

P-VALUE

INTERPRETATION

0.001

0.01

0.02

0.03

0.04

0.049

0.050

0.051

0.06

0.07

0.08

0.09

0.099

≥ 0.1

HIGHLY SIGNIFICANT

SIGNIFICANT

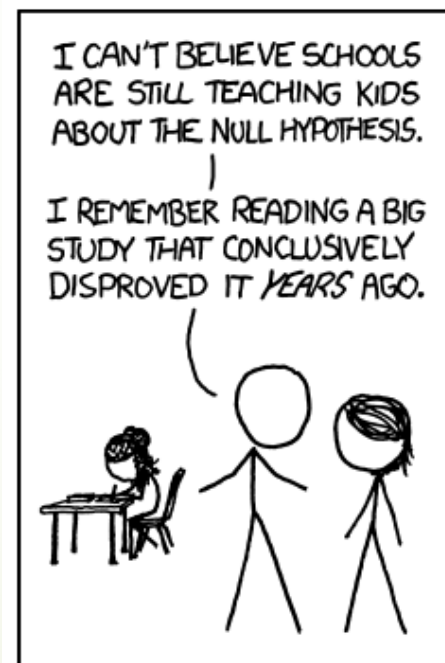
OH CRAP. REDO
CALCULATIONS.

ON THE EDGE
OF SIGNIFICANCE

HIGHLY SUGGESTIVE,
SIGNIFICANT AT THE
 $P < 0.10$ LEVEL

HEY, LOOK AT
THIS INTERESTING
SUBGROUP ANALYSIS

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).

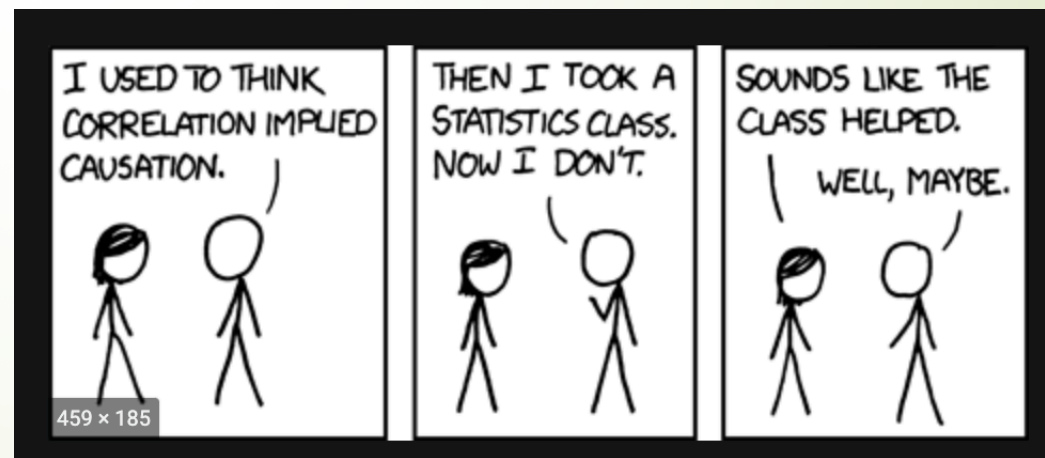


Pitfall 1: over-emphasis on p-values

- 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 video).

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

<u>Source</u>	<u>Example</u>
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

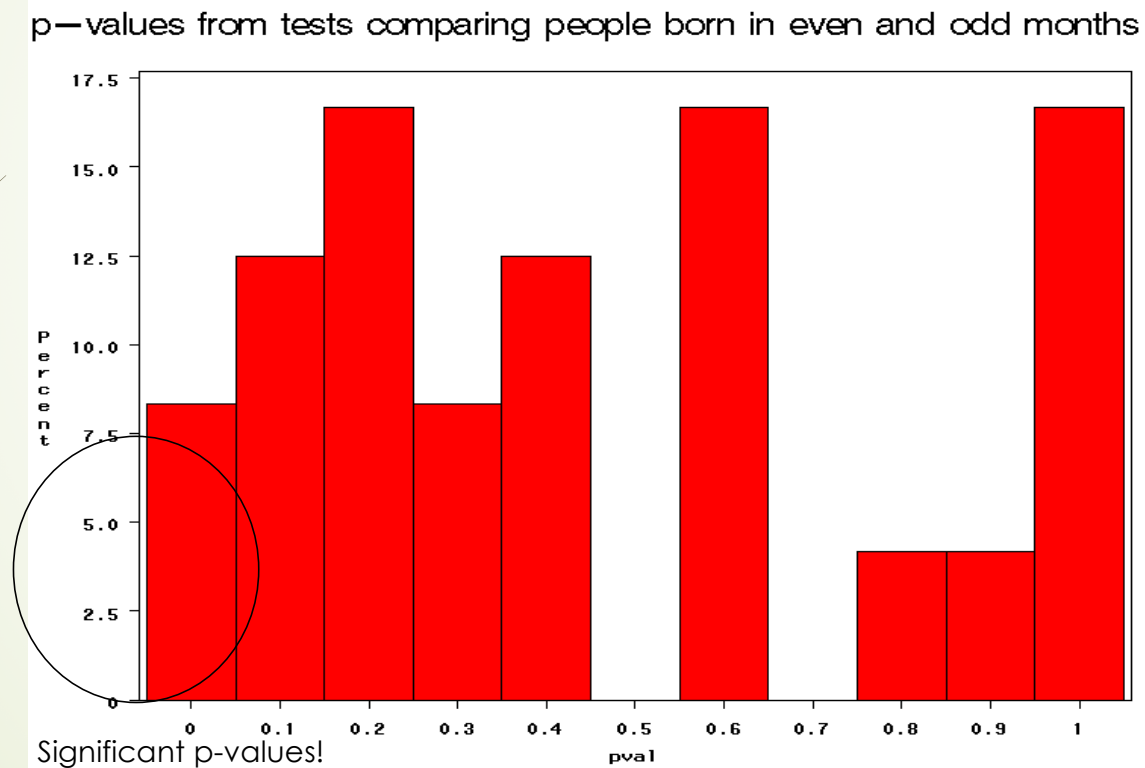
Results from Class survey...

- 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...

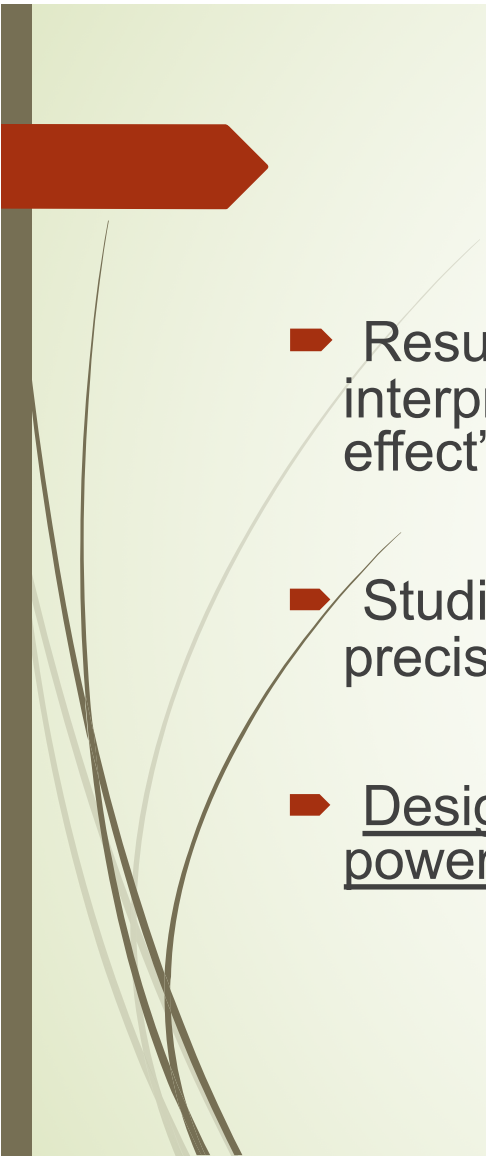
P-value distribution for the 25 tests...



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



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.

