

**Unit 1 – Review of PubHlth 540, *Introductory Biostatistics*
Practice Problems**

SOLUTIONS

1. Recall that variables can be of different types. We learned in introductory biostatistics that appropriate methods of summarization depend on the variable type. And we noticed that, sometimes, a method of summarization is *not* appropriate. For example, it is not appropriate to construct a cumulative frequency graph summary of nominal data.

Complete the following table.

Random Variable				
	Discrete		Continuous	
	Nominal	Ordinal	Interval	Ratio
Descriptive Methods	Bar chart Pie chart - -	Bar chart Pie chart - -	- - Dot diagram Scatter plot (2 variables) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot	- - Dot diagram Scatter plot (2 variables) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot
Numerical Summaries	Freq Table Rel Freq Table	Freq Table Rel Freq Table Cum Freq Table Cum Rel Freq Table	- - - - - means, variances, percentiles	- - - - - means, variances, percentiles

2. Try your hand at the following probability exercises.

- a) Divide a line segment into three parts such that one portion is half the length of original line and the other two portions are each one quarter then length of the original line. Choose a point at random. What is the probability that this point is in the $\frac{1}{2}$ length portion?

Answer: .5

Solution:

As all points along the line segment are equally likely, length is proportional to probability. Thus, $\frac{1}{2}$ length corresponds to $\frac{1}{2}$ of the total probability.

- b) If there is a 14% chance that any person selected at random was born on a Monday, what is the probability that, of any seven people selected at random, exactly one was born on a Monday?

Answer: .40

Solution: This is a binomial probability calculation.

Event of interest is “birthday on a Monday”

$$\pi = \text{Probability}[\text{event}] = .14$$

N = # trials = 7

X = Number of Monday birthdays in sample of 7

Calculate Pr [X = 1] = .3965

<http://faculty.vassar.edu/lowry/binomialX.html>

- c) What are the odds of getting exactly one pair in five card stud poker using a 52 card deck?

Answer: Odds are 42 to 58

Solution: This is a combinatorial calculation that assumes all five card hands are equally likely.

$$\text{probability}[\text{exactly one pair}] = \frac{\# \text{ hands that are exactly one pair}}{\text{Total \# hands possible}}$$

$$\text{total \# hands possible} = \binom{52}{5} = \frac{52!}{5!47!} = \frac{(52)(51)(50)(49)(48)}{(5)(4)(3)(2)(1)} = 2,598,960$$

To solve for the # hands that are exactly one pair, the idea is to think in steps.

1. # choices of a rank (ace or 2 or 3 or ... or queen or king) = 13
2. For selected rank, # choices of suits to be in the pair = $\binom{4}{2} = \frac{4!}{2!2!} = \frac{(4)(3)}{(2)(1)} = 6$.
3. Now think about the remaining 3 cards. They have to be of distinct ranks, else hand will no longer be exactly one pair. # choices of distinct ranks for the other 3 cards from the leftover 12 ranks = $\binom{12}{3} = \frac{12!}{3!9!} = \frac{(12)(11)(10)}{(3)(2)(1)} = 220$
4. Finally, recognize that for each rank, you also get to choose its suit. #choices for suit of 1st of remaining 3 cards = 4
5. Similarly, # choices for suit of 2nd of remaining 3 cards = 4
6. Last but not least, # choices for suit of 3rd of remaining 3 cards = 4

Putting these all together gives us the count of the # hands that are exactly one pair. Count = (13)(6)(220)(4)(4)(4) = 1,098,240. So now we can solve

$$\text{probability[exactly one pair]} = \frac{\# \text{ hands that are exactly one pair}}{\text{Total \# hands possible}} = \frac{1,098,240}{2,598,960} = 0.42257$$

$$\text{Thus, odds [exactly one pair]} = \frac{0.42257}{1-0.42257} = \frac{0.42257}{0.57743}$$

- d) Suppose a quiz contains 20 true/false questions. You know the correct answer to the first 10 questions. You have no idea of the correct answer to questions 11 through 20 and decide to answer each using the coin toss method. Calculate the probability of obtaining a total quiz score of at least 85%.

Answer: .18

Solution: This is also a binomial probability calculation.

