CRIJ 5331

Chapter 8:

Samples and Statistical Inferences

Quantitative Methods in Criminology

Pearson, Chapter 8 PPT

AN EXAMPLE

Imaging that you are given a cup that contains a sample of n=20, drawn from a population 1,440 beans, which were divided equally between white and red (720 beans in each of these two groups)

We drew a sample of 20 beans for every student (note that we typically analyze only one sample drawn from a population in real research)

Will everyone have exactly 10 white and 10 red beans in their sample?

CONTINUED

- If we "randomly" (in a simple way to define it, it means each individual in a population should have the equal chance to be drawn for the research sample.) drawn samples will produce different results as a result of chance or "the luck of the draw"
- The mean of white beans in your sample may not be exactly the same as the mean (average) in the population from which your sample was drawn
- In other words, sampling may not perfectly get a perfect ratio for each subgroup (by variety of criteria) in the population. We will say that the sample should represent the population while be aware of some estimated error ranges the sample may carry from the statistics of the population. This is the basic concept of statistical inference.

CONTINUED

Because of random selection, however, we can employ mathematical laws of probability to

- calculate the chances that the true value in the population resembles the value we measured in our sample

i.e., how **close** is our estimate to the "true" value and how **confident** in this judgment can we be

e.g., President Obama's job approval rating is 56% plus or minus 3% (how close, but not how confident)

ADDITIONAL WONDERS OF RANDOM SAMPLES

(1) if you sampled the same population many, many times, your results (e.g., the mean of white bean means) would be distributed normally. This is called the sampling distribution.

(2) the larger your sample, the greater your chances that your estimate will come close to the true value in the population from which you sampled

(3) The "real" population mean will fall within + or - one s.e. of your (one cup) sample estimate of the mean in 68% of these repeated samples (95% will fall within + or - 2 s.e.)

IN OTHER WORDS, ANY ESTIMATE FROM A RANDOM SAMPLE HAS A CONFIDENCE INTERVAL

What is it?

- The range within which you can be confident to find the true population parameter in, say, 95 of 100 repeated samples of same size and design from the population

To calculate a 95% c.i.: $\bar{x} \pm 1.96\ s.d./\sqrt{n}$

This is the standard error (s.e.)

CONTINUED

- A random sample won't always produce the "true" values of the population parameters you want to estimate.
- No problem! We have a magic machine under the hood that tells us through the use of the s.e. just how close and confident we can be that the statistic we calculated from a single sample will be to the population parameter we're estimating.
- Whew! Do you get it?

THE STANDARD ERROR (S.E.)

The standard error is to the sampling distribution as the standard deviation is to an estimate from a single sample

They are mathematically related:

$$s.e. = \frac{s.d.}{\sqrt{n}}$$

The distinction between s.d. and s.e. is really key to understanding statistical inference. Make sure if you can tell the difference (in a way you can understand).

Randomization, Response Bias, and Sampling Bias: A Famous Example

Literary Digest mailed 10 million surveys in 1936 and asked whether respondents planned to vote for Alfred Landon or FDR for president.

Received 2.4 million replies

Predicted results:

Landon would receive 52% of popular vote and 370 vs. 161 electoral votes for FDR

THE 1936 ELECTION RESULTS

		Electoral College Votes	
		Landon	FDR
Literary Digest	Prediction	370	161
Actual results		8	523

What went wrong here?

1936 LITERARY DIGEST SURVEY

- Response bias: 24% completion rate
- Sampling bias:
  - Lists to whom surveys were sent included:
    - Subscribers to Literary Digest Magazine
    - Car registration records
    - Telephone owners
    - **Therefore, not everyone has equal chance to be selected for survey**
  - The probabilities of poorer folks being in the sample were much less than wealthy ones and socioeconomic class was highly related to the predicted behavior (vote for FDR)

IN SUM

- Our purpose in drawing samples is to make inferences about the populations from which our samples are drawn
- Random selection, however, will produce varying estimates from one sample to another
- But these multiple estimates (if we draw multiple samples) themselves take on the shape of a normal distribution, which we can use to determine how close any single sample's estimate is to the population parameter and with what confidence

BUT STATS ARE ABOUT MORE THAN ESTIMATING A POPULATION MEAN

- It's also about assessing whether two or more groups differ or whether two are more variables are related.
- If based on a sample, these estimates (e.g., a correlation, regression coefficient) too will vary
- We, therefore, want to know whether a relationship we find in our data is a chance event resulting from the luck of the draw.

THE "NULL HYPOTHESIS" AND "STATISTICAL SIGNIFICANCE" WERE INVENTED TO HELP US IN THAT QUEST

R. A. Fisher created these "regrettable" tests in 1932

1. Rejecting the null hypothesis:
   - A formally stated proposition that your statistical test seeks to reject
   - Usually stated that there is no relationship between x and y or no difference in the means of two or more groups (the "nil" hypothesis)
   - You ask whether the evidence you have in a specific sample is sufficiently strong to reject "no relationship" and, therefore, conclude that indeed there is one
2. If you do, Fisher said your relationship was statistically significant (at some level of confidence of being right or wrong).

STATED AGAIN

- Like the different proportion of white beans in a cup (sample) mentioned earlier, tests of statistical significance determine (at some level of confidence) whether group differences or relationships you find are "real" or simply an artifact of the "luck of the draw," i.e., the result of a random sampling process that produced your data.

Significance Testing and the Metaphor of a Jury Trial

The accused is presumed innocent (the null hypothesis)

The prosecutor's job is to assemble enough evidence to make the probability of innocence appear very unlikely.

How unlikely?

Analysts have come to rely on .05 as the standard. In other words, given the strength of the evidence, the jury would only make the mistake of rejecting the null hypothesis in 5 out of 100 such similar situations: convict an innocent man or woman.

P VALUES (OR SIG.)

The probability of rejecting the null hypothesis (i.e., he's guilty, your honor) when false (i.e., he's actually innocent) is reported as the:

P-value

Also known as "sig." on SPSS output, we usually hope to find this value to be small, i.e., less than .05.

N.B.: this is a frequent source of confusion: small is (usually) good for a p value, while we often hope for "big" or strong relationships

The null hypothesis isn't always a zero effect.

WHY DO STATISTICALLY SIGNIFICANT RESULTS OCCUR?

- The actual relationships or group differences in the population are large
- The sample is large

As a consequence, it's important to report confidence intervals and effect sizes

Statistical significance does not necessarily mean that the effects are large!!!

Your evidence can be overwhelming, simply because you've got a big sample

"POWER" IS THE FLIP SIDE OF "SIGNIFICANCE"

In our jury metaphor, without enough "power," you may let a guilty man go free.

Statistically speaking, you fail to reject the null hypothesis (innocence) when it is actually false (guilty)

A rule of thumb calls for tests with power of at least .80. (You'd let a guilty man go free in 20% of such circumstances).

The likely cause of this: too small a sample for the size of the effect in the population.

https://www.danielsoper.com/statcalc/calculator.aspx?id=9 (Post-hoc Statistical Power Calculator for Multiple Regression)

https://www.danielsoper.com/statcalc/calculator.aspx?id=49 (Post-hoc Statistical Power Calculator for a Student t-Test)

DIFFICULT CONCEPTS BUT UBIQUITOUS

So, learn it to be able to read others' works

But try to rely more in our analysis and presentations by also reporting:

- Confidence intervals and
- Effect sizes

EFFECT SIZES

- Pearson's correlation coefficient (r) ranges from -1.0 to +1.0 (|1.0| = perfect effect) for relationships between continuous vars
  - r = .10 small effect
  - r = .30 moderate effect
  - r = .50 large effect
- Differences between group means divided by their pooled s.d.s
  - 0.0 - .20 = small effect
  - .20 - .50 = moderate effect
  - .50 and above = large effect

EFFECT SIZES (CONT)

- Eta square measures the strength of association for group diffs:
  - .01 = small effect
  - .06 = moderate
  - .14 = large
- Other "measures of association" for relationships

SAMPLE SIZE

- Determining sample size is somewhat complex. (Hire a sampling statistician if you're designing a serious sample.)

THE IDEAL SAMPLE SIZE IS A FUNCTION OF

For descriptive statistics like the mean:

- the amount of error in your estimate that you can tolerate (E) (e.g., + or – 3%) How close?
- the confidence you want to have in the error estimate (z = 1.96 if you want to be sure that you're right in 95/100 samples drawn) How confident?
- the amount of variability in the population (as measured empirically by the standard deviation, s.d.)

The more homogeneous the population, the smaller the sample required, i.e., if everyone was the same, you'd only need a sample (n) of one

AN EXAMPLE: % OF WELFARE RECIPIENTS WHO OWN CARS

We want to know the proportion of welfare recipients who own cars within 2% and we want to be 95% certain. Thus:

$$n = \left(\frac{1.96 \times s.d}{.02} \right)^2$$

We need to know the standard deviation.

Fortunately, statisticians have discovered the largest standard deviation of a proportion of .5 is .50.

$$n = \left(\frac{1.96 \times .5}{.02} \right)^2 = \left(\frac{.98}{.02} \right)^2 = 49^2 = 2{,}401$$

Consider a s.d. of .3

$$n = \left(\frac{1.96 \times .3}{.02} \right)^2 = \left(\frac{.59}{.02} \right)^2 = 29.4^2 = 864$$

RULE OF THUMB FOR SAMPLE SIZE

Samples of 1,000 to 1,500 are good for plus or minus about 3% in 95/100 samples of this size when sampling from a population of more than 20,000 members.



SAMPLE SIZES FOR ESTIMATES OF RELATIONSHIPS (E.G., REGRESSION)

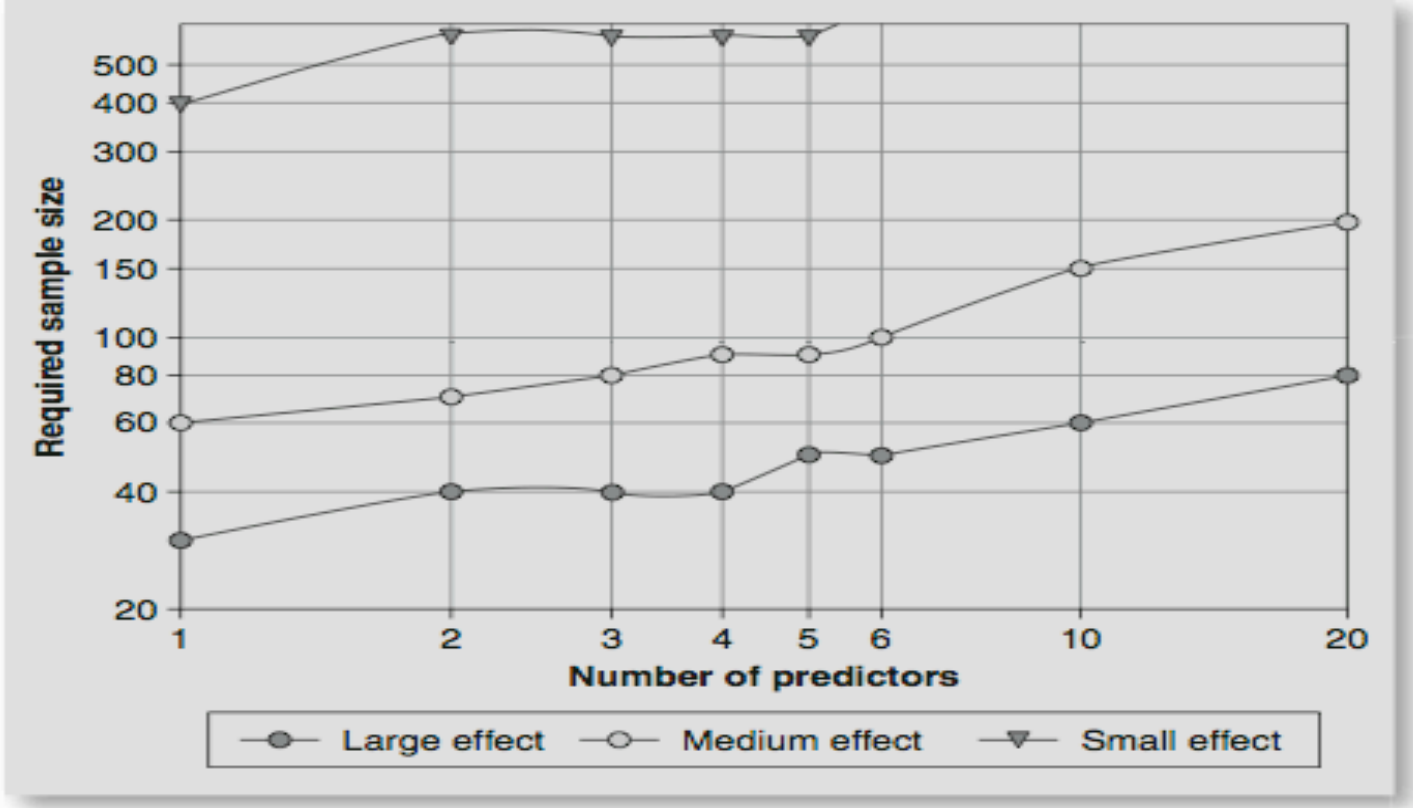
- Again, depend on how close and how confident you want to be to the "true" population parameter

