"I'm afraid that I rather give myself away

when I explain," said he.

"Results without causes are much more impressive."

Sherlock Holmes

The Stock-Broker's Cat

COHEN CHAP 5. HYPOTHESIS TESTS

For EDUC/PSY 6600

TWO GENERAL TYPES OF RESEARCH QUESTIONS

- Do groups significantly differ on 1 or more characteristics?
  - Comparing group means, counts, proportions
  - t-tests, ANOVA, χ^2 tests
- Is there a significant relationship among a set of variables?
  - Testing the association or dependence
  - Correlation, regression

INFERENTIAL STATISTICS

Are differences/associations due to
SAMPLING ERROR
OR
TRUE
experimental/ phenomenological differences/processes?

- Descriptive statistics are limited
  - Rely only on raw data distribution
  - Generally describe 1 variable
  - Do not address accuracy of estimators or hypothesis testing
    - How precise is sample mean or does it differ from a given value?
    - Are there between or within group differences or associations?

- Goals of inferential statistics
  - Hypothesis testing
  - Parameter estimation

- Repeated sampling
  - Estimators will vary from sample to sample
    - Sampling or random error → Variability due to chance

Are our results due to chance?

Not: Are our results important?

Need:
Theoretical knowledge,
measures of effect size,
or
confidence limits

CAUSALITY & STATISTICS

Causality depends on evidence from outside statistics:

1. Phenomenological (educational/behavioral/biological) credibility
2. Strength of association – ruling out occurrence by chance alone
3. Consistency with past research findings
4. Temporality
5. Dose-response relationship
6. Specificity
7. Prevention

Causality is often a judgmental evaluation of combined results from several studies

Z-SCORES AND STATISTICAL INFERENCE

- Probabilities of z-scores used to determine how unlikely or unusual a single case is relative to other cases in a sample

Small probabilities (*p*-values) → Unlikely or unusual scores

- Not frequently interested in whether individual scores are unusual relative to others but whether scores from groups of cases are unusual
- **Sample mean,** $\overline{X}$, summarizes central tendency of a group or sample of subjects

STEPS OF A HYPOTHESIS TEST

1) State the Hypotheses (Null & Alternative)
2) Select the Statistical Test & Significance Level
   - α level
   - One vs. Two tails
3) Select random sample and collect data
4) Find the region of Rejection
   - Based on α & # of tails
5) Calculate the Test Statistic
   - Examples include: z, t, F, χ^2
6) Make the Statistical Decision

p-value
probability of observing
a test statistic
"as extreme or more extreme"
IF the NULL is true

HYPOTHESIS

- Hypotheses are always specified in terms of **population**

- Null (nil) hypothesis
  - H_0: $\mu_1 = \mu_2$

- Research (alternative) hypothesis
  - H_1: $\mu_1 \neq \mu_2$
  - H_1: $\mu_1 < \mu_2$
  - H_1: $\mu_1 > \mu_2$
  - Sometimes H_A is used

REJECTING THE NULL HYPOTHESIS

- Assumption: the NULL hypothesis is TRUE in the POPULATION
- IF the p-value is very SMALL (p-value< α), it is UNLIKELY we would have observed a sample that extreme JUST DUE TO RANDOM CHANCE.
- Can use the **p-value< α** OR **test statistic < Critical Value**
- We either REJECT or FAIL TO REJECT the Null hypothesis
- We NEVER ACCEPT the ALTERNATIVE!!!

"Innocent until proven guilty"

Not enough statistical evidence to reject…

Judgment suspended until further evidence evaluated:

(Inconclusive)

Larger sample?

Insufficient data?

1 OR 2 TAILS?

1-tailed test suggests a directionality in results

- H_1: $\mu_1 < \mu_2$
- H_1: $\mu_1 > \mu_2$
- Rejection region collapsed into 1 end of sampling distribution
  - e.g., $\alpha = 5\%$ in one tail

No computational differences for test statistics

ONLY the p-level differs

- 1-tailed, p of event or more extreme event in one tail
- 2-tailed, p of event or more extreme event in both tails
  - p from 1-tailed*2 (2-sided: $p = .06$, 1-sided: $p = .03$)

Some circumstances may warrant a 1-tailed test, BUT...

We generally prefer and default to a 2-tailed test!!!

More conservative

Rejection region is distributed in both tails

e.g.: $\alpha = 5\%$ distributed across both tails (2.5% in each tail)

If we know outcome, why do study?

Looks suspicious to reviewer's…

"significant results at all costs!"

Cheating to run results then decide?

Yes! $\alpha = 7.5\%$

Can't flip coin, say it will land on tails, then lands on heads, and switch and say that it will land on heads

CHOOSING ALPHA

type I error
We reject the NULL
when we shouldn't have
The risk of "false positive" results

type II error
We FAIL to reject the NULL
when we should have
The risk of "false negative" results

- Alpha = probability of making a type I error
- We want α to be small, but…we can't just make too tiny, since the trade off is increasing the type II error rate
- Default is $\alpha = .05$ (5% = 1 in 20 & seems 'rare' to humans) BUT there is nothing magical about it
- Let it be LARGER value ($\alpha = .10$) IF we'd rather not miss any potential relationship and are okay with some false positives
  - Ex) screening genes, early drug investigation, pilot study
- Set it SMALLER ($\alpha = .01$) IF false positives are costly and we want to be more stringent
  - Ex) changing a national policy, mortgaging the farm

ASSUMPTIONS OF A 1-SAMPLE Z-TEST

1. Sample was drawn at random (at least as representative as possible)

   Nothing can be done to fix NON-representative samples!

   Can not statistically test

2. SD of the sampled population = SD of the comparison population

   Very hard to check

   Can not statistically test

3. Variables have a normal distribution

   Not as important if the sample is large (Central Limit Theorem)

   IF the sample is far from normal &/or small n, might want to transform variables

   Look at plots: histogram, boxplot, & QQ plot (straight 45° line)

   Skewness & Kurtosis: Divided value by its *SE* & > ± 2 indicates issues

   Shapiro-Wilks test (small N): $p < .05$ → not normal

   Kolmogorov-Smirnov test (large N):

APA: RESULTS OF A 1-SAMPLE Z-TEST

State the alpha & number of tails prior to any results

Report exact p-values (usually 2 decimal places), except for "p < .001"

Example Sentence:

A one sample *z* test showed that the difference in the quiz scores between the current sample (N = 9, M = 7.00, SD = 1.23) and the hypothesized value (6.000) were statistically significant, *z* = 2.45, p = .040.

EXAMPLE: 1-SAMPLE Z-TEST

Suppose that an anxiety scale is expressed as a T score, so that $\mu = 50$ & $\sigma = 10$.

After an earthquake hits their town, a **random** sample of townspeople yields the following anxiety score: 72, 59, 54, 56, 48, 52, 57, 51, 64, 67

Cautions About Significance Tests

What statistical significance does not mean…

Statistical significance only says whether the effect observed is likely to be due to chance alone because of random sampling.

Statistical significance may not be practically important. That's because statistical significance doesn't tell you about the **magnitude** of the effect, only that there is one.

An effect could be too small to be relevant. And with a large enough sample size, significance can be reached even for the tiniest effect.

- A drug to lower temperature is found to reproducibly lower patient temperature by 0.4° Celsius (*P*-value < 0.01). But clinical benefits of temperature reduction only appear for a 1° decrease or larger.

STATISTICAL significance does NOT mean PRACTICAL significance!!!

Cautions About Significance Tests

Don’t ignore lack of significance

- Consider this provocative title from the *British Medical Journal*: “Absence of evidence is not evidence of absence.”
- **Having no proof of who committed a murder does not imply that the murder was not committed.**

Indeed, failing to find statistical significance in results is **not rejecting** the null hypothesis. This is very different from actually **accepting** it. The sample size, for instance, could be too small to overcome large variability in the population.

When comparing two populations, **lack of significance** does **NOT** imply that the two samples come from the **same** population. They could represent two very distinct populations with similar mathematical properties.

SPSS: TESTING ASSUMPTIONS

```
* gives a few hints.

FREQUENCIES AGE
   /FORMAT NOTABLE
   /STATISTICS SKEWNESS SESKEW KURTOSIS SEKURT.

* more of a complete picture.

EXAMINE VARIABLES=AGE
  /PLOT NPPLOT
  /STATISTICS NONE.
```

Statistics

AGE Patient's Incoming Age

N	Valid	25
	Missing	0
Skewness		-.348
Std. Error of Skewness		.464
Kurtosis		.584
Std. Error of Kurtosis		.902

Skewness & Kurtosis
Divided by value by its *SE*
> ± 2 indicates issues

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AGE Patient's Incoming Age	.080	25	.200*	.983	25	.940

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction