Statistical Inference: Pitfalls of hypothesis testing

Pitfall 1: over-emphasis on p-values

- Statistical significance does not guarantee clinical significance.
- Example: a study of about 60,000 heart attack patients found that those admitted to the hospital on weekdays had a significantly longer hospital stay than those admitted to the hospital on weekends ($p<.03$), but the magnitude of the difference was too small to be important: 7.4 days (weekday admits) vs. 7.2 days (weekend admits).

Ref: Kostis et al. *N Engl J Med* 2007;356:1099-109.

Pitfall 1: over-emphasis on p-values

- Clinically unimportant effects may be statistically significant if a study is large (and therefore, has a small standard error and extreme precision).

- Pay attention to effect sizes and confidence intervals (see end of this lecture).

Pitfall 2: association does not equal causation

- Statistical significance does not imply a cause-effect relationship.

- Interpret results in the context of the study design.

Pitfall 3: data dredging/multiple testing

- 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 (with three-vessel disease and an abnormal left ventricular contraction) survival of those in "group 1" was significantly different from survival of those in "group 2" ($p<.025$).
- *How could this be since there was no treatment?*

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

Pitfall 3: multiple testing

- The difference resulted from the combined effect of small imbalances in the subgroups
- 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 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 false positive would be…

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

Pitfall 3: multiple testing

Sources of multiple testing

Source	Example
Multiple outcomes	a cohort study looking at the incidence of breast cancer, colon cancer, and lung cancer
Multiple predictors	an observational study with 40 dietary predictors or a trial with 4 randomization groups
Subgroup analyses	a randomized trial that tests the efficacy of an intervention in 20 subgroups based on prognostic factors
Multiple definitions for the exposures and outcomes	an observational study where the data analyst tests multiple different definitions for “moderate drinking” (e.g., 5 drinks per week, 1 drink per day, 1-2 drinks per day, etc.)
Multiple time points for the outcome (repeated measures)	a study where a walking test is administered at 1 months, 3 months, 6 months, and 1 year
Multiple looks at the data during sequential interim monitoring	a 2-year randomized trial where the efficacy of the treatment is evaluated by a Data Safety and Monitoring Board at 6 months, 1 year, and 18 months

Pitfall 3: multiple testing

Results from a previous Class survey…

- The research question was to test whether or not being born on odd or even days predicted anything about your future.
- I discovered that people who born on odd days wake up later and drink more alcohol than people born on even days; they also have a trend of doing more homework ($p=.04$, $p<.01$, $p=.09$).
- Those born on odd days wake up 42 minutes later (7:48 vs. 7:06 am); drink 2.6 more drinks per week (1.1 vs. 3.7); and do 8 more hours of homework (22 hrs/week vs. 14).

Pitfall 3: multiple testing

Results from Class survey…

- I can see the NEJM article title now…
- "Being born on odd days predisposes you to alcoholism and laziness, but makes you a better med student."
- Assuming that this difference can't be explained by astrology, it's obviously an artifact!
- What's going on?…

Pitfall 3: multiple testing

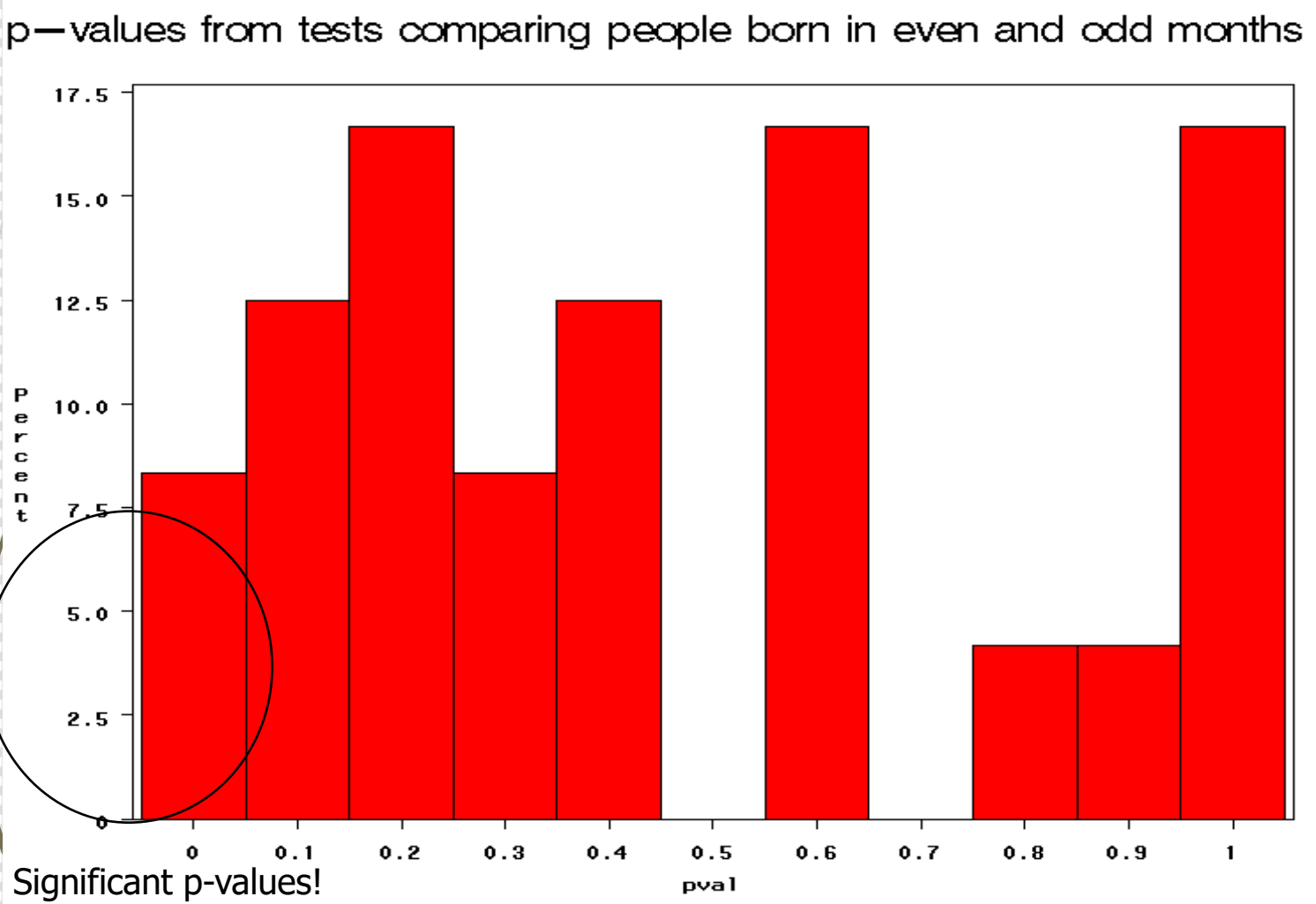
Results from Class survey…

- After the odd/even day question, I asked 25 other questions…
- I ran 25 statistical tests (comparing the outcome variable between odd-day born people and even-day born people).
- So, there was a high chance of finding at least one false positive!

Pitfall 3: multiple testing

Recall: Under the null hypothesis of no associations (which we'll assume is true here!), p-values follow a uniform distribution…

P-value distribution for the 25 tests…

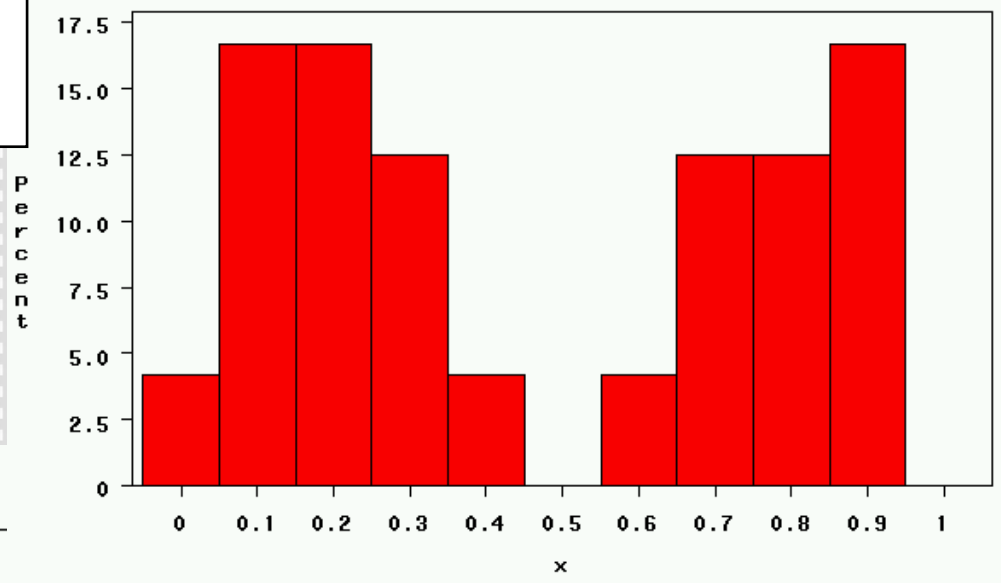


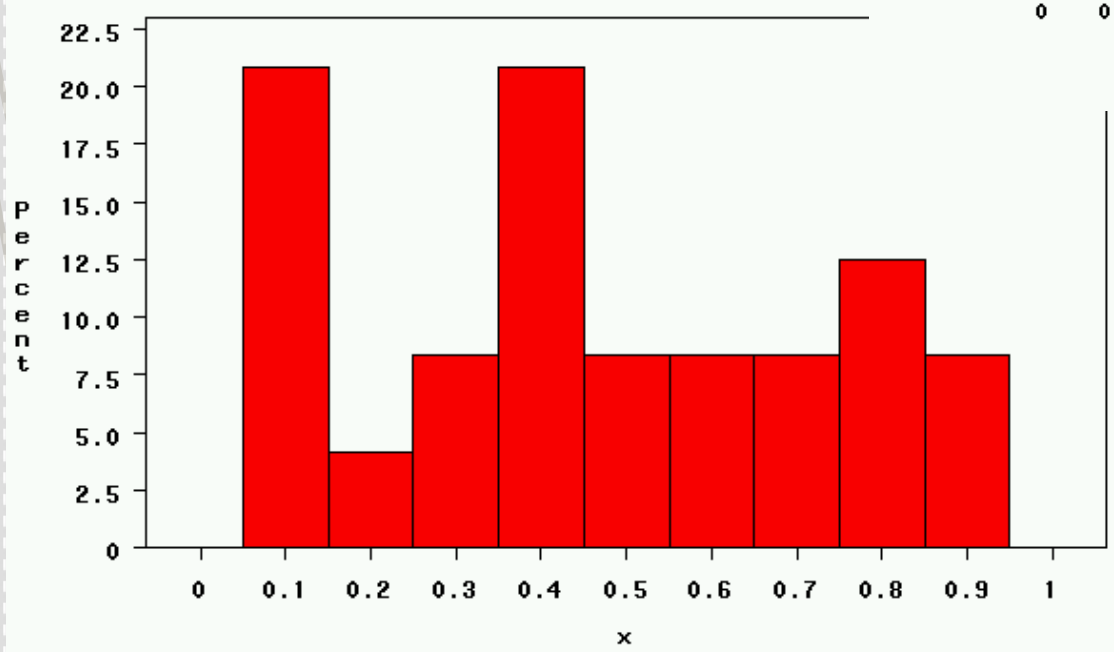
Significant p-values!

