7.1 Statistical Inference

PLSC30500, Fall 2021

co-taught by Molly Offer-Westort & Andy Eggers

(This lecture with references to Aronow & Miller (2019) and Wasserman (2004))

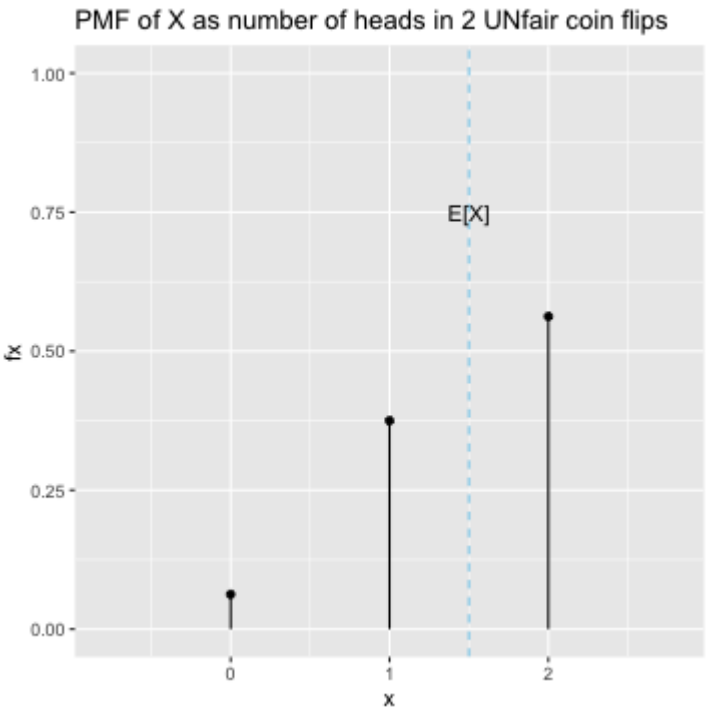
Introduction to estimation

Returning to our example where we flip a coin twice, let X be the number of heads we observe. Our coin is *not* fair, and the probability of getting a heads is 0.8.

The random variable's probability distribution is then:

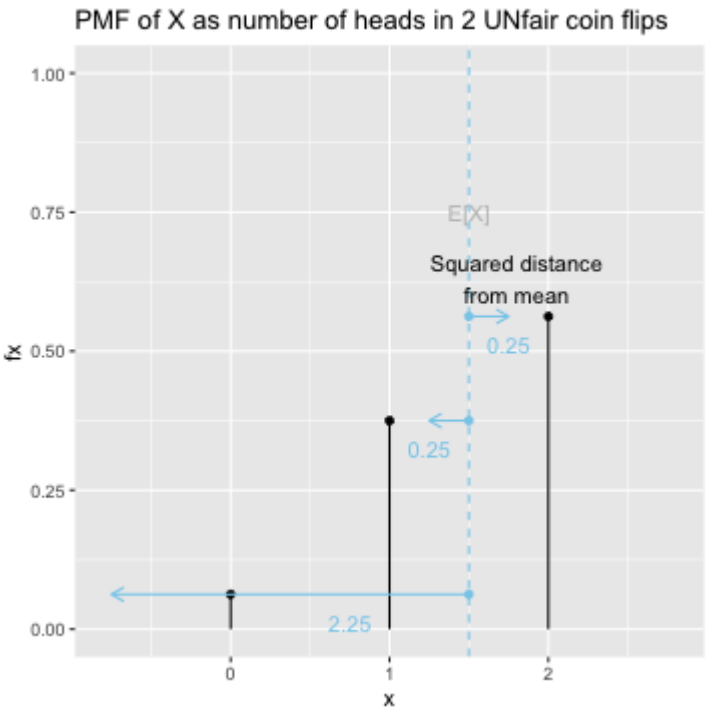
$$f(x) = \begin{cases} 1/16 & x = 0 \\ 3/8 & x = 1 \\ 9/16 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Let's take a look at the mean.



