# Distributions of linear combinations

CE 311S

# MORE THAN TWO RANDOM VARIABLES

The same concepts used for two random variables can be applied to three or more random variables, but they are harder to visualize (triple integrals, triple sums, etc.)

One common thing to do with multiple random variables is to calculate *linear combinations* of them.

# DISTRIBUTIONS OF LINEAR COMBINATIONS

A **linear combination** of the random variables $X_1, \ldots, X_n$ has the form

$$a_1X_1 + a_2X_2 + \ldots a_nX_n$$

That is, we multiply each random variable by a constant coefficient, and add them up.

Examples: $X_1 + X_2$; $X_1 - X_2$; $5X_1 + 10X_2 + 3X_3$

To calculate the **total** of $n$ random variables, we have a linear combination with $a_1 = a_2 = \cdots = a_n = 1$

To calculate the **difference** between 2 random variables, we have a linear combination with $a_1 = 1$ and $a_2 = -1$

I want to calculate the toll revenue on SH-130 today. If $X_1$ is the number of cars and $X_2$ the number of semi trucks, the revenue is $a_1X_1 + a_2X_2$ where $a_1$ and $a_2$ are the toll charged to each car and truck.

# We have the following formulas:

For **any** random variables $X_1, \ldots, X_n$

$$E[a_1X_1 + a_2X_2 + \cdots a_nX_n] = a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n]$$

$$V[a_1X_1 + a_2X_2 + \cdots a_nX_n] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\text{Cov}(X_i, X_j)$$

If $X_1, \ldots, X_n$ are independent, the formula for variance simplifies to

$$V[a_1X_1 + a_2X_2 + \cdots a_nX_n] = a_1^2V[X_1] + a_2^2V[X_2] + \cdots + a_n^2V[X_n]$$

# Example

The toll on SH-130 is \$1.50 for cars and \$4.50 for trucks. The mean and variance of the number of cars is 15,000 and 250,000; and the mean and variance of the number of trucks is 5,000 and 10,000. The number of cars and trucks is correlated, with covariance 50,000. What is the mean and standard deviation of the daily toll revenue?

The toll is $1.50X_1 + 4.50X_2$ where $X_1$ and $X_2$ are the number of cars and trucks.

$E[1.50X_1 + 4.50X_2] = 1.50E[X_1] + 4.50E[X_2] = 45\,000$ dollars

$$
\begin{aligned}
V[1.50X_1 + 4.50X_2] &= 1.50^2\text{Cov}(X_1, X_1) + (1.50)(4.50)\text{Cov}(X_1, X_2) \\
&+ (4.50)(1.50)\text{Cov}(X_2, X_1) + 4.50^2\text{Cov}(X_2, X_2) \\
&= 2.25V[X_1] + 13.5\text{Cov}(X_1, X_2) + 20.25V[X_2] \\
&= 1\,440\,000
\end{aligned}
$$

so the standard deviation is $\sqrt{1440000} = 1200$ dollars

# Example

I run a business where my daily revenue has a mean of 1500 and a standard deviation of 400, while my daily costs have a mean of 1000 and a standard deviation of 300. What is the mean and standard deviation of my daily profit, assuming my daily revenue and costs are independent?

$\Pi = R - X$ so $E[\Pi] = E[R] - E[X] = 500$

$V[R - X] = V[R] + V[X] = 300^2 + 400^2 = 500^2$ so $\sigma_{R-X} = 500$

This is one reason why we use variance even though standard deviation is easier to interpret. Variances add nicely, standard deviations do not.

# Example

I run a business where my daily revenue has a mean of 1500 and a standard deviation of 400, while my daily costs have a mean of 1000 and a standard deviation of 300. What is the mean and standard deviation of my daily profit, assuming my daily revenue and costs have a correlation coefficient of $+0.5$?

$\Pi = R - X$ so $E[\Pi] = E[R] - E[X] = 500$

$V[R - X] = V[R] + V[X] - 2\text{Cov}(R, X) =$
$300^2 + 400^2 - 2(0.5)(300)(400) = 130000$ so $\sigma_{R-X} = 360$

If revenue and costs were negatively correlated, would my daily profit have a higher or lower standard deviation?

Let's try to derive these formulas with $n = 2$ to keep the numbers manageable:

$$E\left[a_1X_1 + a_2X_2\right] = \sum_{x_1}\sum_{x_2}(a_1x_1 + a_2x_2)p(x_1, x_2)$$

$$= \sum_{x_1}\sum_{x_2} a_1x_1p(x_1, x_2) + \sum_{x_1}\sum_{x_2} a_2x_2p(x_1, x_2)$$

$$= a_1\sum_{x_1}\sum_{x_2} x_1p(x_1, x_2) + a_2\sum_{x_1}\sum_{x_2} x_2p(x_1, x_2)$$

$$= a_1E[X_1] + a_2E[X_2]$$

Notice that we did not have to assume independence to derive this formula.

# What about the variance?

$$
\begin{aligned}
V\left[a_1X_1 + a_2X_2\right] &= E[(a_1X_1 + a_2X_2 - \mu_{a_1X_1+a_2X_2})^2] \\
&= E[(a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2))^2] \\
&= E[a_1^2(X_1 - \mu_1)^2 + a_1a_2(X_1 - \mu_1)(X_2 - \mu_2) + \\
&\quad a_2a_1(X_2 - \mu_2)(X_1 - \mu_1) + a_2a_2(X_2 - \mu_2)(X_2 - \mu_2)] \\
&= a_1a_1E[(X_1 - \mu_1)(X_1 - \mu_1)] + a_1a_2E[(X_1 - \mu_1)(X_2 - \mu_2)] + \\
&\quad a_2a_1E[(X_2 - \mu_1)(X_1 0\mu_1)] + a_2a_2E[(X_2 - \mu_1)(X_2 - \mu_1)] \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} a_ia_j\mathrm{Cov}(X_1, X_2)
\end{aligned}
$$

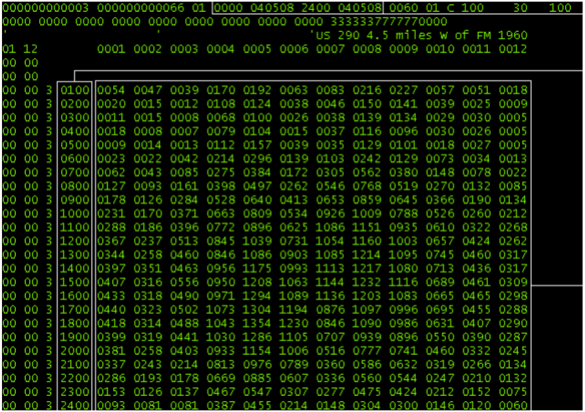
If $X_1$ and $X_2$ are independent, their covariance is zero, so the formula simplifies to

