3 Some Basic Concepts of Statistics

3.1 Introduction

Reminder: Surveys are used to make inferences about populations.

Often, taking a census is expensive and time-consuming, so instead we take a survey.

3.2 Summarizing Information in Populations and Samples: The Infinite Population Case

Example: Fair four-sided die.

Outcomes are 1, 2, 3, and 4.

Let Y be the outcome from one roll.

Find the mean of Y.

Example: Unfair four-sided die.

Assume probabilities are

$p(1) = 0.3,\ p(2) = 0.4,\ p(3) = 0.2,\ p(4) = 0.1.$

Find the mean of Y.

Using *R*:

```
> y = 1 :  4
```

```
> probs = c( 0.3, 0.4, 0.2, 0.1 )
```

Compute the population variance of Y

$\sigma^2 = V(Y) = E(Y-\mu)^2 = \sum_y (y-\mu)^2 \; p(y)$

Alternative formula: $\sigma^2 = \sum_y y^2 \; p(y) - \mu^2$

The population standard deviation is $\sigma = \sqrt{\sigma^2}$.

Suppose the $p(y)$ values are all equal.

Estimate μ, σ^2, and σ from their sample values.

For data set $Y_1, Y_2, \ldots, Y_n$:

How should μ be estimated?

> ### Consider data regarding sports drinks from exercise 3.10 on p. 68.

```
> cost = c( .22, .24, .26, .34, .26, .52, .22, .24, .35 )
```

Compute the sample mean cost.

Estimate σ^2 by the sample variance,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Alternative formula for sample variance,

$$s^2 = \left[\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i \right)^2 / n \right] \; / \; (n-1)$$

The sample *mean* is unbiased for the population mean; i.e., $E\bar{Y} = \mu$ (when μ is finite).

The sample *variance* is unbiased for the population variance; i.e., $Es^2 = \sigma^2$ (when σ^2 is finite and observations are independent).

Estimate σ by the sample standard deviation, $s = \sqrt{s^2}$.

Is s unbiased for σ (when $0 < \sigma^2 < \infty$ and observations are independent)?

Example: Consider the exponential population with mean one. The probability density function is $f(x) = e^{-x}$, for $x > 0$.

(a) Plot the probability density function.

(b) Based on the sample mean for a large sample size, estimate the population mean.

(c) Sample 10,000 values of $\bar{Y}$, where each value of $\bar{Y}$ is based on five independent observations from the population.

(d) Determine the mean of your 10,000 values of $\bar{Y}$.

(e) Based on the sample standard deviation and sample variance for a large sample size, estimate the population standard deviation and population variance of this population.

(f) Sample 10,000 values of s^2, where each value of s^2 is based on five independent observations from the population.

(g) Determine the mean of your 10,000 values of s^2.

(h) Sample 10,000 values of s, where each value of s is based on five independent observations from the population.

(i) Determine the mean of your 10,000 values of s.

(j) Repeat parts (h) and (i) using $n = 50$.

□

Example: Revisit unfair die problem.

```
> # Population is the following:

> pop = c( 1, 1, 1, 2, 2, 2, 2, 3, 3, 4 )
```

```
> # Compute population mean and population variance.
```

```
> # Estimate the population mean and population variance, by using a
   large sample.
```

The mean and variance of the sample mean, $\bar{Y}$

Sampling WITH replacement from this weighted finite population is analogous to sampling independently from an infinite population.

Example:

(a) Roll the unfair die 5 times (i.e., sample with replacement).

```
> pop = c( 1, 1, 1, 2, 2, 2, 2, 3, 3, 4 ) # 'pop' is the population.
```

(b) Obtain the sample mean $\bar{Y}$.

(c) Repeat parts (a) and (b) 10,000 times, and assign your 10,000 realizations of $\bar{Y}$ to a variable.

(d) Determine the (population) mean of $\bar{Y}$.

(e) Determine the (population) variance of $\bar{Y}$.

(f) Determine the sample mean of your 10,000 realizations of $\bar{Y}$.

(g) Determine the sample variance of your 10,000 realizations of $\bar{Y}$.

□

Homework C3.2.1*: Consider an infinite population with 24.7% values of 3; 36.1% values of 5, and 39.2% values of 8.

(a) Determine the mean of this population.

(b) Determine the standard deviation of this population.

(c) If $\bar{Y}$ is based on a simple random sample of size 25 from this population, determine the mean of $\bar{Y}$. *Hint: No random number generation is needed here.*

(d) If $\bar{Y}$ is based on a simple random sample of size 25 from this population, determine the standard deviation of $\bar{Y}$. *Hint: No random number generation is needed here.*

(e) Construct the line graph of the original population.

(f) Construct the line graph of a simple random sample of 100 observations from this population. Compare your graph to the graph in part (e).

(g) Sample 30,000 independent values of $\bar{Y}$, such that each value of $\bar{Y}$ is based on a simple random sample of size 25. *Print just the source code, NOT the 30,000 data values.*

(h) Compute the *mean* of your 30,000 values of $\bar{Y}$, and compare this answer to your answer from part (c).

(i) Compute the *standard deviation* of your 30,000 values of $\bar{Y}$, and compare this answer to your answer from part (d).

End of Homework C3.2.1* □

Homework p. 70: Exercises 3.13, 3.18a (Use first 4 states only, and ignore "taken with probabilities . . . states".), 3.21

3.3 Summarizing Information in Populations and Samples: The Finite Population Case

Unlike the unfair die example, we sample withOUT replacement:

Hence, we are drawing withOUT replacement from a deck of 10 cards.

(a) Draw 5 cards withOUT replacement.

```
> pop = c( 1, 1, 1, 2, 2, 2, 2, 3, 3, 4 )
```

(b) Obtain the sample mean $\bar{Y}$.

(c) Repeat parts (a) and (b) 10,000 times, and assign your 10,000 realizations of $\bar{Y}$ to a variable.

(d) Determine the (population) mean of $\bar{Y}$.

(e) Determine the average of your 10,000 realizations of $\bar{Y}$.

(f) Determine the sample variance of your 10,000 realizations of $\bar{Y}$.

□

Homework C3.3.1: Consider the following population: {2, 2, 2, 6, 6, 9, 9, 11, 12, 13}.

(a) Determine the mean of this population.

(b) Determine the standard deviation of this population.

(c) If $\bar{Y}$ is based on a simple random sample withOUT replacement of size 6 from this population, determine the population mean of $\bar{Y}$. *Hint: No random number generation is needed here.*

(d) Construct the line graph of the original population.

(e) Construct the line graph of a simple random sample of 6 observations withOUT replacement from this population. Compare your graph to the graph in part (d).

(f) Sample 20,000 independent values of $\bar{Y}$, such that each value of $\bar{Y}$ is based on a simple random sample withOUT replacement of size 6. *Print just the source code, NOT the 20,000 data values.*

(g) Compute the *mean* of your 20,000 values of $\bar{Y}$, and compare this answer to your answer from part (c).

(h) Compute the *standard deviation* of your 20,000 values of $\bar{Y}$.

End of Homework C3.3.1 □

Homework p. 71: Exercise 3.18b* (Use first 4 states only, and ignore "taken with probabilities . . . states".)

Hints for homework exercise 3.18b*: We have four states (Connecticut, Maine, Massachusetts, and New Hampshire) with corresponding numbers 42, 17, 69, and 15. Use the first four states only, and ignore "taken with probabilities . . . states." There are six ways to sample exactly two (n=2) of these four (N=4) states. For each of these six ways, estimate the population total number of teachers using `N * y.bar`, where N=4 (the number of states) and `y.bar` is the average number of total teachers for the two states sampled. Hence, you have six estimates, `N * y.bar`, of the population total. Show that the average of your six estimates is equal to the population total (i.e., 42+17+69+15). This shows that your estimator is unbiased for the population total. You should NOT use the function `sample` at all in this exercise. After you complete the math, explain in English what you showed, using complete sentences.

3.4 Sampling Distributions

First consider the **normal distribution**.

In a large sample, approximately what proportion of the observations from a $N(\mu, \sigma)$ distribution fall within 1.96 standard deviations of the mean?

```
> x = rnorm( 50 ) # Generate 50 observations from N(0,1).
> ( -1.96 < x) * (x < 1.96 )
> mean( ( -1.96 < x) * (x < 1.96) )
```

For finite population mean μ, the sample mean is unbiased for μ, regardless of the sample size.

$E\bar{Y} = \mu$

For finite population standard deviation σ, the standard deviation of the sample mean is $\sigma/\sqrt{n}$, when the observations are independent.

$SD(\bar{Y}) = \sigma/\sqrt{n}$

> ### Revisit unfair die problem (i.e., sample with replacement).

