Estimation (from Evidence)

Estimating $E[Y]$ minimizes Mean Squared Error: $E[(Y-a)^2]$

$E[(Y-a)^2] = E[Y^2] - 2aE[Y] + a^2$.

Derivative: $-2E[Y] + 2a = 0$

Minimum at $a = E[Y]$.

Joint Distribution: X, Y

Function: $P[X = x, Y = y]$.

View as Table. Row indexed by x, and column indexed by y.
$\sum_x \sum_y Pr[X = x, Y = y]$? 1.

Conditional expectation: $E[Y|X = x] = \sum_y y \times Pr[Y|X = x]$.

Minimize MSE given $X = x$: Guess $E[Y|X = x]$ when $X = x$.

$E[Y|X]$ denotes a function $f(x) = E[Y|X = x]$

Estimation: linear regression.

Given joint distribution for X, Y.

Predict Y using linear function $f(x) = Ax + b$.

Assume $E[X] = E[Y] = 0$. Find best $f(x) = Ax$.

Minimize $E_X[E_Y[(Y - f(x))^2 | X = x]] = E_{X,Y}[(Y - f(X))^2]$
Also: $\sum_x Pr[X = x] \times \sum_y (y - f(x))^2 Pr[Y = y|X = x]$.

$E[Y^2] - 2E[f(X)Y] + E[f(X)^2] = E[Y^2] - 2AE[XY] + A^2E[X^2]$

Derivative: $-2E[XY] + AE[X^2]$. $A = \frac{E[X,Y]}{E[X^2]}$.

Minimizer: $f(x) = \frac{E[XY]}{E[X^2]}x = \frac{Cov(X,Y)}{Var(X)}x$. (for $E[X] = E[Y] = 0$.

Thus: $LLSE[Y|X] = E[Y] + \frac{Cov(X,Y)}{Var(X)}(X - E(X))$.

Estimation Error

We saw that the LLSE of Y given X is

$$L[Y|X] = \hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X]).$$

How good is estimator? What mean squared estimation error?

$$E[|Y - L[Y|X]|^2] = E[(Y - E[Y] - (cov(X,Y)/var(X))(X - E[X]))^2]$$
$$= E[(Y - E[Y])^2] - 2\frac{cov(X,Y)}{var(X)}E[(Y - E[Y])(X - E[X])]$$
$$+ (\frac{cov(X,Y)}{var(X)})^2 E[(X - E[X])^2]$$
$$= var(Y) - \frac{cov(X,Y)^2}{var(X)}.$$
$$= var(Y)(1 - \frac{cov(X,Y)^2}{var(X)var(Y)}$$

Without observations, the estimate is $E[Y]$. The error is $var(Y)$.
Observing X reduces variance by fraction of $corr(X,Y) = \frac{cov(X,Y)^2}{var(X)var(Y)}$.

By a factor $1 - corr(X,Y)$. Correlation Coefficient is fraction of explained variance.

Poll

Correlation coefficient of 1, what is $E[(Y - L[Y|X])^2]$?

If $Y = X$, what is Correlation Coefficient? 1

If $Y = -X$, what is Correlation Coefficient? 1

If $Y = X/2$, what is Correlation Coefficient? 1.

If $Y = X + B(n,p)$, and $X = B(n,p)$, what is Correlation Coefficient?

$1/2$. Half the variance of Y is explained by X.

Estimation Error: A Picture

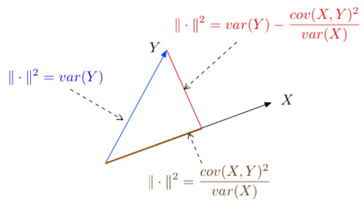
We saw that

$$L[Y|X] = \hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X])$$

and

$$E[|Y - L[Y|X]|^2] = var(Y) - \frac{cov(X,Y)^2}{var(X)}.$$

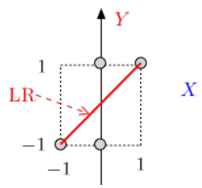
Here is a picture when $E[X] = 0, E[Y] = 0$:
Dimensions correspond to sample points, uniform sample space.



Vector Y at dimension ω is $\frac{1}{\sqrt{\Omega}}Y(\omega)$

Linear Regression Examples

Example:

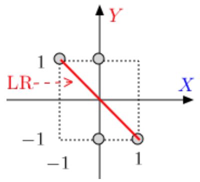


We find:

$E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = 1/2;$
$var[X] = E[X^2] - E[X]^2 = 1/2; cov(X,Y) = E[XY] - E[X]E[Y] = 1/2;$
LR: $\hat{Y} = E[Y] + \frac{cov(X,Y)}{var[X]}(X - E[X]) = X$.

