The **bias** of an estimator $\hat{\theta}$ in estimating θ is defined as
$$\text{Bias} = E(\hat{\theta}) - \theta.$$
The **variance** of an estimator $\hat{\theta}$ θ is defined as
$$\text{Var} = E(\hat{\theta} - E(\hat{\theta}))^2.$$

Definition:
An estimator is **consistent** if its bias goes to zero
as the sample size $n \rightarrow \infty$ (n goes to infinity).

Criteria for a good estimator

A good estimator should have **low bias** and **low variance**.
We can combine these two criteria into one:

the **mean squared error**.

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$
$$= E\left((\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2\right)$$
$$= E\left((\hat{\theta} - E(\hat{\theta}))^2\right) + 2E\left((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)\right) + E\left((E(\hat{\theta}) - \theta)^2\right)$$
$$= var(\hat{\theta}) + 2 \times 0 \times (E(\hat{\theta}) - \theta) + \left(E(\hat{\theta}) - \theta\right)^2$$
$$= var(\hat{\theta}) + bias(\hat{\theta})^2$$

This is a **very important result**.
$MSE = var + bias^2$

MSE is an example of "*expected loss*"
        (in this case, the loss function = squared error).

Neither variance nor bias alone can be interpreted as
an expected loss-- for *any* loss function.

The two main frequentist ways to estimate are:

- **Maximum Likelihood Estimation**
- **Moment Estimation**

Maximum Likelihood Estimator (MLE):

For a model family $\{f_\theta : \theta \in \Theta\}$ and an observation x_{obs}, the **likelihood function** $\ell : \Theta \to \mathbb{R}$ is defined by

$$\ell(\theta) = f_\theta(x_{obs}).$$

A **maximum likelihood estimator** $\hat{\theta}$ satisfies
$$f_{\hat{\theta}}(x_{obs}) = \max_{\theta \in \Theta}\{f_\theta(x_{obs})\}.$$
Often $\hat{\theta}$ is unique.

Example   If $X \sim binom(n, p)$, and we observe $X = x_{obs}$, then

$$f_p(x_{obs}) = \binom{n}{x_{obs}} p^{x_{obs}}(1-p)^{n-x_{obs}}, \text{ therefore } \ell(p) \propto p^{x_{obs}}(1-p)^{n-x_{obs}},$$

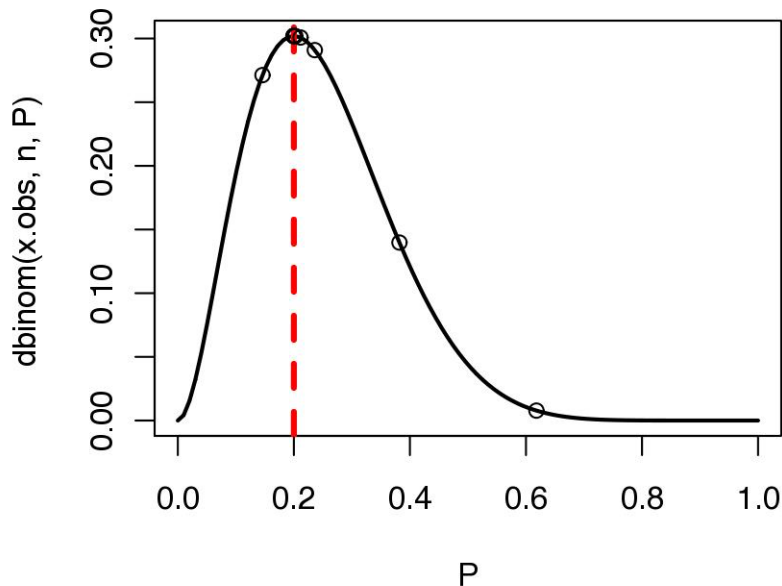and to maximize, we can differentiate and set the slope to zero:

$$\frac{d}{dp}\ell(p) = \frac{d}{dp}\left(p^{x_{obs}}(1-p)^{n-x_{obs}}\right)$$
$$= \left(x_{obs}p^{x_{obs}-1}(1-p)^{n-x_{obs}}\right) + \left(p^{x_{obs}}(-(n-x_{obs})(1-p)^{n-x_{obs}-1}\right)$$
$$\propto x_{obs}p^{-1} - (n-x_{obs})(1-p)^{-1}$$

This derivative is zero when
$$x_{obs}p^{-1} = (n-x_{obs})(1-p)^{-1}$$
$$x_{obs}(1-p) = (n-x_{obs})p$$
$$x_{obs} = (n-x_{obs})p + x_{obs}p = np$$
$$p = x_{obs}/n$$
So $\hat{p} = x_{obs}/n$.

Likelihood function


```
optimize(
    function(arg) {
        result=dbinom(x=2,
            size=10, prob=arg)
        points(arg, result)
        return(result)
    },
    lower=1e-10,
    upper=1-1e-10,
    maximum=TRUE
)
```

(When the parameter has dimension $D > 1$, you can do maximum likelihood estimation on the vector parameter, for example

- with algebra (set the vector of partial derivatives equal to the zero vector, and solve the D equations simultaneously),
- by searching in the D-dimensional space where the parameter lives,
- with the EM algorithm.

