CS70: Lecture 22.

Part I: Confidence Intervals Again | Part II: Linear Regression

1. Confidence?
2. Example
3. Review of Chebyshev
4. Confidence Interval with Chebyshev
5. More examples

Confidence?

- You flip a coin once and get H.

  Do think that $Pr[H] = 1$?
- You flip a coin 10 times and get 5 Hs.

  Are you sure that $Pr[H] = 0.5$?
- You flip a coin 10^6 times and get 35% of Hs.

  How much are you willing to bet that $Pr[H]$ is exactly 0.35?

  How much are you willing to bet that $Pr[H] \in [0.3, 0.4]$?

  Did different exam rooms perform differently? (6 afraid of 7?)

More generally, you estimate an unknown quantity θ.

Your estimate is $\hat{\theta}$.

How much confidence do you have in your estimate?

Confidence?

Confidence is essential is many applications:

- How effective is a medication?
- Are we sure of the milage of a car?
- Can we guarantee the lifespan of a device?
- We simulated a system. Do we trust the simulation results?
- Is an algorithm guaranteed to be fast?
- Do we know that a program has no bug?

As scientists and engineers, be convinced of this fact:

An estimate without confidence level is useless!

Confidence Interval

The following definition captures precisely the notion of confidence.

Definition: Confidence Interval

An interval $[a,b]$ is a 95%-confidence interval for an unknown quantity θ if

$$Pr[\theta \in [a,b]] \geq 95\%.$$

The interval $[a,b]$ is calculated on the basis of observations.

Here is a typical framework. Assume that $X_1, X_2, \dots, X_n$ are i.i.d. and have a distribution that depends on some parameter θ.

For instance, $X_n = B(\theta)$.

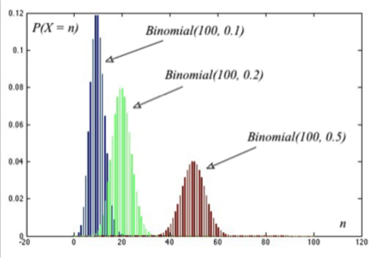
Thus, more precisely, given θ, the random variables X_n are i.i.d. with a known distribution (that depends on θ).

- We observe $X_1, \dots, X_n$
- We calculate $a = a(X_1, \dots, X_n)$ and $b = b(X_1, \dots, X_n)$
- If we can guarantee that $Pr[\theta \in [a,b]] \geq 95\%$, then $[a,b]$ is a 95%-CI for θ.

Confidence Interval: Applications

- We poll 1000 people.
  - Among those, 48% declare they will vote for Trump.
  - We do some calculations ....
  - We conclude that $[0.43, 0.53]$ is a 95%-CI for the fraction of all the voters who will vote for Trump.
- We observe $1,000$ heart valve replacements that were performed by Dr. Bill.
  - Among those, 35 patients died during surgery. (Sad example!)
  - We do some calculations ...
  - We conclude that $[1\%, 5\%]$ is a 95%-CI for the probability of dying during that surgery by Dr. Bill.
  - We do a similar calculation for Dr. Fred.
  - We find that $[8\%, 12\%]$ is a 95%-CI for Dr. Fred's surgery.
  - What surgeon do you choose?

Coin Flips: Intuition



Say that you flip a coin $n = 100$ times and observe 20 Hs.

If $p := Pr[H] = 0.5$, this event is very unlikely.

Intuitively, if is unlikely that the fraction of Hs, say A_n, differs a lot from $p := Pr[H]$.

Thus, it is unlikely that p differs a lot from A_n. Hence, one should be able to build a confidence interval $[A_n - \varepsilon, A_n + \varepsilon]$ for p.

The key idea is that $|A_n - p| \leq \varepsilon \Leftrightarrow p \in [A_n - \varepsilon, A_n + \varepsilon]$.

Thus, $Pr[|A_n - p| > \varepsilon] \leq 5\% \Leftrightarrow Pr[p \in [A_n - \varepsilon, A_n + \varepsilon]] \geq 95\%$.

It remains to find ε such that $Pr[|A_n - p| > \varepsilon] \leq 5\%$.

One approach: Chebyshev.

Confidence Interval with Chebyshev

- Flip a coin n times. Let A_n be the fraction of Hs.
- Can we find ε such that $Pr[|A_n - p| > \varepsilon] \leq 5\%$?

