STA286 Lecture 28

Neil Montgomery

Last edited: 2017-03-30 11:41

comparing two proportions

One variable in the dataset with 0's and 1's; another variable splitting observations into two groups.

The two populations are Bernoulli(p_1) and Bernoulli(p_2). The independent samples are $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$

Patented process:

1. Estimate $\theta = p_1 - p_2$.

comparing two proportions

One variable in the dataset with 0's and 1's; another variable splitting observations into two groups.

The two populations are Bernoulli(p_1) and Bernoulli(p_2). The independent samples are $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$

Patented process:

1. Estimate $\theta = p_1 - p_2$.
2. Estimator $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \overline{X_1} - \overline{X_2}$

comparing two proportions

One variable in the dataset with 0's and 1's; another variable splitting observations into two groups.

The two populations are Bernoulli(p_1) and Bernoulli(p_2). The independent samples are $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$

Patented process:

1. Estimate $\theta = p_1 - p_2$.
2. Estimator $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \overline{X_1} - \overline{X_2}$
3. $\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

comparing two proportions

One variable in the dataset with 0's and 1's; another variable splitting observations into two groups.

The two populations are Bernoulli(p_1) and Bernoulli(p_2). The independent samples are $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$

Patented process:

1. Estimate $\theta = p_1 - p_2$.
2. Estimator $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \overline{X_1} - \overline{X_2}$
3. $\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
4. Whoops! Don't know p_1 or p_2. So use $\hat{p}_1$ and $\hat{p}_2$ instead. Bam. Done.

comparing two proportions

One variable in the dataset with 0's and 1's; another variable splitting observations into two groups.

The two populations are Bernoulli(p_1) and Bernoulli(p_2). The independent samples are $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$

Patented process:

1. Estimate $\theta = p_1 - p_2$.
2. Estimator $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \overline{X_1} - \overline{X_2}$
3. $\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
4. Whoops! Don't know p_1 or p_2. So use $\hat{p}_1$ and $\hat{p}_2$ instead. Bam. Done.

comparing two proportions

One variable in the dataset with 0's and 1's; another variable splitting observations into two groups.

The two populations are Bernoulli(p_1) and Bernoulli(p_2). The independent samples are $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$

Patented process:

1. Estimate $\theta = p_1 - p_2$.
2. Estimator $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \overline{X_1} - \overline{X_2}$
3. $\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
4. Whoops! Don't know p_1 or p_2. So use $\hat{p}_1$ and $\hat{p}_2$ instead. Bam. Done.

Formula for 95% interval:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_2}}$$

two non-robust confidence intervals

Every procedure I have explained so far is *robust* as long as the sample size is large enough (except for the prediction interval formula.)

In principle we could apply the patented procedure to estimate σ^2 with S^2, using a χ^2 distribution.

We could also apply the patented procedure to estimate the ratio σ_1^2/σ_2^2 with S_1^2/S_2^2 using an F distribution.

But the results are well known to be non-robust, even with large sample sizes, so I cannot recommend them for use.

two non-robust confidence intervals

Every procedure I have explained so far is *robust* as long as the sample size is large enough (except for the prediction interval formula.)

In principle we could apply the patented procedure to estimate σ^2 with S^2, using a χ^2 distribution.

We could also apply the patented procedure to estimate the ratio σ_1^2/σ_2^2 with S_1^2/S_2^2 using an F distribution.

But the results are well known to be non-robust, even with large sample sizes, so I cannot recommend them for use.

A modern computationally-intensive technique called *bootstrapping* is a good choice in these and other difficult situations.

chiselling your own estimators onto stone tablets

fun fact from mathematics

Suppose a twice-differentiable function $f(x)$ has a critical value at x_0, and $g(x)$ is strictly increasing and twice-differentiable.

Then $g(f(x))$ also has a critical value at x_0, and the sign of its second derivative at x_0 is the same as the sign of the second derivative of f at x_0.

This can be seen by evaluating the left hand sides at x_0:

$$
\begin{aligned}
(g(f(x)))' &= g'(f(x))f'(x) \\
(g(f(x)))'' &= g''(f(x))(f'(x))^2 + g'(f(x))f''(x)
\end{aligned}
$$

estimating a proportion, from first principles - I

Here's a simulated sequence of 0's and 1's from a Bernoulli(p) distribution. I know what (p), but you don't.

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

What value of p between 0 and 1 is the *most likely* to have produce this sequence of 2 1's and 8 0's?

estimating a proportion, from first principles - I

Here's a simulated sequence of 0's and 1's from a Bernoulli(p) distribution. I know what (p), but you don't.

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

What value of p between 0 and 1 is the *most likely* to have produce this sequence of 2 1's and 8 0's?

