5.4 Joint distributions and independence (discrete)

up until now, only look at probability for 1 r.v. at a time.

Most applications of probability to engineering statistics involve not one but several random variables. In some cases, the application is intrinsically multivariate.

Example 5.32. Consider the assembly of a ring bearing with nominal inside diameter 1.00 in. on a rod with nominal diameter .99 in. If

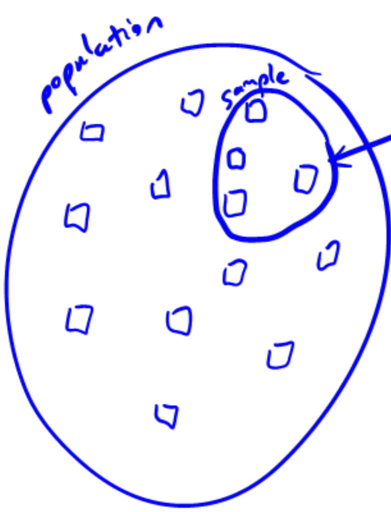
$$X = \text{the ring bearing inside diameter}$$
$$Y = \text{the rod diameter}$$

One might be interested in

$$P[\text{there is an interference in assembly}] = P[X < Y]$$

↳ the assembly can't be made if the rod is thicker than the ring bearing diameter.

Even when a situation is univariate, samples larger than size 1 are essentially always used in engineering applications. The n data values in a sample are usually thought of as subject to chance and their simultaneous behavior must then be modeled.



each of these is subject to chance

⟹ need to talk about their behavior simultaneously

. . .

This is actually a very broad and difficult subject, we will only cover a brief introduction to the topic: **jointly discrete random variables**.

5.4.1 Joint distributions

For several discrete random variable, the device typically used to specify probabilities is a *joint probability function.* The two-variable version of this is defined.

Definition 5.21. A *joint probability function* (joint pmf) for discrete random variables X and Y is a nonnegative function $f(x, y)$, giving the probability that (simultaneously) X takes the values x and Y takes the values y. That is,

$$f(x, y) = P[X = x \text{ and } Y = y]$$

function of two values

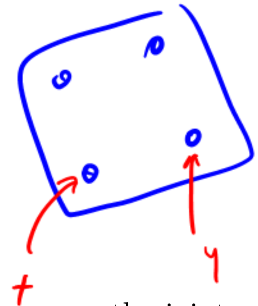
Properties:

1. $f(x,y) \in [0,1]$ for all x,y

2. $\sum_{x,y} f(x,y) = 1$

For the discrete case, it is useful to give $f(x, y)$ in a **table**.

Example 5.33 (Two bolt torques, cont'd). Recall the example of measure the bolt torques on the face plates of a heavy equipment component to the nearest integer. With



X = the next torque recorded for bolt 3
Y = the next torque recorded for bolt 4

} rounded to an integer

the joint probability function, $f(x, y)$, is $= P[X=x, Y=y]$

$x \to$, $y \downarrow$

y\x	11	12	13	14	15	16	17	18	19	20
20	0	0	0	0	0	0	0	2/34	2/34	1/34
19	0	0	0	0	0	0	2/34	0	0	0
18	0	0	1/34	1/34	0	0	1/34	1/34	1/34	0
17	0	0	0	0	2/34	1/34	1/34	2/34	0	0
16	0	0	0	1/34	2/34	2/34	0	0	2/34	0
15	1/34	1/34	0	0	3/34	0	0	0	0	0
14	0	0	0	0	1/34	0	0	2/34	0	0
13	0	0	0	0	1/34	0	0	0	0	0

$P[X = 18 \text{ and } Y = 17]$ $= \frac{2}{34}$

$P[X = 14 \text{ and } Y = 19]$ $= 0$

By summing up certain values of $f(x, y)$, probabilities associated with X and Y with patterns of interest can be obtained.

Consider:

P(bolt 3 requires more torque than bolt 4) →

$$P[X \geq Y] = P(X=13, Y=13) + P(X=14, Y=13) + P(X=14, Y=14) + \dots$$
$$= f(13,13) + f(14,13) + \dots + f(20,13) + P(X=20, Y=13)$$

y\x	11	12	13	14	15	16	17	18	19	20
20										X
19									X	X
18								X	X	X
17							X	X	X	X
16						X	X	X	X	X
15					X	X	X	X	X	X
14				X	X	X	X	X	X	X
13			X	X	X	X	X	X	X	X

