Introduction to Statistical Inference*

Inference is about Parameters (Populations) or general mechanisms -- or future observations. It is not about data (samples) *per se*, although it uses data from samples. Might think of inference as statements about a universe most of which one did not observe, or has not yet observed .

Two main schools or approaches:

Bayesian [not even mentioned by Fletcher]

- Makes direct statements about parameters and future observations
- Uses previous impressions plus new data to update impressions about parameter(s)

e.g.
Everyday life
Medical tests: Pre- and post-test impressions

Frequentist

- Makes statements about observed data (or statistics from data) (used indirectly [but often incorrectly] to assess evidence against certain values of parameter)
- Does not use previous impressions or data outside of current study (meta-analysis is changing this)

  e.g.

  • Statistical Quality Control procedures [for Decisions]
  • Sample survey organizations: Confidence intervals
  • Statistical Tests of Hypotheses

  Unlike Bayesian inference, there is no quantified pre-test or pre-data "impression"; the ultimate statements are about data, conditional on an assumed null or other hypothesis.

  Thus, an explanation of a p-value must start with the conditional "IF the parameter is ... the probability that the data would ..."

Book "Statistical Inference" by Michael W. Oakes is an excellent introduction to this topic and the limitations of frequentist inference.

(Frequentist) Confidence Interval (CI) or Interval Estimate for parameter

Formal definition:

A level 1 - α Confidence Interval for a parameter θ is given by two statistics (i.e.. numbers calculated from data)

$$U_{pper} \text{ and } L_{ower}$$

such that when θ is the true value of the parameter,

$$\text{Prob}\,(L_{ower} \le \theta \le U_{pper}) = 1 - \alpha$$

α	$1 - \alpha$
0.05	**0.95**
0.01	**0.99**

- CI is a **statistic**: a quantity calculated from a sample
- usually use α = 0.01 or 0.05 or 0.10, so that the "level of confidence", 1 - α, is 99% or 95% or 90%. We will also use "α" ("alpha") for tests of significance (there is a direct correspondence between confidence intervals and tests of significance)
- technically, we should say that we are **using a procedure which is guaranteed to cover the true q in a fraction 1 - α of applications**. If we were not fussy about the semantics, we might say that any particular CI has a 1 - α chance of covering θ.
- for a given amount of sample data] the narrower the interval from L to U, the lower the degree of confidence in the interval and vice versa.

Large-sample CI's, based on Standard Error (SE) of statistic

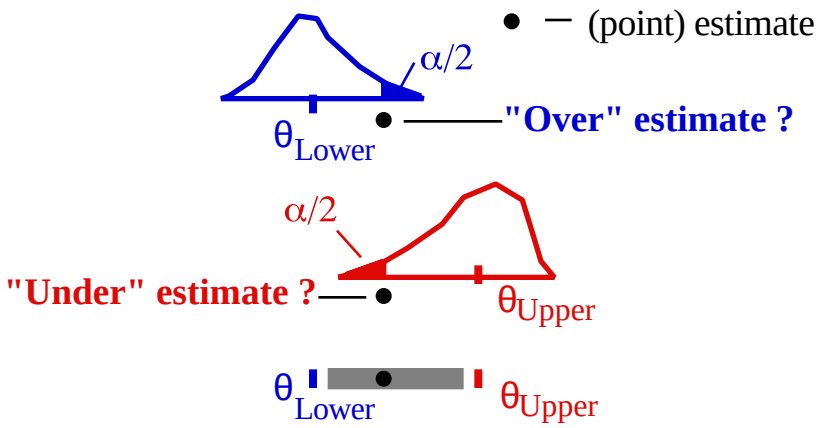
Many large-sample CI's are of the form (hat ^ denotes 'estimate of')

i $\hat{\theta} \pm$ multiple of SE($\hat{\theta}$) or

ii inverse fn. of [fn{$\hat{\theta}$} ± multiple of SE(f{$\hat{\theta}$}]. where fn. is some function of $\hat{\theta}$ which has close to a Gaussian distribution.
e.g. $\hat{\theta}$ = odds or rate ratio; fn. = ***ln*** (natural log) ; inv. fn. = exp.