But also:

- The size of the effect in the population
- The number of independent variables in your model
- The "power" of the test (.80 or above)

REQUIRED SAMPLE SIZE FOR MULTIPLE REGRESSION

Figure 8.3 Required Sample Size (for Multiple Regression) by the Number of Independent Variables and Size of the Effect in the Population



MODEL SPECIFICATION

- Anyone have any idea what this might mean?

Consider an example of gender and promotion

Year of Promotion by Gender, Percent (n)			
Year	Men	Women	Total
First	32.6 (14)	15.4 (8)	23.2
Second	37.2 (16)	26.9 (14)	31.6
Third	16.2 (7)	42.3 (22)	30.5
Fourth	14.0 (6)	15.4 (8)	14.7
			(95)

CONTINUED

These differences in promotion are statistically significant and may provoke us to want to do something to rectify what some of you may view as a problem.

Unfortunately, the finding in itself does not help us determine what to do about the problem.

Q.: Why?

A.: It doesn't explain why the relationship exists or how we might influence change in the dependent variable (rates of promotion) by changing one or more of the conditions that cause this group difference to exist.

Models can be useful for us here

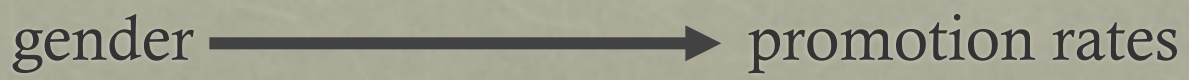
… because they help one think about why relationships (or group differences) exist and to identify those variables that can be changed by, say, taxes, regulations, marketing, etc.

Okay, but what's a "model"?

MODELS OF GENDER/PROMOTION

Model 1

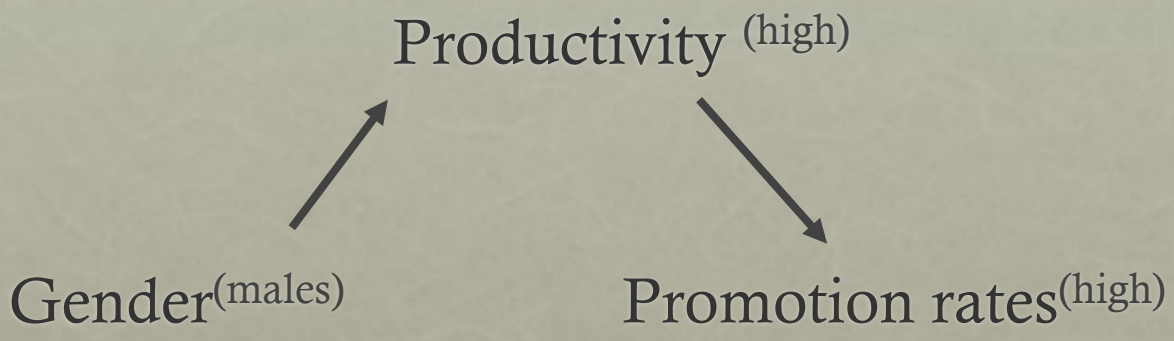
(simple bivariate)