Using Chebyshev, we will see that $\varepsilon = 2.25\frac{1}{\sqrt{n}}$ works. Thus

$$[A_n - \frac{2.25}{\sqrt{n}}, A_n + \frac{2.25}{\sqrt{n}}] \text{ is a 95\%-CI for } p.$$

Example: If $n = 1500$, then $Pr[p \in [A_n - 0.05, A_n + 0.05]] \geq 95\%$.

In fact, $a = \frac{1}{\sqrt{n}}$ works, so that with $n = 1,500$ one has
$Pr[p \in [A_n - 0.02, A_n + 0.02]] \geq 95\%$.

Confidence Intervals: Result

Theorem:
Let X_n be i.i.d. with mean μ and variance σ^2.
Define $A_n = \frac{X_1 + \cdots + X_n}{n}$. Then,

$$Pr[\mu \in [A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]] \geq 95\%.$$

Thus, $[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]$] is a 95%-CI for μ.

Example: Let $X_n = 1\{$ coin n yields $H\}$. Then

$$\mu = E[X_n] = p := Pr[H]. \text{ Also, } \sigma^2 = var(X_n) = p(1-p) \leq \frac{1}{4}.$$

Hence, $[A_n - 4.5\frac{1/2}{\sqrt{n}}, A_n + 4.5\frac{1/2}{\sqrt{n}}]$] is a 95%-CI for p.

Confidence Interval: Analysis

We prove the theorem, i.e., that $A_n \pm 4.5\sigma/\sqrt{n}$ is a 95%-CI for μ.

From Chebyshev:

$$Pr[|A_n - \mu| \geq 4.5\sigma/\sqrt{n}] \leq \frac{var(A_n)}{[4.5\sigma/\sqrt{n}]^2} = \frac{n}{20\sigma^2} var(A_n).$$

Now,

$$\begin{aligned} var(A_n) &= var(\frac{X_1 + \cdots + X_n}{n}) = \frac{1}{n^2} var(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2} \times n.var(X_1) = \frac{1}{n}\sigma^2. \end{aligned}$$

Hence,

$$Pr[|A_n - \mu| \geq 4.5\sigma/\sqrt{n}] \leq \frac{n}{20\sigma^2} \times \frac{1}{n}\sigma^2 = 5\%.$$

Thus,

$$Pr[|A_n - \mu| \leq 4.5\sigma/\sqrt{n}] \geq 95\%.$$

Hence,

$$Pr[\mu \in [A_n - 4.5\sigma/\sqrt{n}, A_n + 4.5\sigma/\sqrt{n}]] \geq 95\%.$$

$\square$

Confidence interval for p in $B(p)$

Let X_n be i.i.d. $B(p)$. Define $A_n = (X_1 + \cdots + X_n)/n$.

Theorem:

$$[A_n - \frac{2.25}{\sqrt{n}}, A_n + \frac{2.25}{\sqrt{n}}] \text{ is a 95\%-CI for } p.$$

Proof:

We have just seen that

$$Pr[\mu \in [A_n - 4.5\sigma/\sqrt{n}, A_n + 4.5\sigma/\sqrt{n}]] \geq 95\%.$$

Here, $\mu = p$ and $\sigma^2 = p(1-p)$. Thus, $\sigma^2 \leq \frac{1}{4}$ and $\sigma \leq \frac{1}{2}$.

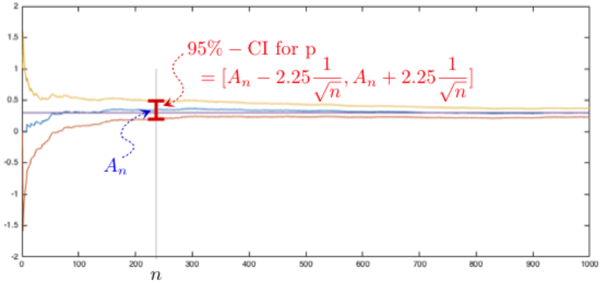
Thus,

$$Pr[\mu \in [A_n - 4.5 \times 0.5/\sqrt{n}, A_n + 4.5 \times 0.5/\sqrt{n}]] \geq 95\%.$$

$\square$

Confidence interval for p in $B(p)$

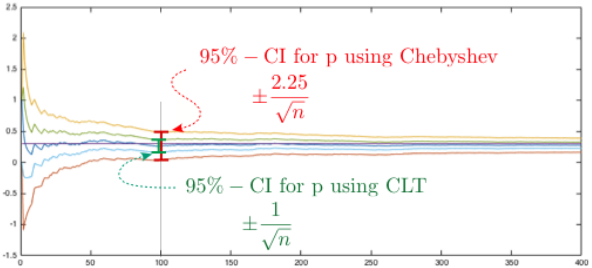
An illustration:



Good practice: You run your simulation, or experiment. You get an estimate. You indicate your confidence interval.

