Math 530/630: CM 4.1

Sampling distributions

Samples

- A **sample** includes the results of n random experiments. A sequence of random variables.

- A **random sample** is the result of a random experiment **repeated** n times. Because experiment can be repeated, and observations are “exchangeable”, iid.

iid sequences of random variables

Number of random variables

		1 rv	> 1 rv
iid?	no	Body weight of Angela {X = 150}	Weight and height of Angela *Data is a sequence of rvs:* {X = 150; Y = 65}

	yes	Body weight of n = 4 randomly selected people
		Data is a sequence of iid rvs:
		{X_1 = 150, X_2 = 125, X_3 = 208, X_4 = 180}
			Weight and height of n = 4 randomly selected people
			Data is a sequence of iid rvs:
			{X_1 = 150, X_2 = 125, X_3 = 208, X_4 = 180}
			{Y_1 = 65, Y_2 = 60, Y_3 = 72, Y_4 = 70}

Both of these are **sequences** of iid random variables

Sequences of rvs

$$X_i \sim N(\mu_i, \sigma_i)$$

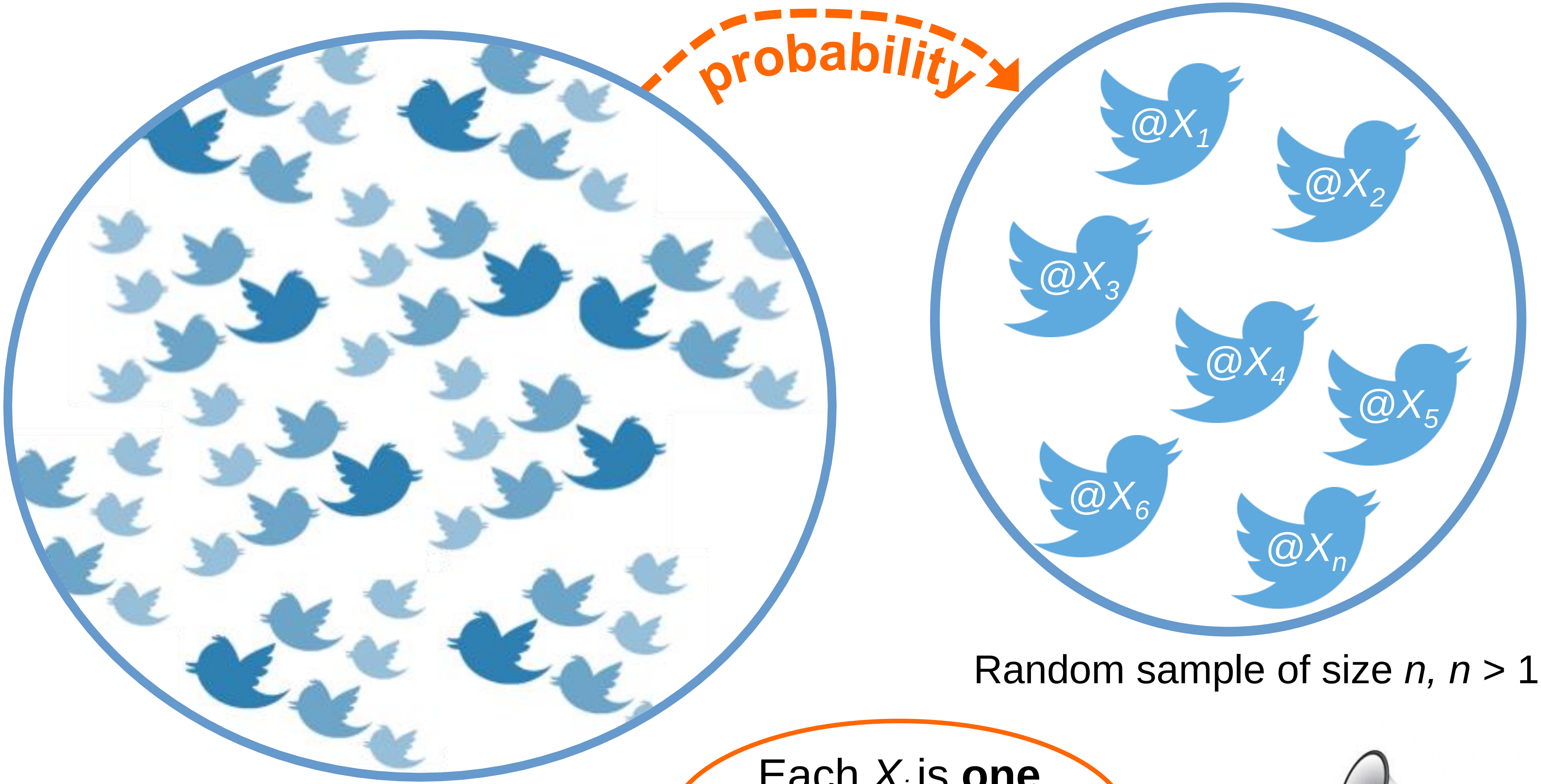
$$X_i \sim N(\mu, \sigma_i)$$

$$X_i \sim N(\mu_i, \sigma)$$

$$X_i \sim N(\mu, \sigma)$$

iid

Random variable, X = # of tweets per day

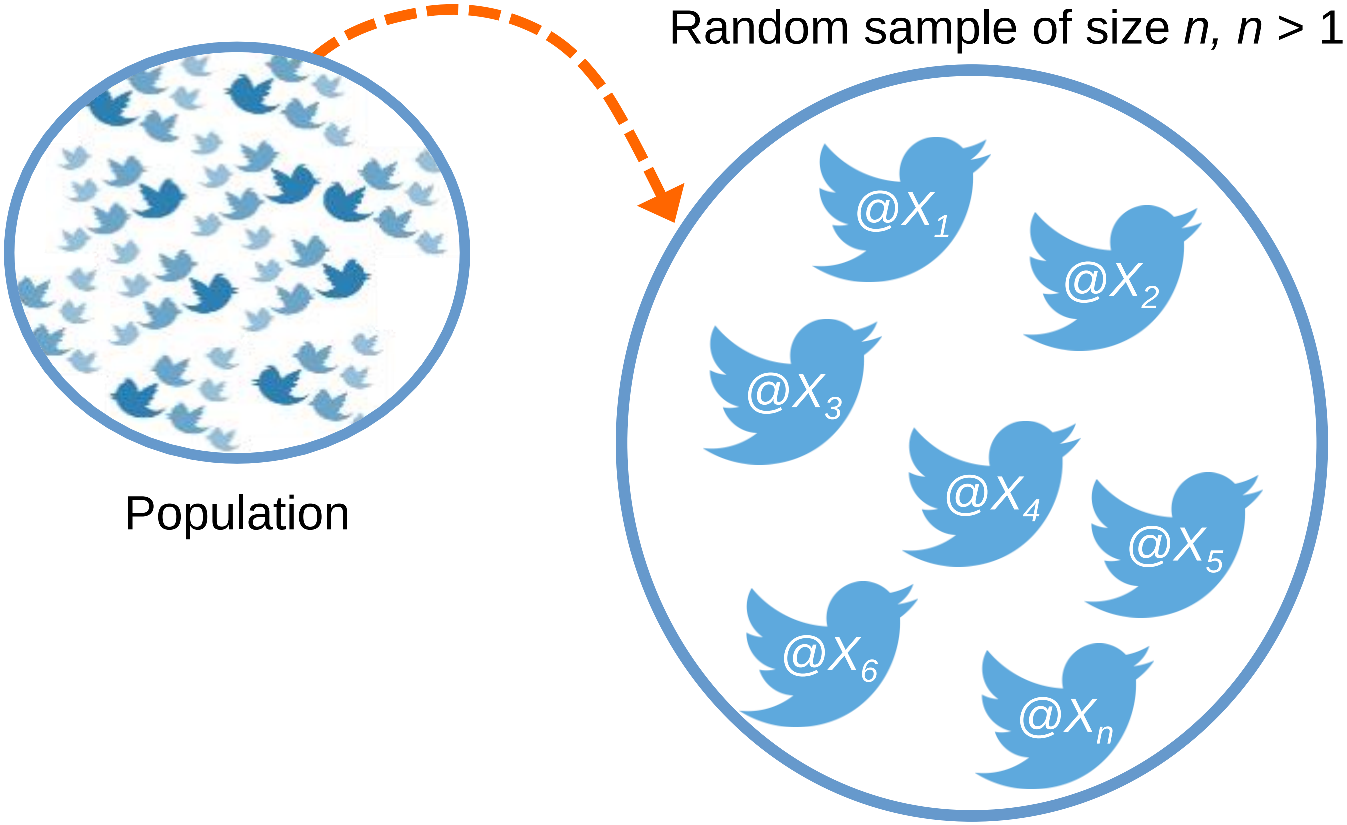


Random sample of size n, $n > 1$

Population

Each X_i is **one** observation on the **same** rv

Random variable, X = # of tweets per day

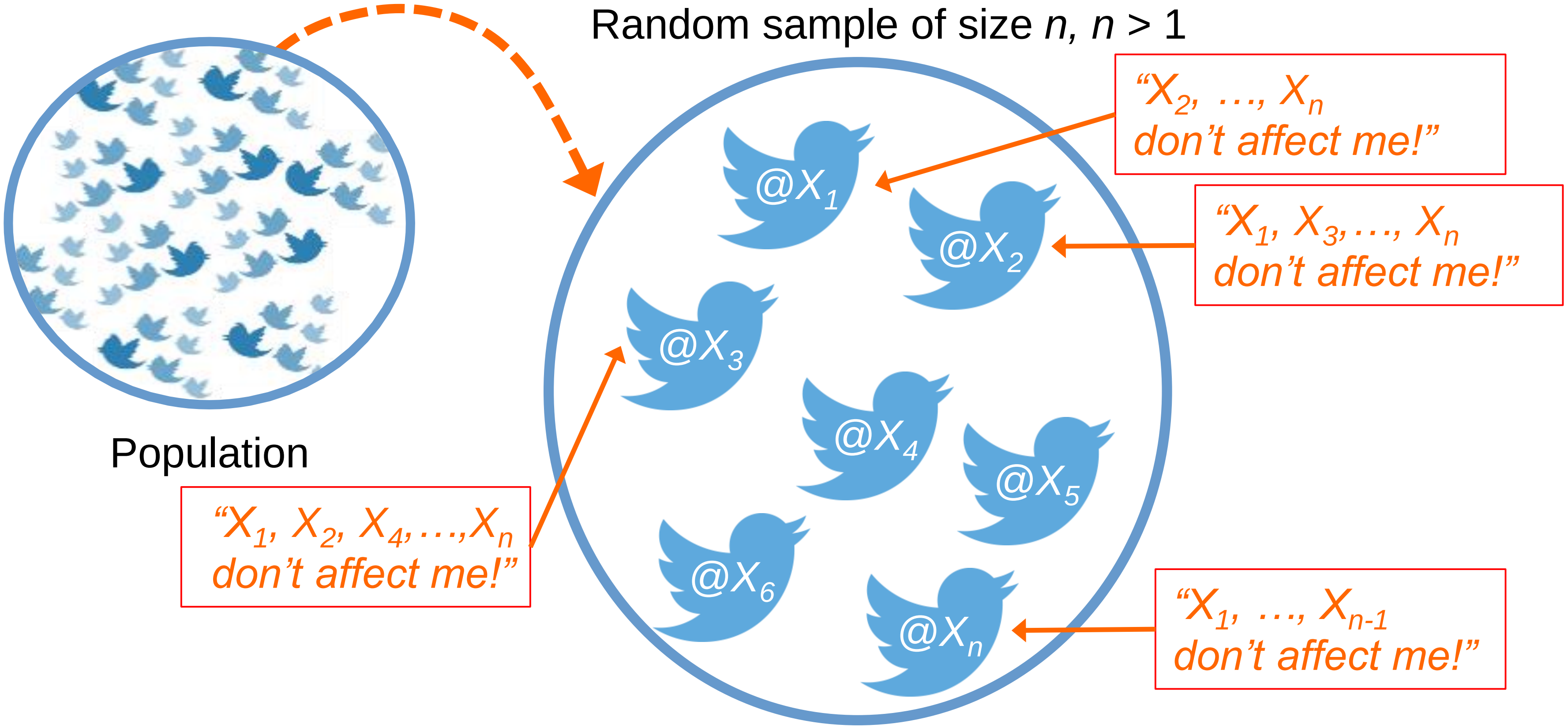


Random sample of size n, $n > 1$

Population

$X_1, X_2, \ldots X_n$
is a sequence of
observations of
the **same** rv

Random variable, X = # of tweets per day



Random sample of size n, $n > 1$

Population

“X_2, …, X_n don’t affect me!”

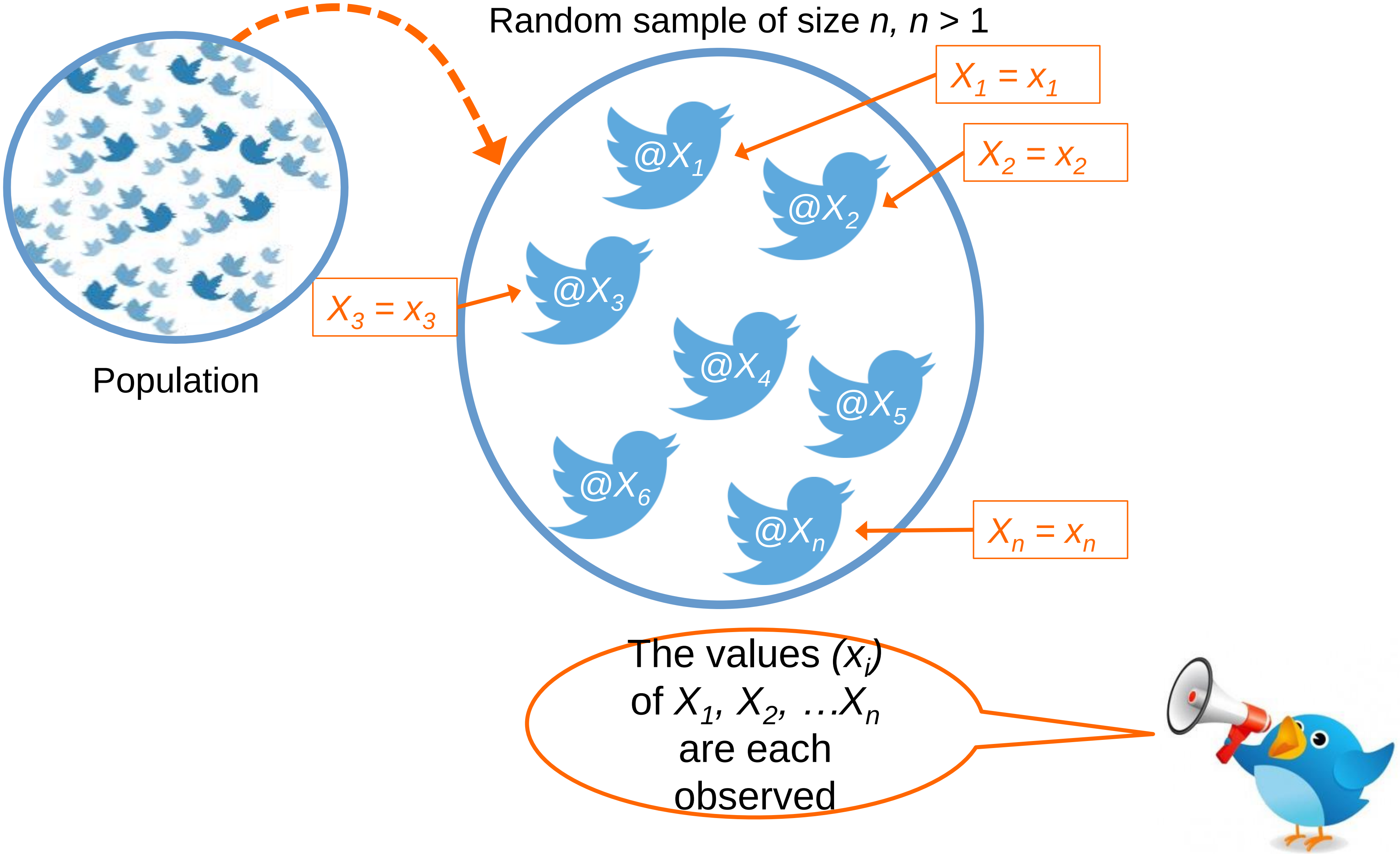
“X_1, X_3, …, X_n don’t affect me!”

“X_1, X_2, X_4, …, X_n don’t affect me!”

“X_1, …, X_{n-1} don’t affect me!”

$X_1, X_2, \ldots X_n$ are **mutually independent**

Random variable, X = # of tweets per day



Random sample of size n, $n > 1$

$X_1 = x_1$

$X_2 = x_2$

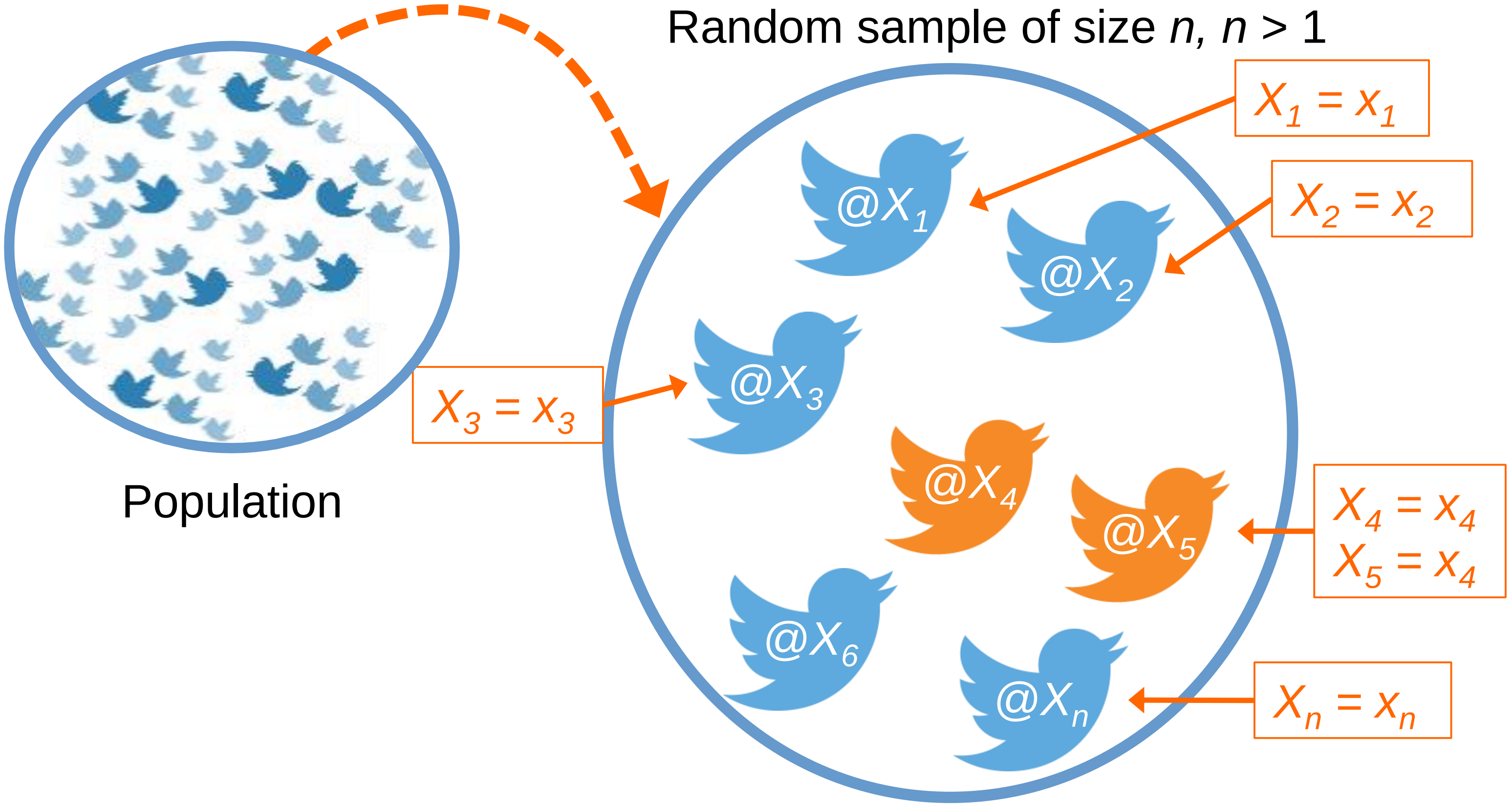
$X_3 = x_3$

$X_n = x_n$

Population

The values (x_i) of $X_1, X_2, \ldots X_n$ are each observed

Random variable, X = # of tweets per day



Random sample of size n, $n > 1$

Population

$X_1 = x_1$

$X_2 = x_2$

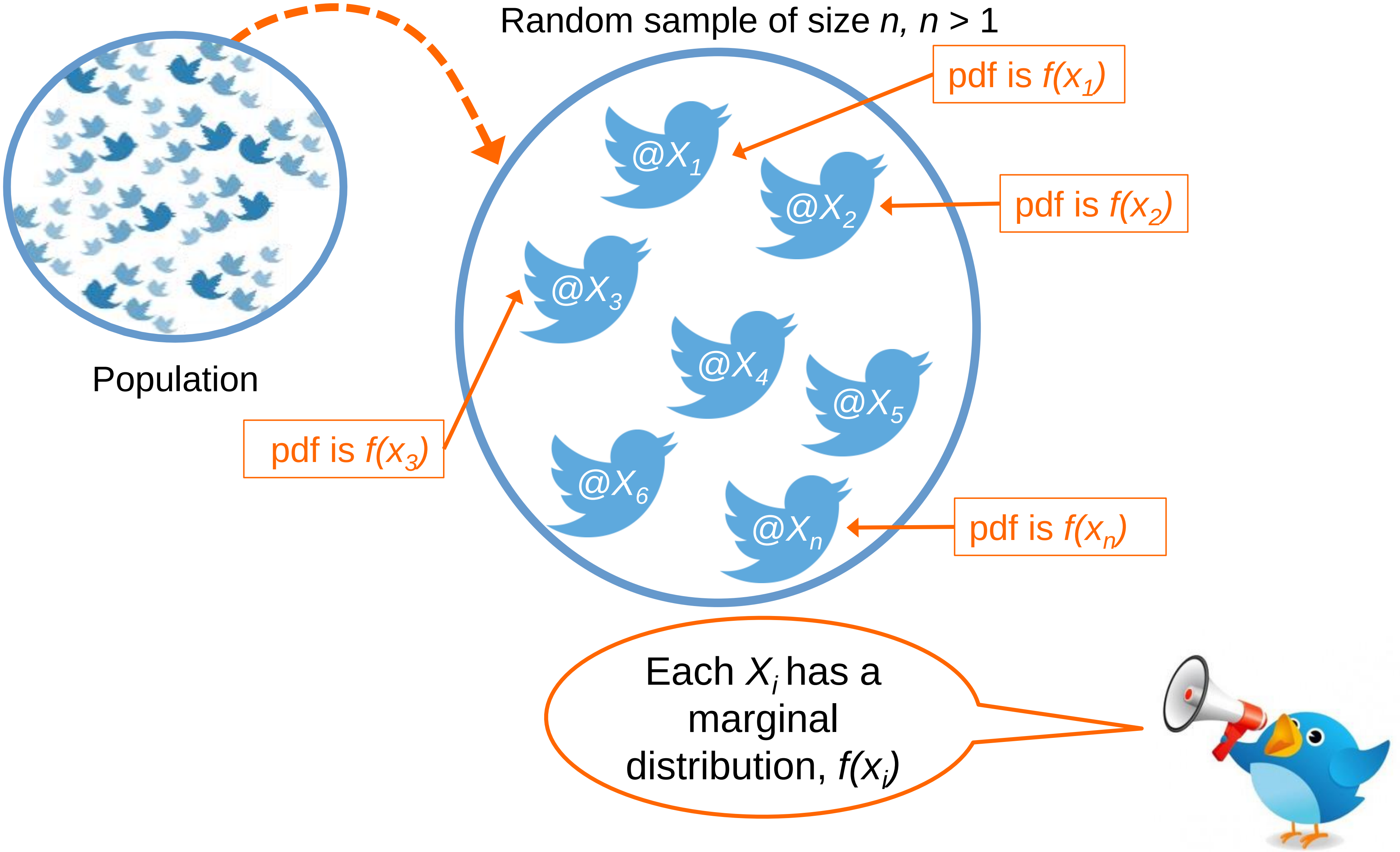
$X_3 = x_3$

$X_4 = x_4$
$X_5 = x_4$

$X_n = x_n$

If two values of x_i are the same…

Random variable, *X* = # of tweets per day



Random sample of size *n*, *n* > 1

Population

pdf is $f(x_1)$

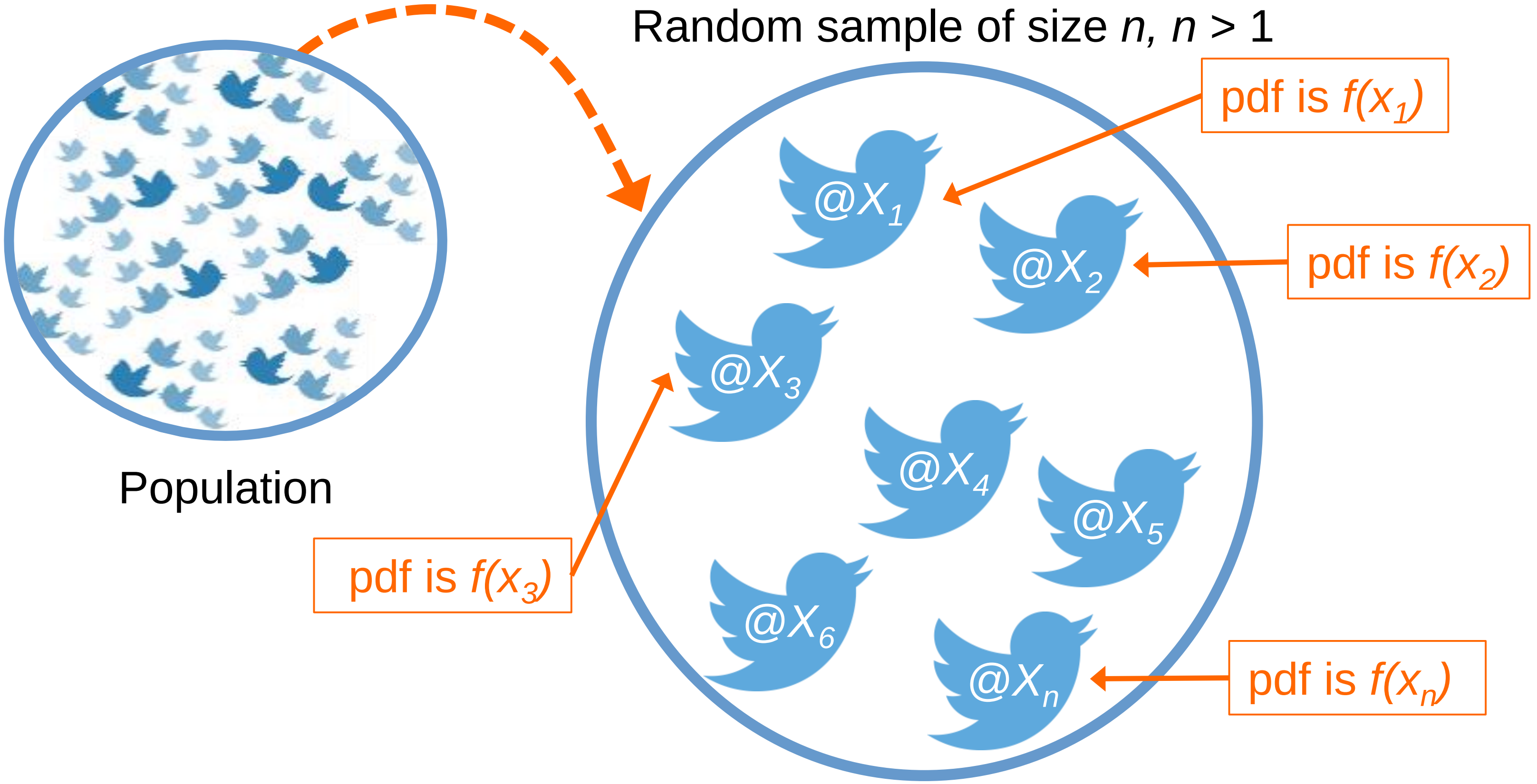
pdf is $f(x_2)$

pdf is $f(x_3)$

pdf is $f(x_n)$

Each X_i has a marginal distribution, $f(x_i)$

Random variable, X = # of tweets per day



Random sample of size n, $n > 1$

Population

pdf is $f(x_1)$

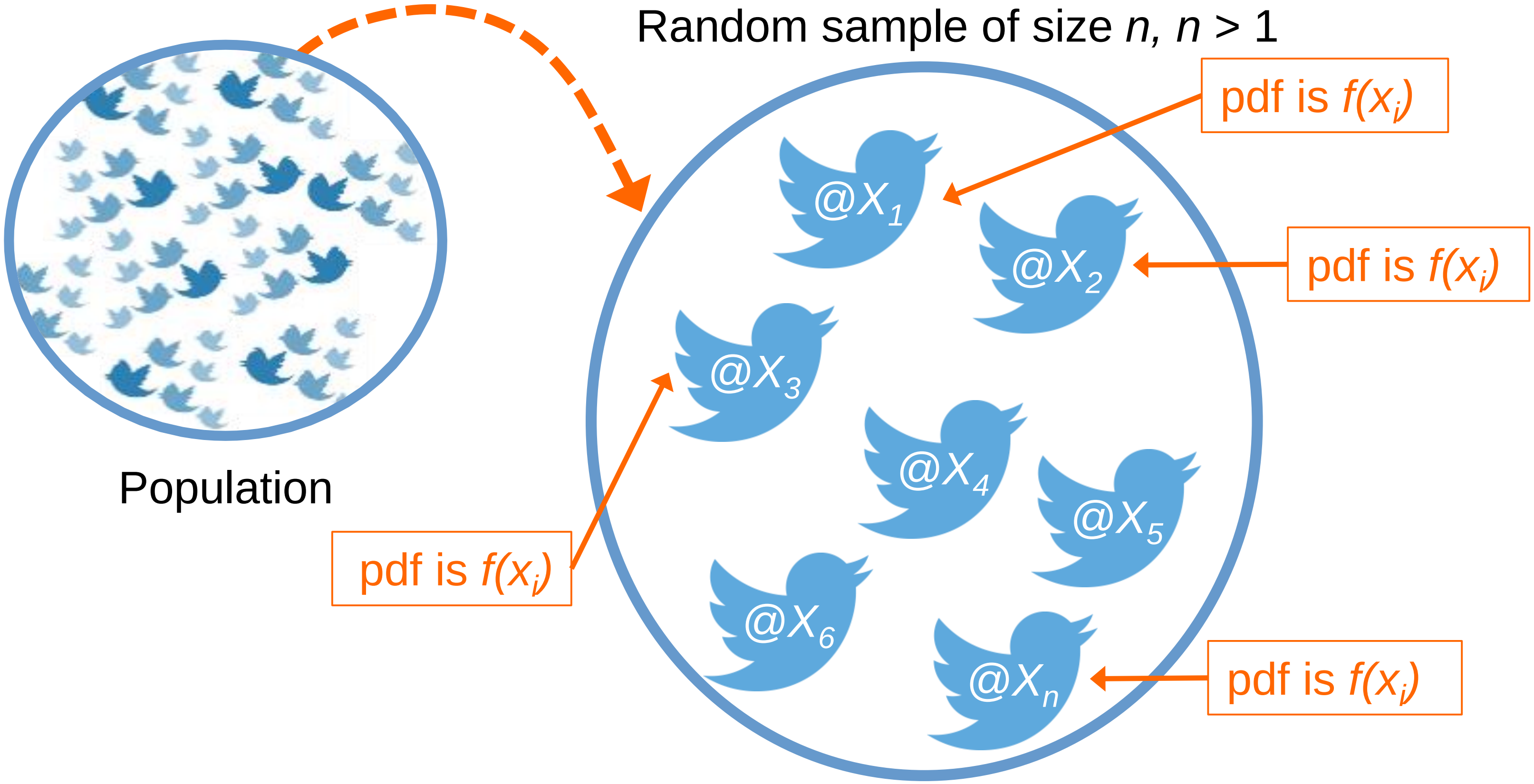
pdf is $f(x_2)$

pdf is $f(x_3)$

pdf is $f(x_n)$

Hooray! Because each observation X_i is **independent**, we can just multiply each $f(x_i)$ get joint pdf!

Random variable, X = # of tweets per day



Random sample of size n, $n > 1$

Population

pdf is $f(x_i)$

pdf is $f(x_i)$

pdf is $f(x_i)$

pdf is $f(x_i)$

And because each observation X_i is **identically distributed**,

$f(x_1) = f(x_2) = \ldots = f(x_n) = f(x_i)$

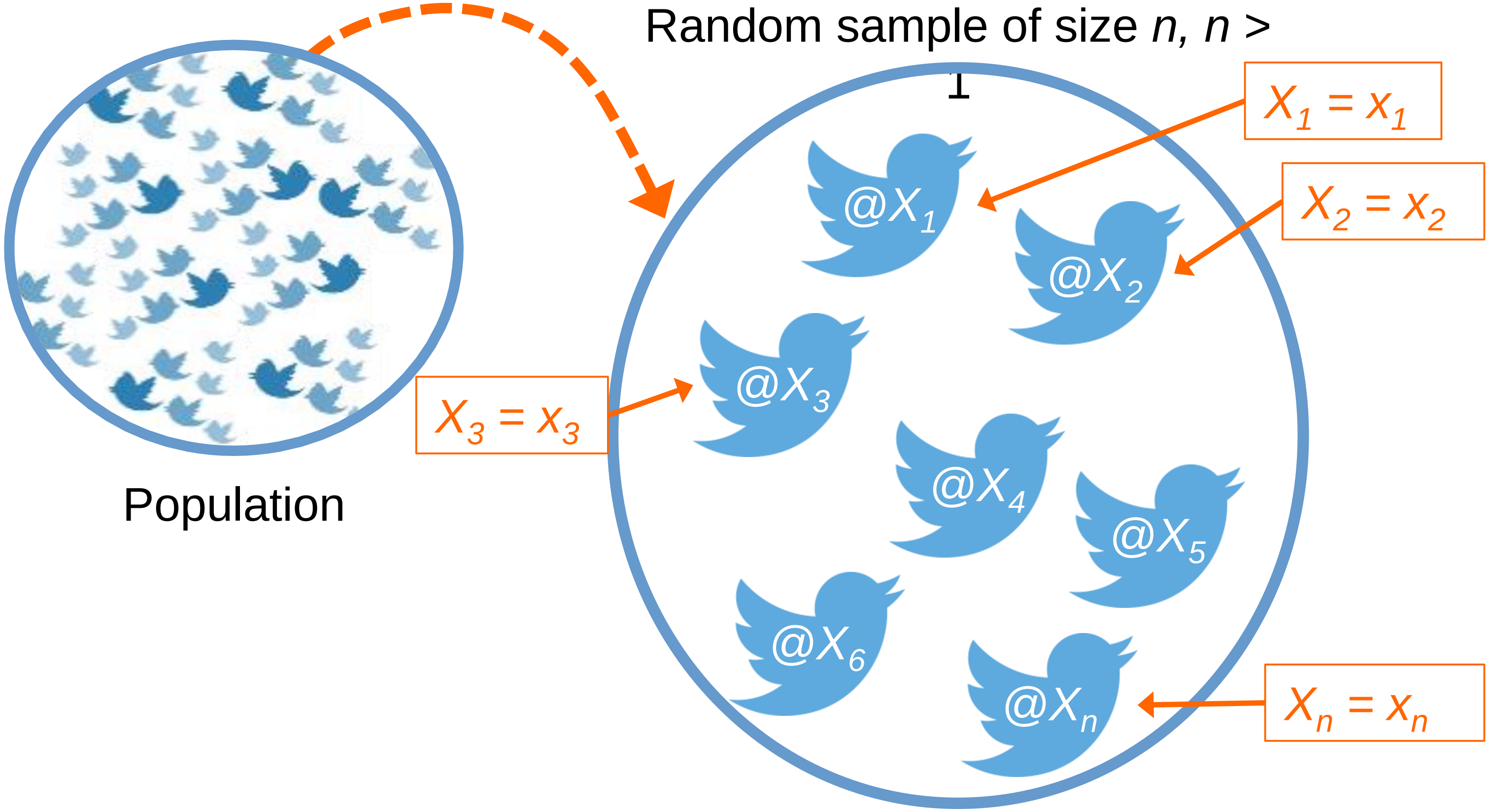
Random variable, *X* = # of tweets per day

$f(x_1, \dots, x_n) = f(x_1) \times f(x_2) \times \dots \times f(x_n) = \prod_{i=1}^{n} f(x_i)$

Since $X_1, X_2, \dots X_n$ are **mutually independent,** the probabilities $P(X_1), P(X_2), \dots P(X_n)$ can simply be multiplied to get the joint pdf of the sequence of observations of the rv of interest

Since $X_1, X_2, \dots X_n$ are **identically distributed**, all marginal densities $f(x_i)$ are the exact same function

Summarizing observed values of an rv



Random sample of size *n*, *n* > 1

Population

$X_1 = x_1$

$X_2 = x_2$

$X_3 = x_3$

$X_n = x_n$

After the sample X_1, X_2, …X_n is drawn, we usually want to create some summary of the x_i values

- Formally, let $X_1, X_2, \ldots X_n$ be a random sample of size n from a population, and let $T(x_1, x_2, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, X_2, \ldots X_n)$.

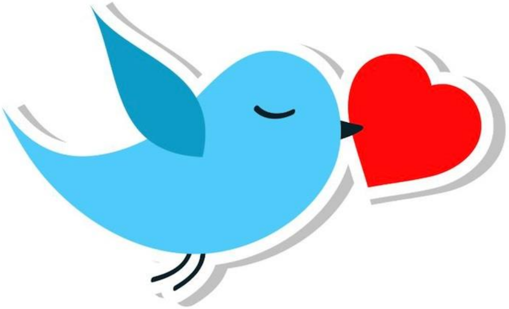
- Then the random variable or random vector $Y = T(x_1, x_2, \ldots, x_n)$ is called a **statistic**.

- The probability distribution of a statistic Y is called the **sampling distribution** of Y.

Statistics

- The definition of a statistic is fairly broad
  - The main restriction is that it cannot be a function of a population parameter
  - Think of a sample as a collection/sequence of rvs: any function of an rv is an rv itself

- 2 main reasons we love them:
  - Sometimes they are **estimators** for population parameters we care about
  - Sometimes they are **test statistics**, i.e. the basis for a hypothesis test



Example statistic: the sample mean

- Formally, $Y = T(x_1, x_2, \ldots, x_n)$ (this just says let's make a new variable, call it *Y*, and make it equal some function, call it *T()*, of the observed values of $X_1, X_2, \ldots X_n$)

- Let's make *Y* the arithmetic average of the values in our random sample. This is a common statistic, so we all tend to call it the same thing (instead of *Y*)…

$$Y = \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

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

It's a **random variable** too

Sample mean of *X* = average # of tweets per day

- *X* versus *x*: It is pretty standard at this point to use uppercase letters to denote that we are talking about a statistic, and to use lowercase letters when we are talking about observed value of a statistic.
- Let's say we observed 6 twitter users whose tweets totaled: *{X_1 = 5; X_2 = 3; X_3 = 0; X_4 = 10; X_5 = 3; X_6 = 2}*
- Each numerical value above is an x_i that we have now observed

$$
\begin{aligned}
\overline{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n} \\
&= \frac{1}{6} \times 23 = 3.833
\end{aligned}
$$

Still a **random variable**

Example statistic: the sample variance

- Formally, $Y = T(x_1, x_2, \ldots, x_n)$ (this just says let's make a new variable, call it *Y*, and make it equal some function, call it *T()*, of the observed values of $X_1, X_2, \ldots X_n$)

- Let's make *Y* the average squared deviations of each value from the mean in our random sample. This is another common statistic, so we all tend to call it the same thing (instead of *Y*)…

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

It's a **random variable** too

Example statistic: the sample variance

- Again, we observed 6 twitter users whose tweets totaled: *{X_1 = 5; X_2 = 3; X_3 = 0; X_4 = 10; X_5 = 3; X_6 = 2}*
- Each numerical value above is an x_i that we have now observed

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{5} \times 58.8333 = 11.766$$

Still a **random variable**

Expectation of a statistic

- Now we have an interesting problem- we have defined the formula for expectation (and the variance) for any rv (either discrete or continuous)

- How do we determine the expectation of a **statistic**, which is both an rv and a function of *other* rvs?

A helpful proof

- Let X_1, X_2, …X_n be a random sample from a population. Let $g(x)$ be a function such that $E[g(X_i)]$ and $Var[g(X_i)]$ exist. Then:

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = nE[g(X_i)]$$

- Since the X_is are identically distributed, the expectation of each individual $g(X_i)$ is the same for all i. We can arbitrarily pick i = 1.

$$nE[g(X_i)] = nE[g(X_1)]$$

For any given X_i, $E(X_i) = \mu$

Expectation of a statistic: the mean

- Let X_1, X_2, …X_n be a random sample from a population with mean μ and variance $\sigma^2 < \infty$.
- We know that the formula for the sample mean is:

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

- To get the expectation of the sample mean statistic, we just take the expectation on each side of this equation.

Expectation of a statistic: the sample mean

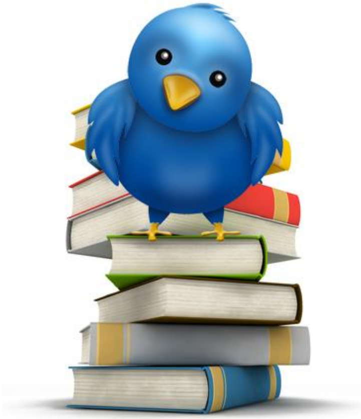
- Let X_1, X_2, …X_n be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then, taking the expectation of both sides of the equation:

$$
\begin{aligned}
E(\overline{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n} n E(X_i) = \mu
\end{aligned}
$$

1. For any given X_i, $E(X_i) = \mu$
2. n can take on any value > 0: it will never change this result!

$$E(\overline{X}) = \mu$$

“The expected value of the sample mean is the population mean”



Variance of a statistic: the sample mean

- So, we know that the expectation of our sample mean is the population mean
- What about the variance of the sample mean?
- We need another helpful proof…

$$Var\left(\sum_{i=1}^{n} g(X_i)\right) = n(Var[g(X_1)])$$

- Again, since the *X*$_i$s are identically distributed, the variance of each individual *g(X*$_i$*)* is the same for all *i*. We can arbitrarily pick *i* = 1.

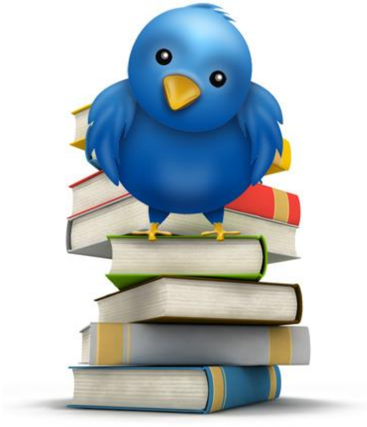
Variance of a statistic: the sample mean

- So, we know that the expectation of our sample mean is the population mean
- What about the variance of the sample mean?

$$
\begin{aligned}
Var(\overline{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} n Var(X_1) = \frac{\sigma^2}{n}
\end{aligned}
$$

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

“The variance of the sample mean is the population variance divided by the number of observations”

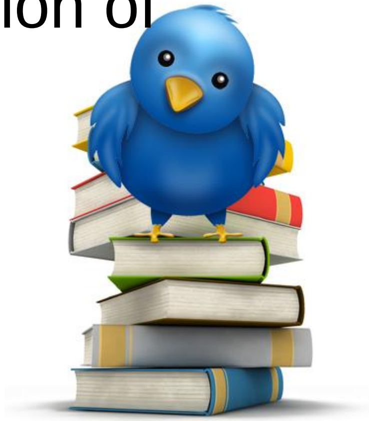


Different from sample variance!

- The sample variance, S^2, is also a statistic, and also has an expectation. Recall this is the formula for the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- To get its expectation, we again take the expectation of both sides of the equation


???

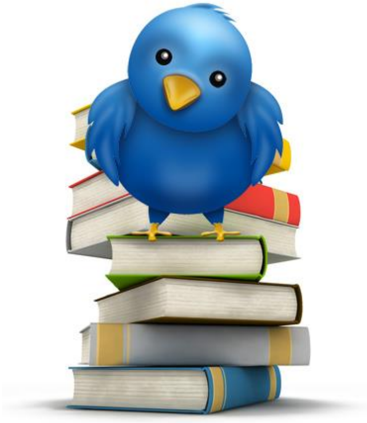
Expectation of a statistic: the sample variance

$$\begin{aligned} E(S^2) &= E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]\right) \\ &= \frac{1}{n-1}[nE(X_1^2) - nE(\overline{X}^2)] \\ &= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \sigma^2 \end{aligned}$$

- If S^2 were defined with n rather than (n – 1) in the denominator, then *E(S²)* would be biased and would not equal the population variance. We use (n – 1) in the denominator to create an unbiased estimator.

$$E(S^2) = \sigma^2$$

“The expectation of the sample variance is the population variance”



Expectation and variance of the sample mean

- Why are these statistics about other statistics helpful to know?

- How likely is it that the true mean of our sample is close to our population mean?

