

# **ESD.86**

# **Hypothesis Testing**

**Dan Frey**

**Associate Professor of Mechanical Engineering and Engineering Systems**



# Gigerenzer's Paper

- "Mindless Statistics"
- Journal of Socio-Economics
- What are its key points?
  - History
  - Research practices
  - Statistics *per se*

# The Null Ritual

1. Set up a statistical null hypothesis of "no mean difference" or "zero correlation." Don't specify predictions of your research hypothesis or any other substantive hypothesis
2. Use a 5% convention for rejecting the null. If significant, accept your hypothesis. Report the result as  $p < 0.005$ ,  $p < 0.01$ , or  $p < 0.001$  (whichever comes next to the obtained p-value)
3. Always perform this procedure

# Is This Really Being Done?

- Quick check – Compendex search for "significant" in abstract and the title of a research journal I read, publish in, review for
- 11 results in 2006, 12 in 2005, etc.

Chusilp, P., and Y. Jin, 2006, "Impact of Mental Iteration on Concept Generation," *ASME J. of Mech. Design* 128:14-25.

Table 7 Correlation matrix of design scores and numbers of iteration

	Best Novelty	Final Novelty	Variety	Quantity	Quality	PR Loop	IS Loop	CR Loop
Best Novelty	1.000							
Final Novelty	0.039	1.000						
Novelty Variety	0.427	-0.456	1.000					
Quantity	0.505 <sup>a</sup>	-0.422	0.663 <sup>a</sup>	1.000				
Quality	0.046	-0.242	0.340	0.465	1.000			
PR Loop	0.160	<b>-0.503<sup>a</sup></b>	<b>0.512<sup>a</sup></b>	<b>0.691<sup>a</sup></b>	<b>0.643<sup>a</sup></b>	1.000		
IS Loop	<b>0.532<sup>a</sup></b>	-0.342	<b>0.553<sup>a</sup></b>	<b>0.839<sup>a</sup></b>	0.572 <sup>a</sup>	0.806 <sup>a</sup>	1.000	
CR Loop	0.249	-0.403	<b>0.587<sup>a</sup></b>	<b>0.781<sup>a</sup></b>	<b>0.725<sup>a</sup></b>	0.935 <sup>a</sup>	0.827 <sup>a</sup>	1.000
Total Loops	0.304	-0.440	<b>0.579<sup>a</sup></b>	<b>0.801<sup>a</sup></b>	<b>0.690<sup>a</sup></b>	0.967 <sup>a</sup>	0.904 <sup>a</sup>	0.978 <sup>a</sup>

The paper's conclusion section states:  
 "The results suggested that (1) increasing number of iteration has positive impact on quality, variety, and quantity, but mixed effect on novelty..."

<sup>a</sup>Denotes significant correlation at  $\alpha=0.05$  (two tailed) Copyright © 2006 by ASME. Used with permission.

# Fisher's Null Hypothesis Testing

1. Set up a statistical null hypothesis. The null need not be a nil hypothesis (i.e., zero difference).
2. Report the exact level of significance ... Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses.
3. Use this procedure only if you know very little about the problem at hand.

# Test for "Goodness of Fit" as Conducted in Weibull's Paper

- Calculates the degrees of freedom  
10 (bins) - 1 - 3 (parameters of the df) = 6
- Calculates the statistic
- States the  $P$ -value
- Comparison to alternative

$$\chi^2 = \sum \frac{(\text{observed} - \text{estimated})^2}{\text{estimated}}$$

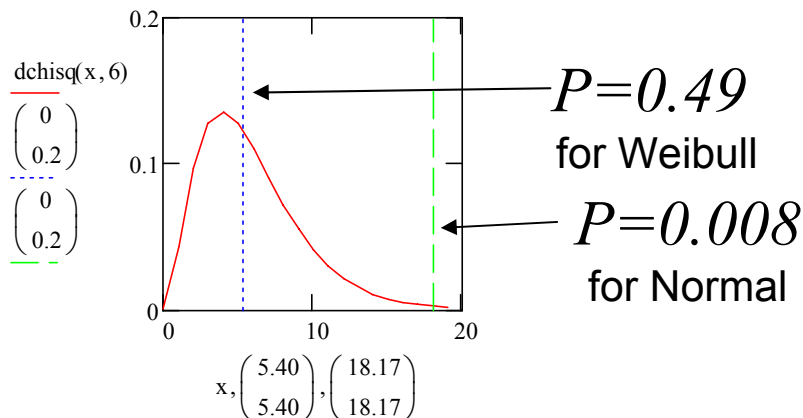


TABLE 1 YIELD STRENGTH OF A BOFORS STEEL  
( $x$  = yield strength in 1.275 kg/mm<sup>2</sup>)

	$x$	Expected values $n$	Observed values $n$	Normal distribution $n$
1	32	10	10	8
2	33	36	33	28
3	34	84	81	71
4	35	150	161	141
5	36	224	224	225
6	37	291	289	301
7	38	340	338	351
8	39	369	369	376
9	40	383	383	386
10	42	389	389	388

Copyright © 2006 by ASME. Used with permission.

