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. \[ \Psi_{X}(t) = \frac{f_{X}(x)}{t} \]

Example

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

\[ f_{X}(x) = \begin{cases} 2e^{-2x}, & x \geq 0 \\ 0, & \text{else} \end{cases} \]

\[ \Psi_{X}(t) = \mathbb{E}(e^{tX}) = \int_{0}^{\infty} e^{tx} 2e^{-2x} \, dx = 2 \int_{0}^{\infty} e^{(t-2)x} \, dx \]

Now this integral is finite only if \( t-2 < 0 \) \[ \rightarrow t < 2 \]

This means (since \( \lambda > 0 \)) that it's definitely finite on an open interval around 0 (e.g., \((-2, 2)\)).
So \( \psi(t) \) exists for \( t < \lambda \) and equals
\[
\psi(t) = \frac{2}{\lambda} \int_0^\infty e^{(t - \lambda)x} dx = \frac{1}{2 - t}.
\]

Now we just crank out the derivatives:
\[
E(X^2) = \left( \frac{d}{dt} \frac{2}{2 - t} \right) \bigg|_{t=0} = \frac{1}{2} \quad \text{So } V(X) = \frac{E(X^2) - (E(X))^2}{E(X)}
\]
\[
E(X^2) = \left( \frac{d^2}{dt^2} \left( \frac{2}{2 - t} \right) \right) \bigg|_{t=0} = 2 \quad \text{and } \text{SD}(X) = \frac{1}{\lambda^2}
\]
\[
E(X^3) = \left( \frac{d^3}{dt^3} \left( \frac{2}{2 - t} \right) \right) \bigg|_{t=0} = 6
\]
\[
E(X^4) = \left( \frac{d^4}{dt^4} \left( \frac{2}{2 - t} \right) \right) \bigg|_{t=0} = \frac{24}{\lambda^4}
\]

Evidently \( E(X^k) = \frac{k!}{\lambda^k} \).
Consequence of the MGF definition

1. \( X \) rv with MGF \( \phi_{X}(t) \)

\[ Z = aX + b \quad (a, b \text{ constants}) \]

Then at every value of \( t \) for which \( \phi_{Z}(at) \) is finite,

\[ \phi_{Z}(t) = e^{bt} \phi_{X}(at). \]

Example

\( Z \sim \text{Binomial} \left( n, p \right) \), \( X = \sum_{i=1}^{n} S_{i} \),

\( S_{i} \sim \text{Bernoulli} \left( p \right) \)

\( i = 1, \ldots, n \)

MGF of \( S_{i} \)

is easy: \( \phi_{S_{i}}(t) = E(e^{tS_{i}}) \)

\[ = e^{tp} \cdot p(S_{i} = 1) + e^{0} \cdot p(S_{i} = 0) \]

This uses the Law of the unconscious statistician (LotUS)

\[ = \left[ p e^{t} + (1-p) \right] \]
\( x_1, \ldots, x_n \) independent \( \mathbb{N}, \mathbb{N} \) of \( \xi \), \( \psi_i(t), \xi = \sum_{i=1}^{n} \xi_i, \)

MGF of \( \xi \) is \( \psi_\xi(t) \) + for every \( t \) such that \( \psi_\xi(t) \) is finite for all \( \xi \).

\( i = 1, \ldots, n, \quad \psi_i(t) = \prod_{i=1}^{n} \psi_i(t) \).

MGF of \( \text{Binomial} \) (18/Any 12)

Since the \( \xi_i \) are \( \text{IID} \),

\( \psi(t) \equiv \prod_{i=1}^{n} \psi_i(t) \).

Now, as before, we just crank out the derivatives:

\( \prod_{i=1}^{n} \left[ p e^t + (1-p) \right] \).
\[ E(\mathbf{X}) = \left( \frac{\partial^2}{\partial t^2} \mathbf{X}(t) \right) \bigg|_{t=0} = \frac{\partial}{\partial t} \left[ \mathbf{p} e^t + (1-\mathbf{p}) \right] \bigg|_{t=0} = \mathbf{y} \mathbf{p} \]}

\[ E(\mathbf{X}^2) = \left( \frac{\partial^2}{\partial t^2} \mathbf{X}(t) \right) \bigg|_{t=0} = \mathbf{y} \mathbf{p} \left( 1 + (n-1) \mathbf{p} \right) \]

