What do *p*-values and confidence intervals really tell us?

Rita Popat, PhD
Clinical Assistant Professor
Division of Epidemiology
Stanford University School of Medicine

August 9, 2007

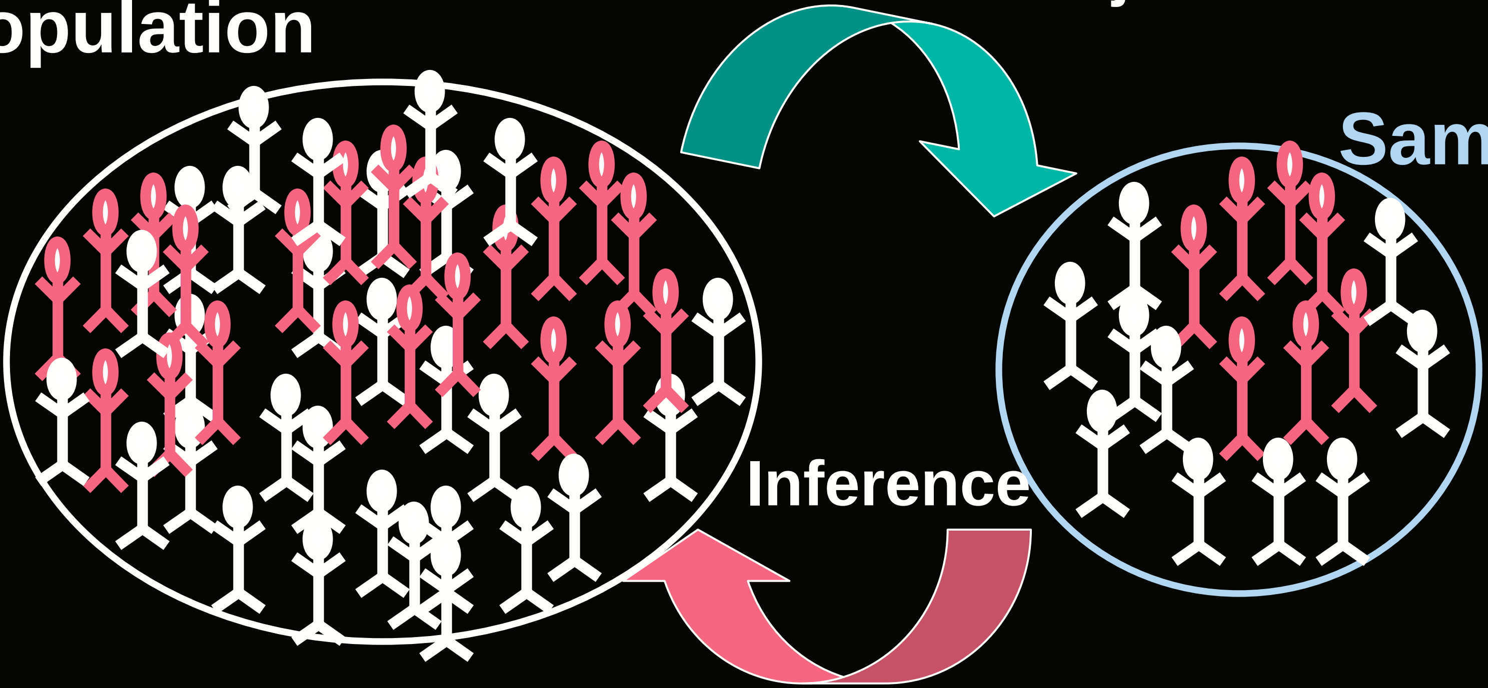
Why use statistics at all?

Average height of all 25-year-old men in North America is a PARAMETER.

The height of the members of a sample of 100 such men are measured; the average of those 100 numbers is a STATISTIC.

Using inferential statistics, we make inferences about population (taken to be unobservable) based on a random sample taken from the population of interest.

Is risk factor X associated with disease Y?



Selection of subjects

Population

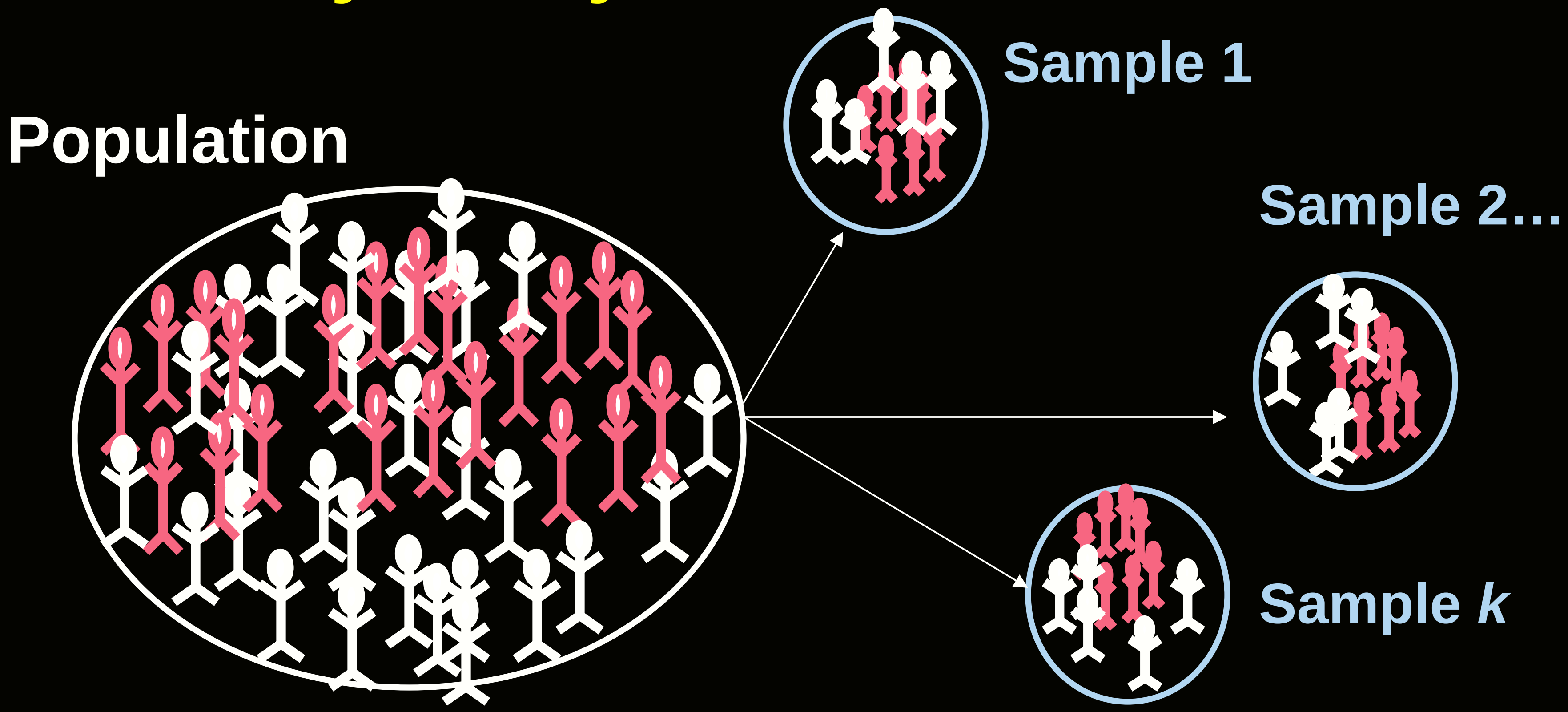
Sample

Inference

From the sample, we compute an estimate of the effect of X on Y (e.g., risk ratio if cohort study):

- Is the effect real? Did chance play a role?

Why worry about chance?



Population

Sample 1

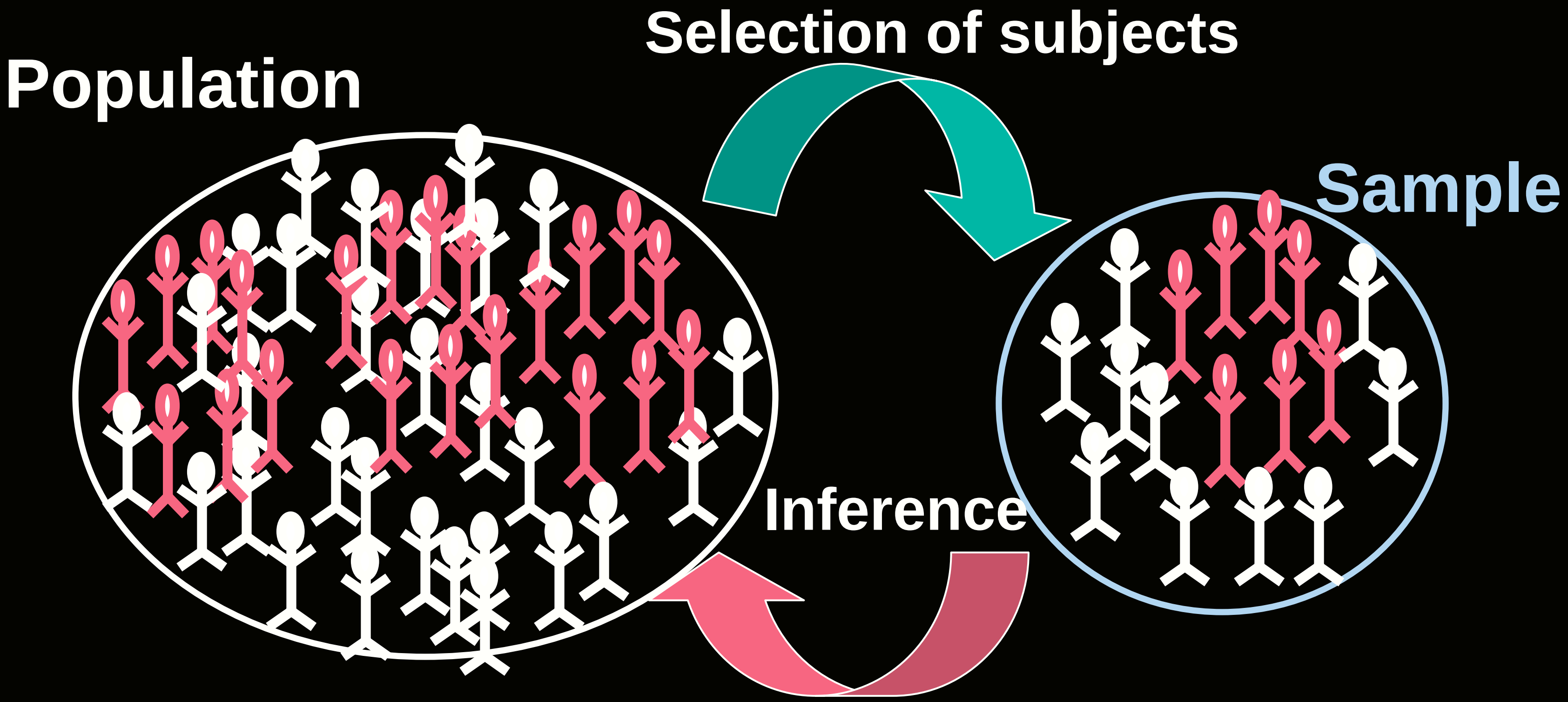
Sample 2…

Sample *k*

Sampling variability…

- you only get to pick one sample!

Interpreting the results



Make inferences from data collected using laws of probability and statistics

- **tests of significance (p-value)**
- **confidence intervals**

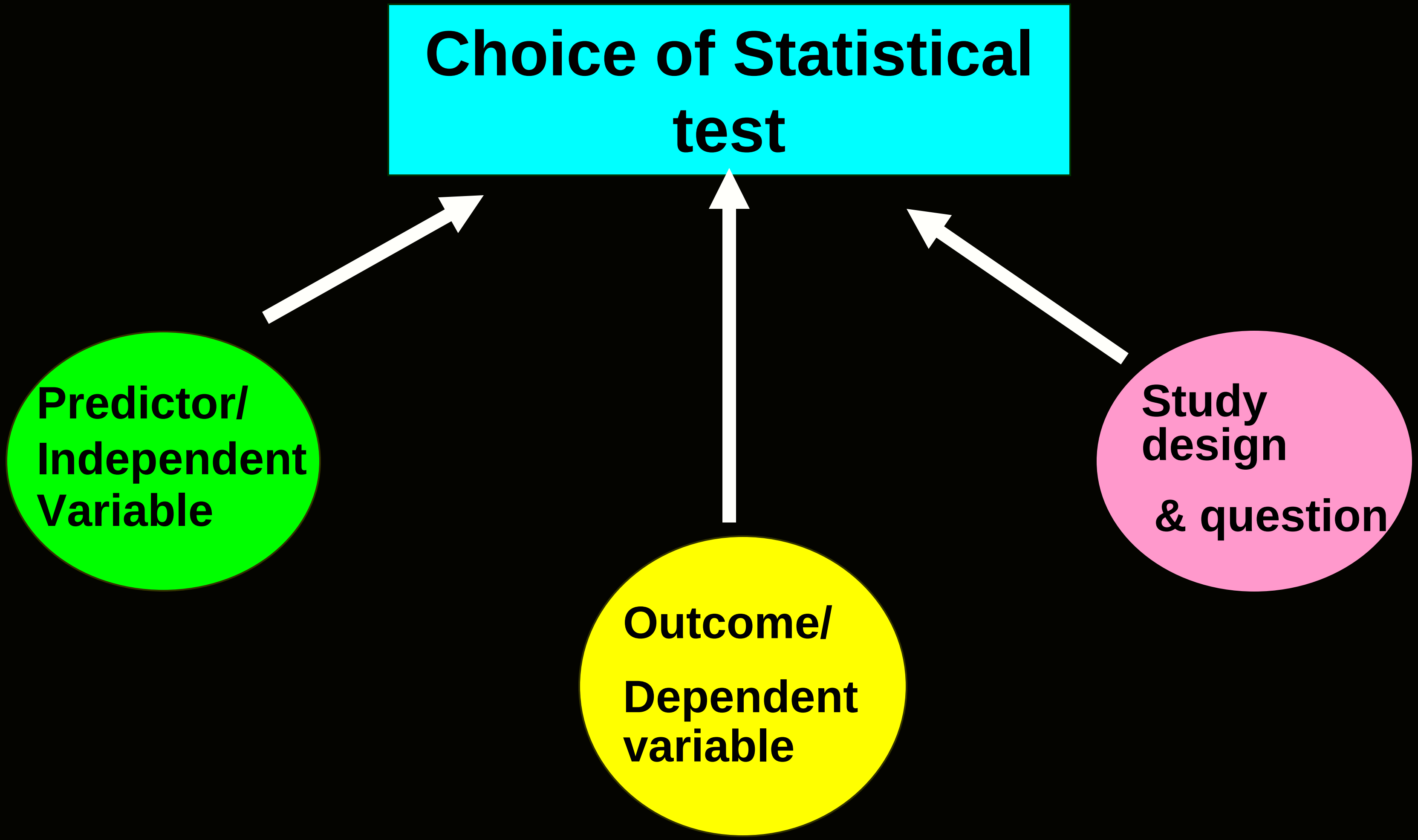
How do we determine if an association is significant?

- Significance is in context of “Statistical” significance
  - p-values
  - Confidence intervals

Significance testing

- The interest is generally in comparing two groups (e.g., risk of outcome in the treatment and placebo group)
- The statistical test depends on the format of the data and the study design

Significance testing



Choice of Statistical test

Predictor/ Independent Variable

Outcome/ Dependent variable

Study design & question

Choice of statistical test when…

Dichotomous outcome (yes/no, alive, dead)

Independent variable
Categorical
(e.g., smoking yes vs no)

Is smoking associated with the outcome?

Statistical test…

- Two sample proportion or
- Chi-square or
- Risk ratio

	Outcome +	Outcome -
Smk (yes)	a (p_{S+})	b
Smk (no)	c (p_{S-})	d

Choice of statistical test when…

Dichotomous outcome (yes/no, alive, dead)

Independent variable
Continuous
(e.g., smoking pack yrs)

Is smoking associated with the outcome?

