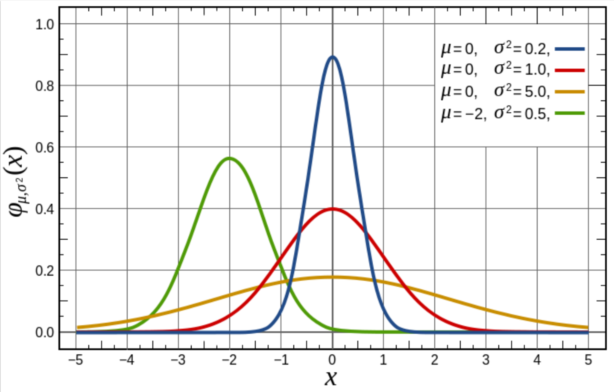
Variance; Continuous Random Variables
18.05 Spring 2017



Variance and standard deviation

X a discrete random variable with mean $E(X) = \mu$.

- Meaning: spread of probability mass about the mean.
- Definition as expectation (weighted sum):

$$\text{Var}(X) = E((X - \mu)^2).$$

- Computation as sum:

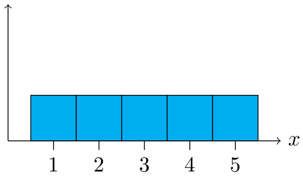
$$\text{Var}(X) = \sum_{i=1}^{n} p(x_i)(x_i - \mu)^2.$$

- Standard deviation $\sigma = \sqrt{\text{Var}(X)}$.
  Units for standard deviation = units of X.

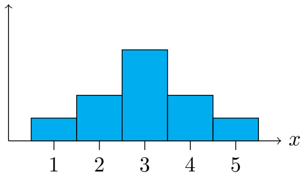
Concept question

The graphs below give the pmf for 3 random variables. Order them by size of standard deviation from biggest to smallest. (Assume x has the same units in all 3.)

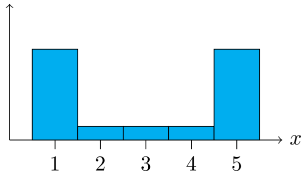
(A)



(B)



(C)



1. ABC 2. ACB 3. BAC 4. BCA 5. CAB 6. CBA

Answer on next slide

Solution

answer: 5. CAB

All 3 variables have the same range from 1-5 and all of them are symmetric so their mean is right in the middle at 3. (C) has most of its weight at the extremes, so it has the biggest spread. (B) has the most weight in the middle so it has the smallest spread.
From biggest to smallest standard deviation we have (C), (A), (B).

Computation from tables

Example. Compute the variance and standard deviation of X.

values x	1	2	3	4	5
pmf $p(x)$	$1/10$	$2/10$	$4/10$	$2/10$	$1/10$

Answer on next slide

Computation from tables

From the table we compute the mean:

$$\mu = \frac{1}{10} + \frac{4}{10} + \frac{12}{10} + \frac{8}{10} + \frac{5}{10} = 3.$$

Then we add a line to the table for $(X-\mu)^2$.

values X	1	2	3	4	5
pmf $p(x)$	1/10	2/10	4/10	2/10	1/10
$(X-\mu)^2$	4	1	0	1	4

Using the table we compute variance $E((X-\mu)^2)$:

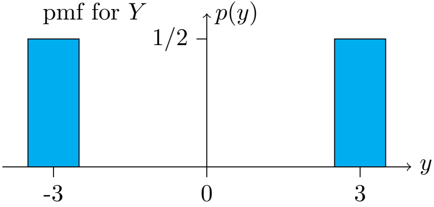
$$\frac{1}{10}\cdot 4 + \frac{2}{10}\cdot 1 + \frac{4}{10}\cdot 0 + \frac{2}{10}\cdot 1 + \frac{1}{10}\cdot 4 = 1.2$$

The standard deviation is then $\sigma = \sqrt{1.2}$.

Concept question

Which pmf has the bigger standard deviation? (Assume *w* and *y* have the same units.)

1. *Y* 2. *W*



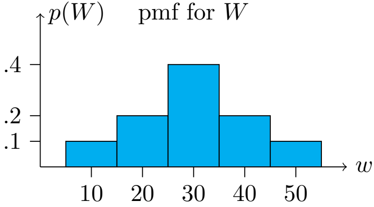


Table question: make probability tables for *Y* and *W* and compute their standard deviations.

Solution on next slide

Solution

answer: We get the table for Y from the figure. After computing $E(Y)$ we add a line for $(Y-\mu)^2$.

Y	-3	3
$p(y)$	0.5	0.5
$(Y-\mu)^2$	9	9

$E(Y) = 0.5(-3) + 0.5(3) = 0.$ $\quad E((Y-\mu)^2) = 0.5(9) + 0.5(9) = 9$
therefore $\text{Var}(Y) = 9 \;\Rightarrow\; \sigma_Y = 3.$

W	10	20	30	40	50
$p(w)$	0.1	0.2	0.4	0.2	0.1
$(W-\mu)^2$	400	100	0	100	400

We compute $E(W) = 1 + 4 + 12 + 8 + 5 = 30$ and add a line to the table for $(W-\mu)^2$. Then

$$\text{Var}(W) = E((W-\mu)^2) = .1(400) + .2(100) + .4(0) + .2(100) + .1(100) = 120$$

$$\sigma_W = \sqrt{120} = 10\sqrt{1.2}.$$

Note: Comparing Y and W, we see that scale matters for variance.

Concept question

True or false: If $\text{Var}(X) = 0$ then X is constant.

1. True 2. False

answer: True. If X can take more than one value with positive probability, than $\text{Var}(X)$ will be a sum of positive terms. So X is constant if and only if $\text{Var}(X) = 0$.

Algebra with variances

If a and b are constants then

$$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X), \qquad \sigma_{aX+b} = |a|\,\sigma_X.$$

If X and Y are independent random variables then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Board questions

1. Prove: if $X \sim \text{Bernoulli}(p)$ then $\text{Var}(X) = p(1-p)$.

2. Prove: if $X \sim \text{bin}(n, p)$ then $\text{Var}(X) = n\,p(1-p)$.

3. Suppose $X_1, X_2, \ldots, X_n$ are independent and all have the same standard deviation $\sigma = 2$. Let $\overline{X}$ be the average of $X_1, \ldots, X_n$.

What is the standard deviation of $\overline{X}$?

Solution on next slide

Solution

1. For $X \sim$ Bernoulli(p) we use a table. (We know $E(X) = p$.)

X	0	1
$p(x)$	$1-p$	p
$(X-\mu)^2$	p^2	$(1-p)^2$

$$\mathrm{Var}(X) = E((X-\mu)^2) = (1-p)p^2 + p(1-p)^2 = p(1-p)$$

2. $X \sim \mathrm{bin}(n, p)$ means X is the sum of n *independent* Bernoulli(p) random variables X_1, X_2, $\ldots$, X_n. For independent variables, the variances add. Since $\mathrm{Var}(X_j) = p(1-p)$ we have

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n) = np(p-1).$$

continued on next slide

Solution continued

3. Since the variables are independent, we have

$$\mathrm{Var}(X_1 + \ldots + X_n) = 4n.$$

$\overline{X}$ is the sum scaled by $1/n$ and the rule for scaling is $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, so

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}(\frac{X_1 + \cdots + X_n}{n}) = \frac{1}{n^2}\mathrm{Var}(X_1 + \ldots + X_n) = \frac{4}{n}.$$

This implies $\sigma_{\overline{X}} = \dfrac{2}{\sqrt{n}}$.

Note: this says that the average of n independent measurements varies less than the individual measurements.

Continuous random variables

- Continuous range of values:

$$[0,1],\ [a,b],\ [0,\infty),\ (-\infty,\infty).$$

- Probability density function (pdf)

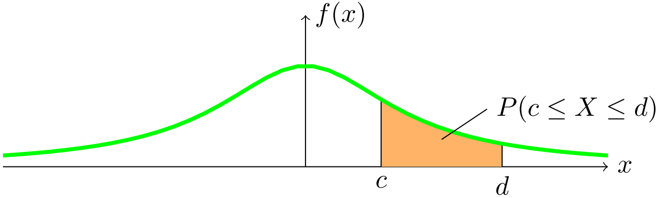
$$f(x) \geq 0; \quad P(c \leq x \leq d) = \int_c^d f(x)\,dx.$$

Units for the pdf are $\dfrac{\text{prob.}}{\text{unit of x}}$

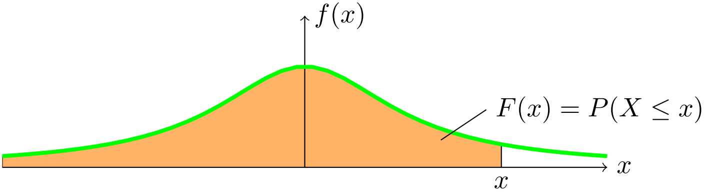
- Cumulative distribution function (cdf)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt.$$

Visualization



pdf and probability



pdf and cdf

Properties of the cdf

(Same as for discrete distributions)

- (Definition) $F(x) = P(X \leq x)$.
- $0 \leq F(x) \leq 1$.
- non-decreasing.
- 0 to the left: $\lim\limits_{x \to -\infty} F(x) = 0$.
- 1 to the right: $\lim\limits_{x \to \infty} F(x) = 1$.
- $P(c < X \leq d) = F(d) - F(c)$.
- $F'(x) = f(x)$.

Board questions

1. Suppose X has range $[0, 2]$ and pdf $f(x) = cx^2$.

(a) What is the value of c.

(b) Compute the cdf $F(x)$.

(c) Compute $P(1 \leq X \leq 2)$.

2. Suppose Y has range $[0, b]$ and cdf $F(y) = y^2/9$.

(a) What is b?

(b) Find the pdf of Y.
Solution on next slide

Solution

1a. Total probability must be 1. So

$$\int_0^2 f(x)\,dx = \int_0^2 cx^2\,dx = c\frac{8}{3} = 1 \;\Rightarrow\; \boxed{c = \frac{3}{8}}.$$

1b. The pdf $f(x)$ is 0 outside of $[0,2]$ so for $0 \le x \le 2$ we have

$$F(x) = \int_0^x cu^2\,du = \frac{c}{3}x^3 = \boxed{\frac{x^3}{8}}.$$

$F(x)$ is 0 fo $x < 0$ and 1 for $x > 2$.

