Hypothesis Testing
Announcements

- **Lab 05 - Assessing Models: Examining the Therapeutic Touch**
  - Due tomorrow (03/11)

- **HW05 - Probability, Simulation, Estimation, and Assessing Models**
  - Due tomorrow (03/11)

- **Project 1:**
  - Almost done grading

- **Project 2 (midterm):**
  - Released after spring break

Flexible grading policy for three projects
Complement: be careful

- A = the event of sampling (with replacement) 5 aces in a row from a deck of card. P(A) = ?
  \[ \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} = \frac{1}{52} \]

- What is the complement of A?
  1. Drawing 5 cards and never getting an ace
  2. Drawing 5 cards and not getting 5 aces
B = the event of sampling (with replacement) 5 cards and never getting an ace. \( P(B) = \ ? \)

\[
\frac{48}{52} \times \frac{48}{52} \times \frac{48}{52} \times \frac{48}{52} \times \frac{48}{52} = \frac{48^5}{52}
\]

\[ P(A) = \frac{1^5}{52} ; P(B) = \frac{48^5}{52} \]

Is \( P(A) = 1 - P(B) \)?

\[
P(A) = \frac{1^5}{52} \approx \frac{1}{380M} \\
P(B) = \frac{48^5}{52} \approx \frac{254M}{380M}
\]
- A = the event of sampling (with replacement) 5 aces in a row from a deck of card. P(A) = ?
  \[ \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} = \frac{1}{52} \]

- The complement of A is:
  1. Drawing 5 cards and never getting an ace
  2. P(not A) = 1 - \( \frac{1}{52} \)^5 \approx \frac{380M - 1}{380M}
Large Random Samples
Why bother sampling?

Probability

Statistics

Sampling
Inference

- **Statistical Inference:**
  - Making conclusions based on data in random samples

- **Example:**
  - Use the data to guess the value of an unknown number
  - Create an estimate of an unknown quantity

Copyright © 2016 Barnard College
Terminology

- **Parameter**
  - Numerical quantity associated with the population

- **Statistic**
  - A number calculated from the sample

- A statistic can be used as an estimator of a parameter
Values of a statistic vary because random samples vary.

“Sampling distribution” or “probability distribution” of the statistic:

- All possible values of a statistic
- And all corresponding probabilities

Can be hard to calculate:

- Either have to do math
- Or generate all possible samples and calculate the statistic based on each sample
Empirical Distribution of a Statistic

- Based on simulated values of a statistic
- Consists of all observed values of the statistic, and the proportion of times each value appeared

- Good approximation to the probability distribution of a statistic
  - If the number of repetitions in the simulation is large
Hypothesis Testing
Choosing Between Two Viewpoints

Based on data:
- “Chocolate has no effect on cardiac disease”
- “Yes, it does”

Questions that we will consider:
- Were data was drawn?
- How the data was drawn?
- What can we conclude from the data?
A model is a set of assumptions about the data

In data science, many models involve assumptions about processes that involve randomness:
- “Chance models”

Key question: does the model fit the data?
Approach to Assessing Models

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts.

- We can compare the model’s predictions to the observed data.

- If the data and the model’s predictions are not consistent, that is evidence against the model.
Swain vs. Alabama, 1965

- Talladega County, Alabama
- Robert Swain, black man convicted of crime
- Appeal: one factor was all-white jury
- Only men 21 years or older were allowed to serve
- 26% of this population were black
- Swain’s jury panel consisted of 100 men
- 8 men on the panel were black
About disparities between the percentages in the eligible population and the jury panel, the Supreme Court wrote:

- “... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negros”

Supreme Court denied Robert Swain’s appeal
**Paraphrase:** 8/100 is less than 26%, but not different enough to show Black men were systematically excluded

**Question:** is 8/100 a realistic outcome if the jury panel selection process were truly unbiased?
Sampling from a Distribution

- Sample at random from a categorical distribution

\[
sample\_proportions(sample\_size, pop\_distribution)
\]

- Samples at random from the population
  - Returns an array containing the distribution of the categories in the sample
Steps in Assessing a Model

- Choose a statistic that will help you decide whether the data support the model or an alternative view of the world.
- Simulate statistic under the assumptions of the model.
- Draw a histogram of the simulated values.
  - This is the model’s prediction for how the statistic should come out.
- Compute the statistic from the sample in the study.
  - If the two are not consistent => evidence against the model.
  - If the two are consistent => data supports the model so far.
Mendel’s genetic model

- Pea plants of a particular kind
- Each one has either purple flowers or white flowers

Mendel’s model:
  - Each plant is purple-flowering with chance 75%, regardless of the colors of the other plants

Question:
  - Is the model good or not?
Choose a Statistic

- Take a sample, see what percent are purple-flowering
- If that percent is much larger or much smaller than 75, that is evidence against the model
- **Distance** from 75 is key
- Statistic:
  - $| \text{sample percent of purple-flowering plants} - 75 |$
- If the statistic is large, that is evidence against the model
Jury Selection:
- **Model:** The people on the jury panels were selected at random from the eligible population
- **Alternative viewpoint:** No, they weren’t

Genetics:
- **Model:** Each plant has a 75% chance of having purple flowers
- **Alternative viewpoint:** No, it doesn’t
Steps in Assessing a Model

- Choose a statistic to measure the “discrepancy” between model and data
- Simulate the statistic under the model’s assumptions
- Compare the data to the model’s predictions:
  - Draw a histogram of simulated values of the statistic
  - Compute the observed statistic from the real sample
- If the observed statistic is far from the histogram, that is evidence against the model
Homework

- Reading 11.2 on your own
  - Multiple Categories

- Tomorrow’s lecture:
  - 11.3 – 11.4
  - A/B Testing (Chapter 12)