The probability of getting this sample exactly is:

$$
\begin{aligned}
&(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot p\cdot(1-p)\cdot p \\
&= p^2(1-p)^8
\end{aligned}
$$

Let's call this function $L(p)$.

estimating a proportion, from first principles - I

Here's a simulated sequence of 0's and 1's from a Bernoulli(p) distribution. I know what (p), but you don't.

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

What value of p between 0 and 1 is the *most likely* to have produce this sequence of 2 1's and 8 0's?

The probability of getting this sample exactly is:

$$
\begin{aligned}
&(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot p\cdot(1-p)\cdot p \\
&= p^2(1-p)^8
\end{aligned}
$$

Let's call this function $L(p)$.

We could maximize $L(p)$, but it's easier to maximize $\ell(p) = \log L(p)$

estimating a proportion, from first principles - II

$$0 = \frac{d}{dp}\ell(p) = \frac{d}{dp}\left(2\log(p) + 8\log(1-p)\right) = \frac{2}{p} - \frac{8}{1-p}$$

estimating a proportion, from first principles - II

$$0 = \frac{d}{dp}\ell(p) = \frac{d}{dp}\left(2\log(p) + 8\log(1-p)\right) = \frac{2}{p} - \frac{8}{1-p}$$

$$\frac{2}{p} = \frac{8}{1-p} \implies p = \frac{2}{10}$$

The second derivative is negative, so this is a maximum.

estimating a proportion, from first principles - II

$$0 = \frac{d}{dp}\ell(p) = \frac{d}{dp}\left(2\log(p) + 8\log(1-p)\right) = \frac{2}{p} - \frac{8}{1-p}$$
$$\frac{2}{p} = \frac{8}{1-p} \implies p = \frac{2}{10}$$

The second derivative is negative, so this is a maximum.

It would have been no harder to work in general, with k 1's out of a sample of size n, and maximizing $L(p) = p^k(1-p)^{n-k}$

estimating a proportion, from first principles - II

$$0 = \frac{d}{dp}\ell(p) = \frac{d}{dp}\left(2\log(p) + 8\log(1-p)\right) = \frac{2}{p} - \frac{8}{1-p}$$

$$\frac{2}{p} = \frac{8}{1-p} \implies p = \frac{2}{10}$$

The second derivative is negative, so this is a maximum.

It would have been no harder to work in general, with k 1's out of a sample of size n, and maximizing $L(p) = p^k(1-p)^{n-k}$

The same calculus gives the maximum at k/n.

estimating a proportion, from first principles - II

$$0 = \frac{d}{dp}\ell(p) = \frac{d}{dp}\left(2\log(p) + 8\log(1-p)\right) = \frac{2}{p} - \frac{8}{1-p}$$
$$\frac{2}{p} = \frac{8}{1-p} \implies p = \frac{2}{10}$$

The second derivative is negative, so this is a maximum.

It would have been no harder to work in general, with k 1's out of a sample of size n, and maximizing $L(p) = p^k(1-p)^{n-k}$

The same calculus gives the maximum at k/n.

This is exactly the same as $\hat{p}$ that was used as "obvious" from before.

“likelihood function” for Bernoulli

The p.m.f. of a Bernoulli(p) is $f(x; p) = p^x(1-p)^{1-x}$ with $x \in \{0, 1\}$.

“likelihood function” for Bernoulli

The p.m.f. of a Bernoulli(p) is $f(x; p) = p^x(1-p)^{1-x}$ with $x \in \{0, 1\}$.

Given a sequence $\{x_1, \ldots, x_n\}$ of 0’s and 1’s, yet another way of constructing $L(p)$ is as follows:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

“likelihood function” for Bernoulli

The p.m.f. of a Bernoulli(p) is $f(x; p) = p^x(1-p)^{1-x}$ with $x \in \{0, 1\}$.

Given a sequence $\{x_1, \dots, x_n\}$ of 0’s and 1’s, yet another way of constructing $L(p)$ is as follows:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

"likelihood function" for Bernoulli

The p.m.f. of a Bernoulli(p) is $f(x;p) = p^x(1-p)^{1-x}$ with $x \in \{0,1\}$.

Given a sequence $\{x_1, \ldots, x_n\}$ of 0's and 1's, yet another way of constructing $L(p)$ is as follows:

$$
\begin{aligned}
L(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \\
&= \prod_{i=1}^{n} f(x_i;p)
\end{aligned}
$$

"likelihood function" for Bernoulli

The p.m.f. of a Bernoulli(p) is $f(x; p) = p^x(1-p)^{1-x}$ with $x \in \{0, 1\}$.

