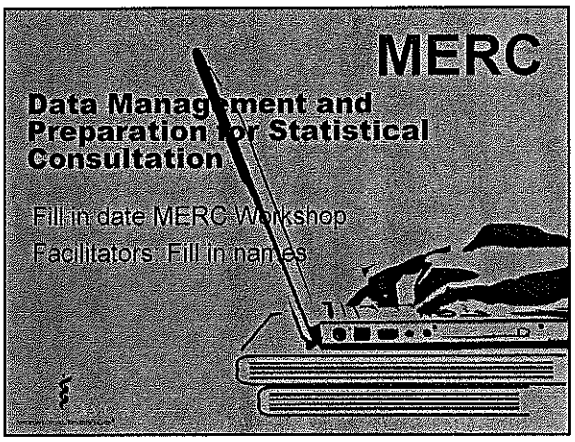


Development Team: Patricia O'Sullivan, Ann Frye, Judy Shea

Last reviewed date: May 2017

Reviewed by: Judy Shea

How Do I...

Prepare for my statistical consultation?

Collect data?

Set up data files?

Enter data?

Check and clean data?

Return to my consultant?

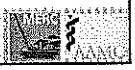
And of course.....

You should consult with a statistician

or research design specialist

while you are *conceptualizing*

the study!



Consulting with a statistician

About what??????

What do you need to ask about?

Addressing statistical issues

Comparing sample to population

The Statistical Consultation

Addressing Statistical Issues

The Statistical Consultation
Addressing Statistical Issues

- Statistical significance
- Type I error
- Type II error
- Power
- Effect size
- Appropriate analyses

Statistical Significance

- What level of uncertainty are you willing to tolerate in making statistical decisions?
- What degree of risk are you willing to take of rejecting a null hypothesis that is true?

  Called "level of significance" or α

Type I Error

Making an incorrect decision = a *type I error*

(rejecting a null hypothesis that is true)

Set value of α

- 0.05 typical
- 0.01 more conservative
- 0.10 more liberal

Type II Error

- Failing to reject a null hypothesis that is false
- Protecting your study against a Type II error
  - Increase the sample size
  - Look at a *meaningful* effect
  - Increase study's alpha

Power

- Power: the probability of identifying a true hypothesis as true (given the structure of your study)
- Power = (1- Type II error)
- Design studies to have a power of 0.8

Maximizing Power

- Design a meaningful intervention
- Increase sample size
- Limit number of variables in the study

Statistical Significance is Not Everything

Large samples
Very small changes or correlations can be statistically significant but essentially without meaning.

Small samples
Important (meaningful) changes or correlations may not be statistically significant.

Statistical Significance

- How important is statistical significance?
- How do you decide if a finding is significant?
- How do you take chance into account in statistical decision making?
- Can something be statistically significant but not meaningful?
- Can statistics solve all problems?

Effect Size Primer

Effect Size (ES)

- A family of indices that measure the treatment effect's magnitude
- Independent of sample size

Effect Size Primer

Two Simple Measures of Effect Size

- The standardized difference between two means
  Cohen's *d* = [(M1 - M2) / pooled std dev]
- The correlation coefficient
- Rough interpretation guidelines
  - 0.2 - small
  - 0.5 - medium
  - 0.8 - large

The Statistical Consultation

Comparing Sample to Population

Who is in your sample?

- Is the sample representative of the population?
  - Single sample test of proportions
    - Percent female representative
    - Percent African American
  - Single sample test of means
    - Sample MCATs equal to population
    - Sample age equal to population
- Expect results to be statistically non-significant

...and who is not?

- Follow the non-respondents
- Compare respondents to non-respondents
  - Same age (t-test)
  - Same genders (chi-square)
  - Same geographic region (chi-square)

And a Reminder!

When you consult with a statistician while you are *conceptualizing* the study, talk about appropriate analyses for your study data.