- 'Multiple' based on desired level of 'confidence'
  e.g. 1.645 for 90% confidence, 1.96 for 95% confidence.
- Standard error (SE) is a function of amount of information on which estimate is based (the more the information, the smaller the SE).
- the '1.645 × SE ' or '1.96 × SE ' called the '**margin of error**'

Method of Constructing a 100(1 - α)% CI (in general):



SE's for "Large Sample" CI's for parameters, and DIFFERENCES thereof [σ : standard deviation (SD) of individuals]

parameter		estimate	SE[estimate]
	θ	$\hat{\theta}$	$SE[\hat{\theta}]$
mean	μ_y	$\bar{y}$	$\frac{\sigma_y}{\sqrt{n}}$
prop.	π	p	$\frac{\sqrt{\pi[1-\pi]}}{\sqrt{n}}$
	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_1$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	$\pi_1 - \pi_2$	$p_1 - p_2$	$\sqrt{\frac{\pi_1[1-\pi_1]}{n_1} + \frac{\pi_2[1-\pi_2]}{n_2}}$

prop. = proportion

In SE, estimated values substituted for unknown ones (in Greek)

EXAMPLES: See exercise on Birthweights and Adult Heights

"Large Sample" CI for Odds Ratio

parameter — data and odds ratio (or rate ratio) estimate

(denominators: full [cohort] or samples ['controls'])

	Exposed(1)		**Not(0)**
Odds Ratio	$\frac{\#\text{cases}}{\text{'denominator'}}$	/	$\frac{\#\text{cases}}{\text{'denominator'}}$

SE[log odds ratio]

$$= \sqrt{\frac{1}{\#\text{exposed cases}} + \frac{1}{\#\text{unexposed cases}} + \frac{1}{\text{exposed 'denomimator'}} + \frac{1}{\text{unexposed 'denominator'}}}$$

EXAMPLE (Kim 2002): No. of CASES of nasal polyposis (**numerators**) among people who live in houses heated by ...

Woodstove ?

Yes (1)	No (0)	
45	10	55 (**CASE** series)

No. of **sampled** (same age-sex) **people** who live in houses heated by...

Woodstove ?

Yes (1)	No (0)	
14	41	55 ('**denominator**' series, **'CONTROLS'**)

Quasi-**rates** in people who live in houses heated with...

Woodstove

Yes (1)	No (0)		MARGIN OF ERROR i.e. multiplier and divisor to be applied to point estimate (i.e. to observed ratio)	95% CI
			$\exp[1.96 \times \sqrt{\frac{1}{45} + \frac{1}{10} + \frac{1}{14} + \frac{1}{41}}\ ]$	13.1 ÷ 2.5 to 13.1 × 2.5
$\frac{41}{14}$	$\frac{10}{41}$	**13.1**	= 2.5	**5.2 to 32.8**

NOTE: If denominator ***much larger*** than # cases (as in cohort study), SE of log odds ratio dominated by # exposed **cases** and # unexposed **cases**. (Control:case ratio of 4 => SE sqrt[1/1 + 1/4] = 1.12 × sqrt[1/1 + 1/ ∞]).

"Large Sample" CI for Rate Ratio AUTISM & MMR vaccinations

No. of CASES of autism (**numerators**) among children who did / did not receive MMR vaccination ... Danish Cohort Study, NEJM Nov 7, 2002

Vaccinated		
Yes (1)	No (0)	
263	53	316 (**CASES**)

No. of **children-years** (c-y) of follow-up [contributed by 0.54 m children]

Vaccinated		
Yes (1)	No (0)	
1.65m	0.48m	2.13m children-years (c-y) (**DENOMINATORS**)

CRUDE Rates ...

Vaccinated				
Yes (1)	No (0)	**rate ratio**	margin of error i.e. multiplier and divisor to be applied to point estimate (i.e. to observed rate ratio)	95% CI*
			$\exp[1.96 \times \sqrt{\frac{1}{263} + \frac{1}{53}}$	1.44 ÷ 1.34 to 1.44 × 1.34
$\frac{263}{1.65m}$	$\frac{53}{0.48m}$	**1.44** *	= 1.34 †	**1.07 to 1.93**

† SE of log rate ratio determined by numbers of **cases.**

*** Note the big difference between the crude (1.44) and adjusted ratio (reason why discussed next). For this reason, I am calculating the CI around the crude ratio in this example simply for didactic purposes. The adjusted ratio was 0.92, 95% CI 0.68 to 1.24 (i.e., 0.92 × ÷ 1.35)**

Rate ratio: "crude"= 1.44; Adjusted (cf. article) = 0.92. WHY ?

"We calculated the relative risk with adjustment for age, calendar period, sex, birthweight, gestational age, mother's education, and socio-economic status"

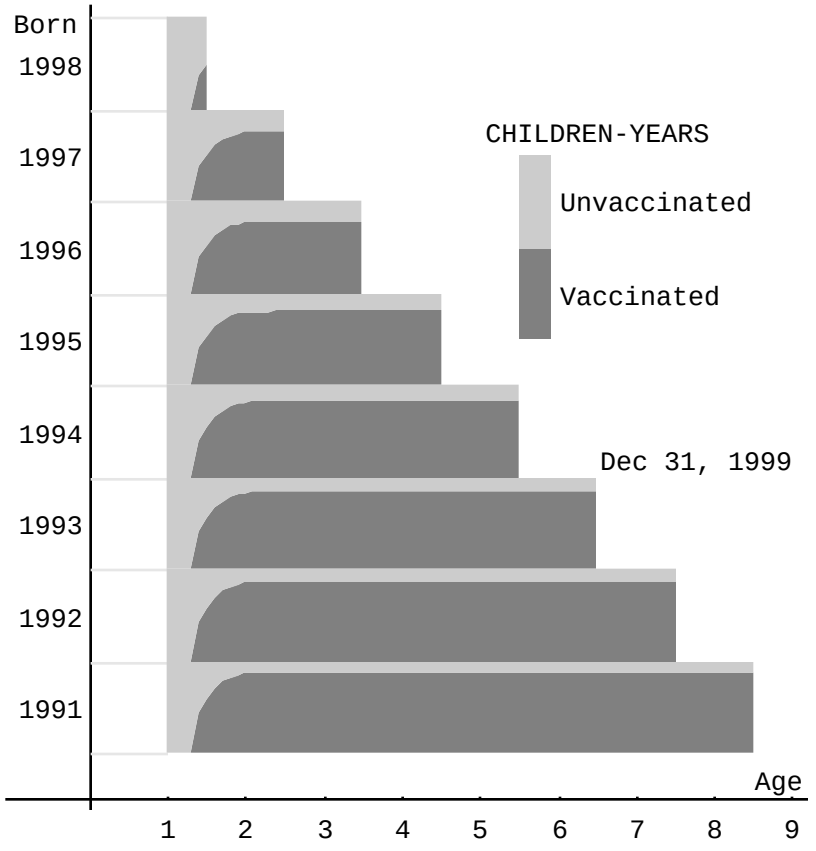
(p 1479, 2nd column, 6 lines from end)

"Except for age, none of these possible confounders changed the estimates. The confounding by age was a function of the time available for follow-up, since much of the follow-up for the unvaccinated group involved young children, in whom autism is often unnoticed"

(p 1481, 2nd column, end 1st paragraph)

Note[JH]: cf. footnote regarding missing gestational ages, Table 1.

(Schematic to help visualize the confounding .. constructed by JH}



Pair-matched Case-Control Study

(e.g. 1st article used in small group session 1, and again in session 2)

The formula on page 1 is for an ***unmatched*** analysis. For a ***matched-pair*** case control analysis, with 1 denoting exposed and 0 unexposed,

$$\text{odds ratio (i.e., point estimate)} = \frac{\#\{\text{case}_1 ; \text{control}_0\} \text{ pairs}}{\#\{\text{case}_0 ; \text{control}_1\} \text{ pairs}}$$

SE[log odds ratio]

$$= \sqrt{\frac{1}{\#\{\text{case}_1 ; \text{control}_0\} \text{ pairs}} + \frac{1}{\#\{\text{case}_0 ; \text{control}_1\} \text{ pairs}}}$$

The (many) ways to (in)correctly describe a CI

Below are my annotated answers to some graduate students' interpretations of a CI

Question: A New York Times poll on women's issues interviewed 1025 women and 472 men randomly selected from the United States excluding Alaska and Hawaii. The poll found that 47% of the women said they do not get enough time for themselves.

(a) The poll announced a margin of error of ±3 percentage points for 95% confidence in conclusions about women. Explain to someone who knows no statistics why we can't just say that 47% of all adult women do not get enough time for themselves.

(b) Then explain clearly what "95% confidence" means.

(c) The margin of error for results concerning men was ± 4 percentage points. Why is this larger than the margin of error for women?

1	True value will be between 43 & 50% in 95% of repeated samples of same size.	• **No**. Estimate will be between μ – margin & μ + margin in 95% of applications.
2	Pollsters say their survey method has 95% chance of producing a range of percentages that includes π.	• **Good**. Emphasize average <u>performance</u> in repeated <u>applications</u> of method.
3	If this same poll were repeated many times, then 95 of every 100 such polls would give a range that included 47%.	• **No!**. See 1.
4	You're pretty sure that the true percentage π is within 3% of 47% . "95% confidence" means that 95% of the time, a random poll of this size will produce results within 3% of π.	• Bayesians would object (and rightly so!) to this use of the "true parameter" as the <u>subject</u> of the sentence. They would insist you use the statistic as the subject of the sentence and the parameter as object.
5	If took 100 different samples, in 95% of cases, the sample proportion will be between 44% and 50%.	• **NO**! The sample proportion will be between truth – 3% & truth + 3% in 95% of them.
6	With this one sample taken, we are sure 95 times out of 100 that 41-53% of the women surveyed do not get enough time for themselves.	• **NO**. 95/100 times the estimate will be within 3% of π, i.e., estimate will be in interval π – margin to π + margin. Method used gives correct results 95% of time.
7	In 95 of 100 comparable polls, expect 44 - 50% of women will give the same answer. Given a parameter, we are 95% sure that the mean of this parameter falls in a certain interval.	• **NO**. Same answer? as what? **Not** given a parameter (ever) . If we were, wouldn't need this course! Mean of a parameter makes no sense in frequentist inference.
8	"using the poll procedure in which the CI or rather the true % is within +/- 3, you cover the true percentage 95% of times it is applied.	• A bit **muddled**... but "correct in 95% of applications" is accurate.
9	Confident that a poll (such) as this one would have measured correctly that the true proportion lies between in 95% .	• **???** [I have trouble parsing this!] In 95% of applications/uses, polls like these come within ± 3% of truth.
10	95% chance that the info is correct for between 44 and 50% of women.	• **???** 95% confidence in the procedure that produced the interval 44-50
11	95% confidence -> 95% of time the proportion given is the good proportion (if we interviewed other groups).	• "Correct in 95% of applications" **Good** to connect the 95% with the long run, not specifically with this one estimate. Always ask yourself: what do I mean by "95% of <u>the time</u>" ? If you substitute "applications" for "time", it becomes clearer.
12	It means that 47% give or take 3% is an accurate estimate of the population mean 19 times out of 20 such samplings. "This result is trustworthy 19 times out of 20" "this poll" : see COMMENT below<----	• **???** 95% of applications of CI give correct answer. How can the <u>same</u> interval 47%±3 be accurate in 19 but not in the other 1? • **???** "<u>this</u>" result: Cf. the distinction between "<u>my</u> operation is successful 19 times out of 20 … " and "operations <u>like the one to be done on me</u> are successful 19 times out of 20"