Linear Regression Examples

Example:

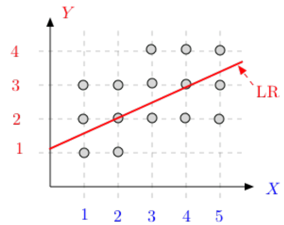


We find:

$E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = -1/2;$
$var[X] = E[X^2] - E[X]^2 = 1/2; cov(X,Y) = E[XY] - E[X]E[Y] = -1/2;$
LR: $\hat{Y} = E[Y] + \frac{cov(X,Y)}{var[X]}(X - E[X]) = -X.$

Linear Regression Examples

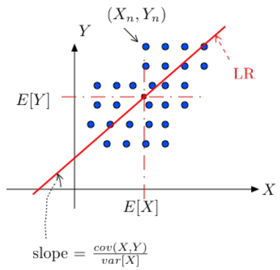
Example:



We find:

$E[X] = 3; E[Y] = 2.5; E[X^2] = (3/15)(1 + 2^2 + 3^2 + 4^2 + 5^2) = 11;$
$E[XY] = (1/15)(1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4;$
$var[X] = 11 - 9 = 2; cov(X,Y) = 8.4 - 3 \times 2.5 = 0.9;$
LR: $\hat{Y} = 2.5 + \frac{0.9}{2}(X - 3) = 1.15 + 0.45X.$

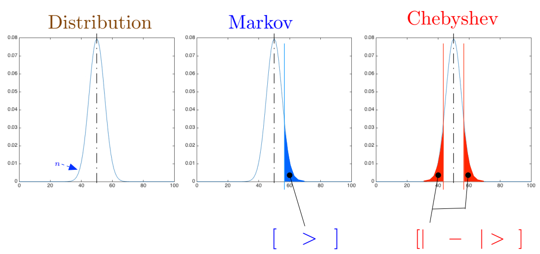
LR: Another Figure



Note that

- the LR line goes through $(E[X], E[Y])$
- its slope is $\frac{cov(X,Y)}{var(X)}$.

Inequalities: An Overview



Andrey Markov

Andrey (Andrei) Andreyevich Markov

Born 14 June 1856 N.S. Ryazan, Russian Empire

Died 20 July 1922 (aged 66) Petrograd, Russian SFSR

Andrey Markov is best known for his work on stochastic processes. A primary subject of his research later became known as Markov chains and Markov processes.

Pafnuty Chebyshev was one of his teachers.

Markov was an atheist. In 1912 he protested Leo Tolstoy's excommunication from the Russian Orthodox Church by requesting his own excommunication. The Church complied with his request.

Lake Woebegone: Poll

What is true?

(A) Everyone is above average (on midterm)
False. Average would be higher.

(B) For a random variable, at most half the people can be more than twice the average.
False. Consider $Pr[X = -2] = 1/3$ and $Pr[X = 1] = 2/3$. $E[X] = 0$.

(C) For the midterm with no negative scores, at most half the people can be more than twice the average.
True. Otherwise average would be higher.

Markov's inequality

The inequality is named for Andrey Markov, though in work by Pafnuty Chebyshev. (Sometimes) called Chebyshev's first inequality.

Theorem Markov's Inequality
Assume $f : \Re \to [0,\infty)$ is nondecreasing. Then,

$$Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)}, \text{ for all } a \text{ such that } f(a) > 0.$$

Proof: First we claim:

$$1\{X \geq a\} \leq \frac{f(X)}{f(a)}.$$

If $X < a$, the inequality reads $0 \leq f(x)/f(a)$, since $f(\cdot) \geq 0$.
If $X \geq a$, it reads $1 \leq f(x)/f(a)$, since $f(\cdot)$ is nondecreasing.

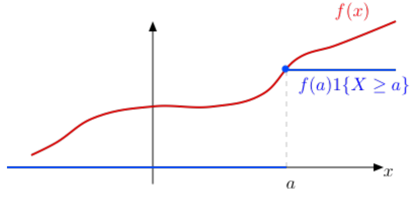
Taking the expectation yields the inequality,
expectation of an indicator is the probability.
and expectation is monotone, e.g., weighted sum of points.

That is, $\sum_v Pr[X = v]1\{v \geq a\} \leq \sum_v Pr[X = v]\frac{f(v)}{f(a)}$.

□

Intuition: $E[f(X)] \geq f(a)Pr[X > a] = f(a)Pr[X > f(a)]$.

