Simple example of this result

(X_1, X_2) joint continuous PDF

$f_{X_1 X_2}(x_1, x_2)$, $Y = a_1 X_1 + a_2 X_2 + b$

with $a_1 \neq 0 \rightarrow Y$ continuous

with PDF $f_Y(y) = \int_{-\infty}^{\infty} f_{X_1, X_2}\left(\frac{y - b - a_2 x_2}{a_1}, x_2\right) \frac{dx_2}{|a_1|}$

Important special case

The simplest thing you can do with two (or more) rvs is to add them.

This is also important in statistics, where the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ plays a key role.

In the result above, take $(a_1, a_2, b) = (1, 1, 0)$ to get $Y = X_1 + X_2$.

Dist. of Y is called the convolution of the dists. of X_1 and X_2

By the above result

$$f_{Y}(y) = \int_{-\infty}^{\infty} f_{X_1}(y-z) f_{X_2}(z)\, dz$$

$$= \int_{-\infty}^{\infty} f_{X_1}(z) f_{X_2}(y-z)\, dz.$$

A more complicated example is defined to be

$X_i \overset{IID}{\sim}$ CDF $F_{X_\bullet}$, PDF $f_{X_\bullet}$ $(i = 1, \dots, n)$ (continuous)

$Y_{(1)} \triangleq \min(X_1, \dots, X_n)$

$Y_{(n)} \triangleq \max(X_1, \dots, X_n)$

These are examples of the order statistics of $(X_1, \dots, X_n)$

Take-Home Test 2 problem 4

$$F_{Y_{(n)}}(t) = P(Y_{(n)} \le t)$$

$\updownarrow$ iff (check)

$$= P(X_1 \le t, X_2 \le t, \dots, X_n \le t)$$

$\overset{IID}{=}$

$$P(X_1 \le t) \cdots P(X_n \le t)$$

$\overset{IID}{=}$

$$\left[F_{X_\bullet}(t)\right]^n$$

so $Y_{(n)}$ has PDF $f_{Y_{(n)}}(t) = \frac{d}{dt}\left[F_{X_0}(t)\right]^n$

$= n\left[F_{X_0}(t)\right]^{n-1} f_{X_0}(t)$

Similarly

$F_{Y_{(1)}}(t) = P(Y_{(1)} \le t) = 1 - P(Y_{(1)} > t)$

$\updownarrow$ iff (check)

$= 1 - P(X_1 > t, \ldots, X_n > t)$

$\overset{\text{IID}}{=} 1 - P(X_1 > t) \cdots P(X_n > t)$

$\overset{\text{IID}}{=} 1 - \left[1 - F_{X_0}(t)\right]^n$

So $Y_{(1)}$ has PDF $f_{Y_{(1)}}(t) = \frac{d}{dt} F_{Y_{(1)}}(t)$

$= n\left[1 - F_{X_0}(t)\right]^{n-1} f_{X_0}(t)$

Multivariate transformations

Generalizing the earlier differentiable & 1-1 result

$\underline{X}_1, \ldots, \underline{X}_n$ continuous joint dist with joint PDF $f_{\underline{X}}(\underline{x})$

Suppose that there is a subset S of $\mathbb{R}^n$ with

$$P[(\underline{X}_1, \ldots, \underline{X}_n) \in S] = 1.$$

define new rvs:

$$\underline{Y}_1 = h_1(\underline{X}_1, \ldots, \underline{X}_n)$$
$$\vdots \qquad \vdots$$
$$\underline{Y}_n = h_n(\underline{X}_1, \ldots, \underline{X}_n)$$

(note some or # of $\underline{X}$s)

Assume that the n functions $h_1, \ldots, h_n$ define a 1-1 differentiable transformation of S onto some subset T of $\mathbb{R}^n$. ← image of $h_1, \ldots, h_n$

Inverse transform:

$$x_1 = h_1^{-1}(y_1, \ldots, y_n)$$
$$\vdots \qquad \vdots$$
$$x_n = h_n^{-1}(y_1, \ldots, y_n)$$

Then the joint PDF $f_{\underset{\sim}{Y}}(\underset{\sim}{y})$ is

$$f_{\underset{\sim}{Y}}(\underset{\sim}{y}) = \begin{cases} f_{\underset{\sim}{X}}[h_1^{-1}(\underset{\sim}{y}), \ldots, h_n^{-1}(\underset{\sim}{y})] \, |J| \\ \qquad \text{for } (y_1, \ldots, y_n) \in T \\ 0 \qquad \text{else} \end{cases}$$

in which

J is the determinant of the matrix

$$\begin{bmatrix} \frac{\partial h_1^{-1}}{\partial y_1} & \cdots & \frac{\partial h_1^{-1}}{\partial y_n} \\ \vdots & & \\ \frac{\partial h_n^{-1}}{\partial y_1} & \cdots & \frac{\partial h_n^{-1}}{\partial y_n} \end{bmatrix}$$

and $|\cdot|$ is absolute value

J is called the Jacobian of the transformation from $\underset{\sim}{X}$ to $\underset{\sim}{Y}$.

(chain rule generalization)

named after the German mathematician Carl Gustav Jacob Jacobi (1804 - 1851)

(died of smallpox at age 46)

Looks like a generalization of the derivative of the inverse in the earlier result.

Example (X_1, X_2) joint (continuous) PDF

$$f_{X_1 X_2}(x_1, x_2) = \begin{cases} 4x_1 x_2 & \text{for } 0 < x_1 < 1 \\ & 0 < x_2 < 1 \\ 0 & \text{else} \end{cases}$$

(check: $\int_0^1 \int_0^1 4x_1 x_2 \, dx_1 \, dx_2$

$= \int_0^1 4x_2 \left(\int_0^1 x_1 \, dx_1 \right) dx_2 = 4 \int_0^1 x_2 \left(\frac{x_1^2}{2} \Big|_0^1 \right) dx_2$

$= 2 \int_0^1 x_2 \, dx_2 = 2 \frac{x_2^2}{2} \Big|_0^1 = 1$)

Let's work out the joint PDF of

$(Y_1, Y_2) \triangleq \left(\frac{X_1}{X_2}, \; X_1 \cdot X_2 \right)$

$y_1 = h_1(x_1, x_2) = \frac{x_1}{x_2}$

$y_2 = h_2(x_1, x_2) = x_1 x_2$

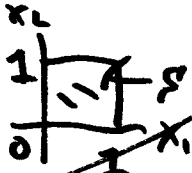
Inverse transform:

solve $\begin{cases} \frac{x_1}{x_2} = y_1 \\ x_1 x_2 = y_2 \end{cases}$ for (x_1, x_2):

$x_1 = h_1^{-1}(y_1, y_2) = \sqrt{y_1 y_2}$

$x_2 = h_2^{-1}(y_1, y_2) = \sqrt{\frac{y_2}{y_1}}$

image: how does $(0 < x_1 < 1, \ 0 < x_2 < 1)$ transform?



