DS (and some other people) call Method 2 the law of the unconscious statistician, because Method 2 looks like a definition but is actually a (difficult) theorem (16 Aug 17) (in full generality) (measure theory: pushforward measure, ...)

Example $X \sim \text{Exponential}(\lambda)$, $(\lambda > 0)$, $E(X) = \frac{1}{\lambda}$

$Y = X^2$

$$E(Y) = \int_0^\infty x^2 \lambda e^{-\lambda x} \quad \text{(integrate by parts twice)} \quad = \frac{2}{\lambda^2}$$

Notice that

$$E(X^2) \neq [E(X)]^2$$

$$\frac{2}{\lambda^2} \neq \left(\frac{1}{\lambda}\right)^2$$

the only functions $Y = h(X)$ for which $E[h(X)] = h[E(X)]$ are linear: $h(x) = a + bx$, as we'll see later

Properties of $E(\underline{Y})$

① If $\underline{Y} = a\underline{X} + b$ then

$E(\underline{Y}) = aE(\underline{X}) + b$ (assuming $E(\underline{X})$ exists)

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

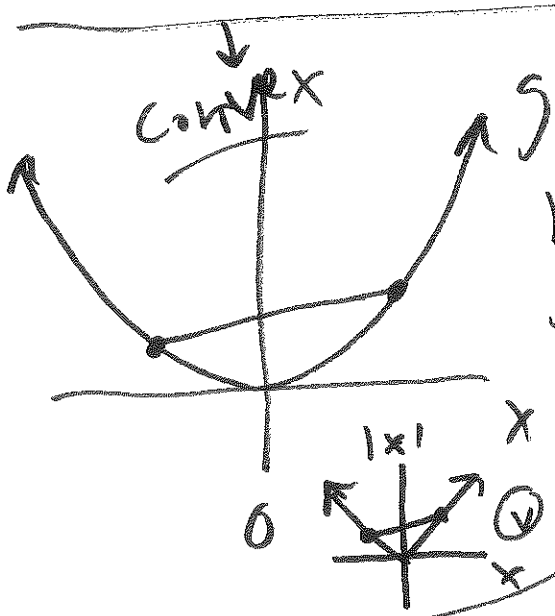
③ If $\underline{X}_1, \ldots, \underline{X}_n$ are n rvs, each with finite $E(\underline{X}_i)$, then $E\left(\sum_{i=1}^{n} \underline{X}_i\right) = \sum_{i=1}^{n} E(\underline{X}_i)$,

④ and $E\left[\sum_{i=1}^{n} (a_i \underline{X}_i + b)\right] = \sum_{i=1}^{n} a_i E(\underline{X}_i) + b$

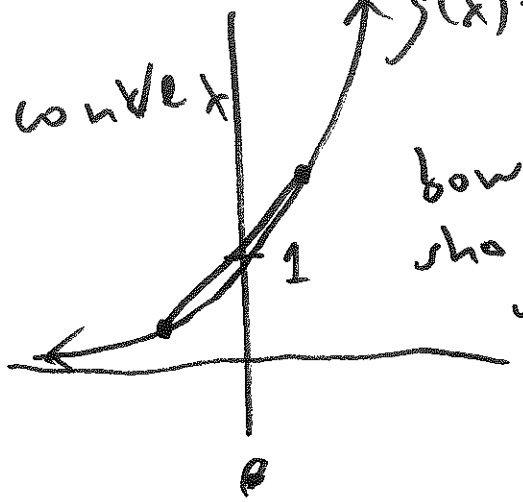
for all constants $(a_1, \ldots, a_n)$ and b.

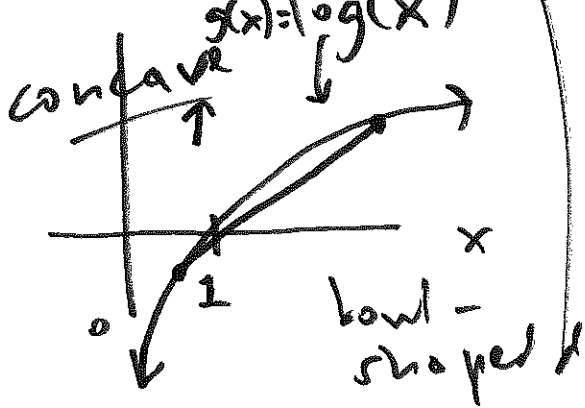
⑤ Def. A function $g: \mathbb{R}^n \to \mathbb{R}$ (this means that $g(\underset{\sim}{x}) = z$, with $\underset{\sim}{x} = (x_1, \ldots, x_n)$ real #s) is convex

if for every $0 < \alpha < 1$ and every $\underset{\sim}{x}$ and $\underset{\sim}{y}$, $g[\alpha \underset{\sim}{x} + (1-\alpha)\underset{\sim}{y}] \leq \alpha g(\underset{\sim}{x}) + (1-\alpha) g(\underset{\sim}{y})$



Graphical version of this: pick any two points on the function & connect them with a line segment; the function is convex if the line segment lies entirely above the function except at the endpoints.





g is concave if

$g[\alpha \underset{\sim}{x} + (1-\alpha)\underset{\sim}{y}] \geq \alpha g(\underset{\sim}{x}) + (1-\alpha) g(\underset{\sim}{y})$

Def. The expectation of a random vector $\underset{\sim}{X} = (X_1, \ldots, X_n)$ is $E(\underset{\sim}{X}) \triangleq [E(X_1), \ldots, E(X_n)]$ (n).

(3) (a) g convex, $\underset{\sim}{X}$ random vector with finite $E(\underset{\sim}{X}) \rightarrow E[g(\underset{\sim}{X})] \geq g[E(\underset{\sim}{X})]$. (Jensen's Inequality)

(b) g concave $\rightarrow E[g(\underset{\sim}{X})] \leq g[E(\underset{\sim}{X})]$.

(attributed to Johan Jensen (1859-1925), Danish mathematician & engineer)

Application of (3)

Suppose that $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p).

Then $E(X_i) = 0 \cdot (1-p) + 1 \cdot p = p$ and

($P(X=0)$, $P(X=1)$)

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np = \text{mean of Binomial}(n,p)$$

($\sum_{i=1}^{n} X_i \sim$ Binomial(n,p))

Expectation of a product when the X_j are independent

$X_1, \ldots, X_n$ independent rv each with finite $E(X_i)$ →

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$$

Contrast this with a sum: $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$ whether the X_i are independent or not;

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

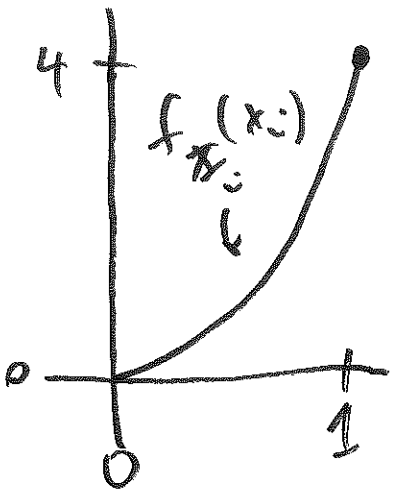
Example You have a (Brita) water filter that you use to improve the taste of Santa Cruz water. How much better would the filter do its job if you filtered the water twice instead of once?

X_1 = proportion of bad stuff removed in the 1st filtering

X_2 = proportion removed in 2nd filtering of what was left from 1st filtering

Reasonable to assume that X_1, X_2 are independent; suppose they're IID with common PDF

$$f_{X_i}(x_i) = \begin{cases} 4x^3 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$



(sensible shape)

Set Y = proportion of original bad stuff remaining after 2 filtrations $= (1-X_1)(1-X_2)$

Then $E(Y) = E[(1-X_1)(1-X_2)] \overset{\text{independence}}{=} E(1-X_1) \cdot E(1-X_2)$

X_1, X_2 independent $\leftrightarrow (1-X_1), (1-X_2)$ independent too

$E(1-X_1) \overset{\text{identical distribution}}{=} E(1-X_2) \triangleq \mu$;

then $E(Y) = \mu^2$.