Note: Table is cumulative,  
 $\chi^2$  test requires frequency in bin

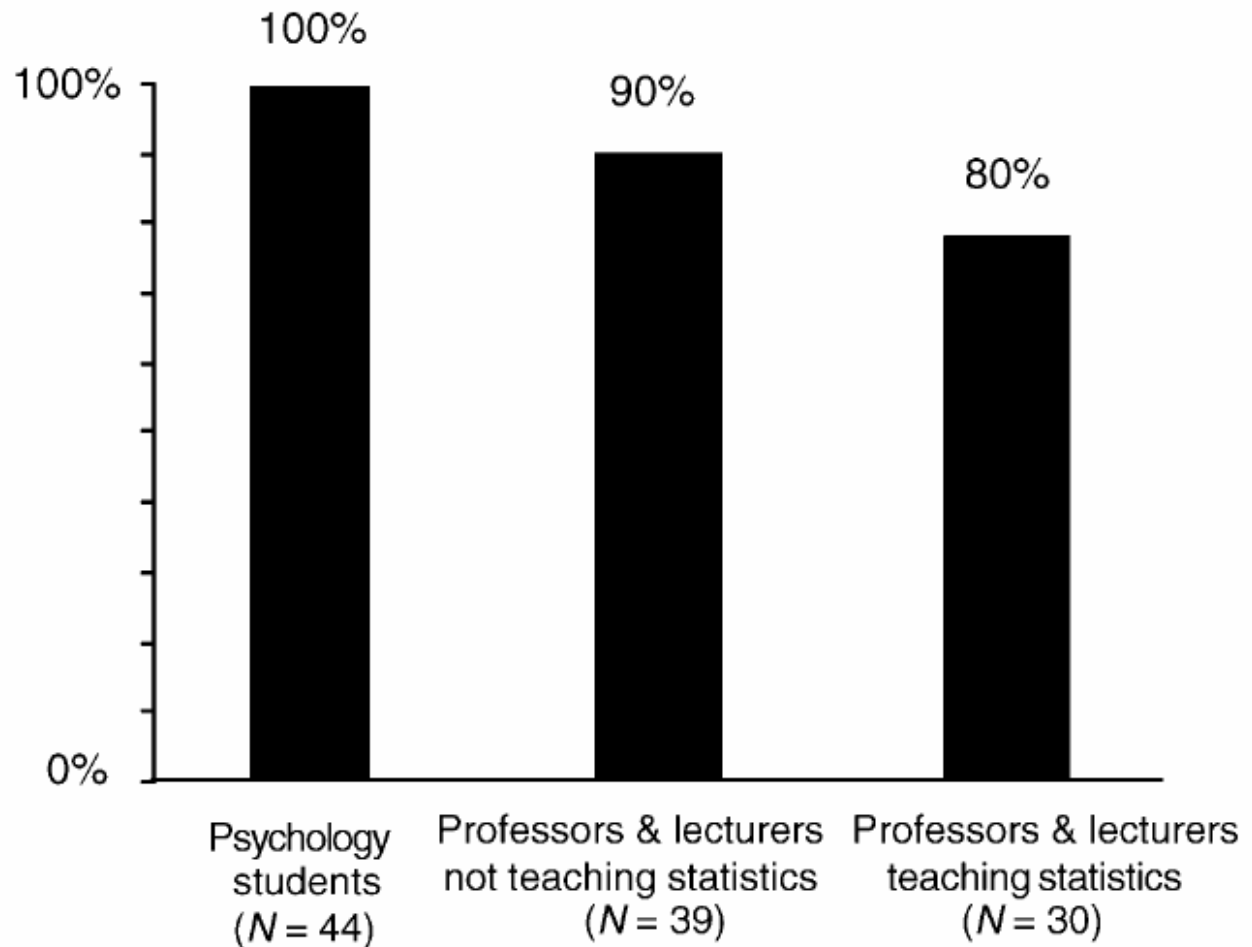
# Gigernezer's Quiz

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” ...

1. You have absolutely disproved the null hypothesis
2. You have found the probability of the null hypothesis being true.
3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
4. You can deduce the probability of the experimental hypothesis being true.
5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

# Quiz Results

The percentages of participants in each group who endorsed one or more of the six false statements regarding the meaning of “ $p = 0.01$ .”





# Matlab's Description of the Two-sample $t$ -test

`h = ttest2(x,y)` performs a  $t$ -test of the null hypothesis that data in the vectors `x` and `y` are independent random samples from normal distributions with equal means and equal but unknown variances, against the alternative that the means are not equal. The result of the test is returned in `h`. `h = 1` indicates a rejection of the null hypothesis at the 5% significance level. `h = 0` indicates a failure to reject the null hypothesis at the 5% significance level.

# Concept Question

- This Matlab code repeatedly generates and tests simulated "data"
- 20 "subjects" in the control and treatment groups
- Both normally distributed with the same mean
- How often will the  $t$ -test reject  $H_0$  ( $\alpha=0.01$ )?

```
for i=1:1000
    control=random('Normal',0,1,1,20);
    trt=random('Normal',0,1,1,20);
    reject_null(i) = ttest2(control,trt,0.01);
end
mean(reject_null)
```

- 1) ~99% of the time
- 2) ~1% of the time
- 3) ~50% of the time
- 4) None of the above

# Concept Question

- This Matlab code repeatedly generates and tests simulated "data"
- 20 "subjects" in the control and treatment groups
- Both normally distributed with the different means
- How often will the  $t$ -test reject  $H_0$  ( $\alpha=0.01$ )?

```
for i=1:1000
    control=random('Normal',0,1,1,200);
    trt=      random('Normal',1,1,1,200);
    reject_null(i) = ttest2(control,trt,0.01);
end
mean(reject_null)
```

- 1) ~99% of the time
- 2) ~1% of the time
- 3) ~50% of the time
- 4) None of the above

# Concept Question

- How do “effect” and “alpha” affect the rate at which the  $t$ -test rejects  $H_0$  ?

```
effect=1;alpha=0.01;  
for i=1:1000  
    control=random('Normal',0,1,1,20);  
    trt=      random('Normal',effect,1,1,20);  
    reject_null(i) = ttest2(control,trt,alpha);  
end  
mean(reject_null)
```

- a) ↑ effect, ↑ rejects
- b) ↑ effect, ↓ rejects
- c) ↑ alpha, ↑ rejects
- d) ↑ alpha, ↓ rejects

- 1) a & c
- 2) a & d
- 3) b & c
- 4) b & d

# What if Assumptions are Violated?

- This Matlab code generates data with a no treatment effect on mean
- But dispersion is affected by treatment
- Type I error rate rises

```
for i=1:1000
    control=random('Normal',0,1,1,20);
    trt=random('Normal',0,2,1,20);
    reject_null(i) = ttest2(control,trt,0.01);
end
mean(reject_null)
```

# What if Assumptions are Violated?

- This Matlab code generates data with a no treatment effect on mean
- The two populations are uniformly dist.
- Type I error rate rises

```
for i=1:1000
    control=random('Uniform',0,1,1,20);
    trt=random('Uniform',0,1,1,20);
    reject_null(i) = ttest2(control,trt,0.01);
end
mean(reject_null)
```

# Model Adequacy Checking

- Normality

- normal probability plot of residuals

- Constant variance

- plot residuals versus time sequence

- plot residuals versus treatments or fitted values

- Bartlett's Test    [ndim, prob] = barttest(x,0.05)

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_a^2$$

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ for at least one } i, j$$

# Randomization Distribution

- The null hypothesis implies that the observations are not a function of the treatments
- If that were true, the allocation of the data to treatments (rows) shouldn't affect the test statistic
- How likely is the statistic observed under re-ordering?

Cotton weight percentage	Observations				
	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11



# Randomization Distribution


- This code reorders the data from the cotton experiment
- Does the ANOVA
- Repeats 5000 times
- Plots the pdf of the F ratio

```
trials=5000;bins=trials/100;  
X=[7 7 15 11 9 12 17 12 18 18 14 18 18  
19 19 19 25 22 19 23 7 10 11 15 11];  
group=ceil([1:25]/5);  
[p,table,stats] = anova1(X, group,'off');  
F=cell2mat(table(2,5))  
  
for i=1:trials  
    r=rand(1,25);  
    [B,INDEX] = sort(r);  
    Xr(1:25)=X(INDEX);  
    [p,table,stats] = anova1(Xr, group,'off');  
    Fratio(i)=cell2mat(table(2,5));  
end  
  
hold off  
[n,x] = hist(Fratio,bins);  
n=n/(trials*(x(2)-x(1)));  
colormap hsv  
bar(x,n)  
hold on  
  
xmax=max(Fratio);  
x=0:(xmax/100):xmax;  
y = fpdf(x,4,20);  
plot(x,y,'LineWidth',2)
```

# Neyman-Pearson "Decision Theory"

1. Set up two statistical hypotheses,  $H_1$  and  $H_2$ , and decide about  $\alpha$ ,  $\beta$ , and sample size before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.
2. If the data falls into the rejection region of  $H_1$ , accept  $H_2$ ; otherwise accept  $H_1$ . Note that accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.
3. The usefulness of the procedure is limited among others to situations where you have a disjunction of hypotheses ... and where you can make meaningful cost-benefit trade-offs for choosing  $\alpha$  and  $\beta$ .

# NP Framework and Two Types of Error

- Set a critical value  $c$  of a test statistic  $T$  or else set the desired confidence level or "size"  $\alpha$
  - Observe data  $X$
  - Reject  $H_1$  if the test statistic  $T(X) \geq c$
  - Probability of Type I Error – The probability of  $T(X) < c \mid H_1$ 
    - (i.e. the probability of rejecting  $H_1$  given  $H_1$  is true)
  - Probability of Type II Error – The probability of  $T(X) \geq c \mid H_2$ 
    - (i.e. the probability of not rejecting  $H_1$  given  $H_2$  is true)
  - The *power* of a test is 1 - probability of Type II Error
  - In the N-P framework, power is maximized subject to Type I error being set to a fixed critical value  $c$  or of  $\alpha$
- or other confidence region (e.g. for "two-tailed" tests)
- 

# The Neyman-Pearson Lemma

- When performing a hypothesis test between two point hypotheses  $H_0$  and  $H_1$ , the most powerful test of size  $\alpha$  is

$$\Lambda(x) = \frac{L(\theta_0|x)}{L(\theta_1|x)} \leq c \text{ where } \Pr(\Lambda(X) \leq c | H_0) = \alpha$$

# A Modern View on the N-P Approach

"The Neyman Pearson approach rests on the idea that, of the two errors, one can be thought of as more important. By convention this is chosen to be the type I error ... In the medical setting, this asymmetry appears reasonable... It has also been argued that, generally in science, announcing a new phenomenon has been observed when in fact nothing has happened is more serious than missing something new that has in fact occurred. We do not find this persuasive..."

# Fisher's Objections to the Neyman-Pearson Framework

- The following concepts apply in acceptance sampling but generally are not meaningful in scientific investigations
  - The distinction between type I and II error
  - The notion of repeated sampling from the same population
  - Inductive behaviour (GG on NP "...accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.")

Fisher, R. A., 1955, "Statistical Methods and Scientific Induction,"  
*Journal of the Royal Statistical Society (B)* **17**:69-77.

# Fisher's Suggestion of How to Think about Results of Hypothesis Tests

"...whenever a test of significance gives us no strong reason for rejecting it (the null hypothesis)... the worker's real attitude in such a case might be, according to the circumstances:"

- a) "The possible deviation from truth of my ...(null) hypothesis ... seems not to be of sufficient magnitude to warrant any immediate modification."
- b) "The deviation is in the direction expected for certain influences... and to this extent my suspicion has been confirmed; but the body of data available so far is not by itself sufficient to demonstrate their reality."

Fisher, R. A., 1955, "Statistical Methods and Scientific Induction," *Journal of the Royal Statistical Society (B)* **17**:69-77.

# Fisher on Smoking

- ~1950 a study at the London School of Hygiene states that smoking is an important cause of lung cancer
- Fisher writes
  - “...an error has been made of an old kind, in arguing from correlation to causation”
  - “For my part, I think it is more likely that a common cause supplies the explanation”

R. A. Fisher, 1958, *The Centennial Review*, vol. II, no. 2, pp. 151-166.

R. A. Fisher, 1958, Letter to the Editor of *Nature*, vol. 182, p. 596.



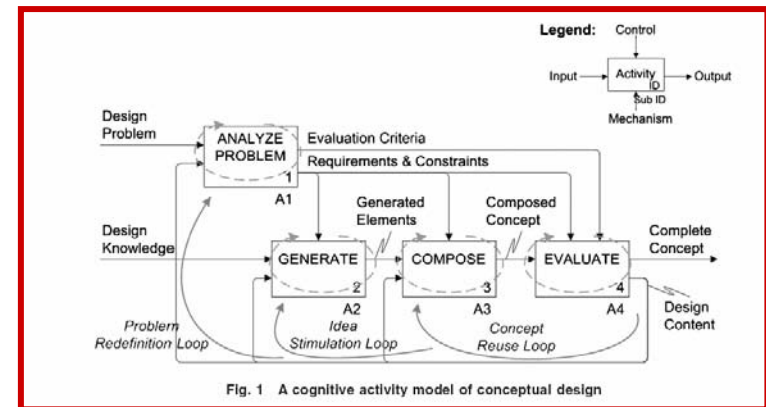
# Similar Issues in Research Today

- Correlation demonstrated in the Table
- Causal link implied in the conclusion
- What was the design of the experiment?
- What mechanisms were proposed?