$$= \frac{1}{34}(3+1+1+2+1+1+2+2+1+2+1)$$
$$= \frac{17}{34}$$

P(bolt 3 and bolt 4 require torques within 1 unit of each other) =

$$P[|X - Y| \leq 1] = f(12,13) + f(13,14) + f(13,13) + \dots + f(20,19)$$
$$= \frac{1}{34}(2+3+1+1+2+1+1+1+2+2+1+1)$$

y\x	11	12	13	14	15	16	17	18	19	20
20									X	X
19								X	X	X
18							X	X	X	
17						X	X	X		
16					X	X	X			
15				X	X	X				
14			X	X	X					
13		X	X	X						

$$= \frac{18}{34}$$

P(bolt 3 requires 17 units of torque no matter what bolt 4 needs) =

$$P[X = 17] = f(17,20) + \dots + f(17,13)$$
$$= \frac{2}{34} + \frac{1}{34} + \frac{1}{34} = \frac{4}{34}$$

y\x	11	12	13	14	15	16	17	18	19	20
20							X			
19							X			
18							X			
17							X			
16							X			
15							X			
14							X			
13							X			

5.4.2 Maginal distributions

this is called "marginalization" because we put the result in the margins of the table (really).

In a bivariate problem, one can add down columns in the (two-way) table of $f(x, y)$ to get values for the probability function of X, $f_X(x)$ and across rows in the same table to get values for the probability distribution of Y, $f_Y(y)$.

Definition 5.22. The individual probability functions for discrete random variables X and Y with joint probability function $f(x, y)$ are called *marginal probability functions*. They are obtained by summing $f(x, y)$ values over all possible values of the other variable.

function of x

$$f_X(x) = \sum_y f(x, y)$$

$$f_Y(y) = \sum_x f(x, y)$$

function of y

Example 5.34 (Torques, cont'd)**.** Find the marginal probability functions for X and Y from the following joint pmf.

y\x	11	12	13	14	15	16	17	18	19	20	$f_Y(y)$
20	0	0	0	0	0	0	0	2/34	2/34	1/34	5/34
19	0	0	0	0	0	0	2/34	0	0	0	2/34
18	0	0	1/34	1/34	0	0	1/34	1/34	1/34	0	5/34
17	0	0	0	0	2/34	1/34	1/34	2/34	0	0	6/34
16	0	0	0	1/34	2/34	2/34	0	0	2/34	0	7/34
15	1/34	1/34	0	0	3/34	0	0	0	0	0	5/34
14	0	0	0	0	1/34	0	0	2/34	0	0	3/34
13	0	0	0	0	1/34	0	0	0	0	0	1/34
$f_X(x)$	1/34	1/34	1/34	2/34	9/34	3/34	4/34	7/34	5/34	1/34	

So

x	$f_X(x)$
11	1/34
12	1/34
13	1/34
14	2/34
15	9/34
16	3/34
17	4/34
18	7/34
19	5/34
20	1/34

y	$f_Y(y)$
13	1/34
14	3/34
15	5/34
16	7/34
17	6/34
18	5/34
19	2/34
20	5/34

are the marginal probability functions for X and Y

Getting marginal probability functions from joint probability functions begs the question whether the process can be reversed. **Can we find joint probability functions from marginal probability functions?** No! (sometimes yes, more later)

Consider X and Y with joint distributions

y \ x	1	2	3	
3	.4	0	0	.4
2	0	.4	0	.4
1	0	0	.2	.2
	.4	.4	.2	

OR

y \ x	1	2	3	
3	.16	.16	.08	.4
2	.16	.16	.08	.4
1	.08	.08	.04	.2
	.4	.4	.2	

Same marginals, but different joints!

$\Rightarrow$ can't necessarily recover joint from marginals alone

(we need additional assumptions)

5.4.3 Conditional distributions

When working with several random variables, it is often useful to think about what is expected of one of the variables, given the values assumed by all others.

For example, in the bolt torque example, you ought to have some expectations for bolt 4 torque if you know bolt 3 took 15 ft lb to loosen.

Definition 5.23. For discrete random variables X and Y with joint probability function $f(x,y)$, the *conditional probability function of X given $Y=y$* is the function of x

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\sum\limits_x f(x,y)} = \frac{\text{joint}}{\text{marginal of } Y}$$

$X|Y$ "X given Y"

and the *conditional probability function of Y given $X=x$* is the function of y

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x,y)}{\sum\limits_y f(x,y)}. = \frac{\text{joint}}{\text{marginal of } X}$$

Example 5.35 (Torque, cont'd)**.** For the torque example with the following joint distribution, find the following:

1. $f_{Y|X}(20|18)$ = $P(Y=20 \text{ given } X=18) = \frac{f(18,20)}{f_X(18)} = \frac{2/34}{7/34} = \frac{2}{7}$
2. $f_{Y|X}(y|15)$
3. $f_{Y|X}(y|20)$
4. $f_{X|Y}(x|18)$

$y\backslash x$	11	12	13	14	15	16	17	18	19	20	$f_Y(y)$
20	0/34	0/34	0/34	0/34	0/34	0/34	0/34	2/34	2/34	1/34	5/34
19	0/34	0/34	0/34	0/34	0/34	0/34	2/34	0/34	0/34	0/34	2/34
18	0/34	0/34	1/34	1/34	0/34	0/34	1/34	1/34	1/34	0/34	5/34
17	0/34	0/34	0/34	0/34	2/34	1/34	1/34	2/34	0/34	0/34	6/34
16	0/34	0/34	0/34	1/34	2/34	2/34	0/34	0/34	2/34	0/34	7/34
15	1/34	1/34	0/34	0/34	3/34	0/34	0/34	0/34	0/34	0/34	5/34
14	0/34	0/34	0/34	0/34	1/34	0/34	0/34	2/34	0/34	0/34	3/34
13	0/34	0/34	0/34	0/34	1/34	0/34	0/34	0/34	0/34	0/34	1/34
$f_X(x)$	1/34	1/34	1/34	2/34	9/34	3/34	4/34	7/34	5/34	1/34	34/34

2. $f_{Y|X}(y|15) = \frac{f(15,y)}{f_X(15)}$

3. $f_{Y|X}(y|20) = \frac{f(20,y)}{f_X(20)}$

y	$f_{Y\|X}(y\|15)$	$f_{Y\|X}(y\|20)$
13	$1/34/(9/34) = \frac{1}{9}$	0
14	$1/34/(9/34) = \frac{1}{9}$	0
15	$3/34/(9/34) = 3/9$	0
16	$2/34/(9/34) = 2/9$	0
17	$2/34/(9/34) = 2/9$	0
18	$0/34/(9/34) = 0$	0
19	0	0
20	0	$1/34/(1/34) = 1$

4. $f_{X|Y}(x|18) = \frac{f(x,18)}{f_Y(18)}$

x	$f_{X\|Y}(x\|Y=18)$
11	0
12	0
13	$1/34/(5/34) = \frac{1}{5}$
14	$1/34/(5/34) = \frac{1}{5}$
15	0
16	0
17	$1/34/(5/34) = \frac{1}{5}$
18	$1/34/(5/34) = \frac{1}{5}$
19	$1/34/(5/34) = \frac{1}{5}$
20	0

5.4.4 Independence

Recall the following joint distribution:

$y \backslash x$	1	2	3	$f_Y(y)$
3	0.08	0.08	0.04	0.20
2	0.16	0.16	0.08	0.40
1	0.16	0.16	0.08	0.40
$f_X(x)$	0.40	0.40	0.20	1.00

What do you notice?

Each $P[X=x, Y=y] = P[X=x]\,P[Y=y]$

Also, $f_{Y|X}(y|3) = \dfrac{f(3,y)}{f_X(3)} \Rightarrow$

y	$f_{Y\mid X}(y\mid 3)$
1	$.08/.2 = .4$
2	$.08/.2 = .4$
3	$.04/.2 = .2$

So, $f_{Y|X}(y|3) = f_Y(y)$. Actually, this is true for all values of x.

i.e. Knowing what value X takes, doesn't matter in questions about Y!

Definition 5.24. Discrete random variables X and Y are *independent* if their joint distribution function $f(x, y)$ is the product of their respective marginal probability functions. This is, independence means that

assume X, Y indep.

$f_{Y|X}(y|x) = \dfrac{f(x,y)}{f_X(x)} = \dfrac{\cancel{f_X(x)} f_Y(y)}{\cancel{f_X(x)}} = f_Y(y).$

$$P[X=x, Y=y] = f(x, y) = f_X(x) f_Y(y) \qquad \text{for all } x, y.$$

$= P[X=x]P[Y=y]$ (for discrete)

If this does not hold, then X and Y are *dependent.*

Alternatively, discrete random variables X and Y are independent if for all x and y,

$$f_{Y|X}(y|x) = f_Y(y) \quad \text{and} \quad f_{X|Y}(x,y) \overset{!}{=} f_X(x)$$

If X and$Y$ are not only independent but also have the same marginal distribution, then they are **independent and identically distributed (iid)**.

5.5 Functions of several random variables

joint distributions

We've now talked about ways to simultaneously model several random variables. An important engineering use of that material is in the analysis of system output that are functions of random inputs.