$\mu = E(1-X_i) = \int_0^1 (1-x_i)\, 4x_i^3\, dx_i = 0.2$,

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

(6) (a) Suppose X is a discrete rv with possible values 0, 1, 2, ...; then $E(X) = \sum_{n=0}^{\infty} P(X \geq n)$.

(b) If X is a continuous rv with possible values $(0, \infty)$, then $E(X) = \int_0^{\infty} [1 - F_X(x)]\, dx$, and CDF $F_X(x)$.

Example of 6(a)

I throw a dart at a dartboard repeatedly, trying to get a bullseye (success). X = # of throw on which I first succeed.

(Ex. throws FFS → $\bar{x}=3$ / F = failure, S = success) Suppose that my success probability is constant across the throws and equals p, & throws are independent. Then $E(\bar{x})$ should be inversely related to p: the worse I am, the longer I expect the (1st draft) process to take; $E(\bar{x})=?$

At least 1 throw always required so $P(\bar{x}\geq 1)=1$; for $n>1$

(at least n tosses required) $\leftrightarrow$ (none of the first $(n-1)$ throws succeeded)

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

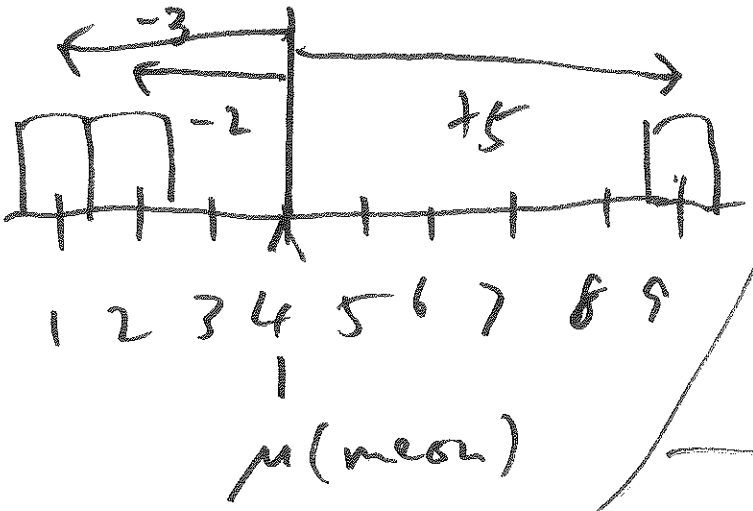
$$E(\bar{x})=\sum_{n=1}^{\infty}(1-p)^{n-1}=1+(1-p)+(1-p)^2+\dots$$

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

(inverse relation ✓)

e.g. If I'm terrible ($p=.01$) I expect to succeed on the $\frac{1}{.01}=100^{\text{th}}$ throw.

Variance and standard deviation



$\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_9 \end{bmatrix}$

mean $\xi_4 = \mu$

X discrete rv, Uniform $\{1, 2, 9\}$; $E(X) = 4 = \mu$

Q: How spread out is the dist. of X around its mean μ?

$(X - \mu) \sim$ Uniform $\{-3, -2, +5\}$ ← deviation from μ

Could try calculating $E(X - \mu)$, but this is 0 for any rv X, because of cancellation of ⊕ and ⊖ deviations; two different easy fixes: $E|X - \mu| \triangleq$ mean average absolute deviation (AAD) (MAD) (Laplace)

or $E(X - \mu)^2 \triangleq$ variance of rv X. (Gauss)

AAD not used much; variance used constantly.

Def X rv with finite mean $E(X) = \mu$;

variance of X = $V(X) \triangleq E\left[(X-\mu)^2\right]$.

Df we Var(X)

If $E(X) = \pm\infty$ or $E(X)$ doesn't exist, $V(X)$ doesn't exist.