Table 7 Correlation matrix of design scores and numbers of iteration

	Best Novelty	Final Novelty	Variety	Quantity	Quality	PR Loop	IS Loop	CR Loop	Total Loops
Best Novelty	1.000								
Final Novelty	0.039	1.000							
Novelty Variety	0.427	-0.456	1.000						
Quantity	0.505 <sup>a</sup>	-0.422	0.663 <sup>a</sup>	1.000					
Quality	0.046	-0.242	0.340	0.465	1.000				
PR Loop	0.160	-0.503 <sup>a</sup>	0.512 <sup>a</sup>	0.691 <sup>a</sup>	0.643 <sup>a</sup>	1.000			
IS Loop	0.532 <sup>a</sup>	-0.342	0.553 <sup>a</sup>	0.839 <sup>a</sup>	0.572 <sup>a</sup>	0.806 <sup>a</sup>	1.000		
CR Loop	0.249	-0.403	0.587 <sup>a</sup>	0.781 <sup>a</sup>	0.725 <sup>a</sup>	0.935 <sup>a</sup>	0.827 <sup>a</sup>	1.000	
Total Loops	0.304	-0.440	0.579 <sup>a</sup>	0.801 <sup>a</sup>	0.690 <sup>a</sup>	0.967 <sup>a</sup>	0.904 <sup>a</sup>	0.978 <sup>a</sup>	1.000

<sup>a</sup>Denotes significant correlation at  $\alpha=0.05$  (two tailed)



Copyright © 2006 by ASME. Used with permission.

The paper's conclusion section states:

"The results suggested that (1) increasing number of iteration has positive impact on quality, variety, and quantity, but mixed effect on novelty..."

# More of Gigerenzer's Concerns

- Meehl's Conjecture
  - In non-experimental settings with large sample sizes, the probability of rejecting the null hypothesis of nil group differences in favor of a directional alternative is about 0.50.
  - Statistical significance  $\neq$  practical significance
- Feynman's Conjecture
  - To report a significant result and reject the null in favor of an alternative hypothesis is meaningless unless the alternative hypothesis has been stated before the data were obtained.
  - (My opinion) Yes, such a result requires independent confirmation, but you should still report it!

# Simpon's Paradox

"Table 1" removed due to copyright restrictions.

It is decisively rejected that male and female applicant had the same probability of being accepted to Berkeley in 1973.

BUT The probability of finding that a department that was biased against women ... is about 57 times in 1000.

Bickel, P. J. , et al., 1975, "Sex Bias in Graduate Admissions: Data from Berkeley," *Science* 187.(4175):398 – 404.

# Simpon's Paradox

Correcting for the tendency of women to apply to graduate departments that are more difficult for applicants of either sex to enter, there is a small but statistically significant bias in favor of women.

**BUT why the tendency?**

Graph removed due to copyright restrictions.

# Type-III error

- "At issue here is the importance of good descriptive and exploratory statistics rather than mechanical hypothesis testing with yes-no answers...**The attempt to give an "optimal" answer to the wrong question** has been called "Type-III error". The statistician John Tukey (e.g., 1969) argued for a change in perspective..."

# Problem Set #5

## 1. Parameter estimation

Make a probability plot

Make an estimate by regression

Make an MLE estimate

Estimate yet another way

Comment on "goodness of fit"

## 2. Hypothesis testing

Find a journal paper using the "null ritual"

Suggest improvements (validity, insight, communication)

# Next Steps

- Between now and next Monday
  - Read Tufte "Visual and Statistical Thinking: Displays of Evidence for Making Decisions"
- Friday, 6 April
  - Recitation to support PS#5
- Monday, 9 April
  - Session on descriptive statistics and graphics
  - PS#5 due, NO NEW HW ASSIGNED
- Wednesday, 11 April
  - Session on Regression (no pre-read)
- Friday, 13 April
  - Session to support the term project
  - Be prepared to stand up and talk for 5 minutes about your ideas and your progress