ISS 305
Evaluating Evidence:
Becoming a Smart Research Consumer

8. Real vs. Illusory Relationships

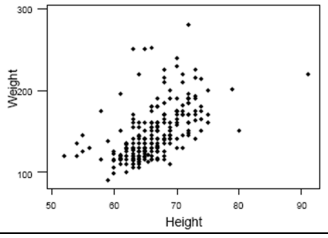
Reminder: Turn on your I<CLICKER

8. Real vs. Illusory Relationships

I. Story of Ivan Steiner
II. The notion of chance
III. Illustrations of difficulty reasoning about chance and probability
   A. Underutilization of baserates
   B. Misperception of chance
   C. Errors with probabilities
   D. Motivated errors: Just World Theory
   E. Insensitivity to sample size
IV. The logic of eliminating chance as a basis for relationships
V. Relationship strength, statistical significance, practical significance, effect size

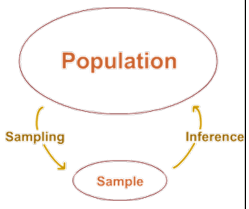
Probabilistically

- Nearly all relationships in psychology are **probabilistic**--as one variable changes, the other variable tends (probabilistically) to change.
  - As height increases, the probability of being heavier increases; it is not certain



Probabilistically

- Also, in most cases, we want to draw a conclusion about everybody (the entire population) from observations of many fewer people.
  - Sometimes, we can be misled because the people we do look at (our sample) may not look like the average person in the full population
  - In order to deal with such problems, we need to be able to reason about **chance** and **probability**



The power of chance and coincidence

The story of Ivan Steiner

- Moral of story:
  - Except for the chance error of an unknown clerk in the CMU registrar's office, Steiner would have ended up as a lawyer, not a psychologist
  - **"Chance" events often determine the course of lives**

The prevalence of uncertainty

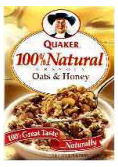
- When we say something is due to **chance**, we **don't** mean that it couldn't be predicted or explained (if we knew enough). Rather, we mean that it is caused by (usually many) factors that we are unaware of and/or are unable to measure.
  - **Something** caused the CMU registrar to make the clerical error in Steiner's records, but we have no idea what, or any reasonable way of finding out.
  - If we knew enough, we should be able to explain and predict which lottery numbers will come up; we don't so we can't and we attribute it to chance factors.

Representativeness heuristic

- Make judgments based on how similar something is to the typical example
- Ex: What cereal is lower in sugar and saturated fats?

A 

or

B 

Underutilization of base rate information

- Example 1:
  - In a group of 100 men, 70 are engineers and 30 are lawyers. Each has written a thumbnail description of himself.
  - Jack's is chosen at random. It reads, "I'm 45 years old. I'm married and have four children. I am generally conservative, careful, and ambitious. I'm not very interested in political and social issues. I spend most of my free time on my hobbies, which include home carpentry, sailing, and mathematical puzzles."
  - What is the probability that Jack is one of the engineers? (0=impossible, 100=certain).
    - **If we didn't have his description, base rate information suggests P(engineer) = .70**
    - **With it, people usually guess that he's very likely to be an engineer (e.g. P=.90)**
    - **BUT, the rating is usually unaffected if we started with 30 engineers and 70 lawyers, and the base rate is P(engineer) = .30.**
- Example 2: Hamill, Wilson, & Nisbett (1980). The prison guard study described earlier in the course:
  - The typicality of the person interviewed had no effect on judgments of the inhumanness of guards, only the content of the interview
- Other examples – see Stanovich (Chapter 10)
- **We tend to ignore or give too little weight to base rate information; preferring vivid, personal information**

Perception of random, chance events

- We've already considered one example
  - We underestimate the likelihood of matched birthdays
- A match is unlikely, but with many people with many birthdays, the probability of an unlikely event can become likely
- We tend not to appreciate the power of large numbers to have such effects
- Other examples...
  - for each, how do you think the average person would respond?

Perception of random, chance events

Question a: Which of the following sequences of coin flips is **most** likely with a fair coin? Which is least likely? (H=HEADS; T=TAILS)

A. HHTHTTHT

B. HHHHTTTT

- Correct Answer: **Both outcomes are equally likely**
- Why do we think otherwise?
  - **Fall victim to the Representativeness Heuristic**
    - **Probability judgments are often based on how representative certain features of events are to what we already know.**
  - **Here, we know that on average, there should be 4 heads and 4 tails in eight flips.**

Perception of random, chance events

Question b: Ozzie and Harriet, a perfectly healthy and normal couple, have had three children, all boys. Harriet is pregnant again. Which is more likely for the fourth child?

A. a boy

B. a girl

C. equally likely

- Correct Answer: **They are equally likely**
- Why do we think otherwise?
  - **Fall victim to the Representativeness Heuristic**
    - **Probability judgments are often based on how representative certain features of events are to what we already know.**
  - **Here, we know that on average, 2 boys and 2 girls in a 4 child family**

Perception of random, chance events

- Question c: On January 1, the winning 3-digit lottery number is 372. Which of the following is least likely to be the winning number on January 2 (that is, which one would you be least willing to choose if you could have any one)?

  A. 248 B. 273 C. 826 D. 372

- Correct Answer: **They are equally likely**
- Why do we think otherwise?
  - **Fall victim to the Gambler's Fallacy**
    - **Gambler's fallacy = the tendency to see links between events in the past and events in the future when the two are really independent**

Perception of random, chance events

- Question d: Suppose that George Johnson's lifetime batting average is .305 (that is, in his professional career, he has got a hit on 30.5% of all his at bats). Over the last two games, George has gone 0 for 8 (that is, got no hits in his last 8 at bats). If we can safely assume that George is not injured or otherwise handicapped as a hitter, what is the probability that he'll get a hit on his next at bat above or below .305?

  A. Above .305 B. .305 C. Below .305

- Correct Answer: **The probability is .305, no higher or lower.**
- Why do we think otherwise?
  - **Fall victim to the Gambler's Fallacy**
    - **long term average "requires" same pattern in the short term**

Illusions of control: Langer (1975)

- People who selected their lottery ticket wanted four times as much for it as people who had someone else select it for them.
- Why?

- Those playing "war" (high card wins) will bet more against an unconfident player than a confident player
- Why?

- People wanting a higher number will roll a die harder than those wanting a low number
- Why?

Trouble reasoning with probabilities: The Conjunction Fallacy

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

What is Linda doing now?

A. Linda is a bank teller.
B. Linda is an elementary school teacher.
C. Lind is a bank teller and is active in the feminist movement.
D. Linda is an insurance salesperson.

Trouble reasoning with probabilities: The Conjunction Fallacy

- Suppose
  - A = Jim is passing this course
  - B = Jim is a German major
- What's reasonable estimate of
  - Probability (A)?
  - Probability (B)?
  - Probability (A and B)?
- **Often, Ps will say that the probability that Jim is both passing the course and a German major is greater than the probability that he is a German major.**
- Conjunction Fallacy:
  - Judging that Probability of (A and B) > min (Prob(A), Prob(B))

New Breast Cancer Treatment

Women can be classified as High or Low Risk. About half (57%) of all women are in the High Risk group, but they account for 92% of all breast cancers. So, Dr. Charles S. Rogers, MD, is doing partial mastectomies on High Risk women before any signs of breast cancer appear.

- What's your best guess at the % of High Risk women who will eventually get breast cancer?
  A. High % (above 67%)
  B. Mod % (between 33% and 67%)
  C. Low % (below 33%)
- If you were a woman, what % would you need to have the preventive surgery (and reconstruction) done?

