# Math 1070-2: Spring 2008

## Lecture 10

Davar Khoshnevisan

Department of Mathematics
University of Utah
http://www.math.utah.edu/~davar

April 2, 2008

# Inference & Testing

- "Does company X practice age discrimination in hiring?"
- "Is this die loaded?"
- "Is this coin fair?" ...
- Consider the last question: [one approach that works here:]
  - Toss the coin 1,000 times [sample of size 1000]
  - Suppose there were 450 heads [$\hat{p} = 450/1000 = 0.45$]
  - SE $\approx \sqrt{\hat{p}(1-\hat{p})/n} \approx 0.0157$
  - A 95% CI for $p$ = true proportion of heads is approximately

$$.45 \pm (1.96 \times 0.0157) = (0.419\,, 0.481)$$

  - "95% sure the coin is *not fair* !" ☹
  - There is a more accurate way [SE is not computed sufficiently accurately for smallish-to-moderate $n$]

# Testing statistical hypotheses

## Buzz words we **must** know

- A [statistical] hypothesis: a statement about a population, usually of the type “parameter such-and-such takes values in this range.” For example:
  - “coin is fair”
  - “hiring practice was fair”
  - “die not loaded” . . .
- A significance test [or “test”] is a method of using data to summarize the evidence regarding a hypothesis.

# Before we begin . . .

. . . your test of significance, you must think about:

1. Assumptions: What are they? Are they met? . . .
2. Hypotheses:
   - (null hyp. or $H_0$): The parameter takes a specific value[s]
   - (alternative hyp. or $H_a$): The parameter takes alternative values
     - Example [proportions]: $H_0: \ p = 1/2$ versus $H_a: \ p > 1/2$
     - Example [means]: $H_0: \ \mu = 2.8937$ versus $H_a: \ \mu \neq 2.8937$
3. Test statistic: $\hat{p}$ for proportions; $\bar{x}$ for means
4. $P$-value: Probab. of rejecting $H_0$ incorrectly
5. Conclusion: report $P$-value and interpret
6. We must do all 5 each time!

# Testing for proportions

## A case study

- $n = 116$ volunteers; astrologer prepared a horoscope for each [Based on date/place of birth etc.]
- Each volunteer also filled out a California Personality Index survey [CPI]
- Each volunteer's birth data and horoscope were shown to an astrologer, together with the results of the CPI for that volunteer and two other randomly-selected volunteers
- The astrologer was asked to find the CPI that matches the birth info/horoscope
- 28 randomly-selected astrologers
- Double-blind study
- **Q:** Are the astrologers doing better than guessing?

# Testing for proportions

## A case study

- Describe $H_0$ and $H_a$
  - $p =$ true proportion of “correct predictions”
  - $H_0 : p = 1/3$ versus $H_a : p > 1/3$
- Results:
  - 116 predictions [one per volunteer]
  - 40 correct predictions
- Assume $H_0$, then proceed with the calculations . . .
- . . . in 5 steps!

# Testing for proportions

## A case study

1. Assumptions:
    - Categorical variables [proportion of correct predictions]
    - random data ✓
    - $n$ large:
        - Assuming $H_0$:
        - $np = 116(1/3) \approx 38.7 > 38 \gg 15$
        - $n(1-p) = 116(2/3) \approx 77.3 > 77 \gg 15$
        - ✓

# Testing for proportions

## A case study

2. Hypotheses:
   - Recall $H_0: p = 1/3$ versus $H_a: p > 1/3$
   - In general $H_0: p = p_0$ versus $H_a: p > p_0$
   - "one-sided" alternative
   - "two-sided" would have been $H_a: p \neq p_0$