	Outcome +	Outcome -
Smoking amount (pk yrs)	$\overline{X}_{o+}$	$\overline{X}_{o-}$

Statistical test…

- Two sample t-test

Significance testing

Subjects with Acute MI

Mortality
IV nitrate
P_N

$\overset{?}{<}$

Mortality
No nitrate
P_C

- Suppose we do a clinical trial to answer the above question
- Even if IV nitrate has no effect on mortality, due to sampling variation, it is very unlikely that $P_N = P_C$
- Any observed difference b/w groups may be due to treatment or a coincidence (or chance)

Obtaining *P* values

	Number dead / randomized				
Trial	Intravenous nitrate	Control	Risk Ratio	95% C.I.	P value
Chiche	**3/50**	**8/45**	**0.33**	**(0.09,1.13)**	**0.08**
Bussman	**4/31**	**12/29**	**0.24**	**(0.08,0.74)**	**0.01**
Flaherty	**11/56**	**11/48**	**0.83**	**(0.33,2.12)**	**0.70**
Jaffe	**4/57**	**2/57**	**2.04**	**(0.39,10.71)**	**0.40**
Lis	**5/64**	**10/76**	**0.56**	**(0.19,1.65)**	**0.29**
Jugdutt	**24/154**	**44/156**	**0.48**	**(0.28, 0.82)**	**0.007**

How do we get this *p*-value?

Table adapted from Whitley and Ball. Critical Care; 6(3):222-225, 2002

Null Hypothesis(H_o)

- There is no association between the independent and dependent/outcome variables
  - Formal basis for hypothesis testing

- In the example, **H_o** :"The administration of IV nitrate has no effect on mortality in MI patients" **or** $\boldsymbol{P_N - P_C = 0}$

Example of significance testing

- **In the Chiche trial:**
  - p_N **= 3/50 = 0.06;** p_C **= 8/45 = 0.178**

- **Null hypothesis:**
  - H_0: $p_N - p_C = 0$ **or** $p_N = p_C$

- **Statistical test:**
  - **Two-sample proportion**

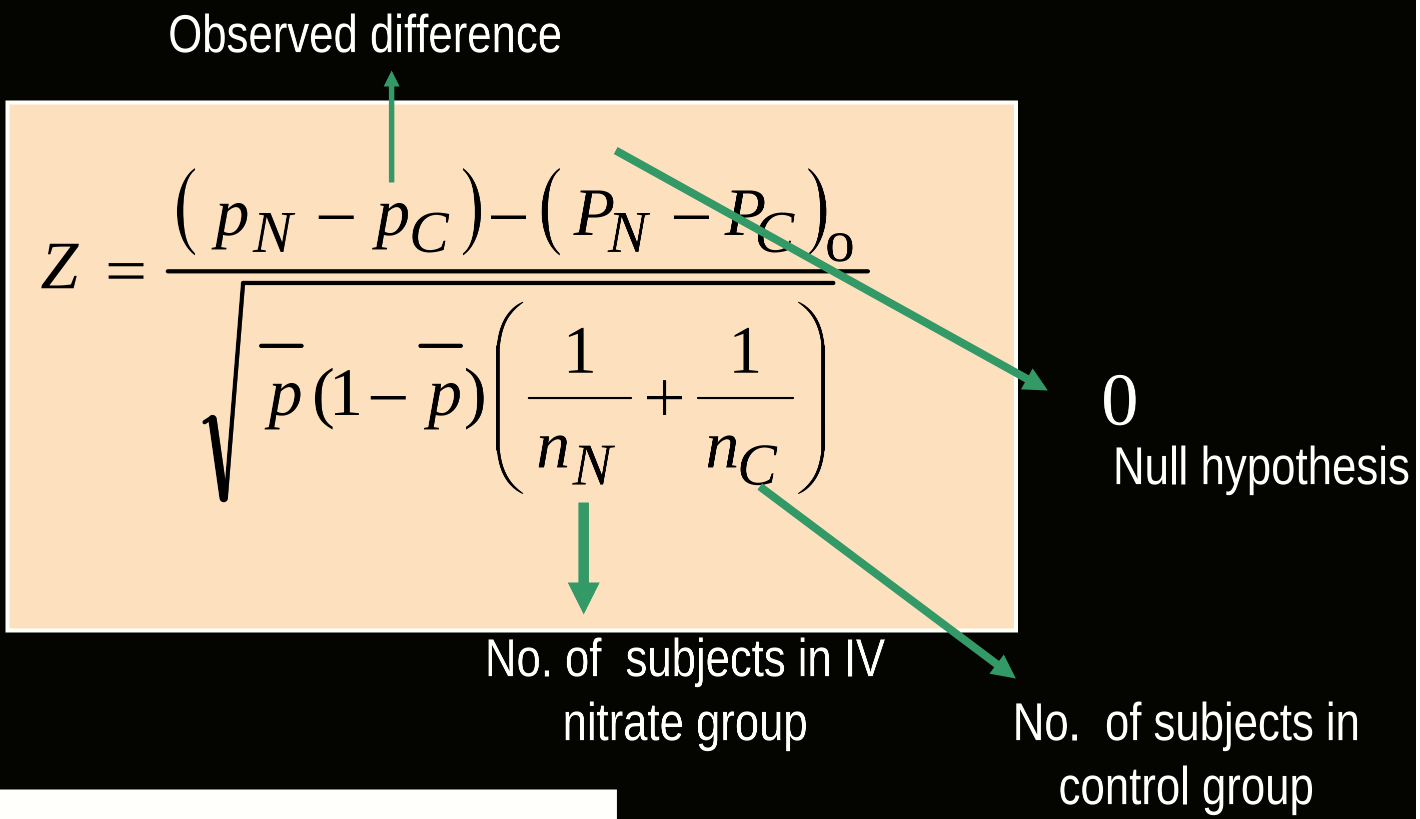
General form of a test statistic

$$\text{test statistic} = \frac{\text{relevant statistic} - \text{hypothesized parameter}}{\text{standard error of the relevant statistic}}$$

$$\bar{p} = \frac{X_N + X_C}{n_N + n_C} \quad , \quad p_N = \frac{X_N}{n_N} \quad , \quad p_C = \frac{X_C}{n_C}$$

Test statistic for Two Population Proportions

The test statistic for $p_1 - p_2$ is a Z statistic:



Observed difference

$$Z = \frac{\left(p_N - p_C\right) - \left(P_N - P_C\right)_0}{\sqrt{\bar{p}\left(1-\bar{p}\right)\left(\frac{1}{n_N} + \frac{1}{n_C}\right)}}$$

0

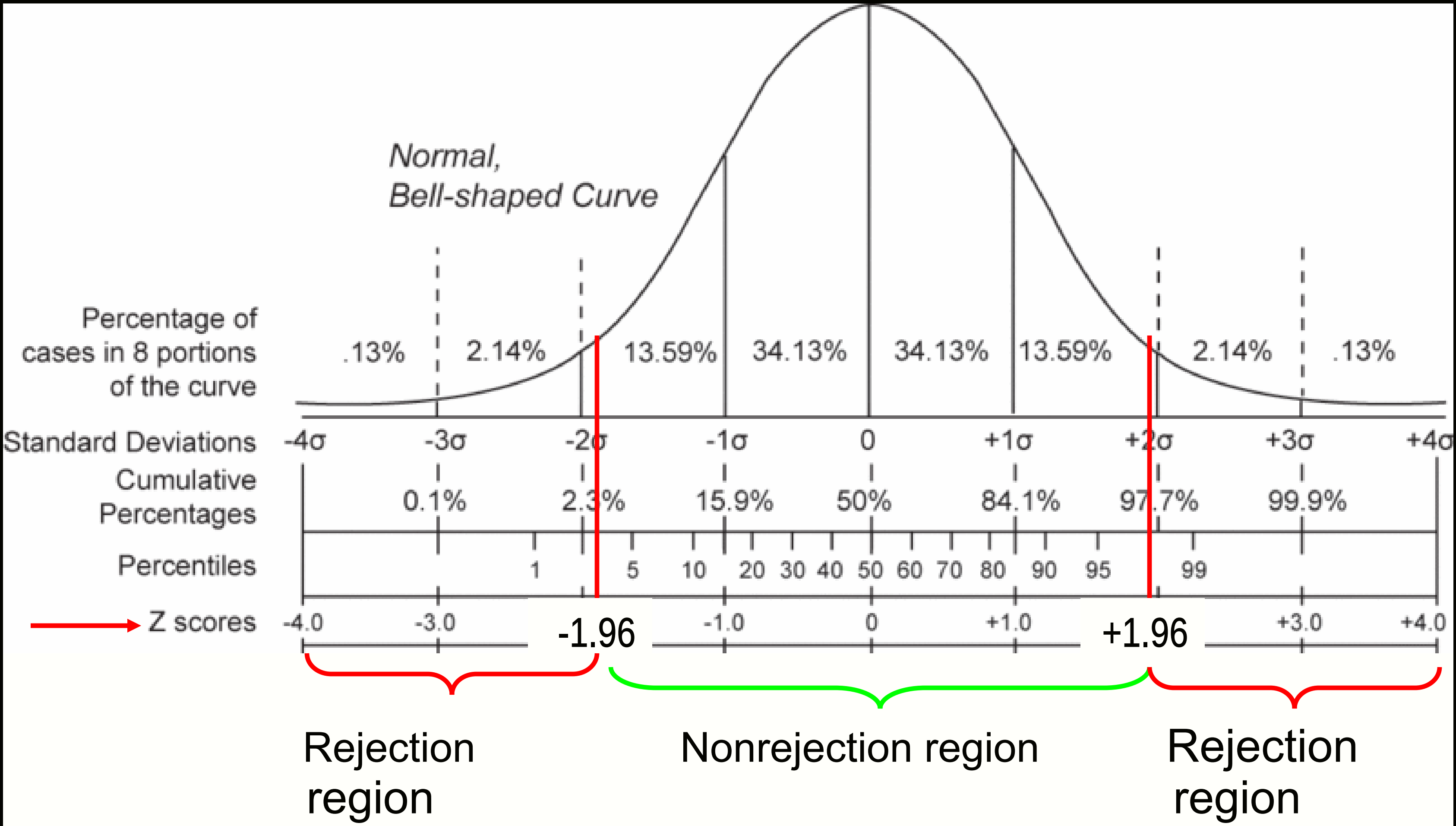
Null hypothesis

No. of subjects in IV nitrate group

No. of subjects in control group

where $\bar{p} = \frac{X_N + X_C}{n_N + n_C}$, $p_N = \frac{X_N}{n_N}$, $p_C = \frac{X_C}{n_C}$

Testing significance at 0.05 level



$Z\alpha/2 = 1.96$

Reject H_0 if $Z < -Z\ \alpha/2$ or $Z > Z\ \alpha/2$

Two Population Proportions

(continued)

$$Z=\frac{\left(0.06-0.178\right)}{\sqrt{0.116(1-.116)\left(\frac{1}{50}+\frac{1}{45}\right)}}=-1.79$$

where

$$\bar{p}=\frac{3+8}{45+50}=0.116\ ,\quad p_N=\frac{3}{45}=0.06\ ,\quad p_C=\frac{8}{50}=0.178$$

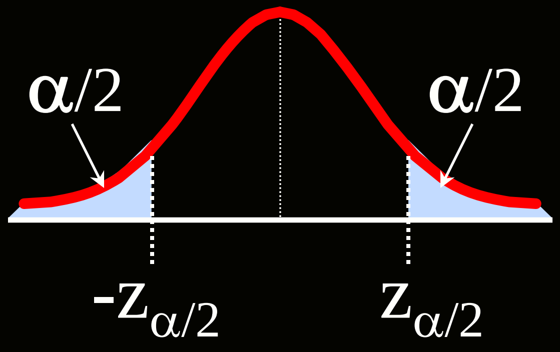
Statistical test for $p_1 - p_2$

Two Population Proportions, Independent Samples

$$Z = \frac{(0.06 - 0.178)}{\sqrt{0.116(1-.116)\left(\frac{1}{50}+\frac{1}{45}\right)}} = -1.79$$

Two-tail test:

$H_0: p_N - p_C = 0$
$H_1: p_N - p_C \neq 0$



$Z_{\alpha/2} = 1.96$

Reject H_0 if $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$

Since -1.79 is > than -1.96, we fail to reject the null hypothesis.

But what is the actual *p*-value?

$P(Z<-1.79) + P(Z>1.79) = ?$

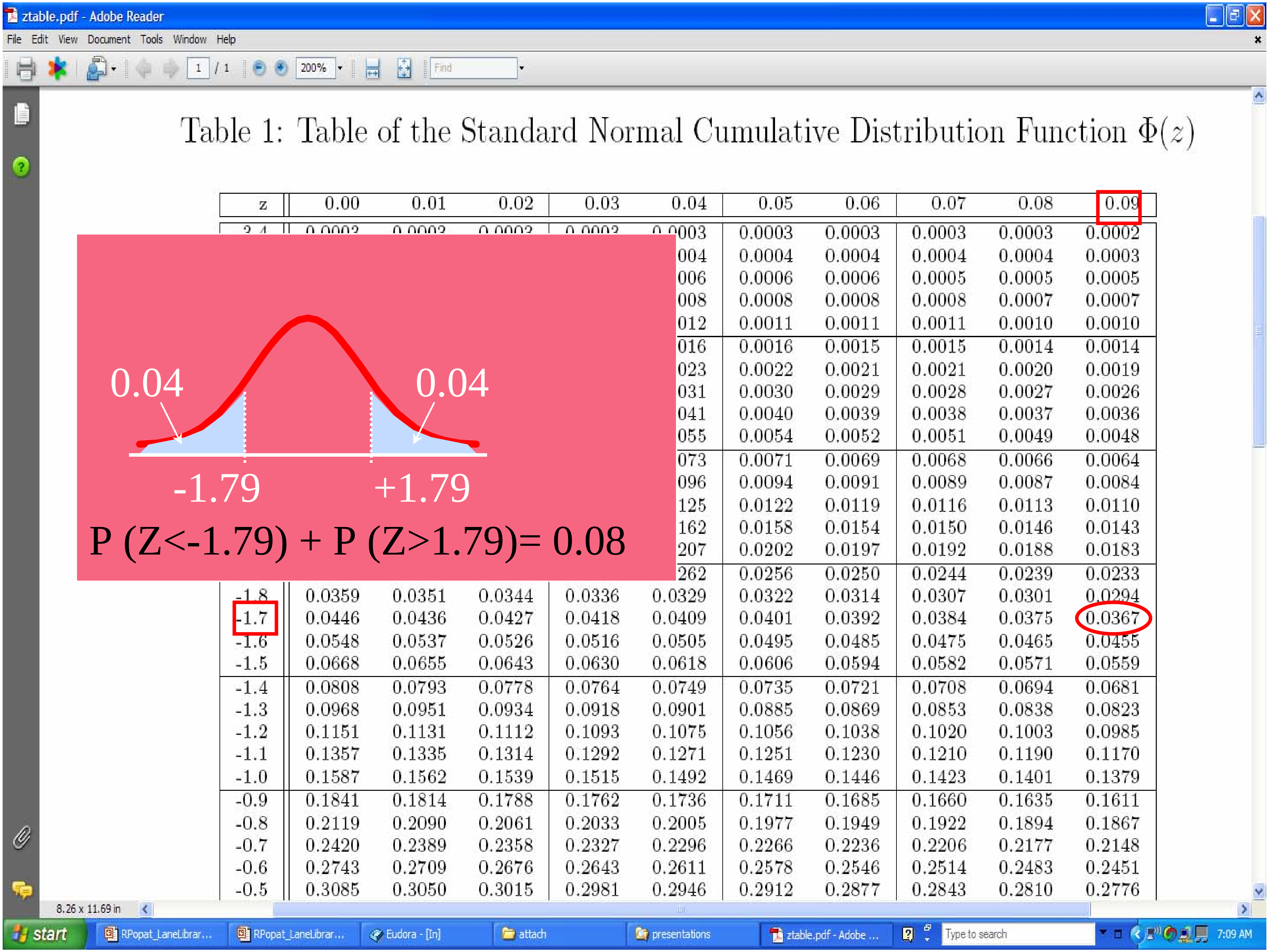


Table 1: Table of the Standard Normal Cumulative Distribution Function $\Phi(z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
					004	0.0004	0.0004	0.0004	0.0004	0.0003
					006	0.0006	0.0006	0.0005	0.0005	0.0005
					008	0.0008	0.0008	0.0008	0.0007	0.0007
					012	0.0011	0.0011	0.0011	0.0010	0.0010
					016	0.0016	0.0015	0.0015	0.0014	0.0014
					023	0.0022	0.0021	0.0021	0.0020	0.0019
					031	0.0030	0.0029	0.0028	0.0027	0.0026
					041	0.0040	0.0039	0.0038	0.0037	0.0036
					055	0.0054	0.0052	0.0051	0.0049	0.0048
					073	0.0071	0.0069	0.0068	0.0066	0.0064
					096	0.0094	0.0091	0.0089	0.0087	0.0084
					125	0.0122	0.0119	0.0116	0.0113	0.0110
					162	0.0158	0.0154	0.0150	0.0146	0.0143
					207	0.0202	0.0197	0.0192	0.0188	0.0183
					262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776

P (Z<-1.79) + P (Z>1.79)= 0.08

What is a *P* value?