Given a sequence $\{x_1, \ldots, x_n\}$ of 0's and 1's, yet another way of constructing $L(p)$ is as follows:

$$
\begin{aligned}
L(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \\
&= \prod_{i=1}^{n} f(x_i; p)
\end{aligned}
$$

$$
\text{Also:} \quad \ell(p) = \log L(p) = \sum_{i=1}^{n} \log f(x_i; p)
$$

likelihood function in general

Given a sequence of observations $\{x_1, \ldots, x_n\}$ ("the data") from a random variable X with pmf or pdf $f(x;\theta)$, a likelihood function $L(\theta) = L(x_1, \ldots, x_n; \theta)$ for the parameter θ is defined as (for any positive g):

$$L(\theta) = \underbrace{g(\mathbf{x}) \prod_{i=1}^{n} f(x_i;\theta)}_{\text{real definition}}$$

likelihood function in general

Given a sequence of observations $\{x_1, \ldots, x_n\}$ ("the data") from a random variable X with pmf or pdf $f(x; \theta)$, a likelihood function $L(\theta) = L(x_1, \ldots, x_n; \theta)$ for the parameter θ is defined as (for any positive g):

$$L(\theta) = \underbrace{g(\mathbf{x}) \prod_{i=1}^{n} f(x_i; \theta)}_{\text{real definition}} \propto \underbrace{\prod_{i=1}^{n} f(x_i; \theta)}_{\text{easy definition}}$$

If X is discrete and f is a pmf, then $L(\theta)$ is literally the probability of the data given θ.

likelihood function in general

Given a sequence of observations $\{x_1, \ldots, x_n\}$ ("the data") from a random variable X with pmf or pdf $f(x;\theta)$, a likelihood function $L(\theta) = L(x_1, \ldots, x_n; \theta)$ for the parameter θ is defined as (for any positive g):

$$L(\theta) = \underbrace{g(\mathbf{x}) \prod_{i=1}^{n} f(x_i;\theta)}_{\text{real definition}} \propto \underbrace{\prod_{i=1}^{n} f(x_i;\theta)}_{\text{easy definition}}$$

If X is discrete and f is a pmf, then $L(\theta)$ is literally the probability of the data given θ.

If X is continuous and f is a pdf, then $L(\theta)$ is not a probability, but it still provides a useful "index" for θ values.

likelihood as “index” in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1, 3, 8; \lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

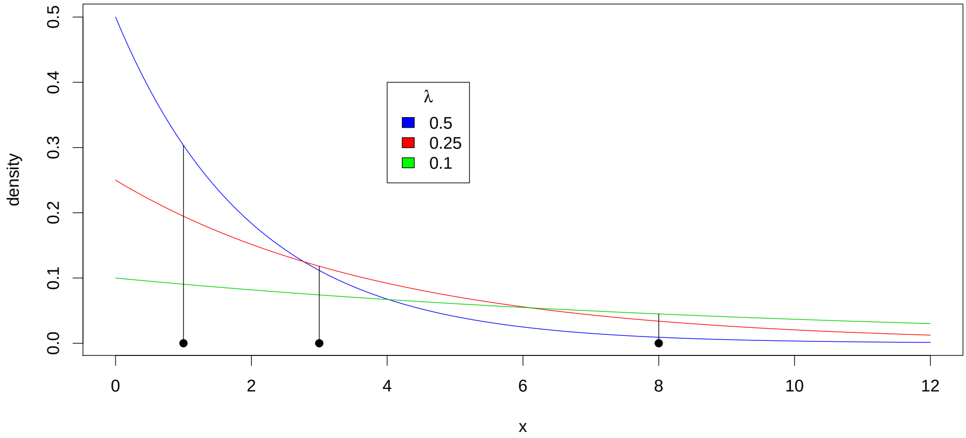
likelihood as "index" in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1,3,8;\lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

Consider three possible candidate guesses for the true value of λ: 0.1, 0.25, and 0.5.

a possibly useless and confusing picture



likelihood as "index" in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1, 3, 8; \lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

likelihood as "index" in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1,3,8;\lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

Consider three possible candidate guesses for the true value of λ: 0.1, 0.25, and 0.5.

$$L(0.5) = 0.5^3 e^{-12} = 3.0984402 \times 10^{-4}$$

likelihood as "index" in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1,3,8;\lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

Consider three possible candidate guesses for the true value of λ: 0.1, 0.25, and 0.5.

$$\begin{aligned} L(0.5) &= 0.5^3 e^{-12} = 3.0984402 \times 10^{-4} \\ L(0.25) &= (0.25)^3 e^{-12 \cdot 0.25} = 7.7792294 \times 10^{-4} \end{aligned}$$

likelihood as “index” in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1,3,8;\lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

Consider three possible candidate guesses for the true value of λ: 0.1, 0.25, and 0.5.