(*) defines 4 inequalities:

$\begin{cases} x_1 > 0, \ x_1 < 1, \\ x_2 > 0, \ x_2 < 1 \end{cases}$

So (a) $\sqrt{y_1 y_2} > 0$, (b) $\sqrt{y_1 y_2} < 1$

(c) $\sqrt{\frac{y_2}{y_1}} > 0$, (d) $\sqrt{\frac{y_2}{y_1}} < 1$

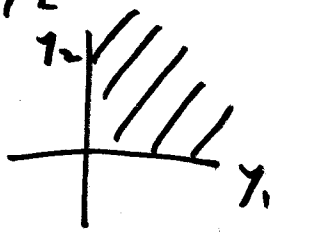
(a) equivalent to $\begin{pmatrix} y_1 > 0 \\ y_2 > 0 \end{pmatrix}$ or $\begin{pmatrix} y_1 < 0 \\ y_2 < 0 \end{pmatrix}$

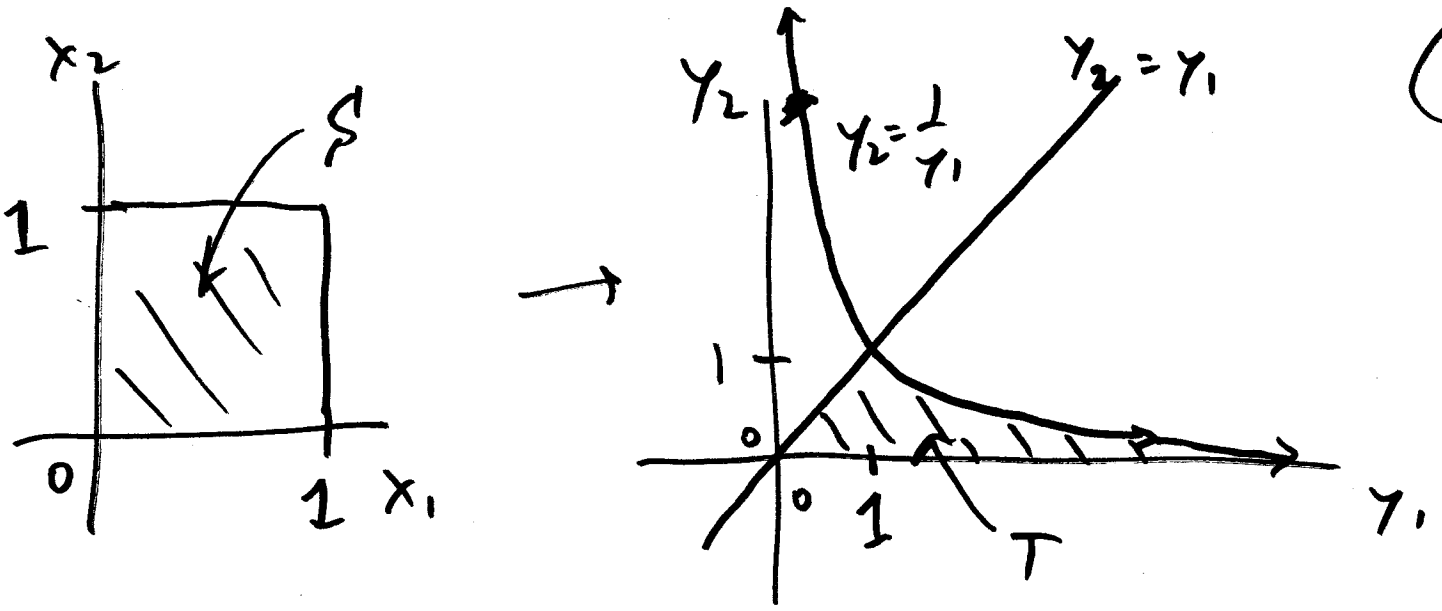
but $y_1 = \frac{x_1}{x_2} > 0$ so it must be $\begin{pmatrix} y_1 > 0 \\ y_2 > 0 \end{pmatrix}$

(c) leads to the same thing

(b) says $y_2 < \frac{1}{y_1}$

(d) says $y_2 < y_1$





$$h_1^{-1}(y_1, y_2) = \sqrt{y_1 y_2}$$

$$h_2^{-1}(y_1, y_2) = \sqrt{\frac{y_2}{y_1}}$$

so $\frac{\partial}{\partial y_1} h_1^{-1} = \frac{1}{2}\sqrt{\frac{y_2}{y_1}}$

$\frac{\partial}{\partial y_2} h_1^{-1} = \frac{1}{2}\sqrt{\frac{y_1}{y_2}}$

$\frac{\partial}{\partial y_1} h_2^{-1} = -\frac{1}{2}\left(\frac{y_2}{y_1^3}\right)^{\frac{1}{2}}$

$\frac{\partial}{\partial y_2} h_2^{-1} = \frac{1}{2}\sqrt{\frac{1}{y_1 y_2}}$

so $J = \det \begin{bmatrix} \frac{1}{2}\left(\frac{y_2}{y_1}\right)^{1/2} & \frac{1}{2}\left(\frac{y_1}{y_2}\right)^{1/2} \\ -\frac{1}{2}\left(\frac{y_2}{y_1^3}\right)^{1/2} & \frac{1}{2}\left(\frac{1}{y_1 y_2}\right)^{1/2} \end{bmatrix} = \frac{1}{2y_1}$

recall $\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$

and (since $y_1 > 0$) $|J| = \frac{1}{2y_1}$

To finish the calculation, in the PDF of $\underset{\sim}{X}$, $f_{\underset{\sim}{X}}(\underset{\sim}{x}) = \begin{cases} 4x_1 x_2 & \begin{pmatrix} 0<x_1<1 \\ 0<x_2<1 \end{pmatrix} \\ 0 & \text{else} \end{cases}$,

substitute $x_1 = \sqrt{y_1 y_2}$, $x_2 = \sqrt{\frac{y_2}{y_1}}$

and bring in the Jacobian:

$$f_{\underset{\sim}{Y}}(\underset{\sim}{y}) = f_{\underset{\sim}{X}}\left[h_1^{-1}(\underset{\sim}{y}), h_2^{-1}(\underset{\sim}{y})\right] |J|$$

$$= 4\left(\sqrt{y_1 y_2}\right)\left(\sqrt{\frac{y_2}{y_1}}\right)\frac{1}{2y_1}$$

$$= \begin{cases} 2\dfrac{y_2}{y_1} & \text{for } (y_1, y_2) \in T \\ 0 & \text{else} \end{cases}$$

A useful trick

start with (X_1, X_2) joint dist.; suppose you're interested only in the dist. of $Y_1 = h_1(X_1, X_2)$ ← univariate. Then one way to compute this dist. is with the following 3 steps.

Step 1: Find another rv $Y_2 = h_2(X_1, X_2)$ such that the transformation $(X_1, X_2) \to (Y_1, Y_2)$ is 1-1 with a differentiable inverse transformation & the calculations are straightforward.

Step 2 work out the joint dist of (Y_1, Y_2).

Step 3 Integrate Y_2 out of the joint dist. (i.e., marginalize over Y_2) to get the marginal dist. of Y_1.

Example of a Y_2 that would not work

$$Y_1 = 2X_1$$

$$Y_2 = 3X_1 = \frac{3}{2}Y_1$$

Here Y_2 is linearly dependent on Y_1, so the rank of the (2×2) Jacobian matrix is only 1 and its determinant is therefore 0.

~~Earlier~~

Example, continued

(X_1, X_2) have joint (continuous) PDF $f_{X_1,X_2}(x_1,x_2) =$

$$\begin{cases} 4x_1x_2 & \\ \cdot\ 0<x_1<1 & \\ \cdot\ 0<x_2<1 & \\ \vdots & \\ 0 & \text{else} \end{cases}$$

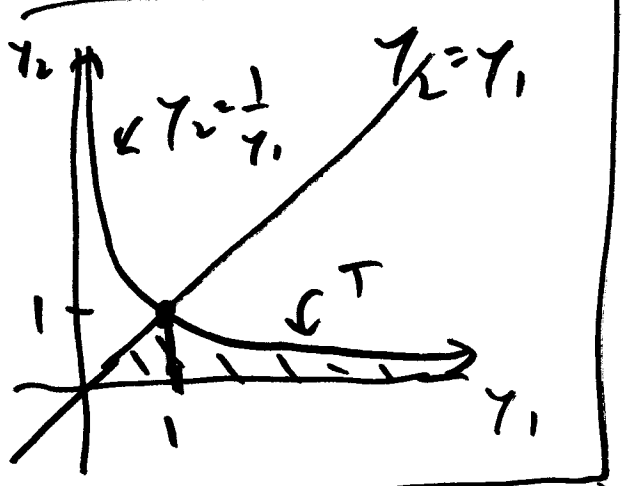
Earlier we found that with $(Y_1, Y_2) = \left(\frac{X_1}{X_2},\ X_1 \cdot X_2\right)$

the transformed PDF was

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} 2\frac{y_2}{y_1} & \text{for } (y_1,y_2)\in T \\ 0 & \text{else} \end{cases}$$

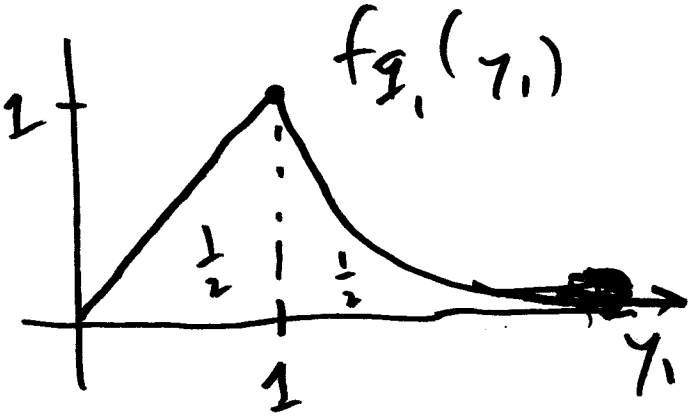
where $T = \{(y_1,y_2): y_1>0,\ y_2 < \min(y_1, \frac{1}{y_1})\}$.

Suppose you were only really interested (marginal) in the dist. of $Y_1 = \frac{X_1}{X_2}$; then all you have to do is integrate Y_2 out of the joint dist.



For $y_1 > 0$, the allowable region for y_2 is in two parts: for $0 < y_1 < 1$, $0 < y_2 < y_1$ and for $y_1 > 1$, $0 < y_2 < \frac{1}{y_1}$

So $f_{Y_1}(y_1) = \begin{cases} \int_0^{y_1} 2\left(\frac{1}{y_1}\right) dy_2 = y_1 & \text{for } 0 < y_1 < 1 \\ \int_0^{1/y_1} 2\left(\frac{y_2}{y_1}\right) dy_2 = y_1^{-3} & \text{for } y_1 > 1 \end{cases}$



weird PDF: not differentiable at $y_1 = 1$

Useful consequence of Jacobian story

$\underset{\sim}{X}_{n,1} = (X_1, \ldots, X_n)$ continuous with joint PDF $f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$,

$\underset{\sim}{Y}_{n,1} = (Y_1, \ldots, Y_n)$ is a linear transformation of $\underset{\sim}{X}$: $\underset{n,1}{\underset{\sim}{Y}} = \underset{n,n}{A} \cdot \underset{n,1}{\underset{\sim}{X}}$

where $\underset{n,n}{A}$ is an invertible (full-rank) matrix.

Then the PDF of $\underset{\sim}{Y}$ is

$$f_{\underset{\sim}{Y}}(\underset{n,1}{y^T}) = \frac{f_{\underset{\sim}{X}}(\underset{n,n}{A^{-1}}\, \underset{n,1}{y^T})}{|\det A|}$$

Example

$Y_1 = X_1 + X_2$

$Y_2 = X_1 - X_2$

$= \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

$\underset{2,2}{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$

$\det A = -2$

$= ad - bc$

$|\det A| = 2$

(recall that)

$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow A^{-1} = \frac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{ad - bc}$

$A^{-1} = \frac{1}{-2}\begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} = \frac{1}{2} A$

Expectation, Variance, Covariance, Correlation

Sec. 4 Example: Tay-Sachs disease (continued)

Earlier we worked out the discrete dist. of the rv

$Y = \begin{pmatrix} \#\text{ of T-S baby in family} \\ \text{of 5, both parents carriers} \end{pmatrix}$

we showed that $(Y) \sim$ Binomial (n, p) with $n = 5$, $p = \frac{1}{4}$

y	$P(Y = y)$
0	$\binom{5}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^5 = 0.2373$
1	$\binom{5}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^4 = 0.3955$
2	$\binom{5}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^3 = 0.2637$
3	$\binom{5}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^2 = 0.0879$
4	$\binom{5}{4}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^1 = 0.0146$
5	$\binom{5}{5}\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^0 = 0.0010$
	1.0000

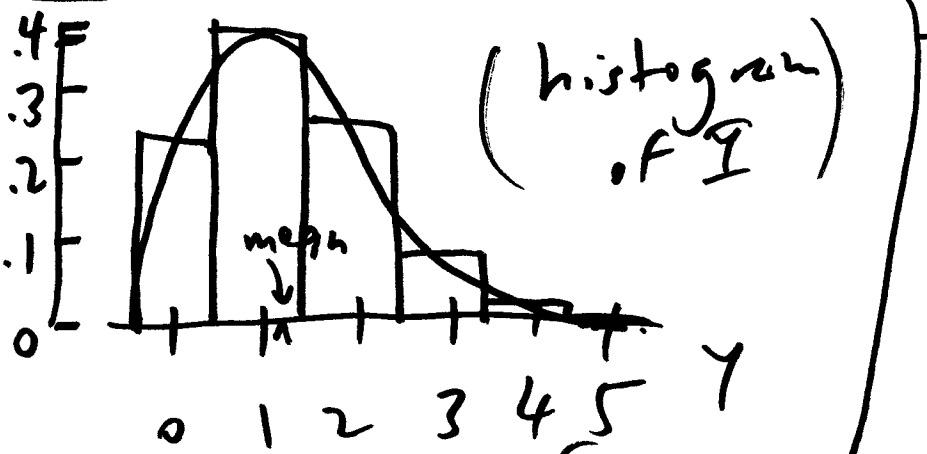
$$P(Y=y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y} & y = 0, 1, \ldots, n \\ 0 & \text{else} \end{cases}$$

Q: Want how many T-S baby should these parents expect to have?

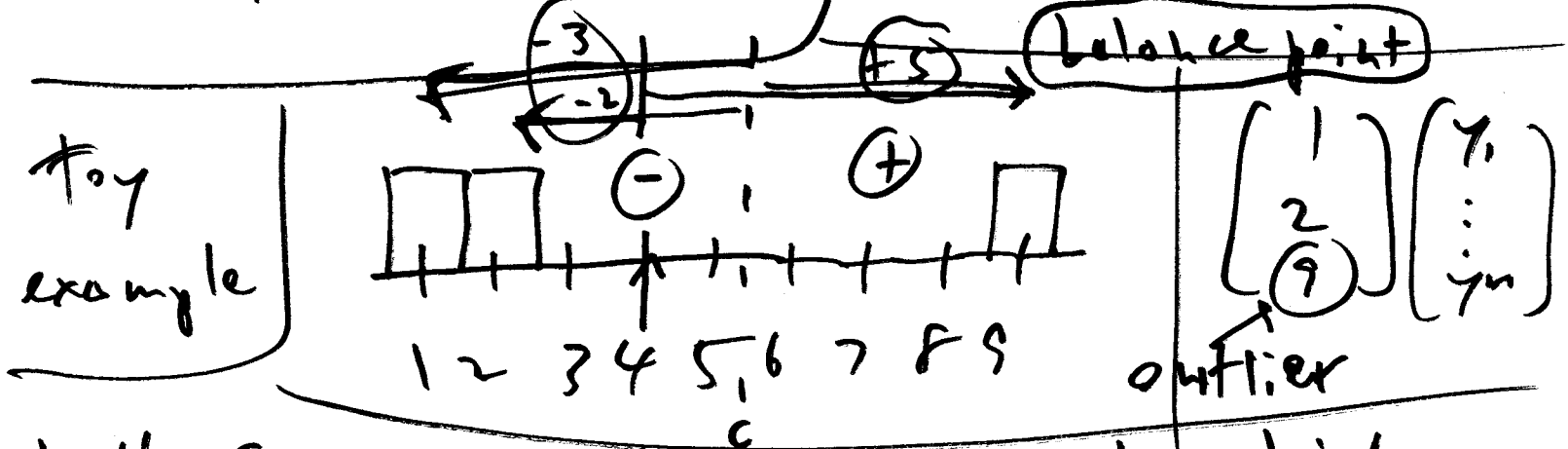
(center of dist. of Y)

A₁ Most likely outcome is 1 T-shirt (mode of the dist. of Y)

A₂ (physics idea) Let's work out the center of mass of the distribution



(histogram of Y)



Let's find the place c where the histogram balances: where (the sum of forces exerted by the histogram bars to the left of c) equals (the sum of forces to the right).