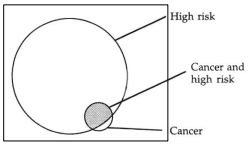
Inverse probabilities

- Conditional probabilities:
  - P(A|B)=Among those where B is true, what is the probability that A is true
  - P(High Risk Group | have breast cancer) **= 71/77 = .92**
  - P(cancer |High Risk group) **= 71/570 = .12**
- We routinely confuse these and assume that P(A|B)≈ P(B|A), but this is usually false.
  - we cannot infer anything about P(A|B) from knowing only P(B|A)

Variable A:	Variable B: Get cancer?		Variable A
Risk Group	No	Yes	totals
High	499	71	570
	87.54%	12.46%	
Low	424	6	430
	98.60%	1.40%	
Variable B totals	923	77	1000

Inverse probabilities

Can better understand this by using Venn diagrams or probability theory:



- Venn diagram:
  - P(Hi R|cancer) high, but
  - P(cancer|Hi R) low
- Probability theory:
  - Ratio rule: P(A|B)/P(B|A)= P(A)/P(B)
    .12 / .92 = .077 / .570 (some rounding error)
    - So, the inverse is the same ONLY if P(A) = P(B)
  - **So, P(cancer|Hi R)= P(Hi R|Cancer) x [P(cancer)/P(Hi R)]**
    .12 = .92 x [.077/.570] = .12

Motivated misperception of probabilities and chance: Belief in a Just World

- The world is a fair place where people get what they deserve and deserve what they get.
  - **Good** people deserve **good** things
  - **Bad** people deserve **bad** things
  - If something **good** happens to you, you must be a **good** person
  - If something **bad** happens to you, you must be a **bad** person
- What actually guarantees this?

Applications of Belief in a Just World

- Court cases: victims who are portrayed as good people

  _____

  –
- Rape victims are

  _____

  –
- An attractive woman's accidental death is viewed as more tragic and unfair than an unattractive woman's

Motivated misperception of probabilities and chance: Belief in a Just World

- Zuckerman (1975): Before exams students who believed more in a ***just world*** (but not those who didn't)
  - (a) volunteered more to serve as Ps in experiments,
  - (b) were more willing to serve as readers for a blind student, and
  - (c) agreed more to participate in a 1-hr study after having already completed an experiment requirement in an introductory psychology course.

  but not after the exam was over
- **Why?**

Insensitivity to Sample size

The hospital problem

A certain town is served by two hospitals. In Large Hospital, about 45 babies are born each day. In Small Hospital, about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

A. ___Large Hospital
B. ___Small Hospital
C. ___About the same (within 5% of each other)

- What's the right answer?
- **Correct answer is the small hospital because there is much more variability in small samples than large samples**
- Why do we make this mistake?
- **Representativeness Heuristic again: Since average % of boys will be the same in the two hospitals, we assume that they'll be about the same in all other regards**
- **Shows the general tendency to pay too little attention to sample size**

More extreme version of the same problem

- Which of the following is **less** likely given a fair coin?
  - A) 2/3 or more Heads out of 3 flips
  - B) 2/3 or more Heads out of 100 flips
- **We'll get 2/3 or more heads half the time with only 3 flips,**
- **But we'll get 2/3 or more heads out of 100 flips very rarely (only ~3 times out of every 10,000 times)**

Insensitivity to sample size and judging relationships

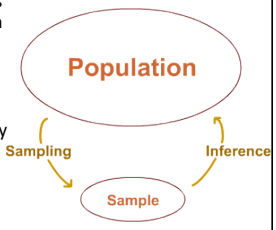
- The failure to take into account the effect of the size of our sample on outcomes is **especially** a problem in evaluating relationships.
- Which of these two 2x2 contingency tables provides the better evidence for a relationship between the variables Beat Up and Eye Color?
  - Note that the difference between beat up (**the effect size**) is the same in both (33%)
  - But, if there really were no difference between all the people beat up and all the people not beat up in the world, a difference this big or bigger would occur, **by chance**,
    - 45% of the time in a randomly selected sample of only 6 people (3 beat up and 3 not beat up)
    - but 0.8% of the time in a random sample of 60 (30 beat up and 30 not beat up)
- So, second table provides much stronger evidence that there is a relationship, because it is much less likely to be due to chance (i.e., sampling error).