COMMENT: Polling companies who say "polls <u>of this size are</u> accurate to within so many percentage points 19 times out of 20" are being statistically correct -- they emphasize the <u>procedure</u> rather than what has happened in this specific instance. Polling companies (or reporters) who say "<u>this</u> poll is accurate .. 19 times out of 20" are talking statistical nonsense -- <u>this specific poll</u> is either "right" or "wrong"!. On average <u>19 polls out of 20 are "correct "</u>. But *this* poll cannot be right on average 19 times out of 20!

Even more ways to (in)correctly describe a CI

The Gallup Poll asked 1571 adults what they considered to be the most serious problem facing the nation's public schools; 30% said drugs. This sample percent is an estimate of the percent of all adults who think that drugs are the schools' most serious problem. The news article reporting the poll result adds, "The poll has a margin of error -- the measure of its statistical accuracy -- of three percentage points in either direction; aside from this imprecision inherent in using a sample to represent the whole, such practical factors as the wording of questions can affect how closely a poll reflects the opinion of the public in general" (The New York Times, August 31, 1987). The Gallup Poll uses a complex multistage sample design, but the sample percent has approximately a normal distribution. Moreover, it is standard practice to announce the margin of error for a 95% confidence interval unless a different confidence level is stated.

a The announced poll result was 30%±3%. Can we be certain that the true population percent falls in this interval? -->

b Explain to someone who knows no statistics what the announced result 30%±3% means. **ANNOTATED ANSWERS next column...** -->

c Does the announced margin of error include errors due to practical problems such as undercoverage and nonresponse? **ANSWER: NO!**

Meta-analysis

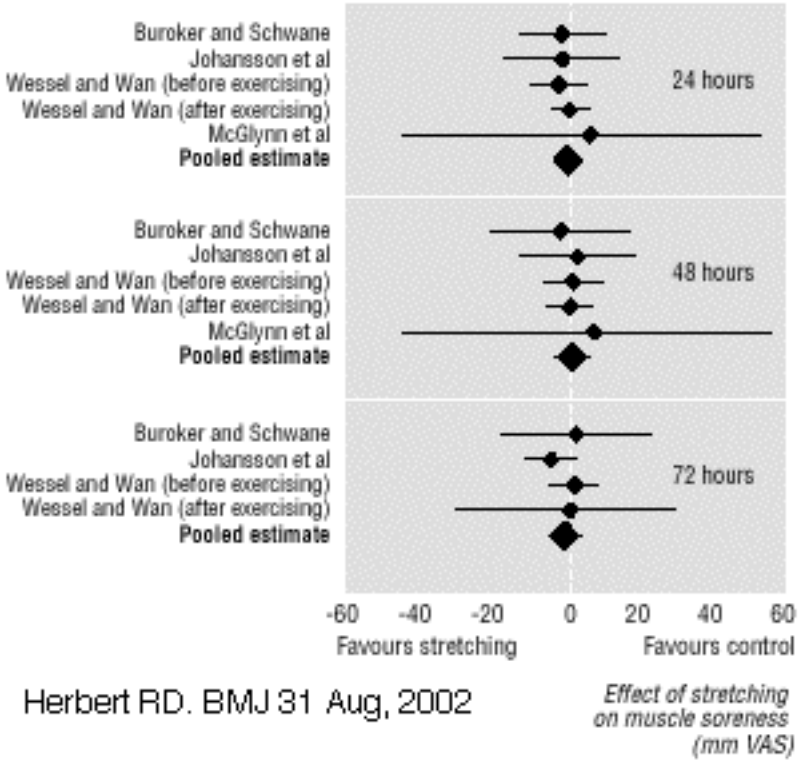


Fig 1 Effects of stretching on delayed onset muscle soreness at 24 hours, 48 hours, and 72 hours after exercise. (VAS=visual analogue scale) [15–17 19]