$$\begin{aligned} L(0.5) &= 0.5^3 e^{-12} = 3.0984402 \times 10^{-4} \\ L(0.25) &= (0.25)^3 e^{-12 \cdot 0.25} = 7.7792294 \times 10^{-4} \\ L(0.1) &= (0.1)^3 e^{-12 \cdot 0.1} = 3.0119421 \times 10^{-4} \end{aligned}$$

likelihood as “index” in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1,3,8;\lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

Consider three possible candidate guesses for the true value of λ: 0.1, 0.25, and 0.5.

$$\begin{aligned} L(0.5) &= 0.5^3 e^{-12} = 3.0984402 \times 10^{-4} \\ L(0.25) &= (0.25)^3 e^{-12 \cdot 0.25} = 7.7792294 \times 10^{-4} \longleftarrow \text{ Highest "likelihood"} \\ L(0.1) &= (0.1)^3 e^{-12 \cdot 0.1} = 3.0119421 \times 10^{-4} \end{aligned}$$

maximum likelihood “estimate”

The value of θ that maximizes $L(\theta)$ is called the *maximum likelihood estimate*.

maximum likelihood “estimate”

The value of θ that maximizes $L(\theta)$ is called the *maximum likelihood estimate*.

In many cases it is more convenient to maximize $\ell(\theta)$.

maximum likelihood “estimate”

The value of θ that maximizes $L(\theta)$ is called the *maximum likelihood estimate*.

In many cases it is more convenient to maximize $\ell(\theta)$.

For example, suppose $x_1, x_2, \ldots, x_n$ are data observed from a $X \sim N(\mu, 1)$ population. A likelihood for μ is:

$$L(\mu) = (2\pi)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

maximum likelihood “estimate”

The value of θ that maximizes $L(\theta)$ is called the *maximum likelihood estimate*.

In many cases it is more convenient to maximize $\ell(\theta)$.

For example, suppose $x_1, x_2, \ldots, x_n$ are data observed from a $X \sim N(\mu, 1)$ population. A likelihood for μ is:

$$L(\mu) = (2\pi)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$\ell(\mu) = \log L(\mu) = C - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

maximum likelihood "estimate"

The value of θ that maximizes $L(\theta)$ is called the *maximum likelihood estimate*.

In many cases it is more convenient to maximize $\ell(\theta)$.

For example, suppose $x_1, x_2, \ldots, x_n$ are data observed from a $X \sim N(\mu, 1)$ population. A likelihood for μ is:

$$L(\mu) = (2\pi)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$\ell(\mu) = \log L(\mu) = C - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To maximize:

$$0 = \frac{d}{d\mu}\ell(\mu) = \sum_{i=1}^{n}(x_i - \mu)$$

maximum likelihood "estimate"

The value of θ that maximizes $L(\theta)$ is called the *maximum likelihood estimate*.

In many cases it is more convenient to maximize $\ell(\theta)$.

For example, suppose $x_1, x_2, \ldots, x_n$ are data observed from a $X \sim N(\mu, 1)$ population. A likelihood for μ is:

$$L(\mu) = (2\pi)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$\ell(\mu) = \log L(\mu) = C - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To maximize:

$$0 = \frac{d}{d\mu}\ell(\mu) = \sum_{i=1}^{n}(x_i - \mu) \Longrightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

the maximum likelihood estimator

A final technicality. When you replace the data $x_1, x_2, \ldots, x_n$ with its "model", the sample: $X_1, X_2, \ldots, X_n$, inside the maximum likelihood estimate, you end up with the *maximum likelihood estimator*, or MLE.

Traditionally the MLE notation is the parameter-with-a-hat.

the maximum likelihood estimator

A final technicality. When you replace the data $x_1, x_2, \ldots, x_n$ with its "model", the sample: $X_1, X_2, \ldots, X_n$, inside the maximum likelihood estimate, you end up with the *maximum likelihood estimator*, or MLE.

Traditionally the MLE notation is the parameter-with-a-hat.

For example, the maximum likelihood estimator for μ using a sample from a $N(\mu, 1)$ population is:

$$\hat{\mu} = \overline{X}$$

Everything so far extends to vector parameters. For example (textbook example 9.21), the maximum likelihood estimates given data $x_1, \ldots, x_n$ from a $N(\mu, \sigma)$ population, the MLE for $\theta = (\mu, \sigma^2)$ are:

$$\hat{\mu} = \overline{X} \qquad \widehat{\sigma^2} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n}$$

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a 1-1 function.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a 1-1 function.
4. it is asymptotically normal.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a 1-1 function.
4. it is asymptotically normal.
5. if $c\hat{\theta}$ is unbiased for some constant c, then $c\hat{\theta}$ is the unbiased estimator with the smallest variance (our "gold standard".)