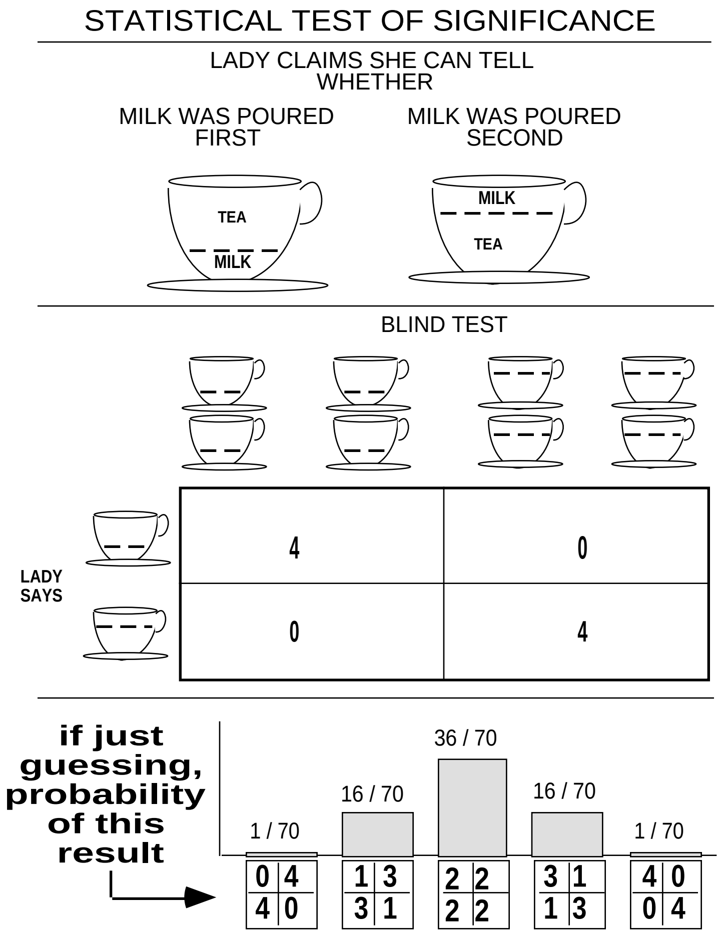
(Frequentist) P-Values and Tests of Hypotheses

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 confidence intervals, tests of significance make use of the concept of a sampling distribution.

Example 1 (see R. A Fisher, Design of Experiments Chapter 2)



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

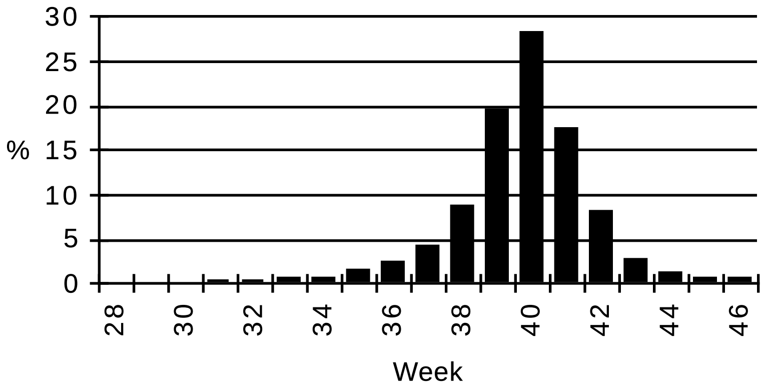
Medical students at the University of Sheffield are asked to provide a recent passport photograph at the start of their paediatric module. The pictures are collated, photocopied, and distributed to the wards, teachers, and hospitals within the paediatric programme. This makes identification easier and 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
TOTAL	366	27	393

Example 3 [Preston-Jones vs. Preston-Jones, English House of Lords]
In 1949 a divorce case was heard in which the sole evidence of adultery was that a baby was born almost 50 weeks after the husband had gone abroad on military service. To quote the court "The appeal judges agreed that the limit of credibility had to be drawn somewhere, but on medical evidence 349 (days) while improbable, was scientifically possible." So the appeal failed.