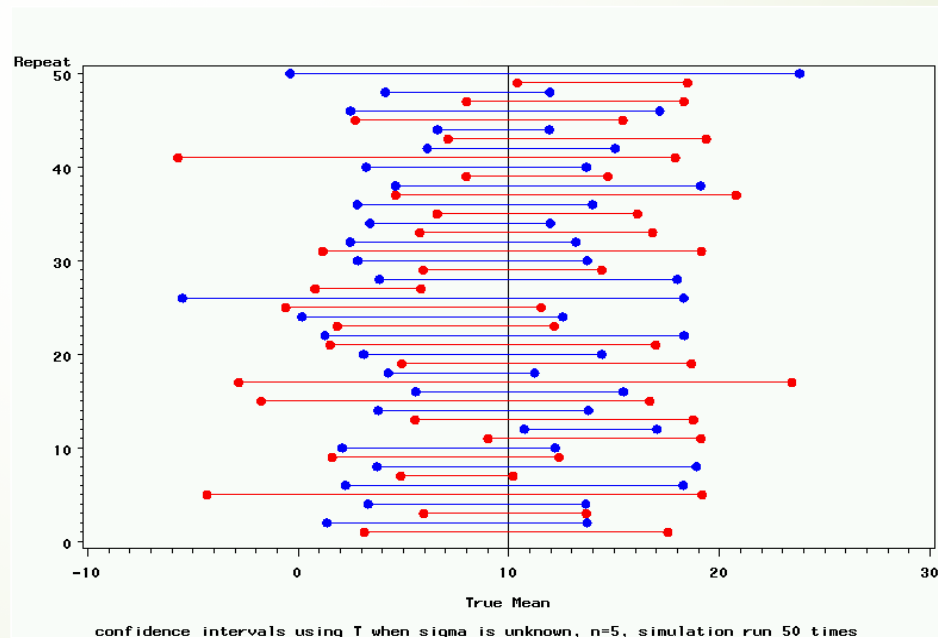
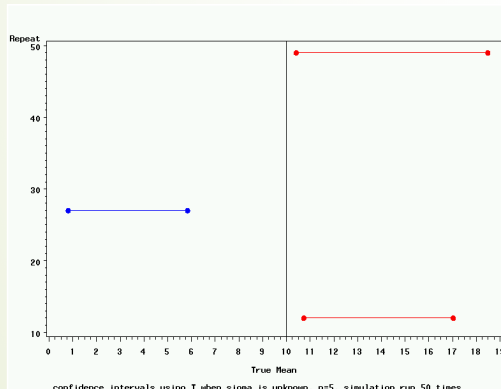


How about Confidence Intervals

- 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...



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

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

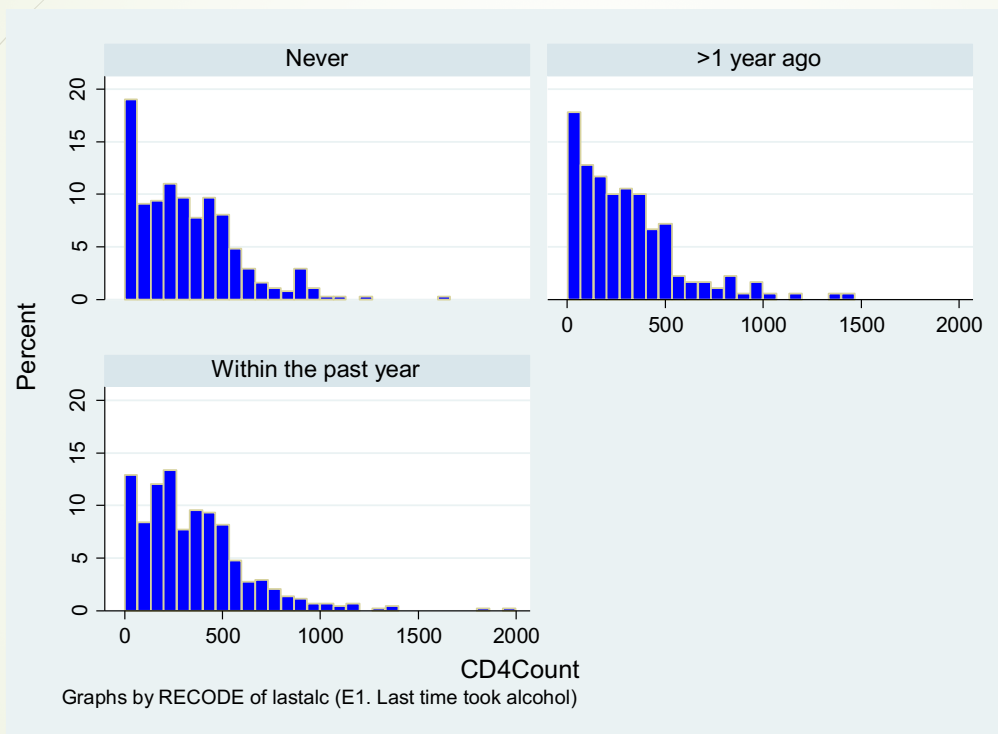


Using ANOVA instead of t-tests

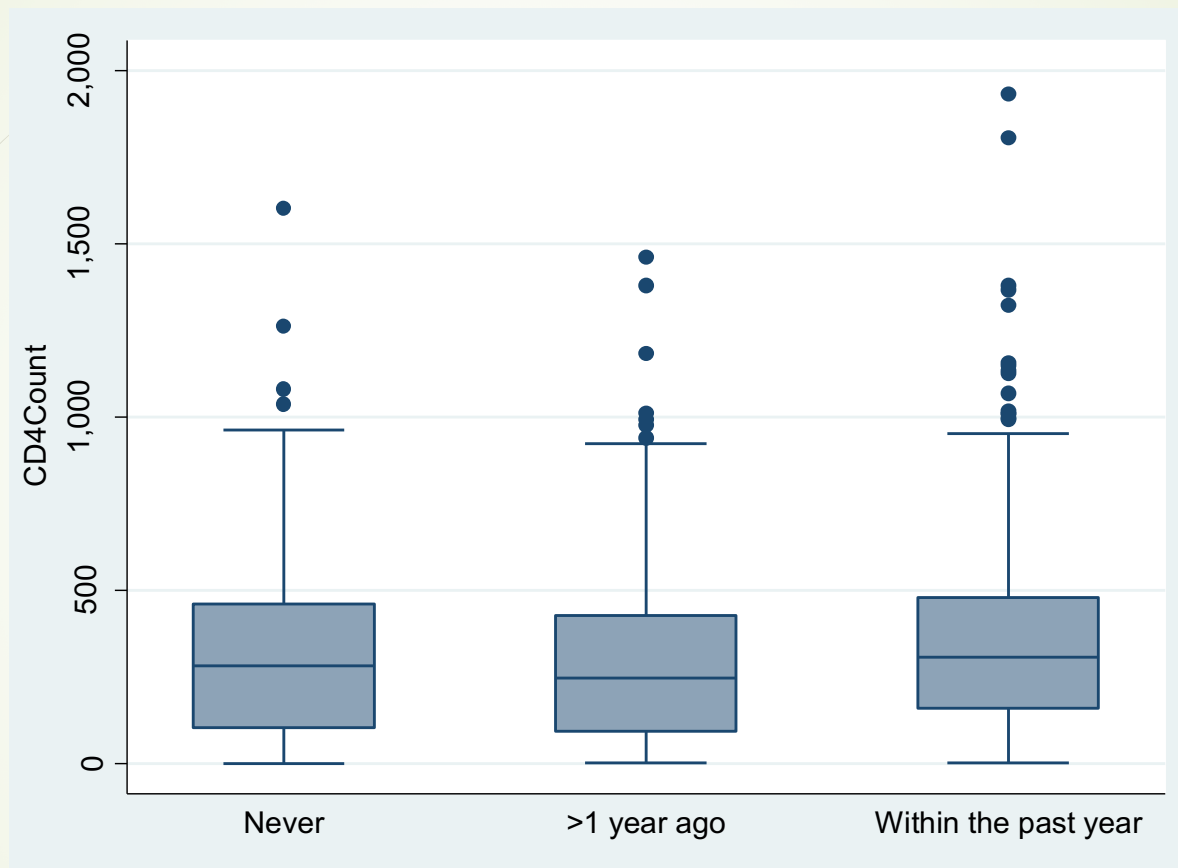
- The extension of the t-test to several independent groups is called analysis of variance or ANOVA
- The underlying data came from normally distributed populations. The observations are independent. The variance within each group over all groups (homoscedasticity)
- Why is it called analysis of variance?
 - 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 $\mu_1 = \mu_2 = \mu_3 = \dots$
 - The alternative H_A is that at least one of the means differs from the others

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 year	441	349.8662	273.9364	3	308	1932
Total	994	329.5322	265.5157	1	285	1932

ANOVA example

- CD4 count, by alcohol consumption category

oneway var groupvar

```
. oneway cd4count lastalc_3
```

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

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: $\chi^2(2) = 2.4514$ Prob> $\chi^2 = 0.294$

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 $(t_{\text{stat}})^2$

```
. 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
```

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 • <code>ttest var1=hypoth val.*</code>
Numerical; Two means, <u>paired data</u>	$H_a: \mu_1 \neq \mu_2$ (two-sided) $H_a: \mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	Paired t-test • <code>ttest var1=var2*</code>
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) • <code>ttest var1, by(byvar) unequal</code>
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 • <code>oneway var1 byvar</code>
Dichotomous; One proportion	$H_a: p \neq p_a$ (two-sided) $H_a: p > p_a$ or $p < p_a$ (one-sided)	Proportion test • <code>prtest var1=hypoth value*</code> • <code>bitest var1=hypoth value</code>
Dichotomous; two proportions	$H_a: p_1 \neq p_2$ (two-sided) $H_a: p_1 > p_2$ (one-sided)	Proportion test (z-test) • <code>prtest var1, by(byvar)</code>
Categorical by categorical (nxk)		