Oracle Corporation et al v. SAP AG et al
Doc. 778 Att. 8

EXHIBIT H

Dockets.Justia.com

STATISTICAL METHODS FOR THE SOCIAL SCIENCES

Fourth Edition

Alan Agresti
University of Florida

Barbara Finlay
Texas A & M University

PEARSON
Prentice Hall

Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Agresti, Alan.
Statistical methods for the social sciences / Alan Agresti, Barbara Finlay.—
4th ed.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-13-027295-9 (alk. paper)
1. Statistics. 2. Social sciences—Statistical methods. I. Finlay, Barbara. II.
Title.
QA276.12.A34 2009
519.5—dc22

2007038723

Editor-in-Chief: *Deirdre Lynch*
Vice President and Editorial Director, Mathematics: *Christine Hoag*
Sponsoring Editor: *Dawn Murrin*
Editorial Assistant/Print Supplements Editor: *Joanne Wendelken*
Senior Managing Editor: *Linda Behrens*
Associate Managing Editor: *Bayani Mendoza de Leon*
Project Manager, Production: *Traci Douglas*
Senior Operations Supervisor: *Diane Peirano*
Executive Marketing Manager: *Kate Valentine*
Marketing Manager: *Wayne Parkins*
Assistant Manager, Marketing Support: *Diana Penha*
Marketing Assistant: *Kathleen DeChavez*
Art Director: *Jayne Conte*
Cover Designer: *Kiwi Design*
Cover Painting: *"Harvest" by Albina Felski*
Composition/Art Studio: *Laserwords*

PEARSON Prentice Hall

© 2009, 1997 by Prentice Hall, Inc.
Pearson Prentice Hall
Pearson Education, Inc.
Upper Saddle River, New Jersey 07458

Previous edition copyright © 1986 and 1979 by Dellen Publishing Company, a division of Macmillan, Inc.

All rights reserved. No part of this book may be reproduced in any form or by any means, without permission in writing from the publisher.

Pearson Prentice Hall™ is a trademark of Pearson Education, Inc.

Printed in the United States of America

10 9 8 7 6 5 4 3 2

ISBN-13: 978-0-13-027295-9
ISBN-10: 0-13-027295-7

Pearson Education Ltd., *London*
Pearson Education Australia Pty. Ltd., *Sydney*
Pearson Education Singapore, Pte. Ltd.
Pearson Education North Asia Ltd., *Hong Kong*
Pearson Education Canada, Inc., *Toronto*
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan, *Tokyo*
Pearson Education Malaysia, Pte. Ltd.

2.2 RANDOMIZATION

Inferential statistical methods use sample statistics to make predictions about population parameters. The quality of the inferences depends on how well the sample represents the population. This section introduces an important sampling method that incorporates ***randomization***, the mechanism for achieving good sample representation.

Simple Random Sampling

Subjects of a population to be sampled could be individuals, families, schools, cities, hospitals, records of reported crimes, and so on. *Simple random sampling* is a method of sampling for which every possible sample has equal chance of selection.

Let n denote the number of subjects in the sample, called the ***sample size***.

> **Simple Random Sample**
>
> A ***simple random sample*** of n subjects from a population is one in which each possible sample of that size has the same probability (chance) of being selected.

For instance, suppose a researcher administers a questionnaire to one randomly selected adult in each of several households. A particular household contains four adults—mother, father, aunt, and uncle—identified as M, F, A, and U. For a simple random sample of $n = 1$ adult, each of the four adults is equally likely to be interviewed. You could select one by placing the four names on four identical ballots and selecting one blindly from a hat. For a simple random sample of $n = 2$ adults, each possible sample of size two is equally likely. The six potential samples are (M, F), (M, A), (M, U), (F, A), (F, U), and (A, U). To select the sample, you blindly select two ballots from the hat.

A simple random sample is often just called a ***random sample***. The *simple* adjective is used to distinguish this type of sampling from more complex sampling schemes presented in Section 2.4 that also have elements of randomization.

Why is it a good idea to use random sampling? Because everyone has the same chance of inclusion in the sample, so it provides fairness. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population. Most inferential statistical methods assume randomization of the sort provided by random sampling.

How to Select a Simple Random Sample

To select a random sample, we need a list of all subjects in the population. This list is called the ***sampling frame***. Suppose you plan to sample students at your school. The population is all students at the school. One possible sampling frame is the student directory.

The most common method for selecting a random sample is to (1) number the subjects in the sampling frame, (2) generate a set of these numbers randomly, and (3) sample the subjects whose numbers were generated. Using *random numbers* to select the sample ensures that each subject has an equal chance of selection.

> **Random Numbers**
>
> ***Random numbers*** are numbers that are computer generated according to a scheme whereby each digit is equally likely to be any of the integers 0, 1, 2, ..., 9 and does not depend on the other digits generated.

- For inference about a population mean, the degrees of freedom equal $df = n - 1$, one less than the sample size.
- The t distribution has thicker tails and is more spread out than the standard normal distribution. The larger the df value, however, the more closely it resembles the standard normal. Figure 5.4 illustrates. When df is about 30 or more, the two distributions are nearly identical.