5.5.1 Linear combinations

For engineering purposes, it often suffices to know the mean and variance for a function of several random variables, $U = g(X_1, X_2, \ldots, X_n)$ (as opposed to knowing the whole distribution of U). When g is **linear**, there are explicit functions.

Proposition 5.1. *If $X_1, X_2, \ldots, X_n$ are n independent random variables and $a_0, a_1, \ldots, a_n$ are $n+1$ constants, then the random variable $U = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$ has mean*

U is a linear combination of $X_1, \ldots, X_n$

$$EU = a_0 + a_1EX_1 + a_2EX_2 + \cdots + a_nEX_n$$

← this holds even if $X_1, \ldots, X_n$ are not independent.

and variance

$$VarU = a_1^2 \overset{\text{Var}}{E}X_1 + a_2^2 \overset{\text{Var}}{E}X_2 + \cdots + a_n^2 \overset{\text{Var}}{E}X_3$$

idea of proof:

For $n=2$ joint probability function $f(x_1,x_2) = P[X_1 = x_1, X_2 = x_2]$.

Define $U = a_0 + a_1X_1 + a_2X_2$

$$EU = E[a_0 + a_1X_1 + a_2X_2]$$

$$= \sum_{x_1}\sum_{x_2}(a_0 + a_1x_1 + a_2x_2) f(x_1,x_2)$$

$$= \sum_{x_1, x_2}\sum a_0 f(x_1,x_2) + \sum_{x_1}\sum_{x_2} a_1 x_1 f(x_1,x_2) + \sum_{x_1}\sum_{x_2} a_2 x_2 f(x_1,x_2)$$

$$= a_0 \underbrace{\boxed{\sum_{x_1}\sum_{x_2} f(x_1,x_2)}}_{=1} + a_1 \sum_{x_1} x_1 \underbrace{\boxed{\sum_{x_2} f(x_1,x_2)}}_{f_{X_1}(x_1)} + a_2 \sum_{x_2} x_2 \underbrace{\boxed{\sum_{x_1} f(x_1,x_2)}}_{f_{X_2}(x_2)}$$

$$= a_0 + a_1 \underbrace{\boxed{\sum_{x_1} x_1 f_{X_1}(x_1)}}_{EX_1} + a_2 \underbrace{\boxed{\sum_{x_2} x_2 f_{X_2}(x_2)}}_{EX_2} = a_0 + a_1EX_1 + a_2EX_2.$$

Check on your own, same ideas hold for $VarU = a_1^2 VarX_1 + a_2^2 VarX_2$. //

Example 5.36. Say we have two independent random variables X and Y with $\mathrm{E}X = 3.3, \mathrm{Var}X = 1.91, \mathrm{E}Y = 25$, and $\mathrm{Var}Y = 65$. Find the mean and variance for

$$U = 3 + 2X - 3Y$$
$$V = -4X + 3Y$$
$$W = 2X - 5Y$$
$$Z = -4X - 6Y$$

$$\begin{aligned} EU &= E(3+2X-3Y) \\ &= 3 + 2EX - 3EY \\ &= 3 + 2(3.3) - 3(25) = -65.4 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}\,U &= \mathrm{Var}(3+2X-3Y) \\ &= 2^2 \mathrm{Var}X + (-3)^2 \mathrm{Var}Y \\ &= 4(1.91) + 9(65) = 592.64 \end{aligned}$$

$$\begin{aligned} EV &= E(-4X+3Y) \\ &= -4EX + 3EY \\ &= -4(3.3) + 3(25) = 61.8 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}\,V &= \mathrm{Var}(-4X+3Y) \\ &= (-4)^2 \mathrm{Var}X + 3^2 \mathrm{Var}Y \\ &= 16(1.91) + 9(65) = 615.56 \end{aligned}$$

$$\begin{aligned} EW &= E(2X-5Y) \\ &= 2EX - 5EY \\ &= 2(3.3) - 5(25) = -118.4 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}\,W &= \mathrm{Var}(2X-5Y) \\ &= 2^2 \mathrm{Var}X + (-5)^2 \mathrm{Var} \\ &= 4(1.91) + 25(65) = 1632.64 \end{aligned}$$

$$\begin{aligned} EZ &= E[-4X-6Y] \\ &= -4EX - 6EY \\ &= -4(3.3) - 6(25) = -163.2 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}\,Z &= \mathrm{Var}(-4X-6Y) \\ &= (-4)^2 \mathrm{Var}X + (-6)^2 \mathrm{Var}Y \\ &= 16(1.91) + 36(65) = \\ &\qquad 2370.56 \end{aligned}$$

and X and Y are independent.

Example 5.37. Say $X \sim Binomial(n = 10, p = 0.5)$ and $Y \sim Poisson(\lambda = 3)$. Calculate the mean and variance of $Z = 5 + 2X - 7Y$.

First note

$EX = n \cdot p = 10 \cdot 0.5 = 5$

$Var X = n \cdot p(1-p) = 10 \cdot 0.5 \cdot 0.5 = 2.5$

$EY = \lambda = 3$

$Var Y = \lambda = 3$

Then $EZ = E(5+2X-7Y)$
$= 5 + 2EX - 7EY$
$= 5 + 2(5) - 7(3)$
$= -6$

$Var Z = Var(5 + 2X - 7Y)$
$= 2^2 Var X + (-7)^2 Var Y$
$= 4(2.5) + 49(3)$
$= 157$

independent + identically distributed

A particularly important use of Proposition 5.1 concerns n iid random variables where each $a_i = \frac{1}{n}$, for $i=1,..,n$

$X_1, .., X_n$ are conceptually equivalent to random selections (with replacement) from a single numerical population

We can find the mean and variance of the random variable

sample mean.

$$\overline{X} = \frac{1}{n}X_1 + \cdots \frac{1}{n}X_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$X_1, .., X_n$ are identically distributed,
$EX_1 = \cdots = EX_n = \mu$
$VarX_1 = \cdots = VarX_n = \sigma^2$

as they relate to the population parameters $\mu = \mathrm{E}X_i$ and $\sigma^2 = \mathrm{Var}X_i$.

For independent variables $X_1, \ldots, X_n$ with common mean μ and variance σ^2,

$$\begin{aligned}\mathrm{E}\overline{X} &= E\left[\tfrac{1}{n}X_1 + \cdots + \tfrac{1}{n}X_n\right] \\ &= \tfrac{1}{n}EX_1 + \cdots + \tfrac{1}{n}EX_n \quad \text{(prop 5.1)} \\ &= \underbrace{\tfrac{1}{n}\mu + \cdots + \tfrac{1}{n}\mu}_{n \text{ terms}} \\ &= \tfrac{1}{n}\cdot n\mu = \mu\end{aligned}$$

So, the expected value of the sample mean is population mean.

$$\begin{aligned}\mathrm{Var}\overline{X} &= Var\left[\tfrac{1}{n}X_1 + \cdots + \tfrac{1}{n}X_n\right] \\ &= \left(\tfrac{1}{n}\right)^2 Var X_1 + \cdots + \left(\tfrac{1}{n}\right)^2 Var X_n \quad \text{(prop 5.1)} \\ &= \underbrace{\tfrac{1}{n^2}\sigma^2 + \cdots + \tfrac{1}{n^2}\sigma^2}_{n \text{ terms}} \\ &= \tfrac{1}{n^2}\cdot n\cdot\sigma^2 = \sigma^2/n\end{aligned}$$

The variance of the sample mean for a sample of size n is the population variance divided by the sample size.

ie. as the sample size grows, the variability of the sample mean decreases.

Example 5.38 (Seed lengths)**.** One botanist measured the length of 10 seeds from the same plant. The seed lengths measurements are $X_1, X_2, \ldots, X_{10}$. Suppose it is known that the seed lengths are iid with mean $\mu = 5$ mm and variance $\sigma^2 = 2$mm.

Calculate the mean and variance of the average of 10 seed measurements.

$\bar{X}$ = average of 10 measurements
$= \frac{1}{10} \sum_{i=1}^{10} X_i$

Since X_i iid with $\mu = 5$, $\sigma^2 = 2$,

$E\bar{X} = \mu = 5$ mm

$\mathrm{Var}\,\bar{X} = \frac{1}{n}\sigma^2 = \frac{2}{10} = 0.2$

5.5.2 Central limit theorem

One of the most frequently used statistics in engineering applications is the sample mean. We can relate the mean (location) and variance (spread) of the probability distribution of the sample mean to those of a single observation when an iid model is appropriate.

* in the case of the sample mean, if the sample size is large enough, we can also approximate the shape of the probability distribution of the sample mean!

Central Limit Theorem (CLT)

Proposition 5.2. *If $X_1, \ldots, X_n$ are* ① *iid random variable (with mean μ and variance σ^2), then for large n, the variable* ② *$\overline{X}$ is approximately normally distributed. That is,*

③ $n \geq 25$

"approximately distributed"

$$\overline{X} \mathrel{\dot\sim} Normal\left(\mu, \frac{\sigma^2}{n}\right).$$

This is one of the **most important** results in statistics.

Example 5.39 (Tool serial numbers)**.** Consider selecting the last digit of randomly selected serial numbers of pneumatic tools. Let

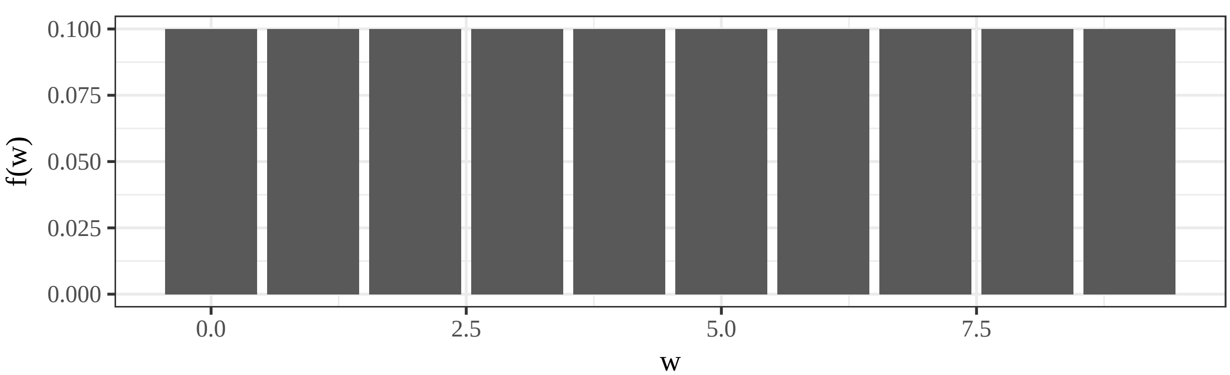
$W_1 =$ the last digit of the serial number observed next Monday at 9am

$W_2 =$ the last digit of the serial number observed the following Monday at 9am

A plausible model for the pair of random variables W_1, W_2 is that they are independent, each with the marginal probability function

$$f(w) = \begin{cases} .1 & w = 0, 1, 2, \ldots, 9 \\ 0 & \text{otherwise} \end{cases}$$

uniform over $0, \ldots, 9$



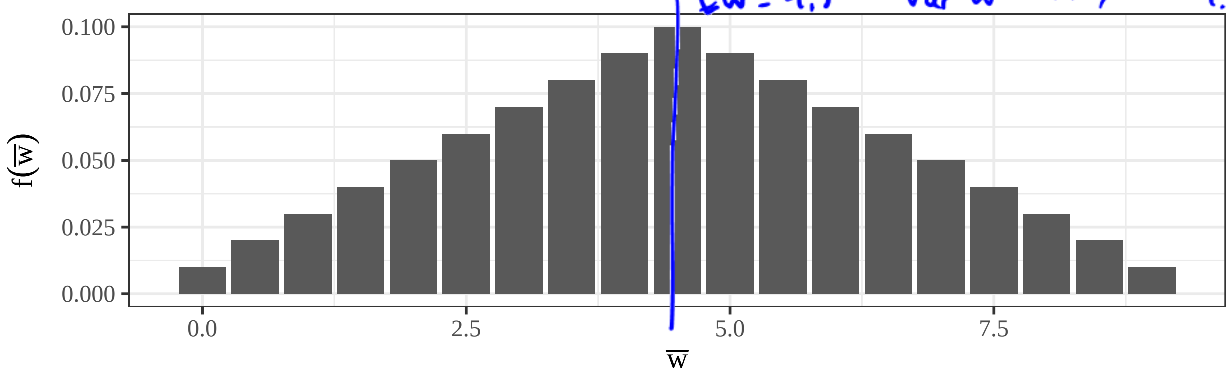
With E$W = 4.5$ and Var$W = 8.25$.

check this on your own

Using such a distribution, it is possible to see that $\overline{W} = \frac{1}{2}(W_1 + W_2)$ has probability distribution

$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$
0.00	0.01	2.00	0.05	4.00	0.09	6.00	0.07	8	0.03
0.50	0.02	2.50	0.06	4.50	0.10	6.50	0.06	8.5	0.02
1.00	0.03	3.00	0.07	5.00	0.09	7.00	0.05	9	0.01
1.50	0.04	3.50	0.08	5.50	0.08	7.50	0.04		

$E\overline{W} = 4.5$ $\quad Var\,\overline{W} = 8.25/2 = 4.125$



Comparing the two distributions, it is clear that even for a completely flat/uniform distribution of W and a small sample size of $n = 2$, the probability distribution of $\overline{W}$ looks more bell-shaped than the underlying distribution.

Now consider larger and larger sample sizes, $n = 1, \ldots, 40$: and look at the distribution of the sample mean for larger and larger samples.

Click for video... online.

$\overline{W}$ will always have $E\overline{W} = 4.5$, $Var\,\overline{W} = \frac{8.25}{n}$

and approaches normality as $n \to \infty$.

Example 5.40 (Stamp sale time)**.** Imagine you are a stamp salesperson (on eBay). Consider the time required to complete a stamp sale as S, and let

$n = 100 \geq 25$

$$\overset{③}{\overline{S}} = \underline{\text{the sample mean}} \text{ time required to complete the next } \overset{②}{100} \text{ sales}$$

Each individual sale time should have an $Exp(\alpha = 16.5s)$ distribution. We want to consider approximating $P[\overline{S} > 17]$.

assumed iid ①

$$S_i \overset{iid}{\sim} Exp(16.5) \quad i = 1, \dots, 100$$

$$\left.\begin{aligned} \mu = ES_i &= 16.5 \\ \sigma^2 = Var S_i &= 16.5^2 = 272.25 \end{aligned}\right\} \text{ for } i = 1, \dots, 100 \quad (\text{iid})$$

$$\Rightarrow \quad E\overline{S} = 16.5$$

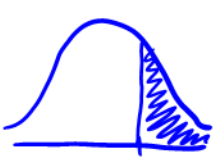
$$Var\,\overline{S} = \frac{272.25}{100} = 2.7225$$

Since $n = 100 \geq 25$,

$$\overline{S} \,\dot{\sim}\, N(16.5,\ 2.7225 = 1.65^2)$$

$$P(\overline{S} > 17) = P\left(\frac{\overline{S} - 16.5}{1.65} > \frac{17 - 16.5}{1.65}\right)$$

$$\approx P(Z > 0.303) \quad \text{where } Z \sim N(0,1) \text{ by CLT}$$



.303

$$= 1 - P(Z \leq 0.303)$$

$$= 1 - \Phi(0.303)$$

$$\approx 1 - 0.6217 \quad (\text{at least})$$

$$= 0.3783$$

Example 5.41 (Cars)**.** Suppose a bunch of cars pass through certain stretch of road. Whenever a car comes, you look at your watch and record the time. Let X_i be the time (in minutes) between when the i^{th} car comes and the $(i+1)^{th}$ car comes for $i = 1, \ldots, 44$. Suppose you know the average time between cars is 1 minute. Find the probability that the average time gap between cars exceeds [redacted] 1.05 minutes.

for next 44 cars

X_i = time in minutes between i^{th} car and $(i+1)^{th}$ car

$\Rightarrow X_i \overset{iid}{\sim} \text{Exp}(\alpha)$, where $\alpha = 1$ for $i = 1, \ldots, 44$

Let $\bar{X} = \frac{1}{44} \sum_{i=1}^{44} X_i$ (average gap between cars for 44 cars)

want $P(\bar{X} > 1.05)$.

$EX_i = \alpha = 1$ for $i = 1, \ldots, 44$

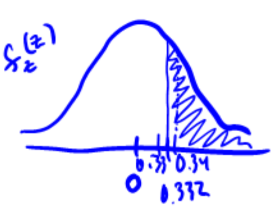
$\text{Var} X_i = \alpha^2 = 1$

iid $\Rightarrow E\bar{X} = 1$

$\text{Var}\bar{X} = \frac{1}{44}$

$n = 44 \geq 25 \Rightarrow \bar{X} \,\dot\sim\, N(1, \frac{1}{44})$

$$P(\bar{X} > 1.05) = P\left(\frac{\bar{X} - 1}{\sqrt{1/44}} > \frac{1.05 - 1}{\sqrt{1/44}}\right)$$



$\approx P(Z > 0.332)$ $Z \sim N(0,1)$ because $n = 44 \geq 25$ by CLT

$\approx P(Z > 0.34)$ (at least)

$= 1 - P(Z \leq 0.34)$

$= 1 - \Phi(0.34)$

$= 1 - 0.6331 = 0.3669$

Example 5.42 (Baby food jars, cont'd)**.** The process of filling food containers appears to have an inherent standard deviation of measured fill weights on the order of 1.6g. Suppose we want to calibrate the filling machine by setting an adjustment knob and filling a run of n jars. Their sample mean net contents will serve as an indication of the process mean fill level corresponding to that knob setting.

You want to choose a sample size, n, large enough that there is an 80% chance the sample mean is within 0.3 g of the actual process mean.

We want to choose n such that

$$0.8 = P[\mu - 0.3 \le \bar{X} \le \mu + 0.3]$$

where X_i = the weight of 1 jar   assume iid

$\bar{X}$ = the sample mean weight of n jars.

For n large enough ($n \ge 25$), we know that

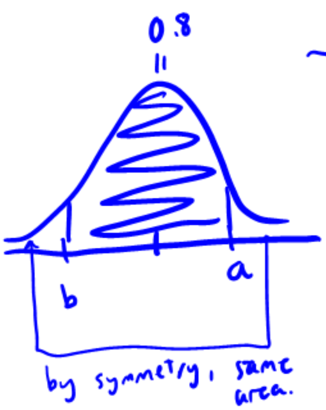
$\bar{X} \dot\sim N(\mu, \frac{\sigma^2}{n})$ where $\sigma^2 = 1.6^2$ (again, we don't know μ)

$\mathrm{Var}\,\bar{X} = \frac{1.6^2}{n} \Rightarrow$ St. dev $\bar{X} = \sqrt{\frac{1.6^2}{n}} = \frac{1.6}{\sqrt{n}}$

$$0.8 = P[\mu - 0.3 \le \bar{X} \le \mu + 0.3]$$

$$= P\left[\frac{\mu - 0.3 - \mu}{\underbrace{1.6/\sqrt{n}}_{\text{st dev of } \bar{X}}} \le \frac{\bar{X} - \mu}{1.6/\sqrt{n}} \le \frac{\mu + 0.3 - \mu}{1.6/\sqrt{n}}\right]$$

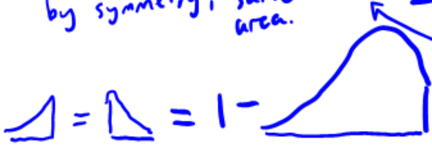
$$\approx P\left[-\frac{0.3}{1.6/\sqrt{n}} \le Z \le \frac{0.3}{1.6/\sqrt{n}}\right]$$ where $Z \sim N(0,1)$ (if $n \ge 25$) by CLT



$$= \Phi\Big(\underbrace{\frac{0.3}{1.6/\sqrt{n}}}_{a}\Big) - \Phi\Big(\underbrace{-\frac{0.3}{1.6/\sqrt{n}}}_{b}\Big)$$

$$= \Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right) - \left[1 - \Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right)\right]$$

$$= 2\Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right) - 1$$



So, $\frac{1.8}{2} = \Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right)$

$\Rightarrow \Phi^{-1}(0.9) = \frac{0.3}{1.6/\sqrt{n}}$

$\left(\Phi^{-1}(0.9) \cdot \frac{1.6}{0.3}\right)^2 = n$

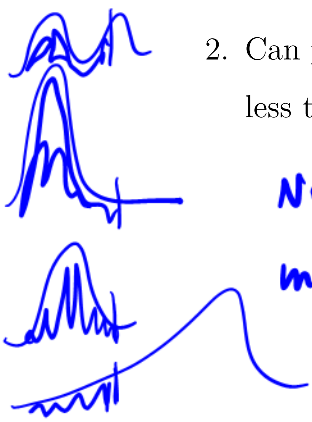
$\left(1.29 \cdot \frac{1.6}{0.3}\right)^2 =$ 47.3344

choose $n = 48$

Example 5.43 (Printing mistakes). Suppose the number of printing mistakes on a page follows some unknown distribution with a mean of 4 and a variance of 9. Assume that number of printing mistakes on a printed page are iid.

1. What is the approximate probability distribution of the average number of printing mistakes on 50 pages?

$\bar{X} \dot{\sim} N(4, \frac{9}{50})$ by CLT since $n = 50 \geq 25$ have iid data.



2. Can you find the probability that the number of printing mistakes on a single page is less than 3.8?

No, because the probability distribution of # of printing mistakes on a single page is unknown.

3. Can you find the probability that the average number of printing mistakes on 10 pages is less than 3.8?

No, because $n = 10 < 25$, so the CLT cannot be used

Thus, the distribution of $\bar{X}$ is unknown.

4. Can you find the probability that the average number of printing mistakes on 50 pages is less than 3.8?

Yes, because $n = 50 \geq 25$, and X_i iid. $\bar{X} \dot{\sim} N(4, \frac{9}{50})$

$$P(\bar{X} < 3.8) = P\left(\frac{\bar{X} - 4}{\sqrt{9/50}} < \frac{3.8 - 4}{\sqrt{9/50}}\right)$$
$$\approx P(Z < -0.4714)$$
$$\approx \Phi(-.48) \text{ "atleast"}$$
$$= 0.3156$$