Pitfall 3: multiple testing

Compare with…

Next, I generated 25 “p-values” from a random number generator (uniform distribution). These were the results from two runs…





Pitfall 3: multiple testing

In the medical literature…

- Hypothetical example:
  - Researchers wanted to compare nutrient intakes between women who had fractured and women who had not fractured.
  - They used a food-frequency questionnaire and a food diary to capture food intake.
  - From these two instruments, they calculated daily intakes of all the vitamins, minerals, macronutrients, antioxidants, etc.
  - Then they compared fracturers to non-fracturers on all nutrients from both questionnaires.
  - They found a statistically significant difference in vitamin K between the two groups ($p<.05$) on the FFQ only.
  - They had a lovely explanation of the role of vitamin K in injury repair, bone, clotting, etc.

Pitfall 3: multiple testing

Factors indicative of chance findings

1. Analyses are exploratory.	The authors have mined the data for associations rather than testing a limited number of *a priori* hypotheses.
2. Many tests have been performed, but only a few p-values are "significant".	If there are no associations present, .05*k significant p-values ($p<.05$) are expected to arise just by chance, where k is the number of tests run.
3. The "significant" p-values are modest in size.	The closer a p-value is to .05, the more likely it is a chance finding. According to one estimate*, about 1 in 2 p-values <.05 is a false positive, 1 in 6 p-values <.01 is a false positive, and 1 in 56 p-values <.0001 is a false positive.
4. The pattern of effect sizes is inconsistent.	If the same association has been evaluated in multiple ways, an inconsistent pattern of effect sizes (e.g., risk ratios both above and below 1) is indicative of chance.
5. The p-values are not adjusted for multiple comparisons	Adjustment for multiple comparisons can help control the study-wide false positive rate.

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

Pitfall 4: high type II error (low power)

- Lack of statistical significance is not proof of the absence of an effect.
- Example: A study of 36 postmenopausal women failed to find a significant relationship between hormone replacement therapy and prevention of vertebral fracture. The odds ratio and 95% CI were: 0.38 (0.12, 1.19), indicating a potentially meaningful clinical effect. Failure to find an effect may have been due to insufficient statistical power for this endpoint.

…the absence of evidence is not evidence of absence

Ref: Wimalawansa et al. *Am J Med* 1998, 104**:**219-226.

Pitfall 4: high type II error (low power)

- Results that are not statistically significant should not be interpreted as "evidence of no effect," but as "no evidence of effect"
- Studies may miss effects if they are insufficiently powered (lack precision).
- Design adequately powered studies and report approximate study power if results are null.

Pitfall 5: the fallacy of comparing statistical significance

- Presence of statistical significance in one group and lack of statistical significance in another group ≠ a significant difference between the groups.

- 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.” However, the improvement in the treatment group was not significantly better than the improvement in the placebo group, so this is actually a null result.