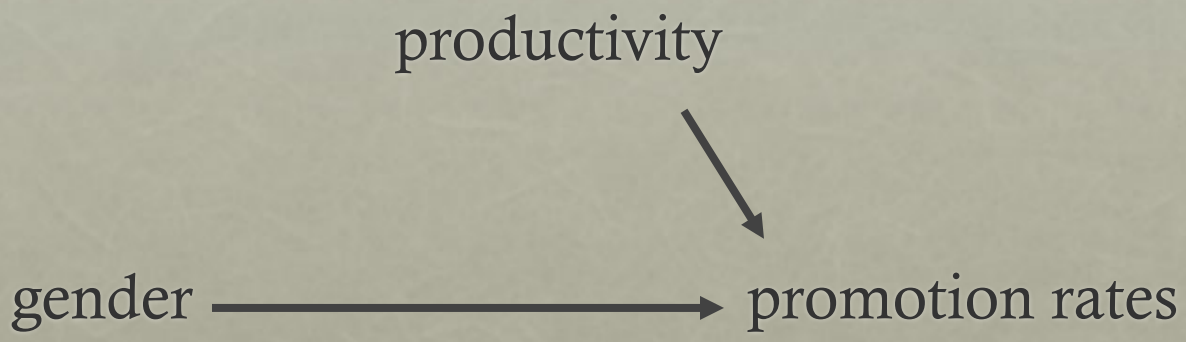
BUT WHAT IF PRODUCTIVITY IS RELATED TO GENDER AND PROMOTION RATES?

- That is to say, what "rival hypotheses" or alternative models might exist to the bivariate model by adding "productivity"?

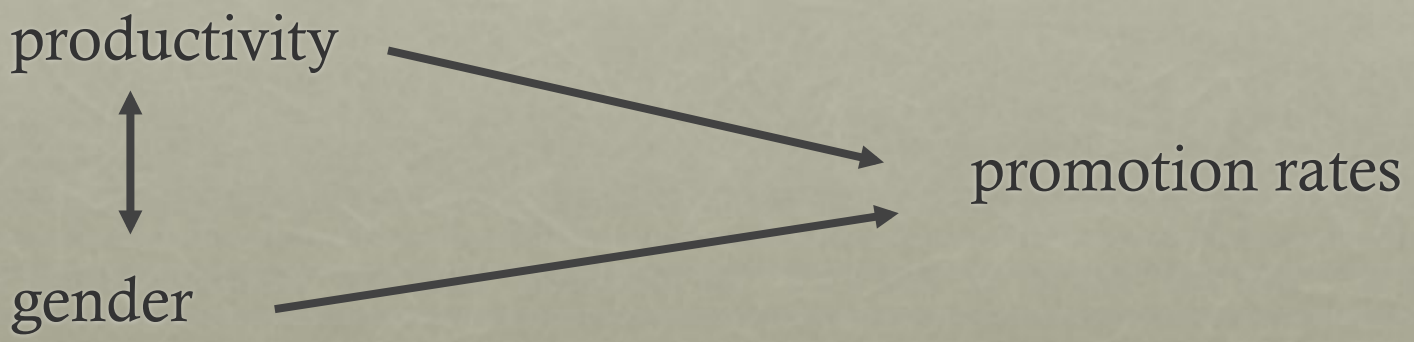
Model 2: Productivity "explains" the relationship



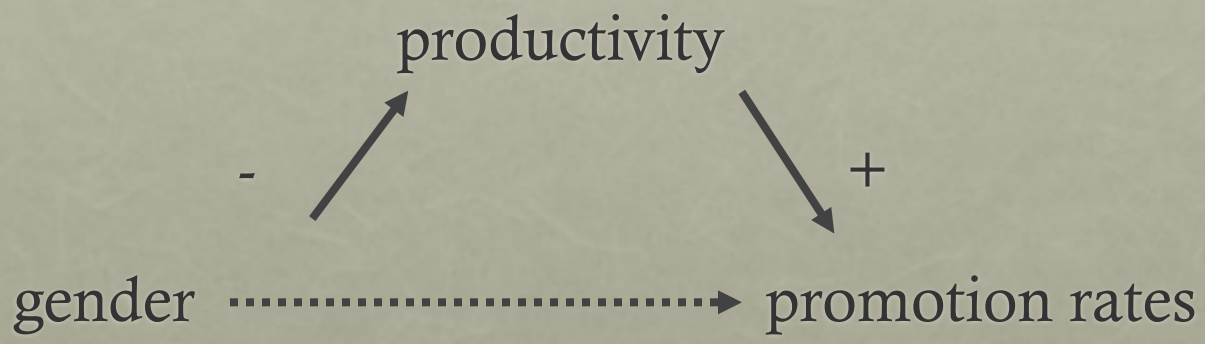
Model 3: Productivity "Replicates" the Relationship



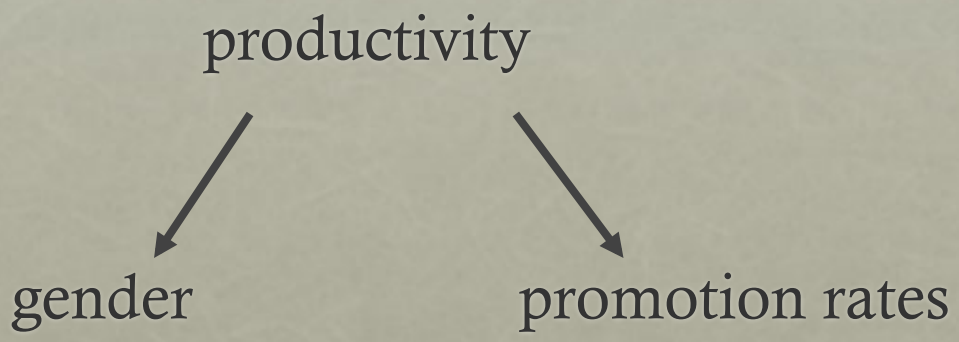
Model 4: Productivity and Gender "Interact"



Model 5: Productivity "Suppresses" the Relationship



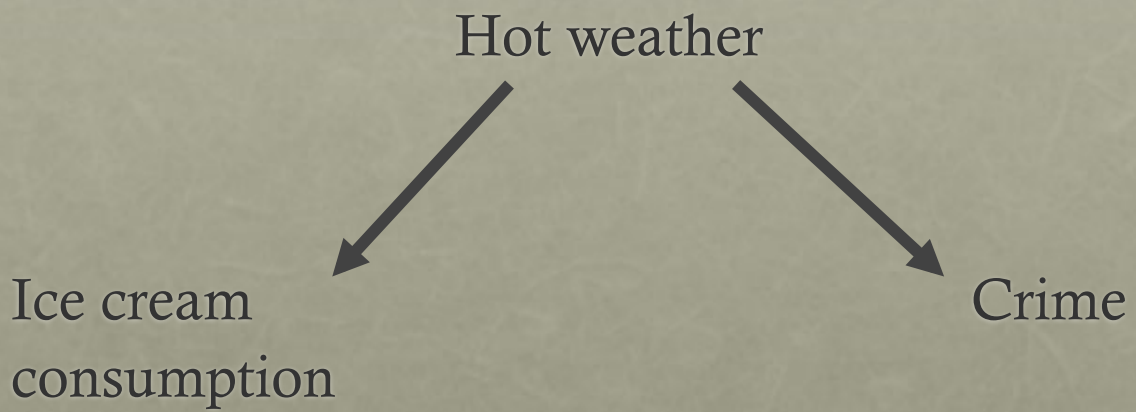
Model 6: The relationship is "spurious"



A FAMOUS SPURIOUS RELATIONSHIP



- The true relation is:



Introducing a 3rd Variable

Year of Promotion by Gender, Controlling for Productivity % (n)				
Year of service	Low Productivity		High Productivity	
	Men %	Women %	Men %	Women %
First or Second	47 (8)	22 (6)	85 (22)	52 (18)
Third or Beyond	53 (9)	78 (12)	15 (4)	47 (16)

- Q.: What plausible alternative explanations exist for better understanding the relationship between gender and promotion rates?

ALTERNATIVE "EXPLAINERS"

Personal Characteristics

- Network relationships
- Height
- Verbal acuity
- Race
- Age
- Accent
- Education
- Credentials

Behavior

- Hours worked
- Arrival/departure time

Attitudes

- Desire, fear of failure
- Willingness to ask for a promotion
- Team spirit

IMPLICATIONS

- Our model will affect what we measure and how large our sample should be, as well as our causal interpretation
- It may be easy to discredit someone's analysis by claiming that the model is "misspecified" by not including plausible explanatory variables