K. Way statistics behave:

1. If statistic is sample average,
   a. then mean of the sampling distribution is population average
   b. then sd of the sampling distribution is population sd/ $\sqrt{n}$.
      i. Variance of sum of independent observations is sum of variances
      ii. Dividing by $n$ divides variance of sum by $n^2$.
   c. shape of sampling distribution is exactly normal when original distribution was exactly normal

2. Law of large numbers:
   a. If
      i. we observe a process many times
      ii. observations don’t influence each other: independent
      iii. Average values of a random variable
   b. then our statistic will get closer and closer to average of population,

3. Shape of sampling distribution: normal: *central limit theorem* (CLT)
4. Example:
   a. Hypothetical population of size 20,000 with mean $\mu$ and $\sigma$
      matching our sample
   b. Draw billion samples with replacement of same size $n$ as
      our sample, and calculate average: histogram will represent
      sampling distribution,
   c. will have mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.
      i. By rule, with a billion samples,
         • 68% of such samples will lie between $\mu - 1 \times \sigma/\sqrt{n}$ and
            $\mu + 1 \times \sigma/\sqrt{n}$
         • 95% of such samples will lie between $\mu - 2 \times \sigma/\sqrt{n}$ and
            $\mu + 2 \times \sigma/\sqrt{n}$

5. Counterexample: Always choose same person.


7. What do we need to make CLT work? Combination of:
   a. Sample size large ($> 30$ unless otherwise noted)
   b. Population not too far off to start with.

8. Example: Fig. 31/.
   a. Note that the more die thrown, the closer the histogram comes
Fig. 31: Central Limit Theorem as Applied to a Discrete Distribution

Distribution of mean of 1 die

Distribution of mean of 2 dice

Total Distribution of mean of 4 dice

Total Distribution of mean of 8 dice
9. Fig. 32/ contains a continuous example.

L. Once we know that the shape is normal, and once we know the center and spread, we can calculate proportions in various ranges.

1. Example: mean of 30 from skewed distribution
   a. mean of population is 2.1
   b. sd of pop is 2.05
   c. mean of sampling distribution of average is 2.1
   d. sd of sampling distribution of average is \( \frac{2.05}{\sqrt{30}} = 0.374 \)
   e. Probability of seeing average between 2 and 3 is given from normal table at \( \frac{3 - 2.10}{0.374} = 2.41 \) and \( \frac{2 - 2.10}{0.374} = -0.27 \).
   f. From normal table, the probability to the left of these two values are .99 and 0.39 resp.
   g. Probability between these is .99 -.39 = .60

2. Example: mean of 4 Dice
   a. mean is 3.5
   b. sd of one die is \( \sqrt{(1 - 3.5)^2 \times \frac{1}{6} + (2 - 3.5)^2 \times \frac{1}{6} + (3 - 3.5)^2 \times \frac{1}{6}} \)
   \( \sqrt{(5/2)^2 \times (1/3) + (3/2)^2 \times (1/3) + (1/2)^2 \times (1/3)} \) =
Fig. 32: Central Limit Theorem as Applied to a Continuous Distribution

Density of raw observations

Potential data values

Distribution of mean of 5 obs

Potential mean values

Distribution of mean of 15 obs

Potential mean values

Distribution of mean of 30 obs
\[
\sqrt{\frac{35}{12}} = 1.71.
\]

c. SD of average is \(\frac{1.71}{\sqrt{4}} = .85\).

d. Normal approximation to probability that average \(\bar{X}\) is between 3 and 4 inclusive:
\[
P(3 \leq X \leq 4) = P\left(\frac{(3 - 3.5)/0.85 \leq Z \leq (4 - 3.5)/0.85}{0.85}\right) = P(-.59 \leq Z \leq .59) = 0.722 - 0.278 = 0.444
\]

D&P: 8.3

M. A particular discrete distribution: sample proportion.

1. Ex, biological process either cures or does not cure people who appear to have the same initial condition.

2. We want to know \(\pi\), proportion of time cure happens: parameter

3. We observe proportion in our sample: statistic

4. Takes values on \(0, 1/n, 2/n, \ldots, 1\).

   a. Could calculate probabilities for each of these probabilities

   b. If we had probabilities, we could use formulas from last class to give mean and variance.

5. For each person, construct random variable with value 1 if cured, value 0 if not.

6. Calculate sampling distribution of mean:
a. mean of random variable is \( 0 \times (1 - \pi) + 1 \times \pi = \pi \):
   proportion is population mean.

b. statistic is \( p = \) sum of random variables divided by sample size:
   sample average

c. By rule from last class, average of sampling distribution for \( p \) is
   the right answer, \( \pi \).

7. Calculate variance for sampling distribution:

a. variance of random variable is
   \[
   \begin{align*}
   & (0 - \pi)^2 \times (1 - \pi) + (1 - \pi)^2 \times \pi = \\
   & \pi^2 (1 - \pi) + (1 - \pi)^2 \pi = \pi (1 - \pi)(1 - \pi + \pi) = \pi (1 - \pi)
   \end{align*}
   \]

b. Recall rule: If statistic is mean of \( n \) observations, independent
   with the same variance, then variance of statistic is variance of
   observations/ \( n \)

i. variance of proportion is \( \pi (1 - \pi)/n \).

ii. sd of proportion is \( \sqrt{\pi(1 - \pi)/n} \).

c. Behavior

i. as sample gets larger, sd gets smaller

ii. To cut sd in half, must quadruple \( n \)

d. Example:

i. Population of 10000, ask people whether they approve or
disapprove of Clinton. Suppose 4800 approve of Clinton

ii. samples with replacement of 10 people; sampling
distribution of proportion approving has mean .48, sd
\[ \sqrt{.48 \times .52/10} = .158. \]

iii. samples with replacement of 100 people; sampling
distribution of proportion approving has mean .48, sd
\[ \sqrt{.48 \times .52/100} = .050. \]

iv. samples with replacement of 1000 people; sampling
distribution of proportion approving has mean .48, sd
\[ \sqrt{.48 \times .52/1000} = .016. \]

v. samples without replacement of 10000 people; sampling
distribution of proportion approving has mean .48, sd 0.

vi. Shows that when taking survey without replacement,
get answer too big if you are asking too large a part of the
population: \( > 10\% \).

vii. Independence rule disobeyed.

8. Example: In polling example above, if you poll 1000 and the true
level of Clinton support is 48\%, you should see results between
44.8 and 51.2\% 95\% of time.
Lecture 11

9. Can use normal approx if $n\pi \geq 5$, $n(1 - \pi) \geq 5$
   
   a. need enough observations
   
   b. Need not too lopsided

   D&P: Appendix I

10. What if we consider count rather than proportion (ie, don’t pre-divide by sample size)?
   
   a. Takes values on 0, 1, ..., $n$.
   
   b. Called binomial distribution (number of draws $n$, success probability $\pi$).
   
   c. Mean is $n\pi$
   
   d. Variance is $n\pi(1 - \pi)$.