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \to \begin{bmatrix} y_1 - c \\ \vdots \\ y_n - c \end{bmatrix}$$

want sum = 0

$$\sum_{i=1}^{n} (y_i - c) = 0 =$$

$$\left(\sum_{i=1}^{n} y_i\right) - nc = 0$$

A_3 median of the dist. of Y (here that's also 1)

here $\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$

mean $\bar{y} = 4$

$\sum_{i=1}^{n} y_i - nc = 0 \iff$

$c = \frac{1}{n}\sum_{i=1}^{n} y_i \triangleq \bar{y}$ = the sample mean of the (sample) data set $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

Here each value of Y occurred only once:

$\bar{y} = \sum_{i=1}^{n} \left(\frac{1}{n}\right) y_i$

If some values are more probable than others, the generalization of $\left(\frac{1}{n}\right)$ weight on each y value would be to weight each y by its probability $P(Y=y)$.

Def. A rv is bounded if all of its possible values are finite.

Def Let Y be a bounded discrete rv with PMF $f_Y(y) = P(Y=y)$. The mean or expected value or expectation of Y,

is $E(I) \triangleq \sum_{\text{all } y} y\, p(I=y) = \sum_{\text{all } y} y f_I(y)$.

T-s example

$E(I) = (0)(.2373) + (1)(.3955) + \ldots + (5)(.0010) = 1.25000000$ ← surprisingly round #

Symbolically if $I \sim \text{Binomial}(n, p)$

then $E(I) = \sum_{y=0}^{n} y \binom{n}{y} p^y (1-p)^{n-y}$

Wa

$= \sum_{y=1}^{n} y \binom{n}{y} p^y (1-p)^{n-y}$ (since summand is 0 for $y=0$)

Xhi assumes that $n > 1$; proof for $n=1$ is on the next page

Wolfram alpha

$= \sum_{y=1}^{n} y \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$ cancel y going to $y \cdot (y-1)!$

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

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

$\binom{n-1}{y-1}$

$$= np \sum_{y=1}^{n} \binom{n-1}{y-1} p^{y-1} (1-p)^{n-1-(y-1)}$$

$$= np \boxed{\sum_{i=0}^{n-1} \binom{n-1}{i} p^{i} (1-p)^{n-1-i}} \quad (\text{substitute } i = y-1)$$

Binomial$(n-1, p)$ dist.

this $= 1$ because binomial probabilities add up to 1

So: if $Y \sim \text{Binomial}(n, p)$

for $n \geq 1$, $E(Y) = np$

When $n = 1$, $\text{Binomial}(1, p) = \text{Bernoulli}(p)$.

In this case $E(Y) = 0 \cdot P(Y=0) + 1 \cdot P(Y=1)$

$= 0 \cdot (1-p) + 1 \cdot p = p$

$= np$ with $n = 1$

So: for all $n \geq 1$ (integer) and $0 < p < 1$, $Y \sim \text{Binomial}(n, p) \rightarrow E(Y) = np$.

T-S example $(h=5, p=\frac{1}{4})$ $E(T) = \frac{5}{4} = 1.25$

If discrete X is unbounded, then either expectation of X may not exist, because

$$\sum_{x<0} x f_X(x) = -\infty \quad \left(\text{and/or } \sum_{x \geq 0} x f_X(x) = +\infty \right)$$

or the distribution "puts too much mass near $\pm\infty$"

Def. X discrete rv with pmf $f_X(x)$; consider $\sum_{x<0} x f_X(x)$ and $\sum_{x \geq 0} x f_X(x)$. If both sums are infinite, $E(X)$ is undefined (or does not exist); if at least one sum is finite, then

$$E(X) = \sum_{\text{all } x} x f_X(x) \quad \text{exists (it may still be infinite)}$$

To create a discrete rv whose mean doesn't exist, you have to play a careful game, because $\sum_{\text{all } x} f_X(x)$ has to be finite (it has to equal 1) but $\sum_{\text{some } x} x f_X(x)$ has to be infinite.

Example The harmonic series $\left(1 + \frac{1}{2} + \frac{1}{3} + \dots\right) = \sum_{x=1}^{\infty} \frac{1}{x}$ was known to the ancient Greeks, because the wavelengths (integers) of the overtones of a vibrating string are $\frac{1}{2}$, $\frac{1}{3}$, ... of the fundamental wavelength of the string. The fact that $\sum_{x=1}^{\infty} \frac{1}{x} = +\infty$ (i.e., the harmonic series diverges) was first shown in the 1300s (!) by the French philosopher Nicole Oresme ($\sim$1320–1382).

It's clear from this divergence that you can't create a rv X with pmf $p(X=x) = \frac{c}{x}$, $x = 1, 2, \ldots$, because the probability would sum to $+\infty$.

But $p(X=x) = \frac{c}{x^2}$ or $p(X=x) = \frac{c}{x(x+1)}$ turns out to work; for example, $\sum_{x=1}^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6}$ (!) and, even more conveniently, $\sum_{x=1}^{\infty} \frac{1}{x(x+1)} = 1$.

Ds use this to construct two pathological discrete distributions, to show what can go wrong with the idea of expectation.

Example 1 $f_X(x) = \begin{cases} \frac{1}{x(x+1)} & x = 1, 2, \ldots \\ 0 & \text{else} \end{cases}$.

$$E(X) = \sum_{x=1}^{\infty} x \cdot \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1} = +\infty$$

so $E(X)$ exists, it's just infinite.

Example 2

$$f_X(x) = \begin{cases} \dfrac{1}{2|x|(|x|+1)} & x = \pm 1, \pm 2, \ldots \\ 0 & \text{else} \end{cases}$$

we already know that $\sum_{\text{all } x} f_X(x) = 1$, so X is a well-defined rv; but $\sum_{x=-1}^{-\infty} x \cdot \frac{1}{2|x|(|x|+1)} = -\infty$

and $\sum_{x=1}^{\infty} x \frac{1}{2x(x+1)} = +\infty$, so $E(X)$ does not exist.

we won't work with pathological rv, mostly.

Expectation for continuous rvs

Def. X bounded continuous rv

with PDF $f_X(x) \rightarrow E(X) \triangleq \int_{-\infty}^{\infty} x f_X(x) dx$.

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

recall that $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{else} \end{cases}$

$$\int_0 E(X) = \int_0^{\infty} \lambda x e^{-\lambda x} dx \overset{\text{integrate by parts}}{=} \frac{1}{\lambda}.$$

For this reason, many people parameterize the exponential distribution differently:

Alternative definition

$X \sim \text{Exponential}(\eta)$ $(\eta > 0)$ (η: eta)

$$\rightarrow f_X(x) = \begin{cases} \frac{1}{\eta} e^{-\frac{x}{\eta}} & x > 0 \\ 0 & \text{else} \end{cases}$$

with this parameterization you can see that $E(X) = \eta$ (easier to interpret).

Nevertheless, to avoid confusion with DS, I'll stick with $\lambda e^{-\lambda x}$.

If continuous rv I is unbounded, a bit of care is once again required to define $E(I)$.

Def. I continuous rv with PDF $f_I(y)$; consider $\int_{-\infty}^{0} y f_I(y)\,dy$ and $\int_{0}^{\infty} y f_I(y)\,dy$. If both integrals are infinite, $E(I)$ is undefined (or does not exist); if at least one of these integrals is finite, $E(I) = \int_{\mathbb{R}} y f_I(y)\,dy$ exists (but it may still be infinite).

Example A dist. that does arise in practical statistical applications is the Cauchy distribution (attributed to Augustin-Louis Cauchy (1789-1857), a French mathematician who wrote 800 published articles & 5 textbooks in a 52-year period (15/year), but actually first studied carefully by Poisson in 1824). $f_Y(y) = \frac{1}{\pi(1+y^2)} (-\infty < y < \infty)$ is the (standard) Cauchy distribution.

It does integrate to 1, but $\int_0^\infty \frac{y}{\pi(1+y^2)}\,dy = +\infty$ and $\int_{-\infty}^0 \frac{y}{\pi(1+y^2)}\,dy = -\infty$, so $E(Y)$ does not exist, because its tails go to 0 extremely slowly.

This is because for large y, $\frac{y}{1+y^2} \doteq \frac{1}{y}$

and $\int_c^\infty \frac{1}{y}\,dy = +\infty$, the continuous analogue of the harmonic series.

(any $c > 0$)

Expectation of a function of a rv

X continuous rv with PDF $f_X(x)$, $Y = h(X)$.

Method 1 Work out PDF $f_Y(y)$;

then $E(Y) = \int_{\mathbb{R}} y\, f_Y(y)\,dy$.

Method 2 (faster) (if this exists)

$$E(Y) = \int_{\mathbb{R}} h(x)\, f_X(x)\,dx.$$

Discrete version:

$$E[h(X)] = \sum_{\text{all } x} h(x)\, f_X(x).$$

(X: discrete)

DS (and some other people) call Method 2 (LOTUS) the law of the Unconscious Statistician, because Method 2 looks like a definition but is actually a (difficult) theorem (in full generality) (16 Aug)

(measure theory: pushforward measure, ...)

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

$Y = X^2$

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

Notice that

$$E(X^2) \neq \left(E(X)\right)^2$$

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

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 $E(\underline{X})$

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 $E(\underline{X}) \leq b$.

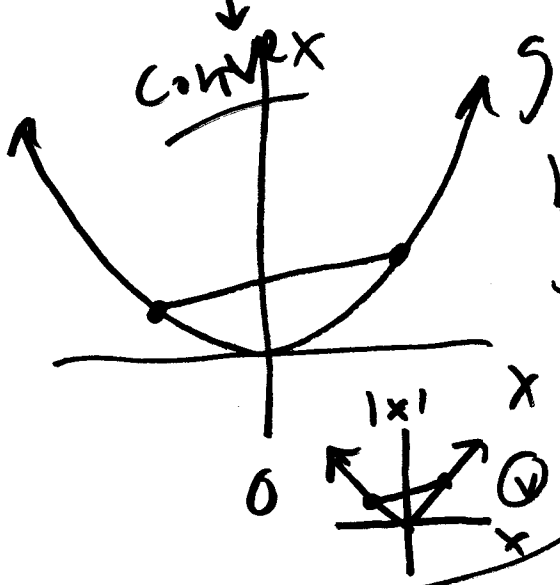
③ If $\underline{X}_1, \ldots, \underline{X}_n$ are n rvs, each with finite $E(\underline{X}_i)$, then $E\left(\sum_{i=1}^{n} \underline{X}_i\right) = \sum_{i=1}^{n} E(\underline{X}_i)$,

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

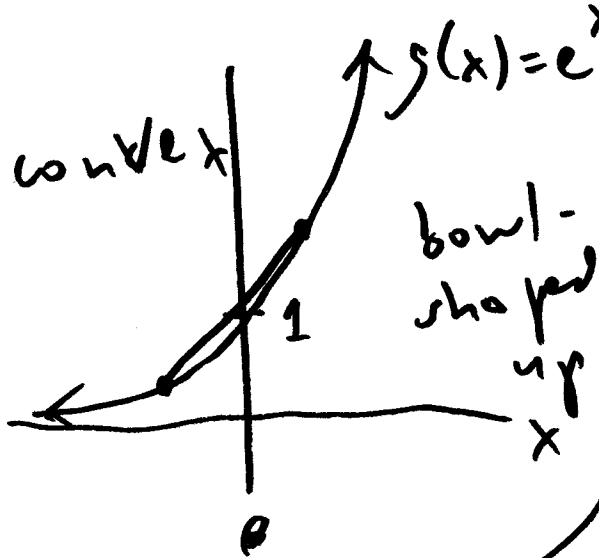
for all constants $(a_1, \ldots, a_n)$ and b.

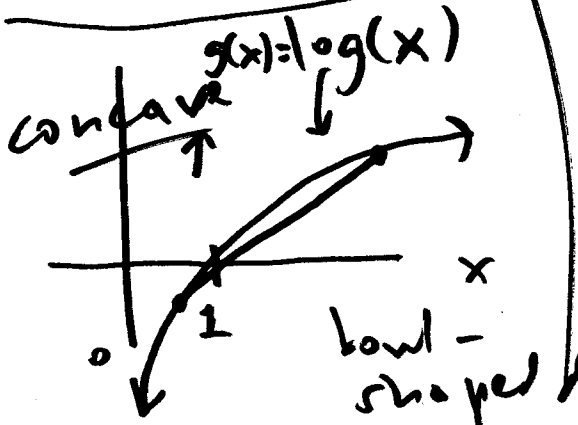
Def. A function $g: \mathbb{R}^n \to \mathbb{R}$ (this means that $g(\underline{x}) = z$, $\underline{x} \leftarrow (x_1, \ldots, x_n)$, real #s) is convex

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 entirely lies above the function except at the endpoints.





g is concave if

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

concave

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)

---

(5) (a) g convex, $\underset{\sim}{X}$ random vector with finite $E(\underset{\sim}{X})$ $\rightarrow E[g(\underset{\sim}{X})] \geq g[E(\underset{\sim}{X})]$.

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

Jensen's Inequality

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

(14 Mar 19)

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

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

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

↖ Binomial(n,p)

Expectation of a product when the X_j are independent

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

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

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.

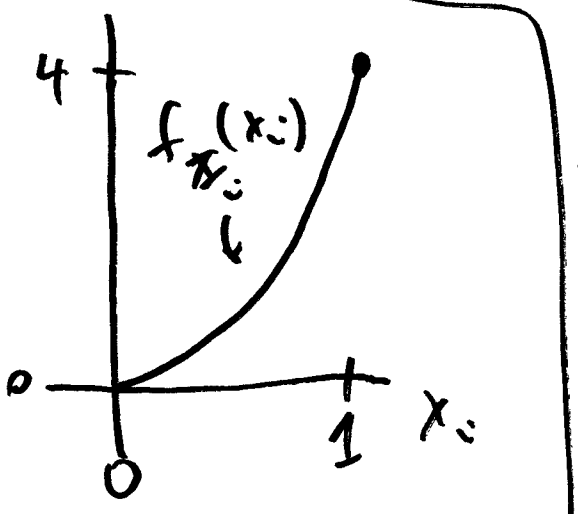
Example You have a (Brita) water filter that you use to improve the taste of Santa Cruz water. How much better would the filter do its job if you filtered the water twice instead of once?

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 $E(X) = \sum_{x=0}^{\infty} P(X \geq x)$.

(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 → $\mathbb{I}=3$) Suppose that my success probability is constant across the throws and equals p; & throws are independent.

F = failure
S = success

Then $E(\mathbb{I})$ should be inversely related to p:

the worse I am, the longer I expect the process to take; $E(\mathbb{I}) = ?$ (1st bullseye)

geometric distribution

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

$$\left(\begin{array}{c}\text{at least } n \\ \text{tosses required}\end{array}\right) \leftrightarrow \left(\begin{array}{c}\text{none of the first } (n-1) \\ \text{throws succeeded}\end{array}\right)$$

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

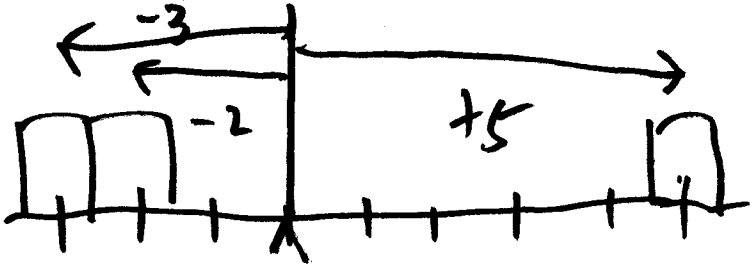
$$E(\mathbb{I}) = \sum_{n=1}^{\infty} (1-p)^{n-1} = 1 + (1-p) + (1-p)^2 + \dots$$

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

(inverse relation ✓)

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

Variance and standard deviation



$\begin{bmatrix} 8 \\ 1 \\ 8_2 \\ 8_9 \end{bmatrix}$ mean $8_4 = \mu$

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

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

deviation from μ

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

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

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

Gauss (1800) K (Laplace) (1785)

AAD not used much; variance used constantly.

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

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

Def 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 $\triangleq \sqrt{V(X)} \triangleq \underline{SD}(X)$.

(K. Pearson [name]) (~1890)

Consequences of the definitions

① $V(X) = E\left[(X-\mu)^2\right]$

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

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

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

this is a different way to compute the variance

So $V(X) =$ (expectation of X^2) $-$ (square of expectation of X)

Toy example

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

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

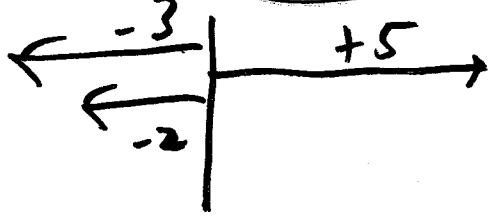
$E(X-\mu)^2 = \frac{1}{3}(1-4)^2 + \frac{1}{3}(2-4)^2 + \frac{1}{3}(9-4)^2 \doteq 12.7$

$= V(X)$.

So

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

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.

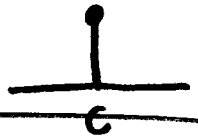
This is a consequence of Jensen's Inequality:

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

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

concave

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, $E(X) \overset{\text{def}}{=} \mu_X$, $V(X) \overset{\text{def}}{=} \sigma_X^2$ (lower-case sigma)

and $SD(X) \overset{\text{def}}{=} \sigma_X$

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 (18) has endured even though the units of the variance are wrong: for independent rvs, variance is additive whereas SD is not. (16 Aug 19)

name of SD: Karl Pearson (1890)

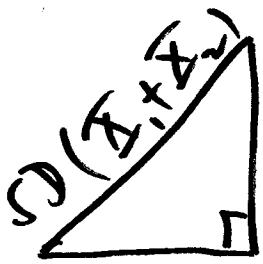
correct units

special case of (5):

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

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

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



so $SD(X_1 + X_2) \geq \max\{SD(X_1), SD(X_2)\}$

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

Immediately, $\max\{SD(X_1), SD(X_2)\} \leq SD(X_1 + X_2) \leq SD(X_1) + SD(X_2)$ (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 $V(S_i) = E(S_i^2) - [E(S_i)]^2$ we're halfway there.

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

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

So $E(S_i^2) = E(S_j) = p$ and finally

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

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

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

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." ($\leftarrow \sigma_X$, $\leftarrow \mu_X$)

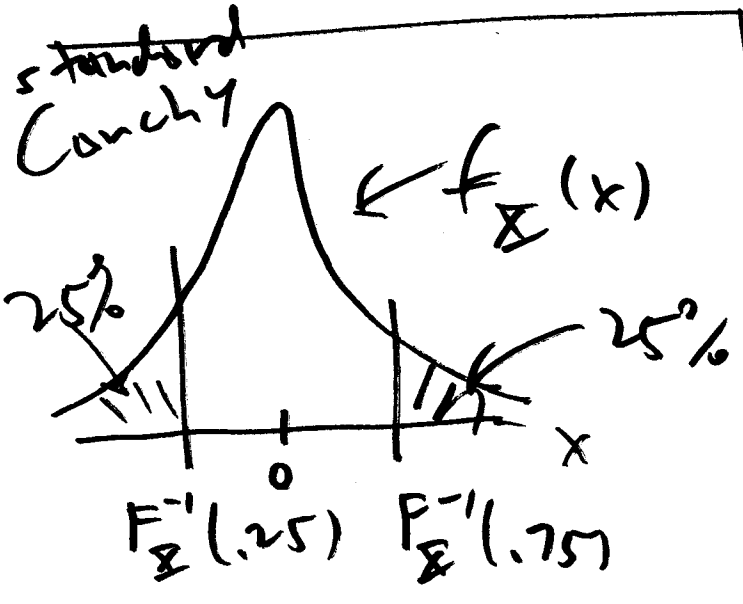
How do you measure the spread of a distribution if the variance doesn't exist?

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 <u>any</u> dist., whether its variance exists or not.



Earlier we defined the <u>interquartile range</u> (IQR) as

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

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

(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

$$F_X(x) = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi} = p \quad \text{for } x;$$

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

$= -\cot(p\pi)$

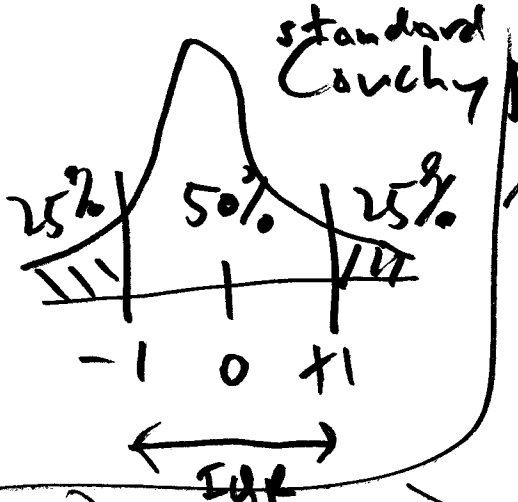
So the IdR for the standard Cauchy distribution is

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

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

$$= 2.$$

Moments of a rv



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

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

$$= E(X-\mu)^2 \leftarrow E(X^1)$$

with the usual mathematical impulse to generalize:

Def. X rv, k integer ≥ 1 →

$E(X^k) \triangleq$ 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 $E(X) \overset{\text{mean}}{=} \mu$, k integer ≥ 1 → $E\left[(X-\mu)^k\right]$ 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 $X = E\left[(X-\mu)^2\right]$

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,

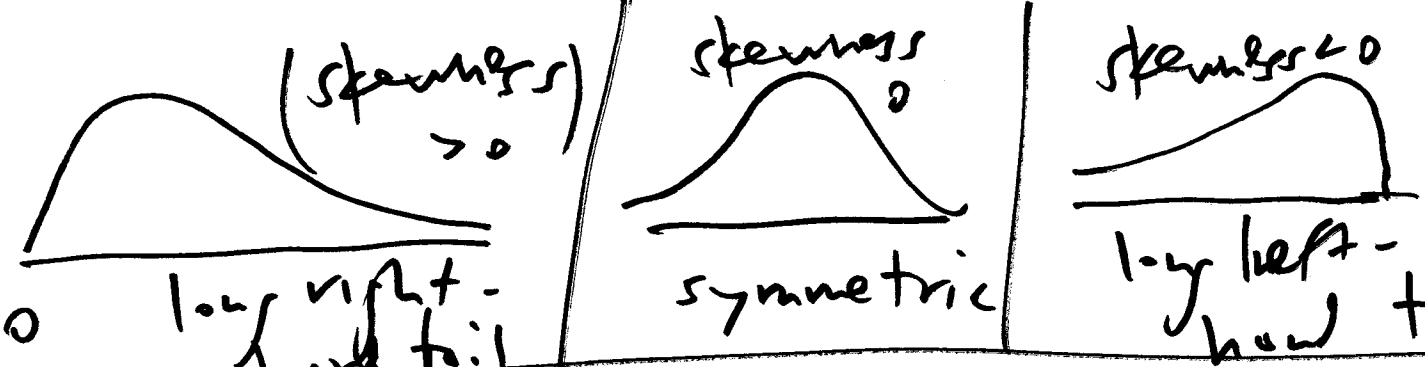
then $E\left[(X-\mu)^k\right] = 0$ for all odd integers k for which $E\left[(X-\mu)^k\right]$ exists

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 $\underline{\text{skewness}(X)} \triangleq E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right]$.

converting X to standard units

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



Moment generating functions

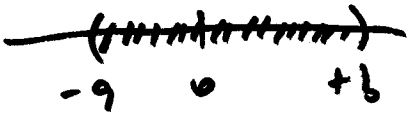
Def. X rv, t a real number

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

↖ 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.

(16 May 19)

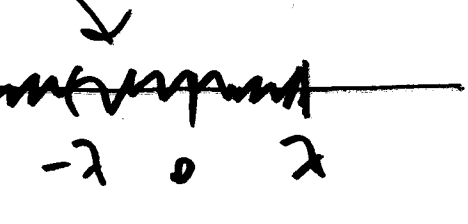
Example

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

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

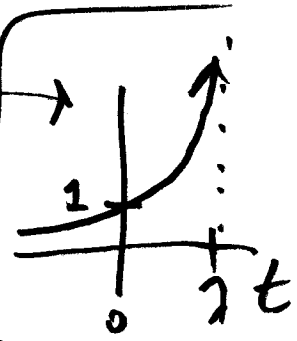
$$\psi_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x} \, dx = \lambda \int_0^\infty e^{(t-\lambda)x} \, dx$$

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

$$E(X^2) = \left[\frac{d^2}{dt^2} \left(\frac{\lambda}{\lambda - t} \right) \right]_{t=0} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

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