MATH 320: Probability

Lecture 9: Summary Measures

Chapters 2 and 3: Distributions (2.2, 2.3, and 3.1)

Expected value

Data reduction

- When we try to interpret numerical information that has a wide range of values, we like to reduce our confusion by looking at a single number which summaries the information.

  Sample of n data points $\longrightarrow$ 1 or 2 summary measures.

Motivating example

- When the quizzes are returned, students are interested in the quiz average as well as the distribution of grades.

  Lets say we have the following quiz scores: 6, 7, 8, 9

  First we are going to calculate the mean like we normally would, then do some rearranging.

  $$\text{Mean} = \frac{6+7+8+9}{4} = \frac{1}{4}(6+7+8+9) = 6\left(\tfrac{1}{4}\right) + 7\left(\tfrac{1}{4}\right) + 8\left(\tfrac{1}{4}\right) + 9\left(\tfrac{1}{4}\right) = 7.5$$

  Written out like this, we can think of $\frac{1}{4}$ as a probability and the numbers as x's, which are particular instances of the random variable X. Then we have a probability function.

Quiz score (x)	6	7	8	9
$f(x)$	1/4	1/4	1/4	1/4

  mean = 7.5

  Now what if we said that a score of 9 is more likely than the other scores. Our new pmf is:

Quiz score (x)	6	7	8	9
$f(x)$	1/6	1/6	1/6	1/2

  The new mean going to increase. Let's calculate it:

  $$\text{Mean} = 6\left(\tfrac{1}{6}\right) + 7\left(\tfrac{1}{6}\right) + 8\left(\tfrac{1}{6}\right) + 9\left(\tfrac{1}{2}\right) = 8 > 7$$

- What we are actually calculating here is called the expected value, which we can think of as a weighted average of the x's where the probabilities $P(X = x_i)$ are the weights.

  This is how we get mean (aka expected value) of a random variable from its pmf, which is usually what we are given.

Defining expected value

- Definitions

  Let X be a discrete random variable. The **expected value** of X is defined by

  $$E(X) = \sum_{x \in X} x f(x)$$

  Let X be a continuous random variable. The **expected value** of X is defined by

  $$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

  The expected value of the random variable X is often denoted by the Greek letter μ.

  $$E(X) = \mu$$

- Examples

  1. Pmf for the random variable X, the number of health insurance claims filed by a policyholder in a year, is given in the table below.

Number of claims (x)	0	1	2	3
$f(x)$	0.28	0.43	0.20	0.09

  Find the expected number of claims.

  $$E(X) = \sum x f(x)$$
  $$= 0(0.28) + 1(0.43) + 2(0.2) + 3(0.09)$$
  $$= 1.1$$

  2. Let X be the loss severity random variable for the warranty policy with density:

  $$f(x) = \begin{cases} 0.02 - 0.0002x & 0 \leq x \leq 100 \\ 0 & \text{otherwise} \end{cases}$$

  Find the expected loss.

  $$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^{100} x(0.02 - 0.0002x) dx$$
  $$= 0.01x^2 - \frac{0.0002}{3}x^3 \Big|_0^{100} \quad F(100) - F(0)$$
  $$= 0.01(100^2) - \frac{0.0002}{3}(100^3)$$
  $$= \frac{100}{3} \approx 33.33$$

3. Expected value of a piecewise density function example: Let X have the following pdf:

$$f_X(x) = \begin{cases} 560x & 0 \le x \le 0.05 \\ -15x + 3.75 & 0.05 < x \le 0.25 \\ 0 & \text{otherwise} \end{cases}$$

Find the expected value.

– *STRATEGY:* Integrate in pieces.

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$$
$$= \int_0^{0.05} x(560x)\,dx + \int_{0.05}^{0.25} x(-15x + 3.75)\,dx$$
$$= 0.0233 + 0.035$$
$$= 0.05833$$

4. Smith is offered the following gamble: he is to choose a coin at random from a large collection of coins and toss it randomly.

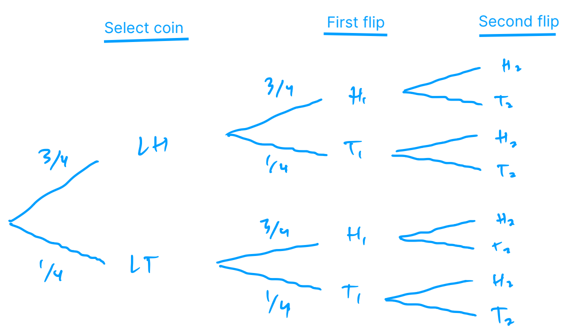
– 3/4 of the coins in the collection are loaded toward head (LH) and 1/4 are loaded towards a tail (LT).

– If a coin is loaded towards a head, then when the coin is tossed randomly, there is a 3/4 probability that a head will turn up and a 1/4 probability that a tail will turn up. Similarly, if the coin is loaded towards tails, then there is a 3/4 chance of tossing a tail on any given toss.

– If Smith tosses a head, he loses \$100 and if he tosses a tail, he wins \$200.

– Smith is allowed to obtain "sample information" about the gamble. When he chooses the coin at random, he is allowed to toss it once before deciding to accept the gamble with that same coin.

Suppose Smith tosses a head on the sample toss. Find Smith's expected gain/loss on the gamble if it is accepted.