Variable A: Beat Up	Variable B: Eye Color		Variable A Totals
	Blue	Brown	
Yes	2	1	3
	66.67%	33.33%	
No	1	2	3
	33.33%	66.67%	
Variable B Totals	3	3	6

Variable A: Beat Up	Variable B: Eye Color		Variable A Totals
	Blue	Brown	
Yes	20	10	30
	66.67%	33.33%	
No	10	20	30
	33.33%	66.67%	
Variable B Totals	30	30	60

The logic of hypothesis testing/inferential statistics

- Basic problem: How to make inferences about large populations from small samples
- If we could observe everyone in the population (without error), then sample size would be a moot point
- But we usually cannot, for practical reasons
- So, there's always the possibility that, by chance, our sample doesn't accurately describe the full population
  - This kind of discrepancy is called "sampling error"
  - And, the smaller our sample, the bigger the risk of sampling error.



Core logic of hypothesis testing

- Hypothesis testing - Systematic procedure for deciding whether the results of a study with a sample support a particular theory or practical innovation
  - Drawing inferences about the population based on information from the sample
- If treatment/manipulation has no effect, "A" should happen; if it has an effect, then anything other than "A" should happen
- Results show that "B" happened, and the chance of "B" happening is very small (e.g., 1%, 5%)
- We say that given treatment/manipulation has no effect, the chance of "B" happening is so small, that we must reject the assumption (hypothesis) that treatment has no effect, and accept the idea that treatment has some effects

Statistical Hypotheses

- A *type* of hypotheses that specify the general nature of the relationships between population **parameters** (e.g., μ).
- Statistical hypotheses are tested using **inferential statistics**
- Experiments often pit two hypotheses against each other: The **null** hypothesis and **alternate/ research** hypothesis.

Statistical Hypotheses

- Null Hypothesis ($\mathbf{H_0}$)
  - The data for the groups are so similar that the scores must have been drawn from the same population.
  - Example: $\mu_1 = \mu_2$ **or** $\mu_1 = \mu_2 = \mu_3$ **…etc.**
  - Often the reverse of what an experimenter actually believes
  - Researchers probably wants (and tries) to reject the null hypothesis – devil's advocate position

Statistical Hypotheses

- Alternate Hypothesis ($\mathbf{H_1}$ **or** $\mathbf{H_a}$)
  - The data for each group came from different populations.
  - Specifies values for the parameter that are **incompatible** with the null hypothesis.
  - A decision to reject $\mathbf{H_0}$ implies an acceptance of $\mathbf{H_1}$
    - Constitutes support of an experimenter's original research hypothesis.
  - **Not all alternate hypotheses are equal**
    - Non-directional: $\mu_1 \neq \mu_2$
    - Directional: $\mu_1 > \mu_2$

Comparing hypotheses

Null Hypothesis (H_0) – "No effect"

- "The data for these groups are so similar that the scores must have been drawn from the same population."

Alternate Hypothesis (H_1 or H_a) – "There's an effect"

- "The data for these groups are so different, they must have come from different populations."

Some other things to know

- **Independent variable (IV)**
  - Variable that is **manipulated** by experimenter. The factor of interest to the experimenter, the one that is being studied to see if it will influence behavior.
- **Dependent variable (DV)**
  - Variable that is **measured** by experimenter. DV "depends upon" the independent variable. AKA the outcome variable.

The generic procedures used in hypothesis testing

- Goal is to choose between these two possibilities (null vs. alternative hypothesis) using the sample evidence you have.
- Assume that the null hypothesis (no relationship, no difference) is true.
  - Figure out what would be an unusually big effect under this assumption.
    - Defines a critical region of unusually big effects
  - If the observed effect is in that region (big enough), then conclude that this is just too unlikely/unusual an event, and that the null hypothesis is probably not true
    - and say that you reject the null hypothesis, or that the observed difference is statistically significant
  - If the observed effect is not in that region (not too big), conclude that the null hypothesis probably is true
    - retain the null hypothesis; the observed difference is non-significant

Critical regions, alpha levels, & sample sizes

- Again, if $\mu_1 = \mu_2$ (no relationship), then a "big" difference between μ_1 and μ_2 is to be unexpected or unusual
- But how big is big enough?
- By convention, the most unusual 5% of possible outcomes are identified
  - this value is called the significance level or the alpha (α) level
  - it is occasionally different than p =.05; almost never larger, though
- The critical regions depend upon α AND on sample sizes
  - the larger the samples, the bigger the critical regions, and hence, the easier it is to conclude that any difference is a significant one.

