Reducing Type 1 and Type 2 Errors

Jeffrey Michael Franc
MD, FCFP.EM, Dip Sport Med, EMDM

Medical Director, E/D Management
Alberta Health Services

Associate Clinical Professor of Emergency Medicine
University of Alberta

Visiting Professor in Disaster Medicine
Universita' Degli Studi del Piemonte Orientale

Objectives

1. Understand the principles of deciding on a sample size to reduce Type 1 and Type 2 errors
2. Understand the impact of multiple hypothesis testing on type-1 risk

Quiz: Question 1

Our researcher wishes to calculate the appropriate sample size to ensure that the study has enough power to capture a significant difference.

Which of the following is a possible solution? (You can choose more than one)

Quiz: Options

A. Use 30 days
B. Do a power based calculation of sample size
C. Use as big of a sample as possible
D. Consult a statistician
E. Calculate sample size using a web based service
F. Use tables of sample size
G. Do a small pilot study and calculate sample size after

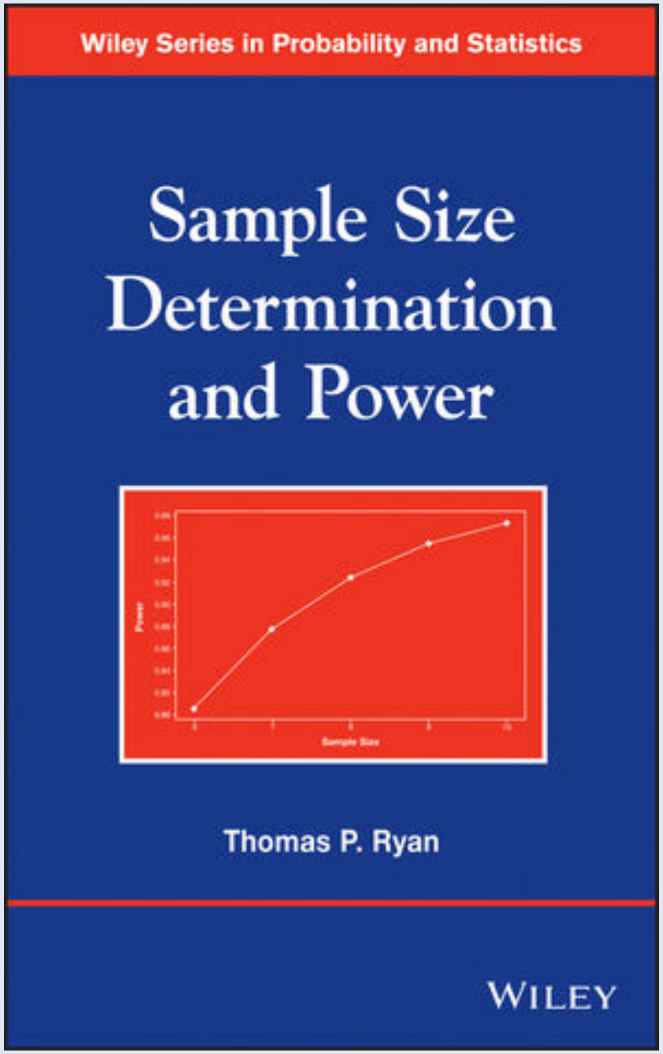
Quiz: Question 2

The researcher decides that in order to optimize his time, he will test all the flavors of gelato with his additive.

Is this a good idea? Why or why not?

Sample Size


CAUTION
CHAOS FIELD
ESTIMATED STRENGTH: 47 KrZ
LIMIT EXPOSURE TO THIS AREA
AND REPORT ABNORMALITIES
IN YOUR LIFE AFTER EXPOSURE


Wiley Series in Probability and Statistics
Sample Size
Determination
and Power
Thomas P. Ryan
WILEY

Sample Size

How do we select the sample size?

1. As big as you can get
   - The whole population
   - Pre-imposed limits
2. Choose 30
3. Power calculation
   - using "generic" effect sizes
   - using variance

Whole Population

Give some examples of when we will examine the whole population:

Whole Population

Give some examples of when we will examine the whole population:

- A researcher wants to examine the success of a course in ultrasound on the ability of residents to detect free fluid. The course will be offered to all residents in the Emergency Medicine Program

Pre-imposed Limits

Give an example when pre-imposed limits determine the sample size:

Pre-imposed Limits

Give an example when pre-imposed limits determine the sample size:

- A study is being done on a new experimental ventilator for ICU patients. The manufacturer will allow the ICU to use the ventilator for only 48 hours.

Just Use 30

When you have no information about the population, researchers will often start with 30 subjects. Why?

Central Limit Theorum

When a mean of 30 observations is used, its distributions approaches the Normal (Gaussian Curve) even if the actual population is non-normal.

(Just makes the statistics easy)

Power Calculation

Who has done this?

Was it easy?

Power Calculation

Steps to find sample size with power:

1. Specify hypothesis and test
2. Specify the significance (α)
3. Specify and effect size that is of *scientific* interest
   - Obtain historical values or estimates of other parameters (usually σ)**
   - "Generic" effect size
4. Specify a target power
5. Calculate

Sample Size Calculation for the Gelato Experiment

1: Specify Hypothesis and Test

H_0: ($\mu_1 = \mu_2$) True daily mean chocolate sales are equal with and without additive.

H_A: ($\mu_1 \neq \mu_2$) True daily mean chocolate sales are not equal.

Test statistic:

$$\bar{X}_2 - \bar{X}_1 = \frac{\sum X_2}{n_2} - \frac{\sum X_1}{n_1}$$

(Test is two-sample t-test)

2: Specify Level of Significance

$\alpha = 0.05$

Probability of committing type 1 error

(The researcher accepts that if the null hypothesis is true, there is about a 5% chance of actually rejecting it)

3: Calculate Effect Size

Who has been asked to
calculate an effect size?

3: Effect Size

- Effect size: Estimated magnitude of the relationship
- This is a scientific (not statistical) question
- Unfortunately involves several unknown quantities

$$\tilde{\theta} = \frac{(\mu_2 - \mu_1)}{\sigma}$$

3: Effect Size

How do we determine μ_2 ?

This is *scientific* question.

What degree of change in the mean is practically (clinically) relevant?

3: Effect Size

How do we determine σ ?

- Previous research
- Pilot study
- Estimate from max and min
  (generally range = 6σ)

3: Effect Size for Gelato Experiment

- The researcher feels that an increase of at least 4 scoops per day would warrant retooling of the factory
- The owner is pretty sure that the range of scoops sold is always between 15 and 45. ($6\sigma=30$)

$$\tilde{\theta} = \frac{(34-30)}{5} = 0.8$$

3: Generic Effect for Gelato Experiment

A POWER PRIMER

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size: Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. $P = .5$ and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Note. ES = population effect size.

4: Specify Target Power

- $\beta=0.2$ Power=0.8
- Researcher accepts the risk that if the CHOSEN alternative hypothesis were true, there would be about an 80% chance of detecting it.
- Usually power of 80% is chosen
- Balance of adequate power with manageable sample size

5. Making the Calculation: Tables

Table 2

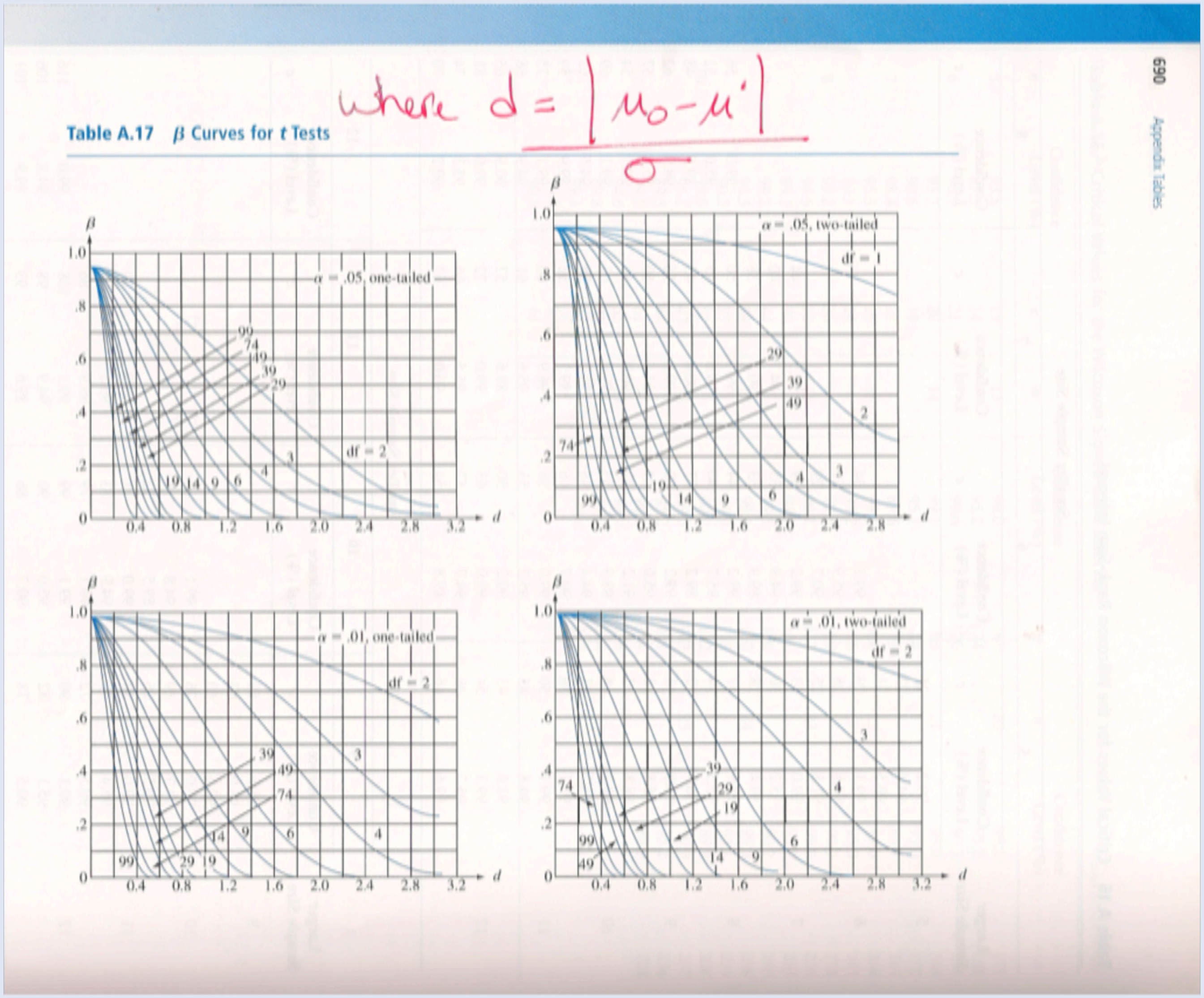
N for Small, Medium, and Large ES at Power = .80 for α = .01, .05, and .10

Test	α .01 Sm	.01 Med	.01 Lg	.05 Sm	.05 Med	.05 Lg	.10 Sm	.10 Med	.10 Lg
1. Mean dif	586	95	38	393	64	26	310	50	20
2. Sig *r*	1,163	125	41	783	85	28	617	68	22
3. *r* dif	2,339	263	96	1,573	177	66	1,240	140	52
4. *P* = .5	1,165	127	44	783	85	30	616	67	23
5. *P* dif	584	93	36	392	63	25	309	49	19
6. χ^2									
1*df*	1,168	130	38	785	87	26	618	69	25
2*df*	1,388	154	56	964	107	39	771	86	31
3*df*	1,546	172	62	1,090	121	44	880	98	35
4*df*	1,675	186	67	1,194	133	48	968	108	39
5*df*	1,787	199	71	1,293	143	51	1,045	116	42
6*df*	1,887	210	75	1,362	151	54	1,113	124	45
7. ANOVA									
$2g^a$	586	95	38	393	64	26	310	50	20
$3g^a$	464	76	30	322	52	21	258	41	17
$4g^a$	388	63	25	274	45	18	221	36	15
$5g^a$	336	55	22	240	39	16	193	32	13
$6g^a$	299	49	20	215	35	14	174	28	12
$7g^a$	271	44	18	195	32	13	159	26	11
8. Mult *R*									
$2k^b$	698	97	45	481	67	30			
$3k^b$	780	108	50	547	76	34			
$4k^b$	841	118	55	599	84	38			
$5k^b$	901	126	59	645	91	42			
$6k^b$	953	134	63	686	97	45			
$7k^b$	998	141	66	726	102	48			
$8k^b$	1,039	147	69	757	107	50			