Study 1: Teaching Method Comparison

After midterm, students randomly assigned to

- Classroom lecture plus slides
- Team learning
- CAI

Measures

- Midterm exam
- Final exam

Studies 2 & 3: What Kind of Analysis?

Data available: MCAT total scores, gender, ethnicity, college GPA, % grades for M1 and M2 years, USMLE Step 1 scores

- Study 2: What predicts USMLE score?
- Study 3: What distinguishes students who drop out of medical school from those who do not?

Collect Data

Collect Data
Paper Instruments

Advantages
- Good control on distribution and return of instruments
- Responders can indicate where they are unclear
- Anonymity
- Can do double data entry for data verification

Disadvantages
- Entering data takes time and chance for error
- Forms need to be hand reviewed prior to entry
- Comments take time to enter
- Entry expensive
- Anonymity

Web-Based Forms

Advantages
- No data entry
- Easy distribution
- Easy completion

Disadvantages
- Physical appearance has limited appeal
- Requires internet access
- Data enterer cannot be verified
- May be hard to correct entry errors
- Still takes file preparation for analysis

Scannable Forms

Advantages
- Efficient data entry
- Forces forms to be clean and clear
- Can be produced easily
- Software can be relatively inexpensive

Disadvantages
- Forms must be reviewed prior to scanning
- Digital scan requires error checking
- Need to review for scanning problems
- Software can be expensive

Direct Entry

Direct data entry into a computer by someone other than the respondent

- Phone survey
- Consumer survey
- Electronic health record

Set up Data Files

Variable Names

Use short, meaningful name

Add a variable label with more information

Include

- An ID variable (even for anonymous data)
- Grouping variable (if applicable)

Data Naming Example – Study 4

You ask residents in a 3-year training program to

1) rate their understanding of 7 categories of complementary or alternative medicine (CAM)

2) indicate their gender and age

3) write how they describe their ethnicity

- How many variables are there?
- What are their names?

Variable Formats

Variable formats

- Numeric
- String

From previous example, which could be

- numeric?
- string?

Variable Values

- Indicates code for possible responses for a variable:
  - 0 = no
  - 1 = yes
- Missing data codes

How would you code the previous example?

Develop a Codebook

- Variable names, labels and values are listed in a codebook
- Develop codebook before data collection
  - Can be done directly on data collection form
- Codebooks can be generated by computer programs

Exercise - Practice Codebook Development

- Using the CAM survey, develop a codebook for the data
- Transfer onto codebook template for your records
  - Variable name
  - Type – string or numeric
  - Label
  - Values – possible values and their meaning

Codebook Template

Variable Name	Type	Label	Values

Enter Data

Software Options

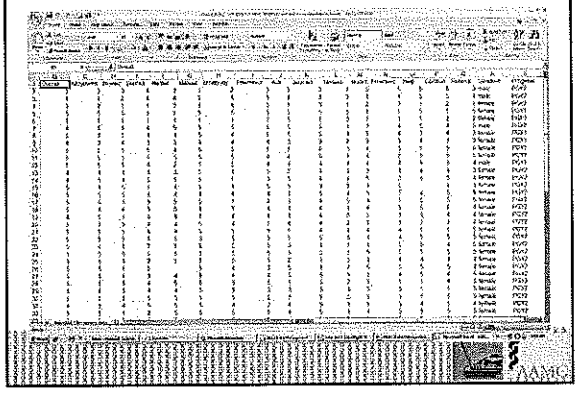
Data entry
- Excel
- Access
- Directly into a statistical package