- ‘P’ stands for probability
  - Tail area probability based on the observed effect
  - Calculated as the probability of an effect as large as or larger than the observed effect (more extreme in the tails of the distribution), assuming null hypothesis is true

- Measures the strength of the evidence against the null hypothesis
  - Smaller P values indicate stronger evidence against the null hypothesis

What is a *P* value?

- Fisher suggested 5% level ($p<0.05$) could be used as a scientific benchmark for concluding that fairly strong evidence exists against H_0
    - Was never intended as an absolute threshold
    - Strength of evidence is on a continuum
    - Simply noting the magnitude of the P-value should suffice
    - Scientific context is critical

- By convention, p-values of <.05 are often accepted as “statistically significant” in the medical literature; but this is an arbitrary cut-off.

What is a *P* value?

- $P<0.05$ is an arbitrary cut-point
  - Does it make sense to adopt a therapeutic agent because P-value obtained in a RCT was 0.049, and at the same time ignore results of another therapeutic agent because P-value was 0.051?

- Hence important to report the exact p-value and not ≤ 0.05 or >0.05

P-values

Trial	Number dead / randomized		Risk Ratio	95% C.I.	P value
	Intravenous nitrate	Control			
Chiche	**3/50**	**8/45**	**0.33**	**(0.09,1.13)**	**0.08**
Flaherty	**11/56**	**11/48**	**0.83**	**(0.33,2.12)**	**0.70**
Lis	**5/64**	**10/76**	**0.56**	**(0.19,1.65)**	**0.29**
Jugdutt	**24/154**	**44/156**	**0.48**	**(0.28, 0.82)**	**0.007**

Some evidence against the null hypothesis

Very weak evidence against the null hypothesis…very likely a chance finding

Very strong evidence against the null hypothesis…very unlikely to be a chance finding

Interpreting *P* values
If the null hypothesis were true…

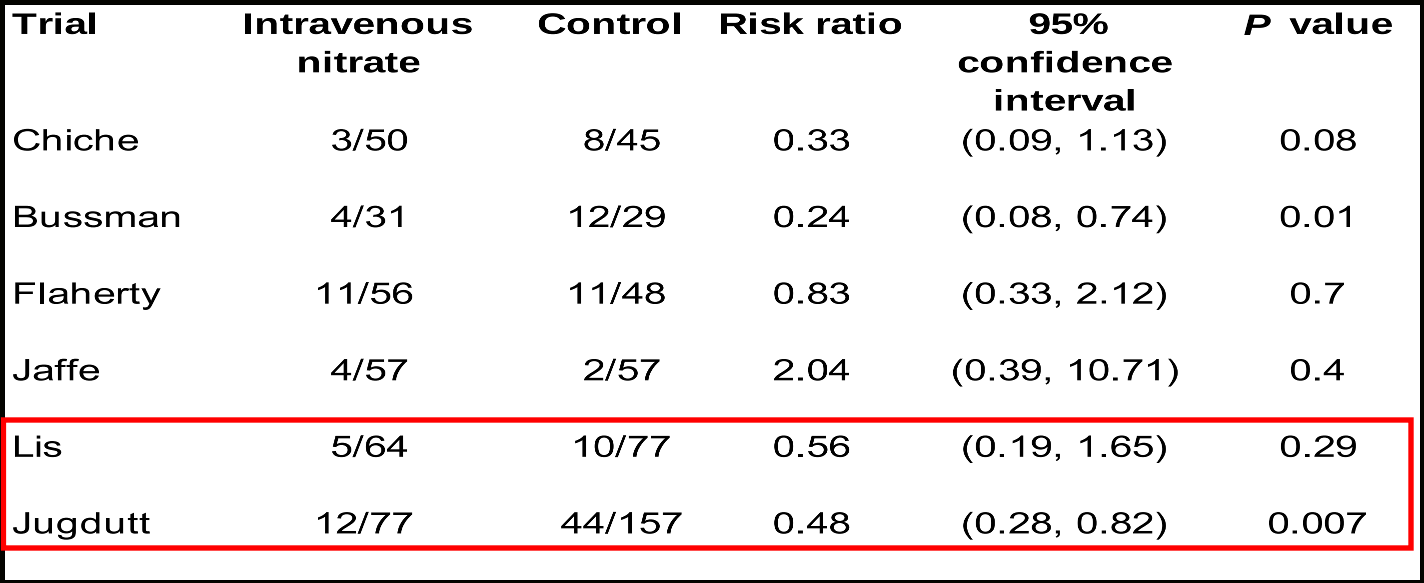
	Number dead / randomized				
Trial	Intravenous nitrate	Control	Risk Ratio	95% C.I.	P value
Chiche	**3/50**	**8/45**	**0.33**	**(0.09,1.13)**	**0.08**
Flaherty	**11/56**	**11/48**	**0.83**	**(0.33,2.12)**	**0.70**
Lis	**5/64**	**10/76**	**0.56**	**(0.19,1.65)**	**0.29**
Jugdutt	**24/154**	**44/156**	**0.48**	**(0.28, 0.82)**	**0.007**

…8 out of 100 such trials would show a risk reduction of 66% or more extreme just by chance

…70 out of 100 such trials would show a risk reduction of 17% or more extreme just by chance…very likely a chance finding

Very unlikely to be a chance finding

Interpreting *P* values



Trial	Intravenous nitrate	Control	Risk ratio	95% confidence interval	*P* value
Chiche	3/50	8/45	0.33	(0.09, 1.13)	0.08
Bussman	4/31	12/29	0.24	(0.08, 0.74)	0.01
Flaherty	11/56	11/48	0.83	(0.33, 2.12)	0.7
Jaffe	4/57	2/57	2.04	(0.39, 10.71)	0.4
Lis	5/64	10/77	0.56	(0.19, 1.65)	0.29
Jugdutt	12/77	44/157	0.48	(0.28, 0.82)	0.007

- Size of the p-value is related to the sample size

- Lis and Jugdutt trials are similar in effect (~ 50% reduction in risk)…but Jugdutt trial has a large sample size

Interpreting *P* values

Trial	**Intravenous nitrate**	**Control**	**Risk ratio**	**95% confidence interval**	***P* value**
Chiche	3/50	8/45	0.33	(0.09, 1.13)	0.08
Bussman	4/31	12/29	0.24	(0.08, 0.74)	0.01
Flaherty	11/56	11/48	0.83	(0.33, 2.12)	0.7
Jaffe	4/57	2/57	2.04	(0.39, 10.71)	0.4
Lis	5/64	10/77	0.56	(0.19, 1.65)	0.29
Jugdutt	12/77	44/157	0.48	(0.28, 0.82)	0.007

- Size of the p-value is related to the effect size or the observed association or difference
- Chiche and Flaherty trials approximately same size, but observed difference greater in the Chiche trial

P values

- P values give no indication about the clinical importance of the observed association

- A very large study may result in very small p-value based on a small difference of effect that may not be important when translated into clinical practice

- Therefore, important to look at the effect size and confidence intervals…

Confidence intervals

"Statistics means never having to say you're certain!"

- P values give no indication about the clinical importance of the observed association
- Relying on information from a sample will always lead to some level of uncertainty.
- Confidence interval is a range of values that tries to quantify this uncertainty:
  - For example , 95% CI means that under repeated sampling 95% of CIs would contain the true population parameter

P-values versus Confidence intervals

- P-value answers the question...
  - "Is there a statistically significant difference between the two treatments?“

- The point estimate and its confidence interval answers the question...
  - "What is the size of that treatment difference?", and "How precisely did this trial determine or estimate the treatment difference?"

Computing confidence intervals (CI)

- General formula:

  (Sample statistic) ± [(confidence level) × (measure of how high the sampling variability is)]

- Sample statistic: observed magnitude of effect or association (e.g., odds ratio, risk ratio)

- Confidence level: varies – 90%, 95%, 99%. For example, to construct a 95% CI, $Z_{\alpha/2} = 1.96$

- Sampling variability: Standard error (S.E.) of the estimate is a measure of variability

$$\boxed{\text{Point estimate} \pm (Z_{\alpha/2} \times \text{S.E.})}$$

Confidence intervals

- The above picture shows 50 realisations of a confidence interval for μ.
- Each 95% confidence interval has fixed endpoints, where μ might be in between (or not).
- There is no probability of such an event!

Confidence intervals

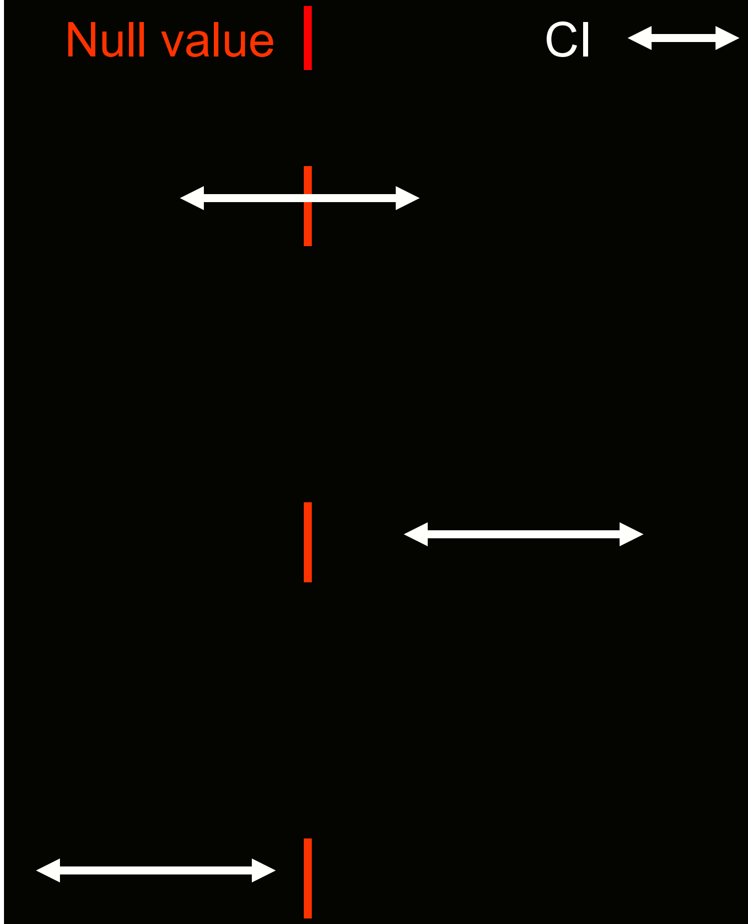
- Suppose α =0.05, we **cannot** say: "with probability 0.95 the parameter μ lies in the confidence interval."
- We only know that by repetition, 95% of the intervals will contain the true population parameter (μ)
- In 5 % of the cases however it doesn't. And unfortunately we don't know in which of the cases this happens.
- That's why we say: with **confidence level** 100(1 − α) % μ lies in the confidence interval."

Interpretation of Confidence intervals

- Width of the confidence interval (CI)
    - A narrow CI implies high precision
    - A wide CI implies poor precision (usually due to inadequate sample size)

- Does the interval contain a value that implies no change or no effect or no association?
    - CI for a difference between two means: Does the interval include 0 (zero)?
    - CI for a ratio (e.g, OR, RR): Does the interval include 1?

Interpretation of Confidence intervals



No statistically significant change

Statistically significant increase

Statistically significant decrease

Connection between P-values and CIs

- If a 95% CI includes the null effect, the P-value is >0.05 (and we would fail to reject the null hypothesis)

- If the 95% CI excludes the null effect, the P-value is <0.05 (and we would reject the null hypothesis)

Interpreting confidence intervals

	Number dead / randomized				
Trial	Intravenous nitrate	Control	Risk Ratio	95% C.I.	P value
Chiche	**3/50**	**8/45**	**0.33**	**(0.09,1.13)**	**0.08**

Wide interval: suggests reduction in mortality of 91% and an increase of 13%

Flaherty	**11/56**	**11/48**	**0.83**	**(0.33,2.12)**	**0.70**
Jaffe	**4/57**	**2/57**	**2.04**	**(0.39,10.71)**	**0.40**

Reduction in mortality as little as 18%, but little evidence to suggest that IV nitrate is harmful

Jugdutt	**24/154**	**44/156**	**0.48**	**(0.28, 0.82)**	**0.007**

Table adapted from Whitley and Ball. Critical Care; 6(3):222-225, 2002

What about clinical importance?

> "A difference, to be a difference, must make a difference." -- Gertrude Stein

- Does the confidence interval lie partly or entirely within a range of clinical indifference?
- Clinical indifference represents values of such a trivial size that you do not want to change your current practice
    - E.g., would you recommend a cholesterol-lowering drug that reduced LDL levels by 2 units in one year?

What about clinical importance?

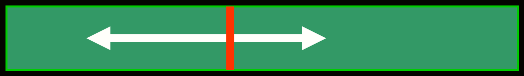
- Clinical importance is a medical judgment, not statistical!

- Clinicians should change practice only if they believe the study has definitively demonstrated a treatment difference and that the treatment difference is large enough to be clinically important.

- Depends on your knowledge of
    - a range of possible treatments
    - their costs
    - their side effects

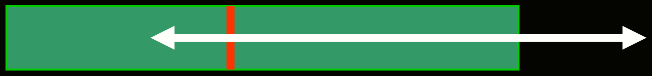
Interpretation of Confidence intervals

Null value |　　CI ↔



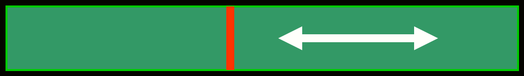
Range of clinical indifference

Keep doing things the same way!



Range of clinical indifference

Sample size too small?

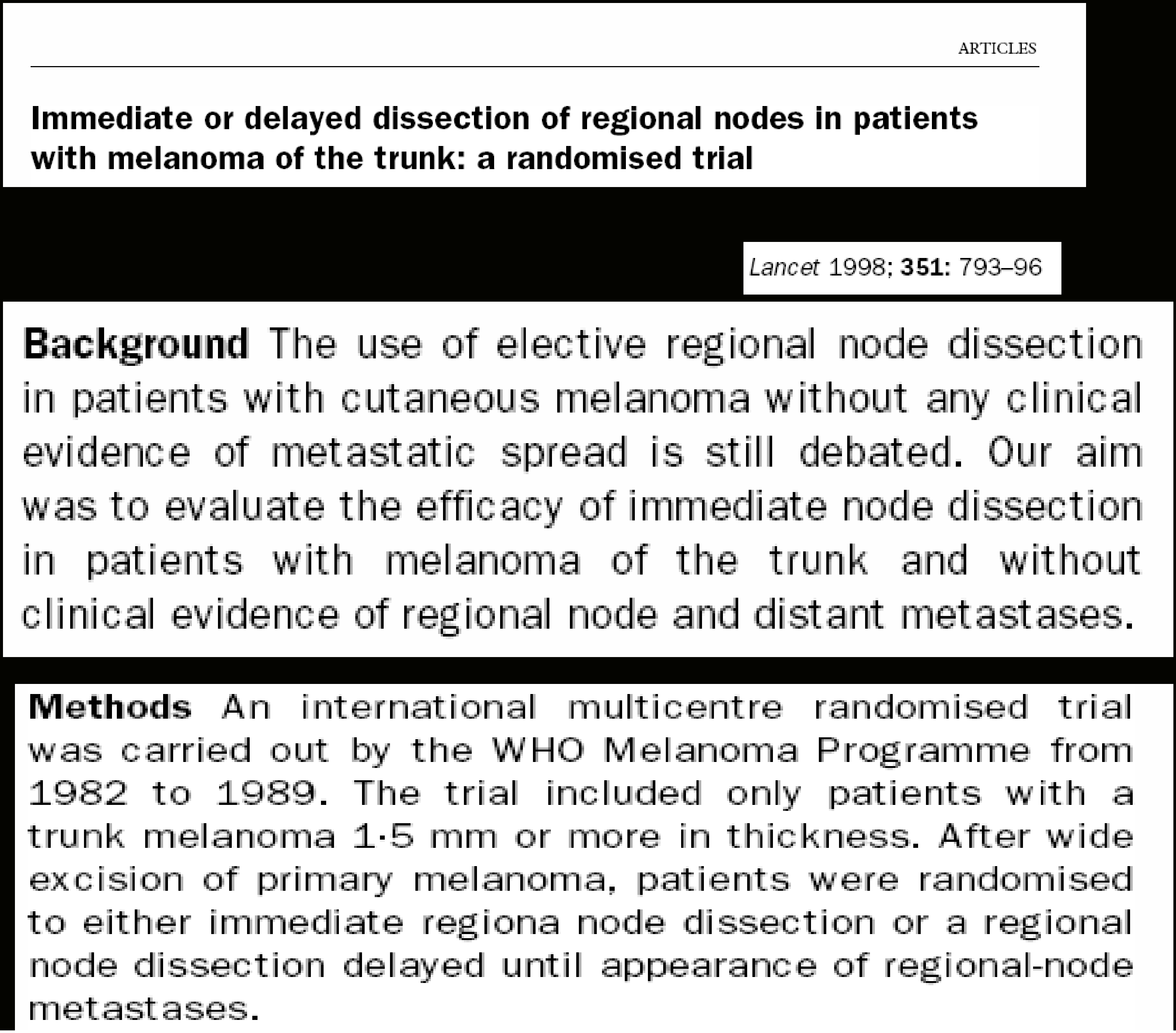


Range of clinical indifference

Statistically significant but no practical significance

Range of clinical indifference

Statistically significant and practical significance



Immediate or delayed dissection of regional nodes in patients with melanoma of the trunk: a randomised trial

Lancet 1998; **351:** 793–96

Background The use of elective regional node dissection in patients with cutaneous melanoma without any clinical evidence of metastatic spread is still debated. Our aim was to evaluate the efficacy of immediate node dissection in patients with melanoma of the trunk and without clinical evidence of regional node and distant metastases.

Methods An international multicentre randomised trial was carried out by the WHO Melanoma Programme from 1982 to 1989. The trial included only patients with a trunk melanoma 1·5 mm or more in thickness. After wide excision of primary melanoma, patients were randomised to either immediate regiona node dissection or a regional node dissection delayed until appearance of regional-node metastases.

Lancet 1998; **351:** 793–96

ARTICLES

Immediate or delayed dissection of regional nodes in patients with melanoma of the trunk: a randomised trial

"Multivariate analysis showed that routine use of immediate node dissection had no impact on survival (hazard ratio 0·72,95% CI 0·5–1·02), whilst the status of regional nodes affected survival significantly (p=0·007)."

Do you agree with the authors interpretation of the results?

Lancet 1998; **351:** 793–96

ARTICLES

Immediate or delayed dissection of regional nodes in patients with melanoma of the trunk: a randomised trial

"Multivariate analysis showed that routine use of immediate node dissection had no impact on survival (hazard ratio 0·72,95% CI 0·5–1·02), whilst the status of regional nodes affected survival significantly (p=0·007)."

-- In isolation,the p value tells us that the result was not ("statistically") significant.

-- The point estimate of 0·72 and p-value of 0·07 suggest that the result (or a result even more extreme) is consistent with a relative survival benefit of 28%, and that the probability of the result being due to chance is small in comparison.

--The 95% CI of 0.49-1.04 helps to shed more light…

- there might be a survival benefit of up to 50% for immediate dissection (a hazard ratio for death of 0·5).
- there might be a survival detriment of up to 2% (a hazard ratio of 1·02).

Lancet 1998; **351:** 793–96

ARTICLES

Immediate or delayed dissection of regional nodes in patients with melanoma of the trunk: a randomised trial

"Multivariate analysis showed that routine use of immediate node dissection had no impact on survival (hazard ratio 0·72,95% CI 0·5–1·02), whilst the status of regional nodes affected survival significantly (p=0·007)."

- The results of the study are therefore inconclusive, but we cannot rule out a clinically relevant survival advantage.

- With a larger study population, a survival benefit would probably be confirmed statistically.

Clinical vs statistical significance

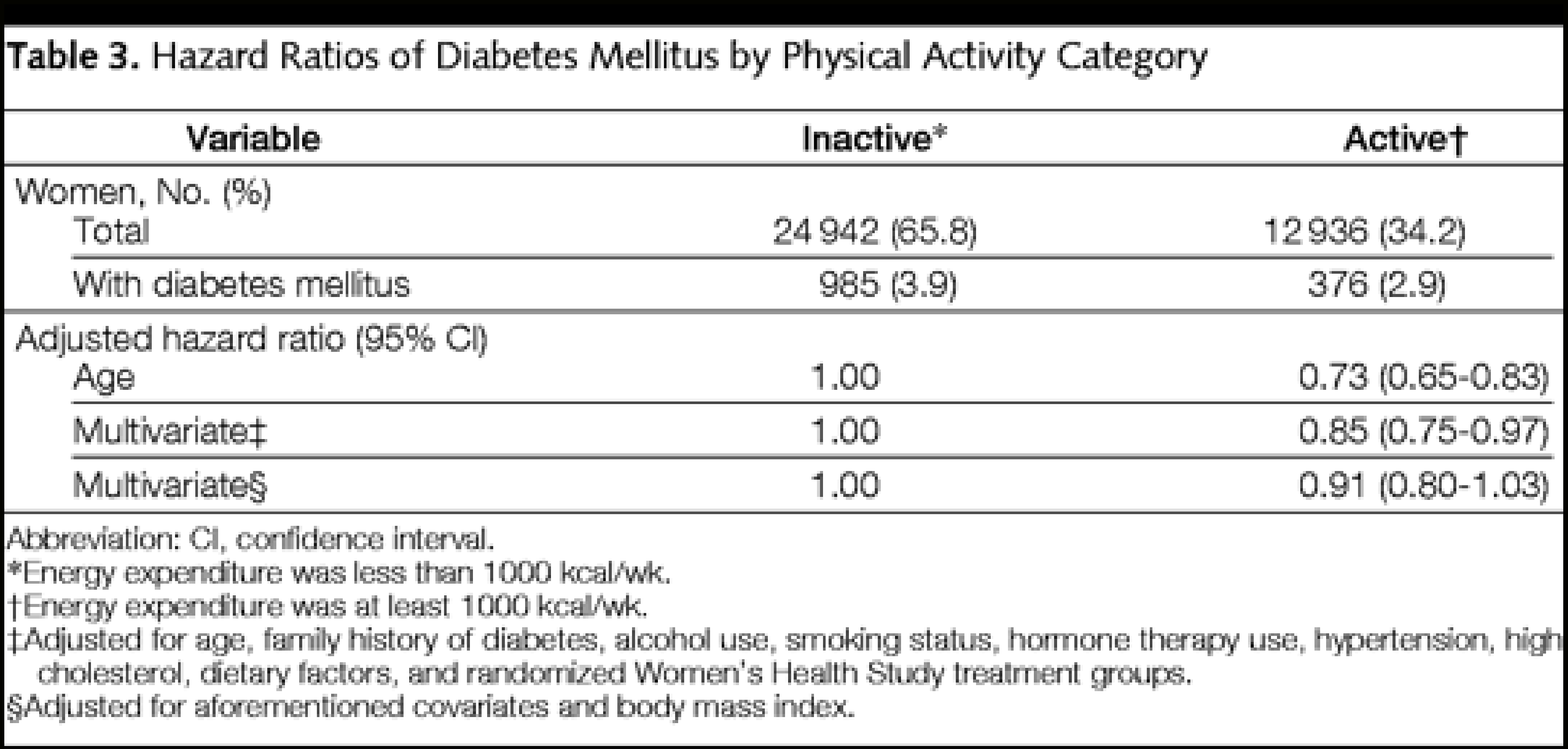


Table 3. Hazard Ratios of Diabetes Mellitus by Physical Activity Category

Variable	Inactive*	Active†
Women, No. (%)		
Total	24 942 (65.8)	12 936 (34.2)
With diabetes mellitus	985 (3.9)	376 (2.9)
Adjusted hazard ratio (95% CI)		
Age	1.00	0.73 (0.65-0.83)
Multivariate‡	1.00	0.85 (0.75-0.97)
Multivariate§	1.00	0.91 (0.80-1.03)

Abbreviation: CI, confidence interval.

*Energy expenditure was less than 1000 kcal/wk.

†Energy expenditure was at least 1000 kcal/wk.

‡Adjusted for age, family history of diabetes, alcohol use, smoking status, hormone therapy use, hypertension, high cholesterol, dietary factors, and randomized Women's Health Study treatment groups.

§Adjusted for aforementioned covariates and body mass index.

Weinstein et al. JAMA 292:1188-94

Is physical activity associated with the risk of T2DM?

Clinical vs statistical significance

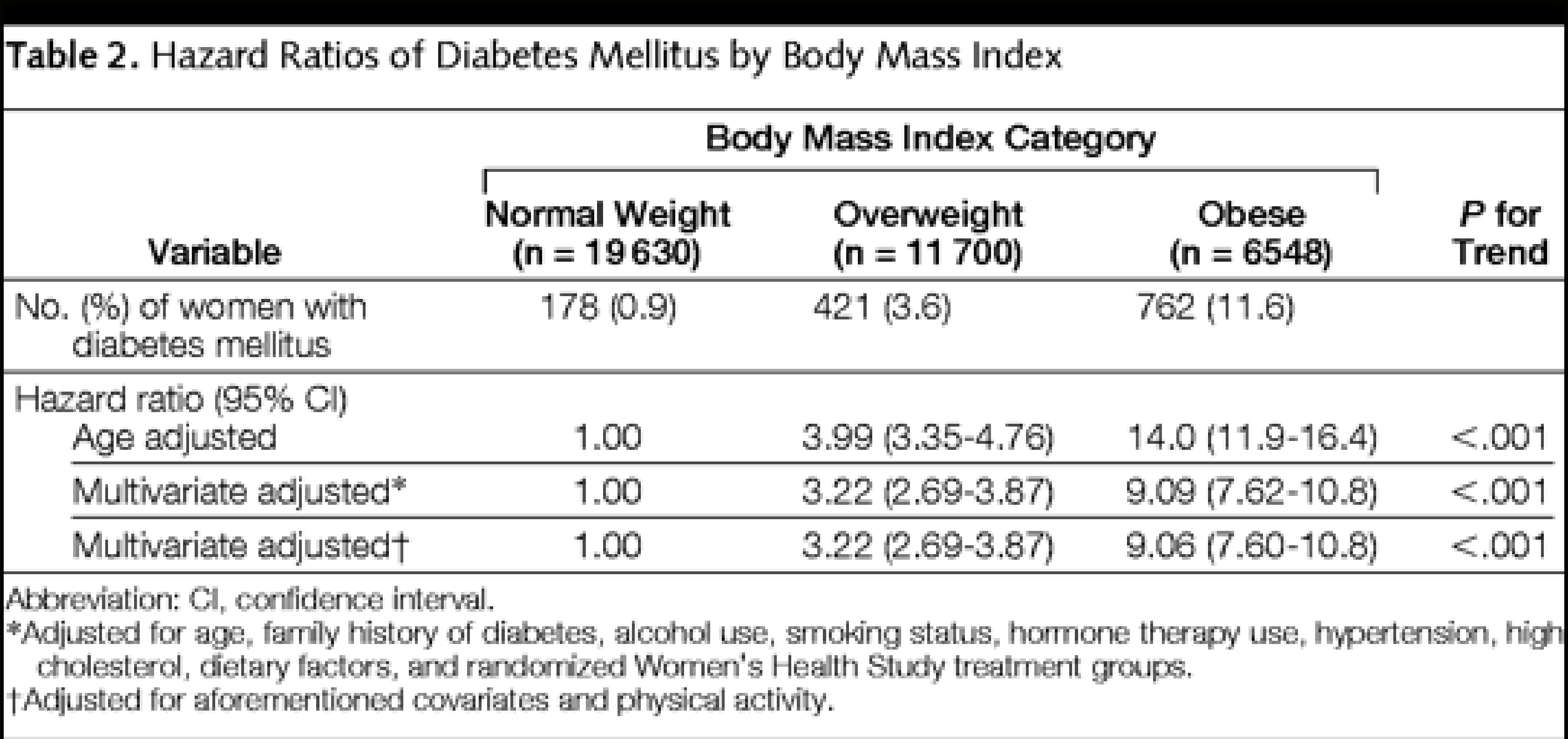


Table 2. Hazard Ratios of Diabetes Mellitus by Body Mass Index

	Body Mass Index Category			
Variable	**Normal Weight (n = 19 630)**	**Overweight (n = 11 700)**	**Obese (n = 6548)**	***P* for Trend**
No. (%) of women with diabetes mellitus	178 (0.9)	421 (3.6)	762 (11.6)	
Hazard ratio (95% CI)				
Age adjusted	1.00	3.99 (3.35-4.76)	14.0 (11.9-16.4)	<.001
Multivariate adjusted*	1.00	3.22 (2.69-3.87)	9.09 (7.62-10.8)	<.001
Multivariate adjusted†	1.00	3.22 (2.69-3.87)	9.06 (7.60-10.8)	<.001

Abbreviation: CI, confidence interval.

*Adjusted for age, family history of diabetes, alcohol use, smoking status, hormone therapy use, hypertension, high cholesterol, dietary factors, and randomized Women's Health Study treatment groups.

†Adjusted for aforementioned covariates and physical activity.

Weinstein et al. JAMA 292:1188-94

Is BMI associated with the risk of T2DM?

Reaction of investigator to results of a statistical significance test

		Statistical significance	
		Not significant	Significant
Practical importance of observed effect	Not important	Happy 🙂	Annoyed 😕
	Important	Very sad 😢	Elated 😃

Which statement(s) is/are correct?

The p-value is:

- The probability that my data are wrong.
- The probability of my data under the null hypothesis.
- The probability that I erroneously find an association.
- The probability that I find an association when one exists.

Which of the following odds ratios for the relationship between various risk factors and heart disease are statistically significant at the .05-significance level? Which are likely to be clinically significant?

	Statistically significant?	**Clinically significant?**
A. Odds ratio for every 1-year increase in age: 1.10 (95% CI: 1.01—1.19)	✓	✓
B. Odds ratio for regular exercise (yes vs. no): 0.50 (95% CI: 0.30—0.82)	✓	✓
C. Odds ratio for high blood pressure (high vs. normal): 3.0 (95% CI: 0.90—5.30)		✓
D. Odds ratio for every 50-pound increase in weight: 1.05 (95% CI: 1.01—1.20)	✓	

Summary of key points

- A P-value is a probability of obtaining an effect as large as or larger than the observed effect, assuming null hypothesis is true
  - Provides a measure of strength of evidence against the H_o
  - Does not provide information on magnitude of the effect
  - Affected by sample size and magnitude of effect: interpret with caution!
  - Cannot be used in isolation to inform clinical judgment

Summary of key points

- Confidence interval quantifies
  - How confident are we about the true value in the source population
  - Better precision with large sample size
  - Corresponds to hypothesis testing, but much more informative than P-value

- Keep in mind clinical importance when interpreting statistical significance!

Example: 95%CI for an odds ratio

	Cases	Control	
Exposure +	**20**	**10**	**30**
Exposure -	**6**	**24**	**30**

OR = (20 * 24) / (6*10) = 8.0

$$\ln(OR) \pm [1.96 \times S.E.(\ln OR)]$$

$$LCL = e^{\ln(8) - [1.96 \times \sqrt{\frac{1}{20}+\frac{1}{6}+\frac{1}{10}+\frac{1}{24}}]}$$

$$UCL = e^{\ln(8) + [1.96 \times \sqrt{\frac{1}{20}+\frac{1}{6}+\frac{1}{10}+\frac{1}{24}}]}$$

$$95\%\ CI = (8.0)e^{-1.96\sqrt{\frac{1}{20}+\frac{1}{6}+\frac{1}{10}+\frac{1}{24}}},\ (8.0)e^{+1.96\sqrt{\frac{1}{20}+\frac{1}{6}+\frac{1}{10}+\frac{1}{24}}} = (2.47,\ 25.8)$$