Inferential Statistics: Basic Concepts

This chapter discusses some of the basic concepts in inferential statistics. Details of particular inferential tests–t-test, correlation, contingency table analysis, etc.–are included in other chapters. This overview is not meant to replace the more-complete information available in a statistics text.

The Goal of Inferential Statistics

Quantitative research in psychology and social science aims to test theories about the nature of the world in general (or some part of it) based on samples of "subjects" taken from the world (or some part of it). When we perform research on the effect of TV violence on children's aggression, our intent is to create theories that apply to all children who watch TV, or perhaps to all children in cultures similar to our own who watch TV. We of course cannot study all children, but we can perform research on samples of children that, hopefully, will generalize back to the populations from which the samples were taken. Recall that external validity is the ability of a sample to generalize to the population (and to other operationalizations that are assumed to represent that same constructs).

Inferential statistics is the mathematics and logic of how this generalization from sample to population can be made. The fundamental question is: can we infer the population's characteristics from the sample's characteristics? Descriptive statistics remains local to the sample, describing its central tendency and variability, while inferential statistics focuses on making statements about the population.

Hypothesis Testing

Inferential stats is closely tied to the logic of hypothesis testing, discussed in other chapters. In hypothesis testing, the goal is usually to reject the null hypothesis. The null hypothesis is the null condition: no difference between means or no relationship between variables. Data are collected that allow us to decide if we can reject the null hypothesis, and do so with some confidence that we're not making a mistake. In psychology, the standard level of confidence, "alpha," is a probability of .05 that we are rejecting the null hypothesis when we shouldn't (Type I Error).

The null hypothesis is a statement of the null condition in the population, not the sample. For example,

H_0: mean of males = mean of females

is formally expressed in terms of population, not sample, means:

$$H_0: \mu_M = \mu_F$$

where μ (mu) signifies the population mean, that is, the mean of all the males or females in the entire population. However, we only have a sample to work with, so we obtain a sample mean, M (or X-bar).

The null hypothesis we are so intent on rejecting concerns the population means. Our task is to determine if the sample means are sufficiently different to *infer* that the population means are different.

Basic Principle of Inferential Statistics

At a certain basic level, all inferential statistics procedures are the same: they seek to determine if the observed (sample) characteristics are sufficiently deviant from the null hypothesis to justify rejecting it. How deviant? Well, this is the whole point of a statistics course.

The ingredients for making this calculation are the same for all statistical procedures:

1. the size of the observed difference(s)
2. the variability in the sample
3. the sample size.

If instead we are looking for relationships (e.g., correlation analysis), the size of the relationship replaces the size of the difference. Therefore, researchers seeks to:

1. find large differences (or relationships)
2. hold down unwanted variability
3. obtain large samples.

(Good research must also worry about internal and external validity, and the ultimate goal: building theories.)

Error Variance

Types of error were introduced in an earlier chapter. Research must deal with two kinds of error variance: systematic and unsystematic. Systematic error means that there is a bias in the data. For example, say your experiment on the personality trait extraversion-introversion was conducted in both morning and afternoon sessions. For some reason, you had all the introverts come to the morning sessions and all the extraverts come to the afternoon sessions. This mistake would produce a confounding of personality trait with time of day, which for this trait dimension is a problem.

Unsystematic error is termed **error variance** or **residual variance**. "Error" may be a misnomer here in that it is not exactly as if the researcher made a mistake. Behavior is complex, multiply determined, and interacts with many unknown or unforeseen intrapersonal and situational influences, so as psychologists we would never expect everyone to act (or respond on questionnaires) in

exactly the same way. Careful research tries to minimize as many of the influences on behavior, besides the independent variable, as possible. However, this is practically impossible so we always see differences in responding within any sample or condition of an experiments.

For example, in an experiment on the effects of alcohol on driving performance, subjects will come to the experiment with different levels of driving skills and in varying moods and dispositions. Therefore, within any condition of such an experiment (no alcohol condition, 3-drinks condition, etc.) we should expect variability in performance. This within-group (within experimental condition) variability is error variance. It is, essentially, what we don't know or haven't tried to find out about the subjects. In the words of my first (rather frightening) statistics teacher, "error is ignorance."

Reducing Ignorance

We can reduce ignorance in many ways, and doing so is one of the goals of researchers. For example, we could have selected only subjects with at least 5 years of driving experience for our alcohol study, assuming that after 5 years people have learned as much as there is to learn about driving, hence reducing differences among the participants within each of the alcohol conditions.

Another way to reduce error variance is to use something we know about the subjects to change the experimental design in a way that "accounts for" or "removes" some error variance. The strategy is to divide up the participants into groups within which people will be more similar to each other. These groups will have lower within-group variability in their behavior. For example, if females are more careful drivers than males, as the probability data used by insurance companies to calculate premiums clearly indicate, then adding a variable "gender" to the design to separate the males and females into their own groups should reduce driving performance variability within the groups. By knowing something about gender, we have reduced our ignorance.

All the Formulae are the Same

In a statistics course, you would learn the several formulae used to calculate the numbers you need in order to determine if you may reject the null hypothesis. A few of these formulae are presented here to point out their similarities. Conceptual, not actual, formulae are used. Just what these formulae mean is explained in

Name of Test	**Formula (conceptual)**	**What it does**
t-test	$t = \frac{\text{Mean Difference}}{\text{Error Variance}}$	Difference between two means
Analysis of Variance	$F = \frac{\text{Differences Among Means}}{\text{Error Variance Within Groups}}$	Differences among many means, with many IVs
Chi-Square Test	$\chi^2 = \frac{\text{Extent to which frequencies are not consistent with the null hyp}}{\text{Size of sample}}$	Differences in frequencies; or relationship between nominal variables

the data analysis chapters.

Statistics

The calculation involving the ratio of size of difference (numerator) to amount of error (denominator), illustrated in the table, produces a number termed a **statistic**. t, F, and χ^2 are statistics. This value becomes larger as the ratio of difference (or relationship) to error increases, as the formulae indicate. A larger statistic obviously is better.

The statistic is compared to a table of statistics to determine if it is large enough to reject the null hypothesis. As the sample size increases, the statistic need not be as large to reject the null hypothesis. The reason for this is that larger samples give more confidence that the obtained sample is not unrepresentative of the population, that is, not a "fluke." Small samples are notoriously "unstable" because every small sample that is drawn from a population will be distinctly different due to chance. Therefore, to be confident that the ratio is sufficiently large to reject the null hypothesis when the sample is small, the statistic must be larger.

p-values

The null hypothesis can only be rejected when the probability of a Type I Error (error of rejecting it when you shouldn't) is less than .05 (the alpha level). If this were, say, 1960, the only way that you could calculate a statistic would be by hand, and you would determine if the statistic was large enough to reject the null hypothesis, relative to the size of the sample, by looking up the required value in a table at the back of a statistics book. The table would have a column of values headed by ".05" in addition to other columns with other alpha levels, such as ".01". If your statistic was larger than the value in the ".05" column for your sample size, you would reject the null hypothesis.

Times have changed. Statistical analysis software calculates a **p-value** that indicates the probability of committing a Type I Error based on the statistic (which it also calculates for you) and the sample size. The p-value can be any number between 0 and 1. If the p-value is less than alpha (generally .05), you may reject the null hypothesis.

The following portion of a t-test from SPSS illustrates a p-value, labeled "Sig. (2-tailed)." The p-value is .001, a probability much lower than .05, indicating that we should reject the null hypothesis. The column "df" is related to the sample size.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
HEADACHE	Equal variances assumed	3.960	.050	-3.306	87	.001	-.76	.229	-1.213	-.302
	Equal variances not assumed			-3.296	80.468	.001	-.76	.230	-1.215	-.300

Procedure for Performing an Inferential Test

Here is a step-by-step procedure for performing inferential statistics.

1. Start with a theory

   (TV violence reduces children's ability to detect violent behavior because it blurs the distinction between real and fantasy violence)

2. Make a research hypothesis

   (Children who experience TV violence will fail to detect violent behavior in other children)

3. Operationalize the variables

4. Identify the population to which the study results should apply

   (Kids in developed nations)

5. Form a null hypothesis for this population

   (H_0: $u_H = u_L$, where u_H is mean accurate detection of violence by children with high prior exposure to TV violence; u_L is low prior exposure)

6. Collect a sample of children from the population and run the study

7. Perform statistical tests to see if the obtained sample characteristics are sufficiently different from what would be expected under the null hypothesis to be able to reject the null hypothesis.

8. Publish the paper, get famous, find a job at Harvard

Details on how to perform step (7) are included in other chapters focusing on the use of SPSS (Statistical Package for the Social Sciences) to analyze data.