Probabilities are

$p(1) = 0.3,\ p(2) = 0.4,\ p(3) = 0.2,\ p(4) = 0.1.$

```
> # Graph the population histogram.
> hist(pop)
```

```
> # Consider the distribution of $\bar{Y}$ when rolling the die 5 times.
```

$E\bar{Y} = ?$ $\quad$ $\mathrm{SD}(\bar{Y}) = ?$

First, roll the die 5 times (i.e., sample with replacement).

```
> sample( pop, 5, TRUE )
```

```
> # Construct the histogram of $\bar{Y}$ (approximately).
```

Simulate 60,000 values of $\bar{Y}$.

Central Limit Theorem: For independent observations and large n (and positive finite σ), the sample mean (and sample sum) is approximately normal.

Also, for independent approximately **normal** *observations and* **any** n*, the sample mean (and sample sum) is approximately normal.*

```
> # Repeat the previous example of unfair die for a sample of size 30.
```

$E\bar{Y} = ?$ $\quad$ $\mathrm{SD}(\bar{Y}) = ?$

What is the approximate distribution of the sample mean when rolling the die 30 times?

```
> hist(die.means) # Construct the histogram.
```

Sample sums also are approximately normal, for large n and positive finite σ.

```
> # Determine the exact proportion of your 60,000 values of $\bar{Y}$ which
  fall within $1.96\sigma/\sqrt{n}$ of $\mu$.
> # In other words, estimate $P(\mu - 1.96\sigma/\sqrt{n} < \bar{Y} < \mu + 1.96\sigma/\sqrt{n})$ using the
  computer but withOUT using a normal approximation.
```

Homework C3.4.1:

(a) Revisit homework C3.2.1. Graph your 30,000 values of $\bar{Y}$ in a histogram. Discuss the shape of your histogram.

(b) Revisit homework C3.3.1. Graph your 20,000 values of $\bar{Y}$ in a histogram. Discuss the shape of your histogram.

End of Homework C3.4.1 □

Homework p. 68: Exercises 3.4, 3.20* (Use 10,000 sample means. Hint: Read in SCHOOLS.txt and use `replicate`.)

3.5 Covariance and Correlation

The *population correlation coefficient*, ρ, measures the linear relationship between two variables x and y.

The *sample correlation coefficient*, $\hat{\rho}$, estimates ρ, and is based on a sample of pairs of observations (x, y).

Example:

```
> x1 = 1:20

> y1 = 4 * x1 - 30

> plot( x1, y1 )

> cor( x1, y1 )

> y2 = -y1


> y3 = y1 + rnorm( 20, 0, 5 )


> y4 = y1 + rnorm( 20, 0, 10 )


> y5 = y1 + rnorm( 20, 0, 20 )


> x2 = 10 * x1 + 15

> cor( x2, y5 )


> x3 = -10 * x1 + 15

> cor( x3, y5 )
```

Interpretation: ρ^2 measures the proportion of variability in y that can be explained by x, when the (x, y) data are linear.

```
> x6 = rnorm( 5000, 20, 3 )

> y6 = rnorm( 5000, -90, 15)


> x7 = rnorm( 5000, 20, 3 )
> y7 = x7 + rnorm( 5000, 10, 3 )

> plot( x7, y7 )


> x8 = 30:80

> y8 = -3 * ( x8 - 55) ^ 2 + 3000 + rnorm( 51, 0, 100 )
```

Summary of $\hat{\rho}$ and ρ

(1) $-1 \leq \hat{\rho} \leq 1$ and $-1 \leq \rho \leq 1$.

(2) $|\hat{\rho}|$ and $|\rho|$ are not affected by linear transformations on the data.

(3) $\hat{\rho}$ and ρ measure linear data only.

(4) When $\hat{\rho} > 0$, the data are *positively correlated.*

(5) When $\hat{\rho} < 0$, the data are *negatively correlated.*

(6) Zero correlation does not imply no association.

Homework p. 71: Exercise 3.15 ('SAT.txt')

3.6 Estimation

A *parameter* is a numerical descriptive measure of a population.

Recall: An *estimator* is a function of observable random variables, and is used to estimate a parameter.

Example: Consider a simple random sample of size n from a population of size N. Let μ be the (unknown) population mean, τ be the (unknown) population total, and $\bar{Y}$ be the sample mean.

(a) How should τ be estimated?

(b) Determine whether or not this estimator is biased.

□

Notation: The estimator $\hat{\theta}$ (based on the data) is used to estimate the unknown parameter θ (based on the population).