Confidence interval for p in $B(p)$

Improved CI: In fact, one can replace 2.25 by 1.



Quite a bit of work to get there: continuous random variables; Gaussian; Central Limit Theorem.

Confidence Interval for $1/p$ in $G(p)$

Let X_n be i.i.d. $G(p)$. Define $A_n = (X_1 + \cdots + X_n)/n$.

Theorem:

$$[\frac{A_n}{1 + 4.5/\sqrt{n}}, \frac{A_n}{1 - 4.5/\sqrt{n}}] \text{ is a 95\%-CI for } \frac{1}{p}.$$

Proof: We know that

$$Pr[\mu \in [A_n - 4.5\sigma/\sqrt{n}, A_n + 4.5\sigma/\sqrt{n}]] \geq 95\%.$$

Here, $\mu = \frac{1}{p}$ and $\sigma = \frac{\sqrt{1-p}}{p} \leq \frac{1}{p}$. Hence,

$$Pr[\frac{1}{p} \in [A_n - 4.5\frac{1}{p\sqrt{n}}, A_n + 4.5\frac{1}{p\sqrt{n}}]] \geq 95\%.$$

Now, $A_n - 4.5\frac{1}{p\sqrt{n}} \leq \frac{1}{p} \leq \frac{1}{p} \leq A_n + 4.5\frac{1}{p\sqrt{n}}$ is equivalent to

$$\frac{A_n}{1 + 4.5/\sqrt{n}} \leq \frac{1}{p} \leq \frac{A_n}{1 - 4.5/\sqrt{n}}.$$

□

Examples: $[0.7A_{100}, 1.8A_{100}]$ and $[0.96A_{10000}, 1.05A_{10000}]$.

Which Coin is Better?

You are given coin A and coin B. You want to find out which one has a larger $Pr[H]$. Let p_A and p_B be the values of $Pr[H]$ for the two coins.

Approach:

- Flip each coin n times.
- Let A_n be the fraction of Hs for coin A and B_n for coin B.
- Assume $A_n > B_n$. It is tempting to think that $p_A > p_B$. Confidence?

Analysis: Note that

$$E[A_n - B_n] = p_A - p_B \text{ and } var(A_n - B_n) = \frac{1}{n}(p_A(1-p_A) + p_B(1-p_B)) \leq \frac{1}{2n}.$$

Thus, $Pr[|A_n - B_n - (p_A - p_B)| > \varepsilon] \leq \frac{1}{2n\varepsilon^2}$, so

$$Pr[p_A - p_B \in [A_n - B_n - \varepsilon, A_n - B_n + \varepsilon]] \geq 1 - \frac{1}{2n\varepsilon^2}, \text{ and}$$

$$Pr[p_A - p_B \geq 0] \geq 1 - \frac{1}{2n(A_n - B_n)^2}.$$

Example: With $n = 100$ and $A_n - B_n = 0.2$, $Pr[p_A > p_B] \geq 1 - \frac{1}{8} = 0.875$.

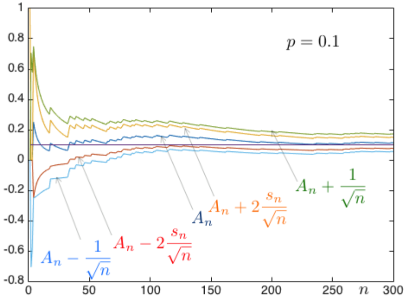
Unknown σ

For $B(p)$, we wanted to estimate p. The CI requires $\sigma = \sqrt{p(1-p)}$. We replaced σ by an upper bound: $1/2$.

In some applications, it may be OK to replace σ^2 by the following sample variance:

$$s_n^2 := \frac{1}{n} \sum_{m=1}^{n} (X_m - A_n)^2.$$

However, in some cases, this is dangerous! The theory says it is OK if the distribution of X_n is nice (Gaussian). This is used regularly in practice. However, be aware of the risk.



Summary

$\boxed{\text{Confidence Intervals}}$

1. Estimates without confidence level are useless!
2. $[a,b]$ is a 95%-CI for θ if $Pr[\theta \in [a,b]] \geq 95\%$.
3. Using Chebyshev: $[A_n - 4.5\sigma/\sqrt{n}, A_n + 4.5\sigma/\sqrt{n}]$ is a 95%-CI for μ.
4. Using CLT, we will replace 4.5 by 2.
5. When σ is not known, one can replace it by an upper bound.
6. Examples: $B(p), G(p)$, which coin is better?
7. In some cases, one can replace σ by the empirical standard deviation.

