w_1,w_2), h_2^{-1}(w_1,w_2))|\det(J)|$$

$$= f_{Y_1,Y_2}\left(\frac{w_1+2w_2}{7}, \frac{3w_1-w_2}{7}\right)(1/7)$$

$$= \frac{1}{7\cdot 7\cdot 15}(w_1+2w_2+3w_1-w_2)$$

$$= \frac{1}{735}(4w_1+w_2)$$

Now we find the support. By plugging in the extremes of Y_1 & Y_2 into W_1 & W_2 accordingly we get
$w_1\in[0,8]$ & $w_2 \in [-3,6]$.

The support can be more complicated

Thus

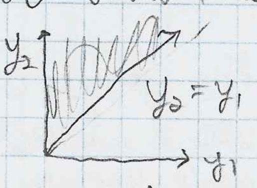
$f_{W_1,W_2}(w_1,w_2) = 1/735(4w_1+w_2)$ $\forall (w_1,w_2)\in[0,8]\times[-3,6]$.

What are the weaknesses of this method? It only gives the you the joint PDF, which sucks if you only care about 1 variable.

Example:

Suppose $f_{Y_1,Y_2}(y_1,y_2) = e^{-y_2}$ for $0 \le y_1 \le y_2 < \infty$. Use the method of Jacobians to find $W = Y_2 - Y_1$.

We draw the support of f



We need to find a second function to make this method work. Let
$W_1 = W = Y_2 - Y_1$ & $W_2 = Y_2$, ← simplifying choice (don't always pick identity)

Since $W_1 = Y_2 - Y_1$ & $W_2 = Y_2$, we know $W_1\ge0$, $W_2\ge 0$, & $W_1 \le W_2$.

We find our inverses.

$Y_1 = Y_2 - W_1 = W_2 - W_1$ & $Y_2 = W_2$

Thus

$Y_1 = h_1^{-1}(W_1,W_2) = W_2 - W_1$ & $Y_2 = h_2^{-1}(W_1,W_2) = W_2$

Doing basic derivatives we get

$J = \begin{bmatrix}-1 & 1\\ 0 & 1\end{bmatrix} \Rightarrow \det(J) = -1$

If you pick a different 2nd variable, then you get a different joint distribution. You should get the same marginal tho.

Now we substitute (next page)

$f_{W_1,W_2}(w_1, w_2) = f_{Y_1,Y_2}(h_1^{-1}(w_1, w_2), h_2^{-1}(w_1, w_2)) \, |\det(J)|$

$= f_{Y_1,Y_2}(w_2 - w_1, w_2) \, |-1|$

$= e^{-w_2}$ for $0 \le w_1 \le w_2 < \infty$

Now to get the distribution of $W_1 = W$, we find the marginal

$f_{W_1}(w_1) = \int_{w_1}^{\infty} e^{-w_2} \, dw_2 = e^{-w_1}$ for $w_1 \ge 0$.

$w_2 \ge w_1$

Order Statistics

Instead of trying to determine value, determine relative order.

Def:

Suppose $\vec{Y}$ is a set of n independent continuous rVs.
We denote the smallest observation by $Y_{(1)}$, 2nd $Y_{(2)}$, etc. Thus
$Y_{(1)} \le Y_{(2)} \le \dots \le Y_{(n)}$

We call $Y_{(k)}$ the <u>k-th order statistic</u>.

Example:

where $Y_1, \dots, Y_n$ are identically distributed & independent

If $W = h(\vec{Y}) = Y_{(n)}$ then h is function returning the maximum observation. How do we find the distribution of W? Our old methods don't work.

Let's consider the CDF $F_W(w)$

$F_W(w) = P(Y_{(n)} \le w)$

$= P((Y_1 \le w) \cap \dots \cap (Y_n \le w))$ ← all smaller

$= P(Y_1 \le w) \dots P(Y_n \le w)$ ← $Y_1, \dots, Y_n$ idependent

$= F_{Y_1}(w) \dots F_{Y_n}(w)$

$= F_Y(w)^n$

Now we find the PDF by taking the derivative

$f_W(w) = n F_Y(w)^{n-1} f_Y(w)$.

This works no matter the distribution as long as $Y_1, \dots, Y_n$ are identical & independent

The above example only works for the maximum tho. Let's go general.

Suppose $W = Y_{(k)}$ for some $Y = Y_1, \dots, Y_n$. Since $W = Y_{(k)}$, we only care about the $n-1$ other variables, in particular we need $k-1$ smaller & $n-k$ larger. This is like a multinomial of our 3 groups: lower, k, upper. We need the following distributions:

lower ($\le w$): $k-1$ variables
k: 1 variable
upper ($> w$): $n-k$ variables

Using the multinomial coefficient we get

$$\binom{n}{k-1,\ 1,\ n-k} = \frac{n!}{(k-1)!\,(1)!\,(n-k)!} = \frac{n!}{(k-1)!\,(n-k)!}$$

For any Y, the probability of being below is $P(Y \le w) = F_Y(w)$. Since we need $k-1$ independent RVs we get $\underline{F_Y(w)^{k-1}}$.

For any Y, the probability of being above is $P(Y > w) = 1 - F_Y(w)$. Since we need $n-k$ independent RVs we get $(1 - F_Y(w))^{n-k}$

We'll now skip all the actual math & go to the result which is

$$f_W(w) = \frac{n!}{(k-1)!\,(n-k)!} f_Y(w)\, F_Y(w)^{k-1} (1 - F_Y(w))^{n-k}.$$

or

$$f_{(k)}(w) = \frac{n!}{(k-1)!\,(n-k)!} f_Y(y)\, F_Y(y)^{k-1} (1 - F_Y(y))^{n-k}$$

independently identically distributed

<u>Example:</u>

Suppose $Y_1, \dots, Y_5 \overset{iid}{\sim} \text{Beta}(2,2)$. Find the distribution of the min, median, & max.

For each part we need the CDF & PDF.

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} y(1-y) = \frac{3!}{1!\,1!}(y - y^2) = 6y - 6y^2 \quad \forall y \in [0,1]$$

$$F(y) = \int_0^y 6t - 6t^2\,dt = [3t^2 - 2t^3]_0^y = 3y^2 - 2y^3 \quad \forall y \in [0,1]$$

Now we find $f_{(1)}(y)$, $f_{(3)}(y)$, $f_{(5)}(y)$ for min, median, max respectively

Min. $f_{(1)}(y) = \frac{5!}{0!\,4!}(6y - 6y^2)\cancel{(3y^2 - 2y^3)^0}(1 - 3y^2 + 2y^3)^4$

$= 5(6y - 6y^2)(1 - 3y^2 + 2y^3)^4 \quad \forall y \in [0,1]$

Med. $f_{(3)}(y) = \frac{5!}{2!\,2!}(6y - 6y^2)(3y^2 - 2y^3)^2(1 - 3y^2 + 2y^3)^2 \quad \forall y \in [0,1]$

Max $f_{(5)}(y) = \frac{5}{4!\,0!}(6y - 6y^2)(3y^2 - 2y^3)^4\cancel{(1 - 3y^2 + 2y^3)^0} = 5(6y - 6y^2)(3y^2 - 2y^3)^4 \quad \forall y \in [0,1]$

Cool! That wasn't too hard. If we picked an even n we'd have a function of 2 order statistics. Same for range. IQR can have even more. We need a way to do functions of order statistics.

Joint Distributions of (2) Order Statistics

Let $Y_1, \dots, Y_n$ be independent continuous RVs all w/ CDF $F(y)$. Then the joint distribution of the jth & kth ($j<k$) order statistics $Y_{(j)}$ & $Y_{(k)}$ is

$$f_{Y_{(j)},Y_{(k)}}(y_1,y_2) = \frac{n!}{\underset{<j}{(j-1)!}\,\underset{j<\cdot<k}{(k-j-1)!}\,\underset{>k}{(n-k)!}} f(y_1)f(y_2)F(y_1)^{j-1}(F(y_2)-F(y_1))^{k-j-1}(1-F(y_2))^{n-k}$$

This follows the same pattern of splitting into a multinomial & the PDFs of your fixed points & CDFs of the variable ones.

<u>Example:</u>
Let $Y_1, \dots, Y_n \sim u(0,1)$. Find the joint distribution of $Y_{(1)}$ & $Y_{(n)}$.
Recall $f(y) = 1 \ \forall y \in [0,1]$ & $F(y) = y \ \forall y \in [0,1]$.

This gives us joint PDF

$$f_{Y_{(1)},Y_{(n)}}(y_1,y_2) = \frac{n!}{(0)!(n-2)!(0)!}(1)(1)\cancel{(y_1)^0}(y_2-y_1)^{n-2}\cancel{(1-y_2)^0}$$

$$= n(n-1)(y_2-y_1)^{n-2} \quad \forall\, 0 \le y_1 \le y_2 \le 1$$

Note $Y_i \perp\!\!\!\perp Y_j$ but $Y_{(i)} \not\perp\!\!\!\perp Y_{(j)}$.

Now find the distribution of the range $W_1 = Y_{(n)} - Y_{(1)}$.
We use the method of Jacobians to find W_1. Let $W_2 = Y_{(n)}$.

We find the inverses

$$Y_{(1)} = h_1^{-1}(W_1, W_2) = W_2 - W_1 \quad \& \quad Y_{(n)} = h_2^{-1}(W_1, W_2) = W_2$$

Thus
$$J = \begin{bmatrix} \frac{dh_1^{-1}}{dw_1} & \frac{dh_1^{-1}}{dw_2} \\ \frac{dh_2^{-1}}{dw_1} & \frac{dh_2^{-1}}{dw_2} \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}$$

Substituting we get

$$f_{W_1,W_2}(w_1,w_2) = f_{Y_{(1)},Y_{(n)}}(h_1^{-1}(w_1,w_2), h_2^{-1}(w_1,w_2))\,|\det(J)|$$

$$= f_{Y_{(1)},Y_{(n)}}(w_2-w_1, w_2)\,\cancel{|-1|}$$

$$= n(n-1)(\cancel{w_2} - \cancel{w_2} + w_1)^{n-2}$$

$$= n(n-1)(w_1)^{n-1} \quad \forall\, 0 \le w_1 \le w_2 \le 1$$

Now we have to find $f_{W_1}(w_1)$ b/c we we asked about the distribution of the range which is W_1

$$f_{W_1}(w_1) = \int_{w_1}^{1} n(n-1)(w_1)^{n-2}\,dw_2$$

$$= (n(n-1)(w_1)^{n-2} w_2 \Big]_{w_2=w_1}^{w_2=1}$$

$$= n(n-1)w_1^{n-2} - n(n-1)w_1^{n-1}$$

$$= n(n-1)w_1^{n-2}(1-w_1) \qquad \forall w_1 \in [0,1]$$

Notice that $f_{W_1}(w_1)$ the the PDF of beta$(n-1, 2)$. Thus
$W_1 = Y_{(n)} - Y_{(1)} \sim$ beta$(n-2, 1)$