A picture



$$f(a)1\{X \geq a\} \leq f(x) \Rightarrow 1\{X \geq a\} \leq \frac{f(X)}{f(a)}$$

$$\Rightarrow Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)}$$

Markov Inequality Example: G(p)

Let $X = G(p)$. Recall that $E[X] = \frac{1}{p}$ and $E[X^2] = \frac{2-p}{p^2}$.

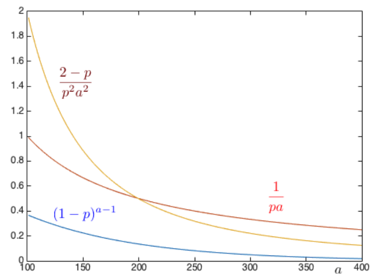
Markov: $Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)}$.

Choose $f(x) = x$, we get

$$Pr[X \geq a] \leq \frac{E[X]}{a} = \frac{1}{ap}.$$

Choose $f(x) = x^2$, we get

$$Pr[X \geq a] \leq \frac{E[X^2]}{a^2} = \frac{2-p}{p^2a^2}.$$



Markov Inequality Example: $P(\lambda)$

Let $X = P(\lambda)$. Recall that $E[X] = \lambda$ and $E[X^2] = \lambda + \lambda^2$.

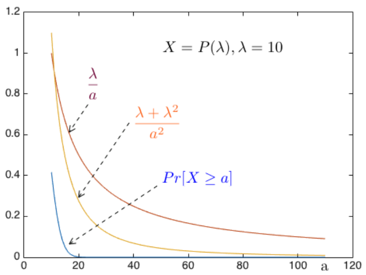
Markov: $Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)}$.

Choosing $f(x) = x$, we get

$$Pr[X \geq a] \leq \frac{E[X]}{a} = \frac{\lambda}{a}.$$

Choosing $f(x) = x^2$, we get

$$Pr[X \geq a] \leq \frac{E[X^2]}{a^2} = \frac{\lambda + \lambda^2}{a^2}.$$



Chebyshev's Inequality

This is Pafnuty's inequality:
Theorem:

$$Pr[|X - E[X]| > a] \leq \frac{var[X]}{a^2}, \text{ for all } a > 0.$$

Proof: Let $Y = |X - E[X]|$ and $f(y) = y^2$. Then,

$$Pr[Y \geq a] \leq \frac{E[f(Y)]}{f(a)} = \frac{var[X]}{a^2}.$$

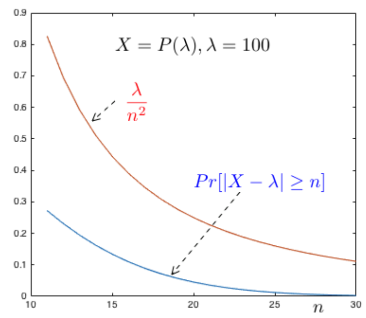
□

This result confirms that the variance measures the "deviations from the mean."

Chebyshev and Poisson

Let $X = P(\lambda)$. Then, $E[X] = \lambda$ and $var[X] = \lambda$. Thus,

$$Pr[|X - \lambda| \geq n] \leq \frac{var[X]}{n^2} = \frac{\lambda}{n^2}.$$



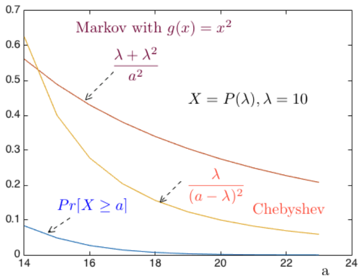
Chebyshev and Poisson (continued)

Let $X = P(\lambda)$. Then, $E[X] = \lambda$ and $var[X] = \lambda$. By Markov's inequality,

$$Pr[X \geq a] \leq \frac{E[X^2]}{a^2} = \frac{\lambda + \lambda^2}{a^2}.$$

Also, if $a > \lambda$, then $X \geq a \Rightarrow X - \lambda \geq a - \lambda > 0 \Rightarrow |X - \lambda| \geq a - \lambda$.

Hence, for $a > \lambda$, $Pr[X \geq a] \leq Pr[|X - \lambda| \geq a - \lambda] \leq \frac{\lambda}{(a-\lambda)^2}$.



Fraction of H's

Here is a classical application of Chebyshev's inequality.

How likely is it that the fraction of H's differs from 50%?

Let $X_m = 1$ if the m-th flip of a fair coin is H and $X_m = 0$ otherwise.