Data analysis
- Major, common packages: SPSS, SAS
- Less well known: STATA, R

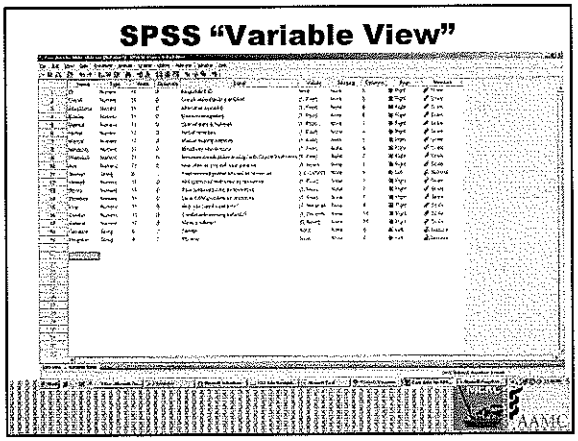
Data Entry

- One line per subject
- Build in data range restrictions
- Double data entry

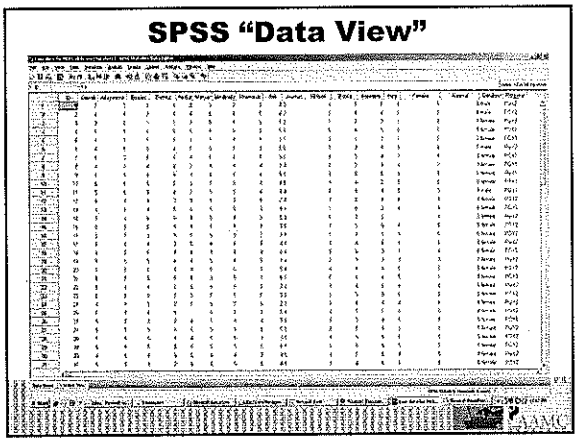
Excel Worksheet



SPSS "Variable View"



SPSS "Data View"



Merging Data

You can bring together data from two files.

- One variable must have the same name and format in both files.
- Files must be sorted in order by that variable.

Check and Clean Data

Cleaning the Data

- Compute frequencies for each variable
  - Check maximum and minimum values to make sure none are out of range
- Clean/fix missing values
- Compute descriptive statistics for each variable
  - Make sure means and standard deviations make sense

Understanding the Data

For continuously scaled variables

- Construct histograms
- Check skewness

Understanding the Data

For bivariate relationships

- Construct scatterplots
- Examine for bivariate outliers
- Examine for linearity/nonlinearity

Practice Looking at Data

- Review descriptive statistics handout
- Review scatterplots

Example – Study 5

This study had, among others, two variables whose relationship was of interest:

- Scale score composed of combined survey items labeled "understanding others' value"
- A post-intervention score of knowledge labeled "post score"

What to do
- Enter data
- Run frequencies
- Look at distributions
- Start to look at inter-relationships of variables

Statistics from Study 5

Descriptive Statistics

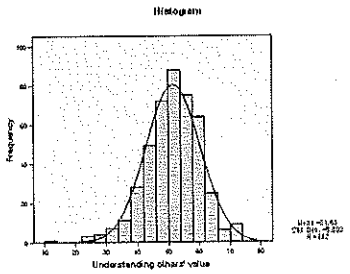
	N	Minimum	Maximum	Mean	Std	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Understanding others' value	442	12	72	61.63	8.802	-.497	.116
Post score	37	7	36	16.19	4.932	1.723	.388
Valid N (listwise)	36						

Statistics from Study 5

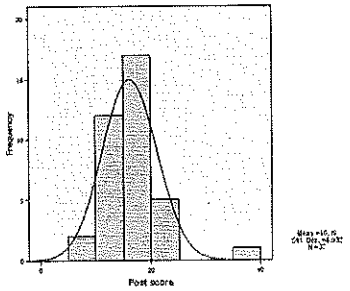
Post score

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7	1	.2	2.7	2.7
	9	1	.2	2.7	5.4
	10	1	.2	2.7	8.1
	12	3	.7	8.1	16.2
	13	4	.9	10.8	27.0
	14	4	.9	10.8	37.8
	15	4	.9	10.8	48.6
	16	4	.9	10.8	59.5
	17	3	.7	8.1	67.6
	18	5	1.1	13.5	81.1
	19	1	.2	2.7	83.8
	20	1	.2	2.7	86.5
	21	1	.2	2.7	89.2
	22	1	.2	2.7	91.9
	23	2	.4	5.4	97.3
	36	1	.2	2.7	100.0
	Total	37	8.2	100.0	
Missing	System	412	91.8		
Total		449	100.0		

