

3 Some Basic Concepts of Statistics

3.1 Introduction

Reminder: Surveys are used to make inferences about populations.

Review graphing and numerical techniques from Math 220 (or Math 318).

Graphing Techniques

STRIP CHARTS & LINE GRAPHS:

```
> x1 = c( 60:99, 70:89, 75:84, 78:81 )
```

```
> x1
```

```
> stripchart(x1, "stack")
```

```
> linegraph(x1) # Somewhat similar to 'stripchart'.
```

A distribution is *symmetric* if the left half of the distribution looks like the mirror image of the right half.

A distribution is *left-skewed* if it has a long left tail.

```
> x2 = c(x1, 51, 56, 42, 35, 20)
```

A distribution is *right-skewed* if it has a long right tail.

```
> x3 = c(3:149, 11:128, 20:98, 30:77, 37:39, 154:158, 167:169, 178)
```

STEM-AND-LEAF PLOTS:

```
> stem(x1)
```

```
> stem(x2) # left-skewed histogram
```

```
> stem(x3) # right-skewed histogram
```

HISTOGRAMS:

Rules for constructing a histogram.

- (1) Divide the interval covered by the range of the data into a number of intervals.
- (2) Count the number of data points in each of these intervals.
- (3) Construct a bar over each segment so that the **area** of the bar is proportional to the number of data points inside the interval.

```
> hist(x1)
```

A *relative frequency* or *probability* histogram has total area = 1 (noting rule #3), so that the proportion of observations within an interval is **equal** to the **area** of that interval.

```
> hist(x1, prob=T) # For a relative frequency or probability histogram.
```

What proportion of the observations are ≤ 70 ?

Example: Normal distribution.

- (a) Plot the probability density function (pdf) of a Normal($\mu = 50$, $\sigma = 10$) population, and discuss skewness in this *population* histogram.
- (b) Shade in the above pdf from 30 to 70 using *R*.
- (c) Generate 1000 random variates from a Normal($\mu = 50$, $\sigma = 10$) population.
- (d) Construct a histogram of your 1000 random variates, and discuss skewness in this *sample* histogram.
- (e) Repeat parts (c) and (d) using 100,000 random variates.

□

Example: Chi-square distribution.

- (a) Plot the probability density function (pdf) of a chi-square population with one degree of freedom, and discuss skewness in this *population* histogram.

> ?dchisq

- (b) Shade in the above pdf from 0 to 4 using *R*.
- (c) Generate 1000 random variates from a chi-square population with one degree of freedom.
- (d) Construct a histogram of your 1000 random variates, and discuss skewness in this *sample* histogram.

(e) Repeat parts (c) and (d) using 100,000 random variates.

□

BOX PLOTS

The *median* of a data set is middle observation when the observations are ordered from smallest to largest. If the number of observations is even, then the *median* is the average of the two middle observations.

Example: Find the median of the observations {4, 7, 5, 3, 1, 13, 10}.

Try the *R* command `sort()`.

Example: Find the median of the observations {4, 7, 3, 1, 13, 10}.

The *lower quartile* of a data set is the median of the ordered observations to the *left* of and including the position of the overall median.

The *upper quartile* of a data set is the median of the ordered observations to the *right* of and including the position of the overall median.

The *five-number summary* is (minimum, lower quartile, median, upper quartile, maximum).

The *interquartile range* (IQR) is the upper quartile minus the lower quartile.

An observation is said to be an *outlier* if it is at least $1.5 \times \text{IQR}$ away from its nearest quartile.

Constructing a boxplot (which includes outliers):

- (1) Draw a number line including the minimum and maximum observations.
- (2) Draw a rectangle connecting the lower and upper quartiles.
- (3) Draw a line through the rectangle at the sample median.
- (4) Draw the whiskers, where the left whisker is a line connecting the left edge of the rectangle to the smallest observation which is not an outlier, and the right whisker is a line connecting the right edge of the rectangle to the largest observation which is not an outlier.
- (5) Draw a circle or star at each outlier.

Example: For the data set below, construct the boxplot: {78, 87, 63, 85, 81, 94, 71, 30, 74, 90, 59, 67}.

Comparing medians and means in previous data sets x1, x2, x3.