Pitfall 5: the fallacy of comparing statistical significance

Misleading significance comparisons

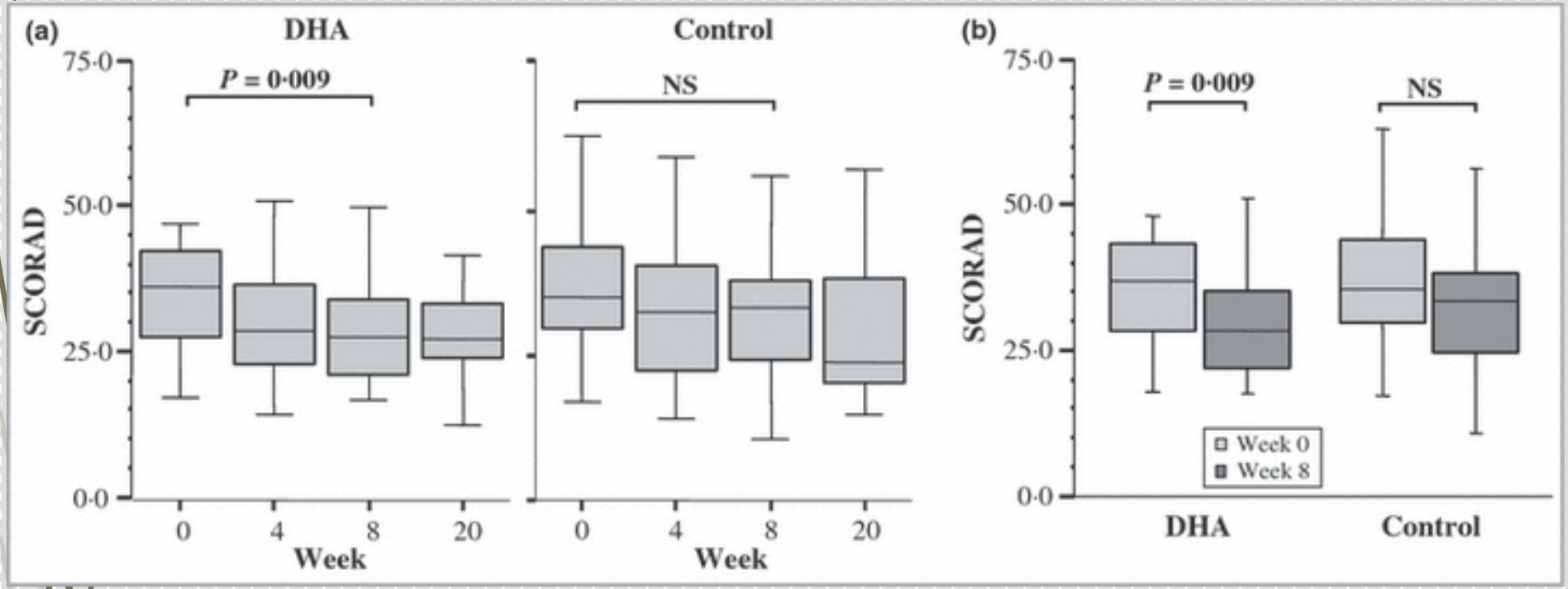


Figure 3 from: Koch C, Dölle S, Metzger M, Rasche C, Jungclas H, Rühl R, Renz H, Worm M. Docosahexaenoic acid (DHA) supplementation in atopic eczema: a randomized, double-blind, controlled trial. *Br J Dermatol*. 2008 Apr;158(4):786-92. Epub 2008 Jan 30.

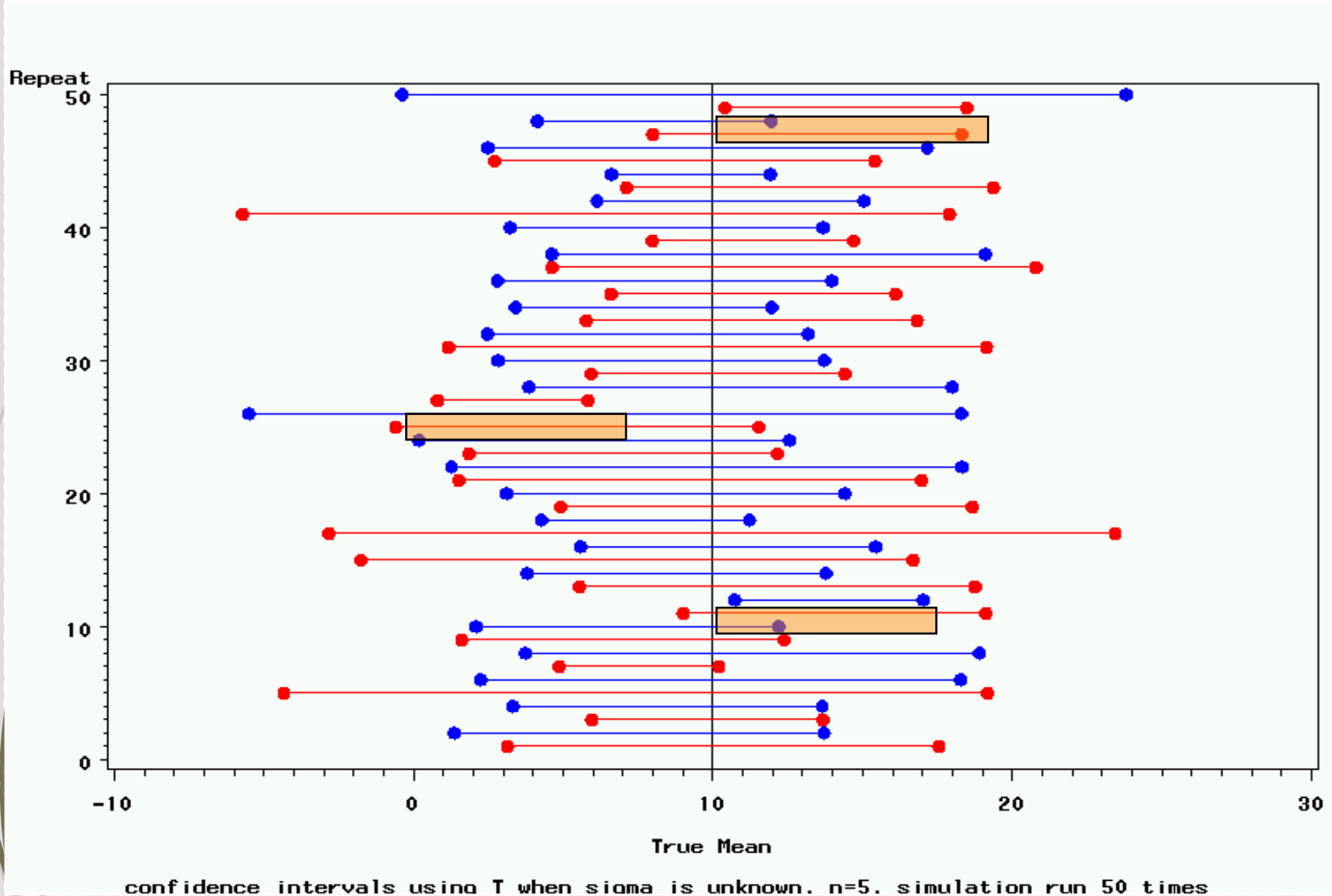
How about Confidence Intervals

- A plausible range of values for a population parameter.
- The precision of an estimate.(When sampling variability is high, the confidence interval will be wide to reflect the uncertainty of the observation.)
- Statistical significance (if the 95% CI does not cross the null value, it is significant at .05)

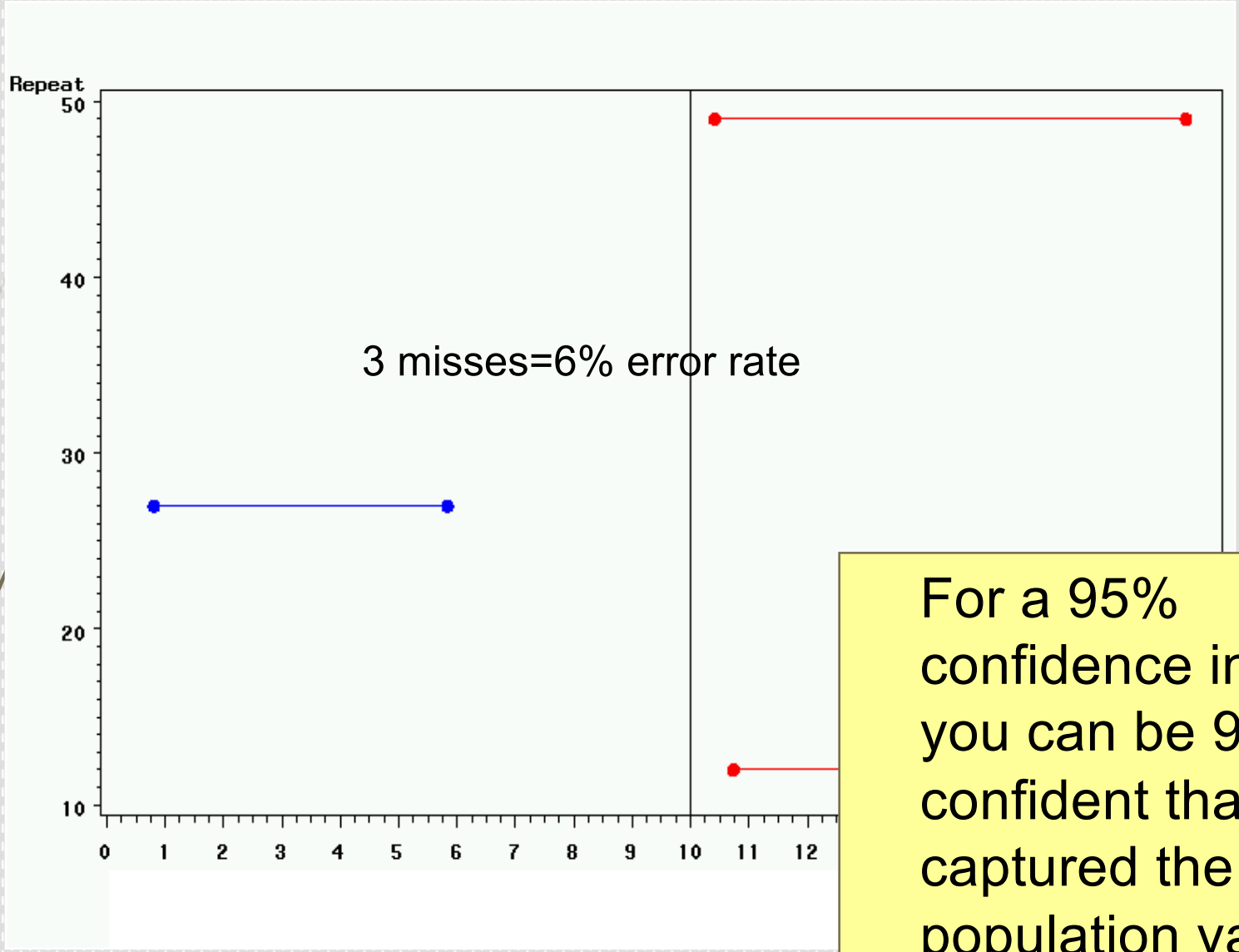
The true meaning of a confidence interval

- Suppose we do a computer simulation:
- Imagine that the true population value is 10.
- Have the computer take 50 samples of the same size from the same population and calculate the 95% confidence interval for each sample.
- Here are the results…

95% Confidence Intervals



95% Confidence Intervals



For a 95% confidence interval, you can be 95% confident that you captured the true population value.