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

\[ = \mathbf{y} \mathbf{p} + n(n-1) \mathbf{p}^2 - \mathbf{y} \mathbf{p}^2 \]

\[ = \mathbf{y} \mathbf{p} + 2 \mathbf{y} \mathbf{p} - \mathbf{y} \mathbf{p}^2 - \mathbf{y} \mathbf{p}^2 \]

\[ = \mathbf{y} \left( \mathbf{p} - \mathbf{p}^2 \right) = \mathbf{y} \mathbf{p} (1-\mathbf{p}) \]

\[ E(\mathbf{X}^3) = \left( \frac{\partial^3}{\partial t^3} \left[ \mathbf{p} e^t + (1-\mathbf{p}) \right] \right) \bigg|_{t=0} = \mathbf{y} \mathbf{p} \left[ (n-2)(n-1) \mathbf{p}^2 + 3\mathbf{p}^2 n(n-1) \right] \]
\[ X \text{ has MGF } \Psi_X(t), \text{ finite in an open interval around } t = 0 \]

iff I, I have then \( \Psi_X(t) = \Psi_I(t) \leftrightarrow \text{identical probability distributions} \)

So the MGF (if it exists) uniquely characterizes a random variable.

We've already made some contrasts versus between the mean and the median of a distribution; here are 2 more things worth saying:

1. If RV with values in an interval I,
   \( h(x) \) 1-1 function on I, \( Y = h(X) \);
if a question is fool

what should you do? How can you tell

not in general true if the mean

\[ \text{median of } X \text{ is } \text{median of } Z \text{, then } h(x) = x \text{ is } \]

\[ \text{if } w = \frac{1}{Z} \text{ then } h(x) = \frac{1}{x}. \text{ This is} \]

\[ E(\mathbb{E}(Z)) = \mathbb{E}(E(Z)) \text{ unless } \mathbb{E}(x) = a + b \]

before it is observed suppose your job

1 if not no call is my best guess so
Let's say you pick the number $\hat{x}$ - (a fixed, known constant) before $X$ is observed.

Then, after $X$ arrives, your prediction error would be $X - \hat{x}$ which might be either positive or negative. One possible criterion for goodness would be to find $\hat{x}$ such that $E(X - \hat{x}) = 0$.

**Def** The bias of $\hat{x}$ as a predictor for $X$ is $bias(\hat{x}) = E(X - \hat{x})$.

**Def** Your prediction $\hat{x}$ is unbiased if $bias(\hat{x}) = 0$. Clearly, to achieve this just choose $\hat{x} = E(X)$. 
Another possible criterion for goodness would be to find $\hat{x}$ such that $E(\hat{x} - \bar{x})^2$ is small. (Gauss)

\[ \text{Def: } E[(\hat{x} - \bar{x})^2] \text{ is called the mean squared error (MSE) of } \hat{x} \text{ as a prediction for } \bar{x}. \]

The $\hat{x}$ that minimizes MSE is $\hat{x} = E(\bar{x})$.

Small proof:

\[ E[(\hat{x} - \bar{x})^2] = \underbrace{E(\hat{x}^2 - 2\hat{x}\bar{x} + \bar{x}^2)}_{\text{small}} \]

\[ = \bar{x}^2 - 2\bar{x}E(\bar{x}) + E(\bar{x}^2) \]

This is a quadratic function of $\hat{x}$,

\[ \frac{d}{d\hat{x}} E[(\hat{x} - \bar{x})^2] = 2\hat{x} - 2\bar{x}E(\bar{x}) = 0 \]

if $\hat{x} = E(\bar{x})$.

\[ \frac{d^2}{d\hat{x}^2} E[(\hat{x} - \bar{x})^2] = 2 > 0 \]

So $E(\bar{x})$ is a minimum.
Also easy to show: \[\text{MSE}(\hat{x}) = E(\hat{x} - \hat{x})^2 = \sigma_x^2 \]

So the choice \( \hat{x} = E(\hat{x}) \) both minimizes MSE(\( \hat{x} \)) and achieves 0 bias, and with this choice \( \text{MSE}(\hat{x}) = \sigma_x^2 \).

A different criterion for a good prediction \( \hat{x} \) would be to find \( \hat{x} \) such that \( E[|\hat{x} - \hat{x}|] \) is small. \( E[|\hat{x} - \hat{x}|] \) is called the mean absolute error (MAE) of \( \hat{x} \) as a prediction for \( \hat{x} \).
Another small theorem

Let \( m \) be (the) median of \( X \), where \( m \) is the \( x \) that minimizes \( \text{MAE}(x) \).

is (the) median \( m \).

Reminder: why?

Careful definition of median

If \( rv \) is every number \( m \) such that

\[ P(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq m) \geq \frac{1}{2} \]

is a median of the dist. of \( X \).