Linear Regression.

Linear Regression

1. Preamble
2. Motivation for LR
3. History of LR
4. Linear Regression
5. Derivation
6. More examples

Linear Regression: Preamble

The best guess about Y, if we know only the distribution of Y, is $E[Y]$.

More precisely, the value of a that minimizes $E[(Y-a)^2]$ is $a = E[Y]$.

Proof:

Let $\hat{Y} := Y - E[Y]$. Then, $E[\hat{Y}] = 0$. So, $E[\hat{Y}c] = 0, \forall c$. Now,

$$\begin{aligned} E[(Y-a)^2] &= E[(Y - E[Y] + E[Y] - a)^2] \\ &= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a \\ &= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2 \\ &= E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2]. \end{aligned}$$

Hence, $E[(Y-a)^2] \geq E[(Y - E[Y])^2], \forall a$. $\square$

Linear Regression: Preamble

Thus, if we want to guess the value of Y, we choose $E[Y]$.

Now assume we make some observation X related to Y.

How do we use that observation to improve our guess about Y?

The idea is to use a function $g(X)$ of the observation to estimate Y.

The simplest function $g(X)$ is a constant that does not depend of X.

The next simplest function is linear: $g(X) = a + bX$.

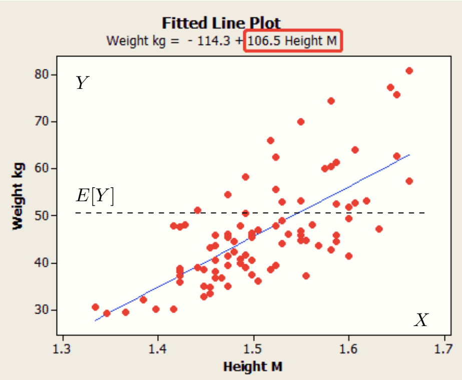
What is the best linear function? That is our next topic.

A bit later, we will consider a general function $g(X)$.

Linear Regression: Motivation

Example 1: 100 people.

Let (X_n, Y_n) = (height, weight) of person n, for $n = 1, \ldots, 100$:



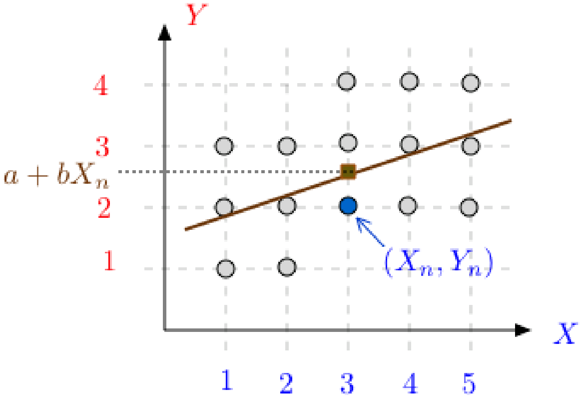
The blue line is $Y = -114.3 + 106.5X$. (X in meters, Y in kg.)

Best linear fit: Linear Regression.

Motivation

Example 2: 15 people.

We look at two attributes: (X_n, Y_n) of person n, for $n = 1, \ldots, 15$:



The line $Y = a + bX$ is the linear regression.

Covariance

Definition The covariance of X and Y is

$$cov(X, Y) := E[(X - E[X])(Y - E[Y])].$$

Fact

$$cov(X, Y) = E[XY] - E[X]E[Y].$$

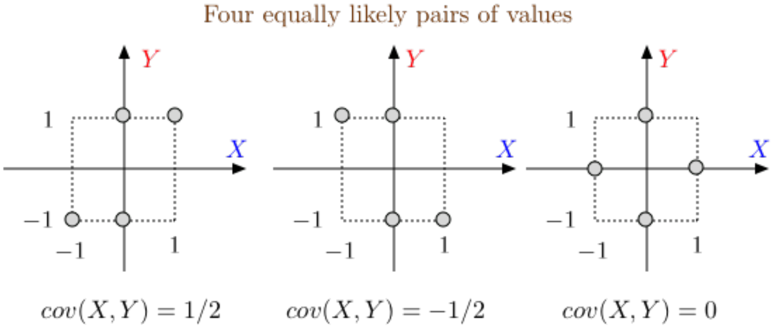
Proof:
Think about $E[X] = E[Y] = 0$. Just $E[XY]$. $\square$ish.

