Properties of \( E(Y) \):

1. If \( Y = aX + b \) then \( E(Y) = aE(X) + b \).

2. If you can find a constant \( c \) with \( P(X = c) = 1 \) then (naturally enough) \( E(X) \geq c \); if \( b \) exists with \( P(X = b) = 1 \) then \( E(X) \leq b \).

3. If \( X_1, \ldots, X_n \) are r.v.s, each with finite \( E(X_i) \), then \( E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) \),

4. and \( E\left( \sum_{i=1}^{n} (aX_i + b) \right) = a\left( \sum_{i=1}^{n} E(X_i) \right) + b \) for all constants \( a_1, \ldots, a_n \) and \( b \).

Def. A function \( g: \mathbb{R}^n \to \mathbb{R} \) is convex means that \( g(\bar{x}) \) is convex over the real numbers, for \( \bar{x} = (x_1, \ldots, x_n) \).
if for every $0 < d < 1$ and every $x$ and $y$, $g[(d \cdot x + (1-d) \cdot y)] \leq d \cdot g(x) + (1-d) \cdot g(y)$

$g(x) = x^2$ Graphical version of this: pick any two points on the function and connect them with a line segment; the function is convex if the line segment entirely lies above the function except at the end points.

$g$ is concave if

$g[(d \cdot x + (1-d) \cdot y)] \geq d \cdot g(x) + (1-d) \cdot g(y)$
Let \( \mathbf{X} = (X_1, \ldots, X_n) \) be a random vector such that \( \mathbf{X} \sim n \). Then the expectation of \( \mathbf{X} \) is \( E(\mathbf{X}) = [E(X_1), \ldots, E(X_n)] \).

Jensen's Inequality

Suppose \( g \) is convex, \( \mathbf{X} \) a random vector with finite \( E(\mathbf{X}) \to E[\, g(\mathbf{X})] \geq g[\, E(\mathbf{X})] \).

(Without loss of generality, \( E[\, g(\mathbf{X})] = g[\, E(\mathbf{X})] \).)

(Historical note: attributed to Johan Jensen (1859 - 1922), Danish mathematician & engineer)

Application of (3)

Suppose \( X_1, \ldots, X_n \) are IID Bernoulli \((p)\).

Then \( E(X_i) = 0 \cdot (1-p) + 1 \cdot p = p \) and \( \Pr(X=0) = (1-p) \) and \( \Pr(X=1) = p \).

Then \( E(\sum X_i) = \sum E(X_i) = np = \text{mean of \( \sum X_i \)} \).
Expectation of a product when the $X_i$ are independent

Contrast this with the sum:

$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$

whether the $X_i$ are independent or not;

$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i)$ only when the $X_i$ are independent.

Example: You have a Brita water filter that you use to improve the taste of Santa Cruz water. How much better would the filter do its job if you filtered the water twice instead of once?
\( \bar{x}_1 \): proportion of bad stuff removed in the 1st filtering

\( \bar{x}_2 \): proportion removed in 2nd filtering of what was left from 1st filtering

Reasonable to assume that \( \bar{x}_1, \bar{x}_2 \) are independent; suppose they're IID with common PDF

\[
\begin{cases} 
4x^3 & 0 < x < 1 \\
0 & \text{else} 
\end{cases}
\]

(Sensible shape)

Set \( \bar{x} = \) proportion of original bad stuff remaining after 2 filtrations \( = (1 - \bar{x}_1)(1 - \bar{x}_2) \) independent

Then \( E(\bar{x}) = E[(1 - \bar{x}_1)(1 - \bar{x}_2)] = E(1 - \bar{x}_1) \cdot E(1 - \bar{x}_2) \)

\( \bar{x}_1, \bar{x}_2 \) independent \( \iff (1 - \bar{x}_1), (1 - \bar{x}_2) \) independent too

then \( E(\bar{x}) = \mu^2 \)
\[ \mu = E(1 - X_i) = \int_0^1 (1 - x_i) 4 x_i^3 \, dx_i = 0.2, \]

so 80\% of bad stuff expected to be removed in 1st filtering; \( E(\bar{X}) = \mu^2 = 0.04, \)
so expect only 4\% of bad stuff to remain after 2 filterings.

\[ \text{(b) Suppose} \]

\( X \) is a discrete rv with possible values \( 0, 1, 2, \ldots \); then \( E(X) = \sum_{n=1}^{\infty} \frac{n}{n} \).

\( \text{(b) If } \bar{X} \text{ is a continuous rv with possible values } (0, \infty), \text{ then } E(\bar{X}) = \int_0^\infty \left[ 1 - F_{\bar{X}}(x) \right] \, dx, \)

and CDF \( F_{\bar{X}}(x) \),

Example of 6 (b)

I throw a dart at a dart board repeatedly, trying to get a bullseye (success).

\( \bar{X} = \# \text{ of throw on which I first succeed.} \)
(Ex. throws $FFS' + S = 3$) Suppose that my $F =$ failure success probability is constant $S =$ success across the throws and equals $p$.

Then $E(S)$ should be inversely related to $p$.

The worse I am, the longer I expect the process to take; $E(S) =$ ?

At least 1 throw always required so $P(S \geq 1) = 1$; for $n > 1$ at least $n$ tosses required.

none of the first $(n-1)$ throws succeeded.

so $P(S \geq n) = (1-p)^{n-1}$ and geometric series

$E(S) = \sum_{n=1}^{\infty} (1-p)^{n-1} = 1 + (1-p) + (1-p)^2 + \ldots$

$= \frac{1}{1 - (1-p)} = \frac{1}{p}$ (inverse relation)

If I'm terrible ($p = .01$) I expect to succeed on the last throw.

$\text{The } \frac{1}{.01} = 100 \text{ is high.}$
Variance and standard deviation

\[ \begin{array}{ccc}
-3 & -2 & +5 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\mu (\text{mean}) \\
\end{array} \]

\[ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \]

mean \( 4 = \mu \)

\( X \) discrete RV, uniform \( \{1, 2, 9\} \); \( E(X) = 4 = \mu \)

1. How spread out is the dist. of \( X \) around its mean \( \mu \)? \( (X - \mu) \sim \text{uniform} \{ -3, -2, +5 \} \)

Could try calculating \( E(X - \mu) \), but this is 0 for any RV \( X \), because of cancellation of + and - deviations; two different easy fixes: \( E|X - \mu| \) mean absolute deviation (MAD), \( E(X - \mu)^2 \) variance of RV \( X \).

\( \text{AAD} \) not used much; variance used constantly.
Def: \( X \) rv with finite mean \( E(X) = \mu \);

variance of \( X \): \( \text{Var}(X) = \text{Var}(X) = E[(X - \mu)^2] \).

If \( E(X) = \pm \infty \) or \( E(X) \) doesn't exist, \( \text{Var}(X) \) doesn't exist.

The units are wrong: if \( X \) is in $\$, \( \text{Var}(X) \) is in $^2$.

Every fix: standard deviation of \( X \):

\[
\text{std. dev. of } X = \sqrt{\text{Var}(X)} = \sigma(X).
\]

Consequences of these definitions:

1. \( \text{Var}(X) = E[(X - \mu)^2] \)
   
   \[
   = E(X^2 - 2\mu X + \mu^2) \\
   = E(X^2) - 2\mu E(X) + \mu^2 \\
   = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2.
   \]

This is a different way to compute the variance.
\[ \text{so } \sigma(X)^2 = \left( \text{expectation of } X^2 \right) - \left( \text{square of expectation of } X \right) \]

**Example**

\[ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \rightarrow \begin{cases} 1 \\ 2 \\ 3 \end{cases} \sim \text{Uniform } \{1, 2, 3\} \]

\[ \text{mean } \mu = 2 \]

\[ E(X - \mu)^2 = \frac{1}{3} (1-2)^2 + \frac{1}{3} (2-2)^2 + \frac{1}{3} (3-2)^2 = \frac{1}{3} \]

\[ \sigma^2(X) = \sqrt{\frac{1}{3}} = 0.577 \]

This is a reasonable summary of the length of the arrows.

2. For any rv \( X \), \( \sigma(X)^2 \geq 0 \); if \( X \) is bounded, \( \sigma(X)^2 \) exists & is finite.

This is a consequence of Jensen's Inequality:

\[ g(x) = x^2 \text{ is convex } \Rightarrow E(X^2) \geq \left[ E(X) \right]^2 \]

\[ \sigma(X)^2 = E(X^2) - \left[ E(X) \right]^2 \geq 0. \]
\( V(X) = 0 \iff P(X = c) = 1 \) for some constant \( c \) (this is a trivial rv)

Notation: In the same way that, by convention, \( E(X) = \mu_X, V(X) = \sigma_X^2 \)

and \( SD(X) = \sigma_X \)

\( \therefore X \) rv, \( \sigma = a \sigma_X + b \)

\( \rightarrow V(Z) = a^2 V(X) = a^2 \sigma_X^2 \) \( \text{and} \)

\( SD(Z) = |a| \sigma_X \).