1	This means that the population result will be between 27% and 33% 19/20 times.	• **NO**! **Population result is wherever it is and it doesn't move**. Think of it as if it were the speed of light.
2	95% of the time the actual truth will be between 30 ± 3% and 5% it will be false.	• It either is or it isn't … the truth doesn't vary over samplings.
3	If this poll were repeated very many times, then 95 of 100 intervals would include 30% .	• **NO**. 95% of polls give answer within 3% of truth, NOT within 3% of the mean in this sample.
4	Interval of true values ranges b/w 27% + 33%.	• **???** There is only one true value. AND, it isn't 'going' or 'ranging' or 'moving' anywhere!
5	Confident that in repeated samples estimate would fall in this range 95/100 times.	• **NO**. Estimate falls within 3% of π in 95% of applications
6	95% of intervals will contain true parameter value and 5% will not. Cannot know whether result of applying a CI to a particular set of data is correct.	• **GOOD**. Say "Cannot know whether CI derived from a particular set of data is correct." ***Know about behaviour of procedure!*** If not from Mars, (i.e. if you use past info) might be able to bet more intelligently on whether it does or not.
7	In 1/20 times, the question will yield answers that do not fall into this interval.	• **No**. In 5% of applications, estimate will be more than 3% away from true answer. See 1,2,3 above.
8	This type of poll will give an estimate of 27 to 33% 19 times out of 20 times.	• **NO**. Won't give 27 ± 3 19/20 times. Estimate will be within ± 3 of truth in 19/20 applications
9	5% risk that μ is not in this interval.	• ??? If an after the fact statement, somewhat inaccurate.
10	95 / 100 times if do the calculations, result 27-33% would appear.	• **No it wouldn't**. See 1,2,3,7.
11	95% prob computed interval will cover parameter.	• Accurate if viewed as a prediction.
12	The true popl'n mean will fall within the interval 27-33 in 95% of samples drawn.	• **NO**. True popl'n mean will not "fall" anywhere. It's a fixed, unknowable constant. Estimates may fall around it.

1200 are hardly representative of 80 million homes /220 million people!!

The Nielsen system for TV ratings in U.S.A.
(Excerpt from article on "Pollsters" from an airline magazine)

"...Nielsen uses a device that, at one minute intervals, checks to see if the TV set is on or off and to which channel it is tuned. That information is periodically retrieved via a special telephone line and fed into the Nielsen computer center in Dunedin, Florida.

With these two samplings, Nielsen can provide a statistical estimate of the number of homes tuned in to a given program. A rating of 20, for instance, means that 20 percent, or 16 million of the 80 million households, were tuned in.

To answer the criticism that 1,200 or 1,500 are hardly representative of 80 million homes or 220 million people, Nielsen offers this analogy:

Mix together 70,000 white beans and 30,000 red beans and then scoop out a sample of 1000. the mathematical odds are that the number of red beans will be between 270 and 330 or 27 to 33 percent of the sample, which translates to a "rating" of 30, plus or minus three, with a 20-to-1 assurance of statistical reliability. The basic statistical law wouldn't change even if the sampling came from 80 million beans rather than just 100,000." ...

Why, if the U.S. has a 10 times bigger population than Canada, do pollsters use the same size samples of approximately 1,000 in both countries?