Desirable properties of estimators (small or no bias, and small variability):

(1) unbiasedness: $E\hat{\theta} = \theta$

> ### Revisit unfair die problem.

Probabilities are

$p(1) = 0.3,\ p(2) = 0.4,\ p(3) = 0.2,\ p(4) = 0.1.$

(a) Roll the die 5 times to obtain $\bar{Y}$. Replicate this procedure to produce 60,000 values of $\bar{Y}$. Average all 60,000 values of $\bar{Y}$.

```
> die.means = replicate( 6e4, mean( sample( 1:4, 5, replace=TRUE,
   prob=c(0.3, 0.4, 0.2, 0.1) ) ) )
> die.means = replicate( 6e4, mean( sample( pop, 5, replace=TRUE ) ) )
   # Alternatively.
> mean( die.means )
```

What concept is being demonstrated here?

(b) Roll the die 5 times to obtain the sample variance s^2. Replicate this procedure to produce 60,000 values of s^2. Average all 60,000 values of s^2.

```
> die.vars = replicate( 6e4, var( sample( 1:4, 5, replace=TRUE,
   prob=c(0.3, 0.4, 0.2, 0.1) ) ) )
> die.vars = replicate( 6e4, var( sample( pop, 5, replace=TRUE ) ) )
   # Alternatively.
> mean(die.vars)
```

What concept is being demonstrated here?

(c) Roll the die 5 times to obtain the sample standard deviation s. Replicate this procedure to produce 60,000 values of s. Average all 60,000 values of s.

```
> die.sds = replicate( 6e4, sd( sample( 1:4, 5, replace=TRUE,
   prob=c(0.3, 0.4, 0.2, 0.1) ) ) )
> die.sds = replicate( 6e4, sd( sample( pop, 5, replace=TRUE ) ) )
   # Alternatively.
> die.sds = sqrt( die.vars ) # Alternatively.
> mean(die.sds)
```

What concept is being demonstrated here?

□

(2) small variability: $V(\hat{\theta}) = \sigma^2_{\hat{\theta}}$

What is $V(\bar{Y})$, for independent observations?

```
> ### Revisit unfair die problem.
```

(a) Roll the die **5** times to obtain the sample mean $\bar{Y}$. Replicate this procedure to produce 60,000 values of $\bar{Y}$. Determine the standard deviation of all 60,000 values of $\bar{Y}$.

```
> die.means1 = replicate( 6e4, mean( sample( 1:4, 5, replace=TRUE,
   prob=c(0.3, 0.4, 0.2, 0.1) ) ) )
```

```
> die.means1 = replicate( 6e4, mean( sample( pop, 5, replace=TRUE ) ) )
   # Alternatively.
> sd( die.means1 )
```

What concept is being demonstrated here?

(b) Roll the die **20** times to obtain the sample mean $\bar{Y}$. Replicate this procedure to produce 60,000 values of $\bar{Y}$. Determine the standard deviation of all 60,000 values of $\bar{Y}$.

```
> die.means2 = replicate( 6e4, mean( sample( 1:4, 20, replace=TRUE,
   prob=c(0.3, 0.4, 0.2, 0.1) ) ) )
> die.means2 = replicate( 6e4, mean( sample( pop, 20, replace=TRUE ) ) )
   # Alternatively.
> sd( die.means2 )
```

What concept is being demonstrated here?

What concept is being demonstrated from comparing `sd( die.means1)` with `sd( die.means2 )`?

(c) Compute the ratio of your two answers from parts (a) and (b).

What concept is being demonstrated here?

□

When determining the quality of an estimator, your textbook typically does not compute bias, since the bias of a *reasonable* estimator typically decreases at a fast rate as the sample size increases.

For many types of situations when the sample size is large, the standard deviation of the estimator is larger than the absolute value of the bias.

The *error of estimation* is $|\hat{\theta} - \theta|$.

Many (but not all) estimators $\hat{\theta}$ are approximately normal for large sample sizes.

Sometimes some other distribution (e.g., t-distribution) gives a better approximation, when the estimator is *standardized* in some sense.

Confidence intervals often may be constructed using z- or t-tables.

Homework p. 68: Exercises 3.2, 3.6, 3.8, 3.10, 3.12

3.7 Summary

The *mean* measures centrality.

The *standard deviation* and *variance* measure spread.

A *probabilist* uses a population to describe properties of samples.

A *statistician* uses a sample to make inferences about a population.