Joint Statistical Meetings
ASA CE Shortcourse 2002-10C
11 August 2002

www.bio.ri.ccf.org/UnifyPow

Sample Size Analysis in Study Planning

(using UnifyPow.sas)

Ralph O'Brien, PhD
Biostatistics and Epidemiology
The Cleveland Clinic Foundation

Examples

Built-in capabilities:

New tools from SAS Institute

SAS Institute developer John Castelloe and his colleagues are building terrific tools in this area that will debut, in part, in V9.0, due out "quietly" in late 2002. The main release of V9 has been delayed to V9.1, scheduled for release in 2003.

First products:

- PROC POWER will handle many of the basics.
- PROC GLMPOWER will handle linear models using syntax and modeling structure congruent with PROC GLM.
- Java-based optional interface that works from your web browser.

These PROCS will grow in functionality and new power procs will be added, oneday making UnifyPow obsolete. I am consulting actively and formally on all this.

What is statistical power?

Statistical power is the **probability that a given inference test will be statistically significant,** i.e. p-value $<$ α-level.

Decision (sample data)	Truth ("infinite data"): no effect	Truth ("infinite data"): some effect
"insufficient evidence"	correct	Type II error (β)
"evidence in favor of effect"	Type I error (α)	correct (power)

If the null hypothesis is true, then "nullpower" $= \alpha$.

What does power buy us?

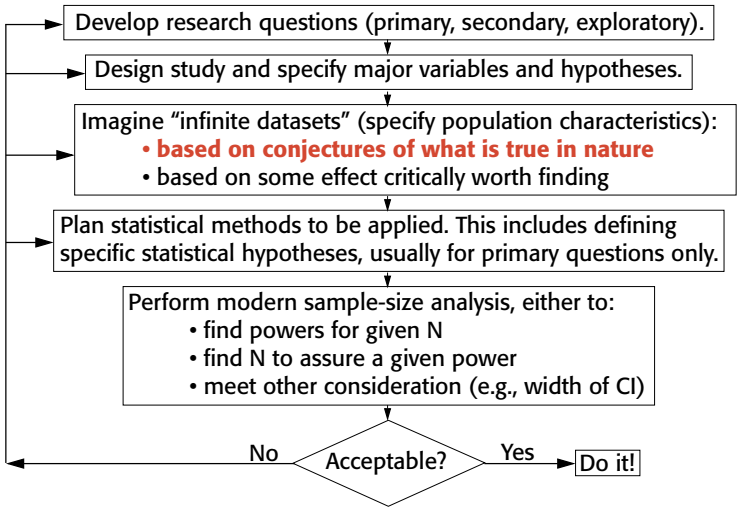
Some think that statistical power is the chance of obtaining a **publishable result.**

If power is high and the test is non-significant, this implies that the effect, if any, is small. So **non-significant studies will be informative** when there is high power to detect critical effects. (higher "negative predictive value")

Studies with high power give researchers greater confidence that **a statistically significant result reflects the truth.** (higher "positive predictive value")

Reviewers require adequate statistical power to assure that animals and patients are being studied ethically and resources are being allocated wisely.

Flowchart of sample-size analysis



Which result is more impressive?

S1. Sam Shotgun is a born empiricist. He runs lots of studies to see what he can discover. Even though he figures that only 20% of his research hypotheses are actually true (**"right idea" rate = 20%**), by running many studies he will make enough discoveries (get enough significant results) to contribute well to his field. Shotgun's last study was underfunded, so **N was small**. For a given conjecture about the "infinite dataset," he computed that **power was 40%** for **$\alpha = 0.05$.** He ran the study anyway and got **$p = .042$** for the primary result.

S2. Same as S1, but suppose that Shotgun was funded well enough to recruit **a larger N** so that **power was 90%** for **$\alpha = 0.05$.** He got **$p = .042$** for the primary result.

Which result is more impressive?

W1. Wanda Wellfocused develops study questions with great care and guesstimates that 60% of her research hypotheses are actually true (**"right idea" rate = 60%**). But she often lacks the resources to study them with sufficient sample sizes. She ran her latest study at a **small N**. For a given conjecture about the infinite dataset, she had computed the **power to be 40%** for **$\alpha = 0.05$.** She got **$p = .042$** for her primary result.

W2. Same as W1, but suppose that Wellfocused got solid funding and had a **larger N,** so that the **power was 90%** at **$\alpha = 0.05$.** He got **$p = .042$** for his primary result.

Which result is more impressive?

Investigator, sample size	"Right Idea" Rate (θ)	Type I Error Rate (α)	Power ($1 - \beta$)	p
Shotgun, lower N	.20	.05	.40	.042
Shotgun, larger N	.20	.05	.90	.042
Wellfocused, lower N	.60	.05	.40	.042
Wellfocused, larger N	.60	.05	.90	.042

Shotgun, underfunded, runs 100 studies. Expectations:

	Outcome	
State of Nature	**evidence for effect (test significant)**	**insufficient evidence for effect (test non-sig)**
no effect (null true) ($1 - \theta = .80$)	5% of 80 = 4	95% of 80 = 76
effect (null false) ($\theta = .20$)	40% of 20 = 8	60% of 20 = 12
	Type I Junk Rate[a] = 4/12 = .333	Type II Junk Rate[b] = 12/88 = .136

a. **Type I Junk Rate:** The test was significant; so Shotgun concludes that there is an effect. What is the chance that this result is "junk," that is, there really is no effect present in nature (the infinite dataset).

b. **Type II Junk Rate:** The test was non-significant. What is the chance that this result is "junk," that is, we missed finding an effect present in nature?

Type I Junk Rate (1 - Positive Predictive Value)

P[effect is null | test is significant] $= 4/(4 + 8) = 0.333$

$$= \frac{P[\text{test sig} \mid \text{no effect}]P[\text{no effect}]}{P[\text{test sig} \mid \text{no effect}]P[\text{no effect}] + P[\text{test sig} \mid \text{specific effect}]P[\text{specific effect}]}$$

$$= \frac{\alpha\bullet(1 - \theta)}{\alpha\bullet(1 - \theta) + \text{power}\cdot\theta}$$

Type II Junk Rate (1 - Negative Predictive Value)

P[true effect | test is not sig] $= 12/(12 + 76) = 0.136$

$$= \frac{P[\text{test not sig} \mid \text{specific effect}]P[\text{spec. effect}]}{P[\text{test not sig} \mid \text{specific effect}]P[\text{spec. effect}] + P[\text{test not sig} \mid \text{no effect}]P[\text{no effect}]}$$

$$= \frac{(1 - \text{power})\bullet\theta}{(1 - \text{power})\bullet\theta + (1 - \alpha)\bullet(1 - \theta)}$$

Investigator, sample size	"Right Idea" Rate (θ)	Type I Error Rate (α)	Power ($1 - \beta$)	Type I Junk Rate	Type II Junk Rate
Shotgun, lower N	.20	.05	.40	.333	.136
Shotgun, larger N	.20	.05	.90	.182	.026
Wellfocused, lower N	.60	.05	.40	.077	.486
Wellfocused, larger N	.60	.05	.90	.036	.136

Conclusions: "Shotgun" studies may operate at a 5% Type I error rate, but they have high Type I junk rates. "Well-focused" studies may operate at a 10% Type II error rate (90% power), but their Type II junk rates will exceed that. Whether shotgun or well-focused, higher power lowers both types of junk rates.

"justification of sample size is crux of design"

"... Ideally, clinical trials should have adequate **power, ≈90%**, to detect a **clinically relevant difference** between the experimental and control therapies. Unfortunately, the power of clinical trials is **frequently influenced by budgetary concerns** as well as pure biostatistical principles. Yet an underpowered trial is, by definition, unlikely to demonstrate a difference between the interventions assessed and may ultimately be considered of little or no clinical value. **From an ethical standpoint**, an underpowered trial may put patients needlessly at risk of a new therapy without being able to come to a clear conclusion."

Topol EJ, et. al. ***Circulation,*** 1997, 95: 1072-1082

Non-directional vs. directional hypotheses

("Two-tailed" vs. "One-tailed" Tests)

- "Two-tailers always" say all hypotheses are really non-directional because researchers will usually rationalize and report any result that goes in the "wrong" direction. Some clinical trialists say that all trials should be non-directional, because one direction tests efficacy and other tests harm.
- "Flexible scientists" (Note my bias!) say that studies have unique backgrounds and goals. If the research hypothesis is directional, then so should be the statistical hypothesis. This will gain power—if you pick the right direction. If the research hypothesis is non-directional, then so should be the statistical hypothesis. This broadens the question, making any statistically significant result appear more "honest."

UnifyPow mission

To develop comprehensive SAS-based tools for research planning, in particular, for choosing sample sizes and assessing statistical power.

Communicating

UnifyPow is copyrighted freeware available at
www.bio.ri.ccf.org/UnifyPow

Ralph O'Brien corresponds best via email to
robrien@bio.ri.ccf.org

Main things about UnifyPow

A shameless boast: UnifyPow is freeware—but you must have SAS—that rivals freestanding commercial Wintel-based applications.

- You do not need to be a SAS expert. **easy to learn**
- Runs in base SAS system, thus under numerous platforms & configurations **cross-platform portability**
- Builds a SAS data set of results. Thus, SAS users can develop customized reports, even merging results from two or more UnifyPow runs. **reporting flexibility**

More main things

- Handles
  - Unbalanced sample sizes in G independent groups for all relevant problems.
  - One- and two-tailed tests when appropriate.
  - Any alpha level.
  - General contrasts (including $df_H > 1$) on cell means, logits, and Fisher's Z-transformed correlations.

  depth: as much as any, but still a long way to go
- Strives hard to use exact or virtually exact computations whenever feasible. **accuracy**

http://www.bio.ri.ccf.org/UnifyPow

- Acrobat (PDF) files:
  - workshop notes
  - SUGI 23 Proceedings paper, updated as UnifyPow evolves. Describes methods covered.
  - other stuff
- Text files:
  - UnifyPow source code. Algorithms described within code.
  - Instructions for implementation.
  - Test set of examples from full-day workshop.
- **Simple way to register name and email address.**

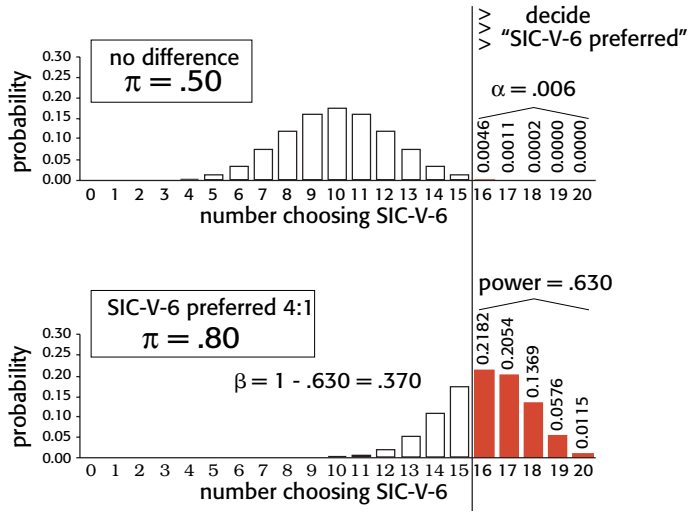
Testing a single proportion (the sign test)

- Julia Chill's Frozen Sensations will have people taste soy vanilla ice cream produced by each of two methods (SIC-V-5 and SIC-V-6). The study is forced choice (must pick one) and blinded.
- SIC-V-6 is the new method, so it must be shown to be better or the project will be stopped.
- N = number of tasters; Y = number that prefer SIC-V-6
  π = true (unknown) preference rate for SIC-V-6.
- The standard hypothesis:
  <SIC-V-6 not better> H_0: $\pi \leq .50$
  <SIC-V-6 better> H_a: $\pi > .50$

- Chill's conjectures that SIC-V-6 is preferred 4:1 (80%).
- N = 20 tasters.
- Decision rule: Conclude
  - "SIC-V-6 preferred" if Y ≥16 (≥80% people choose it)
  - "Not sufficient evidence" if Y ≤ 15

How good is this simple experiment?

Binomial distributions, N = 20



Key probabilities: N = 20

- α = p-value for 16 "SVIC-6" = Pr[Y ≥ 16; π = .50] = .0046 + .0011 + .0002 ≈ .006
- power = Pr[Y ≥ 16; π = .80] = .2182 + .2054 + .1369 + .0576 + .0115 ≈ .630

Truth ("infinite data")

Decision (sample data)	no difference	SVIC-6 preferred
"insufficient evidence"	correct	Type II error (β)
"evidence says SVIC-6 is preferred"	Type I error (α)	correct (power)

UnifyPow: testing single proportion (sign test)

```
%let UnifyPow = your file specification here;

%include "&UnifyPow";

datalines4;

proportion .80 . conjectures 4:1 preference
null .50 . (This is default; not needed)
alpha .01 .05
Ntotal 20 40
// balance sides
// NoNotes
;;;;

%tables
```

See below how to get Ntotals for:

alpha .005
power .990 .995

The // syntax makes these lines comments, so not in effect here.

Scenario: proportion .80 . conjectures 4:1 preference

		ALPHA			
		0.01		0.05	
		Total N		Total N	
		20	40	20	40
		Pow-er	Pow-er	Pow-er	Pow-er
Method	Type				
Exact Binomial	2-side bnml	.630	.912	.804	.981
	1-side bnml	.630	.957	.804	.992

Critical values and actual alpha levels using binomial distribution.

			ALPHA					
			0.01			0.05		
Method	Total N	Type	Actual Alpha	Lower Crit Value	Upper Crit Value	Actual Alpha	Lower Crit Value	Upper Crit Value
Exact Binomial	20	2-side bnml	0.007	3	16	0.041	5	15
		1-side bnml	0.006	.	16	0.021	.	15
	40	2-side bnml	0.006	11	29	0.038	13	27
		1-side bnml	0.008	.	28	0.040	.	26

Tight control of Type I and II errors

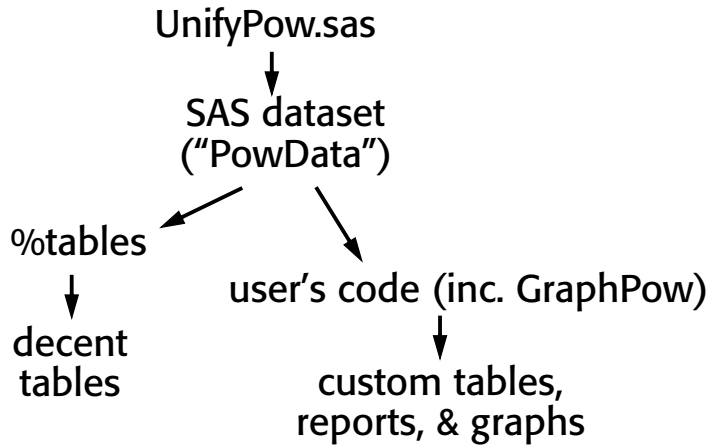
(Note use of UnifyPow comment blocks.)

Julia Chill's is statistically savvy and wants very tight control of Type I and Type II errors, so runs the following.

```
datalines4;
/#
Same problem, but now find minimum N to
achieve specified power at given alphas.
#/
proportion .80
power .99 .995
alpha .005
sides 1
;;;;
%tables
```

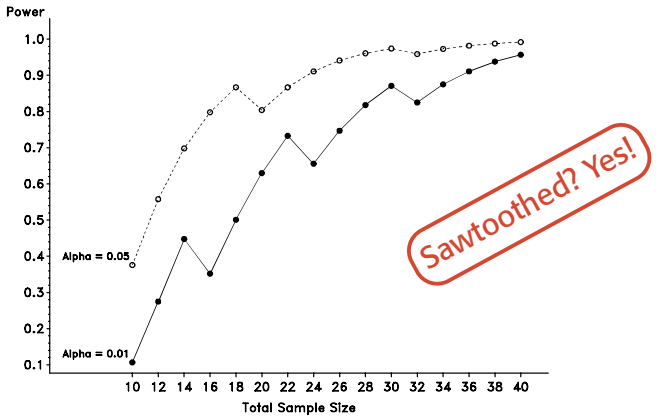
		ALPHA	
		0.005	
		Minimum Power	
		.990	.995
		Total N	Total N
Method	Type		
Exact Binomial	1-side bnml	59	66

UnifyPow.sas has nice reporting flexibility



GraphPow (permanently in beta)

- **GraphPow is a SAS freeware macro that plots UnifyPow results.**
- Developed by Christine Skibinski, Cleveland Clinic Foundation.
- SAS/GRAPH must be installed.
- Write robrien@bio.ri.ccf.org or cskibins@bio.ri.ccf.org if you want to try it.



Testing single proportion (custom null)

```
datalines4;
/#
This is a “Phase II” kind of study.  If it gives
promising results, then a larger, better study
will be conducted.

Julia Chill’s Frozen Sensations produces key lime
pies. Her food engineers have developed a
potentially new way (C12) to manufacture the crust
(hardest component) and they want to compare
it with the current one (C11).

They reason that in order to profitably switch to
the new method, it must shown to be preferred by
at least 65% when taste tested.
```



```
They also reason that a Type I error here will not
have big consequences, because C12 will not go
into production unless it is succeeds in a larger
study. A Type II error is more costly, because
this would shut down a line of research that
should have gone forward. So they are willing to
set an alpha level of 0.20.

They conjecture that 85% prefer the new method.

#/
proportion .85 . subjective conjecture
null .65 . need to say that >65% prefer new
alpha .05 .20
power .95 .90 . 5% and 10% Type II error rate
tails 1
;;;;
%tables
```



```
NOTE: Testing Ho: pi = 0.65

------------------------------------------------------------

The SAS System

Scenario: proportion .85 . subjective conjecture
```

	alpha			
	0.050		0.200	
	Minimum Power		Minimum Power	
	0.950	0.900	0.950	0.900
	Total N	Total N	Total N	Total N
Method	Statistic				
Exact Binomial	1-side bnml	52	42	28	22

Critical values and actual alpha levels using binomial distribution.

		alpha					
		0.050			0.200		
		Actual Alpha	Lower Crit Value	Upper Crit Value	Actual Alpha	Lower Crit Value	Upper Crit Value
Method	Minimum Power	Type					
Exact Binomial	0.950	1-side bnml	0.045	.	40	0.182	.	21
	0.900	1-side bnml	0.043	.	33	0.163	.	17

```
These critical values are part of the rejection region.
The note above describes how they are set.
```

How does UnifyPow set this two-tailed critical region?

```
NOTE: SETTING 2-TAILED CRITICAL REGIONS FOR THE BINOMIAL DISTRIBUTION.
      Denote the critical regions for a 2-tailed test as "major" and
      "minor" depending on which one is consistent with the true pi.
      Thus for Ho: pi = .35 with a conjecture of true pi = .20, the
      major critical region would be in the lower tail of the
      binomial(Ntotal, .35) distribution.  Let alpha_major and
      alpha_minor be the Type I error rates in these tails.  UnifyPow
```

Simple answer: It puts as much of $\alpha/2$ in the minor tail (direction opposite to conjecture) as it can. Then it puts as much of the "unused" α (at least $\alpha/2$) as it can into the major tail, that is, the one consistent with the conjectured effect.

```
      The critical values tabled below are in the rejection region.
      For example, the alpha = .05, two-tailed test of Ho: pi = .35 with
      NTotal = 40 gives lower and upper critical values of 8 and 21
      if the conjectured true pi is less than .35.  Thus, the major
      critical region is r = 0, 1, ..., 8 and the minor one is
      r = 21, 22, ..., 40.  If the conjectured true pi exceeds .35,
      then the major region is r = 20, 21, ..., 40 and the minor one is
      r = 0, 1, ..., 7.
```


Balancing the binomial's tails

- Same "5 vs. 6" tasting study as above, but use "balance tails" to keep each tail's Type I error ≤ alpha/2.
- Most common method, but does not "spend" all of the alpha optimally with respect to the conjectured true vs. null binomial probabilities. Note loss of power for N = 20 and alpha = 0.01.

```
proportion .80 . conjecture is 4:1 for SIC-V-6
null .50 . (This is default; not really needed.)
Ntotal 20 40
alpha .01 .05
balance tails
;;;;
%tables
```



```
Scenario: proportion .80 . conjecture is 4:1
preference for SIC-V-6
```

		alpha			
		0.010		0.050	
		Total N		Total N	
		20	40	20	40
		Pow-er	Pow-er	Pow-er	Pow-er
Method	Statistic				
Exact Binomial	2-tail bnml	.411	.912	.804	.981
	1-tail bnml	.630	.957	.804	.992

From Road Runner Sports

2 out of 3 runners* are injured at any time

- training errors
- muscle imbalances or weaknesses
- wrong shoes

*a statistic that is impossible for me to believe

Comparing 2 independent means

BeeBop Athletic Equipment is testing the "XDM-X," an experimental prototype of a running shoe. Possible successor to popular XDM model.

Research question:

Does the XDM-X shoe reduce injuries?

Design

- High-mileage runners will be randomly assigned to run in either

  XDM-X eXperimental shoe, or
  XDM-S Standard XDM altered cosmetically to make it also look like experimental.

- To get sufficient durability data on the XDM-X:

  2/3 of the runners will get the XDM-X.
  1/3 of the runners will get the XDM-S.

- Runners will train at least 5 times/week, 50 miles/week for 26 weeks.

Main outcome measure

P_i = Proportion of days i-th Runner is injured, including days when he/she runs with the injury.

To be handled as continuous variable.

The "Elicitation Process" gets more complex

What is the infinite dataset? (the population distributions)

Scenario to examine—

Ample experience with the XDM shoe suggests
median of P for XDM-S: 0.09

BeeBop conjectures
median of P for XDM-X: 0.07

Conjectured spread: 95% of XDM-S runners will have P between 0.01 and 0.23. Spread with XDM-X will be similar.

Arcsin transform: To better meet the distributional needs of the ordinary two-group t test, the analysis will use

$$Y_i = \arcsin(P_i^{1/2}).$$

Scenario becomes:

mean of Y for XDM-S: $\mu_S = \arcsin(0.09^{1/2}) = 0.30$

mean of Y for XDM-X: $\mu_X = \arcsin(0.07^{1/2}) = 0.27$

Guesstimating the SD. The 95% range for P_i of 0.01 to 0.23 becomes a 95% range for Y_i of 0.10 to 0.50. Taking Y_i as Normal, this gives $4\sigma = 0.40$, or $\sigma = 0.100$. To assess sensitivity, σ will be bracketed by $\sigma = 0.080$ and $\sigma = 0.125$.

Look reasonable?

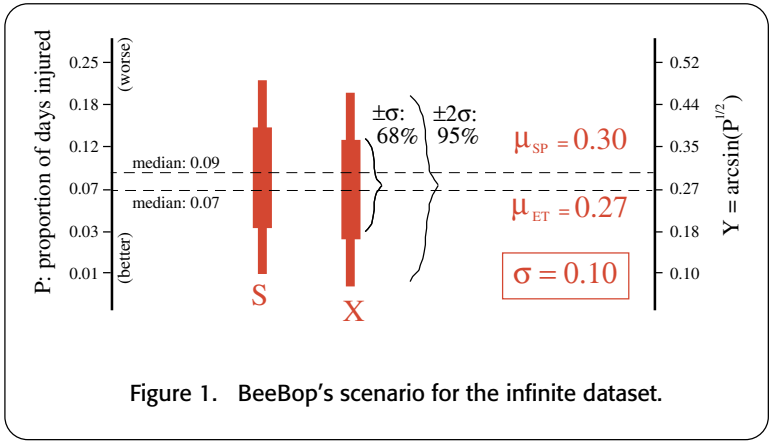


Figure 1. BeeBop's scenario for the infinite dataset.

Planning question

BeeBop plans to study about 200 runners, but could recruit as many as 270. What is the statistical power of the ordinary two-group t test when $\alpha = 0.05$?

Is directional hypothesis appropriate here? Yes, because BeeBob is only interested in whether the XDM-X reduces injuries. If this cannot be established, there is no reason to bring out new model.

H_0: greater or same injury rate ⇒ XDM-X not better
H_A: lower injury rate ⇒ XDM-X better

But we will also look at two-tailed results, just to see them.

UnifyPow statements

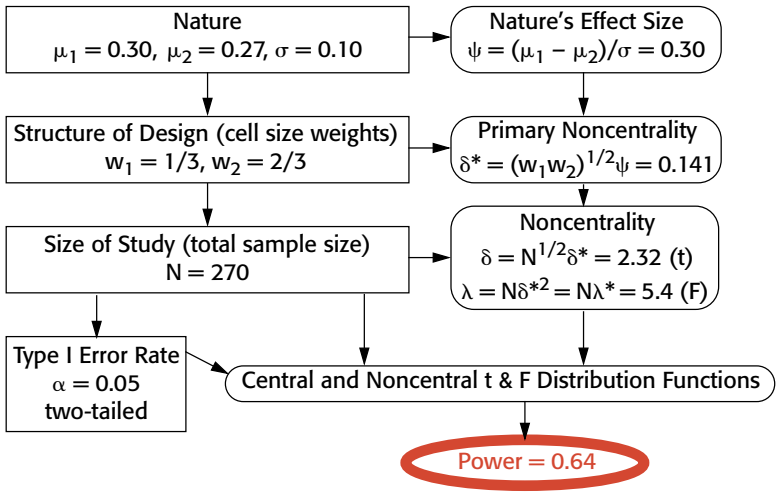
```
%include "&UnifyPow";
datalines4;

means .30 .27 . arcsin[sqrt(p)] for p = 9% vs. 7%
weight 1 2
sd .08 .10 .125
NTotal 201 270
alpha .05 . (default; statement not needed)
;;;;
%tables
```


Scenario: means.30 .27 . arcsin[sqrt(p)] for p = 9% vs. 7%
and Effect: Ordinary t test

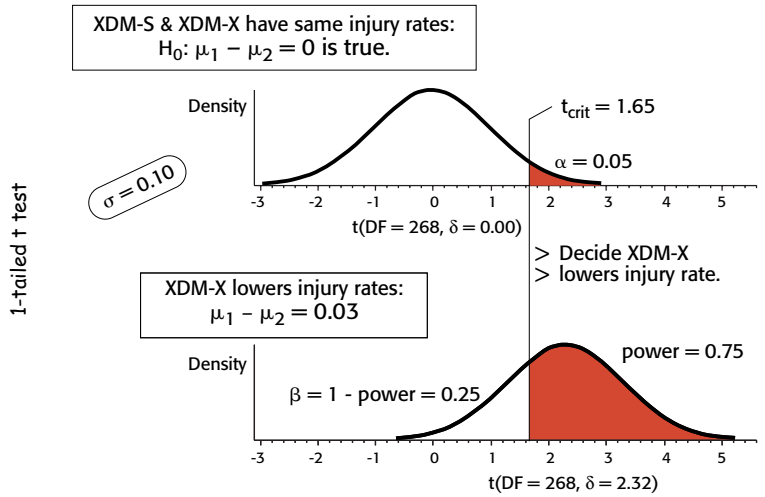
		Standard Deviation					
		0.08		0.1		0.125	
		Total N		Total N		Total N	
		201	270	201	270	201	270
		Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Alpha	Type						
0.05	2-tail t	.703	.825	.514	.639	.358	.457
	1-tail t	.803	.895	.638	.750	.482	.583

^^^^
||||

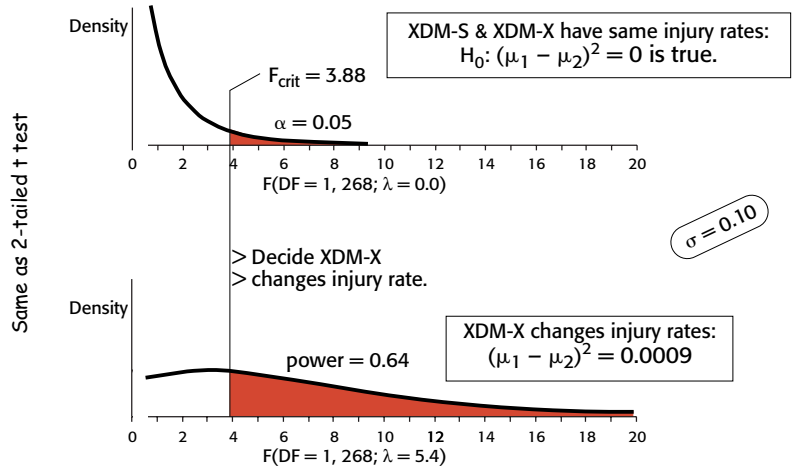
Components of power for two-group t & F = t^2 tests



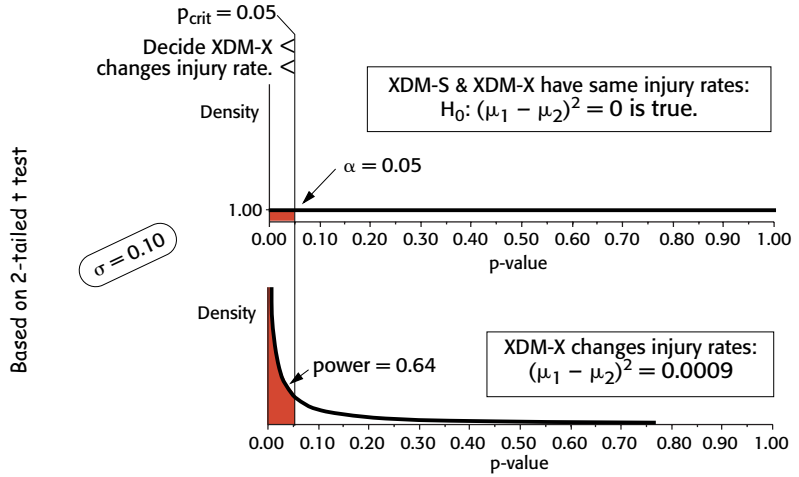
The "tale" of two t distributions (N = 270)



The "tale" of two F = t^2 distributions (N = 270)



The "tale" of two p-value distributions



The concept of composite power

Simple idea: Define a probability distribution over a set of conjectured values for the standard deviation (or any other parameter). Then just compute the expected value of the power over these values.

```
title3 "Arcsin Transform Approach";
datalines4;
means .30 .27 . arcsin[sqrt(p)] for p = 9% vs. 7%
weight 1 2
SD        .07 .08 .09 .10 .11 .12 .13
priors_SD  1   2   3   4   3   2   1
NTotal 201 270
alpha .05 . (default; statement not needed)
;;;;
%tables
```


		Total N	
		201	270
		Power	Power
Alpha	Type		
0.050	2-tail t	.532	.649
	1-tail t	.648	.752

Note: values here are composites formed by taking a weighted average over the power values obtained for the various SD specifications.

Using baseline covariates in randomized designs

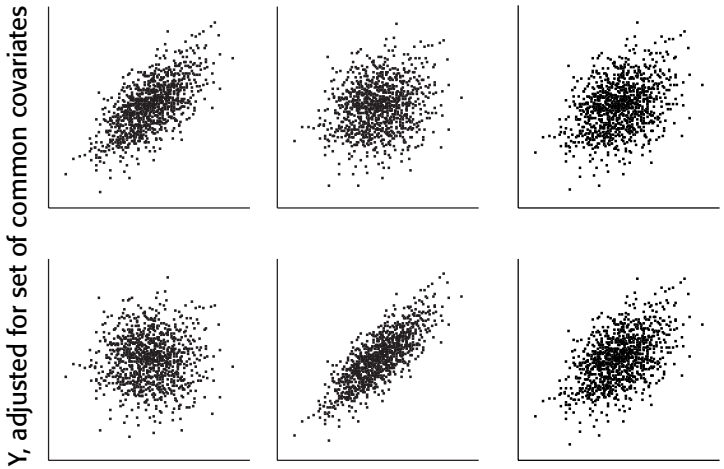
True random assignment assures that the "infinite data" distributions of baseline covariates are unrelated to treatment groups.

However, adding baseline covariates to the model can reduce the error variance and thus increase the power for testing the treatment effects.

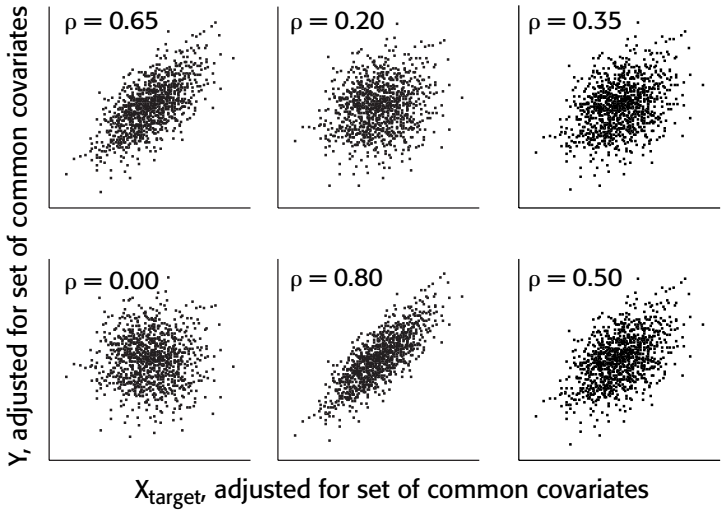
Consider BeeBop's XDM-S and XDM-X design again, but suppose we have data on each runner's prior injury rate.

We need to conjecture what the correlation is between prior injury rate (X, with arcsin transformation) and the future rate (Y), adjusting for shoe type.

Helpful tool for eliciting correlation conjectures



The answers



Suppose that BeeBop believes this *partial correlation* is at least $\rho = 0.50$ and could be as high as $\rho = 0.65$.

Conversion: adjusted variance $= (1 - \rho^2)\sigma^2$

adjusted SD $= (1 - \rho^2)^{1/2}\sigma$

SDReduction factor $= 1 - (1 - \rho^2)^{1/2}$

Thus using X = prior injury rate as covariate reduces the error SD by 13.4% if $\rho = 0.50$ or 24.0% if $\rho = 0.65$.

Conversion: SDreduction $= 1 - (1 - \rho^2)^{1/2}$


```
datalines4;
means .30 .27 . arcsin[sqrt(p)] for p = 9% vs 7%
weight 1 2
SD .10 . You can give more choices.
covariates 1 . (prior injury rate)
PartialCorr 0.00 0.50 0.65
//above same as: SDreduction 0.00 0.134 0.240
Ntotal 201 270
;;;;
%tables
```



```
Comparing 2 groups on location, adjusted for 1 covariate:
   <Parametric>    Ho: mu_adjusted{1} - mu_adjusted{2} = 0


Scenario: means .30 .27 . arcsin[sqrt(p)] for p = 9% vs 7%
-------------------------------------------------------------
|                            |     Standard Deviation       |
|                            |------------------------------|
|                            |             0.1              |
|                            |------------------------------|
|                            |Covariate Partial Correlation|
|                            |------------------------------|
|                            |    0     |   0.5    |  0.65  |
|                            |----------+----------+--------|
|                            | Total N  | Total N  | Total N |
|                            |----------+----------+--------|
|                            |201 |270 |201 |270 |201 |270 |
|                            |----+----+----+----+----+----|
|                            |Pow-|Pow-|Pow-|Pow-|Pow-|Pow-|
|                            | er | er | er | er | er | er |
|----------------------------+----+----+----+----+----+----|
|Effect   |Alpha   |Type     |    |    |    |    |    |    |
|---------+--------+---------|    |    |    |    |    |    |
|Ordinary|0.050    |2-tail t|.514|.639|.635|.762|.747|.861|
|t test  |         |--------+----+----+----+----+----+----|
|        |         |1-tail t|.638|.750|.746|.849|.838|.920|
-------------------------------------------------------------
```

Suppose that 3 covariates are used (gender, age, prior injury rate) and they are conjectured to reduce the SD by 13.4% or 24.0% or 30.0%.

```
datalines4;
means .30 .27 . arcsin[sqrt(p)] for p = 9% vs. 7%
weight 1 2
SD .10 . You can give more choices
covariates 3 . (gender, age, prior injury rate)
SDreduction 0.00 0.134 .240 .300
NTotal 201 270
;;;;
%tables
```


Comparing 2 groups on location, adjusted for 3 covariates:
<Parametric> Ho: mu_adjusted{1} - mu_adjusted{2} = 0

	Standard Deviation							
	0.1							
	Proportion of SD Reduced by Covariates							
	0		0.134		0.24		0.3	
	Total N		Total N		Total N		Total N	
	201	270	201	270	201	270	201	270
	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Effect	Alpha	Type							
Ordinary t test	0.050	2-tail t	.514	.639	.635	.762	.747	.861	.813	.911
		1-tail t	.638	.750	.746	.849	.837	.920	.887	.952

This is for 50 baseline covariates. What is the lesson here?

	Standard Deviation							
	0.1							
	Proportion of SD Reduced by Covariates							
	0		0.134		0.24		0.3	
	Total N		Total N		Total N		Total N	
	201	270	201	270	201	270	201	270
	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Effect	Alpha	Type							
Ordinary t test	0.050	2-tail t	.513	.638	.633	.762	.746	.861	.812	.911
		1-tail t	.637	.749	.745	.849	.837	.920	.886	.952

One-way ANOVA with complex contrasts

Great quote from Milliken & Johnson's *Analysis of Messy Data,* 1992:

> "All of the above treatment structures* can always be considered as a one-way structure for analysis purposes. In particular, when the treatment structure is a complex combination of two or more treatment structures, it is **usually best to consider the set of treatments as a one-way treatment structure."**

*Refers to the following designs (1)one-way, (2)two-way, (3) general factorial, (4) fractional factorial, (5) factorial with one or more "detached" control groups.

One-way ANOVA (nested design)

- High-mileage runners will be randomly assigned to run in either

  XDM-X eXperimental shoe
  (3 subtypes: XDM-X_1, XDM-X_2, XDM-X_3)

  XDM-S Standard XDM altered cosmetically to make it also look like experimental

- To get sufficient durability data on the XDM-X:

  6/9 of the runners will get the XDM-X.
  (2/9 each for XDM-X_1, XDM-X_2, XDM-X_3)

  3/9 of the runners will get the XDM-S.

- Runners will train at least 5 times/wk, 50 miles/wk for 26 wks.

Conjecture for means

XDM-S	XDM-X		
	Subtype 1	Subtype 2	Subtype 3
.300	.275	.270	.265

Planning question: What is power for test of XDM-S vs. average of all XDM-X versions? What is power for test comparing the 3 XDM-X subtypes?

Core UnifyPow statements

```
means .300 .275 .270 .265
// 2/3 of the runners will get some
// version of the experimental shoe.
weight 3  2  2  2
SD .08 .10 .125
NTotal 207 270
/#
Normally, I would be using:
  NoOverall  . Ignore 3 df overall test.
We do it here to make a teaching point about
getting more power from specific contrasts.
#/
contrasts
"XDM-S vs. XDM-X (all versions)"
   3 -1 -1 -1
"Variation among XDM-X subtypes"
   0  1  -1   0
>  0  0   1  -1
```


Scenario: means .300 .275 .270 .265

			Standard Deviation					
			0.08		0.1		0.125	
			Total N		Total N		Total N	
			207	270	207	270	207	270
			Power	Power	Power	Power	Power	Power
Test	Alpha	Type						
Overall test	0.050	Regular F	.569	.699	.382	.489	.253	.325
XDM-S vs. XDM-X (all versions)	0.050	2-tail t	.716	.825	.526	.639	.367	.457
		1-tail t	.813	.895	.649	.750	.491	.583
Variation among XDM-X subtypes	0.050	Regular F	.078	.087	.067	.073	.061	.065

Key concept: specificity breeds power (if you guess right)

Expected value:

$$E[F(df_H, df_E)] = (1 + N\lambda^*/df_H)[df_E/(df_E - 2)]$$
$$\approx 1 + N\lambda^*/df_H$$

Test:	[3 -1 -1 -1] contrast	overall test
.05 critical values:	$F(1, 60) = 4.00$	$F(3, 60) = 2.76$
Primary noncent. (λ^*):	$0.000200/\sigma^2$	$0.000211/\sigma^2$
Key term (λ^*/df_H):	$0.000200/\sigma^2$	$0.000070/\sigma^2$

Estimating power from data (F)

Ref: Chapter by O'Brien and Muller (1993). Available on UnifyPow website. Chapter also includes material on t.

Key concept (again):

$$E[F(df_H, df_E)] = (1 + N\lambda^*/df_H)[df_E/(df_E - 2)]$$

This gives rise to an unbiased estimator

$$\hat{\lambda}^* = [(df_E - 2)/df_E]df_H(F - 1)/N$$

If $\hat{\lambda}^* > 0$, then we could take $\lambda = N_{new}\,\hat{\lambda}^*$ to be the noncentality in order to estimate power for any N_{new}.

The **median estimator** is a better point estimate here. First, solve

$$\hat{\lambda}_{.50} = \Pr[F(df_H, df_E, \hat{\lambda}_{.50}) \geq F_{obs}] = .50$$

using, say, the SAS function FNONCT(F, dfH, dfE, .50). Then $\hat{\lambda}^*_{.50} = \hat{\lambda}_{.50}/N$, where N is the total sample size associated with the observed F statistic.

Then use $\lambda = N_{new}\,\hat{\lambda}^*_{.50}$ to get powers for N_{new} you want.

Some software uses $\hat{\lambda} = df_H F$ and computes "observed power" for the test at hand. Obvious bias. Discuss later.

Confidence limits for power (F)

Similarly, we can find the lower limit of a one-tailed $(1 - \gamma)$ CI for λ^* using

$$\hat{\lambda}_\gamma = \Pr[F(df_H, df_E, \hat{\lambda}_\gamma) \geq F] = \gamma.$$

This can be done with the SAS function

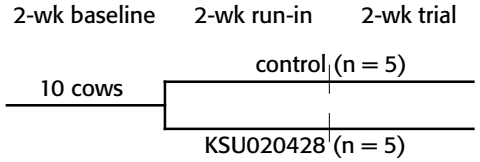
FNONCT(F, dfH, dfE, 1 - gamma).

Then $\hat{\lambda}^*_\gamma = \hat{\lambda}_\gamma/N$, where N is the sample size associated with the F statistic. Then use $\lambda = N_{new}\,\hat{\lambda}^*_\gamma$ to get limits for any N_{new} you desire.

Example of median estmate and lower CIs for power

Does the feed additive KSU020428 alter the fat content of daily milk?

Design:



Outcome: Y = percentage change from baseline.

Data analysis: F(1, 8) = 2.12, p = 0.18.

A crude SAS program for this problem

```
%macro LambdaCI (F, dfH, dfE, ConfLevel);
/*
Find the lower 100*ConfLevel% confidence limit for the
noncentrality (lambda), i.e. given the observed F(dfH,
dfE),
we are 100*ConfLevel% confident that lambda exceeds this
limit. One cannot be more that 100*(1-p)% confident that
lambda exceeds 0.0.
*/
  ConfLevel = &ConfLevel;
  p = SDF('F', F, dfH, dfE, 0);
  if p < (1 - ConfLevel/100) then
    LambdaLimit = fnonct(&F,&dfH, &dfE, &ConfLevel);
  else LambdaLimit = .;
  PrimeLambdaLimit = LambdaLimit/N;
  output;
  %mend LambdaCI;
```



```
data FindCIlimits;
/*
Problem
=======
Does the feed additive KSU020428 alter the fat content
of dairy milk? Study was run with 10 cows total. All
assessed over two-week baseline, then randomized 1:1 to
get KSU020428 or control. Outcome was percentage change
from baseline. Observed F = 2.12, p = 0.18.
*/
  keep F dfH dfE N p ConfLevel LambdaLimit
       PrimeLambdaLimit;
  F = 2.12;
  dfH = 1;
  dfE = 8;
  N = 10;
  %LambdaCI (F, dfH, dfE, .50);
  %LambdaCI (F, dfH, dfE, .70);
  %LambdaCI (F, dfH, dfE, .80);
  %LambdaCI (F, dfH, dfE, .81);
  %LambdaCI (F, dfH, dfE, .82);
run;
proc print data=FindCIlimits; run;
```


Proc Print of results:

F	dfH	dfE	N	Conf level	p	Lambda Limit	Prime Lambda Limit
2.12	1	8	10	0.500	0.18348	1.95717	0.19572
2.12	1	8	10	0.667	0.18348	0.82736	0.08274
2.12	1	8	10	0.800	0.18348	0.08517	0.00852
2.12	1	8	10	0.810	0.18348	0.03345	0.00335
2.12	1	8	10	0.820	0.18348	.	.

I need to write a module for UnifyPow that will handle this problem, but for now it can still be "tricked" to get the powers.

Getting UnifyPow to crunch and table

```
datalines4;
/#
Does the feed additive KSU020428 alter the fat content
of dairy milk? Data F(1, 8) = 2.12 (p = 0.18) based on
5 + 5 cows.

Compute lower confidence bounds for N = 30, 60, 100.
#/
Exemplary SSH
NumParms 2
Nexemplary 10
SD 1 . SD is already incorporated into lambdas
alpha .05
Ntotal 30 50 100
tails 2
effects
"median estimate" 1 1.95717
"lower 67% bound" 1 0.82736
"lower 80% bound" 1 0.08517
;;;;
%tables
```


			Standard Deviation		
			1		
			Total N		
			30	50	100
			Power	Power	Power
Test	Alpha	Type			
median estimate	0.050	2-tail t	.648	.865	.992
lower 67% bound	0.050	2-tail t	.331	.513	.813
lower 80% bound	0.050	2-tail t	.078	.098	.150

What might you write about this?

"Given the results we obtained in our pilot study on 5+5 cows, we estimate that an experiment on 50+50 cows has a 50% chance of having at least 0.99 power (at α = 0.05, two-tailed) and a 67% chance of having at least 0.81 power."

ANOVA on logNormal Y (really cool)

P_i = Proportion of days i-th Runner is injured, including days when he/she runs with the injury.

To be handled as continuous variable with a logNormal distribution: log(P) ~ Normal.

We take the coefficient of variation of P, CV = σ/μ, to be homogeneous across the 4 groups. This makes the SD of log(P) homogeneous across the groups.

logNormal outcomes fit many situations!

This greatly simplifies the "Elicitation Process"

What is the infinite dataset? (the population distributions)

Scenario to examine—
Start with **base mean of P for XDM-S: "100%"**

BeeBop **conjectures:**

XDM-S	XDM-X		
	Subtype 1	Subtype 2	Subtype 3
100%	80%	78%	76%

So the XDM-X_1 is conjectured to average 80% of the injuries as the XDM-S shoe.

Conjectured spread: The SD in each group is 30% of the mean:

CV = 0.30

We will also consider CV = 0.35 and 0.40.

Planning questions

BeeBop plans to study about 200 runners, but could recruit as many as 270.

What is power for the test of XDM-S vs. average of all XDM-X versions?

What is power for test comparing the 3 XDM-X subtypes?

Core UnifyPow statements

```
means:logNormal 1.00 0.84 0.82 0.80
weight 3 2 2 2
CV 0.30 0.35 0.40 . SD{j}/mu{j} homogeneous
// priorsCV 4 2 1
NTotal 207 270
NoOverall . Ignore 3 df overall test.
contrasts
"XDM-S vs. XDM-X (all versions)"
  3 -1 -1 -1
"Variation among XDM-X subtypes"
  0 1 -1  0
> 0 0  1 -1
;;;;
%tables
```


			Coef of Variation					
			0.3		0.35		0.4	
			Total N		Total N		Total N	
			207	270	207	270	207	270
			Power	Power	Power	Power	Power	Power
Test	Alpha	Type						
Compare all means	0.050	Regular F	.982	.997	.932	.980	.852	.937
XDM-S vs. XDM-X (all versions)	0.050	2-sided t	.995	.999	.976	.995	.936	.978
		1-tail t	.998	.999	.989	.998	.967	.990
Variation among XDM-X subtypes	0.050	Regular F	.100	.117	.087	.099	.078	.088

Composite power for logNormal

```
datalines4;
means:logNormal 1.00 0.84 0.82 0.80
weight 3 2 2 2
CV          0.30 0.35 0.40
priors_CV     4    2    1
NTotal 207 270
NoOverall . Ignore 3 df overall test.
contrasts
"XDM-S vs. XDM-X (all versions)"
3 -1 -1 -1
"Variation among XDM-X subtypes"
0 1 -1 0
> 0 0 1 -1
;;;;
%tables
```



```
Scenario: means:logNormal 1.00 0.84
0.82 0.80
-------------------------------------
|                           | Total N |
|                           |---------|
|                           |207 |270 |
|                           |----+----|
|                           |Pow-|Pow-|
|                           | er | er |
|---------------------------+----+----|
|Test         |Type         |    |    |
|-------------+-------------|    |    |
|XDM-S vs.    |2-sided t    |.981|.995|
|XDM-X (all   |-------------+----+----|
|versions)    |1-tail t     |.991|.997|
|-------------+-------------+----+----|
|Variation    |Regular F    |    |    |
|among XDM-X  |             |    |    |
|subtypes     |             |.093|.108|
-------------------------------------
```


Mean difference of paired observations

Example motivated from Manocha, et. al. (1986) as summarized by Altman (Practical Statistics for Medical Research, 1991, p. 189).

Subjects: One group of healthy women, age 22-30.

Measures: Average daily energy intake (in Kilojoules: kJ), pre-menstrual (Y_1) and post-menstrual (Y_2), 10 days each.

Hypotheses: Let $D = Y_1 - Y_2$. Interested in

$H_0: \mu_D = 0$ vs. $H_A: \mu_D \neq 0$

Need to make conjectures on:

$[\mu_1, \mu_2]$, $[\sigma_1, \sigma_2]$, and $\rho = \text{corr}(Y_1, Y_2)$

Suppose

$$[\mu_1, \mu_2] = [6500, 5600],$$

so that

$$\mu_D = 900.$$

Believes that σ will be greater during pre-menstrual days. Suppose

$$[\sigma_1, \sigma_2] = [1500, 1300].$$

The correlation between Y_1 and Y_2 is suspected to be at least

$$\rho = 0.85$$

It is convenient to display these values in an “SD-Corr” matrix,

$$\tilde{\Sigma} = \begin{bmatrix} 1500 & 0.85 \\ 0.85 & 1300 \end{bmatrix}$$

$$\sigma_D = [\sigma_1{}^2 + \sigma_2{}^2 - 2\rho\sigma_1\sigma_2]^{1/2}.$$

$$\sigma_D = [1500^2 + 1300^2 - 2(1500)(1300)(0.85)]^{1/2} = 791.$$

Use one-group t test to assess

$$H_0\text{: } \mu_D = 0$$

assuming D is Normal with

$$\mu_D = 900 \text{ and } \sigma_D = 791.$$

σ_D can be varied according to general equation,

$$\sigma_D(m, \rho) = m[\sigma_1{}^2 + \sigma_2{}^2 \; - 2\rho\sigma_1\sigma_2]^{1/2}$$

m: “SD multiplier” (SDmult)

What is the power ...

- for N = 10 or 15 subjects?
- if the correlation is greater, say $\rho = 0.90$? (Pilot was 0.95.)
- if the standard deviations are, say, 20% larger ($m = 1.20$)?

UnifyPow makes it easy!

Match-pairs t test

```
/*/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/

 Pre-Menstrual vs. Post-Menstrual Dietary Intake
 Mean difference of paired observations via
 the traditional t test

/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/
*/
%include "&UnifyPow";
title2 "Pre-Menstrual vs. Post-Menstrual Dietary Intake";
title3 "Traditional t-Based Tests";
datalines4;

PairedMeans 6500 5600  . reported pre and post kJ/day
SD 1500 1300 . "base" pre and post SD
corr .85 .90 . correlation of pre vs. post (solve for each)
SDMultiplier 1.0 1.2 . solve for each, default = 1.0
alpha .05 .01
TotalPairs 10 15 . Any word with 'total' is OK
;;;;
%tables
```


Testing difference of single pair of correlated measures:
<Parametric> Ho: mu(Y1 - Y2) = 0

		x SD (SD Multiplier)							
		1				1.2			
		Corr(Y1, Y2)				Corr(Y1, Y2)			
		0.85		0.9		0.85		0.9	
		Total Pairs		Total Pairs		Total Pairs		Total Pairs	
		10	15	10	15	10	15	10	15
Alpha	**Type**	Power	Power	Power	Power	Power	Power	Power	Power
0.05	2-sided t	.892	.983	.970	.998	.761	.927	.895	.984
	1-sided t	.952	.994	.990	.999	.868	.967	.954	.995
0.01	2-sided t	.640	.902	.827	.981	.453	.745	.645	.905
	1-sided t	.764	.949	.908	.993	.587	.837	.768	.951

Testing a single Pearson correlation

Dr. Jean Netticks studies ARX enzyme levels in father and son. It has long been assumed that $0.00 < \rho[\text{ARX(dad)}, \text{ARX(son)}] < 0.55$.

Netticks' new theory predicts a correlation of $\rho > .70$.

Test

H_0: $\rho \leq 0.55$
H_a: $\rho > 0.55$.

```
datalines4;
correlation .70
null .55
Ntotal 50 75 100
tails 1
```



```
Testing single correlation coefficient using
Fisher's r-to-Z:

   <Parametric>    Ho: Z(rho) =  Z(0.55)

Scenario: rho .70
-------------------------------------------
|                           |    ALPHA    |
|                           |-------------|
|                           |     0.05    |
|                           |-------------|
|                           |   Total N   |
|                           |-------------|
|                           | 50 | 75 |100|
|                           |----+----+---|
|                           |Pow-|Pow-|Pow-|
|                           | er | er | er |
|---------------------------+----+----+---|
|Method       |Type         |    |    |   |
|-------------+-------------|    |    |   |
|Fisher's     |1-tail Z     |    |    |   |
|r-to-Z test  |             |    |    |   |
|of one rho   |             |.525|.680|.790|
-------------------------------------------
```


Comparing two Pearson correlations

Dr. Jean Netticks again. Question—

> Is there a difference on ρ[ARX(dad), ARX(son)] between obese and normal fathers?";

Scenario—
Obese: $\rho = 0.55$ Normal: $\rho = 0.70$

```
datalines4;
correlation .55 .70
weight 1 3
Ntotal 400 to 1600 by 400
tails 2
;;;;
%tables
```



```
Testing correlations using Fisher's r-to-Z:
   <Parametric>    Ho: Z(rho1) - Z(rho2) = 0

Scenario: rho .55 .70
-------------------------------------------------
|                           |       ALPHA       |
|                           |-------------------|
|                           |       0.05        |
|                           |-------------------|
|                           |     Total N       |
|                           |-------------------|
|                           |400 |800 |1200|1600|
|                           |----+----+----+----|
|                           |Pow-|Pow-|Pow-|Pow-|
|                           | er | er | er | er |
|---------------------------+----+----+----+----|
|Method       |Type         |    |    |    |    |
|-------------+-------------|    |    |    |    |
|Comparing    |2-tail Z     |    |    |    |    |
|two          |             |    |    |    |    |
|correlations |             |    |    |    |    |
|(r-to-Z)     |             |.567|.858|.961|.990|
-------------------------------------------------
```

Multiple regression via partial correlation

Research Question: Are higher levels of plasma homocysteine "independently" associated with atheroclerosis (build-up of plaque in coronary arteries)?

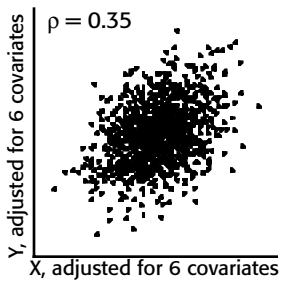
Design: Dr. Ann Ginna will use ordinary regression (after appropriate transformations) to assess the relationship between total homocysteine level (**tHcy**) and a plaque burden index (**PBI**), adjusting for 6 other variables: (1) age; (2) gender; plasma levels of (3) folate, (4) B_6, and (5) B_{12}; and (6) a serum cholesterol index. She states that PartialCorr(tHcy, PBI) must exceed 0.15 to prove clinically meaningful. This gives the one-tailed hypothesis

H_0: PartialCorr(tHcy, PBI) $\leq$ 0.15
H_A: PartialCorr(tHcy, PBI) $>$ 0.15

Dr. Ginna states that relationships of

true PartialCorr(tHcy, PBI) = 0.35

are typical in these kinds of studies.




```
%include "&UnifyPow";
datalines4;
PartialCorr .35
null .15
// NumParms = int + 6
covariates
//               + 1 target risk factor
NumParms 8 .
alpha .005 .025  . special alphas (long story)
Ntotal 180 360
sides 1
;;;;
%tables
```



```
Scenario: PartialCorr .35
-------------------------------------------------
|                           |       ALPHA       |
|                           |-------------------|
|                           |  0.005  |  0.025  |
|                           |---------+---------|
|                           | Total N | Total N |
|                           |---------+---------|
|                           |180 |360 |180 |360 |
|                           |----+----+----+----|
|                           |Pow-|Pow-|Pow-|Pow-|
|                           | er | er | er | er |
|---------------------------+----+----+----+----|
|Method      |Statistic     |    |    |    |    |
|------------+--------------|    |    |    |    |
|Fisher's    |1-tail Z      |    |    |    |    |
|r-to-Z test |              |    |    |    |    |
|of          |              |    |    |    |    |
|PartialRho  |              |.590|.925|.800|.980|
-------------------------------------------------
```


Comparing nested (full vs. reduced) OLS regressions

Mr. Corey Latour is proposing to do OLS modeling to predict job satisfaction. Will adding a 4-level nominal predictor (3 dummy variables) be significant, if that addition "should" increase R^2 from 0.45 to 0.50?

```
R**2 .45 .50
NumParms 5 8
Ntotal 100 125 150
alpha .05 .01
;;;;
%tables
```

Testing Ho: Beta_6 = Beta_7 = Beta_8 = 0.

Scenario: R**2 .45 .50

		ALPHA				
		0.05		0.01		
		Total N		Total N		
		100	125	150	100	125	150
		Power	Power	Power	Power	Power	Power
Method	Type						
Comparing nested R**2 values	Regular F	.741	.843	.909	.506	.648	.762

Testing single β in OLS regression model

Another study by Corey Latour.

Predicting

Y: annual salary (example: 32.4 = $32,400 US).

from

X_3: TeamValu, a 0-4 visual analog (continuous) scale measuring the employee's value to his/her teammates as determined by an external evaluator.

Model:

$$\text{Salary} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(\text{TeamValu}) + \epsilon$$

Focusing on

β_3: TeamValu's (unstandardized) β coefficient. Only β under examination here. β_3 corresponds to the differential in annual salary (in $1000) associated with a 1-point difference in TeamValu, given 2 other unspecified predictors, X_1 and X_2.

Scenarios: $\beta_3 = 0.50$ and 1.00. $\beta_3 = 0.50$ implies that a 1-point difference in TeamValu is associated with a $500 difference in annual salary, given X_1 & X_2.

Tol(X_3): The R^2 from predicting TeamValu from X_1 and X_2 is conjectured to be 0.30-0.40, e.g. TeamValu's tolerance is $1 - R^2 = 0.60$-0.70.

Model:

$$\text{Salary} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(\text{TeamValu}) + \epsilon$$

```
1betaOLS 0.50 1.00 . 0.50 ==> $500
SDX .6 .8  . Conjectures for SD(TeamValu)
Tolerance .6 .7 . Conjectures for Tol(TeamValu)
SD 3.0 3.5 . Conjectures for SD(error)
NumParms 4 .  (= 3 Xs + 1 for intercept)
power .80 .90
tails 1
;;;;
```

```
/*
The default table was OK, but I still decided to
modify it a bit.
*/

proc tabulate format=best4. order=data;
 class
  alpha BetaWt tolernce SDx testtype SD NomPower;
 var Ntotal;
 table
  alpha='Alpha:',BetaWt='Beta for TeamValu'*
  tolernce='Tol(X)' *SDx='SD(X)' *
testtype='Type',
  SD='SD(Resid)' * NomPower='Power'*mean=' ' *
Ntotal='Total N'
/rtspace=38;
```


Alpha: 0.05

				SD(Resid)			
				3		3.5	
				Power		Power	
				.800	.900	.800	.900
				Total N	Total N	Total N	Total N
Beta for TeamValu	Tol(X)	SD(X)	Type				
0.5	0.6	0.6	1-tail t	1033	1432	1405	1948
		0.8	1-tail t	582	806	791	1097
	0.7	0.6	1-tail t	886	1227	1205	1668
		0.8	1-tail t	499	691	678	940
1	0.6	0.6	1-tail t	260	359	353	488
		0.8	1-tail t	147	203	199	276
	0.7	0.6	1-tail t	223	308	303	418
		0.8	1-tail t	126	174	171	236

Comparing 2 independent proportions

Placebo vs. DCA for Lactic Acidosis in Children with Malaria. Treated only with quinine, 28% die. What if one dose of DCA cuts mortality by 1/4, to 21%?

- Actual study of children with severe malaria (3% of cases).
- 28% base mortality rate from meta-analysis of published surveys.
- 1/4 reduction in mortality as per animal-model study.
- Small pilot study had 2/10 deaths in each group.
- 2:1 randomization favoring DCA pleases local health officials.
- One interim look using $\alpha = .01$ & final analysis at .045.
  (UnifyPow does not rigorously handle power for sequential tests.)
- Able to randomize 1500. Should interim look come at 750 or 999?

Design and Scenario

	Lived	Died	
Placebo	72% of n_1	28% of n_1	$n_1 = N/3$
DCA	79% of n_2	21% of n_2	$n_2 = 2N/3$

Digression: Finding O'Brien-Fleming alpha levels

First, from www.biostat.wisc.edu/landemets, downloaded and used the Fortran program "ldbnds" (v. 2, 12/1/99) to get the following output;

look	time	lower	upper	exit prob	cum exit pr
1	0.67	-2.5019	2.5019	0.01235	0.01235
2	1.00	-1.9935	1.9935	0.03765	0.05000

Then, used the following tiny SAS program to find the alpha level associated with for 2.5019 and 1.9935:

```
data _null_;
alpha1 = 2*ProbNorm(-2.5019);
alpha2 = 2*ProbNorm(-1.9935);
put alpha1= alpha2=;
run;
```

This gives: alpha1=0.0123528811 alpha2=0.0462067235

Simplest way for two proportions (not best way)

```
%include "&UnifyPow";
datalines4;
proportions .28 .21
weight 1 2 . 2/3 of patients get DCA
alpha 0.01235 0.04621
Ntotal 999 1500
;;;;
%tables
```


		alpha			
		0.012		0.046	
		Total N		Total N	
		999	1500	999	1500
		Power	Power	Power	Power
Method	Statistic				
Approx Uncondit'l "chi^2"*	2-sided t	.487	.700	.684	.850
	1-sided t	.589	.783	.785	.911
Exact Uncondit'l**	2-sided t	.456	.666	.655	.826
	1-sided t	.558	.754	.761	.894
Fisher's exact conditional	2 sided	.441	.660	.641	.822
	1 sided	.543	.748	.749	.891
Likhd Ratio for Log Odds Ratio	2 sided	.476	.688	.673	.841
	1 sided	.578	.772	.776	.905

Note: We will redo LR method later to introduce how the "exemplary data" method can handle complex linear models.

Footnote to table

*The Approximate Unconditional "chi**2" test corresponds to the common (Pearson) chi-square test for a 2 x 2 table. However, the method employed here uses a regular t test with Y = 0 (no) or 1 (yes), which has been shown to give more accurate p-levels than the common test and can be computed with any t test routine. See D'Agostino, et. al. (1988), Am. Statistician, 42:198-202.

**The Exact Unconditional corresponds to the test proposed by Suissa and Shuster (1985), J Royal Stat Soc A, 148:317-327).

Two proportions (trimming output)

```
proportions .28 .21
weight 1 2 . 2/3 of patients get DCA
alpha 0.01235 0.04621
Ntotal 999 1500
method chi**2
// Can use statements such as
//   method all  (default)
//   method LR
//   method ExactUnconditional
//   method Fishers
// Or any combination, such as
//   method chi**2 ExactUnconditional
;;;;
/* Making the table clearer... */
data PowData;  set PowData;
if alpha = 0.04621 and Ntotal = 999 then delete;
if alpha = 0.01235 and Ntotal = 1500 then delete;
run;
%tables
```


Scenario: proportions .28 .21

	alpha	
	0.0-12	0.0-46
	Total N	Total N
	999	1500
	Power	Power
Method / Statistic		
Approx Uncondit'l "chi^2"* / 2-sided t	.487	.850
Approx Uncondit'l "chi^2"* / 1-sided t	.589	.911

Two proportions: expanded

2PROPORTIONS

- proportion1 vs. proportion2 specified directly
- proportion 1 vs. proportion2 = f(relative risk)
- proportion 1 vs. proportion2 = f(odds ratio)

Examples below are constructed so that the first combination in each defines the same problem, which is

<1> proportion1 = 0.28 vs. proportion2 = 0.21

<2> proportion1 = 0.28, with a RelativeRisk = 0.75

<3> proportion1 = 0.28, with an OddsRatio = 0.6835

Two Proportions specifying several π_1 and several π_2

```
2proportions
proportion_1 .28 .24
proportion_2 .21 .18
weight 1 2 . 2/3 of patients get DCA
alpha 0.01235 0.04621
Ntotal 999 1500
//methods all chi**2 Fishers ExactUnconditional LR
methods chi**2
;;;;
%tables
data PowData;  set PowData;
  if alpha = 0.04621 and Ntotal = 999 then delete;
  if alpha = 0.01235 and Ntotal = 1500 then delete;
%tables
```

Scenario: 2proportions
and Method: Approx Uncondit'l "chi^2"*

		Group 1 Probability							
		0.28				0.24			
		Group 2 Prob				Group 2 Prob			
		0.21		0.18		0.21		0.18	
		Total N		Total N		Total N		Total N	
		999	1500	999	1500	999	1500	999	1500
		Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Alpha	Test Type								
0.012	2-sided t	.487	.	.876	.	.077	.	.396	.
	1-sided t	.589	.	.921	.	.121	.	.497	.
0.046	2-sided t	.	.850	.	.994	.	.251	.	.773
	1-sided t	.	.911	.	.997	.	.359	.	.856

Two proportions using relative risks

```
2proportions
proportion_1 .28 .24
RelativeRisk .75 .80
weight 1 2 . 2/3 of patients get DCA
alpha 0.01235 0.04621
Ntotal 999 1500
;;;;
data PowData;  set PowData;
  if alpha = 0.04621 and Ntotal = 999 then delete;
  if alpha = 0.01235 and Ntotal = 1500 then delete;
%tables
```


Scenario: 2proportions
and Method: Likhd Ratio for Log Odds Ratio

		Group 1 Probability							
		0.28				0.24			
		Relative Risk				Relative Risk			
		0.75		0.8		0.75		0.8	
		Total N		Total N		Total N		Total N	
		999	1500	999	1500	999	1500	999	1500
		Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Alpha	Test Type								
0.012	2 sided	.476	.	.284	.	.386	.	.225	.
	1 sided	.578	.	.377	.	.486	.	.309	.
0.046	2 sided	.	.841	.	.646	.	.763	.	.559
	1 sided	.	.905	.	.753	.	.848	.	.677

Two proportions using odds ratios

```
2proportions
proportion_1 .28 .24
OddsRatio 0.6835 .75
weight 1 2 . 2/3 of patients get DCA
alpha 0.01235 0.04621
Ntotal 999 1500
;;;;
data PowData;  set PowData;
  if alpha = 0.04621 and Ntotal = 999 then delete;
  if alpha = 0.01235 and Ntotal = 1500 then delete;
run;
%tables
```

	Group 1 Probability							
	0.28		0.24					
	Odds Ratio		Odds Ratio					
	0.6835	0.75	0.6835	0.75				
	Total N	Total N	Total N	Total N				
	999	1500	999	1500	999	1500	999	1500
	Power	Power	Power	Power	Power	Power	Power	Power
Alpha	Test Type							
0.012	2 sided	.476	.	.263	.	.423	.	.231	.
	1 sided	.578	.	.352	.	.525	.	.316	.
0.046	2 sided	.	.841	.	.615	.	.798	.	.568
	1 sided	.	.905	.	.727	.	.874	.	.685

Testing goodness of fit to a multinomial distribution

Do the Biostat Boys use fair dice?

A trio of shady characters called the Biostat Boys run a craps game at lunch near the FDA. One of the regular patrons, Lucky Luke, has been losing lately and wonders if the dice are fair. The known distribution for a fair pair of dice is:

X:	2	3	4	5	6	7	8	9	10	11	12
p(X):	.028	.056	.083	.111	.139	.166	.139	.111	.083	.056	.028

Luke believes that one dice is weighted so that the "1" comes up 1/4 of the time, not 1/6. So he computes that the distribution may be:

X:	2	3	4	5	6	7	8	9	10	11	12
p(X):	.042	.067	.092	.116	.142	.166	.125	.100	.075	.050	.025

How many tosses must he observe in order to have 90% power ($\alpha = .05$) to find such a discrepancy, based on the full distribution?

```
//            <2>  <3>  <4>  <5>  <6>  <7>  <8>  <9> <10> <11> <12>
GoodnessOfFit .042 .067 .092 .116 .142 .166 .125 .100 .075 .050 .025
Null          .028 .056 .083 .111 .139 .166 .139 .111 .083 .056 .028
scenario “One dice is loaded with Pr[‘1’] = 1/4, not 1/6”
alpha .05 .10
power .80 .90
;;;;
%tables
```

Same thing, only larger font:

```
//             <2>  <3>  <4>  <5>  <6>
GoodnessOfFit .042 .067 .092 .116 .142 ...
Null          .028 .056 .083 .111 .139 ...
Scenario "One dice is loaded with Pr['1'] = 1/4 "
alpha .05 .10
power .80 .90
```


Scenario: One dice is loaded with Pr[‘1’] = 1/4, not 1/6

		ALPHA			
		0.05		0.1	
		Minimum Power		Minimum Power	
		.800	.900	.800	.900
		Total N	Total N	Total N	Total N
Method	Statistic				
Ordinary Pearson	Chi-square	1109	1402	913	1187
Likhd Ratio	Chi-square	1186	1500	976	1270

Lesson in design: Construct tight hypotheses!

What if Lucky Luke only counts how many "1"s appear on a throw? The distributions are

	X:	0	1	2
fair	p(X):	.694	.278	.028
biased	p(X):	.625	.333	.042

```
scenario “One dice is loaded ...  1/4, not 1/6”
//              <0>  <1>  <2>
GoodnessOfFit .625 .333 .042
Null          .694 .278 .028
alpha .05 .10
power .80 .90
;;;;
%tables
```


Scenario: One dice is loaded with Pr[‘1’] = 1/4, not 1/6

		ALPHA			
		0.05		0.1	
		Minimum Power		Minimum Power	
		.800	.900	.800	.900
		Total N	Total N	Total N	Total N
Method	Statistic				
Ordinary Pearson	Chi-square	390	512	312	424
Likhd Ratio	Chi-square	413	542	330	448

Lesson: It pays to test tighter hypotheses...if you guess right!

Testing association in an R x C contingency table

Variation in sarcoma type by region.

Stimulated by example found in Section 3.4 of Freeman DH (1987). ***Applied Categorical Data Analysis,*** New York: Marcel Dekker.

There are 3 different types of soft-tissue sarcomas of the arms and legs:

- Fibroid
- Lipoid
- Mixed (or other)

Incidence data can be obtained from good cancer registries.

The Question: Do the relative proportions of these 3 types differ among 4 geographic regions?

Region	Sarcoma Soft Tissue Type			% of Total
	Fibroid	Lipoid	Mixed	
A	.50	.20	.30	30%
B	.60	.25	.15	20%
C	.35	.35	.30	25%
D	.45	.20	.35	25%

```
2WayContTable .50 .20 .30
 >            .60 .25 .15
 >            .35 .35 .30
 >            .45 .20 .35
scenario “Numbers made up to generate example!”
weight .30 .20 .25 .25
alpha .01 .05
power .90 .95
;;;;
%tables
```

Scenario: Numbers made up to generate example!

		ALPHA			
		0.01		0.05	
		Minimum Power		Minimum Power	
		.900	.950	.900	.950
		Total N	Total N	Total N	Total N
Method	Statistic				
Ordinary Pearson	Chi-square	500	580	380	440
Likhd Ratio	Chi-square	480	560	360	420

Wilcoxon-Mann-Whitney (W-M-W) test

Recall...

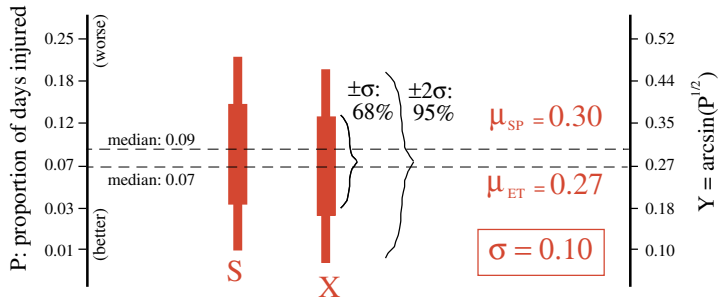


Figure 1. BeeBop's scenario for the infinite dataset.

Core UnifyPow statements

```
%include "&UnifyPow";

title1 "Beebop Shoes: XDM-S vs. XDM-X";
title2 "Median injury rates: 9% XDM-S vs. 7% XDM-X";
title3 "Wilcoxon-Mann-Whitney";

datalines4;
scenario "arcsin[sqrt(p)] for p = 9% vs. 7%"
means .30 .27
weight 1 2
sd .08 .10 .125
NTotal 201 270
alpha .05
tails 2
Wilcoxon
method Lehmann ARE . also "Noether" "all"
;;;;

%tables
```


Scenario: arcsin[sqrt(p)] for p = 9% vs. 7%
and Alpha: 0.050

			Standard Deviation					
			0.08		0.1		0.125	
			Total N		Total N		Total N	
			201	270	201	270	201	270
			Power	Power	Power	Power	Power	Power
Method	Type	Parent						
Wilcoxon Mann-Whitney [Lehmann (p1, p2, p3) aprx]	2-tail W	Normal	.680	.806	.491	.616	.341	.437
		Logistic	.738	.856	.546	.675	.382	.488
		Laplace	.847	.935	.669	.796	.486	.610

Lehmann's approximation proved best most of the time in some simulations I did for fun many years ago. The ARE (asymptotic relative efficiency) and Noether methods are optional. Specify 'all' to get all.

	Standard Deviation					
	0.08		0.1		0.125	
	Total N		Total N		Total N	
	201	270	201	270	201	270
	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Method / Type / Parent						
Wilcoxon Mann-Whitney [aprx via ARE W vs. t] / 2-tail W / Normal	.683	.807	.496	.619	.345	.440
Logistic	.743	.858	.552	.679	.387	.492
Laplace	.864	.944	.687	.810	.499	.622
min ARE	.639	.767	.458	.576	.317	.406
Ordinary t test / 2-tail t / Normal	.703	.825	.514	.639	.358	.457

The "min ARE" results give lower limits for power, assuming that the parent distributions are continuous (not necessarily symmetric) and have bounded densities that differ only in their means.

W-M-W using p_1 short-cut specification

In BeeBop's S vs. X running shoe two-group design,

p_1 = Pr[random Y from S group > random Y from X group]

Greatly simplfies W-M-W elicitation.

Here, assume parent distribution is logistic, i.e. a little heavier tailed than Normal. The Laplace parent is heaviest tailed used here. "SD" is now relative concept (default: 1.0).

```
2Wilcoxon .60              ← specify p1 = 0.60
relativeSD 1.00 1.15 1.25  ← if SD increases by 15% or 25%
weight 1 2
NTotal 201 270
parent logistic  ← The p1 = 0.60 is relative to the logistic parent.
;;;;  /* Other parents are "normal" "Laplace" */
%tables
```



```
Parent Distributions
--------------------
Powers for the Wilcoxon will be approximated assuming Normal,
Logistic, and Laplace parent distributions, thus giving a range of
tail thicknesses (kurtoses) and asymptotic relative efficiencies (ARE):

   Parent      Kurtosis      ARE
  --------     --------     -----
  Normal          0.0       0.955
  Logistic        1.2       1.097
  Laplace         3.0       1.500
  <<Lower limit of ARE>>    0.864

You specified the LOGISTIC distribution and p1 = 0.600.
This equates to psi = (mu1 - mu2 - NullValue)/SD = 0.3349.
```

	Nonparametric Moments								
	p1			p2			p3		
	Relative Std Dev			Relative Std Dev			Relative Std Dev		
	1	1.15	1.25	1	1.15	1.25	1	1.15	1.25
Parent / Normal	.594	.582	.575	.432	.419	.412	.432	.419	.412
Logistic	.600	.587	.580	.439	.425	.418	.439	.425	.418


```
Scenario: 2Wilcoxon .60      . specify p1 = 0.60
AND Alpha: 0.05
```

			Relative Std Dev					
			1		1.15		1.25	
			Total N		Total N		Total N	
			201	270	201	270	201	270
			Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Method	Type	Parent						
Wilcoxon Mann-Whitney [Lehmann aprx]	2-tail W	Normal	.581	.712	.469	.590	.409	.521
		Logistic	.639	.769	.522	.649	.457	.577
		Laplace	.762	.874	.644	.773	.574	.704
	1-tail W	Normal	.701	.811	.595	.708	.536	.645
		Logistic	.752	.855	.646	.759	.584	.697
		Laplace	.850	.929	.755	.858	.694	.804

continued ...

Ordinary t test	2-tail t	Normal	.605	.734	.491	.613	.429	.543
	1-tail t	Normal	.721	.827	.616	.727	.555	.665

Difference of paired observations: Wilcoxon signed-rank

Again, match-pairs problem of pre-menstrual vs. post-menstrual dietary intake. If you want to consider the Wilcoxon test ...

```
datalines4;
PairedMeans 6500 5600
SD 1500 1300 . give exactly 2 SDs
corr .85 .90
SDMultipl 1.0  1.2 . solve for each; default 1.0
alpha .05 .01
TotalPairs 10 15 . Any word with 'total' is OK
Wilcoxon
NoNotes
;;;;
%tables
```


From UnifyPow output:

```
Comparing locations of pair of correlated
measures:
   <Parametric>    Ho: mu(Y1 - Y2) = 0

Testing location of a single group:
  <Nonparametric>  Ho: p1 = .50
     where p1 = Pr[D{i} > 0 ] + .50*Pr[D{i} = 0 ].
           D{i} = Y1{i} - Y2{i}.
```


Scenario: PairedMeans 6500 5600 . reported pre and post KJ/day & SD 1500 1300
and Alpha 0.050

			x SD (SD Multiplier)					
			1			1.2		
			Corr(Y1, Y2)			Corr(Y1, Y2)		
			0.85	0.9		0.85	0.9	
			Total Pairs	Total Pairs		Total Pairs	Total Pairs	
			10	15	10	15	10	15	10	15
			Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Method	Type	Parent								
Wilcoxon Signed Rank [Lehmann aprx]	2-tail W	Normal	.845	.994	.982	.999	.650	.923	.850	.995
		Logistic	.854	.994	.974	.999	.682	.936	.858	.994
		Laplace	.853	.992	.962	.999	.710	.944	.857	.993
	1-tail W	Normal	.963	.999	.999	.999	.836	.979	.965	.999
		Logistic	.965	.999	.999	.999	.858	.983	.967	.999

Matched-pairs Wilcoxon signed rank using p_1

```
/*
Example motivated from Manocha, et. al. (1986) as
summarized by Altman (Practical Statistics for Medical
Research, p 189). Question: Do women report eating
different amounts of food on pre- vs. post-menstral
days?
*/
1Wilcoxon .80  . 80% report pre > post
NoNotes
/#
Next line tells UnifyPow to compute for some "base" SD,
plus 10% larger, 20% larger.
#/
RelativeSD 1.00 1.10 1.20
alpha .05 .01
TotalPairs 10 15 . Any word containing 'total' is OK
parent Laplace . parent dist'n of difference score
```


			Relative Std Dev					
			1		1.1		1.2	
			Total N		Total N		Total N	
			10	15	10	15	10	15
			Power	Power	Power	Power	Power	Power
Method	Type	Parent						
Wilcoxon Signed Rank [Lehmann aprx]	2-tail W	Normal	.345	.569	.295	.487	.257	.422
		Logistic	.381	.617	.328	.535	.286	.466
		Laplace	.436	.682	.383	.606	.339	.540
	1-tail W	Normal	.510	.727	.447	.646	.396	.577
		Logistic	.552	.769	.486	.692	.433	.623
		Laplace	.611	.820	.548	.755	.496	.694
t test of one mean	2-tail t	Normal	.448	.646	.384	.565	.333	.495
	1-tail t	Normal	.598	.770	.530	.700	.473	.634

Wilcoxon-Mann-Whitney: 2 groups, ordered categorical

```
%include "&UnifyPow";
title2 "Dr. Alalgia's 2-Group Design";
title3 "Treatments/ ET: Enhanced Thermal Biofeedback";
title4 "            SP: Sham Placebo";
title5 "7-pt Likert Scale for Improvement in Headache";
/*
Much    Somewhat  A Little    No     A Little  Somewhat   Much
Worse    Worse      Worse    Change    Better    Better    Better
 -3       -2         -1        0         1         2         3
*/
datalines;
2Wilcoxon
.03 .04 .08 .10 .27 .28 .20
.10 .10 .15 .25 .20 .15 .05
weight 2 1
power .90 .95
alpha .05 .01
%tables
```

Note: New method by Kolassa (1995) offers better approximation over the method UnifyPow currently uses for this problem.


```
Testing location difference between 2 groups:
  <NonParametric>  Ho: p1 = .50
      where p1 = Pr[Y{i,1} - Y{i',2} > 0 ] + .50*Pr[Y{i,1} - Y{i',2} = 0 ].
            Y{i,g} = Y for case i in group g.

Nonparametric Moments (if no ties possible)
-------------------------------------------
Let Y{i,g} be the outcome score for case i in group g.  (For the
PairedMu problem, Y{i,g} is a difference score.)  Then,
  p1 = Pr[Y{i,1} - Y{i',2} > Null]
  p2 = Pr[(Y{i,1} - Y{k,2} > Null) and (Y{i',1} - Y{k,2} > Null)]
  p3 = Pr[(Y{i,1} - Y{k,2} > Null) and (Y{i,1} - Y{k',2} > Null)]


--------------------------------------------------------------------------
```

	Nonparametric Moments		
	p1	p2	p3
Parent			
Custom	.706	.560	.566

Dr. Alalgia's 2-Group Design
7-pt Likert Scale for Improvement in Headache

Scenario: 2Wilcoxon,
{.03 .04 .08 .10 .27 .28 .20},
{.10 .10 .15 .25 .20 .15 .05}

	Alpha			
	0.05		0.01	
	Minimum Power		Minimum Power	
	.900	.950	.900	.950
	Total N	Total N	Total N	Total N
Method / Type				
Wilcoxon-Mann-Whitney / 2-tail W	87	105	126	147
[Lehmann (p1, p2, p3) aprx] / 1-tail W	72	87	108	129

Wilcoxon signed-rank/ 1 group, interval-level categorical outcome

```
/*
Dr. Alalgia's cross-over design: biofeedback vs. sham control
Question: Does this biofeedback therapy relieve severe headache?
Measure: 7-pt Likert Scale for Improvement in Headache

   Much    Somewhat  A Little    No     A Little  Somewhat    Much
   Worse    Worse      Worse    Change    Better    Better    Better
   <-3>      <-2>      <-1>      <0>       <1>       <2>       <3>


Possible outcomes...

   Most favorable:
                    "much better" under biofeedback  .....  +3
                    "much worse" under sham control  .....  -3
                     ------------------------------------
                                    net effect ...........  +6
```



```
Least favorable:
                    "much worse" under biofeedback   .....  -3
                    "much better" under sham control .....  +3
                     ------------------------------------
                                    net effect ...........  -6

Scenario:

<-6> <-5> <-4> <-3> <-2> <-1> <00> <+1> <+2> <+3> <+4> <+5> <+6>
---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ----
.02  .03  .05  .05  .05  .06  .09  .14  .14  .14  .09  .08  .06

*/
datalines4;
1Wilcoxon
.02 .03 .05 .05 .05 .06 .09 .14 .14 .14 .09 .08 .06
Limits -6 6 .  Default: 1, 2, 3, ..., NumCat
Null 0      .  Null *must* be specified.
Ntotal 50 75 100
alpha .05
;;;;
%tables
```



```
Testing location of single group:
  <Nonparametric>  Ho: p1 = .50
      where p1 = Pr[Y{i} > 0 ] + .50*Pr[Y{i} = 0 ].

Category values (interval scale):
  -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6


Nonparametric Moments (if no ties possible)
-------------------------------------------
Let Y{i} be the outcome score for case i.  (For the PairedMu
problem, Y{i} is a difference score.)  Then,
  p1 = Pr[Y{i} > Null]
  p2 = Pr[Y{i} + Y{i'} > 2*Null]
  p3 = (p2 + p1**2)/2
  p4 = Pr[(Y{i} + Y{i'} > 2*Null) and (Y{i} + Y{i''} > 2*Null)]

UnifyPow will reverse ordering relations in order to force p1 > .5
Ties are handled by partitioning probabilities appropriately, e.g.,
  p1 = Pr[Y{i,1} - Y{i',2} > Null] + .50*Pr[Y{i,1} - Y{i',2} = Null]
```

	Nonparametric Moments		
	p1	p2	p4
Parent			
Custom	.695	.713	.574

	Alpha		
	0.05		
	Total N		
	50	75	100
Method	Type	Power	Power	Power
Wilcoxon Signed Rank	2-sided W	.765	.916	.974
[Lehmann (p1, p2, p3, p4) aprx]	1-sided W	.859	.959	.989

2-group (AB/BA) cross-over design via t tests on differences

```
/*
Asthma (Active) Treaments A vs. B, Outcome is FEV1

Two-Group (AB/BA) Cross-Over Design with Continuous Outcome

This follows Patel (1983) example (2.1) in Jones B, Kenward MG
(1989, Design and Analysis of Cross-Over Trials, Chapman and
Hall). Two active drugs, patients have acute bronchial asthma,
design is ordinary AB/BA crossover, outcome measure is FEV1.
Higher is better.

Scenario:                                  Mean FEV1
                                        Drug A        Drug B
                                   |   ------------------------
                              AB   |   mu11 = 1.6     mu12 = 1.7
                   Order           |
                              BA   |   mu21 = 1.9     mu22 = 2.3


    SD(A) = .60, SD(B) = .75, SDMultiplier = {1.0, 1.2}

    corr(A,B) = {.60, .70, 80}
*/
```



```
datalines;
scenario "B superior in both orders, but more so in BA"
//          <A>   <B>
PairedMeans 1.6   1.7   .  AB order
>           1.9   2.3   .  BA order
SD          .60   .75   .  "base" SD for A & B
Corr .6 .7 .8     . conjectures for corr(A, B)
SDMultipliers 1.0 1.2    . 100% and 120% of base SDs
sides 2
TotalPairs 50
NoOverall
contrast
"Drug x Order (= Period)" 1 -1 . compare A-B diff for AB vs BA
"Drug A vs. Drug B" 1 1  . average A-B diff over both orders
;;;;
%tables
```


Scenario: B superior in both orders, but more so in BA
and Effect: Drug x Order (= Period)

	x SD (SD Multiplier)					
	1			1.2		
	Corr(Y1, Y2)			Corr(Y1, Y2)		
	0.6	0.7	0.8	0.6	0.7	0.8
	Total Pairs	Total Pairs	Total Pairs	Total Pairs	Total Pairs	Total Pairs
	50	50	50	50	50	50
	Power	Power	Power	Power	Power	Power
Alpha	Type						
0.05	2-tail t	.390	.485	.637	.288	.360	.486

Drug A vs. Drug B

Alpha	Type						
0.05	2-tail t	.800	.893	.971	.646	.761	.894

Non-inferiority testing, also known as Blackwelder's one-sided "equivalency" testing strategy

Background. Abciximab plus a low-dose of heparin reduces complications and increases 30-day survival rates in patients undergoing high-risk coronary angioplasty or other kinds of revascularization (EPILOG Invetigators, 1997 [N Engl J Med, 336:1689]). But abciximab is expensive. Giving bivalirudin (Bittl et. al., 1995 [N Engl J Med, 333:764-9]) albeit with provisional ("bail-out") use of abciximab may give equivalent efficacy and safety at lower costs.

2-group parallel design, with 2:1 weighting:

- B+provA: Bivalirudin + provisional abciximab [n = (2/3)N]
- A+lowH: abciximab + a low-dose of heparin [n = (1/3)N]

Primary end-point is binary: Death by any cause, or an MI, or an urgent revascularization within 30 days. (same as EPILOG trial)

Non-inferiority testing: Same as one-sided equivalency trial handled using Blackwelder's method (1982 [Controlled Clinical Trials, 3:345-53]). This tests

H_0: B+provA appreciably inferior to A+lowH
H_A: B+provA not appreciably inferior to A+lowH

which becomes

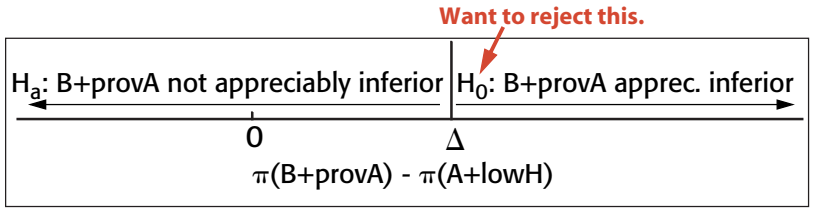
H_0: π(B+provA) - π(A+lowH) $\geq \Delta$

H_A: π(B+provA) - π(A+lowH) $< \Delta$

$\Delta > 0$ sets the "region of equivalency." Take it to be within 25% of 0.052, the 30-day event rate of found in the EPILOG trial. Thus, use $\Delta = 0.012 < 0.013 = 0.052/4$.

Assume: π(B+provA) = π(A+lowH) = 0.052.

A graphic really helps



We seek a power of 0.90 for an $\alpha = .05$ test of this one-tailed hypothesis. Use either the traditional Pearson ("approximate") unconditional test or its exact counterpart (Suissa & Shuster, 1985 [J Royal Stat Soc A, 48:317-27]), whichever is more powerful. See O'Brien and Muller (1993).]


```
proportion .052 .052  . rate from EPILOG study
scenario "B+provA equal"
weight 2 1
method chi-square ExactUnconditional
// other options for methods are 'FishersExact',
// 'LR', and 'ALL'.
null .012  . a bit less than .013 (25% of 0.052)
Ntotal 8001 9999 12000 14001
tails 1
;;;;

/* store results to build custom table */
data StorePow; set PowData;  run;
```

Sponsor actually believes that B+provA protocol may cut primary events by at least 5%. That is, they think that

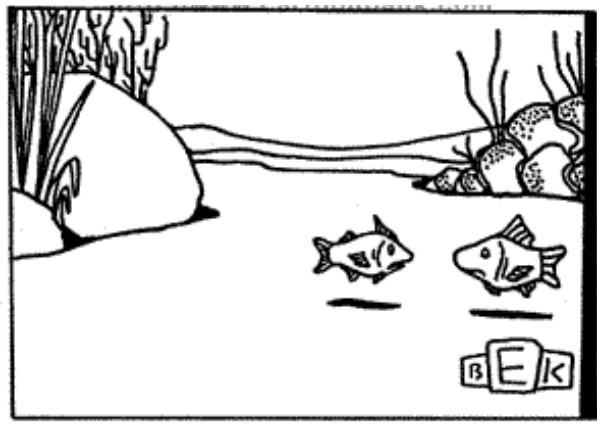
$$\pi(\text{B+provA}) = .95*.052 = .0494$$

```
proportion .0494 .052
scenario "B+provA 5% better"
weight 2 1
method chi-square ExactUnconditional
null .012
Ntotal 8001 9999 12000 14001
tails 1
;;;;

/* merge results with first set */
data PowData; set StorePow PowData;
```



```
/* build custom table */
proc tabulate format = 6.3 order=data;
 class alpha effctitl NullValu NTotal scenario;
 var power;
 table
  scenario="True State" * NullValu="Delta" *
effctitl = "Method",
  alpha = "Alpha" * NTotal ="Total N" *
power="Power"*mean=" "
    /rtspace=45; run;
```


			Alpha			
			.050			
			Total N			
			8001	9999	12000	14001
			Power	Power	Power	Power
True State	Delta	Method				
B+provA equal	0.012	Approximate Uncondit'l "chi^2*"	0.737	0.817	0.874	0.915
		Exact Uncondit'l**	0.737	0.817	0.874	0.915
B+provA 5% better	0.012	Approximate Uncondit'l "chi^2*"	0.880	0.934	0.965	0.981
		Exact Uncondit'l**	0.875	0.930	0.962	0.980



"Now they're saying that shiny things attached to hooks are bad for you."

McNemar's test of 2 correlated proportions

A matched pairs case/control study with binary outcome:
Is Thromboembolism[1] related to using The Pill?

Example follows study by Sartwell PE, et. al. (1969, Thromoembolism and oral contraceptives: an epidemiologic case-control study. Am J Epi, 90: 365-380.)

Cases — Women of childbearing age who have some kind of thromoembolism.

Controls — Matched one per case—treated in the same hospital; matched on age, number of prior pregnancies, income level.

1. blocking of a blood vessel by a blood clot dislodged from its site of origin

Scenario:

		Did Control Use The Pill?	
		No	**Yes**
Did Thromboembolism Case Use The Pill?	No	.54	.08
	Yes	.32	.06

```
McNemar .32 .08
Ntotal 50 to 150 by 25
;;;;
%tables
```



```
Testing Ho: pi12 - pi21 = 0 using McNemar's test.


Scenario: McNemar .32 .08
-----------------------------------------------------
|                           |          ALPHA          |
|                           |-------------------------|
|                           |          0.05           |
|                           |-------------------------|
|                           |          Pairs          |
|                           |-------------------------|
|                           |  50  |  75  | 100  | 125  | 150  |
|                           |------+------+------+------+------|
|                           | Pow- | Pow- | Pow- | Pow- | Pow- |
|                           |  er  |  er  |  er  |  er  |  er  |
|---------------------------+------+------+------+------+------|
|Method       |Type         |      |      |      |      |      |
|-------------+-------------|      |      |      |      |      |
|McNemar      |2-tailed     |.798  |.931  |.980  |.995  |.999  |
|(exact)      |-------------+------+------+------+------+------|
|             |1-tailed     |.839  |.954  |.988  |.997  |.999  |
-----------------------------------------------------
```


Complex yet practical power analyses

1. Formulate the problem completely. This is the hardest part.
2. Create exemplary data set for the problem. Straightforward programming (if you are a decent progammer).
3. "Analyze" exemplary data set using ordinary software. Think of this as a mock run of the real data analysis.
4. Use UnifyPow to convert test statistics (SSH or χ^2 values) to powers (for given Ns) or sufficient Ns (for given powers).

Power for general linear models using exemplary data: a simple t-test the tedious way

Simple example to introduce EXEMPLARY SSH idea and methodology. BeeBob running shoe study revisited.

1. Artificially set up groups of subjects so that key sample statistics are identical to conjectured true values, and their sample sizes reflect the cell weights. This creates an **exemplary data set.**

For ANOVA-type problems, an exemplary data set is one having:
sample means = "true" means
sample variance = anything, even 0.00.

2. Perform "ordinary" GLM or REG analyses on the exemplary data.

3. Use UnifyPow to convert exemplary SSH statistics to noncentralities in order to do power and sample-size calculations.

The theory justifying this computational trick is described in O'Brien and Muller (1993, Section 8.3; full manuscript available in PDF at www.bio.ri.ccf.org/UnifyPow). Note: The EXEMPLARY SSH problem type in UnifyPow replaces the PowSetUp module described in O'Brien and Muller.

```
data;
input shoetype $ arsin_P  n_exemp  ;  datalines4;
        XDM-X      .27        20
        XDM-S      .30        10
;;;;
proc glm order=data;
  class shoetype;  freq n_exemp;
  model arsin_P = shoetype;  ← regular analysis on exemplary data
```



```
Key part of output
------------------

Dependent Variable: ARSIN_P
Frequency:          N_EXEMP

Source     DF   Type III SS  Mean Square      Value   Pr > F

SHOETYPE   1    0.00600000   0.00600000       Infty   <.0001
```

UnifyPow commands. Note that ***now*** you put in conjectured error variance.

```
exemplary SSH
NumParms 2
Nexemplary 30
sd .08 .10 .125
NTotal 201 270
effects
"XDM-X vs. XDM-S" 1  0.00600000
;;;;
%tables
```

copied from GLM output


```
Scenario: exemplary SSH
```

	Standard Deviation					
	0.08		0.1		0.125	
	Total N		Total N		Total N	
	201	270	201	270	201	270
	Power	Power	Power	Power	Power	Power
Test	Alpha	Type						
XDM-X vs. XDM-S	0.05	2-tail t	.703	.825	.514	.639	.358	.457
		1-tail t	.803	.895	.638	.750	.482	.583

Same results as the easy way!

Nobody should use UnifyPow this way to do a traditional t test! We do it here to motivate the "exemplary SSH" method to compute and table power or sample sizes for complex general linear models.

Power for general linear models using exemplary data: a complex ANCOVA, with contrasts

Dr. Mindy Bowdy: 3-Group Design + Covariate

Groups—Personality types:
(D) Dominators (R) Regulars (F) Friendlies

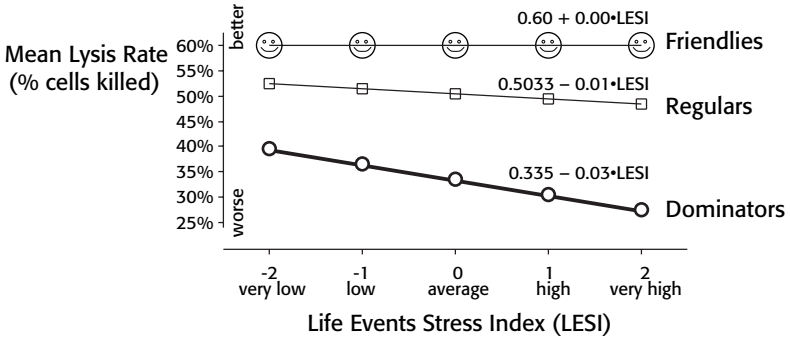
Covariate— Life Events Stress Index (LESI):
(-2) very low (-1) low (0) average (1) high (2) very high.

Outcome—Lysis rate:
% of Q843 human leukemia cells destroyed by immune response in blood (in vitro)

Model with separate simple linear regressions (unequal slopes) for each group is (i-th case in group j):

$$lysis_{ij} = \tau_j + \beta_j(LESI)_{ij} + \epsilon_{ij}$$

Conjectured means




```
data; input lysis0 DRFgrp $ LESI n;

/* create exemplary data */
if DRFgrp = 'D' then beta = -.03;
if DRFgrp = 'R' then beta = -.01;
if DRFgrp = 'F' then beta =  .00;
lysis = lysis0 + beta*LESI;
*    lysis0    DRF   LESI    n     ; datalines;
     .3350     D     -2     02
     .3350     D     -1     03
     .3350     D      0     04
     .3350     D      1     05
     .3350     D      2     06
     .5033     R     -2     12
     .5033     R     -1     12
     .5033     R      0     12
     .5033     R      1     12
     .5033     R      2     12
     .6000     F     -2     04
     .6000     F     -1     04
```



```
     .6000     F      0     04
     .6000     F      1     04
     .6000     F      2     04
;


/* "analyze" exemplary data to get SSH values */
proc glm order=data; class DRFgrp; freq n;   ← regular data analysis on exemplary data
model lysis = DRFgrp DRFgrp*LESI/noint solution;
contrast 'DvRvF main | LESI=0' DRFgrp 1 -1 0, DRFgrp 0 1 -1;
contrast 'Ave LESI slope' DRFgrp*LESI .333 .333 .333;
contrast 'DvRvF x LESI' DRFgrp*LESI 1 -1 0, DRFgrp*LESI 0 1 -1;
contrast 'DvR | LESI=0' DRFgrp 1 -1 0;
contrast 'RvF | LESI=0' DRFgrp 0 1 -1;
contrast 'Slopes: DvR' DRFgrp*LESI 1 -1 0;
contrast 'Slopes: RvF' DRFgrp*LESI 0 1 -1;
```

Key parts of the output:

```
--------------------------------------------------------------------------
                Number of observations in data set = 100

Contrast                 DF    Contrast SS  Mean Square  F Value    Pr > F

DvRvF main | LESI=0       2      0.6722149    0.3361074 99999.99    0.0001
Ave LESI slope            1      0.0258462    0.0258462 99999.99    0.0001
DvRvF x LESI              2      0.0175385    0.0087692 99999.99    0.0001
DvR | LESI=0              1      0.3837566    0.3837566 99999.99    0.0001
RvF | LESI=0              1      0.1402634    0.1402634 99999.99    0.0001
Slopes: DvR               1      0.0108387    0.0108387 99999.99    0.0001
Slopes: RvF               1      0.0030000    0.0030000 99999.99    0.0001
--------------------------------------------------------------------------
```


This information is then used in UnifyPow as follows.

```
%include "&UnifyPow";
datalines;
Exemplary SSH
Nexemplary 100
alpha .05
SD .12 .15
Ntotal 200 300 500
NumParms 6
tails 2
effects
"DvRvF main | LESI=0"  2   0.6722149
"Ave LESI slope"       1   0.0258462
"DvRvF x LESI"         2   0.0175385
```

copied from GLM output


```
/* store these .05 results; */
data PowData1; set PowData;

/* now run .025 tests (Bonferroni adjustment) */
%include "&UnifyPow";
datalines;
Exemplary SSH
Nexemplary 100
alpha .025
SD .12 .15
Ntotal 200 300 500
NumParms 6
sides 2
effects
"DvR | LESI=0"     1     0.3837566
"RvF | LESI=0"     1     0.1402634
"Slopes: DvR"      1     0.0108387
"Slopes: RvF"      1     0.0030000

/* concantenate datasets and produce one table */
data PowData;  set PowData1 PowData;
%tables
```

copied from GLM output

Scenario: Exemplary SSH

			Std Dev					
			0.12			0.15		
			Total N			Total N		
			200	300	500	200	300	500
			Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Test	Alpha	Type						
DvRvF main | LESI=0	0.05	Regular F	.999	.999	.999	.999	.999	.999
Ave LESI slope	0.05	2-tail t	.470	.638	.848	.326	.456	.667
DvRvF x LESI	0.05	Regular F	.264	.380	.588	.182	.256	.404
DvR | LESI=0	0.025	2-tail t	.999	.999	.999	.999	.999	.999
RvF | LESI=0	0.025	2-tail t	.984	.999	.999	.897	.981	.999
Slopes: DvR	0.025	2-tail t	.154	.228	.380	.103	.148	.244

Power for logit analyses using exemplary data: comparing 2 independent proportions (the tedious way)

Remember this?

Design and Scenario

	Lived	Died	
Placebo	72% of n_1	28% of n_1	$n_1 = N/3$
DCA	79% of n_2	21% of n_2	$n_2 = 2N/3$

A trick to handle any situation involving likelihood ratio tests in categorical modeling.

```
title2 "Lactic Acidosis in Children with Malaria (revisited)";
title3 "28% die untreated. What if DCA cuts this by 25%?";
/*
Theory: O'Brien (1986, SUGI-11 Proceedings, 778-784) and summarized in
Agresti (1990, Analysis of Categorical Data, Wiley, p 243). Also: Shieh G,
O'Brien RG (1998, unpublished JSM invited talk by O'Brien), available from
www.bio.ri.ccf.org/UnifyPow. Dr. Shieh furtively published this joint work
on his alone: Shieh G (2000, Biometrics, 56:1192-1196).
*/

data exemplary;
  input treatment $ outcome $ n_exemplary;
  DCA = 0*(treatment="placebo") + 1*(treatment="DCA");
  died = 0*(Outcome = "lived") + 1*(Outcome = "died");
/* treatment outcome  n_exemplary */  datalines4;
    placebo   lived      72
    placebo   died       28  /* 28% mortality */
    DCA       lived     158
    DCA       died       42  /* 21% mortality */
;;;;
run;
```


Proc Print of exemplary data set:

Obs	treatment	outcome	n_exemplary	DCA	died
1	placebo	lived	72	0	0
2	placebo	died	28	0	1
3	DCA	lived	158	1	0
4	DCA	died	42	1	1

Regular data analysis on the exemplary data:

```
proc logistic;
 weight n_exemplary;  model died = DCA;
```


Key part of output:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.7903	1	0.1809

```
exemplary chi**2
Nexemplary 300
alpha .01 .045
NTotal 750 999 1500
effects
"DCA vs. Placebo"  1  1.7903
%tables
```

← *copied from LOGISTIC output*

Scenario: exemplary chi**2

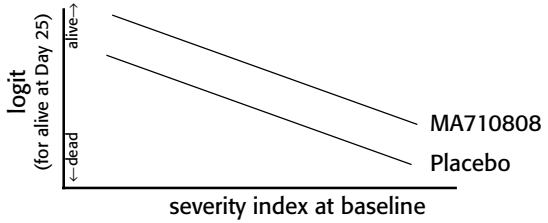
	alpha	
	0.0-12	0.0-46
	Total N	Total N
	999	1500
	Power	Power
Effect: DCA vs. Placebo; Statistic: 2-sided	.476	.841
Effect: DCA vs. Placebo; Statistic: 1-sided	.577	.905

Same results as the easy way!

Logistic regression

Design and Rough Scenario

- Monoclonal antibody MA710808 is a possible treatment for patients in critical condition with severe sepsis.
- Randomized, placebo controlled, blinded Phase III trial
- 2/3 will get MA710808; 1/3 will get placebo



Predictor Variables

- **SevIndx0.** Note: Several baseline measures are to be collected and combined with a baseline severity index, SevIndx0. SevIndx0 is a continuous measure developed in the previous Phase II trial and in this population ranges from 1.5 (least severe) to 9.5 with a tri-modal shape. For the sake of pragmatism, we shall simplify somewhat and pretend that its distribution is discrete:

SevIndx0:	2	3	4	5	6	7	8	9
Probability:	.10	.10	.05	.25	.25	.05	.10	.10

- **MA710808.** Code the dummy variable:
  MA710808 = 0, if received placebo
  MA710808 = 1, if received MA710808
- Thus, Pr[SevIndx0 = 2 & MA710808 = 0] =
  1/10 * 1/3 = 1/30 = 2/60.

Assumed Model for Sample-Size Analysis

- From the Phase II results we take the linear predictor to be

  logit = 2.5 - 0.5*SevIndx0 + ln(3)*MA710808.

  The specification "ln(3)*MA710808" sets an odds ratio of 3 in favor of MA710808. This is a large effect, but actually less than the odds ratio of almost 5.0 observed in the Phase II study.

Create exemplary data set for the problem:

skip all detail in workshop

```
data sepsis;
keep SevIndx0 MA710808 probX Alive25 probY ExmpCnt;
input SevIndx0 ProbSevI;
TotNExmp = 10002;  *Total N of exemplary data;
RanRateP = 1/3;  *Randomization Rate for the Placebo group;
betaInt = 2.5;
betaSev = -0.5;
betaMA = log(3);  *Note: ln(3) = 1.0986;
/*
Each dataline gives one SevIndx0 group. From each we generate
"exemplary" counts, ExmpCnt, for 4 types of cases based on the total
exemplary sample size, TotNExmp:
    (1) Cases on placebo who are not alive at Day 25.
    (2) Cases on placebo who are alive at Day 25.
    (3) Cases on MA710808 who are not alive at Day 25.
    (4) Cases on MA710808 who are alive at Day 25.
*/
do MA710808 = 0 to 1;  * 1 = randomized to MA710808;
  logit = betaInt + betaSev*SevIndx0 + betaMA*MA710808;
  PrAliv25 = 1/(1 + exp(-logit));  *Prob that case is alive at Day 25;
  if MA710808 = 0 then RandRate = RanRateP;
  else RandRate = 1 - RanRateP;
  probX = RandRate*ProbSevI;
  do Alive25 = 0 to 1;  * 1 = alive at Day 25;
    if Alive25 = 0 then probY = (1-PrAliv25);
    else probY = PrAliv25;
    ExmpCnt = round(TotNExmp*probX*probY);  *Exemplary count;
    output;
```



```
    end;
  end;
*SevIndx0  ProbSevI  ;     datalines;
     2       .10
     3       .10
     4       .05
     5       .25
     6       .25
     7       .05
     8       .10
     9       .10
```

32 prototypical cases comprise the exemplary data set. The variable ExmpCnt gives the exemplary counts based on N_{total} = 10,002 (= 10,005 due to rounding).

Case	SevIndx0	MA710808	ProbX	Alive25	ProbY	ExmpCnt
1	2	0	0.03333	0	0.18243	61
2	2	0	0.03333	1	0.81757	273
3	2	1	0.06667	0	0.06923	46
4	2	1	0.06667	1	0.93077	621
5	3	0	0.03333	0	0.26894	90
/\						
cases 6-30 not shown						
/\						
30	9	0	0.03333	1	0.11920	40
31	9	1	0.06667	0	0.71123	474
32	9	1	0.06667	1	0.28877	193

"Analyze" exemplary data set using ordinary software.

- Input (SAS):

```
proc genmod;
  model Alive25 = SevIndx0 MA710808 /
            dist = binomial
            link = logit
            type3;
    freq ExmpCnt;
```


- Key lines from output:

```
Sum Of Frequency Weights          10005  <--

      LR Statistics For Type 3 Analysis

     Source          DF    ChiSquare   Pr>Chi

     SEVINDX0         1      1820.89   <0.0001
 --> MA710808         1       528.95   <0.0001
```

Convert test statistics to powers or sufficient Ns.

```
%include "&UnifyPow";
datalines4;
Exemplary chi**2
Nexemplary 10005 . Not 10002 due to rounding.
alpha .01 .05
power .80 .90 .95
ForceN 3 . Forces Ntotal to be multiple of 3
effects
"MA710808 vs. Placebo"  1   528.9458
```


Scenario: Exemplary chi**2

		alpha					
		0.010			0.050		
		Minimum Power			Minimum Power		
		0.800	0.900	0.950	0.800	0.900	0.950
		Total N	Total N	Total N	Total N	Total N	Total N
Effect	Statistic						
MA710808 vs. Placebo	2-tail Z	222	282	339	150	201	249
	1-tail Z	192	249	300	117	162	207

Poisson regression

- **8 large identical computers.** The Reliable Web Server Company is relying on 8 large identical computers to support all of its clients.
- **Computers crash.** Since Reliable upgraded its operating system 12 months ago to JupiterOS v3.0, it has been experiencing a number of unexplained system crashes, averaging over 1 crash per week per computer, but with some variation over computers:

Computer:	A	B	C	D	E	F	G	H
Crashes/week:	1.2	2.0	0.6	0.8	1.4	1.2	1.0	0.4

This crash rate is barely acceptable.

- **New release, JupiterOS v3.1.** Supposed to greatly reduce the number of unexplained crashes. Reliable's systems managers are skeptical of such claims and worried that the costs involved in upgrading to v3.1 may not justify the increased reliability, or, worse, that the new version may be even more crash prone.

Experiment!

- **Upgrade only 2 of the 8 machines** so they can compare v3.0 and v3.1 over a number of weeks. Any other changes to the computers will be done identically to all 8.

- **How long should they test v3.1?**

Outcome Measure

- **Number of crashes per week.** Let Y_{it} be the number of crashes for Computer i in week t.
- **Y_{it} ~ Poisson(m_i).** Assume crashes are independent events. Then Y_{it} is distributed as a Poisson random variable with mean m_i.

Predictor variables

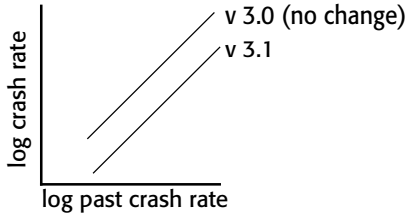
- **PastCrRt.** Past crash rates given above.
- **V31.** Code the dummy variable:
  V31 = 0, if running v3.0
  V31 = 1, if running v3.1

Assumed Model for Sample-Size Analysis

- Take the linear predictor to be

$$\ln\{m_i\} = 0.0 + 1.0*\ln[\text{PastCrRt}_i] - \ln(3)*V31_i.$$

  - Makes makes future crash rates for v3.0 equal to past crash rates.
  - But makes v3.1 have 1/3 the number of crashes as v3.0 (given the same past crash rate).



Create exemplary data set for the problem.

Do not bother with during presentation!

```
data crashes;
title2 "Get exemplary LR test value for computer crashing problem";
keep PastCrRt V31 m Ncrashes lnPCR lnTotWks;
input Computer $ PastCrRt V31;
TotWeeks = 100;
betaInt = 0.0;
betaPCR = 1.0;
betaV31 = -log(3);  *Note: -ln(3) = = -1.0986;
m = exp(betaInt + betaPCR*log(PastCrRt) + betaV31*V31);
Ncrashes = TotWeeks*m;
lnPCR = log(PastCrRt);
lnTotWks = log(TotWeeks);

* Computer  PastCrRt  V31;      datalines;
     A        1.2      0
     B        2.0      1
     C        0.6      0
     D        0.8      1
     E        1.4      0
     F        1.2      0
     G        1.0      0
     H        0.4      0
```


8 prototypical computers comprise the exemplary data set. The variable Ncrashes gives the exemplary number of crashes for 100 weeks.

Case	PastCrRt	v31	m	Ncrashes
1	1.2	0	1.20000	120.000
2	2.0	1	0.66667	66.667
3	0.6	0	0.60000	60.000
4	0.8	1	0.26667	26.667
5	1.4	0	1.40000	140.000
6	1.2	0	1.20000	120.000
7	1.0	0	1.00000	100.000
8	0.4	0	0.40000	40.000

"Analyze" exemplary data set using ordinary software.

- Key parts of input (SAS):

```
/*
   From data step:
                  TotWeeks = 100;
                  lnTotWks = log(TotWeeks);
                  lnPCR = log(PastCrRt);
*/
     proc genmod;
       model Ncrashes = lnPCR V31 /
          dist = Poisson
          link = log
          offset = lnTotWks
          type3;
```


Key lines from output:

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
LNPCR	1	104.1605	0.0001
V31	1	104.5144	0.0001

Convert test statistics to powers or sufficient Ns.

```
title2 "Sample-size analysis for JupiterOS
v3.1 evaluation";
title3 "Units = weeks (all eight machines)";
%include "&UnifyPow";
datalines;
Exemplary chi**2
Nexemplary 100 .  weeks (all eight machines)
alpha .05 .10
power .90 .95
tails 2
effects
"v3.1 vs. v3.0"  1  104.5144
%tables
```


Scenario: Exemplary chi**2

		ALPHA			
		0.05		0.1	
		Minimum Power		Minimum Power	
		.900	.950	.900	.950
		Total N	Total N	Total N	Total N
Effect	Statistic				
v3.1 vs. v3.0	2-tail Z	11	13	9	11

A glimpse at some cool general theory

Source: R. O'Brien's invited talk at the 1998 JSM; handout on the UnifyPow website. G. Shieh collaborated on this, then soon unilaterally published the essence of the work in *Biometrics* (December, 2000). Nuf said.

Consider any generalized linear model where the hypothesis of interest is formed by comparing a model with rank r_{full} parameters versus one with $r_{reduced}$ parameters. The usual LR test statistic has a noncentral χ^2 distribution with $(r_{full} - r_{reduced})$ degrees of freedom and noncentrality parameter,

$$\lambda \cong N\lambda^* + \zeta$$

where λ^* and ζ are functions of the design matrix and true population parameters, **but not N.**

λ^* is relatively easy to understand and compute. (This is what we have been doing.)

ζ is unwieldy, but fortunately it is usually inconsequential in value compared to $N\lambda^*$. ζ can be either positive or negative.

Thus in practice we can use:

$$\lambda \cong N\lambda^*$$

How much difference does it make to leave out ζ?

Logistic Regression Example ($\alpha = .01$)

Total N	Nominal Power (2-sided test)		95% Confidence Limits (5000 simulations)		$\frac{\zeta}{N\lambda^* + \zeta}$
	$N\lambda^* + \zeta$	$N\lambda^*$	lower	upper	
222	.806	.802	.791	.813	.0083
282	.903	.901	.898	.915	.0066
339	.952	.951	.954	.964	.0055

Poisson Regression Example ($\alpha = .10$)

Total Weeks	Nominal Power (2-sided test)		95% Confidence Limits (5000 simulations)		$\frac{\zeta}{N\lambda^* + \zeta}$
	$N\lambda^* + \zeta$	$N\lambda^*$	lower	upper	
9	.928	.923	.933	.946	.0251
11	.962	.960	.969	.978	.0205

Another logit analysis ("historical importance?")

Old example from O'Brien (1986, SUGI-11 Proceedings, 778-784); later summarized in Agresti (1990, Analysis of Categorical Data, Wiley, p 243).

"PrNRcare" stands for Probability that "Non-Routine" cardiac care is given at birth. The question is whether a new diagnostic test ("Lyons") provides additional value in predicting NRcare beyond that provided by using only the Standard test.

Outcome at birth: "NonRoutn"

0 = no non-routine care was given
1 = non-routine care was given

Predictors before birth

"Standard" test

1 = worrisome result
2 = reassuring result

"Lyons" test

1 = very worrisome result
2 = somewhat worrisome result
3 = reassuring result

```
data babies;
input Standard Lyons CellProb PrNRcare;

* expected number of routine outcomes;
NonRoutn=0;
  ExempCnt = (1-PrNRcare)*CellProb*1000;
  output;
* expected number of nonroutine outcomes;
```



```
NonRoutn=1;
  ExempCnt = PrNRcare*CellProb*1000;
  output;

*Standard  Lyons  CellProb  PrNRcare  ;     datalines;
     1       1       .04       .40
     1       2       .08       .32
     1       3       .04       .27
     2       1       .02       .30
     2       2       .18       .22
     2       3       .64       .15
data babies; set babies;
*build dummy vars for PROC LOGIST;
XStd = (Standard=2) - (Standard=1);
XLyons1 = (Lyons=1) - (Lyons=3);
XLyons2 = (Lyons=2) - (Lyons=3);
XSL1 = XStd*XLyons1;
XSL2 = XStd*XLyons2;
```

Proc Print of exemplary data set:

OBS	STANDARD	LYONS	CELLPROB	PRNRCARE	NONROUTN	EXEMPCNT
1	1	1	0.04	0.40	0	24.0
2	1	1	0.04	0.40	1	16.0

3	1	2	0.08	0.32	0	54.4
4	1	2	0.08	0.32	1	25.6
5	1	3	0.04	0.27	0	29.2
6	1	3	0.04	0.27	1	10.8
7	2	1	0.02	0.30	0	14.0
8	2	1	0.02	0.30	1	6.0
9	2	2	0.18	0.22	0	140.4
10	2	2	0.18	0.22	1	39.6
11	2	3	0.64	0.15	0	544.0
12	2	3	0.64	0.15	1	96.0

```
/* "analyze" exemplary data to get -2lnL values */

proc logistic;  weight ExempCnt;
  model NonRoutn = Xstd;

proc logistic;  weight ExempCnt;
  model NonRoutn = Xstd XLyons1;

proc logistic;  weight ExempCnt;
  model NonRoutn = Xstd XLyons1 XLyons2;

proc logistic; weight ExempCnt;
  model NonRoutn = Xstd XSL1 XSL2;

proc logistic;  weight ExempCnt;
  model NonRoutn = Xstd XLyons1 XLyons2 XSL1 XSL2;
```

Key parts of proc logistic output

Criterion	Intercept Only	Intercept and Covariates	
-2 LOG L	983.943	964.437	← Model: Standard main only
-2 LOG L	983.943	956.319	← Model: Standard main + Lyons(lin)
-2 LOG L	983.943	956.277	← Model: Standard main + Lyons(2 df) main
-2 LOG L	983.943	960.910	← Model: Standard main + Standard*Lyons(2 df) interaction
-2 LOG L	983.943	955.990	← Model: Standard + Lyons(2 df) main + Standard*Lyons(2 df) interaction

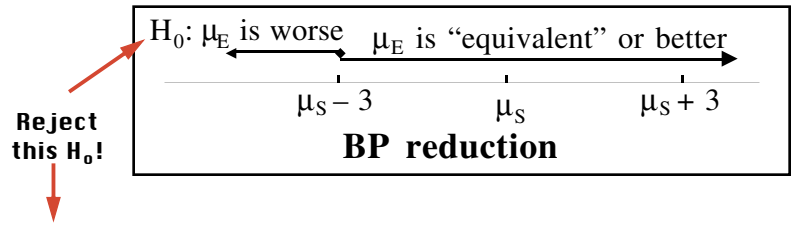

```
%include "&UnifyPow";
datalines;
Exemplary chi**2
Nexemplary 1000
alpha .01 .05 .
Ntotal  400 600 1000 .
effects
"Standard Only"  1  983.943  964.437
"Lyons(lin) Given Standand"  1  964.437  956.319
"Lyons Given Standard"  2  964.437  956.277
;
```


		ALPHA					
		0.01			0.05		
		Total N			Total N		
		400	600	1000	400	600	1000
		Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Effect	Type						
Standard Only	2-tail Z	.586	.801	.967	.798	.928	.993
	1-tail Z	.680	.863	.982	.875	.962	.997
Lyons(lin) Given Standand	2-tail Z	.220	.356	.608	.437	.598	.813
	1-tail Z	.300	.452	.699	.562	.713	.886
Lyons Given Standard	Chi-square	.156	.266	.498	.347	.495	.727

Composite power + Non-inferiority testing

Complex example extended from one supplied by nQuery Advisor developer Janet Elashoff.

- Experimental vs. Standard drug
- Outcome: BP reduction (means: μ_E, μ_S; higher is better)
- SD = 6 mmHg
- Want to show E is "at least as good" as S:
- μ_E is "equivalent" to μ_S (within ±3)
- or μ_E is better than μ_S ($\mu_E > \mu_S$)



- H_0: $\mu_E - \mu_S \leq -3$ (E worse than S)
  H_A: $\mu_E - \mu_S > -3$ (E "equivalent" or better)
- Scenario: $\mu_E - \mu_S = 0$ (E the same)
- Sampling ratio: $n_S/n_E = 4$ (only 20% get E)
- Desired power: 90%


```
%include "&UnifyPow";
datalines4;
mu 0 0    . Experimental and Standard are the same
null -3   . Experimental must be no worse than 3
sd 5 6 8
weight 1 4   . 20% randomized to Experimental
alpha .01 .05
power .90 .95
tails 1   . Experimental just as good or better
;;;;
%tables

%include "&UnifyPow";
datalines4;
mu 1 0    . Experimental 1 mmHg better than Standard
null -3   . Experimental must be no worse than 3
sd 5 6 8
weight 1 4   . 20% randomized to Experimental
alpha .01 .05
power .90 .95
tails 1   . Experimental just as good or better
;;;;
%tables
```


2-group one-sided equivalence test
ENAR 99

Scenario: mu 0 0 . Experimental and Standard are the same
Ordinary t test

		Standard Deviation					
		5		6		8	
		Minimum Power		Minimum Power		Minimum Power	
		.900	.950	.900	.950	.900	.950
		Total N	Total N	Total N	Total N	Total N	Total N
Alpha	Type						
0.01	1-tail t	230	280	330	400	585	710
0.05	1-tail t	155	190	220	275	385	485

Scenario: mu 1 0 . Experimental 1 mmHg better than Standard
Ordinary t test

		Standard Deviation					
		5		6		8	
		Minimum Power		Minimum Power		Minimum Power	
		.900	.950	.900	.950	.900	.950
		Total N	Total N	Total N	Total N	Total N	Total N
Alpha	Type						
0.01	1-tail t	130	160	190	225	330	400
0.05	1-tail t	90	110	125	155	220	275

Composite solutions: UnifyPow + a *little* SAS programming

UnifyPow builds SAS datasets of the results, which gives it extraordinary flexibility if you know a little SAS.

Strategy. In this problem, we do not know the true difference between the group means or the common SD. If we make educated guesses or have actual "pilot" estimates of those values, then we can put prior probabilities on those guesses/estimates. Suppose we have the following priors on the true differences between the groups and the true SD:

	Conjecture for mean difference: $\mu_E - \mu_S$				
	−1	**0**	**1**	**2**	**3**
Prior Prob	0.10	0.30	0.30	0.20	0.10

	Conjecture for SD				
	5.0	**5.5**	**6.0**	**7.0**	**8.0**
Prior Prob	0.10	0.20	0.30	0.25	0.15

This defines 5•5 = 25 scenarios! Consider $\mu_E - \mu_S = 0.0$ and SD = 5.0. It has a prior probability of 0.30•0.10 = 0.03.

For $\alpha = 0.05$ and N = 150, the power is 0.900 (in output below).

"Composite power" =

$$\sum_i \sum_j \text{Prob(mean diff = i)•Prob(SD = j)•Power(i, j)}$$

The SAS code below is supplied for completeness only. We will not go over this in the workshop! To SAS users this is basic stuff. To others, it is not hard to figure it out and just mimic. The point here is that because UnifyPow is embedded within SAS, you can tailor your output to do cool things.

```
%include "&UnifyPow";
datalines4;
mu 0 1    . Experimental 1 mmHg worse than Standard
null -3   . Experimental must be no worse than 3 worse
sd 5 5.5 6 7 8
weight 1 4   . 20% randomized to Experimental
```



```
alpha .01 .05
Ntotal 150 200 300
tails 1   . Experimental just as good or better
;;;;
data AllRslts; set powdata;
scenwgt = .1;
scenario = "Exper'tl 1 worse  (prior=0.1)  ";
run;

%include "&UnifyPow";
datalines4;
mu 0 0    . Experimental = Standard
null -3   . Experimental must be no worse than 3 worse
sd 5 5.5 6 7 8
weight 1 4   . 20% randomized to Experimental
alpha .01 .05
Ntotal 150 200 300
tails 1   . Experimental just as good or better
;;;;
data powdata; set powdata;
scenwgt = .3;
scenario = "Exper'tl is same (prior=0.3) ";
data AllRslts; set AllRslts powdata;
run;

%include "&UnifyPow";
datalines4;
mu 1 0    . Experimental  1 better
null -3   . Experimental must be no worse than 3 worse
sd 5 5.5 6 7 8
weight 1 4   . 20% randomized to Experimental
alpha .01 .05
```

```
Ntotal 150 200 300
tails 1  . Experimental just as good or better
;;;;
data powdata; set powdata;
scenwgt = .3;
scenario = "Exper'tl 1 better (prior=0.3)  ";
data AllRslts; set AllRslts powdata;
run;

%include "&UnifyPow";
datalines4;
mu 2 0   . Experimental  2 better
null -3  . Experimental must be no worse than 3 worse
sd 5 5.5 6 7 8
weight 1 4  . 20% randomized to Experimental
alpha .01 .05
Ntotal 150 200 300
tails 1  . Experimental just as good or better
;;;;
data powdata; set powdata;
scenwgt = .2;
scenario = "Exper'tl 2 better (prior=0.2)  ";
data AllRslts; set AllRslts powdata;
run;

%include "&UnifyPow";
datalines4;
mu 3 0   . Experimental  3 better
null -3  . Experimental must be no worse than 3 worse
sd 5 5.5 6 7 8
weight 1 4  . 20% randomized to Experimental
alpha .01 .05
```



```
Ntotal 150 200 300
tails 1  . Experimental just as good or better
;;;;
data powdata; set powdata;
scenwgt = .1;
scenario = "Exper'tl 3 better (prior=0.1)  ";
data AllRslts; set AllRslts powdata;
run;

/* Table all scearios at once for SD = 5, 6, 8 */
  data trimSD; set AllRslts;
  if (SD = 5) or (SD = 6) or (SD = 8);
  proc tabulate  format = 4.3 order=data;
    class alpha scenario SD NTotal;
    var power;
    table scenario="Scenario" * alpha="Alpha",
      SD="Standard Deviation BP change (mmHg)"
      * Ntotal="Total N" * Power="Power"*mean=" "
            /rtspace=28;
```


		Standard Deviation BP change (mmHg)								
		5			6			8		
		Total N			Total N			Total N		
		150	200	300	150	200	300	150	200	300
		Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Scenario	Alpha									
Exper'tl 1 worse (prior=0.1)	0.01	.350	.468	.667	.239	.325	.489	.133	.178	.274
	0.05	.620	.729	.869	.492	.593	.745	.335	.407	.533
Exper'tl is same (prior=0.3)	0.01	.721	.852	.965	.540	.685	.869	.306	.413	.603
	0.05	.900	.959	.994	.786	.880	.965	.473	.681	.828
Exper'tl 2 better (prior=0.2)	0.01	.994	.999	.999	.954	.991	.999	.760	.882	.976
	0.05	.999	.999	.999	.992	.999	.999	.920	.970	.996
Exper'tl 3 better (prior=0.1)	0.01	.999	.999	.999	.994	.999	.999	.906	.970	.998
	0.05	.999	.999	.999	.999	.999	.999	.978	.995	.999

Who can digest all this?

SAS code to compute composite power

```
/* Weighted average of scenarios gives one composite table */
  data AllRslts; set AllRslts;
  SDwgt = (SD=5)*.10 +
          (SD=5.5)*.20 +
          (SD=6)*.30 +
          (SD=7)*.25 +
          (SD=8)*.15;
  ScenWgt = ScenWgt*SDwgt;
  CompScen = "Composite Using Priors";

  proc tabulate  format = 4.3 order=data;
    class alpha CompScen testtype NTotal;
    var power;
    weight scenwgt;
    table CompScen="Scenario" * alpha="Alpha" * testtype="Type",
      Ntotal="Total N" * Power="Power"*mean=" "
            /rtspace=31;
```

Boiling it all down...

		Total N		
		150	200	300
Scenario	Alpha	Type	Power	Power	Power
Composite Using Priors	0.01	1-tail t	.685	.777	.878
	0.05	1-tail t	.840	.896	.949

Thanks for Coming!

Rebuild the Tribe!