Define

$$Y_n = \frac{X_1 + \cdots + X_n}{n}, \text{ for } n \geq 1.$$

We want to estimate

$$Pr[|Y_n - 0.5| \geq 0.1] = Pr[Y_n \leq 0.4 \text{ or } Y_n \geq 0.6].$$

By Chebyshev,

$$Pr[|Y_n - 0.5| \geq 0.1] \leq \frac{var[Y_n]}{(0.1)^2} = 100 var[Y_n].$$

Now,

$$var[Y_n] = \frac{1}{n^2}(var[X_1] + \cdots + var[X_n]) = \frac{1}{n} var[X_1] \leq \frac{1}{4n}.$$

$$Var(X_i) = p(1-p) \leq (.5)(.5) = \frac{1}{4}$$

Fraction of H's

$$Y_n = \frac{X_1 + \cdots + X_n}{n}, \text{ for } n \geq 1.$$

$$Pr[|Y_n - 0.5| \geq 0.1] \leq \frac{25}{n}.$$

For $n = 1,000$, we find that this probability is less than 2.5%.

As $n \to \infty$, this probability goes to zero.

In fact, for any $\varepsilon > 0$, as $n \to \infty$, the probability that the fraction of Hs is within $\varepsilon > 0$ of 50% approaches 1:

$$Pr[|Y_n - 0.5| \leq \varepsilon] \to 1.$$

This is an example of the Law of Large Numbers.

We look at a general case next.

Weak Law of Large Numbers

Theorem Weak Law of Large Numbers

Let $X_1, X_2, \ldots$ be pairwise independent with the same distribution and mean μ. Then, for all $\varepsilon > 0$,

$$Pr[|\frac{X_1 + \cdots + X_n}{n} - \mu| \geq \varepsilon] \to 0, \text{ as } n \to \infty.$$

Proof:
Let $Y_n = \frac{X_1 + \cdots + X_n}{n}$. Then

$$Pr[|Y_n - \mu| \geq \varepsilon] \leq \frac{var[Y_n]}{\varepsilon^2} = \frac{var[X_1 + \cdots + X_n]}{n^2 \varepsilon^2} = \frac{n var[X_1]}{n^2 \varepsilon^2} = \frac{var[X_1]}{n \varepsilon^2} \to 0, \text{ as } n \to \infty.$$

□

Summary

Variance; Inequalities; WLLN

- Variance: $var[X] := E[(X - E[X])^2] = E[X^2] - E[X]^2$
- Fact: $var[aX + b] = a^2 var[X]$
- Sum: X, Y, Z pairwise ind. $\Rightarrow var[X + Y + Z] = \cdots$
- Markov: $Pr[X \geq a] \leq E[f(X)]/f(a)$ where ...
- Chebyshev: $Pr[|X - E[X]| \geq a] \leq var[X]/a^2$
- WLLN: X_m i.i.d. $\Rightarrow \frac{X_1 + \cdots + X_n}{n} \approx E[X]$

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?

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 mileage 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 and voters, 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, $X_n = B(\theta)$.

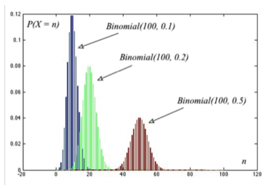
Thus, more precisely, given θ, the random variables X_n are i.i.d. with a known distribution (that depends on θ).

- 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 Candidate A.
  - 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 Candidate A.
- We observe 1,000 heart valve replacements that were performed by Dr. Bill.
  - Among those, 35 patients had to be redone.
  - 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,

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

□

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 $\sigma \leq \frac{1}{2}$.

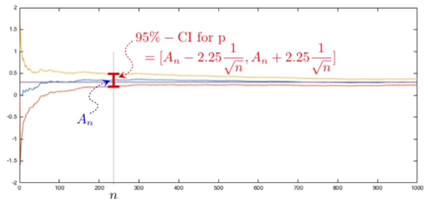
Thus,

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

□

Confidence interval for p in $B(p)$

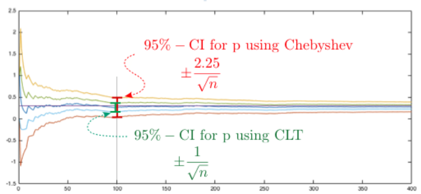
An illustration:



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?

Given coins A and B. Which has higher $Pr[H]$? 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, $Pr[|A_n - B_n - (p_A - p_B)| > \varepsilon] \leq \frac{1}{2n\varepsilon^2}$,

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

or

"p-value" of "$p_A \geq p_B$" $\leq \frac{1}{2n(A_n - B_n)^2}$.

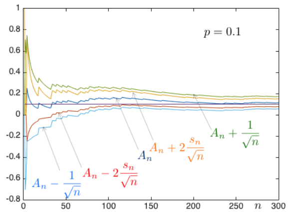
Unknown σ

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

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.