Special cases: \( a = 0 \):

\( V(X + c) = V(X) \)

\( SD(X + c) = SD(X) \)

\( V(aX) = a^2 V(X) \) \( \text{if } X, \ldots, X \)

\( b = 0 \)

\( SD(aX) = |a| SD(X) \)

are independent rv with finite means, \( V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i) \).
This is why the concept of variance (\(\sigma^2\)) has endured even though the units of the variance are wrong: for independent rvs, variance is additive, whereas \(\sigma\) is not. Special case of (5):

\[ \sigma(X_1, X_2) = \sqrt{\sigma(X_1)^2 + \sigma(X_2)^2} \]

\[ \text{SD}(X_1, X_2) \]

\[ \text{SD}(X_1 + X_2) \]

\[ \sqrt{\text{SD}(X_1)^2 + \text{SD}(X_2)^2} \]

i.e., SD grows like the hypotenuse of a right triangle.

Furthermore, \(\max\{\text{SD}(X_1), \text{SD}(X_2)\} < \text{SD}(X_1 + X_2) < \text{SD}(X_1) + \text{SD}(X_2)\).
Consequence \( \sum_{i=1}^{n} \sum_{i=1}^{n} \) independent \( n \) of each \( \alpha_i, \ldots, \alpha_n, b \) constants \( \rightarrow \)

\[
\sqrt{\left( \sum_{i=1}^{n} \alpha_i \cdot X_i \right) + b} = \sum_{i=1}^{n} \alpha_i \cdot \sqrt{V(X_i)}.
\]

**Example** \( X \sim \text{Binomial}(n, p) \); we already know that \( E(X) = np \); what about \( V(X) \) and \( SD(X) \)?

Let \( S_i = \begin{cases} 1 & \text{if success on } i^{\text{th}} \text{ success} \\ 0 & \text{else} \end{cases} \)

for \( (i=1, \ldots, n) \) and suppose as usual that \( S_1, \ldots, S_n \) are IID \( \text{Bernoulli}(p) \) —

then \( \sum_{i=1}^{n} S_i \) and we can work out its variance without difficulty.
\[ V(X) = V\left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} V(X_i) \]  

So, we need to work out the variance of a Bernoulli rv. We already know that \( E(X_i) = p \), so if we use the formula \( V(X_i) = E(X_i^2) - [E(X_i)]^2 \), we're halfway there.

Bernoulli rvs are funny:

\[ X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1-p) \end{cases} \]

So, \( E(X_i^2) = E(X_i) = p \).  

and finally:

\[ V(X_i) = E(X_i^2) - [E(X_i)]^2 = p - p^2 = p(1-p) \]
and $V(\bar{X}) = \frac{1}{n} V(S_i) = \frac{n}{n} p(1-p) = \sqrt{n p(1-p)}$.

Example: TS disease

$X = \# T-S$ babies in family of $n = 5$.

both parents carriers so $p = P(T-S) = \frac{1}{4}$

$\sim \text{Binomial}(np) = \text{Binomial}(5, \frac{1}{4})$

We already worked out that $E(X) = np = 1.25$

Now $SD(X) = \sqrt{np(1-p)} = \sqrt{5(\frac{1}{4})(\frac{3}{4})} = 0.97$

It's useful to summarize this by saying "The number of T-S babies this couple will have will be around 1.25, give or take about $\frac{1}{2} \sigma_x = 0.49$"
Example \( X \sim \text{Cauchy} \)

Earlier we defined the quantiles on any distribution. But we can't use the idea of the quantiles on any distribution.

If the distribution of \( X \) has a variance, then \( \text{Var}(X) \) does exist.

So clearly \( \text{Var}(X) \) doesn't exist. Therefore \( \text{Var}(X) \) doesn't exist.

\[
f_X(x) = \begin{cases} 
    a(x+b) & \text{if } x > 0 \\
    0 & \text{otherwise}
\end{cases}
\]
The Cauchy CDF is:

\[ F_X(x) = \int_\infty^x \frac{1}{\pi(1+t^2)} \, dt \]

The arctangent function is what's called the principal inverse of \( \tan(x) \), varying from \(-\frac{\pi}{2}\) to \(\frac{\pi}{2}\) as \(-\infty < x < \infty\)

\[ F_X(x) = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi} \]

So, the IDF for the Cauchy distribution is

\[ \text{IDF}_X(x) = \tan^{-1}(\frac{x}{\pi}) - \frac{1}{\pi} \]

Thus,

\[ \text{IDF}_X(x) = \tan^{-1}(\frac{3}{4}) - \tan^{-1}(\frac{1}{4}) \]

\[ = \tan\left(\frac{3\pi}{4}\right) - \tan\left(-\frac{\pi}{4}\right) \]

\[ = 2 \]
Cauchy PDF Moments of a rv

\[ E(X) = E(X^2) \]
\[ \sqrt{\text{Var}(X)} = E(X^2) - \left[ E(X^2) \right]^2 \]

with the usual mathematical impulse to generalize:

**Def.** \( X \) rv, \( k \) integer \( \geq 1 \) →

\[ E(X^k) = \text{the } k^{th} \text{ moment of } X \]

Of course \( E(X^k) \) may not exist, and if it does it may be infinite, but the idea is still useful. You can show that (kth moment of \( X \) exists) \( \leftrightarrow \ E(|X|^k) < \infty \)
Consequence of the moment definition:

1. If \( E(|X|^k) < \infty \) for some integer \( k \geq 1 \), then \( E(|X|^j) < \infty \) for all integers \( j < k \).

In other words, if the \( k \)th moment of \( X \) exists, so do the \((k-1)\)st, \((k-2)\)nd, \ldots, moments.

Definition: \( X \) rv with expectation \( E(X) = \mu \), \( k \) integer \( \geq 1 \) \( \rightarrow \) \( E[(X-\mu)^k] \) is called the \( k \)th central moment of \( X \) or the \( k \)th moment of \( X \) around its mean. Clearly this idea generalizes the variance of \( X = E[(X-\mu)^2] \).
\[ E[(X - \mu)^2] = E(X) - \mu = \mu - \mu = 0, \]

i.e., every rv has 2nd central moment 0.

3) If \( X \) is symmetric around \( \mu \), then \( E[(X - \mu)^k] = 0 \) for all odd integers \( k \) for which \( E[(X - \mu)^k] \) exists.

This motivates a new definition:

Let \( X \) rv with mean \( \mu \), \( \mu \in \mathbb{R} \);

if the third moment of \( X \) exists and is finite, then skewness \( (X) = \frac{\text{E}\left(\frac{X - \mu}{\sigma_X}\right)^3}{\text{E}\left(\frac{X - \mu}{\sigma_X}\right)} \).

All symmetric distributions with finite 3rd moment have skewness 0.
Def. \( X \) rv, \( t \) a real number.

\[ \psi(t) = E(e^{tX}) \text{ is called the moment generating function of } X \]

The reason for this definition.

Theorem. \( X \) rv with MGF \( \psi_X(t) \), finite for all values of \( t \) in an open interval \((-a, b)\) around \( 0 \) \((a > 0)\).

Then for all integers \( n > 0 \),

\[ E(X^n) = \frac{d^n}{dt^n} \psi_X(t) \bigg|_{t=0} \]

is the derivative of \( \psi_X \), evaluated at \( t = 0 \).
This is a handy theorem: if its premise is satisfied & the calculations are manageable, you get all the moments of \( X \) just by computing \( \psi_X(t) \) and differentiating it over & over. Example

\( X \sim \text{Exponential}(\lambda) \)

\[
\begin{aligned}
f_X(x) &= \begin{cases} 
\lambda e^{-\lambda x}, & x \geq 0 \\
0 & \text{else}
\end{cases} \\
\end{aligned}
\]

\[
\begin{aligned}
\psi_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} \, dx = \lambda \int_0^\infty e^{(t-\lambda)x} \, dx \\
\end{aligned}
\]

Now this integral is finite only if \( t - \lambda < 0 \), is for \( t < \lambda \), but

This means (since \( \lambda > 0 \))

that it's definitely finite in an open interval around 0 (e.g. \((-\lambda, \lambda))\).
So \( \psi(t) \) exists for \( t < 0 \) and equals

\[
\psi(t) = 2 \int_0^\infty e^{(t-2)x} \, dx = \frac{2}{2-t}.
\]

Now we just crank out the derivatives:

\[
E(x^4) = \left( \frac{d^4}{dt^4} \left( \frac{2}{2-t} \right) \right)_{t=0} = \frac{24}{2^4}.
\]

\[
E(x^3) = \left( \frac{d^3}{dt^3} \left( \frac{2}{2-t} \right) \right)_{t=0} = \frac{6}{2^3}.
\]

\[
E(x^2) = \left( \frac{d^2}{dt^2} \left( \frac{2}{2-t} \right) \right)_{t=0} = \frac{2}{2^2}.
\]

\[
E(x) = \left( \frac{d}{dt} \left( \frac{2}{2-t} \right) \right)_{t=0} = \frac{1}{2}.
\]

So \( \psi(\infty) = E(x^2) - (E(x)^2) \) and

\[
\text{Evident \( \text{E}(x^k) = \frac{k!}{2^k} \cdot \binom{\frac{k}{2}}{\frac{k}{2}} \).}
\]