Example of nonunique median

All \( 2 \leq x \leq 3 \) have \( F_X(x) = \frac{1}{2} \).

\( X \) discrete on \( \{1, 2, 3, 4\} \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( P(X=x) )</th>
<th>( F_X(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

CDF \( F_X(x) \)

Most people would say median = 2.5
Which is a better criterion, MSE or MAE? There is no right answer to this question: it depends on the real-world consequences of your prediction errors $(\hat{y} - y)$; quantifying these consequences involves the creation of a utility function briefly which we'll examine later.

Independence of 2 or more RVs is a special case of a more general reality in which (your uncertainty about something) and (your uncertainty about something else) are related. Let's see how to quantify such relationships.
Def. \( X, Y \) rv with finite means \( \mu_X \) and \( \mu_Y = E(Y) \). The covariance of \( E(X) \) and \( Y \), written \( \text{cov}(X,Y) \), is defined as

\[
\text{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)],
\]

as long as this expectation exists.

Consequence of this definition:

\[
1. \quad (\mu_X - E(Y))(E(X) - \mu_X) = E(XY) - E(X)E(Y) + \mu_X \mu_Y
\]

So \( \text{cov}(X,Y) = E(XY) - E(X)E(Y) + \mu_X \mu_Y \)
2) Sufficient condition for \( C(x, z) \to 0 \) to exist: \( \frac{x^2}{\varepsilon} < \infty \) and \( \frac{z^2}{\varepsilon} < \infty \). 

Covariance is a good start at measuring strength of relationship, but it has a big flaw: its value depends on the units of measurement of \( x \) and \( z \).

Example: \( x \) = education level (years of schooling completed), \( z \) = yearly income ($)

\( x \) = temperature in °C, \( z \) = relative humidity (%)

If you change your mind & measure temperature in °F = \( \frac{9}{5} \) °C + 32,

\( C(x', z) = C(\frac{9}{5}x + 32, z) \neq C(x, z) \)
Easy to show that if $a$, $d$ are constants then $c(x+y, z) = a c(x, z)$. So
\[ c(x', z) = 1.8 \cdot c(x, z), \] i.e., you can make the association between temperature & relative humidity seem linear just by switching from °C to °F!

Easy fix:

**Def**: The process of converting a rv $X$ to standard units (SU) is achieved with the linear transformation $X' = \frac{X - E(X)}{SD(X)}$. (as long as $\text{std}(x) < \infty$, this is a meaningful definition)

$E(X') = 0$, $V(X') = 1 = SD(X')$
Def. If $X, Y$ have finite variances, then

$$\sigma_X^2 \text{ and } \sigma_Y^2 \quad \text{(and therefore finite means } \mu_X \text{ and } \mu_Y) \rightarrow \text{ the correlation of } X, Y \equiv \rho (X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

and $\rho$ is

$$\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

With this definition, the correlation is invariant to linear transformation of either variable (i.e., for any constants $a, c \geq 0$ and $b, d$,

$$\rho(aX + b, cY + d) = \rho(X, Y)$$

(If $a < 0$, $\rho(aX + b, Y) = -\rho(X, Y)$.)
Consequences of the correlation definition:

1. **Cauchy-Schwarz inequality**:
   For all random variables $X, Y$ for which $E(XY)$ exists, $\left( E(XY) \right)^2 \leq \left( E(X)^2 \right) \cdot \left( E(Y)^2 \right)$.

   From which, $\left( \rho(X,Y) \right)^2 \leq \frac{\sigma_X^2 \cdot \sigma_Y^2}{E(X^2) \cdot E(Y^2)}$.
   
   
   \[ -1 \leq \rho(X,Y) \leq 1 \]

   Karl Schwar (1843 - 1921), German mathematician
   (associated)

\[ \rho(X,Y) > 0 \iff X, Y \text{ positively correlated} \]

\[ \rho(X,Y) < 0 \iff X, Y \text{ negatively correlated} \]

\[ \rho(X,Y) = 0 \iff X, Y \text{ uncorrelated} \]

2. $X, Y$ independent r.v. with $\{ X \sim N(0, \sigma_X^2), Y \sim N(0, \sigma_Y^2) \}$

   $\rightarrow \rho(X,Y) = 0$
So independence implies correlation, but (interestingly) not the converse:

**Example:**

Let $X \sim \text{Uniform} \{-1, 0, 1\}$, $E(X) = 0$

$X$ and $X^2$ are clearly dependent since $X$ completely determines $X^2$, but $E(X^2) = E(X^3) = E(X) = 0$

(since $X$ and $X^3$ are identically distributed) and thus

$$C(X, X^2) = E(X \cdot X^2) - E(X) \cdot E(X^2) = 0$$

$$\rho(X, X^2) = \frac{C(X, X^2)}{\sigma_X \sigma_{X^2}} = 0$$

so $X$ and $X^2$ are uncorrelated!