Sample size and statistical significance

- Even the smallest observed effect will be judged to be statistically significant IF the sample size is large enough.
- For example,

Sample size	Critical Value
6 Ps	.80
60 Ps	.253
500 Ps	.088
1,000 Ps	.062
5,000 Ps	.028
10,000 Ps	.020

- Note that it takes more and more participants to get an equivalent drop in the critical level

Outcomes: What you find in your study (IV is "significant" or not)

1. Significant
   - *Reject* Null Hyp. (H_0)
   - *Support* for research/alternate hypothesis (H_1)
   - Conclude that IV does influence DV
   - "There is sufficient evidence to conclude … $p < .05$"
2. Not significant
   - *Fail to reject* Null Hyp. (H_0)
   - *No support* for research/alternate hypothesis (H_1)
   - Conclude that IV does NOT influence DV
   - "There is insufficient evidence to conclude… *n.s.*"

What is TRUE in the real world?

1. IV Factor does influence DV
   - There is an effect of your factor on your DV
   - For example, the new drug does actually reduce depression.
2. IV Factor does not influence DV
   - There is NO effect of your factor on your DV
   - For example, the new drug does not have any effect on depression.

Decision errors

- Situations where right procedures lead to wrong decisions
  - Like deciding H_0 is false when it is really true
- Possible in hypothesis testing because we are making decisions about the population using only information from the sample
  - The premise of hypothesis testing is based on probabilities
- Small chance ≠ impossible

Decision errors

- Type I error
  - Reject the null hypothesis when in fact it is true
    - *(i.e., there is no treatment effect, but you say there is one)*
  - Conventional level of significance: 5% or 1%
    - If the chance of getting the results in the sample assuming the null hypothesis is true is smaller than 5% (1%), then this chance is too small, and we reject the null hypothesis and call the results statistically significant
    - Again, small chance ≠ impossible
  - Alpha (α)
    - Probability of making a Type I error
    - Usually the same as the level of significance

Decision errors

- Type I error
  - We can never know if we are making a Type I error.
  - However, we can try to reduce the chance of making a Type I error.
  - Ways to reduce Type I error
    - Set the significance level (α) lower (e.g., 5% or 1%, instead of 20%)

Decision errors

- Type II error
  - Retain the null hypothesis when in fact the null hypothesis is false
    - *(i.e., there IS a treatment effect, but you say there isn't one)*
  - Accept the null hypothesis when in fact the research hypothesis is true
  - Beta (β)
    - The probability of making type II error

Decision errors

- Tradeoffs between Type I and II errors
- Choose a stringent α (e.g., 1% rather than 20%) → lower Type I error;
  - But a stringent α makes it more difficult to reject the null hypothesis
    - Our results need to be stronger or more extreme to be significant
  - This increases Type II error
    - May end up concluding that a new treatment is not effective when it can help, just not quite as much

Contingency Table Outcomes

		Real Situation (unknown to us)	
		Null hypothesis retained **IV does not Influence DV**	Research hypothesis supported **IV does Influence DV**
Hypothesis Testing Conclusions	Reject the null and support the research hypothesis	**FALSE ALARM** TYPE I ERROR Prob. = α .05	**HIT** Prob. = Power: (1-β) .80
	Inconclusive (retain the null hypothesis and unable to support the research hypothesis)	**CORRECT NEGATIVE** Prob. = (1 – α) .95	**MISS** TYPE II ERROR Prob. = β .20

Contingency Table Outcomes - Judicial System

H_0: person is innocent		What is TRUE in the Real World	
		Defendant is Innocent	**Defendant is Guilty**
What the Jury decides	***Guilty Verdict*** Reject presumption of Innocence	**FALSE ALARM** TYPE I ERROR Prob. = α .05	**HIT** Prob. = Power: (1-β) .80
	Not Guilty Verdict Fail to reject presumption of Innocence	**CORRECT NEGATIVE** Prob. = (1 – α) .95	**MISS** TYPE II ERROR Prob. = β .20

Effect Size

- Strength of association between variables
  - In statistical test, strength of association between IV & DV.
  - A.K.A treatment magnitude, magnitude of effect

- Effect sizes should always be reported
  - Disagreement as to how
- If DV is well understood, *M* & *SD* are enough to indicate effect size.

Effect Size

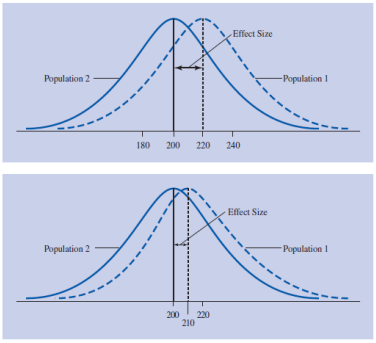
In statistical tests refers to
"how strongly the factor/IV affects the DV"

Effect size is *separate* from the *statistical* test.

- *Statistical* test answers the question,
  "Is there an effect (yes/no)?" or "Are the two samples different (yes/no)?"
- Effect size answers the question,
  "How strong is the effect?" or "How different are they?"

Effect size is important for power analysis, but it's also scientifically interesting in its own right; should be reported

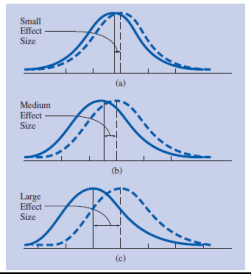
Effect size



Effect size

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- Conventional evaluations of effect sizes in psychology
  - Small effects: .20
  - Medium effects: .50
  - Large effects: .80



Meta-analysis

- Combines results from different studies
- Provides an overall effect size
- Common in the more applied areas of psychology
- This technique allows researchers to average effect sizes across many studies
  - While each individual study may have its own limitations, when results average across them, it is assumed that the limitations will average out and provide a better estimate of the effect size
- **PROBLEMS?**

Power:
A major design consideration

- The probability of getting a statistically significant result ***if*** the research hypothesis is true (rejecting the null, when it is in fact false)
  - i.e., there IS a treatment effect, and you say so.
  - **1 – β**, preferred value = **.80**. (80% chance of finding significant effect when effect is truly there.)
  - NOT an error. You want power!

"If there *is* a treatment effect, do I have enough power to detect it?"

Power of around .80 or higher is desirable
(80% chance of detecting an effect, if one is present)

Determinants of statistical power

- Effect size
  - Larger effect size → less overlap between null (population 2) and research (population 1) populations
  - Result is more likely to be significant and more powerful
- Sample size
  - Larger sample size → smaller standard error
  - Distributions are narrower, and less likely to overlap, which makes the power larger
- α level
  - Smaller α (more stringent, from .05 to .01), smaller power
- One-tailed versus two-tailed test
  - One-tailed test is more powerful
- Type of hypothesis testing procedure
  - Parametric versus nonparametric
  - Parametric is more powerful

Practical vs. statistical significance

- When we say a difference is ***statistically significant***, we mean that it is **unlikely** that a difference/effect/relationship this large would occur simply due to chance (i.e., sampling error)
  - Q: How unlikely?
  - **A: with probability α (usually .05), the *level of significance***
- It doesn't guarantee that it is a large difference
  - For sufficiently large sample sizes, even a tiny difference may be judged to be statistically "significant"
  - {Conversely, when we say an observed difference is **not** statistically significant, we do not necessarily mean that the difference is small.
    - Even quite large differences will not be judged to be statistically "significant" if the sample size is sufficiently small.}
- AND statistical significance does not guarantee that the difference is *practically significant*
  - that the difference is large enough to matter to you
  - often, practical significance is not an empirical question at all, but a value or attitude question.