Pregnancy Duration: 17000 cases > 27 weeks
(quoted in Guttmacher's book)



Example 4 Occupational driving and lumbar disc degeneration: a case-control *** study [Battié et al. THE LANCET • Published online October 15, 2002]

****** Comment (JH) . It is NOT a case-control study.***

Background Back problems are reported more by occupational drivers than by any other occupational group. One explanation is that whole-body vibration caused by the vehicle leads to accelerated disc degeneration, herniation, and associated symptoms. We aimed to investigate the effects of lifetime driving exposure on lumbar disc degeneration in monozygotic twins with very different histories of occupational driving during their life.

Methods We assessed 45 male monozygotic twin pairs from the population-based Finnish Twin Cohort who had greatly different patterns of occupational driving during their life. Data were obtained for driving exposures and potential confounding factors through an extensive, structured interview. We assessed disc degeneration with lumbar MRI.

Findings Disc degeneration did not differ between occupational drivers and their twin brothers. We also did not identify any overall tendency for greater degeneration or pathology in occupational drivers than their twin brothers.

Interpretations Although driving may exacerbate symptoms of back problems, it does not damage the disc. Our inability to identify structural damage should be encouraging to those employed in occupations involving motorised vehicles and operation of heavy equipment.

Table 1: Exposures to occupational driving and possible confounding

Variable	Drivers (n=45)	Co-twins (n=45)	Paired difference	p*
Lifetime occupational driving hours†	34 000 (20 500)	6300 (7000)	27 700 (18 500)	<0·0001
Vibration-weighted driving hours‡	23 900 (16 400)	4200 (5000)	19 700 (14 900)	<0·0001
Job code (1–4 scale)§	2·7 (0·8)	2·4 (0·9)	0·2 (1·2)	0·2471
Occupational lifting/day (kg lifted frequency/day)§	1700 (3900)	1000 (2200)	700 (4600)	0·1010
Time working twisted/bent (h/day)§	1·8 (1·8)	1·6 (1·5)	0·2 (2·2)	0·7306
Commute time (min)	26·7 (24·2)	34·2 (25·6)	7 (29)	0·0507
Cigarette smoking (pack-years)	17·8 (20·8)	15·5 (19·4)	2·3 (19·5)	0·2667

Values are mean (SD). *Calculated with Wilcoxon rank test. †A work year of 220 days of 8 h/day of occupational driving would equal 1760 driving hours. ‡This measure is the number of hours spent in driving jobs multiplied by an estimate of the vibration frequency for the corresponding vehicle type, as measured in the 1980s by ISO 2631 vibration measurement standards. §Means are weighted by number of months spent in each job over the participant's lifetime work history.

Table 2: Differences in lumbar degenerative findings in upper and lower lumbar regions between drivers and their co-twins (Upper lumbar)

Disc outcome	Scale*	Driver	Co–twin	Difference (95% CI)	p
Signal intensity (digital)	CSF-adjusted	–0·26 (0·05)	–0·27 (0·05)	0·01 (–0·01 to 0·03)	0·1840
Annular tears	Disks with tears	0·30 (0·59)	0·33 (0·64)	–0·05 (–0·27 to 0·17)	1·0000
Bulging	0–3 score	0·93 (0·53)	1·00 (0·59)	–0·07 (–0·21 to 0·07)	0·6079
Herniations	Disks with herniations	0·07 (0·25)	0·07 (0·25)	0·00 (–0·09 to 0·09)	1·0000
Disc height narrowing	0–3 score	0·22 (0·26)	0·26 (0·33)	–0·04 (–0·12 to 0·04)	0·3053

Values are mean (SD). CSF=cerebrospinal fluid. *All variables are scaled so that a **higher score represents more degeneration.**

Example 5 Do infant formula samples shorten the duration of breast-feeding? Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 May 21;1(8334):1148-51.

RCT which withheld free formula samples [normally given by baby-food companies to 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%)
not breast feeding	52	35	87
	227	217	444

P=0.07. So, the difference is "Not Statistically Significant" at 0.05 level.

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.

Elements of a Statistical Test The ingredients and the methods of procedure in a statistical test are:

Elements / STEPS	**Preston-Jones**	**"free formula samples" study**	**Tea Tasting / Med Students**
1. A claim about a parameter (or about the shape of a distribution, or the way a lottery works, etc.). Note that the null and alternative hypotheses are usually stated using Greek letters, i.e. in terms of population parameters, and in advance of (and indeed without any regard for) sample data. *[Some have been known to write hypotheses of the form H: $\bar{y}$ = ... , thereby ignoring the fact that the whole point of statistical inference is to say something about the population in general, and not about the sample one happens to study. It is worth remembering that statistical inference is about the individuals one DID NOT study, not about the ones one did. This point is brought out in the absurdity of a null hypothesis that states that in a triangle taste test, exactly p=0.333.. of the n = 10 individuals to be studied will correctly identify the one of the three test items that is different from the two others.]*	Parameter (unknown) : DATE OF CONCEPTION Claim about parameter H_0 DATE ≤ HUSBAND LEFT (use = as 'best case') H_a DATE > HUSBAND LEFT	π: prop'n B'Fed at 1 month. Difference $\pi_{FREE} - \pi_{NOT}$ Claim about parameter H_0 $\pi_{FREE} - \pi_{NOT} = 0$ H_a $\pi_{FREE} - \pi_{NOT} \neq 0$	π: proportion correct π: proportion failing $\pi_{COMPLIERS} - \pi_{NON\text{-}C}$ $\pi_{COMPLIERS} - \pi_{NON\text{-}C} = 0$ $\pi_{COMPLIERS} - \pi_{NON\text{-}C} \neq 0$
2. A probability model (in its simplest form, a set of assumptions) which allows one to predict how a relevant statistic from a sample of data might be expected to behave under H_0.	***Empirical***	individual 'responses' are Bernoulli; difference in 2 proportions (statistics) follows ***Gaussian*** distrn. (prop'ns based on large n's)	Number correct ~ **HyperGeometric** distrn. Among those who fail, number who had not complied ~ **HyperGeometric** distrn.
3. A probability level or threshold or dividing point below which (i.e. close to a probability of zero) one considers that an event with this low probability 'is unlikely' or 'is not supposed to happen with a single trial' or 'just doesn't happen'. This pre-established limit of extreme-ness is referred to as the "α (alpha) level" of the test. ***THIS STEP NOT APPLICABLE IF SIMPLY WISH TO CALCULATE AND REPORT THE P-VALUE ITSELF***	(a priori ??) "limit of extreme-ness" relative to H_0 - for judge to decide Note extreme-ness measured as conditional probability, not in days	α = **0.05**	• Tea: See below (Fisher) • Medical Students: P-VALUE REPORTED Cannot tell whether 0.001 was a pre-specified threshold (α), or (as I suspect) the most extreme p-value in the Table consulted, or rounded to this by the editors