$$
\begin{aligned}
\mathrm{E}[X] &= \sum_x x f x \\
&= 0 \times \frac{1}{16} + 1 \times \frac{3}{8} + 2 \times \frac{9}{16} \\
&= \frac{24}{16} \\
&= 1.5
\end{aligned}
$$

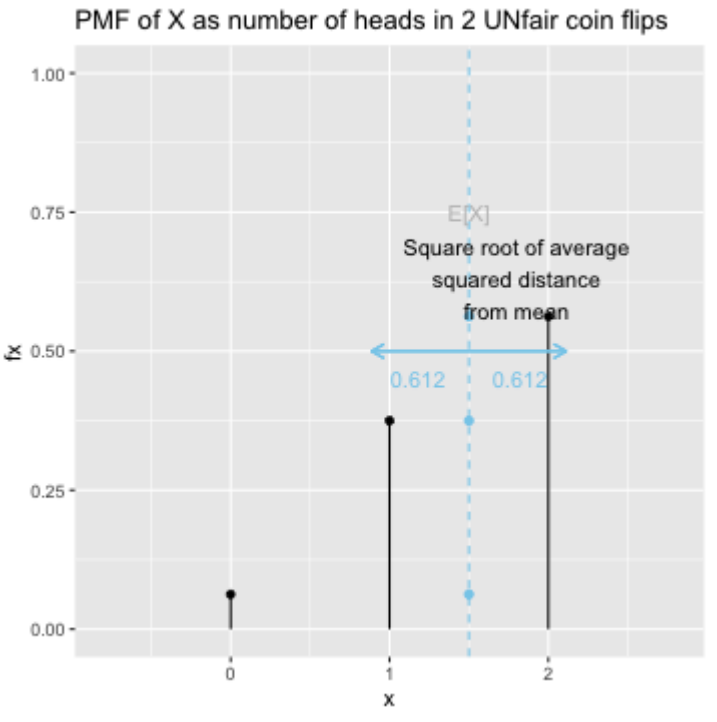
And the spread.



Variance = average squared distance from the mean

$$\begin{aligned}\text{Var}[X] &= \text{E}[(X - \text{E}[X)^2] \\ &= 2.25 \times \frac{1}{16} + 0.25 \times \frac{3}{8} + 0.25 \times \frac{9}{16} \\ &= 0.375\end{aligned}$$

And the spread.



SD = square root of variance

$$= \sqrt{0.375} = 0.612$$

We can check our calculations of the expectation and spread in R.

First, we'll want to simulate the random process. We'll run the simulation a large number of times, so we'll get an accurate calculation.

```
n <- 10000
X <- c(0, 1, 2)
probs <- c(1/16, 3/8, 9/16)
x.observed <- replicate(sample(X, prob = probs, replace = TRUE, size = 1), n = n)

head(x.observed)
```

```
## [1] 1 0 1 1 2 2
```

```
mean(x.observed)
```

```
## [1] 1.4998
```

```
var(x.observed)
```

```
## [1] 0.3782378
```

```
sd(x.observed)
```

```
## [1] 0.6150104
```

The process that we just did -- sampling and estimation based on observed data -- is a very common process in empirical research.

Sampling

Very often, we only observe a limited number of observations, which are drawn from a large population.

We can summarize the data we observe, but we would might like to *make inferences* about the larger population--i.e., to summarize what we know about that population based on the data we observe.

In our two coin flip example, suppose we don't know whether the coin is fair or not. We can observe the results of a large number of coin flips, and make an educated guess.

Formally, that educated guess is called *estimation*.

Random samples

We say our data is a *random sample* if our observations are *independent and identically distributed*.

Formally, if we have n draws, $X_1, \ldots, X_n$, these draws are i.i.d. if they are independent from each other, and all have the same CDF.

$$X_1, \ldots, X_n \sim F_X$$

Notational aside: $\sim$ is read as "distributed," and means that the random variable X has the distribution function F.

Sample mean

Let's repeat our random sampling from the double coin flip, but we'll consider a smaller sample, of size $n = 100$.

```
n <- 100
x.observed <- replicate(sample(X, prob = probs, replace = TRUE, size = 1), n = n)

head(x.observed)
```

```
## [1] 2 1 2 2 2 1
```

Our *sample mean* is the mean we observe in our data.

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

```
mean(x.observed)
```

```
## [1] 1.49
```

We differentiate the *sample mean* from the *population mean* because the sample mean will vary with every new sample we draw.

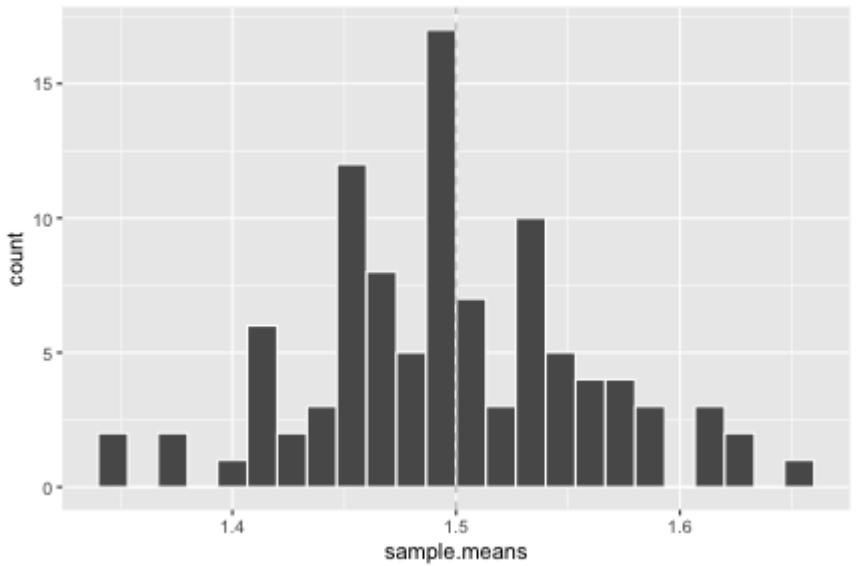
We'll use a loop to see what would happen if we took a sample of size $n = 100$ from the population distribution many times.

```
n.reps <- 100

x.mat <- matrix(ncol = n.reps, nrow = n)

for(i in 1:n.reps){
  x.mat[,i] <- replicate(sample(X, prob = probs, replace = TRUE, size = 1), n = n)
}
```

```
ggplot(tibble(sample.means = colMeans(x.mat)), aes(x = sample.means)) +
  geom_histogram(bins = 25, position = 'identity', color = 'white') +
  geom_vline(xintercept = Ex, color = 'grey', lty = 'dashed')
```



We see the sample means are roughly distributed around the mean of the underlying population.

The expected value of the sample mean is the population mean.

See Aronow & Miller for proof.

De moivre la place theorem

Sample variance

The unbiased sample variance is

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

This looks a little bit different from the population variance formula,

$$\operatorname{Var}[X] = \operatorname{E}[(X - \operatorname{E}[X])^2]$$

Why do we divide by $n-1$, instead of n?

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

The sample mean, $\bar{X}_n$, has an expected value of $\mathrm{E}[X]$. However, because it is made up of the $1, \ldots, n$ X_i that we actually observe, the expected difference between $(X_i - \bar{X}_n)$ is a little bit smaller than the expected difference between $(X_i - \mathrm{E}[X])$.

To account for this, we divide by $n-1$, instead of n.

The sample mean is itself a random variable, and so it has its own mean and variance. The mean of the sample mean is the population mean. The variance of the sample mean is:

$$\mathrm{Var}[\bar{X}] = \frac{\mathrm{Var}[X]}{n}$$

Let's check this in our simulation as well. We saw that mathematically, $\mathrm{Var}[X]$ was 0.375. So

$$\frac{\mathrm{Var}[X]}{n} = \frac{0.375}{100} = 0.00375$$

```
var(colMeans(x.mat))
```

```
## [1] 0.003746909
```

R has built in functions to calculate the sample mean and variance, but we can create functions of our own to check this.