X = Count of correct answers among questions 11-20

N = # trials = 10

A grade of 85% or better corresponds to .85[20] = 17 correct or more. Thus, among questions 11-20, you must be correct 7 or more times.

π = Probability[correct answer] = .50

Calculate Pr [X \geq 7] = .171875

<http://faculty.vassar.edu/lowry/binomialX.html>

3. Refresh your memory of the elements of a confidence interval. There are three:
 (1) point estimate; (2) standard error of the point estimate; and (3) confidence coefficient.
 Complete the following summary table

Normal Distribution: Confidence Interval for [$\mu_1 - \mu_2$] (Two Independent Groups) CI = [point estimate] \pm (conf.coeff)SE[point estimate]			
	σ_X^2 and σ_Y^2 are both known	σ_X^2 and σ_Y^2 are both NOT known but are assumed EQUAL	σ_X^2 and σ_Y^2 are both NOT known and NOT Equal
Estimate	$\bar{X}_{Group 1} - \bar{X}_{Group 2}$	$\bar{X}_{Group 1} - \bar{X}_{Group 2}$	$\bar{X}_{Group 1} - \bar{X}_{Group 2}$
SE to use	$SE[\bar{X}_{Group 1} - \bar{X}_{Group 1}] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$S\hat{E}[\bar{X}_{Group 1} - \bar{X}_{Group 1}] = \sqrt{\frac{S_{pool}^2}{n_1} + \frac{S_{pool}^2}{n_2}}$ where you already have obtained: $S_{pool}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$	$S\hat{E}[\bar{X}_{Group 1} - \bar{X}_{Group 1}] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
Confidence Coefficient Use Percentiles from	Normal	Student's t	Student's t
Degrees freedom	Not applicable	(n₁ - 1) + (n₂ - 1)	$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{\left[\frac{S_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2}\right]^2}{n_2 - 1}\right)}$

4. See if you can recall some of the important concepts that are discussed in introductory biostatistics.
- (a) What is a sampling distribution?

Answer:

A sampling distribution is a probability distribution for a random variable that is itself a statistic. It is the distribution of all possible values of that statistic that is obtained as a result of generating all possible simple random samples of the same size from the same population.

- (b) What does the central limit theorem tell us? Why is it so useful to us? What is a z-score? What is a t-score?

Answer:

The central limit theorem tells us that as the sample size increases, the sampling distribution of the sample mean approaches normality.

This is very useful to us because it allows us, even for moderate sample size, to assume that the distribution of the sample average is normal

$$\text{z-score} = \frac{\text{random variable} - E[\text{random variable}]}{\text{SE}[\text{random variable}]} \text{ is assumed Normal}(0,1)$$

$$\text{t-score} = \frac{\text{random variable} - E[\text{random variable}]}{\text{Estimated SE}[\text{random variable}]} \text{ is assumed Student's t}$$

- (c) In a sentence or two, explain the meaning of a 95% confidence interval for a population mean that has lower limit 35.6 and upper limit 52.8

Answer:

This interval allows the investigator to say that he/she is 95% confident that the unknown true population mean is between 35.6 and 52.8. Of course, the true mean is either in this interval or not; we don't actually know. Recall that it is NOT CORRECT to say that the probability is .95 that the population mean is between 35.6 and 52.8.

- (d) Define p-value. Interpret $p < .05$ and $p < .01$. Given identical study conditions, which gives stronger evidence against the null hypothesis?

Answer:

A p-value is a chance statement. A p-value is the probability that the test statistic of interest attains a value as extreme, or more extreme (relative to the null hypothesis), under the null hypothesis probability model.

All other things being equal, a $p < .01$ gives stronger evidence against the null hypothesis.

- (e) Suppose a two sided hypothesis test of treatment benefit in a randomized controlled trial of placebo versus active treatment yields a p-value of 0.045. What are the possible explanations for this result?

Answer:

There are several possible explanations and, ultimately, we do not know which one is the correct explanation. Among them are

- The active treatment is truly different than the placebo.**
- The active treatment is equivalent to the placebo but an event of low probability has occurred.**
- The active treatment is equivalent to the placebo but one or more biases have biased the data in the direction of statistical significance (albeit marginal)**