Elements / STEPS (Continued)	**Preston-Jones**	"free formula samples" study	**Tea Tasting / Med Students**
4. A sample of data, which under H_0 is expected to follow the probability laws in (2).	date of delivery.	Breastfeeding behaviours of 444 individuals.	The 8 responses
5. The most relevant statistic (e.g. $\bar{y}_1$ $\bar{y}_2$ i if interested in inference about the comparative parameter $\mu_1 - \mu_2$)	The most relevant statistic (date of delivery; same as raw data: n=1)	difference of two proportions or %'s i.e., 77 - 84	The number of (say) the 'milk-1st' that were correctly identified
6. The probability of observing a value of the statistic as extreme or more extreme (relative to that hypothesized under H_0) than we observed. This is used to judge whether the value obtained is either 'close to' i.e. 'compatible with' or 'far away from' i.e. 'incompatible with', H_0. The 'distance from what is expected under H_0' is usually measured as a tail area or probability and is referred to as the "P-value" of the statistic in relation to H_0.	We calculate **probability** of a gestation of **349 days OR LONGER**. Even the most likely gestation (280 days) is rare (~ 4%) P-value = *Upper* tail area : Prob[349 **or 350 or 351 ...**] : quite small	P-Value = 0.07	• Tea: P-Value = 2/ 70 (sum of 2-tails) • Students: P-Value quite small
IF DO NOT PLAN TO DICHOTOMIZE P-VALUE, STOP HERE			
7. A comparison of this "extreme-ness" or "unusualness" or "amount of evidence against H_0 " or P-value with the agreed-on "threshold of extreme-ness". I f it is beyond the limit, H_0 is said to be "rejected". If it is not-too-small, H_0 is "not rejected". These two possible decisions about the claim are reported as "the null hypothesis is rejected at the P= α significance level" or "the null hypothesis is not rejected at a significance level of 5%".	Judge didn't tell us his threshold, but it must have been smaller than that calculated in 6. ***Note: the p-value does NOT take into account any other 'facts', prior beliefs, testimonials, etc.. in the case. But the judge probably used them in his overall decision (just like the jury did in the OJ case).***	P-Value (= 0.07) is NOT < 0.05 H_0 is NOT REJECTED	• Tea: P < 0.05 ; H_0 is REJECTED. ***See next page for origins of the '0.05' significance level***

WHY 0.05 ?? R. A. Fisher (from **Mathematics of a Lady Tasting Tea**)

It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him. Thus, if he wishes to ignore results having probabilities as high as 1 in 20—the probabilities being of course reckoned from the hypothesis that the phenomenon to be demonstrated is in fact absent—then it would be useless for him to experiment with only 3 cups of tea of each kind. For 3 objects can be chosen out of 6 in only 20 ways, and therefore complete success in the test would be achieved without sensory discrimination, i.e., by "pure chance," in an average of 5 trials out of 100.

It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate the whole of the possible effects of chance coincidence, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 trials is decidedly "significant," in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result

LINK BETWEEN CONFIDENCE INTERVALS & TESTS OF HYPOTHESES

- **Calculate a 95% CI as 'point estimate ± 1.96 SE'**
- **Test H_0 (at $\alpha = 0.05$) by calculating how many SE's the point estimate is from H_0. i.e.,**

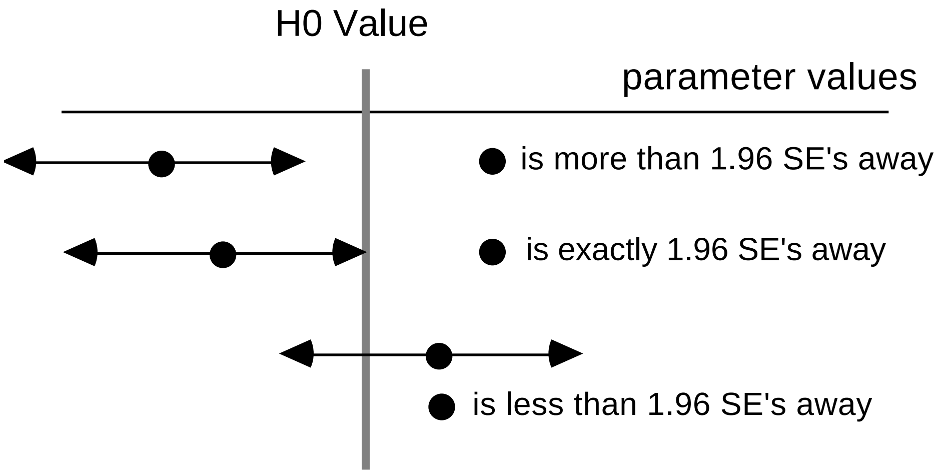
  - calculate $\frac{\textbf{point estimate - null value}}{\textbf{SE}}$