1c. We could compute the probability as $\displaystyle\int_1^2 f(x)\,dx$, but rather than redo the integral let's use the cdf:

$$P(1 \le X \le 2) = F(2) - F(1) = 1 - \frac{1}{8} = \boxed{\frac{7}{8}}.$$

Continued on next slide

Solution continued

2a. Since the total probability is 1, we have

$$F(b) = 1 \Rightarrow \frac{b^2}{9} = 1 \Rightarrow \boxed{b = 3}.$$

2b. $f(y) = F'(y) = \dfrac{2y}{9}$.

Concept questions

Suppose X is a continuous random variable.

(a) What is $P(a \leq X \leq a)$?

(b) What is $P(X = 0)$?

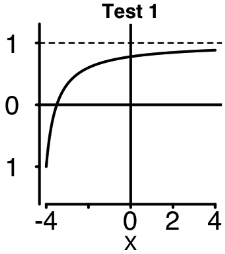
(c) Does $P(X = a) = 0$ mean X never equals a?

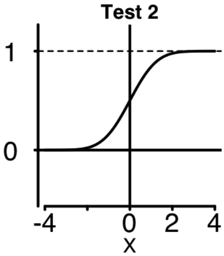
answer: **(a)** 0

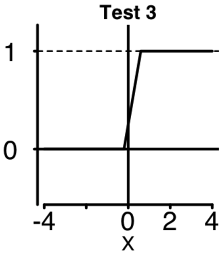
(b) 0
(c) No. For a continuous distribution any single value has probability 0. Only a range of values has non-zero probability.

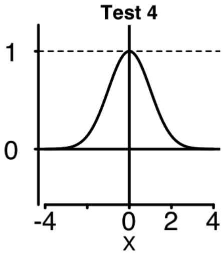
Concept question

Which of the following are graphs of valid cumulative distribution functions?









Add the numbers of the valid cdf's and click that number.

answer: Test 2 and Test 3.

Solution

Test 1 is not a cdf: it takes negative values, but probabilities are positive.

Test 2 is a cdf: it increases from 0 to 1.

Test 3 is a cdf: it increases from 0 to 1.

Test 4 is not a cdf because it decreases. A cdf must be non-decreasing since it represents *accumulated* probability.

Exponential Random Variables

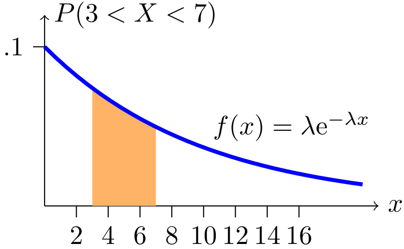
Parameter: λ (called the rate parameter).

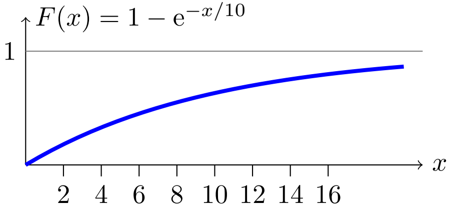
Range: $[0, \infty)$.

Notation: exponential(λ) or exp(λ).

Density: $f(x) = \lambda e^{-\lambda x}$ for $0 \le x$.

Models: Waiting time





Continuous analogue of geometric distribution –memoryless!

Board question

I've noticed that taxis drive past 77 Mass. Ave. on the average of once every 10 minutes.

Suppose time spent waiting for a taxi is modeled by an exponential random variable

$$X \sim \text{Exponential}(1/10); \qquad f(x) = \frac{1}{10}\mathrm{e}^{-x/10}$$

(a) Sketch the pdf of this distribution

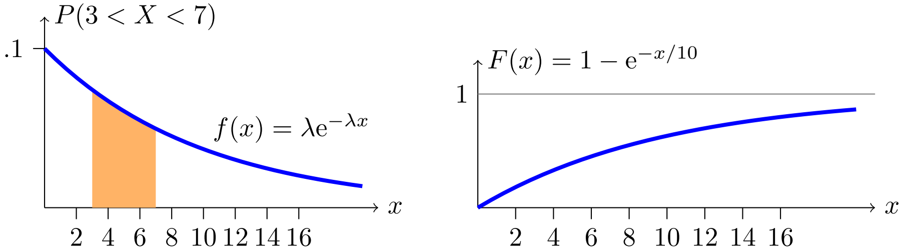
(b) Shade the region which represents the probability of waiting between 3 and 7 minutes

(c) Compute the probability of waiting between between 3 and 7 minutes for a taxi

(d) Compute and sketch the cdf.

Solution

Sketches for (a), (b), (d)



(c)

$$(3 < X < 7) = \int_3^7 \frac{1}{10} e^{-x/10}\, dx = -e^{-x/10}\Big|_3^7 = e^{-3/10} - e^{-7/10} \approx 0.244$$