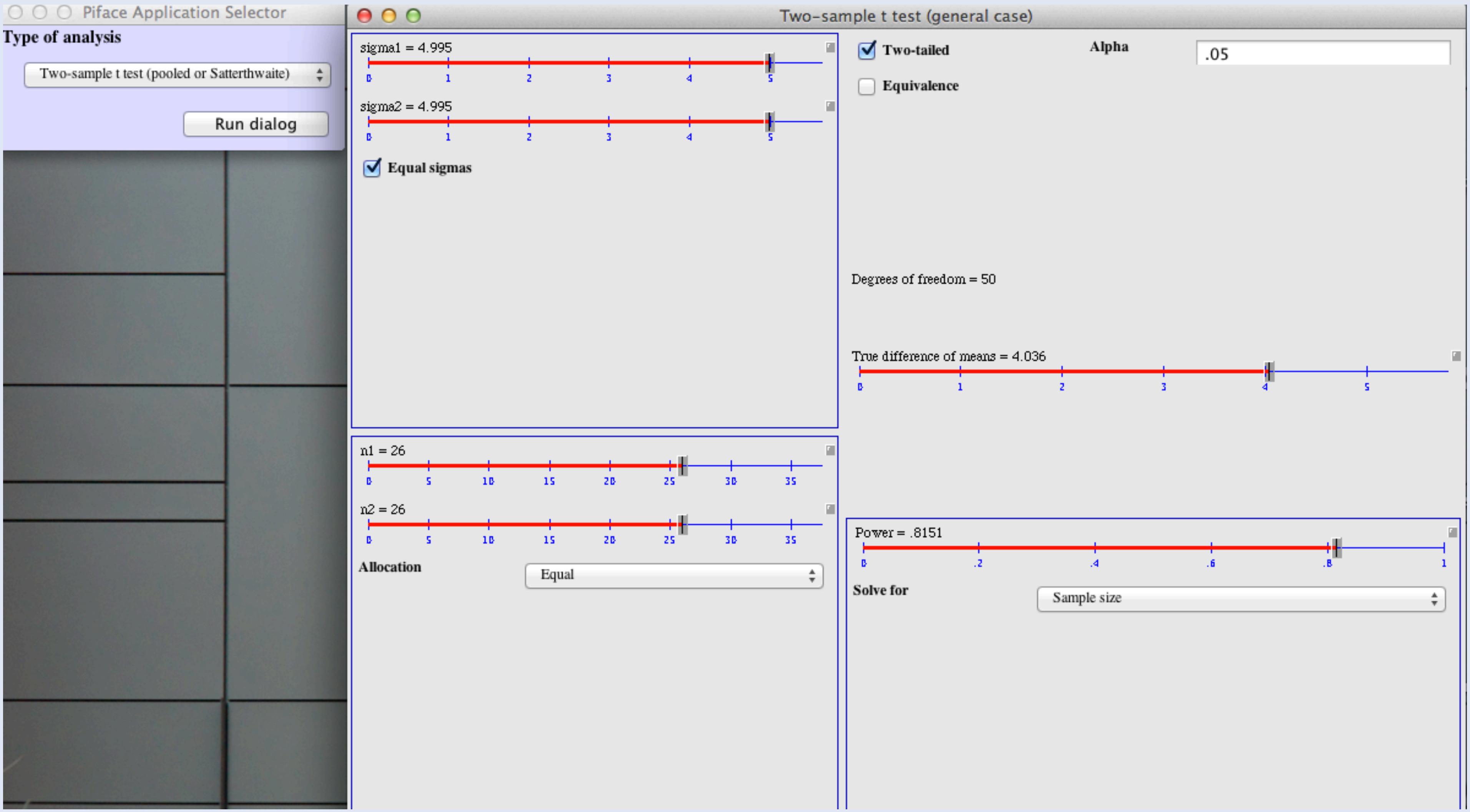
Note. ES = population effect size, Sm = small, Med = medium, Lg = large, diff = difference, ANOVA = analysis of variance. Tests numbered as in Table 1.

[a] Number of groups. [b] Number of independent variables.

Making the Calculation: ROC



5: Making the Calculation: Software



Summary: Sample Size

I receive quite a few questions that start with something like this: "I'm not much of a stats person, but I tried [details...] -- am I doing it right?"

Please compare this with: "I don't know much about heart surgery, but my wife is suffering from ... and I plan to operate ... can you advise me?”

(Russ Lenth, 2006)

Sample Size Calculation

Questions?

Multiple Hypothesis Testing

Multiple Hypothesis Testing

When the boss is away…our researcher decides that he will add the flavor supplement to ALL flavors.

He will test the same hypothesis on each of the flavors.

Gelato Experiment

	Flavor	Test Result ($\alpha=0.05$)
1	Chocolate	Rejected
2	French Vanilla	Rejected
3	Chocolate Chip	Not Rejected
4	Peanut Butter Cup	Not Rejected
5	Raspberry	Not Rejected
6	Banana	Not Rejected
7	Butter Pecan	Not Rejected
8	Cherry	Not Rejected
9	Strawberry	Rejected
10	Mint	Not Rejected
11	Cookie Dough	Not Rejected
12	Neapolitan	Not Rejected
13	Cookies and Cream	Not Rejected
14	M & M	Not Rejected
15	Pistachio	Not Rejected
16	Mint Chocolate Chip	Not Rejected
17	Chocolate Almond	Not Rejected
18	Lemon	Not Rejected
19	Caramel	Not Rejected
20	Coffee	Not Rejected

Quiz

	Flavor	Test Result (α=0.05)
1	Chocolate	Rejected
2	French Vanilla	Rejected
3	Chocolate Chip	Not Rejected
4	Peanut Butter Cup	Not Rejected
5	Raspberry	Not Rejected
6	Banana	Not Rejected
7	Butter Pecan	Not Rejected
8	Cherry	Not Rejected
9	Strawberry	Rejected
10	Mint	Not Rejected
11	Cookie Dough	Not Rejected
12	Neapolitan	Not Rejected
13	Cookies and Cream	Not Rejected
14	M & M	Not Rejected
15	Pistachio	Not Rejected
16	Mint Chocolate Chip	Not Rejected
17	Chocolate Almond	Not Rejected
18	Lemon	Not Rejected
19	Caramel	Not Rejected
20	Coffee	Not Rejected

Who has done this?

What's the problem?

Multiple Hypothesis Tests

Chance of 'Winning': 5/6 (83%)

Chance of an error: 1/6 (17%)

Multiple Hypothesis Tests

Who wants to pull the trigger 6 times?

Multiple Hypothesis Tests

Who wants to pull the trigger 6 times?

$$P(survival) = \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = 33\%$$

Chance of an error is 77%

Multiple Hypothesis Tests

What is the probability of rejecting one or more hypothesis just by chance?

	Flavor	Test Result ($\alpha=0.05$)
1	Chocolate	Rejected
2	French Vanilla	Rejected
3	Chocolate Chip	Not Rejected
4	Peanut Butter Cup	Not Rejected
5	Raspberry	Not Rejected
6	Banana	Not Rejected
7	Butter Pecan	Not Rejected
8	Cherry	Not Rejected
9	Strawberry	Rejected
10	Mint	Not Rejected
11	Cookie Dough	Not Rejected
12	Neapolitan	Not Rejected
13	Cookies and Cream	Not Rejected
14	M & M	Not Rejected
15	Pistachio	Not Rejected
16	Mint Chocolate Chip	Not Rejected
17	Chocolate Almond	Not Rejected
18	Lemon	Not Rejected
19	Caramel	Not Rejected
20	Coffee	Not Rejected

Experiment Wide Risk of Type I Error

$$P(\alpha) = 1 - 0.95^{20} = 0.64$$

There is a 64% of rejecting at least one hypothesis just by chance

Gelato Experiment

Our researcher goes to his boss with paper in hand.

The boss states: “With all those significance tests, there it is likely you committed a type I error”

Gelato Experiment

What can our researcher do?

What Can You Do?

Multiple Hypothesis Testing

MAXIM: Testing multiple hypotheses in a single experiment requires careful planning from the start

Multiple Hypotheses

Methods to correct for multiple hypotheses

1. Bonferroni Correction **
2. Tests for multiple hypotheses:
   - Contingency Tables (χ^2 test)
   - Linear Regression (ANOVA)**
3. Tukey-Kramer
4. Sheffe Method
5. Dunnett's Test
6. Duncan's New Range Test

Bonferroni Correction

Corrects the alpha value for individual tests to control for experiment wide error.

This means it is much harder to reject the null hypothesis!!!

$$p_i < \frac{\alpha}{m}$$

$$p_i < \frac{0.05}{20}$$

$$p_i < 0.0025$$

Gelato Experiment

	Flavor	Test Result ($\alpha=0.05$)	p-value
1	Chocolate	Rejected	0.01
2	French Vanilla	Rejected	0.01
3	Chocolate Chip	Not Rejected	>0.05
4	Peanut Butter Cup	Not Rejected	>0.05
5	Raspberry	Not Rejected	>0.05
6	Banana	Not Rejected	>0.05
7	Butter Pecan	Not Rejected	>0.05
8	Cherry	Not Rejected	>0.05
9	Strawberry	Rejected	0.03
10	Mint	Not Rejected	>0.05
11	Cookie Dough	Not Rejected	>0.05
12	Neapolitan	Not Rejected	>0.05
13	Cookies and Cream	Not Rejected	>0.05
14	M & M	Not Rejected	>0.05
15	Pistachio	Not Rejected	>0.05
16	Mint Chocolate Chip	Not Rejected	>0.05
17	Chocolate Almond	Not Rejected	>0.05
18	Lemon	Not Rejected	>0.05
19	Caramel	Not Rejected	>0.05
20	Coffee	Not Rejected	>0.05

Contingency Tables (Chi Square Test)

```
table(Y$correct,Y$operator);
```

	C1	C10	C2	C3	C4	C5	C6	C7	C8	C9	S1	S10	S2	S3	S4	S5	S6	S7	S8	S9
FALSE	2	3	3	1	3	0	2	2	2	2	1	2	2	1	4	2	1	5	2	5
TRUE	7	6	6	8	6	9	7	7	7	7	8	7	7	8	5	7	8	4	7	4

```
> chisq.test(table(Y$correct,Y$operator));
   Pearson's Chi-squared test
data:  table(Y$correct, Y$operator)
X-squared = 18.8148, df = 19, p-value = 0.4688
```

ANOVA

```
>summary(aov(time~operator,data=Y));

              Df Sum Sq Mean Sq F value Pr(>F)
operator      19 556474   29288   51.14 <2e-16 ***
Residuals    160  91635     573
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1
```

Multiple Hypothesis Testing

Questions?

Quiz: Options

A. Use 30 samples✔
B. Do a power based calculation of sample size✔
C. As big of a sample as possible✔
D. Consult a statistician✔
E. Calculate sample size using a web based service✔
F. Use tables of sample size✔
G. Do a small pilot study and calculate sample size after✔

Quiz: Question 2

Multiple significance tests:

- Consider carefully pros and cons
- Plan methodology carefully at onset
- Expect to sacrifice power

Objectives

1. Understand the principles of deciding on a sample size to reduce Type 1 and Type 2 errors
2. Understand the impact of multiple hypothesis testing on type-1 risk

Math Lesson (Optional)

What is the potential problem with using the "generic" effect sizes.