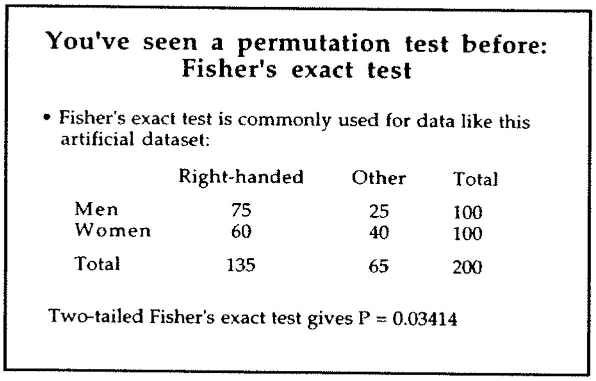
Some Uses of Permutation Tests and the Bootstrap in Craniofacial Research

(PI) The next presentation will be by ***Jim Hodges, University of Minnesota***. Jim is a biostatistician and is Director of the Biostatistical Core of the Minnesota Oral Health Clinical Research Center.He will present ***Some Uses of Permutation Tests and the Bootstrap in Craniofacial Research***

(HD)This talk is a friendly rejoinder to the talk Melissa Begg gave a year ago. Her talk was about analyzing binary outcomes in periodontal trials, where the outcomes for teeth within a mouth are correlated. There was nothing wrong with her talk; the problem I had with it was that she was comparing six different methods for a fairly specific problem, and I have a mathematician's memory, which is to say hardly any, and I can't remember different methods for specific problems. I'm not much of a mathematician either, so I got stiffed on both ends of that, I would think.

I'm hoping to convince you that permutation tests provide a general approach you can use on many different problems, including Melissa's, and you don't have to remember specific tests. Briefly , I will start out by convincing you that you have already seen a permutation test called Fisher's exact test. Then I will have a brief theological interlude so Larry [Laster] doesn't have a stroke. I will show you a couple of examples, then address the practical question of how many random permutations you need to use, and then I'll conclude.



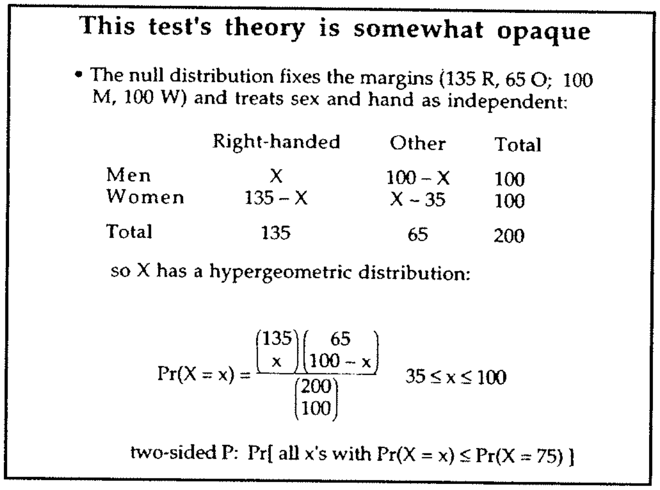
You've seen a permutation test before: Fisher's exact test

- Fisher's exact test is commonly used for data like this artificial dataset:

	Right-handed	Other	Total
Men	75	25	100
Women	60	40	100
Total	135	65	200

Two-tailed Fisher's exact test gives P = 0.03414

Figure 1



This test's theory is somewhat opaque

- The null distribution fixes the margins (135 R, 65 O; 100 M, 100 W) and treats sex and hand as independent:

	Right-handed	Other	Total
Men	X	100 – X	100
Women	135 – X	X – 35	100
Total	135	65	200

so X has a hypergeometric distribution:

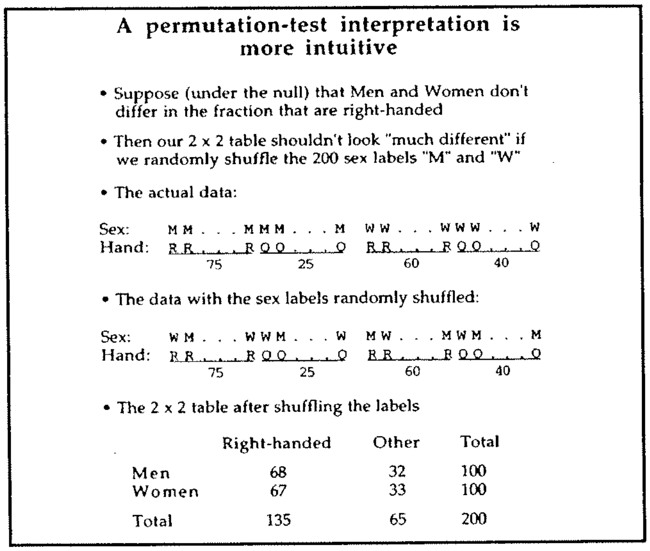
$$\Pr(X = x) = \frac{\binom{135}{x}\binom{65}{100-x}}{\binom{200}{100}} \qquad 35 \le x \le 100$$

two-sided P: Pr[all x's with $\Pr(X = x) \le \Pr(X = 75)$]

Figure 2

Here **(Figure 1)** is a typical data set you might insert into Fisher's exact test. I made up these data, so don't go tell your spouse about this one. The issue here is: Is gender related to handedness, right-handed *vs.* other. In this "study" we have 100 men and 100 women in the sample, and men are more likely to be right-handed. If you compute the two-tailed Fisher exact test you get a P-value of 0.034 plus some more decimal places.

The theory behind Fisher's exact test is opaque as usually presented **(Figure 2)**. In this test, you condition on the margins of the table and then derive the distribution of the test statistic as if there is in fact no relationship between gender and handedness. Conditioning on the margins means you fix that there are 100 men and 100 women and you fix that for men and women together there are 135 right-handed and 65 other-handed people. Fixing these margins leaves only one free value in the table; without loss of generality, I use the upper left cell of the table as the free value and call it X. Because of the marginal restrictions, X can only take values between 35 and 100, and under the null distribution of no relationship X has a hypergeometric distribution, as in Figure 2. You might, at this point, go to a text book to get some



A permutation-test interpretation is more intuitive

- Suppose (under the null) that Men and Women don't differ in the fraction that are right-handed
- Then our 2 x 2 table shouldn't look "much different" if we randomly shuffle the 200 sex labels "M" and "W"
- The actual data:

```
Sex:   M M . . . M M M . . . M   W W . . . W W W . . . W
Hand:  R R . . . R O O . . . O   R R . . . R O O . . . O
            75          25             60          40
```

- The data with the sex labels randomly shuffled:

```
Sex:   W M . . . W W M . . . W   M W . . . M W M . . . M
Hand:  R R . . . R O O . . . O   R R . . . R O O . . . O
            75          25             60          40
```

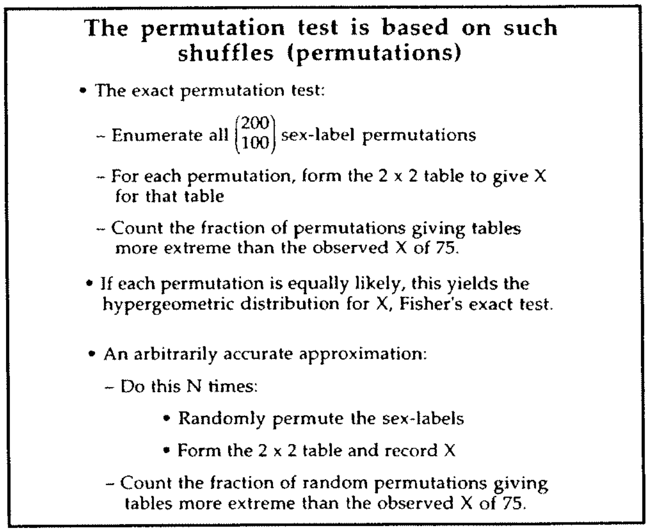
- The 2 x 2 table after shuffling the labels

	Right-handed	Other	Total
Men	68	32	100
Women	67	33	100
Total	135	65	200

Figure 3

intuition about the hypergeometric distribution and you'll find some song and dance about pulling balls out of urns. This has no intuitive content for me at all.

The permutation test interpretation is more intuitive **(Figure 3)** and now I'll tell you about it. Under the null hypothesis that men and women do not differ in the fractions that are right-handed, our two-by-two table shouldn't look much different if we randomly shuffle the 200 gender labels. To be more specific about that, here is the actual fake data in the middle of Figure 3. I have 100 men labeled by M and 100 women labeled by W and for each subject right-handedness is labeled by R and other-handedness by O. As in the two-by-two table, the men have 75 Rs and 25 Os, and the women have 60 Rs and 40 Os. In Fisher's exact test, you hold constant the numbers of Ms and Ws and the numbers of Rs and Os. I represented that here by showing, in the data with the sex labels randomly shuffled, that we still have 75 Rs and 25 Os, and 60 Rs and 40 Os, and because we just permuted (shuffled) the sex labels, we still have 100 Ms and 100 Ws. The only thing that's being changed from the actual (fake) data and the data with shuffled sex labels is where the 100 Ms and the 100 Ws are relative to the Rs and Os.



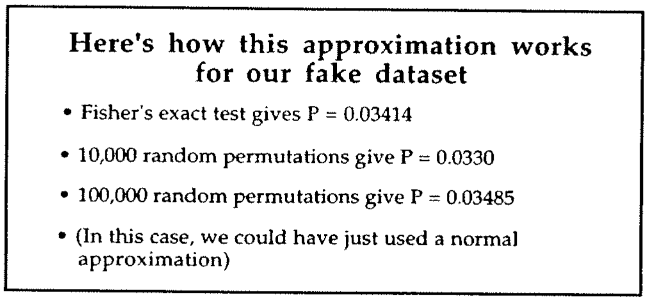
The permutation test is based on such shuffles (permutations)

- The exact permutation test:
  - Enumerate all $\binom{200}{100}$ sex-label permutations
  - For each permutation, form the 2 x 2 table to give X for that table
  - Count the fraction of permutations giving tables more extreme than the observed X of 75.
- If each permutation is equally likely, this yields the hypergeometric distribution for X, Fisher's exact test.
- An arbitrarily accurate approximation:
  - Do this N times:
    - Randomly permute the sex-labels
    - Form the 2 x 2 table and record X
  - Count the fraction of random permutations giving tables more extreme than the observed X of 75.

Figure 4

Having shuffled the Ms and Ws, we can now construct the two-by-two table for the shuffled data; it is at the bottom of **Figure 3**. In the shuffled dataset, we have deliberately destroyed whatever relationship there might be in these data between handedness and gender; this is what a draw from the null distribution might look like.

The permutation test is based on such shuffles or permutations **(Figure 4).** An exact permutation test would enumerate all 200-choose-100 ways of permuting the sex labels, and for each such permutation would extract X, the number of right-handed men. If you add the assumption that each permutation is equally likely, X has the hypergeometric distribution shown in Figure 2, and the P-value for Fisher's exact test could be obtained by counting the permutations giving more extreme values of X than the observed X of 75. I find this a more intuitive derivation of Fisher's test because it directly involves the relationship between handedness and gender. The problem with doing the exact permutation test is that there are 10^{59} permutations, so in this case you wouldn't actually do it. However, you can approximate the exact test arbitrarily well by using N random permutations of the sex labels. That is, N times you randomly permute the sex labels, and form the two-by-two table from the pseudo data and record the number of right-handed males. Then



Here's how this approximation works for our fake dataset

- Fisher's exact test gives P = 0.03414
- 10,000 random permutations give P = 0.0330
- 100,000 random permutations give P = 0.03485
- (In this case, we could have just used a normal approximation)

Figure 5

you count the fraction of random permutations giving tables more extreme than the observed 75, and that gives you a P-value approximating the exact one.

I computed this approximation for my artificial example **(Figure 5)**; the first 10,000 random permutations gave a P of 0.0330, and 100,000 random permutations gave P = 0.03485, both quite close to the exact value. **Figure 6** plots the exact distribution of X as well as the approximations based on 10,000 and 100,000 random permutations, and a normal approximation with the same mean and variance as X has in the 100,000 random permutations. As you can see, the exact and approximate distributions are very close to the normal distribution.

Now for the theological interlude **(Figure 7)**. As with anything else in statistics, the test statistic and P-value are just functions of the data, and their interpretation is a separate matter. There are three different interpretations of permutation tests, and it's a legitimate metaphor to call them Orthodox, Conservative and Reform, or maybe Reconstructionist, referring to the different tendencies in Judaism. The oldest interpretation arises in a randomized trial. If you have done a randomized trial or a randomized selection of a survey sample, then the rationale for using random permutations arises from the randomization itself. The permutation test just simulates the randomization in the actual study while forcing the simulated study to conform to the null hypothesis.

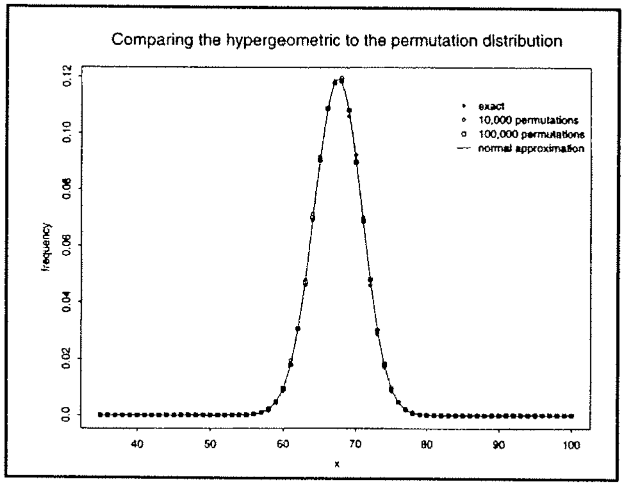
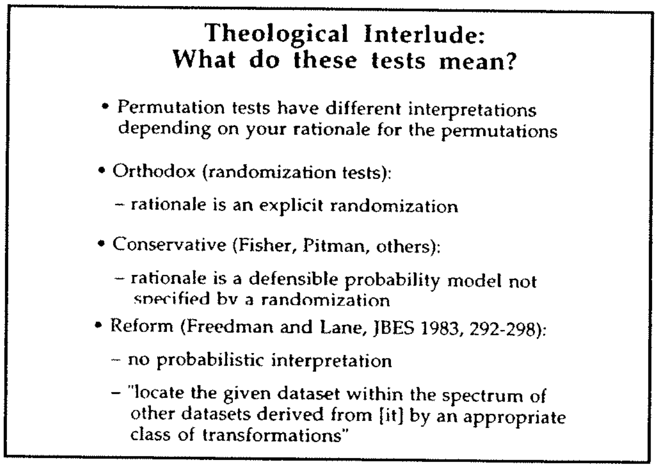


Figure 6



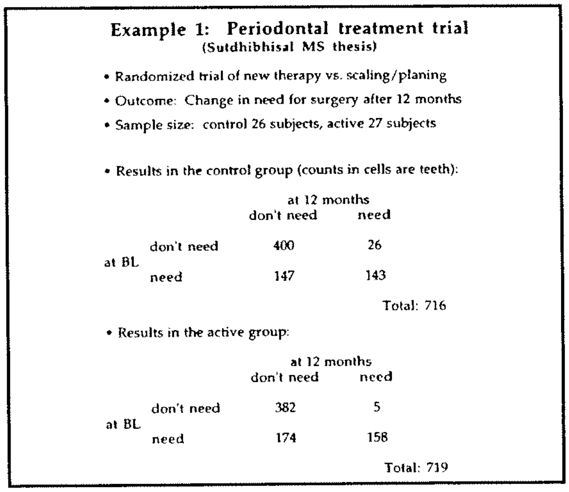
Theological Interlude: What do these tests mean?

- Permutation tests have different interpretations depending on your rationale for the permutations
- Orthodox (randomization tests):
  - rationale is an explicit randomization
- Conservative (Fisher, Pitman, others):
  - rationale is a defensible probability model not specified by a randomization
- Reform (Freedman and Lane, JBES 1983, 292-298):
  - no probabilistic interpretation
  - "locate the given dataset within the spectrum of other datasets derived from [it] by an appropriate class of transformations"

Figure 7

If you are a hard-core randomization person, the test I have shown you is not merely a test but is *the test*, the only legitimate one. The second interpretation, the Conservative interpretation, arises in cases in which you don't have an explicit randomization but you do have a highly defensible probability model. For example, Fisher used a lot of probability models arising from genetic theory. These are strongly motivated models and they can provide a rationale for a permutation test that does not arise from a randomization. Of course, those of the Orthodox persuasion don't accept this, and thirty years ago if you wanted to see statisticians scream at each other, you could make it happen by putting people of these two persuasions in the same room.

Now, there are instances in which neither the randomization rationale nor the defensible probability model are available, but some people want to use permutation tests in those cases anyway. David Freedman and David Lane have provided such an interpretation in the ***Journal*** *of Business and Economic Statistics* article noted on **Figure 7.** They rationalize the permutation test as being a way to "locate the given dataset within the spectrum of other datasets derived from [it] by an appropriate class of transformations." That's nearly uninterpretable out of context; The article is actually not bad.



Example 1: Periodontal treatment trial
(Sutdhibhisal MS thesis)

- Randomized trial of new therapy vs. scaling/planing
- Outcome: Change in need for surgery after 12 months
- Sample size: control 26 subjects, active 27 subjects

- Results in the control group (counts in cells are teeth):

		at 12 months	
		don't need	need
at BL	don't need	400	26
	need	147	143

Total: 716

- Results in the active group:

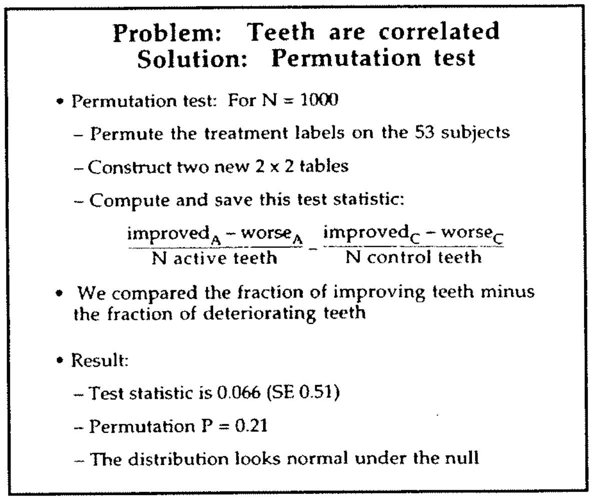
		at 12 months	
		don't need	need
at BL	don't need	382	5
	need	174	158

Total: 719

Figure 8

Now I will show you two examples **(Figure 8)**. The first is from a periodontal treatment trial, a data set we analyzed as part of the masters thesis of one of our periodontal residents, Sanutm Sutdhibhisal. The data come from a randomized trial of a new therapy versus scaling/planing. Forgive me if I'm vague about the nature of the study, but these data haven't been published yet. The outcome is the change in the need for surgery after twelve months of therapy.

There were 53 total subjects, 26 in the control group and 27 receiving active therapy. The two tables in **Figure 8** show you the results. The counts are teeth; this was a highly dentulous subject population and the average number of teeth per subject was just under 28. In the control group there were, at baseline, 426 teeth that did not need surgery (400 + 26), while 290 teeth did need surgery (147 + 143). Twelve months later, 169 teeth still needed surgery (26 + 143). So 147 teeth improved: they needed surgery at baseline but not at the 12 month follow-up. Also, 26 teeth got worse: they didn't need surgery at baseline but did at the 12 month follow-up. In the active group, the table is laid out the same way; 174 teeth improved and 5 teeth got worse. It looks like the subjects did better in the active group. But that's why statisticians are hired, to turn apparent successes into failures.



Problem: Teeth are correlated
Solution: Permutation test

- Permutation test: For N = 1000
  - Permute the treatment labels on the 53 subjects
  - Construct two new 2 x 2 tables
  - Compute and save this test statistic:

$$\frac{\text{improved}_A - \text{worse}_A}{\text{N active teeth}} - \frac{\text{improved}_C - \text{worse}_C}{\text{N control teeth}}$$

- We compared the fraction of improving teeth minus the fraction of deteriorating teeth

- Result:
  - Test statistic is 0.066 (SE 0.51)
  - Permutation P = 0.21
  - The distribution looks normal under the null

Figure 9

The problem here **(Figure 9)** is that the teeth within are mouth are, or may be, correlated. The solution is a permutation test. For N = 1,000 random permutations, we permuted the treatment labels among the 53 subjects. This is just like the earlier example of Fisher's exact test: each of the 53 subjects is labeled control or active; we permute those labels and create a pseudo data set of the same form as the original, but in which we've broken the relationship in the actual data between treatment and outcome. You could use any test statistic; we used the one in the figure because it had some intuitive value, comparing the two groups according to the net number of improved teeth as a fraction of the total number of people in the group.

We observed a test statistic of 0.066; the standard error associated with it, under the null, is 0.051, and the permutation-test P-value is 0.21. So as promised, I succeeded in making the result go away.

For 1,000 random permutations, **Figure 10** (page 5) is a QQ plot (normal quantile plot) of the 1,000 test statistics. In a QQ plot, if the points fall on a straight line, you can argue that the data look like a sample from a normal

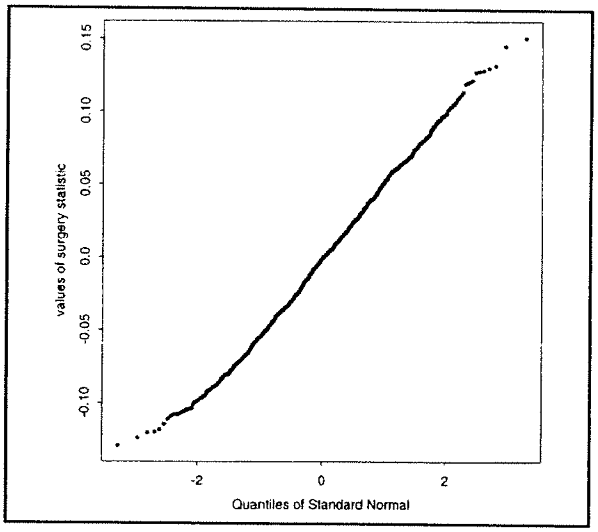
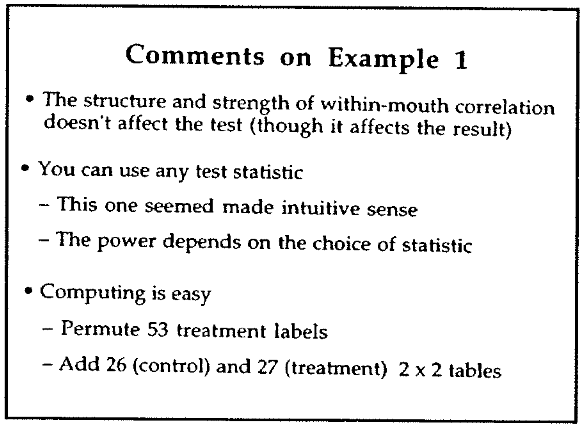


Figure 10

distribution. As you can see, this distribution looks pretty normal, a little short-tailed down in the lower tail.

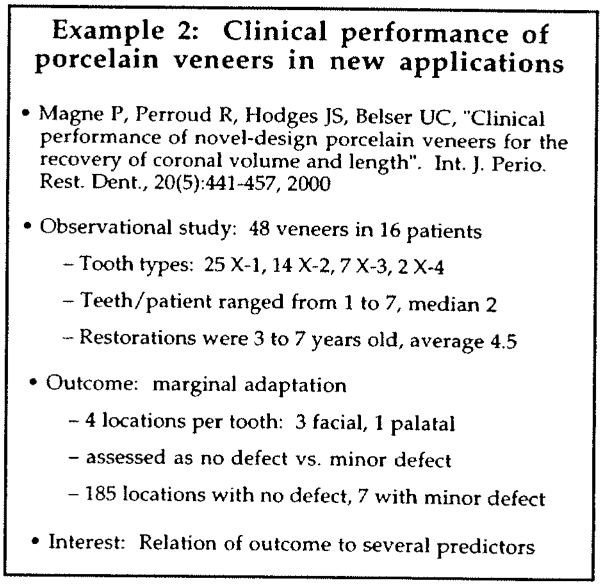
(Figure 11) Note that I didn't have to worry about the structure or strength of the within-mouth correlation; whatever it is, I use exactly the same test. The structure and strength of the correlation does affect the result, of course; the greater the correlation within mouth, the less information you get from all the extra teeth. Second, you can use any test statistic you want. We picked this one because it was easy and was readily interpretable. Obviously the power of the test depends on which statistic you use. Finally, the computing is a piece of cake. For each permutation you permute 53 treatment labels; for each subject you have a two-by-two table, so after permuting the labels and you just add up 26 and 27 two-by-two tables to get tables in the form of the actual data.



Comments on Example 1

- The structure and strength of within-mouth correlation doesn't affect the test (though it affects the result)
- You can use any test statistic
  - This one seemed made intuitive sense
  - The power depends on the choice of statistic
- Computing is easy
  - Permute 53 treatment labels
  - Add 26 (control) and 27 (treatment) 2 x 2 tables

Figure 11

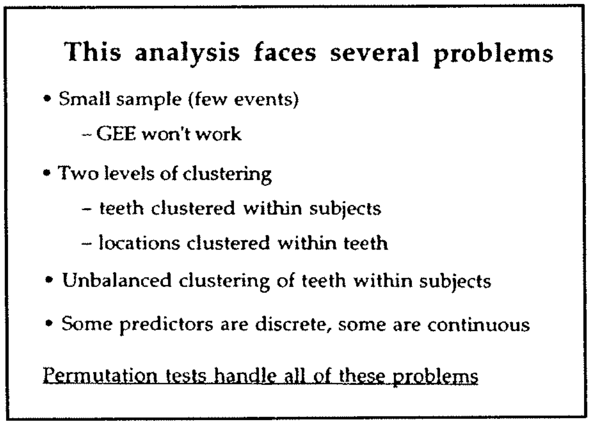


Example 2: Clinical performance of porcelain veneers in new applications

- Magne P, Perroud R, Hodges JS, Belser UC, "Clinical performance of novel-design porcelain veneers for the recovery of coronal volume and length". Int. J. Perio. Rest. Dent., 20(5):441-457, 2000
- Observational study: 48 veneers in 16 patients
  - Tooth types: 25 X-1, 14 X-2, 7 X-3, 2 X-4
  - Teeth/patient ranged from 1 to 7, median 2
  - Restorations were 3 to 7 years old, average 4.5
- Outcome: marginal adaptation
  - 4 locations per tooth: 3 facial, 1 palatal
  - assessed as no defect vs. minor defect
  - 185 locations with no defect, 7 with minor defect
- Interest: Relation of outcome to several predictors

Figure 12

The second example **(Figure 12)** is more complicated and, referring to my theological interlude, it is not Kosher for Passover, that is, there's no randomization to justify the random permutations. This study measured the clinical performance of porcelain veneers in new applications. By "new applications", I mean applications in which there is substantial tooth loss, that is, the veneer does more than just replace damaged enamel, for example. This work has been published; the citation is on the figure. The study was an observational study of the first 16 patients treated in this manner in the first author's practice. There were 48 teeth total; 25 of the teeth were central incisors, 14 were lateral incisors, and so on. The number of teeth per patient varied considerably; one patient had 7 teeth, 3 patients had one tooth and the median was two teeth per patient. The restorations were 3 to 7 years old. The outcome that I will discuss here is the marginal adaptation of the restoration. I will apologize right now that I can't discuss marginal adaptation intelligently; I'm not a dentist, much less a prosthodontist. If you have any dental questions, there are a couple of dentists here. Marginal adaptation was assessed at 4 locations per tooth, 3 on the facial side and 1 on the palatal side, and was assessed as "no



This analysis faces several problems

- Small sample (few events)
  - GEE won't work
- Two levels of clustering
  - teeth clustered within subjects
  - locations clustered within teeth
- Unbalanced clustering of teeth within subjects
- Some predictors are discrete, some are continuous

Permutation tests handle all of these problems

Figure 13

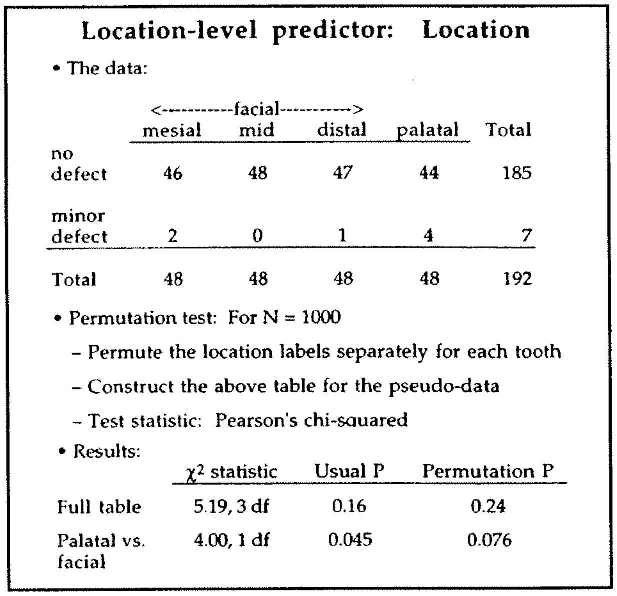
defect" or "minor defect". So with 48 teeth and four locations per tooth, that's 192 assessments. Of those 192, 7 had a minor defect. The interest here is the relation of this outcome to several predictors, such as restoration age and other things.

This analysis presents several difficulties **(Figure 13)**. It has a small sample in the sense there are very few events, only 7 minor defects. Thus, generalized estimating equations (GEE) won't work, because they rely on large-sample approximations. Indeed, with so few events, GEE algorithms will usually not even converge. We also have two levels of clustering: teeth clustered within subjects and locations clustered within teeth. We have unbalanced clustering of teeth within subjects, fairly substantial imbalance. Finally, some of our predictors are discrete and some are continuous. Permutation tests take care of all of these problems: I can use one type of test for everything. As you will see, though, I do have to do a little bit of thinking for each test.

I could show you examples of tests at each of the three levels, that is, at the patient, tooth and location levels. I will just show you two or one depending on how late I am.

(PI) You are fine.

(HD) In the first test I'll show you **(Figure 14)** the predictor is the location on the tooth. That is, we are interested in whether the four locations differed in their chance of having a defect. The data are in the table just above the middle of the figure. In the mesial facial location, out of 48 teeth, 46 had no defect and 2 had a minor defect. Mid-facial and distal facial had 0 and 1 defects respectively, while palatal had 4. Thus, the palatal location may be worse than the others.



Location-level predictor: Location

- The data:

	<-----------facial-----------		---------->		
	mesial	mid	distal	palatal	Total
no defect	46	48	47	44	185
minor defect	2	0	1	4	7
Total	48	48	48	48	192

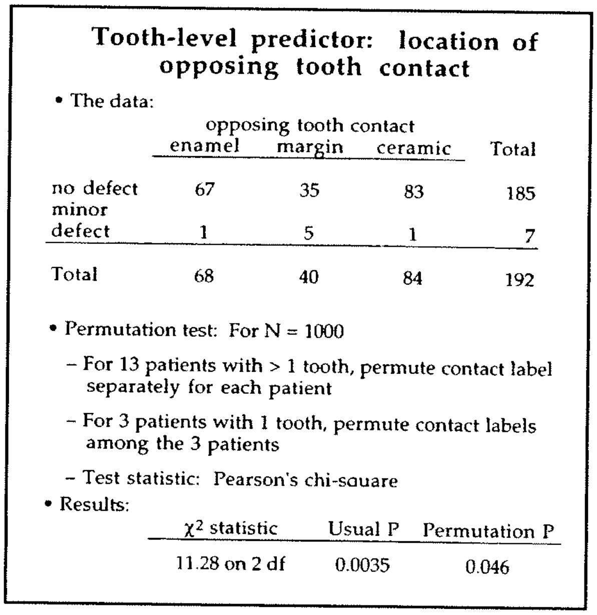
- Permutation test: For N = 1000
  - Permute the location labels separately for each tooth
  - Construct the above table for the pseudo-data
  - Test statistic: Pearson's chi-squared
- Results:

	χ^2 statistic	Usual P	Permutation P
Full table	5.19, 3 df	0.16	0.24
Palatal vs. facial	4.00, 1 df	0.045	0.076

Figure 14

In our permutation test, we did the following for each of 1,000 random permutations. First, permute the location labels within each tooth. You don't permute teeth within mouth and you don't make permutations across subjects because those permutations would have no effect at all on the pseudo data and wouldn't affect the results of the test. By permuting the location label separately for each tooth, you break the relationship in the actual data between tooth location and marginal adaptation, while preserving the aggregate result for each tooth and the correlation of the teeth within a mouth.

Having done the permutation, you construct the above table for the pseudo data and compute the test statistic. I used Pearson's χ^2 statistic. If you do this test on the full table (see "Results" at the bottom of the figure), with all four locations, the $\chi^2 = 5.19$ on 3 degrees of freedom, the usual P-value is 0.16, and the permutation test P is 0.24. The null distribution produced by this permutation test emphatically does not look like a normal distribution: it took only nine distinct values, of which five were 5.19 or larger. Naturally, my co-author asked, what if



Tooth-level predictor: location of opposing tooth contact

- The data:

	opposing tooth contact enamel	margin	ceramic	Total
no defect	67	35	83	185
minor defect	1	5	1	7
Total	68	40	84	192

- Permutation test: For N = 1000
  - For 13 patients with > 1 tooth, permute contact label separately for each patient
  - For 3 patients with 1 tooth, permute contact labels among the 3 patients
  - Test statistic: Pearson's chi-sоuare
- Results:

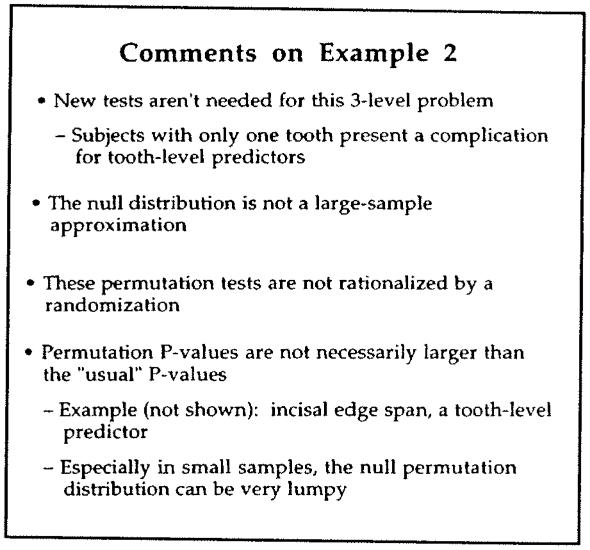
χ^2 statistic	Usual P	Permutation P
11.28 on 2 df	0.0035	0.046

Figure 15

we combined the three facial sites and just compared facial to palatal? If you do the usual test, you have a temptation into sin here with a P-value of 0.045, but if you do the permutation test you are rescued, because the P-value comes out to be greater than 0.05. Once again, the statistician has made the result go away. Again, the null distribution was highly non-normal, with probability on only six values, two of which were 4.00 or larger.

Now **(Figure 15)** consider a predictor at the tooth level. For each tooth restored, you can identify where it contacted the opposing tooth: was it on the enamel, the margin, or the ceramic? The table near the top of thefigure shows the data relevant to that issue. We still have 192 locations at which marginal adaptation was assessed, four per tooth. For restored teeth where the opposing tooth contact was on the enamel, there was one minor defect out of 68 sites (17 teeth); where the contact was on the margin, there were five minor defects out of 40 sites (10 teeth); and where the contact was on the ceramic, there was one minor defect out of 84 sites (21 teeth).

For the permutation test, we again did 1,000 random permutations, but we have a new complication here. We want to break the relationship between a tooth-specific measure (location of opposing tooth contact) and the result, marginal adaptation. So the natural thing to do is to permute teeth within patients, that is, to do a different permutation for each patient. However, three of the patients only had a single tooth, so there's nothing to permute. Thus, for the three patients that had a single tooth, we permuted the contact labels among those three patients. This permutation scheme breaks the relationship in the actual data between where the contact is on the tooth and what the results were for that tooth, while maintaining all of the other correlation structure in the data. Again, we used Pearson's chi-squared statistic, with two degrees of freedom, with the result shown in the Figure. The usual P-value is 0.0035, which makes people salivate, but here I was unable to be successful; they got a significant result from the permutation test (P = 0.046). I guess I will get kicked out of the statistician's union for that. For this permutation test, the null distribution took 24 distinct values, of which five were at least as large as the observed chi-squared statistic.

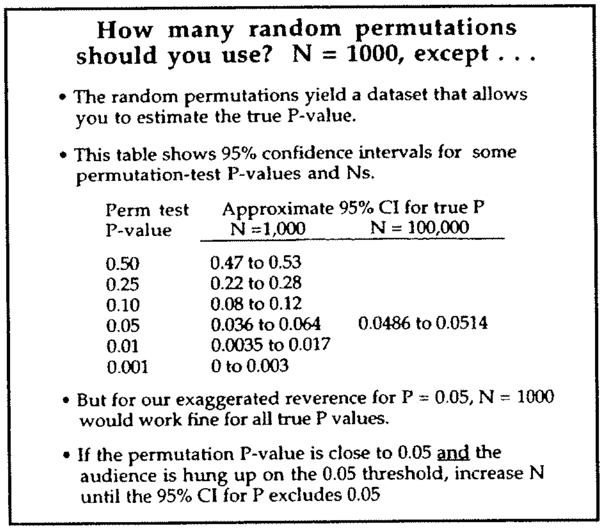


Comments on Example 2

- New tests aren't needed for this 3-level problem
  - Subjects with only one tooth present a complication for tooth-level predictors
- The null distribution is not a large-sample approximation
- These permutation tests are not rationalized by a randomization
- Permutation P-values are not necessarily larger than the "usual" P-values
  - Example (not shown): incisal edge span, a tooth-level predictor
  - Especially in small samples, the null permutation distribution can be very lumpy

Figure 16

Some comments on this example **(Figure 16)**. This is a three-level example, so we had to think up different permutation schemes for the different levels, but we didn't need to find or think up an entirely new test. We also had the complication (for the tooth level predictor) that some subjects had only a single tooth, but that



How many random permutations should you use? N = 1000, except . . .

- The random permutations yield a dataset that allows you to estimate the true P-value.
- This table shows 95% confidence intervals for some permutation-test P-values and Ns.

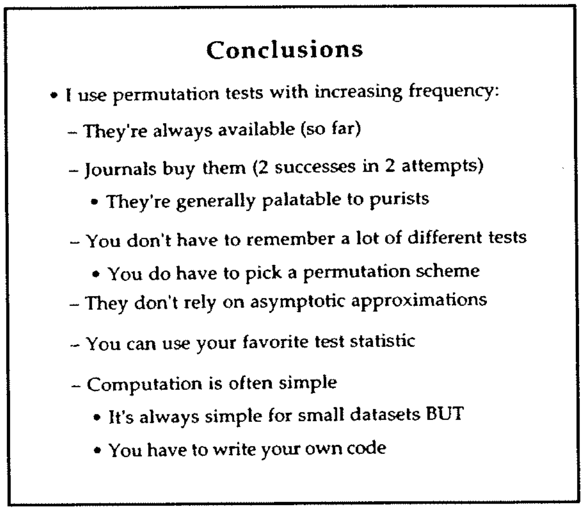
Perm test P-value	Approximate 95% CI for true P N =1,000	 N = 100,000
0.50	0.47 to 0.53	
0.25	0.22 to 0.28	
0.10	0.08 to 0.12	
0.05	0.036 to 0.064	0.0486 to 0.0514
0.01	0.0035 to 0.017	
0.001	0 to 0.003	

- But for our exaggerated reverence for P = 0.05, N = 1000 would work fine for all true P values.
- If the permutation P-value is close to 0.05 and the audience is hung up on the 0.05 threshold, increase N until the 95% CI for P excludes 0.05

Figure 17

was no big deal. Second, you don't get anything that looks like a chi-squared null distribution for any of these analyses: for some of the permutation tests I did, the probability in the permutation distribution lives on only five values of the test statistic. This is not surprising, because there were few events. So the advantage of the permutation test is that it gives a null distribution that is not a large-sample approximation. There was no randomization, and I can't rationalize by a randomization or a probability model, so we have to rely on the Freedman-Lane rationale. Finally, in another test I didn't show you the permutation P-value, in spite of the clustering, was actually smaller than the usual P-value. This can happen because the null permutation distribution is so discrete and lumpy in small examples. Thus, you can't do the usual test and assume the permutation test will give you a bigger P-value, a strategy which would save a lot of trouble if it worked.

Finally, for a practical issue **(Figure 17)**: how many random permutations should you use? I used 1,000 for all of the examples, and of course I should justify that. The key thing to understand is that when you are computing the permutation P-value, you're in effect asking how many pseudo datasets gave test statistics as large as or larger than the one actually observed. So you are estimating the P value by creating a dataset of test statistics drawn from the null.



Conclusions

- I use permutation tests with increasing frequency:
  - They're always available (so far)
  - Journals buy them (2 successes in 2 attempts)
    - They're generally palatable to purists
  - You don't have to remember a lot of different tests
    - You do have to pick a permutation scheme
  - They don't rely on asymptotic approximations
  - You can use your favorite test statistic
  - Computation is often simple
    - It's always simple for small datasets BUT
    - You have to write your own code

Figure 18

(LA) Would your unique applications here prohibit you from using things like Stat Exact, which does the permutation test.

(HD) Let me get back to that later. This is exactly the problem of computing a confidence interval for the proportion in a binomial distribution. Thus, if you have an observed permutation-test P-value, it is quite straightforward to get an approximate 95% confidence interval for the true P-value. **Figure 17** shows those confidence intervals for several permutation-test P-values for N = 1,000. So if you did 1,000 permutations and observed a P-value of 0.25, then with 95% confidence the true P-value is between 0.22 and 0.28. In this case, 1,000 random permutations is plenty: anywhere in that interval (and substantially below it) you get the same substantive result, not significant. That's true for every permutation test P-value in this table except for 0.05 or anything close to it: if your permutation test P-value is 0.05, then the actual value could readily be as small as 0.036 or as large as 0.064. If you use 100,000 permutations instead of 1,000, you have better resolution and usually will avoid this quandary.

Now, this 95% confidence interval of 0.036 to 0.064 doesn't bother me. I interpret any P-value in that interval to mean the same thing, that is, that you have a borderline result. But journal editors don't think that way, because they're in the business of suppressing hanky-panky. So if you are in a situation where the P-value is close to 0.05 and the audience is hung up on the 0.05 threshold, then you need to use a bigger N until the 95% confidence interval excludes 0.05.

My conclusion **(Figure 18)** is that I find myself using permutation tests more often; I've done it in other problems as well as the examples. I use it in preference to the bootstrap. If you are interested in this issue, I can talk to you about it at lunch. If you have a problem where you are interested in doing a permutation test, I will be happy to discuss that with you also.

Next, journals do buy permutation tests: I have tried twice and been accepted twice. A big advantage for me is that I don't have to remember a lot of different tests, although I do have to work something out for each application, namely the permutation scheme. These tests do not rely on asymptotic approximations. The computation is often simple.

I have a confession to make about the first example I showed you. I have a high workload and I should spend more time thinking. Because I don't, I didn't actually use the simple calculation I showed you in Figure 10, but rather permuted labels on the entire data set instead of just the labels on 28 two-by-two tables. Subject to Larry's comment, you do have to write your own code. I am probably the least computer-literate PhD statistician in my age cohort, and I can do it, so it can't be too hard. That's all I have to say on this. Can we have lights up?

(applause)

(PI) Questions.

(HD) Larry [Laster] asked, can you do this in StatXact? I don't know; do we have anybody here that knows a lot about StatXact?

(LA) I do.

(KI) In most of the applications that you showed, the sample sizes were large enough that StatXact will default to the asymptotic result. In the new version, I don't think that's true. You have more flexibility to specify that you want the exact computation. It might take a while, though.

(HD) It's important to keep in mind when the asymptotic approximation works well. In my second example, I had 7 events out of 192 possibilities. The relevant number for determining whether I have a "large" sample is 7, not 192. Does StatXact allow you to condition in the way I have in that example?

(LA) You might have to do some reformulating to get ready for StatXact. But in most cases, not all, it might save you a lot of time. You won't have to write any code.

(HD) You have to write the StatXact code.

(LA) It's almost nothing. StatXact is simple. There is no code.

(KI) For a continuous outcome, it might be more complicated. My question to you as a fellow permutation lover, is that we can do all of these permutation tests but we can pick any statistic we want. The theory will hold whatever is done. But you alluded to it earlier, the power for these tests will not be the same for all statistics. My question is, how do we intelligently pick a test statistic to summarize the data the most meaningful way?

(HD) I don't have anything intelligent to say about that except what I said earlier, namely that I have come to prefer continuous over discrete test statistics. Now, I know what Larry will say -- I sat next to him in dinner last night -- so I will anticipate him. In large samples, in simple cases the permutation test is equivalent to the t-test which is known to be the uniform uniformly most powerful etc, etc.

(LA) And similar to distribution-free.

(HD) I guess don't have anything useful to tell you on that. I just try to stick to popular statistics.

(PL) It is time for lunch.

(The luncheon recess is taken)