Odds Ratio example: Antidepressant use & Heart Disease

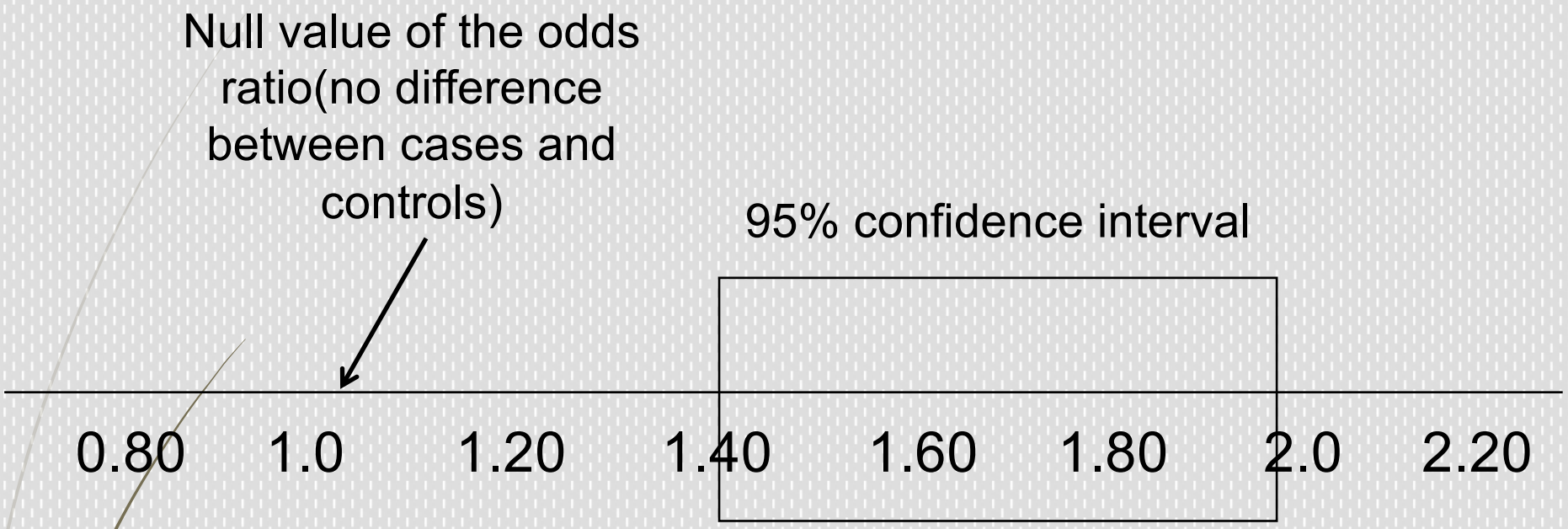
	Heart disease case	Control
antidepressants	217	871
No exposure	716	4645

$$\text{odds ratio} = \frac{\frac{217}{716}}{\frac{871}{4645}} = \frac{217 * 4645}{716 * 871} = 1.62$$

1.62 (1.41 to 1.99)

•"Antidepressants as risk factor for ischaemic heart disease: case-control study in primary care"; Hippisley-Cox et al. *BMJ* 2001; 323; 666-669

IS this a statistically significant association? YES



Null hypothesis: Proportions of cases who used antidepressants equals proportion of controls who used antidepressants.

Alternative hypothesis: Proportions are not equal.

P-value < .05

Multiple Comparisons: How do we control the family-wise error rate

- Least Significant Differences (LSD)
- Tukey (or Tukey-Kramer)
- Bonferroni
- Scheffe
- False Discovery Rate

Multiple Comparisons: controlling the family-wise error rate

- Least Significant Differences (LSD)
- Has high power but less control of the family-wise error rate
- Least conservative when doing many (>6) comparisons

Multiple Comparisons: controlling the family-wise error rate

- Tukey (or Tukey-Kramer)
- Specifies the exact significance level for your comparisons
- Controls the α level but less power

Multiple Comparisons: controlling the family-wise error rate

- Bonferroni 
- Divide your level of α by the number of tests.
- Very conservative EXCEPT if you have the exact number of comparisons you are going to do.
- Assumes all null hypotheses are true

Multiple Comparisons: controlling the family-wise error rate

- Scheffe
- Most conservative
- Always controls for all possible comparisons
- OK if you are ‘data snooping’

Multiple Comparisons: controlling the family-wise error rate

- False Discovery Rate
- A new (and rational) approach
- Gives the expected proportion of Type 1 errors among the rejected hypotheses
- Basic idea is ranking the p-values and only choosing the top *X* comparisons