$$a_1^2 \text{Cov}(X_1, X_1) + a_2^2 \text{Cov}(X_2, X_2)$$

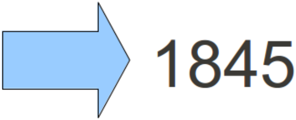
or simply $a_1^2 V[X_1] + a_2^2 V[X_2]$

# WHAT ARE STATISTICS?

Remember the measures of location and variability from Chapter 1?



```
000000000003 000000000066 01 0000 040508 2400 040508 0060 01 C 100     30    100
0000 0000 0000 0000 0000 0000 0000 0000 0000 3333337777770000
'                  '                          'US 290 4.5 miles W of FM 1960
01 12        0001 0002 0003 0004 0005 0006 0007 0008 0009 0010 0011 0012
00 00
00 00
00 00 3 0100 0054 0047 0039 0170 0192 0063 0083 0216 0227 0057 0051 0018
00 00 3 0200 0020 0015 0012 0108 0124 0038 0046 0150 0141 0039 0025 0009
00 00 3 0300 0011 0015 0008 0068 0100 0026 0038 0139 0134 0029 0030 0005
00 00 3 0400 0018 0008 0007 0079 0104 0015 0037 0116 0096 0030 0026 0005
00 00 3 0500 0009 0014 0013 0112 0157 0039 0035 0129 0101 0018 0027 0005
00 00 3 0600 0023 0022 0042 0214 0296 0139 0103 0242 0129 0073 0034 0013
00 00 3 0700 0062 0043 0085 0275 0384 0172 0305 0562 0380 0148 0078 0022
00 00 3 0800 0127 0093 0161 0398 0497 0262 0546 0768 0519 0270 0132 0085
00 00 3 0900 0178 0126 0284 0528 0640 0413 0653 0859 0645 0366 0190 0134
00 00 3 1000 0231 0170 0371 0663 0809 0534 0926 1009 0788 0526 0260 0212
00 00 3 1100 0288 0186 0396 0772 0896 0625 1086 1151 0935 0610 0322 0268
00 00 3 1200 0367 0237 0513 0845 1039 0731 1054 1160 1003 0657 0424 0262
00 00 3 1300 0344 0258 0460 0846 1086 0903 1085 1214 1095 0745 0460 0317
00 00 3 1400 0397 0351 0463 0956 1175 0993 1113 1217 1080 0713 0436 0317
00 00 3 1500 0407 0316 0556 0950 1208 1063 1144 1232 1116 0689 0461 0309
00 00 3 1600 0433 0318 0490 0971 1294 1089 1136 1203 1083 0665 0465 0298
00 00 3 1700 0440 0323 0502 1073 1304 1194 0876 1097 0996 0695 0455 0288
00 00 3 1800 0418 0314 0488 1043 1354 1230 0846 1090 0986 0631 0407 0290
00 00 3 1900 0399 0319 0441 1030 1286 1105 0707 0939 0896 0550 0390 0287
00 00 3 2000 0381 0258 0403 0933 1154 1006 0516 0777 0741 0460 0332 0245
00 00 3 2100 0337 0243 0214 0813 0976 0789 0360 0586 0632 0319 0266 0134
00 00 3 2200 0286 0193 0178 0669 0885 0607 0336 0560 0544 0247 0210 0132
00 00 3 2300 0153 0126 0137 0467 0547 0307 0277 0475 0424 0212 0152 0075
00 00 3 2400 0093 0081 0081 0387 0455 0214 0148 0304 0300 0146 0120 0060
```




What was the purpose of these?

We wanted to use a single number to describe the data set in some way. (This is the definition of a **statistic**. In mathematical terms:

Consider a sample of $n$ elements, and let $X_i$ describe the variable of the $i$-th member of the sample. A **statistic** is a random variable $Y$ which is determined from the random variables $X_1, \ldots, X_n$

Examples:

Sample mean: $Y = \sum_{i=1}^{n} X_i / n$

Maximum value: $Y = \max_{i=1}^{n}\{X_i\}$

Total: $Y = \sum_{i=1}^{n} X_i$

The important thing to notice is that **the statistics are random variables themselves.**

Let's say I roll a die five times, and take the average of the values.

$$6, 6, 3, 1, 1 \to 3.4$$

$$5, 1, 1, 5, 2 \to 2.8$$

$$5, 6, 3, 2, 4 \to 4.0$$

$$3, 3, 2, 4, 5 \to 3.4$$

$$2, 6, 6, 1, 3 \to 3.6$$

Each sample could have a different mean, so the sample means form a random variable (taking the values 3.4, 2.8, 4.0, 3.4, 3.6, and so on). Can we say anything meaningful about its distribution?

# Yes, we can.

In fact, we will shortly see that if $n$ is large, the sample mean has a normal distribution **no matter what the distribution of the $X_i$ is**. Furthermore, its mean is simply the mean of the individual random variables, and its variance is the variance of the individual random variables, divided by $n$.