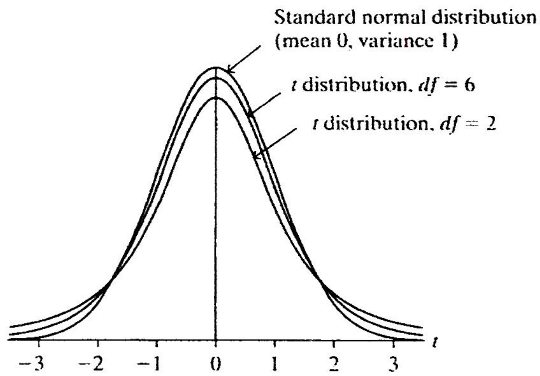

FIGURE 5.4: t Distribution Relative to Standard Normal Distribution. The t gets closer to the normal as the degrees of freedom (df) increase, and the two distributions are practically identical when $df > 30$.

- A t-score multiplied by the estimated standard error gives the margin of error for a confidence interval for the mean.

Table B at the end of the text lists t-scores from the t distribution for various right-tail probabilities. Table 5.1 is an excerpt from that table. The column labelled $t_{.025}$ has probability 0.025 in the right tail and a two-tail probability of 0.05. This is the t-score used in 95% confidence intervals.

To illustrate, when the sample size is 29, the degrees of freedom are $df = n - 1 = 28$. With $df = 28$, we see that $t_{.025} = 2.048$. This means that 2.5% of the t distribution falls in the right tail above 2.048. By symmetry, 2.5% also falls in the left tail below $-t_{.025} = -2.048$. See Figure 5.5. When $df = 28$, the probability equals 0.95 between

TABLE 5.1: Part of Table B Displaying t-Scores. The scores have right-tail probabilities of 0.100, 0.050, 0.025, 0.010, 0.005, and 0.001.

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.657	318.3
10	1.372	1.812	2.228	2.764	3.169	4.144
28	1.313	1.701	2.048	2.467	2.763	3.408
30	1.310	1.697	2.042	2.457	2.750	3.385
100	1.290	1.660	1.984	2.364	2.626	3.174
infinity	1.282	1.645	1.960	2.326	2.576	3.090

The α-Level: Using the P-Value to Make a Decision

A significance test analyzes the strength of the evidence against the null hypothesis, H_0. We start by presuming that H_0 is true. We analyze whether the data would be unusual if H_0 were true by finding the P-value. If the P-value is small, the data contradict H_0 and support H_a. Generally, researchers do not regard the evidence against H_0 as strong unless P is very small, say, $P < 0.05$ or $P < 0.01$.

Why do smaller P-values indicate stronger evidence against H_0? Because the data would then be more unusual if H_0 were true. When H_0 is true, the P-value is roughly equally likely to fall anywhere between 0 and 1. By contrast, when H_0 is false, the P-value is more likely to be near 0 than near 1.

In practice, it is sometimes necessary to decide whether the evidence against H_0 is strong enough to reject it. The decision is based on whether the P-value falls below a prespecified cutoff point. It's most common to reject H_0 if $P \leq 0.05$ and conclude that the evidence is not strong enough to reject H_0 if $P > 0.05$. The boundary value 0.05 is called the ***α-level*** of the test.

α-Level

The ***α-level*** is a number such that we reject H_0 if the P-value is less than or equal to it. The α-level is also called the ***significance level***. In practice, the most common α-levels are 0.05 and 0.01.

Like the choice of a confidence level for a confidence interval, the choice of α reflects how cautious you want to be. The smaller the α-level, the stronger the evidence must be to reject H_0. To avoid bias in the decision-making process, you select α *before* analyzing the data.

EXAMPLE 6.5 Adding Decisions to Previous Examples

Let's use $\alpha = 0.05$ to guide us in making a decision about H_0 for the examples of this section. Example 6.2 (page 149) tested H_0: $\mu = 4.0$ about mean political ideology. With sample mean $\bar{y} = 4.075$, the P-value was 0.50. The P-value is not small, so if truly $\mu = 4.0$, it would not be unusual to observe $\bar{y} = 4.075$. Since $P = 0.50 > 0.05$, there is insufficient evidence to reject H_0. It is believeable that the population mean ideology was 4.0.

Example 6.4 tested H_0: $\mu = 0$ about the mean weight gain for teenage girls suffering from anorexia. The P-value was 0.017. Since $P = 0.017 < 0.05$, there is sufficient evidence to reject H_0 in favor of H_a: $\mu > 0$. We conclude that the treatment results in an increase in mean weight. Such a conclusion is sometimes phrased as, "The increase in the mean weight is *statistically significant* at the 0.05 level." Since $P = 0.017$ is *not* less than 0.010, the result is *not* significant at the 0.010 level. In fact, *the P-value is the smallest level for α at which the results are significant*. So, with P-value $= 0.017$, we reject H_0 if $\alpha = 0.02$ or 0.05 or 0.10, but not if $\alpha = 0.015$ or 0.010 or 0.001. ■

Table 6.3 summarizes significance tests for population means.

Robustness for Violations of Normality Assumption

The t test for a mean assumes that the population distribution is normal. This ensures that the sampling distribution of the sample mean $\bar{y}$ is normal (even for small n) and, after using s to estimate σ in finding the se, the t test statistic has the t distribution. As the sample size increases, this assumption of a normal population becomes less

Main Title
Publication Info
, David Ray, Statistics for business and economics / David R. Anderson, Dennis J. Sweeney ...
Library of
Call No.