Using ANOVA instead of t-tests

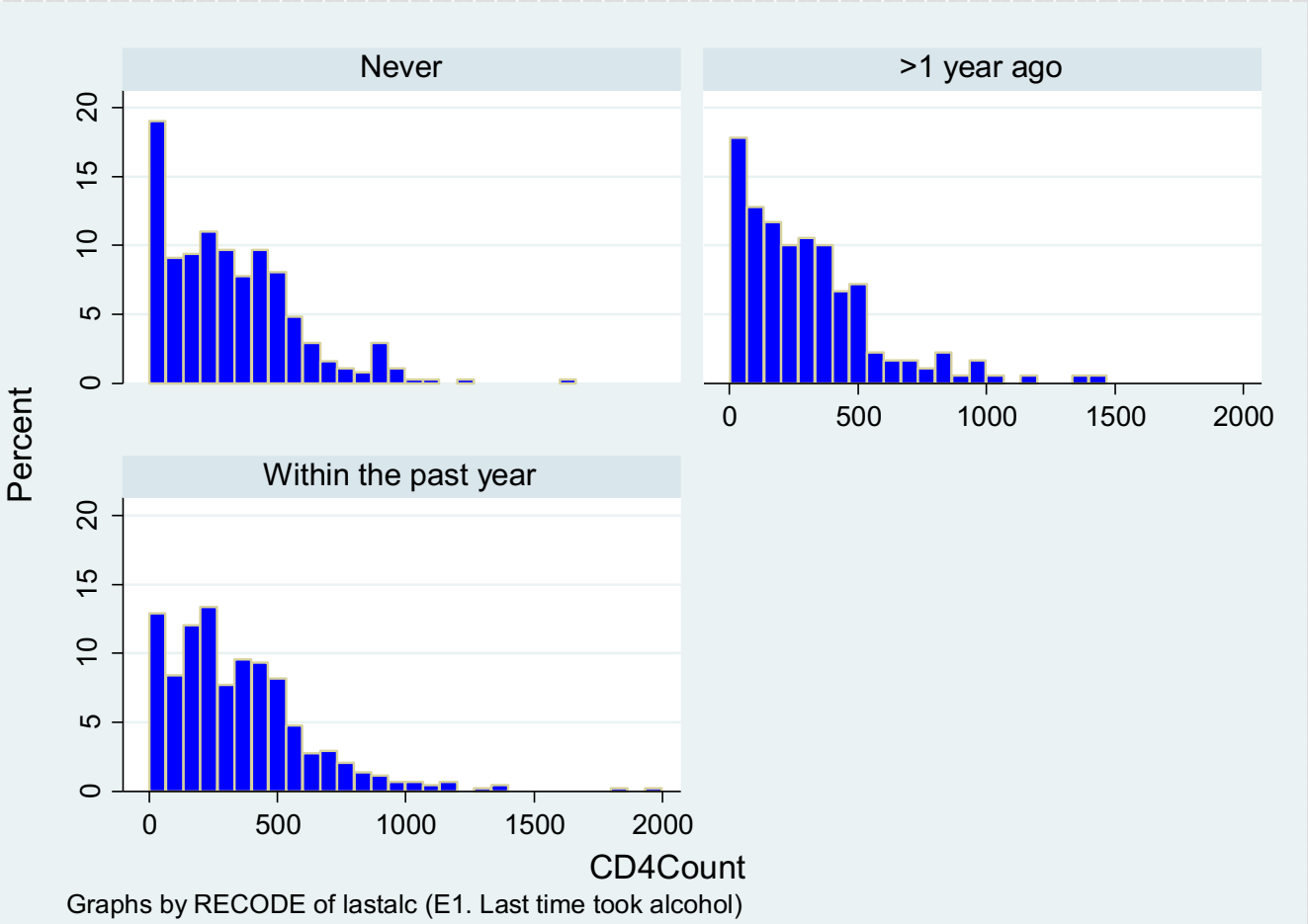
- The extension of the t-test to several independent groups is called analysis of variance or ANOVA
- Why is it called analysis of variance?
  - Even though your hypothesis is about the means, the test actually compares the variability between groups to the variability within groups
- The null hypothesis is:
  - H_0: all equal means $\mu_1=\mu_2=\mu_3=\ldots$
  - The alternative H_A is that at least one of the means differs from the others

ANOVA assumptions

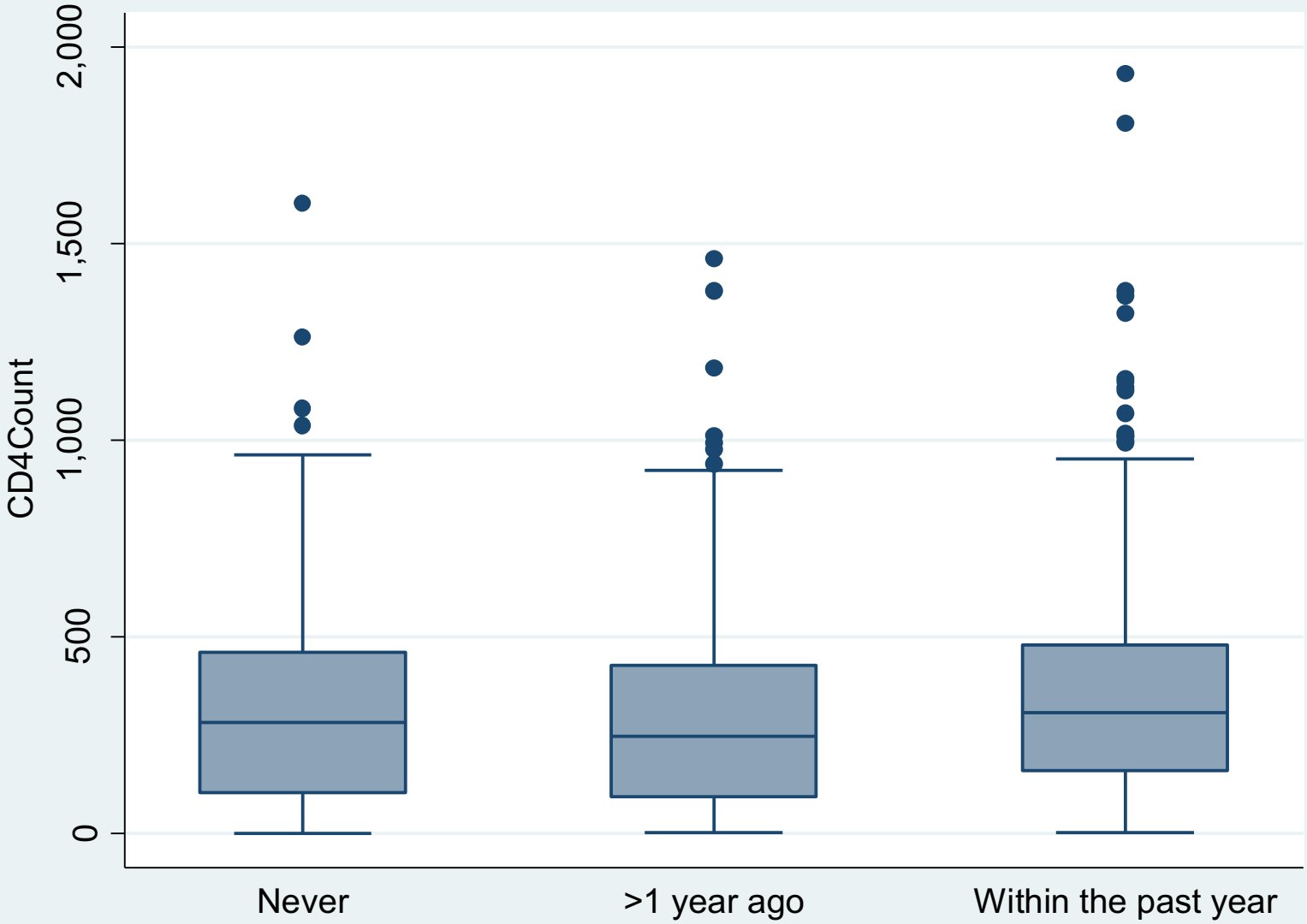
- The underlying data came from normally distributed populations
- The observations are independent
- The variance within each group over all groups (homoscedasticity)

