

Basic Statistics I: Hypothesis Testing

Clinical question

- Can brief physician counseling affect drinking behavior in problem drinkers?

Hypothesis Testing

- **Fact: MD advice affects drinking behavior**
- **Question: How much?**
- **Possibilities:**
 - **Big improvements**
 - **Small improvements**
 - **No improvement (zero effect)**
 - **Behavior exacerbated**

Hypothesis Testing (cont)

- **Possibilities:**
 - Big improvements
 - Small improvements
 - No improvement (zero effect)
 - Behavior exacerbated
- **Hypothesis: Effect is X**
- **Question: Do our observations support the hypothesis?**

What do we mean by “effect”?

- Outcome measure: Reduction in drinks/wk
- Will pts respond identically? (NO)
- Need to consider the effect in a “typical” patient
- How to measure?
 - Average — over many appropriate pts
 - Median

True vs Estimated Effects

- “True” effect is defined to be average of effects for all patients to whom treatment might be applied (population)
- A study gives us the chance to estimate the true effect
- This estimate is called the “observed effect” (in the sample)
- Example: 7.6 dr/wk for Advice vs 3.4 dr/wk for No Advice

True vs Estimated Effects

- Example: 7.6 dr/wk for Advice vs 3.4 dr/wk for No Advice
- “Observed effect” = $7.6 - 3.4 = 4.2$ dr/wk greater reduction with Advice?
- “True effect” = ???
- 4.2 dr/wk is pretty good, but this is only what we observed.
- How small could the true effect *really* be?

Null Hypotheses

- Could the true effect be zero?
- That is, advice doesn't help reduce drinking
- Idea: compare observed outcome to this scenario to answer this question.
- H_0 : True diff = 0 dr/wk

Do Investigators Believe H_0 ?

- The null hypothesis is a starting point that accounts for the observed differences in the simplest possible way—attributing the result to chance alone and to nothing systematic
- Investigator may believe H_0
- Investigator may expect data to disprove H_0

Inference

- How strongly to the data confirm (or contradict) H_0 ?
- Intuitively, the larger the observed effect, the less credible is the “chance” explanation.
- As observed effects increase in size (get further from zero), the hypothesis of “no true effect” becomes decreasingly plausible

The P-value

- ... is a measure of support
- ... for the null hypothesis
- ... derived from the observed data
- ... by comparing the observed effect (4.2 dr/wk)
- ... to the null effect (0 dr/wk)

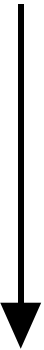
The P-value scale

- **$p = 1.0$: The data are as much in agreement with the null hypothesis as they could possibly be.**
 - Example: If our measured effect were 0 dr/wk, then $p=1.0$
- **$p = 0.3$: Imperfect agreement of data with null, but certainly not inconsistent**
- **$p = 0.05$: The “conventional” boundary for chance implausibility**
- **$p = 0.01$: The data, while not impossible under H_0 , is largely inconsistent with H_0**
- **$p = 0.0$: The data are as far from agreement with the null hypothesis as they could possibly be.**

What is P?

- P depends on the observed outcome
- P = fraction of studies which, by chance alone, would produce data more discrepant from H_0 than that observed in this particular study.

How often is $P < 0.05$?

True Effect	Fraction with $P < 0.05$
none	0.05
tiny	0.06
	
huge	0.999...

A good study

True Effect	Fraction with $P < 0.05$
none	0.05
tiny	0.06
important	reasonably large
↓	↓
huge	0.999...

Effect of sample size

- **Small studies can detect only the largest effects**
- **Enormous studies can detect even the tiniest effects**
- **WANT: studies large enough to detect clinically important effects**
- **NOTE: No relation of statistical significance to clinical importance**

At the time of the 12-month follow-up, there were significant reductions in 7-day alcohol use (mean number of drinks in previous 7 days decreased from 19.1 at baseline to 11.5 at 12 months for the experimental group vs 18.9 at baseline to 15.5 at 12 months for controls; $t=4.33$; $P<.001$),

- Outcome measure. Here comparing use via average change in drinking behavior
- Effect of “screening” + “advice”: $19.1 - 11.5 = 7.6$ drinks/wk
- Effect of “screening”: $18.9 - 15.5 = 3.4$ drinks/wk
- Effect of “advice”: additional reduction of 4.2 dr/wk
- Is this a real effect?

At the time of the 12-month follow-up, there were significant reductions in 7-day alcohol use (mean number of drinks in previous 7 days decreased from 19.1 at baseline to 11.5 at 12 months for the experimental group vs 18.9 at baseline to 15.5 at 12 months for controls; $t=4.33$; $P<.001$),

- Is this a real effect?
- Probably....
- P-values measure strength of the evidence
- ... but not the importance of the result.

Baseline differences

- **Statistical Adjustment**
- **If smoking behavior affects drinking behavior change,**
- **...we want to assess Advice holding smoking behavior constant**

Table 3.—Logistic Regression Model of 20% or More Reduction in Drinking*

Characteristics	Adjusted Odds Ratio (95% Confidence Interval)
Women	1.08 (0.77-1.51)
Smoking in last 6 mo	0.73† (0.53-1.01)
Age, y	
18-30	1.07 (0.83-1.39)
31-40	0.94 (0.73-1.22)
41-50	0.88 (0.66-1.17)
51-65	1.12 (0.83-1.51)
Depressed in last 30 days	0.73 (0.49-1.09)
Experimental	2.15‡ (1.58-2.93)
Child conduct disorder	1.30 (0.87-1.94)
Adult antisocial personality	1.18 (0.67-2.06)

*Overall predictive power, 60%.

† $P \leq .06$.

‡ $P \leq .001$.