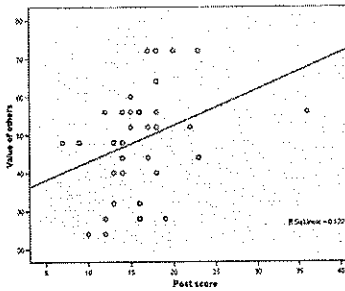
Graph from Study 5



Graph from Study 5



Graph from Study 5



Now it is your turn

Example 6 - Study set-up

This study was a randomized trial of Oral Case Presentation feedback. About 160 students were randomized to 2 groups – usual feedback and structured intentional feedback.

1. Medical students randomized to the treatment (about 20 a block) were given OCP booklets and asked to gather 1 feedback form a week
2. Data for each student were linked to other data, (e.g., ratings of clinical competence, shelf exam scores)

DO STUDENTS WHO GET STRUCTURED FEEDBACK DO BETTER ON END-OF-CLERKSHIP PRESENTATION?

Your job

1. Make a code book
2. Comb through the data

What to do
- Enter data
- **Run frequencies**
- **Look at distributions**
- **Link files**
- **Start to look at inter-relationships of variables**

Look at the data

Check N's

		BLOCK	Student Number	Student Name	Evaluator Type	DATE	Setting	Patient Type
N	Valid	568	568	568	551	498	551	328
	Missing	0	0	0	17	70	17	240

BLOCK

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	138	24.3	24.3	24.3
	2	161	28.3	28.3	52.6
	3	135	23.8	23.8	76.4
	4	134	23.6	23.6	100.0
	Total	568	100.0	100.0	

20 students a block 20 * 4 = 80
each asked to do 9 cards 80 * 9 = 720 = 79%
Do we understand why block 2 is higher?

What values do we see?

BLOCK

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	138	24.3	24.3	24.3
	2	161	28.3	28.3	52.6
	3	135	23.8	23.8	76.4
	4	134	23.6	23.6	100.0
	Total	568	100.0	100.0	

Setting

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ambulatory	193	29.0	25.0	25.0
	Inpatient	380	66.9	75.0	100.0
	Total	551	97.0	100.0	
Missing	99	13	2.5		
	System	2	.4		
	Total	17	3.0		
Total		568	100.0		

What values do we see?

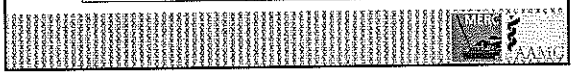
Patient Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	new	238	41.9	72.6	72.6
	follow-up	86	15.7	27.1	99.7
	5	1	.2	.3	100.0
	Total	328	57.7	100.0	
Missing	99	238	41.9		
	System	2	.4		
	Total	240	42.3		
Total		568	100.0		

What values do we see?

OCP_Ht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4	1	.2	.2	.2
	5	6	1.1	1.1	1.3
	6	34	6.0	6.1	7.3
	7	173	30.5	30.9	38.3
	7.5	1	.2	.2	38.5
	8	264	35.9	36.6	75.0
	8.5	1	.2	.2	75.1
	9	139	24.5	24.8	100.0
	Total	560	98.4	100.0	
Missing	0	7	1.2		
	System	2	.4		
	Total	9	1.6		
Total		569	100.0		



What do we observe?

Time to Observe

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	3	.5	.5	.5
	3	3	.5	.5	1.1
	4	2	.4	.4	1.4
	5	41	7.2	6.8	8.1
	60	2	.4	.4	70.8
	90	125	22.7	23.8	89.8
	120	2	.4	.4	100.0
	Total	560	98.6	100.0	
Missing	System	2	.4		
Total		569	100.0		

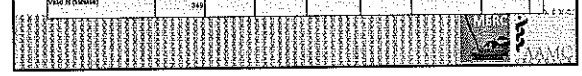
CHECK

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	560	100.0	100.0	100.0

Move on to distributions...

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Skewness
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
OCP_Ht	560	5.00	4.00	9.00	7.7732	.93894	-.424
OCP_PrelMedHX	545	5.00	4.00	9.00	7.7345	.92634	-.340
OCP_SocFam RK	560	5.00	4.00	9.00	7.7232	.95712	-.424
OCP_PE	567	4.00	5.00	9.00	7.6773	.84378	-.275
OCP_Studies	427	5.00	4.00	9.00	7.6370	.99237	-.341
OCP_AssessSkills	552	5.00	4.00	9.00	7.6187	1.03166	-.354
OCP_Plan	560	5.00	4.00	9.00	7.5696	.96561	-.391
OCP_Organization	562	5.00	4.00	9.00	7.6334	.99019	-.384
OCP_Oral Speak Ability	561	15.00	5.00	20.00	7.9800	1.07326	2.073
Valid N (listwise)	349						

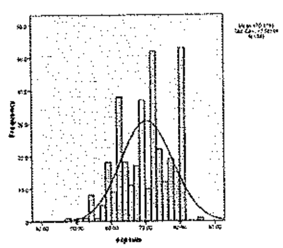


Start to put all of the items together..

	N	%
Valid	349	61.4
Excluded	219	38.6
Total	568	100.0

Cronbach's Alpha	N of Items
.951	9

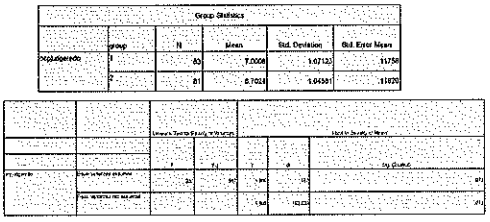
And then we look at distributions..



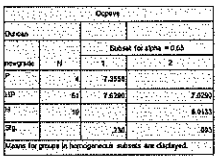
And then we aggregate to other scores, etc.

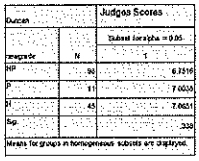
Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
summaryinpt	164	4.57	8.65	6.0877	.45016
presinpt	164	3.75	7.00	6.0649	.53142
summaryoutpt	160	4.43	7.00	6.0158	.55491
presoutpt	160	3.00	7.00	6.0244	.76016
nbdf	160	62.00	96.00	78.1875	6.56543
ceptime	77	6.00	46.25	16.6077	7.49537
evaltime	78	1.60	14.50	7.0633	2.65946
evaltot	77	4.25	9.00	7.3610	1.09250
cardq1	76	6.50	9.00	7.7609	.46042
cardq2	77	6.00	9.00	7.7167	.48431
cardq3	78	5.00	9.00	7.7313	.58880
cardq4	78	6.00	8.00	7.6565	.55453
cardq5	77	6.33	9.00	7.6681	.55411
cardq6	78	4.00	8.83	7.5402	.63633
cardq7	76	4.00	8.67	7.5378	.64420
cardq8	78	6.00	8.78	7.6136	.55179
cardq9	78	7.00	9.00	7.5424	.49842
Valid N (listwise)	70				

Bottom line – is there a difference?



What do scores relate to?





Summary

Check n's – for each analyses

Check all values used – be sure you understand and expect each one

Get to know your data

We've reviewed how to...

- Prepare for your statistical consultation
  - Compare sample to population
  - Address statistical issues
- Collect data
- Set up data files
- Enter data
- Check and clean data

Returning to the Consultant

- What understanding do the researcher and consultant have about their roles with analysis, writing about results, and authorship?
- Recommended (at the very least): a visit to make sure analyses and interpretations are correct!

Questions?

MERC Evaluation Link

Please go to the link below and complete the evaluation:

http://goo.gl/8pwl6B