Biostatistics [SBE304] (Fall 2019)
Tutorial 4

Sampling Distributions and Point Estimations

Prof. Ayman M. Eldeib   Asem Alaa

Tuesday 22nd October, 2019

1 Tutorial facts

The problems in this tutorial comprises:

(A) Programming Works:

- **>_ 0** programming in-class demos.
- **4** programming homework.
- **0** self-practicing programming works.

(B) Problem Set:

- **8** problems to be solved in-class.
- **8** problems homework.
- **1** self-practicing problems.

Join this GitHub assignment page to create a repository for your submissions: https://classroom.github.com/a/I3Z5Z8D2

2 Sampling Distributions and Point Estimations

2.1 Pre-class reading

1. Lecture notes of "Random Sampling and Sampling Distributions" by Prof. Ayman M. Eldeib

2. From Chapter 7 of Montgomery's textbook, read (pp. 148-159)

2.2 Chapter overview

2.2.1 Stats & Their Distributions

2.2.2 Fxns of Observed Sample

Obs Sample Mean	$\bar{x} = \frac{1}{n} \sum x_i$
Obs Sample Var	$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
Obs Sample Max	$x_{(n)} = max(x_i)$

A statistic is a random variable and the most common are listed above.

Simple Random Samples: The random variables $X_1, \ldots, X_n$ are said to form a simple random sample of size n if each X_i is an independent random variable, every X_i has the same probability distribution.

Sampling Distribs: Every statistic has a probability distribution (a pmt or pdf) which we call its sampling distribution. To determine its distrib can be hard but we use simulations and the CLT to do so.

Simulation Experiments: we must specify the statistic of interest, the population distribution, the sample size(n) and the number of samples (k). Use a computer to simulate each different simple random sample, construct a histogram which will give approx sampling distribution of the statistic.

2.3 The Dist of Sample Mean

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then $E(\bar{X}) = \mu_{\bar{X}} = \mu$ and $V(\bar{X}) = \sigma_{\bar{X}}^2 = \sigma^2/n$. Also if $S_n = X_1 + \ldots + X_n$ then $E(S_n) = n\mu$ and $V(S_n) = n\sigma^2$.

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from a normal distribution with mean μ and variance σ^2. Then for any n, $\bar{X}$ is normal distributed with mean μ and variance σ^2/n. Also S_n is normal distributed with mean $n\mu$ and variance $n\sigma^2$.

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from Bernoulli(p), then $S_n \sim$ Binomial(n,p).

2.3.1 Distribution of The Sample Mean $\bar{X}$

Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then $E(\bar{X}) = \mu_{\bar{X}} = \mu$ and $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$

The standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is often called the **standard error of the mean**.

For a NORMAL random sample with the same mean and std as above, then for any n, $\bar{X}$ is normally distributed with the same mean and std.

Central Limit Theorem: Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\bar{X}$ has approximately a normal dis with mean μ and variance σ^2/n. Also S_n is normal distributed with mean $n\mu$ and variance $n\sigma^2$. No matter which population we sample from, the probability histogram of the sample mean follow closely a normal curve when n is sufficiently large. **Rule of thumb: if $n \geq 30$ CLT can be used.**

2.3.2 Approximate Sampling Distribution of a Difference in Sample Means

If we have two independent populations with means μ_1 and μ_2 and variances σ_1 and σ_2 and if $\bar{X}_1$ and $\bar{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is approximately standard normal if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of Z is exactly standard normal.

2.3.3 Estimators

Parameter of Interest (θ) true, yet unknown, population parameter
Point Estimate: $(\hat{\theta})$ Our guess for θ based on sample data
Point Estimator: $(\hat{\theta})$ statistic selected to get a sensible point estimate

A sensible way to quantity the idea of $\hat{\theta}$ being close to θ is to consider the least squared error $(\hat{\theta} - \theta)^2$. A good measure of the accuracy is the expected or mean square error MSE = $E[(\hat{\theta} - \theta)^2]$. It is often not possible to find the estimator with the smallest MSE so we often restrict our attention to **unbiased** estimators and find the best estimator of this group.
Unbiased: Point Estimate $\hat{\theta}$ if $E(\hat{\theta}) = \theta$ for all θ.
Then $\hat{\theta}$ has a prob distribution that is always "centered" at the true θ value.