For the sake of completeness.

$$\begin{aligned} E[(X - E[X])(Y - E[Y])] &= E[XY - E[X]Y - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

$\square$

Examples of Covariance



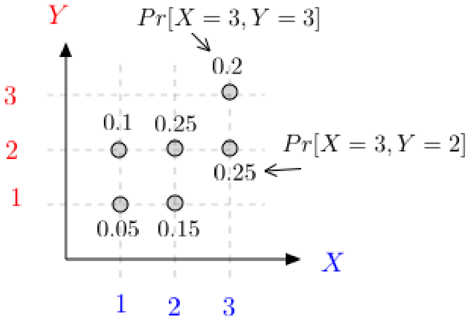
Note that $E[X] = 0$ and $E[Y] = 0$ in these examples. Then $cov(X, Y) = E[XY]$.

When $cov(X, Y) > 0$, the RVs X and Y tend to be large or small together. X and Y are said to be positively correlated.

When $cov(X, Y) < 0$, when X is larger, Y tends to be smaller. X and Y are said to be negatively correlated.

When $cov(X, Y) = 0$, we say that X and Y are uncorrelated.

Examples of Covariance



$$E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9$$
$$E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8$$
$$E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2$$
$$E[XY] = 1 \times 0.05 + 1 \times 2 \times 0.1 + \cdots + 3 \times 3 \times 0.2 = 4.85$$
$$cov(X, Y) = E[XY] - E[X]E[Y] = 1.05$$
$$var[X] = E[X^2] - E[X]^2 = 2.19.$$

Properties of Covariance

$$cov(X,Y) = E[(X-E[X])(Y-E[Y])] = E[XY] - E[X]E[Y].$$

Fact

(a) $var[X] = cov(X,X)$

(b) X, Y independent $\Rightarrow cov(X,Y) = 0$

(c) $cov(a+X, b+Y) = cov(X,Y)$

(d) $cov(aX+bY, cU+dV) = ac \cdot cov(X,U) + ad \cdot cov(X,V)$
$\qquad + bc \cdot cov(Y,U) + bd \cdot cov(Y,V).$

Proof:

(a)-(b)-(c) are obvious.

(d) In view of (c), one can subtract the means and assume that the RVs are zero-mean. Then,

$$\begin{aligned} cov(aX+bY, cU+dV) &= E[(aX+bY)(cU+dV)] \\ &= ac \cdot E[XU] + ad \cdot E[XV] + bc \cdot E[YU] + bd \cdot E[YV] \\ &= ac \cdot cov(X,U) + ad \cdot cov(X,V) + bc \cdot cov(Y,U) + bd \cdot cov(Y,V). \end{aligned}$$

$\square$

Linear Regression: Non-Bayesian

Definition
Given the samples $\{(X_n, Y_n), n = 1, \ldots, N\}$, the Linear Regression of Y over X is

$$\hat{Y} = a + bX$$

where (a, b) minimize

$$\sum_{n=1}^{N} (Y_n - a - bX_n)^2.$$

Thus, $\hat{Y}_n = a + bX_n$ is our guess about Y_n given X_n.
The squared error is $(Y_n - \hat{Y}_n)^2$.
The LR minimizes the sum of the squared errors.

Why the squares and not the absolute values?
Main justification: much easier!

Note: This is a non-Bayesian formulation: there is no prior.

Linear Least Squares Estimate

Definition
Given two RVs X and Y with known distribution $Pr[X = x, Y = y]$, the Linear Least Squares Estimate of Y given X is

$$\hat{Y} = a + bX =: L[Y|X]$$

where (a, b) minimize

$$g(a,b) := E[(Y - a - bX)^2].$$

Thus, $\hat{Y} = a + bX$ is our guess about Y given X.
The squared error is $(Y - \hat{Y})^2$.
The LLSE minimizes the expected value of the squared error.

Why the squares and not the absolute values?
Main justification: much easier!

Note: This is a Bayesian formulation:
there is a prior $Pr[X = x, Y = y]$.

LR: Non-Bayesian or Uniform?

Observe that

$$\frac{1}{N}\sum_{n=1}^{N}(Y_n - a - bX_n)^2 = E[(Y - a - bX)^2]$$

where one assumes that

$$(X, Y) = (X_n, Y_n), \text{ w.p. } \frac{1}{N} \text{ for } n = 1, \ldots, N.$$

That is, the non-Bayesian LR is equivalent to the Bayesian LLSE that assumes that (X, Y) is uniform on the set of observed samples.

Thus, we can study the two cases LR and LLSE in one shot.

However, the interpretations are different!

LLSE

Next Time.