- This probability is determined by the sampling distribution

The sampling distribution of sample means

- Let’s perform the following random experiment:
  - Toss *n* fair dice;
  - Observe the number of dots (“pips”) showing for each die as x_i;
  - Calculate the mean number of dots across observed *n* values (note: here, *n* = 1)

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- Let’s start with *n* = 1 die

Three distributions to keep in mind simultaneously

1. The distribution of *X* in the population

2. The distribution of *x* in the particular sample

3. The sampling distribution of sample means across all possible samples

The population distribution

- Here the population is infinite ($n \rightarrow \infty$) and X has this probability distribution

x_i	p_i
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
sum	1

Mean of the population distribution

x_i	p_i
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
sum	1

What is the mean of this distribution?

Mean of the population distribution

x_i	p_i	$E(x_i)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6
sum	**1**	**3.5**

What is the mean of this distribution?

$$E(X) = \mu_X = \sum x_i \times p_i = 3.5$$

Variance of population distribution

x_i	p_i	$E(x_i)$	$[x_i – E(X)]^2$	× p_i
1	1/6	1/6	6.25	1.042
2	1/6	2/6	2.25	0.375
3	1/6	3/6	0.25	0.042
4	1/6	4/6	0.25	0.042
5	1/6	5/6	2.25	0.375
6	1/6	6/6	6.25	1.042
sum	**1**	**3.5**	**not yet!**	**2.917**

What is the variance/standard deviation of this distribution?

$$Var(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 p(x_i)$$

$$= 2.917$$

$$sd(X) = \sqrt{2.917} = 1.708$$

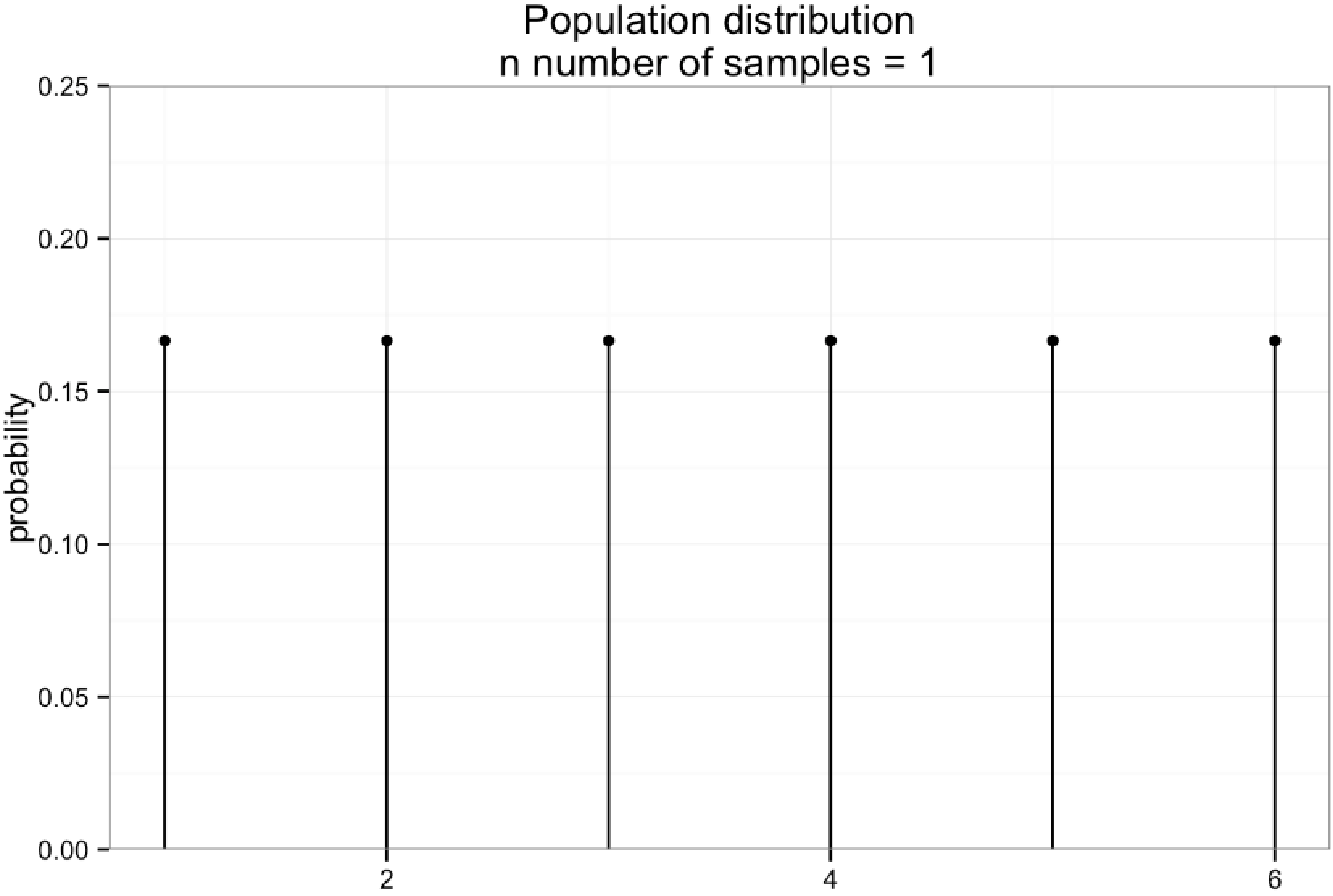
n = 1

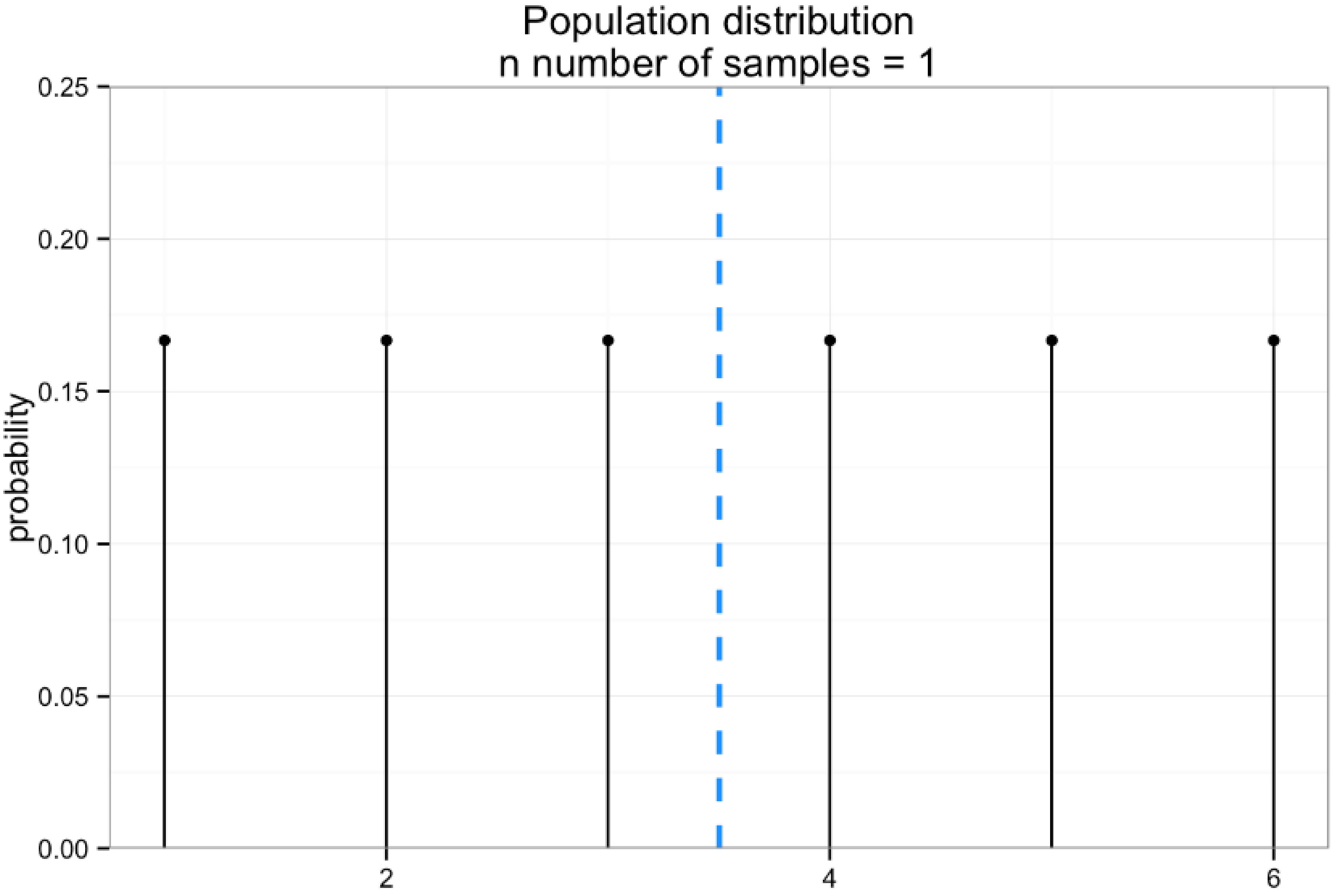
x_i	p_i
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
sum	1

$$E(X) = \mu_X = 3.5$$

$$Var(X) = \sigma_X = 2.917$$

We can also consider this as the sampling distribution of means from samples of size
n = 1
(where the samples are independent draws from an infinite population)


Population distribution
n number of samples = 1
probability
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6


Population distribution
n number of samples = 1
probability
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6

- Now let’s add a second die to our random experiment. Our random variable, X, is the mean number of pips across both die. What is the population mean for this sample?

(a) 2
(b) 3.5
(c) 2.5
(d) I have no idea!

Step 1: what is the sample space? How can we count the number of elements in it?

- $n = 2$
- $6 \times 6 = 36$ possible combinations

```
omega <- expand.grid(rep(list(1:6), 2))
omega <- omega %>%
  data.frame()
names(omega) <- c("die1", "die2") #just renaming my 2 variables
```

```
> omega
   die1 die2
1     1    1
2     2    1
3     3    1
4     4    1
5     5    1
6     6    1
7     1    2
8     2    2
9     3    2
10    4    2
11    5    2
12    6    2
13    1    3
14    2    3
15    3    3
16    4    3
17    5    3
18    6    3
19    1    4
20    2    4
21    3    4
22    4    4
23    5    4
24    6    4
25    1    5
26    2    5
27    3    5
28    4    5
29    5    5
30    6    5
31    1    6
32    2    6
33    3    6
34    4    6
35    5    6
36    6    6
```

Step 2: what are the sample means of all possible samples in sample space?

```
omega <- omega %>%
  mutate(xbar_i = (die1 + die2)/2)
```

```
> omega
   die1 die2 xbar_i
1     1    1    1.0
2     2    1    1.5
3     3    1    2.0
4     4    1    2.5
5     5    1    3.0
6     6    1    3.5
7     1    2    1.5
8     2    2    2.0
9     3    2    2.5
10    4    2    3.0
11    5    2    3.5
12    6    2    4.0
13    1    3    2.0
14    2    3    2.5
15    3    3    3.0
16    4    3    3.5
17    5    3    4.0
18    6    3    4.5
19    1    4    2.5
20    2    4    3.0
21    3    4    3.5
22    4    4    4.0
23    5    4    4.5
24    6    4    5.0
25    1    5    3.0
26    2    5    3.5
27    3    5    4.0
28    4    5    4.5
29    5    5    5.0
30    6    5    5.5
31    1    6    3.5
32    2    6    4.0
33    3    6    4.5
34    4    6    5.0
35    5    6    5.5
36    6    6    6.0
```

Step 3a: how many unique values of xbar?

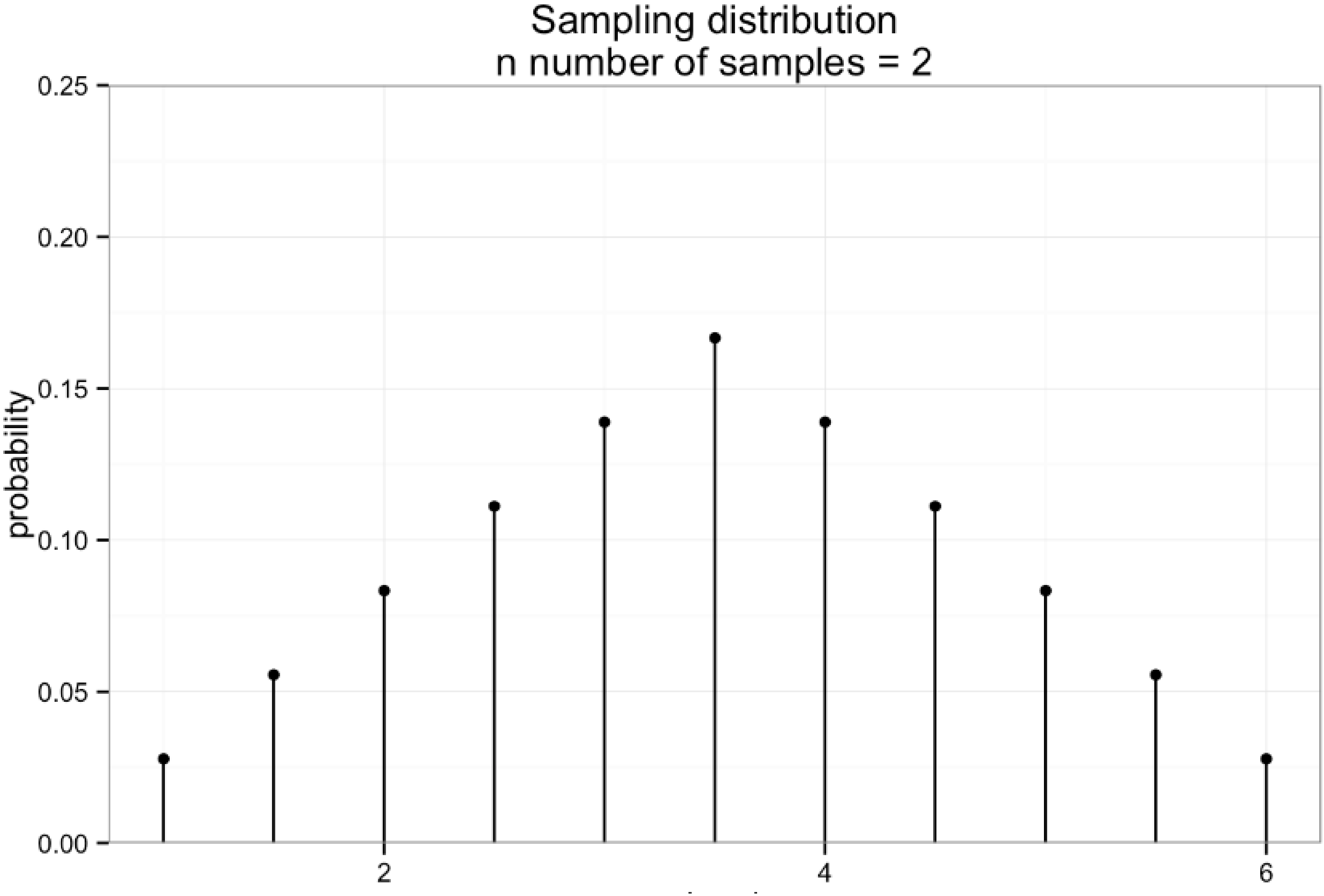
```
> omega %>%
    select(xbar_i) %>%
    distinct()

   xbar_i
1     1.0
2     1.5
3     2.0
4     2.5
5     3.0
6     3.5
7     4.0
8     4.5
9     5.0
10    5.5
11    6.0
```