$\Rightarrow X: \$ \Rightarrow f(x) = \begin{cases} P(H_2 \mid H_1) & x = -100 \\ P(T_2 \mid H_1) & x = 200 \end{cases}$

think about range then probs

$\Rightarrow f(-100) = P(H_2 \mid H_1) = \dfrac{P(H_1 \cap H_2)}{P(H_1)} = \dfrac{\overset{1}{P(LH)P(H_1 \cap H_2 \mid LH)} + \overset{5}{P(LT)P(H_1 \cap H_2 \mid LT)}}{P(LH)P(H_1 \mid LH) + P(LT)P(H_1 \mid LT)}$

$= \dfrac{\frac{3}{4}\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) + \frac{1}{4}\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)}{\frac{3}{4}\left(\frac{3}{4}\right) + \frac{1}{4}\left(\frac{1}{4}\right)}$

$= 0.7$

$\Rightarrow f(200) = 1 - f(-100) = 1 - 0.7 = 0.3$

x	-100	200
$f(x)$	0.7	0.3

$\Rightarrow E(X) = \sum x f(x)$

$= -100(0.7) + 200(0.3)$

$= -10$

Expected value of a function of a random variable

Motivating example

- Suppose X is a random variable, but we are actually interested in a function of the random variable $g(X)$ (which is another random variable).

  One common application of this is when $g(X) = aX + b$.

- Now suppose the table from the previous Example 1 is for a type of policy which guarantees a fixed payment of \$100 dollars for each claim.

  Then the amount paid to a policy holder in a year is just \$100 multiplied by the number of claims. The total claim amount is a new random variable $\underline{\quad 100X \quad}$.

  We now have two random variables, X and Y, and each has their own pmf.

Number of claims (x)	0	1	2	3
Total claim amount (y)	0	100	200	300
$f(y)$ $= f(x)$	0.28	0.43	0.20	0.09

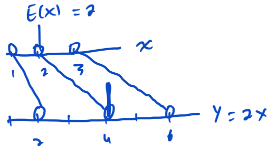
(a) Find the expected total claim amount.

$$E(Y) = 0(0.28) + 100(0.43) + 200(0.2) + 300(0.09)$$
$$= 100[\underbrace{0(0.28) + 1(0.43) + 2(0.2) + 3(0.09)}_{E(X)}]$$
$$= 100\,\underbrace{E(X)}_{=1.1}$$
$$= 110$$

Theorems

ex

$E(X) = 2$



$E(Y) = 4$

- $\boxed{\text{Theorem: For any constant } a \text{ and random variable } X, \quad E(aX) = a\,E(X)}$

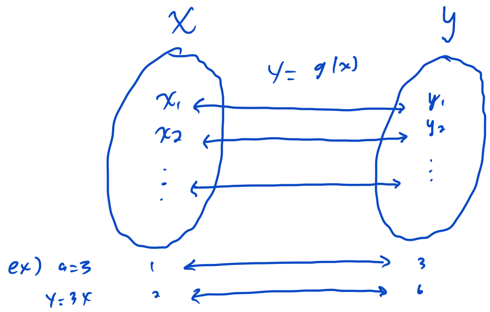
  We can derive this!

$$E_Y(Y) = \sum y\, f_Y(y)$$
$$= \sum a x\, f_Y(ax) \qquad \text{(substitute } y = ax \text{)}$$
$$= \sum a x\, f_X(x) \qquad \text{(if } Y = aX \Rightarrow f_Y(ax) = f_X(x) \text{)}$$

ex) $Y = 3 \iff X = 1$
same as

$$= a \sum x\, f_X(x)$$
$$= a\, E(X)$$

- Transformation mapping
  - If $Y = g(X)$ is a one-to-one function, then the inverse of $g(X)$ exists. So we can go "backwards" in our mapping.



- We can also extend this rule for adding a constant. Continuing example:

  (b) Lets say the insurance company has a yearly fixed cost of \$20 per policyholder for administering the insurance policy. So the total cost in a year for a policy is the sum of the claim payments and the administrative cost.

  Write our new random variable for the total cost and find the expected cost per policy per year.

x	$f(x)$	$z =$ $100X+20$	$f(z)$
0	0.28	20	0.28
1	0.43	120	0.43
2	0.20	220	0.20
3	0.09	320	0.09

Total cost　　total payment amount

$\rightarrow Z = Y + 20$

$= 100X + 20$

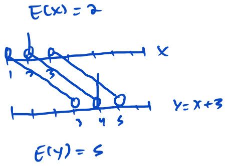
↳ # of claims

$\rightarrow E(Z) = E(100X + 20)$

$= 100E(X) + E(20)$

$= 100(1.1) + 20$

$= 130$



- Theorem: For any constants a and b and random variable X,

$$E(aX + b) = aE(X) + b$$

Discrete derivation:

⟨ follows same pattern as deriving $E(aX)$ ⟩

$$E(Y) = \sum y f_Y(y) = \sum (ax+b) \underbrace{f_Y(ax+b)}_{= f_X(x)} = \sum (ax+b) f_X(x) = a\underbrace{\sum x f(x)}_{=E(X)} + b\underbrace{\sum f_X(x)}_{=1}$$

If $Y = aX+b$　　substitute

$$\stackrel{\checkmark}{=} aE(X) + b$$

Continuous derivation:

$$\text{If } Y = aX + b \rightarrow E(Y) = E(aX+b) = \int_{-\infty}^{\infty} (ax+b) f(x)\,dx$$

$$= a \underbrace{\int_{-\infty}^{\infty} x f(x)}_{=E(X)} + b \underbrace{\int_{-\infty}^{\infty} f(x)\,dx}_{=1}$$

$$\checkmark = aE(X) + b$$

- Simple example: Continuing previous example 2. Suppose that due to inflation the losses on the warranty policy are expected to increase by 5% [redacted] plus fixed cost of $30. Find the expected value for the following year.

$$Y = 1.05X + 30 \Rightarrow E(Y) = E(1.05X + 30) = 1.05\underbrace{E(X)}_{\frac{100}{3}} + 30 = 65$$

- Theorem: For some constant random variable $X = a$,

$$E(X) = E(a) = a$$

(Discrete) Derivation:

$$f(x) = \begin{cases} 1 & x = a \\ 0 & \text{elsewhere} \end{cases} \Rightarrow E(X) = \sum x f(x)$$

$$= a f(a)$$

$$\checkmark = a$$

Generalizing expected value of a function of a random variable

- Now suppose we are working with functions that are more complex than $Y = g(X) = aX + b$, specifically functions that are not one-to-one function. Let's see how to find the expected value when this is the case.
- Example: Let X be a discrete random variable with the pmf given below. Find the expected value of $Y = g(X) = X^2$.

x	-1	0	1
$f(x)$	0.2	0.6	0.20
$g(x) = x^2$	$(-1)^2$	0^2	1^2

same $f(x)$

$$\rightarrow E(g(X)) = E(X^2) = \sum x^2 f(x)$$

$$= 1(0.2) + 0(0.6) + 1(0.2)$$

$$= 0.4$$

OR

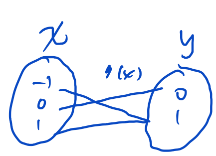
$y = g(x)$	0	1
$f(y)$	0.6	0.4

$$\rightarrow E(Y) = \sum y f(y)$$

$$= 0(0.6) + 1(0.4)$$

$$\checkmark = 0.4$$

More complex because $g(x)$ is not a one-to-one function.



- This example illustrates two major points.
  1. The distribution table for X can be converted into a preliminary table for $g(X)$ with entries for $g(x)$ and $f(x)$, but some regrouping may be necessary to get the actual distribution table for $Y = g(X)$.
  2. Even though the tables are not the same, they lead to the same results for the expected value of $Y = g(X)$.

- Theorem:

  Let X be a discrete random variable. The **expected value** of $Y = g(X)$ is given by

$$E(Y) = \sum_{y} y f_Y(y) = E(g(x)) = \sum_{x} g(x) f_X(x)$$

simplest way

  Let X be a continuous random variable. The **expected value** of $Y = g(X)$ is given by

$$E(Y) = \int_{-\infty}^{\infty} y \underbrace{f_Y(y)}_{=??} \, dy = E(g(x)) = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx$$

only way so far

- Examples:
  1. Let the random variable X have the pmf $f(x) = \frac{x}{10}$ for $x = 1, 2, 3, 4$.

     Find $E(X^2)$ and $E[X(5 - X)]$. $\rightarrow E(x) = 3$

$$\Rightarrow E(x^2) = \sum_{x=1,2,3,4} x^2 \left(\frac{x}{10}\right) = \sum_{x=1,2,3,4} \frac{x^3}{10} = 10$$

$$\Rightarrow E[x(5-x)] = E(5x - x^2) = 5\underbrace{E(x)}_{=3} - \underbrace{E(x^2)}_{=10} = 5$$

  2. Let $f(x) = 3x^{-4}$ for $x > 1$. Find $E(2X^2)$.

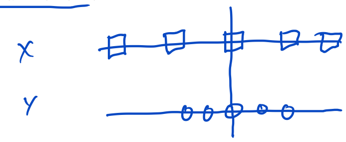
$$E(2x^2) = \int_1^{\infty} 2x^2 (3x^{-4}) \, dx = \int_1^{\infty} 6x^{-2} \, dx = -6x^{-1} \Big|_1^{\infty} = 6$$

.

Variance and standard deviation

Measures of spread

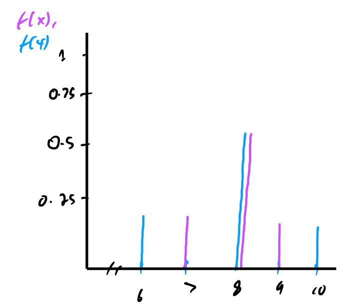
- The mean of a random variable gives a nice single summary number to measure central tendency. However two different random variables can have the same mean and still be quite different.



same center, different spread

Motivating example

- Below is the pmfs of quiz scores for two different classes.



(a) Find the mean of each.

Class 1: RV X

Score (x)	7	8	9
$f(x)$	0.2	0.6	0.2

Class 2: RV Y

Score (y)	6	8	10
$f(y)$	0.2	0.6	0.2

$\mu_X = E(X) = \sum x f(x) = 8$

$\mu_Y = E(Y) = \sum y f(y) = 8$

Means are the same, but obviously the two random variables are quite different. There is much more variation or dispersion in Y than X. The question is, how to measure that variation?

(b) We could look at the distance between each individual value of x or y from the mean of its distribution. $X - \mu_X \rightarrow E(X-\mu_X) = E(X) - \mu_X = 0$

Always true for any random variable

This is always true for any random variable.

(c) However, if we look at the square of the distance between the mean, this problem does not occur. Now find the expected values of each of these new pmfs.

RV $= (X - \mu_X)^2$

$(x-8)^2$	$(7-8)^2 = 1$	$(8-8)^2 = 0$	$(9-8)^2 = 1$
$f(x)$	0.2	0.6	0.2

RV $= (Y - \mu_Y)^2$

$(y-8)^2$	$(6-8)^2 = 4$	$(8-8)^2 = 0$	$(10-8)^2 = 4$
$f(y)$	0.2	0.6	0.2

$E[(X-\mu_X)^2] = \sum (x-\mu_X)^2 f(x) = 1(0.2) + 0(0.6) + 1(0.2) = 0.4$

$E[(Y-\mu_Y)^2] = \sum (y-\mu_Y)^2 f(y) = 4(0.2) + 0(0.6) + 4(0.2) = 1.6$

- This is the single measure of variation that is most widely used in probability theory.

Defining variance and standard deviation

- Definition: The **variance** of a random variable X is defined to be

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

- Definition: The **standard deviation** of a random variable is the square root of its variance. It is denoted by the greek letter σ.

$$\sigma = \sqrt{V(X)}$$

- Notes
  - The variance is also written as σ^2.
  - $SD(X)$ is in the same units as X and $V(X)$ is in units^2.
  - Variance is just a special expected value with $g(X) = (X - \mu)^2$.

- Definitions:

Discrete	Continuous
$V(X) = \sum_x (x - \mu_X)^2 f(x)$	$V(X) = \int_{-\infty}^{\infty} (x-\mu_x)^2 f(x)\,dx$

- Examples:
  1. Let X be the random variable with $f(x) =$ [redacted] $\frac{3}{8}x^2$ for $0 < x < 2$. Find σ_X^2.

① $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^2 x\left(\frac{3}{8}x^2\right)dx$

$= \frac{3}{32} x^4 \Big|_0^2 \; F(2) - F(0)$

$= 1.5$

② $V(X) = E[(X - \underbrace{\mu_X}_{1.5 = 1.5})^2] = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x)\,dx$

$= \int_0^2 (x - 1.5)^2 \frac{3}{8} x^2\,dx$

$= \frac{3}{8}\int_0^2 \left(x^2 - 3x + \frac{9}{4}\right) x^2\,dx$

$= \frac{3}{8}\int_0^2 \left(x^4 - 3x^3 + \frac{9}{4}x^2\right)dx$

$= \frac{3}{8}\left(\frac{1}{5}x^5 - \frac{3}{4}x^4 + \frac{9}{12}x^3\right)\Big|_0^2 \; F(2) - F(0)$

$= 0.15$

2. Let $f(x) = \dfrac{3x+4}{22}$ for $x = -1, 0, 1, 2$.

   Find the mean, variance and standard deviation of X.

① $E(X) = \sum_{x} x f(x) = \sum_{x=-1,0,1,2} x\left(\frac{3x+4}{22}\right) = \frac{1}{22}\sum_{x=-1,0,1,2}(3x^2+4x) = \frac{1}{22}\left[(3(-1)^2+4(-1)) + \cdots + (3(2^2)+4(2))\right] = \frac{13}{11}$

② $V(X) = \sum_{x}(x-E(X))^2 f(x) = \sum_{x=-1,0,1,2}\left(x-\frac{13}{11}\right)^2\left(\frac{3x+4}{22}\right) = \frac{1}{22}\left[\left(\left(-1-\frac{13}{11}\right)^2(3(-1)+4)\right) + \cdots + \left(\left(2-\frac{13}{11}\right)^2(3(2)+4)\right)\right] = \frac{95}{121}$

③ $SD(X) = \sqrt{V(X)} = \sqrt{95/121}$

Expectation as a linear operator

- Now lets say we want to find the expected value of $g(X) = (X-3)^2 = X^2 - 6X + 9$.
- This is a more complicated function of X because it has multiple X's. To find the expected value of $g(X)$ in this case, we can do what intuitively makes sense.

$c_1 = 1$, $g_1(x) = x^2$ $\quad c_2 = 6$, $g_2(x) = x$ $\quad c_3 = 9$, $g_3(x) = 1$

$E(g(X)) = E[(X-3)^2] = E(X^2 - 6X + 9) = E(X^2) - 6E(X) + 9$

can algebra within $E(\cdot)$ $\quad$ Separate & simplify

- Theorem: The reason this works is because of the following property of expectation:

$$E\left[\sum_{i=1}^{k} c_i g_i(X)\right] = \sum_{i=1}^{k} c_i E[g_i(X)]$$

  In fancy words, the expected value of a linear combination equals a linear combination of expected values.

- Because of the property, expectation is often called a **linear (or distributive) operator**.

  Integration and derivation are linear operators as well:

  $\int [g(x) + f(x)]\mathrm{d}x = \int g(x)\,dx + \int f(x)\,dx$

  $\frac{\mathrm{d}}{\mathrm{d}x}[g(x) + f(x)] = \frac{d}{dx} g(x) + \frac{d}{dx} f(x)$

Another way to calculate variance

- Using the property we just showed for expectation, we can arrive at another way to calculate the variance of a random variable.

$$V(X) = E[(X-\mu)^2] \quad \text{(definition)}$$
$$= E[X^2 - 2\mu X + \mu^2] \quad \text{(expand)}$$
$$= E(X^2) - 2\mu \underbrace{E(X)}_{=\mu} + \mu^2 \quad \text{(d istribute \& simplify)}$$
$$= E(X^2) - \mu^2$$

Theorem:

- Another way to calculate variance:

$$V(X) = E(X^2) - \underbrace{\mu^2}_{\text{m an}^2}$$

- Examples:

1. Below is a pmf table for X.

x	0	1	2	3
$f(x)$	0.5	0.3	0.06	0.14

Find $V(X)$ using the new formula.