ANOVA example

- Does CD4 count at time of testing differ by drinking category?




```
*Using vct_baseline_biostat200_v1.dta  **
hist cd4count, by(lastalc_3) percent fcolor(blue)
```




```
graph box cd4count, over(lastalc_3)
```

ANOVA example

```
tabstat cd4count, by(lastalc_3) s(n mean sd min median max)


Summary for variables: cd4count
     by categories of: lastalc_3 (RECODE of lastalc (E1. Last time took
   alcohol))

      lastalc_3 |         N      mean        sd       min       p50       max
----------------+--------------------------------------------------------------
          Never |       373  317.1475  253.4013         1       283      1601
    >1 year ago |       180  305.3778  266.9453         2     248.5      1461
Within the past |       441  349.8662  273.9364         3       308      1932
           year |
----------------+--------------------------------------------------------------
          Total |       994  329.5322  265.5157         1       285      1932
--------------------------------------------------------------------------------
```

ANOVA example

- CD4 count, by alcohol consumption category

oneway var groupvar

$$F = \frac{s_B^2}{s_W^2}$$

```
. oneway cd4count lastalc_3

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      344571.162      2    172285.581       2.45     0.0867
 Within groups      69660550.3     991    70293.189
------------------------------------------------------------------------
    Total           70005121.5     993    70498.6118

Bartlett's test for equal variances:  chi2(2) =   2.4514  Prob>chi2 = 0.294
```

s_B^2 s_W^2 s^2

k=3 groups, n=994 total observations. n-k=991

```
. di Ftail(2,991,2.45)
.08681613
```

ANOVA (t-test vs. F test)

- Note that if you only have two groups, you will reach the same conclusion running an ANOVA as you would with a t-test
- The test statistic F_{stat} will equal $(t_{stat})^2$

```
. oneway cd4count sex

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      521674.035      1   521674.035      7.41     0.0066
 Within groups      70155332.6    997   70366.4319
------------------------------------------------------------------------
    Total           70677006.7    998   70818.6439

Bartlett's test for equal variances:  chi2(1) =   0.0472  Prob>chi2 =
0.828
```

ANOVA (t-test vs. F test)

```
. ttest cd4count, by(sex)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |     374    299.6925     13.6301    263.5935     272.891     326.494
       2 |     625    346.9104    10.65047    266.2618    325.9953    367.8255
---------+--------------------------------------------------------------------
combined |     999    329.2332    8.419592    266.1177    312.7111    345.7554
---------+--------------------------------------------------------------------
    diff |           -47.21789    17.34162               -81.24815   -13.18762
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  -2.7228
Ho: diff = 0                                     degrees of freedom =      997

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0033         Pr(|T| > |t|) = 0.0066          Pr(T > t) = 0.9967
```

Statistical hypothesis tests

Data and comparison type	Alternative hypotheses	Test and Stata command
Numerical; One mean	H_a: $\mu \neq \mu_a$ (two-sided) H_a: $\mu > \mu_a$ or $\mu < \mu_a$ (one-sided)	Z or t-test • ttest var1=hypoth val.*
Numerical; Two means, paired data	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	Paired t-test • ttest var1=var2*
Numerical; Two means, independent data	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	T-test (equal or unequal variance) • ttest var1, by(byvar) unequal
Numerical, Two or more means, independent data	H_a: $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or $\mu_2 \neq \mu_3$ etc.	ANOVA •oneway var1 byvar
Dichotomous; One proportion	H_a: $p \neq p_a$ (two-sided) H_a: $p > p_a$ or $p < p_a$ (one-sided)	Proportion test • prtest var1=hypoth value* • bitest var1=hypoth value
Dichotomous; two proportions	H_a: $p_1 \neq p_2$ (two-sided) H_a: $p_1 > p_2$ (one-sided)	Proportion test (z-test) • prtest var1, by(byvar)
Categorical by categorical (nxk)		