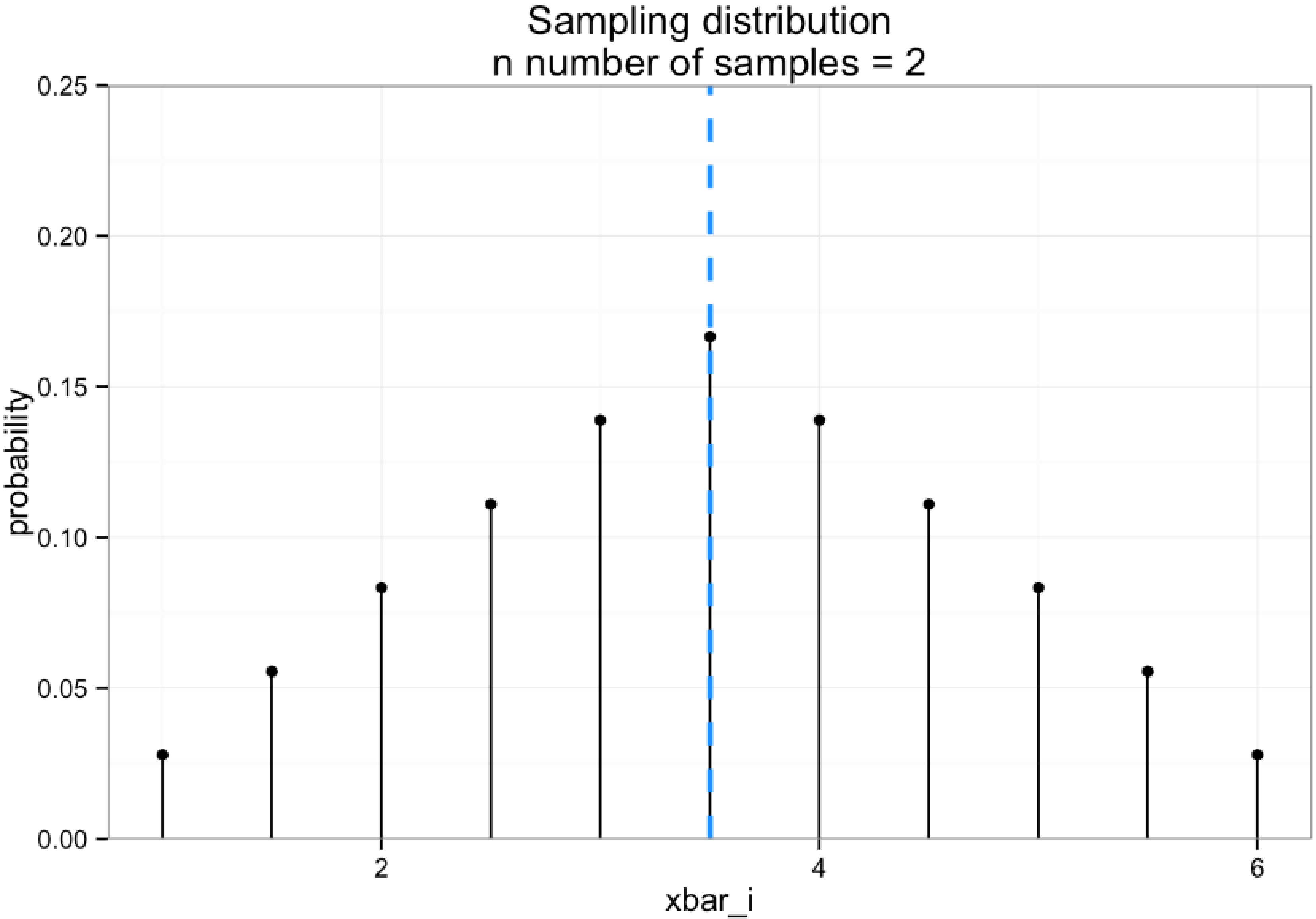
Step 3b: what is the probability of each?

```
> xbar_prob <- omega %>%
    group_by(xbar_i) %>%
    summarise(count = n(), p_i = count/36) %>%
    arrange(xbar_i)

> xbar_prob
Source: local data frame [11 x 3]

   xbar_i count        p_i
1     1.0     1 0.02777778
2     1.5     2 0.05555556
3     2.0     3 0.08333333
4     2.5     4 0.11111111
5     3.0     5 0.13888889
6     3.5     6 0.16666667
7     4.0     5 0.13888889
8     4.5     4 0.11111111
9     5.0     3 0.08333333
10    5.5     2 0.05555556
11    6.0     1 0.02777778
```

What is the sum here?

```
> xbar_prob %>%
   summarise(sum_of_ps =
sum(p_i))
Source: local data frame [1 x 1]

  sum_of_ps
1  1
```


Sampling distribution
n number of samples = 2
probability
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6


Sampling distribution
n number of samples = 2
probability
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6
xbar_i

Step 4: what is the expectation?

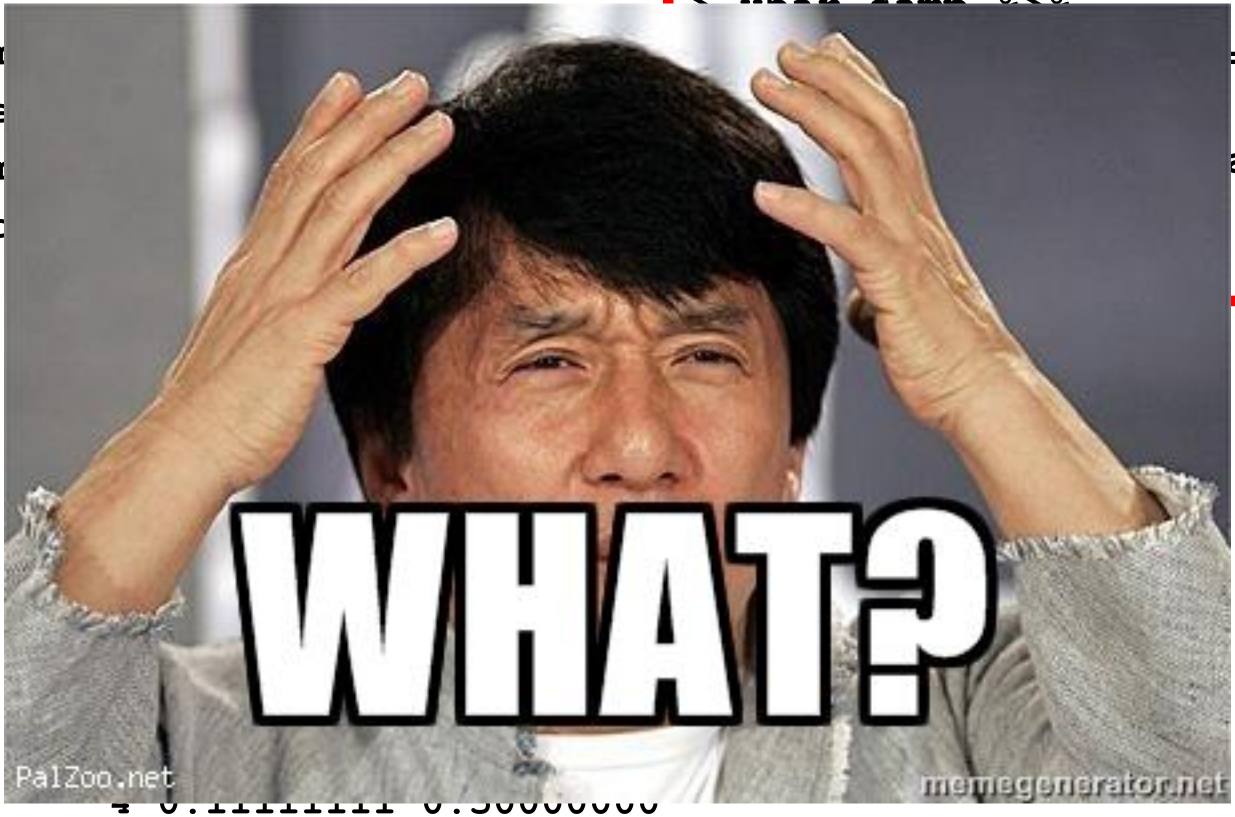
```
> xbar_samp <- xbar_prob %>%
    mutate(e_x_i = xbar_i * p_i)
> xbar_samp
Source: local data frame [11 x 4]

   xbar_i count        p_i      e_x_i
1     1.0     1 0.02777778 0.02777778
2     1.5     2 0.05555556 0.08333333
3     2.0     3 0.08333333 0.16666667
4     2.5     4 0.11111111 0.27777778
5     3.0     5 0.13888889 0.41666667
6     3.5     6 0.16666667 0.58333333
7     4.0     5 0.13888889 0.55555556
8     4.5     4 0.11111111 0.50000000
9     5.0     3 0.08333333 0.41666667
10    5.5     2 0.05555556 0.30555556
11    6.0     1 0.02777778 0.16666667
```

```
> xbar_samp %>%
+   summarise(e_xbar =
sum(e_x_i))
Source: local data frame [1 x 1]

 e_xbar
1    3.5
```

Step 4: what is the expectation?

```
> xbar_sam
    mutate
> xbar_sam
Source: lo


   xbar_i
1       1.0
2       1.5
3       2.0
4       2.5
5       3.0
6       3.5
7       4.0
8       4.5      4 0.11111111 0.50000000
9       5.0      3 0.08333333 0.41666667
10      5.5      2 0.05555556 0.30555556
11      6.0      1 0.02777778 0.16666667
```

```
> xbar_samp %>%
                                   =

                                ame [1 x 1]
```



What did we just show?

Distribution: population

Statistic: mean

$$E(X) = \mu_X = \sum x_i \times p_i = 3.5$$

Distribution: sampling distribution of the sample mean

Statistic: mean

$$E(\overline{x}) = \sum \overline{x_i} p_i = 3.5$$

$$E(X) = \mu_X = 3.5$$

$$E(\overline{x}) = \sum \overline{x_i} p_i = 3.5$$

What if we now roll 3 dice ($n = 3$)?

I think I got it... but just in case
tell me the whole thing again.
I wasn't listening.
DIGITAL MOVIE

To the in class exercise!

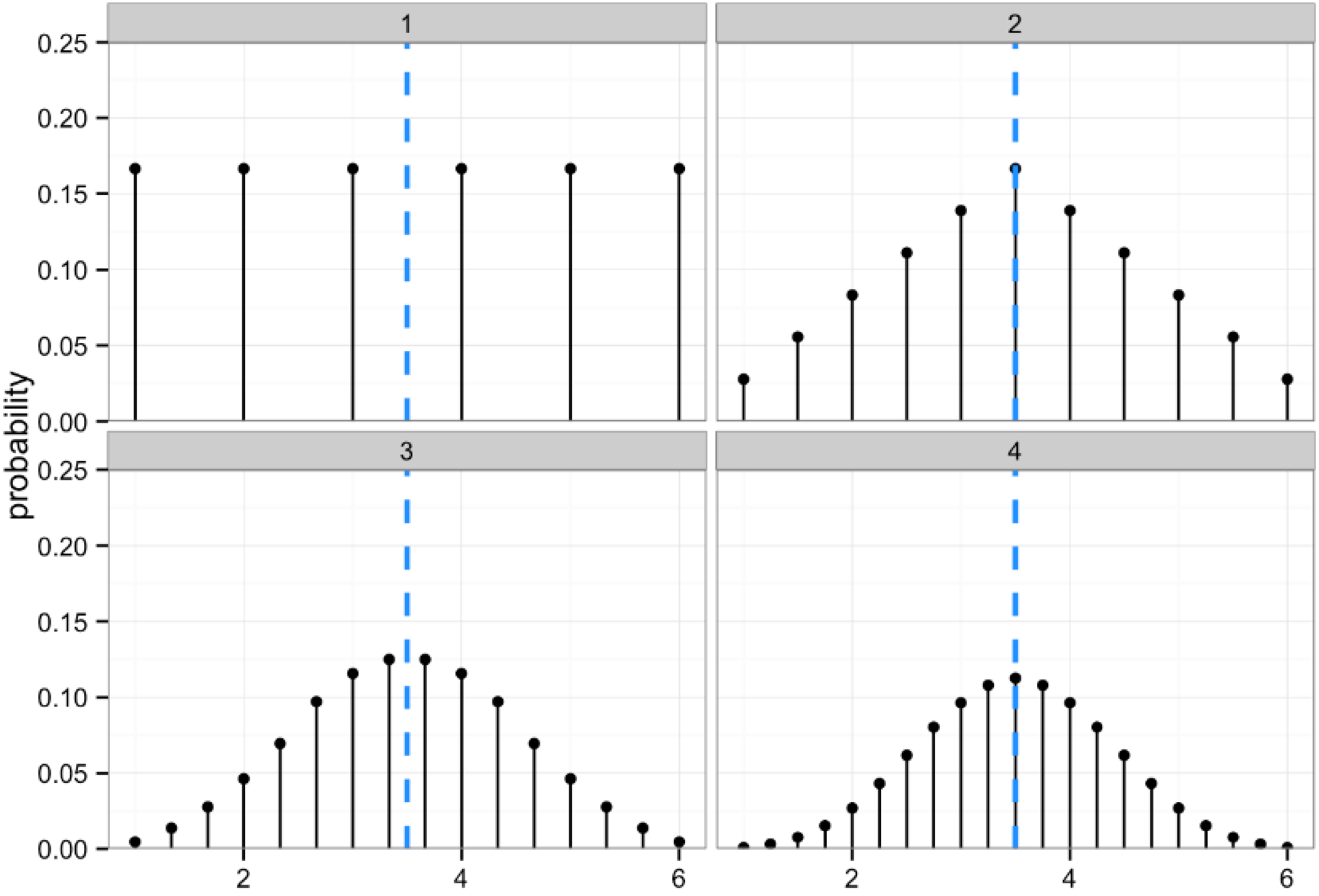
Did you get…

```
Source: local data frame [16 x 4]

      xbar_i count         p_i        e_x_i
1   1.000000     1 0.00462963 0.00462963
2   1.333333     3 0.01388889 0.01851852
3   1.666667     6 0.02777778 0.04629630
4   2.000000    10 0.04629630 0.09259259
5   2.333333    15 0.06944444 0.16203704
6   2.666667    21 0.09722222 0.25925926
7   3.000000    25 0.11574074 0.34722222
8   3.333333    27 0.12500000 0.41666667
9   3.666667    27 0.12500000 0.45833333
10  4.000000    25 0.11574074 0.46296296
11  4.333333    21 0.09722222 0.42129630
12  4.666667    15 0.06944444 0.32407407
13  5.000000    10 0.04629630 0.23148148
14  5.333333     6 0.02777778 0.14814815
15  5.666667     3 0.01388889 0.07870370
16  6.000000     1 0.00462963 0.02777778
```

```
  e_xbar
1    3.5
```

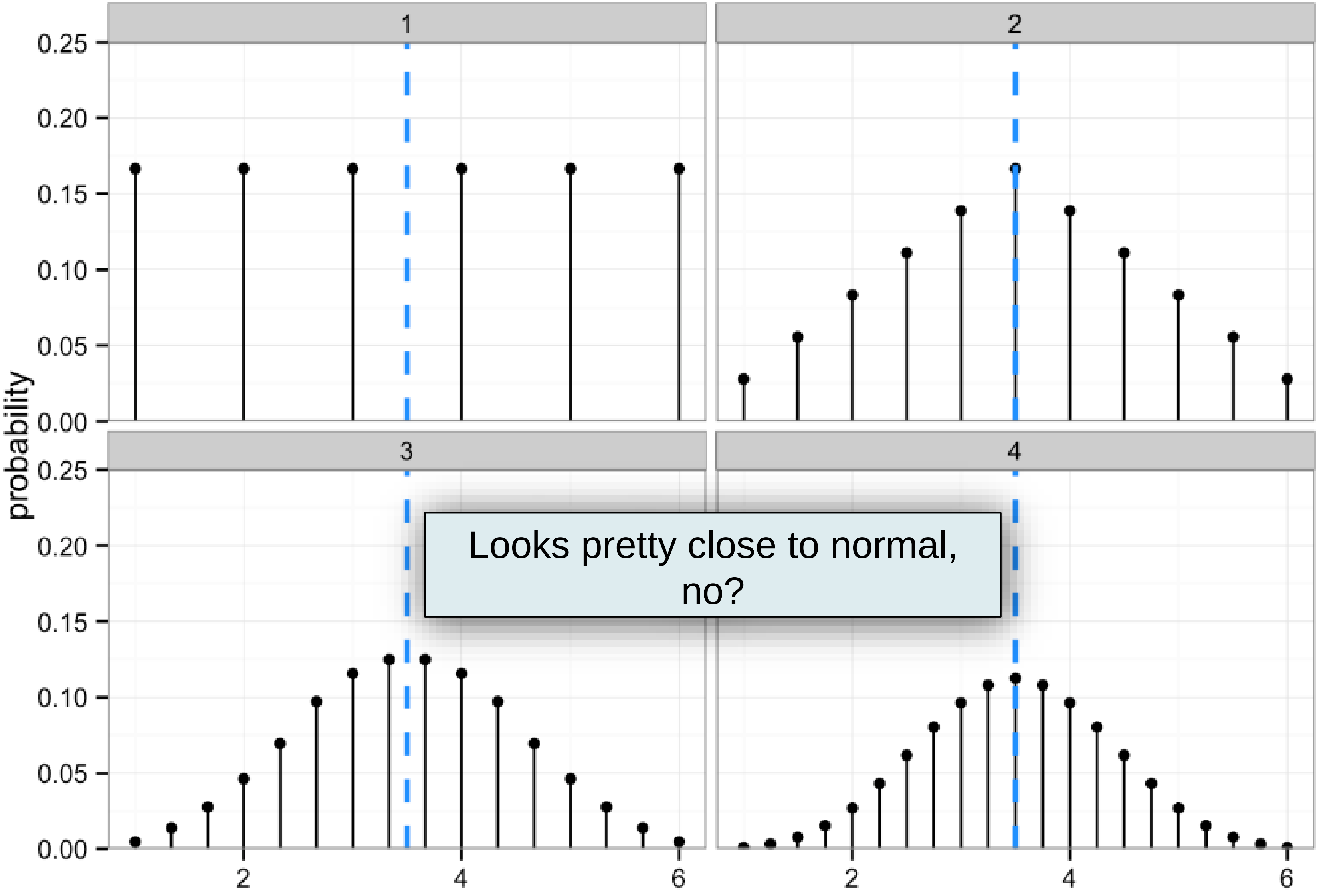
With 3 dice, this count column should sum to the total number of elementary events in this sample space: 6 × 6 × 6 = 216

What if we now increase the sample size to *n* = 4?


1
2
3
4
probability
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6

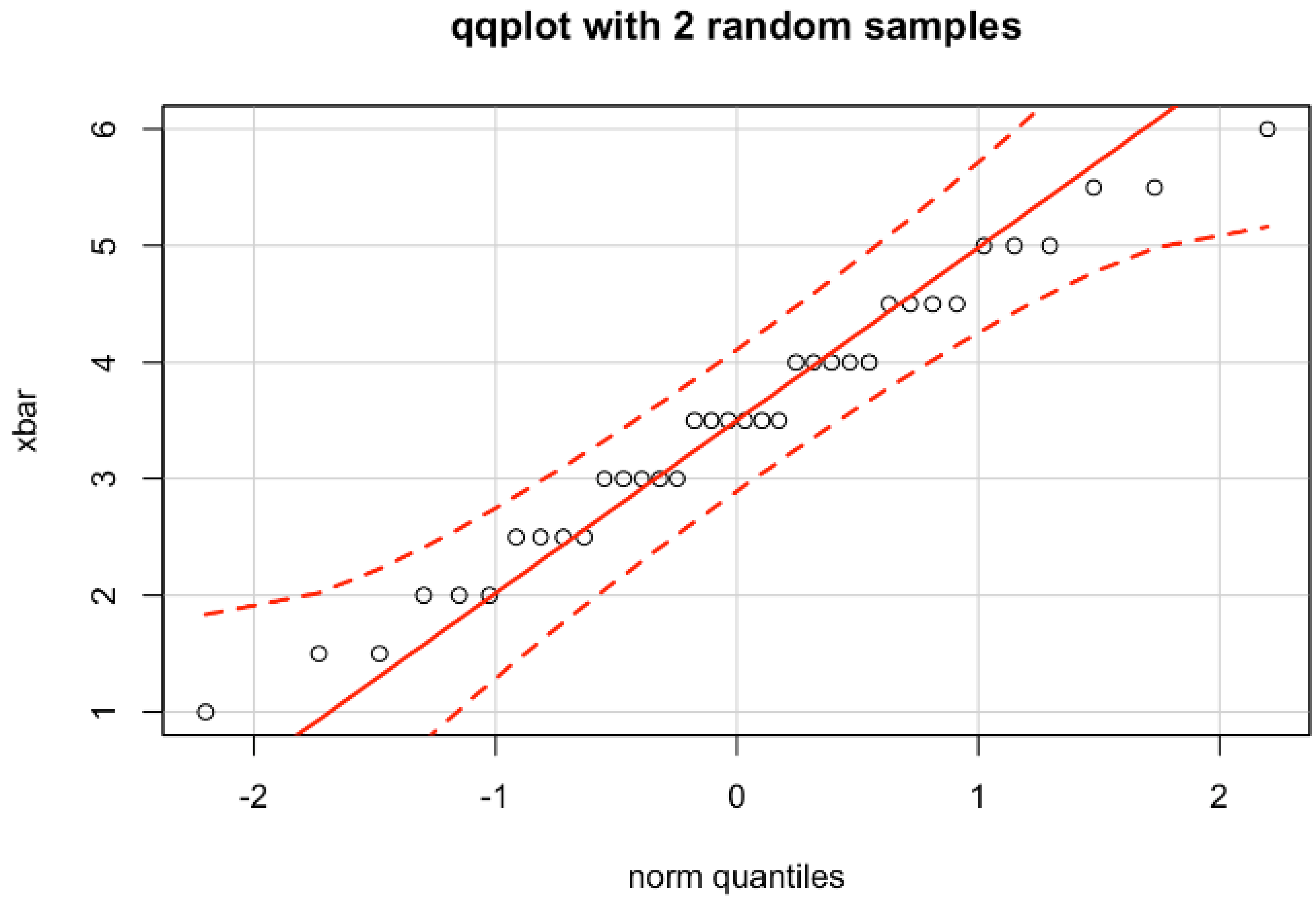
All of the below are a sequence of iid rvs

- The previous result applies whether:
  - I roll *n* dice all at once
    - *n* = number of dice
  - I roll a single die *n* times (if I have a specific “style” or “method” for rolling, could argue this is not iid…)
    - *n* = number of rolls
  - *n* people each roll a single die once
    - *n* = number of people
  - *n* fair dice are rolled; each by a different person
    - *n* = number of each person/die combination

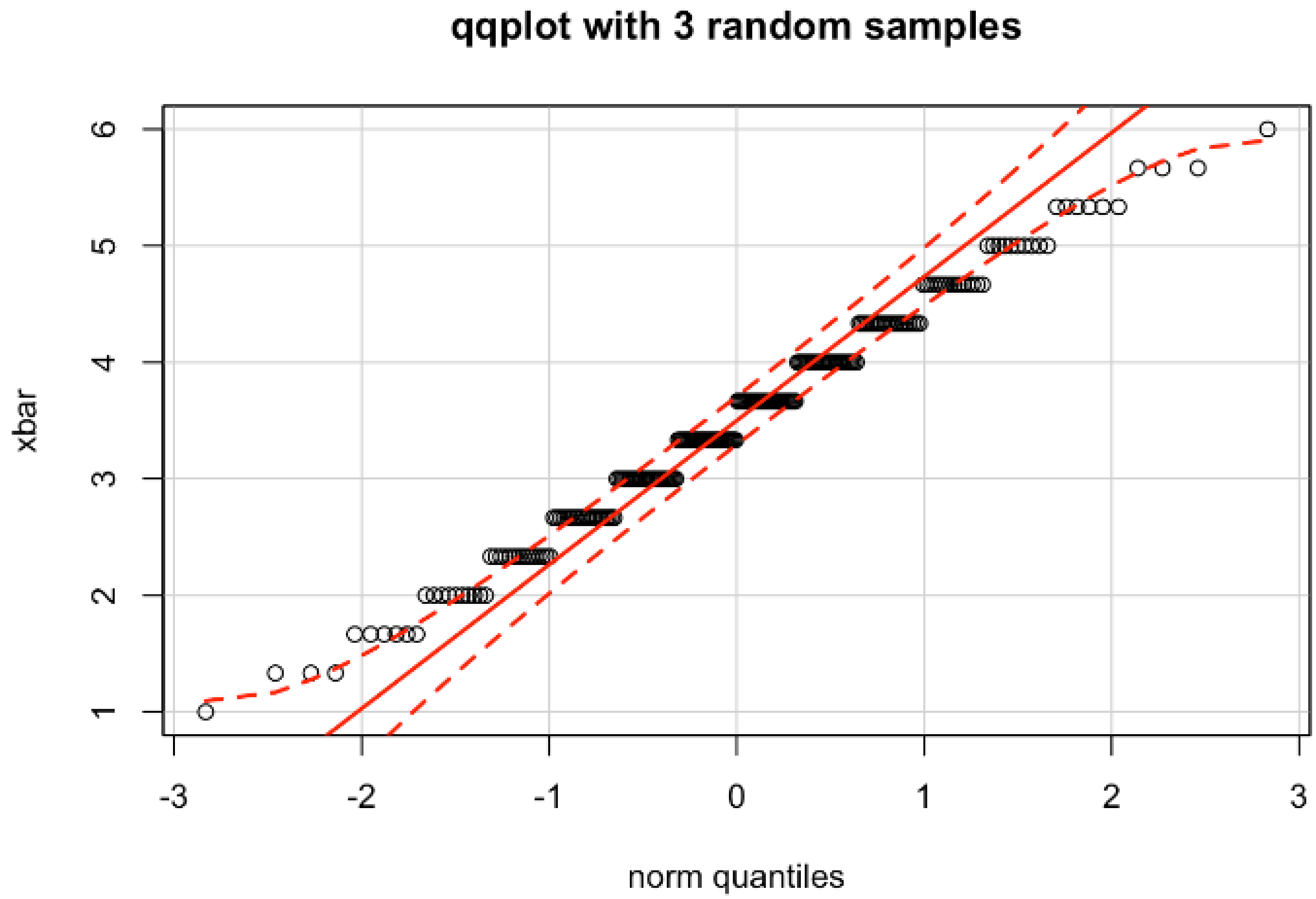


Looks pretty close to normal, no?

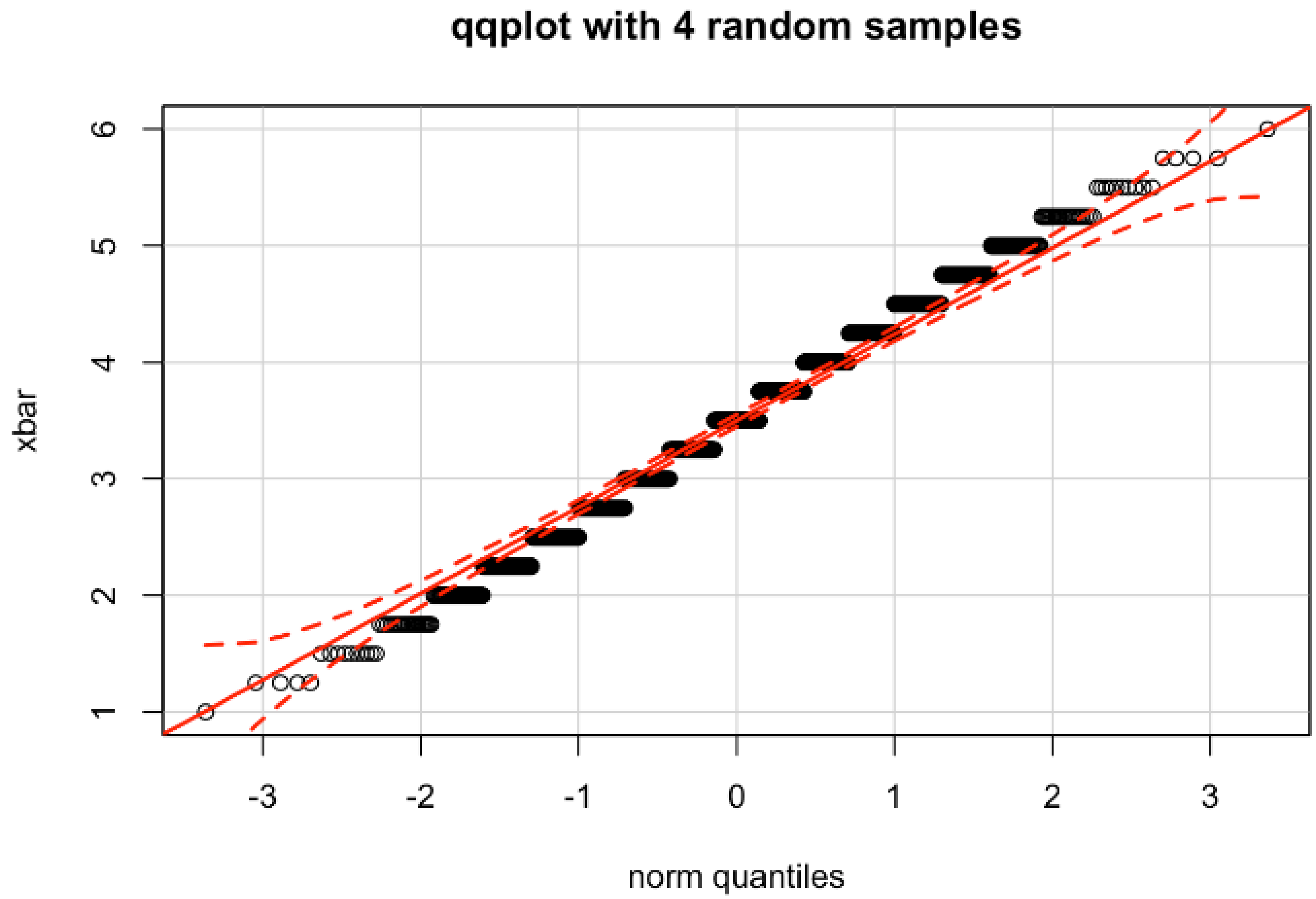
“Close to normal?”



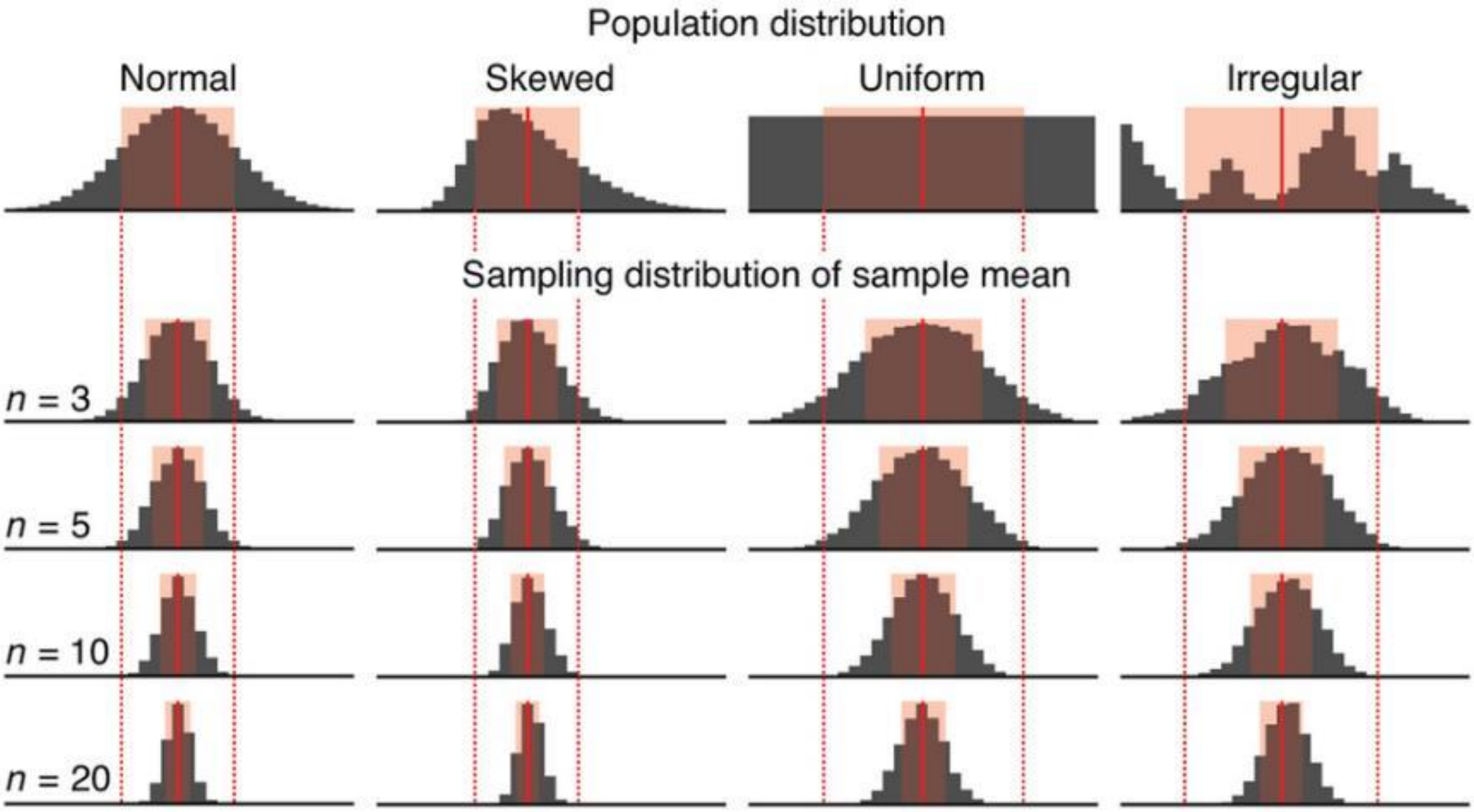
“Close to normal?”



“Close to normal?”



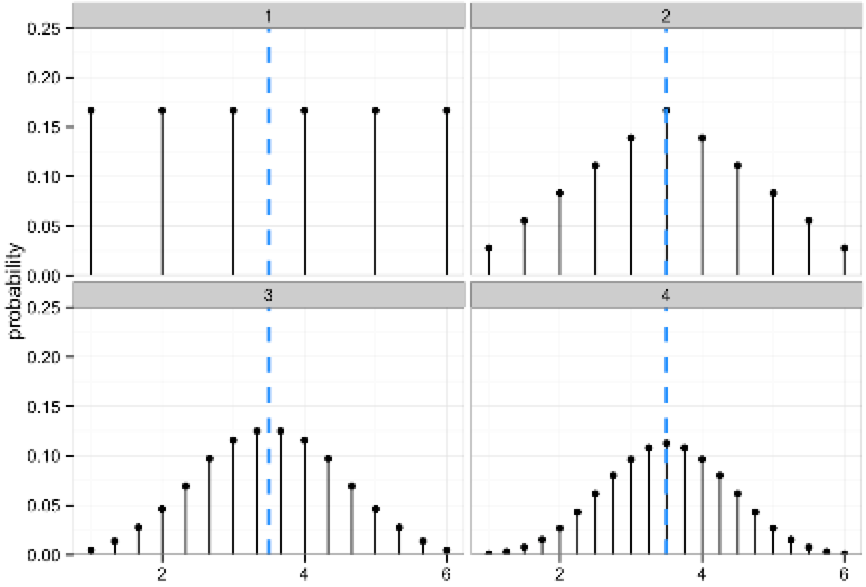
Sample means for 10,000 samples

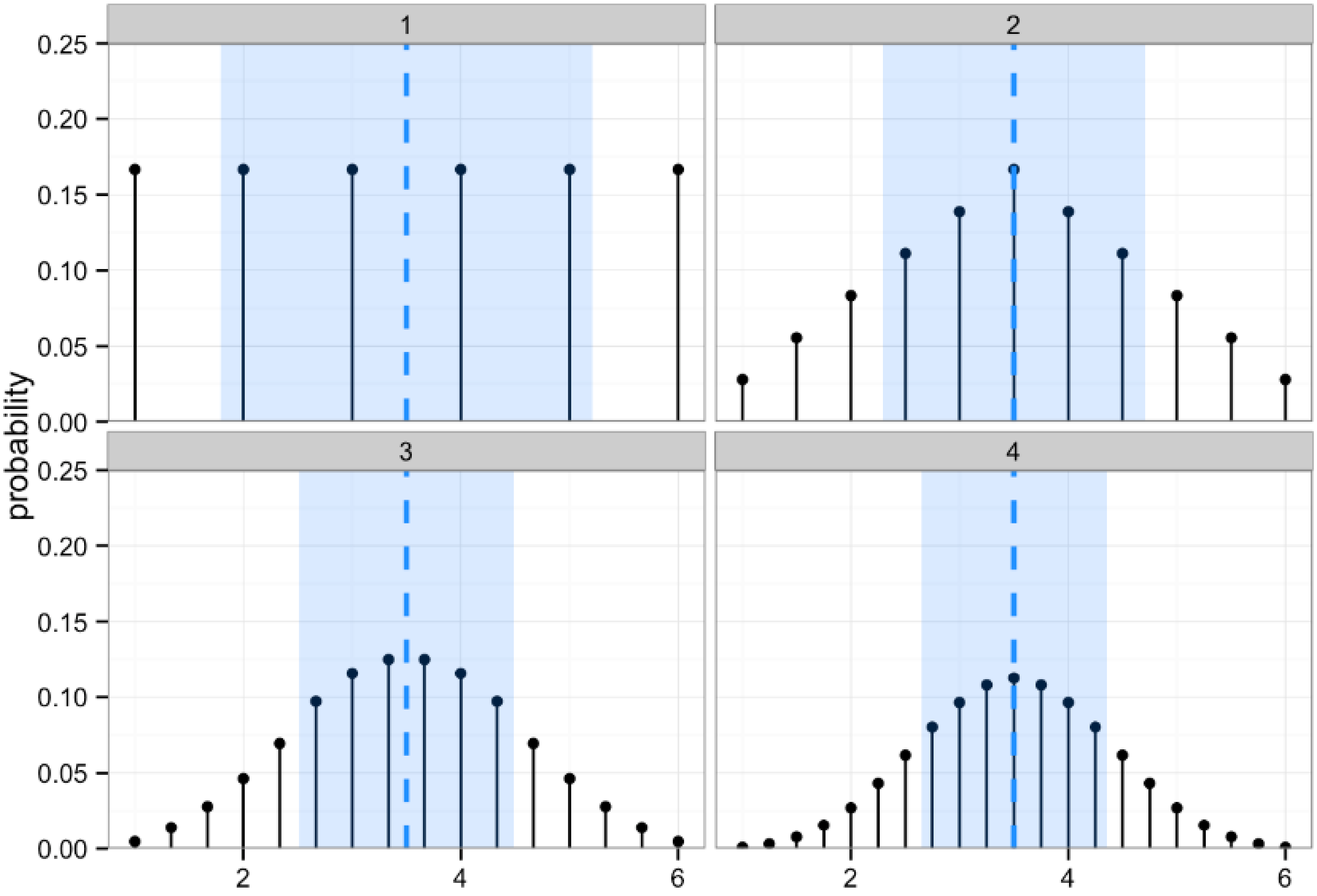


http://www.nature.com/nmeth/journal/v10/n9/full/nmeth.2613.html

What did change?

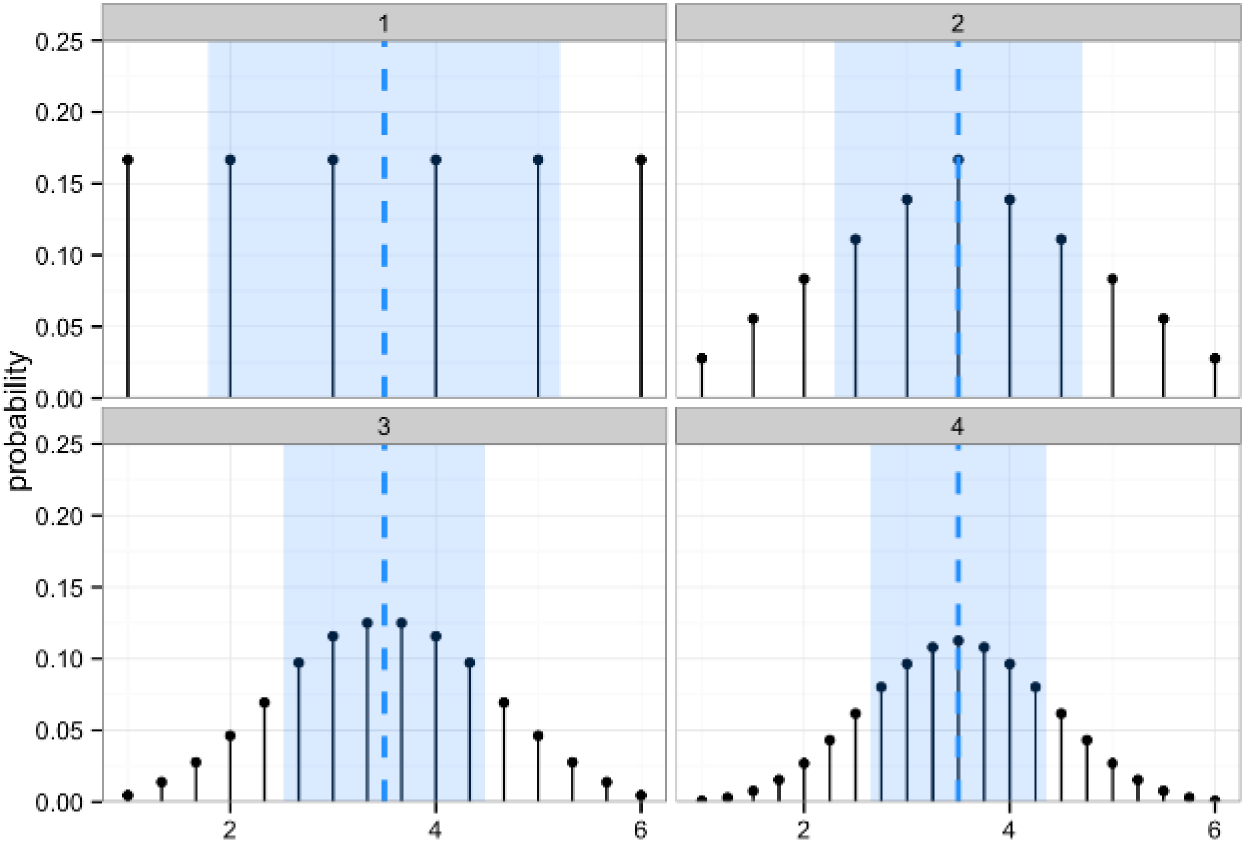
- Hopefully I’ve convinced you that no matter how many random samples we take, the expectation of our sample mean is the population mean.
- But clearly something else changed every time we added another random sample/observation. What was it?




1
2
3
4
probability
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6

In other words…

- The sample mean is obviously a good guess at *E(X)* = μ, but how good is it?



Spread of the sampling distribution of means

# of dice	$\mu_{\overline{X}}$	$\sigma^2_{\overline{X}}$	$\sigma_{\overline{X}}$
1	3.5	2.917	1.708
2	3.5	1.458	1.207
3	3.5	0.972	0.986
4	3.5	0.729	0.854

$$\sigma_{\overline{X},n=2} = \frac{1.708}{\sqrt{2}} = 1.207$$

$$\sigma_{\overline{X},n=3} = \frac{1.708}{\sqrt{3}} = 0.986$$

$$\sigma_{\overline{X},n=4} = \frac{1.708}{\sqrt{4}} = 0.854$$

If we know the spread for the population distribution, can we calculate the spread of the sampling distributions of the mean?