  - diff. "stat. sig" if 'number of SE's away' > 1.96 (P < 0.05)

If one edge of the 95% CI just touches the H_0 value, the point estimate must be exactly 1.96 multiples of the SE away from H_0, so the p-value is exactly 0.05

If the 95% CI includes the H_0 value, the point estimate must be less than 1.96 multiples of the SE away from H_0, so the p-value is > 0.05

If the 95% CI does not iclude the H_0 value, the point estimate must be more than 1.96 multiples of the SE away from H_0, so the p-value is < 0.05



● is more than 1.96 SE's away

● is exactly 1.96 SE's away

● is less than 1.96 SE's away

-> Direct link between 100(1-α)% CI and test with level α

"Operating" Characteristics of a Statistical Test

As with diagnostic tests, there are 2 ways a statistical test can be wrong:

1) **The null hypothesis was in fact correct but [because of purely random variation] the sample was genuinely extreme and the null hypothesis was therefore (wrongly) rejected.**
2) **The alternative hypothesis was in fact correct but the sample was not found to be incompatible with the null hypothesis and so it was not ruled out.**

The probabilities of the various test results can be put in the same type of 2x2 table used to show the characteristics of a diagnostic test.

Result of Statistical Test

		"**Negative**" (do not reject H_0)	"**Positive**" (reject H_0 in favour of H_a)
TRUTH	H_0	$1 - \alpha$	α
	H_a	β	$1 - \beta$

Ftetcher and Fletche p 48 (medical tests) and p 188 (statitical tests) have the rows and columns the opposite way. There is no standard way of displaying 2 x 2 tables.

There are also other possibilities, such as that the data were wrong, or the summaries or statistical calculations were done incorrectly. These systematic errors would not be reversed or minimized by increasing the sample sizes.

The quantities ($1 - \beta$) and ($1 - \alpha$) are the "sensitivity (power)" and "specificity" of the statistical test. Statisticians usually speak instead of the complements of these probabilities, the false positive fraction (α) and the false negative fraction (β) as "Type I" and "Type II" errors respectively [It is interesting that those involved in diagnostic tests emphasize the correctness of the test results, whereas statisticians seem to dwell on the errors of the tests; they have no term for $1-\alpha$].

Note that all of the probabilities start with (i.e. are conditional on knowing) the truth. This is exactly analogous to the use of sensitivity and specificity of diagnostic tests to describe the performance of the tests, conditional on (i.e. given) the truth.

As such, they describe performance in a "**what if**" or artificial scenario, just as sensitivity and specificity are determined under theoretical 'lab' conditions.

So just as we cannot interpret the result of a diagnostic test simply on basis of sensitivity and specificity, likewise we cannot interpret the result of a statistical test in isolation from what one already believes about the null/alternative hypotheses.

Interpretation of a "positive statistical test"

It should be interpreted in the same way as a "positive diagnostic test" i.e. in the light of the characteristics of the subject being examined. The lower the prevalence of disease (or the credibility of the alternative hypothesis), the lower is the post-test probability that a positive diagnostic test is a "true positive". Similarly with statistical tests. We are now no longer speaking of sensitivity = Prob(test + | H_a) and specificity = Prob(test - | H_0) but rather, the other way round, of Prob(H_a | test +) and Prob(H_0 | test -), i.e. of positive and negative predictive values, both of which involve the "background" from which the sample came.