① $E(X) = \sum_x x f(x) = 0(0.5) + \cdots + 3(0.14) = 0.84$

② $E(X^2) = \sum_x x^2 f(x) = 0^2(0.5) + \cdots + 3^2(0.14) = 1.8$

③ $V(X) = E(X^2) - (E(X))^2 = 1.8 - 0.84^2 = 1.0944$

2. Find the variance of the warranty loss random variable using the alternate formula.

$$f(x) = \begin{cases} 0.02 - 0.0002x & 0 \leq x \leq 100 \\ 0 & \text{otherwise} \end{cases}$$

① $E(X) = \frac{100}{3}$

② $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^{100} x^2(0.02 - 0.0002x)\,dx$

$$= \frac{0.02}{3}x^3 - \frac{.0002}{4}x^4 \Big|_0^{100} \quad F(100) - F(0)$$

$$= 1666.67$$

③ $V(X) = E(X^2) - (E(X))^2 = 1666.67 - \left(\frac{100}{3}\right)^2 = 555.56$

Calculating the variance from the definition would require evaluation of the integral:

$$V(X) = \int_0^{100} \left(x - \frac{100}{3}\right)^2 (0.02 - 0.0002x)\, dx$$

This would be straightforward, but time consuming relative the the other way if done by hand. Of course with computing, calculation time is not an issue.

- Important practical note: The calculation of variance can be done more easily by hand using this alternate form rather than the definition.

  But using the definition is more efficient for computers when large values of X are present (problems with overflow due to the magnitude of X^2).

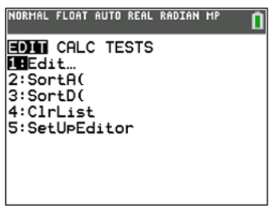
Technology session

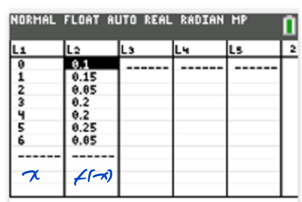
- Using TI-84 (and TI-30XS MultiView) to calculate $E(X)$ and $SD(X)$.

1) Enter data
STAT

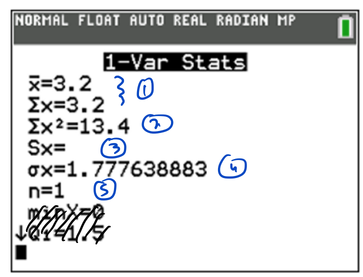
Pmf

x	$f(x)$
0	0.10
1	0.15
2	0.05
3	0.20
4	0.20
5	0.25
6	0.05

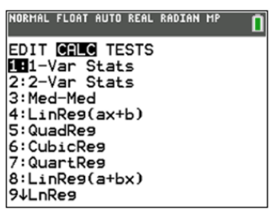


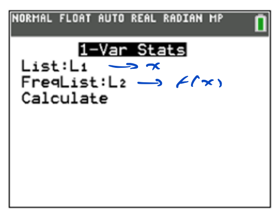


Results



2) Calculate
STAT → CALC





(1) $\bar{x} = E(X) = \sum x f(x)$
"$\sum x$"

(2) $\sum x^2 = E(x^2)$

(5) $\sigma_x = SD(x)$

If we typed in probabilities correctly

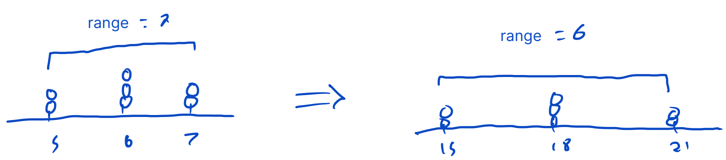
(3) S_x = BLANK or Error
↳ data standard deviation

(5) $n = 1$ ✓ → $\sum f(x)$
↳ Check ✓

Variance and standard deviation of a function of a random variable

Variance and standard deviation of $Y = aX + b$

- Previously, we saw how multiplying by a coefficient a and adding a constant b affected the expected value of our new random variable $Y = aX + b$. Now let's study the affect they have on the variance and standard deviation.
- First we will look at only the affect of the coefficient.



Derivation of $V(aX)$:

If $Y = aX \rightarrow V(Y) = E[(Y-\mu_Y)^2]$

↳ $\mu_Y = a\mu_X$

$$= E[(aX - a\mu_X)^2]$$
$$= E[a^2(X-\mu)^2]$$
$$= a^2 E[(X-\mu)^2]$$
$$\overset{\checkmark}{=} a^2 V(X)$$

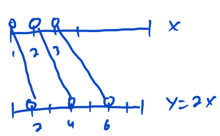
- Theorem: For any constant a and random variable X, (alternate notation)

$$V(aX) = a^2 V(X) \qquad \sigma^2_{aX} = a^2 \sigma^2_X$$

The standard deviation of aX can now be obtained by taking the square root.

$$SD(aX) = \sqrt{a^2 V(X)} = |a|\, SD(X) \qquad \sigma_{aX} = |a|\, \sigma_X$$

- Intuitively, here's why the a is squared for $V(aX)$:



$SD(Y) = 2\,SD(X) \rightarrow V(Y) = [SD(Y)]^2$

"2 fold more spread out"

$$= [2\,SD(X)]^2$$
$$= 2^2\, V(X)$$

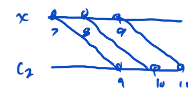
- Example: Using the previous quiz scores from class 1, lets investigate two different types of curves the instructor could use and their impact on the expected value and variance.
  - Curve 1: Each individual score is raised by 25%. $C_1 = 1.25X$

$E(C_1) = E(1.25X)$
$= 1.25\, \underbrace{E(X)}_{=8}$
$= 10$

$V(C_1) = V(1.25X)$
$= 1.25^2\, \underbrace{V(X)}_{=0.4}$
$= 0.625$

– Curve 2: Every score gets and additional 2 points. $c_2 = x + 2$

new center
same spread



Original score (x)	7	8	9
Curved score (c_2)	9	10	11
$f(c_2)$ $= f(x)$	0.2	0.6	0.2

<u>using calculator (showing work)</u>

→ Entered data → $L_1 = c_2$
$L_2 = f(c_2)$

→ TI 84 → 1 var stats (List = L_1, FreqList = L_2) → $(\sigma_{c_2} = 0.632)^2$ ⇒ $V(c_2) \approx 0.4$
TI 30XS → 1 var stats (Data = L_1, FRQ = L_2) ↓ $= V(x)$

- This example shows the intuitive idea that if all values are shifted by exactly b units, the mean changes but the dispersion around the new mean is exactly the same as before.

- Theorem: For any constants a and b and random variable X,

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

- Example: Let's say your Test 1 grades X had the pmf given below (grades were out of 100 points).

  The test corrections gave you 1/2 of the missed points back. Find the expected value and standard deviation of the grades after test corrections. Compare this to the original expected value and standard deviation.

  (a) Define a new random variable Y = grade after test corrections.

Grade (x)	$f(x)$	$Y = 0.5x + 50$ (L_1)	$f(y) = f(x)$
60	0.02	80	.02
65	0.04	82.5	.04
70	0.17	85	.17
75	0.23	87.5	.23
80	0.21	90	.21
85	0.15	92.5	.15
90	0.12	95	.12
95	0.05	97.5	.05
100	0.01	100	.01

$Y = X + (100 - X)0.5$
$= X + 50 - 0.5X$
$= 0.5X + 50$

(b) Calculate $E(Y)$ and $SD(Y)$.

Short way → $E(Y)$ & $V(Y)$ Indirectly

① Find $E(X)$ & $V(X)$ first.

Entered data → $L_1 = x$, $L_2 = f(x)$

1-var stats (List $= L_1$, FreqList $= L_2$) (TI 84)

$\Rightarrow E(X) = 79.05$
$\bar{x}$ in calc

$SD(X) = 8.357 \Rightarrow V(X) = 8.357^2$
σ_x in calc $\downarrow = 69.839$

② Use formulas to find $E(Y)$ & $V(Y) \Rightarrow SD(Y)$

→ $E(Y) = E(0.5X + 50) = 0.5\,\underbrace{E(X)}_{=79.05} + 50 = \boxed{89.525}$

→ $V(Y) = V(0.5X + 50) = 0.5^2\,\underbrace{V(X)}_{=69.839} = 17.464 \Rightarrow SD(Y) = \sqrt{V(Y)}$
$\downarrow \boxed{= 4.179}$

Long way → $E(Y)$ & $SD(Y)$ directly

① finding new pmf of Y

② Then find $E(Y)$ & $SD(Y)$

→ Entered data → $L_1 = y$, $L_2 = f(y)$, $L_3 = 0.5L_1 + 50$ (TI 30XS)

→ 1-var stats (Data $= L_3$, FRQ $= L_2$)

$\Rightarrow \boxed{E(Y) = 89.525}$ check $n = 1$ ✓
$\bar{x}$

$\boxed{SD(Y) = 4.179}$
σ_x

Final comparison: Test corrections have higher mean and less variability.

- Note that we have learned and practiced how to find the expected value of non-linear functions of X such as $Y = X^2 - 7$, but not necessarily for variance. Here's why:

→ $E(Y) = E(X^2 - 7) = E(X^2) - 7 = \left(\int_{-\infty}^{\infty} x^2 f(x)\,dx\right) - 7$ } fairly easy

→ $V(Y) = V(X^2 - 7) = V(X^2)$

→ options → 1) using definition: $V(X^2) = E[(X^2 - E(X^2))^2]$

$\downarrow = \int_{-\infty}^{\infty} (x^2 - E(X^2))^2 f(x)\,dx$

2) using alternate formula

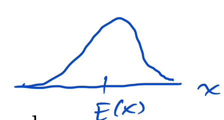
$V(X^2) = E[(X^2)^2] - [E(X^2)]^2$

$\downarrow = E(X^4) - (\mu_{X^2})^2$

} possible, just more involved

3) Find $f_Y(y)$ & find $V(Y)$ like normal → Haven't learned this yet!

Evaluating expected value (as a predictor of X)



- Expected value is a measure of center (i.e. where a distribution is located).

  In other words, we can interpret $E(X)$ as a good guess at a value of X. Here's why this interpretation makes sense:

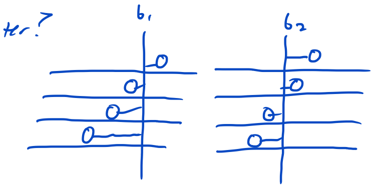
  – Suppose that we measure the distance between a random variable X and a constant b by $(X - b)^2$. The closer b is to X, the smaller this quantity is.

    $(X - b)^2$ is a random variable, which is not a good measure because X changes every time.

    $E[(X - b)^2]$ is constant, which makes it a good measure because it never changes.

b_1 or b_2 "closer" to center?

$\Rightarrow$ minimize collective distance to every point?



- Find the value of b that minimizes $E[(X-b)^2]$ and, hence will provide us with a good predictor of X (i.e. the optimal value of b).

  Steps:

  (a) Rearrange $g(b) = E[(X-b)^2] = E(x^2 - 2xb + b^2)$

  $\downarrow \quad = E(x^2) - 2bE(x) + b^2$

  (b) Goal: Find b to minimize step result of step (a).

  $\rightarrow g'(b) = \frac{d}{db}\left[\cancel{E(x^2)} - 2bE(x) + b^2\right]$

  $\downarrow \quad = -2E(x) + 2b$

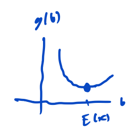
  $\rightarrow 0 = -2E(x) + 2b$

  $\Rightarrow b = E(x)$

  (c) Confirm minimum with second derivative test:

  $g''(b) = \frac{d}{db}[-2E(x) + 2b]$

  $\downarrow \quad = 2 > 0 \Rightarrow$ minimum ✓



$\Rightarrow E(x)$ minimizes $E[(x-b)^2]$

$\Rightarrow E(x)$ is the center

Other measures of center

Introduction

- The mean of a random variable is the most widely used single measure of central tendency.

  There are other measures which are also informative. Each of these has their own interpretation, advantages and shortcomings.

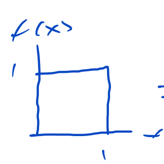
Mode

- Definition: The **mode** is the x value(s) which maximizes the distribution function $f(x)$.

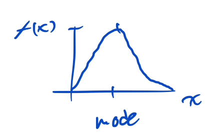
  For discrete random variables, the mode is the x with the highest probability (we can think of this as the most likely); there can be multiple modes.

  For continuous random variables it is the x value where $f(x)$ is the highest.

$f(x) = 1, \; 0 \le x \le 1$



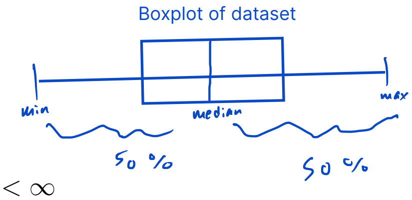
$\Rightarrow$ multiple modes



Median

- The **median** m of a continuous random variable X is the solution of the equation
$$F(m) = P(X \leq m) = 0.5$$

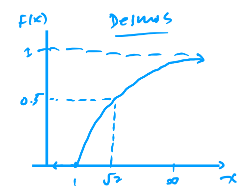
  Thus, we can think of the median m as the "equal areas point", which means the x value dividing the lower 50% probability and the upper 50%.



- Examples

  1. Let X have the following cdf:
$$F(x) = P(X \leq x) = 1 - \frac{1}{x^2}, \qquad 1 \leq x < \infty$$

     Find the median m.



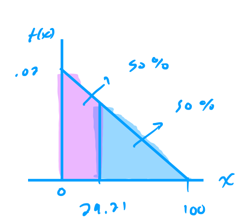
$$0.5 = F(m) = 1 - \frac{1}{m^2}$$
$$0.5 = \frac{1}{m^2}$$
$$m^2 = 2$$
$$m = \pm\sqrt{2} \to \text{in range } 1 < x < \infty$$
$$m = \sqrt{2}$$

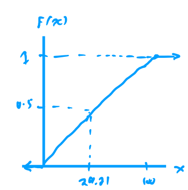
  2. The straight-line density example about losses on a warranty insurance policy had the following pdf and cdf (which we now know how to find):
$$f(x) = \begin{cases} 0.02 - 0.0002x & 0 \leq x \leq 100 \\ 0 & \text{otherwise} \end{cases}$$
$$F(x) = \int_0^x 0.02 - 0.0002t \, dt \;=\; 0.02x - 0.0001x^2, \qquad 0 \leq x \leq 100$$

     Find the median m.

$$\to 0.5 = F(m) = 0.02m - 0.0001m^2$$
$$0 = \underbrace{-0.0001}_{a}m^2 + \underbrace{0.02}_{b}m \underbrace{- 0.5}_{c}$$
$$\to m = \frac{-0.02 \pm \sqrt{0.02^2 - 4(-0.0001)(-0.5)}}{2(-0.0001)}$$
$$\Rightarrow m = 29.29 \quad \text{or} \quad 170.71$$
$$m \in X \qquad m \notin X$$

quadratic formula
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$





This has a nice intuitive interpretation. Half of the losses will be less than \$ 79.21 and the other half will be greater.

3. Symmetric densities: If the density function is symmetric, the median can be found without calculation.

Result: For symmetric densities, the median m is equal to the point of symmetry.

Example: Find the expected value and the median for the ROI example, where

$$f(x) = \begin{cases} 0.75(1-x^2) & -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

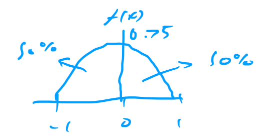
$$E(x) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-1}^{1} x(0.75(1-x^2))\,dx$$
$$= \frac{3}{8}x^2 - \frac{3}{16}x^4 \Big|_{-1}^{1} \quad F(1) - F(-1)$$
$$= \left(\frac{3}{8} - \frac{3}{16}\right) - \left(\frac{3}{8} - \frac{3}{16}\right)$$
$$= 0$$

median

① find cdf $F(x)$

② set $0.5 = F(m)$

③ solve for m



Result: In general, the median and mean are not equivalent.