2.4 Problem Set

1. Problem

A synthetic fiber used in manufacturing carpet has tensile strength that is normally distributed with mean 75.5 psi and standard deviation 3.5 psi. Find the probability that a random sample of n = 6 fiber specimens will have sample mean tensile strength that exceeds 75.75 psi.

1. Solution

2. Problem

A normal population has mean 100 and variance 25. How large must the random sample be if you want the standard error of the sample average to be 1.5?

2. Solution

3. Problem

The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean 8.2 minutes and standard deviation 1.5 minutes. Suppose that a random sample of n = 49 customers is observed. Find the probability that the average time waiting in line for these customers is:

(a) Less than 10 minutes

(b) Between 5 and 10 minutes

(c) Less than 6 minutes

3. Solution

4. Problem

A random sample of size $n_1 = 16$ is selected from a normal population with a mean of 75 and a standard deviation of 8. A second random sample of size $n_2 = 9$ is taken from another normal population with mean 70 and standard deviation 12. Let $\bar{X}_1$ and $\bar{X}_2$ be the two sample means. Find:

(a) The probability that $\bar{X}_1 - \bar{X}_2$ exceeds 4

(b) The probability that $3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5$ probability?

4. Solution

5. Problem

Scientists at the Hopkins Memorial Forest in western Massachusetts have been collecting meteorological and environmental data in the forest data for more than 100 years. In the past few years, sulfate content in water samples from Birch Brook has averaged 7.48 mg/L with a standard deviation of 1.60 mg/L.

(a) What is the standard error of the sulfate in a collection of 10 water samples?

(b) If 10 students measure the sulfate in their samples, what is the probability that their average sulfate will be between 6.49 and 8.47 mg/L?

(c) What do you need to assume for the probability calculated in (b) to be accurate?

5. Solution

6. Problem

Data on the oxide thickness of semiconductor wafers are as follows: 425, 431, 416, 419, 421, 436, 418, 410, 431, 433, 423, 426, 410, 435, 436, 428, 411, 426, 409, 437, 422, 428, 413, 416.

(a) Calculate a point estimate of the mean oxide thickness for all wafers in the population.

(b) Calculate a point estimate of the standard deviation of oxide thickness for all wafers in the population.

(c) Calculate the standard error of the point estimate from part (a).

(d) Calculate a point estimate of the median oxide thickness for all wafers in the population.

(e) Calculate a point estimate of the proportion of wafers in the population that have oxide thickness of more than 430 angstroms.

6. Solution

7. Problem

A random sample of 36 observations has been drawn from a normal distribution with mean 50 and standard deviation 12. Find the probability that the sample mean is in the interval $47 \leq X \leq 53$. Is the assumption of normality important? Why?

7. Solution

8. Problem

A procurement specialist has purchased 25 resistors from vendor 1 and 30 resistors from vendor 2. Let $X_{1,1}, X_{1,2}, ..., X_{1,25}$ represent the vendor 1 observed resistances, which are assumed to be normally and independently distributed with mean 100 ohms and standard deviation 1.5 ohms. Similarly, let $X_{2,1}, X_{2,2}, ..., X_{2,30}$ represent the vendor 2 observed resistances, which are assumed to be normally and independently distributed with mean 105 ohms and standard deviation of 2.0 ohms. What is the sampling distribution of $\bar{X}_1 - \bar{X}_2$? What is the standard error of $\bar{X}_1 - \bar{X}_2$?

8. Solution

9. Problem

Forest canopy closure. The determination of the percent canopy closure of a forest is essential for wildlife habitat assessment, watershed runoff estimation, erosion control, and other forest management activities. One way in which geoscientists estimate percent forest canopy closure is through the use of a satellite sensor called the Landsat Thematic Mapper. A study of the percent canopy closure in the San Juan National Forest (Colorado) was conducted by examining Thematic Mapper Simulator (TMS) data collected by aircraft at various forest sites (IEEE Transactions on Geoscience and Remote Sensing, Jan. 1986). The mean and standard deviation of the readings obtained from TMS Channel 5 were found to be 121.74 and 27.52, respectively.

a. Let $\bar{Y}$ be the mean TMS reading for a sample of 32 forest sites. Assuming the figures given are population values, describe the sampling distribution of $\bar{Y}$.

b. Use the sampling distribution of part a to find the probability that $\bar{Y}$ falls between 118 and 130.

9. SOLUTION

---

10. PROBLEM

Like hurricanes and earthquakes, geomagnetic storms are natural hazards with possible severe impact on the earth. Severe storms can cause communication and utility breakdowns, leading to possible blackouts. The National Oceanic and Atmospheric Administration beams electron and proton flux data in various energy ranges to various stations on the earth to help forecast possible disturbances. The following are 25 readings of proton flux in the 47-68 kEV range (units are in p/(cm2-sec-ster-MeV)) on the evening of December 28, 2011:

2310 2320 2010 10800 2190 3360 5640 2540 3360 11800 2010 3430 10600 7370 2160 3200 2020 2850 3500 10200 8550 9500 2260 7730 2250

a. Find a point estimate of the mean proton flux in this time period.

b. Find a point estimate of the standard deviation of the proton flux in this time period.

c. Find an estimate of the standard error of the estimate in part (a).

d. Find a point estimate for the median proton flux in this time period.

e. Find a point estimate for the proportion of readings that are less than 5000 p/(cm2-sec-ster-MeV).

10. SOLUTION

---

11. PROBLEM

Suppose that we have a random sample $X_1, X_2, ..., X_n$ from a population that is $N(\mu, \sigma^2)$ We plan to use $\hat{\Theta} = \sum_{i=1}^{n} (X_i - \bar{X})^2/c$ to estimate σ^2. Compute bias in $\hat{\Theta}$ as an estimator of σ^2 as a function of the constant c.

11. SOLUTION

12. PROBLEM

Data on pull-off force (pounds) for connectors used in an automobile engine application are as follows: 79.3, 75.1, 78.2, 74.1, 73.9, 75.0, 77.6, 77.3, 73.8, 74.6, 75.5, 74.0, 74.7, 75.9, 72.9, 73.8, 74.2, 78.1, 75.4, 76.3, 75.3, 76.2, 74.9, 78.0, 75.1, 76.8.

   a. Calculate a point estimate of the mean pull-off force of all connectors in the population. State which estimator you used and why.

   b. Calculate a point estimate of the pull-off force value that separates the weakest 50% of the connectors in the population from the strongest 50%.

   c. Calculate point estimates of the population variance and the population standard deviation.

   d. Calculate the standard error of the point estimate found in part (a). Interpret the standard error.

   e. Calculate a point estimate of the proportion of all connectors in the population whose pull-off force is less than 73 pounds.

12. SOLUTION

13. PROBLEM

Suppose that X is the number of observed "successes" in a sample of n observations where p is the probability of success on each observation.

   1. Show that $\hat{P} = \frac{X}{n}$ is an unbiased estimator of p.

   2. Show that the standard error of $\hat{P}$ is $\sqrt{\frac{p(1-p)}{n}}$. How would you "estimate" the standard error?

13. SOLUTION

14. PROBLEM

The accompanying data on IQ for third-graders at a university lab school was:

   • for males: 117 103 121 112 120 132 113 117 132 149 125 131 136 107 108 113 136 114

   • for females: 114 114 102 109 113 102 131 114 124 127 117 127 120 103 90

Prior to obtaining data, denote the male values by $X_1, ..., X_m$ and the female values by $Y_1, ..., Y_n$. Suppose that the X_i's constitute a random sample from a distribution with mean μ_1 and standard deviation σ_1 and that the Y_i's form a random sample (independent of the X_i's) from another distribution with mean μ_2 and standard deviation σ_2.

   a. Use rules of expected value to show that $X - Y$ is an unbiased estimator of $\mu_1 - \mu_2$. Calculate the estimate for the given data.

   b. Use rules of variance from Lecture notes to obtain an expression for the variance and standard deviation (standard error) of the estimator in part (a), and then compute the estimated standard error.

   c. Calculate a point estimate of the ratio σ_1/σ_2 of the two standard deviations.

d. Suppose one male third-grader and one female third-grader are randomly selected. Calculate a point estimate of the variance of the difference $X - Y$ between male and female IQ.

14. Solution

🏠 ## 15. Problem

Each of 150 newly manufactured items is examined and the number of scratches per item is recorded (the items are supposed to be free of scratches), yielding the following data:

No. of scratches per item	0	1	2	3	4	5	6	7
Observed Frequency	18	37	42	30	13	7	2	1

Let $X =$ the number of scratches on a randomly chosen item, and assume that X has a Poisson distribution with parameter λ.

1. Find an unbiased estimator of λ and compute the estimate for the data [Hint: $E(X) = \lambda$ for X Poisson, so what could be used as an estimator for $E(x)$?].

2. What is the standard deviation (standard error) of your estimator? Compute the estimated standard error. [Hint: $\sigma_X^2 = \lambda$ for X Poisson.]

15. Solution

🏠 ## 16. Problem

Using a long rod that has length μ, you are going to lay out a square plot in which the length of each side is μ. Thus the area of the plot will be μ^2. However, you do not know the value of m, so you decide to make n independent measurements $X_1, X_2, ..., X_n$ of the length. Assume that each X_i has mean μ (unbiased measurements) and variance σ^2.

1. Show that X^2 is not an unbiased estimator for μ^2. [Hint: For any rv Y, $E(Y^2) = V(Y) + E(Y)^2$. Apply this with $Y = \bar{X}$.]

2. For what value of k is the estimator $\bar{X}^2 - kS^2$ unbiased for μ^2 [Hint: Compute $E(\bar{X} - kS^2)$.]

16. Solution

---

17. PROBLEM

Consider a random sample $X_1, ..., X_n$ from the pdf

$$f(x; \theta) = 0.5(1 + \theta x) \quad \text{and} \quad -1 \leq x \leq 1$$

where $-1 \leq \theta \leq 1$ (this distribution arises in particle physics). Show that $\hat{\theta} = 3\bar{X}$ is an unbiased estimator of θ. [Hint: First determine $\mu = E(X) = E(\bar{X})$.]

17. SOLUTION

3   ℝ Programming Language

3.1   ℝ Programming Problems

For these exercises, we will be using the following dataset:

```
library(downloader)
url <- "https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extd\
ata/mice_pheno.csv"
filename <- basename(url)
download(url, destfile=filename)
dat <- na.omit( read.csv(filename) )
```

---

1. PROBLEM

If a list of numbers has a distribution that is well approximated by the normal distribution:

(a) what proportion of these numbers are within one standard deviation away from the list's average?

(b) What proportion of these numbers are within two standard deviations away from the list's average?

(c) What proportion of these numbers are within three standard deviations away from the list's average?

---

2. PROBLEM

Define `y` to be the weights of males on the control diet?

(a) What proportion of the mice are within one standard deviation away from the average weight (remember to use `popsd` for the population `sd`)

(b) What proportion of these numbers are within two standard deviations away from the list's average?

(c) What proportion of these numbers are within three standard deviations away from the list's average?

---

3. PROBLEM

Note that the numbers for the normal distribution and our weights are relatively close. Also, notice that we are indirectly comparing quantiles of the normal distribution to quantiles of the mouse weight distribution. We can actually compare all quantiles using a qqplot. Which of the following best describes the qq-plot comparing mouse weights to the normal distribution?

A. The points on the qq-plot fall exactly on the identity line.

B. The average of the mouse weights is not 0 and thus it can't follow a normal distribution.

C. The mouse weights are well approximated by the normal distribution, although the larger values (right tail) are larger than predicted by the normal. This is consistent with the differences seen between question 3 and 6.

D. These are not random variables and thus they can't follow a normal distribution.

Create the above qq-plot for the four populations: male/females on each of the two diets. What is the most likely explanation for the mouse weights being well approximated? What is the best explanation for all these being well approximated by the normal distribution?

A. The CLT tells us that sample averages are approximately normal.

B. This just happens to be how nature behaves. Perhaps the result of many biological factors averaging out.

C. Everything measured in nature follows a normal distribution.

D. Measurement error is normally distributed.

---

4. PROBLEM

Define `y` to be the weights of males on the control diet?

a. What proportion of the mice are within one standard deviation away from the average weight (remember to use `popsd` for the population `sd`)

b. What proportion of these numbers are within two standard deviations away from the list's average?

c. What proportion of these numbers are within three standard deviations away from the list's average?