# Testing for proportions

## A case study

### 3. Test statistic:

- The test statistic for proportions [$np > 15$ and $n(1-p) > 15$]:

$$z = \frac{\hat{p} - p_0}{\text{SE}_0}$$

- $p_0 =$ proportion if $H_0$ were true
- $\text{SE}_0 =$ SE if $H_0$ were true
- Here:
  - $\hat{p} = 40/116 \approx 0.3448$ [$n = 116$]
  - $p_0 = 1/3$
  - $\text{SE}_0 = \sqrt{p_0(1-p_0)/n} \approx 0.0438$
  - $z = (0.3448 - \frac{1}{3})/0.0438 \approx 0.2618$
  - $z =$ how far is observed prop. [$\hat{p}$] from what $H_0$ predicts [$p$]

# Testing for proportions

## A case study

4. *P*-value:
   - $H_0 : p = \frac{1}{3}$ vs. $H_a : p > \frac{1}{3}$
   - $z \approx 0.26$
   - *P*-value = probab. of seeing this *z*-score or even more
   - Want the right tail of 0.2618 on a normal table [draw the region!]
   - $P\text{-value} \approx 1 - \underbrace{0.6026}_{\text{normal table}} = 0.3974$
   - $\checkmark$

5. Conclusion:
   - Report $P$-value: $0.3974 \approx 40\%$
   - Interpret:
     - If $H_0$ were true, then chances are $\approx 40\%$ that the astrologers would correctly predict 40 [out of 116] or more
     - 40% is a [very] reasonable chance
     - Do not conclude $H_a$
   - If $H_a$ were two-sided [$H_a: \; p \neq \frac{1}{3}$] then double the one-sided $P$-value [$P$-value $\approx 80\%$ . . . draw the normal curve]

# One more buzz word we need to know

## Significance level

- A predescribed number [commonly $0.05 = 1 - 0.95$]
- If $P$-value $\leq$ significance level then $H_0$ makes the observed values "too unlikely"
  - Reject $H_0$
- If $P$-value $>$ significance level then $H_0$ makes the observed values "statistically possible"
  - Do not reject $H_0$
  - $\neq$ "accept" $H_0$ unless $P$-value $\gg$ significance level
  - Same as the court of law
    [if there isn't enough evidence, then "not guilty"]

# $P$-values

- Always report your $P$-value together with


  conclusion/decision
- In astrology case study $P$-value $\approx 0.40 \gg 0.05$; support $p > \frac{1}{3}$

# Summary

1. Assumptions:
   - Categorical var with population proportion $p$
   - Need a random sample
   - Need $n$ large enough
2. Hypothesis ($H_0$ vs. $H_a$)
3. Test statistic $z = (\hat{p} - p_0)/\text{SE}_0$
4. $P$-value:
   - If $H_a: \ p > p_0$ then right-tailed probab
   - If $H_a: \ p < p_0$ then left-tailed probab
   - If $H_a: \ p \neq p_0$ then two-tailed
5. Conclusion: If $P$-value $\leq$ significance level then reject $H_0$. Else, do not. Always explain your decision.

# On the significance levels

Quotation from Freedman, Pisani, Purves (1997):

> *How small does P[-value] have to get before you reject the null hypothesis? . . . many statisticians draw lines at 5% and 1%. If P is less than 5%, the result is "statistically significant, and the null hypothesis is rejected at 5% level"; if P is less than 1%, the result is "highly significant." However, the question is almost like asking how cold it has to get before you are entitled to say, "It's cold." A temperature of* $70^\circ$ *F is balmy,* $-20^\circ$ *F is cold indeed, and there is no sharp dividing line.*
>
> *Logically it is the same with testing. There is no sharp dividing line between probable and improbable results. A P-value of 5.1% means just about the same thing as* 4.9% . . .

# On the significance levels

Quotation from Freedman, Pisani, Purves (1997):

> *. . . However, these two P-values can be treated quite differently, because many journals will only publish results which are "statistically significant"—the 5% line. Some of the more prestigious journals will only publish results which are "highly significant"—the 1% line. These arbitrary lines are taken so seriously that many investigators only report their results as "statistically significant" or "highly signifiant." They don't even bother telling you the value of P, let alone what test they used.*

Interestingly, this absurd practice started because of the way statistical tables were originally [and currently] laid out.