One problem with variance

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

Easy fix: standard deviation of X $\triangleq \sqrt{V(X)} \triangleq SD(X)$.

Consequences of the definitions

① $V(X) = E\left[(X-\mu)^2\right]$

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

$= E(X^2) - 2\mu \underbrace{E(X)}_{\mu} + \mu^2$

$= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$.

this is a different way to compute the variance

so $V(X) = \left(\text{expectation of } X^2\right) - \left(\text{square of expectation of } X\right)$.

Toy example: $\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ mean $\mu = 4$

$X \sim \text{Uniform}\{1, 2, 9\}$

$E(X-\mu)^2 = \frac{1}{3}(1-4)^2 + \frac{1}{3}(2-4)^2 + \frac{1}{3}(9-4)^2 = 12.7$

$= V(X)$.

$SD(X) = \sqrt{12.7} \doteq 3.6$

This is a reasonable summary of the length of the arrows

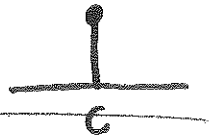
② For any rv X, $V(X) \geq 0$; if X is bounded, $V(X)$ exists & is finite.

This is a consequence of Jensen's Inequality:

$g(x) = x^2$ is convex so $E(X^2) \geq [E(X)]^2$,

i.e. $V(X) = E(X^2) - [E(X)]^2 \geq 0$.

③ $V(X) = 0 \iff P(X = c) = 1$ for some constant c



(this is a trivial rv)

Notation In the same way that, by convention, $E(X) = \mu_X$, $V(X) \triangleq \sigma_X^2$ and $SD(X) \triangleq \sigma_X$ (lower-case sigma)

④ X rv, $Y = aX + b$

$\rightarrow V(Y) = a^2 V(X) = a^2 \sigma_X^2$ and

$SD(Y) = |a| \sigma_X$. (for any constants a, b)

Special cases $a = 1$: $V(X + c) = V(X)$

$SD(X + c) = SD(X)$

$V(aX) = a^2 V(X)$

$(b = 0)$ $SD(aX) = |a| SD(X)$

⑤ If $X_1, \ldots, X_n$ are independent rv with finite means, $V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$.

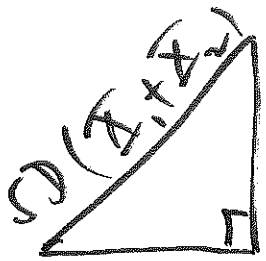
This is why the concept of variance has endured even though the units of the variance are wrong: for independent rvs, variance is additive, whereas SD is not. (wrong units)

name of SD: Karl Pearson (1890)

special case of (5):

X_1, X_2 independent $\Rightarrow V(X_1 + X_2) = V(X_1) + V(X_2)$

$= [SD(X_1)]^2 + [SD(X_2)]^2$



$SD(X_1 + X_2) =$

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

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

Immediately, $\max\{SD(X_1), SD(X_2)\} < SD(X_1 + X_2) < SD(X_1) + SD(X_2)$ (indep)

Consequence of (5) $X_1, \ldots, X_n$ independent rv, $a_1, \ldots, a_n, b$ constants $\rightarrow$

$$V\left[\left(\sum_{i=1}^{n} a_i X_i\right) + b\right] = \sum_{i=1}^{n} a_i^2 V(X_i).$$

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

Let $S_i = \begin{cases} 1 & \text{if success on } i^{\text{th}} \text{ success/failure trial} \\ 0 & \text{else} \end{cases}$

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

then $X = \sum_{i=1}^{n} S_i$ and we can work out its variance without difficulty.

$$V(X) = V\left(\sum_{i=1}^{n} S_i\right) \overset{\text{independence}}{=} \sum_{i=1}^{n} V(S_i)$$