Moment estimators

A moment estimator is obtained by setting an observed value to its expected value (as a function of the parameter) and solving for the parameter.

Example   For the binomial, $E(X) = np$. So we solve $x_{obs} = n\hat{p}$, to get $\hat{p} = x_{obs} / n$.

So for the binomial,
MLE and moment estimator are the same,
$$\hat{p} = x_{obs} / n.$$

Example: The MLE is different from the Moment estimator

Estimating a normal distribution's variance:

Suppose we observe i.i.d. data $(x_1,...,x_n) \sim N(0,\sigma^2)$. We know that the mean is zero, but we don't know the variance.

Goal: to estimate the sampling variance σ^2.

First, let's try the MLE, the maximum likelihood estimator:

$$f(x_1,...,x_n \mid \sigma^2) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_i^2}{2\sigma^2}\right)\right)$$

$$\ell(\sigma^2) = \left(\frac{1}{\sqrt{\sigma^2}}\right)^n \exp\left(-\frac{\sum x_i^2}{2\sigma^2}\right) = \left(\sigma^2\right)^{-n/2}\exp\left(-\frac{\sum x_i^2}{2\sigma^2}\right)$$

Maximizing the likelihood is the same as maximizing the log-likelihood.

Differentiate the log-likelihood to find the maximizer:

$$\frac{d}{d\sigma^2}\log\ell(\sigma^2) = \frac{d}{d\sigma^2}\left(-\frac{n}{2}\log(2\pi\sigma^2)-\frac{\sum x_i^2}{2\sigma^2}\right)$$

$$= -\frac{n}{2}\frac{1}{\sigma^2}-\frac{\sum x_i^2}{2}\left(-(\sigma^2)^{-2}\right)$$

$$\propto -n\sigma^2 + \sum x_i^2$$

Setting this equal to zero, we get the MLE is $\widehat{\sigma^2} = n^{-1}\sum x_i^2$. That's logical!!

Now, how about the moment estimator?

$\sigma^{-2}\sum x_i^2$ has a "chi-square distribution on *n* degrees of freedom":

$$\frac{x_1^2 + ... + x_n^2}{\sigma^2} \sim \chi_n^2.$$

(χ_n^2 is the same as a gamma distribution $G(n/2, 1/2)$.)

The mean of χ_n^2 equals *n*. ← **important fact!**
So the moment estimator comes from setting $\sigma^{-2}\sum x_i^2 = n$:
$$\widehat{\sigma^2} = n^{-1}\sum x_i^2$$
just like the MLE.

But wait!
What if we don't know the true population *mean* μ either?
$$(x_1, ..., x_n) \sim N(\mu, \sigma^2)$$

How about trying **MLE**? Start with
$$\ell(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum(x_i - \mu)^2}{2\sigma^2}\right).$$
Maximize the likelihood over μ for fixed σ^2: $\hat{\mu} = \bar{x} = n^{-1}\sum x_i$,
the sample mean. Then maximizing over σ^2 gives
$$\widehat{\sigma^2} = n^{-1}\sum(x_i - \bar{x})^2,$$
like before,
but we replace the unknown mean μ by its estimate $\hat{\mu} = \bar{x}$.

How about **Method of Moments**? Things get different.
$\frac{1}{\sigma^2}\sum(x_i - \bar{x})^2$ has a chi-square distribution, but this time on
n – 1 degrees of freedom: χ_{n-1}^2. You can think of it as
"*losing a degree of freedom*" (losing a chunk of
information) because we have to use the extra information
to estimate μ.

So the moment estimator comes from setting
$\sigma^{-2}\sum(x_i - \bar{x})^2 = n - 1$, to get

$$\widehat{\sigma^2} = (n-1)^{-1}\sum\left(x_i - \bar{x}\right)^2.$$

This is BIGGER than the MLE,
by a factor of $n/(n-1) = 1 + 1/(n-1)$.

We say that the MLE in this case is **biassed**.
The **bias** of an estimator $\hat{\theta}$ in estimating θ is defined as

$\text{Bias} = E(\hat{\theta}) - \theta.$

$\text{Bias (MLE)} = E(\widehat{\sigma^2}) - \theta = E\left(n^{-1}\sum(x_i - \bar{x})^2\right) - \theta$

$\qquad\qquad = \dfrac{n-1}{n}E\left((n-1)^{-1}\sum(x_i - \bar{x})^2\right) - \theta$

$\qquad\qquad = \left(\dfrac{n-1}{n}\right)\theta - \theta = -n^{-1}\theta$

This bias is less than zero.
This means we are **overfitting**;
because we can tinker with a free parameter (μ),
we can be fooled into thinking the noise (variance)
is less than it is,
.... and thinking that the parameter estimates are more
accurate than they are.

Well, at least the bias goes to zero as n goes to ∞.