A Popular Misapprehension: It is not uncommon to see or hear seemingly knowledgeable people state that

> "the P-value (or alpha) is the probability of being wrong if, upon observing a statistically significant difference, we assert that a true difference exists"

Glantz (in his otherwise excellent text) and Brown (Am J Dis Child 137: 586-591, 1983 -- on reserve) are two authors who have made statements like this. For example, Brown, in an otherwise helpful article, says (italics and strike through by JH) :

> *~~"In practical terms, the alpha of .05 means that the researcher, during the course of many such decisions, accepts being wrong one in about every 20 times that he thinks he has found an important difference between two sets of observations"~~* [1]

But if one follows the analogy with diagnostic tests, this statement is like saying that

> *"1-minus-specificity is the probability of being wrong if, upon observing a positive test, we assert that the person is diseased".*

We know [cf Dx tests] that we cannot turn Prob(test | H) into Prob(H | test) without some knowledge about the unconditional or a-priori Prob(H) ' s.

Also ask yourself 'how many 'tests' or 'comparisons' were carried out?' Maybe this is the one positive test out of many that the authors carried out but did not report?

[1] [Incidentally, there is a second error in this statement : it has to do with equating a "statistically significant" difference with an important one... minute differences in the means of large samples will be statistically significant]

MULTIPLE MEDICAL / STATISTICAL TESTS (e.g. SMAC18 / SMAC24)

An automated clinical chemistry analyzer can give 18 routinely ordered chemical determinations on one order (glucose, BUN, creatinine, ..., iron). The "normal" values for these 18 tests are established by the concentrations of these chemicals in the sera of a large sample of healthy volunteers. Say that the normal range is defined so that an average of 3% of the values found in these healthy subjects fell outside.

The binomial distribution is not perfectly applicable in this instance (since elements may be correlated) but as a first approximation it gives the following probabilities that out of the 18 tests, a healthy subject will have

# abnormal tests:	0	1	2	>2
probability that this many tests will be abnormal:	57%	32%	8.5%	2.5%

Even if H_0 ('patient is healthy') is true, there is >40% chance that AT LEAST ONE of the 18 test results will be 'abnormal'

Capitalization on chance... 'multiple looks' at data

Statistics question from Ingelfinger et al's text Clinical Biostatistics

Mrs A has mild diabetes controlled by diet. Her morning urine sugar test is negative 80% of the time and positive (+) 20% of the time [It is never graded higher than +].

At her regular visit she tells her physician that the test has been + on each of the last 5 days.

What is the probability that this would occur if her condition has remained unchanged? Does this observation give reason to think that her condition has changed?

Is the situation different if she observes, between visits, that the test is positive on 5 successive days and phones to express her concern?

The more medical or statistical tests or 'looks' that are performed on healthy patients (or when the null hypothesis is true), the higher the chance of at least one false positive result. SO... INTERPRET POSITIVE RESULTS CAREFULLY

"Definitive" and "INCONCLUSIVE" **Negative Studies**

EXAMPLE of "Definitive" Negative Study

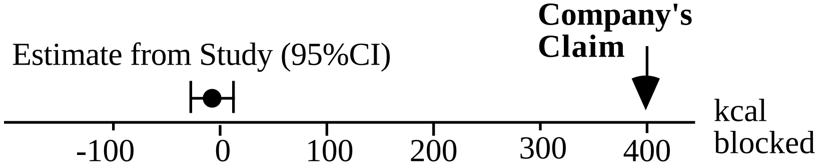
Starch blockers--their effect on calorie absorption from a high-starch meal.

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 antiamylase 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 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 ± S.E.M., 80 ± 4 as compared with 78 ± 2). We conclude that starch-blocker tablets do not inhibit the digestion and absorption of starch calories in human beings.**

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

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
	kcal	*kcal*	*%*	*kcal*	*kcal*	*%*
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 !

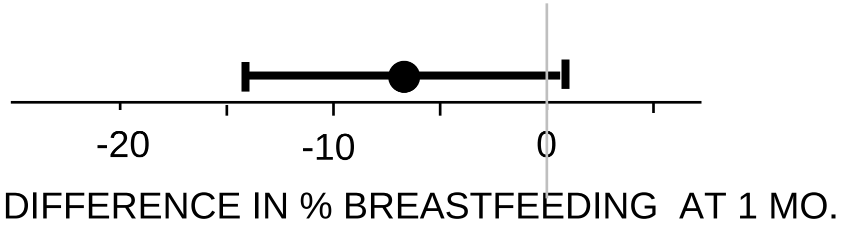
EXAMPLE of "INCONCLUSIVE" Negative Study

Do infant formula samples shorten the duration of breast-feeding? Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 May 21;1(8334):1148-51.

RCT which withheld free formula samples [normally given by baby-food companies to 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%)
not breast feeding	52	35	87
	227	217	444

* P=0.07. So, the difference is "Not Statistically Significant" at 0.05 level.



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 PRECISION WITH WHICH IT IS MEASURED

For an **helpful paper on topic of 'negative' studies**, see Powell-Tuck J "A defence of the small clinical trial: evaluation of three gastroenterological studies." British Medical Journal Clinical Research Ed..292(6520):599-602, 1986 Mar 1.

"SIGNIFICANCE"

notes prepared by FDK Liddell, ~1970

And then, even if the cure should be performed, how can he be sure that this was not because the illness had reached its term, or a result of chance, or the effect of something else he had eaten or drunk or touched that day, or the merit of his grandmother's prayers? Moreover, even if this proof had been perfect, how many times was the experiment repeated? How many times was the long string of chances and coincidences strung again for a rule to be derived from it?

Michel de Montaigne 1533-1592

The same arguments which explode the Notion of Luck may, on the other side, be useful in some Cases to establish a due comparison between Chance and Design. We may imagine Chance and Design to be as it were in Competition with each other for the production of some sorts of Events, and may calculate what Probability there is, that those Events should be rather owing to one than to the other... From this last Consideration we may learn in many Cases how to distinguish the Events which are the effect of Chance, from those which are produced by Design.

Abraham de Moivre: 'Doctrine of Chances' (1719)

If we... agree that an event which would occur by chance only once in (so many) trials is decidedly 'significant', in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us.

R A Fisher 'The Design of Experiments'
(First published 1935)

The difference between two treatments is 'statistically significant' if it is sufficiently large that it is unlikely to have risen by chance alone. The level of significance is the probability of such a large difference arising in a trial when there really is no difference in the effects of the treatments. (But the lower the probability, the less likely is it that the difference is due to chance, and so the more highly significant is the finding.)

- Statistical significance does not imply clinical importance.
- Even a very unlikely (i.e. highly significant) difference may be unimportant.
- Non-significance does not mean no real difference exists.
- A significant difference is not necessarily reliable.
- Statistical significance is not proof that a real difference exists.
- There is no 'God-given' level of significance. What level would you require before being convinced:

  a to use a drug (without side effects) in the treatment of lung cancer?

  b that effects on the foetus are excluded in a drug which depresses nausea in pregnancy?

  c to go on a second stage of a series of experiments with rats?
- Each statistical test (i.e. calculation of level of significance, or unlikelihood of observed difference) must be strictly independent of every other such test. Otherwise, the calculated probabilities will not be valid. This rule is often ignored by those who:

  - measure more than on response in each subject
  - have more than two treatment groups to compare
  - stop the experiment at a favourable point.

WHAT A P-VALUE IS NOT [some of 101 ways to say it wrong!]

In the following situation, a significance test for a population mean μ is called for. **State the null hypothesis H_o and the alternative hypothesis H_a in each case.**

Experiments on learning in animals sometimes measure how long it takes a mouse to find its way through a maze. The mean time is 18 seconds for one particular maze. A researcher thinks that a loud noise will cause the mice to complete the maze faster. She measures how long each of 10 mice takes with a noise as stimulus.