Standard error

- The standard error (SE) is the standard deviation of the sampling distribution of a statistic
- Here, the statistic is the mean, so the standard deviation of the sampling distribution of the mean is the standard error of the mean (often called SEM)

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

- But realize that this is just a specific standard error -- it’s a more general concept and they won’t always have this exact form

LAW OF LARGE NUMBERS

Weak LLN

- The average of a large, iid sample will be close to the true mean

$$\overline{X}_n \xrightarrow{P} \mu$$

Strong LLN

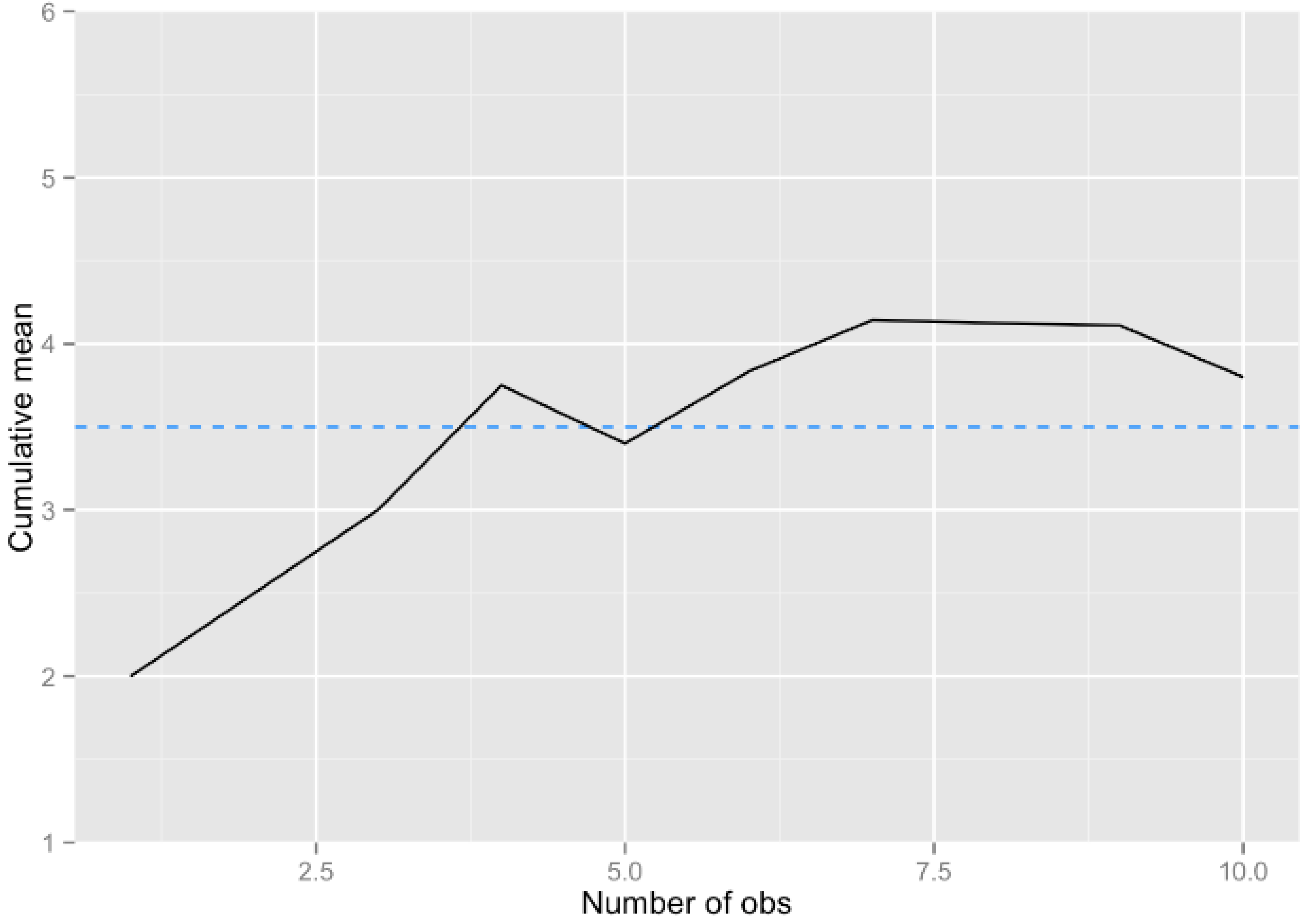
- *Let X_1, X_2, … be iid rvs with $E[X_i]=\mu < \infty$. Then:*

$$P\left(\lim_{N\to\infty} \frac{X_1 + X_2 + \cdots + X_N}{N} = \mu\right) = 1$$

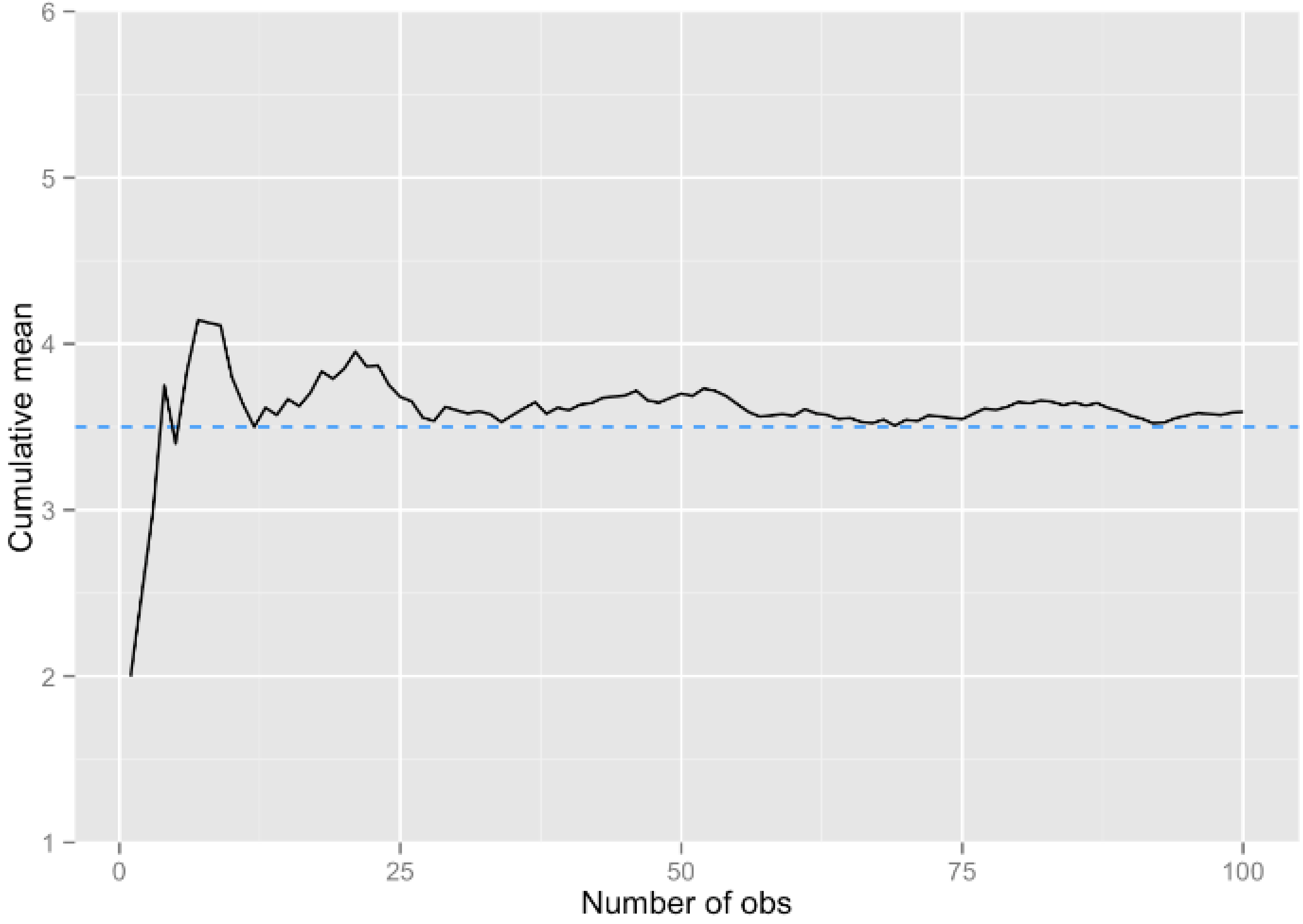
- Works by swamping, not compensation.
  - If you flip a fair coin and get 80 heads in the first 100 flips, then over the next 100 flips, the expected number of heads is 50—not 20*

*see gambler's fallacy; but caveat- Bayesian inference

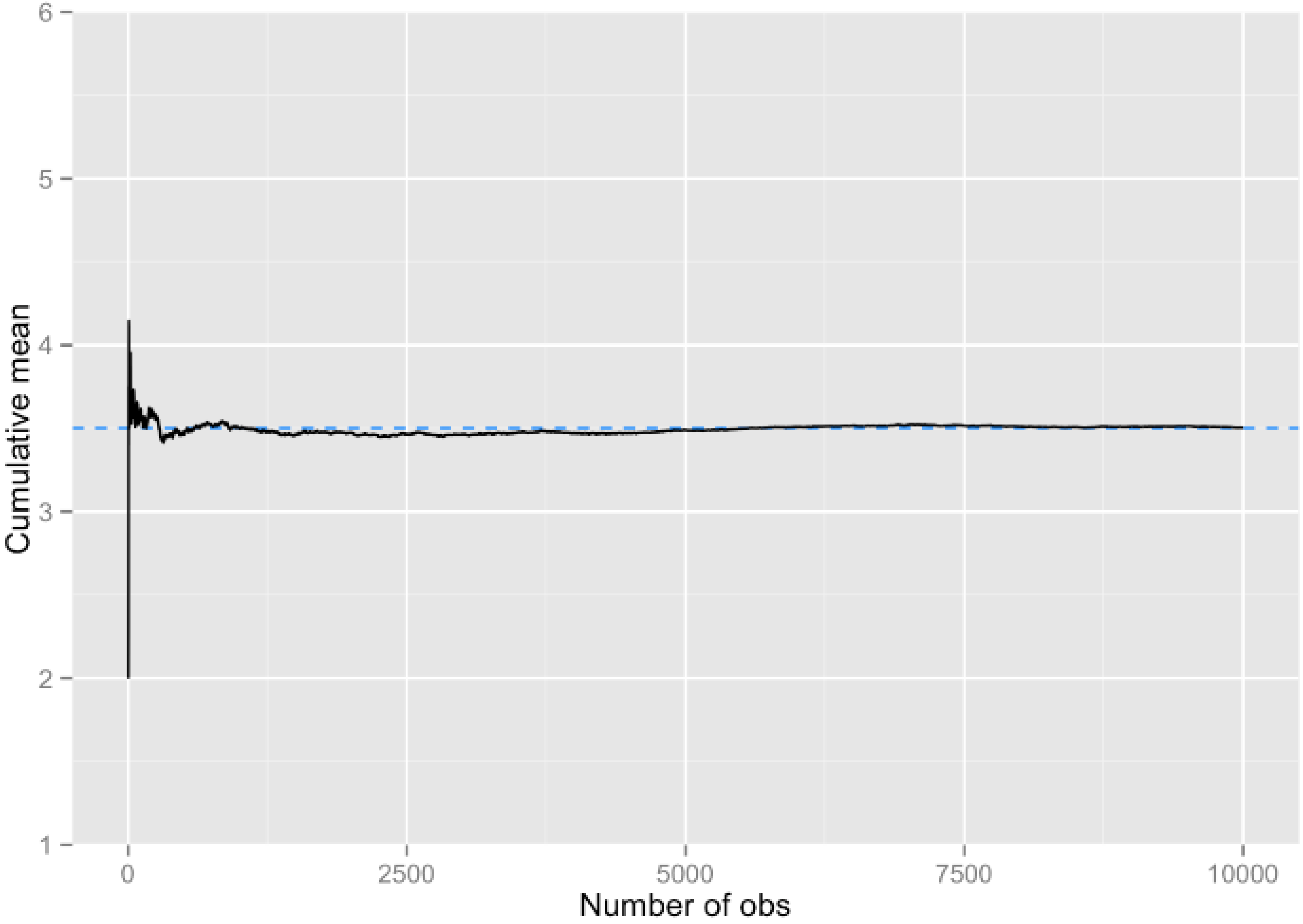
LLN in action: 10 dice to estimate the known population mean



LLN in action: 100 dice to estimate the known population mean



LLN in action: 10,000 dice to estimate the known population mean



What LLN does not say

- *“If I can just make my sample big enough, I won’t have to worry about error.”*
- There is no sample that is “big enough” in an unqualified sense
- Data still must be iid
- In stats, there are precious few fundamental constants, like there are in math (think: *e*) or physics (think: speed of light)
- Context and goals always matter

CENTRAL LIMIT THEOREM

CLT

- The sampling distribution for the mean of a large, iid sample will be approximately a normal distribution, regardless of the shape of the population distribution of *X*:

$$\overline{x} \to N(\mu, \sigma/\sqrt{n})$$

- How large *n* needs to be for the approximation to be good enough depends upon how far from normal the population distribution is, but *n* ≥ 100 almost always suffices.

- If the population distribution of X is itself normal, then, regardless of n, the sampling distribution of x is exactly normal.

CLT

- *“If we average more and more independent random quantities with a common distribution, and that common distribution isn’t too pathological, then the average becomes closer and closer to a Gaussian”* – Cosma Shalizi

Applying the CLT

- *“I can average any large-ish bunch of numbers and divide by the sd and call it a z-score. Then I can compare it to a N(0,1) to determine statistical significance. I’ve got a hit if the number’s greater than 1.96!”*

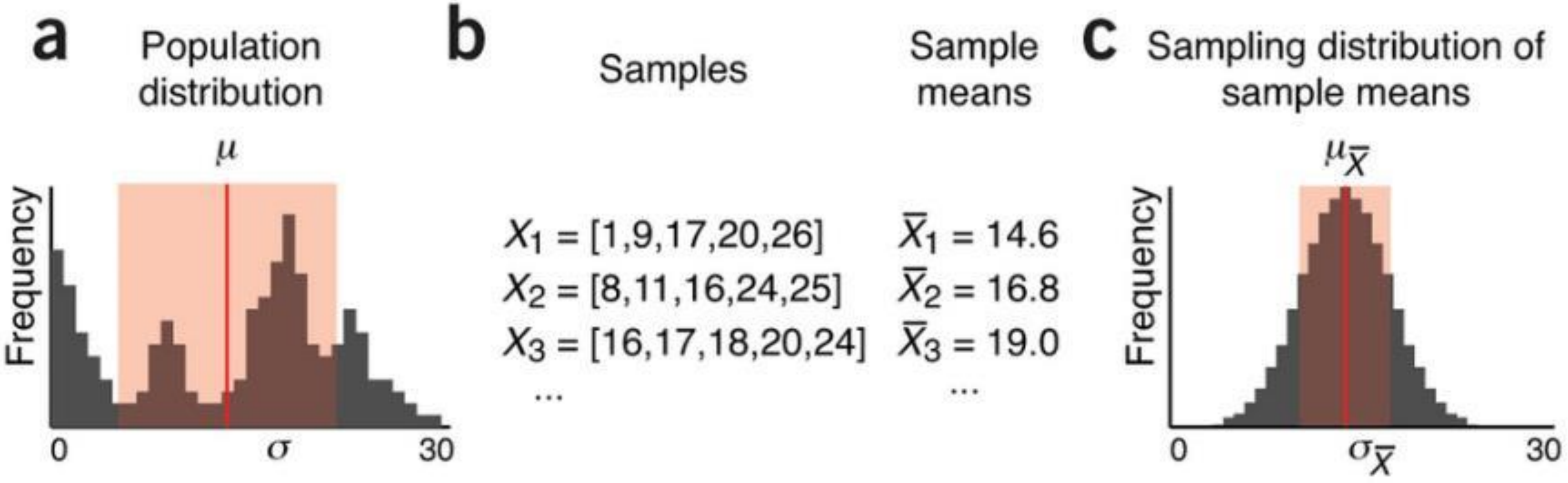
- the CLT assumes you’re averaging observations that are **iid**

- CLT applies in the case of:
  - Averaging gene expression for 1 gene across exchangeable subjects
  - Averaging disfluency use for 1 conversation task across exchangeable subjects

- CLT does not apply if:
  - Averaging gene expression for 1 subject across genes
  - Averaging disfluency use for 1 subject across conversation tasks

Done-ish

Population parameters are estimated by sampling



$X_1 = [1,9,17,20,26]$ $\overline{X}_1 = 14.6$

$X_2 = [8,11,16,24,25]$ $\overline{X}_2 = 16.8$

$X_3 = [16,17,18,20,24]$ $\overline{X}_3 = 19.0$

... ...

http://www.nature.com/nmeth/journal/v10/n9/full/nmeth.2613.html

QQplots in R

```
qqnorm(dataframe$variable)
qqline(dataframe$variable)

library(car)
qqPlot(dataframe$variable)
```

Do this!:
https://xiongge.shinyapps.io/QQplots/

Read this!:
http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot