Statistics for Health Care

Unit 6:

Overview/Teasers

Overview

- Type I and Type II errors and statistical power; pitfalls of p-values
- Overview of statistical tests

Teaser 1, Unit 6

- A prospective cohort study of 34,079 women found that women who exercised >21 MET hours per week (≈60 minutes moderate -intensity exercise daily) gained **significantly less** weight than women who exercised <7.5 MET hours **(p<.001)**
- Widely covered in the media. Headlines:
  - “To Stay Trim, Women Need an Hour of Exercise Daily.”
  - “New Exercise Goal: 60 Minutes a Day”

- Physical Activity and Weight Gain Prevention. JAMA 2010;303:1173-1179.

How big was the effect?

- How much less weight do you think the high exercise group gained compared with the low exercise group over 3 years?
- Write down a guess!

Teaser 2, Unit 6

Abstract

OBJECTIVES:

The aim of the pilot study was to determine the efficacy of dietary n-3 PUFA docosahexaenoic acid (DHA) in patients with atopic eczema.

METHODS:

Fifty-three patients suffering from atopic eczema aged 18-40 years were recruited into this randomized, double-blind, controlled trial and received either DHA 5.4 g daily (n = 21) or an isoenergetic control of saturated fatty acids (n = 23) for 8 weeks. At weeks 0, 4, 8 and 20 the clinical outcome was assessed by the SCORAD (severity scoring of atopic dermatitis) index.

RESULTS:

DHA, but not the control treatment, resulted in a significant clinical improvement of atopic eczema in terms of a decreased SCORAD [DHA: baseline 37.0 (17.9-48.0), week 8 28.5 (17.6-51.0); control: baseline 35.4 (17.2-63.0), week 8 33.4 (10.7-56.2)].

What should we conclude from these results? Did DHA beat placebo?

Statistics in Medicine

Module 1:

Type I and type II errors

Hypothesis Testing

The Steps:

1. Define your hypotheses (null, alternative)
2. Specify your null distribution
3. Do an experiment
4. Calculate the p-value of what you observed
5. Reject or fail to reject the null hypothesis

Follows the logic: If A then B; not B; therefore, not A.

Summary: The Underlying Logic of hypothesis tests…

Follows this logic:

Assume A (Null hypothesis is true).

If A, then B.

Not B (If observation is not B…).

Therefore, Not A. (null hypothesis is not true)

But throw in a bit of uncertainty…If A, then probably B…

Error and Power

Note the sneaky conditionals

- Type-I Error (False positive) (also known as "α"):
  - Rejecting the null when the effect (observation) isn't real.
- Type-II Error (False negative) (also known as "β"):
  - Failing to reject the null when the effect (observation) is real.
- POWER (the flip side of type-II error: 1- β):
  - The probability of seeing a true effect if one exists.

…

Type I and Type II Error in a box

Your Statistical Decision	**True state of null hypothesis**	
	H_0 True (example: the vaccine doesn’t work)	H_0 False (example: the vaccine works)
Reject H_0 (ex: you conclude that the vaccine works)	***Type I error (α)*** ***(False positive)***	*Correct* *(True positive)*
Do not reject H_0 (ex: you conclude that there is insufficient evidence that the vaccine works, does not mean vaccine does not work)	*Correct* *(true negative)*	***Type II Error (β)*** ***(False negative)***

Error and Power

- Type I error rate (or significance level): the probability of finding an effect that isn't real (false positive).
    - If we require p-value<.05 for statistical significance, this means that we are permitting a false positive rate of 5% (1 in 20).
- Type II error rate: the probability of missing an effect (false negative).
- Statistical power: the probability of finding an effect if it is there (the probability of not making a type II error).
    - When we design studies, we typically aim for a power of 80% (allowing a false negative rate, or type II error rate, of 20%).

Statistical power (~ sensitivity in diagnosis)

Statistical power is the probability of finding an effect *if one exists*.

Note: In diagnosis (e.g. cancer imaging), sensitivity is often more important

Factors Affecting Power

1. Size of the effect (observation – mean)
2. Standard deviation of the characteristic
3. Bigger sample size
4. Significance level desired

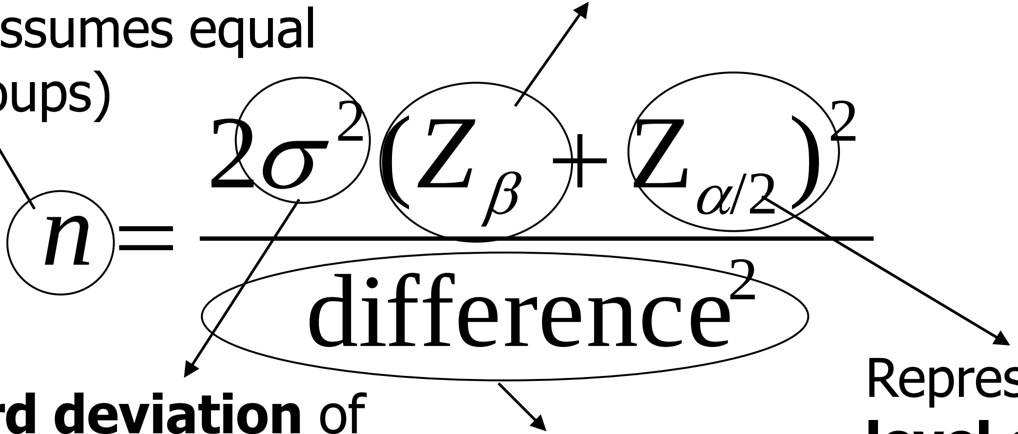
Sample size calculations

- Based on these elements, you can write formal mathematical equations that relates power, sample size, effect size, standard deviation, and significance level…

Example: formula for difference in means

Represents the **desired power** (typically .84 for 80% power).

Sample size in each group (assumes equal sized groups)



$$n = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

Standard deviation of the outcome variable

Effect Size (the difference in means)

Represents the desired **level of statistical significance** (typically 1.96).

Statistics in Medicine

Sample size formulas, derivations

Distribution, difference in means

- T-distribution (Z for n>100)
- Mean=true difference in means
- Standard error: $\boxed{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}$

Distribution, difference in proportions

- Z-distribution (normal(0, 1) distribution)
- Mean=true difference in proportions
- Standard error: $\boxed{\sqrt{\frac{p(1-p)}{n}+\frac{p(1-p)}{m}}}$

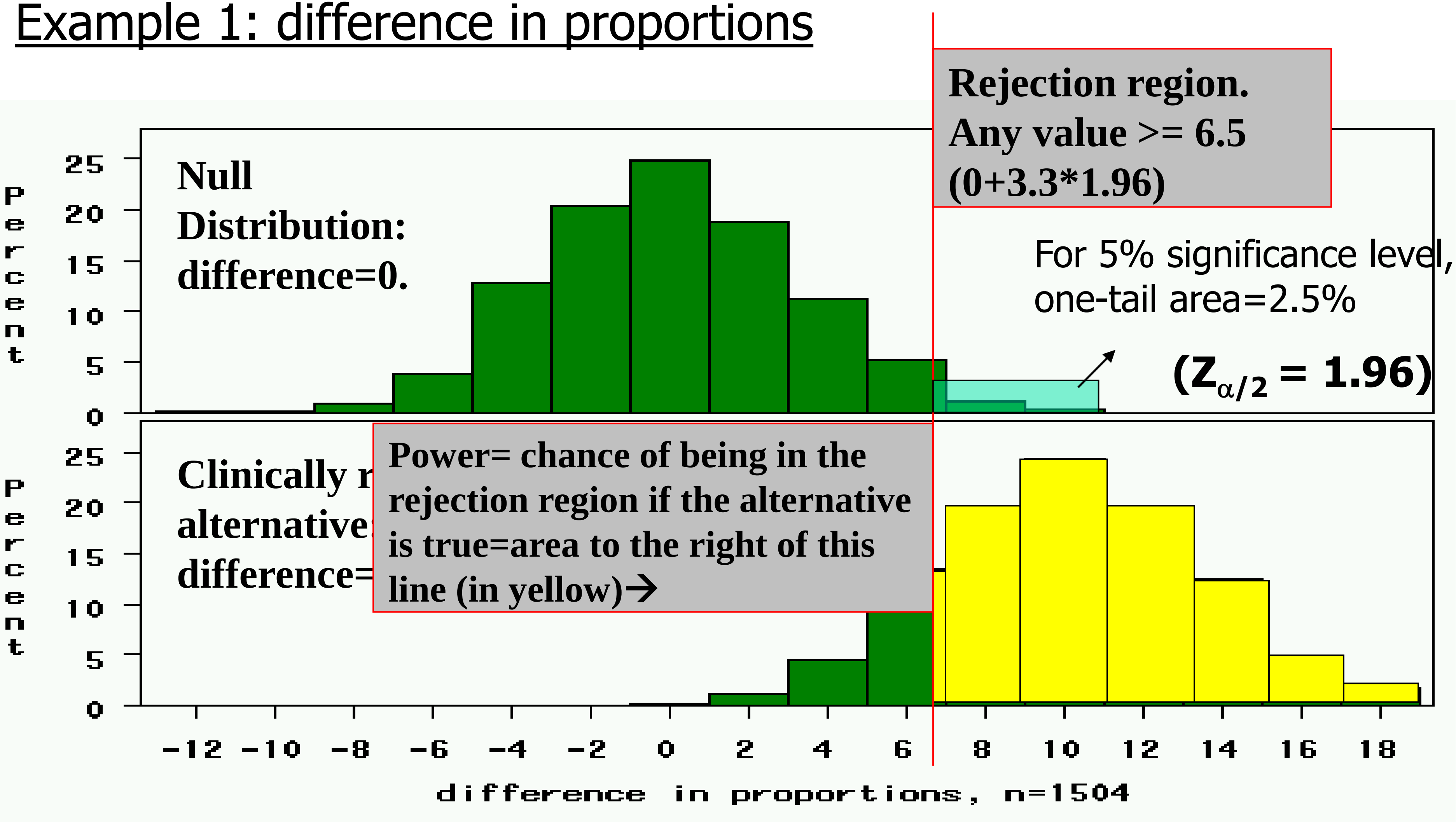
Power and sample size

Power = What’s the probability that we will correctly reject the null hypothesis when the alternative hypothesis is in fact true?

I.e., what’s the probability of detecting a real effect?

Can we quantify how much power we have for given sample sizes?

Example 1: difference in proportions



Null Distribution: difference=0.

Rejection region. Any value >= 6.5 (0+3.3*1.96)

For 5% significance level, one-tail area=2.5%

($Z_{\alpha/2}$ = 1.96)

Clinically r alternative: difference=

Power= chance of being in the rejection region if the alternative is true=area to the right of this line (in yellow)→

Example 1: difference in proportions

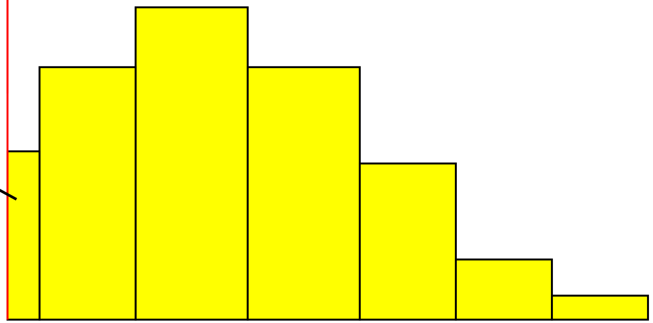
Rejection region.
Any value >= 6.5
(0+3.3*1.96)

Power= chance of being in the rejection region if the alternative is true=area to the right of this line (in yellow)

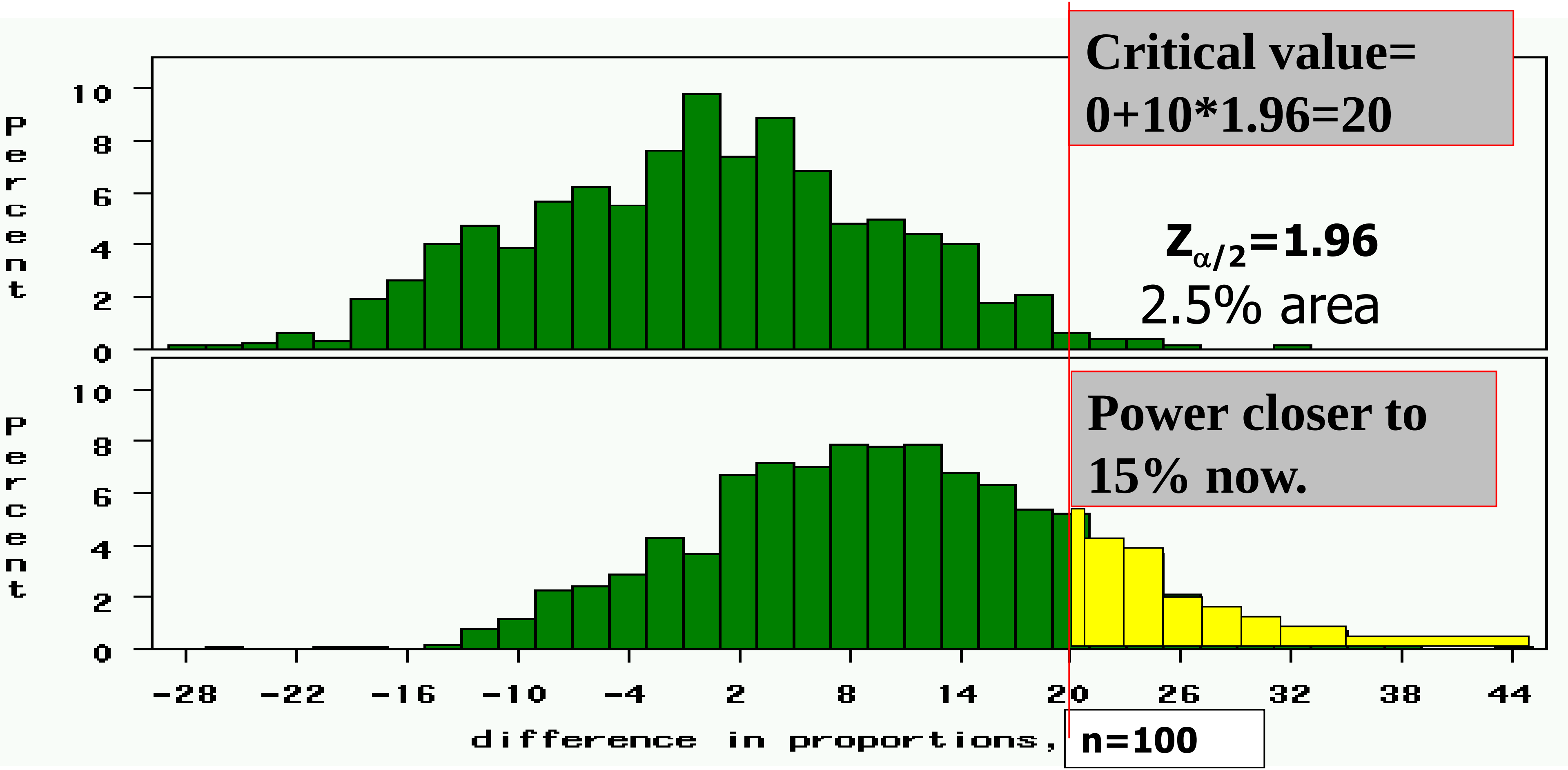
Power here:

$$P(Z > \frac{6.5-10}{3.3}) =$$

$$P(Z > -1.06) = 85\%$$



study 1: difference in proportions, smaller sample size



Critical value= 0+10*1.96=20

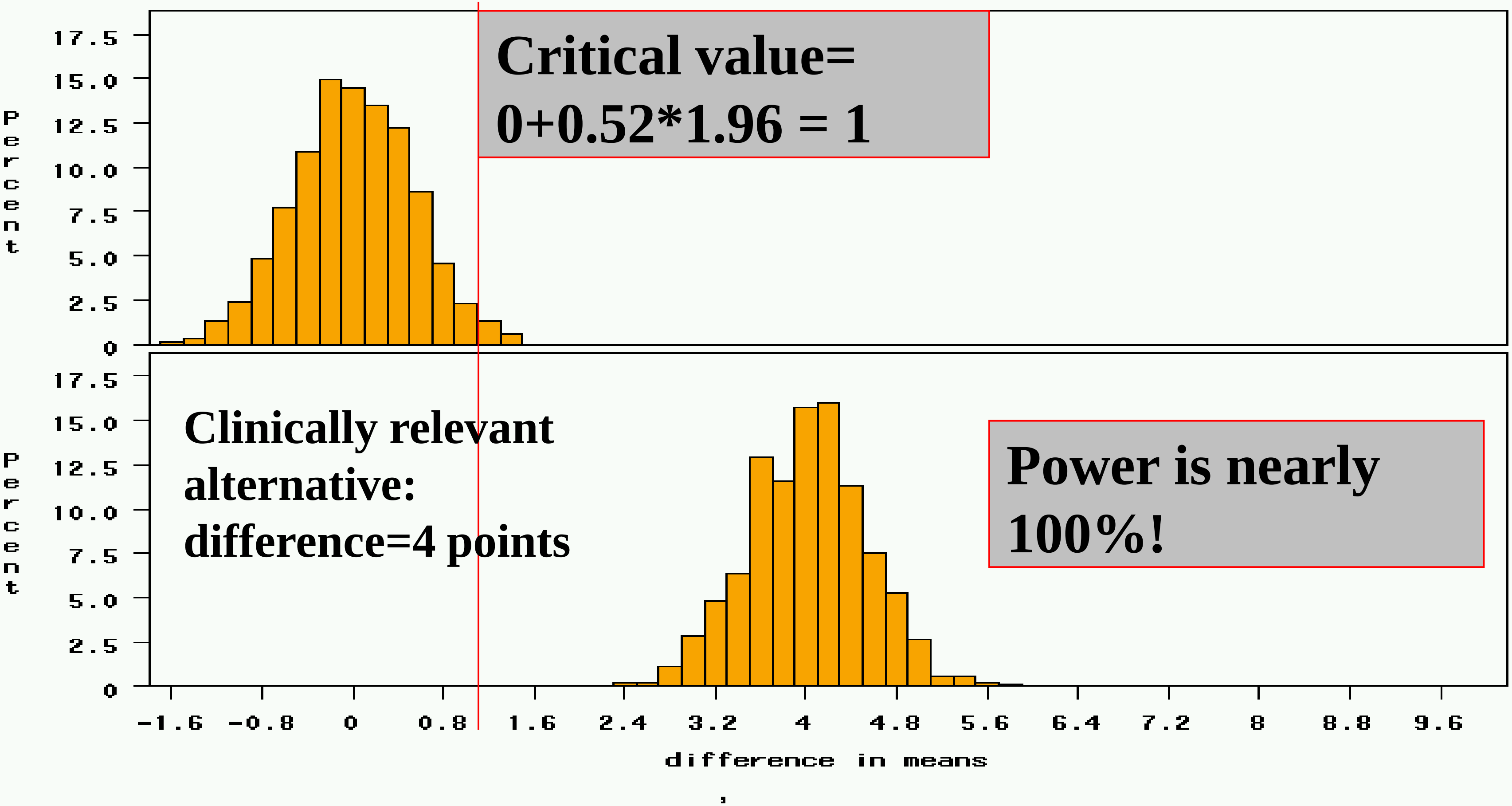
$\mathbf{z_{\alpha/2}=1.96}$

2.5% area

Power closer to 15% now.

n=100

Example 2: difference in means

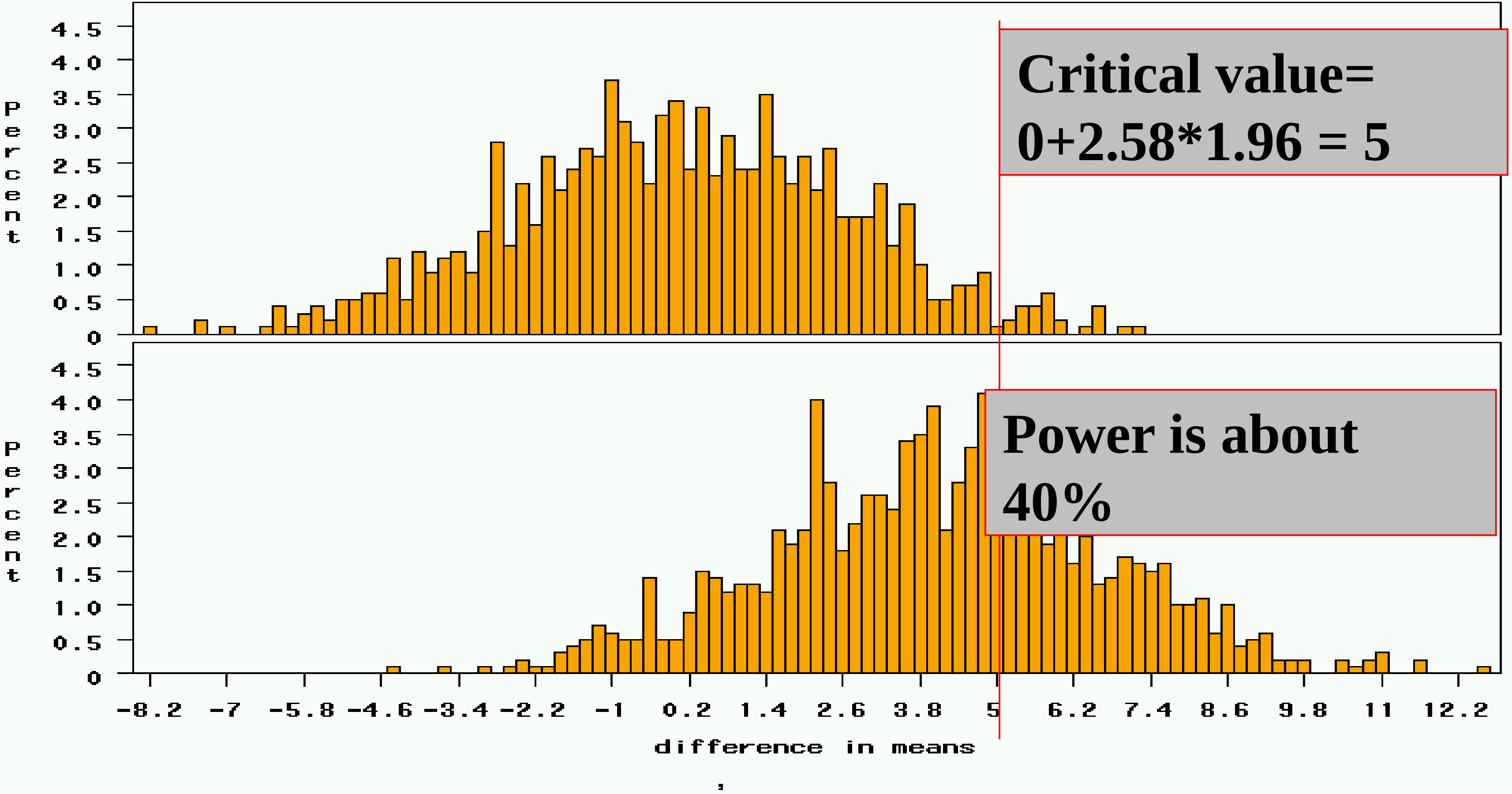


Critical value= 0+0.52*1.96 = 1

Clinically relevant alternative: difference=4 points

Power is nearly 100%!

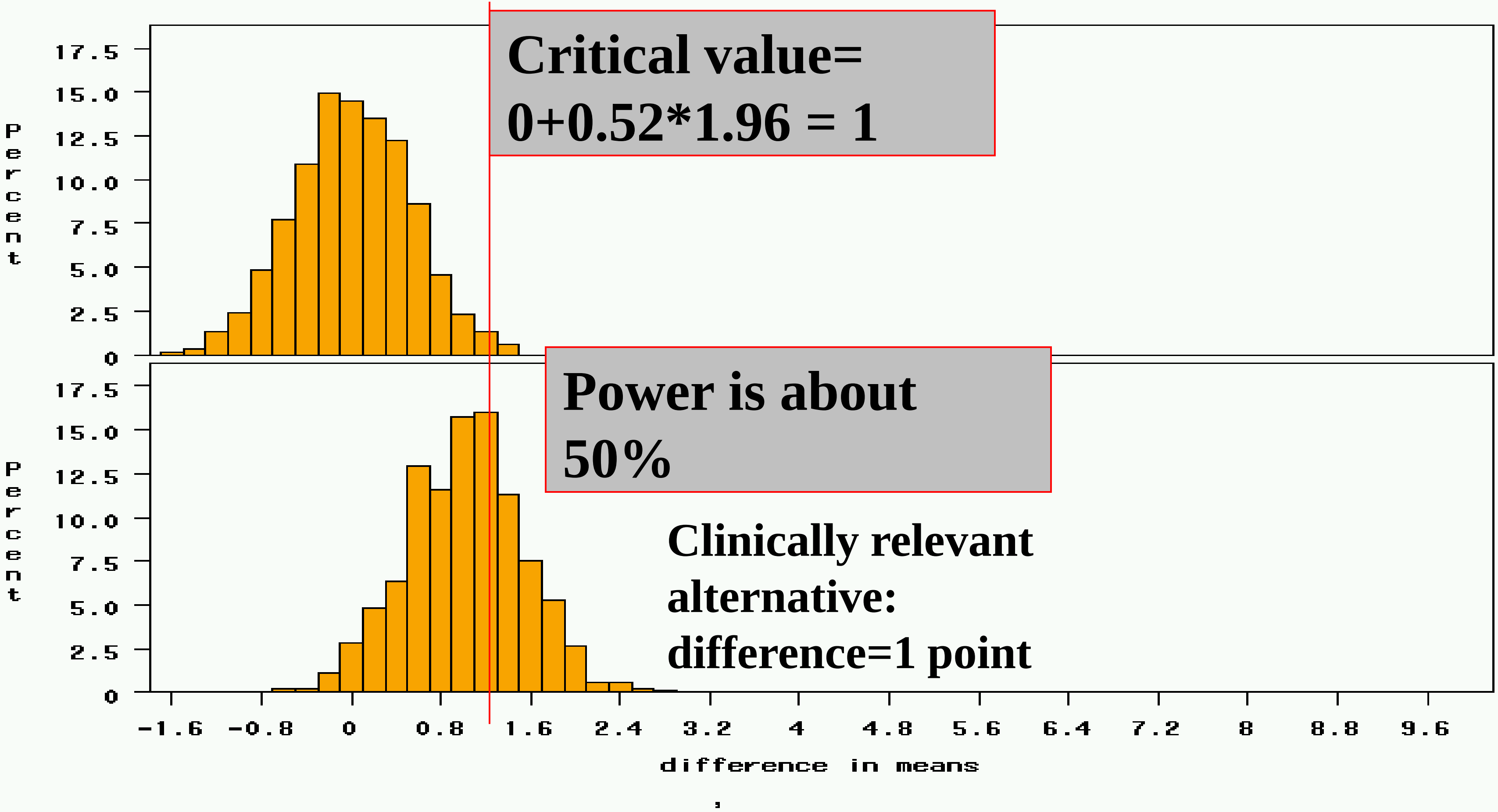
Example 2: difference in means, greater outcome variability



Critical value= 0+2.58*1.96 = 5

Power is about 40%

Example 2: difference in means, smaller effect size



Critical value= 0+0.52*1.96 = 1

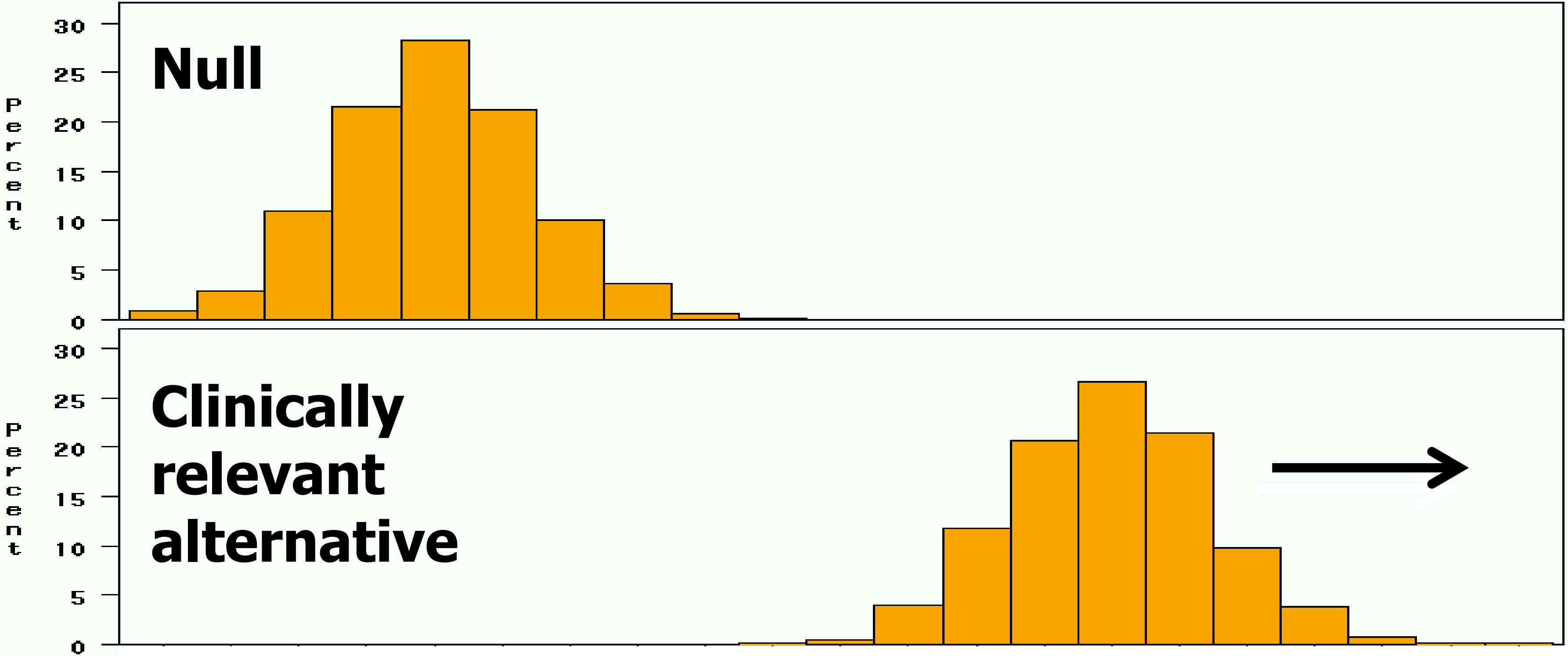
Power is about 50%

Clinically relevant alternative: difference=1 point

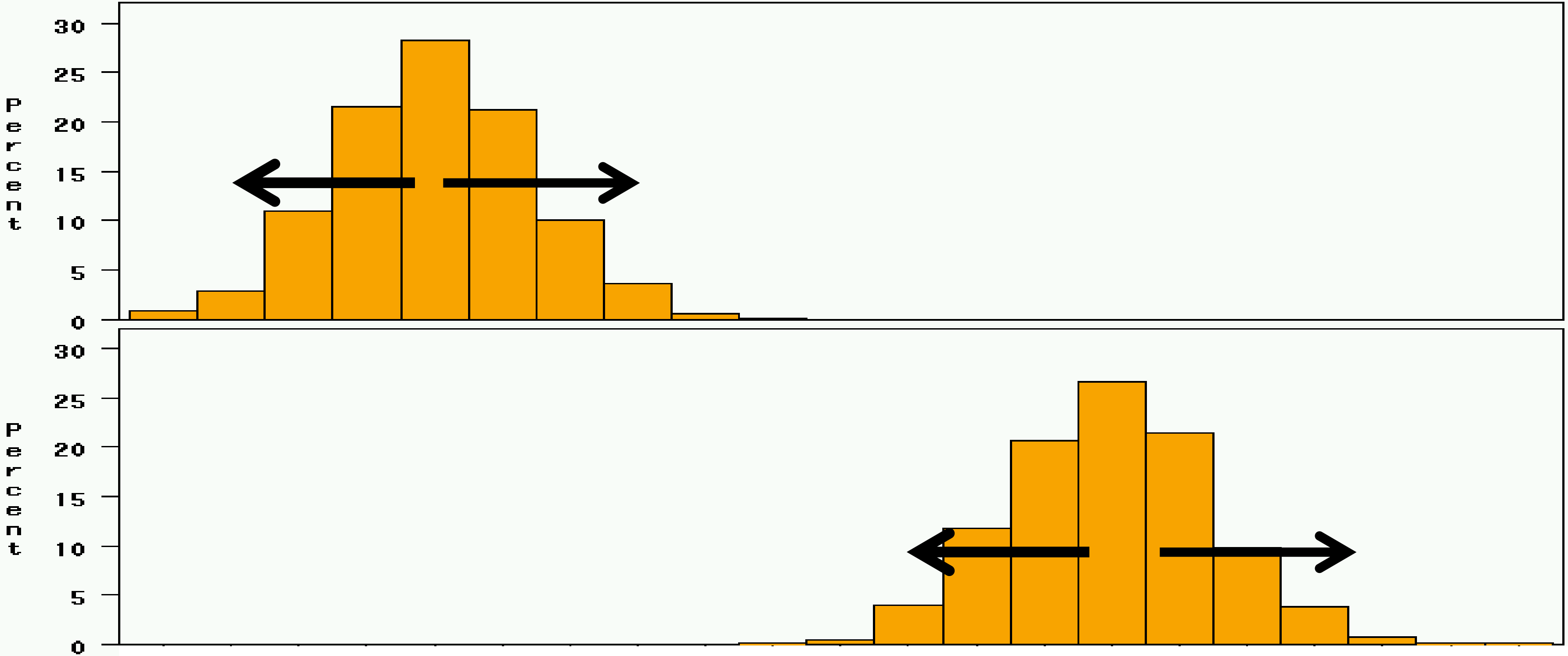
Factors Affecting Power

1. Size of the effect ↑
2. Standard deviation of the characteristic ↓
3. Bigger sample size ↑
4. Significance level desired ↓ (increase in number)

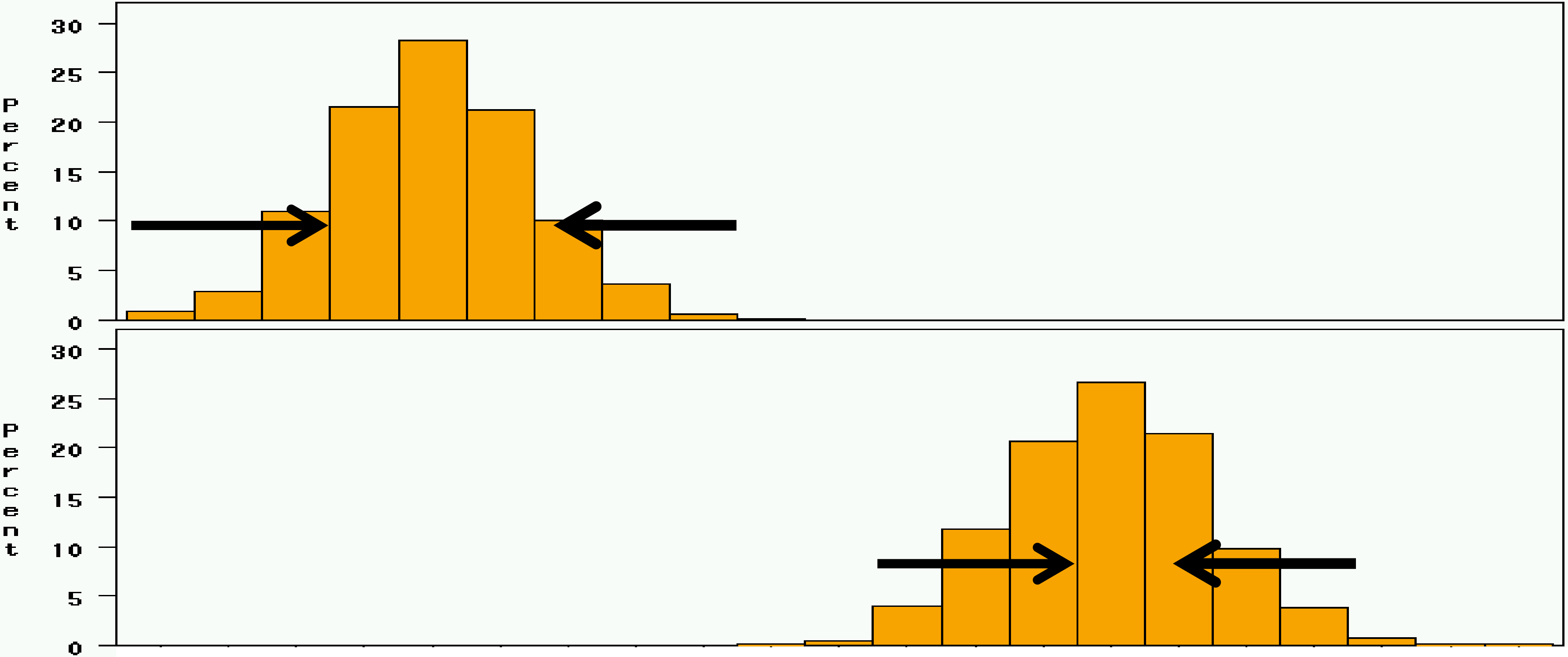
1. Bigger difference from the null mean



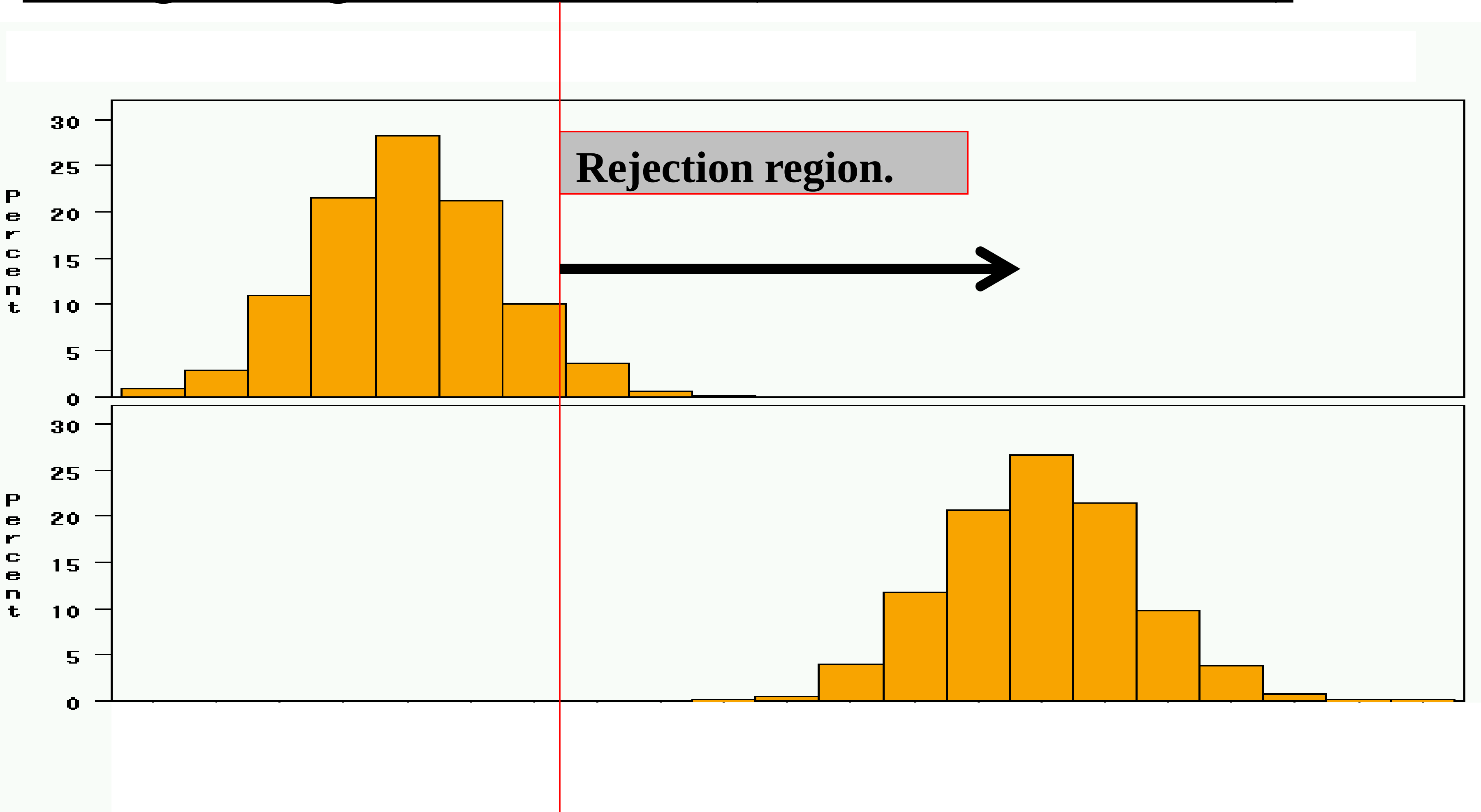
2. Bigger standard deviation



3. Bigger Sample Size



4. Higher significance level (decrease in number)



Sample size calculations

- Based on these elements, you can write a formal mathematical equation that relates power, sample size, effect size, standard deviation, and significance level…

Example: formula for difference in means

Sample size in each group (assumes equal sized groups)

Represents the **desired power** (typically .84 for 80% power).

$$n = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

Standard deviation of the outcome variable

Effect Size (the difference in means)

Represents the desired **level of statistical significance** (typically 1.96).

Example: formula for difference in proportions

Sample size in each group (assumes equal sized groups)

Represents the **desired power** (typically .84 for 80% power).

$$n = \frac{2\overline{p}(1-\overline{p})(Z_{\beta}+Z_{\alpha/2})^2}{(p_1-p_2)^2}$$

A measure of **Variability** of a proportion

Effect Size (difference in proportions)

Represents the desired **level of statistical significance** (typically 1.96).

Statistics in Medicine

Module 2:

P-value pitfalls: statistical vs. clinical significance

Statistical vs. clinical significance

- Trivial effects may achieve statistical significance if the sample size is large enough.

Example

- A prospective cohort study of 34,079 women found that women who exercised >21 MET hours per week (≈60 minutes moderate -intensity exercise daily) gained **significantly less** weight than women who exercised <7.5 MET hours **(p<.001)**
- Widely covered in the media. Headlines:
  - “To Stay Trim, Women Need an Hour of Exercise Daily.”
  - “New Exercise Goal: 60 Minutes a Day”

- Physical Activity and Weight Gain Prevention. JAMA 2010;303:1173-1179.

How big was the effect?

- How much less weight do you think the high exercise group gained compared with the low exercise group over 3 years?
- Guesses?

Mean (SD) Differences in Weight Over Any 3-Year Period by Physical Activity Level, Women's Health Study, 1992-2007a

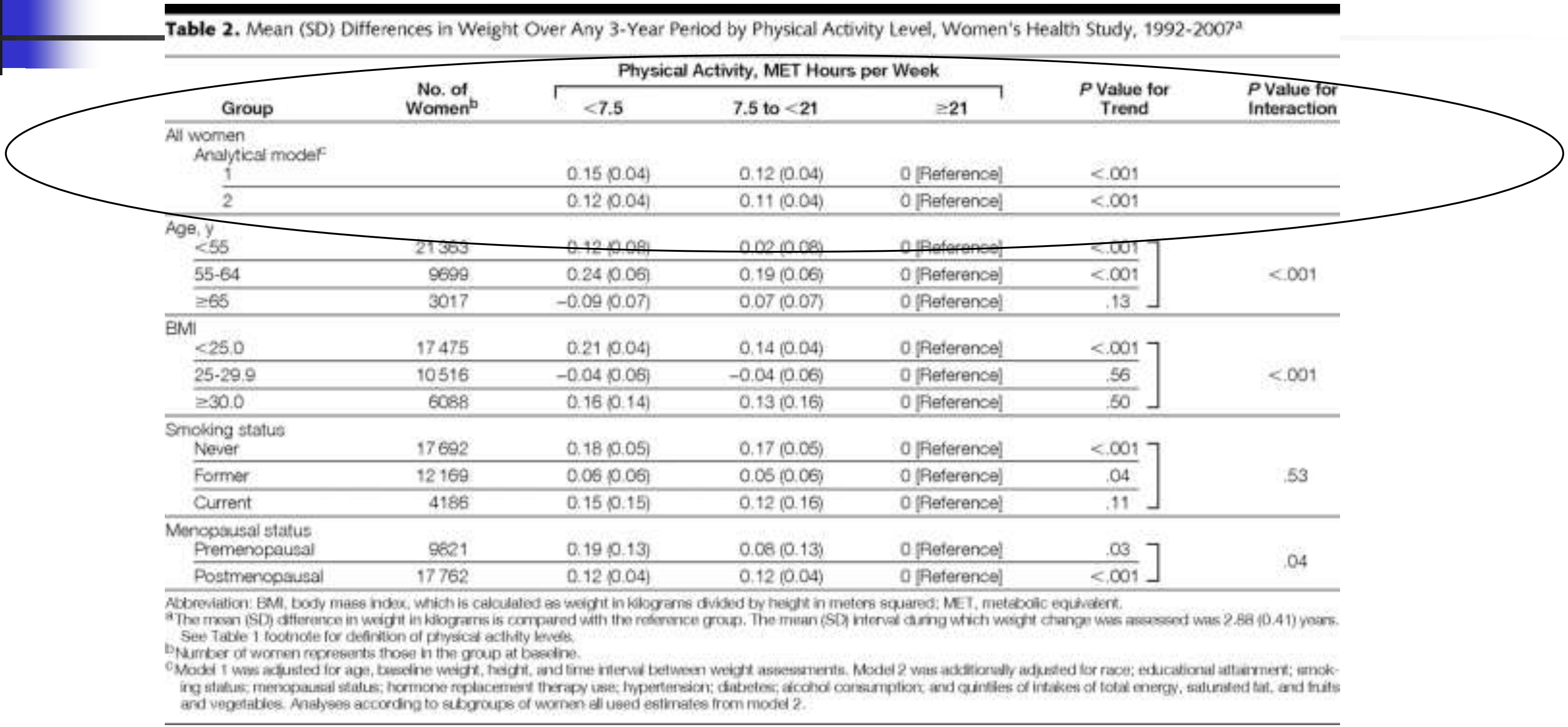


Table 2. Mean (SD) Differences in Weight Over Any 3-Year Period by Physical Activity Level, Women's Health Study, 1992-2007[a]

		Physical Activity, MET Hours per Week				
Group	No. of Women[b]	<7.5	7.5 to <21	≥21	*P* Value for Trend	*P* Value for Interaction
All women						
Analytical model[c]						
1		0.15 (0.04)	0.12 (0.04)	0 [Reference]	<.001	
2		0.12 (0.04)	0.11 (0.04)	0 [Reference]	<.001	
Age, y						
<55	21 363	0.12 (0.08)	0.02 (0.08)	0 [Reference]	<.001	<.001
55-64	9699	0.24 (0.06)	0.19 (0.06)	0 [Reference]	<.001	
≥65	3017	−0.09 (0.07)	0.07 (0.07)	0 [Reference]	.13	
BMI						
<25.0	17 475	0.21 (0.04)	0.14 (0.04)	0 [Reference]	<.001	<.001
25-29.9	10 516	−0.04 (0.06)	−0.04 (0.06)	0 [Reference]	.56	
≥30.0	6088	0.16 (0.14)	0.13 (0.16)	0 [Reference]	.50	
Smoking status						
Never	17 692	0.18 (0.05)	0.17 (0.05)	0 [Reference]	<.001	.53
Former	12 169	0.06 (0.06)	0.05 (0.06)	0 [Reference]	.04	
Current	4186	0.15 (0.15)	0.12 (0.16)	0 [Reference]	.11	
Menopausal status						
Premenopausal	9821	0.19 (0.13)	0.08 (0.13)	0 [Reference]	.03	.04
Postmenopausal	17 762	0.12 (0.04)	0.12 (0.04)	0 [Reference]	<.001	

Abbreviation: BMI, body mass index, which is calculated as weight in kilograms divided by height in meters squared; MET, metabolic equivalent.

[a] The mean (SD) difference in weight in kilograms is compared with the reference group. The mean (SD) interval during which weight change was assessed was 2.88 (0.41) years. See Table 1 footnote for definition of physical activity levels.

[b] Number of women represents those in the group at baseline.

[c] Model 1 was adjusted for age, baseline weight, height, and time interval between weight assessments. Model 2 was additionally adjusted for race; educational attainment; smoking status; menopausal status; hormone replacement therapy use; hypertension; diabetes; alcohol consumption; and quintiles of intakes of total energy, saturated fat, and fruits and vegetables. Analyses according to subgroups of women all used estimates from model 2.

Reproduced with permission from: Lee, I. M. et al. JAMA 2010;303:1173-1179.

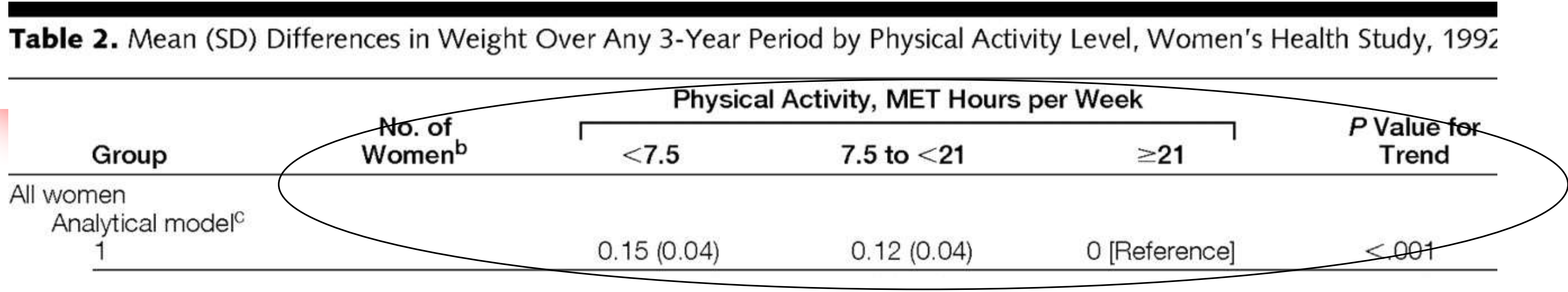


Table 2. Mean (SD) Differences in Weight Over Any 3-Year Period by Physical Activity Level, Women's Health Study, 1992

Group	No. of Women[b]	Physical Activity, MET Hours per Week			*P* Value for Trend
		<7.5	7.5 to <21	≥21	
All women					
Analytical model[c]					
1		0.15 (0.04)	0.12 (0.04)	0 [Reference]	<.001

- What was the effect size? Those who exercised the least gained **<u>0.15 kg</u>** more than those who exercised the most **<u>over 3 years</u>**.

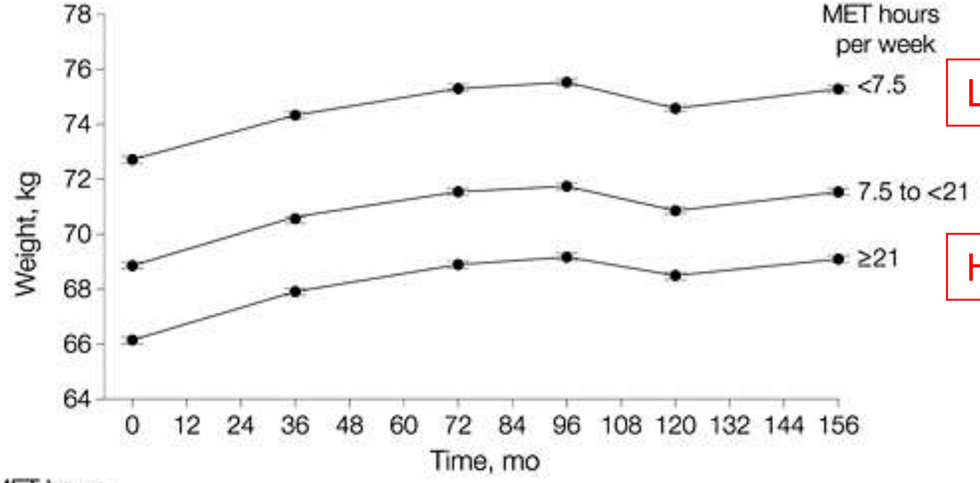
- Extrapolated **<u>over 13 years</u>** of the study, the high exercisers **<u>gained 0.65 kg (0.15 x 13/3) less</u>** than the low exercisers!

- *Classic example of a statistically significant effect that is not clinically significant.*

95% confidence interval

- Point estimate: 0.15 kg (over 3 years)
- 95% confidence interval: 0.07 to 0.23 kg
- Interpretation: The effect could plausibly be as large as a 0.23 kg reduction in weight gain over 3 years.

A picture is worth…



Low exercise group

Medium exercise group

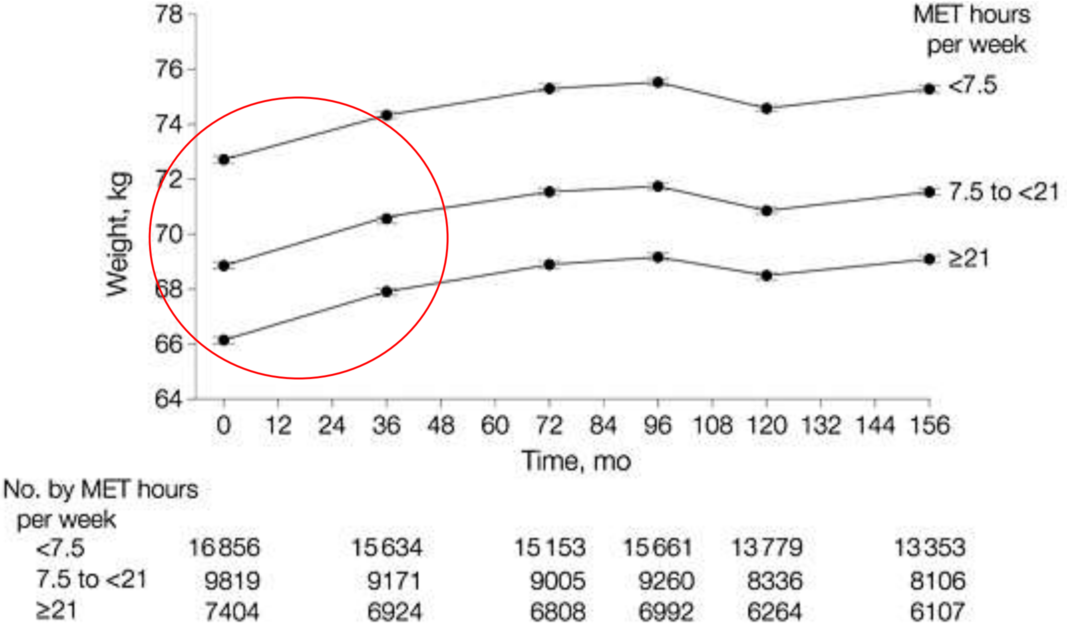
High exercise group

No. by MET hours per week

	0	36	72	96	120	156
<7.5	16856	15634	15153	15661	13779	13353
7.5 to <21	9819	9171	9005	9260	8336	8106
≥21	7404	6924	6808	6992	6264	6107

The heaviest exercisers weigh less to start, *but the weight gain curves between the three baseline groups are almost identical*.

The authors say: “Figure 2 shows the trajectory of weight gain over time by baseline physical activity levels. When classified by this single measure of physical activity, all 3 groups showed similar weight gain patterns over time.”



No. by MET hours per week	0	36	72	96	120	156
<7.5	16856	15634	15153	15661	13779	13353
7.5 to <21	9819	9171	9005	9260	8336	8106
≥21	7404	6924	6808	6992	6264	6107

But baseline physical activity should predict weight gain in the first three years…do those slopes look different to you?

Factors that affect p-values/statistical significance:

Effect size

Sample size

Variability

$$\boxed{\text{statistical significance} \propto \frac{\text{Effect size} \times \text{Sample size}}{\text{Variability}}}$$

Sample size and statistical significance, correlation coefficient

The minimum correlation coefficient that will be statistically significant for various sample sizes. Calculated using the approximation, $r = \frac{2}{\sqrt{n}}$

Sample Size	Minimum correlation coefficient that will be statistically significant, p<.05
10	0.63
100	0.20
1000	0.06
10,000	0.02
100,000	0.006
1,000,000	0.002

Sainani KL. Clinical versus statistical significance. *PM&R*. 2012;4:442-5.

Another headline

Drinkers May Exercise More Than Teetotalers

Activity levels rise along with alcohol use, survey shows

"MONDAY, Aug. 31 (HealthDay News) -- Here's something to toast: Drinkers are often exercisers"…

"In reaching their conclusions, the researchers examined data from participants in the 2005 Behavioral Risk Factor Surveillance System, a yearly telephone survey of about **230,000 Americans**."…

For women, those who imbibed exercised **7.2 minutes more per week** than teetotalers**.** The results applied equally to men…

Take-home points

- P-values help us distinguish between real effects and random fluctuation. If the sample size is large enough, random fluctuation is not an issue and p-values are irrelevant.
- When the sample size is 10s or 100s of thousands (except in the case of rare outcomes), you should ignore p-values.
- Pay attention to the effect size and the confidence interval.
  - Are any of the effect sizes within the confidence interval big enough to care about?

Statistics in Medicine

Module 3:

P-value pitfalls: multiple testing

Multiple testing problem

- In 1980, researchers at Duke randomized 1073 heart disease patients into two groups, but treated the groups equally.
- Not surprisingly, there was no difference in survival.
- Then they divided the patients into 18 subgroups based on prognostic factors.
- In a subgroup of 397 patients survival of those in "group 1" was significantly different from survival of those in "group 2" (p<.025).
- *How could this be?*
- *Results from a chance imbalance in the subgroups.*

Lee et al. "Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease," *Circulation*, 61: 508-515, 1980.

Multiple comparisons

- A significance level of 0.05 means that your false positive rate for one test is 5%. If you run more than one test, your false positive (type I error) rate will be higher than 5%.
- If we compare survival of "treatment" and "control" within each of 18 subgroups, that's 18 comparisons.
- If these comparisons were independent, the chance of at least one type I error (false positive) would be…

$$\boxed{1-(.95)^{18}=.60}$$

Multiple testing

- “If you torture your data long enough they will confess to something”
- If there are no effects, you will still get p-values <.05 just by chance (about 1 in 20 times).
- The more tests you run, the more opportunities there are for chance findings.

Multiple testing example…

- Researchers examined the relationship between intakes of caffeine/coffee/tea and breast cancer overall and within multiple subgroups (50 tests)
- Overall, there was no association
- But there were 4 "significant" or near-significant p-values in subgroups:
  - coffee intake was linked to increased risk in women with benign breast disease (p=.08)
  - caffeine intake was linked to increased risk of estrogen/progesterone negative tumors and tumors larger than 2 cm (p=.02, p=.02)
  - decaf coffee was linked to reduced risk of BC in postmenopausal hormone users (p=.02)

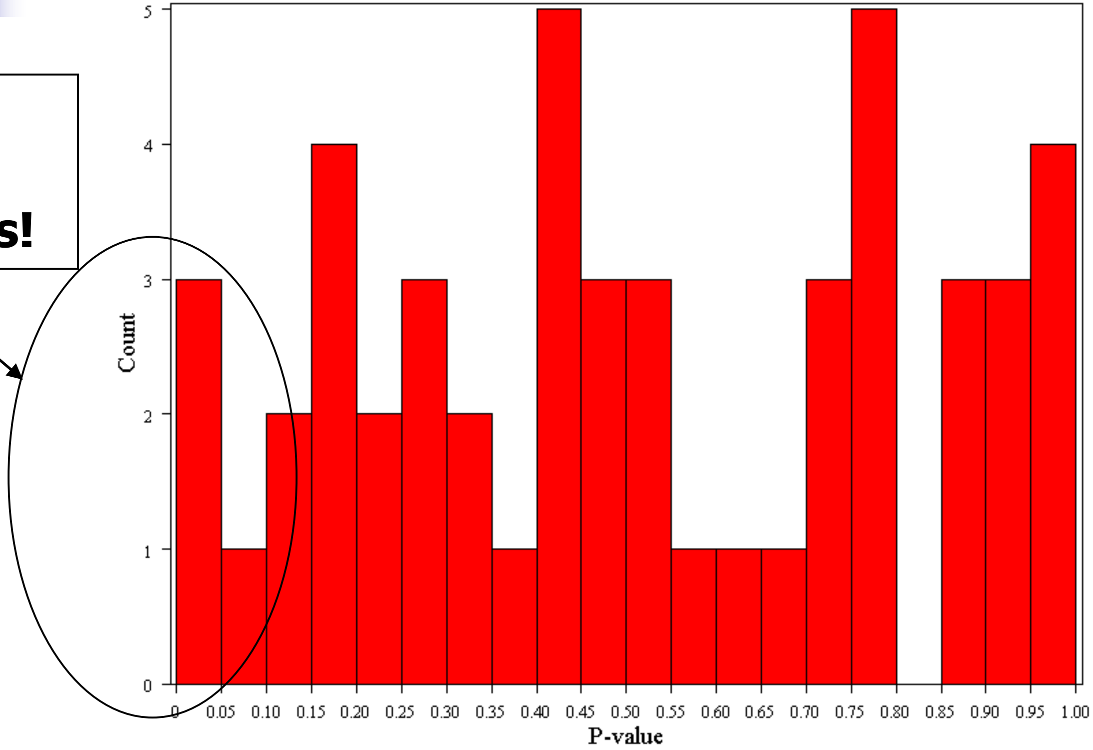
Ishitani K, Lin J, PhD, Manson JE, Buring JE, Zhang SM. Caffeine consumption and the risk of breast cancer in a large prospective cohort of women. *Arch Intern Med.* 2008;168:2022-2031.

Media Coverage:

- "Caffeine consumption was, however, associated with hormone-negative breast cancers and breast tumors larger than 2 cm."
- "But the study did uncover an increased risk of cancer for women with benign breast disease who drank four or more cups of coffee a day. Caffeine consumption was also linked to an increased risk of tumors that are hormone-receptor negative or larger than two centimeters."

Distribution of the p-values from the 50 tests

Likely chance findings!



Also, effect sizes showed no consistent pattern.

The risk ratios:

-were close to 1.0 (ranging from 0.67 to 1.79)

-indicated protection (<1.0) about as often harm (>1.0)

-showed no consistent dose-response pattern.

You may experience a chance finding if:

- Analyses are exploratory (many tests)
- Many tests have been performed but only a few are significant
- The significant p-values are modest in size (between p=0.01 and p=0.05)
- The pattern of effect sizes is inconsistent
- The p-values are not corrected for multiple comparisons

Another example: Microarrays!

- Compare the expression of 30,000 genes between 2 groups.
- If there are no differences between the groups, 1500 (=5%) of the genes will still achieve conventional statistical significance (p<.05).

How common are false positives in the literature?

- According to one estimate:
  - about 1 in 2 p-values <.05 is a false positive (50% chance!)
  - 1 in 6 p-values <.01 is a false positive
  - 1 in 56 p-values <.0001 is a false positive

Sterne JA and Smith GD. Sifting through the evidence—what's wrong with significance tests? *BMJ* 2001; 322: 226-31.

Take-home points

- Look at the totality of the evidence.
- Expect 1 significant p-value (<0.05) about 1 in every 20 tests.

P-value corrections

- One way to control the type-I error (false positive) rate is to apply a correction for multiple comparisons
- Bonferroni is the simplest, but also the most conservative correction procedure

Bonferroni

To make a Bonferroni correction, divide your desired alpha cut-off level (usually .05) by the number of comparisons you are making. Assumes complete independence between comparisons, which is way too conservative.

Obtained P-value	Original Alpha	# tests	New Alpha	Significant?
.001	.05	10	.005	Yes
.011	.05	4	.013	Yes
.019	.05	3	.017	No
.032	.05	2	.025	No
.019	.05	1	.050	Yes

Alternatives to Bonferroni: Holm and Hochberg

- Arrange all the resulting p-values in order from smallest (most significant) to largest: p_1 to p_T

Note: Holm and Hochberg should give you the same results. Use Holm if you anticipate few significant comparisons; use Hochberg if you anticipate many significant comparisons.

Holm

1. Start with p_1, and compare to Bonferroni p (=α/T).
2. If $p_1 < \alpha/T$, then p_1 is significant and continue to step 2. If not, then we have no significant p-values and stop here.
3. If $p_2 < \alpha/(T-1)$, then p_2 is significant and continue to step. If not, then p_2 thru p_T are not significant and stop here.
4. If $p_3 < \alpha/(T-2)$, then p_3 is significant and continue to step If not, then p_3 thru p_T are not significant and stop here.

Repeat the pattern…

Hochberg

1. Start with largest (least significant) p-value, p_T, and compare to α. If it's significant, so are all the remaining p-values and stop here. If it's not significant then go to step 2.
2. If $p_{T-1} < \alpha/(T-1)$, then p_{T-1} is significant, as are all remaining smaller p-vales and stop here. If not, then p_{T-1} is not significant and go to step 3.

Repeat the pattern…

Practice Problem

A large randomized trial compared an experimental drug and 9 other standard drugs for treating motion sickness. An ANOVA test revealed significant differences between the groups. The investigators wanted to know if the experimental drug ("drug 1") beat any of the standard drugs in reducing total minutes of nausea, and, if so, which ones. The p-values from the pairwise ttests (comparing drug 1 with drugs 2-10) are below.

Drug 1 vs. drug …	2	3	4	5	6	7	8	9	10
p-value	.05	.3	.25	.04	.001	.006	.08	.002	.01

Which differences would be considered statistically significant using a Bonferroni correction? A Holm correction? A Hochberg correction?

Answer

Bonferroni makes new α value = $\alpha/9 = .05/9 = .0056$; therefore, using Bonferroni, the new drug is only significantly different than standard drugs 6 and 9.

Arrange p-values:

6	9	7	10	5	2	8	4	3
.001	.002	.006	.01	.04	.05	.08	.25	.3

Holm: $.001<.0056$; $.002<.05/8=.00625$; $.006<.05/7=.007$; $.01>.05/6=.0083$; therefore, new drug only significantly different than standard drugs 6, 9, and 7.

Answer

Arrange p-values:

6	9	7	10	5	2	8	4	3
.001	.002	.006	.01	.04	.05	.08	.25	.3

Hochberg: .3>.05; .25>.05/2; .08>.05/3; .05>.05/4; .04>.05/5; .01>.05/6; .006<.05/7; therefore, drugs 7, 9, and 6 are significantly different.

Statistics in Medicine

Module 4:

P-value pitfalls: Don’t compare p-values!

Within-group vs. between-group comparisons

- Within-group effect
  - Did group A improve compared with itself at baseline?
- Between group effect
  - Did group A improve more than group B?

For controlled studies, only the between-group effects are relevant.

- “The effect was significant in group A (p<.05), but not significant in group B (p>.05)” does not imply that the groups differ significantly.

Example

- In a placebo-controlled randomized trial of DHA oil for eczema, researchers found a statistically significant improvement in the DHA group but not the placebo group.
- The abstract reports: “DHA, but not the control treatment, resulted in a significant clinical improvement of atopic eczema.”

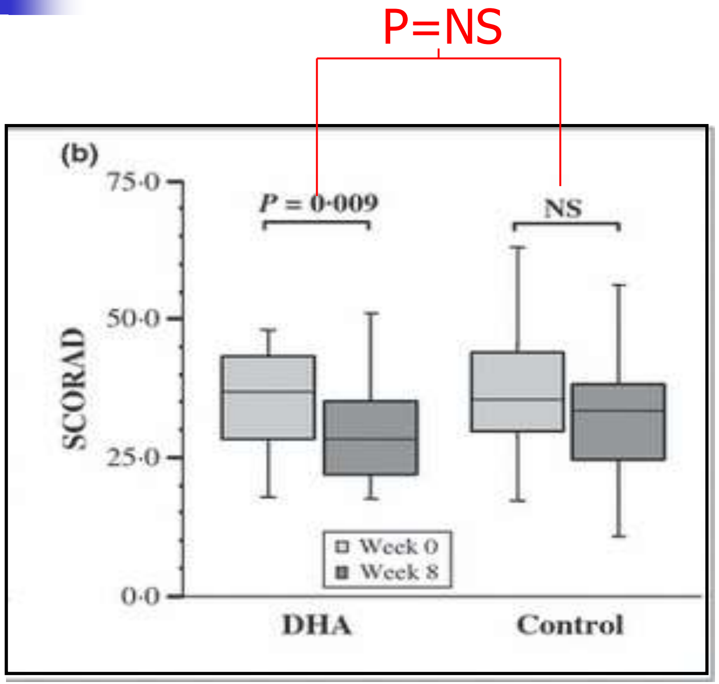
Media coverage…

- “DHA supplementation improves Eczema”
- “Omega-3 can help eczema”

However…

- Buried in the discussion section: "*The improvement in the treatment group was not significantly better than the improvement in the placebo group.*"
- This is a null result—DHA was statistically indistinguishable from placebo!

“P-value comparisons” are meaningless…



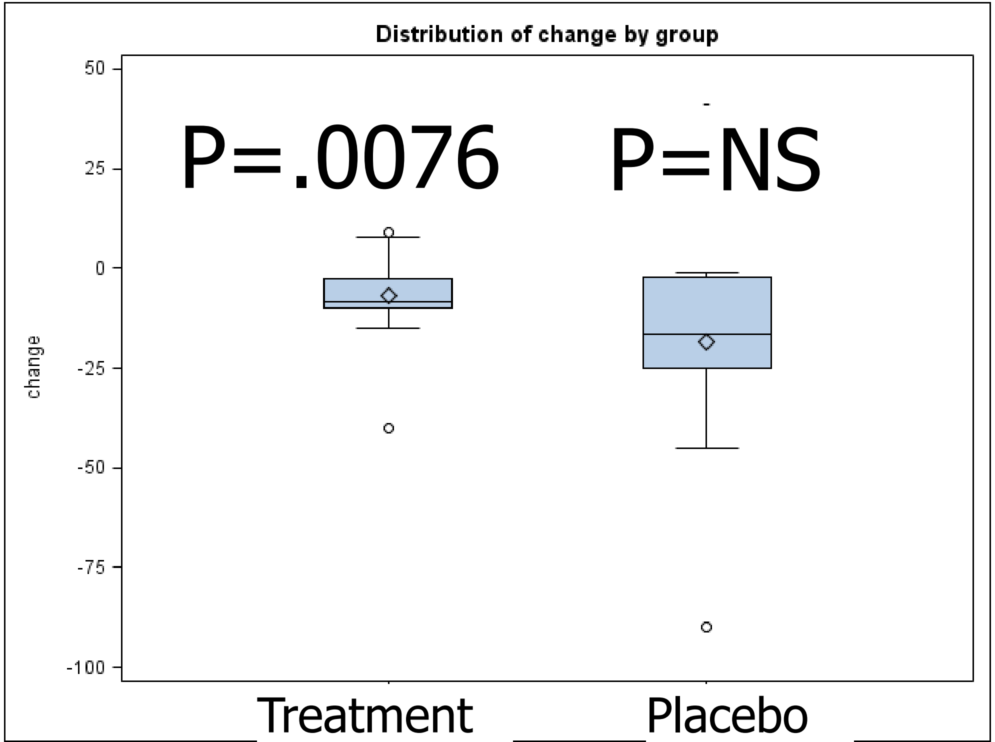
P=NS

The improvement in the DHA group (18%) is not significantly greater than the improvement in the control group (11%).

The authors omitted the relevant p-value from the graphic.

Reproduced with permission from: Koch C, Dölle S, Metzger M, et al. Docosahexaenoic acid (DHA) supplementation in atopic eczema: a randomized, double-blind, controlled trial. Br J Dermatol 2008;158:786-792.

Hypothetical data …



The average and median improvements are *bigger* in the placebo group! But only the improvement in the treatment group is significant.

The between-group difference is not significant.

Propagation of bad statistics!

- That DHA and eczema study has now been cited 33 times, as evidence of a positive effect of DHA on eczema.
- <u>Example, 2009 review article:</u>
- "Koch *et al.* (2008) undertook a randomised, double-blind controlled pilot study. Patients clinically diagnosed with atopic eczema were asked to consume either 5.4 g DHA/day (*n* =21) or a placebo (*n* =23) for eight weeks. Although only a preliminary study, the results indicated that atopic eczema symptoms improved significantly in the DHA compared with the control group." C.H.S. Ruxton; Derbyshire, E. Latest evidence on omega-3 fatty acids and health *Nutrition and Food Science 2009; 39: 423-438.*

Tests for within-group effects vs. tests for between-group effects

Statistical tests for within-group effects	Statistical tests for between-group effects
Paired ttest	Two-sample ttest
Wilcoxon sign-rank test	Wilcoxon rank-sum test (equivalently, Mann-Whitney U test)
Repeated-measures ANOVA, time effect	ANOVA; repeated-measures ANOVA, group*time effect
McNemar′s test	Risk difference, chi-square test, or relative risk

Take-home points

- All randomized, controlled trials should report between-group comparisons as their primary outcome.
  - It is fine to additionally present within-group changes, but these results should be secondary.

Statistics in Medicine

Module 5:

P-value pitfalls: Failure to prove an effect is not proof of no effect.

You can’t prove the null!

- If you fail to reject the null hypothesis, this is not proof of no effect.

Example

- What’s wrong with this statement?
    - “There was no significant effect of treatment (p =0.058), nor treatment by velocity interaction (p = 0.19), **indicating that the treatment and control groups did not differ** in their ability to perform the task.”
- P-values >.05 indicate that we have insufficient evidence of an effect (no evidence to say they are the same); they do not constitute proof of no effect.

Smoking cessation trial

- Weight-concerned women smokers were randomly assigned to one of four groups:
  - Weight-focused or standard counseling plus bupropion or placebo
- Outcome: biochemically confirmed smoking abstinence

Levine MD, Perkins KS, Kalarchian MA, et al. Bupropion and Cognitive Behavioral Therapy for Weight-Concerned Women Smokers. *Arch Intern Med* 2010;170:543-550.

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

Data excerpted from Tables 2 and 3 of Levine MD, Perkins KS, Kalarchian MA, et al. Bupropion and cognitive behavioral therapy for weight-concerned women smokers. *Arch Intern Med* 2010;170:543-550.

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

Counseling methods appear equally effective in the placebo groups.

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

Clearly, bupropion improves quitting rates in the weight-focused counseling group.

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

What conclusion should we draw about the effect of bupropion in the standard counseling group?

Authors’ conclusions/Media coverage…

- “Among the women who received standard counseling, bupropion did not appear to improve quit rates or time to relapse.”
- “For the women who received standard counseling, taking bupropion didn't seem to make a difference.”

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

Buprpion does appear to improve quitting rates over placebo, though it just misses statistical significance.

Correct message from data…

- Bupropion improves quitting rates over counseling alone.
    - Main effect for drug is significant.
    - Main effect for counseling type is NOT significant.
    - Interaction between drug and counseling type is NOT significant.

New concept: interaction!

- A significant interaction means that the treatment effect differs significantly in different subgroups.
  - E.g., the drug works significantly better in the weight-focused counseling group compared with the standard counseling group.
- To prove interaction, we must compare treatment effects not p-values between the groups.
  - The "drug effect was significant (p-value)" in weight-focused counseling but "not significant in standard counseling (p-value)" is NOT proof of interaction.
  - The "drug effect in weight-focused counseling" was significantly greater than the "drug effect in standard counseling" would be proof of interaction.

Interaction:

Months after quit target date	Weight-focused counseling		Standard counseling group		P-value for interaction between bupropion and counseling type
	Bupropion group (n=106)	Placebo group (n=87)	Bupropion group (n=89)	Placebo group (n=67)	
3	41%	18%	33%	19%	.42
6	34%	11%	21%	10%	.39
12	24%	8%	19%	7%	.79

Sainani KL. Misleading comparisons: the fallacy of comparing statistical significance. *PM&R* 2010; 2 (3): 209-13.

Take-home points

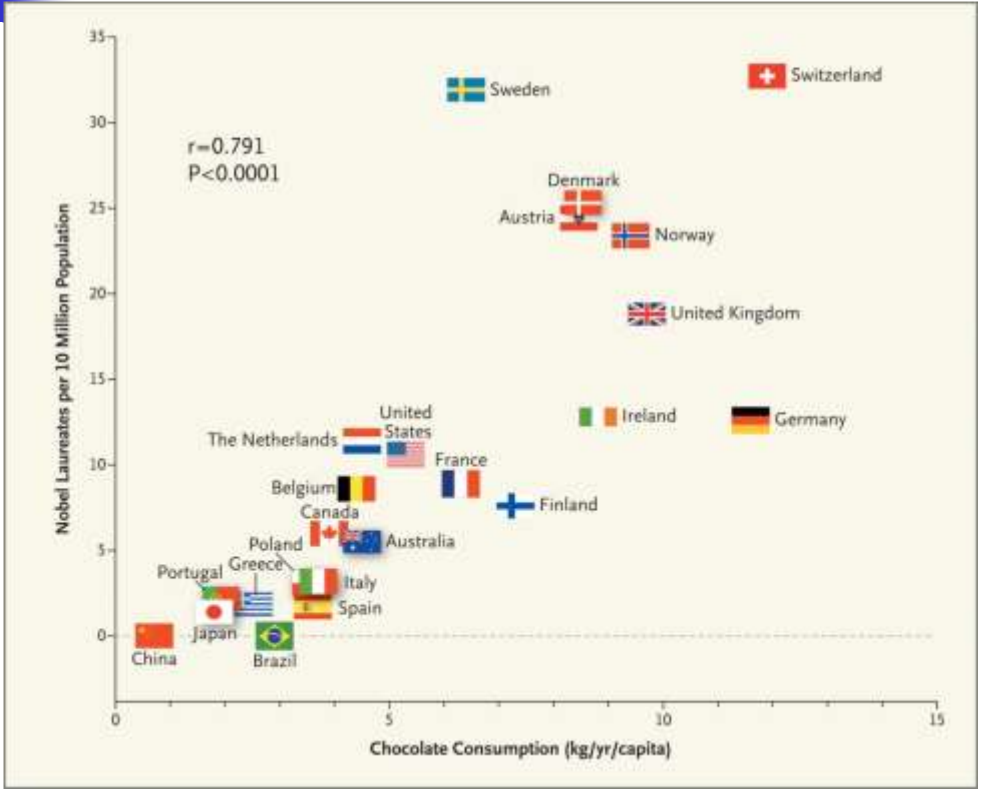
- Most statistical tests are designed to disprove the null hypothesis, not to prove it.
  - If you want to “prove” the null, the best you can do is a non-inferiority or equivalence trial.

Statistics in Medicine

Module 6:

P-value pitfalls: correlation is not causation

Chocolate and Nobel prize winners!



Reproduced with permission from: Messerli FH. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population. *N Engl J Med* 2012;367:1562-1564.

Statistics in Medicine

Module 7:

Introduction to Correlated Data

What are correlated data?

Correlated data arise when pairs or clusters of observations are related and thus are more similar to each other than to other observations in the dataset.

Examples:

-same subject measured at multiple time points

-two eyes or hands from the same person

-siblings

-twin pairs

-husband-wife pairs

-matched case-control pairs

-cluster-randomized trials

Introduction to Correlated Data

- Example: Collateralized debt obligations

From: The Signal and the Noise, Nate Silver

Example: CDOs

- Collateralized debt obligation (CDO)
- Simplified example of a CDO: pool five subprime mortgages together. The CDO pays out unless all five mortgages default.
- Assume 20% chance of default per mortgage.
- Assume the mortgages are INDEPENDENT; then the probability that the CDO defaults = P(A defaults)*P(B defaults)*P (C defaults) *P(D defaults) * P(E defaults) = (.20)**5 = 1/3125
- Seems like a good bet, right?

From: The Signal and the Noise, Nate Silver

Example

- BUT is the assumption of INDEPENDENCE valid?
- Might be valid in a healthy economy.
- But when a massive housing bubble bursts, these mortgages become highly correlated!

- S&P predicted: 0.12% (1 in 850) that a certain CDO would fail to pay out over the next five years
- In fact, 28% defaulted.
- The default rate was 200-times than S&P predicted because they failed to account for correlation!

From: The Signal and the Noise, Nate Silver

Are the observations correlated?

1. What is the unit of observation?
   - person* (most common)
   - limb
   - hand
   - knee
   - half a face
   - clinical center
2. Are the observations independent or correlated?
   - Independent: observations are unrelated (usually different, unrelated people)
   - Correlated: some observations are related to one another, for example: the same person over time (repeated measures), two legs from the same person, two knees from the same person

Correlations

- Ignoring correlations will:
  - *overestimate* p-values for within-person or within-cluster comparisons (increase type2 error or false negative)
  - *underestimate* p-values for between-person or between-cluster comparisons (increase type-1 error or false positive)

1. Within-person comparison: example

- Split-face trial:
    - Researchers assigned 56 subjects to apply SPF 85 sunscreen to one side of their faces and SPF 50 to the other prior to engaging in 5 hours of outdoor sports during mid-day. The outcome is sunburn (yes/no).
    - Unit of observation = side of a face
    - Are the observations correlated? Yes.

Russak JE et al. *JAAD* 2010; 62: 348-349.

Results ignoring correlation:

Table I -- Dermatologist grading of sunburn after an average of 5 hours of skiing/snowboarding (*P* = .03; Fisher's exact test)

Sun protection factor	**Sunburned**	**Not sunburned**
85	1	55
50	8	48

Fisher's exact test compares the following proportions: 1/56 versus 8/56. Note that individuals are being counted twice!

Reproduced with permission from: Russak JE et al. *JAAD* 2010; 62: 348-349.

Correct analysis of data:

Correct presentation of the data from: Russak JE et al. *JAAD* 2010; 62: 348-349. (*P* = .016; McNemar's exact test).

	SPF-50 side	
SPF-85 side	Sunburned	Not sunburned
Sunburned	1	0
Not sunburned	7	48

McNemar's exact test evaluates the probability of the following: In all 7 out of 7 cases where the sides of the face were discordant (i.e., one side burnt and the other side did not), the SPF 50 side sustained the burn.

2. Between-person comparison: example

- Hypothetical trial in which 50 patients with bilateral eye disease are randomly assigned to receive an active drug or a placebo solution in both eyes.
- Treatment is considered a success if symptoms improve by more than 50% in a given eye.
- Unit of observation = eye
- Are the observations correlated? YES

Example: between-patient comparison

Results from a hypothetical trial in which 50 subjects were randomized to receive active drug (n=25) or placebo (n=25) in both eyes.

Analysis	N (%) of eyes improving in the treatment group	N (%) of eyes improving in the control group	p-value	Odds ratio and 95% CI
Assuming eyes are independent*	27/50 (54%)	17/50 (34%)	.046	2.28 (1.02, 5.11)
Correcting for within-subject correlation**	27/50 (54%)	17/50 (34%)	.11	2.28 (0.83, 6.28)

*Data were analyzed with unconditional logistic regression.

**Data were analyzed using a generalized estimating equation, correcting for within-subject correlation.

Reprinted from Table 3 of: Sainani K. The importance of accounting for correlated observations. *PM&R* 2010 Sep;2:858-61.

Between-cluster example: Exercise labels study again…

- This was a cluster-randomized trial
  - Investigators randomly assigned the 4 stores to interventions (not individuals)
- But authors used people (n=1600), not stores (n=4), as the unit of observation
- People are correlated within-store, but authors ignored these correlations
- Thus, the p-values may be under-estimated (overly optimistic)
- Fixes:
  - Change the unit of observation: analyze data at the store level
  - Account for correlations using GEE modelling

Statistics in Medicine

<u>Module 8:</u>

Overview of Statistical Tests:

What test do I use?

Common statistics for various types of outcome data

Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Fisher's exact test McNemar's exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model (beyond the scope of this course)	Time-varying effects (beyond the scope of this course)

1. What is the outcome (dependent variable)?

Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar’s test Conditional logistic regression GEE modeling	Fisher’s exact test McNemar’s exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model	Time-varying effects

2. Are the observations correlated?

Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Fisher's exact test McNemar's exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model	Time-varying effects

3. Are key model assumptions met?

Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	McNemar's exact test Fisher's exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model	Time-varying effects

Continuous outcome (means)

Outcome Variable	**Are the observations independent or correlated?**		**Alternatives if the normality assumption is violated and small sample size:**
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**Ttest** (2 groups) **ANOVA** (2 or more groups) **Pearson's correlation coefficient** (1 continuous predictor) **Linear regression** (multivariate regression technique)	**Paired ttest** (2 groups or time-points) **Repeated-measures ANOVA** (2 or more groups or time-points) **Mixed models/GEE modeling**: (multivariate regression techniques)	Non-parametric statistics **Wilcoxon sign-rank test** (alternative to the paired ttest) **Wilcoxon rank-sum test** (alternative to the ttest) **Kruskal-Wallis test** (alternative to ANOVA) **Spearman rank correlation coefficient** (alternative to Pearson's correlation coefficient)

Binary or categorical outcomes (proportions)

Outcome Variable	**Are the observations correlated?**		**Alternatives if sparse data:**
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	**Risk difference/relative risks** (2x2 table) **Chi-square test** (RxC table) **Logistic regression** (multivariate regression technique)	**McNemar's chi-square test** (2x2 table) **Conditional logistic regression** (multivariate regression technique) **GEE modeling** (multivariate regression technique)	**McNemar's exact test** (alternative to McNemar's chi-square, for sparse data) **Fisher's exact test** (alternative to the chi-square, for sparse data)

Time-to-event outcome (survival data)

Outcome Variable	**Are the observation groups independent or correlated?**		**Modifications if assumptions violated:**
	independent	correlated	
Time-to-event (e.g., time to fracture)	**Rate ratio** (2 groups) **Kaplan-Meier statistics** (2 or more groups) **Cox regression** (multivariate regression technique)	**Frailty model** (multivariate regression technique)	**Time-varying effects**