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

Bernoulli rvs are funny: $S_j = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{---------- } (1-p) \end{cases}$

so $S_j^2 = \begin{cases} 1^2 = 1 & \text{with probability } p \\ 0^2 = 0 & \text{---------- } (1-p) \end{cases}$

so $E(S_j^2) = E(S_j) = p$ and finally

$$V(S_j) = E(S_j^2) - [E(S_j)]^2 = p - p^2 = p(1-p)$$

and $V(\bar{X}) = \sum_{i=1}^{n} V(S_i) = \sum_{i=1}^{n} p(1-p) = \boxed{np(1-p)}$

and $SD(\bar{X}) = \sqrt{np(1-p)}$

Example: T-S disease

X = (# T-S babies in family of $n=5$, both parents carriers so $p = P(\text{T-S baby}) = \frac{1}{4}$)

$\sim$ Binomial(n, p) = Binomial$(5, \frac{1}{4})$

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

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

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 1" $\leftarrow \sigma_X$

$\uparrow \mu_X$

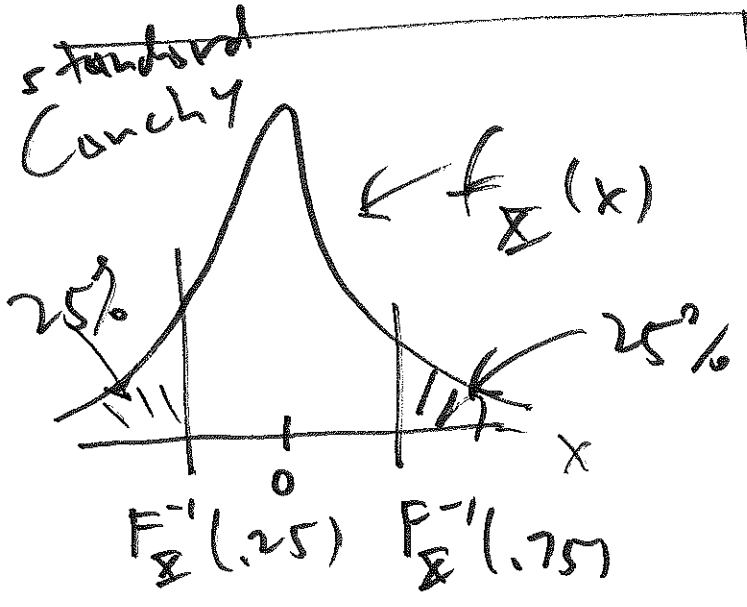
How do you measure the spread of a distribution if the variance doesn't exist?

Example $X \sim$ (standard) Cauchy

$$f_X(x) = \left\{ \frac{1}{\pi(1+x^2)} \text{ for all } -\infty < x < \infty \right\}$$

Earlier we saw that $E(X)$ doesn't exist, so clearly $V(X)$ doesn't exist either.

But we can use the idea of quantiles on any dist., whether its variance exists or not.



Earlier we defined the interquartile range (IQR) as

IQR = $F_X^{-1}(.75) - F_X^{-1}(.25)$

Standard Cauchy CDF is $F_{\Sigma}(x) = \int_{-\infty}^{x} \frac{1}{\pi(1+t^2)}\,dt$

(calculus) $= \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi}$

Here $\tan^{-1}(x)$ (arctangent) is what's called the principal inverse of $\tan(x)$, ranging from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$ as $-\infty < x < \infty$

Need to solve

$F_{\Sigma}(x) = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi} = p$ for x;

result is $x = F_{\Sigma}^{-1}(p) = \tan\left(\frac{p-\frac{1}{2}}{\pi}\right)$,

$= -\cot(p\pi)$

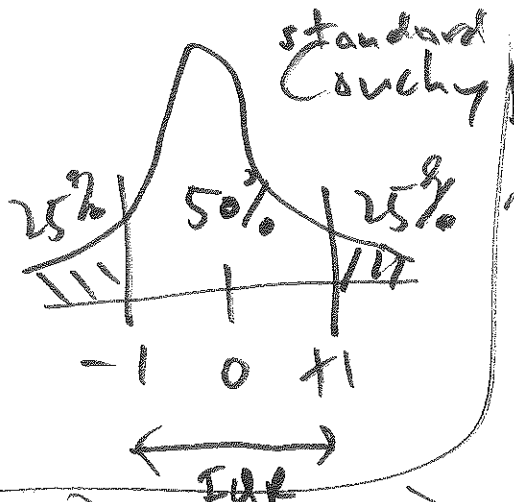
So the Idr for the standard Cauchy distribution is

$IdR = F_{\Sigma}^{-1}\left(\frac{3}{4}\right) - F_{\Sigma}^{-1}\left(\frac{1}{4}\right)$

$= \tan\left(\frac{\pi}{4}\right) - \tan\left(-\frac{\pi}{4}\right)$

$= 2.$

Moments of a rv



$$E(X) = E(X^1)$$

$$V(X) = E(X^2) - [E(X^1)]^2$$

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

with the usual mathematical impulse to generalize:

Def. X rv, k integer ≥ 1 $\rightarrow$

$E(X^k) \triangleq$ the k^{th} 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

$$\left(\begin{array}{c}\text{kth moment} \\ \text{of } X \text{ exists}\end{array}\right) \leftrightarrow E(|X|^k) < \infty$$

Consequences 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 $(k-1)^{st}$, $(k-2)^{nd}$, ..., moments.

Definition

X rv with expectation $E(X) = \mu$, k integer ≥ 1 → $E[(X-\mu)^k]$ 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 $X = E[(X-\mu)^2]$

② $E\left[(X-\mu)^1\right] = E(X) - \mu = \mu - \mu = 0$,

i.e., every rv has 1st central moment 0.

---

③ If X is symmetric around μ_X,

~~the dist. of~~

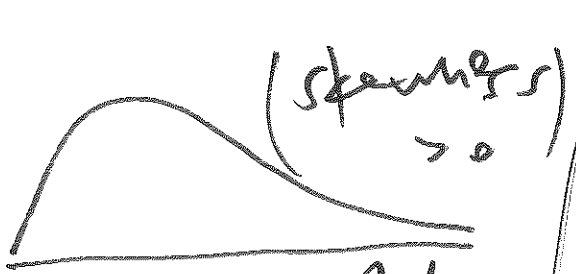
then $E\left[(X-\mu)^k\right] = 0$ for all odd integers k for which $E\left[(X-\mu)^k\right]$ exists

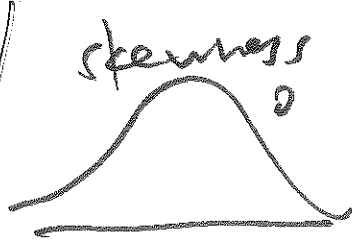
---

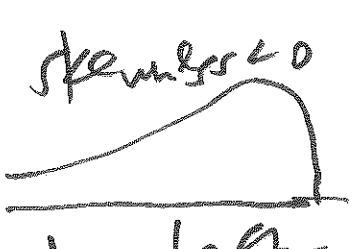
This motivates a new definition:

Def | X rv with mean μ_X, SD σ_X; if the third moment of X exists and is finite, then $\underline{\text{skewness}(X)} \triangleq E\left(\frac{X-\mu_X}{\sigma_X}\right)^3$.

All symmetric distributions with finite 3rd moment have skewness 0.







Moment generating functions

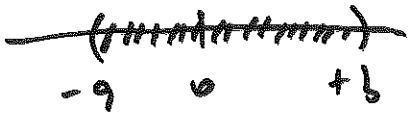
Def. X rv, t a real number

$\psi_X(t) \triangleq E(e^{tX})$ is called the moment generating function of X (MGF)

ψ: psi

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 ($a > 0$, $b > 0$);



then for all integers $n > 0$,

$$E(X^n) = \frac{d^n}{dt^n} \psi_X(t) \Big|_{t=0}$$

← n^{th} derivative of ψ_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.

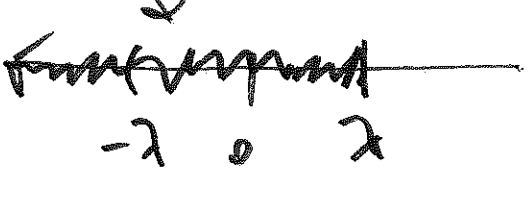
Example

$X \sim \text{Exponential}(\lambda)$ $\quad (\lambda > 0)$

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

$$\psi_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x} \, dx = \lambda \int_0^\infty e^{(t-\lambda)x} \, dx$$

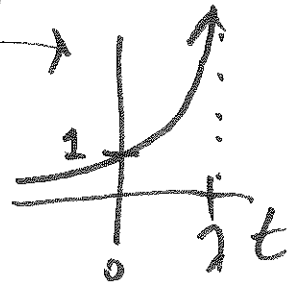
Now this integral is finite only if $t - \lambda < 0$, i.e. for $t < \lambda$, but this means (since $\lambda > 0$)



that it's definitely finite in an open interval around 0 (e.g. $(-\lambda, \lambda)$).

So $\psi_X(t)$ exists for $t < \lambda$ and equals

$$\psi_X(t) = \lambda \int_0^\infty e^{(t-\lambda)x} dx = \frac{\lambda}{\lambda - t}$$



Now we just crank out the derivatives:

$$E(X^1) = \left(\frac{d}{dt} \frac{\lambda}{\lambda - t} \right)\Bigg|_{t=0} = \frac{1}{\lambda}$$

So $V(X) = E(X^2) - [E(X)]^2$

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

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

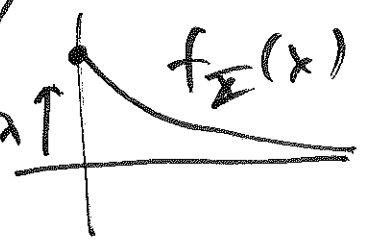
and $SD(X) = \frac{1}{\lambda}$

$$E(X^3) = \left(\frac{d^3}{dt^3} \left(\frac{\lambda}{\lambda - t} \right) \right)_{t=0} = \frac{6}{\lambda^3}$$

positive skew (long right-hand tail)

$$E(X^4) = \left(\frac{d^4}{dt^4} \left(\frac{\lambda}{\lambda - t} \right) \right)_{t=0} = \frac{24}{\lambda^4}$$

Evidently $E(X^k) = \frac{k!}{\lambda^k}$.



Consequences of the MGF definition

① X rv with MGF $\psi_X(t)$, $Y = aX + b$, (a, b constants)

Then at every value of t for which $\psi_X(at)$ is finite,

$$\psi_Y(t) = e^{bt}\psi_X(at).$$

Example

$X \sim$ Binomial (n, p), $X = \sum_{i=1}^{n} S_i$,

$S_i \overset{IID}{\sim}$ Bernoulli (p) $(i = 1, \ldots, n)$

MGF of S_i

is easy: $\psi_{S_i}(t) = E(e^{tS_i})$

$= e^{t\cdot 1} p(S_i = 1)$

$+ e^{t\cdot 0} p(S_i = 0)$

$= [pe^t + (1-p)]$

This uses the Law of the Unconscious Statistician (LotUS)

(2) $X_1, \ldots, X_n$ independent rv, mgf of X_i is $\psi_{X_i}(t)$, $Y = \sum_{i=1}^{n} X_i$,

mgf of Y is $\psi_Y(t)$ + for every t such that $\psi_{X_i}(t)$ is finite for all $i = 1, \ldots, n$, $\psi_Y(t) = \prod_{i=1}^{n} \psi_{X_i}(t)$.

---

mgf of Binomial, continued (18 Aug 17)

Since the S_i are IID,

$$\psi_X(t) \overset{IID}{=} \prod_{i=1}^{n} \psi_{S_i}(t)$$

$$\overset{IID}{=} \prod_{i=1}^{n} \left[p e^t + (1-p) \right]$$

$$\overset{IID}{=} \left[p e^t + (1-p) \right]^n$$

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