SCATTER PLOTS:

Example: Enter data from exercise 3.10 on p. 68, regarding calories and cost for sports drinks.

```
> calories = c(60, 70, 60, 70, 50, 66, 60, 67, 80)
```

```
> cost = c(0.22, 0.24, 0.26, 0.34, 0.26, 0.52, 0.22, 0.24, 0.35)
```

```
> plot(calories, cost)
```

Homework C3.1.1: Using the USPOP data set from appendix D, pp. 446-447, and using *R*, list your source code and your output.

- (a) Scan all of the data into a matrix. You do NOT need to print this matrix.
- (b) Create a new variable for “percent in poverty”, and print the values of this new variable.
- (c) Construct a *stripchart* for percent in poverty.
- (d) Construct a *linegraph* for percent in poverty.
- (e) Construct a *stem-and-leaf plot* for percent in poverty.
- (f) Construct a *frequency histogram* for percent in poverty using *R*.
- (g) Construct a *relative frequency histogram* for percent in poverty.
- (h) Based on the above histogram, do the data appear *symmetric, somewhat left-skewed, or somewhat right-skewed*? **EXPLAIN.**
- (i) Construct a *boxplot* for percent in poverty using *R*.
- (j) Plot population total (i.e., the *y*-axis) versus percent in poverty (i.e., the *x*-axis) for all 50 states.
- (k) Based on your answer to part (j), explain whether or not population size seems to be related to the percent in poverty.

End of Homework C3.1.1. □

Homework C3.1.2:

- (a) Plot the probability density function (pdf) of a chi-square population with five degrees of freedom, and discuss skewness in this *population* histogram.
- (b) Shade in the above pdf from 0 to 10 using *R*.
- (c) Generate 300 random variates from a chi-square population with five degrees of freedom, and print these 300 random variates.

- (d) Construct a histogram of your 300 random variates, and discuss skewness in this *sample* histogram.
- (e) Repeat parts (c) and (d) using 100,000 random variates, but do NOT print these 100,000 random variates.

End of Homework C3.1.2. \square

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, \dots, Y_n :

How should μ be estimated?

```
> ### Revisit 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).

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

Is s unbiased for σ ?

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) What is the mean of this population?

(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) What are the standard deviation and 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.

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

Rolling the above unfair die many times is conceptually equivalent to sampling from what type of 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 then average your 10,000 realizations of \bar{Y} .

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

- (e) Determine the (population) variance of \bar{Y} .
- (f) 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. *Just print 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 □

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

SIMILAR to unfair die example, but 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 then average your 10,000 realizations of \bar{Y} .

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

(e) 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. *Just print 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 □

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 σ/\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.
```

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

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

> sample(pop, 5, T)

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

Simulate 10,000 values of \bar{Y} .

Central Limit Theorem: For independent observations and large n (and 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.

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

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 finite σ .

- > hist(die.sums) # Construct the histogram.
- > # Determine the exact proportion of your 10,000 values of \bar{Y} which fall within $1.96\sigma/\sqrt{n}$ of μ .
- > # 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 □

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 - 15
```

```
> plot(x1, y1)
```

```
> cor(x1, y1)
```

```
> y2 = -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.

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

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

```
die.means = replicate( 1e4, mean( sample( 1:4, 5, replace=T, prob=c(0.3, 0.4, 0.2, 0.1)
) ) )
```

```
> mean(die.means)
```

What concept is being demonstrated here?

```
> die.vars = replicate( 1e4, var( sample( 1:4, 5, replace=T, prob=c(0.3, 0.4, 0.2, 0.1) )
) )
```

```
> mean(die.vars)
```

What concept is being demonstrated here?

```
> die.sds = replicate( 1e4, sd( sample( 1:4, 5, replace=T, prob=c(0.3, 0.4, 0.2, 0.1) ) ) )
```

```
> mean(die.sds)
```

What concept is being demonstrated here?

(2) **small variability:** $V(\hat{\theta}) = \sigma_{\hat{\theta}}^2$

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

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

```
> sd(die.means1)
```

What concept is being demonstrated here?

```
> die.means2 = replicate( 1e4, mean( sample( 1:4, 20, replace=T, prob=c(0.3, 0.4, 0.2,
0.1) ) ) )
```

```
> sd(die.means2)
```

What concept is being demonstrated here?

What concept is being demonstrated from comparing $\text{sd}(\text{die.means1})$ with $\text{sd}(\text{die.means2})$?

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 pace 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.

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.