1 $\bar{x}$ = ave. time of 10 mice w/ loud noise.

H_o: mu - $\bar{x}$ = 0 or mu = $\bar{x}$

• No! Ho must be in terms of parameter(s).

One of our (now) faculty once wrote it this way once in a grant application, and the Math/Statistic dept. called us on it!

2 Ho : a loud noise has no effect on the rapidity of the mouse to find its way through the maze.

• It is about mean of a population, i.e. about **mice (plural; generic!)**. It is not about the 10 mice in the study!

A randomized comparative experiment examined whether a calcium supplement in the diet reduces the blood pressure of healthy men. The subjects received either a calcium supplement or a placebo for 12 weeks. The statistical analysis was quite complex, but one conclusion was that "the calcium group had lower seated systolic blood pressure **(P=0.008)** compared with the placebo group."

Explain this conclusion, especially the P-value, as if you were speaking to a doctor who knows no statistics.

(From R.M. Lyle et al., "Blood pressure and metabolic effects of calcium supplement in normotensive white and black men," J American Med Assoc, 257 (1987), pp. 1772-1776.)

1 The P-value is a probability: "P=0.008" means 0.8% . It is the probability, **assuming** the null hypothesis is true, that a sample (similar in size and characteristics as in the study) would have an average BP this far (or further) below the placebo group's average BP. In other words, if the null hypothesis is really true, what's the chance 2 group of subjects would have results this different or more different?

• Not bad!

2 Only a 0.008 chance of finding this difference by chance if, in the population there really was no difference between treatment and central groups.

• Good! But say ≥ this diff.
not
> this diff.

3 The p-value of .008 means that the probability of the observed results if there is, in fact, no difference between "calcium" and "placebo" groups is 8/1000 or 1/125.

• Good, but should change to "the observed results **or results more extreme**"

4 The p-value measures the probability or chance that the calcium supplement had no effect.

• No. First, Ho and Ha refer not just to the n subjects studied, but to all subjects like them. They should be stated in the present tense. **Second, the p-value is about data, under the null H, NOT about the credibility of H_o or H_a.**

5 There is strong evidence that Ca supplement in the diet reduces the blood pressure of healthy men.

The probability of this being wrong conclusion according to the procedure and data corrected is only 0.008 (i.e. 0.8%) .

• Stick to "official wording".. **IF** Ca makes no Δ to average BP, chance of getting ... Notice the illegal attempt to make the p-value into a predictive value -- about as dangerous as a statistician trying to interpret a medical test that gave a reading in the top percentile of the 'healthy' population.

6 Only 0.8% chance that the lower BP in Calcium group is lower than placebo due to chance.

• **If** Ca++ does nothing, then prob. of obtaining a result ≥ this extreme is
- **Illegal.as written** [and almost as muddled and difficult to parse as Bill Clinton's English]

7 The chance that the supplement made no change or raised the B/P is very slim.

• p-value is a conditional statement, predicated (calculated on supposition that) Ca makes no difference to μ. Often stated in present tense. p-value is more 'after the data' in 'past-conditional' tense. **Illegal**.as written

8 There is 0.8% that this difference is due to chance alone and 99.2% chance that this difference is a true difference.

• Not really.. Just like the previous statements, this type of language is crossing over into "Bayes-Speak".

9 There has been a significant reduction in the BP of the treated group...

there's only a probability of 0,8% that this is due to chance alone.

• Cannot talk about the cause... Can say "**IF** no other cause than chance, then prob. of getting ≥ a difference of this size is ...

Notice how 'hamstrung' we are by the 'Frequentist' approach. The Bayesian Bayesian approach is easier and direct.

10 If the P-value 0.008 is true, then 99.2% of the time calcium supplements will work. Only 0.8% of the time, a blood pressure level will be more or less the expected ameliorative effect.

NO! NO! NO!

See earlier pages on what a P-Value is