

## Learning Objectives

- o Understand the different *summary numbers* used to describe the *pattern of variation* in a characteristic or measurement --- from individual to individual, or from one measurement to another of the same individual -- and be able to identify which summaries are more appropriate in which circumstances ⇒ **“Descriptive Statistics.”**
- o Understand (i) the concept of a *Margin of Error*, used to quantify by how much (say) a mean level, or a proportion, observed in a *sample* of individuals might -- just because of *sampling variation* -- be an under-, or an over-estimate, of the level/proportion of interest (ii) the factors that affect this Margin of Error, and (iii) how it is used to construct a **“Confidence Interval.”**
- o Understand the concepts of, and the proper interpretation of, (statistical) **“P-value”**; **“Statistically significant”**; **test of hypothesis**; **“statistical power.”**

# 1. Example of **Descriptive Statistics**

---

Extended Work Shifts and the Risk of Motor Vehicle Crashes among Interns  
NEJM 2005;352;125-134

## **Prospective nationwide, Web-based survey**

2737 residents in 1st postgraduate year (interns) completed monthly reports:-

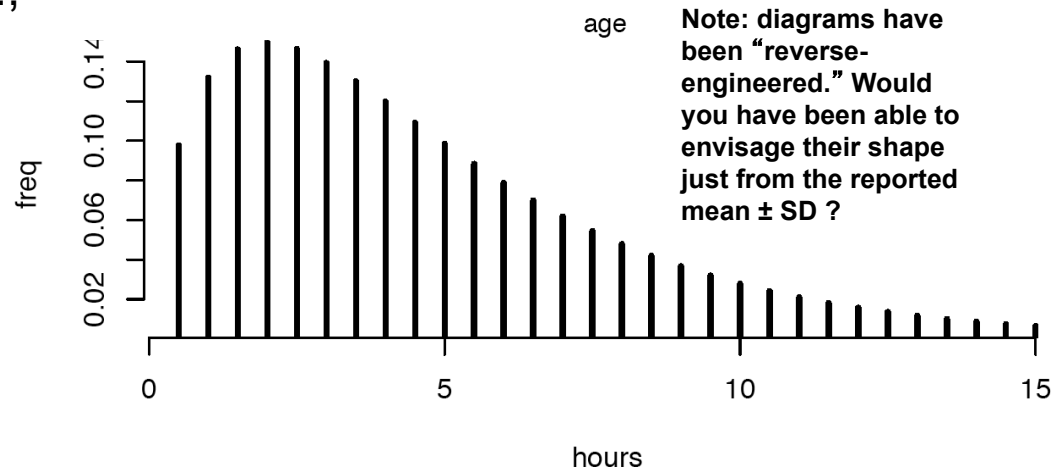
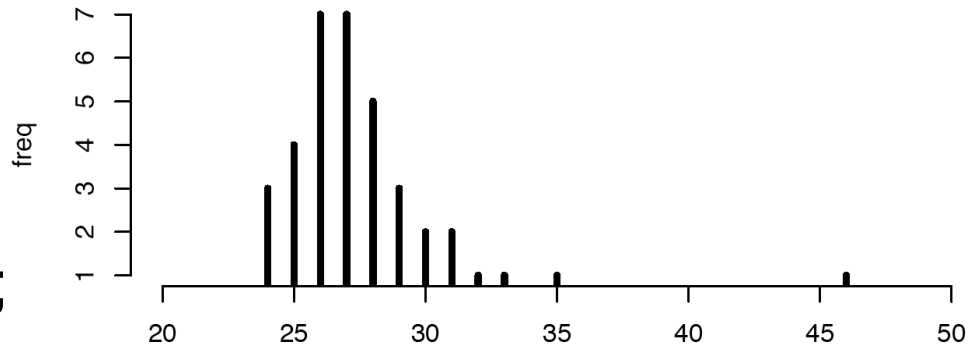
### **Detailed information about...**

- work hours,
- work shifts of an extended duration,
- documented motor vehicle crashes,
- near-miss incidents,
- incidents involving involuntary sleeping.

# Demographic data

Similar to all interns matched through National Resident Matching Program in 2002.

**53% female,**  
**mean age:  $28.0 \pm 3.9$  years;**  
**79% med, 11 % surg, 10% other/ns.;**  
**69% commuted by car,**  
**average weekly commute:**  
 **$91.6 \pm 96.2$  miles,**  
 **$4.4 \pm 3.4$  hours.**



“data reported as means  $\pm$  SD.” [see note]

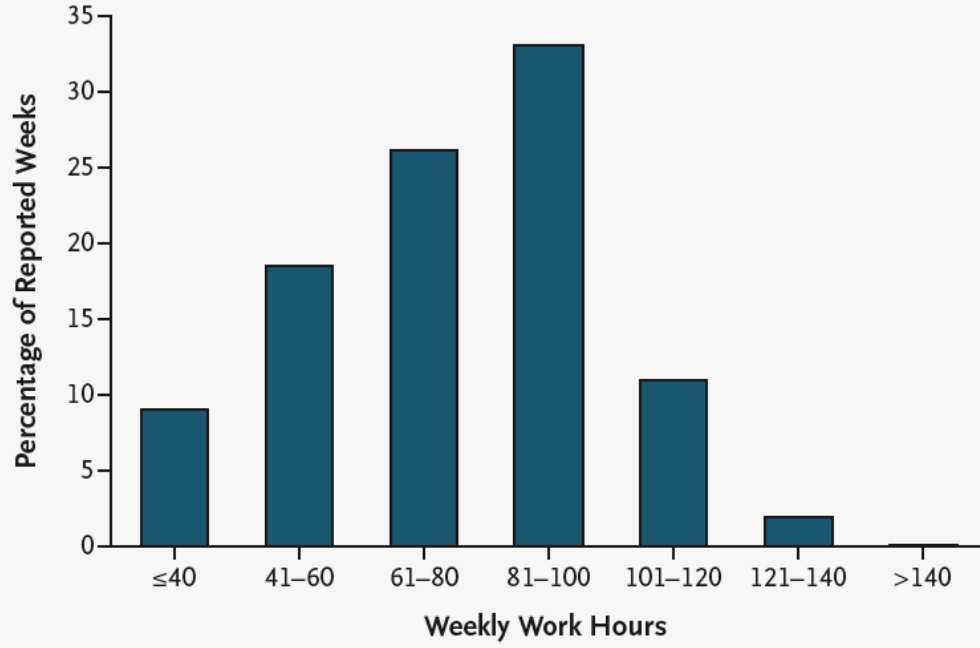
**Mean  $\pm$  SD not very descriptive if non-Gaussian\* pattern:**

**Use Median, and 1st & 3rd Quartiles: 1/4th & 3/4ths**

(\* Gaussian  $\Leftrightarrow$  “Normal”  $\Leftrightarrow$  “Bell Curve”: 68%, 95%, 99.7% within 1, 2, 3 SD's of mean)



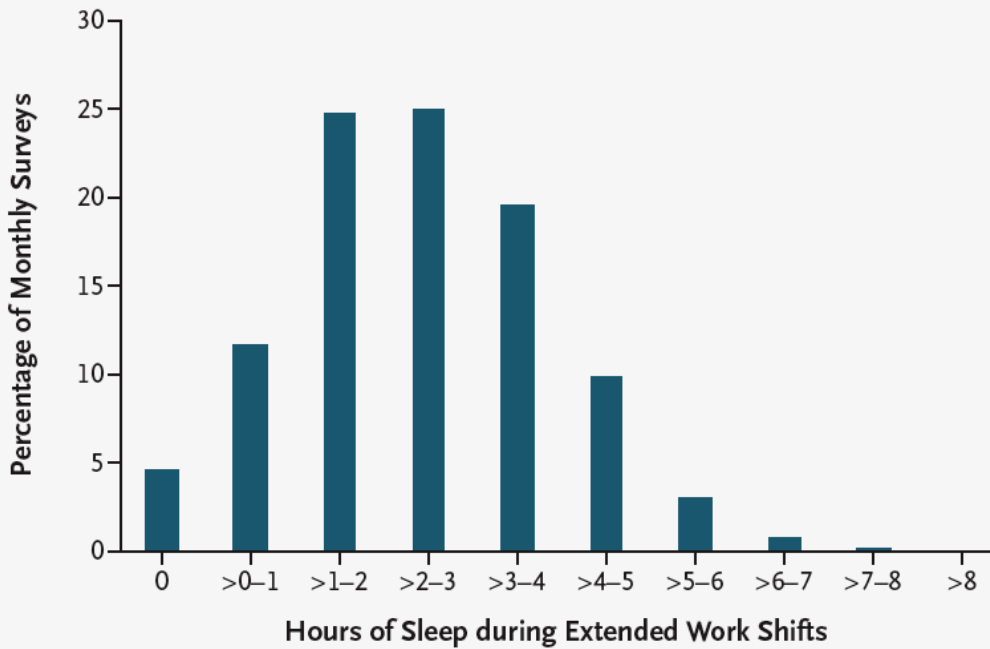
A



Interns averaged  $70.7 \pm 26.0$  hours in the hospital weekly;

they were awake  $67.4 \pm 24.4$  of those hours.

The mean monthly number of extended work shifts that were reported was  $3.9 \pm 3.4$ , with an average duration of  $32.0 \pm 3.7$  hours.



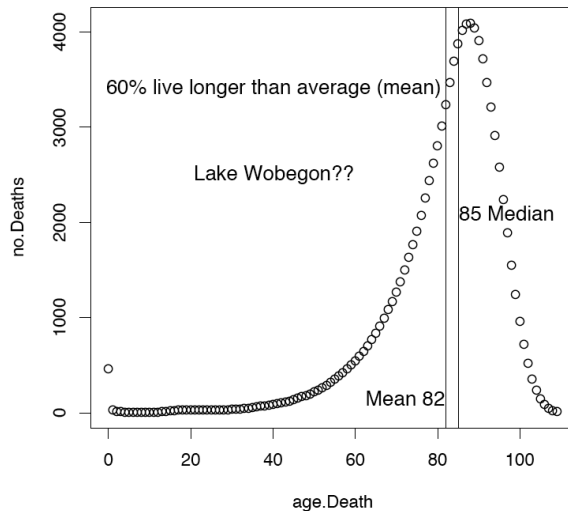
We will see data on crashes a few slides later.

# Descriptive Statistics

(usually in Table 1 of article)

## Central tendency

- Mean (centre of gravity)
- Median (50th percentile)
- Mode (most popular value)



## Dispersion/scatter

- Standard Deviation\*
- Percentiles, e.g.
  - 25th ( $Q_1$ ) & 75th ( $Q_2$ )
  - 5th & 95th..

[ Q = “Quartile” ]

\*Think of Standard Deviation as (approx.) the average of the absolute deviations from the mean.

# Descriptive Statistics

(their use here is descriptive, & to show comparison is “fair”)

A CONTROLLED TRIAL OF A HUMAN PAPILLOMAVIRUS TYPE 16 VACCINE *N Engl J Med* 2002;347:1645-51.

**Background:** Approximately 20 percent of adults become infected with HPV-16. Although most infections are benign, some progress to anogenital cancer. A vaccine that reduces the incidence of HPV-16 infection may provide important public health benefits.

**Methods:** In this double-blind study, we randomly assigned 2392 females\* (16 to 23 years of age) to receive three doses of placebo or HPV-16 virus-like-particle vaccine (40 µg per dose), given at day 0, month 2, and month 6. Genital samples to test for HPV-16 DNA were obtained at enrollment, one mo. after the third vaccination, and every six mo. thereafter. Women were referred for colposcopy according to a protocol. Biopsy tissue was evaluated for cervical intraepithelial neoplasia and analyzed for HPV-16 DNA with use of the polymerase chain reaction. The primary end point was persistent HPV-16 infection, defined as the detection of HPV-16 DNA in samples obtained at two or more visits. The primary analysis was limited to women who were negative for HPV-16 DNA and HPV-16 antibodies at enrollment and HPV-16 DNA at month 7.

TABLE 2. CHARACTERISTICS OF WOMEN INCLUDED IN THE PRIMARY ANALYSIS.

CHARACTERISTIC	HPV-16 VACCINE (N=768)	PLACEBO (N=765)	TOTAL (N=1533)
Age — yr			
Mean ±SD	20.0±1.63	20.1±1.61	20.0±1.62
Range	16–25*	16–23	16–25*
Race or ethnic group — no. (%)			
White	601 (78.3)	561 (73.3)	1162 (75.8)
Black	41 (5.3)	63 (8.2)	104 (6.8)
Hispanic	56 (7.3)	66 (8.6)	122 (8.0)
Asian	49 (6.4)	46 (6.0)	95 (6.2)
Other	21 (2.7)	29 (3.8)	50 (3.3)
Current smoker — no. (%)	183 (23.8)	190 (24.8)	373 (24.3)
Lifetime no. of sex partners — no. (%)			
0	38 (4.9)	34 (4.4)	72 (4.7)
1	218 (28.4)	200 (26.1)	418 (27.3)
2	173 (22.5)	173 (22.6)	346 (22.6)
3	138 (18.0)	131 (17.1)	269 (17.5)
4	105 (13.7)	144 (18.8)	249 (16.2)
5	96 (12.5)	83 (10.8)	179 (11.7)
Thin-layer Papanicolaou test results on day 0 — no. (%)			
Normal	666 (86.7)	656 (85.8)	1322 (86.2)
Abnormal†	84 (10.9)	96 (12.5)	180 (11.7)
ASCUS or AGUS‡	47 (6.1)	58 (7.6)	105 (6.8)
LSIL	35 (4.6)	34 (4.4)	69 (4.5)
HSIL	2 (0.3)	4 (0.5)	6 (0.4)
Unsatisfactory	18 (2.3)	13 (1.7)	31 (2.0)

\* recruited through advertisements on US college campuses and surrounding communities. Not pregnant, reported no prior abnormal Papanicolaou smears, and reported that they had had no more than five male sex partners during their lifetime. Virgins were enrolled if they were seeking contraception. At enrollment, the women provided written informed consent. The institutional review board at each center approved the protocol. Compensation for subjects was determined independently at each center; amounts ranged from \$20 to \$225 per visit.

# Intra-Individual Variation

(helps to assess if “response” in individual is genuine /artifact)

---

- **Sources**

- (short-term) biological variation
- measurement errors

- **Describe amount of variation using..**

- Standard Deviation (SD)
- Coefficient of Variation (c.v.): SD as % of mean
- InterQuatile Range (ie  $Q_1$  to  $Q_3$ )
- etc.

# 2. Sampling Variation

(prelude to **Confidence Intervals**)

---

- o Estimate based on sample  $\neq$  Estimate based on ‘Universe’
- o Cannot 100% *guarantee* it, but there is *a greater chance* of a smaller “error” if use larger sample. Can *quantify the chances* of being “off target” by various amounts.
- o Note: If instrument is faulty, estimate based on ‘Universe’ may not be correct either: e.g., when measuring % of adult Canadians who are bilingual in both official languages. *Non-sampling errors* do not diminish in size (probabilistically) even if use very large sample size (even if use a census).



# Confidence Interval:

OPPOSITE: Different estimates of “% positive” in samples (of size 100) from a “Universe” in which 61% of individuals are “positive”  $\Rightarrow$

To allow for **under/over** estimates, we **add/subtract** a given amount (a **Margin of Error**, i.e., a multiple† of the Standard Error\*) to/from the estimate so as to obtain an “**interval estimate**” rather than a “point-estimate.”

\* Standard Error = SE = SD /  $\sqrt{n}$

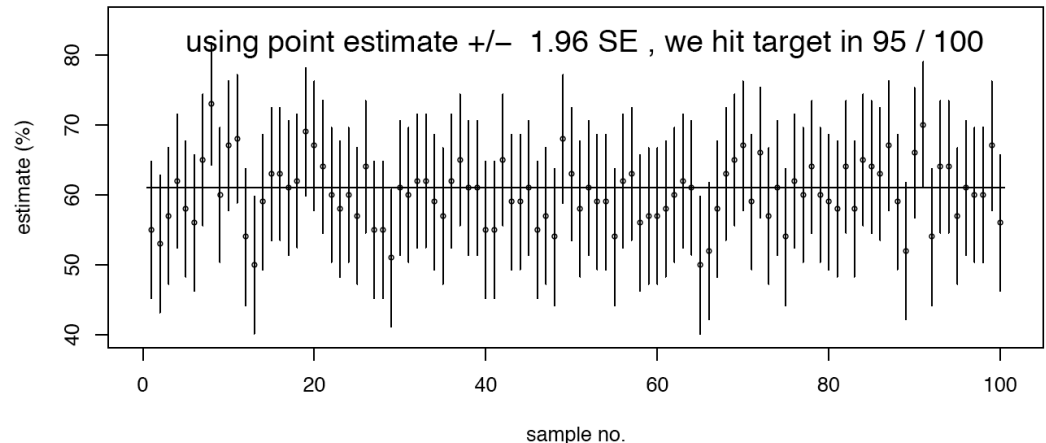
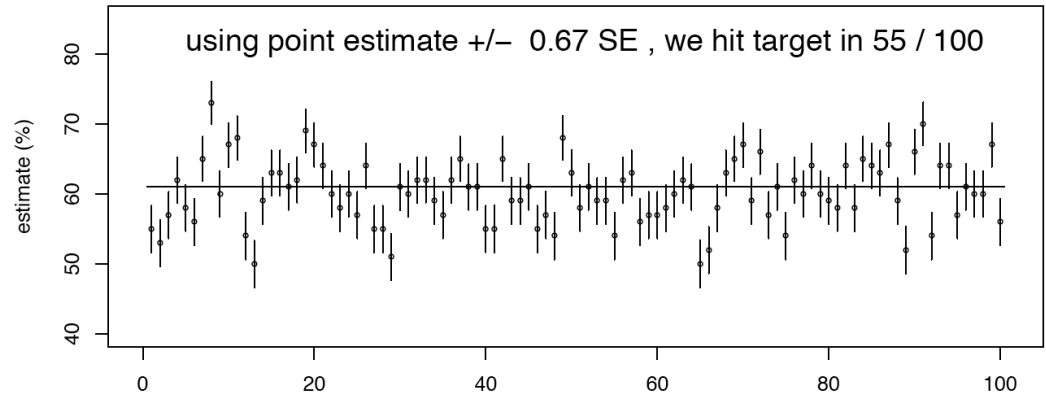
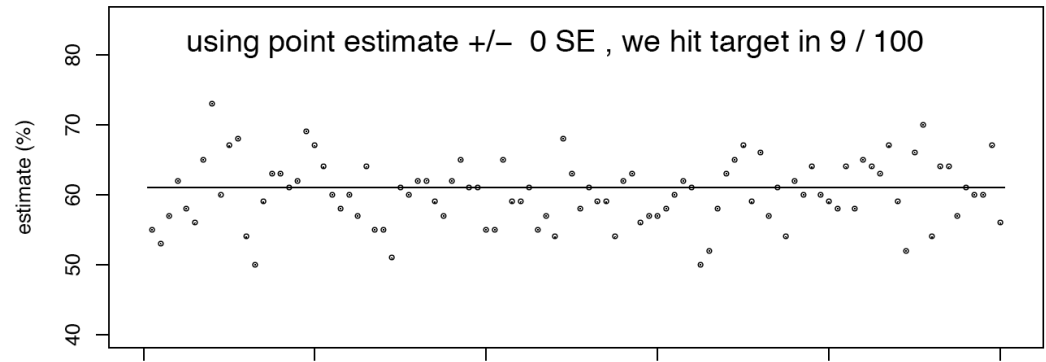
*inversely proportional to  $\sqrt{\text{sample size}}$ ;*

*\* directly proportional to individual-to-individual variability.*

Of course, in real life, we only get to observe 1 sample, but knowing the statistical behaviour of the *procedure* gives us a certain amount of faith in the interval estimate

**(“Confidence” Interval)**

**Increasing the Margin of Error (i.e., the “multiple of SE”) increases the probability of hitting the target (including the true value). But “no free lunch.”**



† 1.64  $\Rightarrow$  90% confidence;  
1.96  $\Rightarrow$  95%, 2.58  $\Rightarrow$  99%

# 95%CI? IC? ... Comment dit on... ?

[La Presse, Montréal, 1993] 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 question a été posée par Gallup à 16 reprises entre 1973 et 1990, et ne l'est qu'une seule fois, en 1973, que la proportion des Canadiens qui se disaient insatisfaits de la façon dont le gouvernement gère 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 moindre 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.**

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

**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.***

The "Margin of Error blurb" for news media was introduced (legislated) in the mid 1980's.

## 2. Confidence Interval

**Definition:** Range of values for a parameter (universe), calculated from the point estimate, the sample size and the observed variation b/w individuals in sample. Choice of degree of confidence up to producer/user, but tradeoff between margin of error and the degree confidence.

**Uses:** Not just for a single proportion or mean; also calculable for comparative parameters such as differences (absolute or ratio) in means, proportions (e.g. risks), incidence rates, x-year or median survival, etc

**Notes:** For *rate ratios*...

(1) CI is usually *asymmetric*: Next example: point estimate: 2.3, lower/upper limits 1.6 & 3.3. 2.3 is 43% higher than 1.6, and 3.3 is 43% higher than 2.3; Often put on log scale  $\Rightarrow$  CI symmetric around the point estimate.

(2) *Width of CI* determined by *numbers of events*, not amounts of experience *per se*.

# Example of Confidence Interval (CI) for Rate Ratio

**Table 1.** Risk of Motor Vehicle Crashes and Near-Miss Incidents after Extended Shifts.\*

Variable	Extended Work Shifts (≥24 hr)	Nonextended Work Shifts (<24 hr)
Crashes		
No. reported	58	73
No. of commutes	54,121	180,289
Rate (per 1000 commutes)	1.07	0.40
<u>Rate ratio (95% CI)</u>	<u>2.3 (1.6–3.3)</u>	1.0
Near-miss incidents		
No. reported	1,971	1,156
No. of commutes	54,121	180,289
Rate (per 1000 commutes)	36.42	6.41
<u>Rate ratio (95% CI)</u>	<u>5.9 (5.4–6.3)</u>	1.0

**TABLE 3. EFFICACY ANALYSES OF A HUMAN PAPILLOMAVIRUS TYPE 16 (HPV-16) LI VIRUS-LIKE-PARTICLE VACCINE.**

	HPV-16 VACCINE				PLACEBO				OBSERVED EFFICACY (95% CI)*	P VALUE
	NO. OF WOMEN	CASES OF INFECTION	WOMAN-YR AT RISK	INFECTION RATE PER 100	NO. OF WOMEN	CASES OF INFECTION	WOMAN-YR AT RISK	INFECTION RATE PER 100		
				WOMAN-YR AT RISK %				WOMAN-YR AT RISK %		
<b>A</b>	768	0	1084.0	0	765	41	1076.9	3.8	100 (90–100)	<0.001
<b>B</b>	800	0	1128.0	0	793	42	1109.7	3.8	100 (90–100)	—§
<b>C</b>	768	6	1084.0	0.6	765	68	1076.9	6.3	91.2 (80–97)	—§

**Efficacy\*:** Maximum potential benefit, in ideal world, with 100% adherence, etc., etc..

**A** Persistent HPV-16 Infection: The per-protocol population included women who received the full regimen of study vaccine and who were seronegative for HPV-16 and negative for HPV-16 DNA on day 0 and negative for HPV-16 DNA at month 7 and in any biopsy specimens obtained between day 0 and month 7; who did not engage in sexual intercourse within 48 hours before the day 0 or month 7 visit; who did not receive any nonstudy vaccine within specified time limits relative to vaccination; who did not receive courses of certain oral or parenteral immunosuppressive agents, immune globulin, or blood products; who were not enrolled in another study of an investigational agent; and who had a month 7 visit within the range considered acceptable for determining the month 7 HPV-16 status.

**B** Persistent HPV-16 Infection:( including women with general protocol violations): The population includes women who received the full regimen of study vaccine and who were seronegative for HPV-16 and negative for HPV-16 DNA on day 0 and negative for HPV-16 DNA at month 7 and in any biopsy specimens obtained between day 0 and month 7.

**C** As in A, but Transient or Persistent HPV-16 Infection:

\* Contrast with benefit in real world, with < 100% adherence, etc. {referred to as “effectiveness”}

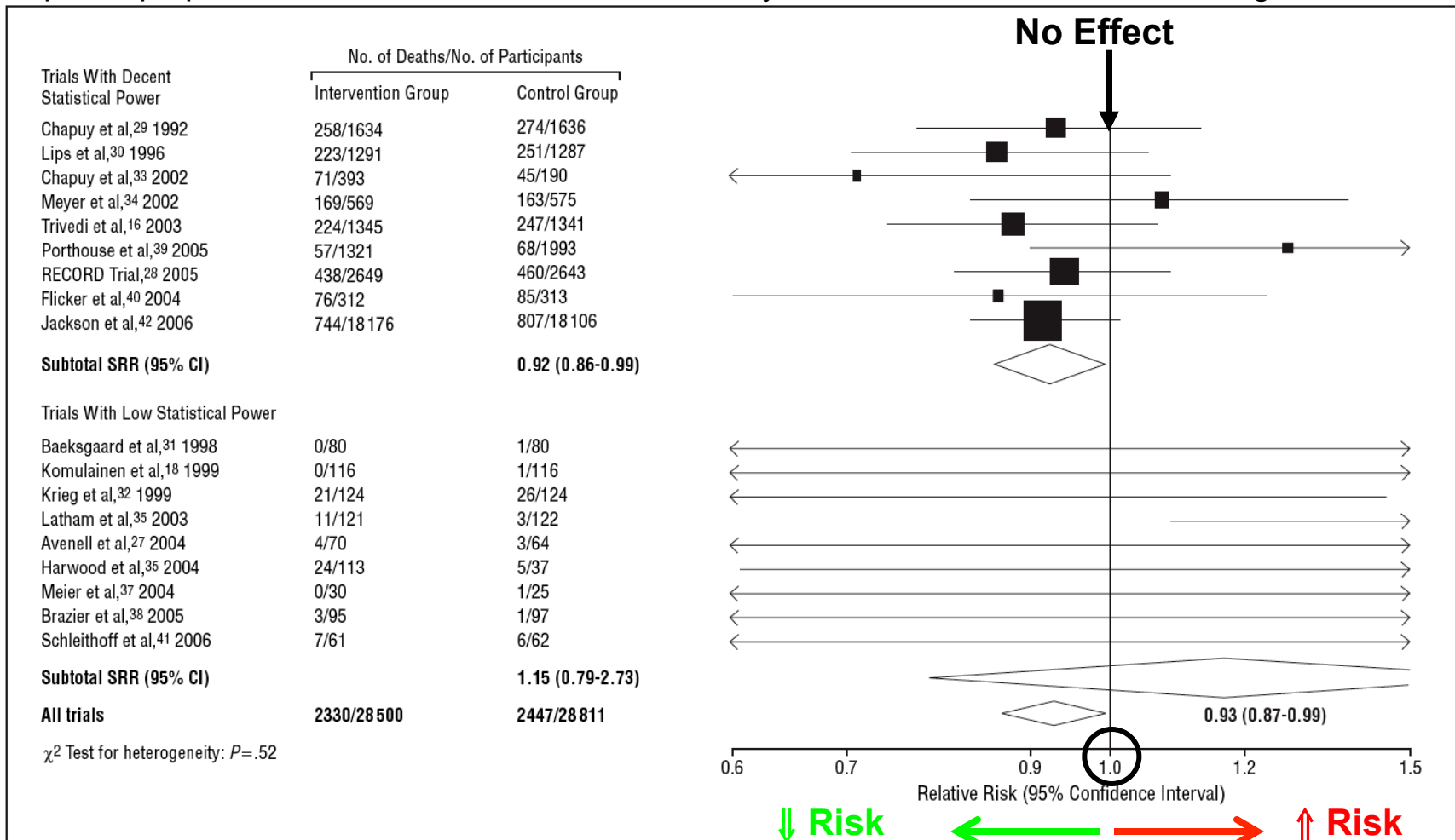
# Vitamin D Supplementation and Total Mortality

A Meta-analysis of Randomized Controlled Trials

Arch Intern Med. 2007;167(16):1730-1737

Each horizontal line, and diamond, is a 95% CI.

Squares proportional to amount of information in study. Note that Relative Risk is on a log scale.



**Figure.** Meta-analysis of data on all-cause mortality in 18 randomized controlled trials with vitamin D. SRR indicates summary relative risk.

# 3. P-Value

(basis for statistical significance levels, tests of hypotheses)

---

**What it is:** A probability, calculated under a “Null Hypothesis” assumption, i.e., assuming that the **only** factor operating is sampling variation.

**Use:** To assess the evidence provided by sample data in relation to a pre-specified claim or 'hypothesis' concerning some parameter(s) or data-generating process. As with a confidence interval, it makes use of the concept of a sampling distribution.

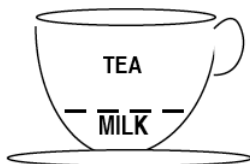
**Examples:** ...

Example 1 (see R. A Fisher, Design of Experiments Chapter

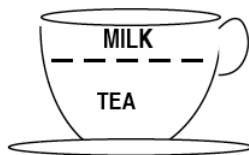
## STATISTICAL TEST OF SIGNIFICANCE

LADY CLAIMS SHE CAN TELL  
WHETHER

MILK WAS POURED  
FIRST





MILK WAS POURED  
SECOND

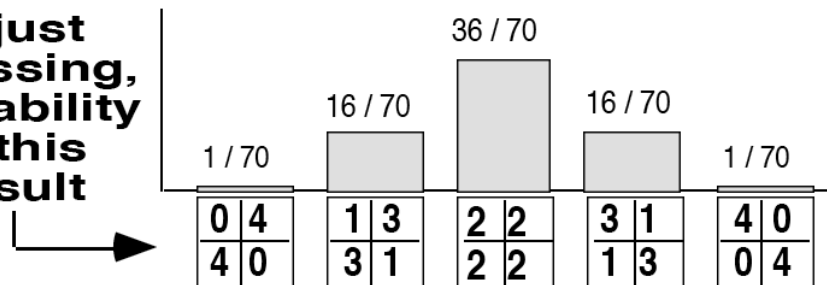


BLIND TEST



LADY SAYS		4	0
		0	4

if just  
guessing,  
probability  
of this  
result



**“Null” Hypothesis**

She cannot tell  
(just guessing)

**“Alternative” Hypothesis**

She can tell

**Any Other Hypotheses**

???



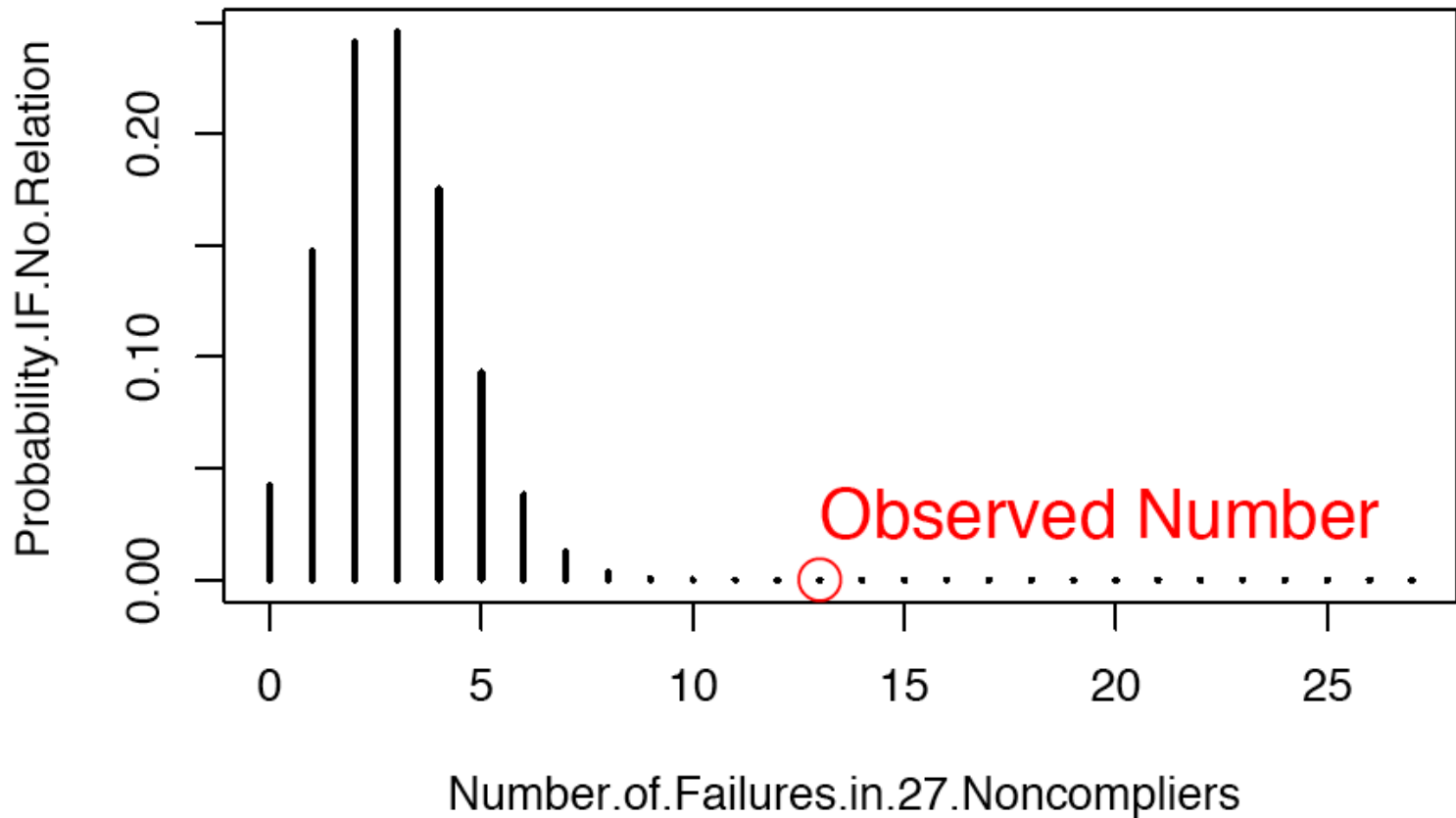
## Example 2 Medical students' compliance with simple administrative tasks and success in final examinations: retrospective cohort study. [BMJ 2002;324:1554–5]

Medical students at the University of Sheffield are asked to provide a recent passport photograph at the start of their paediatric module. The pictures are collated, photocopied, and distributed to the wards, teachers, and hospitals within the paediatric programme. This makes identification easier and facilitates pastoral support and assessment. We studied whether students who were unable to comply with simple administrative tasks—for example, supplying a photograph—were more likely to struggle and subsequently to fail their end of year examinations. All students received a written and a verbal request for a photograph at registration for the module. In the introductory week, verbal reminders were given twice, and a list of those who had not supplied a photograph was displayed on a notice board, downstairs from the venue for the introductory course, on which the week's timetable is also posted. A photograph booth was situated in the students' union building, about 120 metres away from the venue for the introductory week. In 1998 and 1999, a total of 393 students started their paediatric module. Passing the final examinations at the end of the year is prerequisite to entering the final year of the course.

### We checked whether or not a photograph had been provided by the end of the introductory week against the pass and fail lists.

A total of 366 (93%) students handed in photographs, and of these 29 (8%, 95% confidence interval 6% to 11%) failed or were disqualified from sitting the examination at the first attempt because they did not satisfactorily complete the clinical component of the course. Of the 27 students who failed to provide a photograph, 13 (48%, 29% to 67%;  $P < 0.001$ , Fisher's exact test) failed the end of year examinations.

Outcome	Handed In Photo	Did Not Hand In Photo	Total
Passed	337	14	351
Failed	29 ( 8%)	13 ( 48%)	42 ( 11%)
TOTAL	366	27	393



**P-Value** = Probability, **IF NO RELATIONSHIP**, that 13 or more of the 42 failures would be among the 27 Non-Compliers

P-Value =  $2.5 \times 10^{-07}$

**P-Value (In General) = Probability of observing this much, or more, evidence against the Null Hypothesis, IF IN FACT THE NULL HYPOTHESIS IS TRUE.**

## The NULL HYPOTHESIS



"Find out who set up this experiment. It seems that half of the patients were given a placebo, and the other half were given a different placebo" American Scientist 1982; 70:25.

### Example 3: Vaccine Study

P-Value = the Probability,

**CALCULATED UNDER THE ASSUMPTION THAT THERE IS NO RELATIONSHIP (IN THE LARGER UNIVERSE),**

that NONE of the 41 persistent infections would be among the (50%) who were vaccinated

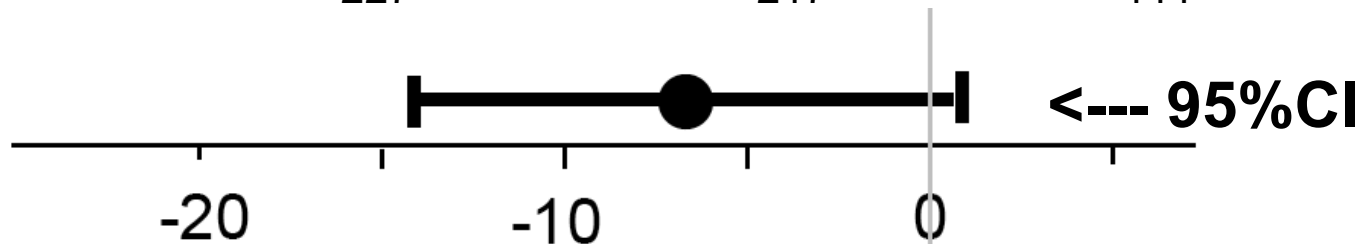
P-Value =  $(1/2)^{41}$  (1-sided)

#### Example 4 Do infant formula samples shorten the duration of breastfeeding?

Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 May 21;1(8334):1148-51.

Randomized Clinical Trial (RCT) which withheld free formula samples [given by baby-food companies to breast-feeding mothers leaving hospital with their infants] from a random half of those studied.

At 1 month	Mothers given sample	Mothers not given sample	Total	
Still breast feeding	175 (77%)	182 (84%)	357 (80.4%)	<b>P=0.07.</b> So, difference is <b>"Not Statistically Significant"</b> at 0.05 level.
Not breast feeding	52	35	87	
TOTAL	227	217	444	



**NO MATTER WHETHER THE P-VALUE IS STATISTICALLY SIGNIFICANT OR NOT, ALWAYS LOOK AT THE LOCATION AND WIDTH OF THE CONFIDENCE INTERVAL. IT GIVES YOU A BETTER AND MORE COMPLETE INDICATION OF THE MAGNITUDE OF THE EFFECT AND OF THE PRECISION WITH WHICH IT WAS MEASURED.**

**THIS IS AN EXAMPLE OF AN INCONCLUSIVE NEGATIVE STUDY**

## EXAMPLE 5: a "Definitive" Negative Study

### Starch blockers -- their effect on calorie absorption from a high-starch meal.

Bo-Linn GW. et al New England Journal of Medicine. 307(23): 1413-6, 1982 Dec 2

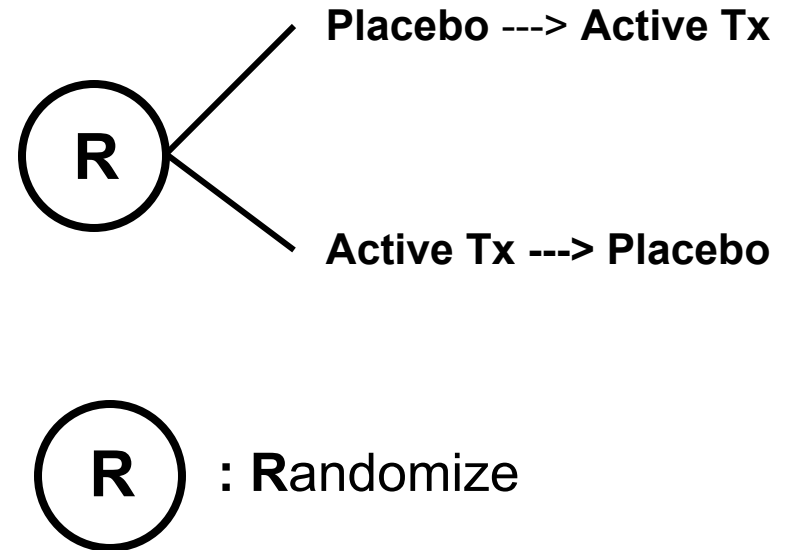
Abstract: It has been known for more than 25 years that certain plant foods, such as kidney beans and wheat, contain a substance that inhibits the activity of salivary and pancreatic amylase. More recently, this anti-amylase has been purified and marketed for use in weight control under the generic name "starch blockers." **Although this approach to weight control is highly popular, it has never been shown whether starch-blocker tablets actually reduce the absorption of calories from starch.**

Using a one-day calorie-balance technique and a high starch (100 g) meal (spaghetti, tomato sauce, and bread), we measured the excretion of fecal calories after normal subjects had taken either placebo or starch-blocker tablets. **If the starch-blocker tablets had prevented the digestion of starch, fecal calorie excretion should have increased by 400 kcal.**

**However, fecal calorie excretion was the same on the two test days (mean  $\pm$  S.E.M.,  $80 \pm 4$  as compared with  $78 \pm 2$ ).**

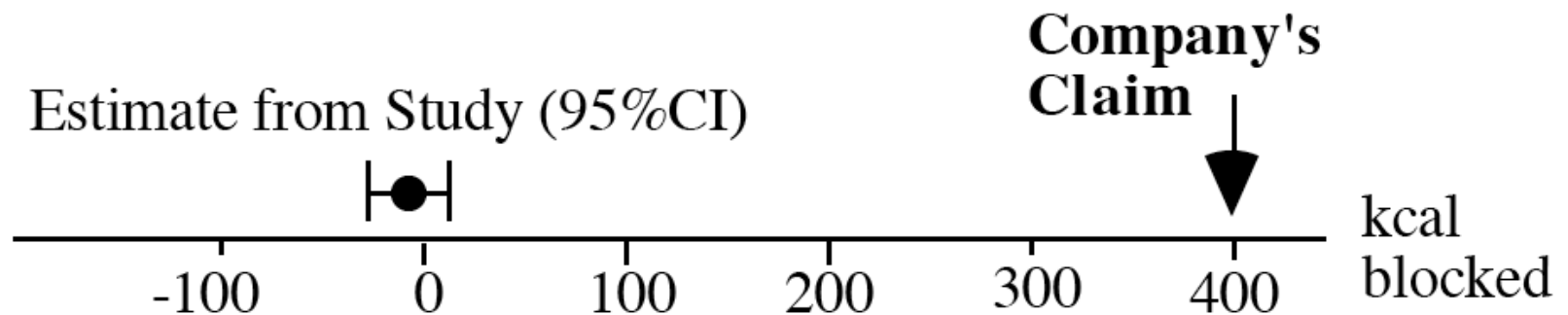
**We conclude that starch blocker tablets do not inhibit the digestion and absorption of starch calories in human beings.**

## Crossover Study



**Table 2. Results in Five Normal Subjects on Days of Placebo and Starch-Blocker Tests.**

	Placebo Test Day			Starch-Blocker test Day		
	DUPLICATE TEST MEAL*	RECTAL EFFLUENT	MARKER RECOVERY	DUPLICATE TEST MEAL	RECTAL EFFLUENT	MARKER RECOVERY
	<i>kcal</i>	<i>kcal</i>	%	<i>kcal</i>	<i>kcal</i>	%
1	664	81	97.8	665	76	96.6
2	675	84	95.2	672	84	98.3
3	682	80	97.4	681	73	94.4
4	686	67	95.5	675	75	103.6
5	676	89	96.3	687	83	106.9
Means	677	80	96.4	676	78	100
±S.E.M.	±4	±4	±0.5	±4	±2	±2



**EFFECT IS MINISCULE (AND ESTIMATE IS QUITE PRECISE!)  
AND VERY FAR FROM COMPANY'S CLAIM !**

# SUMMARY

---

- Descriptive statistics should be descriptive.
- Confidence intervals are preferable to P-values, since they are expressed in terms of the (comparative) parameter of interest; they allow us to judge magnitude, and see if certain parameter values are ruled in / out.
- A statistically significant differences does not necessarily imply a clinically important difference.
- A ‘not-statistically-significant’ difference does not necessarily imply that we have ruled out a clinically important difference.
- Ultimately, P-values, CI’ s and other evidence from a study need to be combined with other information bearing on the parameter or process.
- We should not treat any one study as the last word on the topic.
- We need to worry about *distortions of a non-sampling kind* that are not minimized by having a large ‘*n.*’