Answer : it depends on WHAT IS IT THAT IS BEING ESTIMATED. With n=1,000, the SE or uncertainty of an estimated PROPORTION 0.30 is indeed 0.03 or 3 percentage points. However, if interested in the NUMBER of households tuned in to a given program, the best estimate is 0.3N, where N is the number of units in the population (N=80 million in the U.S. or N=8 million in Canada). The uncertainty in the 'blown up' estimate of the TOTAL NUMBER tuned in is blown up accordingly, so that e.g. the estimated NUMBER of households is

```
U.S.A. 80,000,000[0.3 ± 0.03] = 24,000,000 ± 2,400,000
Canada. 8,000,000[0.3 ± 0.03] =  2,400,000 ±   240,000
```

2.4 million is a 10 times bigger absolute uncertainty than 240,000. Our intuition about needing a bigger sample for a bigger universe probably stems from absolute errors rather than relative ones (which in our case remain at 0.03 in 0.3 or 240,000 in 2.4 million or 2.4 million in 24 million i.e. at 10% irrespective of the size of the universe.

It may help to think of why we do not take bigger blood samples from bigger persons: the reason is that we are usually interested in concentrations rather than in absolute amounts and that concentrations are like proportions.

The "Margin of Error blurb" introduced (legislated) in the mid 1980's

Montreal Gazette August 8, **1981**
NUMBER OF SMOKERS RISES BY FOUR POINTS: GALLUP POLL

Compared with a year ago, there appears to be an increase in the number of Canadians who smoked cigarettes in the past week - up from 41% in 1980 to 45% today. The question asked over the past few years was: **"Have you yourself smoked any cigarettes in the past week?"** Here is the national trend:

Year	'74	'75	'76	'77	'78	'79	'80	'81
Smoked cigarettes in past week (%)	52	47	??	45	47	44	41	45

Today's results are based on 1,054 personal in-home interviews with adults, 18 years and over, conducted in June.

The Gazette, Montreal, Thursday, June 27, **1985**
39% OF CANADIANS SMOKED IN PAST WEEK: GALLUP POLL

Almost two in every five Canadian adults (39 per cent) smoked at least one cigarette in the past week - down significantly from the 47 percent who reported this 10 years ago, but at the same level found a year ago. Here is the question asked fairly regularly over the past decade: "**Have you yourself smoked any cigarettes in the past week?**" The national trend shows:

Year	'75	'76	'77	'78	'79	'80	'81	'82	'83	'84	'85
Smoked cigarettes in past week (%)	47	??	45	47	44	41	45	42 ^^	41	39	39

^^ Smoked regularly or occasionally? [JH: larger n won't reduce 'non-sampling' variation]

Results are based on 1,047 personal, in-home interviews with adults, 18 years and over, conducted between May 9 and 11. ***A sample of this size is accurate within a 4-percentage-point margin, 19 in 20 times.***

La Presse, Montréal, 1993
95%CI? IC? ... Comment dit on... ?

L'Institut Gallup a demandé récemment à un échantillon représentatif de la population canadienne d'évaluer la manière dont le gouvernement fédéral faisait face à divers problèmes économiques et général. Pour 59 pour cent des répondants, les libéraux n'accomplissent pas un travail efficace dans ce domaine, tandis que 30 pour cent se déclarent de l'avis contraire et que onze pour cent ne formulent aucune opinion.

La même question a été posée par Gallup à 16 reprises entre 1973 et 1990, et ne n'est qu'une seule fois, en 1973, que la proportion des Canadiens qui se disaient insatisfaits de la façon dont le gouvernement gérait l'économie a été inférieure à 50 pour cent.

Les conclusions du sondage se fondent sur **1009** interviews effectuées entre le 2 et le 9 mai 1994 auprès de Canadiens âgés de 18 ans et plus. **Un échantillon de cette ampleur donne des résultats exacts à 3,1 p.c., près dans 19 cas sur 20.** La marge d'erreur est plus forte pour les régions, par suite de l'importance moidre de l'échantillonnage; par exemple, les **272** interviews effectuées au Québec ont engendré une marge d'erreur de **6 p.c**. dans **19 cas sur 20**.