To begin, we need to make some assumptions about the $X_i$

The random variables $X_1, \ldots, X_n$ are a **random sample** if they are independent and identically distributed.

(This is often abbreviated as the "iid" property.)

Assume that $X_1, \ldots, X_n$ are a random sample, and let $\overline{X}$ represent the sample mean:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

What is $E[\overline{X}]$?

In the dice example above, this is asking what the average is *of the average of five dice rolls*. This is conceptually different from asking what the average is of each roll of the die, although we might think the answers should be the same.

Notice that $\overline{X}$ is a *linear combination* of $X_1, \ldots, X_n$:

$$\overline{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

So therefore

$$E[\overline{X}] = \frac{1}{n}E[X_1] + \cdots + \frac{1}{n}E[X_n]$$

Since $X_1, \ldots, X_n$ are identically distributed, they all have the same mean (call it $\mu$):

$$E[\overline{X}] = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu$$

So, $E[\overline{X}] = \mu$ as well: **the expected value of the sample mean is the expected value of the original random variable**.

So in the dice example, over a long time the average of the sample means (3.4, 2.8, 4.0...) will be very close to 3.5 (the expected value of a single roll).

We can repeat the same idea for variance. Because $\overline{X}$ is a linear combination with weights $1/n$, and because $X_1, \ldots, X_n$ are independent, we have

$$V[\overline{X}] = \frac{1}{n^2} V[X_1] + \cdots + \frac{1}{n^2} V[X_n]$$

Since $X_1, \ldots, X_n$ are identically distributed, they all have the same variance (call it $\sigma^2$):

$$V[\overline{X}] = \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

So, $V[\overline{X}] = \sigma^2/n$: **the variance of the sample mean is NOT the variance of the original random variable, but is smaller by a factor of** $n$.

In the dice example, the variance of the sample mean rolls (3.4, 2.8, 4.0...) will be smaller than the variance of the roll of an individual dice.

(The variance is smaller by a factor of $n$, so the standard deviation is smaller by a factor of $\sqrt{n}$.)

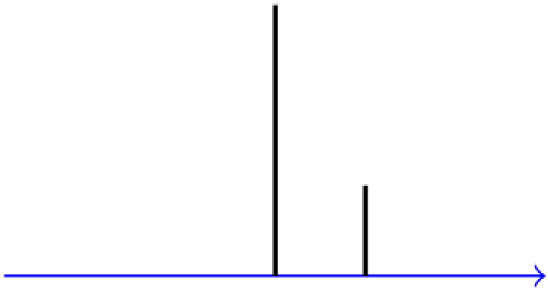
The **central limit theorem** goes one step further and specifies what type of distribution the sample mean has:

Let $X_1, \ldots, X_n$ be a random sample. Then if $n$ is sufficiently large, $\overline{X}$ has approximately a normal distribution, with mean and standard deviation given on the previous slide.

This is true **no matter what distribution the $X_i$ are taken from**. As a practical rule of thumb, if $n > 30$ it is safe to use the Central Limit Theorem.

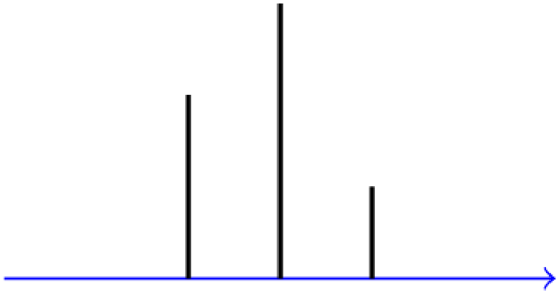
# Visualizing the central limit theorem

This is the PMF for a random variable:



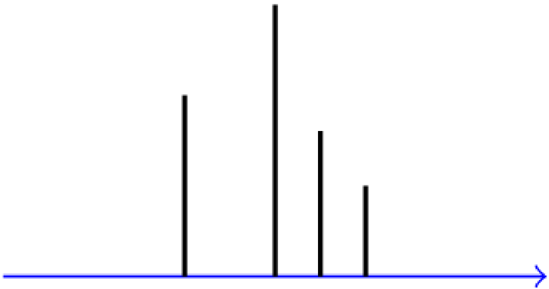
# Visualizing the central limit theorem

This is the PMF of the *average* of two independent draws of the same random variable:



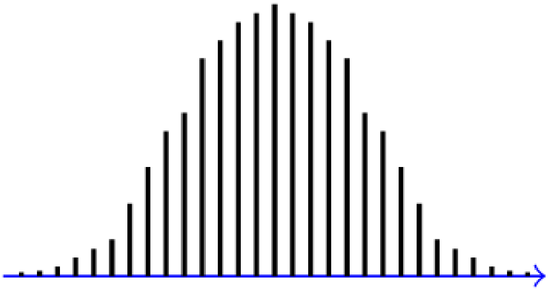
# Visualizing the central limit theorem

This is the PMF of the average of *three* independent draws of the same random variable:



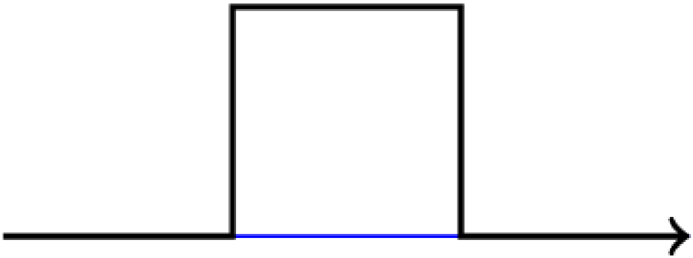
# Visualizing the central limit theorem

This is the PMF of the average of *thirty* independent draws of the same random variable:



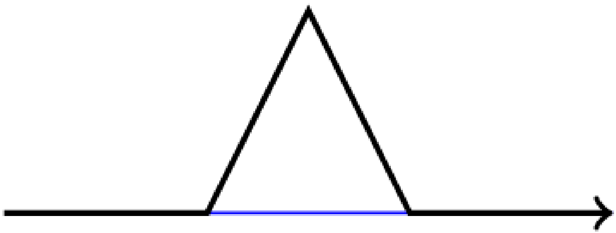
# Visualizing the central limit theorem

This is the PDF for a random variable:



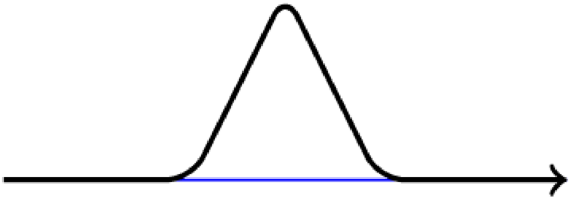
# Visualizing the central limit theorem

This is the PDF of the *average* of two independent draws of the same random variable:



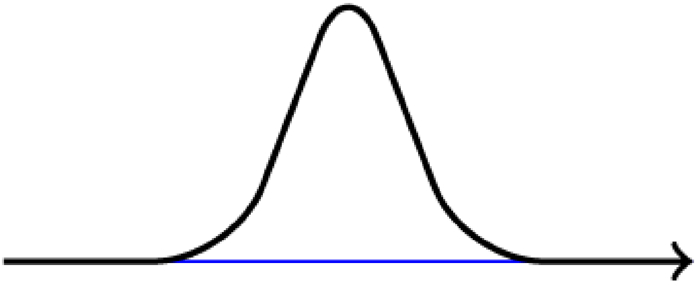
# Visualizing the central limit theorem

This is the PDF of the average of *three* independent draws of the same random variable:



# Visualizing the central limit theorem

This is the PDF of the average of *thirty* independent draws of the same random variable:



# Example

I flip a coin 49 times, and calculate the proportion of flips which were heads. What is the probability that this proportion is between 0.49 and 0.51?