But when the density function is symmetric, these two measures of center are equal.

Percentiles

- For the previous example we worked out about warranty losses, the median could be interpreted as separating the top 50% of losses from the bottom 50% of losses. For this reason, the median is called the 50$^{\text{th}}$ percentile.

  Other percentiles can be defined using similar reasoning. For example, the 90$^{\text{th}}$ percentile separates the top 10% and the bottom 90%. Below is a general definition of percentile.

- Definition: Let X be a continuous random variable and $0 \le p \le 1$. The $\mathbf{100p^{th}}$ **percentile** of X is the number x_p defined by

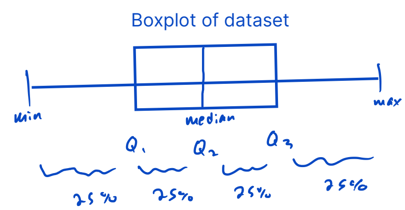
$$F(x_p) = p$$

- Special percentiles: Quartiles (split area under density curve into quarters)

  $x_{0.25} = 25^{th}$ percentile → first, Q_1

  $x_{0.50} = 50^{th}$ percentile → second, Q_2 = median

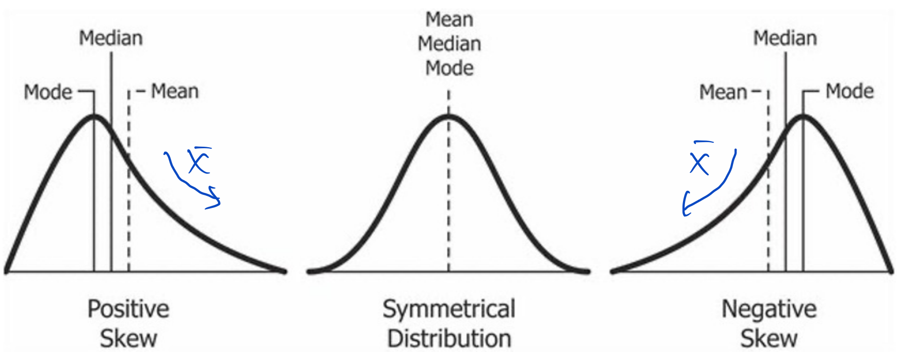
  $x_{0.75} = 75^{th}$ percentile → Third, Q_3



- Another measure of spread (variation):
  - Generally in probability theory, we only talk about standard deviation and variance.
  - In practice (and elementary statistics courses), another measure is often used: inter quartile range (IQR).
  - Definition: The **inter quartile range (IQR)** is equal to the difference of the third and first quartiles:
  $$IQR = Q_3 - Q_1$$

    In probability theory, the IQR is measuring the range of the middle 50% of probability. For datasets, it measures how far data is spread out around the median.
  - In studies, The IQR is typically reported in favor of the standard deviation when the distribution of data is skewed.

    This is because the SD gets inflated due to skewness and outliers, whereas the IQR does not (i.e. it is a **resistant measure**).



- Examples:

4. Find the IQR of X using the following cdf:
$$F(x) = P(X \leq x) = 1 - \frac{1}{x^2}, \qquad 1 \leq x < \infty.$$

Q_1

$0.25 = \frac{1}{4} = F(x_{0.25})$

$\Downarrow \quad = 1 - \frac{1}{x^2}$

$\frac{4}{3} = x^2$

$x_{0.25} = +\sqrt{4/3}$

Q_3

$0.75 = \frac{3}{4} = F(x_{0.75})$

$\downarrow \quad = 1 - \frac{1}{x^2}$

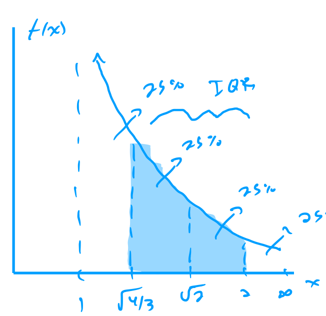
$4 = x^2$

$x_{0.75} = 2$

$\to IQR = Q_3 - Q_1 = x_{0.75} - x_{0.25}$

$= 2 - \sqrt{4/3}$

$= 0.85$



5. The time X in months until failure of a certain product has the pdf

   $f(x) = (1/2)\, x\, \mathrm{e}^{-(x/2)^2}$ for $0 < x < \infty$.

   Find the first and second quartiles of X.

(1) $F(x) = \int_0^x \frac{1}{2} t e^{-(\frac{t}{2})^2} dt$

$= \int_0^x \frac{1}{2} t e^{-t^2/4} dt \longrightarrow$ let $u = t^2/4 \Rightarrow du = \frac{1}{2} t\, dt$

$= \int_0^{x^2/4} e^{-u} du$

$= -e^{-u} \Big|_0^{x^2/4}$

$= e^{-x^2/4} - (-e^{-0})$

$= 1 - e^{-x^2/4}, \quad 0 < x < \infty$

(2) $Q_1 \longrightarrow 0.25 = F(x_{0.25})$

$= 1 - e^{-x^2/4}$

$\ln(e^{-x^2/4}) = \ln(0.75)$

$-x^2/4 = \ln(3/4)$

$x = \sqrt{-4\ln(3/4)} \approx 1.0727$

$Q_2 \longrightarrow 0.5 = F(m)$

$= 1 - e^{-m^2/4}$

$\vdots$

$m = \sqrt{-4\ln(0.5)} \approx 1.6651$