### Additional Note

$\sigma_X^2 \leq \sigma_X^2 < \infty$, $\sigma_X = a \sqrt{b}$

for $a \neq 0$, constants $a > 0$ and $b > 0$.
\( (x > 0) \Rightarrow p(x, y) = 1 \) so \( p(x, y) \) measures the strength of linear association between \( X \) and \( Y \).

Important:

If \( X, Y \) rv, \( \sigma_X^2 < \infty, \sigma_Y^2 < \infty \), then

\[
\sqrt{\text{Var}(X + \beta Y)} = \sqrt{\text{Var}(X)} + \sqrt{\text{Var}(\beta Y)} + 2 \sqrt{\text{Cov}(X, \beta Y)}
\]

\( \text{(direct data science formula)} \)

\( \text{Special case:} \)

\[
\sqrt{\text{Var}(X - \beta Y)} = \sqrt{\text{Var}(X)} + \sqrt{\text{Var}(\beta Y)} - 2 \sqrt{\text{Cov}(X, \beta Y)}
\]
6. \( \mathbb{E}_i, \ldots, \mathbb{E}_n \) such that \( (\mathbb{E}_i, \mathbb{E}_j) \) uncorrelated for all \( 1 \leq i, j \leq n \) → \( \sqrt{\sum_{i=1}^{n} \mathbb{E}_i^2} = \sum_{i=1}^{n} \sqrt{\mathbb{E}_i^2} \) (then)

7. \( \rho(\mathbb{E}, \mathbb{E}) = +1 \)

---

Case 1

points in scatterplot sample from \( f_{\mathbb{E}, \mathbb{E}}(x, y) \) all fall on line with negative slope (necessarily)

Case 2

points in scatterplot sample from \( f_{\mathbb{E}, \mathbb{E}}(x, y) \) outlier

Case 3

non-linearity
Conditional Expectation

If $X, Y$ are related random variables (not necessarily independent), then there is information in $X$ for predicting $Y$; i.e., we should be able to find some function $d: \mathbb{R} \to \mathbb{R}$ such that $d(X)$ is "close" in some sense to $Y$ — what is the optimal $d$?

Galton example:

Galton divided the elliptical scatterplot up into a bunch of vertical strips, e.g., the one over $x_i^*$ or the other one over $x_2^*$. The points in the vertical strip over $x_2^*$ are a random sample from the conditional...
distribution of } \mathbb{Y} \text{ given } X = x^*_2, \quad f_{Z|X=x^*_2} (y | x=x^*_2)

Galton knew about the small theorem.

Fact on p. 107: the number \( \hat{w} \) that minimizes \( \text{(MSE)} \)
the mean square error \( E[(\hat{w} - \hat{\theta})^2] \) of \( \hat{w} \)
as a prediction for \( \hat{\theta} \) is \( \hat{w} = E(\hat{\theta}) \).

So he adopted MSE as his measure of "close"
and concluded that the \( \hat{\theta} \) that minimizes
the MSE \( E[(\hat{\theta} - \hat{\theta})^2] \) in the vertical strip
defined by \( x = x^*_2 \) must be the conditional
mean, or conditional expectation, of the
\[ \mathbb{W} (\mathbb{Y} | X = x^*_2) \] \[ \text{Def.} \mathbb{W}, \mathbb{X}, \mathbb{Z}, \text{finite mean} \]
\[ \{ \text{conditional expectation} \} \quad \mathbb{E}(\mathbb{W} | x) \text{ is just} \]
\[ \{ \text{mean} \text{ of } \mathbb{Z} \text{ given } X = x \} \]
The expectation of the conditional distribution gives \( X \),

\[
\frac{f(y|x)}{f_{X|Y}(y|x)} = \frac{1}{f_{X|Y}(x)} \int f(y|x) dy
\]

for continuous \((Y|X=x)\)

\[
E(Y|X=x) = \sum_y f_{Y|x}(y|x) y
\]

for discrete \((Y|X=x)\)

So far, \( E(Y|X) \) is just a constant, equal to the conditional mean of \( Y \) when \( X \) is \( x \).

Def. \( h(x) = E(Y|X=x) \)

Then the rv \( E(Y|X) = h(X) \) is the conditional expectation of \( Y \) given \( X \). \( \text{(21)} \)