

## Confidence Intervals

Suppose we want to get not only an estimate for some unknown parameter, but also a concrete measure of how good or bad that estimate is.

**Example:** A politician is considering running for office in a district with 50,000 registered voters. He conducts a poll of 2000 ‘randomly selected’ voters, and  $X = 1127$  favor him over his opponent. In other words,

$$\frac{1127}{2000} \times 100\% = 56.35\%$$

voters favor him. Does he win?

### Considerations:

1. The tally  $X$  can be viewed as a binomial random variable, with parameters  $N = 2000$  and unknown  $p$ . Our estimate for  $p$ , based on this measurement, is  $\hat{p} = 0.5635$ .
2. Alternately, each person interviewed can be viewed as a single “Bernoulli” random variable, taking values 1 (favors candidate) or 0 (favors opponent). Then the 200 people sampled are IID random variables from a distribution modelling the district population.

3. Was the sample properly chosen? If the pollster chose the 2000 people carelessly, for example all on a college campus, the sample might not reflect the greater community.
4. Even with a perfect pollster, the *actual* value of  $p$  could be very different from our estimate  $\hat{p}$ . The opponent could even be strongly favored, but by pure chance a sample selected with the majority favoring our candidate.
5. The estimate  $\hat{p}$  is almost certainly at least *a little* wrong.

What we *can* try to do is find a small range of percentages such that the true percentage is in this range with a sufficiently high probability.

Normally, we specify the probability (aka *Confidence Level*) first, then use that to compute the range of percentages. Common choices for the confidence level are 0.90, 0.95, and 0.99 (or 90%, 95%, 99%).

Suppose we specify a confidence level of 95% in this example. We want to use our data to find a “95% confidence interval” for  $p$ .

Want:  $a, b$  such that  $P(a < p < b) = 0.95$

Recall: the CLT applies to the random variable  $X$ . Put

$$Z = \frac{X - Np}{\sqrt{Np(1-p)}}$$

then  $Z$  is approximately standard normal.

From table, we find that if  $Z$  is standard normal,  
 $P(-1.96 < Z < 1.96) = 0.95$

(We sometimes write:  $z_{.025} = 1.96$ )

So:  $0.95 \approx P(-1.96 < \frac{X-Np}{\sigma(X)} < 1.96)$

(where  $\sigma(X)$  =standard deviation of  $X$ )

Equivalently,  $0.95 \approx P(-1.96 < \frac{\hat{p}-p}{\sigma(\hat{p})} < 1.96)$

(where  $\sigma(\hat{p})$  =standard deviation of  $\hat{p}$ )

Solve:

$$0.95 \approx P(\hat{p} - 1.96\sigma(\hat{p}) < p < \hat{p} + 1.96\sigma(\hat{p}))$$

We don't know  $\sigma(\hat{p})$  for the actual (unknown)  $p$ ,  
because

$$\sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{N}}$$

and we don't know  $p$

However, we *do* know  $\hat{p}$ , and so we can approximate  $\sigma(\hat{p})$  by

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

The interval

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}, \quad \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right)$$

is the (approximate) 95% confidence interval  
on  $p$

We sometimes write

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

For our example,  $N = 2000$ ,  $\hat{p} = 0.5335$ , so the interval is

$$\begin{aligned} & 0.5335 \pm 1.96 \sqrt{\frac{(.5335)(.4665)}{2000}} \\ & = 0.5335 \pm (1.96)(0.011) \\ & = 0.5335 \pm .0219 \\ & = (.5116, .5553) \end{aligned}$$

So: we are 95% confident that the voters favor our candidate by between 51.16% and 55.53%

In poll reporting, they normally don't mention the confidence level, and just report something like "This voter is favored by 53.35%, with a margin of error of 2.19 percentage points.

## Another example (without so much algebra)

25 patients with fever are given a new experimental pain reliever; the before and after body temperature is measured, as well as the difference  $X$ . Here are the values:

Before	:	100.6	101.8	101.5	100.3	101.4	101.0	101.4	101.5	100.5	100.9	99.5	101.8	
After	:	98.7	100.4	100.9	100.0	100.4	100.7	99.4	99.5	100.5	100.9	98.7	101.7	
Difference:		1.9	1.4	0.6	0.3	1.0	0.3	2.0	2.0	0.0	0.0	0.8	0.1	
Before	:	99.6	101.3	101.6	101.6	101.5	100.5	101.4	102.7	100.6	100.1	101.3	99.5	101.6
After	:	98.9	100.5	101.4	99.6	101.5	100.5	100.9	102.6	100.2	100.1	101.2	98.9	100.5
Difference:		0.7	0.8	0.2	2.0	0.0	0.0	0.5	0.1	0.4	0.0	0.1	0.6	1.1

Does the pain reliever work?

Assumption:  $x$  has a distribution with unknown mean  $\mu$ , variance  $\sigma^2$ .

To determine: is  $\mu > 0$ ?

Get a confidence interval on  $\mu$ :

We know that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal, and that estimating the  $\sigma$  by  $s$  doesn't change this much.

Arguing as above, a 95% confidence interval for  $\mu$  is

$$\bar{X} \pm 1.96 \frac{s}{\sqrt{n}}$$

Compute:  $\bar{X} = 0.676$ ,  $s = 0.69$

In this case, the 95% confidence interval is

$$0.676 \pm 1.96 \frac{0.69}{\sqrt{25}}$$

or (0.406, 0.946)

So: at the 95% confidence level, there is a drop of temperature of between .4 and .95 degrees.

What if we want to be more confident? Say, 99%?

The only difference is that the 1.96 needs to be changed to  $z_{0.01/2} = z_{0.005}$ , which equals 2.57. Then the interval is

$$0.676 \pm 2.57 \frac{0.69}{\sqrt{25}} = (0.321, 1.03)$$

**Question:** What if the question was, “Does the pain reliever drop temp by at least half a degree?”

## **A final application of probability: DNA testing**

### **Statements from the news:**

“The stain on White House intern Monica Lewinsky’s dress was tested for DNA. Only 1 in 8 trillion people have this DNA profile.”

“The most astronomical figures involved a pair of socks found near Simpson’s bed. Cotton said one sock contained the DNA type of Simpson’s slain ex-wife, Nicole Brown Simpson. Asked how many other whites shared that DNA type, Cotton said one in 9.7 billion. Prosecutor George Clark noted the figure was larger than the Earth’s population, estimated at 5.5 billion, meaning that Ms. Simpson was literally the only person whose blood could be on that sock.” (*USA Today*, 10-18-96)

What do such statements mean? Can we really use them to determine guilt from a DNA match?

**Overview of DNA testing:** DNA, Loci, markers (VNTR, STR, etc), profile

**Procedure:** Choose sites.

Estimate probability distributions at sites based on broad data (blood banks, etc)

Assumption: sites distant enough so measurements are independent.

Use independence (product rule) to determine probability of any given profile.

**Example.** If specify 4 sites, and each has 100 equally probable allele values, then probability of any given profile is  $(1/100)^4$ , or one in 100 million. If specify 10 sites, each with 20 equally probably values, then probability of any given profile is  $(1/20)^{10}$ , or one in  $10^{26}$ .

**Note:** The probabilities for any locus are actually confidence intervals, so you might get a range of values, eg between  $.009^4$  and  $.011^4$  (or between one in 68, 301, 346 and one in 152, 415, 790).

## The “Prosecutor’s Fallacy”

A DNA sample from a crime is typed, and profile computed to have a one in one million probability. Then DNA from the defendant is typed, and has this same profile. Consider two statements:

- (1) “There is only a 1 in a million chance the defendant is innocent”
- (2) “The probability of obtaining this DNA profile from a randomly selected individual is 1 in a million.”

These are *not* the same!

Statement (2) is correct (if our other estimates and assumptions are correct). Statement (1) is false.

Actual computation requires *conditional probability*:

$P(A|B)$  = The probability of A given that you know B is true

Statement (1) is really the assertion that

$P(\text{Defendant innocent} \mid \text{blood profile matched}) = 1/1000000$

but we don’t yet know this conditional probability.

**Bayes Theorem** If  $A, B$  are events, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Let:

$A$  =Defendant is innocent

$A^c$  =Defendant is guilty

$B$  =Defendant's blood matches the crime scene profile

and let  $p = P(A^c)$  (which we don't know, but might have a preconception about). We know that  $P(B|A) = 1/1000000$ ,  $P(B|A^c) = 1$ , so plug into Bayes Theorem, get:

$$\begin{aligned} P(A|B) &= \frac{(1/1000000)(1-p)}{(1/1000000)(1-p) + p} \\ &= \frac{1-p}{1+99999p} \end{aligned}$$

this varies depending on what is the 'prior' probability  $p$  of guilt: